Information Theoretic Regularization in Diffuse Optical Tomography

Christos Panagiotou

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

of the

University College London

Department of

Medical Physics & Bioengineering

2011

Abstract

Diffuse optical tomography (DOT) retrieves the spatially distributed optical characteristics of a medium from external measurements. Recovering these parameters of interest involves solving a non-linear and severely ill-posed inverse problem. In this thesis we propose methods towards the regularization of DOT via the introduction of spatially unregistered, *a priori* information from alternative high resolution anatomical modalities, using the information theory concepts of joint entropy (JE) and mutual information (MI). Such functionals evaluate the similarity between the reconstructed optical image and the prior image, while bypassing the multi-modality barrier manifested as the incommensurate relation between the gray value representations of corresponding anatomical features in the modalities involved. By introducing structural *a priori* information in the image reconstruction process, we aim to improve the spatial resolution and quantitative accuracy of the solution.

A further condition for the accurate incorporation of *a priori* information is the establishment of correct alignment between the prior image and the probed anatomy in a common coordinate system. However, limited information regarding the probed anatomy is known prior to the reconstruction process. In this work we explore the potentiality of spatially registering the prior image simultaneously with the solution of the reconstruction process.

We provide a thorough explanation of the theory from an imaging perspective, accompanied by preliminary results obtained by numerical simulations as well as experimental data. In addition we compare the performance of MI and JE. Finally, we propose a method for fast joint entropy evaluation and optimization, which we later employ for the information theoretic regularization of DOT. The main areas involved in this thesis are: inverse problems, image reconstruction & regularization, diffuse optical tomography and medical image registration.

Statement of intellectual contribution

I, Christos Panagiotou, declare that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been properly indicated in the thesis.

Publications

- Arridge, S. R., Panagiotou, C., Schweiger, M., and Kolehmainen, V. (2008b). Multimodal diffuse optical tomography: Theory. In *Translational multimodality optical imaging*, chapter 5, pages 101–123. Artech Press

- Panagiotou, C., Somayajula, S., Gibson, A. P., Schweiger, M., Leahy, R. M., and Arridge, S. R. (2009b). Information theoretic regularization in diffuse optical tomography. *J. Opt. Soc. Am. A*, 26(5):1277–1290

- Panagiotou, C., Somayajula, S., Gibson, A. P., Schweiger, M., Leahy, R. M., and Arridge, S. R. (2009a). Diffusion optical tomography using entropic priors. In *Proceedings of the Sixth IEEE International Symposium on Biomedical Imagting (ISBI'09)*, pages 165–168

- Somayajula, S., Panagiotou, C., Rangarajan, A., Quanzheng, L., Arridge, S. R., and Leahy, R. M. (2010). PET image reconstruction using information theoretic anatomical priors. *IEEE Trans. Med. Im.*, 30(3):537–49

- Pedemonte, S., Cardoso, M. J., Bousse, A., Panagiotou, C., Kazantsev, D., , Arridge, S., Hutton, B. F., and Ourselin, S. (2010a). Class conditional entropic prior for MRI enhanced SPECT reconstruction. In *proceedings of IEEE Nuclear Science Symposium and Medical Imaging Conference*, volume M18-294

- Kazantsev, D., Pedemonte, S., Bousse, A., Panagiotou, C., Arridge, S., Hutton, B. F., and Ourselin, S. (2010). PET bayesian reconstruction using automatic bandwidth selection for joint entropy optimization. In *proceedings of IEEE Nuclear Science Symposium and Medical Imaging Conference*, volume M18-299

Acknowledgements

My ineffable gratitude to my two supervisors Dr. Adam Gibson and Prof. Simon Arridge. Prof. Arridge independently conceived the novel idea of information theoretic regularization with un-registered priors, at a time when no related articles had appeared in the relevant literature. He generously provided me with the gift of an idea to work with, which I could not possibly claim to be my own. During the entirety of this work, he amply provided me with his subject expertise in the most pedagogical manner and without his ongoing motivation through out difficult periods, I would not be now in the position to write this paragraph. I consider him as a friend, my teacher and an inspiration on how meticulous research should be conducted. Although an understatement, thank you so much indeed. I am most indebted to Dr. Gibson for his principal supervision, for his trust in providing me with the much needed funding, as well as for embracing my work with the most genuine interest and for providing me with invaluable suggestions and constant motivation to reach completion. You are an exemplary teacher, mentor, a gifted researcher, a friend and I sincerely thank you for your support and guidance.

Although my supervision was formally limited to the aforementioned, in reality the support by Dr. Martin Schweiger effortlessly qualifies for a supervision on its own. I do not need to state Martin's scientific credentials as his reputation in the field precedes him. I am most grateful for his explicit technical suggestions on the subject but especially for his constant encouragement and moral support. Martin's ability to transmit his knowledge with clarity is truly unmatched and he has gently spent time to respond to the most outrageous queries from my part, without ever showing a glimpse of desperation. I am humbled by the level of your contribution and your support. Thank you Martin.

I want to thank my co-authors Dr. Somayajula and Prof. Leahy of University of Southern California for our collaboration, their hospitality during my visit and the pleasure of working with them. I would also want to thank my co-authors Dr. Kazantsev and Mr. Pedemonte for our recent collaboration. Their input on the subject of information theoretic regularization has been invaluable.

I would like to explicitly thank few members from the Centre of Medical Imaging Computing in the Computing Science Dept. of UCL. I consider myself privileged to have worked with exemplary researchers such as Drs. Abdel Douiri, Juha Heiskala, Jason Kastanis, Vadim Soloviev, Athanasios Zacharopoulos, Vasileios Zografos, Eleutheria Panagiotaki and Mr. Tim Rudge.

I have only positive experiences and good memories from my interaction with the people working in the Biomedical Optics Research Laboratory in the Medical Physics & Bioengineering Dept. of UCL. I would like to explicitly thank Prof. Jem Hebden, Dr. Ben Cox and Dr. Murad Banaji for their explicit suggestions and for all the interesting discussions. I would like to (alphabetically) thank true friends who I have found in Drs. Thomas Allen, Chris Colley, Teresa Correia, Louise Enfield, Nick Everdell, Julian Henty, Tryphon Lambrou, Jan Laufer, Terence Leung, Salavat Magazov, Caroline Reid, Ilias Tachtsidis, Bradley Treeby, Anne Vanhoestenberghe, Edward Zhang, Robert Cooper, Ben Price, Ifung Lu as well as Mr. Paul Burke, and Ms. Maria Papademetriou. They have effortlessly created simply by their presence an inspirational as well as pleasant environment to work within, bursting with genuine interest in science. Thank you very much indeed for all the pleasant time we have spent together.

I would like to thank Dr. Sebastien Ourselin for his useful suggestions regarding image registration - specifically regarding the solution reset in the proposed simultaneous reconstruction/registration scheme, as well as Prof. Daniel Alexander for all the useful discussions.

I want to thank Drs. Ville Kolehmainen of University of Kuopio and Petri Hiltunen of Aalto University of Technology for their suggestions and explicit contributions in optical tomography reconstruction and regularization. In addition I want to thank Dr. Ilkka Nissila and Mr. Atte Lajunen of Aalto University for providing the experimental data which was used to test the proposed methods.

I cannot express enough gratitude to all my friends for so generously providing me with constant moral support, as well as for making my life beautiful so often. Although it is is unfair to explicitly single out people, I feel obliged to (alphabetically) state few names from the UK: K. Chatziarapis, A. Kanakis, G. Katrakalidis, Dr. P. Konstantopoulos, A. & A. Lambridis, M. Manos, K. Papadopoulos, S. Pasxalidis, K. Tornaros, Dr. T. Tsagaris also from Greece: M. Charalambi, A. Kolimenakis, V. Kovanis, Dr. P. Pandis, K. Pappa, S. Paraskeuaidis, V. Pikoulas, G. Retsoulis and J. Varelas who have patiently bore with me during my effort.

This work would have never been realized without the financial support by the Engineering and Physical Sciences Research Council (EPSRC) EP/D502330/1.

Finally, I am ineffably indebted to my parents Magda & Yiannis and my sister Rallou for their ever-present love and care; for raising me in an inspirational and thought-provoking environment with high values and scientific inclination; for financially supporting my entire education; as well as for the patience and genuine interest which they have repeatedly displayed during this work, especially in times when I have been re-iterating my struggle and explaining over and over again the specifics of Optical Tomography and Information Theory. The following is a dedication to you.

Contents

Abstract **3**

Statement of intellectual contribution **4**

Publications **5**

Contents **8**

List of Figures **15**

Acronyms **21**

I Introduction **24**

1 Prologue **26**

1.1 Introduction 26

1.2 Problem statement, aims and contributions 28

1.3 Structure of the thesis 32

II Theory **34**

2 Inverse Problems **36**

2.1 Introduction 36

2.2 Parameter and data space 38

2.3 Ill-posedness 38

2.4 Linear case 39

2.4.1 Singular Value Decomposition 39

2.4.2 Linear least squares 41

2.4.2.1 Mostly one solution 41

2.4.2.2 Infinite solutions 43

2.4.2.3 Numerically rank deficient system 44

2.4.2.4 Discrete ill-posed problems with ill-determined rank 45

2.4.3 Generalized Tikhonov regularization 46
2.5 Non-linear case 48
2.6 optimization 48
2.6.1 Non-gradient based optimization 48
2.6.2 Gradient based optimization: First order methods 49
2.6.2.1 Gradient descent 49
2.6.2.2 Conjugate gradients 50
2.6.3 Gradient based optimization: Second order methods 52
2.6.4 Line search 54
2.7 Summary 55

3 Diffuse optical tomography 57

3.1 Introduction 57
3.2 Fundamental optical quantities 57
3.2.1 Radiance 57
3.2.2 Absorption 58
3.2.3 Scattering 59
3.2.4 Index of refraction 60
3.3 Imaging schemes 60
3.4 Forward problem in DOT 61
3.4.1 Radiative Transfer Equation 62
3.4.2 Diffusion approximation to the Radiative Transfer Equation 64
3.4.2.1 Validity of the Diffusion Approximation 66
3.4.2.2 Boundary conditions 67
3.4.2.3 Source conditions 67
3.4.2.4 Measurable quantities 68
3.4.3 Finite element method 69
3.4.3.1 Forward model and FEM 72
3.4.4 Alternative approaches 73
3.4.4.1 Analytical methods 73
3.4.4.2 Monte Carlo methods 73
3.5 Inverse problem 74
3.5.1 Linear case 75
3.5.2 Non-linear case 75
3.5.3 Jacobian computation 76
3.5.4 Solution scaling and positivity 76
3.5.4.1 Parameter normalization 77
3.5.4.2 Positivity of solution 78
3.5.4.3 Data scaling 79

3.6 Regularization and multi-modality imaging 79
3.6.1 Quadratic penalties (Gaussian priors) 79
3.6.2 Non-quadratic penalties 80
3.6.3 Edge preserving regularization 80
3.6.4 Hard constraints and other types of prior information 82
3.7 Reference images: anatomical and functional correspondence 84
3.8 Ill-posedness, noise and resolution 85
3.8.1 Ill-posedness 85
3.8.2 Noise 86
3.8.3 Resolution 87
3.9 Summary 88

4 Information theory 89

4.1 Introduction 89
4.2 Random processes, probabilities and random variables 90
4.2.1 Random process 90
4.2.2 Probabilities 91
4.2.3 Random variables 91
4.2.4 Probability distribution and density functions 92
4.2.5 Images and randomness 93
4.2.6 Expectation, variance and standard deviation of a random variable and sample statistics 94
4.2.7 Systems of two random variables 95
4.3 The Normal density 96
4.4 Density estimation 97
4.4.1 Parametric techniques 98
4.4.1.1 Maximum likelihood estimator 98
4.4.1.2 Bayesian estimation 99
4.4.2 Non parametric kernel density estimation 100
4.5 Information theoretic functionals 102
4.5.1 Entropy 102
4.5.2 Joint entropy, conditional entropy and mutual information 104
4.5.3 Differential entropy 105
4.5.4 Empirical entropy 106
4.5.5 Images, information theory and multi-modality data 107
4.5.5.1 Minimum/maximum entropy images 107
4.5.5.2 Information theoretic functionals and multi-modality 108
4.5.5.3 Discrete vs differential entropy in the imaging context 111
4.6 Summary 111

5 Medical image registration 113
5.1 Introduction . . . 113
5.2 Applications in medical imaging . . . 113
5.3 Definitions and notation . . . 115
5.4 Classifications of registration problems . . . 116
5.5 Spatial Transformations . . . 118
5.5.1 Linear transformations . . . 118
5.5.1.1 Rigid transformations . . . 119
5.5.1.2 Non-rigid transformations . . . 119
5.5.1.3 Composition of multiple transformations . . . 120
5.5.1.4 Origin of transformation . . . 120
5.5.2 Non-linear transformations . . . 122
5.5.2.1 Demon's registration . . . 122
5.5.2.2 Thin plate splines . . . 123
5.5.2.3 B-Spline Free form deformations . . . 123
5.5.3 Interpolation . . . 125
5.6 Image similarity evaluation . . . 126
5.6.1 Sum of squared differences . . . 126
5.6.2 Correlation techniques . . . 128
5.6.3 Ratio-Image Uniformity and Partitioned Intensity Uniformity . . . 129
5.6.4 Information theoretic similarity measures in image registration . . . 129
5.6.4.1 Joint entropy vs Mutual information . . . 129
5.6.4.2 Normalized mutual information . . . 131
5.7 Optimization . . . 133
5.7.1 Multi-resolution pyramids . . . 133
5.7.2 Interpolation artefacts and blurring . . . 134
5.8 Conclusion . . . 134

III Methods and Results 136

6 Efficient entropy and derivative computation 138
6.1 Computationally efficient marginal entropy evaluation . . . 139
6.1.1 Efficient marginal PDF estimation . . . 140
6.1.1.1 Binning range and spacing . . . 144
6.1.2 Fast marginal entropy estimation (Shannon formulation) . . . 145
6.1.3 Fast marginal entropy estimation (empirical formulation) . . . 146
6.1.4 Comparison of the entropic estimators . . . 146
6.2 Computationally efficient marginal entropy derivative estimation . . . 150
6.2.1 Shannon's entropy derivatives . . . 151

6.2.2 Evaluation of the efficient Shannon entropy estimator derivatives 152
6.2.3 Empirical entropy derivatives 154
6.2.4 Evaluation of the efficient empirical entropy estimator derivatives 157
6.2.5 Comparison of the derivatives of the marginal entropy estimators 158
6.3 Computationally efficient joint entropy evaluation 162
6.3.1 Efficient joint PDF estimation 163
6.3.2 Fast joint entropy estimation (Shannon formulation) 165
6.3.3 Validation and comparison of the joint entropic estimators 166
6.4 Efficient joint entropy derivative computation 167
6.4.1 Shannon's joint entropy derivatives 167
6.4.2 Empirical joint entropy derivatives 169
6.4.3 Comparison of the derivatives of the joint entropy estimators 170
6.4.4 Run time tests 171
6.5 Efficient mutual information evaluation via marginalization of the joint probability density function 171
6.6 Summary 172

7 Information theoretic regularization in diffuse optical tomography 175

7.1 Introduction 175
7.2 Formulation of the inverse problem 177
7.2.1 Objective function minimization scheme 178
7.3 Comparison between JE and MI for regularization purposes: Theoretical intuition 179
7.3.1 JE vs MI: Image pair {1,1} 180
7.3.2 JE vs MI: Image pair {1,2} 181
7.3.3 JE vs MI: Image pair {1,3} 182
7.3.4 JE vs MI: Image pair {1,4} 183
7.3.5 JE vs MI: Image pair {1,5} 184
7.4 KDE binning range/ kernel standard deviation 184
7.4.1 Binning range 184
7.4.2 Kernel standard deviation 185
7.5 Case study 1: 2D numerical simulation 185
7.5.1 Simulation description 185
7.5.2 Comparison between Shannon and empirical entropy implementation 186
7.6 Case study 2: 2D MRI-derived target, MRI prior 196
7.6.1 Simulation description 196
7.6.2 Meshes and reconstruction initialization 196
7.6.3 Information theoretic regularization 197
7.7 Case study 3: 3D reconstruction of a software phantom 201
7.7.1 Simulation description 201

7.7.2 Reconstruction and initialization 201
7.7.3 Information theoretic regularization: correct prior 201
7.7.4 Information theoretic regularization: incorrect prior 201
7.8 Case study 4: DOT/MRI experimental phantom multi-modality study 207
7.8.1 Phantom description 207
7.8.2 Data acquisition process 208
7.8.3 Reconstruction setting and initialization 208
7.8.4 Information theoretic regularization: original prior 211
7.8.4.1 Joint entropy reconstruction 211
7.8.4.2 Mutual information reconstruction 211
7.8.5 Information theoretic regularization: alternative prior 212
7.8.5.1 Joint entropy reconstruction 213
7.8.5.2 Mutual information reconstruction 213
7.9 Summary and discussion 215

8 Information theoretic regularization in diffuse optical tomography with unregistered structural priors 216

8.1 Introduction 216
8.2 Formulation of the inverse problem 219
8.2.1 Regularization functional 220
8.3 Objective function minimization scheme 222
8.3.1 Derivatives with respect to optical solution 223
8.3.2 Derivatives with respect to transformation parameters 224
8.3.3 Simultaneous reconstruction/registration (SRR) 225
8.3.3.1 Initialization 225
8.3.3.2 Iteration in simultaneous reconstruction/registration 225
8.3.3.3 B-Spline grid refinements and solution resets 226
8.3.3.4 Discussion 227
8.4 Testing framework and results 229
8.4.1 Case study 1 230
8.4.1.1 Case 1: Initialization 230
8.4.1.2 Case 1: SRR results 231
8.4.2 Case study 2 237
8.4.3 Case study 3 242
8.5 Comparison among cases 246
8.6 Discussion 247
8.7 Summary 247

IV Conclusion 250

9 Summary and future directions 252

9.1 Potential improvements . . . 254

9.1.1 Kernel density estimation . . . 254

9.1.2 Modeling of intra-image spatial dependency among gray values . . . 254

9.1.3 Optimization . . . 255

9.1.4 Application to other modalities . . . 255

Bibliography 257

List of Figures

1.1 Presentation of selected diffuse optical tomography clinical studies illustrating the type of contrast and resolution found in practice. 27

1.2 Aim 1: Introduction of structural information in DOT from high-resolution images with *incommensurate* gray values 29

1.3 Aim 2: Introduction of structural information in DOT from high-resolution, *unregistered* images with *incommensurate* gray values 30

2.1 Action of a linear forward operator with full column rank and data not being inside the range of the forward operator 42

2.2 Action of a linear rank-deficient forward operator and data not being inside the range of the forward operator 44

2.3 Successive descent directions produced by the gradient descent method 50

2.4 Successive descent directions produced by the conjugate gradient method 51

2.5 Inverse quadratic interpolation computed on successive line search λ estimates 55

3.1 Radiance of light is reduced during its traversal of a non-scattering but homogeneously absorbing medium 58

3.2 Radiance of light is reduced during its traversal of a non-scattering but homogeneously absorbing medium 58

3.3 Radiance of light is reduced during its traversal of a non-absorbing but homogeneously scattering medium 59

3.4 The imaging process in DOT involves non-invasive imaging by a configuration of sources/detectors places on the surface of the probed domain 60

3.5 Graphical illustration of imaging schemes in optical tomography 61

3.6 Light incident on an elementary volume. Part of the incident light can be absorbed and part of the light can be scattered. In the case that the volume encloses a source, emission can also occur 63

3.7 Visualization of the photon density fields (amplitude and phase) for the frequency domain case, for a given optical parameter distribution 66

3.8 (a)-(b) Absorption/scattering distributions (c) source Q and detectors $D1\dots4)$ (d) log-temporal point spread function (TPSF) obtained from each detector for irradiation from Q in the noise free time domain (TD) case. 70

3.9 An unstructured mesh produced by the triangulation of a circular domain Ω. It consists of 1021 nodes and 1920 elements 71

3.10 Linear shape function Ξ_i centered on node $\mathcal{N}_i$ 71

3.11 Photon measurement density functions 77

3.12 The imaging process in DOT involves non-invasive imaging by a configuration of sources/detectors places on the surface of the probed domain 87

4.1 Images and randomness 93

4.2 Examples of probability distribution and density functions 94

4.3 Joint probability distribution and mass/density functions between two images 97

4.4 Examples of normal densities. 4.4a Uni-variate densities $\mathcal{N}(0,0.5), \mathcal{N}(0,1), \mathcal{N}(0,2)$. 4.4b Isotropic bi-variate density with $\mu_x = 0, \mu_y = 0, \sigma_x = 1, \sigma_y = 1$. 4.4c Anisotropic bi-variate density with $\mu_x = 0, \mu_y = 0, \sigma_x = 0.5, \sigma_y = 2$ 98

4.5 Non-parametric kernel density estimation 101

4.6 Entropy and independence to grouping of events 102

4.7 Relationships among information theory concepts expressed in the form of a Venn diagram 105

4.8 Minimum/maximum entropy images 108

4.9 Minimum/maximum joint entropy images 109

4.10 Entropy, multi-modality and spatial dependence 110

4.11 Differential entropy and images 112

5.1 Medical image registration and difference imaging 114

5.2 Effects of the choice of origin of linear transformations 121

5.3 Non-rigid transformation based on B-Spline Free Form Deformations 124

5.4 Local support of B-Spline free form deformations 125

5.5 Control point grid refinement 126

5.6 Forward interpolation 127

5.7 Backward interpolation 127

5.8 Joint entropy vs Mutual Information in image registration 130

5.9 Framework for testing the response of mutual information to the ratio of the pixels in the overlap domain which are labeled as background, over the number of pixels in the same domain depicting the main feature 132

5.10 Response of the information theoretic functionals to the various alignment configurations of the test case described by Fig. 5.9. 133

5.11 Information theoretic, intensity based similarity measures and image resolution levels . . 134

6.1 The continuous gray values $\alpha \in A$ are interpolated to a regularly spaced grid $\breve{x}$. The originally frequency encoded density of A is now reflected by the weights $w(\breve{x})$ fixed over the equispaced $\breve{x}$. . . . 142

6.2 Graphical representation of the entries in array $\mathfrak{b}$, which is used for fast interpolations between $\breve{x}$ and α. Array $\mathfrak{b}$ stores the normalized distance between a continuous α_k and its immediate neighbour $\breve{x}_i$ for which $\breve{x}_j < \alpha_i$. The array $\mathfrak{i}$ stores the index i of the left neighbouring grid point. . . . 142

6.3 Graphical illustration of the binned kernel density estimator . . . 144

6.4 Binning range for PDF sampling . . . 145

6.5 Comparison between Shannon & empirical entropy estimators for optimum kernel width u^{opt}. . . . 148

6.6 Comparison graph among the mean entropy estimates of **x** $\sim \mathcal{N}(0,1)$ for variable kernel u and of the true entropy $h(\mathcal{N}(0,1))$. . . . 149

6.7 Comparison graph among the mean entropy estimates of **x** $\sim \mathcal{N}(0,1)$ for variable grid size $\breve{N}$ . . . 150

6.8 Shannon entropy estimator derivatives . . . 153

6.9 Fast probability density derivative estimation . . . 156

6.10 Empirical entropy estimator derivatives. . . . 157

6.11 Fitting empirical estimator derivatives to the Shannon derivative, as a function of kernel standard deviation u . . . 159

6.12 Visual comparison between Shannon & empirical entropy estimator derivatives of two samples of different size. . . . 161

6.13 Bi-linear interpolation in JPDF estimation. Graphical representation of involved quantities 164

6.14 Example of normally distributed joint sample used in joint entropy evaluation. . . . 166

6.15 Comparison between joint Shannon & empirical entropy estimates of a sample drawn from normally distributed RVs **x**, **y**. . . . 167

6.16 Derivatives of the Shannon and empirical joint entropy estimators. . . . 172

6.17 Computational time for joint entropy and derivative computation . . . 174

7.1 Test images and corresponding PDFs/JPDFs for comparing JE and MI . . . 180

7.2 Demonstration of JE partial structural invariance . . . 184

7.3 Target μ_a and μ_s' images for 2D simulations . . . 186

7.4 Comparison between Shannon & empirical entropy estimators for variable kernel width u^{opt}. . . . 187

7.5 Target distributions and the 4 reference images pairs, incommensurately related to the target gray values. Ref 1 displays full correspondence between its features and the ones in the target distributions. Ref. 2 contains features not existing in the target space. Ref. 3 is missing features. The gradient in Ref. 4 is enforced by centring a 2D Gaussian (σ: 50 pixels) on top of Ref. 1 and multiplying the pixels values underneath. We also add 5% Gaussian multiplicative noise. 190

7.6 μ_a reconstructions by introducing the available reference image pairs with joint entropy or mutual information. The converged TK1 reconstructions are provided for comparison along with the initialization guess. 190

7.7 μ_s' reconstructions by introducing the available reference image pairs with joint entropy and mutual information. 190

7.8 Effect of information theoretic regularization on JPDFs 191

7.9 Profiles for μ_a at $y = 40$ or $y = 100$ for reconstructions using TK1, JE and MI. We also provide the profiles for $x^\star$ and x_{ref}. The choice of y targets the profiling of different features and was made taking into account the feature correspondence between $x^\star$ and x_{ref} space for each case. 192

7.10 Profiles for μ_s' at $y = 40$ or $y = 100$. 193

7.11 Reconstruction errors (against true solution) for the images depicted in Fig. 7.12. 194

7.12 Reconstructions using prior pair #2 for 40 $\log$-spaced τ values. 195

7.13 Case study 2: Initial images 197

7.14 Case study 2: Reconstructions using 40 $\log$-spaced τ values. 198

7.15 Case study 2: Reconstructed images 199

7.16 Cylindrical software phantom with embedded absorption (red) and scattering (blue) perturbations 202

7.17 Converged TV reconstruction: Visualized using optimal display range to enhance features 203

7.18 Converged TV reconstruction: Visualized using the same display range as the true solution 203

7.19 Prior image 1: The prior has one-to-one feature correspondence with the true solution. Gray values are incommensurately related to the true solution. 204

7.20 Converged JE reconstruction using the structurally correct prior 1 204

7.21 Converged MI reconstruction using the structurally correct prior 1 205

7.22 Prior image 2: The prior has extra features when considering the true μ_a and true μ_s' solution. Gray values are incommensurately related to the true solution. The displayed isosurface image is not representative of the gray values as the high background value dominates the isosurface rendering. It is solely displayed as an indicator of the locations of the various features. 205

7.23 Converged JE reconstruction using the structurally incorrect prior 2 206

7.24 Converged MI reconstruction using the structurally incorrect prior 2 206

7.25 Case study 3: Photograph of probed phantom 209

7.26 Case study 3: Structure of MRI/DOT phantom 209

7.27 Case study 3: TV-reconstructed volume at selected views 210

7.28 Case study 3: MRI prior volume at selected views 211

7.29 Case study 3: JE-reconstructed volume at selected views 212

7.30 Case study 3: MI-reconstructed volume at selected views 212

7.31 Case study 3: MRI prior volume at selected views 213

7.32 Case study 3: JE-reconstructed volume at selected views for the case of rotated prior. . . 213

7.33 Case study 3: MI-reconstructed volume at selected views for the case of rotated prior. . . 214

8.1 Depiction of bias in reconstructions induced by the introduction of spatially unregistered priors. 217

8.2 Conditional entropy in translational transformations 222

8.3 Information theoretic functionals against different overlap configurations manifesting from rotational misalignment and for various sizes of field of view: Discrete entropy case 223

8.4 Information theoretic functionals against different overlap configurations manifesting from rotational misalignment and for various sizes of field of view: Continuous entropy case. 223

8.5 Simultaneous reconstruction/registration: Alternating optimization scheme 227

8.6 Data flow diagram in the proposed simultaneous registration/reconstruction scheme . . . 228

8.7 Simultaneous registration/reconstruction test images 229

8.8 Case 1: limited initialization for the simultaneous reconstruction/registration scheme . . 231

8.9 Case 1: results obtained by the simultaneous reconstruction/registration scheme. 232

8.10 Case 1: Simultaneous registration/reconstruction error plots 234

8.11 Case 1: Simultaneous reconstruction/registration output at the various B-Spline FFD grid refinement/reset levels: Optical solution estimates. 235

8.12 Case 1: Simultaneous reconstruction/registration output at the various B-Spline FFD grid refinement/reset levels: Optical solution estimates. 236

8.13 Case 2: limited initialization for the SRR scheme. 237

8.14 Case 2: results obtained by the simultaneous reconstruction/registration scheme. 239

8.15 Case 2: Simultaneous reconstruction/registration output at the various B-Spline FFD grid refinement/reset levels: Optical solution estimates. 240

8.16 Case 2: Simultaneous reconstruction/registration output at the various B-Spline FFD grid refinement/reset levels: Optical solution estimates. 241

8.17 Case 3: results obtained by the simultaneous reconstruction/registration scheme. 242

8.18 Case 3: Simultaneous reconstruction/registration output at the various B-Spline FFD grid refinement/reset levels: Optical solution estimates. 244

8.19 Case 3: Simultaneous reconstruction/registration output at the various B-Spline FFD grid refinement/reset levels: Optical solution estimates. 245
8.20 Regions of interest: Cases 1 & 2 248
8.21 Region of interest: Bias and Variance error 249

Acronyms

TK_0 zeroth-order Tikhonov.

TK_1 first-order Tikhonov.

BFGS Broyden-Fletcher-Goldfarb-Shanno.

c.o.m. centre of mass.

CC cross correlation.

CE conditional entropy.

CG conjugate gradients.

CS collimated source.

CT X-ray computed tomography.

CW continuous wave.

DA diffusion approximation.

DBC Dirichlet boundary condition.

DDOT difference diffuse optical tomography.

DE differential entropy.

DOT diffuse optical tomography.

DS diffuse source.

FEM finite element method.

FFD free form deformation.

FFT fast Fourier transform.

fMRI functional magnetic resonance imaging.

FOV field of view.

i.i.d. independent and identically distributed.

IT information theory.

JE joint entropy.

JPDF joint probability density function.

KDE kernel density estimation.

L-BFGS limited memory BFGS.

LS least squares.

MEM maximum entropy method.

MI mutual information.

MISE mean integrated square error.

ML maximum likelihood.

MRI magnetic resonance imaging.

NCC normalized cross correlation.

NIR near infrared.

NMI normalized mutual information.

PDF probability density function.

PET positron emission tomography.

PMDF photon measurement density function.

RBC Robin boundary condition.

ROI region of interest.

RTE radiative transfer equation.

RV random variable.

SPECT single photon emission computed tomography.

SRR simultaneous reconstruction/registration.

SSD sum of squared differences.

SVD singular value decomposition.

TD time domain.

TPS thin plate spline.

TPSF temporal point spread function.

TSVD truncated singular value decomposition.

TV total variation.

UI ultrasound imaging.

Part I

Introduction

Chapter 1

Prologue

1.1 Introduction

Diffuse optical tomography (DOT) [Arridge, 1999; Boas et al., 2001; Gibson et al., 2005a] is a non-invasive, functional medical imaging modality that utilizes the translucency of tissue to light at near infrared (NIR) wavelengths. The imaging process involves the trans-illumination of the probed anatomy from multiple source sites and the subsequent measurement of the light exiting the anatomy at detector sites. Sources and detectors are exclusively located on the surface of the anatomy. By probing with NIR light, DOT aims to retrieve quantitatively and spatially accurate estimates of the physical properties which govern the propagation of light inside the probed anatomy. In optical imaging, the optical properties of interest are usually expressed in terms of light absorption and scattering or in terms of absorption and diffusion. These quantities have direct physiological relevance. Regarding the former, the main absorber of NIR light in tissue is haemoglobin - the principal oxygen carrier in blood. The retrieval of accurate estimates of the absorption distribution in tissue can be interpreted in terms of medical importance such as blood volume, blood flow and blood oxygen saturation in multi-spectral studies. The properties of light scattering differ among the various tissue types, hence its spatial distribution can be regarded as an indicator of the underlying anatomical structure [Boas et al., 2001]. In addition, scattering changes have been related with functional responses, such as underlying neural activity [Gratton et al., 1997]. It is evident that the physiological information related with the optical properties of tissue can have high medical significance. Potential applications of optical imaging include screening for breast cancer as tumours are optically visible due to the increased vascularization; tumour malignancy classification from the relative levels of oxy- and deoxy- haemoglobin - denoted as $\mathrm{HbO_2}$ and HHb respectively - which can be retrieved by multi-spectral optical imaging [Boas et al., 2001]; neonatal brain imaging for the detection of haemorghages and brain oxygenation levels; muscle haemodynamics and more [Gibson et al., 2005a].

Figure 1.1 showcases two selected DOT clinical imaging studies in order to familiarize the reader with the resolution levels of the modality. The presented cases regard a neonatal brain haemorrhage [Hebden et al., 2002] and a breast imaging study for cancer screening [Yates et al., 2005]. Corresponding intra-subject images obtained from alternative high-resolution modalities - used to validate the DOT results in the aforementioned studies - are also presented for visual comparison.

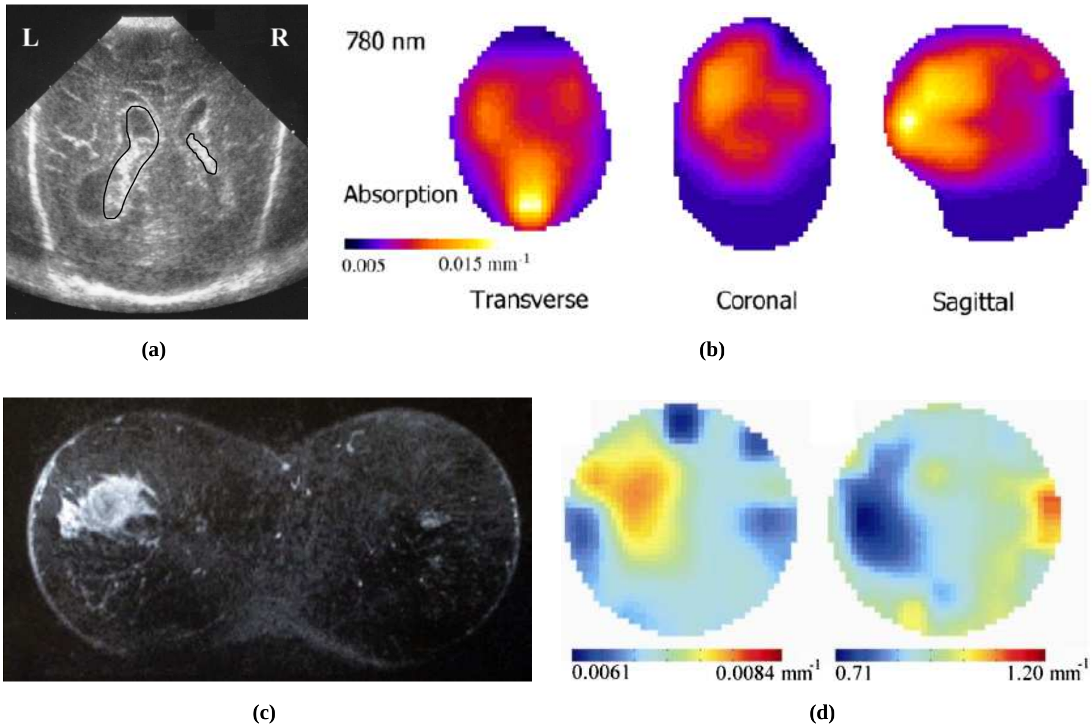


Figure 1.1: Presentation of selected DOT clinical studies to illustrating the type of contrast and resolution found in practice. **Subfig. 1.1a** Ultrasound scan of infant brain revealing a large haemorrhage in the left ventricle. **Subfig. 1.1b** DOT 3D scan of the same infant acquired at 780mm NIR wavelength. The absorption image is solely presented. **Subfig. 1.1c** MRI scan (difference image between post- and pre- contrast agent image application) revealing a tumour in the right breast **Subfig. 1.1d** Corresponding DOT images of the right and left breast respectively. Only the absorption images are presented.

The low resolution of DOT would be immediately apparent to any reader accustomed with alternative high resolution modalities, such as magnetic resonance imaging (MRI)/functional magnetic resonance imaging (fMRI) and X-ray computed tomography (CT) or even compared to the lower resolution positron emission tomography (PET), single photon emission computed tomography (SPECT) and ultrasound imaging (UI). Although the low resolution is a weakness of DOT compared to other modalities, nevertheless there are clear motives justifying the ongoing effort for its improvement as DOT presents a number of advantages compared to its alternatives, in terms of the medical significance of the retrieved information as well as the practicalities of the imaging process. Regarding the former, DOT's sensitivity to both $\mathrm{HbO_2}$ and HHb is unique, with the exception of psychoacoustic imaging [Xu and Wang, 2006] which compared to DOT has superior spatial resolution but very small penetration depth. The modality has very fast temporal resolution $(< 1s)$ (decreasing with the thickness of the probed medium) which enables continuous monitoring of the physiological processes. We already mentioned that in addition to the sensitivity of $\mathrm{HbO_2}$, HHb, blood flow and volume, DOT is also sensitive to neural activation as it affects the scattering of light. In the practical setting, DOT is relatively inexpensive modality $(< £300K)$ compared to the costs of MRI/fMRI and PET $(> £1m)$ or even SPECT and CT $(£300K - £1m)$ [Correia, 2010]. In addition, the instrumentation is highly portable, similar to UI, which enables continuous monitoring of a patient at bedside and does not require the immobilisation of the patient - dictated by MRI/fMRI, PET or SPECT - as the NIR probe is attached directly to the skin of the anatomy. Breast imaging with DOT does not require the compression of the breast, compared to what is required by X-ray mammography. Finally, its operation does not entail any safety risks for the patient as, compared to CT, PET and SPECT, NIR light has no ionising effects. In addition and contrary to fMRI, DOT imaging does not require the application of contrast agents, which their suitability for neonatal studies is still under investigation. The potential benefits of DOT imaging in medicine are too many to be ignored. The ongoing research towards the improvement of modality - in terms of the quantitative and spatial resolution as well as the consistency of the retrieval of accurate results - aims to render DOT a significant technology on its own or in conjunction with established imaging technologies. More detailed comparison between DOT and additional modalities can be found in [Boas et al., 2001; Correia, 2010].

1.2 Problem statement, aims and contributions

Problem statement and aims Diffuse optical tomography recovers the optical properties of the probed anatomy by utilizing the measured light exiting its surface. These properties are represented in the form of images, hence the computation scheme resulting in these images is termed *image reconstruction* and the actual recovered properties as *solution*. It is essential that the solution is spatially and quantitatively accurate, in the sense that it reflects the *true* and unknown solution, that is the true optical properties of the probed anatomy. From a mathematical perspective, the recovery of the solution is formulated as a non-linear inverse problem, as the quantities of interest are not directly accessible for measurement due to the non-invasive nature of DOT. Unfortunately the inverse problem is severely ill-posed, a condition which compromises the consistent retrieval of accurate solutions.

The aim of this work is to improve the accuracy of the obtained reconstructions by utilizing *a priori* information regarding the true solution. This information is incorporated into the inverse problem via a process known as *regularization*, which effectively attempts to alleviate the negative effects of ill-posedness. The nature of prior information considered in this work regards the *structure* of the true solution or in other terms, the *spatial* distribution of the true optical parameters. It should be emphasized that the prior information is strictly limited to structure. In simpler terms, we have prior knowledge on structural features that should appear in the optical reconstruction, for example tumours or distinct anatomical classes, however we have no information regarding the optical quantities which correspond to these features.

Potential sources of such structural *a priori* information are medical images of the same anatomy, obtained from high-resolution imaging modalities. We refer to these images as *reference* or *prior* images. These secondary modalities consistently recover highly accurate spatial and quantitative estimates of the physical quantities which govern their operation - for example the magnetic properties of tissue targeted by MRI. Given that both DOT and the secondary modality probe the same anatomy, they should return - to some extent - *structurally* similar images. However, the actual values which populate corresponding anatomical areas in the optical and secondary image, are *incommensurately* related, due the very different physical nature between of the optical and secondary quantities.

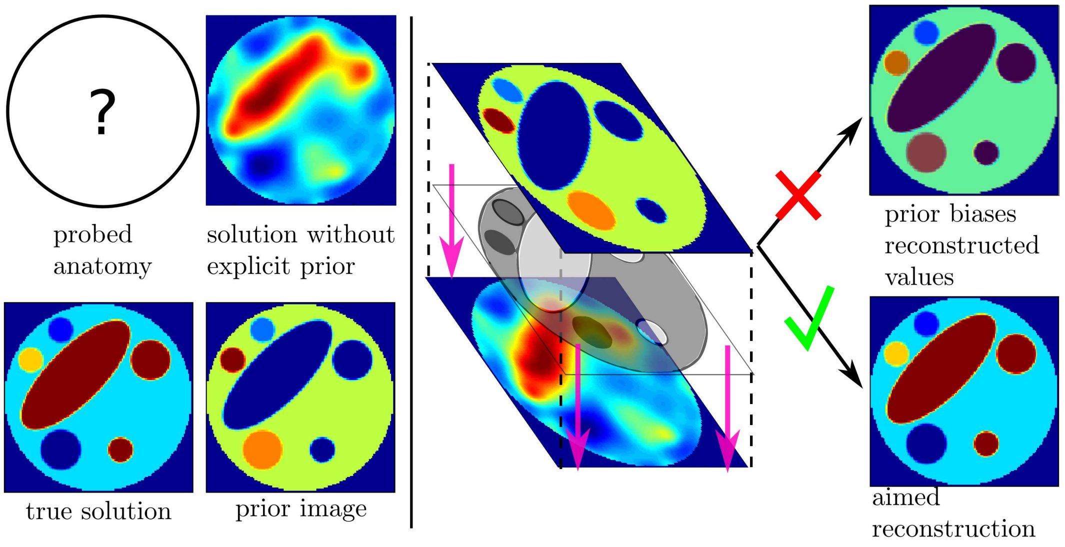

?
probed anatomy
solution without explicit prior
true solution
prior image
prior biases reconstructed values
aimed reconstruction

Figure 1.2: Aim 1: Introduction of structural information in DOT from high-resolution images with *incommensurate* gray values

Figure 1.2 schematically showcases the task at hand. The prior image accurately reflects the structure of the true solution. In effect, the prior provides explicit information regarding the features which should appear in the reconstruction and also to 'where' these features should appear. In accordance to the point noted earlier, the values which populate the true solution and the prior image are incommensurately related. The principal problem which this thesis seeks to solve, is to isolate the structural information in the prior image from the actual values populating it and then utilize it for improving the optical solution. Ideally, the method should be able to retrieve the same solution regardless of the choice of values in the prior image. The values in the prior should have minimal biasing effect on the reconstructed optical values.

One should notice that we emphasized that it is the prior which dictates 'where' features should appear. The prior is superimposed onto the solution - while the latter undergoes reconstruction - and it enforces its structure. The process can be vaguely related to 'carbon copying'. In the previous discussion we implicitly assumed that the features between the prior and the true solution were accurately spatially aligned or co-registered. However, in the real world this condition is not always guaranteed. If the probed anatomy is subjected to a spatial deformation between the data acquisition processes of the two modalities or if the modalities image the anatomy from a different angle or distance, then the prior image - although correct in its content (one-to-one feature correspondence with the true solution) - cannot be trusted in terms of 'where' its features should appear. A deformed tumour due to compression in Mammography of the breast would be the same tumour imaged by DOT. However because the latter images the breast without compression, the size and location between the two tumour representations would differ, rendering the prior information inaccurate.

Figure 1.3 showcases this problematic case. The prior image has the correct content, but it is translated, rotated, scaled and locally deformed. By blindly enforcing its mis-registered structure to

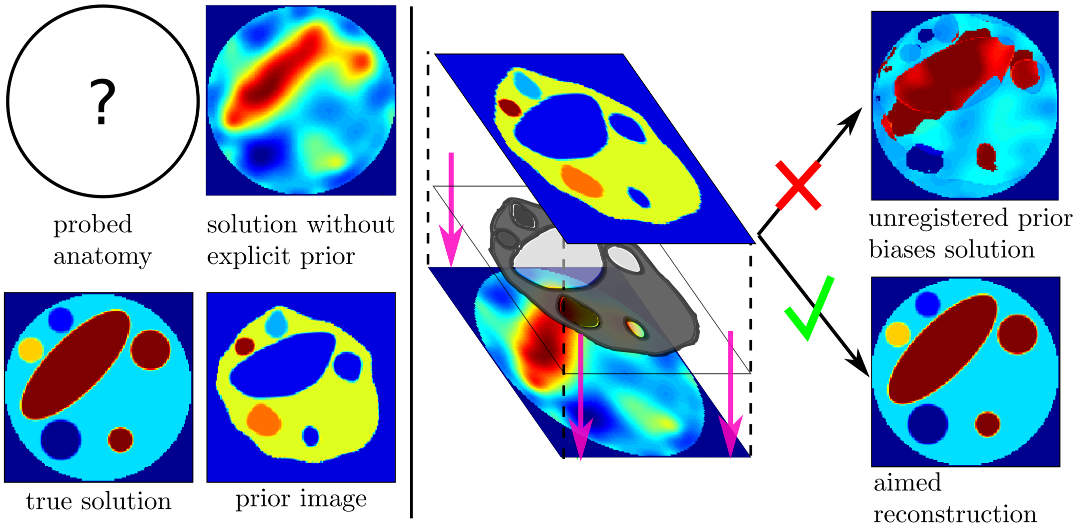

?
probed anatomy
solution without explicit prior
true solution
prior image
unregistered prior biases solution
aimed reconstruction

Figure 1.3: Aim 2: Introduction of structural information in DOT from high-resolution, *unregistered* images with *incommensurate* gray values

DOT, the obtained reconstruction is evidently biased. This gives rise to the secondary aim of this work, that is the incorporation of unregistered prior information in DOT, while minimizing the bias due to spatial mis-alignment.

The notions of image structure or structural information are extensively used throughout this thesis. Natural images, such as the ones depicting anatomy, are structured. In this context structure is manifested as strong dependencies in the values populating spatially proximal pixels [Wang et al., 2004]. In contrast, a grainy - for example white noise - is considered unstructured. However, a widely accepted definition of structure has been proven too elusive to find. This does not come as a surprise as according to Nielsen and Lillholm [2006] 'image structure is a collection of operationally defined image features'. Operationally defined features can be edges, corners or even local texture patterns. As the set of these features varies according to the task-in-hand, image structure can be perceived differently in different applications. One definition which approaches the perception of structure in this work, is given by Wang et al. [2004]. They define structural information as features in the image which are invariant to changes in global intensity (luminance) and contrast among regions. Such features can for example be the *presence* edges formed between regions. We emphasize presence, as when contrast between regions changes, the magnitude of the gradients usually employed to identify edges, changes as well. The structure of an anatomy depicted in the image should be invariant to the actual gray values populating the pixels. In a more general note, image structure should be invariant to any *bijective* transformation of the images' gray values.

In this work, structure is not modeled explicitly. We do not explicitly identify operationally defined structural features. We are not interested in the individual structure of the considered images. We are interested in assessing the structural *similarity* between two images. Information theoretic functionals which are exhibit a level of invariance to bijective transformations of the pixel values of one image and thus they can assess structural similarity without the need of explicitly defining structure or the features which comprise it.

Contributions We approach the first task by proposing an *information theoretic* regularization scheme for DOT. Information theoretic functionals such as joint entropy (JE) and mutual information (MI) enable the measurement of structural similarity between multi-modal images, due to their inherent capacity of by-passing the multi-modality barrier of the incommensurate relation between their values. During the reconstruction process, these functionals penalize all potential solutions which are not structurally similar to the prior image, effectively enforcing the structure of the latter on the obtained solution. The same functionals, due to their structural similarity evaluation capacity, can distinguish two spatially aligned images from two unregistered ones. We propose a simultaneous reconstruction/registration (SRR) scheme which treats the spatial location/shape of the prior as an unknown quantity, which we attempt to recover simultaneously with the optical solution.

In addition, we characterize the ability of the functionals to incorporate registered structural information from reference images with *incorrect content* - for example prior with missing or extra features.

Finally, we propose an extension to an already available scheme which enables the efficient manipulation of the marginal entropy of a single image, to the case of the JE and ultimately MI between two images. The scheme is crucial for achieving information theoretic regularization in tractable run-times. In addition, we characterize two possible formulations and implementations of entropy, regarding their accuracy and speed.

The next point is of importance. In the introductory section of this chapter, we outlined the advantages of DOT compared to alternative high resolution modalities. In our list we included cost and portability. If the robustness of DOT imaging could only be guaranteed in the case of multi-modal simultaneous imaging in conjunction with a high resolution modality such as MRI or CT, the low cost and the portability of DOT is irrelevant. The ultimate aim in multi-modal DOT imaging would be to use a single intra-subject image from a high resolution modality as prior for all subsequent DOT studies, given that the prior information does not compromise the new information obtained by each DOT scan. Such scheme can only work if we can compensate for deformations of the probed anatomy, which can take place between the acquisition of the high-resolution image and the subsequent DOT studies. This capability would re-establish the viability of the high resolution prior image to provide information for the individual follow-up DOT studies. In the most ideal case, one could use a *probabilistic high-resolution atlas* as a x_{ref} for DOT studies, completely removing the need for a high-resolution scan prior to DOT imaging. Again, compensating for mis-registration between the atlas and the anatomy probed by DOT would be an essential requirement. This work contributes towards this ultimate aim.

The information theoretic regularization of DOT and the efficient computation scheme has been already published [Arridge et al., 2008b; Panagiotou et al., 2009a,b]. Parts of this work have also contributed to publications in other modalities (PET/SPECT) [Kazantsev et al., 2010; Pedemonte et al., 2010a; Somayajula et al., 2010] as well as one undergoing completion [Pedemonte et al., 2010b]. We aim to publish the SRR scheme for DOT in the near future.

1.3 Structure of the thesis

The structure of this document is as follows. The next part is comprised of three chapters and introduces the relevant theory. The discussion is self-contained with emphasis on providing intuition via the use of examples which were produced specifically for this work, unless stated otherwise.

- Chapter 2 introduces the general concepts of inverse problem theory. These include the formal definition of the quantities, spaces and operators involved in an inverse problem, its forward counterpart, the problematic condition of ill-posedness and the approaches which can alleviate it - with emphasis on regularization. The discussion addresses both linear and non-linear inverse problems. In addition we briefly introduce non-linear optimization which is later employed in this work.

- Chapter 3 focuses on DOT. Firstly, the discussion introduces the fundamental physical concepts on which the operation of the modality is based. The forward problem of DOT is presented and involves the physical models of light propagation in tissue as well as approaches towards their implementation. The discussion proceeds by formulating the inverse problem of image reconstruction as an objective function minimization task. Finally, the chapter concludes by presenting selected literature, specifically focused on regularization methods proposed for DOT with emphasis on multi-modality.

- Chapter 4 introduces *information theory*, within which the two core concepts of this work - entropy and MI, are defined. The chapter starts by a brief introduction to probability theory and random variables which are completely fundamental to information theory. Entropy and MI are formally introduced, along with a discussion from an imaging perspective, on their inherent multi-modal image similarity evaluation capacity. Finally, we discuss the differences between the discrete and continuous formulation of the aforementioned concepts.

- Chapter 5 concludes the theoretical section by introducing medical image registration. The discussion is focused on the three major algorithmic parts comprising every registration algorithm, specifically *spatial transformations* with a reference to interpolation, *similarity measures* and registration specific *optimization* approaches. Regarding the first, we touch the concepts of linear and non-linear transformations. In the discussion regarding similarity measures, we revisit the information theoretic concepts of chapter 4, to comment on their capacity on measuring image dissimilarity due to spatial misalignment and not simply due to differences between the information depicted in two spatially aligned images. The discussion regarding optimization only refers to registration specific approaches often employed to improve convergence.

The third part encompasses the contribution of this work and introduces the proposed methods and the obtained results.

- In chapter 6 we present the efficient joint entropy evaluation and derivative computation scheme. In addition, the chapter evaluates the accuracy of two entropic formulations by comparing them against gold standards. Computational complexity estimates and run-time performance charts are explicitly provided.

- In chapter 7 we present the information theoretic regularisation of DOT via the incorporation of spatially registered, high resolution images with incommensurate values. An intuitive discussion regarding the regularizing capacity of JE and MI is provided. The chapter presents results obtained from specifically designed numerical simulations - both 2D and 3D - as well as an experimental study.

- Finally, in Chapter 8 we present the extension of information theoretic regularization of DOT in order to enable the incorporation of unregistered priors. The proposed SRR scheme compensates for potential global and also local (non-rigid) mis-registration between the optical solution and the prior images. Preliminary results on 2D simulations are presented as indicators of the validity of the method. The discussion re-examines the capacity of information theoretic functionals to be used for the combined purpose of regularizing DOT and driving a registration scheme.

Finally, the thesis concludes in Chapter 9, which summarizes the findings of this work and also it suggests potential future extensions towards the improvement on the current performance of the method.

Part II

Theory

Chapter 2

Inverse Problems

2.1 Introduction

This chapter introduces the theory of *inverse problems*, refers to the specific challenges surrounding them and outlines a sample of the basic approaches towards their solution. Inferring information regarding unobservable parameters of interest from secondary measurable quantities, by utilizing some known relation between the two, is the definition of an inverse problem. Considering the following example, inverse problems are not exclusively a science matter.

Selecting fruits from the market involves the *qualitative* solution of an inverse problem. One is firstly drawn to a fruit by its appearance. Obviously, one is mostly if not utterly interested in consuming its interior, however not all markets are gracious enough to allow the customer to try the product before paying. In the latter case, the information of interest is definitely unobservable. The consumer though, not willing to compromise the quality of the meal, attempts to infer information regarding the condition in the interior of the fruit, by making numerous external 'measurements', such as assessing its smell, texture, weight, plumpness or even echo to knocks on its boundary. Understandably, the customer relates the outcome of these 'measurements' to the unobservable interior state of the probed fruit, based on past experience. For example, the reader might be familiar with the fact that by thumping a ripe watermelon, the produced effect is a dull hollow sound. In the market one is required to invert this process and assess the watermelon by the produced sound. The process simply involves thumping the fruit and if the produced sound *matches* that of a ripe one (recalled by past experience), the purchase is finalized - always conditioned to fair pricing.

Considering the fact that most readers have tried bad fruits, this constitutes an initial indicator that the retrieval of an accurate solution of an inverse problem is in many cases a non-trivial task. To the surprise of the author, the inverse problem in the non-invasive *quantitative* assessment of the interior quality of watermelons has already been studied [Diezma-Iglesias et al., 2004].

In a more general note, assume a physical system under study. Tarantola [2004] divided the study of such a system in three parts:

1. The *parametrization of the system*, which regards the discovery of a minimal set of model parameters - denoted as x; which can completely characterize the system

2. the *forward modeling*, which regards the discovery of the physical laws enabling, for given values of the model parameters $x \in \mathbb{X}$, predictions on values of some secondary observable quantities $y \in \mathbb{Y}$. The physical relation between x and y is realized in the form of an operator $\mathcal{F}$ enabling the uni-directional mapping

$$\mathcal{F} : \mathbb{X} \mapsto \mathbb{Y} \tag{2.1}$$

3. and the *inverse problem*, which regards the inference of specific values of the model parameters x given that the secondary quantities y have been observed. Ultimately, the inverse problem is expressed as finding the mapping from $\mathbb{Y}$ to $\mathbb{X}$

$$\mathcal{F}^{-1} : \mathbb{Y} \mapsto \mathbb{X} \tag{2.2}$$

The reason inverse problems arise - or equivalently the need to infer x indirectly from y, is solely due to the inaccessibility of the former to be subjected to direct measurement. This can be either due to physical limitations or because the time instance at which x characterized the system, has already elapsed.

Retrieving a quantitatively accurate estimate of x can prove to be a challenging task. The are numerous causes which can potentially compromise the effort including i) the insufficient amount of useful data y being available, either due to high levels of noise contamination or due to the limited number of data acquisition events, ii) the complexity of the physical process itself, iii) the inherent sensitivity of some forward operators to the numerical manipulation applied to them during their inversion. Regarding the latter, this sensitivity can be a product of the discretisation process employed to approximate a naturally continuous problem [Hansen, 1998; Vogel, 2002]. These factors contribute to a condition known as *ill-posedness*, which is introduced in Sec. 2.3 and explicitly describes the characteristics of a problematic inverse problem.

Alleviating the effects of ill-posedness requires specialized treatment and the employment of advanced numerical methods. Methods specifically designed to treat ill-posedness are commonly referred to as *regularization methods*.

The discussion continues in Section 2.2 where the spaces involved in the inverse problem are formally defined. Section 2.3 outlines the Hadamard postulates of ill-posedness, central in inverse theory. The discussion continues in Sec. 2.4 by describing the linear inverse problem. Although the main topic of this Thesis involves a non-linear inverse problem, the analysis of the linear case provides detailed intuition by explicitly revealing the manifestation of ill-posedness and the actual effect of the various methods employed to address it. Section 2.5 generalizes on the linear case and formulates the non-linear inverse problem. Finally, Section 2.6 describes a sample of iterative objective function optimization methods which can be utilized by both the linear and non-linear inverse problems.

It should be noted that preceding sections are based on the deterministic setting, defined with point-based solution estimators. For a statistical approach to the inverse problem, the reader is redirected to the excellent texts [Kaipio and Somersalo, 2005; Kolehmainen, 2001; Tarantola, 2004].

2.2 Parameter and data space

Considering the definition of a physical system of the previous section, let $x = \{x_1, x_2, \ldots, x_N\}^{\mathrm{T}}$ be its parametrization with $x \in \mathbb{X}$. Focusing on the inverse problem, $\mathbb{X}$ is then referred to as the *parameter space* or *solution space*. Similarly let $y = \{y_1, y_2, \ldots, y_M\}^{\mathrm{T}}$ denote the observable quantities where $y \in \mathbb{Y}$ and $\mathbb{Y}$ is known as the *data space*. Let $\mathbb{X} \subset \mathbb{R}^N$, $\mathbb{Y} \subset \mathbb{R}^M$. The top index $x^{(k)}$ is used to distinguish between different vectors x, populated by different values $\mathrm{x}_i,\ i = 1 \ldots, N$.

Both vector spaces $\mathbb{X}$ and $\mathbb{Y}$ are considered to be metric spaces. These are spaces where the distances between points lying in the space - for example $x_i, x_j \in \mathbb{X}$ - can be measured using some distance function $\mathcal{D}(x_i, x_j) \in \mathbb{R}^+$, $\forall i, j$, for which it holds 1) $\mathcal{D}(x_i, x_j) \geq 0$ with the equality holding only when $x_i = x_j$ 2) $\mathcal{D}(x_i, x_j) = \mathcal{D}(x_j, x_i)$ and 3) $\mathcal{D}(x_i, x_j) \leq \mathcal{D}(x_i, x_k) + \mathcal{D}(x_k, x_j)$. In practice, $\mathbb{X}$ and $\mathbb{Y}$ are typically N and M dimensional Hilbert spaces $\mathbb{H}$ or Banach spaces $\mathbb{B}$. A Hilbert space $\mathbb{H}$ is a vector space where the embedded norm used to measure the distances between points is expressed in terms of the inner product. Let vectors $u, v \in \mathbb{R}^N$ each comprised by $\mathrm{u}_i, \mathrm{v}_i,\ i = 1 \ldots N$. Then $\mathbb{R}^N$ is a Hilbert space under the Euclidean inner product $\langle u, v \rangle_2 = u^{\mathrm{T}} v = \sum_{i=1}^{n} \mathrm{u}_i \mathrm{v}_i$ and the induced norm is the Euclidean norm $\|u\|_2 = \sqrt{\langle u, u \rangle_2} = \sqrt{\sum_{i=1}^{n} \mathrm{u}_i^2}$ [Vogel, 2002]. Banach spaces are more general as the embedded norm does not have to be strictly defined in terms of the inner product. Banach spaces are a generalization of Hilbert spaces.

2.3 Ill-posedness

Hadamard [1902] outlined the following three postulates which are considered as prerequisites in order a system to be well-behaved [Hansen, 1998; Vogel, 2002]:

(i) *Existence of the solution*: For each $y \in \mathbb{Y}$ there exists a solution x for which $\mathcal{F}(x) = y$

(ii) *Uniqueness of the solution*: The solution x is unique

(iii) *Continuity*: Small changes in y should not result in arbitrarily large perturbations in x. Hence, assuming $\mathcal{F}(x_1) = y_1$ and $\mathcal{F}(x_2) = y_2$, then $x_1 \rightarrow x_2$ whenever $y_1 \rightarrow y_2$

When all three postulates are met, the problem is classified as *well-posed*. In any other case the problem is *ill-posed*. The forward problem in this work is well-posed as it meets all the above postulates. The inverse problem however is very often ill-posed due to the numerous factors, some of them briefly outlined in Sec. 2.1. It should be mentioned that in the strict sense, a finite dimensional inverse problem cannot be *ill-posed* with respect to the third postulate (continuity), as the ratio $\|\Delta x\| \,/\, \|\Delta y\|$ stays always bounded, hence no “arbitrarily” large perturbations occur [Hansen, 1998; Kaipio and Somersalo, 2005; Kolehmainen, 2001]. However unless the discretisation is very coarse, $\|\Delta x\| \,/\, \|\Delta y\|$ - although bounded, can be significantly large. Therefore the discrete inverse problem will be far too sensitive to errors in y during numerical manipulation and effectively ill posed. These problems are often referred to as *discrete ill-posed inverse problems* [Hansen, 1998].

2.4 Linear case

This section is based on a collection of excellent sources, the most prominent ones being [Golub and Van Loan, 1996; Hansen, 1998; Kaipio and Somersalo, 2005; Kolehmainen, 2001; Strang, 1988; Vogel, 2002] and Section 2.6 in [Press et al., 1992b]. Let $\mathcal{F} \in \mathbb{R}^{M \times N}$ be a linear operator formed as a $M \times N$ matrix. Specifically for the linear case, the forward mapping of Eq. 2.1 is realized for specific values of x as

$$\mathcal{F}x = y. \tag{2.3}$$

Important definitions regarding operators such as $\mathcal{F}$ include the *range* and the *null space*. The range $\mathcal{R}(\mathcal{F})$ of $\mathcal{F}$ is defined as

$$\mathcal{R}(\mathcal{F}) = \{y \in \mathbb{Y} | y = \mathcal{F}x, \ \forall x \in \mathbb{X}\}, \mathcal{R}(\mathcal{F}) \subseteq \mathbb{Y} \tag{2.4}$$

and the null space Null($\mathcal{F}$)

$$\text{Null}(\mathcal{F}) = \{x \in \mathbb{X} | \mathcal{F}x = 0\}, \text{Null}(\mathcal{F}) \subseteq \mathbb{X}. \tag{2.5}$$

Another crucial concept is the *rank* r of $\mathcal{F}$. It is the number of linearly independent rows and columns in $\mathcal{F}$, with $r < \min(N, M)$ [Strang, 1988]. The *nullity* is defined as the number of the linearly dependent rows in $\mathcal{F}$ and equals the dimensionality of the null space or $dim(\text{Null}(\mathcal{F})) = M - r$. This is based on the *rank and nullity theorem* where *rank* + *nullity* = M [Farenick, 2000; Strang, 1988]. We also define the *left-null space* $\text{Null}(\mathcal{F}^{\mathrm{T}})$ and the *row-space* $\mathcal{R}(\mathcal{F}^{\mathrm{T}})$. For these spaces it holds that

$$\mathcal{R}(\mathcal{F}) = \text{Null}(\mathcal{F}^{\mathrm{T}})^{\perp} \tag{2.6}$$

$$\mathcal{R}(\mathcal{F}^{\mathrm{T}}) = \text{Null}(\mathcal{F})^{\perp} \tag{2.7}$$

where $\perp$ denotes the *orthogonal complement*.

2.4.1 Singular Value Decomposition

The singular value decomposition (SVD) is a powerful factorization method providing insight regarding potential problems during the attempt to solve $\mathcal{F}x = y$ with respect to x. In addition it enables the computation of $\mathcal{F}^{-1}$ if that exists or in any other case it can provide the best compromise solution. Under SVD, $\mathcal{F}$ is represented as the product of three matrices:

$$\mathcal{F} = U\Sigma V^{\mathrm{T}} = \sum_{i=1}^{N} u_i \lambda_i v_i^{\mathrm{T}} \tag{2.8}$$

where $\Sigma \in \mathbb{R}^{M \times N}$ is the diagonal matrix

$$\Sigma = \begin{pmatrix} \Sigma_r & 0_{r\times(N-r)} \\ 0_{(M-r)\times r} & 0_{(M-r)\times(N-r)} \end{pmatrix} \in \mathbb{R}^{M\times N}, \tag{2.9}$$

where $\Sigma_r = diag(\lambda_1, \lambda_2, \ldots, \lambda_r)$ and $\lambda_1 > \lambda_2 > \cdots > \lambda_r > 0$. The diagonal entries λ_i are known as the *singular values*. Rank r equals the number of non-zero entries in the diagonal of Σ. $U \in \mathbb{R}^{M\times M}$ and $V \in \mathbb{R}^{N\times N}$ are matrices defined as $U = (u_1, u_2, \ldots, u_M)$ and $V = (v_1, v_2, \ldots, v_N)$. The column vectors $u_i \in \mathbb{R}^M$ and $v_j \in \mathbb{R}^N$ are known as the *left* and *right singular vectors* and satisfy

$$u_i^{\mathrm{T}} u_j = \delta_{ij} \quad , \quad v_i^{\mathrm{T}} v_j = \delta_{ij} \tag{2.10}$$

$$\mathcal{F} v_i = \lambda_i u_i \quad , \quad \mathcal{F}^T u_i = \lambda_i v_i \tag{2.11}$$

where δ_{ij} denotes the Kronecker delta defined as:

$$\delta_{ij} \equiv \begin{cases} 0, & \forall i \neq j \\ 1, & \forall i = j \end{cases} \tag{2.12}$$

The condition number of the matrix $\mathcal{F}$ is defined as the ratio between the largest and smallest singular value:

$$cond(\mathcal{F}) = \frac{\lambda_1}{\lambda_r} \tag{2.13}$$

and it is an indicator of the sensitivity of $\mathcal{F}$ to numerical manipulation.

A square $\mathcal{F}$ is an essential condition in order for an inverse $\mathcal{F}^{-1}$ to exist which would satisfy both solution existence & uniqueness. In most inverse problems $\mathcal{F}$ is rectangular with $M \neq N$, hence there is an immediate breach of the first two postulates of well-posedness (see Sec. 2.3). More specifically in the *under-determined* case $(N > M)$ there are infinite solutions (Null$(\mathcal{F}) \neq \{\emptyset\}$), with $\{\emptyset\}$ denoting the empty set. In the *over-determined* case $(N < M)$ there will be no solution - except for some very “thin” subset of potential y, specifically if the latter is expressed as a linear combination of the columns of $\mathcal{F}$ [Strang, 1988]. Finally, even when $M = N$, the problem can have infinite solutions if $r < \min(M, N)$. The condition $M \neq N$ can be easily identified by the size of $\mathcal{F}$. Rank $r < \min(M, N)$ can be evaluated by SVD. In all cases, the adherence of the inverse problem to the first two postulates have to be restored and SVD can provide the next best compromise solution. SVD can also identify the reasons for breaches of the third postulate. These identified cases are classified in three distinct categories:

(i) The problem is rank-deficient with $r < min(N, M)$.

(ii) Assume the case of maximum possible rank $r = \min(M, N)$. Let Σ_r be comprised by two distinct sets $\{\lambda_1, \lambda_2, \ldots, \lambda_k\}$ and $\{\lambda_{k+1}, \lambda_{k+2}, \ldots, \lambda_r\}$, where the entries in the latter tend to zero. Then the problem is known to be *numerically* rank deficient with numerical rank k, as $r - k$ rows and columns are *almost* linearly dependent. The $r - k$ equations form the numerical nullspace of $\mathcal{F}$.

(iii) The singular values in Σ_r decay gradually to almost zero without the existence of a distinct gap as in (ii). This is the case of discrete ill-posed problems with ill-determined rank.

The following sections outline the most popular methods to approach the problematic cases outlined above and also provide an intuitive interpretation.

2.4.2 Linear least squares

The method of least squares (LS) is used to restore solution existence by attempting to find the best compromise solution $\hat{x}_{LS}$ to the true solution $x^\star$ when the latter is unreachable or inseparable from a set of possibly infinite solutions.

2.4.2.1 Mostly one solution

The system $\mathcal{F}x = y$ has *at most* one solution when $r = N \leq M$ (full column rank), as there are no linearly dependent columns hence Null$(\mathcal{F}) = \{\emptyset\}$. In the practical setting however, the measured data y would unavoidably be contaminated with some form of random noise, which can render it unreachable by $\mathcal{F}$ or equivalently $y \notin \mathcal{R}(\mathcal{F})$. In that case Eq. 2.3 does not have an exact solution. Hence there is a breach of the Hadamard postulate of well-posedness regarding solution existence. However, the need to find a "practical" solution estimate $\hat{x}$ to the given problem still exists. The effort in this cases focuses on the retrieval of the *best compromise solution estimate* $\hat{x}_{LS} \in \mathcal{R}(\mathcal{F})$, where the modeled data $\mathcal{F}\hat{x}_{LS}$ is maximally proximal to the true measurements y. This proximity is assessed by a *similarity measure* $\mathcal{D}(\hat{x})$ also known as *data discrepancy functional* or *data fit term*. In the case of LS, the data fit term is expressed by the squared Hilbert norm (see Section 2.2). Consider the following term

$$\mathcal{D}(x) = \|\mathcal{L}_W(y - \mathcal{F}x)\|^2 . \tag{2.14}$$

Under the assumption that individual observation $y \in \mathbb{Y}$ are uncorrelated, then $\mathcal{L}_W$ is a diagonal matrix which can be used to assign different weights to individual measurements, when the latter are not trusted equally. Essentially, when $\mathcal{L}_W = I$ then $\hat{x}$ minimizes the average error in all M equations and Eq. 2.14 is the standard LS. If the diagonal elements in $\mathcal{L}_W$ vary, then the weighted average distance of $\mathcal{F}\hat{x}_{LS}$ from y might be closer to some individual measurements $y \in \mathbb{Y}$ than others. For $\mathcal{L}_W \neq I$, Eq. 2.14 is known as the *weighted least squares* functional.

The retrieval of a solution estimate $\hat{x}_{LS}$ satisfying Eq. 2.14 can be formulated as a problem of error minimisation

$$\hat{x}_{LS} = \arg\min_{x} \left[\mathcal{E}(x) = \|\mathcal{L}_W(y - \mathcal{F}x)\|^2\right] . \tag{2.15}$$

In that case $\hat{x}_{LS}$ should lie in the bottom of some multidimensional basin in the solution space, therefore satisfying the condition $\partial\mathcal{E}(\hat{x}_{LS})/\partial x = 0$. Therefore, Eq. 2.15 is expanded and the corresponding derivatives are set to zero

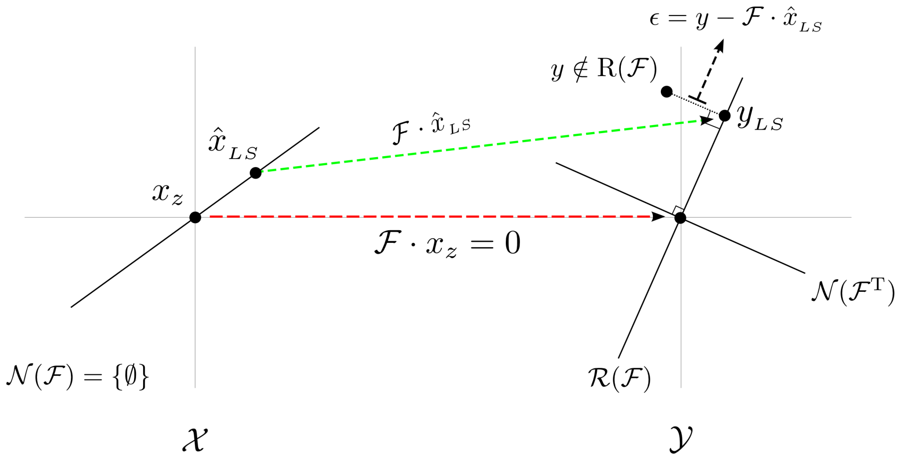
$$\partial \mathcal{E}(\hat{x}_{LS})/\partial x = 0 \tag{2.16}$$

$$\frac{\partial}{\partial x}\left[(y - \mathcal{F}\hat{x}_{LS})^{\mathrm{T}} W (y - \mathcal{F}\hat{x}_{LS})\right] = 0 \tag{2.17}$$

$$\mathcal{F}^{\mathrm{T}} W (y - \mathcal{F}\,\hat{x}_{LS}) = 0 \tag{2.18}$$

where $W = \mathcal{L}_W^{\mathrm{T}}\mathcal{L}_W$. Eq. 2.18 is known as the *normal equations*. The normal equations have a geometrical interpretation. Let $\epsilon = (y - \mathcal{F}\,\hat{x}_{LS})$. As ϵ sets Eq. 2.18 to 0, then $\epsilon \in \mathrm{Null}(\mathcal{F}^{\mathrm{T}})$. Considering Eq. 2.6, then $\epsilon \perp \mathcal{R}(\mathcal{F})$ which is the shortest distance between the noisy data $y \notin \mathcal{R}(\mathcal{F})$ and the range $\mathcal{R}(\mathcal{F})$ of reachable data.

Figure 2.1 schematically shows the action of $\mathcal{F}$ and the involved spaces in the discussed case. We have assumed $W = I$. Least squares computes an $\hat{x}_{LS}$ mapped by $\mathcal{F}$ to the orthogonal projection of y in $\mathcal{R}(\mathcal{F})$. The distance $\epsilon \perp \mathcal{R}(\mathcal{F})$ is minimal. Because all columns in $\mathcal{F}$ are linearly independent ($r = N$), it holds $\mathrm{Null}(\mathcal{F}) = \emptyset$. In addition if $r < M$, then $\mathrm{Null}(\mathcal{F}^{\mathrm{T}}) \neq \emptyset$.



$\epsilon = y - \mathcal{F}\cdot\hat{x}_{LS}$
$y \notin \mathrm{R}(\mathcal{F})$
$\mathcal{F}\cdot\hat{x}_{LS}$
y_{LS}
$\hat{x}_{LS}$
x_z
$\mathcal{F}\cdot x_z = 0$
$\mathcal{N}(\mathcal{F}^{\mathrm{T}})$
$\mathcal{N}(\mathcal{F}) = \{\emptyset\}$
$\mathcal{R}(\mathcal{F})$
$\mathcal{X}$
$\mathcal{Y}$

Figure 2.1: Action of a linear forward operator with full column rank $r = N \leq M$ and data not being inside the range of the forward operator or $y \notin \mathcal{R}(\mathcal{F})$. In this case a solution does not exist as no $x \in \mathbb{X}$ maps - via the application of $\mathcal{F}$ to y. Least squares establish solution existence (first Hadamard postulate) by retrieving the best compromise solution $\hat{x}_{LS}$. The retrieved $\hat{x}$ maps to $y_{LS} \in \mathcal{R}(\mathcal{F})$ which is minimally distant from the measured $y \notin \mathcal{R}(\mathcal{F})$.

When $r = N \leq M$ then $(\mathcal{F}^{\mathrm{T}} W \mathcal{F})$ is *square*, *symmetric* and *invertible*. After some rearrangement Eq. 2.18 returns the LS solution estimate

$$\hat{x}_{LS} = (\mathcal{F}^{\mathrm{T}} W \mathcal{F})^{-1}\mathcal{F}^{\mathrm{T}} W y, \tag{2.19}$$

where $\mathcal{F}^{\dagger} = (\mathcal{F}^{\mathrm{T}} W \mathcal{F})^{-1}\mathcal{F}^{\mathrm{T}} W$ is the weighted *Moore-Penrose* pseudo-inverse originally defined for $W = I$ [Moore, 1920; Penrose, 1955]. The linear system of Eq. 2.18 can be solved by a choice of methods such LU or QR decompositions [Golub and Van Loan, 1996; Press et al., 1992b] or by more advanced iterative methods with reduced memory requirements such as the *generalized minimum residual method* (GMRES) [Saad and Schultz, 1986] or the linear conjugate gradients (CG) method [Shewchuk, 1994].

2.4.2.2 Infinite solutions

Let again $\mathcal{F}x = y$, but with $r < N$ leading to Null$(\mathcal{F}) \neq \{\emptyset\}$. The problem falls in the first of the cases outlined in Section 2.4.1 of rank deficiency due to linearly dependent columns. The LS solution $\hat{x}_{_{LS}}$ still exists, but is now not unique. This constitutes a breach of the Hadamard postulate of well-posedness regarding solution uniqueness. In fact, there are infinite $\hat{x}_{_{LS}}$ that map to $y_{_{LS}}$. To restore the second Hadamard condition, SVD can be employed. From all solutions $\hat{x}_{_{LS}}$ minimizing $\|\mathcal{F}x - y\|^2$, SVD singles out as a unique solution $\hat{x}_{_{SVD}}$, the one which has shortest length $\|\hat{x}\|$. This is expressed as

$$\hat{x}_{_{SVD}} = \arg\min_{x} \left\{ \mathcal{E}(x) = \|\Xi(x)\| \;\middle|\; \Xi(x) = \|\mathcal{L}_W(\mathcal{F}x - y)\|^2 \right\}. \tag{2.20}$$

Figure 2.2 graphically shows the action of $\mathcal{F}$ for this case. As $r < N$ then Null$(\mathcal{F}) \neq \emptyset$. Any $x \in \mathbb{X}$ can now be split into two components $x_r \in \mathcal{R}(\mathcal{F}^{\mathrm{T}})$ and $x_n \in$ Null$(\mathcal{F})$, which are orthogonal to each other. Effectively, $\|x\|^2 = \|x_r\|^2 + \|x_n\|^2$. Every $y \in \mathbb{Y}$ comes from a unique x_r [Strang, 1988]. The infinite solutions $\hat{x}_{_{LS}}$ are comprised by that x_r and the infinite x_n. In the schematic, the component x_r of the solutions that map to $y_{_{LS}}$ is denoted as $\hat{x}_{_{SVD}}$ and is in fact the SVD solution. The infinite solutions form the line $\hat{x}_{_{LS}}$ parallel to Null$(\mathcal{F})$ passing from $\hat{x}_{_{SVD}}$. From all solutions that project to $x_{_{LS}}$, $\hat{x}_{_{SVD}}$ is the shortest one as $\|\hat{x}_{_{SVD}}\| < \|\hat{x}_{_{LS}}\| \, \forall \hat{x}_{_{LS}}$. In effect, $\hat{x}_{_{SVD}}$ is singled out by zeroing the null component $\|\hat{x}_n\| = 0$.

When $r < N$ then $\nexists (\mathcal{F}^{\mathrm{T}}\mathcal{F})^{-1}$ [Strang, 1988]. The LS pseudo-inverse $\mathcal{F}^{\dagger}$ is hence not defined and $\hat{x}_{_{SVD}}$ or $\hat{x}_{_{SVD}}$ cannot be obtained by Eq. 2.19. Fortunately a *shortest length* LS solution vector can be computed using the SVD pseudo-inverse $\mathcal{F}^{\dagger}$[1] given by

$$\mathcal{F}^{\dagger} = V\Sigma^{+}U^{\mathrm{T}} \tag{2.21}$$

where Σ^{+} is a $N \times M$ matrix formed by replacing all non-zero entries in the diagonal of the Σ by their reciprocal or

$$\Sigma^{+} = \begin{pmatrix} \Sigma_r^{-1} & 0_{r\times(M-r)} \\ 0_{(N-r)\times r} & 0_{(N-r)\times(M-r)} \end{pmatrix}. \tag{2.22}$$

Considering Eq. 2.8, the solution estimate using the SVD pseudo inverse is given by [Strang, 1988]

[1] the same notation is adopted for both LS and SVD pseudo-inverses

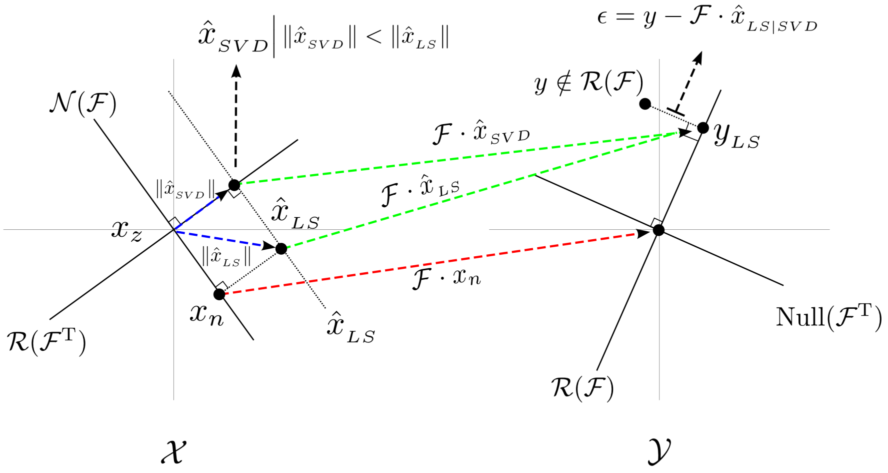

$\hat{x}_{SVD} \Big| \, \|\hat{x}_{SVD}\| < \|\hat{x}_{LS}\|$
$\epsilon = y - \mathcal{F} \cdot \hat{x}_{LS|SVD}$
$y \notin \mathcal{R}(\mathcal{F})$
$\mathcal{N}(\mathcal{F})$
$\mathcal{F} \cdot \hat{x}_{SVD}$
y_{LS}
$\|\hat{x}_{SVD}\|$
$\mathcal{F} \cdot \hat{x}_{LS}$
$\hat{x}_{LS}$
x_z
$\|\hat{x}_{LS}\|$
$\mathcal{F} \cdot x_n$
x_n
$\hat{x}_{LS}$
$\mathrm{Null}(\mathcal{F}^{\mathrm{T}})$
$\mathcal{R}(\mathcal{F}^{\mathrm{T}})$
$\mathcal{R}(\mathcal{F})$
$\mathcal{X}$
$\mathcal{Y}$

Figure 2.2: Action of a linear rank-deficient forward operator with $r < N$ and data not being inside the range of the forward operator $y \notin \mathcal{R}(\mathcal{F})$. In this case $\mathrm{Null}(\mathcal{F}) \neq \emptyset$. The infinite set of solutions $\hat{x}_{LS}$ are comprised by the $\hat{x}_{SVD}$ part and the infinite null components $x_n \in \mathrm{Null}(\mathcal{F})$. The SVD establishes solution uniqueness by singling out a solution estimate $\hat{x}_{SVD}$ from the infinite $\hat{x}_{LS}$, by setting the null component to zero via $\hat{x}_{SVD} = \underset{\hat{x}_{LS}}{\arg\min} \{ \|\hat{x}_{LS}\| \}$

$$\hat{x}_{SVD} = \mathcal{F}^{\dagger} y \tag{2.23}$$

$$= \sum_{i=1}^{N} \frac{u_i^T y}{\lambda_i} v_i \tag{2.24}$$

The above minimizes $\|\mathcal{F}x - y\|^2$ only for the r linearly independent equations. For the rest nothing can be done as they have a form

$$U_{(M-r)\times N} \Sigma_{(M-r)\times N} \left(V_{(M-r)\times N}\right)^{\mathrm{T}} \cdot \hat{x}_{SVD} = y_{(M-r)\times 1} \tag{2.25}$$

$$0_{(M-r)\times N} \cdot \hat{x}_{SVD} = y_{(M-r)\times 1} \tag{2.26}$$

due to the 0 population in the Σ sub-matrix $\Sigma_{(M-r)\times N}$. Any of the infinite LS solutions $\hat{x}_{LS}$ satisfies the $M - r$ equations. To summarize, the SVD solution $\hat{x}_{SVD}$ is the unique, shortest length LS solution.

2.4.2.3 Numerically rank deficient system

From the previous sections it is evident that solution existence and uniqueness can be restored. The inverse problem can still violate the third Hadamard postulate of continuity. This is the case of the second and third categories outlined in Section 2.4.1. Regarding the case (ii), a subset $\lambda_{(r-k)}$ of the singular values tend to zero with a clear gap from the remaining λ_k. The almost $r - k$ linearly dependent columns/rows can cause numerical instability during the computation of $\hat{x}_{SVD}$. Considering Eq. 2.24, as λ_i becomes increasingly small it amplifies potential errors y (for example due to noise contamination) which can significantly affect $\hat{x}_{SVD}$ [Hansen, 1990b, 1998]. Problems characterized by this gap in the singular values can be numerically treated by the *truncated* SVD or truncated singular value decomposition (TSVD) [Hansen, 1987, 1990b, 1998; Varah, 1979]. The method simply replaces the problematic small λ_i with 0. Effectively it removes the almost linearly dependent parts of the linear system. In practice a threshold τ is involved to separate $\lambda_{(r-k)}$ from λ_k. The TSVD solution can then be expressed by the standard SVD solution (Eq. 2.28) multiplied by a *filter function* $w_\tau(\lambda^2)$ [Vogel, 2002]

$$w_\tau(\lambda^2) = \begin{cases} 1, & \lambda^2 > \tau \\ 0, & \lambda^2 \leq \tau \end{cases} \tag{2.27}$$

Then the TSVD solution is given by

$$\hat{x}_{TSVD} = \sum_{\lambda_i}^{N} w_\tau(\lambda_i^2) \frac{u_i^T y}{\lambda_i} v_i. \tag{2.28}$$

By setting the cluster of small λ to 0, TSVD removes measured data. However, the removed data does not contain useful information but on the contrary would have rendered the system numerically unstable. TSVD effectively is an effort towards the numerical treatment of a problematic linear system and to re-establish the problem's adherence to the Hadamard postulates of well-posedness. In this sense, TSVD improves all three Hadamard's postulates as a best compromise solution in the LS sense always exists, it is unique as SVD chooses the one with minimum length and it is stable by zeroing the error generating $\lambda_{(r-k)}$. As TSVD restores well-posedness it is considered to be a regularization method [Hansen, 1992b]. Alternative methods exist to filter the small λ, such as the Tikhonov regularization introduced Sec. 2.4.3.

2.4.2.4 Discrete ill-posed problems with ill-determined rank

Regarding case (iii) of Section 2.4.1 where the singular values gradually approach zero, regularization is again required as similar to the previous section small singular values result in instability. One of the problems in this category is that it is now not obvious where to truncate Σ (this is the matrix containing λ) prior to inversion. Using a low threshold α in TSVD, might not remove enough $\lambda_i \rightarrow 0$ to adequately improve the stability of the linear system. In contrast, a high α would stabilize the system but could also remove essential, information rich data measurements. Approaches towards the numerical treatment of this kind of problems include TSVD, although this time the choice of threshold α should preferably be based on a more elaborate strategy.

Regarding the TSVD case, it has been suggested that a reasonable index i to apply the truncation in the diagonal of Σ would be the one where the *discrete Picard condition* cease to be satisfied. This was firstly understood by [Varah, 1979] and discussed in further detail in [Hansen, 1990a,b]. Considering Eq. 2.28, the discrete Picard condition states that in order for the system to be numerically stable then the rate of decay of $\left|u_i^\mathrm{T} y\right|$ should be in average faster than λ_i. In any other case $\lambda_i \rightarrow 0$ would significantly affect the solution. Alternative popular regularization methods include the Tikhonov regularization [Phillips, 1962; Tikhonov, 1963], the *maximum entropy method* [Burch et al., 1983; Cover and Thomas, 1991;

Jaynes, 1982] and many more. The method of Tikhonov regularization will be used as an example in order to discuss the effect of regularization in both the linear and non-linear setting.

2.4.3 Generalized Tikhonov regularization

Let an objective function be comprised by a data discrepancy measure such as the LS in Eq. 2.15. Generalized Tikhonov regularization [Vogel, 2002] is applied by introducing a penalty term $\Psi(x)$ to the objective function according to

$$\hat{x}_{\text{TK}} = \arg\min_{x} \left[\mathcal{E}(x) = \|\mathcal{L}_W(\mathcal{F}x - y)\|^2 + \tau\Psi(x) \right]. \tag{2.29}$$

The term $\Psi(x)$ is a scalar function. It effectively imposes soft constraints to $\hat{x}$, by variably penalizing candidate solutions x, depending on how much they deviate from the solutions satisfying $\Psi(x)$. More specifically, the solution of Eq. 2.29 is related to the solution of the following constrained problem [Björck, 1996]

$$\arg\min_{x} \left[\|\mathcal{L}_W(\mathcal{F}x - y)\|^2 \right], \qquad \text{subject to } \Psi(x) < \tau, \tau \in \mathbb{R}^+ \tag{2.30}$$

The actual form of $\Psi(x)$ is usually problem specific. The scalar $\tau \in \mathbb{R}^+$ in Eq. 2.29 is known as the *regularization parameter* and weights the penalization imposed by $\Psi(x)$. The selection of τ is of high-importance and is a field of research on its own. Small values for τ lead to reduced regularization, therefore the obtained solution estimates can be expected to be noisy and incorrect due to the untreated ill-posedness conditions. In contrast, high τ can lead to the over-regularization of the problem where $\Psi(x)$ dominates the solution, rendering the measured data y insignificant. A sample of schemes facilitating elaborate approaches towards the selection of τ include the L-curve, the generalized cross validation and the discrepancy principle. For a detailed discussion the reader is referred to Chapter 7 in [Vogel, 2002], [Hansen, 1998] and the references within.

Ordinary Tikhonov regularization [Kolehmainen, 2001; Vogel, 2002] refers to a particular type of $\Psi(x)$ - specifically quadratic functionals. These are widely adopted regularizing schemes for many inverse problems. The generic form for these penalties is expressed as

$$\Psi(x) = \left\| L(x - \hat{x}^{(0)}) \right\|^2 \tag{2.31}$$

where $L \in \mathbb{R}^{M\times N}$ is a regularization operator and $\hat{x}^{(0)}$ is the initial estimate of $x^\star$. The specifics of L distinguish between the Tikhonov regularization methods. $L = I$ results in the zeroth-order Tikhonov (TK$_0$) regularization scheme. When the diagonal entries in L differ, the scheme corresponds to weighted least squares where variable penalty weights the individual entries x in x. L can also be a differential operator of various orders. The first order differential operator results in the popular first-order Tikhonov (TK$_1$) regularization, which penalizes for non-smooth solutions. We then expand $\mathcal{E}(x)$ of Eq. 2.29, with the term $\Psi(x)$ being given by Eq. 2.31, derive its partial derivatives and set them to **0**, in a manner similar to the derivation of Eq. 2.18. Then, after a rearrangement of the terms, one obtains the augmented normal equations [Kolehmainen, 2001]

$$(\mathcal{F}^{\mathrm{T}}W\mathcal{F}+\tau L^{\mathrm{T}}L)x=\mathcal{F}^{\mathrm{T}}Wy+\tau L^{\mathrm{T}}L\hat{x}^{(0)}. \tag{2.32}$$

where $W=\mathcal{L}_W^{\mathrm{T}}\mathcal{L}_W$. Revealing some of the effects of the regularization is possible by expressing Eq. 2.29 with $\Psi(x)$ of Eq. 2.31 in a stacked form [Hansen, 1998; Kolehmainen, 2001; Varah, 1979]

$$\mathcal{E}(x)=\left\|\begin{pmatrix}\mathcal{F}\mathcal{L}_W\\ \tau^{1/2}L\end{pmatrix}x-\begin{pmatrix}\mathcal{L}_W y\\ \tau^{1/2}L\hat{x}^{(0)}\end{pmatrix}\right\|^2 \tag{2.33}$$

Using this notation the contribution of the regularization function towards the alleviation of ill-posedness becomes apparent. Given that $\mathcal{F}$ has linearly independent columns, then the augmented $\mathcal{F}$ of Eq. 2.15 is now a full column rank resulting in a unique solution [Björck, 1996].

By setting $L=I$ and $\hat{x}^{(0)}=\mathbf{0}$ one obtains the standard TK_0 regularization scheme

$$\hat{x}_{\text{TK0}}=\arg\min_x\left[\mathcal{E}(x)=\|\mathcal{L}_W(\mathcal{F}x-y)\|^2+\tau\|x\|^2\right]. \tag{2.34}$$

with corresponding normal equations

$$(\mathcal{F}^{\mathrm{T}}W\mathcal{F}+\tau I)\hat{x}_{\text{TK0}}=\mathcal{F}^{\mathrm{T}}Wy. \tag{2.35}$$

After rearranging Eq. 2.35 to $\hat{x}_{\text{TK0}}=(\mathcal{F}^{\mathrm{T}}W\mathcal{F}+\tau I)^{-1}\mathcal{F}^{\mathrm{T}}Wy$, it can be expanded using the SVD expression of Eq. 2.8. In addition, by considering Eqs. 2.10 and 2.11 [Vogel, 2002], it results in

$$\hat{x}=\sum_{i=1}^{N}\frac{\lambda_i(u_i^{\mathrm{T}}y)}{\lambda_i^2+\tau}v_i \tag{2.36}$$

By considering the SVD pseudo-inverse solution (Eq. 2.24), then Eq. 2.36 reveals that the TK_0 regularization effectively acts as a filter for small singular values. Similar to the TSVD notation (Eq. 2.27), the TK_0 filter function is given:

$$w_\tau(\lambda^2)=\frac{\lambda^2}{\lambda^2+\tau} \tag{2.37}$$

It should be noted that all objective functions expressed in the linear case can also be solved using iterative minimisation schemes. We describe some prominent optimization methods in the context of the non-linear inverse problem in Sec. 2.6. For the case of $L=\nabla$, one obtains the TK_1. It penalizes for non-smooth features - specifically the gradients in x. Its SVD filter analogue can be derived in a manner similar to TK_0.

2.5 Non-linear case

The classification of an inverse problem as non-linear mainly refers to the nature of the forward operator which now describes a more complex relationship between x and y. The non-linear realization of the forward mapping of Eq. 2.1 for specific values x is expressed as

$$y = \mathcal{F}(x). \tag{2.38}$$

Assuming Gaussian noise contamination of y [Arridge, 1999; Viola, 1995], the LS data fit term can be employed as an objective function

$$\mathcal{E}(x) = \|y - \mathcal{F}(x)\|^2 + \tau\Psi(x) \tag{2.39}$$

It should be noted that in the non-linear case, an inverse or pseudo-inverse operator $\mathcal{F}^{-1}$ is not realized. The final solution $\hat{x}$ is obtained by an iterative minimization of $\mathcal{E}(x)$ expressed as

$$\hat{x} = \arg\min_{x} \left[\mathcal{E}(x) = \|y - \mathcal{F}(x)\|^2 + \tau\Psi(x)\right]. \tag{2.40}$$

The above minimizes the discrepancy between the modeled data and the measured data subject to the soft constraints imposed by $\Psi(x)$. By itself, the LS functional guarantees solution existence by obtaining the best compromise solution, however this solution might not be unique and the continuity between x and $\mathcal{F}(x)$ might not be guaranteed (2nd and 3rd Hadamard postulates of well-posedness). Although, the explicit action of $\Psi(x)$ cannot be demonstrated in a manner similar to the linear case, its purpose is to help meet the last two of the Hadamard postulates. $\Psi(x)$ can either be a least squares based functional, similar to the ones described in the linear case or it can have alternative forms. For example see the diffuse optical tomography (DOT) relevant regularization functionals introduced in Sec. 3.6.

2.6 optimization

2.6.1 Non-gradient based optimization

Approaches towards the minimisation of Eq. 2.40 can be classified as those which utilize the gradient of $\mathcal{E}(x)$ and those which do not. An example of a non-gradient optimization scheme is based on evolution strategies, for example differential evolution [Storn and Price, 1997]. These are population based, stochastic global function minimisers, which track the simultaneous evolution - or equivalently improvement - of multiple initial estimates. The fitness of each evolved estimate is constantly tracked and the fitter survive. These methods have better chances in identifying global optima than their gradient based analogues, which update a single parametrization, due to the large number of initialization states. For example, the method of differential evolution - termed as a global optimization method - the suggested number of simultaneously tracked estimates is $10N$ with N being the number of unknowns [Storn and Price, 1997]. Consequently, in order for such schemes to be computationally tractable, they require fast evaluations of $\mathcal{E}(x)$ in order to maintain a low overall computational cost.

Another non-gradient based method is the downhill simplex method by Nelder and Mead [1965]. According to Press et al. [1992b], 'a simplex is the geometrical figure consisting, in N dimensions, of $N + 1$ points (or vertices) and all their interconnecting line segments, polygonal faces, etc. In two dimensions, a simplex is a triangle.' The method is initialized by defining an initial simplex. This is done by choosing its $N + 1$ points. The method now proceeds by evaluating the objective function at all $N + 1$ points of the simplex. Once the evaluation has taken place, the method continues by identifying the point with the highest objective function value and moves it to the opposite face - via a reflection through the centroid of the remaining N points. If the new point corresponds to an improved estimate, the simplex is additionally stretched across the direction of the initial reflection. If the objective function does not return a lower at the new point, the simplex is contracted. These moves allow the simplex to move throughout the multi-dimensional solution space, as well as change its shape, in order to move towards the minimum and eventually bracket it. The latter takes place when the whole simplex moves to some basin of attraction and starts contracting until its points reach the bottom of the basin. The method terminates when all the simplex points are in close proximity, according to some threshold criteria. The final estimate is a function of all the simplex points, for example the the simplex's centroid.

2.6.2 Gradient based optimization: First order methods

Optimization methods can utilize gradient information of various orders, provided that the objective function is sufficiently smooth and differentiable up to the required order. They proceed in an iterative fashion, where an initial solution estimate $x^{(0)}$ is sequentially improved until some convergence criteria are satisfied.

2.6.2.1 Gradient descent

The elementary optimization method utilizing the gradient of $\mathcal{E}(x)$ is the steepest descent method [Press et al., 1992b; Shewchuk, 1994]. Given some solution estimate $x^{(k)}$, then next update $x^{(k+1)}$ is obtained by taking a step of size λ across a line passing from $x^{(k)}$ and pointing to the direction where $\mathcal{E}(x^{(k)})$ decreases most rapidly. This direction is given by the negative gradient of Eq. 2.50,

$$d^{(k)} = -\,\partial\mathcal{E}(x^{(k)})/\partial x \tag{2.41}$$

$$= -2\left(\sum_i \left(y_i - \mathcal{F}_i(x^{(k)})\right)\right) J^{(k)} \tag{2.42}$$

where $J^{(k)} = \partial\mathcal{F}(x^{(k)})/\partial x$ is the *Jacobian*. The solution update is then expressed as

$$x^{(k+1)} = x^{(k)} + \lambda^{(k)} d^{(k)} \tag{2.43}$$

where step $\lambda^{(k)}$ denotes the length traversed over $d^{(k)}$ and can be computed by a line-search approach (see 2.6.4). Starting from $x^{(k)}$ and assuming that the minimisation across $d^{(k)}$ is accurate, then the improved solution $x^{(k+1)}$ must reside on a new minimum with respect to λ. Hence $\partial\mathcal{E}(x^{(k+1)})/\partial\lambda = 0$. This implies that if the minimum is not global and an another direction $d^{(k+1)}$ of decrease exists at $x^{(k+1)}$, it should have a zero directional component across $x^{(k)}$. This can only happen if $d^{(k+1)} \perp d^{(k)}$.

Briefly, in every iteration gradient descent searches for the next update at a direction orthogonal to the descent direction of the previous iteration. Consequently, throughout the entire minimisation process the method can revisit many times at the same direction. This can lead to slow convergence rates. Figure 2.3 graphically shows search directions obtained by the method of gradient descent.

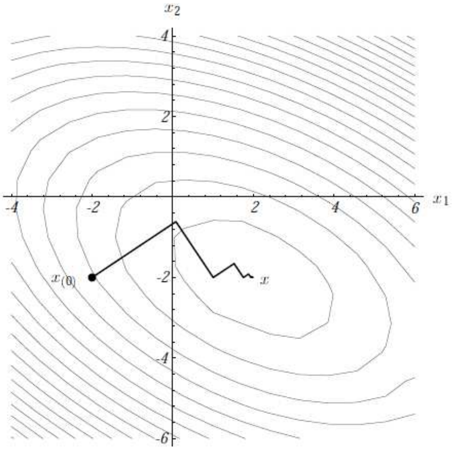


Figure 2.3: Successive descent directions produced by the gradient descent method. Courtesy of Jonathan Shewchuk. Source [Shewchuk, 1994].

2.6.2.2 Conjugate gradients

Conjugate gradients [Nocedal and Wright, 1999; Press et al., 1992b; Shewchuk, 1994] is an alternative optimization method which also utilizes first order derivative information. In Section 2.6.2.1 it was shown that gradient descent exhibits slow convergence as it can re-visit previously searched directions. Ideally one would seek to visit each direction only once. Consider again the solution space of Figure 2.3 and the first descent direction starting from $x^{(0)}$. Ideally, the size of the step across the depicted direction should be larger so the next orthogonal direction would lead directly to solution in the middle of the basin.

Unfortunately, the computation of such step requires the solution to be known *a priori*, in which there would be no reason to search for it in the first place [Shewchuk, 1994]. Conjugate gradients was motivated by the fundamental concept that descent directions should not be revisited more than once. It achieves this effect by enforcing directions to be A-orthogonal or *conjugate*, rather than orthogonal. In the linear case, A corresponds to the linear forward operator $\mathcal{F}$. The A-orthogonality between two vectors $d^{(j)}, d^{(k)}$ is satisfied when

$$\left(d^{(j)}\right)^{\mathrm{T}} A d^{(k)} = 0 \tag{2.44}$$

The geometrical meaning of A-orthogonality is the following. Any two A-orthogonal vectors in the solution space of $\mathcal{E}(x)$, would be orthogonal in the transformed solution space transformed by A. By ensuring such condition among descent directions, the method leads to convergence in a number of iterations equivalent to the dimensionality of the solution space [Shewchuk, 1994].

Figure 2.4 shows the successive improvement of the solution using the CG method. In a 2D space, the solution is obtained in two successive improvements - one for each direction. The depicted vectors are A-orthogonal

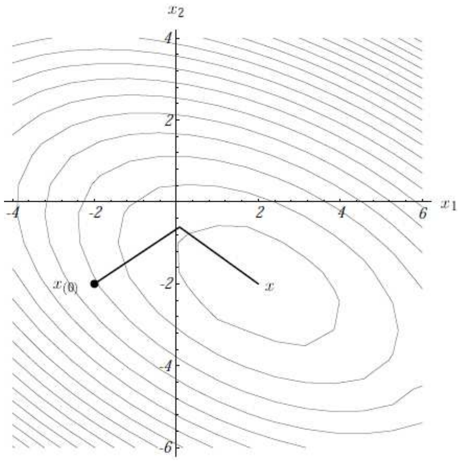


Figure 2.4: Successive descent directions produced by the CG method. Courtesy of Jonathan Shewchuk. Source [Shewchuk, 1994].

In the non-linear case first order CG method, the operator A is not explicitly required . The successive solution updates are produced according to $x^{(k+1)} = x^{(k)} + \lambda^{(k)} d^{(k)}$. The CG update direction for the first iteration is set to the negative gradient

$$d^{(0)} = g^{(0)}. \tag{2.45}$$

The updates at subsequent iterations need to establish conjugacy with all the previous directions. Conveniently, one does not need to explicitly store all previous directions in memory to achieve this condition. An elaborate update scheme can ensure conjugacy of a new search direction with all the previous ones, by using information solely from the last visited direction. The new directions are computed according to

$$d^{(k)} = -g^{(k)} + \beta^{(k)} d^{(k-1)}. \tag{2.46}$$

Proposed schemes for computing the term $\beta^{(k)}$ include the Fletcher-Reeves update [Fletcher and Reeves, 1964]

$$\beta^{(k)} = \frac{\left(g^{(k)}\right)^{\mathrm{T}} \cdot \left(g^{(k)} - g^{(k-1)}\right)}{\left(g^{(k-1)}\right)^{\mathrm{T}} \cdot g^{(k-1)}} \tag{2.47}$$

and the Polak-Ribière update [Polak and Ribière, 1969].

$$\beta^{(k)} = \frac{\left(g^{(k)}\right)^{\mathrm{T}} \cdot g^{(k)}}{\left(\left(g^{(k-1)}\right)^{\mathrm{T}} \cdot g^{(k-1)}\right)}. \tag{2.48}$$

It should be noted that the condition of conjugacy among the produced updates is degraded through out the minimisation process. To deal with this problematic behavior, the algorithm requires restarts which involve setting the update direction to the negative gradient $d^{(k)} = -\partial \mathcal{E}(x^{(k)})/\partial x$. These restarts can either be repeated every fixed number of iterations or whenever an explicit test of the level of conjugacy fails. Details about restarting and testing conjugacy can be found in [Shewchuk, 1994]. The first order CG minimisation method with fixed restarts is described in Algorithm 2.1.

As a final note, there is second order CG method where the analogue of the operator A of the linear CG, is explicitly formed as $\partial^2\mathcal{F}(x)/\partial x$ - known as the *Hessian*. As the Hessian changes throughout the minimisation this results in loss of conjugacy between successive update directions and the restarting mechanism need to be applied. Details of the second order CG can be found in in [Shewchuk, 1994].

2.6.3 Gradient based optimization: Second order methods

Under the condition that $\mathcal{E}(x)$ is twice differentiable, then at each iteration k, $\mathcal{E}(x^{(k)})$ can be expanded in the vicinity of $x^{(k)}$ by utilizing the Taylor series expansion [Kastanis, 2007]

$$\mathcal{E}(x) = \sum_{n=0}^{\infty} \frac{1}{n!} \frac{\partial^n \mathcal{E}(x^{(k)})}{\partial x^n} (x - x^{(k)})^n. \tag{2.49}$$

Consider the second order approximation, obtained by setting $n = 2$ and given by

$$\tilde{\mathcal{E}}(x) = \mathcal{E}(x^{(k)}) + \left(\frac{\partial \mathcal{E}(x^{(k)})}{\partial x} + \frac{1}{2} \left(x - x^{(k)}\right)^{\mathrm{T}} \frac{\partial^2 \mathcal{E}(x^{(k)})}{\partial x^2} \right) (x - x^{(k)}) \tag{2.50}$$

By definition, the minimizer $x^{(k+1)}$ of Eq. 2.50 is a stationary point with zero gradient

$$\frac{\partial \tilde{\mathcal{E}}(x^{(k+1)})}{\partial x} = \frac{\partial \mathcal{E}(x^{(k)})}{\partial x} + \left(x^{(k+1)} - x^{(k)}\right)^{\mathrm{T}} \frac{\partial^2 \mathcal{E}(x^{(k)})}{\partial x^2} = 0 \tag{2.51}$$

Under the assumption that the inverse of the Hessian $\left(\partial^2 \mathcal{E}(x^{(k)})/\partial x^2\right)^{-1}$ exists, Eq. 2.51 after rearranging leads to

$$x^{(k+1)} = x^{(k)} - \left(\frac{\partial^2 \mathcal{E}(x^{(k)})}{\partial x^2} \right)^{-1} \times \frac{\partial \mathcal{E}(x^{(k)})}{\partial x} \tag{2.52}$$

The gradient of the objective function used in Eq. 2.52 is given by

$$\frac{\partial \mathcal{E}(x^{(k)})}{\partial x} = -2 \left(\sum_i \left(y_i - \mathcal{F}_i(x^{(k)}) \right) \right) \frac{\partial \mathcal{F}(x^{(k)})}{\partial x} + \tau \frac{\partial \Psi(x^{(k)})}{\partial x} \tag{2.53}$$

```
Set thresholds: $\epsilon_{global}, \epsilon_{iter}, iterMax > 0$
Set initial line-search step length: $\lambda^{(0)} > 0$
Set iteration counter: $k \leftarrow 0$
Set number of iterations before restart: $N_k$
Compute: $\mathcal{E}(x^{(0)})$
Compute gradient: $g^{(0)} = \partial\mathcal{E}(x^{(0)})/\partial x^{(0)}$
Define: $\mathcal{E}(x^{(-1)}) = -\infty$
while $\mathcal{E}(x^{(k)}) \geq \epsilon_{global}$ && $\left(\mathcal{E}(x^{(k)}) - \mathcal{E}(x^{(k-1)})\right) \geq \epsilon_{iter}$ && $k \leq iterMax$ do
    if $k == 0 \,||\, N_k$ iterations elapsed since last restart then
        $p^{(0)} = -g^{(0)}$
    else
        $\beta^{(k)} = \begin{cases} \left(\left(g^{(k)}\right)^{\mathrm{T}} \cdot \left(g^{(k)} - g^{(k-1)}\right)\right) \Big/ \left(\left(g^{(k-1)}\right)^{\mathrm{T}} \cdot g^{(k-1)}\right) & \text{Fletcher-Reeves OR} \\ \left(\left(g^{(k)}\right)^{\mathrm{T}} \cdot g^{(k)}\right) \Big/ \left(\left(g^{(k-1)}\right)^{\mathrm{T}} g^{(k-1)}\right) & \text{Polak-Ribière update} \end{cases}$
        $p^{(k)} = -g^{(k)} + \beta^{(k)} p^{(k-1)}$
    end
    Perform line search for $x^{(k)}$ and compute step-size: $\lambda^{(k)}$
    Parameter update: $x^{(k+1)} = x^{(k)} + \lambda^{(k)} p^{(k)}$.
    Increment counter: $k \leftarrow k + 1$
end
```

Algorithm 2.1: The non-linear Conjugate Gradients algorithm.

The Hessian term $\partial^2\mathcal{E}(x^{(k)})/\partial x^2$ is given by

$$\frac{\partial^2\mathcal{E}(x^{(k)})}{\partial x^2} = -2\left(\sum_i \left(y_i - \mathcal{F}_i(x^{(k)})\right)\right)\frac{\partial^2\mathcal{F}(x^{(k)})}{\partial x^2} + 2\left\{J^{(k)}\right\}^{\mathrm{T}} J^{(k)} + \tau\frac{\partial^2\Psi(x^{(k)})}{\partial x^2} \tag{2.54}$$

By setting $S^{(k)} = \sum_i \left(y_i - \mathcal{F}_i(x^{(k)})\right) J^{(k)}$, the Newton-Raphson update is then expressed as

$$x^{(k+1)} = x^{(k)} + \lambda^{(k)}\left(S^{(k)} + \left\{J^{(k)}\right\}^{\mathrm{T}} J^{(k)} + \tau\frac{\partial^2\Psi(x^{(k)})}{\partial x^2}\right)^{-1} \times \tag{2.55}$$

$$\left(S^{(k)} - \tau\frac{\partial\Psi(x^{(k)})}{\partial x}\right) \tag{2.56}$$

The term $S^{(k)}$ is computationally costly to compute and can also lead to loss of the positive-definiteness of the Hessian [Press et al., 1992b; Schweiger et al., 2005]. By ignoring this term this leads to the *damped* Gauss-Newton update [Nocedal and Wright, 1999; Schweiger et al., 2005]

$$x^{(k+1)} = x^{(k)} + \lambda^{(k)} \left(\left\{ J^{(k)} \right\}^{\mathrm{T}} J^{(k)} + \tau \frac{\partial^2 \Psi(x^{(k)})}{\partial x^2} \right)^{-1} \times \tag{2.57}$$

$$\left(S^{(k)} - \tau \frac{\partial \Psi(x^{(k)})}{\partial x} \right) \tag{2.58}$$

Another alternative is to replace $S^{(k)}$ with $c^{(k)}I$, where $c^{(k)} > 0$ is a control parameter and I the identity matrix. This leads to the Levenberg-Marquardt update [Levenberg, 1944; Marquardt, 1963]

$$x^{(k+1)} = x^{(k)} + \left(\left\{ J^{(k)} \right\}^{\mathrm{T}} J^{(k)} + \tau \frac{\partial^2 \Psi(x^{(k)})}{\partial x^2} + c^{(k)} I \right)^{-1} \times \tag{2.59}$$

$$\left(S^{(k)} - \tau \frac{\partial \Psi(x^{(k)})}{\partial x} \right) \tag{2.60}$$

Notice that $\lambda^{(k)}$ is not included in this update. The magnitude of the update is controlled by $c^{(k)}$. The Levenberg-Marquardt belongs to a category of methods known as *trust-region* methods. For $c^{(k)} >> 0$ the method behaves as gradient descent whereas for $c^{(k)} = 0$ it behaves as the Gauss-Newton scheme.

There is another category of methods, termed *quasi-Newton* methods which do not require the explicit computation of the Hessian matrix. These methods successively build approximations of the Hessian matrix or its inverse [Press et al., 1992b] using the gradient of the objective function. Effectively they utilize first-order derivative information however - close to the minimum - these methods enjoy the quadratic convergence rates of the Newton method. One such method is known as Broyden-Fletcher-Goldfarb-Shanno (BFGS). A limited memory version known as limited memory BFGS (L-BFGS), is used for the optimization of problems with large number of variables, where the approximation of the Hessian is never realized in full, hence there is no need to store it. This method is employed for non-rigid registration in Chapter 8.

2.6.4 Line search

A line search describes a minimisation across a line. The term usually refers to a scheme employed by gradient based optimization schemes in order to estimate the length of $\lambda^{(k)}$ which needs to be traversed in the update direction at every iteration. The estimated step is the minimizer [Nocedal and Wright, 1999]

$$\lambda^{(k)} = \underset{\lambda > 0}{\arg\min} \left[\Xi(\lambda) = \mathcal{E}(x^{(k)}) + \lambda d^{(k)} \right]. \tag{2.61}$$

There are various approaches which produce estimates of $\lambda^{(k)}$. One attempt to find the exact minimizer of the above function, however this can prove to be an expensive process given that it has to be repeated at every iteration. Many approaches compromise on the accuracy of the step length and employ an *inexact* line search which retrieves an approximation of the exact minimizer $\lambda^{(k)}$. For a review of this methods the reader is directed to [Bazaraa et al., 1993; Nocedal and Wright, 1999; Press et al., 1992b].

One such method which is employed for the purposes of this work is described here which utilizes an *inverse quadratic interpolation*[1] [Press et al., 1992b]. It uses multiple evaluations of the global

[1] it is called inverse because ultimately it retrieves the abscissa and not the ordinate [Press et al., 1992b]

objective $\mathcal{E}(x^{(k)})$ for different λ sizes, but no gradient information.

Briefly, let $\mathcal{E}(x^{(k)})$ be the error at iteration k. Having computed the update direction $d^{(k)}$ for k, the step λ is set to some initial value and the potential update $x_\lambda = x^{(k)} + \lambda d^{(k)}$ for the given λ is computed. If $\Xi(\lambda) < \mathcal{E}(x^{(k)})$ in Eq. 2.61, the method doubles the step and re-evaluates the objective. At all times, the method stores in memory the three *more recent* successive step estimates $\lambda_\alpha < \lambda_m < \lambda_\beta$ as well as their corresponding objective errors $\Xi(\lambda_\alpha),\ \Xi(\lambda_m)$ and $\Xi(\lambda_\beta)$. The process keeps on producing new $\lambda_\beta = 2\lambda_\beta$ and updating $\lambda_\alpha, \lambda_m$ to be its immediate predecessors until an increase in the error occurs or $\Xi(\lambda_\beta) > \Xi(\lambda_m)$. The method assumes that the solution space forms a quadratic basin of attraction in the neighbourhood of the true step. It thus uses an inverse quadratic interpolation by fitting a quadratic $Q(\lambda) = c_1\lambda^2 + c_2\lambda + c_3$ to the three $\lambda_\alpha,\ \lambda_m$ and λ_β. These produces three equations from which c_1, c_2 and c_3 are computed. The minimum of the quadratic, corresponding to the final step estimate is then $\lambda^{(k)} = -\frac{c_2}{2c_1}$. The quadratic interpolation is graphically shown in Figure 2.5.

A similar process takes place if the error corresponding to the initialization step is higher than the initial error. In that case, the steps are not doubled but halved. The full algorithm is described in Algorithm 2.2.

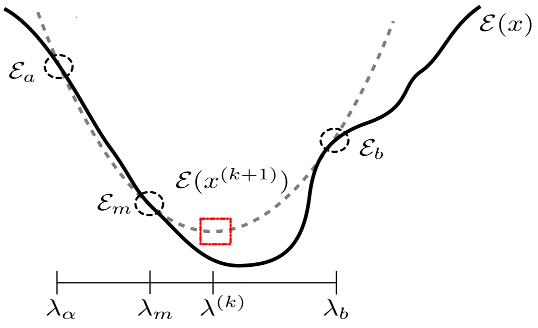


Figure 2.5: Inverse quadratic interpolation computed on successive line search λ estimates. The final estimate $\lambda^{(k)}$ is located at the minimum of the quadratic curve fitted on the λ_α, λ_β estimates bracketing the minimum, in the interim λ_m.

2.7 Summary

We have presented a brief introduction to the basic concepts in inverse problem theory. Specifically, we have outlined the three Hadamard postulates, whose breach constitutes a case of ill-posedness manifestation in an inverse problem. The case of the linear inverse problem was described as it explicitly reveals the effects of ill-posedness. In addition, we introduced the concept of regularization which operates toward the alleviation of these problematic characteristics which plague an inverse problem and its effects can be explicitly observed in the linear case. We have also formulated the case for the non-linear inverse problem and the generalized Tikhonov regularisation scheme. The non-linear inverse problem is ultimately formulated as an optimization problem. A sample of optimization schemes was briefly discussed.

```
Set threshold: $\epsilon_{search} > 0$
Set initial step sizes: $\lambda_\alpha \leftarrow 0$, $\lambda_\beta \leftarrow \lambda^{(k-1)},$ where $\lambda^{(k-1)}$ is the step used in iteration
    $k-1$. If $k=1$ then $\lambda_\beta \leftarrow 1$
Compute objective function updates: $\mathcal{E}_\alpha = \mathcal{E}(x^{(k)} + \lambda_\alpha d^{(k)})$ and $\mathcal{E}_\beta = \mathcal{E}(x^{(k)} + \lambda_\beta d^{(k)})$
Bracket the minimum: if $\mathcal{E}_\beta > \mathcal{E}_\alpha$ then
    Set: $\lambda_m \leftarrow \lambda_\beta/2$ and compute $\mathcal{E}_m = \mathcal{E}(x^{(k)} + \lambda_m d^{(k)}$
    while $(\mathcal{E}_m - \mathcal{E}_\alpha) > \epsilon_{search}$ do
        $\lambda_m \leftarrow \lambda_m/2$, $\lambda_\beta \leftarrow \lambda_\beta/2,$ compute $\mathcal{E}_m$
    end
else
    Set: $\lambda_m \leftarrow \lambda_\beta,\ \lambda_\beta \leftarrow 2\lambda_\beta$ and compute $\mathcal{E}_m$
    while $(\mathcal{E}_m - \mathcal{E}_\beta) > \epsilon_{search}$ do
        $\lambda_\alpha \leftarrow \lambda_m$, $\lambda_m \leftarrow \lambda_\beta$, $\lambda_\beta \leftarrow 2\lambda_\beta,$ compute $\mathcal{E}_m$
    end
end
Obtain $\lambda^{(k)}$ via quadratic interpolation of $\lambda_\alpha,\ \lambda_m, \lambda_\beta$:
$\alpha \leftarrow \left(\frac{\mathcal{E}_\alpha - \mathcal{E}_\beta}{\lambda_\alpha - \lambda_\beta} - \frac{\mathcal{E}_\alpha - \mathcal{E}_m}{\lambda_\alpha - \lambda_m}\right)(\lambda_\beta - \lambda_m)^{-1}$
$c_2 = \frac{\mathcal{E}_\alpha - \mathcal{E}_\beta}{\lambda_\alpha - \lambda_\beta} - c_1(\lambda_\alpha + \lambda_\beta)$
$\lambda^{(k)} = -c_2/(2c_1)$
```

Algorithm 2.2: Pseudo code for inexact line-search based on inverse quadratic interpolation.

Chapter 3

Diffuse optical tomography

3.1 Introduction

Imaging with diffuse optical tomography (DOT) is a notoriously non-trivial problem. There are challenges in all aspects of its operation namely data acquisition, forward modelling and the solution of the underlying inverse problem which is severely ill-posed.

The structure of this chapter is as follows; Section 3.2 briefly introduces the fundamental optical quantities of interest. Section 3.3 touches the imaging schemes. The forward problem in DOT is discussed in Section 3.4. It refers to the radiative transfer equation (RTE) as a model of light propagation, provides physical insight and outlines the derivation of the diffusion approximation (DA) to the RTE which is used in DOT. Approaches towards the solution of the DA are also outlined, with emphasis given in the description of the finite element method (FEM), which is the method employed in this work. The formulation of the inverse problem in DOT and approaches towards its solution are outlined in Section 3.5. The discussion continuous in Sec. 3.6 which discusses various regularization schemes which have been utilized in DOT, with emphasis on multi-modality applications. Section 3.7 discusses the validity of using anatomical images as priors in functional medical imaging modalities. Having introduced the physics of DOT imaging and the inverse problem, the chapter finally concludes with Sec. 3.8 where we explicitly discuss the sources of ill-posedness in DOT as well as noise and resolution issues.

3.2 Fundamental optical quantities

As light traverses through a turbid medium, its propagation is affected by a number of physical events rising from the interaction of light with the anatomical structure. Two of these events are light absorption μ_a and scattering μ_s. Before the introduction of these concepts consider the definition of *radiance*.

3.2.1 Radiance

The radiance $I(r, \hat{s}, t)$ is a fundamental quantity in optics. It is defined in units of power, per area, per unit solid angle or $Wm^{-2}sr^{-1}$. Consider Fig. 3.1. Radiance is defined so that the amount of radiant power dP, of some specified frequency interval $\{\nu, \nu + \mathrm{d}\nu\}$, transported by photons passing through an elementary area da defined by its normal unit vector $\hat{s}_n$, at time instance t, towards a direction $\hat{s}$ and confined to an element of solid angle d$\hat{s}$ is given by [Ishimaru, 1978]

$$\mathrm{d}P = I(r, \hat{s}, t)\cos\vartheta \; \mathrm{d}\nu \; \mathrm{d}a \; \mathrm{d}\hat{s} \tag{3.1}$$

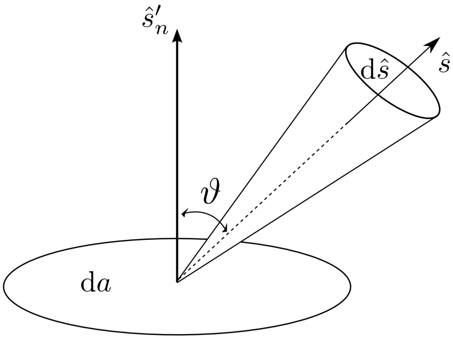


Figure 3.1: Radiance of light is reduced during its traversal of a non-scattering but homogeneously absorbing medium

3.2.2 Absorption

Absorption occurs when the interaction between matter and the incident electromagnetic radiation is such that this leads to partial or complete conversion of the latter to thermal energy [Bohren and Huffman, 1983; Hecht and Zajac, 2002].

Let an absorbing compound be dissolved in a non-scattering medium, resulting in a homogeneous solution. Given collimated radiation through such a medium, then the reduction dI in radiance (or intensity) I as light travels an infinitesimal distance dl - see Fig. 3.2 - depends on a material constant called the *absorption coefficient* μ_a (units mm^{-1})

$$\mathrm{d}I = -\mu_a I \mathrm{d}l \tag{3.2}$$

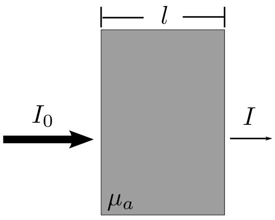


Figure 3.2: Radiance of light is reduced during its traversal of a non-scattering but homogeneously absorbing medium

Figure 3.2 graphically illustrates this effect. Given incident radiance I_0, the radiance after some distance l can be computed by rearranging Eq. 3.2 and integrating over l results in

$$I = I_0 e^{-\mu_a l} \tag{3.3}$$

Eq. 3.3 reveals an exponential attenuation of I with respect to the traversed distance. From a physical point of view, the reciprocal $1/\mu_a$ (mm) corresponds to the *mean free path* that a photon can travel without being subjected to an absorption event [Schmidt, 1999].

3.2.3 Scattering

Scattering occurs when the interaction between matter and the incident electromagnetic radiation is such that it leads to a change in the direction of propagation of the latter. DOT considers elastic scattering which dictates that the energy of the incident radiation - as well as its wavelength - is preserved resulting only in a change of direction [Hecht and Zajac, 2002].

Similar to the case of absorption, assume a non-absorbing medium of length l which is trans-illuminated by incident light of intensity I_0. The relation between input and output intensities I_0 and I depends on the *scattering coefficient* μ_s (mm^{-1}):

$$I = I_0 e^{-\mu_s l} \tag{3.4}$$

Figure 3.3 graphically illustrates this effect. The reciprocal $1/\mu_s$ $(mm^{-1})$is the mean free path which a photon can travel without being subjected to a scattering event.

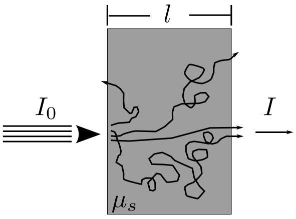


Figure 3.3: Radiance of light is reduced during its traversal of a non-absorbing but homogeneously scattering medium

The notions of directionality and angles are fundamental in the description of scattering phenomena. The normalized *phase function* $\Theta(\hat{s}, \hat{s}\,')$ is interpreted as the probability of a photon initially traveling at $\hat{s}$ being scattered into a direction $\hat{s}\,'$, hence it holds [1]

$$\int_{4\pi} \Theta(\hat{s}\,', \hat{s}) \mathrm{d}\hat{s} = 1. \tag{3.5}$$

Assuming that the scattering is axially symmetric relative to the initial direction of propagation, then $\Theta(\hat{s}\,', \hat{s})$ only depends on the scattering angle ϑ between $\hat{s}\,', \hat{s}$. Effectively, ϑ is the polar angle considering a spherical coordinate system, for which it holds $u = \cos(\vartheta) = \hat{s}\,' \cdot \hat{s}$, with the latter denoting the *dot*

[1] see Sec. 4.2 for an introduction to probabilities

product. The dependence of the phase function to ϑ can be explicitly denoted by $\Theta(\hat{s}\,', \hat{s}) = \Theta(\hat{s}\,' \cdot \hat{s})$. The scattering characteristic of the medium is described by the mean cosine of the scattering angle defined as

$$\bar{\Theta} = \int_{-1}^{1} u\Theta(u)\mathrm{d}u. \tag{3.6}$$

where $\bar{\Theta} = 0$ corresponds to a uniform angular distribution of the scattering angle (isotropic medium), $\bar{\Theta} > 0$ corresponds to forward scattering $(\vartheta < 90°)$ and $\bar{\Theta} < 0$ to backward scattering $(\vartheta > 90°)$[Schweiger, 1994].

3.2.4 Index of refraction

The refractive index $\mathfrak{r}_1$ of a medium Ω_1 describes the reduction in the speed c_{Ω_1} that light propagates within it, with respect to the speed c of propagation in a vacuum.

$$\mathfrak{r}_1 = \frac{c}{c_{\Omega_1}} \tag{3.7}$$

3.3 Imaging schemes

DOT is a non-invasive modality, thus the imaging process probes the medium from a configuration of sources/detector fibres placed on its surface. The end tips of the fibres are attached directly to the surface of the anatomy, for example see [Hebden et al., 2004; Hillman et al., 2001] or via a fluid-coupled patient interface, for example [Enfield et al., 2007]. The spatial arrangement and the number of sources/detectors on the surface both play a pivotal role in the spatial and quantitative accuracy of the reconstructed images[Pogue et al., 1999a]. Figure 3.4 graphically illustrates a test medium with its spatial domain denoted by Ω and its surface with $\partial\Omega$.

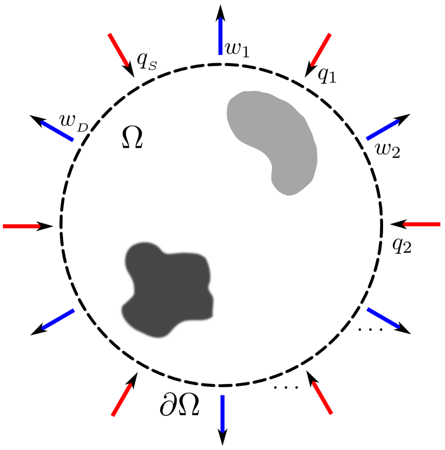


Figure 3.4: The imaging process in DOT involves non-invasive imaging by a configuration of sources/detectors placed on the surface $\partial\Omega$ of the probed domain Ω.

Irrespective of the geometry, imaging with DOT can be classified to three distinct schemes, *(i)* continuous wave (CW) *(ii)* time domain (TD) and *(iii)* frequency domain. CW systems involve the trans-illumination of the medium by a source emitting light of constant intensity. By measuring the exiting light, the level of attenuation of the incident light is obtained. In the TD case, the medium is irradiated by ultra short picoseconds pulses of light. The photons comprising each pulse follow different paths across the medium and exit the medium at different times. Photon counting detectors record for a short interval the individual photons exiting the medium and measure their flight times relative to a reference pulse [Hillman et al., 2000; Schmidt et al., 2000]. This results in a build up of a histogram of photon flight times for each source-detector pair, known as the temporal point spread function (TPSF) extending over several nano-seconds [Delpy et al., 1988]. UCL has developed a time resolved optical tomography system known as MONSTIR, employing 32 source fibers and 32 detector optodes [Hebden et al., 1999; Schmidt et al., 2000]. Finally frequency domain systems involve amplitude modulated sources, usually in the order of few hundred MHz. Measurements of the reduction in amplitude and the phase shift are obtained at the detector locations [Gibson et al., 2005a]. For more information about recent developments in imaging systems please see Gibson et al. [2005a] and the references within. Figure 3.5 graphically illustrates the input/output of the three imaging schemes.

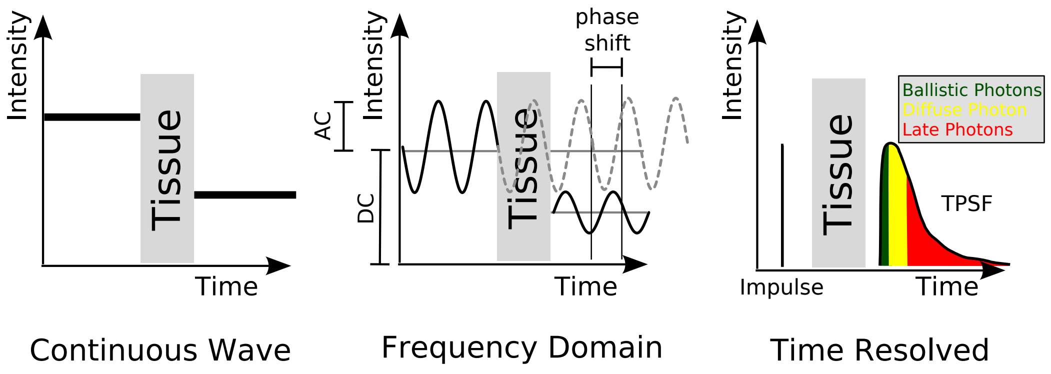


Figure 3.5: Graphical illustration of imaging schemes in optical tomography. CW systems measure the difference between incident/measurable light intensity. Frequency domain systems measure the phase shift and amplitude reduction as a response to light of modulated amplitude. Finally TD systems emit pico-second pulses and build a TPSF. Each part of the TPSF corresponds to the probability of photons arriving at a specific time.

3.4 Forward problem in DOT

Chapter 2 defined the forward problem as a mapping from the space $\mathbb{X}$ of the parameters x describing the physical system of interest, to the space $\mathbb{Y}$ of the measurable quantities y. Let $\Omega \subset \mathbb{R}^n$ be a simply connected domain denoting the medium to be probed with DOT, with $\partial\Omega$ explicitly denoting its boundary and $r \in \Omega$ arbitrary spatial locations. In the context of optical tomography, $\mathbb{X}$ consists of two *independent* spaces or $\mathbb{X} = \left\{\Xi^{(a)}(\Omega), \Xi^{(b)}(\Omega)\right\}$. The parameters of interest are then expressed as a pair of continuous scalar functions $\big(a(r), b(r)\big) \in \mathbb{X}$. The physical quantity represented by $a(r)$ and

$b(r)$ depends on the employed forward model $\mathcal{F}$ which models the propagation of light. In the case of this model being the RTE, the two quantities are $\big(\mu_a(r), \mu_s(r)\big)$ denoting the light absorption and scattering, whereas in the case of the simpler DA employed by DOT, $\mu_s(r)$ is replaced with the diffusion coefficient $\kappa(r)$ introduced in Sec. 3.4.2.

The forward problem regards the modelling of the imaging process by computational means. It commences by simulating the trans-illumination of the medium from N_q given sources $q(\boldsymbol{m}), \boldsymbol{m} \in \partial\Omega$. Light enters Ω, interacts with $\big(a, b\big)$ - where the latter are assigned with known values - and propagates according to the physical laws dictated by the forward model. The portion of the incident radiation which is not converted to thermal energy due to absorption events exits the medium. A simulated measurement process captures the exiting light at N_w detectors $w(\boldsymbol{m})$, from which the measurable quantities y are derived. Effectively, the forward problem solves the equation

$$y_s(\boldsymbol{m}) = \mathcal{F}_s(a(r), b(r)) \tag{3.8}$$

where s denotes the index of some specific source.

3.4.1 Radiative Transfer Equation

The RTE is a physical model which describes the transport of energy in random media comprised by particles which can absorb, emit and scatter radiation [Chandrasekhar, 1950; Ishimaru, 1978]. It does not model any phenomena described by the wave nature of light such as interference or diffraction and it requires a selection of wavelengths that are smaller than the objects under study [Gibson et al., 2005a; Ishimaru, 1978]. The standard derivation assumes that the refractive index is constant within the medium, although modifications allowing the latter to spatially vary have been proposed, for example see [Ferwerda, 1999; Martí-López et al., 2003]. The RTE is widely considered as an adequate model for light propagation within the tissue [Arridge, 1999].

Let $\Omega \subseteq \mathbb{R}^3$ be the domain of the medium, a position vector $r \in \Omega$ and time variable t. Considering the flow of light at r, then the equation of transfer is expressed as

$$\begin{aligned}&\left(\frac{1}{c}\frac{\partial}{\partial t} + \ \hat{s}\cdot\nabla + \mu_a(r) + \mu_s(r)\right) I(r, \hat{s}, t)\\ &= \mu_s(r)\int_{4\pi} \Theta(\hat{s}\,', \hat{s}) I(r, \hat{s}\,', t)\mathrm{d}\hat{s}\,' + q(r, \hat{s}, t)\end{aligned} \tag{3.9}$$

where $I(r, \hat{s}, t)$ is the *radiance* at position r, towards a direction defined by the unit vector $\hat{s}$ and at time t; $\Theta(\hat{s}\,', \hat{s})$ is the normalized scattering phase function; $q(r, \hat{s}, t)$ denotes a source term located at r and emitting radiation in direction $\hat{s}$; $\cdot\nabla$ is the divergence operator; c is the speed of light propagation in the medium and finally μ_a and μ_s are the absorption and scattering coefficients. [Boas, 1996].

The RTE can be interpreted as an energy conservation equation. Consider an elementary volume centered at r such as the one depicted in Fig. 3.6. For time instance t, the term $\big(\mu_a(r) + \mu_s(r)\big)I(r, \hat{s}, t)$ in the LHS of the RTE describes the loss of energy in $\hat{s}$, due to absorption and scattering events taking place in r. The RHS of the RTE corresponds to the gain in energy in $\hat{s}\,'$. The first term denotes the portion of light scattered in $\hat{s}\,'$, from the total light incident in r from all $\hat{s}$. The second term describes gains from emission within the elementary volume in r, if the latter encloses a source term.

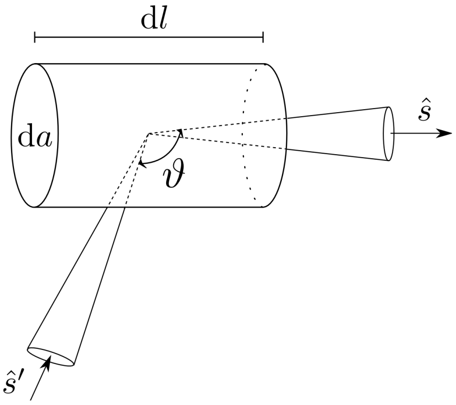


Figure 3.6: Light incident on an elementary volume. Part of the incident light can be absorbed and part of the light can be scattered. In the case that the volume encloses a source, emission can also occur

Finally, three quantities of interest derived from I include

$$\text{photon density}: \quad \Phi(r,t) = \int_{4\pi} I(r,\hat{s},t)\mathrm{d}\hat{s} \tag{3.10}$$

$$\text{photon current}: \quad \boldsymbol{J}(r,t) = \int_{4\pi} \hat{s}I(r,\hat{s},t)\mathrm{d}\hat{s} \tag{3.11}$$

$$\text{exitance}: \quad \Gamma(r,t) = \int_{4\pi} \hat{s}I(r,\hat{s},t)\cdot\hat{s}_n\mathrm{d}\hat{s} \tag{3.12}$$

where *photon density* (or *fluence rate*) is the total radiant energy in course of transit at some unit volume in r due to the radiation from all angles. In contrast *photon current* (or *net flux*) defines the rate of flow of radiant energy across an elementary volume [Chandrasekhar, 1950]. *Exitance* describes the energy transfer per unit time, through a unit area defined by its normal $\hat{s}_n$, integrated over the solid angle [Kolehmainen, 2001].

Analytical solutions to RTE are non-trivial to derive, except for very simple geometries. In the case of complex geometries, numerical methods have to be employed which require the discretisation of the involved quantities. One can easily see that due to their angular dependence, I, q and Θ are effectively *spherical functions* mapping spherical coordinates (ϑ, ϕ) to some scalar value. A natural way to represent functions of this type, is via a spherical harmonics basis expansion. This enables their approximation as a linear combination of a finite number of basis functions and coefficients.

$$I(r,\hat{s},t) = \sum_{l=0}^{\infty}\sum_{m=-l}^{l}\left(\frac{2l+1}{4\pi}\right)^{1/2} I_{l,m}(r,t)\ Y_{l,m}(\hat{s}) \tag{3.13}$$

$$q(r,\hat{s},t) = \sum_{l=0}^{\infty}\sum_{m=-l}^{l}\left(\frac{2l+1}{4\pi}\right)^{1/2} q_{l,m}(r,t)\ Y_{l,m}(\hat{s}) \tag{3.14}$$

$$\begin{aligned}\Theta(\hat{s}\,',\hat{s}) &= \sum_{l=0}^{\infty}\sum_{m=-l}^{l}\left(\frac{2l+1}{4\pi}\right)^{1/2} \Theta_l\ P_l(\cos\theta) \\ &= \sum_{l}^{\infty}\sum_{m=-l}^{l} \Theta_l\ Y^*_{l,m}(\hat{s}\,')Y_{l,m}(\hat{s})\end{aligned} \tag{3.15}$$

where $Y_{l,m}$ are the spherical harmonic basis functions; $I_{l,m}, q_{l,m}$ and Θ_l are the basis coefficients and Y^* denotes the complex conjugate. The approximations of the RTE using spherical harmonic expansion of the three quantities are known as P_N approximations and are obtained by assuming that $I_{l,m} = 0$ for all $l > N$ [Arridge, 1999]. Apparently the accuracy of the approximation increases with l.

Due to difficulties in computing numerical solutions for the RTE in complex geometries and for arbitrary parameter distributions, most of the current approaches in optical imaging using the DA to model the light propagation inside a diffusive medium, giving rise to DOT.

3.4.2 Diffusion approximation to the Radiative Transfer Equation

The DA is an approximation of the RTE. In diffusion theory, the radiance (or diffuse radiance in this case) is assumed to encounter many particles and undergo a high number of scattering events. This results in an *almost* uniform phase function $\Theta(\hat{s}\,',\hat{s})$. However a slight directionality in the propagation should still exist, otherwise for a completely uniform phase function, the net diffuse flux would be zero [Ishimaru, 1978]. Derivations of the DA can be found in [Arridge, 1999; Ishimaru, 1978]. Briefly, consider the spherical harmonic expansion of the previous section. The DA is a special case of the P_1 approximation of the RTE [Tarvainen et al., 2008]. The P_1 approximation of the RTE is obtained by setting the terms $I_{l,m}(r,t), q_{l,m}(r,t)$ and Θ_l in Eqs. 3.13-3.15 to zero, for all $|l| > 1$. After some lengthy computations and assuming that $\Theta_0 = 1$ [Arridge, 1999], we arrive at the approximations [Ishimaru, 1978; Kolehmainen, 2001; Schweiger, 1994]

$$I(r,\hat{s},t) \approx \frac{1}{4\pi}\Phi(r,t) + \frac{3}{4\pi}\hat{s}\cdot\boldsymbol{J}(r,t) \quad \text{and} \tag{3.16}$$

$$q(r,\hat{s},t) \approx \frac{1}{4\pi}q_0(r,t) + \frac{3}{4\pi}\hat{s}\cdot q_1(r,t), \tag{3.17}$$

where $\Phi(r,t)$, $\boldsymbol{J}(r,t)$ denote the diffuse photon density and diffuse photon current and are given by Eqs. 3.10-3.11; q_0 and q_1 are the isotropic and anisotropic components of the source term expressed respectively as

$$q_0(r,t) = \int_{4\pi} q(r,t,\hat{s})\mathrm{d}\hat{s} \tag{3.18}$$

and

$$q_1(r,t) = \int_{4\pi} \hat{s} \cdot q(r,t,\hat{s}) \mathrm{d}\hat{s}. \tag{3.19}$$

Inserting Eqs. 3.16 and 3.17 in the RTE (Eq. 3.9) one arrives at the pair of equations

$$\left(\frac{1}{c}\frac{\partial}{\partial t} + \mu_a(r)\right) \Phi(r,t) + \nabla \cdot \boldsymbol{J}(r,t) = q_0(r,t) \quad \text{and} \tag{3.20}$$

$$\left(\frac{1}{c}\frac{\partial}{\partial t} + \mu_a(r) + \mu_s'(r)\right) \boldsymbol{J}(r,t) + \frac{1}{3}\nabla I(r,t) = q_1(r,t). \tag{3.21}$$

where $\Phi(r,t)$ and $\mu_s'(r) = (1-\Theta_1)\mu_s$ is the *reduced scattering coefficient* and $\Theta_1 = \bar{\Theta}$ is the mean scattering phase function defined in Eq. 3.6.

As mentioned in the start of the section, DA is a special case of the $P1$ approximation which led to the pair of Equations 3.20 and 3.21. It is special as it should satisfy the conditions [Arridge, 1999]

$$q_1(r,t) = 0, \quad \frac{\partial \boldsymbol{J}(r,t)}{\partial t} = 0. \tag{3.22}$$

The first condition states that the anisotropic component of the source term $q(r,t)$ is not considered. The second condition implies that the photon current does not change, which is clearly erroneous in the time dependent case as it would require the light propagation to be instantaneous within the medium. The condition however can be justified by specifying that $\mu_a \ll \mu_s'$ inside the medium. [Arridge, 1999; Hillman, 2002]. Then Eq. 3.21 is simplified into Fick's Law [Arridge, 1999]

$$\boldsymbol{J}(r,t) = -\kappa(r)\nabla\Phi(r,t) \tag{3.23}$$

where

$$k(r) = \frac{1}{3\big(\mu_a(r) + \mu_s'(r)\big)} \tag{3.24}$$

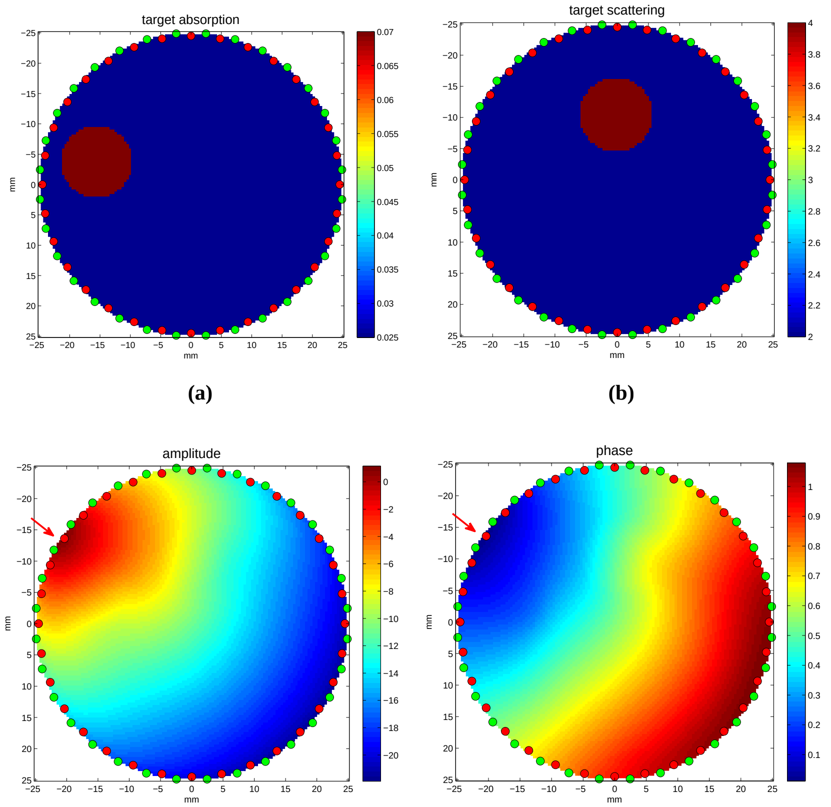
is the *diffusion coefficient*. The DA in the time domain is obtained by inserting Eq. 3.23 in Eq. 3.20 leading to

$$-\nabla \cdot \big(\kappa(r)\nabla\Phi(r,t)\big) + \mu_a(r)\Phi(r,t) + \frac{1}{c}\frac{\partial\Phi(r,t)}{\partial t} = q_0(r,t), \quad r \in \Omega. \tag{3.25}$$

The frequency domain case is obtained via the Fourier Transform of Eq. 3.25 giving rise to

$$-\nabla \cdot (\kappa(r)\nabla\ddot{\Phi}(r,\omega)) + \mu_a(r)\ddot{\Phi}(r,\omega) + \frac{i\omega}{c}\ddot{\Phi}(r,\omega) = \ddot{q}_0(r,\omega), \quad r \in \Omega. \tag{3.26}$$

where $\ddot{\Phi}$ denotes the complex photon density and $\ddot{q}_0$ is a source at r from which an amplitude modulated input signal of frequency $\nu = \omega/2\pi$ is introduced [Schweiger and Arridge, 1997] and expressed using complex notation for amplitude and phase. Figure 3.7 shows the complex photon density fields obtained from the trans-illumination of the circular Ω, from an amplitude modulated source (100MHz), populated by the depicted optical distribution of $\big(\mu_a(r), \mu_s'(r)\big)$.



(a) **(b)**

(c) **(d)**

Figure 3.7: Visualization of the photon density fields (amplitude and phase) for the frequency domain case, for a given optical parameter distribution. (a)-(b): μ_a/μ_s' distributions. The optical values for background | perturbation are: μ_a = 0.025 | 0.07 mm^{-1} and μ_s = 2 | 4 mm^{-1}. Sources and detectors are shown with green and red colours respectively. (c)-(d): For all $r \in \Omega$ the fields represent the $\log$-amplitude and phase difference with respect to the incident radiation.

3.4.2.1 Validity of the Diffusion Approximation

The approximation of the RTE by the DA requires that $\mu_a \ll \mu_s$ and the light propagation to be only weakly anisotropic [Groenhuis et al., 1983; Schweiger et al., 1995]. The diffuse light propagation should be slightly anisotropic otherwise the photon current would be zero and there would be no net light propagation [Ishimaru, 1978]. Condition $\mu_a \ll \mu_s$ holds for many anatomical regions of interest, for example breast or the brain. The presence of non-scattering regions, namely the CSF layer around the brain and in the ventricles does not allow the DA to accurately model the propagation of light [Dehghani et al., 1999]. The ability of DA to image media with non-scattering regions has been evaluated by a number of groups. Okada and Delpy [2003] investigated the effects of thickness of the CSF regions on the accuracy of the reconstructed images whereas Gibson et al. [2005b] commented on the extent that different reconstruction techniques cope with the problematic non-scattering regions. A number of methods have been proposed to cope with this specific issue such as the radiosity-diffusion hybrid model [Firbank et al., 1996] or the hybrid model of Tarvainen et al. [2008] which utilizes the RTE in those regions for which DA does not hold and the latter is used everywhere else. Simpler schemes include the proposal of Koyama et al. [2005] which showed that by assigning a value of $\mu_s' = 0.3mm^{-1}$ to *a priori* identified low-scattering regions in the probed medium, enables an adequate modelling of light propagation by the DA. Similar findings were recently published by Oki et al. [2009]. Another problem regards the unsuitability of DA to model light propagation in the vicinity of the light sources as propagation is highly anisotropic. Tarvainen et al. [2005b] proposed a hybrid model, employing RTE near the source locations and DA everywhere else.

3.4.2.2 Boundary conditions

The boundary conditions in the RTE are obtained by noting that no photons at the boundary travel inwards except from source terms, where $\hat{s}$ is the normal to $\partial\Omega$ at $\boldsymbol{m}$. [Arridge, 1999; Ishimaru, 1978; Schweiger et al., 1995]

$$I(\boldsymbol{m}, \hat{s}, t) = 0, \quad \text{for } \hat{s} \cdot \hat{s}_n < 0, \tag{3.27}$$

The diffusion equation cannot satisfy this condition exactly due to the simple angular distribution assumed for the diffuse radiation [Ishimaru, 1978]. Instead, there are a number of boundary conditions which can be adopted. One is the Dirichlet boundary condition (DBC) which physically is equivalent to a perfect absorbing medium surrounding the domain Ω and is expressed as

$$\Phi(\boldsymbol{m}, t) = 0, \quad \forall \boldsymbol{m} \in \partial\Omega. \tag{3.28}$$

A more realistic alternative in physical terms is imposed by the Robin boundary condition (RBC), which models the effects of a non-scattering medium surrounding Ω [Schweiger et al., 1995] and is expressed as

$$\Phi(\boldsymbol{m}, t) + 2\kappa(\boldsymbol{m})\hat{s}_n \cdot \nabla\Phi(\boldsymbol{m}, t) \overset{\text{from Eq. 3.23}}{=} \Phi(\boldsymbol{m}, t) + 2\hat{s}_n \cdot \boldsymbol{J}(\boldsymbol{m}, t) = 0. \tag{3.29}$$

If one needs to model the case of refractive index discrepancy within Ω and the surrounding medium, one can employ a modified RBC [Schweiger et al., 1995]

$$\Phi(\boldsymbol{m}, t) + 2A\hat{s}_n \cdot \boldsymbol{J}(\boldsymbol{m}, t) = 0 \tag{3.30}$$

where $A = (1+R)/(1-R)$ and where R is the internal reflection of uniformly diffuse radiation [Groenhuis et al., 1983; Schweiger et al., 1995]. Suggested values for R were computed in [Egan and Hilgeman, 1979] Alternative formulas for A have been proposed, for example see [Aronson, 1993; Keijzer et al., 1988].

3.4.2.3 Source conditions

Two possible type of sources include the pencil-beam collimated source (CS) and the diffuse source (DS) [Arridge, 1999; Schweiger et al., 1995]. The first cannot be modelled exactly by the DA. Instead, its effects can be approximated by an isotropic point source

$$q_0(r,t) = \delta(r - r_s)Q(t), \quad r \in \Omega \tag{3.31}$$

where Q is the source strength [Schweiger and Arridge, 1997] and r_s is situated at a depth of one scattering length $1/\mu_s'$ below $\boldsymbol{m} \in \partial\Omega$ where light is truly incident. The diffuse source scheme is expected to produce accurate results at distances greater than one mean free path length from r_s but not closer. In contrast, DS sources can naturally be modelled under the DA as an inward directed photon current distributed across a boundary segment $\partial\Omega_s \subset \partial\Omega$ [Schweiger et al., 1995]. This requires the modification of the DBC condition (see Eq. 3.28) to

$$\Phi(\boldsymbol{m},t) = 0, \quad \forall \boldsymbol{m} \in \{\partial\Omega \setminus \partial\Omega_s\} \tag{3.32}$$

$$\kappa(\boldsymbol{m})\hat{s}_n \cdot \nabla\Phi(\boldsymbol{m},t) = -\Gamma_s(\boldsymbol{m},t)w(\boldsymbol{m}), \quad \forall \boldsymbol{m} \in \partial\Omega_s \tag{3.33}$$

where $\Gamma_s(\boldsymbol{m},t)$ is the source flux along the boundary section $\partial\Omega_s$ and $w \in [0,1]$ is a function weighting the flux according to each distance from $\boldsymbol{m}$ [Schweiger, 1994]. The negation of Γ_s explicitly shows that flux is directed inwards to Ω [Kolehmainen, 2001]. Similarly, implementing the DS under the RBC, results in the modification of Eq. 3.29 along $\partial\Omega_s$ resulting to [Kolehmainen, 2001]

$$\Phi(\boldsymbol{m},t) + 2\kappa(\boldsymbol{m})A\hat{s}_n \cdot \nabla\Phi(\boldsymbol{m},t) = \begin{cases} -4\Gamma_s(\boldsymbol{m},t)w(\boldsymbol{m}) & \boldsymbol{m} \in \partial\Omega_s \\ 0 & \boldsymbol{m} \in \{\partial\Omega \setminus \partial\Omega_s\} \end{cases} \tag{3.34}$$

.

3.4.2.4 Measurable quantities

The quantity physically measured on the boundary - or equivalently the one modelled by the forward problem - is the diffuse outgoing radiation or exitance (Eq. 3.12). It is given by

$$\ddot{\Gamma}(\boldsymbol{m},\omega) = \ddot{\Gamma}(\boldsymbol{m},\omega = 0) + \mathcal{A}(\boldsymbol{m})e^{i[\omega t + \phi(\boldsymbol{m})]} \tag{3.35}$$

where $\ddot{\Gamma}^{(DC)}(\boldsymbol{m},\omega)\ddot{\Gamma}(\boldsymbol{m},\omega = 0)$ is the DC component and $\ddot{\Gamma}^{(AC)}(\boldsymbol{m},\omega) = \mathcal{A}(\boldsymbol{m})e^{i[\omega t + \phi(\boldsymbol{m})]}$ is the AC. It can also be re-written as the *Neumann* data

$$\ddot{\Gamma}^{(AC)}(\boldsymbol{m},\omega) = -c\kappa(\boldsymbol{m})\hat{s}_n \cdot \nabla\ddot{\Phi}(\boldsymbol{m},\omega) \tag{3.36}$$

The measurement types in the frequency domain case are the *phase shift* ϕ between source and measurement signal [Schweiger and Arridge, 1997]

$$\mathrm{y}_\phi(\boldsymbol{m}) = \arg \ddot{\Gamma}^{(AC)}(\boldsymbol{m},\omega) = \tan^{-1} \frac{\mathrm{Im}\left[\ddot{\Gamma}^{(AC)}(\boldsymbol{m},\omega)\right]}{\mathrm{Re}\left[\ddot{\Gamma}^{(AC)}(\boldsymbol{m},\omega)\right]} \tag{3.37}$$

and the modulation depth $m_{\mathcal{A}}$, defined as amplitude of the measured signal normalized by the DC component

$$\mathrm{y}_{\mathcal{A}}(\boldsymbol{m},\omega) = \frac{\mathrm{mod}\left(\ddot{\Gamma}^{(AC)}(\boldsymbol{m},\omega)\right)}{\ddot{\Gamma}(\boldsymbol{m},\omega=0)} = \frac{\left|\ddot{\Gamma}^{(AC)}(\boldsymbol{m},\omega)\right|}{\ddot{\Gamma}(\boldsymbol{m},\omega=0)} \tag{3.38}$$

In the TD case, the TPSF corresponds to the temporally dependent diffuse exitance $\Gamma(t)$. It can be interpreted as the distribution of the arrival times of the detected photons. Figure 3.8 shows a circular domain Ω with known μ_a and μ'_s distributions which is irradiated from a single source Q. The corresponding TPSFs built at four different detectors sites are shown and as expected, vary for different source-detector pairs [Delpy et al., 1988]. Commonly, one derives more practical data types which usually integrate out the temporal dependency and reduce the measurement to a single scalar value [Schweiger, 1994]. Example measurement types include [Arridge, 1999; Kolehmainen, 2001]

$$\mathrm{y}_{E[\Gamma(t)]} = \frac{1}{E[\Gamma(t)]}\int_0^{\infty}\Gamma(t)\mathrm{d}t \quad \text{normalized integrated intensity} \tag{3.39}$$

$$\mathrm{y}_{\langle t^n\rangle} = \langle t^n\rangle = \frac{1}{E[\Gamma(t)]}\int_{-\infty}^{\infty} t^n\Gamma(t)\mathrm{d}t \quad n^{th} \text{ temporal moment,} \tag{3.40}$$

$$\mathrm{y}_{c_n} = \frac{1}{E[\Gamma(t)]}\int_{-\infty}^{\infty}(t-\langle t^n\rangle)\Gamma(t)\mathrm{d}t \quad n^{th} \text{ central moment} \tag{3.41}$$

where $E[\Gamma(t)] = \int_0^{\infty}\Gamma(t)\mathrm{d}t$. For further discussion about data types see [Arridge, 1999; Arridge and Schweiger, 1993b; Hebden et al., 1998; Pineda et al., 2006].

3.4.3 Finite element method

The FEM is a general technique for the numerical solution of differential and integral equations which appear in science and engineering [Johnson, 1987]. Some of its principal characteristics are that it supports domains of arbitrary shape and quite arbitrary boundary conditions, it is robust and it has solid mathematical foundation [Ciarlet and Lions, 1991]. The existing literature covering FEM is vast and the reader can easily find extensive and dedicated reviews regarding general FEM, for example see [Johnson, 1987; Zienkiewicz and Taylor, 2000]. In the field of biomedical optics, FEM was introduced for the purpose of solving the DA in complex geometries in the early 1990s [Arridge et al., 1993; Schweiger, 1994; Schweiger et al., 1993, 1992, 1995] and since then it has been adopted by numerous research groups for the same purpose. For the purpose of modelling the DA in this work, we have employed the implementation in the TOAST[1] software package developed by Schweiger and Arridge at UCL.

In this case FEM is based on *method of weighted residuals* [Segerlind, 1984; Zienkiewicz and Taylor, 2000] to construct an integral formulation of the differential equation to be solved. Consider the frequency-domain formulation of DA described by Eq. 3.26. The FEM method for the time domain is based on a similar derivation [Schweiger and Arridge, 1997]. Suppose that $\ddot{\Phi}^{\star}$ is the solution to Eq. 3.26, therefore satisfying

[1] TOAST stands for Time-resolved Optical and Scattering Tomography. As of today, additional information and downloads can be found at http://web4.cs.ucl.ac.uk/research/vis/toast/intro.html.

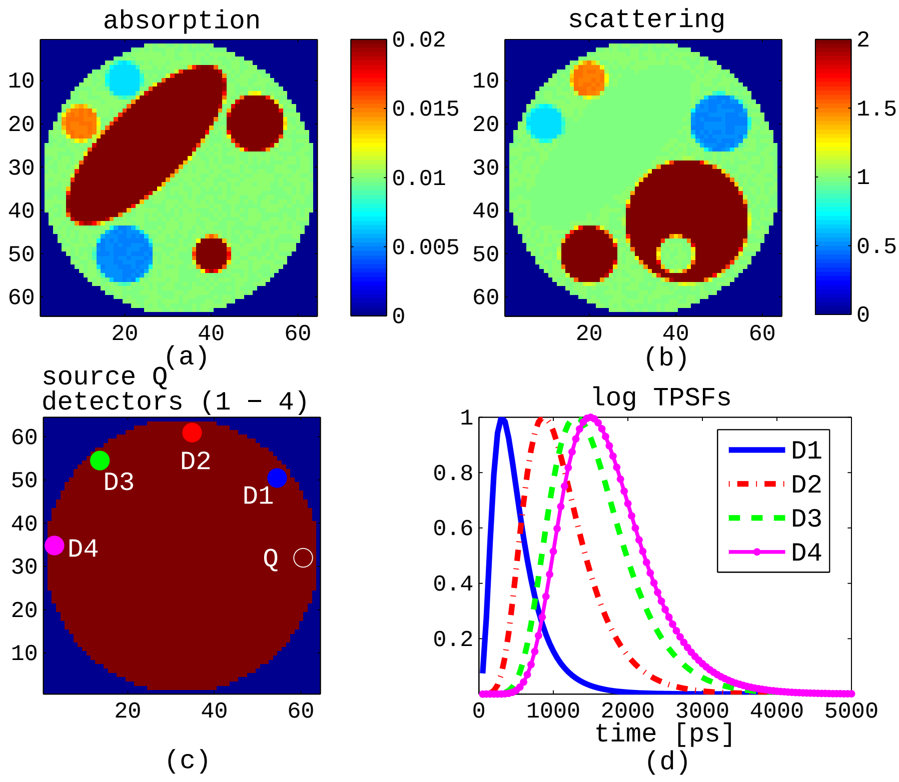


Figure 3.8: (a)-(b) Absorption/scattering distributions (c) source Q and detectors $D1\ldots4)$ (d) log-TPSF obtained from each detector for irradiation from Q in the noise free TD case.

$$\left\{\frac{i\omega}{c} - \nabla\cdot\kappa(r)\nabla + \mu_a(r)\right\}\ddot{\Phi}^{\star}(r,\omega) - \ddot{q}_0(r,\omega) = 0, \quad \forall r \in \Omega. \tag{3.42}$$

$\ddot{\Phi}^{\star}$ is continuous over Ω. We seek a solution estimate $\ddot{\Phi}(r,\omega) \to \ddot{\Phi}^{\star}(r,\omega)$. In order to render the problem manipulable by a computer, the problem under consideration must be completely defined by a finite number of unknowns. One way to achieve this is by approximating $\ddot{\Phi}$ with a finite basis representation, where it is expressed by linear combination of a finite number of basis functions $\Xi(r)$ and basis coefficient $\ddot{\Phi}^b \in \mathbb{R}^N_\Xi$ or

$$\ddot{\Phi}^h(r,\omega) = \sum_{i=1}^{N_\Xi} \ddot{\Phi}^b_i(\omega)\Xi_i(r) \tag{3.43}$$

Ω is now divided into a finite number N_e of non-overlapping elements e, forming an unstructured mesh which completely covers Ω so $\Omega = \cup_{j=1}^{N_e} e_j$. Elements are joined at $N_\mathcal{N}$ node locations $\mathcal{N}$ defined at the vertices of the elements. The elements are usually of simple shape such as triangles in the 2D case or tetrahedra in 3D, however higher order shapes can be used. An example 2D mesh is shown in Fig. 3.9. Given this configuration, a simple approximation of the true solution can be obtained by adopting piecewise linear Ξ. The basis functions used to approximate the solution are known as shape functions

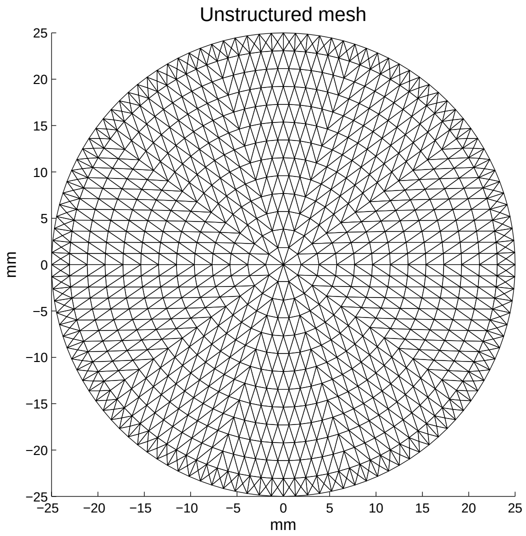


Figure 3.9: An unstructured mesh produced by the triangulation of a circular domain Ω. It consists of 1021 nodes and 1920 elements

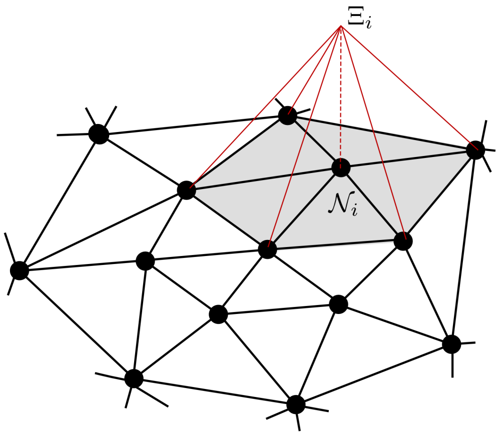


Figure 3.10: Linear shape function Ξ_i centered on node $\mathcal{N}_i$

in the FEM, due to the shape of their support area. The support of Ξ_j, is finite and extends to all all e containing $\mathcal{N}_j$ (see Fig. 3.10).

Given arbitrary function $\mathcal{W}$ satisfying the same boundary condition as $\ddot{\Phi}$ [Arridge et al., 1993], then the method of weighted residuals requires that

$$\int_{\Omega} \mathcal{W}(r) \left[\left\{ \frac{i\omega}{c} - \nabla \cdot \kappa(r) \nabla + \mu_a(r) \right\} \ddot{\Phi}^h(r,\omega) - \ddot{q}_0(r,\omega) \right] \mathrm{d}\Omega = 0. \tag{3.44}$$

Integrating Eq. 3.44 by parts using the Green's first identity [Zienkiewicz and Taylor, 2000]

$$\int_{\Omega} a\nabla \cdot (b\nabla c)\mathrm{d}\Omega = -\int_{\Omega} \nabla a(b\nabla c)\mathrm{d}\Omega + \oint_{\partial\Omega} a(b\nabla c)\mathrm{d}(\partial\Omega) \tag{3.45}$$

results in

$$\begin{aligned}\int_{\Omega} \left[\kappa(r)\nabla\mathcal{W}(r) \cdot \nabla\ddot{\Phi}^h(r,\omega) + \mu_a(r)\mathcal{W}(r)\ddot{\Phi}^h(r,\omega) + \frac{i\omega}{c}\mathcal{W}(r)\ddot{\Phi}^h(r,\omega) \right] \mathrm{d}\Omega = \\ \int_{\Omega} \mathcal{W}(r)\ddot{q}_0(r,\omega)\mathrm{d}\Omega + \oint_{\partial\Omega} -\Gamma(\boldsymbol{m},\omega)\mathcal{W}(\boldsymbol{m},\omega)\mathrm{d}(\partial\Omega)\end{aligned} \tag{3.46}$$

Eq. 3.46 is the *weak form* of Eq. 3.44 as it is comprised by lower order derivatives of $\ddot{\Phi}^h$, at the price of requiring the first derivative of $\mathcal{W}$. In effect the weak form is more realistic than the original formulation which implied excessive 'smoothness' of $\ddot{\Phi}^h$ by incorporating its second derivatives [Zienkiewicz and Taylor, 1987]. Regarding exact form of $\mathcal{W}$, there are a number of possible alternatives [Zienkiewicz and Taylor, 2000]. A very common choice is the same shape function Ξ used for the approximation in Eq. 3.43. This is known as the *Galerkin* formulation. Eq. 3.46 can be re-written in a matrix form ([Schweiger, 1994]) as

$$\left(\boldsymbol{K}(\kappa)+\boldsymbol{C}(\mu_a)+\frac{i\omega}{c}\boldsymbol{M}\right)\ddot{\boldsymbol{\Phi}}^b=\boldsymbol{Q}+\boldsymbol{\beta} \tag{3.47}$$

where

$$\begin{aligned}
\boldsymbol{K}_{ij} &= \int_\Omega \kappa(r)\nabla\Xi_j(r)\cdot\Xi_i(r)\mathrm{d}\Omega, & \boldsymbol{Q}_j(t) &= \int_\Omega \Xi_j(r)\ddot{q}_0(r,\omega)\mathrm{d}\Omega \\
\boldsymbol{C}_{ij} &= \int_\Omega \mu_a(r)\Xi_j(r)\Xi_i(r)\mathrm{d}\Omega, & \boldsymbol{\beta}_j(t) &= \oint_{\partial\Omega} -\Xi_j(\boldsymbol{m})\Gamma(\boldsymbol{m},\omega)\mathrm{d}(\partial\Omega) \\
\boldsymbol{M}_{ij} &= \int_\Omega \Xi_j(r)\Xi_i(r)\mathrm{d}\Omega, & \ddot{\boldsymbol{\Phi}}^b &= \left[\ddot{\Phi}_1^b(\omega),\ddot{\Phi}_2^b(\omega),\ldots,\ddot{\Phi}_{N_\Xi}^b(\omega)\right]^{\mathrm{T}}.
\end{aligned} \tag{3.48}$$

The time domain analogue of Eq. 3.47 can prove more complex to solve due to its dependency on the time variable t. Methods for modelling time dependency under the FEM setting are discussed in [Arridge et al., 1993; Schweiger, 1994]. For the implementation of the boundary and source conditions in FEM see [Schweiger et al., 1995].

3.4.3.1 Forward model and FEM

The complex photon density nodal coefficients $\ddot{\boldsymbol{\Phi}}_s^b$ rising from illumination of Ω from source s, with $s=1\ldots,S$ and for specific scalar functions $\mu_a(r),\kappa(r)$, are retrieved by solving Eq. 3.47

$$\ddot{\boldsymbol{\Phi}}^b=\left(\boldsymbol{K}(\kappa)+\boldsymbol{C}(\mu_a)+\frac{i\omega}{c}\boldsymbol{M}\right)^{-1}(\boldsymbol{Q}+\boldsymbol{\beta}) \tag{3.49}$$

The continuous photon density $\ddot{\Phi}_s^h$ across the domain is retrieved by Eq. 3.43. This gives rise to the outgoing measurement distributions $\mathrm{y}_s(\boldsymbol{m},\omega)$ corresponding to the Neumann data (Eq. 3.36). The forward mapping from μ_a,κ to $\mathrm{y}_s(\boldsymbol{m},\omega)$ for a specific source s is denoted as

$$\mathrm{y}_s(\boldsymbol{m},\omega)=\mathcal{F}_s(\mu_a,\kappa;\omega). \tag{3.50}$$

Following the notation of Arridge [1999]; Schweiger et al. [2005], a measurement model $\mathcal{M}$ is defined to sample the outgoing distributions in the boundary at M detector sites indexed by d

$$y_{s,d}(\omega)=\mathcal{M}_d\left[\mathrm{y}_s(\boldsymbol{m},\omega)\right]=\int_{\partial\Omega} w_d(\boldsymbol{m})\mathrm{y}_s(\boldsymbol{m},\omega) \tag{3.51}$$

where $w_{\mathrm{d}}(\boldsymbol{m})$ represents the finite aperture of detector d. It should be noted that $M\leq D$ where D is the number of the full set of detector sites. The reason for this distinction is that in the inverse problem it is possible to use a subset of the detectors for each activated source. For example, one can deactivate detectors too close in proximity with an activated source as the detected signal corresponds to superficial photon paths.

Effectively this concludes the forward operator $\mathcal{F}$ definition mapping $\mu_a,\kappa\in\mathbb{X}$ to $\mathrm{y}_{s,d}\in\mathbb{Y}$ for all source/effective detector combinations, denoted by

$$\mathrm{y}(\omega) = \begin{bmatrix} \left[\mathrm{y}_{1,1}(\omega), \mathrm{y}_{1,2}(\omega) \ldots \mathrm{y}_{1,D}(\omega)\right]^{\mathrm{T}} \\ \left[\mathrm{y}_{2,1}(\omega), \mathrm{y}_{2,2}(\omega) \ldots \mathrm{y}_{2,D}(\omega)\right]^{\mathrm{T}} \\ \vdots \\ \left[\mathrm{y}_{S,1}(\omega), \mathrm{y}_{S,2}(\omega) \ldots \mathrm{y}_{S,D}(\omega)\right]^{\mathrm{T}} \end{bmatrix} = \mathcal{F}(\mu_a, \kappa; \omega) = \begin{bmatrix} \mathcal{M}_d\left[\mathcal{F}_1(\mu_a, \kappa; \omega)\right] \\ \mathcal{M}_d\left[\mathcal{F}_2(\mu_a, \kappa; \omega)\right] \\ \vdots \\ \mathcal{M}_d\left[\mathcal{F}_S(\mu_a, \kappa; \omega)\right] \end{bmatrix} \tag{3.52}$$

It should be noted that $y \in \mathbb{R}^{M \times S}$ (vectorized form) and each complex measurement $\mathrm{y}_{s,d}$ consists of the amplitude and phase part (see Section 3.4.2.4).

3.4.4 Alternative approaches

3.4.4.1 Analytical methods

Analytic solutions to the diffusion equation 3.26 can be obtained by applying the appropriate Green's operator [Arridge, 1999].

$$\ddot{\Phi}(r, \omega) = \ddot{\mathcal{G}}(r, \omega)\left[\ddot{q}_0\right] \tag{3.53}$$

$$= \int_{\Omega} G^{(\ddot{\Phi})}(r, r', \omega)\ddot{q}_0(r', \omega)\mathrm{d}^n r. \tag{3.54}$$

In the case that the source term $\ddot{q}_0$ (its time-domain representation) is a δ-function the solution becomes, from the convolution of Eq. 3.54, just the Green's function itself. The pulsed sources used in optical imaging can be considered a sufficient approximation to a δ-function [Arridge and Hebden, 1997]. The exact form of Green's functions $G^{(\ddot{\Phi})}$ for simple geometries (spheres, slabs and cylinders) with homogeneous optical distributions can be found in specialized literature, for example see [Arridge et al., 1992].

3.4.4.2 Monte Carlo methods

Monte Carlo methods have evolved to be the "gold standard" technique for modelling of light in tissue [Binzoni et al., 2008]. Introductory texts specifically for the field of biomedical optics include [Flock et al., 1989; Jacques and Wang, 1995; Prahl et al., 1989; Wilson and Adam, 1983]. In brief, the rules which govern photon propagation inside Ω with known $\mu_a(r), \mu_s'(r)$, are perceived as random variables (RVs). Such rules represent the mean free path between absorption and scattering events and the scattering angle in the case of the latter. The RVs are accompanied by well defined probability density functions (PDFs) derived from the given $\mu_a(r), \mu_s(r)$. The propagation path of individual photons is modelled by repeatedly instantiating the RVs to compute the next position of the photon. The instantiation of RVs is accomplished by employing random number generators which produce samples from the PDFs. The disadvantage of these methods is the increased computational complexity rising especially for complex $\mu_a(r), \mu_s'(r)$. Accuracy depends on sufficient statistics and the latter requires the tracking of a high number of individual photons, where each individual path is characterized by multiple RVs instantiations. Rendering the task computationally tractable is crucial and has driven dedicated studies such as [Alerstam et al., 2008], which introduces a parallel processing approach utilizing modern multi-core GPUs. Monte Carlo methods are mainly used for forward modelling when DA does not hold [Boas et al., 2002; Okada and Delpy, 2003] or for validating newly proposed models due to their "gold standard" status, for example [Heino et al., 2003; Schweiger et al., 1995; Sikora et al., 2004].

3.5 Inverse problem

The inverse problem in DOT seeks to recover continuous scalar functions $\mu_a(r), \kappa(r)$ which resemble the *true* optical parameter distributions $\mu_a^\star(r), \kappa^\star(r)$ within Ω, given measured data y and known incoming radiation from sources q on $\partial\Omega$.

The continuous representation of the strictly positive scalar functions $\mu_a(r)$ and $\kappa(r)$ cannot be handled by computers. Following the paradigm of Schweiger and Arridge [2003]. the solution can be expressed as an expansion of a finite set of basis functions $b^{(i)}(r), i = 1 \rightarrow N/2$ defined in Ω, along with a set of basis coefficients $\mu_{a_i}, \kappa_i \in \mathbb{R}^{N/2}$ [1]. Given the coefficients, the continuous solution can be defined anywhere via

$$\mu_a(r) \approx {\mu_a}^b(r) = \sum_{i=1}^{N/2} \mu_{a_i} b^{(i)}(r), \quad \text{and} \tag{3.55}$$

$$\kappa(r) \approx \kappa^b(r) = \sum_{i=1}^{N/2} \kappa_i b^{(i)}(r) \tag{3.56}$$

where $x^b(r) \neq x(r)$ due to the approximation nature of the basis representation. The basis used to represent the reconstructed images will be referred to as *solution basis* as the number of coefficients define the dimensionality of the problem. Schweiger & Arridge [Schweiger and Arridge, 2003] investigate the effects of different basis representations in the reconstruction. The same paper provides information of how to switch between different basis representations. This inter-basis mapping is used for example in TOAST, to map from the solution basis to the nodal basis used in FEM forward solver. The problem now reduces to the retrieval of the finite basis coefficients denoted by

$$\mu_a = \left[\mu_{a_1}, \mu_{a_2}, \ldots, \mu_{a_{N/2}}\right]^{\mathrm{T}} \tag{3.57}$$

$$\kappa = \left[\kappa_1, \kappa_2, \ldots, \kappa_{N/2}\right]^{\mathrm{T}} \tag{3.58}$$

To simplify the notation, both unknown quantities are combined under a common variable name, giving rise to

$$x_+ := \begin{pmatrix} \mu_a \\ \kappa \end{pmatrix}, \; x_+ \in \mathbb{R}^N. \tag{3.59}$$

The term ${x_+}^\star$ denotes the true unknown solution comprised by $\mu_a^\star$ and $\kappa^\star$. The aim is to find the optical parameter estimate $\hat{x}_+$ which best fit the experimentally measured data y. In the deterministic paradigm this is expressed as

[1] Note that the absence of the positional vector r differentiates the basis coefficients μ_a, κ from the continuous scalar functions defined earlier

$$\hat{x}_+ = \underset{x_+}{\text{arg min}} \left[\mathcal{E}(x_+) = \mathcal{D}(y, \mathcal{F}(x_+)) + \tau\Psi(x_+)\right] \tag{3.60}$$

where $\mathcal{D}(y, \mathcal{F}(x_+))$ measures the discrepancy between measured data and modelled data and $\Psi(x_+)$ is the regularization function weighted by τ (see Chapter 2). Regularization is crucial as it alleviates ill-posedness and increases the chances that the retrieved $\hat{x}_+$ matches $x^\star$. The most common choice for $\mathcal{D}(y, \mathcal{F}(x_+))$ is the squared L_2 norm $\|y - \mathcal{F}(x_+)\|^2$, valid when the noise in the data follows a Gaussian distribution [Arridge, 1999; Arridge and Schotland, 2009; Viola, 1995]. Indeed, photon detection can be modelled by Poisson statistics. With a sufficiently large number of detected photons, the Poisson statistics can be approximated by a Gaussian distribution, with a variance proportional to the magnitude of the measurements [Arridge and Schotland, 2009; Guven et al., 2005; Oh et al., 2002; Ye et al., 2001].

3.5.1 Linear case

One needs to differentiate between the linear and non-linear case in DOT. The linear case which can be referred to as difference diffuse optical tomography (DDOT) uses the difference between two acquired data sets or $y^\delta = y - y^{(0)}$ and attempts to recover the difference in the corresponding optical properties $x_+^\delta = x_+ - {x_+}^{(0)}$. Linearisation of $\mathcal{F}$ in the vicinity of ${x_+}^{(0)}$ is given by the Taylor series

$$y = y^{(0)} + \mathcal{F}'({x_+}^{(0)})(x_+ - {x_+}^{(0)}) + \mathcal{F}''({x_+}^{(0)})(x_+ - {x_+}^{(0)})^2 + \ldots \tag{3.61}$$

where $\mathcal{F}'$ and $\mathcal{F}''$ are the first and second order *Fréchet* derivatives [Arridge, 1999; Schweiger et al., 2005]. Retaining only the first order terms in Eq. 3.61 then the problem is expressed as

$$y^\delta = Jx_+^\delta. \tag{3.62}$$

where $J \in \mathbb{R}^{2M \times N}$ is a discrete approximation of the $\mathcal{F}'$ matrix known as the *Jacobian* or *sensitivity matrix*. In practice, the DOT inverse problem can be assumed to be linear when initial estimate ${x_+}^{(0)}$ is close to ${x_+}^\star$ and the measured data y are close to the simulated measurements $y^{(0)} = \mathcal{F}({x_+}^{(0)})$. This is typically the case in difference imaging where measurements are taken before and after a small change in the optical properties [Gibson et al., 2005a]. Solving Eq. 3.62 in conjunction with some adopted regularization technique, can be approached by direct or iterative methods such as the ones discussed in Chapter 2. DDOT example studies include [Everdell et al., 2004; Gaudette et al., 2000; Gibson et al., 2005b].

3.5.2 Non-linear case

The non-linear case simply seeks to solve Eq. 3.60 when the assumptions about linear dependence between data and optical parameters do not hold. Numerical non-linear optimization techniques such as the ones introduced in Section 2.6 can be used. Most studies in DOT employ gradient based schemes, although non-gradient methods have been also proposed. For example Hielscher et al. [2000] employed an evolution strategy for reconstructing the optical parameters of homogeneous targets. The method however becomes non-tractable as the number of unknowns increases as it requires multiple evaluations of the costly forward operator. The most common first derivative method is the non-linear conjugate gradients, for example see [Arridge and Schweiger, 1998; Hielscher et al., 1999; Klose and Hielscher, 2002; Panagiotou et al., 2009b; Roy and Sevick-Muraca, 2001] with most studies employing a *Polak - Ribière* update scheme. Methods utilizing the second derivative are usually characterized by faster convergence rates, given that the basin of attraction at the minimum can be approximated by a quadratic. However, they also require more memory in order to explicitly store the Hessian matrix. Schweiger & Arridge [Schweiger et al., 2005] applied a Gauss-Newton scheme and investigated *Levenberg-Marquardt* and *damped Gauss-Newton* schemes for restoring global convergence (see Sec. 2.6.3). They also proposed the use of a generalized minimal residual (GMRES) Krylov method. The Hessian is sequentially accessed via matrix-vector products and never explicitly realized in full, hence the need of storage is converted to additional but manageable computational cost for the extra product computations. Klose and Hielscher [2003] investigated the performance of the Quasi-Newton methods and compared them with the GN approach whereas Roy and Sevick-Muraca [1999] proposed the use of truncated Newton methods.

3.5.3 Jacobian computation

Arridge [1999] classifies the available methods for computing J in three categories: *(i)* semi-analytic *(ii)* Monte-Carlo and *(iii)* numerical partial differential equation methods. This work makes use of *(iii)*, specifically the FEM based implementation found in TOAST. This is based on the concept of photon measurement density functions (PMDFs) introduced in its analytic form by Arridge [1995] and with a detailed analysis of the FEM based analogue discussed by Arridge and Schweiger [1995a].

For any s^{th} source and d^{th} detector the forward and adjoint photon density fields $\ddot{\Phi}_s(\Omega,\omega)$ and $\ddot{\Phi}^{\ddagger}_d(r,\omega)$ are computed. The PMDFs are then defined for any source/detector pair as

$$\ddot{\varrho}^{\mu_a}_{s,d} = -\frac{\ddot{\Phi}_s \ddot{\Phi}^{\ddagger}_d}{\mathrm{y}_{s,d}}, \quad \ddot{\varrho}^{\kappa}_{s,d} = -\frac{\nabla\ddot{\Phi}_s \cdot \nabla\ddot{\Phi}^{\ddagger}_d}{\mathrm{y}_{s,d}} \tag{3.63}$$

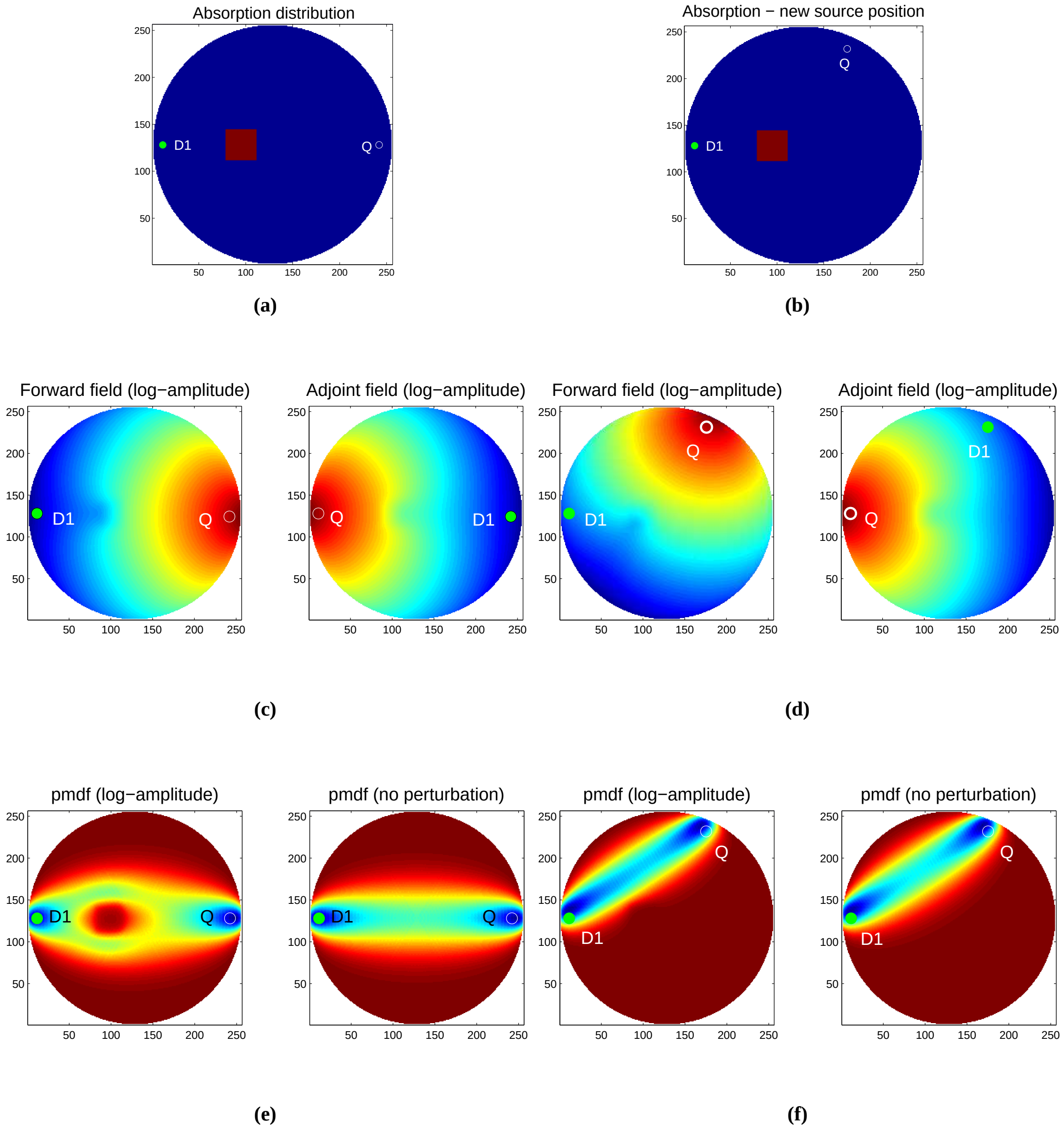
Figure 3.11 graphically shows the concepts of the PMDFs and their relation to the forward and adjoint fields for two absorption distributions. Effectively, the PMDFs depict the sensitivity of the data to perturbations at any μ_a, κ. The vectors in Eq. 3.63 are mapped into the solution basis and are then split into real and imaginary parts to construct the rows of the Jacobian $J \in \mathbb{R}^{2M\times N}$.

$$J = \begin{pmatrix} \partial \mathrm{y}^{(\mathcal{A})}/\partial\mu_a, & \partial \mathrm{y}^{(\mathcal{A})}/\partial\kappa \\ \partial \mathrm{y}^{(\phi)}/\partial\mu_a, & \partial \mathrm{y}^{(\phi)}/\partial\kappa \end{pmatrix} \tag{3.64}$$

$$= \begin{pmatrix} J^{(\mathcal{A},\mu_a)}_{s,d} & = \mathrm{Re}\left[\ddot{\varrho}^{\mu_a}_{s,d}\right] & , & J^{(\mathcal{A},\kappa)}_{s,d} & = \mathrm{Re}\left[\ddot{\varrho}^{\kappa}_{s,d}\right] \\ J^{(\phi,\mu_a)}_{s,d} & = \mathrm{Im}\left[\ddot{\varrho}^{\mu_a}_{s,d}\right] & , & J^{(\phi,\kappa)}_{s,d} & = \mathrm{Im}\left[\ddot{\varrho}^{\kappa}_{s,d}\right] \end{pmatrix}. \tag{3.65}$$

3.5.4 Solution scaling and positivity

Obtaining a solution estimate $\hat{x}_+$ by solving Eq. 3.60 requires essential transformations in order to improve the chances of the optimization algorithm to retrieving an accurate $\hat{x}_+$ and ensuring its positivity.



(a)

(b)

(c)

(d)

(e)

(f)

Figure 3.11: TOAST generated PMDFs using FEM. For simplicity, only absorption images are included and only the amplitude $\mathcal{A}$ part of the PMDFs. 3.11a-3.11b Two μ_a distributions which differ regarding the location of the source Q. 3.11c-3.11d forward (or direct) and adjoint fields corresponding to $\log\left(\mathrm{Re}\left[\ddot{\Phi}\right]\right)$. The perturbation can be seen in all fields. 3.11e-3.11f PMDFs corresponding to $\log\left(\mathrm{Re}\left[\ddot{\varrho}^{\mu_a}_{s,d}\right]\right)$. Two additional PMDFs are included where the perturbation in each case was removed, for comparison reasons. Each of the twoPMDFs (1^{st} and 3^{rd}) corresponds to a row in J.

3.5.4.1 Parameter normalization

Optical imaging differs from most modalities as it simultaneously reconstructs not one, but two physical quantities μ_a and κ. The range of values of each quantity differs as they represent different physical quantities with $\kappa > \mu_a$. In an iterative optimization scheme unless the parameter update for each κ and μ_a is proportional to the range of the optimized quantities, one of the quantities can potentially dominate the optimization process. Such problematic behavior can be treated by rendering both quantities dimensionless, effectively normalizing their individual contributions during the search for the optimum. Schweiger and Arridge [1999a] proposed the normalization of each of the two parts of x_+ with the mean value of the distributions used for initialization, giving rise to $\breve{\mu}_a = \mu_a/\bar{\mu}_a$ and $\breve{\kappa} = \kappa/\bar{\kappa}$ or by making use of the combined notation $\breve{x}_+ = x_+/\bar{x}_+$. This can also be achieved by optimizing the $\log(x_+)$, a choice which we also employ for the reasons described below.

3.5.4.2 Positivity of solution

The solution estimate $\hat{x}_+$ in Eq. 3.60 must be strictly positive. Unconstrained optimization can lead to negative optical quantities which is physically impossible. Approaches which can address this problem and can be inspired by other modalities such as positron emission tomography (PET), include the general theory of constrained optimization [Nocedal and Wright, 1999; Press et al., 1992a], augmentation of the objective functional to include penalties for negative solutions (same principle as generalized Tikhonov regularization) [Mumcuoglu et al., 1994], modifications of the line search process such as the bent-line search [Kaufman, 1987; Mumcuoglu and Leahy, 1994] and active sets [Kaufman, 1993]. PET solutions however are not required to be strictly positive but only non-negative. This enables DOT to adopt simpler approach described in [Schweiger and Arridge, 1999a; Schweiger et al., 2005]. The parameter vector is logarithmically transformed during the optimization of Eq. 3.60 giving rise to $x = \log(x_+)$. When an estimate $\hat{x}$ is reached, a strictly positive solution is obtained via exponentiation. Combining the parameter normalization described earlier, the full parameter transformations are given by

$$x = \mathcal{S}(\breve{x}_+) = \log(x_+/\bar{x}_+) \tag{3.66}$$

$$x_+ = \mathcal{S}^{-1}(x) = \exp(x)\bar{x}_+ \tag{3.67}$$

The optimization is now performed with respect to logarithmically and normalized vector of optical parameters x. By adopting an L_2 data discrepancy functional, the updated objective function of Eq. 3.60 becomes:

$$\hat{x} = \arg\min_{x} \left[\mathcal{E}(x) = \left\| \frac{\acute{y} - \mathcal{F}\big(\mathcal{S}^{-1}(x)\big)}{c_1} \right\|^2 + \tau\Psi(x) \right] \tag{3.68}$$

The normalizing constant c_1 is usually employed in an iterative optimization scheme to set the initial error to a value of 2, regardless of the considered $\acute{y}$ [Arridge, 1999]. This enables a consistent choice of the regularizing weight τ, whose value now only reflects the percentage of the regularization required for the provided data set. The actual form of c_1 is

$$c_1 = \left[\big(\acute{y} - \mathcal{F}(\mathcal{S}^{-1}({\mu_a}^{(0)})\big), \big(\acute{y} - \mathcal{F}(\mathcal{S}^{-1}(\kappa^{(0)}))\big) \right]^{\mathrm{T}} \tag{3.69}$$

where ${\mu_a}^{(0)}, \kappa^{(0)}$ denote the estimates at initialization. The μ_a, κ parts in c_1 normalize the corresponding parts of $\big(\acute{y} - \mathcal{F}(\mathcal{S}^{-1}(x)\big)$ in $\mathcal{E}(x)$. The final solution estimate obtained by Eq. 3.68, which adheres to the positivity condition, is obtained by $\hat{x}_+ = \mathcal{S}^{-1}(\hat{x})$. As it was already noted, the logarithmic transformation has the additional effect of rendering the μ_a and κ components in x dimensionless, which further helps in the parameter normalization.

3.5.4.3 Data scaling

Data y is subjected to transformation

$$
\acute{y} = \begin{pmatrix} \acute{\mathrm{y}}_{\mathcal{A}} \\ \acute{\mathrm{y}}_{\phi} \end{pmatrix} = \begin{pmatrix} c_{\text{re}}\mathrm{Re} \\ c_{\text{im}}\mathrm{Im} \end{pmatrix} \begin{pmatrix} \log\left(\mathrm{y}_{\mathcal{A}}(\boldsymbol{m}, \omega)\right) \\ \mathrm{y}_{\phi}(\boldsymbol{m}) \end{pmatrix} \tag{3.70}
$$

The logarithmic transformation of the amplitude part of the data is widely used in DOT. The measured signal is exponentially attenuated as the source/detector separation increases. Consequently, distant source/detector pairs are heavily penalized compared to the the ones in closer proximity. It is therefore natural to apply the logarithmic scaling as it reduces the impact of this erroneous effect. Constants c_{re}, c_{im} effectively normalize the $\log$-amplitude and phase which have completely different ranges [Schweiger et al., 2005]. See Schweiger and Arridge [1999a] for a similar discussion regarding the time-resolved case.

3.6 Regularization and multi-modality imaging

Due to the severe ill-posedness of DOT, the choice of the regularization functional $\Psi(x)$ is an essential matter as it drastically affects the the spatial and quantitative accuracy of the retrieved optical solution $\hat{x}_+$. The basic theory and concepts behind regularization were briefly covered in Chapter 2. This section provides a small sample of the various regularization schemes which have been published in the DOT specific literature. The classification of regularization methods can follow various schemes. Kaipio and Somersalo [2005] approach the regularization according to the Bayesian formulation, giving rise to prior densities. The main identified categories are Gaussian densities which correspond to a least squares functional in the deterministic setting as well as non-Gaussian densities, soft and hard-priors.

3.6.1 Quadratic penalties (Gaussian priors)

The most common regularization term in DOT is the quadratic functional

$$
\Psi(x) = \left\| L(x - x^{(0)}) \right\|^2 \tag{3.71}
$$

which in the Bayesian formulation corresponds to Gaussian densities [Kaipio and Somersalo, 2005; Kolehmainen, 2001]. In the case of $L = I$, this is known as the *white noise* prior or zeroth-order Tikhonov (TK$_0$) introduced in Sec. 2.4.3. It has been used in multiple studies in DOT including [Boverman et al., 2005; Model et al., 1997; Pogue et al., 1995].

In the case of $L \neq I$ but still a diagonal matrix, the regularization penalizes different regions of the image with different weighting. Such spatially varying regularization scheme was initially proposed in [Arridge and Schweiger, 1993a] and has been also adopted in [Arridge and Schweiger, 1995b; Li et al., 2003; Pogue et al., 1999b; Zhang et al., 2005a]. Specifically, Li et al. [2003]; Zhang et al. [2005a] used variable regularization weights for different tissue types, designated by prior X-Ray anatomical images. Lower regularization was applied to areas corresponding to lesions, as according to the authors, a smaller τ value reduces the penalty for the reconstruction of optical contrast and thus increases the probability of finding contrast in the designated region. Co-registration was guaranteed due to the DOT/X-Ray simultaneous probing apparatus.

Smoothness priors can be introduced by the first-order Tikhonov (TK$_1$) (see Sec. 2.4.3). This is one of the most common regularization schemes in DOT employed for example in [Arridge, 1993; Arridge and Schweiger, 1995b; Schweiger et al., 2005].

Considering the nature of the linear operator L, recent studies have defined more advanced forms in order to include *a priori* information from alternative modalities [Brooksby et al., 2005a; Dehghani et al., 2007; Yalavarthy et al., 2007a,b]. The methods displayed edge preserving characteristics where a smoothing operator is firmly applied within regions identified to belong to the same tissue type but with a reduced effect near region borders. The identification of the regions is accomplished via a labeled anatomical prior. These type of regularization effectively is a form of edge preserving prior.

Heiskala et al. [2009] employed both TK$_0$ and TK$_1$ regularization in their simulated brain imaging studies where they used the RTE as a propagation model under a Monte Carlo forward solving scheme. In addition, they have made use of unregistered *a priori* information provided by a probabilistic atlas, in order to improve the accuracy of the forward solver. The probabilistic atlas was firstly non-rigidly registered to match the boundary in the domain where light propagation would be modelled. The MRI based atlas was assigned with optical values according to the literature and finally provided the probability of each tissue type at each location inside the head for the Monte Carlo simulations.

3.6.2 Non-quadratic penalties

L_1-norm The incorporation of the L_1 norm as regularization functional, defined as

$$\Psi_{L_1}(x) = \|x\|_1 \,, \quad \text{where, } \|x\|_1 = \sum_{i=1}^{N} |x_i| \,, \tag{3.72}$$

was studied for DDOT as well as for the case of fluorescence DOT in [Cao et al., 2007; Mohajerani et al., 2007]. For information regarding fluorescence DOT see Milstein et al. [2003]. The above functional is known to promote the reconstruction of sparse solutions with as few as possible non-zero elements. Note that both DDOT and fluorescence reconstruct difference images hence non-zero pixels can be a part of the solution. The ideal form of a functional that would penalize non-sparse solutions would be the L_0 norm $\|x\|_0 = |\{i : x_i \neq 0\}|$. However its minimization is classified as an NP-hard problem [Natarajan, 1995] hence a compromise is achieved by the L_1 norm which is convex and can be optimized. More information about L_0, L_1 and sparsity can be found in Candès et al. [2007].

3.6.3 Edge preserving regularization

Total variation The total variation (TV) was initially introduced in image processing as an image enhancement tool, for example for the purpose of de-noising and de-blurring [Dobson and Santosa, 1996;

Rudin et al., 1992; Vogel and Oman, 1998]. Since then it has been proven as an effective regularizer in imaging inverse problems, across modalities including DOT. Its functional form is given by $\Psi_{\text{TV}}(x) = \int_\Omega |\nabla x(r)| \,\mathrm{d}r$. However, due to the discontinuity of the absolute value function at the origin, TV is usually approximated by

$$\Psi_{\text{TV}}(x) = \int_\Omega \xi\big(\nabla x(r)\big)\,\mathrm{d}r, \quad \text{where} \tag{3.73}$$

$$\xi(t) = \sqrt{t^2+\beta^2} - \beta, \quad \beta \to 0^+. \tag{3.74}$$

Additional choices for $\xi(t)$ can be found in [Vogel, 2002]. TV measures the *total lateral surface area* of the graph of x [Vogel, 2002]. It promotes piece-wise constant reconstructions as it smooths out fast intensity oscillations or weak edges - possibly belonging to artefacts - but allows the formation of *jump discontinuities*. TV measures the *total lateral surface area* of the graph of x [Vogel, 2002]. Reviews which refer to TV include [Kaipio and Somersalo, 2005; Vogel, 2002] whereas DOT studies include [Kolehmainen et al., 2000; Paulsen and Jiang, 1996; Tarvainen et al., 2005a].

Arridge et al. [2008a] recently proposed a modified TV, enabling the explicit designation of regions in the reconstructed image based on *a priori* information, where data-driven edge formation would be least penalized. A reference image x_{ref} providing the edge information can be obtained by applying edge detection methods on higher resolution images of the target medium obtained by alternative imaging modalities. The modified TV$_{\text{ref}}$ is defined as

$$\Psi_{\text{TV}_{\text{ref}}}(x) = \int_\Omega \xi(|\nabla x(r)|_{\mathcal{D}})\mathrm{d}r, \quad \text{where} \tag{3.75}$$

$$\xi(t) = \beta\sqrt{[t^2+\beta^2]} - \beta^2, \tag{3.76}$$

$$|\nabla x(r)|_{\mathcal{D}} = \sqrt{(\nabla x)^{\mathrm{T}}\mathcal{D}(r)\nabla x}, \quad \text{is an image to image mapping and} \tag{3.77}$$

$$\mathcal{D}(r) = \exp\left\{-\frac{|\nabla x_{\text{ref}}(r)|}{\beta_{\text{ref}}}\right\} I, \quad \text{is a symmetric tensor function.} \tag{3.78}$$

with I being the identity matrix and β_{ref} being a threshold controlling the influence of the edges in x_{ref}. Eq. 3.78 returns small values in r with high $|\nabla x_{\text{ref}}(r)|$ which propagate $\Psi_{\text{TV}_{ref}}(x)$ and reduce the penalty for edge formation in $x(r)$.

Anisotropic diffusion regularization Anisotropic diffusion regularization was proposed for DOT by Douiri et al. [2005a,b]. In the imaging paradigm, diffusion can be interpreted as a process of equilibration of the grey value concentrations in the image. In the case of directionally invariant or simply isotropic diffusion, “equalizing“ the concentration of grey values results in noise removal but this results in smoothing of important structural features such as the interfaces between physiologically different regions. The proposed anisotropic scheme preserved edges by blocking the diffusion process in directions orthogonal to the edge but not parallel to it. The regularization functional is identical to Eq. 3.73 but with $\xi(\cdot)$ being the *Hubert* function

$$\xi(|\nabla x(r)|) = \begin{cases} \frac{|\nabla x(r)|^2}{2}, & \text{if } |\nabla x(r)| \leq \beta \\ \beta\, |\nabla x(r)| - \frac{\beta^2}{2}, & \text{otherwise,} \end{cases} \tag{3.79}$$

and where β is now a scale parameter adjusted in each iteration. Similar to the transition from TV to TV$_{\text{ref}}$, Douiri et al. [2007] proposed a modification to the anisotropic diffusion regularization, in order to enable the introduction of *a priori* information from a reference image x_{ref}. x_{ref} would indicate regions in Ω where the diffusion process would be blocked, therefore edge formation would be promoted.

Anisotropic smoothness regularization Kaipio *et al.* [1999] proposed an edge preserving regularization method for the severely ill-posed problem of *electrical impedance tomography* [Webster, 1990]. Although this Chapter reviews DOT related methods, we refer to it as it is formulated in a similar manner with TV, but with

$$\xi(x) = x^2, \quad \text{and} \tag{3.80}$$

$$\mathcal{D}(r) = I - (1 + \|\nabla x_{\text{ref}}(r)\|^2)^{-1} \nabla x_{\text{ref}}(r) \nabla x_{\text{ref}}(r)^{\text{T}}. \tag{3.81}$$

It is worth noting that the authors refer to the TV and its dependency on the total length of the edges to be reconstructed which is not the case in this method. A comparison between this method and TV$_{\text{ref}}$ does not exist in the literature.

Hiltunen et al. [2008] proposed an alternative approach for anisotropic smoothness regularization. The approach was based on an alternating optimization scheme were at each iteration k an improvement to optical estimate $x^{(k)}$ was firstly computed. $x^{(k)}$ was perceived as the best current smooth approximation estimate of the true underlying distribution $x^\star$. The second part of the iteration used $x^{(k)}$ as a pilot in order to improve the estimate of a secondary scalar function $\lambda^{(k)}(r)$. $\lambda^{(k)}(r)$ was responsible for scaling the effect of regularization at pre-specified directions - effectively providing anisotropic smoothing. $\lambda^{(k)}(r)$ computation was based on the already available estimate $x^{(k)}$, which provided information about the presence of edges in the image.

3.6.4 Hard constraints and other types of prior information

Regularization provided by penalty functionals under the generalized Tikhonov regime are effectively soft constraints. They dictate a preference to some specific solution subspaces by penalizing the rest but they don't strictly enforce these constraints. On the contrary hard constraints must be satisfied by the totality of the problem's parameters. A regularization scheme involving hard constraints was proposed by Schweiger and Arridge [1999b] for DOT. They employed a magnetic resonance imaging (MRI) slice of an adult head, pre-segmented into four distinct regions corresponding to skin, bone, grey and white matter. The regions provided by the MRI spatially corresponded to the true regions in $x^\star$. Rather than using literature values as an initialization for the inverse solver, they approached the reconstruction in two separate parts. In the first part the solver was constrained to retrieve a single optical estimate x_k for all pixels identified to belong to the same anatomical region k. Considering all regions, this approach drastically reduced the dimensionality of the problem from potentially thousands of degrees of freedom (that is one for each pixel) to just eight - a single μ_a and μ_s' value for each region. The reduction of the dimensionality rendered a usually *under-determined* problem to an *over-determined* one, removing effectively the need of regularization. The piecewise constant estimate was used as a initialization guess for a second optimization using soft constraints, therefore resulting in a more realistic, continuously varying final optical estimate.

Dehghani et al. [2003] used the above scheme to analyse the resolution of the images in small animal imaging. Similarly, Ntziachristos et al. [2002] studied the performance of MRI guided diffuse optical spectroscopy of breast lesions. The apparatus of the system allowed scanning with both modalities without change in the positioning of the patient, thus enforcing co-registration of the involved medical signals. This allowed the voxels in the solution domain of the optical image to be labeled according to the tissue type of the breast to which they were superimposed. The determination of the tissue type was accomplished by the co-registered MRI signal. As in [Schweiger and Arridge, 1999b], by restricting all pixels in the optical domain belonging to the same tissue type to be completely correlated, the under-determined problem of DOT becomes overdetermined as the number of unknowns drops dramatically. Jiang et al. [2008]; Xu et al. [2008] used the method to incorporate prior information in trans-rectal ultrasound imaging (UI) driver, DOT imaging.

Pogue and Paulsen [1998] used the information from a segmented coronal MRI slice acquired from a rat, in order to create a realistic finite element mesh based numerical phantom to be probed. The three distinct regions of bone, muscle and brain tissue in the phantom, were assigned with optical values from the literature, known to correspond with each of the identifiable tissue types. The numerical phantom was then used for the simulation of the data acquisition process. The solution of the inverse problem was regularized by restricting the reconstruction of the optical parameters in regions which would correspond to brain or bone structures, hence reducing the dimensionality of the problem. The initial guess for the reconstruction was based on literature suggested values. The method differed from the scheme of Schweiger and Arridge [1999b] which was discussed earlier, as it would not apply hard constraints on the brain and bone regions but allowed the optical parameters to vary freely.

Guven et al. [2005] proposed a method towards the incorporation of a multi-modal structural prior image x_{ref} in DOT, explicitly designed to reduce undesirable, erroneous bias due to different features between x_{ref} and $x^{\star}$. The method involved a hierarchical Bayesian approach defining multi-stage priors. The first prior conditioned the optical solution initially with respect to the unknown hyper-parameters, namely the the mean and standard deviation of the optical solution. A second stage prior - or hyper-prior - conditioned these secondary hyper-parameters with respect to the anatomical image.

An alternative form of priors proposed in DOT regards *spectral priors*. It is known that the optical properties of tissue vary as a function of wavelength λ. The absorption coefficient of a mixture of chromophores can be expressed as the sum of the products of the concentration of each chromophore c_i with its extinction coefficient ϵ_i or $\mu_a(\lambda) = \sum_i \epsilon_i(\lambda)c_i$. The wavelength dependent extinction coefficient of chromophore represents the level of absorption per μmol of the chromophore, per liter of solution, per mm. Li et al. [2005] proposed the use of multiple spectral and spatial priors in optical tomography. Making use of MRI images, spatial prior information for the distribution chromophores of water and lipid could be incorporated into the optical inverse problem. The remaining chromophores targeted by DOT, namely oxy- and deoxy- haemoglobin were retrieved by the acquired optical data. The reconstruction of the same chromophores from different optical wavelength data acquisitions comprised a form of spectral prior. Brooksby et al. [2005b] compared structural with spectral priors concluding that the former improve spatial resolution, the latter improve quantitative accuracy and their combined usage returns superior images overall. Similar studies include [Corlu et al., 2005, 2003]. Intes et al. [2004], extended the flexible hierarchical Bayesian scheme by Guven et al. [2005] to introduce physiological information from chromophore concentrations obtained from multi-wavelength probing. He reported results of optical absorption parameter reconstruction of a slab, mimicking typical values encountered in a human breast, assisted by MRI derived priors.

Zhu *et al.* [1999; 2003; 2005] proposed the simultaneous probing of breast tissue with UI and DOT, for lesion detection. Their apparatus consisted of a hand-held probe combining on the same application surface 12 optical source fibers, 4 optical detector fibers and 20 piezoelectric crystals comprising the ultrasound array. The combined apparatus guaranteed sufficient co-registration between the optical and UI signal. While the UI signal provided very accurate localization of a tumour, the optical data provided information regarding the haemoglobin concentration of the tumour, allowing classification between malignant cancers and benign lesions. Regularization was provided via the method of reducing the unknowns and therefore improve the under-determined condition of the problem. They used the UI signal to identify the location of the lesions and proceeded by increasing the resolution of the optical solution at the identified region while reducing the resolution at the background.

3.7 Reference images: anatomical and functional correspondence

Introducing structural information in DOT from reference requires the underlying true optical parameters to be distributed in a manner similar to the secondary quantities depicted in the reference images.

Reference images x_{ref} are usually supplied from alternative imaging modalities probing for anatomical characteristics (MRI, X-ray computed tomography (CT), UI) or physiological function (functional magnetic resonance imaging (fMRI) and PET). These modalities consistently retrieve images of higher spatial resolution compared to that of DOT. In this sense, the secondary quantity depicted by the images refers to the underlying physical quantity targeted by these alternative modalities such as the magnetic properties of tissue and possible contrast agents, X-Ray attenuation, acoustic properties of tissue or radioactive tracer uptake characteristics of the probed anatomy. In order for the structural similarity between $x^{\star}$ and x_{ref} to hold, it is understandable that features comprising the optically probed medium should also exist in the medium probed by the secondary modality. Understandably, the probability of the latter condition being satisfied increases when both DOT and reference modality probe the exact same organ, in an intra-subject study and preferably at the exact same time. One should aim to match the field of view of both modalities during the set up of the experimental data acquisition process so spatial co-registration of the features potentially visible in both modalities is maximized.

Even in the ideal case described above, difficulty arises as DOT is a functional imaging modality whereas a number of the high-resolution modalities return only anatomical information. Variable amounts of blood, oxygenation levels or even electrical activation Gratton et al. [1997]; Stepnoski et al. [1991] can locally dominate the distribution of the optical properties while being undetected by other modalities. This undoubtedly creates structural differences between $x^\star$ and x_{ref} and compromises the quality of the prior information introduced by the latter. It is known however that distinct anatomical areas corresponding to different tissue types, are characterized by different optical properties - at least at a baseline level. The literature which provides suggested values for the optical coefficients of various tissue types comprising distinct anatomical regions is extensive, for example see Cheong et al. [1990]; Durduran et al. [2002]; Okada and Delpy [2003]; Troy et al. [1996] and the references within. Because different tissue types have different optical values (for example in brain all CSF, white, grey-matter, skin, bone correspond to different optical properties), the true optical solution $x^\star$ must structurally resemble a high-resolution anatomical image x_{ref} from another modality at least at that baseline level, justifying the use of the latter as *a priori* information. Studies which successfully use anatomical prior information in functional imaging modalities such as DOT or PET include Ardekani et al. [1996]; Boverman et al. [2005]; Brooksby et al. [2005a, 2004]; Comtat et al. [2002]; Gindi et al. [1993]; Li et al. [2003]; Ntziachristos et al. [2000, 2002]; Rangarajan et al. [2000]; Som et al. [1998]; Zaidi et al. [2003]; Zhu et al. [2005]. Introducing functional *a priori* information on a functional modality is another possibility. Spatial correspondence between various functional signals has been established, for example in the DOT/fMRI case in Zhang et al. [2005b], in PET/fMRI case Judenhofer et al. [2008]. Culver et al. [2008] discusses the possibility of introducing prior information in DOT from reference images obtained by PET. The finding of these studies are encouraging towards multi-modal functional imaging using methods such as the one proposed in this work.

3.8 Ill-posedness, noise and resolution

3.8.1 Ill-posedness

There are various factors which contribute to the ill-posedness of DOT. One contributing factor is the multiple sources of noise which can potentially contaminate the measured data y. Noisy measurements can be unreachable by the forward operator of DOT - in the sense that they do not live in $\mathcal{R}\left(\mathcal{F}(x)\right)$ for all feasible x. An additional reason for $y \notin \mathcal{R}\left(\mathcal{F}(x)\right)$ is that the DA is not valid very close to the sources [Arridge, 1999]. Effectively these factors lead to a breach of the first postulate of well-posedness - that is the existence of the true solution of $\mathcal{F}(x) = y$ (see Sec. 2.3). This case is usually treated by accepting the best unique compromise solution, such as $\hat{x}_{LS}$ (see the intuitive discussion of Sec. 2.4.2 for the linear case analogue of $\hat{x}$ which is based on the same principle).

Solution uniqueness, the second postulate by Hadamard, is compromised as the problem is in most cases under-determined. This case usually manifests in 3D studies where the dimensionality of the solution, determined by the number of voxels comprising it, is substantially larger than the dimensionality of the measured data. Regarding the uniqueness of the solution of the CW-DOT, it is compromised by the fact that we solve simultaneously for the distribution of two physical processes (absorption and diffusion/scattering) rather than one, proved even in the limit where the complete data is measured on the boundary [Arridge and Lionheart, 1998; Harrach, 2009]. The non-uniqueness condition is also not met in the frequency domain DOT when the distribution of the refractive index in the probed domain is considered as an unknown [Arridge and Lionheart, 1998]. Furthermore, an inherent and unavoidable source of ill-posedness in DOT is the diffuse nature of light propagation in tissue [Boas et al., 2001].

Additional practical issues which further complicate DOT and can potentially affect the well-posedness of the system include the approximation of the true physical propagation of light by the various light transport models such as the DA; the non-complete collection of the exiting light from the medium, which effectively renders the measured data, a sample of the complete data; the non-linear relationship between unknowns and data in the light transport models; and discretisations such as the representation of the continuous solution by a finite number of unknowns.

3.8.2 Noise

In Subsec. 3.8.1 we refered to the effects of noise in the well-posedness of the system. There a various source of noise in DOT discussed in detail in [Schmidt, 1999].

When imaging across mediums of large thickness $(> 6cm)$, the intensity of the exiting light is several orders of magnitude lower than the one of the incident radiation. Only few photons exit the medium. In this case, DOT imaging is performed using the powerful pulsed laser sources and photon counting techniques incorporated into TD systems [Gibson et al., 2005a]. In that setting, the data is contaminated by Poisson distributed noise arising from the stochastic nature of the photon counting process. This results to signal-to-noise ratio which only increases with the square root of the number of traced photons [Schmidt, 1999]. Hebden et al. [1998] notes that a measurement arising from a TPSF built by 10^6 photons has less than 0.2% noise. One should consider that 99% of the probing laser pulses (modulated in the order to 10 of MHz) do not produce a photon detection event [Schmidt, 1999]. The efficiency of the MONSTIR optical tomography system of UCL, that is percentage of generated photon count events given the number of photons arriving at the detector sites, is 0.04 [Schmidt, 1999]. Then, to ensure that the measured data at the sector sites is based on 10^6 photons counts, each source has to be activated for a certain amount of time, producing pulses of light. All the photon counting events from this prolonged activation result to a single TPSF which then results to a single measurement (see Subsec. 3.4.2.4). For example, probing through a $9cm$ object with $\mu_a = 0.01mm^{-1}$ and $\mu_s' = 1mm^{-1}$ this would result to a required activation of $4s$ [Hillman, 2002]. One can conclude that the shot noise dependence to photon counts dictates the temporal resolution of the system, which depends on the size of the probed object.

There are additional sources of noise in DOT, other than the aforementioned. The measurements can be contaminated by noise arising from the coupling of the fibres with the skin and accuracy of the representation of the fibres location during the FEM modelling; random noise due to detection of stray room light and from thermally induced emission in the *photon multiplier tubes* used by the detectors; systematic noise due to detection of stray laser light, internal laser reflections in the system etc. [Schmidt, 1999].

3.8.3 Resolution

The temporal resolution in DOT is dictated from the number of photons required to achieve an acceptable signal-to-noise ratio as well as instrumentation issues. Studies targeting function in the superficial layers of the anatomy can achieve very fast temporal resolutions $(< 1s)$, as the source/detector array is usually located in a surface region close to the targeted area, effectively measuring the reflected light. High depth measurements require transmission measurements, arising from probing configurations where the sources and detectors are located on opposite surfaces of the medium. The intensity of the exiting light in the transmission measurements is lower than that of reflection measurements, effectively limiting the temporal resolution (see Subsec. 3.8.2).

The spatial resolution of DOT reconstructions cannot be easily quantified as it is affected by many factors, both physical and study dependent. The resolution is inherently limited by the diffuse nature of light propagation in tissue. Its effect is schematically showcased in Fig. 3.12 where a homogeneously absorbing medium - both in X-Rays and NIR light -with an embedded perturbation is probed by CT and DOT. The X-Rays travel in straight lines, thus the profile of the detected intensity consists of high-frequency components which reflect the ability of CT to resolve spatial details. In contrast, the diffuse nature of near infrared (NIR) light results to spread out profiles carrying little spatial information regarding the perturbations location. By taking into consideration the multiple sources of noise and the ill-posedness of the inverse problem, the already limited spatial information is further compromised.

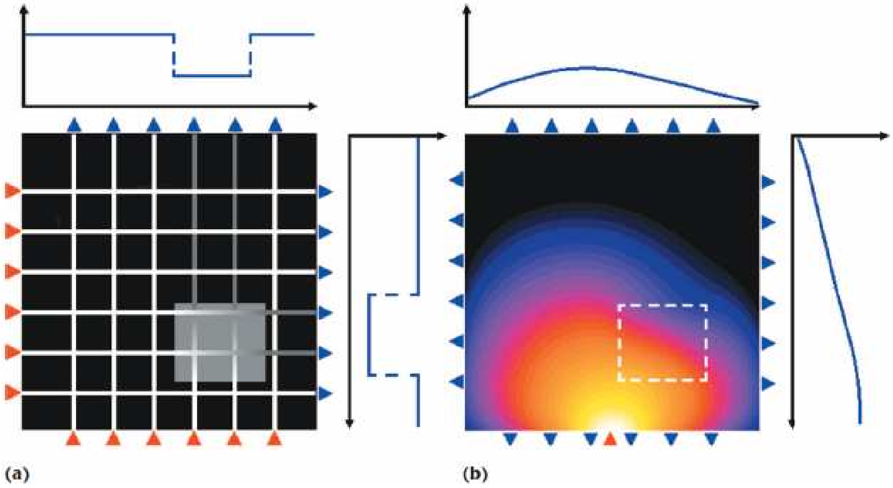


Figure 3.12: Diffuse light transport. (a) A schematic view of computed tomography (CT)-like projections along straight lines, where included objects cast âĂIJshadowsâĂİ on the opposite detector array. (b) Photon density wave from a single source propagating through a diffuse medium with an embedded object. Detectors placed around the surface maximize data information. Courtesy of Schweiger et al. [2003]

Another important characteristic of the resolution of DOT, is that it is spatially dependent as it deteriorates in higher depths, where the distance from both sources and detectors is increased. Photons traveling across the centre of the medium are more likely to be absorbed due to the higher number of interactions with the molecular structure of the tissue. High depth imaging requires transmission measurements resulting from sources/detectors located in diametrical locations. In contrast, features close to the surface can be more accurately resolved, due to the higher number of photons following superficial paths. The are photons emitted and later detected by spatially proximal source/detector pairs. Effectively this constitutes a reflection[1] measurement. However, regions in the reconstruction close to the boundary are susceptible to artefacts. Consider the case where photons emitted from a source travel in superficial layers and are detected by an adjacent detector. Superficially traveling photons however will not travel the full circumference of an object as the path would be very large and they will probably get absorbed. Superficially travelling photons will only be detected at locations proximal to the source. This means that the reconstruction of the pixel values at regions close to the surface depend on a limited number of measurements from the few contributing source/detector pairs, located close to these regions. This can be thought of as a locally under-determined problem and artefacts can manifest. The intuition regarding the depth dependent resolution of DOT has lead to spatially varying regularization schemes discussed in Sec. 3.6.1, where boundary regions are over-regularized to suppress artefacts whereas deep locations are under-regularized to preserve the limited information coming from these regions.

In addition, the resolution of DOT is dependent on the actual target it attempts to reconstruct. Some regions can be completely invisible if they can be surrounded by high absorbing layers which do not allow light to travel inside them. A similar behaviour can be observed in scattering, where a region might become invisible when a surrounding layer scatters light to directions other than its interior. The resolution of DOT increases with the number of sources/detectors employed [Ntziachristos et al., 2001].

It is evident that quantifying the resolution of DOT is a complicated task as there is not standard imaging apparatus as well as due to its dependency to the specifics of each case. Resolution levels regarding the position of a perturbation, have been reported to be as low as $2mm$ in simulated studies - see for example [Ntziachristos et al., 2001]. The overall size of a perturbation however can be underestimated. In a more general note we refer to the resolution levels of DOT reported by Correia [2010], which vary between $1 - 3cm$.

3.9 Summary

This chapter has introduced the main concepts of DOT imaging. The discussed topics varied from intuition on the physical aspect of light propagation in media, the review of the mathematical formulation of the forward problem, the light propagation models of RTE and its DA with weight on the FEM for the practical implementation of the latter. In addition, the inverse problem was formulated with emphasis to regularization and multi-modality imaging. Solving the inverse problem of DOT is a very challenging task and this drives dedicated research from numerous labs around the world. Additional details regarding DOT can be found in the suggested topical reviews [Arridge, 1999; Arridge and Schotland, 2009; Boas et al., 2001; Gibson et al., 2005a] and the references within.

[1] In this context reflection means that photons are exiting the medium from the same side of the medium they entered.

Chapter 4

Information theory

4.1 Introduction

The aim of this chapter is to introduce two of the fundamental concepts of information theory (IT), namely *information entropy* H and mutual information (MI). In the context of this work, both concepts will be approached from an imaging perspective.

Probably the most notable contribution to the development of information theory was made in the context of telecommunications. Its mathematical foundations were laid by Claude E. Shannon in his 1948 seminal paper *"A Mathematical Theory of Communication"* [Shannon, 1948]. The main problem addressed by Shannon's work regarded the efficient encoding of messages produced by an information source, in order to transmit them over a communication channel. Shannon defined the channel's transfer capacity considering its noise characteristics - as they can compromise the accuracy of the communicated message at the receivers' end - and the rate of information transmission from the source. He then provided a ground-breaking proof which stated that if the source's rate of generating information was less than the maximum capacity of the channel, a message could then be encoded, communicated and finally received with arbitrarily small uncertainty about the accuracy of its contents. By also defining the absolute maximum rate of transmission of information in a noisy channel, Shannon effectively defined the limits of efficient, error-free communication.

One question which needs to be answered regards to what exactly constitutes information and how is it defined. Shannon was prompt to declare that the semantic aspect of the information produced by a source or in a less rigorous and more abstract terminology - its *meaning*, is irrelevant to the engineering problem (telecommunications in that context). Information in IT is defined as 'the reduction in *uncertainty*, from the level prior to the receipt of the message to the level after it has been received'. The uncertainty after the receipt of the message, reflects on the possibility that the received message differs from the actual message sent from the source, due to noise contamination during transit. In this case, the uncertainty after the arrival of the message is a measure of noise.

The next obvious question which arises regards the uncertainty - or equivalently the choice - among the outcomes of the involved random processes - for example the message generating source or the uncertainty in the receiver's end due to noise contamination during transit - and how it can be measured. Shannon identified three postulates (see Sec. 4.5.1) which should be met by any uncertainty measure.

The only measure which meets all three postulates was answered by Shannon conclusively and unequivocally and is *entropy* H. Entropy can be considered as a measure of *uncertainty, randomness, choice or disorder*, as all terms are equivalent in this context.

Consider the following example. In the single performance of a coin toss, there is uncertainty involved regarding the outcome or its prediction. When the coin toss is performed and the outcome *heads* is recorded, the observer has received *information*. As it was noted, information in the context of IT has nothing to do with the semantic aspects of the message. Assuming that there are no errors during the observation of the toss outcome, there is zero uncertainty about the outcome heads being realized and the information gain (uncertainty reduction) is $H_{before} - H_{observed} = H_{before}$.

In the unlikely case of a coin with *head* on both sides, there is no uncertainty involved prior to the coin toss. The outcome will always be heads. In this case, the performance of the experiment is completely irrelevant. There is no information to be gained by the observer as the outcome is known *a priori*. Observing the outcome carries no information whatsoever as uncertainty is already zero.

The beauty of IT is that its rules and theorems operate on the abstract level. Quoting Kullback [Kullback, 1959], "*information theory is a branch of the mathematical theory of probability and statistics. As such, its abstract formulations are applicable to any probabilistic or statistical system of observations*". Indeed, IT has found widespread use in numerous scientific disciplines such as engineering, computer science, physics, economics and many more; see for example [Cover and Thomas, 1991] where each chapter elaborates on a different application of IT. Recommended references for IT include the original Shannon's paper [Shannon, 1948] as well as [Ash, 1990; Cover and Thomas, 1991; Mackay, 2002] and the *excellent* introductory primer [Schneider, 1995].

The structure of this chapter is as follows: Sec. 4.2 introduced fundamental concepts of probability theory such as random processes, random variables (RVs), probability density functions (PDFs) and sample statistics in the uni-variate and bi-variate setting. Sec. 4.3 defines the Normal - or Gaussian density - used extensively in this work. Sec. 4.4 describes parametric and non-parametric density estimation techniques. Sec. 4.5 introduces the information theoretic functionals of entropy, conditional entropy, joint entropy (JE) and MI. In addition it discusses the concepts of empirical and differential entropy (DE). The entire discussion draws analogues to the imaging context to assist clarity and provide further intuition.

4.2 Random processes, probabilities and random variables

4.2.1 Random process

A random process is a process whose outcome cannot be predicted with certainty. Examples include the toss of a coin resulting to heads (h) or tails (t) or the arrival of a person in the train station at a certain time t within a specific time interval $t_1 \leq t \leq t_2$. Given such a process, it is possible to describe some of its aspects using *set theory* terminology [Papoulis and Pillai, 2001]. The *sample space* $\Omega_\zeta = \{\zeta_1, \zeta_2, \ldots, \zeta_N\}$ of a random experiment, is the set comprised by all N possible *experimental outcomes* ζ_i, $i = 1, 2, \ldots N$. It is also known as the *certain event*, as one of its elements will always be realized. The empty-set $\{\emptyset\}$ is known as the *impossible event*. Ω_ζ consists of 2^N subsets known as the *events*, comprising the events' set $\mathfrak{G}^\star = \{\{\emptyset\},\{\zeta_1\},\{\zeta_2\},\dots\{\zeta_1,\zeta_2\},\{\zeta_1,\zeta_2\},\dots,\{\zeta_1,\zeta_2,\zeta_3\},\dots,\Omega^\star\}$, $\Omega^\star = \Omega \setminus \{\emptyset\}$. For example in the single coin toss process, the events are $\{\{\emptyset\},\{h\},\{t\},\{h,t\}\}$. We denote the individual events in $\mathfrak{G}^\star$ by ω_i.

The sample space of a random process can be *finite*, *countable infinite* or *uncountable infinite*. The coin toss repeated a fixed number of times falls in the first category, the repetition of the coin toss until heads are realized falls to the second category (possible events are $\mathfrak{G}^\star = \{\{h\},\{th\},\{tth\},\dots\}$) whereas the event of arriving at the station at some time instance t fall to the third category, due to the continuous nature of time.

4.2.2 Probabilities

The potential events $\omega \in \mathfrak{G}^\star$ can be assigned with a probability $Pr(\omega)$ which can be interpreted as measure of the uncertainty regarding their occurrence [Papoulis and Pillai, 2001]. $Pr(\omega)$ should satisfy

$$Pr(\omega) \geq 0\ , \tag{4.1}$$

$$Pr(\Omega) = 1 \text{ and} \tag{4.2}$$

$$Pr(\omega_i \cup \omega_j) = Pr(\omega_i) + Pr(\omega_j),\ \text{if } \omega_i \cap \omega_j \text{ for } \ i \neq j \tag{4.3}$$

where $\cap$ and $\cup$ denote set operations of intersection and union respectively. For mutually exclusive events it holds that

$$Pr(\omega_1 \cup \omega_2 \cup \dots) = Pr(\omega_1) + Pr(\omega_2) + \dots \tag{4.4}$$

It should be noted that in practical situations one usually considers a subset $\mathfrak{G}$ of $\mathfrak{G}^\star$. This practice removes the need of explicitly assigning a probability for every single $\omega \in \mathfrak{G}^\star$. It rather focuses on the ones which correspond to the information attempted to be inferred by the observer of the random process. For example, considering an experiment defined by two consecutive coin tosses, one might be interested in the probability of tails being realized in the second toss. The event corresponding to this outcome is $\omega_1 = \{ht, tt\}$. Considering the information of interest, a reduced set of events which considered for this experiment are $\mathfrak{G} = \{\omega_1, \overline{\omega_1}\}$, where $\overline{\omega_1} \cup \omega_1 = \Omega$ with $\overline{\omega_1}$ being the complement of ω_1. The triplet $(\Omega, \mathfrak{G}, Pr)$ constitutes the *probability space* of the random process [Kaipio and Somersalo, 2005].

4.2.3 Random variables

In the heart of probabilistic methods lies the concept of the random variable (RV). In this work RVs are denoted by boldfaced letters, for example **x**. Given an experiment with a sample space $\Omega_{\mathbf{x}}$, **x** is defined as a function which maps events $\mathfrak{G}^\star$ to numbers $x \in \mathcal{R}(\mathbf{x})$, or

$$\mathbf{x}(\omega) = x. \tag{4.5}$$

If $\mathcal{R}(\mathbf{x}) = \mathbb{R}$, the RV is continuous whereas if $\mathbf{x}$ is discrete if it is only defined at given values $\mathcal{R}(\mathbf{x}) = \{x_i\}, i = 1, 2, \ldots$. Consider also two additional concepts, the *trial* and the *sample*. A trial refers to a single performance of a random experiment where an outcome ω_j gives rise to a particular value $x_k = \mathbf{x}(\omega_j)$. A sample A (denoted by capital letters) is comprised for a collection of N trials

$$A = \{a_1, a_2, \ldots, a_N\} \tag{4.6}$$

where $a_i = x_k$ denotes the value of $\mathbf{x}$ realized in the i^{th} trial. It should be noted that different trials can be assigned with equivalent values of x.

4.2.4 Probability distribution and density functions

Probability distributions Let $\omega_i,\ i \in I$ denote all events for which $\mathbf{x}(\omega_i) \leq x$ or under the more compact notation $\mathbf{x} \leq x$ (the event ω_i is not explicitly shown). The probability distribution function of of $\mathbf{x}$ is defined as

$$\mathtt{F}_{\mathbf{x}}(x) \equiv Pr(\mathbf{x} \leq x),\ \forall x \in [a, b] \tag{4.7}$$

where $Pr(\cdot)$ denotes the probability of some event, $\mathtt{F}_{\mathbf{x}}(a) = 0$, $\mathtt{F}_{\mathbf{x}}(b) = 1$ and $\mathtt{F}_{\mathbf{x}}(x + h) \geq \mathtt{F}_{\mathbf{x}}(x),\ \forall h \geq 0$.

Probability mass and density functions The probability density function (PDF) $p_{\mathbf{x}}(x)$ of an RV $\mathbf{x}$ is a function whose integral within some interval $[x_a, x_b] \in \Omega$ equals $Pr(x_a \leq \mathbf{x} \leq x_b)$. It is defined as the derivative of $\mathtt{F}_{\mathbf{x}}(x)$ with respect to x

$$p_{\mathbf{x}}(x) = \frac{\mathrm{d}\mathtt{F}_{\mathbf{x}}(x)}{\mathrm{d}x} \tag{4.8}$$

$$= \lim_{\Delta x \to 0} \frac{\mathtt{F}_{\mathbf{x}}(x + \Delta x) - \mathtt{F}_{\mathbf{x}}(x)}{\Delta x} \geq 0,\ \forall x \tag{4.9}$$

where $p_{\mathbf{x}}(x) \geq 0,\ \forall x$ due to the monotonically increasing nature of $\mathtt{F}_{\mathbf{x}}(x)$. For any PDF, $\int_{\mathcal{R}(\mathbf{x})} p_{\mathbf{x}}(x)\mathrm{d}x = 1$ should always hold.

Discrete case In the case of a discrete RVs $\mathbf{x}$ the concept of a continuous density is not defined. It is replaced by a collection of discrete masses centered at discrete RV values. The probability mass function is defined as [Papoulis and Pillai, 2001]

$$P_{\mathbf{x}}(x) = \sum_{x_i \in \mathcal{R}(\mathbf{x})} Pr(\mathbf{x} = x_i)\delta(x - x_i) \tag{4.10}$$

where $\delta(x - x_i)$ is the Kronecker's delta defined in Eq. 2.12. Note that the capital $P(\cdot)$ is explicitly used for the discrete case. The discrete x_i correspond to the jump-discontinuity points of $\mathtt{F}_{\mathbf{x}}(x)$ in 4.2a. In addition it holds that $\sum_{x_i \in \mathcal{R}(\mathbf{x})} P_{\mathbf{x}}(x_i) = 1$.

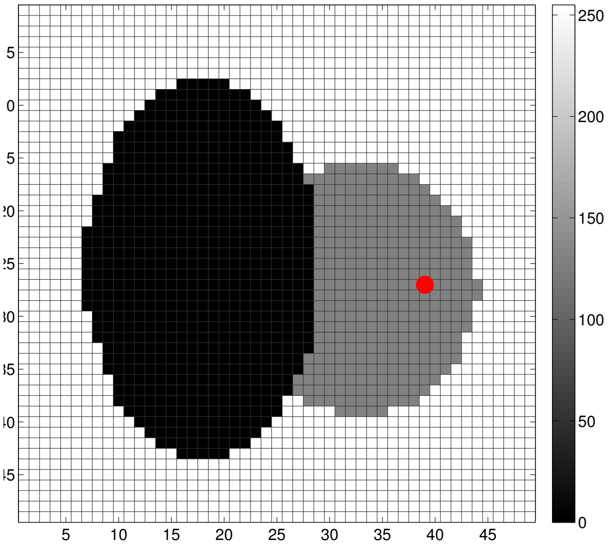


Figure 4.1: Images and randomness. A randomly generated point is depicted (see text)

4.2.5 Images and randomness

We pause the reference to the underlying theory in order to draw an analogy between the concepts described in this section and the imaging setting. Consider the image of size N depicted in Fig. 4.1. It is printed on the page, its grey values, structure and size are fixed and there is no obvious connection to the notion of randomness. In order to enable the application of statistical tools - such as the information theoretic concepts introduced in this chapter - one is required to introduce the notion of randomness. As there is no apparent random process, it is *up to us* to devise one.

Consider the following. The reader can cover the image with a piece of paper which only only depicts the pixel grid of the image, but not the actual grey value information. Then he/she can ask a second person to randomly choose a pixel location $r_i,\ i = 1, 2, \ldots N$ - such as the one depicted by the red dot. In this context, the random process regards the generation of r_i. This description gives rise to a sample space $\Omega_\zeta = \{\{r_1\}, \{r_2\}, \ldots, \{r_N\}\}$ comprised by all possible outcomes r_i. There are 2^N potential events in $\mathfrak{S}^\star$, for example $\{r_1\}$, $\{r_{13}\}$ or $\{r_7, r_4, r_N\}$, where the latter regards the case of r_i belonging to a collection of pixels locations. Because we do not want to compute the probability for all these events, we select the subset $\mathfrak{S} \subset \mathfrak{S}^\star$ comprised solely with the same events of Ω_ζ

A random variable **x** is now devised which maps events r to a number x. In this case *we* choose the number x to take three discrete values, equivalent to three distinct grey levels of the image. These are $x = \{0 \text{ (black)}\ ,\ 128 \text{ (grey)}\ , 255 \text{ (white)}\ \}$. We could have chosen a x of different nature, for example the arbitrary $\mathbf{x}(r_i) = r_i^2$, however grey values are preferable in this case.

One can now ask the question, what is the probability $\{\mathbf{x}(r_i) = 255\}$?. It is sensible to expect that as the number of white pixels is greater than the other two grey values, the event[1] $\mathbf{x}(r_i) = 255$ would be more probable. Note that the choice of r_i must be completely random. If the person who chooses r_i is biased and selects locations in the middle of the image, then the probability of $\mathbf{x}(r_i) = \{\{0\}, \{128\}\}$

[1] the actual event is the generation of r_i and however we now refer to the value assigned by **x** as the event

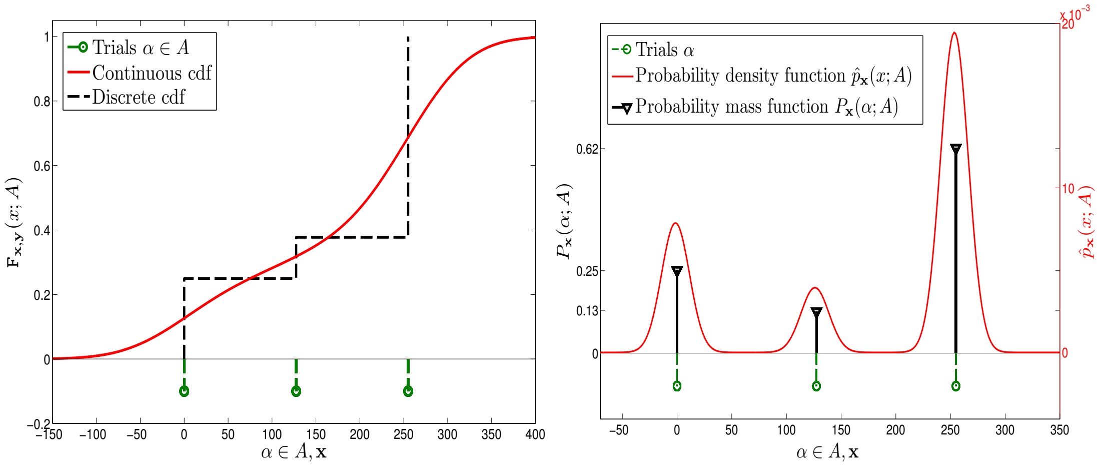


Figure 4.2: Examples of probability distribution - (discrete/continuous) (Fig. 4.2a) and probability mass/density functions (Fig. 4.2b). The continuous distribution and density are estimated from a finite sized sample A. The trials α depict the values of the trials. Each depicted trial actually corresponds to numerous trials sharing the same value x

becomes dominant. To avoid possible bias a random process involving the generation of two or more r, we would require the sampled locations to be *independent*, and *identically distributed* (i.i.d.) in a uniform manner.

Assuming the i.i.d. condition, the probability $Pr(\mathbf{x} = x_k)$ is given by the probability mass function depicted in Fig. 4.2b. Apparently, the amplitude of the individual masses reflect on the size of the three regions. The $Pr(\mathbf{x} = x_k)$ values have a frequency interpretation, corresponding to the number of pixels x_k considering the full size N of the image. Effectively, $Pr(\mathbf{x} = x_k) = N_{x_k}/N$.

The continuous $p_{\mathbf{x}}(x)$ assumes that $x \in \mathbb{R}$. In that case, although not reflected on the image itself, one would expect that values $x \rightarrow 255$ would be more probable than values close to 0 or 128. Finally, the image itself can considered as a sample A of trials $\alpha = \mathbf{x}(r_i) = x_k$, where all possible r_i are considered.

4.2.6 Expectation, variance and standard deviation of a random variable and sample statistics

Expected value Predicting the outcome of a single trial of an RV $\mathbf{x}$ cannot be done with certainty. There are however secondary quantities expressed in terms of $\mathbf{x}$, which are not considered to be random. An example of such quantity is the long term average value of $\mathbf{x}$ known as the *expected value*. Using the fair coin toss paradigm, let $\mathbf{x}$ describe the realization of *heads* such as $\mathbf{x}(\omega = heads) = 1,\ \mathbf{x}(\omega \neq h) = 0$. For a number of coin tosses $N \gg 0$, the number of $\mathbf{x} = 1$ would approach $N/2$. For an RV $\mathbf{x} \in \mathcal{R}(\mathbf{x})$ the expected value or its *mean* $E[\mathbf{x}]$ (also denoted as $\bar{\mathbf{x}}$) is defined as

$$E[\mathbf{x}] = \begin{cases} \int_{x \in \mathcal{R}(\mathbf{x})} x\ p_{\mathbf{x}}(x)\mathrm{d}x, & \text{continuous case} \\ \sum_{x_i \in \mathcal{R}(\mathbf{x})} x_i\ P_{\mathbf{x}}(x_i), & \text{discrete case.} \end{cases} \tag{4.11}$$

.

Variance and standard deviation The *variance* $Var(\mathbf{x})$ of an RV- also denoted as $\sigma^2_{\mathbf{x}}$ - is a measure of the dispersion of the random variable around its mean $\bar{\mathbf{x}}$. It is defined as the average squared difference of every value of the possible realizations of $\mathbf{x}$ from the mean $\bar{\mathbf{x}}$ or

$$Var\left(\mathbf{x}\right) = \sigma^2_{\mathbf{x}} = E\left[(\mathbf{x} - \bar{\mathbf{x}})^2\right] = \begin{cases} \int_{x \in \mathcal{R}(\mathbf{x})} \left(x - E_{\mathbf{x}}\left[\mathbf{x}\right]\right)^2 p_{\mathbf{x}}(x)\mathrm{d}x, & \text{continuous case} \\ \sum_{x_i \in \mathcal{R}(\mathbf{x})} \left(x_i - E_{\mathbf{x}}\left[\mathbf{x}\right]\right)^2 P_{\mathbf{x}}(x_i), & \text{discrete case.} \end{cases} \tag{4.12}$$

Taking the square of the difference ensures that the amplitude of the deviation from the mean is measured irrespective of the direction of the deviation. The exponent however propagates to the actual unit of variance, which now equals the squared unit of $\mathbf{x}$. By taking the positive root of $Var\left(\mathbf{x}\right)$, a measure of spread is obtained in the original units. This measure is the *standard deviation* denoted by $\sigma_{\mathbf{x}}$ given by

$$\sigma_{\mathbf{x}} = \sqrt{Var\left(\mathbf{x}\right)} \tag{4.13}$$

Sample statistics The true expectation $E\left[\mathbf{x}\right]$ and variance $Var\left(\mathbf{x}\right)$ of $\mathbf{x}$ are *population parameters*. As the population in its totality can be unavailable or too large to process, an estimate is usually retrieved from an N_A sized sample A of $\mathbf{x}$. If the sample based estimate matches the true population parameter, the estimator is called *unbiased* [Spiegel and Stephens, 2008]. The sample mean is defined as

$$E_A\left[\mathbf{x}\right] = \frac{1}{N_A}\sum_{i}^{N_A} a_i \tag{4.14}$$

Contrary to $E\left[\mathbf{x}\right]$, the sample mean is an RV [Viola, 1995]. For large N_A

$$\lim_{N_A \to \infty} E_A\left[\mathbf{x}\right] = \lim_{N_A \to \infty} \frac{1}{N_A}\sum_{i}^{N_A} a_i \to E\left[\mathbf{x}\right] \tag{4.15}$$

The sample mean $E_A\left[\mathbf{x}\right]$ (also denoted as $\bar{A}$) is an unbiased estimator of the true expectation $E_{\mathbf{x}}\left[\mathbf{x}\right]$ given that $N_A \to \infty$ [Viola, 1995]. Similarly, an unbiased estimator for the sample variance - given that $\bar{A}$ has been estimated from sample and is not know *a priori*- is given by [Papoulis and Pillai, 2001]

$$Var_A\left(\mathbf{x}\right) = \frac{1}{N_A - 1}\sum_{i=1}^{N_A}(x_i - \bar{A})^2 \tag{4.16}$$

4.2.7 Systems of two random variables

Often the physical system under investigation involves more than one random processes. Each process is described by a dedicated RV. Assume the simplest case of a multivariate system, one that is described by two RVs $\mathbf{x}$ and $\mathbf{y}$. Although random to some degree, the outcome of one process can be dependent on already realized outcomes of the other. This implies a functional relationship $\mathbf{y} = f(\mathbf{x})$. Dependency among RVs cannot be determined by the distributions and densities defined in Sec. 4.2.4. One has to consider their *joint* behavior. The bi-variate setting gives rise to the joint event

$$\{\mathbf{x} = x, \mathbf{y} = y\}\,, \quad \text{discrete RV case} \tag{4.17}$$

$$\{\mathbf{x}, \mathbf{y}\} \in D\,,\ D = \big\{\{x_a, x_b\}, \{y_a, y_b\}\big\} \subseteq \Omega_{\mathbf{x},\mathbf{y}}, \quad \text{continuous RV case.} \tag{4.18}$$

The bi-variate analogues of Sec. 4.2.4 are the joint distribution $\mathtt{F}_{\mathbf{x},\mathbf{y}}(x, y)$ and joint probability mass/density functions (JPDF) $P_{\mathbf{x},\mathbf{y}}(x, y)$, $p_{\mathbf{x},\mathbf{y}}(x, y)$. They correspond to the probability of the joint event being realized, for discrete and continuous RVs respectively. Two RVs are statistically independent [Papoulis and Pillai, 2001] if

$$p_{\mathbf{x},\mathbf{y}}(x, y) = p_{\mathbf{x}}(x) \cdot p_{\mathbf{y}}(y). \tag{4.19}$$

In the context of the joint setting, the statistics of the individual RVs are called *marginal*. This gives rise to the marginal PDFs

$$p_{\mathbf{x}}(x) = \sum_{y \in \mathcal{R}(\mathbf{y})} p_{\mathbf{x},\mathbf{y}}(x, y), \quad \text{and} \quad p_{\mathbf{y}}(y) = \sum_{x \in \mathcal{R}(\mathbf{x})} p_{\mathbf{x},\mathbf{y}}(x, y). \tag{4.20}$$

Another function which arises in the joint setting is probability density of **x**, *conditioned* on the fact that some event $\mathbf{y} = y$ has already been observed. The conditional probability $p_{\mathbf{x},\mathbf{y}}(x \mid y)$ is expressed as

$$p_{\mathbf{x},\mathbf{y}}(x \mid y) = \frac{p_{\mathbf{x},\mathbf{y}}(x, y)}{p_{\mathbf{y}}(y)}. \tag{4.21}$$

Finally, the *Bayes'* formula relates the two conditional probability functions rising from **x** and **y**, according to

$$p_{\mathbf{x},\mathbf{y}}(x \mid y) = \frac{p_{\mathbf{y},\mathbf{x}}(y \mid x) p_{\mathbf{x}}(x)}{p_{\mathbf{y}}(y)} \tag{4.22}$$

Randomness in pairs of images In an analogy to the discussion of Sec. 4.2.5, a random process in a system of two images randomly probes two images **x**, **y** at corresponding spatial locations r_i. There is now uncertainty in predicting joint events of type $\{\mathbf{x}(r_i) = x_k, \mathbf{y}(r_i) = y_l\}$. The probability of specific outcomes arising is provided by $p_{\mathbf{x},\mathbf{y}}(x, y)$. Figure 4.3 depicts the formed probability distribution and mass/density functions by the two images. Apparently, the probability of the outcome $\{\mathbf{x} = 255, \mathbf{y} = 0\}$ is dominant due to the size of the overlap between the regions populated with these values.

4.3 The Normal density

Of the various density functions that have been investigated, none has drawn more attention other than the normal density - also known as *Gaussian* density [Duda et al., 2001]. A justification for the increased

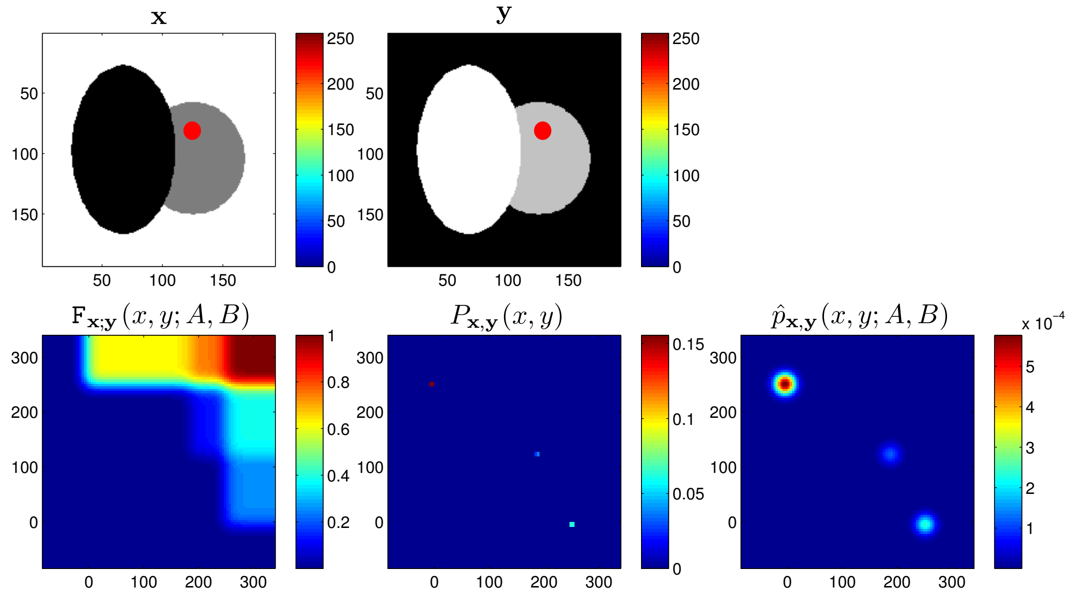


Figure 4.3: Joint probability distribution and mass/density functions between two images

attention of the normal density comes from the *central limit theorem*. Given N independent random variables $\mathbf{x}_i$ of arbitrary densities $p_{\mathbf{x}}(x)$, then the sum $\mathbf{x} = \sum_i^N \mathbf{x}_i$ constitutes an RV with mean $\mu = \sum_i^N \mu_i$ and variance $\sigma^2 = \sum_i^N \sigma_i^2$, where μ_i, σ_i are the means and standard deviations of the individual RVs. The central limit theorem states that as N increases, $p_{\mathbf{x}}(x)$ approaches the normal density [Papoulis and Pillai, 2001; Rice, 2001]. In essence the central limit theorem states that the aggregate effect of the sum of a large number of small, independent random disturbances will lead to a normal density [Duda et al., 2001]. The normal density is defined as

$$\mathcal{N}(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{4.23}$$

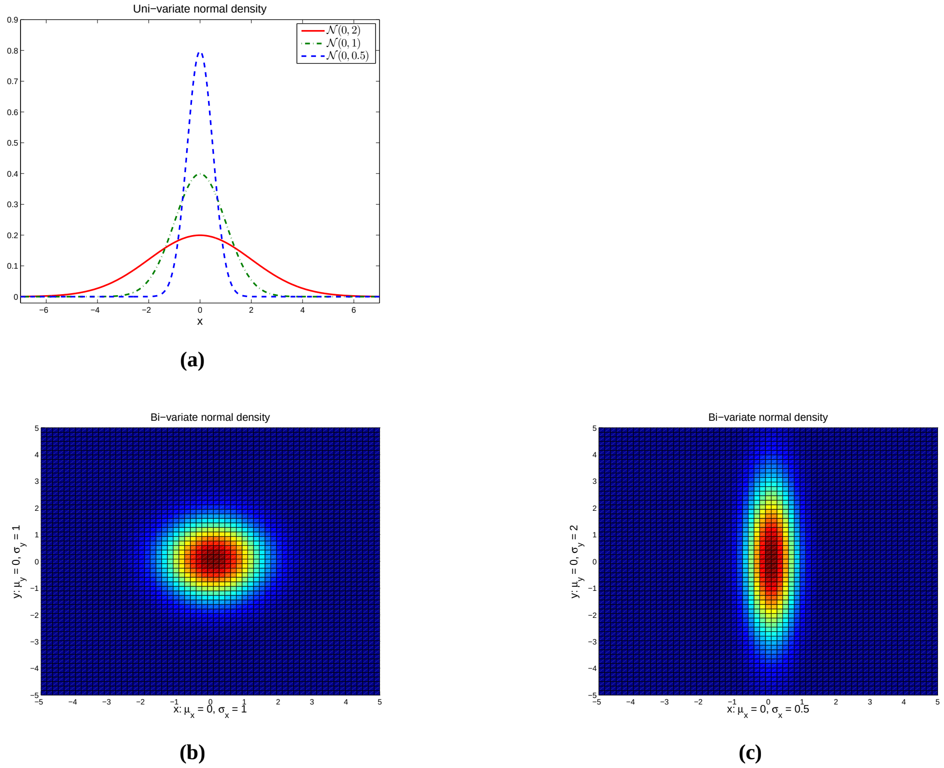
where μ and σ^2 denote the mean and variance of the density. $\mathcal{N}(\mu, \sigma^2)$. $\mathbf{x} \sim \mathcal{N}(\mu, \sigma^2)$ denotes that $\mathbf{x}$ is normally distributed. Examples of uni-variate normal densities are shown in Fig. 4.4a. The multivariate normal density in n dimensions is defined as

$$\mathcal{N}(\mathbf{M}, \mathbf{\Sigma}) = p_{\mathbf{X}}(X) = \frac{1}{(2\pi)^{d/2}\,|\mathbf{\Sigma}|} \exp\left[-\frac{1}{2}(\mathbf{X}-\mathbf{M})^{\mathrm{T}}\mathbf{\Sigma}^{-1}(\mathbf{X}-\mathbf{M})\right] \tag{4.24}$$

where $\mathbf{X}$ now refers to a $d-$dimensional column vector of the individual RVs $\mathbf{x}_i$, $\mathbf{M}$ is a $d-$component vector of consisting of μ_i, $\mathbf{\Sigma}$ is a $d \times d$ covariance matrix with $|\mathbf{\Sigma}|$ and $\mathbf{\Sigma}^{-1}$ being its determinant and inverse [Duda et al., 2001]. If $\forall i, j$ it holds that $\mathbf{x}_i$ and $\mathbf{x}_j$ are independent, then all the off diagonal components of $\mathbf{\Sigma}$ are zero. Examples of multivariate normal densities are shown in Fig. 4.4b-4.4c.

4.4 Density estimation

The task of estimating PDFs is central to probabilistic inference. Two of the different approaches towards density estimation include *parametric* and *non parametric* techniques. The following sections briefly



(a)

(b)

(c)

Figure 4.4: Examples of normal densities. 4.4a Uni-variate densities $\mathcal{N}(0, 0.5), \mathcal{N}(0, 1), \mathcal{N}(0, 2)$. 4.4b Isotropic bi-variate density with $\mu_x = 0, \mu_y = 0, \sigma_x = 1, \sigma_y = 1$. 4.4c Anisotropic bi-variate density with $\mu_x = 0, \mu_y = 0, \sigma_x = 0.5, \sigma_y = 2$

introduce them.

4.4.1 Parametric techniques

Parametric techniques assume that the true $p_{\mathbf{x}}(x)$ approaches some already known, parametrically defined PDF class. For example assume that **x** is normally distributed or $p_{\mathbf{x}}(x) \sim \mathcal{N}(\mu, \sigma^2)$. In that case the task of estimating $p_{\mathbf{x}}(x)$ simply reduces to the estimation of μ and σ^2. Under a more general setting the parameters which completely describe the assumed parametric form of $p_{\mathbf{x}}(x)$ are denoted as θ. The dependence of **x** to θ is explicitly expressed in the form of a conditional density as $p_{\mathbf{x}}(x \mid \theta) \sim \mathcal{N}(\mu, \sigma^2)$.

4.4.1.1 Maximum likelihood estimator

Let $\theta^\star$ denote the true unknown parameters. The maximum likelihood (ML) estimator assumes that $\theta^\star$ is a fixed, non-random quantity [Papoulis and Pillai, 2001]. Let $A = \{a_1, a_2, \ldots, a_N\}$ be a sample of **x**, comprised by i.i.d. trials drawn from $p_{\mathbf{x}}(x)$. As $\theta^\star$ uniquely determines $p_{\mathbf{x}}(x)$, it should also determine the distribution of the trials $a_i \in A$. The dependence of A to θ is expressed by $L(A|\theta) = p_{\mathbf{x}}(A \mid \theta)$ interpreted as the *likelihood of the sample* A [Viola, 1995]. It corresponds to the probability of A being realized for given θ. Due to the i.i.d. and the multivariate version of Eq. 4.19, the joint probability of the realizations in the sample is expressed as

$$L(A|\theta) = \prod_{a_i \in A} p_{\mathbf{x}}(a_i \mid \theta) \tag{4.25}$$

By definition, the ML estimate $\hat{\theta}_{ML}$ of $\theta^{\star}$ is the one that maximizes the likelihood of the sample [Papoulis and Pillai, 2001]

$$\hat{\theta}_{ML} = \arg\max_{\theta} \left[L(A|\theta) = p_{\mathbf{x}}(A \mid \theta)\right]. \tag{4.26}$$

It is typical to maximize the $\log$ of Eq. 4.25 as the product is turned to a sum and the optima are preserved due to the monotonic transformation by the $\log$ transform

$$\hat{\theta}_{ML} = \arg\max_{\theta} \left[l(A|\theta) = \log\left(p_{\mathbf{x}}(A \mid \theta)\right)\right] \tag{4.27}$$

$$= \arg\max_{\theta} \left[l(A|\theta) = \sum_{a_i \in A} p_{\mathbf{x}}(a_i \mid \theta)\right] \tag{4.28}$$

where we have used the same notation $\hat{\theta}_{ML}$ for the estimates retrieved by both standard ML and its logarithmically transformed variant. As it was noted, parametric methods assume that $p_{\mathbf{x}}(A \mid \theta)$ is given. To give an example of ML estimation in action, assume that $\mathbf{x} \sim \mathcal{N}(\mu, \sigma^2)$. Then $l(A; \mu, \sigma^2) = \sum_{a_i \in A} \log\left(\frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(a_i-\mu)^2}{2\sigma^2}\right]\right)$. By differentiating and setting the derivatives to zero, the obtained parameter estimates are $\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} a_i$ and $\hat{\sigma^2} = \frac{1}{N}\sum_{i=1}^{N}(a_i - \hat{\mu})$ [Duda et al., 2001]. These expressions show that the estimated mean and variance of $\mathbf{x}$ based on a sample A is the mean and variance of the true distribution $\mathcal{N}(\mu, \sigma^2)$.

4.4.1.2 Bayesian estimation

Contrary to the ML estimator, the Bayesian approach regards θ as an RV- hence denoted as $\boldsymbol{\theta}$ - which are distributed according to some PDF $p_{\boldsymbol{\theta}}(\theta)$. The term $p_{\boldsymbol{\theta}}(\theta)$ constitutes *a priori* information which the Bayesian approach enables its incorporation in the estimation. From Bayes Law (Eq. 4.22) and by considering a sample A, the estimator is defined as

$$p_{\boldsymbol{\theta}}(\theta \mid A) = \frac{p_{\mathbf{x}}(A \mid \theta) p_{\boldsymbol{\theta}}(\theta)}{p_{\mathbf{x}}(A)} \tag{4.29}$$

where $p_{\boldsymbol{\theta}}(\theta \mid A)$ is known as the *posterior density* expressing the probability of the parameters of interest given a sample A of $\mathbf{x}$, $p_{\mathbf{x}}(A \mid \theta)$ is the likelihood function defined earlier, $p_{\boldsymbol{\theta}}(\theta)$ is the *prior density* and $p_{\mathbf{x}}(A) = \int p_{\mathbf{x}}(A \mid \theta) p_{\boldsymbol{\theta}}(\theta) \mathrm{d}\theta > 0$ is the total probability of A [Duda et al., 2001; Papoulis and Pillai, 2001]. The second conceptual difference between the Bayesian and ML approach is that the former retrieves a density function - the posterior - and not a point estimate. Given $p_{\boldsymbol{\theta}}(\theta \mid A)$, it is now possible to derive point based estimators such as the *maximum a posteriori* defined as

$$\hat{\theta}_{MAP} = \arg\max_{\theta} \left[p_{\boldsymbol{\theta}}(\theta \mid A) \propto p_{\mathbf{x}}(A \mid \theta) p_{\boldsymbol{\theta}}(\theta)\right] \tag{4.30}$$

where the denominator $p_{\mathbf{x}}(A)$ in Eq. 4.29 can be omitted as it simply acts as a normalising constant [Kaipio and Somersalo, 2005] and does not depend on θ. In addition to point based estimators, the Bayesian approach enables interval estimates such as

$$p\left\{\theta_1 < \boldsymbol{\theta} < \theta_2\right\} = \gamma \tag{4.31}$$

where $1-\gamma$ is the confidence level [Papoulis and Pillai, 2001]. More detailed explanation can be found in [Duda et al., 2001; Gelman et al., 2003; Papoulis and Pillai, 2001].

4.4.2 Non parametric kernel density estimation

Contrary to the methods introduced in the previous section, non parametric techniques do not assume that the form of the underlying densities is known. The estimation of the density is based solely on the available sample. This results in more generic estimators which can model densities of arbitrary shape. One such method which is also employed in this work is the non-parametric kernel density estimation (KDE) - also commonly referred to as the Parzen window estimator [Duda et al., 2001; Izenman, 1991; Parzen, 1962; Silverman and Green, 1986; Simonoff, 1996; Wand and Jones, 1995].

Let $p_{\mathbf{x}}(x)$ be the true PDF of **x** and A a drawn sample. For the uni-variate case, an estimate $\hat{p}_{\mathbf{x}}(x; A)$ of the true $p^{\star}_{\mathbf{x}}(x)$ can be obtained at some x, by utilizing a $N-$sized sample A of **x** via

$$\hat{p}_{\mathbf{x}}(x; A) = \frac{1}{N}\sum_{i=1}^{N} K_u(x-\alpha_i). \tag{4.32}$$

Its bi-variate analogue, utilizes two N-sized samples A and B and the estimated joint probability density function (JPDF) is given by

$$\hat{p}_{\mathbf{x},\mathbf{y}}(x, y; A, B) = \frac{1}{N}\sum_{i=1}^{N_{\mathrm{y}}} K_{\mathbf{\Sigma}}(x-\alpha_i, y-\beta_i). \tag{4.33}$$

In both expressions, $K_u(\cdot)$, $K_{\mathbf{\Sigma}}(\cdot,\cdot)$ are *kernels* or *window functions* of width u or covariance $\mathbf{\Sigma}$ respectively. In essence, the window function weights the contribution of its trial $a_i \in A$ according to each distance from x. If $K_u(\cdot)$ is smooth and differentiable and satisfies $\int K_u(x)dx = 1$, then the estimate $\hat{p}_{\mathbf{x}}(\mathrm{x}; A)$ will be continuous and differentiable Duda et al. [2001]. A common choice is the normal density. Alternative distributions which have been used as kernels in the literature include the Uniform, Triangle, Epanechnikov and Cosine kernels [Silverman and Green, 1986].

Regarding the retrieval of accurate PDF estimates, the choice of the kernel width u is more important than the choice of the actual kernel type [Silverman and Green, 1986]. Various methods for u selection are discussed in literature such as the L-curve or generalized cross-validation. For a detailed introduction on kernel width selection methods, the reader is is redirected to specialized literature and the references within [Hall et al., 1991; Hansen, 1992a, 1998; Jones et al., 1996; Raykar and Duraiswami, 2006; Silverman and Green, 1986; Turlach, 1993; Vogel, 2002]. In addition, many optimization problems which involve a KDE incorporate the bandwidth of the kernel in the overall set of optimized parameters - see for example [Kazantsev et al., 2010; Viola, 1995].

Finally, KDE can be characterized by high levels of computational complexity. Estimating the $\hat{p}_{\mathbf{x}}(x; A)$ for M values of x (see Eq. 4.32) results in a complexity of $\mathcal{O}(NM)$, which is dependent on the sample size. Numerous methods have been proposed to reduce of complexity. These include data reduction methods which utilize binning or clustering strategies to reduce the size of the available data and approximate kernel decompositions which decouple the point x where the density is estimated from the remaining sample points A, so the summation over A in Eq. 4.32 can be performed separately in a manner akin to orthogonal series density estimators [Izenman, 1991]. See [Girolami and He, 2003] and the references within for fast KDE estimation. Viola [Viola, 1995] used a stochastic approach to optimize entropy which requires a PDF estimate. In each each iteration of the optimization, the KDE utilized a randomly selected sample drawn from the RV sample A. Although this constitutes a data reduction strategy considering a single iteration, throughout the entire optimization routine the majority of the sample A was utilized.

Figure 4.5 shows a uni-variate density estimator in action computing an estimate from a finite sample from a continuous RV. The advantage of continuous density estimates over discrete ones - for example the depicted normalized histograms - is the potential differentiability of the former.

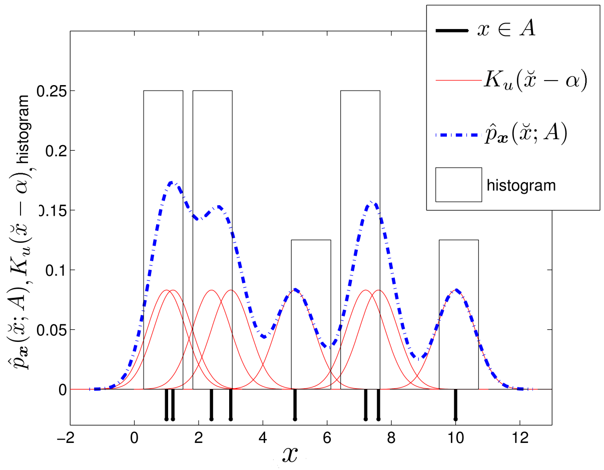


Figure 4.5: Non-parametric kernel density estimation. The black lines under the horizontal axes denote N samples a_i drawn from some continuous RV $\mathbf{x}$, forming the finite sample A. $p_{\mathbf{x}}(\mathrm{x}; A)$ denotes the continuous probability density estimate of $\mathbf{x}$, at regular spaced locations $\breve{x}$. $K_u(\breve{x} - a_i)$ denotes a Gaussian kernel centered at each sample point a_i. Effectively, the PDF estimate $\hat{p}_{\mathbf{x}}(\breve{x}_j; A)$ at a point $\breve{x}_j$, equals the sum of the contributions from all kernels $K_u(\breve{x}_j - a_i)$, $\forall i$.

4.5 Information theoretic functionals

4.5.1 Entropy

Section 4.1 introduced entropy as measure of uncertainty in the context of IT. Let $\mathbf{x} \sim P_{\mathbf{x}}(x)$ be a discrete RV, with $x = \{x_1, x_2, \dots, x_N\}$ and $p_{\mathbf{x}}(x)$.

Shannon in [Shannon, 1948] identified three properties which should characterize a measure $H\big(P_{\mathbf{x}}(x_i)\big)$ of the uncertainty in $\mathbf{x}$. These are:

1. H should continuously depend on $P_{\mathbf{x}}(x_i)$
2. If all events are equi-probable or equivalently $P_{\mathbf{x}}(x_i) = \frac{1}{N}$, $\forall i$, then H should be a monotonically increasing function of N. In other words given that all possible outcomes are equally likely to be realized, then as the number of possible outcomes increases so does the uncertainty of predicting their realization.
3. H should be independent on the grouping of events. For example let there be choice from three events $\Omega_\zeta = \{\{x_1\}, \{x_2\}, \{x_3\}\}$ with corresponding probabilities $P_i = P_{\mathbf{x}}(x_i), i = 1, 2, 3$ (see Fig. 4.6). The choice can broken down to two successive choices. The first is between $\big\{\{x_1\}, \{x_{2,3}\}\big\}$ with $x_{2,3} = \{x_2, x_3\}$ and corresponding probability $P_{2,3}$. On the condition that $x_{2,3}$ is selected, the second choice is between $\{\{x_2\}, \{x_3\}\}$. H should then equal the weighted sum of the individual values of H or

$$H(P_1, P_2, P_3) = H(P_1, P_{2,3}) + P_{2,3} \cdot H(P_{2|\{2,3\}}, P_{3|\{2,3\}}) \tag{4.34}$$

where $P_{i|\{2,3\}} = P_{\mathbf{x}}(x_i \mid x_{2,3})$

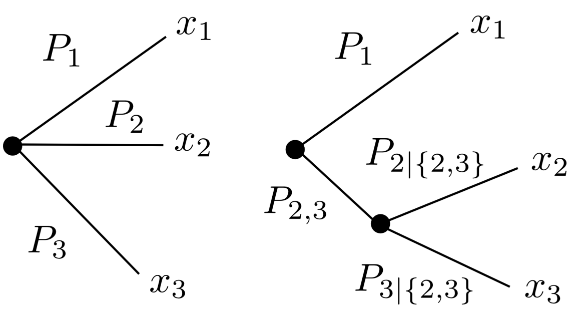


Figure 4.6: Entropy and independence to grouping of events. **Left schematic**: ungrounded events **Right schematic:** grouped events

Shannon proved that the only function satisfying all of the above postulates is the *entropy* of $\mathbf{x}$ defined as

$$H(\mathbf{x}) = -\sum_{x \in \mathcal{R}(\mathbf{x})} p_{\mathbf{x}}(x) \log\big(P_{\mathbf{x}}(x)\big) \tag{4.35}$$

$$= -\, E_{\mathbf{x}}\big[\log\big(P_{\mathbf{x}}(x)\big)\big] \tag{4.36}$$

An alternative derivation of the entropic functional can be found in [Schneider, 1995]. We briefly reproduce it in the context of imaging, due to its highly intuitive nature.

Let an image A of size N populated by three distinct grey values $\mathcal{R}(\mathbf{x}) = \{\{x_1\}, \{x_2\}, \{x_3\}\}$. Probing A at a randomly generated location r_i, can lead to three possible outcomes $\mathbf{x}(r) \in \mathcal{R}(\mathbf{x})$. We call this *uncertainty of three outcomes*. Consider now that the image is also probed in a second location r_j, $i \neq j$ and we seek to predict the outcome of the joint event $\{\mathbf{x}(r_i), \mathbf{x}(r_j)\}$. The joint event of this case now has nine potential outcomes, three $\mathbf{x}(r_j)$ for every one of the three potential $\mathbf{x}(r_i)$. This corresponds to *9-uncertainty*. If we would introduce a third process $\mathbf{x}(r_k)$, the number of joint outcomes would rise to twenty seven. To generalize, for q processes each with K_i outcomes, the number of joint outcomes (or joint uncertainty) is $\prod_i^q K_i$. It would be more convenient if the joint uncertainty would rise linearly with the total number of individual outcomes $K = \sum_i^q K_i$. The latter can be achieved by measuring the uncertainty with a logarithmic measure $u(K) = log(K)$.

Consider again the case of a single process with $K = 3$ equi-probable outcomes. Applying the uncertainty measure results in

$$u(K) = \log(K) \tag{4.37}$$

$$= -\log\left(\frac{1}{K}\right) \tag{4.38}$$

$$= -\log\big(P_{\mathbf{x}}(x_i)\big), \forall i = 1, \ldots, K \tag{4.39}$$

where we have employed the equality $\log(K^p) = p \cdot \log(K)$ and the frequency interpretation of probability. We know however that $0 \leq P_{\mathbf{x}}(x_i) \leq 1$, $\forall i$ (see Sec. 4.2.2). For highly probable x_i, $P_{\mathbf{x}}(x_i) \rightarrow 1$, leading to minimal uncertainty $u(P_{\mathbf{x}}(x_i)) \rightarrow 0^+$. On the contrary, for highly improbable x_i, $P_{\mathbf{x}}(x_i) \rightarrow 0^+$ and $u(P_{\mathbf{x}}(x)) \rightarrow \infty$. The nature of $u(P_{\mathbf{x}}(x_i))$ as a measure of uncertainty is now apparent. The term $u(P_{\mathbf{x}}(x_i))$ is also known as *surprisal*, in terms of observing x_i [Schneider, 1995; Tribus, 1961].

To complete the intuitive example, one needs to consider the case where the various x_i are not equi-probable. In that case it is reasonable to evaluate the *average uncertainty*, which can be expressed as an expectation (see Sec. 4.2.6) of the individual uncertainties of the possible events weighted by their probability of occurrence. This results in the entropy term of Eq. 4.35 and completes the derivation. The entropy $H\big(P_{\mathbf{x}}(x)\big)$ - or simply $H(\mathbf{x})$ - is a measure of the average uncertainty per outcome of $\mathbf{x}$. The units of $H(\mathbf{x})$ depends on the on the base b of the logarithm. In the case of $b = 2$, uncertainty is measured in *bits* and in the case of the natural logarithm, the unit is nats [Cover and Thomas, 1991]. For $H(\mathbf{x})$ it always holds

$$H(\mathbf{x}) \geq 0 \tag{4.40}$$

4.5.2 Joint entropy, conditional entropy and mutual information

The concept of entropy extends in the multi-variate setting giving rise to additional quantities. Assume the simplest case of a bi-variate system with RVs **x** and **y**.

Joint entropy JE $H(\mathbf{x},\mathbf{y})$ [Cover and Thomas, 1991] corresponds to the average uncertainty in predicting joint events $\{\mathbf{x}=x,\mathbf{y}=y\}$. It is defined as

$$\begin{aligned} H(\mathbf{x},\mathbf{y}) &= -\sum_{x\in\mathcal{R}(\mathbf{x})}\sum_{y\in\mathcal{R}(\mathbf{y})} P_{\mathbf{x},\mathbf{y}}(x,y)\log\left(P_{\mathbf{x},\mathbf{y}}(x,y)\right) && (4.41)\\ &= -E_{\mathbf{x},\mathbf{y}}\left[\log\left(P_{\mathbf{x},\mathbf{y}}(x,y)\right)\right] && (4.42)\end{aligned}$$

where $E_{\mathbf{x},\mathbf{y}}[\cdot]$ is the bi-variate expression of the expected value. For all cases the following holds:

$$H(\mathbf{x},\mathbf{y}) \leq H(\mathbf{x}) + H(\mathbf{y}) \tag{4.43}$$

with the equality being realized in the case of complete independence between the RVs. Similar to the marginal case, $H(\mathbf{x},\mathbf{y}) \geq 0$ always holds.

Conditional entropy Conditional entropy $H(\mathbf{x}|\mathbf{y})$ - also known as *equivocation* - corresponds to the average uncertainty of **x** given that **y** has been observed. Conditional entropy is expressed as

$$\begin{aligned} H(\mathbf{x}|\mathbf{y}) &= -\sum_{y\in\mathcal{R}(\mathbf{y})} p_{\mathbf{x},\mathbf{y}}(x,y)\log\left(P_{\mathbf{x},\mathbf{y}}(x|y)\right) && (4.44)\\ &= -E_{\mathbf{x},\mathbf{y}}\left[\log\left(P_{\mathbf{x}|\mathbf{y}}(x,y)\right)\right] && (4.45)\end{aligned}$$

If **x** depends on **y**, then by observing **y** the uncertainty of **x** should always decrease. This results to the inequality

$$H(\mathbf{x}\mid\mathbf{y}) \geq H(\mathbf{x}), \tag{4.46}$$

with equality being realized in the case of complete independence between **x** and **y**. In that case, knowing **y** does not reduce the initial uncertainty $H(\mathbf{x})$.

It should be emphasized that $H(\mathbf{x}\mid\mathbf{y})$ is not a measure of dependency [Viola, 1995]. Low $H(\mathbf{x}\mid\mathbf{y})$ values can be observed either due to high dependency between **x** and **y** or simply because **x** is inherently characterized by low uncertainty on its own - or equivalently $H(\mathbf{x})$ is low. In order to measure uncertainty, one needs to consider the relative decrease in uncertainty, from the initial value $H(\mathbf{x})$ to $H(\mathbf{x}\mid\mathbf{y})$. This is accomplished by MI introduced next.

Mutual information Mutual information was introduced by Shannon [1948] under the name *rate of transmission*. It is a *measure* of dependency and is expressed as

$$MI(\mathbf{x},\mathbf{y}) = MI(\mathbf{y},\mathbf{x}) \tag{4.47}$$
$$= H(\mathbf{x}) - H(\mathbf{x} \mid \mathbf{y}) \tag{4.48}$$
$$= H(\mathbf{y}) - H(\mathbf{y} \mid \mathbf{x}) \tag{4.49}$$
$$= H(\mathbf{x}) + H(\mathbf{y}) - H(\mathbf{x},\mathbf{y}) \tag{4.50}$$

The first two equations define the information between **x**, **y** as the reduction of the initial uncertainty after the observation of the second RV has taken place. This is consistent with the definition of information in the introductory section. Considering Eq. 4.43, then $MI(\mathbf{x},\mathbf{y}) \geq 0$ for all cases.

The relation between information theoretic functionals can be depicted a *Venn diagram* of Figure 4.7 [Cover and Thomas, 1991].

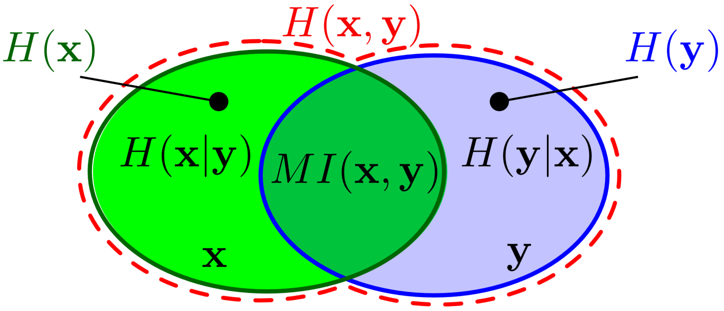

$H(\mathbf{x})$
$H(\mathbf{x},\mathbf{y})$
$H(\mathbf{y})$
$H(\mathbf{x}|\mathbf{y})$
$MI(\mathbf{x},\mathbf{y})$
$H(\mathbf{y}|\mathbf{x})$
x
y

Figure 4.7: Relationships among information theory concepts expressed in the form of a Venn diagram

4.5.3 Differential entropy

In the case of continuous RVs, then the sums of Eqs. 4.36, 4.42 and 4.45 are replaced by integrals and the probability masses with probability densities. The continuous entropy - also known as DE - is defined for both marginal and joint cases respectively as

$$h(\mathbf{x}) = \int_{x\in\mathcal{R}(\mathbf{x})} p_{\mathbf{x}}(x)\log\big(p_{\mathbf{x}}(x)\big)\mathrm{d}x, \quad \text{and} \tag{4.51}$$
$$h(\mathbf{x},\mathbf{y}) = \int_{x\in\mathcal{R}(\mathbf{x})}\int_{y\in\mathcal{R}(\mathbf{y})} p_{\mathbf{x},\mathbf{y}}(x,y)\log\big(p_{\mathbf{x},\mathbf{y}}(x,y)\big)\mathrm{d}x\,\mathrm{d}y. \tag{4.52}$$

Most inequalities of the discrete case hold in the continuous analogue, with some exceptions [Cover and Thomas, 1991; Shannon, 1948; Viola, 1995]. The most notable is that contrary to Eq. 4.40, $h(\mathbf{x})$ can be negative. This is due to the fact that $p_{\mathbf{x}}(x)$ in $u(\mathbf{x}) = -\log\big(p_{\mathbf{x}}(x)\big)$, are continuous densities which can take values greater than 1, resulting to negative entropy values. An extreme example regards the entropy of the Dirac delta function $\delta(x)$, for which it holds $\int_{-\infty}^{\infty}\delta(x)\mathrm{d}x = 1$ and

$$\delta(x) = \begin{cases} \infty, & x = 0 \\ 0, & x \neq 0 \end{cases}. \tag{4.53}$$

In this case, $h_{\mathbf{x}}\big(\delta(x)\big) = -\infty$. The continuous entropy may attain values of $+\infty$ or $-\infty$ [Ash, 1990]. In addition, contrary to the discrete case, continuous entropy does not measure uncertainty *in an absolute way* [Shannon, 1948]. Changes in the coordinate system - for example the spacing $\mathrm{d}x$ affects the entropic values. Contrary to the discrete case, DE is not invariant to all kind of transformations. It is invariant to translations or $h(\mathbf{x}+c) = h(\mathbf{x})$. However, it is not invariant to linear changes with $h(c\mathbf{x}) = h(\mathbf{x}) + \log|c|$ or in the case where $\mathbf{x}$ is a random vector subjected to a linear transformation by an operator F, then $h(F\mathbf{x}) = h(\mathbf{x}) + \log|\det(F)|$ [Cover and Thomas, 1991]. See Sec. 4.5.5.3 for a case study. For more general transformations $\mathbf{y} = f(\mathbf{x})$, the value of DE changes according to $h(\mathbf{y}) = h(\mathbf{x}) + \int p_{\mathbf{x}}(x)\log\left|\frac{\partial \mathbf{y}}{\partial x}\right|\mathrm{d}x$, with $\left|\frac{\partial \mathbf{y}}{\partial x}\right|\mathrm{d}x$ being the Jacobian of the transformation $\mathbf{y} = f(\mathbf{x})$ [Reza, 1994]. It has to be noted however that by assuming a *fixed* coordinate system, DE constitutes a measure of uncertainty relative to that system [Shannon, 1948; Viola, 1995]. It should be emphasized that mutual information is always positive even in the case that is expressed in differential entropy terms.

4.5.4 Empirical entropy

Consider the case of DE, as it is employed later in this Thesis for the purpose of information theoretic regularisation, mainly due to the continuous nature and differentiability of the functional. The entropic definitions assume that $p_{\mathbf{x}}(x)$ is known. In practice this is often not the case and the information one has regarding $\mathbf{x}$ is contained in some available sample. Let $A = \{a_1, a_2, \ldots, a_N\}$ be N-size sample of $\mathbf{x}$. One attempts to retrieve a continuous estimate $\hat{p}_{\mathbf{x}}(\mathbf{x}; A)$ (see Section 4.4.2) from A. Using the PDFs/ JPDFs, the *empirical* entropy $\hat{h}(\mathbf{x}; A)$ of the sample can be estimated. Firstly, one can approximate the integral formulation with numerical integration schemes such as the the trapezoidal rule. In that case $\hat{p}_{\mathbf{x}}(\mathbf{x}; A)$ is estimated in regularly spaced location x_i which partition $\mathcal{R}(\mathbf{x})$ in equal intervals of spacing Δx. The Shannon's integral formulation of empirical entropy can then be approximated by

$$\hat{h}_s(\mathbf{x}; A) = -\sum_{x_i \in \mathcal{R}(\mathbf{x})} \hat{p}_{\mathbf{x}}(x_i; A)\log\big(\hat{p}_{\mathbf{x}}(x_i; A)\big)\Delta x. \tag{4.54}$$

An alternative approach is to employ the expectation formulation of Eq. 4.36, which in this case corresponds to the sample mean (see Eq. 4.14) expressed as

$$\hat{h}_e(\mathbf{x}; A) = \sum_{i=1}^{N} \log\big(\hat{p}_{\mathbf{x}}(a_i; A)\big) \tag{4.55}$$

The discrete empirical entropy uses discrete estimates $P_{\mathbf{x}}(x; A)$, for example a normalized histogram. Finally, a crucial detail should be emphasized regarding Eq. 4.55. Assume that one employs a KDE which optimizes the kernel width u simultaneously with a minimization of the expectation formulation of the entropy. Viola [Viola, 1995] explained in p. 47 of his Thesis, that if the same sample is used for the estimation of $\hat{p}_{\mathbf{x}}(a_i; A)$ as well as for Eq. 4.55, then u will always converge to 0^+. This is expected, as the obtained $\hat{p}_{\mathbf{x}}(a_i; A)$ would consist of *Dirac* deltas centered at each a_i, a configuration which corresponds to the minimum $h(\mathbf{x}) = -\infty$. To bypass this problem Viola proposed splitting A in two parts, one for obtaining the PDF and the other for entropy estimation. In this work we do not optimize for u. The reader is directed to our recent work of applying information theoretic regularization in positron emission tomography (PET) simultaneously with a search for u [Kazantsev et al., 2010].

4.5.5 Images, information theory and multi-modality data

Section 4.2.5 introduced randomness in the imaging setting, in the context of predicting the grey value $x_k = \mathbf{x}(r_i)$[1] of an image at some random location r_i. Section 4.5.1 introduced the concept of *surprisal* or *uncertainty* in observing some x_k, given its probability of occurrence and the entropy which is the average uncertainty given possible events and weighted by their individual probability of occurrence. We now describe the information theoretic concepts in the imaging setting.

4.5.5.1 Minimum/maximum entropy images

Consider Fig. 4.8. It depicts a homogeneous image and an image with uniformly distributed grey values, both of size N. The $P_{\mathbf{x}}(x; \text{Image})$ are also depicted for each case. Regarding the first image, if one attempts to probe the image at random r_i, the resulting value will always be $x_k = 128$. This is depicted by $P_{\mathbf{x}}(x; \text{Image } 1) = 1$. The surprisal for $x_k = 128$ is $u(128) = -\log\left(P_{\mathbf{x}}(128; \text{Image } 1)\right) = 0$ and as $x = 128$ is the only possible outcome, $u(128)$ is also equivalent to the average surprisal or simply the empirical entropy $\hat{H}(\mathbf{x}; \text{Image } 1)$. There is no uncertainty in the outcome of this random process.

On the contrary, most pixels of the second image are assigned with different values. The probability of any outcome x_k is $P_{\mathbf{x}}(x_k; \text{Image } 2) = 1/N$. This is the case of highest uncertainty and Image 2 corresponds to the maximum entropy configuration with maximum surprisal $u(1/N)$. If one would consider a larger image with uniformly distributed values, the entropic content would be even higher as $1/N \downarrow$ which would lead to an increase in surprisal.

In general the following holds. Images with few features and small number of distinct grey values have low entropic content. Their PDFs are highly clustered, with few modes corresponding to high probability entries. On the contrary, highly disordered images with many distinct grey values have high entropic content. Their PDFs are highly dispersed, with many entries of low probability.

It is essential at this point to draw an analogy with the popular maximum entropy method (MEM) [Burch et al., 1983; Gull and Daniell, 1978; Jaynes, 1982, 1957a,b; Skilling and Bryan, 1984]. In the imaging context, MEM was proposed as a regularization method for the inverse problem in astronomy by Gull and Daniell [1978]. The favoured reconstructed image was the one which fitted the data but was also characterized by maximum entropy. However, in that context and contrary to what has been described in this section, the maximum entropy image was homogeneous. The reason for this complete reversal of entropic interpretation, is due to the employment of a different random process, devised to introduce randomness in the image. The random process in that setting described the formation of the image and not its probing in random locations. To describe the process, they assumed the proverbial team of monkeys with N photons in their disposal [Jaynes, 1984]. The photons were then thrown *randomly* by the monkeys, on a dark image defined by a grid of pixels of equal area. This could lead to the formation of various images, where N_i denotes the photons - or the grey value - in the i_{th} pixel. The entropy of

[1] It should be noted that k is different than i as the former indexes gray values, whereas the latter pixel locations. It is possible that $x_k = \mathbf{x}(r_i) = \mathbf{x}(r_j)$ for $i \neq j$

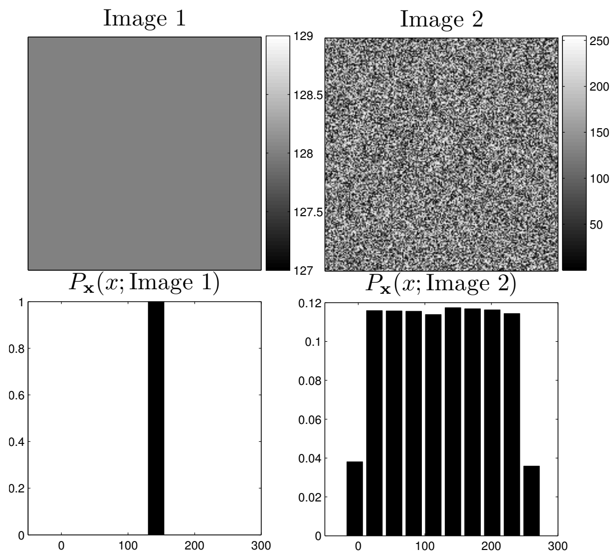


Figure 4.8: Minimum/maximum entropy images: Image 1 is homogeneous with a highly clustered PDF and corresponds to the minimum entropy configuration. Image 2 has uniformly distributed grey values with a highly dispersed PDF and corresponds to the maximum entropy configuration for an image of this size.

a given image was defined as the natural logarithm of the number of ways the monkeys could generate it, with the probability of a specific image being proportional to its multiplicity $N!/\prod_i N_i$. The most probable image and the one favoured by MEM, was the one with maximum multiplicity, which results in $N_i = 1/N, \forall i$. This image is formed in a setting of total uncertainty regarding the landing position of each photon, hence *maximum entropy method*. Using our interpretation of randomness, we could rename the method as *minimum entropy method*. It is worth to note that MEM is the *only* method which does not introduce correlations in the reconstructed image, beyond those that are required by the data. It is the *least-biasing* regularization method and *maximally non-committal* about what is not known regarding the solution [Skilling and Bryan, 1984].

Similar to the case of a single image, the minimum and maximum entropy image pairs are depicted in Fig. 4.9. In the first image pair, any randomly selected r_i returns a certain outcome $\{\mathbf{x}(r_i) = 128, \mathbf{y}(r_i) = 128\}$. The JPDF is maximally clustered, populated with one entry corresponding to the certain probability of the above event. On the contrary, two images with uniformly distributed grey values have the highest uncertainty with each outcome $\{\mathbf{x}(r), \mathbf{y}(r)\}$ having a probability $1/N$ - with N being the number of pixels. The JPDF is maximally dispersed as expected.

4.5.5.2 Information theoretic functionals and multi-modality

From the entropy definition of Eqs. 4.35, 4.51 and their multivariate counterparts, it is evident that both marginal and JE and consequently their derived functional MI, do not depend directly on the actual grey values of the images under consideration but rather on their corresponding probability masses/densities. Consequently, two images with different levels of *absolute* grey values, which however are characterized by similar probability distributions, are expected to have similar entropic content. It is exactly this non-direct dependence on the grey values which attributes to the functionals a level of inherent invariance

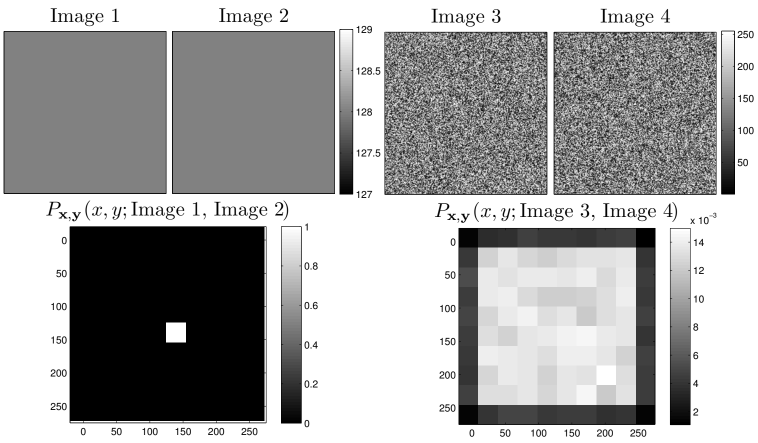


Figure 4.9: Minimum/maximum joint entropy images: The first pair is comprised by two homogeneous images with a highly clustered JPDF and corresponds to the minimum joint entropy configuration. The second pair consists of two images with uniformly distributed grey values and a highly dispersed (close to uniform) JPDF. It corresponds to the maximum joint entropy configuration for image pairs of this size.

to the incommensurate grey values between multi-modal images and makes them ideal as multi-modal structural similarity measures.

We now emphasize the differences between marginal entropy, JE and MI. Consider Figure 4.10.

Marginal entropy Image 2 is created by transforming the grey values of Image 1 according to an arbitrarily selected, non-linear, injective transformation $Im2 = -2(Im1)^2 + 3(Im1) - \log((Im1) + 1)$. The marginal entropy depends solely on the marginal PDFs depicted in the last row (reflected across the x-axes for visualization purposes). The two images have identical structure but the grey values populating the corresponding features are non-linearly related. However, the PDFs of both images are identical in entropic terms, as they assign equal probabilities to the same number of events. Entropy does not depend on the actual value of the events x - or equivalently on the location of the clusters in the PDF - but only on their probability. The entropic levels of the images are equal with $H(Im1) = H(Im2) = 0.9035\ nats$.

The third image in Fig. 4.10 reveals the lack of intra-image spatial dependence of entropy. Image 3 is created by applying a random spatial permutation on the pixels of Image 1, hence they both share identical PDFs and consequently entropic value. In this case, two structurally dissimilar images have equal entropic content. One can conclude that *the difference between the entropy of two images* cannot be used as a measure of their structural similarity.

JE/MI Consider now the JE and MI of three image pairs $\{\text{Image 1, Image Z}\}$, $Z = [1, 2, 3]$ in Fig. 4.10, as a measure of similarity between incommensurate images. Both JE and MI depend on the JPDFs of the formed image pairs, however the latter depends also on the the marginal terms (see Eq. 4.50). The axis of the JPDFs correspond to the grey value of the images involved in a pair and the entries describe the probability of $Pr(\{\mathbf{x}(r) = x, \mathbf{y}(r) = y\})$. The JPDF is constructed by sequentially accessing the N pixel positions of the involved images and for every position r, the algorithm retrieves the grey value pair $\{\mathbf{x}(r), \mathbf{y}(r))\}$ and increment its probability in the initially empty JPDF plane. This implies that both JE

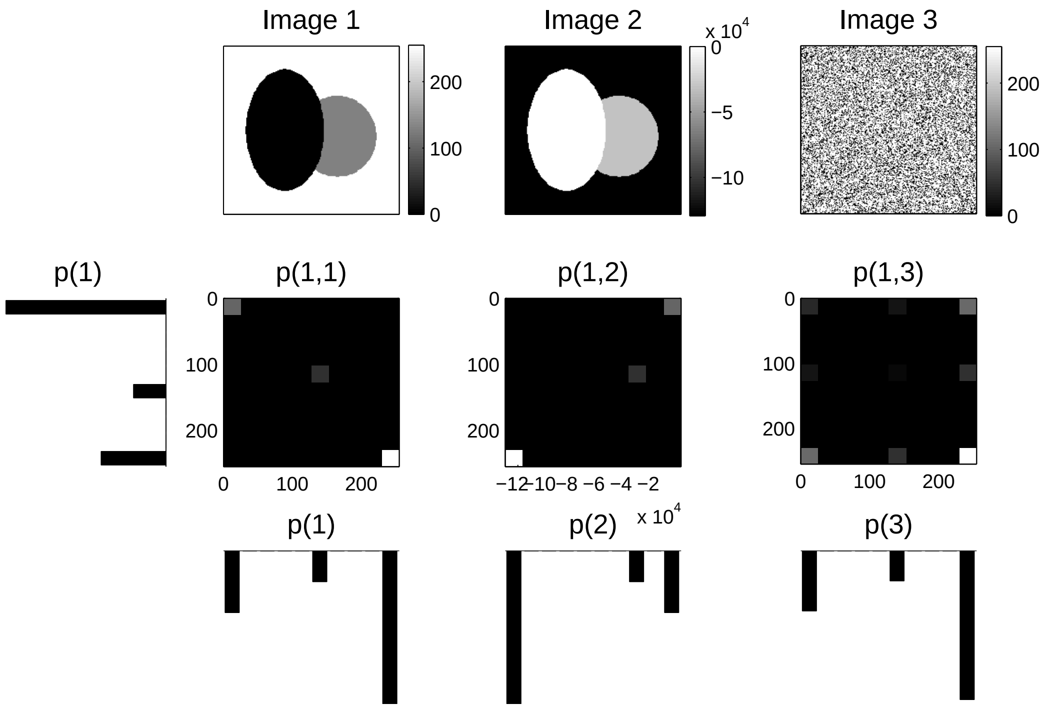


Figure 4.10: Entropy, multi-modality and spatial dependence. **Top row:** Three images of equal size. Image 2 was created by applying an arbitrary non-linear injective transformation, in this case $Im2 = -2(Im1)^2 + 3(Im1) - \log((Im1) + 1)$. Image 3 is created via a random permutation of the *pixel locations* of Image 1, consequently it shares an identical PDF with Image 1. **Middle row:** Marginal PDF of Image 1 and three JPDFs formed by image pairs $\{\text{Image 1}, \text{Image } Z\}$, $Z = [1, 2, 3]$. **Bottom row** Marginal PDFs of the three images, shown with their vertical axis inverted.

and MI - due to their dependence on the JPDF - consider spatially corresponding pixels between images. This is the source of spatial inter-image pixel-wise dependence in JE and MI. It should be emphasized however that the functionals do not imply any spatial dependence among the various pixel values *within* a single image.

The JE of the first two image pairs is $H(Im1, Im1) = H(Im1, Im2) = 0.9035$. JE is invariant to grey value transformations on images, given that the JPDFs are populated with the same number of clusters and with equivalent probability between cases for each cluster. The same holds for the MI case as $MI(Im1, Im1) = MI(Im1, Im2) = 0.9035$. In this case one takes into account the similarity between the marginal PDFs of the images. The PDF of Image 1 is constant. The PDF of Image Z corresponds to same levels of $H(Z)$, $\forall Z$ (same number of clusters, equal amplitude).

Finally, consider the image pair $\{\text{Image 1}, \text{Image 3}\}$. According to the earlier discussion regarding the marginal entropy approach of Fig.. 4.10, its entropy is equivalent to the previous images. Considering the JE of the formed pairs, it is evident that $P(\text{Image 1, Image 3})$ has additional clusters, hence more outcomes are possible and with increasing choice the uncertainty rises. This reflects to JE which is now increased to $H(Im1, Im3) = 1.8070$ as well as to MI which decreases to $MI(Im1, Im3) = 0.0001$, as the images in the third pair are structurally independent. Therefore, JE and MI can measure structural similarity between multi-modal images, contrary to the marginal terms.

Additional differences among the JE and MI functionals exist. We revisit both similarity measures in Sec. 5.6.4.1 in the context of image registration, where the information depicted in the assessed images can be altered due to spatial transformations and variable overlap regions. Finally, in Sec. 7.3 we compare the capacity of JE and MI, to act as means of introducing *a priori* structural information in the image inverse problem of diffuse optical tomography (DOT).

4.5.5.3 Discrete vs differential entropy in the imaging context

Consider 4.11. It demonstrates the absence of full invariance in the case of differential entropy under linear transformations (see Sec. 4.5.3). The first two images are the ones depicted in Fig. 4.10 whereas Image 3 is created by rescaling Image 2 in the grey value range of Image 1 via

$$Im3 = min(Im1) + \frac{Im2 - min(Im2)}{max(Im2) - min(Im2)}(max(Im1) - min(Im1)). \tag{4.56}$$

We seek to compare the empirical DE of the depicted images. To do so we employ a KDE which produces the continuous density estimates depicted in the third row.

The DE of the three images are $h(Im1) = 8.83443685$, $h(Im2) = 15.06297344$ and $h(Im3) = 8.83441969$ nats. $h(Im3)$ and $h(Im1)$ differ exactly by

$$\log\left(\left|c = \frac{max(Im1) - min(Im1)}{max(Im2) - min(Im2})\right|\right), \tag{4.57}$$

where c is the slope of the linear transformation, consistent with the theory in Sec. 4.5.3. DE is invariant to the additive component of the linear transformation. $h(3) \approx h(1)$ mainly because they are in a common range with equal binning intervals. Hence, prior to comparing the differential entropy of images, it is preferable to compensate for linear changes either by explicitly compensating for $\log(|c|)$ or equivalently re-scaling one of the images in the range of the second. The reason for which $h(3)$ does not completely match $h(1)$ is due to the difference in the corresponding PDFs, specifically the partial overlap of the two clusters in the third PDF, which is absent in the first. Note that such need does not exist in the case of discrete entropy. All probability masses are normalized hence their sum is invariant to the coordinate system (see Fig. 4.10). Finally, it should again be emphasized that when the coordinate system is kept fixed, entropy is a valid measure of uncertainty between random variables defined on that coordinate system.

4.6 Summary

This chapter has introduced the main concepts of information theory, specifically marginal, joint and conditional entropy as well as MI. In addition, the empirical entropy of a sample was also introduced. This discussion was preceded by a brief introduction to the fundamental probability theory concepts that information theory is based upon. The differences between the entropy of discrete and continuous RVs were specifically outlined. The subject of PDFs estimation was briefly covered by outlining a sample of parametric and non-parametric methods, focusing on the latter. The discussion was held from an

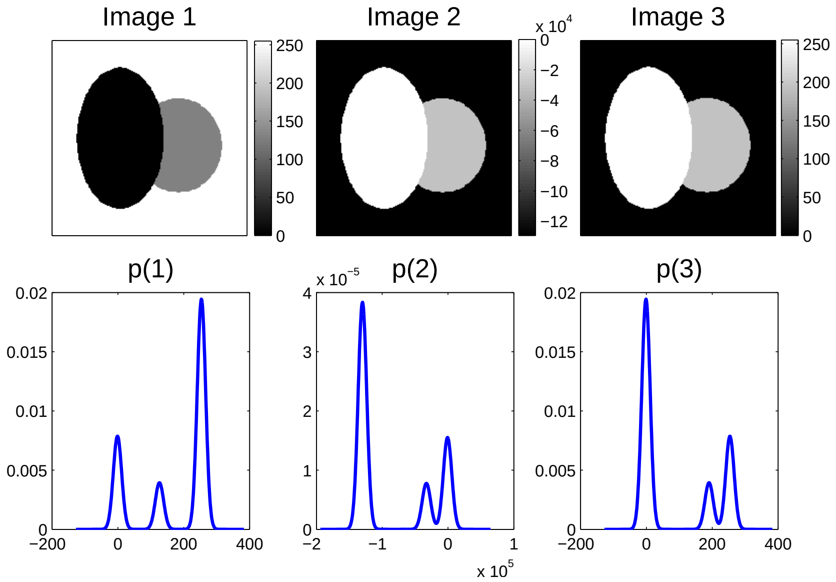


Figure 4.11: Differential entropy and images. Top row shows three images all of the same size. Image 1 has three distinct grey values. Image 2 was created by applying an arbitrary non-linear injective transformation, in this case $Im2 = -2(Im1)^2 + 3(Im1) - \log((Im1) + 1)$. Image 3 is created by re-scaling Image 2 to the same grey value range of Image 1.

imaging perspective with the appropriate schematics to assist intuition. Finally, the inherent capacity of the information theoretic functionals to serve as multi-modal similarity measures was discussed.

Chapter 5

Medical image registration

5.1 Introduction

Image registration is the collective term for all methods aiming to establish the accurate alignment of images, so that their corresponding features become spatially superimposed, given a common coordinate system. The alignment of images has proven to be a highly sought after capacity and it has found numerous applications in various scientific disciplines including art, astronomy, astro-physics, biology, chemistry, criminology, genetics, physics, remote sensing, security, machine vision and medicine [Fischer and Modersitzki, 2008].

The structure of this chapter is as follows: Section 5.2 briefly introduces a sample of applications of image registration to medical imaging. In 5.3 the three fundamental parts of an image registration algorithm are introduced - namely *similarity measures*, *spatial transformations* and *optimization*; alongside with the corresponding notation. Section 5.4 briefly outlines the classification criteria of the various registration problems. The last sections (5.5-5.7) revisit the parts of spatial transformations, similarity measures and optimization in more detail by presenting a sample of the various algorithmic and conceptual choices existing in the literature, on these topics.

5.2 Applications in medical imaging

In the context of medical imaging, image registration is a highly active subject of research and this reflects on the size of the corresponding specialized literature. The increased attention is not surprising as the alignment of medical images is of widespread interest across the full spectrum of the available imaging modalities. Image registration applications are numerous.

One of its applications regards difference imaging, to which image registration has been established as an indispensable part. Difference imaging simply describes the process of subtracting two intramodal images of the same subject, resulting in a third image which depicts the differences between the considered images. The differences which have high significance from a medical perspective include changes in physiology for example disease progression considering individuals or a population group where statistical outcomes are retrieved (cohort studies) [Fox et al., 2001, 1996; Freeborough and Fox, 1998; Mahanand and Kumar, 2009]; response of disease to therapy in follow up studies [Li et al., 2009]; and changes in contrast in pre- and post- contrast agent application in diagnostic studies [Rueckert et al., 1999] (see Fig. 5.1). One expects the difference image to solely depict changes in physiology. However the identification of changes in physiology between the images, is usually compromised by unavoidable extra features appearing in the difference images due the spatial misalignment of the involved scans. Medical image registration can compensate for the latter and aid towards the easier extraction of the information of interest.

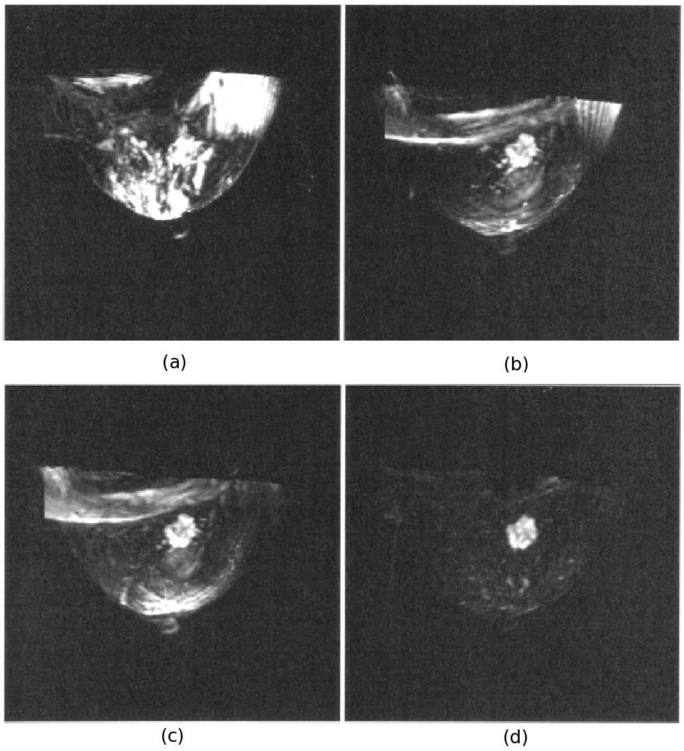


Figure 5.1: Medical image registration and difference imaging. Images depict difference images formed by pre- and post- contrast 3D MRI scans of the breast, visualized via *maximum intensity projection*. **(a)** Difference between pre- and post- contrast images where no registration has been applied. Motion of the breast during the scans results in spatial misalignment of the corresponding features. **(b)** improvement after rigid registration (translation/rotation) **(c)** improvement after affine registration (rigid transformations + scaling, shearing etc) **(d)** improvement after non-rigid registration (local deformations). The vast majority of corresponding features present in both pre- and post- contrast images become spatially aligned and hence they cancel each other in the difference image. The only prominent feature corresponds to a tumour, as it corresponds to a difference due to contrast variability and not spatial misalignment, hence it cannot be compensated by image registration (source: [Rueckert et al., 1999])

Automated image registration facilitates the execution of these much needed alignments. The non-rigid spatial alignment of huge data-sets in cohort studies, such as serial 3D volume scans - effectively 4D datasets - acquired from multiple subjects, is a highly intensive process even for modern computers and from a numerical perspective is characterized by thousands of degrees of freedoms. The execution of such tasks would be infeasible without automated image registration schemes. Even simpler tasks - such as affine registrations of relatively small data-sets - would require significantly increased amounts of time, effort and cost. In some tasks, arguably the automated registration schemes frequently perform better than the human threshold of assessing mis-registration, which has been identified to be limited to spatial displacements equal or greater than 0.2mm [Fitzpatrick et al., 1998; Früwald et al., 2009; Holden et al., 2000].

Image registration is also extensively used in surgical planning and interventions. Intra-operative low resolution images can be registered to high resolution pre-operative ones, in order to update the latter to the current state of the operated anatomy after deformation due to surgical intervention. In addition, surgical equipment can be tracked in real time using ultrasound or visual markers and subsequently registered against the medical images. Images and surgical equipment are visualized in a single view, allowing the surgeons to perform the surgery knowing exactly the position of the equipment inside the anatomy and its proximity to vital organs or the target abnormality. The registration of patient, anatomy and surgical equipment to a common coordinate system is also necessary for robotic surgery and remote surgery. [Camarillo et al., 2004; Gering et al., 2001; Pott and Schwarz, 2002; Satava, 1999; Stoll and Dupont, 2005]. Other tasks which would be difficult to perform without image registration include real time compensation for target anatomy deformation due to physiological function, for example the compensation for lung motion due to respiratory function, during radiotherapy [Coselmon et al., 2004; Murphy, 2004] or focused ultrasound treatment [Ries et al., 2010].

Articles reviewing medical image registration specific literature include [Brown, 1992; Hill et al., 2001; Maintz and Viergever, 1998; Maurer, Jr., C.R. and Fitzpatrick, 1993; Van den Elsen et al., 1993]. A sample of books focusing on the topic is [Fitzpatrick et al., 2000; Hajnal, 2001]. For non-medical applications please see [Zitova, 2003].

5.3 Definitions and notation

As in previous chapters, a digital image **x** is considered to be a mapping from discrete spatial positions r in some domain $\Omega_{\mathbf{x}}$, to grey values $x \in \mathbb{R}$. This mapping is explicitly expressed as

$$\mathbf{x} : r \in \Omega_{\mathbf{x}} \mapsto x = \mathbf{x}(r) \in \mathbb{R} \tag{5.1}$$

An image registration process spatially transforms one image - termed as *moving* or *source* - so it becomes aligned with a *target* image which remains unaffected by the transformation. Consider in the 2D case, the static image **x** and the moving image **y**. Let **x** be defined over the simplest discrete geometrical arrangement, that is a rectangular pixel grid g with spacing Δx, Δy in the direction of each of the the Cartesian axes, with N pixels. The nodal locations on the grid correspond to the pixel locations $r = \{r_1, r_2, \ldots, r_k, \ldots, r_N\}$. Each r_k is a vector and can be expressed and/or indexed in terms of its axial components $r_j = r_{ij} = \{\mathrm{x}_i, \mathrm{y}_j\}$, with $i = 1, 2, \ldots, N_{\mathrm{x}},\ j = 1, 2, \ldots, N_{\mathrm{y}}$ and $N_{\mathrm{x}} \cdot N_{\mathrm{y}} = N$. The moving $\mathbf{y}(r'),\ r' \in \Omega_{\mathbf{y}}$ is defined in a similar manner. However it can differ from **x** with respect to the number, locations and spacing between the pixels.

Image registration algorithms can differ with respect to various algorithmic details, however they all share three distinct algorithmic parts. These are:

1. A *spatial transformation scheme*

$$r'' = \mathcal{T}(r';\theta), \tag{5.2}$$

which is applied on the spatial coordinates $r' \in \Omega_{\mathbf{y}}$ of the moving image **y**, resulting in the transformed $r'' \in \Omega_{\mathbf{y}^{\mathcal{T}}}$ and which is controlled by a finite number of parameters θ. The grey values $y = \mathbf{y}(r')$ are considered to be spatially coupled with the locations r', upon which y are initially defined prior to the transformation. As each r' is subjected to the spatial transformation resulting in a new r'', the latter is assigned with the gray value $\mathbf{y}(r')$ of the original location r'. The transformed moving image is denoted as $\mathbf{y}^{\mathcal{T}}(r'')$, or equivalently $\mathbf{y}^{\mathcal{T}}\big(\mathcal{T}(r';\theta)\big)$.

Spatial transformations operate in conjunction with an *interpolation scheme*. The application of the spatial transformation of Eq. 5.2 would most often result in continuous locations r''. However, as was already noted earlier, image similarity measures can only operate between images defined at common spatial locations. Given that r or equivalently **x** is unaffected from the transformation, then $\mathbf{y}^{\mathcal{T}}(r'')$ has to be re-defined over $r \in \Omega_{\mathbf{x}}$. The latter is accomplished by the interpolation scheme. The transformed image, after the application of the interpolation scheme is denoted as $\mathbf{y}^{\mathfrak{T}}(r)$. The different superscript between $\mathbf{y}^{\mathfrak{T}}(r)$ and $\mathbf{y}^{\mathcal{T}}(r'')$ emphasizes that the gray values of the former are a product of interpolation.

2. A *similarity metric* $\Psi\big(\mathbf{x}(r), \mathbf{y}^{\mathfrak{T}}(r)\big)$ which quantitatively evaluates the level of similarity between two images and ideally attains its maximum value when accurate alignment is established.

For any given alignment configuration based on some estimate of θ, the similarity is evaluated between the overlapping regions of **x** and $\mathbf{y}^{\mathfrak{T}}$, commonly denoted for both images as $\Omega_{\mathbf{x};\mathbf{y}^{\mathfrak{T}}}$.

3. A *solver* which seeks a parameter estimate $\hat{\theta}$ for which $\mathbf{y}^{\mathfrak{T}}\big(\mathcal{T}(r';\hat{\theta})\big) = \mathbf{y}^{\mathfrak{T}}(r)$ becomes aligned with $\mathbf{x}(r)$ and similarity between the two images is maximized. Most often, the solver is comprised by an iterative optimization scheme which maximizes the similarity measure - or equivalently minimizes a dissimilarity measure (negative similarity) with respect to θ. This is expressed as

$$\hat{\theta} = \underset{\theta}{\arg\min}\left[\mathcal{E}(\theta) = -\Psi\Big(\mathbf{x}(r), \mathbf{y}^{\mathfrak{T}}\big(\mathcal{T}(r';\theta)\big)\Big)\right], \quad r \in \Omega_{\mathbf{x};\mathbf{y}^{\mathfrak{T}}}. \tag{5.3}$$

5.4 Classifications of registration problems

Image registrations methods can be classified according to a number of criteria characterizing the underlying task. Unfortunately, there is no single, universally accepted image registration methodology which can perform equally well for all given problems. Probably the most notable attempt to classify registration problems was by Van den Elsen et al. [1993], which was also adopted by Maintz and Viergever [1998] in his extensive literature review. The identified classification criteria are worth noting as they bring to some perspective the full extent of medical registration applications. Extensive literature for each of the categories can be found in the review paper by Maintz and Viergever [1998]. The classification categorizes registration problems according to:

- *The dimensionality of each of the involved images*, with cases including 2D/2D [Jacquet et al., 2009], 3D/3D [Rueckert et al., 1999], 2D/3D [Huang et al., 2009; Turgeon et al., 2005], studies involving time series and more.

- *The nature of the spatial transformation* which needs to be employed to bring images to alignment. *Rigid* transformations perform translational and rotational realignments [Zuo et al., 1996], *affine* which perform rigid transformations as well as scaling and shearing [Jenkinson and Smith, 2001] and *non-rigid* transformations which compensate for local deformations. Reviews of non-rigid registration problems include [Crum et al., 2004; Lester and Arridge, 1999].

- *The domain of the transformation* is classified as *i) global* where the transformation applies to the full image or *ii) local* when it only affects a smaller region.

- *The nature of registration basis*. The task of aligning images can in some cases be converted to a task of aligning secondary features. Achieving alignment between corresponding inter-image features is assumed to establish registration between the entire images. Similarity measures usually operate on the feature space. This class refers to the categorization of registration problems depending on the type of features employed. These can be further categorized as *i) extrinsic*, involving features explicitly added to the probed anatomy either *a)* invasively - for example stereotactic frames or implantable markers [Maurer, Jr., C.R. et al., 1997], or *b)* non-invasively - for example fiducial markers attached to the skin [Breeuwer et al., 1998; Walvoord et al., 2008] and more. *ii) Intrinsic* methods involve image-derived features such as *a)* sets of independent point landmarks [Chui and Rangarajan, 2003], which they not necessarily have to represent explicit anatomical sites, but can be any matching locations which can be confidently and unequivocally identified as spatially corresponding in the involved images [Likar and Pernuš, 1999]. Such locations usually have *geometrical significance*, for example curvature extrema such as local peaks, pits or saddle points [Audette et al., 2000]. *b)* Segmentation derived, higher order structures such as lines or surfaces, expressed in full (dense-points or parametrically) or in reduced form, for example crest-lines identified in surfaces [Thirion, 1996]. Audette *et al.* provides a specialized review on surface based registration [Audette et al., 2000]. *c)* Finally, other than independent points or surfaces, one can employ the actual pixel values directly as features giving rise to intensity based methods [Lemieux and Barker, 1998; Maes et al., 1997; Woods et al., 1993] or can further reduce the utilized data by using derived statistical quantities from the gray values, as for example is done by the method of moments [Alpert et al., 1990; Faber and Stokely, 1988].

- *the relationship between the modalities of the involved images*, giving rise to *i)* mono-modal *ii)* multi-modal *iii)* modality to model [D'Agostino et al., 2007] -for example the alignment of a medical image with a model of statistical nature such as a probabilistic atlas of anatomy [Mazziotta et al., 1995] or function [Lancaster et al., 2000]. *iv)* The last category involves registration of high-resolution pre-operative images of the probed anatomy to lower-resolution intra-operative images of the probed anatomy [Gering et al., 2001; Raabe et al., 2002]. This enables the information in the pre-operative images to be updated and match the intra-operative patient's anatomy, potentially altered due to the surgical intervention. Given also the knowledge of the exact physical position of the patient in space via stereotaxy one can achieve registration of real-time tracked surgical instruments [Gering et al., 2001]. The combination of the above methodologies gives rise to what is known as image-guided surgery

- *The subject of the images* refers to the actual patient subjected to scanning, giving rise to *i)* intra-subject *ii)* inter-subject and *iii)* atlas, where a modality-atlas registration takes place. In the last two cases, the notion of exact anatomical correspondence does not exist, however it is expected that a level of correspondence can be recovered due to homology [Crum et al., 2003]. This can also be the case in the intra-subject case, when the images are parts of time-series and the anatomy changes due to pathology or medical intervention.

- *The object depicted in the images*, refering to the actual anatomical part depicted in the images. Registration between different anatomical parts involves different degree of complexity. Registration between head images [Studholme et al., 1996] can potentially be established using rigid/affine transformations in the intra-subject case, or non-rigid for inter-subject. However registration between for example cardiac images [Makela et al., 2002] or thoracic images [Goerres et al., 2002] can prove more complex due to the dynamic movement of the heart or the lungs respectively.

- The level of interaction/supervision by an expert, during the registration process.

5.5 Spatial Transformations

This section briefly introduces a subset of the spatial transformations $r'' = \mathcal{T}(r'; \theta)$ employed in registration applications. The discussion assumes the case of 2D images.

5.5.1 Linear transformations

Linear transformations are a fundamental part of most registration algorithms. They are usually employed to recover the potentially large initial misalignment between **x** and **y**, before employing more accurate methods, for example see [Rueckert et al., 1999]. One of the main characteristic of these transformations is that they are applied uniformly to all pixels of the transformed region. In the majority of the cases, the transformation is global as it involves the entire **y**. However, there have been studies which divide the involved images to regions or *blocks*, which are subsequently matched via local linear transformations. The final transformation is constructed by combining the effect of the local transformations, usually by employing an interpolation scheme. For example see the locally affine transformation discussed in [Commowick et al., 2006; Pitiot et al., 2006].

The general form of a 2D linear transformation consists of six control parameters $\theta = \{\theta_1, \theta_2, \ldots, \theta_6\}$, which uniformly transform the pixel locations $r' = [\mathrm{x}'; \mathrm{y}']^{\mathrm{T}}$ to the resulting coordinates $r'' = [\mathrm{x}''; \mathrm{y}'']^{\mathrm{T}}$ according to

$$\mathsf{x}'' = \theta_1 \mathsf{x}' + \theta_2 \mathsf{y}' + \theta_3, \tag{5.4}$$

$$\mathsf{y}'' = \theta_4 \mathsf{x}' + \theta_5 \mathsf{y}' + \theta_6. \tag{5.5}$$

The above can be also expressed in a matrix form

$$\begin{bmatrix} \mathsf{x}'' \\ \mathsf{y}'' \\ 1 \end{bmatrix} = \begin{bmatrix} \theta_1 & \theta_2 & \theta_3 \\ \theta_4 & \theta_5 & \theta_6 \\ 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} \mathsf{x}' \\ \mathsf{y}' \\ 1 \end{bmatrix} \tag{5.6}$$

or even more compactly as

$$r'' = \mathrm{M} \cdot r' \tag{5.7}$$

Variants of linear transformations are defined for different configurations of matrix M and are discussed in the following sections.

5.5.1.1 Rigid transformations

Rigid transformations are transformations which preserve distances, non-zero angles and straightness of lines [Fitzpatrick et al., 1998]. They consist of *translations* and *rotations*.

Translations Translations regard the displacement of all $r' = [\mathsf{x}', \mathsf{y}']^{\mathrm{T}}$ in straight lines and at the same direction, by fixed distance $t = [\mathsf{T}_\mathsf{x}, \mathsf{T}_\mathsf{y}]^{\mathrm{T}}$, for each spatial component. The transformation is expressed as $r'' = r' + t$ or in its equivalent matrix form $r'' = \mathsf{T} \cdot r'$, with the matrix being defined as

$$\begin{bmatrix} \mathsf{x}'' \\ \mathsf{y}'' \\ 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & \mathsf{T}_\mathsf{x} \\ 0 & 1 & \mathsf{T}_\mathsf{y} \\ 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} \mathsf{x}' \\ \mathsf{y}' \\ 1 \end{bmatrix} \tag{5.8}$$

Rotations Rotational transformations in the 2D case depend on a single parameter, that is the angle Θ of rotation, expressed in *radians*. In the 2D case the rotation is performed around the axis orthogonal to the Cartesian plane at the origin $[0, 0]^{\mathrm{T}}$. The matrix M in Eq. 5.6 is replaced by the rotation matrix R, giving rise to $r'' = \mathsf{R} \cdot r'$, where R is explicitly defined as

$$\begin{bmatrix} \mathsf{x}'' \\ \mathsf{y}'' \\ 1 \end{bmatrix} = \begin{bmatrix} \cos\Theta & -\sin\Theta & 0 \\ \sin\Theta & \cos\Theta & 0 \\ 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} \mathsf{x}' \\ \mathsf{y}' \\ 1 \end{bmatrix} \tag{5.9}$$

5.5.1.2 Non-rigid transformations

Scaling A scaling transformation effectively multiples each of the Cartesian coordinates determining a pixel location by some scalars S_x and S_y. The transformation is performed by a scaling matrix S, giving rise to $r'' = \mathsf{S} \cdot r'$ or equivalently

$$\begin{bmatrix} \mathsf{x}'' \\ \mathsf{y}'' \\ 1 \end{bmatrix} = \begin{bmatrix} \mathsf{S_x} & 0 & 0 \\ 0 & \mathsf{S_y} & 0 \\ 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} \mathsf{x}' \\ \mathsf{y}' \\ 1 \end{bmatrix} \tag{5.10}$$

where the structure of S is revealed. In the case of $\mathsf{S_x} = \mathsf{S_y}$, the scaling is called *isotropic* and preserves the straightness of lines and angles but not distances.

Shearing Shearing transformations which preserve straight lines and parallelism but they don preserve angles among lines. Shearing transformations are defined by a matrix Sh giving rise to $r'' = \mathsf{Sh} \cdot r'$ or equivalently

$$\begin{bmatrix} \mathsf{x}'' \\ \mathsf{y}'' \\ 1 \end{bmatrix} = \begin{bmatrix} 1 & \mathsf{Sh}_x & 0 \\ \mathsf{Sh}_y & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} \mathsf{x}' \\ \mathsf{y}' \\ 1 \end{bmatrix} . \tag{5.11}$$

Affine transformations Affine transformations are the most general of the linear transformations described by Eq. 5.7. The parameters θ in Eqs. 5.5-5.6 are not restricted. Affine transformations preserve the straightness of lines and parallelism but not angles [Fitzpatrick et al., 1998].

5.5.1.3 Composition of multiple transformations

Conveniently, the nature of the linear transformations enables several transformations to be performed by the application of a single matrix. The matrix is constructed by multiplying the matrices of the individual transformations in the reverse order which the latter are applied to the image. For example, if one seeks to firstly scale, then rotate and finally translate $\mathbf{y}$, then the transformation matrix which achieves the sought outcome is defined as $\mathrm{M} = \mathsf{T} \cdot \mathsf{R} \cdot \mathsf{S}$ and is applied to the moving coordinates according to Eq. 5.7.

5.5.1.4 Origin of transformation

The choice of the origin of transformation can affect the resulting $\mathbf{y}^{\mathfrak{T}}$. The default origin for rotation, scaling and shearing is the Cartesian origin $[0, 0]^{\mathrm{T}}$. This detail should be considered especially in cases where the main feature depicted in the image is not centered at the origin. In that case, small transformations can result in large displacements of the main feature and drastically reduce the similarity between $\mathbf{x}$ and $\mathbf{y}^{\mathfrak{T}}$. The effect on applying transformations about the Cartesian origin is depicted in the top row of Fig. 5.2. The image $\mathbf{y}$ to be subjected to transformation is depicted at the top-left. The dark background depicts the extended spatial domain in order to emphasize the transformation effects. The second image show $\mathbf{y}^{\mathfrak{T}}$ resulting from a rotation about the origin, situated at the lower-left corner of $\mathbf{y}$. The gray region corresponds to the initial location of $\mathbf{y}$ and is included to assist visual comparison. For a small rotation of $30°$ the overlap reduces significantly. Apparently, a rotation of $\Theta = 90°$ would lead to a $\Omega_{\mathbf{x};\mathbf{y}^{\mathfrak{T}}} = \{\emptyset\}$, where $\{\emptyset\}$ is the empty set. Similarly, considering the Cartesian origin as the origin of transformations can result in large displacements for relatively small scaling and shearing transformations.

One way to alleviate this problem, is to choose a different origin of the transformations. One can use the actual geometric centre of the image - that is the centre of the pixel grid however a better choice is the centre of mass (c.o.m.) known as the image's zeroth order geometrical moment. The usage of c.o.m. is a better choice when the main feature depicted in an image is not centered at the geometric centre of the image. In order to choose the new origin, the c.o.m. of the image is translated at the Cartesian origin prior to the transformations, then the transformations are applied and finally the resulting image is translated back to the initial c.o.m. location. The c.o.m. of an image is denoted as $\bar{r}' = [\bar{\mathsf{x}}', \bar{\mathsf{y}}]$, with

$$\bar{\mathsf{x}}' = \frac{\sum_i \sum_j i \cdot \mathbf{y}(r'_{ij})}{\sum_i \sum_j \mathbf{y}(r'_{ij})}, \quad \bar{\mathsf{y}}' = \frac{\sum_i \sum_j j \cdot \mathbf{y}(r'_{ij})}{\sum_i \sum_j \mathbf{y}(r'_{ij})}. \tag{5.12}$$

Considering the composition of transformations (see Sec. 5.5.1.3), then the execution of a rotation, scaling and shearing with respect to c.o.m. at the given order, would be expressed as $r'' = \mathrm{M} \cdot r'$ with $\mathsf{M} = \mathsf{T}_{\bar{r}'}^{-1} \cdot \mathsf{Sh} \cdot \mathsf{S} \cdot \mathsf{R} \cdot \mathsf{T}_{\bar{r}'}$ and $\mathsf{T}_{\bar{r}'} = [\bar{\mathsf{x}}', \bar{\mathsf{y}}']^{\mathrm{T}}$, $\mathsf{T}_{\bar{r}'}^{-1} = [-\bar{\mathsf{x}}', -\bar{\mathsf{y}}']^{\mathrm{T}}$.

The effect on $\mathbf{y}^{\mathfrak{T}}$ of applying transformations with respect to c.o.m. is depicted in the bottom row of Fig.. 5.2. The resulting overall displacement is apparently smaller than the previous case.

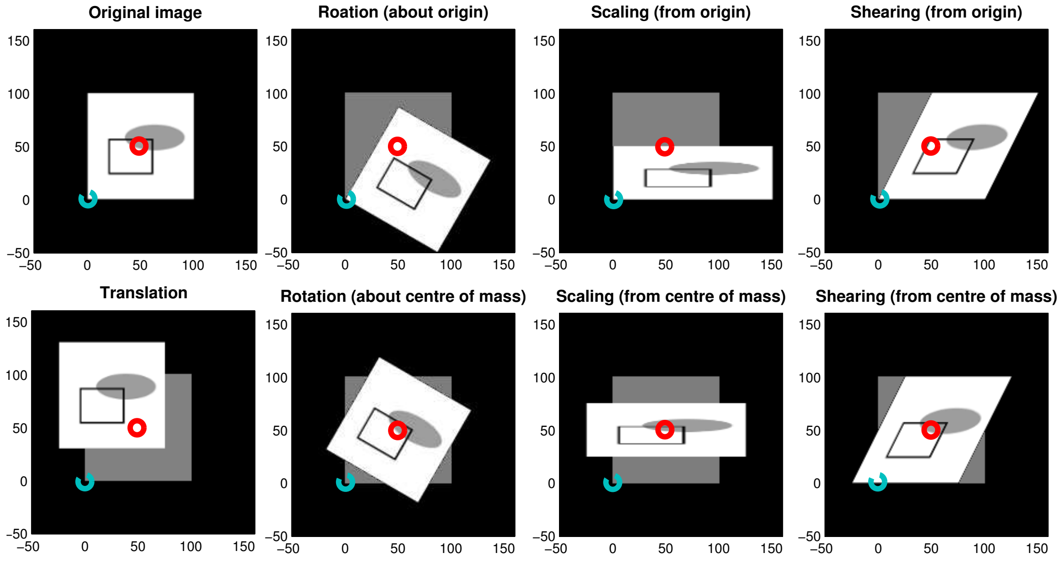


Figure 5.2: Effects of the choice of origin of linear transformations. Highlighted features: c.o.m. (red), Cartesian axes origin (blue). **Column 1** Top: Moving image **x**. Bottom: Translated by $t = [-25, 30]^{\mathrm{T}}$. Gray region depicts location of initial **x** **Column 2** Top: Rotation about origin $\Theta = -30°$, Bottom: Rotation about c.o.m. - same Θ Anisotropic scaling from origin with $\mathsf{S}_\mathsf{x} = 1.5$ and $\mathsf{S}_\mathsf{y} = 0.5$. Bottom: Scaling from c.o.m. - same scaling factors. **Column 4** Shearing from origin, with $\mathsf{Sh}_x = 0.5$ and $\mathsf{Sh}_y = 0$. Bottom: Shearing from c.o.m. - same shearing factors.

The c.o.m. as well as higher order image moments can be used as features for driving image alignment. Although not very robust, they can provide an initial registration estimate for more accurate methods. Firstly, **y** is transformed so its c.o.m. coincides with the c.o.m. of **x**. This compensates for translational misalignment. Rotational misalignment between **x** and $\mathbf{y}^{\mathfrak{T}}$ can be compensated by aligning the principal axes of the involved images which are returned by the first order moments. Higher order moments can compensate for additionally types of misalignment such as scaling. See [Alpert et al., 1990; Lester and Arridge, 1999; Maintz and Viergever, 1998] and the references within, regarding principal moments and registration.

5.5.2 Non-linear transformations

Non-linear transformations differ from the linear global transformations as they affect the pixels in **y** in a non-uniform manner and the resulting transformations are non-rigid. Various schemes have been proposed to facilitate such transformations. Elastic transformations assume the the moving image can be modeled as a deformable elastic body [Bajcsy and Kovačič, 1989]. Under this assumption, the image can then be deformed by the application of some external force. This external force is opposed by the internal forces of the elastic body, resisting the deformation. In the image registration paradigm, the driving force is proportional to the gradient of the similarity measure with respect to the deformation of **y**. Consequently, the external forces tries to alter **y** in a manner which increases its similarity with the target **x**. The registration converges when external and internal forces cancel each other, establishing a state of equilibrium. Elastic transformations can prove restrictive in retrieving local transformations, due to the internal forces which act antagonistically to the external force [Crum et al., 2004; Lester and Arridge, 1999].

An alternative approach which does not carry this restriction, involves the modeling of **y** as a viscous fluid [Christensen et al., 1996]. The motion of the fluid's particles, in this context $r \in \Omega_{\mathbf{y}}$, is described by a velocity field evolving over time and which satisfies the Navier-Stokes partial differential equation of conservation of momentum [D'Agostino et al., 2003]. Once again the force which drives the registration is the derivative of the similarity measure.

5.5.2.1 Demon's registration

Thirion [1998] approached the registration of intra-modal images as a diffusion process. He introduced the concept of *demon's* in imaging, originally conceived by Maxwell to illustrate a paradox of Thermodynamics (see Thirion's article). In the context of imaging, demons are defined on image feature's interfaces, which are perceived as semi-permeable membranes. Demons act as actuators, based on the concept of polarity which dictates if region in **y** should move *inside* or *outside* a spatially neighbouring region in **x**, via the latter's interface. Hence, demons drive regions of **y** to *diffuse* through contours of **x**. To determine polarity and compute the demons forces, Thirion used a similarity measure known as *optical flow*, which in temporal studies assumes that the intensity of a pixel, displaced between successive time frames, remains constant. In non-temporal studies, one simply assumes that inter-image features indicating corresponding physiological locations must be characterized by similar intensities, hence the method applies only to the intra-modal case. Additional information on demons' algorithm can be found in [Pennec et al., 1999].

5.5.2.2 Thin plate splines

Splines are smooth piecewise polynomial functions which enable the spatial transformation to be expressed as a basis expansion. The expansion requires a set of points defined in the moving image which act as the basis coefficients. As these control points are displaced, a basis function of choice models a smooth displacement of all locations defined among these control points. One of the two spline methods introduced in this paper is the thin plate spline (TPS) [Bookstein, 1989]. Employing TPS implies that the image is modeled as a thin metal plate. Let $z(\mathsf{x},\mathsf{y})$ define the surface of the thin plate as an elevation from the Cartesian plane. Then, the spline will take a form which minimizes its bending energy [Eriksson and Astrom, 2006]

$$J(z) = \int\int_{\mathbb{R}^2} \left(\left(\frac{\partial^2 z}{\partial \mathsf{x}^2}\right)^2 + 2\left(\frac{\partial^2 z}{\partial \mathsf{x}\partial \mathsf{y}}\right)^2 + \left(\frac{\partial^2 z}{\partial \mathsf{y}^2}\right)^2 \right) \mathrm{d}x\mathrm{d}y. \tag{5.13}$$

The function which minimizes the above equation is $\mathcal{E}(\mathsf{x},\mathsf{y}) = \sum_i c_i u\left(\left\|[\mathsf{x},\mathsf{y}]^{\mathrm{T}} - \varphi_i\right\|\right)$, where c_i are mapping coefficients, φ are the control points and $u(\mathsf{x},\mathsf{y}) = -r^2 \cdot \log(r^2)$ is the TPS radial basis function with $r = \sqrt{\mathsf{x}^2 + \mathsf{y}^2}$ [Bookstein, 1989]. The basis function in TPS has global support which can prove costly to compute for all control point locations. Registration with TPS uses landmarks in both **x** and **y** at corresponding physiological locations. One then attempts to match the landmarks, whereas the rest of the space is bent as little as possible. The next spline based transformation is defined in more detail as it is employed later in this work.

5.5.2.3 B-Spline Free form deformations

B-Spline Free form deformations (FFDs) were introduced by Sederberg and Parry [1986] and have been employed for various image processing tasks, for example image metamorphosis [Lee et al., 1996]. Rueckert et al. [1999] proposed their employment as a spatial transformation scheme for the purpose of non-rigid medical image registration. Modeling non-rigid transformations of **y** with FFDs is accomplished by manipulating a lattice of control points, arranged over the domain $\Omega_{\mathbf{y}}$. As these control points are displaced, their surrounding pixels are also subjected to a displacement, weighted according to their distance from the perturbed control point via B-Spline functions. Figure 5.3 depicts a control point grid overlaid on an image and the resulting transformed image due to control point perturbation. Let φ_{ij} denote the control points indexed by i, j uniformly spaced with spacing $\Delta\varphi$ in both Cartesian directions. Then, the transformation of r' can be expressed as a linear combination of the B-Spline basis functions and the control points φ_{ij}, expressed as [Lee et al., 1996]

$$\begin{bmatrix} \mathsf{x}'' \\ \mathsf{y}'' \end{bmatrix} = \mathcal{T}\left(\begin{bmatrix} \mathsf{x}' \\ \mathsf{y}' \end{bmatrix} ; \varphi \right) \tag{5.14}$$

$$= \sum_{k=0}^{3}\sum_{l=0}^{3} B_k(s)B_l(t)\varphi_{(i+k)(j+1)}, \tag{5.15}$$

where $i = \lfloor \mathsf{x}' \rfloor - 1$, $j = \lfloor \mathsf{y}' \rfloor - 1$, $s = \mathsf{x}' - \lfloor \mathsf{x}' \rfloor$ and $t = \mathsf{y}' - \lfloor \mathsf{y}' \rfloor$. The terms B_k and B_l correspond to the uniform cubic B-spline functions defined as

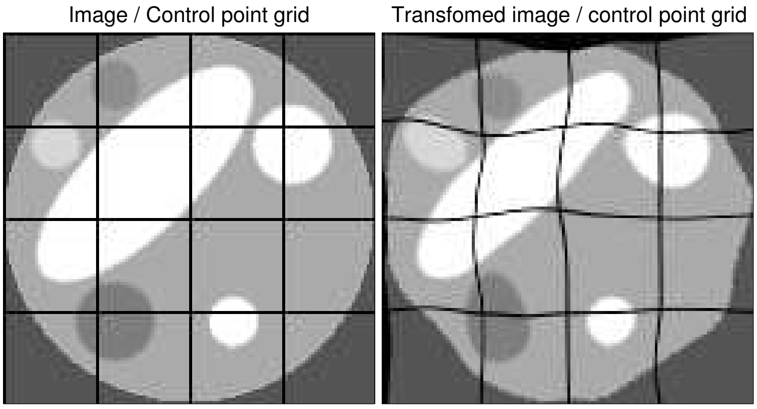


Image / Control point grid
Transfomed image / control point grid

Figure 5.3: Non-rigid transformation based on B-Spline Free Form Deformations. **Left:** Image with superimposed control point grid **Right:** Transformed image due to perturbations on the control point locations

$$B_0(t) = (1-t)^3/6, \tag{5.16}$$

$$B_1(t) = (3t^3 - 6t^2 + 4)/6, \tag{5.17}$$

$$B_2(t) = (-3t^3 + 3t^2 + 3t + 1)/6, \tag{5.18}$$

$$B_3(t) = t^3/6, \tag{5.19}$$

with $0 \leq t < 1$.

Considering Eq. 5.15, FFD transformations are locally controlled in the sense that the B-spline basis functions have a finite support of $4\ \Delta\varphi$ [Ino et al., 2005]. This means that if a φ_{ij} is perturbed, it will only affect pixels under the B-Spline support. This is demonstrated in Fig. 5.4.

The local nature support has a number of implications. Firstly it leads to a reduced computational cost in estimating the pixel transformations resulting from the control point perturbations, compared with spline methods using basis of global support such as the TPS [Rueckert et al., 1999]. One only needs to compute the transformations for the pixels in the local neighbourhood. The other implication is crucial to the registration paradigm. Assume that the same feature is depicted in **x** and **y** in different spatial locations, with an in-between distance $> 4\Delta\varphi$. Although the alignment of the features would maximize the similarity measure, the transformation applied to **y** cannot facilitate a displacement of pixels to a distance greater than $4\Delta\varphi$. To deal with this inherent constraint and increase the chances of global convergence, FFD based registration methods employ a series of control point lattices of increasing resolution. Initial coarse lattices enable alignment of corresponding features which are further away, as the corresponding $4\Delta\varphi$ is now large and is potentially greater than the spatial displacement which needs to be recovered. When registration using the coarse lattice has been established, the algorithm refines it to a higher resolution allowing highly localized misalignments to be matched more accurately. Lattice refinement has been described by Forsey and Bartels [1988]. Figure 5.5 depicts an example coarse lattice and its refined version. The final transformation applied to each r', is constructed by summing the $k = 1, \ldots, L$ transformations $\mathcal{T}^k\left(r'; \varphi^k\right)$ applied to the same point, from all L control-point lattices φ^k

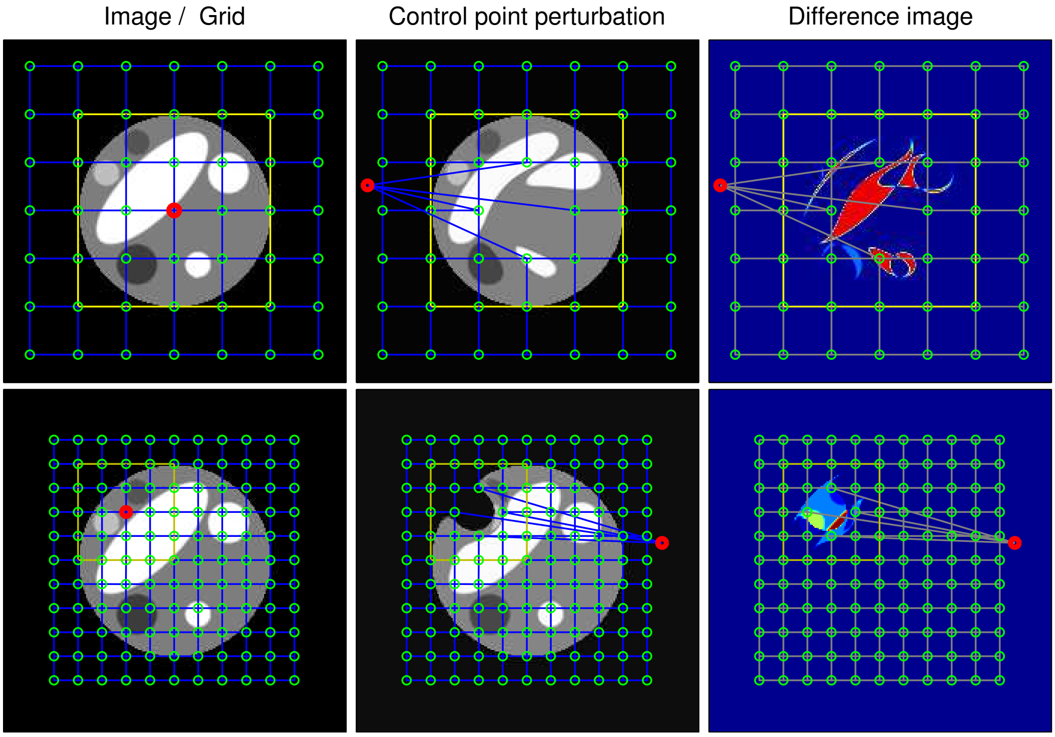


Figure 5.4: Local support of B-Spline free form deformations **Top:** Coarse grid **Bottom:** Dense grid. **Left:** Initial image. Control point and local support of $4\Delta\varphi$ highlighted **Middle:** Extreme perturbation of control point and transformed image **Right:** Squared difference between initial and transformed image. The perturbation is confined in the region of local support only.

[Rueckert et al., 1999] or

$$r'' = \sum_{k=1}^{L} \mathcal{T}^k \left(r'; \varphi^k\right) \tag{5.20}$$

Finally, in order to enforce smoothness in the deformations modeled by FFD, Rueckert et al. [1999] proposed the usage of the TPS as a regularization in the FFD registration framework. The registration increases the similarity by perturbing the control points, but similar to TPS registration, the movement of pixels located among the control points is constrained so they minimize the thin plate bending energy. The corresponding objective function is defined as

$$\hat{\theta} = \arg\min_{\theta} \left[\mathcal{E}(\theta) = -\Psi\Big(\mathbf{x}(r), \mathbf{y}^{\mathfrak{T}}\big(\mathcal{T}(r';\theta)\big)\Big) + \tau J\big(\mathcal{T}(r';\theta)\big) \right], \quad r \in \Omega_{\mathbf{x};\mathbf{y}^{\mathfrak{T}}}. \tag{5.21}$$

where τ is the regularization parameter weighting the contribution of the TPS.

5.5.3 Interpolation

There exist various interpolation schemes which can be employed in order to re-define the transformed image $\mathbf{y}^{\mathcal{T}}(r'')$ - having continuous locations - to $\mathbf{y}^{\mathfrak{T}}(r), r \in \Omega_{\mathbf{x};\mathbf{y}^{\mathfrak{T}}}$, with r coinciding with the regular grid g where the target $\mathbf{x}$ is defined upon. Schemes include bi-linear, bi-cubic, spline, zero padding

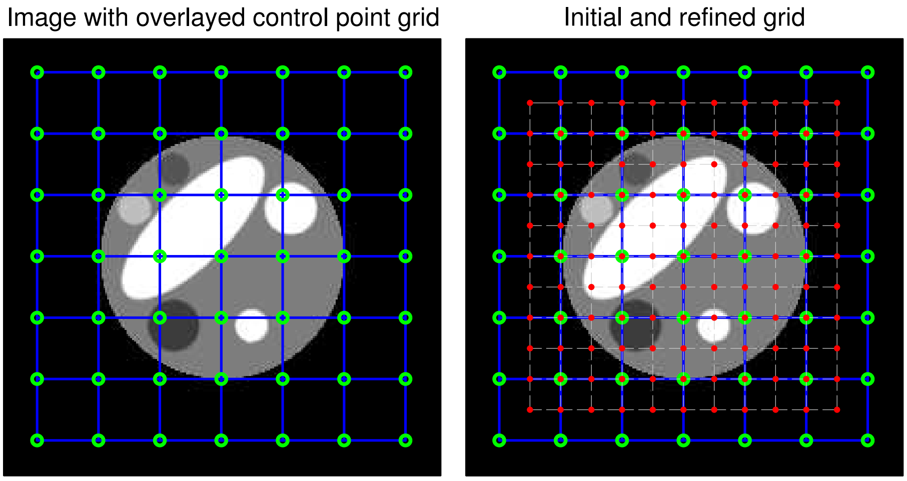


Image with overlayed control point grid

Initial and refined grid

Figure 5.5: Control point grid refinement. **Left:** Initial coarse lattice to model global transformations. **Right:** Refined lattice to model highly localized transformations.

fast Fourier transform (FFT) based interpolation [Bracewell, 1999; Press et al., 1992b] and more. Apart from the choice of the actual interpolation algorithm, one has the choice of a *forward* or a *backward* interpolation scheme. Consider the case of bi-linear interpolation (defined in Sec. 6.3.1). Consider Fig. 5.6. A forward interpolation scheme, directly populates $\mathbf{y}^{\mathfrak{T}}(r)$ from the values $\mathbf{y}^{\mathcal{T}}(r'')$ defined at the continuous r''. This case however can be problematic. It is possible that some r might not be assigned with a value $\mathbf{y}^{\mathfrak{T}}(r)$, if they do not lie inside the support of the interpolation kernels. On the contrary, the backward interpolation scheme (see Fig. 5.7) guarantees that such a problem does not occur. All locations r are explicitly visited. For any r, the inverse transformation $\mathcal{T}^{-1}(r;\theta)$ - given that it exists - results in the continuous $\tilde{r}' \in \Omega_{\mathbf{y}}$. The scheme continues by assigning $\tilde{r}'$ with a $\mathbf{y}^{\mathfrak{I}}(\tilde{r}')$ via interpolation - hence the $\mathfrak{I}$ notation as $\mathbf{y}$ is not defined in continuous locations. In this scheme, interpolation takes place in $\Omega_{\mathbf{y}}$ and not in $\Omega_{\mathbf{y}^{\mathcal{T}}}$ of the transformed coordinates. The value is finally transfered to the considered r, resulting in $\mathbf{y}^{\mathfrak{T}}(r) = \mathbf{y}^{\mathfrak{I}}(\tilde{r}')$. In the case that $\tilde{r}' \notin \Omega_{\mathbf{y}}$, then $\mathbf{y}^{\mathfrak{T}}(r)$ can be assigned with a neutral value, for example one which can be perceived as background.

5.6 Image similarity evaluation

The choice of the metric $\Psi\big(\mathbf{x}(r), \mathbf{y}^{\mathfrak{T}}(r)\big)$ is crucial to the performance of the registration process. The main reason is that different metrics can have - by design - different capacities, rendering them suitable for specific categories of registration tasks.

5.6.1 Sum of squared differences

In the case of intensity-based registration basis, the sum of squared differences (SSD) between the scalar gray values x and y of spatially corresponding pixel locations in images **x** and **y** is defined as

$$\Psi_{SSD}(\mathbf{x}(r), \mathbf{y}^{\mathfrak{T}}(r)) = \frac{1}{N_{\mathbf{x};\mathbf{y}^{\mathfrak{T}}}} \sum_{r \in \Omega_{\mathbf{x};\mathbf{y}^{\mathfrak{T}}}} \big(\mathbf{x}(r) - \mathbf{y}^{\mathfrak{T}}(r)\big)^2, \quad \forall r_k \in \Omega_{\mathbf{x};\mathbf{y}^{\mathfrak{T}}}. \tag{5.22}$$

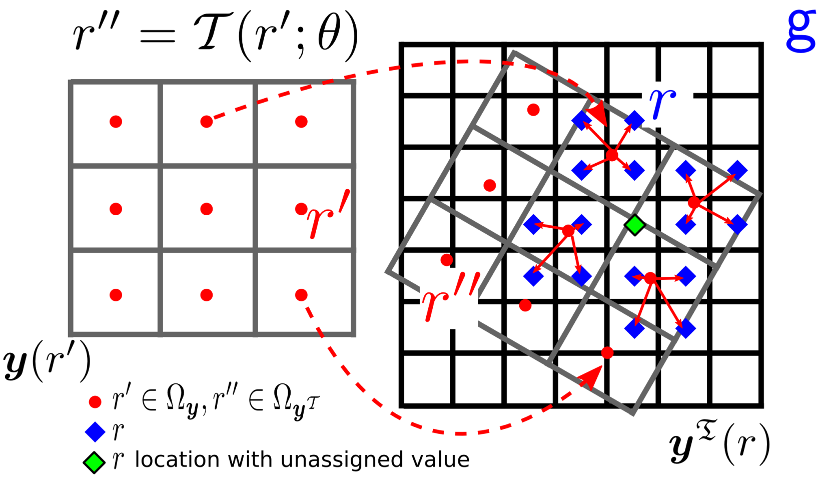

$r'' = \mathcal{T}(r';\theta)$
g
r
r'
r''
$\boldsymbol{y}(r')$
$\boldsymbol{y}^{\mathfrak{T}}(r)$
● $r' \in \Omega_{\boldsymbol{y}}, r'' \in \Omega_{\boldsymbol{y}^{\mathcal{T}}}$
◆ r
◇ r location with unassigned value

Figure 5.6: Forward interpolation. Locations $r' \in \Omega_{\mathbf{y}}$ are transformed to the continuous $r'' \in \Omega_{\mathbf{y}}$. The transformed image $\mathbf{y}^{\mathfrak{T}}$ has to be defined on the same grid g with $\mathbf{x}$, to enable similarity evaluation. Forward interpolation extrapolates $\mathbf{y}^{\mathcal{T}}(r'')$ to grid locations $r \in \Omega_{\mathbf{x};\mathbf{y}^{\mathfrak{T}}}$, resulting in $\mathbf{y}^{\mathfrak{T}}(r)$. Due to limited support some r are not assigned with a value.

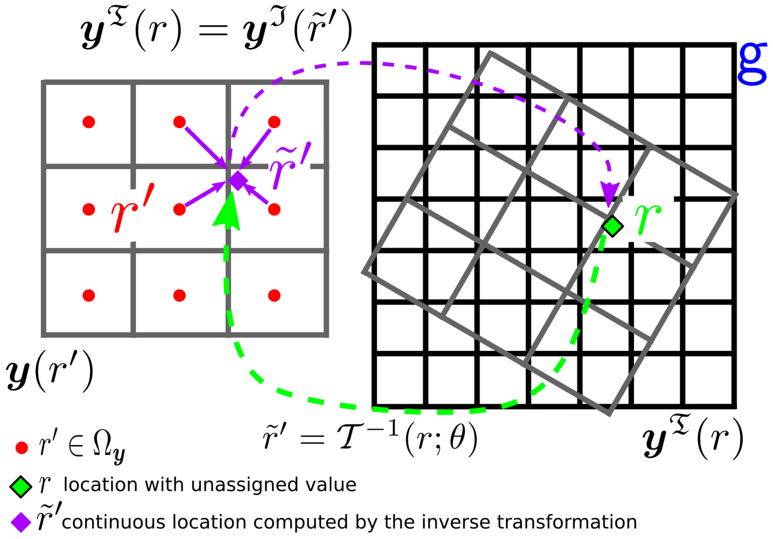

$\boldsymbol{y}^{\mathfrak{T}}(r) = \boldsymbol{y}^{\mathfrak{I}}(\tilde{r}')$
g
$\tilde{r}'$
r'
r
$\boldsymbol{y}(r')$
$\tilde{r}' = \mathcal{T}^{-1}(r;\theta)$
$\boldsymbol{y}^{\mathfrak{T}}(r)$
● $r' \in \Omega_{\boldsymbol{y}}$
◇ r location with unassigned value
◆ $\tilde{r}'$continuous location computed by the inverse transformation

Figure 5.7: Backward interpolation. All locations r are explicitly visited. For any r, the inverse transformation results in the continuous $\tilde{r}'$. Then, $\tilde{r}'$ is assigned with a value $\mathbf{y}^{\mathfrak{I}}(\tilde{r}')$ via interpolation - hence the $\mathfrak{I}$. This value is finally transfered to the considered r, resulting in $\mathbf{y}^{\mathfrak{T}}(r) = \mathbf{y}^{\mathfrak{I}}(\tilde{r}')$.

SSD can be expected to effectively evaluate the similarity between the involved images in cases where the gray values of corresponding structures present in both images, differ only by Gaussian noise [Viola, 1995]. Understandably the above condition can only be met in the intra-modality case, but even when the latter is the case, still it cannot be guaranteed. For example, in magnetic resonance imaging (MRI) images, although the noise in the area corresponding to the probed medium is approximately Gaussian, in the low intensity areas - corresponding to the air between probed medium and scanner - it follows a Rician distribution [Fitzpatrick et al., 2000; Gudbjartsson and Patz, 1995]. One way to bypass this problem is by excluding the pixels contaminated with non-Gaussian noise - if the latter can be identified - as done by Hajnal [2001]. In addition, SSD has been consistently employed as a similarity metric in corresponding points in landmark- and surface- based methods. In that case, similarity is interpreted by means of spatial distance between corresponding points. In that sense, SSD measures the squared Euclidean distance by replacing $\mathbf{x}(r)$ and $\mathbf{y}^{\mathcal{T}}(r)$ in Eq. 5.22 with the individual, corresponding landmarks coordinates, identified in both target and source images. Relevant work includes the *orthogonal Procrustes* method [Arun et al., 1987; Fitzpatrick et al., 1998], the *head-and-hat* algorithm [Pelizzari et al., 1989] and the *iterative closest point* [Besl and McKay, 1992; Maurer, Jr., C.R. et al., 1998].

5.6.2 Correlation techniques

Cross correlation (CC) or cross-covariance [Hill et al., 2001] is an intensity based measure which exhibits additional flexibility with respect to SSD. Specifically, it can evaluate the similarity between images to which their corresponding gray values are linearly related. It is expressed as:

$$\Psi_{CC}\big(\mathbf{x}(r), \mathbf{y}^{\mathfrak{T}}(r)\big) = \sum_{r \in \Omega_{\mathbf{x};\mathbf{y}^{\mathfrak{T}}}} \mathbf{x}(r)\mathbf{y}^{\mathfrak{T}}(r) \tag{5.23}$$

A normalized variant of CC, is the normalized cross correlation (normalized cross correlation (NCC)) - also known as correlation coefficient. NCC is equal to 1 if the relationship between the two signals is perfectly linear and zero if the relationship is random. It is insensitive to differences in mean signal intensity and it is also insensitive to noise in the low-intensity areas which cause problems to SSD [Lemieux and Barker, 1998; Lemieux et al., 1994]. NCC is defined as:

$$\Psi_{NCC}(\mathbf{x}(r), \mathbf{y}^{\mathcal{T}}(r)) = \frac{\sum_{r_i \in \Omega_{\mathbf{x};\mathbf{y}^{\mathfrak{T}}}} (\mathbf{x}(r_i) - \bar{\mathbf{x}})(\mathbf{y}^{\mathcal{T}}(r_i) - \overline{\mathbf{y}^{\mathcal{T}}})}{\left(\sum_{r_i \in \Omega_{\mathbf{x};\mathbf{y}^{\mathfrak{T}}}} (\mathbf{x}(r_i) - \bar{\mathbf{x}})^2\right)^{1/2} \left(\sum_{r_i \in \Omega_{\mathbf{x};\mathbf{y}^{\mathfrak{T}}}} (\mathbf{y}^{\mathcal{T}}(r_i) - \overline{\mathbf{y}^{\mathcal{T}}})^2\right)^{1/2}} \tag{5.24}$$

where $\bar{\mathbf{x}}, \overline{\mathbf{y}^{\mathcal{T}}}$ denote the mean gray values in each image. The correlation techniques can be used in multi-modal applications [Junck et al., 1990]. Multi-modal images are usually non-linearly related which renders both functionals ineffective. Still, there have been studies succeeding to register multi-modal images, by applying the functionals on secondary images, comprised by the edges, ridges or segmented regions of the original ones [Arbel et al., 2001; Maintz et al., 1995]. It is important to emphasize on this concept that even when $\mathbf{x}$ and $\mathbf{y}^{\mathcal{T}}$ are multi-modal, it is possible to derive secondary images for which similarity is more trivial to evaluate.

Finally, a most recent correlation related similarity measure was proposed by Roche et al. [1998a,b] known as correlation ratio. It is inherently more suitable for multi-modal applications as it is not limited to image pairs related by a linear relationship, but by a more generic *functional* relationship.

5.6.3 Ratio-Image Uniformity and Partitioned Intensity Uniformity

Ratio-Image Uniformity and Partitioned Intensity Uniformity are intra-modal and multi-modal similarity measures respectively, proposed by Woods et al. [1992, 1993]. RIU involved the computation of a ratio image $z = \mathbf{x}/\mathbf{y}^{\mathcal{T}}$ and a subsequent computation of the normalized standard deviation of z. When the involved images were aligned, the normalized standard deviation would be minimized. Partitioned intensity uniformity is in effect a generalisation of the ratio-image uniformity concept, but for multi-modal images. It is based in the assumption that although trans-image, corresponding regions of specific anatomical structure or functional activity were depicted by different gray values due to the different modalities, within each image the gray values populating a single region would be similar. Partitioned intensity uniformity requires a segmentation process in order to identify distinct regions in **x**. The value of pixels populating this region would comprise an intensity partition. The algorithm would then attempt to estimate a $\mathbf{y}^{\mathcal{T}}$ for which, the gray values of its pixels that spatially corresponded to a single region of **x**, would need to exhibit minimal standard deviation from their mean value.

5.6.4 Information theoretic similarity measures in image registration

Information theoretic functionals, such as the ones introduced in Chapter 4, have been the dominant choice for image registration similarity measures. Section 4.5.5 discussed their inherent invariance - partial in absolute terms for the case of differential entropy (see Sec. 4.5.3) - to the gray values of the involved images. Understandably, this capacity has made them ideal candidates for medical image registration problems, especially multi-modality ones. joint entropy (JE) was proposed for image registration purposes in [Collignon et al., 1995; Studholme et al., 1995]. Mutual information which is the most widely employed registration functional was independently proposed for this purpose by Viola [1995]; Wells et al. [1996] (MIT, USA) and Collignon et al. [1995]; Maes et al. [1997] (University of Leuven, Belgium). A survey for mutual information (MI) based registration was conducted by Pluim et al. [2003]. Rueckert et al. [2000] proposed the usage of higher order mutual information, in order to introduce the inter-pixel spatial dependence within a single image, which lacks in most information theoretic implementations. Somayajula et al. [2008] introduced spatial information implicitly, by considering the mutual information between the intensities of the images as well as an extended feature space derived by the application of differential and low-pass filter operators. The information theoretic registration literature is vast. In this section we revisit the concepts introduces in Chapter 4 and we explain the differences among information theoretic functionals in the registration framework.

5.6.4.1 Joint entropy vs Mutual information

The JE $h\big(\mathbf{x}(r), \mathbf{y}^{\mathcal{T}}(r)\big)$, $r \in \Omega_{\mathbf{x};\mathbf{y}^{\mathcal{T}}}$ is expected to attain a minimum when two images depicting similar information are correctly aligned. It can be shown however that for specific types of images this minimum is not global. Lower JE minima exist for incorrect alignment configurations between the test images, compromising the overall accuracy of the registration via JE.

Specifically, the problematic behavior manifests in images where the main depicted feature does not occupy the entire image area, allowing a substantial number of pixels to comprise the *background* region. The background pixels usually differ only by noise and appear to be almost uniformly distributed considering the dynamic range of the entire image. The incorrect minima in the JE case manifest when the overlap domain $\Omega_{\mathbf{x};\mathbf{y}^{\mathfrak{T}}}$ between $\mathbf{x}$ and $\mathbf{y}^{\mathfrak{T}}$ is comprised by a large number of these background pixels and with partial or total absence of the the main feature.

Fig. 5.8 showcases the problematic JE case, which does not affect MI and consequently gives a clear advantage of the latter over JE. An image depicting a transverse slice of an MRI volume is translated over its identical copy, creating $\Omega_{\mathbf{x};\mathbf{y}^{\mathfrak{T}}}$ with variable depicted information.

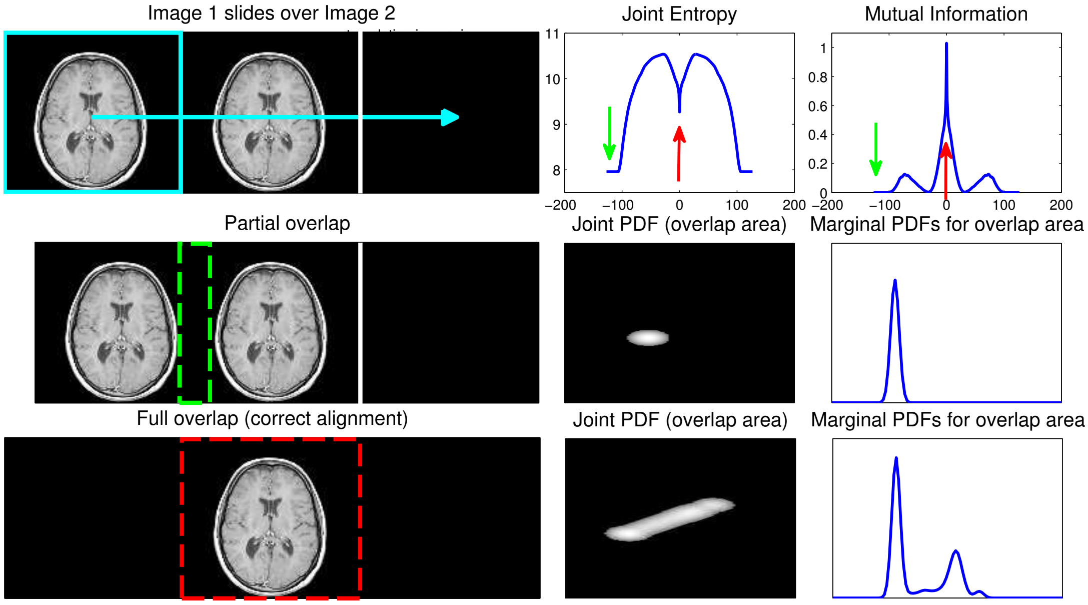


Figure 5.8: JE vs MI in image registration. **Top row - left:** Image 1 slides over identical Image 2 creating different overlap areas $\Omega_{\mathbf{x};\mathbf{y}^{\mathfrak{T}}}$. **Right:** Plots of JE and MI versus translation of moving image. The errors corresponding to the following cases are highlighted. **Middle row - left:** Partial image overlap. **Right:** JPDF and PDFs (only one is depicted as they are identical for both images) for the highlighted $\Omega_{\mathbf{x};\mathbf{y}^{\mathfrak{T}}}$. This is the JE global minimum case and evidently does not correspond to the correct alignment. **Bottom row - left:** Correct alignment. $\Omega_{\mathbf{x};\mathbf{y}^{\mathfrak{T}}}$ comprises by the entire image area. MI accurately attains a global maximum. JE attains a local minimum. **Right:** JPDF and PDFs for overlap area are also depicted.

The overlap configuration depicted in the middle row, corresponds to the global minimum of JE which manifests when $\Omega_{\mathbf{x};\mathbf{y}^{\mathfrak{T}}}$ is populated entirely by background pixels. The JPDF of the overlap area is maximally clustered, comprised by a single high probability cluster corresponding to the event $\{\mathbf{x} = \text{background}, \mathbf{y} = \text{background}\}$ (see Sec. 4.2.5 for a description of random events in the imaging context). On the contrary $MI\big(\mathbf{x}(r), \mathbf{y}^{\mathfrak{T}}(r)\big) = h\big(\mathbf{x}(r)\big) + h\big(\mathbf{y}^{\mathfrak{T}}(r)\big) - h\big(\mathbf{x}(r), \mathbf{y}^{\mathfrak{T}}(r)\big)$, $r \in \Omega_{\mathbf{x};\mathbf{y}^{\mathfrak{T}}}$, depends on JE as well as on the marginal entropy terms. The global maximum of MI - attained for correct alignment - corresponds to a balance between the minimization of $h\big(\mathbf{x}(r), \mathbf{y}^{\mathfrak{T}}(r)\big)$ and the maximization of $h\big(\mathbf{x}(r)\big)$ and $h\big(\mathbf{y}^{\mathfrak{T}}(r)\big)$. In the depicted case (middle row), the marginal PDFs of **x** and **y** for $\Omega_{\mathbf{x};\mathbf{y}^{\mathfrak{T}}}$, are maximally clustered as the $\Omega_{\mathbf{x};\mathbf{y}^{\mathfrak{T}}}$ is uniform. These PDFs correspond to the minimum marginal entropies, contrary to what is needed for *MI* maximization. Actually, the value attained by MI in this case approaches the global minimum of *MI*. The uniform images have zero marginal uncertainty (see Sec. 4.5.5). From Eq. 4.48, given that the marginal uncertainty (or entropy) is already zero, it cannot be further decreased by the conditional uncertainty, hence the zero MI.

The bottom row depicts the case of correct alignment. JE attains a minimum as the JPDF is highly clustered, however not as clustered as the JPDF between the uniform images. Hence JE attains a minimum however it is not global. MI benefits by the reduced JE value, however its maximum is additionally amplified by the dispersed PDFs, corresponding to increased $H\big(\mathbf{x}(r)\big)$ and $H\big(\mathbf{y}^{\mathfrak{T}}(r)\big)$.

To conclude, MI attempts to register areas which are similar but are also populated by pixels which exhibit high variation in intensities. The condition is met when the main feature in the image is depicted in $\Omega_{\mathbf{x};\mathbf{y}^{\mathfrak{T}}}$. On the contrary, JE attempts to register areas which are similar but are comprised by pixels of ideally uniform intensity, a behavior which can compromise the alignment effort. It has to be noted that MI contains two local maxima which can compromise the recovery of alignment if the initial misregistration is high.

5.6.4.2 Normalized mutual information

The previous section established that MI is more robust than JE as a similarity measure for image registration. However, there are still cases where MI fails to score maximum similarity for the correct registration and favours inaccurate alignment configurations. These problematic cases depend on the ratio of the pixels in $\Omega_{\mathbf{x};\mathbf{y}^{\mathfrak{T}}}$ which are labeled as background, over the number of pixels in $\Omega_{\mathbf{x};\mathbf{y}^{\mathfrak{T}}}$ depicting the main feature. Consider Fig. 5.9.

The left column shows the target image with variable sized field of view (FOV). As the FOV scale increases, the background of the images is extended only in the horizontal direction. Let the target images act also as the moving images, by rotating them about their c.o.m. in a range of angles $\Theta = -30^\circ \rightarrow 30^\circ$ with increments of $\Delta\Theta = 5^\circ$. The right column of Fig. 5.9, depicts two alignment configurations for FOV scale 3, highlighting the corresponding $\Omega_{\mathbf{x};\mathbf{y}^{\mathfrak{T}}}$. The schematic depicts one of the two cases of highest misalignment considered in this test $(\Theta = -30^\circ)$, whereas the bottom schematic depicts the correct alignment configuration, where both images are correctly superimposed $(\Theta = 0^\circ)$. We proceed by computing the discrete marginal entropies $H(\mathbf{x})$, $H(\mathbf{y})$, JE and MI on the $\Omega_{\mathbf{x};\mathbf{y}^{\mathfrak{T}}}$, formed by the transformed moving image and the target image of the corresponding FOV scale, for all angles of rotation and for all FOV scales. The results of these evaluations are depicted in the five plots of Fig. 5.10.

As expected, the entropic functionals attain smaller values for higher FOV scales, as the background increases. The fourth plot showcases the problem regarding MI. At the first FOV scale, MI correctly attains a maximum for correct alignment case of $\Theta = 0^\circ$. This is not the case however for FOV scale 3, where the maximum is now replaced by a minimum. Registration by maximization of MI would not be able to recover the rotational misalignment for the images of FOV scale 3.

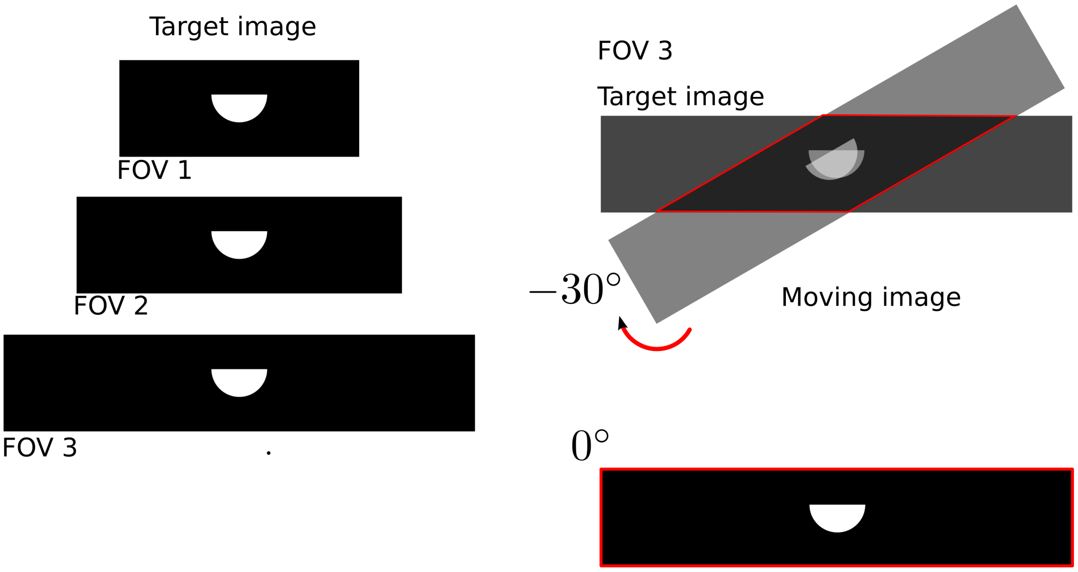


Figure 5.9: Framework for testing the response of MI to the ratio of the pixels in the $\Omega_{\mathbf{x};\mathbf{y}^{\mathfrak{T}}}$ which are labeled as background, over the number of pixels in the same domain depicting the main feature. **Left:** Target images at three FOV scales (see text). **Right:** Maximum mis-alignment configuration at $\Theta = -30^\circ$ and correct alignment at $\Theta = 0^\circ$. $\Omega_{\mathbf{x};\mathbf{y}^{\mathfrak{T}}}$ is highlighted for clarity.

The explanation of this unexpected behavior can be given by considering the marginal and JE terms in MI. The correct alignment depicted in Fig. 5.9 has minimum JE due to alignment. However, the marginal entropies attain very low values due to the large background in $\Omega_{\mathbf{x};\mathbf{y}^{\mathfrak{T}}}$, which reduces uncertainty. In the case of maximum misalignment JE increases by definition, as the JPDF becomes more dispersed (not depicted). However, due to the smaller background size, the corresponding marginal entropies are now higher than before hence the higher overall MI. To conclude, the MI optima are obtained by a fine balance between the sum of the marginal entropy terms and the JE term. For cases such as the one discussed above, this balance favours $\Omega_{\mathbf{x};\mathbf{y}^{\mathfrak{T}}}$ configurations which do not correspond to the true alignment.

To alleviate this problematic effect, various modified MI functionals have been proposed, which normalize the effects of the background size. These are collectively known as normalized mutual information (NMI). The reader is refered to [Hill et al., 2001] for the various expressions. We refer to the most prominent one proposed by Studholme et al. [1999] which is defined as

$$NMI(\mathbf{x}(r), \mathbf{y}^{\mathfrak{T}}(r)) = \frac{h\big(\mathbf{x}(r)\big) + h\big(\mathbf{y}^{\mathfrak{T}}(r)\big)}{h\big(\mathbf{x}(r), \mathbf{y}^{\mathfrak{T}}(r)\big)}, \quad \forall r \in \Omega_{\mathbf{x};\mathbf{y}^{\mathfrak{T}}}. \tag{5.25}$$

For a mathematical explanation of why the above functional successfully normalizes MI see [Melbourne et al., 2009]. The NMI values for simulation of Fig. 5.9 are shown in the fifth plot in Fig. 5.10. Evidently NMI exhibits increased invariance to the size of the background in $\Omega_{\mathbf{x};\mathbf{y}^{\mathfrak{T}}}$ and successfully attains a maximum for $\Theta = 0^\circ$ at all FOV scales.

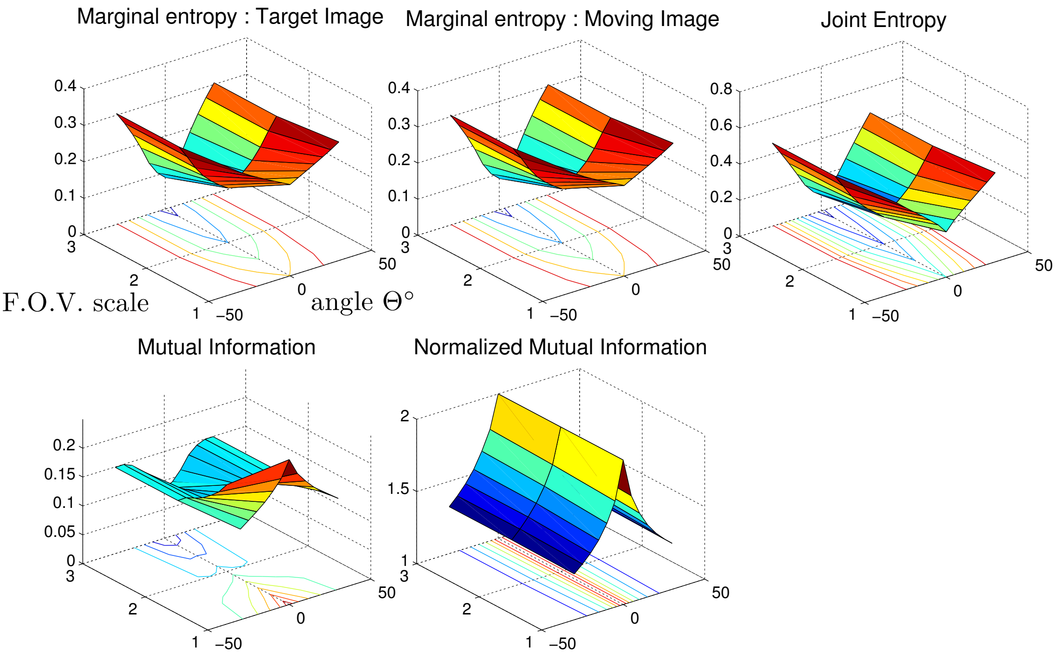


Figure 5.10: Response of the information theoretic functionals to the various alignment configurations of the test case described by Fig. 5.9.

5.7 Optimization

Optimization in the image registration context regards the retrieval of an accurate estimate of the spatial transformation parameters θ applied on $\mathbf{y}^{\mathfrak{T}}$, so that the latter becomes similar to $\mathbf{x}$ with respect to some similarity measure. Many approaches from the optimization related literature have been adopted in order to perform the above task, including ordinary gradient based schemes such as the one used in [Rueckert et al., 1999]; methods which perform gradient based optimization in a stochastic framework [Viola, 1995]; genetic and evolution based optimization [Price et al., 2005; Rouet et al., 2000] and more. A sample of similar optimization schemes were introduced in a general context in Sec. 2.6.

5.7.1 Multi-resolution pyramids

More interesting to review are image registration specific approaches which are often employed to improve global convergence. Intensity based similarity metrics are often plagued by local minima which compromise global convergence. See for example the MI plot in Fig. 5.8. One approach towards the alleviation of these local minima is to employ a multi-resolution pyramid scheme [Maes et al., 1999; Thï£¡venaz et al., 1998; Thï£¡venaz and Unser, 2000]. The basic principle of such schemes is quite simple. The images to be registered are sub-sampled to coarser resolution levels. Registration is firstly performed on the coarser resolutions where local minima are hopefully smoothed out due to the lesser details depicted in the involved coarse images. When the optimization has converged and registration has been established, the scheme continues by registering images at a finer resolution level. However, the initial transformation parameter estimate for each resolution level is obtained by the converged registration of the previously visited coarser level. The process continues until the algorithm registers the images at their original resolutions. Registration in coarser resolution usually compensates for large spatial discrepancies whereas local misalignments are handled in the finer resolution levels. Figure 5.11 repeats the test of Fig. 5.8, but in a coarser resolution. The involved images were down-sampled from the original $128x128$ to $33x33$. Apparently, the local minima at the MI plot have been smoothed out.

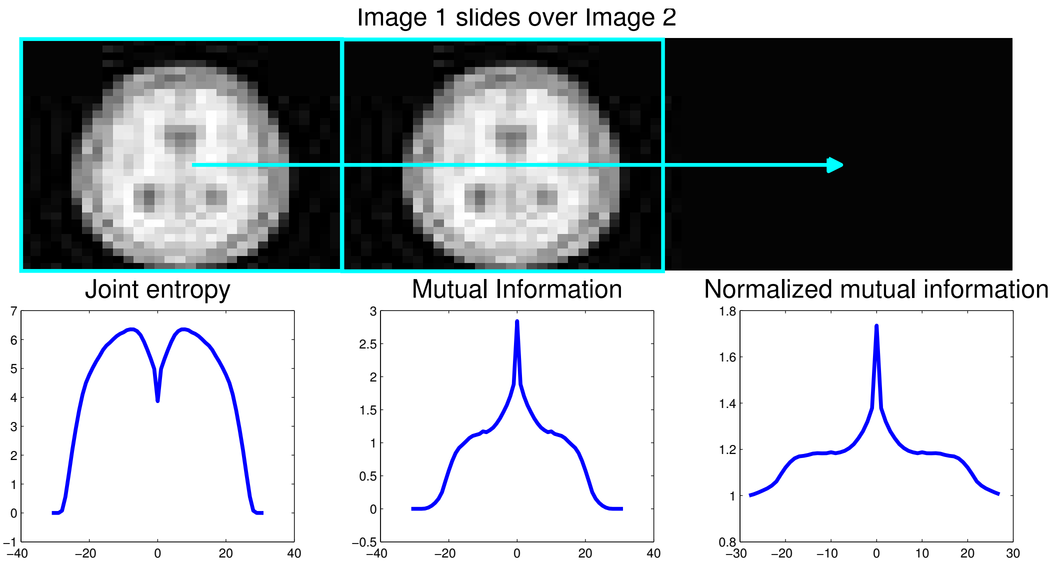


Figure 5.11: Information theoretic, intensity based similarity measures and image resolution levels. Image 1 is translated across Image 2. Similarity measures are computed for the various overlap configurations. Due to the coarse resolution of the images, the global optima have a large basin of attraction, whereas local optima have been largely smoothed out.

5.7.2 Interpolation artefacts and blurring

As was already noted at the start of this chapter, interpolation is required to transfer the transformed $\mathbf{y}^{\mathcal{T}}(\tilde{r};\theta)$ to the r, where the target image $\mathbf{x}(r)$ is defined, in order to enable similarity evaluation. The gray values of the resulting image $\mathbf{y}^{\mathfrak{T}}(r,\theta)$ are a product of interpolation. According to Hill et al. [2001], subversion interpolation can result in blurring of $\mathbf{y}^{\mathcal{T}}$. Details are removed from the image which leads to an entropy reduction and potential introduction of local extrema in the solution space. One way to alleviate this problem is to apply a low-pass filter on the images prior to registration. Hill argues that although the loss of resolution that results from this pre-emptive blurring is a disadvantage, the registration errors caused by the interpolation errors can be greater than the loss of precision resulting from blurring. To conclude, pre-blurring the images prior to registration can increase the chances of global convergence.

5.8 Conclusion

This chapter has briefly introduced the concept of image registration. It introduced the three fundamental algorithmic parts comprising a registration scheme, namely the spatial transformations, similarity measures and optimization. Regarding spatial transformations, both linear and non-linear schemes were discussed with emphasis on the B-Spline free form deformations adopted later in this work. A brief reference to inverse interpolation as a part of spatial transformations was also given. A sample of intensity-based similarity measures was introduced, however more emphasis was given to the ones based on information theoretic functionals, which are the core concepts of this work. Intuitive examples comparing the capacity of each of the functionals and showcasing the strengths and weaknesses were also provided. Finally, we referred to a small sample of approaches which aid optimization convergence in the image registration setting.

Part III

Methods and Results

Chapter 6

Efficient entropy and derivative computation

In this chapter we propose a computationally efficient implementation of the entropic terms involved in the information theoretic regularisation of diffuse optical tomography (DOT), introduced in Chapter 7. The regularisation scheme is incorporated in the general framework of the inverse problem which employs a gradient based optimization for the search of the optical quantities of interest. Consequently, the efficient evaluation of the information theoretic functionals - namely joint entropy (JE) or mutual information (MI) - as well as their derivatives with respect to the optimized quantities is required. The proposed implementation achieves computational savings in the evaluation of both functionals and their derivatives.

The method is inspired by the work of Shwartz et al. [2005] who proposed a scheme for the efficient computation of the empirical marginal entropy (see Section 4.5.4) and its derivative. The contributions of the work presented in this chapter include:

I) The extension of the fast marginal entropic functional and gradient evaluation in order to enable the efficient evaluation of JE and its derivatives.

II) The derivation of the fast marginal and JE derivatives for the case of the Shannon integral entropy formulation (Shwartz et al. [2005] reported the derivatives with respect to the expectation based formulation of entropy)

III) By utilizing the same principle which leads to computational complexity reduction for the case of empirical entropy, we also apply the same scheme on the computation of the classic Shannon's entropy integral formulation, as well as its derivatives. We finally compare both entropic estimators - the empirical and classic Shannon integral - for accuracy and speed and discuss the findings.

The method is published in [Panagiotou et al., 2009b] where it was used to incorporate structural priors in DOT. In addition, it has been used for information theoretic regularization of positron emission tomography (PET) using structural priors [Kazantsev et al., 2010; Pedemonte et al., 2010a; Somayajula et al., 2010] as well as Pedemonte et al. [2010b] which is work-in-progress.

The structure of this chapter is as follows: Section 6.1 describes the method for efficient evaluation of the marginal entropy of two entropic estimators. The method is based on a scheme for fast probability density function (PDF) evaluation, via the interpolation of continuous quantities on a regular grid and then utilizing the fast Fourier transform (FFT) to efficiently compute the kernel density estimation (KDE) due to the convolution structure of the latter. The same section presents a series of validation simulations, designed and performed to test the accuracy of the entropic estimators. In addition we test how the entropic estimates are affected by different choices of parameters, in the employed non-parametric KDE process, such as the kernel width and numbers of quantization bins. Sec. 6.2 describes a computationally efficient scheme for computing the derivatives of the two marginal entropy estimators. In a similar manner, sections 6.3 and 6.4 extend the efficient entropy evaluation and derivative estimation to the JE case. Finally, Sec. 6.5 comments on the efficient estimation of the MI between two random variables (RVs).

6.1 Computationally efficient marginal entropy evaluation

Let **x** be the underlying RV governing the gray values of an image. The image itself is considered to be a sample of **x**, where the gray value of each pixel is considered an independent trial of **x** at some location $r_i \in \Omega_{\mathbf{x}}$. In effect this implies that all gray values are assumed to be independent and identically distributed (i.i.d.) - hence no spatial inter-pixel dependence within a single image is considered. This clearly an erroneous assumption which does not reflect reality as natural images - including medical images - are structured. Pixels which belong to the same anatomical class are more likely to attain similar values. Hence, the PDF describing their gray values is conditioned with respect to spatial location, or similarly the values of their neighbours. However, the independent and identically distributed (i.i.d.) assumption is commonly adopted by entropy estimation methods in medical imaging.

We define a sample A as $A = \{\alpha_1, \alpha_2, \ldots, \alpha_N\}$, where N denotes its size and $\alpha_k = \mathbf{x}(r_k)$ denotes the gray values for pixel locations $r_k \in \Omega_{\mathbf{x}}$. We note that $\alpha_k \in \mathbb{R}, \forall i$.

The aim is to estimate the marginal entropy of A. Firstly, the Shannon's entropy of **x**, re-iterated here for convenience, is expressed as

$$h_s(\mathbf{x}) = -\int_{\mathbb{R}} p^{\star}_{\mathbf{x}}(x) \log\left(p^{\star}_{\mathbf{x}}(x)\right) \mathrm{d}x, \tag{6.1}$$

where $p^{\star}_{\mathbf{x}}$ is the true PDF of **x**. The expression can be approximated using standard numerical integration techniques, such as Riemann integration using the trapezoidal rule. One can utilize the sample A to obtain estimates $\hat{p}_{\mathbf{x}}(\breve{x}_i; A)$ of $p^{\star}_{\mathbf{x}}(x)$, at regularly spaced locations $\breve{x} = \{\breve{x}_1, \breve{x}_2, \ldots, \breve{x}_{\breve{N}}\}$, $\breve{x}_i \in \mathbb{R}$. The spacing between the sample points is $\Delta\breve{x} = \breve{x}_{i+1} - \breve{x}_i$, $i = 1 \ldots \breve{N} - 1$. The approximation of Eq. 6.1 is expressed as

$$\hat{h}_s(\mathbf{x}; A) = -\sum_{i=1}^{\breve{N}} \hat{p}_{\mathbf{x}}(\breve{x}_i; A) \log\left(\hat{p}_{\mathbf{x}}(\breve{x}_i; A)\right) \Delta\breve{x}, \tag{6.2}$$

Consider also the expectation formulation of Eq. 6.1 which will be referred for convenience as empirical entropy. For the case of the finite sample A it is given by

$$\hat{h}_e(\mathbf{x}; A) = -\frac{1}{N} \sum_{k=1}^{N} \log\left(\hat{p}_{\mathbf{x}}(\alpha_k; A)\right). \tag{6.3}$$

Both formulations include a summation over PDF entries, either at $\breve{x}$ for the case of Shannon entropy or directly at α for expectation formulation. The next section describes the adopted approach for computationally efficient marginal PDF estimation proposed by [Silverman, 1982; Silverman and Green, 1986].

6.1.1 Efficient marginal PDF estimation

For the purpose of PDF estimation, we employ a non-parametric kernel density estimation (KDE) which returns a continuous estimate and does not require *a priori* assumptions regarding the form of the underlying unknown density (see Section 4.4.2).

Consider a KDE such as the one introduced in Section 4.4.2 which utilizes the entire N-sized sample A in order to obtain continuous estimates. We provide the expression here for convenience

$$\hat{p}_{\mathbf{x}}(x; A) = \frac{1}{N}\sum_{k=1}^{N} K_u(x - \alpha_k). \tag{6.4}$$

We employ a Gaussian kernel of standard deviation u

$$K_u(\breve{x}_j - \alpha_k) = \frac{1}{\sqrt{2\pi u}} \exp\left(-\frac{(\breve{x}_j - \alpha_k)^2}{2u^2}\right). \tag{6.5}$$

The computation of the PDF at a single point requires N kernel evaluations. Hence, populating the entire $\breve{N}$ grid points in order to perform the numerical integration of Shannon integral of Eq. 6.2 leads to a computational complexity of $\mathcal{O}(\breve{N}N)$. In the case of empirical entropy of Eq. 6.3, the complexity is $\mathcal{O}(N^2)$.

The complexity of both estimators can be improved by reducing the complexity of the employed KDE. As will become apparent, the KDE is a convolution operation. Before proceeding, consider the following definitions.

The convolution between two continuous functions $f(x),\ g(x)$ is expressed as

$$f(x) \star g(x) = \int_{-\infty}^{\infty} f(x-s)g(s)\mathrm{d}s = \int_{-\infty}^{\infty} f(s)g(x-s)\mathrm{d}s. \tag{6.6}$$

The convolution of a signal $f(x)$ with a shifted Dirac delta function $\delta(x-a)$ (see Eq. 4.53 for the definition of $\delta(x-a)$) returns the shifted original signal $f(x-a)$ [Chui, 2008] or

$$f(x) \star \delta(x-a) = \int_{-\infty}^{\infty} f(x-s)\delta(s-a)\mathrm{d}s = f(x-a). \tag{6.7}$$

The convolution structure of the KDE becomes now apparent as

$$\begin{aligned}\hat{p}_{\mathbf{x}}(x;A) &= \frac{1}{N}\sum_{k=1}^{N} K_u(x-\alpha_k) &(6.8)\\ &\overset{6.7}{=} \frac{1}{N}\sum_{k=1}^{N}\int_{-\infty}^{\infty} K_u(x-s)\delta(s-\alpha_k)\mathrm{d}s &(6.9)\\ &\overset{6.7}{=} K_u(x)\star\frac{1}{N}\sum_{k=1}^{N}\delta(x-\alpha_k) &(6.10)\\ &= K_u(x)\star\frac{1}{N}\mathrm{III}(x-\alpha_k) &(6.11)\end{aligned}$$

where $\mathrm{III}(x-\alpha)=\sum_{i=1}^{N}\delta(x-\alpha)\mathrm{d}x$ is a continuous impulse train.

In addition, the convolution theorem [Bracewell, 1999] states that the convolution between two functions $f(x)$, $g(x)$, is equivalent to the product of their Fourier transforms, subsequently subjected to an inverse Fourier transform. Let $\mathfrak{F}\left\{f(x)\right\}$, $\mathfrak{F}\left\{g(x)\right\}$ denote the Fourier transforms of each function and $\mathfrak{F}^{-1}\left\{\cdot\right\}$ be the inverse Fourier transform . Then the convolution theorem is expressed as

$$f(x)\star g(x)=\mathfrak{F}^{-1}\Big\{\mathfrak{F}\left\{f(x)\right\}\times\mathfrak{F}\left\{g(x)\right\}\Big\} \tag{6.12}$$

Silverman *et al.* [1982; 1986] was the first to propose the employment of the Fourier transform in order to perform a convolution with significantly reduced cost. It is known that in the case of discrete signals - meaning that $\alpha\in A$ are regularly spaced - one can employ the FFT implementation of the Discrete Fourier transform, which transforms an N signal in the frequency domain in $\mathcal{O}(N\log N)$ [Bracewell, 1999; Cooley and Tukey, 1965].

We employ Silverman's approach, where the initial non-equispaced N-size sample A is re-sampled on a grid of $\breve{N}$ regularly spaced locations $\breve{x}$ with spacing $\Delta\breve{x}$. Note that this is the same grid of locations where the numerical integration of Shannon entropy will be performed. The resulting regularly spaced version of A is expressed as $\breve{A}=\{\breve{\alpha}_1,\breve{\alpha}_2,\ldots,\breve{\alpha}_{\breve{N}}\}$, where $\breve{\alpha}_i=w(\breve{x}_i)$. The weight $w(\breve{x}_i)$ represents the density of the original continuous sample A in the vicinity of $\breve{x}_i$. To understand the nature of $w(\breve{x})$ consider the following. In the case of the continuous A, each $\alpha_k\in A$ corresponds to an impulse $\delta(x-\alpha_k)$. The density of A would be encoded in the frequency which these non-equispaced impulses appeared in the continuous impulse train. However in the case of $\breve{A}$, the comprising $\breve{\alpha}$ are defined over fixed equispaced locations $\breve{x}$, hence the density cannot be represent in terms of the frequency for which they appear in the domain. Instead, the density is encoded by amplitude so each $\breve{x}_i$ is associated with the weight $w(\breve{x}_i)$ which reflects the density of the *original* sample points α in the vicinity of $\breve{x}_i$. Each weight $w(\breve{x}_i)$ is computed via a linear interpolation

$$w(\breve{x}_i)=\frac{1}{N}\sum_{j=1}^{N}\wedge(\breve{x}_i-\alpha_j), \tag{6.13}$$

where $\wedge(\cdot)$ is a triangular kernel defined as

$$\wedge(u) = \begin{cases} 1 - \left(|u| / \Delta\breve{x}\right), & \text{if } |u| < \Delta\breve{x} \\ 0, & \text{otherwise.} \end{cases} \tag{6.14}$$

Figure 6.1 graphically shows the re-sampling of a continuous sample A with frequency-encoded density, to the equispaced $\breve{A}$ with amplitude encoded density.

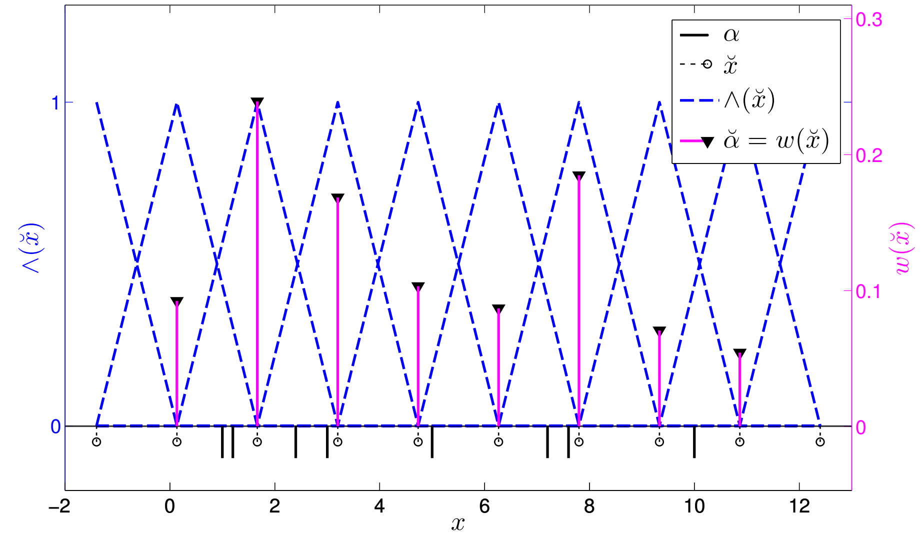


Figure 6.1: The continuous gray values $\alpha \in A$ are interpolated to a regularly spaced grid $\breve{x}$. The originally frequency encoded density of A is now reflected by the weights $w(\breve{x})$ fixed over the equispaced $\breve{x}$.

It is apparent from the nature of $\wedge(\cdot)$ that each $\alpha_k \in \left[\breve{x}_j, \breve{x}_{j+1}\right]$ contributes solely to the weights of its neighbouring $\breve{x}_j, \breve{x}_{j+1}$ with the individual weights being $w(\breve{x}_j) = (1 - \mathfrak{b}_i)/N$ and $w(\breve{x}_{j+1}) = \mathfrak{b}_i/N$, with $\mathfrak{b}_i = \left(\left|\alpha_k - \breve{x}_j\right| / \Delta\breve{x}\right)$. The values in $\mathfrak{b}_k, \forall k$ are the normalized distance between each continuous $\alpha_k \in A$ and its left neighbouring $\breve{x}_i$. Shwartz et al. [2005] proposed that the values $\mathfrak{b}_k$ can be stored into an array, in order to reduce the complexity of subsequent interpolations, as they are later used to transfer quantities defined in α to $\breve{x}$ and the opposite. For that purpose, one additional quantity needs to be stored in an array, which is the index of left neighbouring $\breve{x}_i$ of each α_k. This is explicitly stored in a second array $\mathfrak{i}_k = i$. Figure 6.2 graphically represents the nature of the entries in $\mathfrak{b}$.

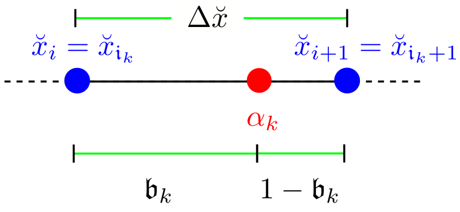

$\Delta\breve{x}$
$\breve{x}_i = \breve{x}_{\mathfrak{i}_k}$
$\breve{x}_{i+1} = \breve{x}_{\mathfrak{i}_k+1}$
α_k
$\mathfrak{b}_k$
$1 - \mathfrak{b}_k$

Figure 6.2: Graphical representation of the entries in array $\mathfrak{b}$, which is used for fast interpolations between $\breve{x}$ and α. Array $\mathfrak{b}$ stores the normalized distance between a continuous α_k and its immediate neighbour $\breve{x}_i$ for which $\breve{x}_j < \alpha_i$. The array $\mathfrak{i}$ stores the index i of the left neighbouring grid point.

A probability density estimate $\hat{p}^b_{\mathbf{x}}(\breve{x}_i; \breve{A})$ is retrieved for $\breve{x}_i, \forall i$, by weighting the contribution from all $w(\breve{x})$ using the kernel function $K_u(\cdot)$. This gives rise to the *binned kernel density estimator* [Silverman and Green, 1986] which effectively is equivalent to the discrete version of Eq. 6.11 where the continuous impulse train is replaced by a discrete impulse train comprised by weights $w(\breve{x})$. The binned kernel estimator is expressed as

$$\hat{p}^b_{\mathbf{x}}(\breve{x}; \breve{A}) = \frac{1}{N}\sum_{j=1}^{\breve{N}} K_u(\breve{x} - \breve{x}_j)w(\breve{x}_j) \tag{6.15}$$

$$= K_u(\breve{x}) \star \frac{1}{N} w(\breve{x}) \tag{6.16}$$

which is effectively the discretized analogue of Eq. 6.8. It should be noted that the binned estimator $\hat{p}^b_{\mathbf{x}}(\breve{x}; \breve{A})$ becomes an arbitrarily good approximation to $\hat{p}_{\mathbf{x}}(\breve{x}; A)$ as $\breve{N}$ increases [Wand, 1994]. We thus write:

$$\hat{p}^b_{\mathbf{x}}(\breve{x}; \breve{A}) \rightarrow \hat{p}_{\mathbf{x}}(\breve{x}; A), \text{ as } \breve{N} \uparrow \tag{6.17}$$

In practice, $\breve{N}$ between 100 and 500 is adequate for the retrieval of an accurate PDF estimate [Hall and Wand, 1996; Wand, 1994]. The binned KDE of Eq. 6.15 has a complexity of $\mathcal{O}(\breve{N}^2)$ which is smaller than the complexities $\mathcal{O}(\breve{N}N)$ and $\mathcal{O}(N^2)$ reported in the start of this section. It is important to emphasize that the new complexity is independent of the sample size. It has a reduced value as $\breve{N} \ll N$, where the latter holds especially in the 3D case. However, it is important to note that as all entries in A contribute to $\breve{A}$ via the interpolation scheme, so effectively the entire information in A is utilized by the binned KDE.

Given that $w(\breve{x})$ and $K_u(\breve{x})$ are now defined at equispaced points, the most significant reduction in complexity is achieved by performing the convolution in the Fourier domain using the FFT [Silverman, 1982; Silverman and Green, 1986].

$$\hat{p}^b_{\mathbf{x}}(\breve{x}; \breve{A}) \quad = \quad K_u(\breve{x}) * \frac{1}{N} w(\breve{x}) \tag{6.18}$$

$$= \quad \mathcal{F}^{-1}\left\{ \mathcal{F}\left\{K_u(\breve{x})\right\} \times \mathcal{F}\left\{\frac{1}{N}w(\breve{x})\right\} \right\} \tag{6.19}$$

We choose to limit the support $supp(K_u)$ of the kernel $K_u(\breve{x})$ to a distance of $6u$ from its mean $\mu = 0$. We thus pre-sample K_u on a regular grid $\breve{g}$ of size N_K and with the same spacing $\Delta\breve{x}$ separating the equidistant $\breve{x}$. Then $\breve{g} = \left\{-\frac{N_K-1}{2}, -\frac{N_K-1}{2}+1, \ldots, 0, \ldots, \frac{N_K-1}{2}-1, \frac{N_K-1}{2}\right\}\Delta\breve{x}$, with $-6u \leq -\frac{N_K-1}{2}$ and $\frac{N_K-1}{2} \leq 6u$. The actual size N_K is not user-defined as it depends on u and $\Delta\breve{x}$. Its size is computed in real-time according to the values of u and $\Delta\breve{x}$.

We can now comment on the complexity of Eq. 6.19 where $K_u(\breve{x})$ is replaced with its finite support analogue $K_u(\breve{g})$. Both discrete signals $K_u(\breve{g})$ and $w(\breve{x})$ are padded with zeros up to a size

$N_{pad} = \breve{N} + N_K - 1$ prior to the FFT to ensure that no spurious frequencies would result by wrapping effects due to circular convolution [Bracewell, 1999]. Hence the complexity of Eq. 6.19 is $\mathcal{O}\big(N_{pad}\log(N_{pad})\big)$. However most papers for example [Shwartz et al., 2005; Silverman, 1982; Silverman and Green, 1986] do not consider convolution complexity with respect to the extended (via padding) domain and simply report the dominant complexity (as $\breve{N} > N_K$) of $\mathcal{O}\big(\breve{N}\log\big(\breve{N}\big)\big)$ for Eq. 6.19 . In this work we adopt the same approach. This is a significant reduction from $\mathcal{O}(\breve{N}^2)$ of the non-FFT binned KDE - given the relatively small $\breve{N}$, and a vast reduction from the complexities reported in the start of this section. Figure 6.3 shows the binned KDE in action. The continuous trials $\alpha \in A$ are interpolated to the regular grid $\breve{x}$ producing the amplitude encoded weights $w(\breve{x})$. Evidently, the entries of $\breve{x}$ which are assigned with a weight $w(\breve{x})$ are the ones immediately neighbouring a continuous α. This results in a sparsely populated impulse train. The weights $w(\breve{x})$ sum to one, so one can perceive them as discrete probabilities. Effectively, $w(\breve{x})$ constitute a normalized histogram where the kernel is the triangle function $\wedge(\cdot)$ and not the standard box function. The convolution - via FFT - of $w(\breve{x})$ with the sampled Gaussian kernel $K_u(\breve{y})$ results in the density $\hat{p}^b_{\mathbf{x}}(\breve{x};\breve{A})$ and populates all $\breve{x}$ which are within a distance of $6u$ from a continuous entry α. The density estimate obtained by the explicit (non-binned, non-FFT) KDE $\hat{p}_{\mathbf{x}}(\breve{x};A)$ are also shown for comparison. In this instance the normalized error between the two sets of estimates is 0.3% which is considered acceptable.

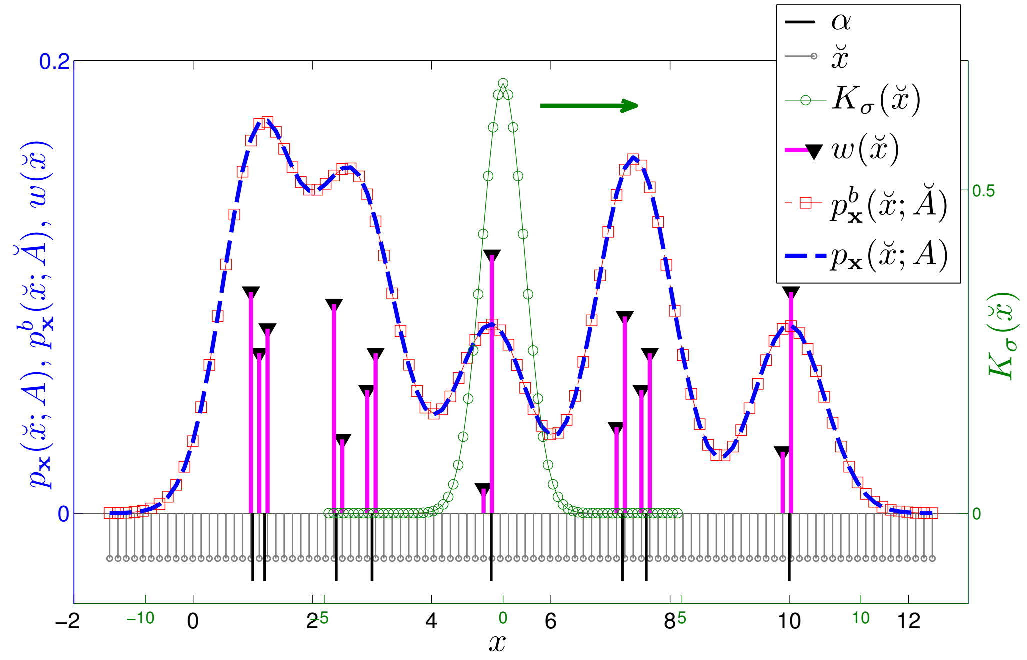


Figure 6.3: Graphical illustration of the binned kernel density estimator. See text for the corresponding discussion.

6.1.1.1 Binning range and spacing

One is required to define the regularly spaced grid $\breve{x}$ where PDF is estimated. Our approach of setting the grid of $\breve{x}$ is graphically illustrated in Fig. 6.4. Prior to setting $\breve{x}$, the binning range $\mathcal{R}(\breve{x})$ has to be defined. For this purpose, we firstly identify the range $\mathcal{R}(A)$ of the initial sample A. According to the discussion in the previous section, the standard KDE returns $\hat{p}_{\mathbf{x}}(x; A) \neq 0$ for all x inside the support of the kernels K_u centered at all regularly spaced trials $\alpha \in A$. Consequently the estimated $\hat{p}_{\mathbf{x}}(x; A)$ is expected to be non-zero for a distance $|x - \alpha_i| \leq 0.5 supp(K_u) = 6u$ from the two extreme continuous trials α_1, α_N. In the case of Fig. 6.4, the kernels K_u centered at the extreme α are highlighted. Hence, the final binning range $\mathcal{R}(\breve{x})$ equals the range $\mathcal{R}(A)$ plus a further extension at both ends by $6.5u$, in order to accommodate the area of non-zero kernel support.

After computing $\mathcal{R}(\breve{x})$, the spacing $\Delta\breve{x}$ between $\breve{x}$ is given by $\Delta\breve{x} = (\breve{x}_{\breve{N}} - \breve{x}_1)/\breve{N}$ where $\breve{N}$ is the *a priori* defined number of regularly spaced $\breve{x}$.

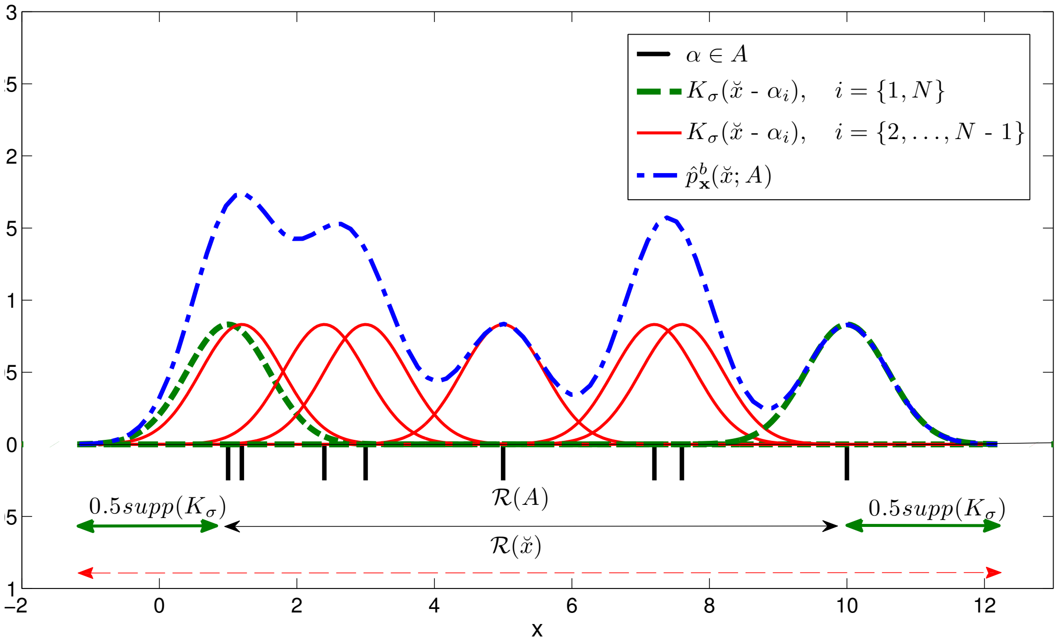


Figure 6.4: Binning range for PDF sampling. Regular grid $\breve{x}$ is not visualized. The Gaussian kernels centered at the extreme sample points are highlighted.

6.1.2 Fast marginal entropy estimation (Shannon formulation)

Utilizing the binned, FFT enabled KDE of Section 6.1.1, it is possible to efficiently obtain an estimate of the entropy of some underlying RV $\mathbf{x}$ from an available continuous sample A. The Shannon's classic integral formulation of entropy (Eq. 6.2) using the efficient KDE is expressed as

$$\hat{h}^b_s(\mathbf{x}; A) = -\sum_{i=1}^{\breve{N}} \hat{p}^b_{\mathbf{x}}(\breve{x}_i; \breve{A}) \log(\hat{p}^b_{\mathbf{x}}(\breve{x}_i; \breve{A}))\, \Delta\breve{x} \tag{6.20}$$

The governing complexity in Eq. 6.20 is the one of the KDE which is $\mathcal{O}\big(\breve{N}\log(\breve{N})\big)$. This is a vast reduction from the initial $\mathcal{O}(\breve{N}N)$ of the initial entropy estimator of Eq. 6.2 which is based on the slow KDE estimator of Eq. 6.8. In addition, it does not depend on the size of the image - only indirectly during the interpolation steps. As an indicator of the achieved computation cost reduction, consider a 3D image with $N = 100^3$ voxels. Assume a grid $\breve{x}$ of size $\breve{N} = 1000$ used for the numerical integration of the PDF. Then the computational cost of the fast Shannon entropy estimator is six orders of magnitude smaller than the non-FFT based estimator.

6.1.3 Fast marginal entropy estimation (empirical formulation)

The empirical entropy formulation utilizing the fast KDE is

$$\hat{h}_e^b(\mathbf{x}; \breve{A}) = -\frac{1}{N}\sum_{k=1}^{N} \log\left(\tilde{p}_{\mathbf{x}}^b(\alpha_k; \breve{A})\right), \tag{6.21}$$

where $\tilde{p}_{\mathbf{x}}^b(\alpha_k; \breve{A})$ is defined over the original continuous locations $\alpha_k \in A$ and is computed via an extra interpolation step from the equispaced $\hat{p}_{\mathbf{x}}^b(\breve{x}; \breve{A})$. Specifically, the interpolation is based on Eq. 6.14 and is expressed as

$$\tilde{p}_{\mathbf{x}}^b(\alpha_k; \breve{A}) = \sum_{j=1}^{N} \wedge(\alpha_k - \breve{x}_{_j})\, \hat{p}_{\mathbf{x}}^b(\breve{x}_{_j}; \breve{A}) \tag{6.22}$$

One can use the array $\mathfrak{b}$ defined in Section 6.1.1 for interpolation purposes and compute the entire $\tilde{p}_{\mathbf{x}}^b(\alpha_k; \breve{A}), \forall k$ as an inner product of vectors, by utilizing efficient linear algebra libraries. The interpolation is then expressed as

$$\tilde{p}_{\mathbf{x}}^b(\alpha_k; \breve{A}) = \left(1 - \mathfrak{b}_{\mathfrak{i}_k}\right) \cdot \hat{p}_{\mathbf{x}}^b(\breve{x}_{\mathfrak{i}_k}; \breve{A}) + \mathfrak{b}_{\mathfrak{i}_k} \cdot \hat{p}_{\mathbf{x}}^b(\breve{x}_{(\mathfrak{i}_k+1)}; \breve{A}), \forall k = 1, 2, \ldots, N \tag{6.23}$$

Although the extra interpolation step has a complexity of $\mathcal{O}(N)$, the overall complexity of Eq. 6.21 is based on the dominant $\mathcal{O}\big(\breve{N}\log(\breve{N})\big)$ of the computation of $\hat{p}_{\mathbf{x}}^b(\breve{x}; \breve{A})$. Hence, the empirical formulation is directly comparable in computational cost terms with Shannon's formulation of Eq. 6.20.

6.1.4 Comparison of the entropic estimators

We seek to evaluate the Shannon's entropy formulation $\hat{h}_s^b(\mathbf{x}; \breve{A})$ of Eq. 6.20 and the empirical entropy $\hat{h}_e^b(\mathbf{x}; \breve{A})$ of Eq. 6.21, both utilizing the binned FFT accelerated KDE.

Entropic estimators compared to a 'gold standard' Evaluating the accuracy of the two entropic estimators can be accomplished by comparing their estimates against some 'gold standard' entropic value of a known density. As a gold standard, we employ the Normal density $\mathcal{N}(0, \sigma)$ which for given standard deviation σ, its entropy can be analytically derived and expressed as [Cover and Thomas, 1991; Viola, 1995]

$$h(\mathbf{x} \sim \mathcal{N}(\mu, \sigma)) = \frac{1}{2}\log(2\pi e\sigma^2) \tag{6.24}$$

Let $\mathbf{x} \sim \mathcal{N}(0, \sigma)$. Before proceeding to entropy estimation with both methods, it is important to choose the width u of the kernel used in the KDE employed by the entropic estimators. The optimum u^{opt} would minimize a distance measure between the PDF estimate $\hat{p}_{\mathbf{x}}^b(\breve{x}; \breve{A})$ obtained by the KDE and the true PDF $p^{\star}_{\mathbf{x}}(\breve{x}) = \mathcal{N}(0,\sigma)$. A commonly used distance measure of similarity between PDFs is the mean integrated square error (MISE) [Silverman and Green, 1986]. By replacing integrals with their numerical approximations MISE is given by

$$\begin{aligned} MISE(u,\breve{A}) &= \sum_{\breve{x}} E\left[\hat{p}^b_{\mathbf{x}}(\breve{x};\breve{A},u) - p^{\star}_{\mathbf{x}}(\breve{x})\right]^2 \Delta\breve{x} && (6.25)\\ &= \sum_{\breve{x}} E\left[\hat{p}^b_{\mathbf{x}}(\breve{x};\breve{A},u) - E\left[\hat{p}^b_{\mathbf{x}}(\breve{x};\breve{A},u)\right] + E\left[\hat{p}^b_{\mathbf{x}}(\breve{x};\breve{A},u)\right] - p^{\star}_{\mathbf{x}}(\breve{x})\right]^2 \Delta\breve{x} && (6.26)\\ &= \sum_{\breve{x}} \left[\hat{p}^b_{\mathbf{x}}(\breve{x};\breve{A},u) - E\left[\hat{p}^b_{\mathbf{x}}(\breve{x};\breve{A},u)\right]\right]^2 + \left(E\left[\hat{p}^b_{\mathbf{x}}(\breve{x};\breve{A},u)\right] - p^{\star}_{\mathbf{x}}(\breve{x})\right)^2 \Delta\breve{x} && (6.27)\\ &= \sum_{\breve{x}} Var\left(\hat{p}^b_{\mathbf{x}}(\breve{x};\breve{A},u)\right)\Delta\breve{x} + \sum_{\breve{x}} bias^2\left(\hat{p}^b_{\mathbf{x}}(\breve{x};\breve{A},u)\right)\Delta\breve{x} && (6.28) \end{aligned}$$

The expectation terms in MISE formula denote that the error is averaged over multiple realizations of A, in order to obtain a more robust statistical estimate.

For the case of a target PDF $p^{\star}_{\mathbf{x}}(\breve{x}) = \mathcal{N}(0,\sigma_i)$, the optimum kernel standard deviation u^{opt} for which $\hat{p}^b_{\mathbf{x}}(\breve{x};\breve{A},u^{\text{opt}})$ minimizes MISE, can be analytically derived and is given by [Parzen, 1962; Silverman and Green, 1986]

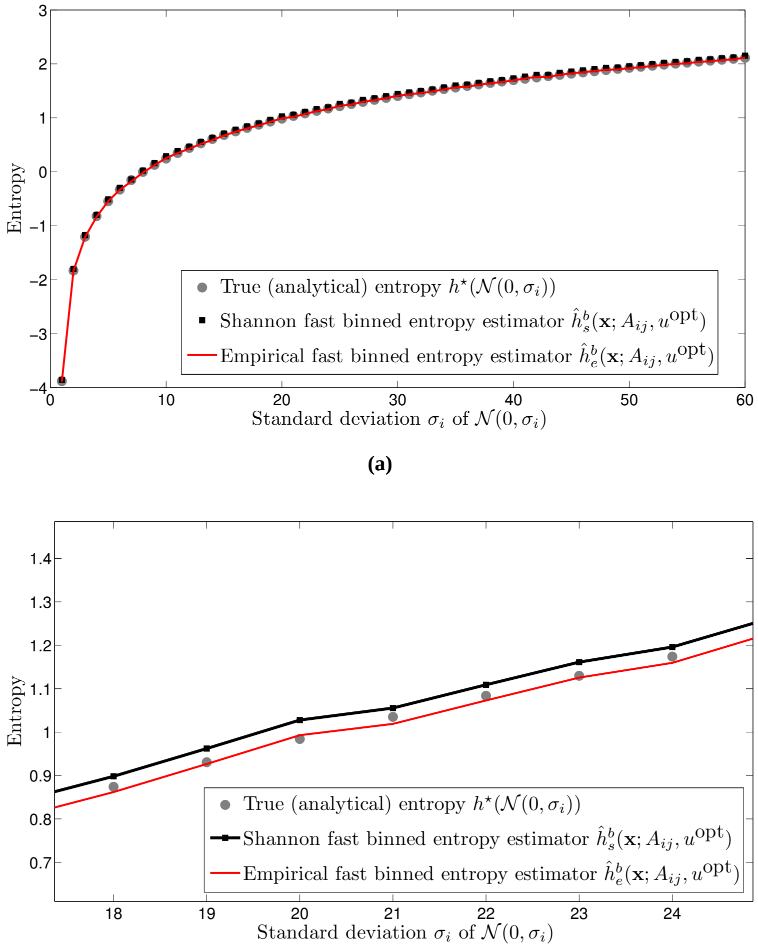
$$u_i^{\text{opt}} = 1.06\sigma_i N^{-1/5} \tag{6.29}$$

where N the size of the sample used for the PDF estimation.

Now that we have a formula for the optimum u^{opt}, we proceed with the evaluation of the entropic estimators. We define 60 RVs $\mathbf{x}_i \sim \mathcal{N}(0,\sigma_i)$ with $\sigma_i = 0.005 \rightarrow 2$, for $i = 1,2,\ldots,60$. For each $\mathbf{x}_i$ we obtain 50 samples A_{ij} indexed by $j = 1,2,\ldots,50$, each consisting of $N = 1000$ continuous trials. For sample size $N = 1000$, the optimum kernel standard deviation is $u^{\text{opt}} = 0.2663$ (Eq. 6.29).

Figure 6.5a shows the differential entropy estimates obtained for both Shannon & empirical binned entropy estimators, each utilizing $\breve{N} = 400$ regularly spaced locations $\breve{x}$ and the optimum u^{opt}. The entropic estimates for each σ_i are averaged over the 50 samples A_{ij} . In addition, the true entropy of $\mathcal{N}(0,\sigma_i)$ is also plotted for the variable σ_i with $h^{\star}(\mathcal{N}(0,\sigma_i)) = 0.5\log(2e\pi\sigma_i)$.

The errors of both estimators compared to the true entropy are computed via a normalized MISE. For the current case of $N = 1000$ sized samples, the errors are presented in the third row (columns 1 & 2) of Table 6.1. The empirical estimator returns an error of less than 1%, whereas the Shannon implementation returns an increased error of 2%. Figure 6.5b depicts a magnification of Figure 6.5a to enhance details. Interestingly, the Shannon estimates appear to overestimate the true entropy values and deviate from the empirical ones by an almost constant offset or $\hat{h}^b_s(\mathbf{x}_i;\breve{A}_{ij},u_i^{\text{opt}}) = \hat{h}^b_e(\mathbf{x}_i;\breve{A}_{ij},u_i^{\text{opt}}) + \epsilon_i$, with $\epsilon_i \approx c, \forall i$ and $c \in \mathbb{R}^+$. We test the total normalized mean squared error between $\epsilon_i, \forall i$ and the mean value $\bar{\epsilon}$, which returned an error of 0.34% (last column of Table 6.1). The relatively small value is an indicator that the entropic estimates obtained by the Shannon estimator deviate from the ones obtained by the empirical estimator by the almost constant value $\bar{\epsilon}$ for all $\mathbf{x}_i$. We repeat the same test for samples



(a)

(b)

Figure 6.5: Comparison between Shannon & empirical entropy estimators for optimum kernel width u^{opt}. 6.5a Entropic estimates are obtained from samples drawn from $\mathbf{x}_i \sim \mathcal{N}(0;\sigma_i)$ for variable $\sigma_i \in [0.005, 2]$. The analytically derived entropy $h^\star(\mathcal{N}(0,\sigma_i))$ is also plotted. The monotonically ascending graph is expected as $\mathcal{N}(0,\sigma_i)$ becomes more clustered for smaller σ_i, hence it is characterized by smaller entropy. 6.5b Detail of 6.5a

comprised of either lesser or more trials and we present the corresponding errors in the other entries of Table 6.1. We observe a progressive improvement in the accuracy of both estimators as samples become larger. In addition, for larger samples the Shannon estimator converges to the empirical estimator and the offset between them approaches a constant value.

From this study we conclude that both estimators can return acceptable entropy estimates given that the utilized samples are large enough. Empirical entropy returned more accurate results for all considered sample sizes.

Entropy estimators and kernel width The next test assesses the effect of the kernel width u on the accuracy of the obtained entropic estimates. Once again we use $\breve{N} = 400$ regularly spaced locations used by the binned estimators. A single target RV $\mathbf{x} \sim \mathcal{N}(0,1)$ is used, from which we retrieve 2000 samples A_i, $(i = 1 \rightarrow 2000)$, each consisting of $N = 1000$ trials. Estimates $\hat{h}_s^b(\mathbf{x}; \breve{A}_i, u_j)$, $\hat{h}_e^b(\mathbf{x}; \breve{A}_i, u_j)$ - where

Table 6.1: Normalized error of the Shannon & empirical entropy estimators against "gold standard" entropy values

Entropy Estimator	Empirical	Shannon	$\bar{\epsilon}$ (MSE%)
Normalized error ($N = 100000$)	0.008%	0.036%	0.058 (0.04%)
Normalized error ($N = 10000$)	0.09%	0.86%	0.0143 (0.14%)
Normalized error ($N = 1000$)	0.4%	2%	0.0358 (0.34%)
Normalized error ($N = 100$)	1.63%	4.2%	0.088 (0.9%)

$\breve{A}_i$ is obtained by re sampling A_i in $\breve{x}$ - are obtained for each A_i. For the kernel standard deviations we use 60 linearly spaced values $u_j = 0.005 \rightarrow 0.45$. We recall that the optimum u^{opt} for $\mathcal{N}(0,1)$ is 0.2663. Figure 6.6 shows the mean entropy estimates for each u_j, averaged over all estimates obtained by using the 2000 samples A_i. The empirical estimator returns its best estimate for a value of $u_e = 0.3776 > u^{\text{opt}}$ whereas the Shannon estimator for $u_s = 0.1313 < u^{\text{opt}}$. The empirical estimator demonstrates higher invariance to changes in u compared to the Shannon estimator and obtains more accurate estimates for a higher range of u. For $u = u^{\text{opt}}$, the true entropy is $h^{\star}(\mathbf{x}) = 1.4189$, whereas the two estimators return $\hat{h}^b_s(\mathbf{x}; A_i, u^{\text{opt}}) = 1.447$ and $\hat{h}^b_e(\mathbf{x}; A_i, u^{\text{opt}}) = 1.412$.

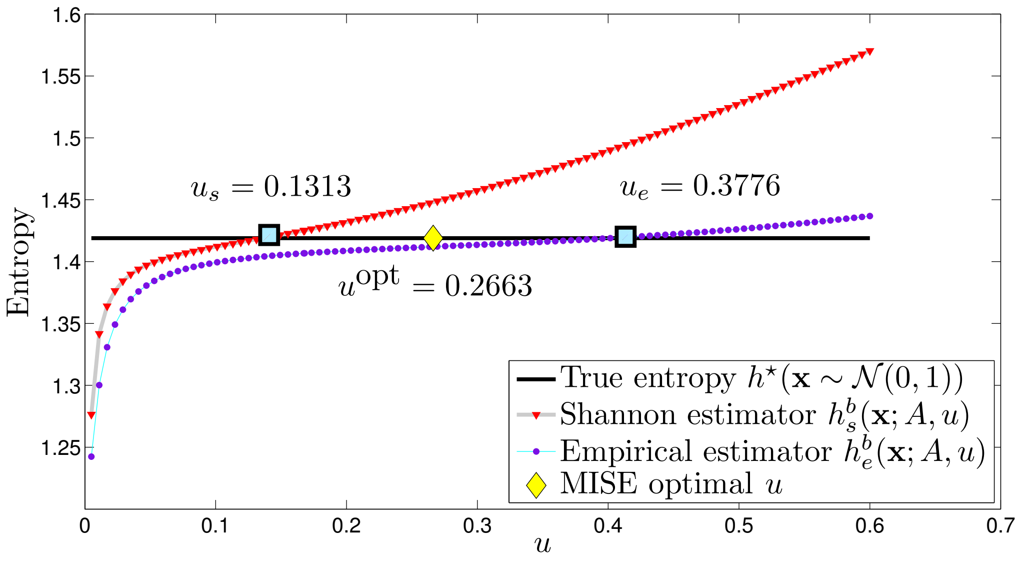


Figure 6.6: Comparison graph among the mean entropy estimates of $\mathbf{x} \sim \mathcal{N}(0,1)$ for variable kernel width u and of the true entropy $h(\mathcal{N}(0,1))$.

Entropy estimators and regular grid size We now assess the effect of the size of the regular grid $\breve{x}$ used by the binned KDE. Consider the underlying RV $\mathbf{x}_i \sim \mathcal{N}(0,1)$. We instantiate 1000 samples A_i, each comprised by 1000 trials. We estimate the entropy of each sample using both estimators, for variable grid sizes $\breve{N}_i \in [100, 10000]$ in 20 logarithmically spaced intervals. The standard deviation u of the kernels used for each estimator, are the ones which returned the best entropic estimates in Fig. 6.6. Figure 6.7

presents the normalized mean integrated square error of each estimator when compared against the true entropy of $\mathbf{x}_i \sim \mathcal{N}(0,1)$. Table 6.2 shows the percentage of the error reduction achieved for different $\breve{N}_i$. We see that the upper limit of the bin number range $(0 \rightarrow 500)$ which was suggested by Wand *et al.* [1996; 1994] for the accurate PDF estimation via a KDE method (see Sec. 6.1.1), retrieves 96% of the total error. We note that by using the optimal kernel widths for each estimator, both estimators return comparable errors.

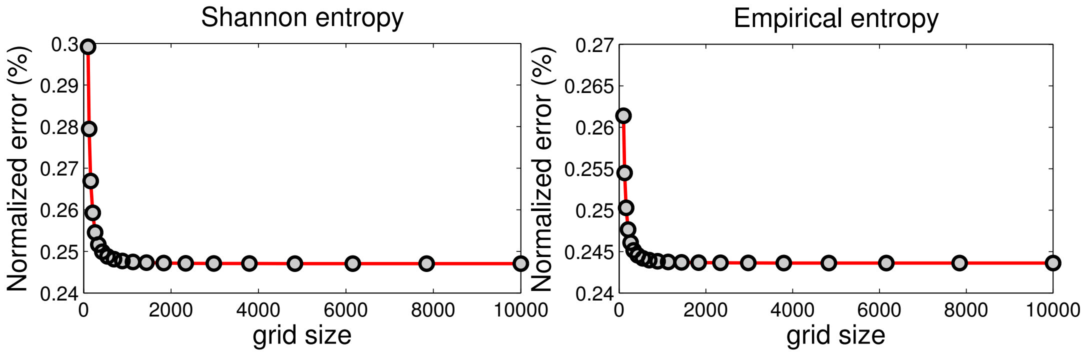


Figure 6.7: Comparison graph among the mean entropy estimates of $\mathbf{x} \sim \mathcal{N}(0,1)$ for variable grid size $\breve{N}$

Table 6.2: Percentage of error recovered vs number of bins

Percentage retrieve	94%	95%	96%	97%	98%	99%
$\breve{N}$ (Empirical)	428	546	546	695	886	1129
$\breve{N}$ (Shannon)	428	546	546	695	886	1129

As a concluding remark of this section, it is evident that both entropic estimators return accurate estimates, given that an optimal kernel width is used. The empirical estimator exhibits less variation to changes in u. However, it should be mentioned that even for non-optimal u, the obtained entropic estimates can still operate as a measure of image uncertainty, or equivalently clustering between different PDFs corresponding to different RVs. For example see the Shannon estimator in Fig. 6.5, which monotonically increases with the increasing σ_i of the considered RVs $\mathbf{x}_i \sim \mathcal{N}(0,\sigma)$. If we are interested to compare the uncertainty of two samples obtained by the RVs, the absolute accuracy of the entropic estimate corresponding to each sample is not as useful as the accuracy of difference between the two estimates.

6.2 Computationally efficient marginal entropy derivative estimation

In this section we describe the used scheme for the efficient estimation of the derivatives of the two entropic estimators.

6.2.1 Shannon's entropy derivatives

By applying the chain rule wherever necessary, the derivative of Eq. 6.2 with respect to the continuous $\alpha_i \in A$ is

$$\frac{\partial \hat{h}_s(\mathbf{x};A)}{\partial \alpha_i} = -\Delta\breve{x}\sum_{j=1}^{\breve{N}}\left[\frac{\partial \hat{p}_{\mathbf{x}}(\breve{x}_j;A)}{\partial \alpha_i}\log\left(\hat{p}_{\mathbf{x}}(\breve{x}_j;A)\right) + \hat{p}_{\mathbf{x}}(\breve{x}_j;A)\frac{\partial \log\left(\hat{p}_{\mathbf{x}}(\breve{x}_j;A)\right)}{\partial \hat{p}_{\mathbf{x}}(\breve{x}_j;A)}\frac{\partial \hat{p}_{\mathbf{x}}(\breve{x}_j;A)}{\partial \alpha_i}\right] \tag{6.30}$$

$$= -\Delta\breve{x}\sum_{j=1}^{\breve{N}}\left(\log\left(\hat{p}_{\mathbf{x}}(\breve{x}_j;A)\right)+1\right)\frac{\partial \hat{p}_{\mathbf{x}}(\breve{x}_j;A)}{\partial \alpha_i} \tag{6.31}$$

$$\overset{\text{Eq. 6.8}}{=} -\frac{\Delta\breve{x}}{N}\sum_{j=1}^{\breve{N}}\left(\log\left(\hat{p}_{\mathbf{x}}(\breve{x}_j;A)\right)+1\right)\frac{\partial K_u(\breve{x}_j-\alpha_i)}{\partial \alpha_i} \tag{6.32}$$

where after considering the Gaussian formulation of $K_u(\breve{x}_j - \alpha_i)$ (see Eq. 6.5)

$$\frac{\partial K_u(\breve{x}_j-\alpha_i)}{\partial \alpha_i} = -K_u(\breve{x}_j-\alpha_i)\left(\frac{\breve{x}_j-\alpha_i}{u^2}\right) \tag{6.33}$$

The complexity of Eq. 6.32 is $\mathcal{O}(\breve{N})$. However, the derivatives are required for all N continuous α_i, leading to a total complexity of $\mathcal{O}(\breve{N}N)$. Once again Eq. 6.32 has a convolution structure. By utilizing the FFT one can achieve reduction in the computational cost. In order to do so, all involved quantities have to be regularly arranged on a common grid. For that reason we firstly compute the derivative not with respect to the continuous α_i but rather with respect to the regularly spaced gray values $\breve{x}$ corresponding to the regular grid. Equation 6.32 then becomes

$$\frac{\partial \hat{h}^b_s(\mathbf{x};\breve{A})}{\partial \breve{x}_i} = -\frac{\Delta\breve{x}}{N}\sum_{j=1}^{\breve{N}}\left(\log\left(\hat{p}^b_{\mathbf{x}}(\breve{x}_j;A)\right)+1\right)\frac{\partial K_u(\breve{x}_j - w(\breve{x}_i))}{\partial \breve{x}_i} \tag{6.34}$$

$$=\left(\log\left(\hat{p}^b_{\mathbf{x}}(\breve{x}_i;A)\right)+1\right)\star\left(-\frac{\Delta\breve{x}}{N}\frac{\partial K_u\left(\breve{x}_i\right)}{\partial \breve{x}_i}\right) \tag{6.35}$$

$$=\mathcal{F}^{-1}\left\{\mathcal{F}\left\{\log\left(\hat{p}^b_{\mathbf{x}}(\breve{x}_i;A)\right)+1\right\}\times\mathcal{F}\left\{-\frac{\Delta\breve{x}}{N}\frac{\partial K_u(\breve{x}_i)}{\partial \breve{x}_i}\right\}\right\} \tag{6.36}$$

The complexity of Eqs. 6.34-6.35 is $\mathcal{O}(\breve{N}^2)$ considering that the derivative is computed for all $i = 1 \rightarrow \breve{N}$ and also given the fact that $\hat{p}^b_{\mathbf{x}}(\breve{x}_i;A)$ is already pre-computed $\forall \breve{x}_i$ during the entropy evaluation. Hence, computational cost reduction has been achieved considering the $\mathcal{O}(\breve{N}N)$ of Eq. 6.32 simply due to the computation of the derivative with respect to $\breve{x}$ rather than α. The FFT based convolution further reduces the complexity to $\mathcal{O}(\breve{N}\log(\breve{N}))$.

Eq. 6.36 computes the derivative at $\breve{x}$. However, the derivative is required at the continuous locations $\alpha \in A$. For that purpose we interpolate the computed derivatives at the continuous locations $\alpha_i \in A$, $\forall i = 1,\ldots,N$ via

$$\frac{\partial \hat{h}^b_s(\mathbf{x};\breve{A})}{\partial \alpha_i} = \left(1-\mathfrak{b}_i\right)\frac{\partial \hat{h}^b_s(\mathbf{x};\breve{A})}{\partial \breve{x}_{\mathfrak{q}_i}} + \mathfrak{b}_i\frac{\partial \hat{h}^b_s(\mathbf{x};\breve{A})}{\partial \breve{x}_{(\mathfrak{q}_i+1)}}, \forall i = 1,2,\ldots,N \tag{6.37}$$

6.2.2 Evaluation of the efficient Shannon entropy estimator derivatives

It is essential to evaluate the accuracy of the FFT accelerated analytic derivatives of Eqs. 6.34 - 6.37. To do so, we compare the latter against the following derivatives:

1. The non-FFT accelerated analytic derivatives of Eq. 6.32 which do not employ an intermediate regularly spaced sample $\breve{A}$ for the purpose of PDF estimation. They rather use the full KDE representation of Eq. 6.8 which ensures that there are no errors propagating from the linear interpolations of the binned KDE. In addition, the FFT accelerated derivatives utilize Gaussian kernels K_u of finite support $supp(K_u) \approx 12u$. The non-FFT derivatives use kernels of infinite support, hence $K_u(\breve{x}_j - \alpha_i)$ of Eq. 6.32 is non-zero for any input. Hence, all trials $\alpha \in A$ contribute directly for the PDF estimate at any continuous location x. This test reveals the effects of the interpolation in the binned entropy estimators, as well as the effects of reducing the continuous N-sized sample to the regularly spaced $\breve{N}$-sized sample

2. The numerical derivatives obtained by finite differences (). Let $A^{i-} = \{\alpha_1, \alpha_2, \ldots, \alpha_i - h, \ldots, \alpha_N\}$ and $A^{i+} = \{\alpha_1, \alpha_2, \ldots, \alpha_i + h, \ldots, \alpha_N\}$ denote the original sample A including a perturbation of a specific trial α_i by some $h \to 0^+$. For α_i, $\forall i = 2 : N-1$, the central derivatives for the fast Shannon estimator of Eq. 6.20 are given by

$$\frac{\partial \hat{h}_s^b(\mathbf{x}; A)}{\partial \alpha_i} = \frac{\hat{h}_s^b(\mathbf{x}; A^{i+}) - \hat{h}_s^b(\mathbf{x}; A^{i-})}{2h}. \tag{6.38}$$

   whereas for $i = 1$ or $i = N$ we use the forward and backward rules $\left(\hat{h}_s^b(\mathbf{x}; A^{i+}) - \hat{h}_s^b(\mathbf{x}; A)\right)/h$ and $\left(\hat{h}_s^b(\mathbf{x}; A) - \hat{h}_s^b(\mathbf{x}; A^{i-})\right)/h$ respectively.

A $N = 1000$ sized sample A is drawn from $\mathbf{x} \sim \mathcal{N}(0,1)$.We compute the Shannon estimator partial derivatives with respect to $\alpha \in A$ using all three mentioned derivatives. The size of the grid remains $\breve{N} = 400$. We set a kernel width for the Shannon estimator of $u_s = 0.1313$, which returns the best entropic estimate with respect to the true entropy of $\mathcal{N}(0,1)$ (see Fig. 6.6). Figure 6.8 shows the graphs of the computed partial derivatives superimposed. The normalized error among the depicted derivatives are given in Table 6.3. We consider the error between the binned, FFT enabled analytic derivatives and the ones obtained by the full KDE implementation to be acceptable. We note that the full KDE implementation - no FFT, unlimited kernel support - is the textbook definition of the KDE. The substantial match between the two is an encouraging outcome. However, the error between the analytic derivatives and the derivatives is significant. However, it shows a dramatic reduction as the size of the sample increases. Although the graphs appear to be matching, the detail in Fig. 6.8 shows that the derivatives are not as smooth as their analytic analogues, possibly due to the binned nature of the estimator. Shwartz et al. [2005] mentioned that an entropy estimator utilizing the binned locations can result in the staircase effect, whereas its analytic derivative exhibits higher invariance. We will see however in the next section that the empirical estimator which is computed directly at the continuous locations, is also affected by the binning process.

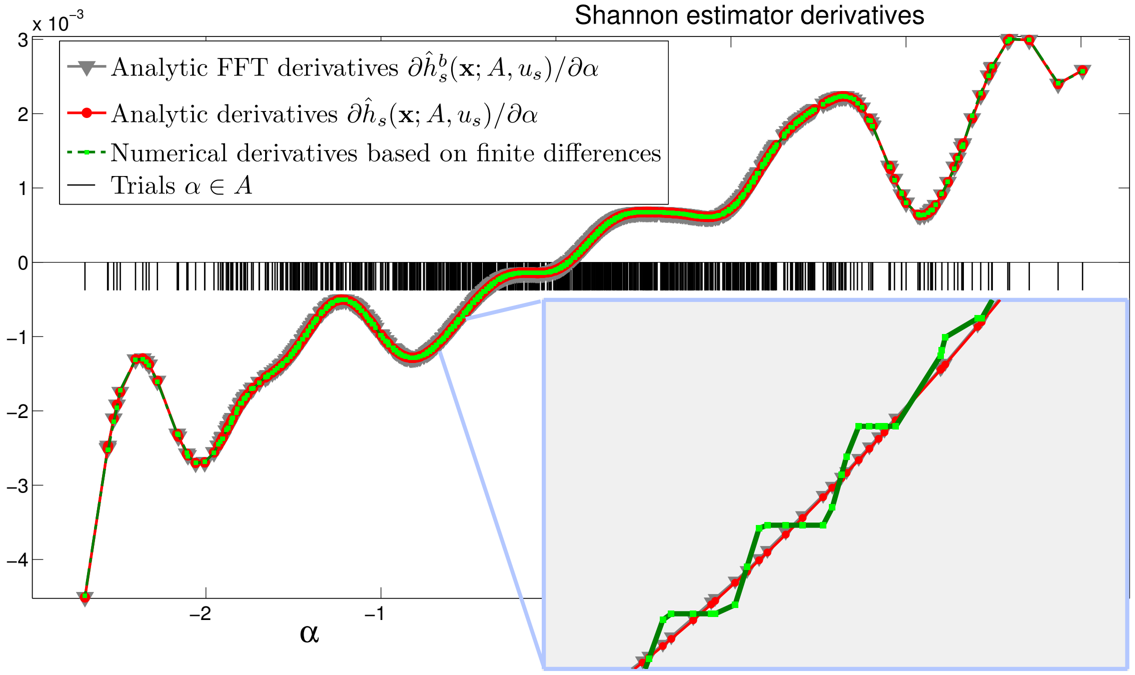


Figure 6.8: Shannon entropy estimator derivatives. Analytic partial entropic derivatives with respect to individual trials (FFT and full KDE, non-FFT implementations), as well as the numerical derivatives based on . Detail of the main graph also provided

Table 6.3: Errors between Shannon entropic estimator analytic derivatives

Compared derivative quantities	Mean squared error $(\%)$
Sample size $N=100$	
Analytic FFT vs analytic non-FFT	0.13%
Analytic FFT vs numerical	36.07%
Sample size $N=1000$	
Analytic FFT vs analytic non-FFT	0.11%
Analytic FFT vs numerical	9.5%
Sample size $N=10000$	
Analytic FFT vs analytic non-FFT	0.04%
Analytic FFT vs numerical	2.45%

6.2.3 Empirical entropy derivatives

The derivative of the empirical entropy formulation $\hat{h}_e^b(\mathbf{x}; A)$ of Eq. 6.21 with respect to the continuous samples $\alpha \in A$ was derived by Shwartz et al. [2005]. The derivation is given here for completeness

$$\frac{\partial \hat{h}_e(\mathbf{x}; A)}{\partial \alpha_r} = -\frac{\partial}{\partial \alpha_r}\left(\frac{1}{N}\sum_{i=1}^{N} \log\left(\hat{p}_{\mathbf{x}}(\alpha_i; A)\right)\right) \tag{6.39}$$

$$= -\frac{1}{N}\sum_{i=1}^{N} \frac{1}{\hat{p}_{\mathbf{x}}(\alpha_i; A)} \frac{\partial}{\partial \alpha_r}\Big(\hat{p}_{\mathbf{x}}(\alpha_i; A)\Big) \tag{6.40}$$

$$= -\frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{\frac{1}{N}\sum_{j=1}^{N} K_u(\alpha_i - \alpha_j)}\right)\left(\frac{\partial}{\partial \alpha_r}\left(\frac{1}{N}\sum_{j=1}^{N} K_u(\alpha_i - \alpha_j)\right)\right) \tag{6.41}$$

$$= -\frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{\frac{1}{N}\sum_{j=1}^{N} K_u(\alpha_i - \alpha_j)}\right)\left(\frac{1}{N}\sum_{j=1}^{N} K_u'(\alpha_i - \alpha_j)\Big(\breve{\delta}(i,r) - \breve{\delta}(j,r)\Big)\right) \tag{6.42}$$

$$\begin{aligned} &= -\frac{1}{N}\frac{\frac{1}{N}\sum_{j=1}^{N} K_u'(\alpha_r - \alpha_j)}{\hat{p}_{\mathbf{x}}(\alpha_r; A)} \\ &\quad + \frac{1}{N}\sum_{i=1}^{N} \frac{\frac{1}{N}K_u'(\alpha_i - \alpha_r)}{\hat{p}_{\mathbf{x}}(\alpha_i; A)} \end{aligned} \tag{6.43}$$

where $K_u'(\cdot) = \partial K_u(\cdot)/\partial \alpha_r$ is defined according to Eq. 6.33. Considering Eq. 6.33 then $K_u'(\alpha_i - \alpha_r) = -K_u'(\alpha_r - \alpha_i)$. Finally $\breve{\delta}(i,r)$ is the Kronecker delta, corresponding to the discrete analogue of the Dirac delta function and is given by

$$\breve{\delta}(x) = \begin{cases} 1, & x = 0 \\ 0, & x \neq y \end{cases}, \tag{6.44}$$

The complexity of estimating $\partial \hat{h}_e(\mathbf{x}; A)/\partial \alpha_r$ via Eq. 6.43 is $\mathcal{O}(N)$. As the derivative needs to be computed for all N trials $\alpha_r \in A$, the overall complexity is $\mathcal{O}(N^2)$.

Both parts of Eq. 6.43 can be computed efficiently by utilizing a regular grid and FFT. We re-iterate its derivation for completeness and also in order to reveal an implementation detail which is important for the purpose of comparison with the Shannon derivatives. Following notation, Equation 6.43 can be rewritten as

$$\frac{\partial \hat{h}_e(\mathbf{x}; A)}{\partial \alpha_r} = \frac{1}{N}\frac{F_1(\alpha_r; A)}{\hat{p}_{\mathbf{x}}(\alpha_r; A)} - F_2(\alpha_r; A) \tag{6.45}$$

where

$$F_1(\alpha_r; A) = \frac{1}{N}\sum_{j=1}^{N} K_u'(\alpha_r - \alpha_j) \quad \text{and} \tag{6.46}$$

$$F_2(\alpha_r; A) = \frac{1}{N}\sum_{i=1}^{N} \frac{K_u'(\alpha_i - \alpha_r)}{\hat{p}_{\mathbf{x}}(\alpha_i; A)}. \tag{6.47}$$

Fast computation of $\frac{1}{N}\frac{F_1(\alpha_r;A)}{\hat{p}_{\mathbf{x}}(\alpha_r;A)}$

Considering the KDE formulation of Eq. 6.8, then $F_1(\alpha_r; A) = \partial \hat{p}_{\mathbf{x}}(\alpha_r; A)/\partial \alpha_r$. However, we have established that $\hat{p}_{\mathbf{x}}(\alpha_r; A)$ can be also computed using the utilization of the binned KDE of Eq. 6.15, which computes $\hat{p}^b_{\mathbf{x}}(\breve{x}; A)$ at $\breve{N}$ regularly spaced locations, followed by an interpolation step which transfers the binned PDF estimate to the continuous locations α. Computing $\hat{p}^b_{\mathbf{x}}(\breve{x}; A)$ for all $\breve{N}$ regular locations $\breve{x}$, is the convolution process expressed by $\hat{p}^b_{\mathbf{x}}(\breve{x}; \breve{A}) = (\frac{1}{\breve{N}}\breve{K}_\sigma * w)(\breve{x})$ (Eq. 6.18).

Computing $F_1(\breve{x}; A)$ simultaneously for all $\breve{x}$ can be accomplished by utilizing a property of the convolution of two functions f, g, which says that the derivative of the convolution equals the convolution of one of the functions with the derivative of the second [Bracewell, 1999]. The latter is expressed as

$$\Big(f(x) \star g(x)\Big)' = f'(x) \star g(x) = f(x) \star g'(x). \tag{6.48}$$

Utilizing this property enables the computation of F_1 as

$$\begin{aligned} F_1(\breve{x}; A) =& \frac{\partial \hat{p}^b_{\mathbf{x}}(\breve{x}; A)}{\partial \breve{x}} && (6.49)\\ =& \frac{1}{N}{K_u}'(\breve{x}) * w(\breve{x}) && (6.50)\\ =& \mathcal{F}^{-1}\left\{\mathcal{F}\Big\{w(\breve{x})\Big\} \times \mathcal{F}\left\{\frac{1}{N}\frac{\partial K_u(\breve{x})}{\partial \breve{x}}\right\}\right\} && (6.51) \end{aligned}$$

Similar to Eq. 6.19, the complexity of the above equation is $\mathcal{O}\big(\breve{N}\log\big(\breve{N}\big)\big)$. Figure 6.9 depicts a sample A; the PDF $\hat{p}^b_{\mathbf{x}}(\breve{x}; A)$; the computed weights $w(\breve{x})$ and the regularly sampled kernel derivative ${K_u}'(\breve{x})$, which when convolved with $w(\breve{x})$ results in $F_1(\breve{x}; A) = \partial \hat{p}^b_{\mathbf{x}}(\breve{x}; A)/\partial \breve{x}$.

Finally, the computation of $F_1(\alpha; A)$ from $F_1(\breve{x}; A)$ is achieved by an interpolation similar to the one of Eq. 6.37.

$$\tilde{F}_1(\alpha_i; A) = \Big(1 - \mathfrak{b}_i\Big)F_2\Big(\breve{x}_{\mathfrak{q}_i}; A\Big) + \mathfrak{b}_i F_1\Big(\breve{x}_{(\mathfrak{q}_i+1)}; A\Big), \forall i = 1, 2, \ldots, N \tag{6.52}$$

To complete the computation of the first term in Eq. 6.45, the $\hat{p}_{\mathbf{x}}(\alpha_r; A)$ in the denominator of $\frac{1}{N}\frac{F_1(\alpha_r;A)}{\hat{p}_{\mathbf{x}}(\alpha_r;A)}$ is replaced by $\tilde{p}^b_{\mathbf{x}}(\alpha_r; A)$ which is already precomputed for the purpose of entropy evaluation in Eq. 6.22.

Fast computation of $F_2(\alpha_r; A)$

The computation of $F_2(x; A)$ can also benefit from fast convolutions via FFT. In a manner analogous to Eq. 6.18 - on which the fast computation of F_1 was based on - $F_2(\breve{x}; A)$ can be expressed as

$$\begin{aligned} F_2(\breve{x}; A) =& \frac{1}{N}K'^{\,\text{mirror}}_u(\breve{x}) * \frac{w(\breve{x})}{\hat{p}^b_{\mathbf{x}}(\breve{x}; A)} && (6.53)\\ =& \mathcal{F}^{-1}\left\{\mathcal{F}\left\{\frac{1}{N}K'^{\,\text{mirror}}_u(\breve{x})\right\} \times \mathcal{F}\left\{\frac{w(\breve{x})}{\hat{p}^b_{\mathbf{x}}(\breve{x}; A)}\right\}\right\}. && (6.54) \end{aligned}$$

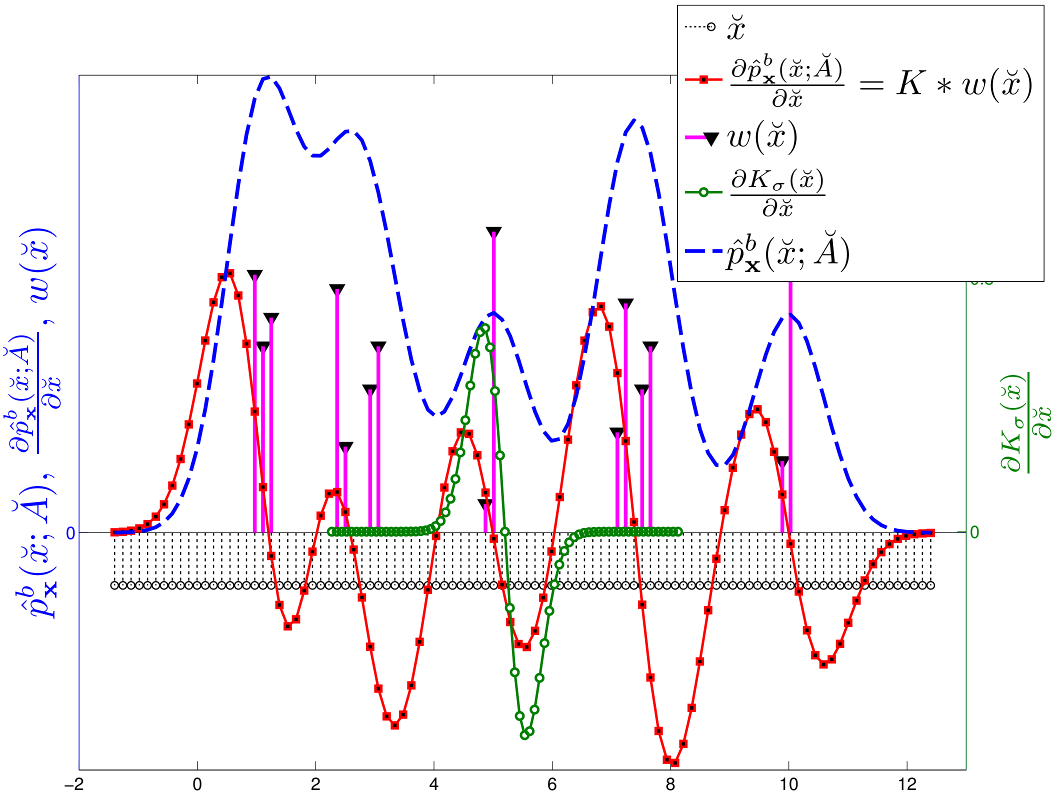


Figure 6.9: Fast PDF derivative estimation. A PDF estimate using the binned estimator is shown by $\hat{p}^b_{\mathbf{x}}(\breve{x};\breve{A})$. Its derivative at $\breve{x}$ is computed by convolving - in the Fourier domain - the regularly spaced sample $\breve{A}$ consisting of $w(\breve{x})$ with the analytic kernel derivative $K_u{}'(\breve{x})$. As a visual qualitative assessment, note that all optima of the $\hat{p}^b_{\mathbf{x}}(\breve{x};\breve{A})$ correspond to zero crossings in $\partial\hat{p}^b_{\mathbf{x}}(\breve{x};A)/\partial\breve{x}$.

where terms $w(\breve{x}), \hat{p}^b_{\mathbf{x}}(\breve{x};A)$ have been precomputed for the entropy evaluation and K_u' is the derivative of the kernel sampled at $\breve{x}$, computed in in Eq. 6.51 and visualized in Fig. 6.9. The sampled $K_u'(\breve{x})$ is mirrored due to the reversal of inputs in Eq. 6.47. An interpolation step similar to Eq. 6.52 transfers $F_2(\breve{x};A)$ to the continuous $\tilde{F}_2(\alpha, A)$.

$$\tilde{F}_2(\alpha_i;A) = \Big(1-\mathfrak{b}_i\Big)F_2\Big(\breve{x}_{\mathfrak{q}_i};A\Big) + \mathfrak{b}_i F_2\Big(\breve{x}_{(\mathfrak{q}_i+1)};A\Big), \forall i = 1,2,\ldots,N \tag{6.55}$$

It is important to note that the implementation proposed by Shwartz et al. [2005] differed in the computation of $F_2(\alpha_i;A)$. In his approach, the factor of Eq. 6.53 was computed directly at the continuous α. The factor was then interpolated back at the regular $\breve{x}$, and finally subjected to the convolution of Eq. 6.53 and the interpolation of Eq. 6.55. We improve on that approach by directly computing the factor at $\breve{x}$. Our approach of empirical entropy estimation removes the need for the first interpolation hence reduces the complexity by $\mathcal{O}(N)$. Shwartz et al. [2005] used explicit *for-loops* for entropy and derivative computation whereas in our implementation all involved quantities are stored in vectors and matrices which enable the computation of entropy and its derivative by utilizing fast linear algebra libraries. Later, when we evaluate the derivative computation approaches we will see that the error between the two approaches is insignificant - see for example Table 6.4. The complexity of Eq. 6.55, computed in a manner similar to Eq. 6.51, is $\mathcal{O}(\breve{N}\log(\breve{N}))$.

6.2.4 Evaluation of the efficient empirical entropy estimator derivatives

Having efficiently computed $\tilde{F}_1(\alpha_i;\breve{A})$ and $\tilde{F}_2(\alpha_i;\breve{A})$, the fast empirical entropy derivatives are obtained by computing the sum of Eq. 6.45. An important difference between the Shannon and empirical derivatives is that the latter requires two 2D FFT in order to compute each of two derivative terms in Eq. 6.45. The combined complexity is $\mathcal{O}(2\breve{N}\log(\breve{N}))$ which is twice the complexity of the Shannon derivatives.

We now test the empirical derivatives using the same approach discussed at the end of Section 6.2.1 and using the *same* $N = 1000$-sized sample A drawn from $\mathbf{x} \sim \mathcal{N}(0,1)$ and used for the Shannon derivative evaluation. Again we use $\breve{N} = 400$ and we use the optimum $u_e = 0.3776$ (see Fig. 6.6). The FFT based derivatives are tested against

i) the ones obtained by numerical ,

ii) the ones obtained by explicitly computing Eq. 6.43 without the employment of binned KDE and FFT as well as

the ones by Shwartz et al. [2005] implementation. The obtained derivatives are depicted in Fig. 6.10. The normalized errors between the analytic derivatives, provided in Table 6.4, are less than 1% for two different sample sizes. However, once again there is significant discrepancy between the analytic derivatives and the . The detail in Fig. 6.10 shows that the stair case effect is also a characteristic of the empirical entropy estimator. We note that the derivatives of the $N = 1000$ sized sample are not depicted. It is encouraging to see that the derivative error decreases as the sample increases.

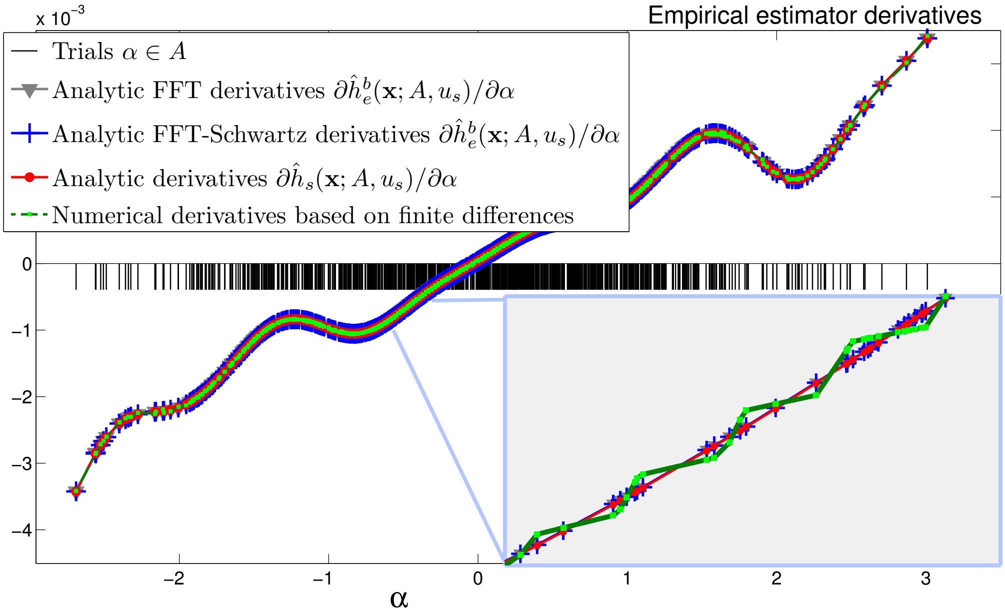


Figure 6.10: Empirical entropy estimator derivatives. Analytic partial entropic derivatives with respect to individual trials (FFT; full KDE, non-FFT; and Shwartz implementation) as well as the numerical derivatives based on . Detail of the main graph also provided

Table 6.4: Errors between empirical entropic estimator analytic derivatives

Compared derivative quantities	Mean squared error $(\%)$
Sample size $N = 100$	
Analytic FFT vs analytic non-FFT	0.05%
Analytic FFT vs numerical	17%
Analytic FFT vs analytic Shwartz FFT	0.008%
Sample size $N = 1000$	
Analytic FFT vs analytic non-FFT	0.02%
Analytic FFT vs numerical	3.13%
Analytic FFT vs analytic Shwartz FFT	0.0077%
Sample size $N = 10000$	
Analytic FFT vs analytic non-FFT	0.012%
Analytic FFT vs numerical	0.7%
Analytic FFT vs analytic Shwartz FFT	0.0078%

6.2.5 Comparison of the derivatives of the marginal entropy estimators

The previous sections evaluated the derivatives of both estimators independently. It is important to see how the derivatives compare to each other. We only consider the FFT enabled analytic derivatives which are of interest in this work. The derivatives depicted in Figs. 6.8-6.10 differ by a very significant 20% (see Fig. 6.12a for their visual superposition). For the estimation of both empirical and Shannon derivatives we have used kernel standard deviations $u_e = 0.3444$ and $u_s = 0.1372$, which produced the best entropic estimates with respect to the true $h^\star(\mathcal{N}(0,1))$. These values differ significantly as well as the resulting PDFs obtained by the KDE (not depicted). In order to check if the discrepancy between the derivatives is due to the choice of u, we keep u_s fixed and we search in a range of values $u_e = c \cdot u_s$ with $c = 1 \rightarrow 2$ for an empirical derivative which produces a better match against the Shannon one. The derivative error with respect to the various c is shown in the left graph of Fig. 6.11. The graph on the right is similar to the one of Fig. 6.6, where the entropy estimates are plotted for variable u. The values of u_s, u_e and the estimated $\hat{u}_e = c \cdot u_s$, for c corresponding to the minimum in the left graph of Fig. 6.11, are highlighted. Interestingly, the $\hat{u}_e$ is very close to the u which minimizes the MISE value discussed in Sec. 6.1.4. Using the newly estimated $\hat{u}_e$ we recompute the derivatives for the same sample with size $N = 1000$ and for a bigger sample with $N = 10000$. The obtained derivatives are depicted in Fig. 6.12. The match between the derivatives vastly improves visually as well the quantitatively. The various errors are summarized in Table 6.5.

As it was already noted, the true derivative is not available to provide the 'gold standard' point of reference in order to assess the absolute accuracy of the two estimators' derivatives. However, the inability to conduct a test against a 'gold standard' is not disabling. Even if this gold standard existed, the estimated derivatives could largely deviate from it for incorrect u. In practice, what we are really interested in is that the *estimated* derivative is an accurate descriptor of the slope of the *estimated* entropy. One would expect that for accurate u, the empirical (expectation based) entropy formulation would approach the true entropy for very large samples and that the Shannon entropy would approach the true entropy for very large samples and dense discretisation. The fact that the error decreases as the sample size increases is an encouraging indicator regarding the comparable nature of the two estimators. In an imaging perspective, the image itself is the maximum sized sample which can be possibly used, hence its estimated entropy is the best possible value which we can obtain. The concept of true entropy does not hold in that case as there is not *true* underlying probability density which describes the gray values. Thus, in the imaging perspective, even the search for the 'best' u does not really exist due to the luck of a 'gold standard' PDF, but it is now a subjective choice given what we seek to accomplish. Using small u in the KDE, captures details in the image but simultaneously does not assume high correlation between different gray values. The opposite happens for large u.

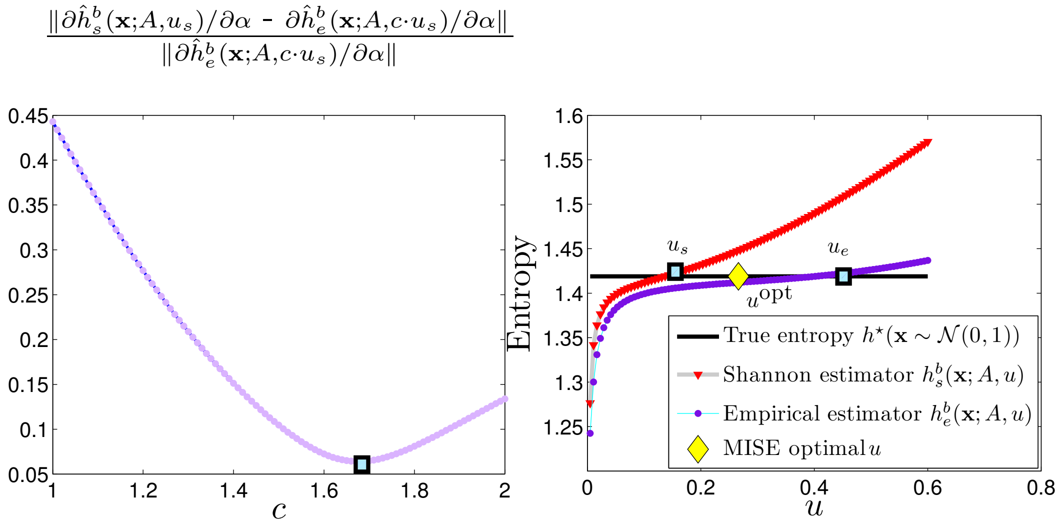


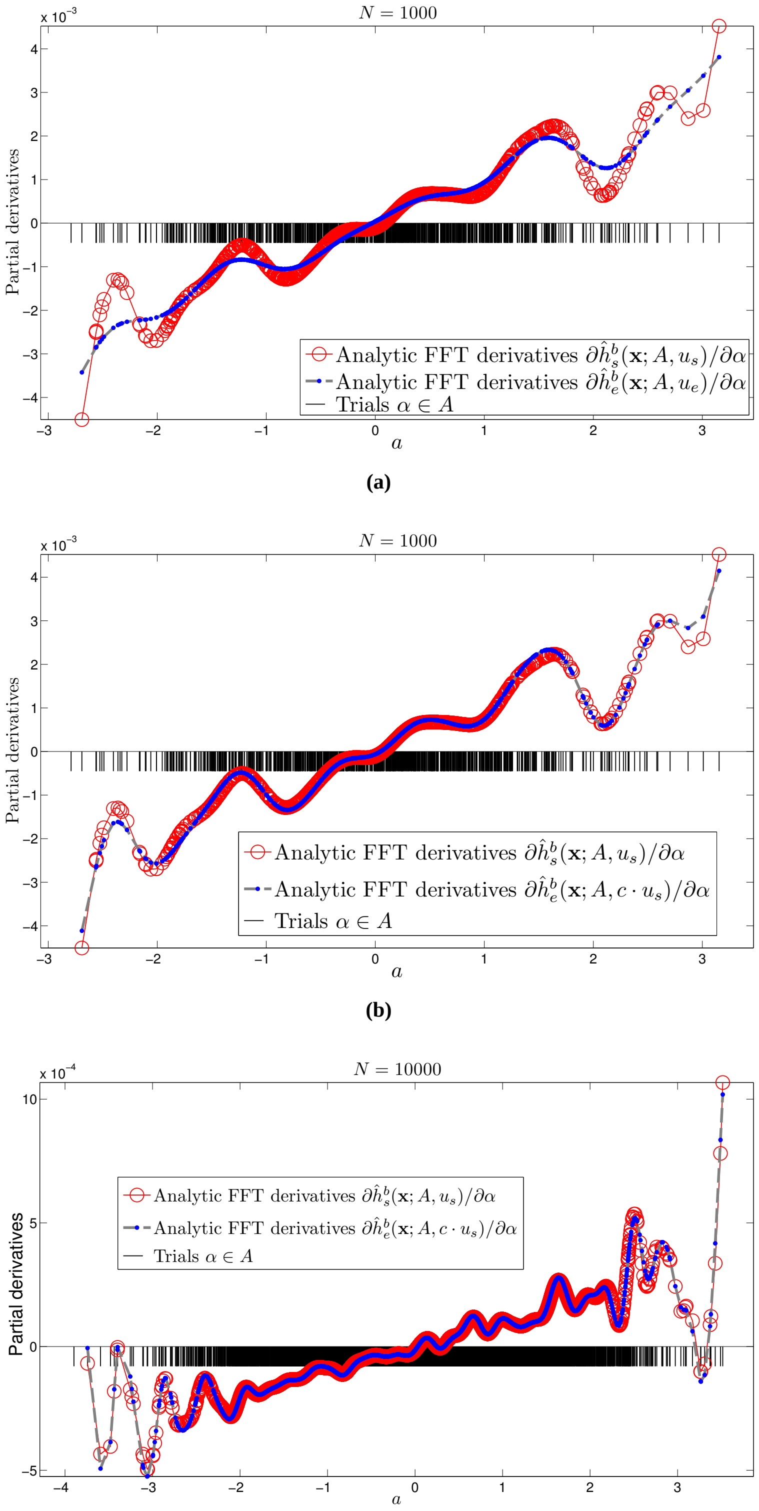
Figure 6.11: Fitting empirical estimator derivatives to the Shannon derivative, as a function of kernel standard deviation u. See text for description. **Left :** Kernel width $u_e = c \cdot u_s$ used for empirical entropy derivative estimation. The minimum of the graph corresponds to the value of c which minimizes the derivative difference. **Right :** Shannon and empirical entropy for various values of u_s and u_e. The values which return the best entropy estimate as well as the $u_e = c \cdot u_s$ which minimizes the derivative difference are highlighted

Table 6.5: Comparison between the Shannon and empirical entropy derivatives estimates obtained from the same sample. Two different sized samples are considered.

Analytic FFT Shannon vs analytic FFT empirical	Normalized error
Ideal kernel u for both estimators	
Sample size $N = 1000$	20.1%
Sample size $N = 10000$	8%
Ideal kernel u for Shannon estimator and fitted u for empirical estimator	
Sample size $N = 1000$	7.26%
Sample size $N = 10000$	2.6%

Table 6.6: Marginal entropy evaluation and derivative estimation: Computational complexity of explicit as well as the FFT based implementations

Non-FFT	Order of Complexity	FFT	Order of complexity
$\hat{h}_s(\mathbf{x}; C)$	$\mathcal{O}(\breve{N}N)$	$\hat{h}^b_s(\mathbf{x}; C)$	$\mathcal{O}\big(\breve{N}\log(\breve{N})\big)$
$\hat{h}_e(\mathbf{x}; C)$	$\mathcal{O}(N^2)$	$\hat{h}^b_e(\mathbf{x}; \breve{A})$	$\mathcal{O}\big(\breve{N}\log(\breve{N})\big)$
$\partial\hat{h}_s(\mathbf{x}; \breve{A})/\partial\alpha_i$	$\mathcal{O}(\breve{N}N)$	$\partial\hat{h}^b_s(\mathbf{x}; \breve{A})/\partial\alpha_i$	$\mathcal{O}\big(\breve{N}\log(\breve{N})\big)$
$\partial\hat{h}_e(\mathbf{x}; C)/\partial\alpha_r$	$\mathcal{O}(N^2)$	$\partial\hat{h}^b_e(\mathbf{x}; \breve{A})/\partial\alpha_i$	$\mathcal{O}\big(\breve{N}\log(\breve{N})\big)$



(a)

(b)

(c)

Figure 6.12: Visual comparison between Shannon & empirical entropy estimator derivatives of two samples of different size. 6.12a $N = 1000$, incorrect widths u_s, u_e; whereas the following subfigures use the optimal widths for derivative matching u_s, $c \cdot u_s$ for two different sample sizes 6.12b $N = 1000$ and 6.12c $N = 10000$.

6.3 Computationally efficient joint entropy evaluation

This section builds on the concepts previously discussed in this chapter and extends them to the case of JE between two RVs. Similar to the definition of $\mathbf{x}$ (see start of Section 6.1), $\mathbf{y}$ is introduced as the underlying RV describing the gray values of a second image. The second image is effectively a sample B of $\mathbf{y}$, expressed as $B = \{\beta_1, \beta_2, \ldots, \beta_N\}$ where $\beta_i = \mathbf{y}(r_i)$ for all pixel locations $r_i \in \Omega_{\mathbf{y}}$. We consider $\Omega_{\mathbf{x}} \equiv \Omega_{\mathbf{y}}$ and also that both images A and B have equal size N.

When considered jointly, the gray values of images A and B forms a joint sample C which consists of N trial pairs $\{\alpha_i, \beta_i\}$, $\forall i = 1, 2, \ldots, N$. It should be emphasized that each of the joint trial pairs is comprised by *spatially corresponding* gray values from both images - that is the gray values sampled at the same pixel location. This constitutes the source of the pixel-wise spatial correspondence between A and B, which is captured by joint entropy. However, it should be noted that within a single image, gray values are considered to be i.i.d.. Hence, no spatial dependence among the pixels of a single image is considered via this formulation.

The approximation of Shannon's joint entropy integral via means of numerical integration is given by

$$\hat{h}_s(\mathbf{x}, \mathbf{y}; C) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \hat{p}_{\mathbf{x},\mathbf{y}}(x, y; C) \log \Big(\hat{p}_{\mathbf{x},\mathbf{y}}(x, y; C) \Big) \mathrm{d}x \, \mathrm{d}y \tag{6.56}$$

$$\approx - \sum_{i,j=1}^{\breve{N}} \hat{p}_{\mathbf{x},\mathbf{y}}(\breve{x}_i; \breve{y}_j; C) \log \Big(\hat{p}_{\mathbf{x},\mathbf{y}}(\breve{x}_i, \breve{y}_j; C) \Big) \, \Delta\breve{x}\Delta\breve{y}, \tag{6.57}$$

where $\breve{y} = \{\breve{y}_1, \breve{y}_2, \ldots, \breve{y}_{\breve{N}}\}$, $\breve{y}_i \in \mathbb{R}$ is a discretisation of $\mathbf{y}$ in regularly spaced intervals with spacing $\Delta\breve{y} = \breve{y}_{i+1} - \breve{y}_{i+1}$. Throughout this work, we set an equal number of grid locations $\breve{N}$ for both $\breve{x}$, $\breve{y}$. The $\breve{N}^2$ nodal positions of the grid can also be indexed via the shorter notation $\tilde{r}_{i,j} = \{\breve{x}_i, \breve{y}_j\}$.

Similar to its marginal analogue of Eq. 6.3, the joint empirical entropy is expressed in terms of the continuous joint trials $\{\alpha_i, \beta_i\}$ and is expressed as

$$\hat{h}_e(\mathbf{x}, \mathbf{y}; C) = -\frac{1}{N} \sum_{i=1}^{N} \log \Big(\hat{p}_{\mathbf{x},\mathbf{y}}(\alpha_i, \beta_i; C) \Big). \tag{6.58}$$

Consider now the joint KDE $\hat{p}_{\mathbf{x},\mathbf{y}}(x, y; C)$, which estimates the joint density for $\{x, y\}$, $x, y \in \mathbb{R}^+$, by utilizing the joint sample C. In an analogy with the marginal KDE, the joint KDE with its 2D convolution (denoted by $\star\star$) structure made apparent is expressed by

$$\hat{p}_{\mathbf{x},\mathbf{y}}(x, y; C) = \frac{1}{N} \sum_{i=1}^{N} K_{\Sigma}(x - \alpha_i, y - \beta_i) \tag{6.59}$$

$$= \frac{1}{N} \sum_{i=1}^{N} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} K_{\Sigma}(x - s, y - t) \delta(s - \alpha_i, t - \beta_i) \mathrm{d}s \, \mathrm{d}t \tag{6.60}$$

$$= K_{\Sigma}(x, y) \star\star \frac{1}{N} \sum_{i=1}^{N} \delta(x - \alpha_i, y - \beta_i) \tag{6.61}$$

where $K_\Sigma(x,y)$ is a bi-variate Gaussian kernel re-iterated here for convenience as

$$K_\Sigma(x,y) = \frac{1}{2\pi \left|\Sigma\right|^{1/2}} \exp\left(-\frac{1}{2}\left[x,y\right]^{\mathrm{T}} \Sigma^{-1} \left[x,y\right]\right), \tag{6.62}$$

with $\Sigma = \begin{bmatrix} u_{\mathbf{x}}^2 & 0 \\ 0 & u_{\mathbf{y}}^2 \end{bmatrix}$ being the covariance matrix and $\delta(x,y)$ is the 2D Dirac delta function for which $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \delta(x,y)\mathrm{d}x\ \mathrm{d}y = 1$, $\delta(x,y) = 0$ for $x^2+y^2 \neq 0$ and $\delta(x,y) = \infty$ otherwise. Regarding Eq. 6.59, it should be emphasized that the estimation of $\hat{p}_{\mathbf{x},\mathbf{y}}(x,y;C)$ utilizes a joint sample comprised by the joint trials $\{\alpha_i, \beta_i\}$, $\forall i = 1,2,\ldots,N$.

At this point, we should emphasize a very crucial detail. The dimensionality of the images constituting C does not affect in any way the computation, computational complexity or implementation of joint entropy which assumes i.i.d. trials. Due to the spatial independence, both images are considered by the functional as vectors, where their entries at corresponding index locations constitute the joint trials. Both joint entropy estimators solely depend on the JPDF whose dimensionality is independent of the dimensionality of the images (1D/1D, 2D/2D or 3D/3D). The JPDF is always 2D; that is one dimension for describing the marginal PDF of the gray values of each of the two images. Hence, the computational complexity estimates which will be derived apply equally to the 2D/2D or 3D/3D case and depend solely on the number of pixels in the image.

6.3.1 Efficient joint PDF estimation

The complexity of the joint KDE of Eq. 6.59 for retrieving an estimate for all $\breve{N}^2$ continuous $\{x,y\}$ of the joint Shannon entropy (Eq. 6.57) is $\mathcal{O}(\breve{N}^2 N)$ whereas for the empirical formulation (Eq. 6.58) is $\mathcal{O}(N^2)$. Both costs however are significantly high and they can be reduced by following an approach similar to Section 6.1.1, which reduces the complexity of the KDE. The continuous joint trials $\{\alpha_i, \beta_i\}$, $\forall i = 1,2,\ldots,N$ are interpolated to a regular 2D grid with grid locations $\left\{\breve{x}_{_i}, \breve{y}_{_j}\right\}$ with $i,j = 1,2,\ldots,\breve{N}$ and with spacings $\Delta\breve{x}$, $\Delta\breve{y}$. In a manner similar to the marginal case, the weights $w(\breve{x},\breve{y})$ assigned to the grid locations, represent the density of the continuous joint trials $\{\alpha,\beta\}$ in the vicinity of $\{\breve{x},\breve{y}\}$ and are computed via a bi-linear interpolation

$$w(\breve{x}_{_i}, \breve{y}_{_j}) = \frac{1}{N}\sum_{k=1}^{N} \wedge(\breve{x}_{_i} - \alpha_k, \breve{y}_{_j} - \beta_k) \tag{6.63}$$

where

$$\wedge(u,v) = \begin{cases} \Big(1 - \left(\left|u\right|/\Delta\breve{x}\right)\Big)\Big((1 - \left(\left|v\right|/\Delta\breve{y}\right)\Big), & \text{if } |u| < \Delta\breve{x} \text{ and } |v| < \Delta\breve{y}, \\ 0, & \text{otherwise.} \end{cases} \tag{6.64}$$

We term the collection of $\left\{\breve{x}_{_i}, \breve{y}_{_j}\right\}$ as $\breve{C}$ which corresponds to the regularly arranged version C. Figure 6.2 graphically shows the quantities involved in the bi-linear interpolation process. In a manner similar to its 1D analogue (see Figure 6.2), after the first application of the interpolation process, a number of the involved quantities are stored in order to be re-used in subsequent interpolations and save computational time. For all joint trials $\{\alpha_k, \beta_k\}, \forall k = 1, 2, \ldots, N$, the stored quantities are **a)** the normalized distance of both α_k and β_k from their nearest regular location $\{\breve{x}_i, \breve{y}_j\}$, with $\breve{x}_i < \alpha_k$ and $\breve{y}_j < \beta_k$. The normalized distances are given by $\mathfrak{b}^x_k = (\alpha_k - \breve{x}_i)/\Delta\breve{x}$ and $\mathfrak{b}^y_k = (\beta_k - \breve{y}_j)/\Delta\breve{y}$. **b)** the indices i, j indicating the nearest regular location $\{\breve{x}_i, \breve{y}_j\}$ which was mentioned above. The indices are stored in the arrays $\mathfrak{i}_k = i$ and $\mathfrak{j}_k = j$. The three remaining grid locations enclosing the joint trial can the easily be recovered via $\{\mathfrak{i}_k + 1, \mathfrak{j}_k\}, \{\mathfrak{i}_k, \mathfrak{j}_k + 1\}, \{\mathfrak{i}_k + 1, \mathfrak{j}_k + 1\}$

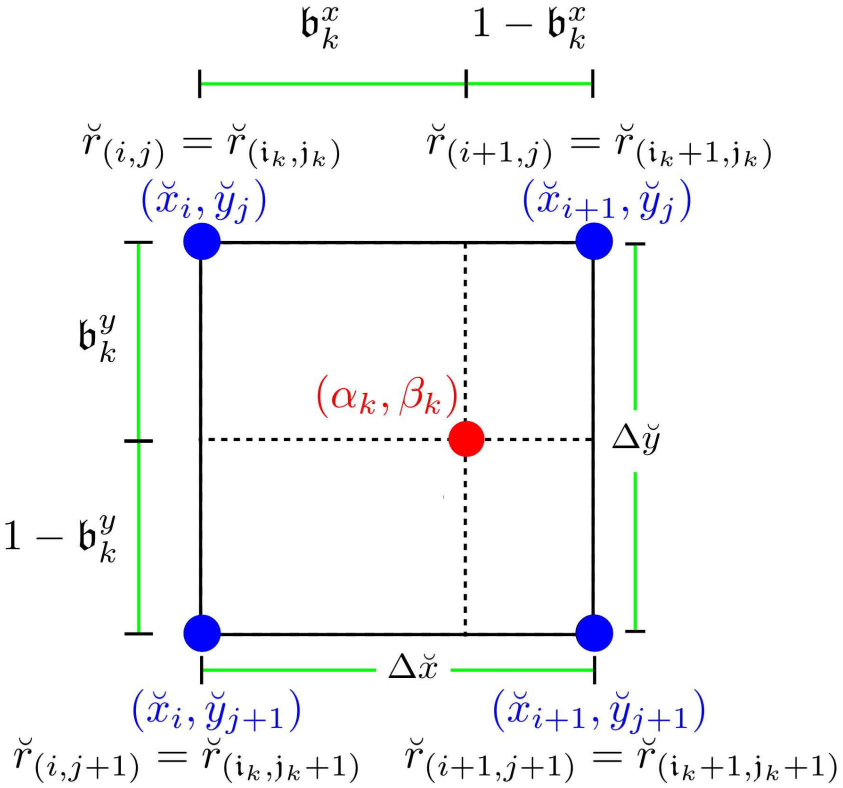

$\mathfrak{b}^x_k$
$1 - \mathfrak{b}^x_k$
$\breve{r}_{(i,j)} = \breve{r}_{(\mathfrak{i}_k,\mathfrak{j}_k)}$
$\breve{r}_{(i+1,j)} = \breve{r}_{(\mathfrak{i}_k+1,\mathfrak{j}_k)}$
$(\breve{x}_i, \breve{y}_j)$
$(\breve{x}_{i+1}, \breve{y}_j)$
$\mathfrak{b}^y_k$
(α_k, β_k)
$\Delta\breve{y}$
$1 - \mathfrak{b}^y_k$
$\Delta\breve{x}$
$(\breve{x}_i, \breve{y}_{j+1})$
$(\breve{x}_{i+1}, \breve{y}_{j+1})$
$\breve{r}_{(i,j+1)} = \breve{r}_{(\mathfrak{i}_k,\mathfrak{j}_k+1)}$
$\breve{r}_{(i+1,j+1)} = \breve{r}_{(\mathfrak{i}_k+1,\mathfrak{j}_k+1)}$

Figure 6.13: Bi-linear interpolation in JPDF estimation. Graphical representation of involved quantities. These quantities are normalized distances $\mathfrak{b}^x$ and $\mathfrak{b}^y$ between $\{\alpha_k, \beta_k\}$ and $\{\breve{x}_i, \breve{y}_j\}$, which are stored in the homonymous arrays ($\mathfrak{b}^x$ and $\mathfrak{b}^y$) and the corresponding indices i, j stored in $\mathfrak{i}_k$ and $\mathfrak{j}_k$ respectively.

In an analogy to its 1D counterpart of Eq. 6.16, the joint binned kernel estimator which retrieves a density estimate at the regular locations $\{\breve{x}, \breve{y}\}$ and can utilize the 2D FFT[1] is defined by

[1] For the purpose of this work we use the open-source Matlab based implementation of 1D/2D FFT based convolution, implemented by Luigi Rosa. The source can be found in http://www.mathworks.com/matlabcentral/fileexchange/4334

$$\hat{p}^b_{\mathbf{x},\mathbf{y}}(\breve{x},\breve{y};\breve{C}) = \frac{1}{N}\sum_{l,m=1}^{\breve{N}} K_{\boldsymbol{\Sigma}}(\breve{x}-\breve{x}_l,\breve{y}-\breve{y}_m)w(\breve{x}_l,\breve{y}_m) \tag{6.65}$$

$$= K_{\Sigma}(\breve{x},\breve{y}) \star\star \frac{1}{N}w(\breve{x},\breve{y}) \tag{6.66}$$

$$= \mathcal{F}^{-1}\left\{\mathcal{F}\left\{K_{\Sigma}(\breve{x},\breve{y})\right\} \times \mathcal{F}\left\{\frac{1}{N}w(\breve{x},\breve{y})\right\}\right\} \tag{6.67}$$

The FFT based convolution of Eq. 6.67 populates the discrete $\breve{N}^2$ grid with a cost of $\mathcal{O}\left(\breve{N}^2\log(\breve{N})\right)$.

6.3.2 Fast joint entropy estimation (Shannon formulation)

The efficient joint entropy binned estimator (integral formulation) is obtained by simply replacing the slow $\hat{p}_{\mathbf{x},\mathbf{y}}(\breve{x},\breve{y};C)$ in Eq. 6.57, with the binned version introduced in the previous section. We thus arrive to

$$\hat{h}^b_s(\mathbf{x},\mathbf{y};\breve{C}) = -\sum_{i,j=1}^{\breve{N}} \hat{p}^b_{\mathbf{x},\mathbf{y}}(\breve{x}_i;\breve{y}_j;\breve{C})\log\left(\hat{p}^b_{\mathbf{x},\mathbf{y}}(\breve{x}_i,\breve{y}_j;C)\right)\Delta\breve{x}\Delta\breve{y}. \tag{6.68}$$

The order of complexity is $\mathcal{O}\big(\breve{N}^2\log(\breve{N})\big)$, largely dominated by the joint KDE. This is a huge reduction from the initial $\mathcal{O}(\breve{N}^2N)$ especially when considering the 3D case with large N.

Regarding the empirical JE (expectation formulation) in Eq. 6.58, we recall that it utilizes the joint sample C which is expressed in terms of the continuous $\hat{p}_{\mathbf{x},\mathbf{y}}(\alpha_i,\beta_i;C)$. FFT enabled $\hat{p}^b_{\mathbf{x},\mathbf{y}}(\breve{x},\breve{y};\breve{C})$ is expressed in terms of the regularly spaced $\{\breve{x},\breve{y}\}$. The utilization of the latter in the empirical joint entropy estimator requires interpolation which enable the transitions $\hat{p}^b_{\mathbf{x},\mathbf{y}}(\breve{x},\breve{y};C) \rightarrow \tilde{p}^b_{\mathbf{x}}(\alpha_k,\beta_k;\breve{C})$. The interpolation is effectively the inverse of the scheme expressed in Eq. 6.63. We utilize the pre-computed quantities discussed in the text relevant to Figure 6.2 to efficiently compute the interpolation as

$$\begin{aligned}\tilde{p}^b_{\mathbf{x},\mathbf{y}}(\alpha_k,\beta_k;C) =& \left(1-\mathfrak{b}^x_k\right)\cdot\left(1-\mathfrak{b}^y_k\right)\cdot\hat{p}^b_{\mathbf{x},\mathbf{y}}(\breve{x}_{\mathfrak{i}_k},\breve{y}_{\mathfrak{j}_k};\breve{C}) + \\ & \left(\mathfrak{b}^x_k\right)\cdot\left(1-\mathfrak{b}^y_k\right)\cdot\hat{p}^b_{\mathbf{x},\mathbf{y}}(\breve{x}_{\mathfrak{i}_{k+1}},\breve{y}_{\mathfrak{j}_k};\breve{C}) + \\ & \left(1-\mathfrak{b}^x_k\right)\cdot\left(\mathfrak{b}^y_k\right)\cdot\hat{p}^b_{\mathbf{x},\mathbf{y}}(\breve{x}_{\mathfrak{i}_k},\breve{y}_{\mathfrak{j}_{k+1}};\breve{C}) + \\ & \left(\mathfrak{b}^x_k\right)\cdot\left(\mathfrak{b}^y_k\right)\cdot\hat{p}^b_{\mathbf{x},\mathbf{y}}(\breve{x}_{\mathfrak{i}_{k+1}},\breve{y}_{\mathfrak{j}_{k+1}};\breve{C}), \quad \forall k=1,2,\ldots,N.\end{aligned} \tag{6.69}$$

The final FFT enabled empirical joint entropy estimator is given by

$$\hat{h}^b_e(\mathbf{x},\mathbf{y};\breve{C}) = -\frac{1}{N}\sum_{k=1}^{N}\log\tilde{p}^b_{\mathbf{x},\mathbf{y}}(\alpha_k,\beta_k;C). \tag{6.70}$$

Considering the interpolations and the summation, the full complexity is $\mathcal{O}\left(2N+\breve{N}^2\log(\breve{N})\right)$ which reduces to the dominant $\mathcal{O}\big(\breve{N}^2\log(\breve{N})\big)$.

6.3.3 Validation and comparison of the joint entropic estimators

Similar to the 1D case (see Section 6.1.4), both joint entropy estimators are validated by comparing their performance in computing the joint entropy of two RVs $\{\mathbf{x}, \mathbf{y}\}$ which follow a bi-variate normal density $\mathcal{N}(\mu, \Sigma^\star)$. The joint entropy of $\mathcal{N}(\mu, \Sigma^\star)$ is analytically derived [Cover and Thomas, 1991] and given by

$$h^\star(\mathcal{N}(\mu, \Sigma^\star)) = 0.5 \log \left((2\pi e)^2 \left| \Sigma^\star \right| \right) \tag{6.71}$$

where $|\Sigma^\star|$ denotes the determinant of $\Sigma^\star$. The estimators obtain the estimates by utilizing a joint sample C drawn from $\{\mathbf{x}, \mathbf{y}\}$. Again, in an analogy to the 1D case, the values of $u_\mathbf{x}$ and $u_\mathbf{y}$ in Σ of the K_Σ can affect the estimates. Their analytically derived optimum values which minimize the $MISE$ between $\hat{p}_{\mathbf{x},\mathbf{y}}(x, y; C)$ and $\mathcal{N}(\mu, \Sigma^\star)$ can be found in Silverman [Silverman and Green, 1986] and are given by $v_\mathbf{x}^{\text{opt}} = v_\mathbf{y}^{\text{opt}} = 0.96 N^{-1/6} \times 0.5(\sigma_\mathbf{x}^2 + \sigma_\mathbf{y}^2)$, where $\sigma_\mathbf{x}, \sigma_\mathbf{y}$ are the true standard deviations in $\mathcal{N}(\mu, \Sigma^\star)$. The optimum covariance matrix for K_Σ is denoted as Σ^{opt}.

For the validation of the estimators, we consider multiple isotropic normal densities $\mathcal{N}(0, \Sigma_i^\star)$, where $(\sigma_\mathbf{x})_i = (\sigma_\mathbf{y})_i$ in $\Sigma_i^\star$ vary from 0.005 to 2. For each $\mathcal{N}(\mu, \Sigma_i^\star)$ we instantiate multiple samples C_j. Figure 6.14 graphically shows such a sample C drawn from $\mathcal{N}(0, u_\mathbf{x} = u_\mathbf{y} = 0.005)$.

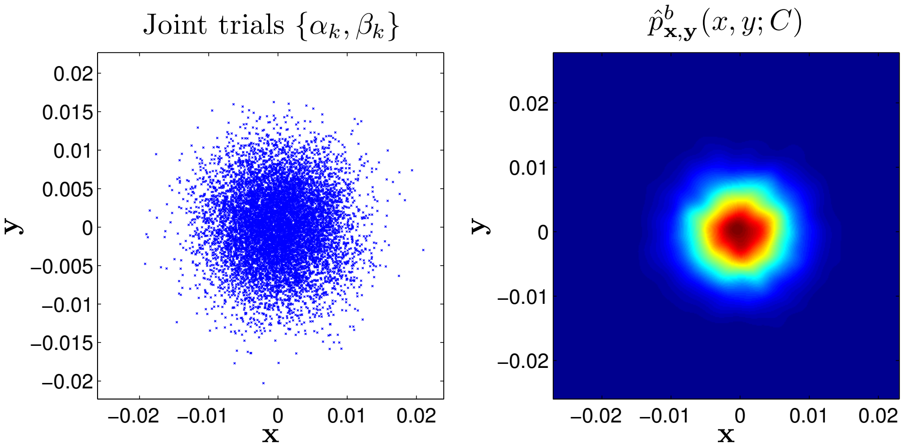


Figure 6.14: Example of a normally distributed joint sample used in joint entropy evaluation. Visualization of joint trials $\{\alpha_k, \beta_k\}$ constituting the joint sample C drawn from the jointly distributed RVs $\{\mathbf{x}, \mathbf{y}\} \sim \mathcal{N}(\mu, \Sigma^\star)$. Regarding the specifics of this example, $\sigma_\mathbf{x} = \sigma_\mathbf{y} = 0.005$ and $\mu = [0, 0]^\text{T}$. The joint PDF estimate is also visualized showing the normally distributed nature of the RVs.

Figure 6.15a shows the plots of both $\hat{h}^b_s(\mathbf{x}, \mathbf{y}; C)$ and $\hat{h}^b_e(\mathbf{x}, \mathbf{y}; C)$ for the various target $\mathcal{N}(0, \Sigma_i^\star)$. For each different $\Sigma_i^\star$, the depicted value corresponds to the average of multiple entropic estimates. Figure 6.15b shows a detail of 6.15a. The normalized mean squared errors between the estimators and the true entropic value, for two different sample sizes N is shown in Table 6.7. In addition similar to the marginal case, the Shannon estimates deviate from the more accurate empirical ones by an almost constant offset $\bar{\epsilon}$.

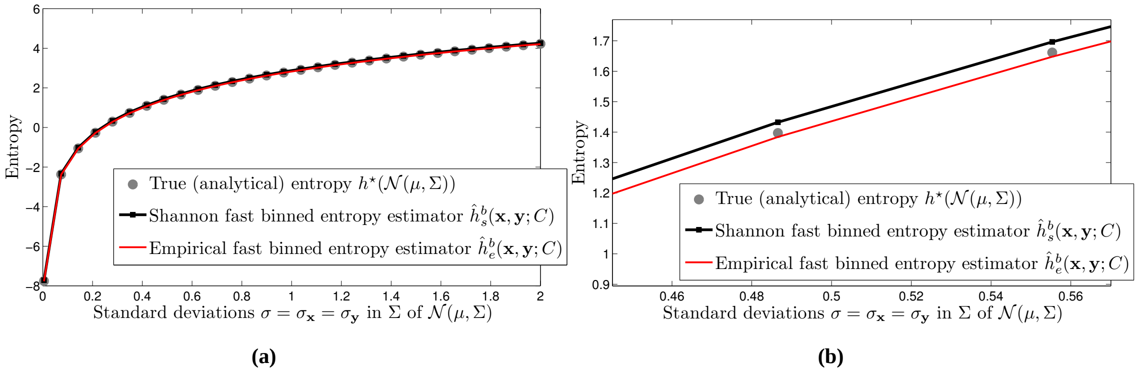


Figure 6.15: Comparison between joint Shannon & empirical entropy estimates of a sample drawn from normally distributed RVs $\mathbf{x}, \mathbf{y}$. 6.15a: Entropic estimates are obtained for multiple C (see text). The analytically derived entropy $h^{\star}(\mathcal{N}(\mu, \Sigma_i))$, $\mu = [0, 0]^{\mathrm{T}}$ is also plotted. The monotonically ascending graph is expected as $\mathcal{N}(\mu, \Sigma_i)$ becomes more clustered as $\sigma_{\mathbf{x}} = \sigma_{\mathbf{y}}$ become smaller. Clustered JPDFs are characterized by lower values of entropy. The size of the sample for this case is $N = 100 \times 100$. 6.15b: Detail of 6.15a.

Table 6.7: Normalized error of the joint Shannon & empirical entropy estimators against the "gold standard" analytic joint entropy of the bi-variate Normal density

Entropy Estimator	Empirical	Shannon	(MSE%) $\bar{\epsilon}$
Normalized error $(N = 100 \times 100)$	0.3%	1%	1% $(\bar{\epsilon} = 1\%)$
Normalized error $(N = 1000 \times 1000)$	0.02%	0.2%	0.9% $(\bar{\epsilon} = 0.3\%)$

6.4 Efficient joint entropy derivative computation

As in the 1D case, the computation of the partial derivatives of both $\hat{h}^b_e(\mathbf{x}, \mathbf{y}; C)$ and $\hat{h}^b_s(\mathbf{x}, \mathbf{y}; C)$ with respect to one of the images - that is trials α_k or β_k - can be efficiently computed via 2D FFT based convolutions. We consider the derivatives with respect to α_k. The derivatives with respect to β_k can be obtained by a simple change of variable due to the symmetry of JE.

6.4.1 Shannon's joint entropy derivatives

The derivatives of the Shannon estimator $\hat{h}_s(\mathbf{x}, \mathbf{y}; C)$ with respect to the continuous gray values α_k of image $\mathbf{x}$ and for all trial pairs $\{\alpha_k, \beta_k\}$ are

$$
\begin{aligned}
\frac{\partial \hat{h}_s(\mathbf{x},\mathbf{y};C)}{\partial \alpha_k} &\overset{\text{Eq. 6.68}}{=} -\Delta\breve{x}\Delta\breve{y}\sum_{i,j=1}^{\breve{N}}\left[\frac{\partial \hat{p}_{\mathbf{x},\mathbf{y}}(\breve{x}_i,\breve{y}_j;C)}{\partial \alpha_k}\log\left(\hat{p}_{\mathbf{x},\mathbf{y}}(\breve{x}_i,\breve{y}_j;)\right)+\right.\\
&\qquad \left.\hat{p}_{\mathbf{x},\mathbf{y}}(\breve{x}_i,\breve{y}_j;C)\frac{\partial \log\left(\hat{p}_{\mathbf{x},\mathbf{y}}(\breve{x}_i,\breve{y}_j;C)\right)}{\partial \hat{p}_{\mathbf{x},\mathbf{y}}(\breve{x}_i,\breve{y}_j;C)}\frac{\partial \hat{p}_{\mathbf{x},\mathbf{y}}(\breve{x}_i,\breve{y}_j;C)}{\partial \alpha_k}\right] &\quad (6.72)\\
&= -\Delta\breve{x}\Delta\breve{y}\sum_{i,j=1}^{\breve{N}}\left(\log\left(\hat{p}_{\mathbf{x},\mathbf{y}}(\breve{x}_i,\breve{y}_j;C)\right)+1\right)\frac{\partial \hat{p}_{\mathbf{x},\mathbf{y}}(\breve{x}_i,\breve{y}_j;C)}{\partial \alpha_k} &\quad (6.73)\\
&\overset{\text{Eq. 6.59}}{=} -\frac{\Delta\breve{x}\Delta\breve{y}}{N}\sum_{i,j=1}^{\breve{N}}\left(\log\left(\hat{p}_{\mathbf{x},\mathbf{y}}(\breve{x}_i,\breve{y}_j;C)\right)+1\right)\frac{\partial K_\Sigma(\breve{x}_i-\alpha_k,\breve{y}_j-\beta_k)}{\partial \alpha_k} &\quad (6.74)
\end{aligned}
$$

with

$$
K'_\Sigma(\breve{x}_i-\alpha_k,\breve{y}_j-\beta_k) \overset{\text{Eq. 6.62}}{=} -K_\Sigma(\breve{x}_i-\alpha_k,\breve{y}_j-\beta_k)\left(\frac{\breve{x}_i-\alpha_j}{u_{\mathbf{x}}^2}\right) \quad (6.75)
$$

Evaluating Eq. 6.74 for all N trials α, has a complexity of $\mathcal{O}(\breve{N}^2N)$. The first stage towards the reduction of the complexity, requires the replacement of the derivative estimator with its binned analogue $\hat{h}_s^b(\mathbf{x},\mathbf{y};C)$ utilizing the binned KDE $\hat{p}^b_{\mathbf{x},\mathbf{y}}(\breve{x}_i,\breve{y}_j;C)$. The joint trials considered in this case are the regular grid nodal positions $\{\breve{x}_k,\breve{x}_l\}$, $\forall k,l=1\rightarrow\breve{N}$. It should be understood that the derivative with respect to the trials of **x**, has to be computed with respect to the $\breve{x}_k$ part of all $\breve{N}^2$ binned trial pairs $\{\breve{x}_k,\breve{x}_l\}$. The binned version of Eq. 6.74 is

$$
\frac{\partial \hat{h}_s^b(\mathbf{x},\mathbf{y};C)}{\partial \breve{x}_{k,l}} = -\frac{\Delta\breve{x}\Delta\breve{y}}{N}\sum_{i,j=1}^{\breve{N}}\left(\log\left(\hat{p}^b_{\mathbf{x}}(\breve{x}_i,\breve{y}_j;C)\right)+1\right)\frac{\partial K_\Sigma(\breve{x}_i-\breve{x}_k,\breve{y}_j-\breve{y}_l)}{\partial \breve{x}_{k,l}} \quad (6.76)
$$

where $\breve{x}_k=\breve{x}_{k,l}$, $\forall l$. Hence, the last term of Eq. 6.76 is effectively $\partial K_\Sigma(\breve{x}_i-\breve{x}_k,\breve{y}_j-\breve{y}_l)/\partial\breve{x}_k$. The only reason we employ a double index is to emphasize that the derivative is computed in $\breve{N}^2$ 2D grid locations $\{\breve{x}_k,\breve{x}_l\}$ and not in $\breve{N}$ 1D grid locations $\breve{x}_k$.

Eq. 6.76 has a 2D convolution structure and the derivatives at all $\breve{N}^2$ locations can be computed efficiently as the product of the 2D FFT of the involved quantities

$$
\begin{aligned}
\frac{\partial \hat{h}_s^b(\mathbf{x},\mathbf{y};C)}{\partial \breve{x}} &= \left(\log\left(\hat{p}^b_{\mathbf{x},\mathbf{y}}(\breve{x},\breve{y};C)\right)+1\right)\star\star\left(-\frac{\Delta\breve{x}\Delta\breve{y}}{N}\frac{\partial K_\Sigma(\breve{x},\breve{y})}{\partial\breve{x}}\right) &\quad (6.77)\\
&= \mathcal{F}^{-1}\left\{\mathcal{F}\left\{\log\left(\hat{p}^b_{\mathbf{x},\mathbf{y}}(\breve{x},\breve{y};C)\right)+1\right\}\times\mathcal{F}\left\{-\frac{\Delta\breve{x}\Delta\breve{y}}{N}\frac{\partial K_\Sigma(\breve{x},\breve{y})}{\partial\breve{x}}\right\}\right\} &\quad (6.78)
\end{aligned}
$$

Similar to the 1D version, the 2D Gaussian kernel is chosen to have a limited support of $12u_{\mathbf{x}}$ and $12u_{\mathbf{y}}$ in each direction. Again both quantities are padded with zeros up to a size of N_{pad}^2 with $N_{pad}=\breve{N}+N_K-1$. This leads to a complexity in Eq. 6.78 of $\mathcal{O}\big(N_{pad}^2\log(N_{pad})\big)\rightarrow\mathcal{O}\big(\breve{N}^2\log(\breve{N})\big)$. This is a huge reduction compared to the initial complexity $\mathcal{O}(\breve{N}^2N)$ of Eq. 6.74. To comprehend the computational savings, assume the case of two 2D images of $N = 100^2$. Let $\breve{N} = 400$ and also assume a realistic $N_K = 100$. The achieved reduction in the order of complexity is three orders of magnitude. Assuming moderate size 3D images of $N = 100^3$, then the achieved cost reduction is a massive five orders of magnitude.

The final derivative is required at the initial A continuous locations α_k and is obtained via the interpolation utilizing the pre-saved quantities depicted in Fig. 6.13. The interpolation is expressed as

$$\begin{aligned}\frac{\partial \hat{h}^b_s(\mathbf{x},\mathbf{y};C)}{\partial \alpha_k} =& \left(1-\mathfrak{b}^x_k\right)\cdot\left(1-\mathfrak{b}^y_k\right)\cdot\frac{\partial \hat{h}^b_s(\mathbf{x},\mathbf{y};C)}{\partial \breve{x}_{(\mathfrak{i}_k,\mathfrak{j}_k)}} + \\ & \left(\mathfrak{b}^x_k\right)\cdot\left(1-\mathfrak{b}^y_k\right)\cdot\frac{\partial \hat{h}^b_s(\mathbf{x},\mathbf{y};C)}{\partial \breve{x}_{(\mathfrak{i}_k+1,\mathfrak{j}_k)}} + \\ & \left(1-\mathfrak{b}^x_k\right)\cdot\left(\mathfrak{b}^y_k\right)\cdot\frac{\partial \hat{h}^b_s(\mathbf{x},\mathbf{y};C)}{\partial \breve{x}_{(\mathfrak{i}_k,\mathfrak{j}_k+1)}} + \\ & \left(\mathfrak{b}^x_k\right)\cdot\left(\mathfrak{b}^y_k\right)\cdot\frac{\partial \hat{h}^b_s(\mathbf{x},\mathbf{y};C)}{\partial \breve{x}_{(\mathfrak{i}_k+1,\mathfrak{j}_k+1)}}, \quad \forall k = 1,2,\ldots,N. \end{aligned} \tag{6.79}$$

6.4.2 Empirical joint entropy derivatives

The derivatives of the empirical joint entropy with respect to the continuous α_k are given by

$$\begin{aligned}\frac{\partial \hat{h}_e(\mathbf{x},\mathbf{y};C)}{\partial \alpha_k} &= -\frac{\partial}{\partial \alpha_k}\left(\frac{1}{N}\sum_{i=1}^{N}\log\left(\hat{p}_{\mathbf{x},\mathbf{y}}(\alpha_i,\beta_i;C)\right)\right) && (6.80)\\ &= -\frac{1}{N}\sum_{i=1}^{N}\frac{1}{\hat{p}_{\mathbf{x},\mathbf{y}}(\alpha_i,\beta_i;C)}\frac{\partial}{\partial \alpha_k}\left(\hat{p}_{\mathbf{x},\mathbf{y}}(\alpha_i,\beta_i;C)\right) && (6.81)\\ &= -\frac{1}{N}\sum_{i=1}^{N}\Bigg[\left(\frac{1}{\frac{1}{N}\sum_{j=1}^{N}K_\Sigma(\alpha_i-\alpha_j,\beta_i-\beta_j)}\right)\times \\ &\qquad \frac{\partial}{\partial \alpha_k}\left(\frac{1}{N}\sum_{j=1}^{N}K_\Sigma(\alpha_i-\alpha_k,\beta_i-\beta_k)\right)\Bigg] && (6.82)\\ &= -\frac{1}{N}\sum_{i=1}^{N}\Bigg[\left(\frac{1}{\frac{1}{N}\sum_{j=1}^{N}K_\Sigma(\alpha_i-\alpha_j,\beta_i-\beta_j)}\right)\times \\ &\qquad \left(\frac{1}{N}\sum_{j=1}^{N}K'_\Sigma(\alpha_i-\alpha_j,\beta_i-\beta_j)\left(\breve{\delta}(i,k)-\breve{\delta}(j,k)\right)\right)\Bigg] && (6.83)\\ &= -\frac{1}{N}\frac{\frac{1}{N}\sum_{j=1}^{N}K'_\Sigma(\alpha_k-\alpha_j,\beta_k-\beta_j)}{\hat{p}_{\mathbf{x},\mathbf{y}}(\alpha_k,\beta_k;C)} \\ &\quad +\frac{1}{N}\sum_{i=1}^{N}\frac{\frac{1}{N}K'_\Sigma(\alpha_i-\alpha_k,\alpha_j-\beta_k)}{\hat{p}_{\mathbf{x},\mathbf{y}}(\alpha_i,\beta_i;C)} && (6.84)\end{aligned}$$

Equation 6.84 can be re-written as

$$\frac{\partial \hat{h}_e(\mathbf{x},\mathbf{y};C)}{\partial \alpha_k} = \frac{1}{N}\frac{F_1(\alpha_k,\beta_k;C)}{\hat{p}_{\mathbf{x},\mathbf{y}}(\alpha_i,\beta_i;C)} - F_2(\alpha_k,\beta_k;C) \tag{6.85}$$

with

$$F_1(\alpha_k,\beta_k;C) = \frac{1}{N}\sum_{j=1}^{N} K'_\Sigma(\alpha_k-\alpha_j,\beta_k-\beta_j), \quad \text{and} \tag{6.86}$$

$$F_2(\alpha_k,\beta_k;C) = \frac{1}{N}\sum_{i=1}^{N} \frac{\frac{1}{N}K'_\Sigma(\alpha_i-\alpha_k,\alpha_j-\beta_k)}{\hat{p}_{\mathbf{x},\mathbf{y}}(\alpha_i,\beta_i;C)}. \tag{6.87}$$

Similar to the 1D case, $F_1(\alpha_k,\beta_k;C)$ is computed efficiently in the 2D regular grid by utilizing the 2D analogue of the convolution property expressed in Eq. 6.48 which can be computed via the 2D FFT

$$F_1(\breve{x},\breve{y};C) = \frac{1}{N}K_\Sigma{}'(\breve{x},\breve{y}) \star\star w(\breve{x},\breve{y}) \tag{6.88}$$

$$= \mathcal{F}^{-1}\left\{\mathcal{F}\left\{w(\breve{x},\breve{y})\right\} \times \mathcal{F}\left\{\frac{1}{N}\frac{\partial K_\Sigma(\breve{x},\breve{y})}{\partial \breve{x}}\right\}\right\} \tag{6.89}$$

The scheme continues by interpolation $F_1(\breve{x},\breve{y};C)$ back to the continuous locations in a manner similar to the interpolation of Eq. 6.69, where $\hat{p}^b_{\mathbf{x},\mathbf{y}}(\cdot;)$ in that Equation is replaced by $F_1(\cdot)$. Finally, the denominator of Eq. 6.86 is already available from Eq. 6.69.

In a similar manner, $F_2(\breve{x},\breve{y};C)$ is firstly computed over the regular grid according to

$$F_2(\breve{x},\breve{y};C) = \frac{1}{N}K'_\Sigma{}^{\text{mirror}}(\breve{x},\breve{y}) * \frac{w(\breve{x},\breve{y})}{\hat{p}^b_{\mathbf{x}}(\breve{x},\breve{y};C)} \tag{6.90}$$

$$= \mathcal{F}^{-1}\left\{\mathcal{F}\left\{\frac{1}{N}K'_\Sigma{}^{\text{mirror}}(\breve{x},\breve{y})\right\} \times \mathcal{F}\left\{\frac{w(\breve{x},\breve{y})}{\hat{p}^b_{\mathbf{x}}(\breve{x},\breve{y};C)}\right\}\right\}, \tag{6.91}$$

and it is subsequently interpolated onto the continuous locations in a manner similar to Eq. 6.79.

Eq. 6.84 has a complexity of $\mathcal{O}(N^2)$ whereas both $F_1(\cdot)$, $F_2(\cdot)$ have $\mathcal{O}\Big(N^2_{pad}\log(N_{pad})\Big) \rightarrow \mathcal{O}\Big(\breve{N}^2\log(\breve{N})\Big)$. Actually, the empirical derivative requires two FFT convolutions - one for each $F_1(\cdot)$, $F_2(\cdot)$ term and is double of the Shannon derivatives. However, the final order of complexity is again reduced to simply $\mathcal{O}\Big(\breve{N}^2\log(\breve{N})\Big)$. Having efficiently computed the terms in the empirical JE expression of Eq. 6.85 at the regular locations, the derivative at the continuous α is obtained via an interpolation similar to the one of Eq. 6.79. Considering a 2D $N = 100^2$ image $\breve{N} = 400$, then the reduction is two orders of magnitude whereas for a 3D $N = 100^3$ and the same sized grid, it is a massive six orders of magnitude.

Table 6.6 summarizes the complexities of the joint estimators and their derivatives.

6.4.3 Comparison of the derivatives of the joint entropy estimators

In order to assess the accuracy of the derivatives of the JE we consider the two images depicted in Fig. 6.16. The second row depicts the FFT enabled analytic derivatives of each estimator with respect to the individual pixels of image 1. The derivatives are obtained for kernel standard deviations values $u_{\mathbf{x}}$ and $u_{\mathbf{y}}$ which returned the best derivative match. The derivatives are plotted on a common scale. The last image shows the normalized squared difference between the two images. The last row depicts the derivatives obtained by individual perturbations in the gray value of each pixel, following a similar scheme to the

Table 6.8: Joint entropy evaluation and derivative estimation: Computational complexity of explicit as well as the FFT based implementations

Non-FFT	Order of Complexity	FFT	Order of complexity
$\hat{h}_s(\mathbf{x},\mathbf{y};C)$	$\mathcal{O}(N\breve{N}^2)$	$\hat{h}_s^b(\mathbf{x},\mathbf{y};C)$	$\mathcal{O}\Big(\breve{N}^2\log(\breve{N})\Big)$
$\hat{h}_e(\mathbf{x},\mathbf{y};C)$	$\mathcal{O}(N^2)$	$\hat{h}_e^b(\mathbf{x},\mathbf{y};C)$	$\mathcal{O}\Big(\breve{N}^2\log(\breve{N})\Big)$
$\partial\hat{h}_s(\mathbf{x},\mathbf{y};C)/\partial\alpha_k$	$\mathcal{O}(N\breve{N}^2)$	$\partial\hat{h}_s^b(\mathbf{x};\breve{A})/\partial\alpha_i$	$\mathcal{O}\Big(\breve{N}^2\log(\breve{N})\Big)$
$\partial\hat{h}_e(\mathbf{x},\mathbf{y};C)/\partial\alpha_k$	$\mathcal{O}(N^2)$	$\partial\hat{h}_e^b(\mathbf{x};\breve{A})/\partial\alpha_i$	$\mathcal{O}\Big(\breve{N}^2\log(\breve{N})\Big)$

one described in Sec. 6.2.1. The errors associated with Fig. 6.16 are provided in Table 6.9. Although the error between the two estimator is significant, the errors with respect to the derivatives can be considered acceptable.

We have also attempted to convert image 1 to image 2 by following a line-search enabled iterative gradient descent scheme, where in each iteration, the derivative of image 1 was subtracted by the image itself. However, both estimators stuck in local minima due the null spaces characterizing the JE functional. The null spaces are present, due to the gray value invariance of the JE functional.

Table 6.9: Errors regarding the derivatives of the joint Shannon & empirical entropy estimators

Compared derivative quantities	Normalized error(%)
Analytic Shannon vs analytic empirical	9%
Analytic Shannon vs Shannon	2.2%
Analytic Empirical vs Empirical	2.7%

6.4.4 Run time tests

We compute JE and and its derivative for images of different sizes, using the non-FFT based estimator as well as the FFT enabled Shannon and empirical estimators. The slow empirical estimator is not plotted as it is even slower than the non-FFT Shannon analogue. The computational times are presented in Fig. 6.17b. Although the order of computational complexity is similar in both images, the exact complexity of the empirical estimator is higher than the Shannon one and this difference is depicted in the plots.

6.5 Efficient mutual information evaluation via marginalization of the joint probability density function

The MI functional (Eq. 4.50) requires the estimation of both marginal and JE terms as well as their derivatives. One can compute all involved terms with the methods discussed in the previous sections. Regarding the functional evaluation, the main source of computational complexity reduction is the efficient PDF and joint probability density function (JPDF) estimation via the linear interpolations on the

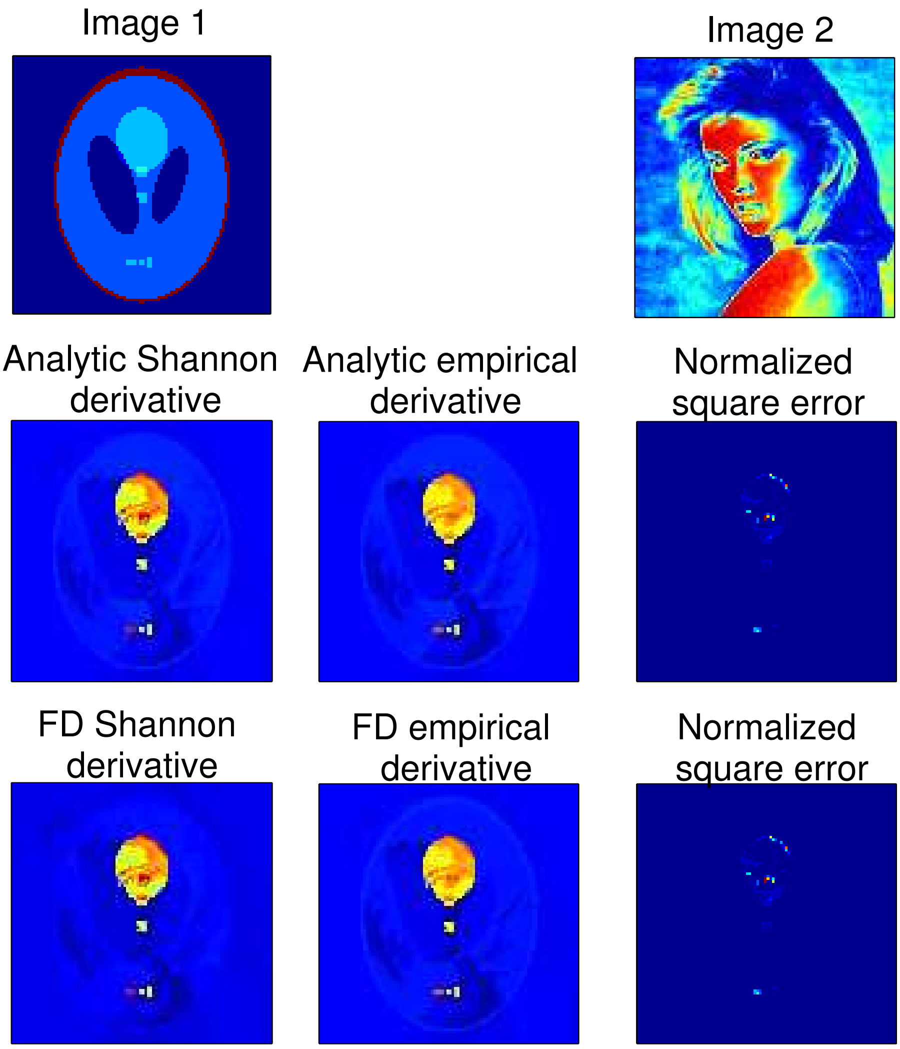
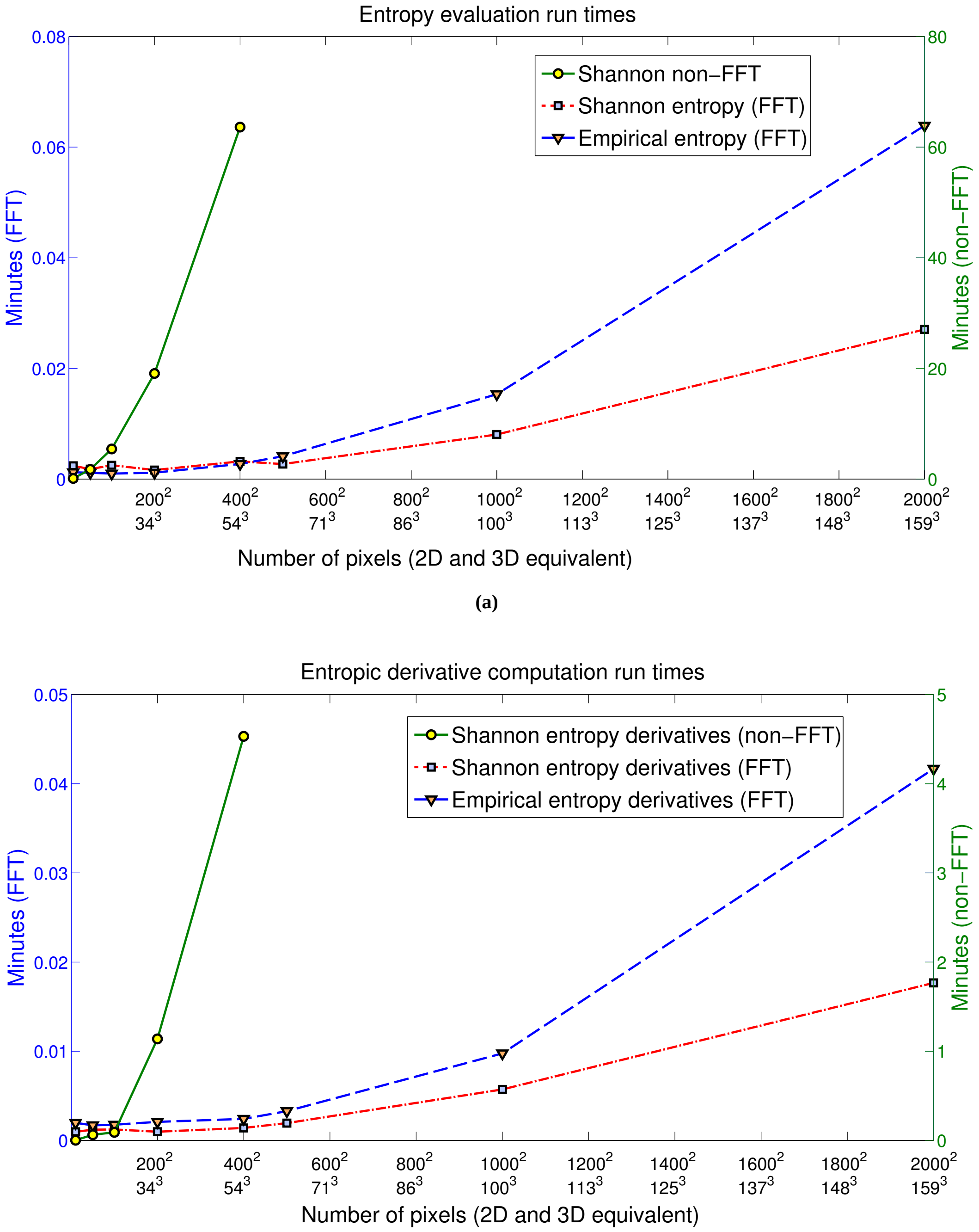


Figure 6.16: Derivatives of the Shannon and empirical joint entropy estimators. The figure depicts the analytic and derivatives of the two estimators, with respect to perturbations applied on the pixels of image 1.

regular grid and and the application of the FFT. The best approach for this is firstly to compute the JPDF and then subsequently derive the marginal PDF via the integral analogue of Eq. 4.20.

6.6 Summary

In this chapter we have described in detail the marginal and joint entropy estimators expressed either as an integral formulation (Shannon entropy) or as an expectation (empirical entropy). We have extended the work of Shwartz et al. [2005] enabling the efficient evaluation of the joint entropy estimators and their derivatives. We have tested both entropic formulations and their derivatives against ‘gold standards’ and discussed the response of the former with respect to the number of bins and the choice of the kernel’s standard deviation used in the non-parametric KDE. The evaluation against ‘gold standards’ is necessary to validate the correctness of the entropic estimators’ implementation. The validated methods are employed later in this work, for estimating the entropy of images - as well as the derivative of the entropy with respect to the images’ gray values - all done for the purpose of regularizing the inverse problem in optical imaging. Finally we derived the orders of complexity of the discussed concepts and provided examples showcasing the run time efficiency of each estimator and its derivative. The suitability of the implementations for image reconstruction are evaluated in Sec. 7.5.



(a)

(b)

Figure 6.17: Run times for *JE* and derivative computation between square images of different size. As *JE* does not depend on image dimensionality, we plot the 2D size of the images as well as the approximate 3D volume equivalent. 6.17a Computational time for non-FFT Shannon entropy evaluation and FFT Shannon and empirical entropy evaluation. The slow method is evaluated up to a size of $400x400$ due to increased complexity. 6.17b Similar plots for derivative estimation.

Chapter 7

Information theoretic regularization in diffuse optical tomography

7.1 Introduction

In this chapter we propose a regularization scheme for diffuse optical tomography (DOT) based on information theory (IT). The proposed method and part of the results presented in this discussion, has been published in [Arridge et al., 2008b; Panagiotou et al., 2009a,b]. The scheme enables the incorporation of *a priori* structural information in the inverse problem, aiding towards the alleviation of the negative effects of ill-posedness. As a consequence, it results in an improvement in the quantitative accuracy as well as in the spatial resolution of the obtained optical solution. The scheme addresses the first of the aims set in chapter 1.

The *a priori* information is provided in the form of pairs of reference images $x_{\text{ref}}^{\mu_a}$ and $x_{\text{ref}}^{\mu'_s}$, expressed also by a combined notation

$$x_{\text{ref}} := \begin{pmatrix} x_{\text{ref}}^{\mu_a} \\ x_{\text{ref}}^{\mu'_s} \end{pmatrix}. \tag{7.1}$$

The reference images depict some secondary quantities of different physical meaning - for example the magnetic properties of tissue obtained by a magnetic resonance imaging (MRI) - which under a fundamental assumption, are expected to be spatially distributed in a manner similar to the true - and initially unknown, optical quantities of interest $\mu_a^\star$ and $\mu_s'^\star$ respectively. The two latter quantities can also be referenced using a combined notation $x^\star$ in a manner similar to Eq. 7.1. The pre-requisite regarding the similarity between the reference images and the true optical solutions, is solely limited to their spatial distribution patterns - also interpreted as *structure*. No assumptions are made regarding the relationship between the gray values, which populate spatially corresponding locations or features, between prior and optical images. It should be noted that in practice a single image $x_{\text{ref}} = x_{\text{ref}}^{\mu_a} = x_{\text{ref}}^{\mu'_s}$ will be available, as most of the high resolution imaging modalities return a single image - although one can consider the possibility of using images obtained from different modalities for μ_a and μ'_s for example functional magnetic resonance imaging (fMRI) as priors for absorption and anatomical MRI for scattering. In the numerical simulations of this chapter, we will use both combinations - different and common priors for μ_a and μ_s'.

Knowing *a priori* the expected structure of $x^\star$ - at least to some extent, allows the penalization of the optical solutions according to their level of structural dissimilarity with the reference images or equivalently, the regularization scheme favours solutions similar to the reference images. By assigning different penalties to different solutions, the regularization effectively changes the solution space. Ideally, this leads to an improvement in the definition of the optima and hopefully to a treatment of the null-spaces, which plague the solution retrieval process of the un-regularised inverse problem. In this work we consider joint entropy (JE) and mutual information (MI) as the functionals of choice.

Information theoretic regularization with explicit *a priori* information has recently drawn increased attention. Somayajula et al. [2005] proposed the usage of MI based priors for the linear inverse problem in positron emission tomography (PET). The method was further extended to enable a scale space approach. Specifically, the method regularized the PET solution by assessing its MI similarity with the prior images at various scales. The priors at their various scales were considered simultaneously, thus the final PET solution was the one most similar to all scale-priors. The scales at that study were i) the images at their initial resolution ii) after being subjected to low-pass filtering and iii) after the application of differential operators. The pixel values at the last two scales incorporate neighbourhood information - for example the gradient at a pixel location depends on the values of its neighbouring pixels. Thus, via the consideration of these scales, the method implicitly modelled spatial dependency between neighbouring pixels. The latter is missing from the standard IT functional implementations, based on kernel density estimation (KDE) methods operating under the independent and identically distributed (i.i.d.) assumption. Nuyts [2007] followed by commenting on the inferiority of the MI functional compared to JE for the purpose of reconstruction regularization, although the study did not entail an in-depth comparison between the two. Van de Sompel and Sir., Brady, M. [2009b] incorporated JE priors in the linear inverse problem of limited view tomography. In a manner similar to Somayajula et al. [2007] they studied the effects of introducing inter-pixel spatial dependency in the JE priors, but their approach differed as now the pixel dependency was explicitly enforced using a Markov random field smoothness prior, effective incorporated as an extra regularization functional. Kazantsev et al. [2010] proposed an optimization scheme for JE priors in PET where the regularization weight was optimized in real-time, simultaneously with the image reconstruction. Pedemonte et al. [2010a] proposed a class-conditional JE scheme for introducing priors in single photon emission computed tomography (SPECT) reconstruction. Finally, Tang et al. [2010] investigated the use of JE prior in reconstruction of 4-D datasets.

The contributions of the work presented in this chapter include:

I) the first application of IT regularization in a severely ill-posed, non-linear inverse problem such as DOT,

II) an in-depth theoretical analysis and comparison regarding the capacity of both JE and MI to act as regularizing functionals, accompanied by custom-made simulations to support the findings,

III) the use of the efficient scheme for marginal and joint entropy evaluation and gradient computation discussed in chapter 6 for the purpose of IT regularization. The accuracy of the presented results in this section, is a strong indicator regarding the validity of the efficient entropic estimators and gradients. IV) finally, the application of the priors to 3D experimental data obtained from a phantom study.

The structure of this chapter is as follows: Section 7.2 re-formulates the inverse problem in DOT in order to enable IT regularization and discusses the adopted optimization scheme. Section 7.3 provides theoretical intuition between the differences between JE and MI regularization. Section 7.4 revisits the choice of binning range during the KDE as well as the standard deviation of the used kernels. Finally sections 7.5-7.7 present 2D and 3D numerical simulations, whereas a study based on experimental data is presented in Sec. 7.8.

7.2 Formulation of the inverse problem

The proposed objective function $\mathcal{E}(x)$ enabling the IT regularization of DOT is defined as

$$\mathcal{E}(x) = \left\| \frac{\acute{y} - \mathcal{F}(\mathcal{S}^{-1}(x))}{c_1} \right\|^2 + \tau\Psi(x, x_{\text{ref}}) \tag{7.2}$$

where $\mathcal{S}^{-1}(x) = x_+ = [\mu_a, \kappa]^{\text{T}}$ denotes the optical parameters estimates $x := [\breve{\mu}_a, \breve{\kappa}]^{\text{T}}$ are the logarithmically transformed and normalized analogues (see Sec. 3.5.4); the data fit term is the L_2 norm; $\Psi(x, x_{\text{ref}})$ is an IT regularization functional assessing the similarity between x and x_{ref} weighted by the regularization parameter τ; c_1 is the normalizing constant defined in Eq. 3.69. The evaluation of the forward operator $\mathcal{F}\left(x_+\right)$ is approached using a finite element method (FEM) based approach. We utilize the TOAST software package mentioned in Sec. 3.4.3.

The solution $\hat{x}$ of the scheme is obtained via a minimization scheme

$$\hat{x} = \underset{x}{\arg\min} \left[\mathcal{E}(x) = \left\| \frac{\acute{y} - \mathcal{F}\big(\mathcal{S}^{-1}(x)\big)}{c_1} \right\|^2 + \tau\Psi(x, x_{\text{ref}}) \right]. \tag{7.3}$$

The proposed scheme is realised by replacing $\Psi(x, x_{\text{ref}})$ with either the differential JE or the negative mutual information -MI. Then considering that $x := [\breve{\mu}_a, \breve{\kappa}]^{\text{T}}$, $\Psi(x, x_{\text{ref}})$ is defined for the case of JE regularization as

$$\Psi(x, x_{\text{ref}}) = \hat{h}(x, x_{\text{ref}}) \tag{7.4}$$

$$= \hat{h}(\breve{\mu}_a, x_{\text{ref}}^{\mu_a}) + \hat{h}(\breve{\kappa}, \kappa_{\text{ref}}), \tag{7.5}$$

whereas for the case of MI regularization as

$$\begin{aligned}
\Psi(x, x_{\text{ref}}) &= - \mathit{MI}(x, x_{\text{ref}}) && (7.6)\\
&= - \left(\hat{h}(x) + \hat{h}(x_{\text{ref}}) - \hat{h}(x, x_{\text{ref}})\right) && (7.7)\\
&= - \mathit{MI}(\breve{\mu}_a, x_{\text{ref}}^{\mu_a}) - \mathit{MI}(\breve{\kappa}, \kappa_{\text{ref}}) && (7.8)\\
&= - \left(\hat{h}(\breve{\mu}_a) + \hat{h}(x_{\text{ref}}^{\mu_a}) - \hat{h}(\breve{\mu}_a, x_{\text{ref}}^{\mu_a})\right)\\
&\quad - \left(\hat{h}(\breve{\kappa}) + \hat{h}(\kappa_{\text{ref}}) - \hat{h}(\breve{\kappa}, \kappa_{\text{ref}})\right). && (7.9)
\end{aligned}$$

Note that the MI functional is negated. We know that the value of MI increases as solution estimates x become more similar to the priors x_{ref}. Considering that $\Psi(x, x_{\text{ref}})$ is introduced in minimization framework such as the one described by Eq. 7.2, the functional should attain a minimum value for solutions x maximally similar to x_{ref}, in order for these solutions to be minimally penalized. The negation of the MI is hence essential in order to achieve the desired behavior.

The retrieved optical solution estimate $\hat{x}$ is subsequently subjected to a transformation $\mathcal{S}^{-1}(x)$ (see Eq. 3.67) which results in the positive and un-normalized estimate of the optical solution $\hat{x}_+$.

7.2.1 Objective function minimization scheme

The minimization of Eq. 7.2 is approached using the iterative non-linear conjugate gradients (CG) optimization method utilizing the *Polak - Ribière* updating scheme described in Sec. 2.6.2.2. CG-based optimization has been used before in diffuse optical imaging - for example see Arridge and Schweiger [1998]. Methods based on higher order derivatives (Sec. 2.6.3) exhibit faster convergence however they are not considered in this context, as the entropic gradients discussed in Chapter 6 have only been analytically derived up to the first order. To further speed up convergence, we adopt the inexact line-search algorithm described in Sec. 2.6.4. An inexact line-search of this kind has been used by Schweiger et al. [2005] for the purposes of DOT reconstruction, although the principal optimization in that work is a second order method. The pseudo code for the CG and the inexact line search algorithms are provided in Algorithms 2.1 and 2.2.

The gradient $g^{(k)} = \partial\mathcal{E}(x^{(k)})/\partial x^{(k)}$ of Eq. 7.2 at iteration k and with respect to the parameter vector x is derived using the chain rule according to

$$\frac{\partial\mathcal{E}\left(x^{(k)}\right)}{\partial x^{(k)}} = -\frac{2}{c_1}\left(\acute{y} - \mathcal{F}\left({x_+}^{(k)}\right)\right)\frac{\partial\mathcal{F}\left({x_+}^{(k)}\right)}{\partial {x_+}^{(k)}}\frac{\partial {x_+}^{(k)}}{\partial x^{(k)}} + \tau\frac{\partial\Psi\left(x^{(k)}, x_{\text{ref}}\right)}{\partial x^{(k)}} \qquad (7.10)$$

Regarding the terms on the RHS of Eq. 7.10, the partial derivative of the forward operator $\partial\mathcal{F}\left({x_+}^{(k)}\right)/\partial x^{(k)}$ is the Jacobian matrix introduced in Sec. 3.5.3. Its computation can be accomplished efficiently by utilizing the photon measurement density functions (PMDFs) based approach discussed in the same section. The term $\partial{x_+}^{(k)}/\partial x^{(k)}$ is expanded by considering the parameter transformation discussed in Sec. 3.5.4 and results in

$$\begin{aligned}
\frac{\partial {x_+}^{(k)}}{\partial x^{(k)}} &\overset{\text{Eq. 3.67}}{=} \frac{\partial \exp(x)\bar{x}_+}{\partial x} && (7.11)\\
&\overset{\text{Eq. 3.66}}{=} x. && (7.12)
\end{aligned}$$

Finally, the derivative of $\Psi\left(x^{(k)}, x_{\text{ref}}\right)$ depends on which IT functional is adopted. In the case of JE it is simply equivalent to

$$
\begin{aligned}
\frac{\partial \Psi\left(x^{(k)}, x_{\text{ref}}\right)}{\partial x^{(k)}} &= \frac{\partial \hat{h}(x, x_{\text{ref}})}{\partial x}, &(7.13)\\
&= \left[\frac{\partial \hat{h}(\breve{\mu}_a, x_{\text{ref}}^{\mu_a})}{\partial \breve{\mu}_a}, \frac{\partial \hat{h}(\breve{\kappa}, \kappa_{\text{ref}})}{\partial \breve{\kappa}}\right]^{\mathrm{T}} &(7.14)
\end{aligned}
$$

which can be efficiently computed using the fast Fourier transform (FFT) enabled analytic derivative estimation scheme proposed in Sec. 6.4. In the case of MI the derivative is

$$
\begin{aligned}
\frac{\partial \Psi\left(x^{(k)}, x_{\text{ref}}\right)}{\partial x^{(k)}} &= -\frac{\partial \hat{h}(x)}{\partial x} + \frac{\partial \hat{h}(x, x_{\text{ref}})}{\partial x}, &(7.15)\\
&= \left[-\frac{\partial \hat{h}(\breve{\mu}_a)}{\partial \breve{\mu}_a} + \frac{\partial \hat{h}(\breve{\mu}_a, x_{\text{ref}}^{\mu_a})}{\partial \breve{\mu}_a}, -\frac{\partial \hat{h}(\breve{\kappa})}{\partial \kappa} + \frac{\partial \hat{h}(\breve{\kappa}, \kappa_{\text{ref}})}{\partial \breve{\kappa}}\right]^{\mathrm{T}} &(7.16)
\end{aligned}
$$

Once again, the efficient estimation of $\partial \hat{h}(x)/\partial x$ is discussed in Sec. 6.2. One can notice that derivative of the term $\hat{h}(x_{\text{ref}})$ of Eq. 7.9 is not included in the MI derivative, as it is independent of x. Also note that the presence of $\partial \hat{h}(x)/\partial x$ in the derivative of MI differentiates it from its JE analogue. In the following section where we compare JE with MI, the effects of the marginal term in the capacity of the functionals for the purpose of regularization will become clearer.

7.3 Comparison between JE and MI for regularization purposes: Theoretical intuition

Section 4.5.5 introduced JE and MI as similarity measures in the multi-modal setting. In Sec. 5.6.4, we revisited the functionals to comment on their differences in the image registration setting, where the functionals depend on the assessed images but also on the variable overlap domain between the images. Here, we revisit the functionals to assess the similarity between x_{ref} and a continuously varying x. In this section we assess the behaviour of the functionals, considering not only the differences in their gray values but also differences on the the structural features which are depicted by them. The reason for this analysis is to extract intuition regarding the bias expected to be introduced in the solution by JE and MI regularization, due to x_{ref} not having a 'one-to-one' structural feature correspondence with the true $x^{\star}$.

Before proceeding it is important to reiterate that the minimization of -MI (or equivalently maximization of MI) corresponds to maximization of the marginal $\hat{h}(x)$ and minimization of $\hat{h}(x, x_{\text{ref}})$. The former results in increased variation in the probability density function (PDF) whereas the minimization of JE results in increased clustering in the joint probability density function (JPDF).

Figure 7.1 is central to this discussion. The first row shows five test images, all under a common scale. In order to compare MI and JE, pairs of images are needed. For this reason we consider five image pairs $\{1, Z\}$, where $Z = 1, 2, \ldots, 5$. As the functionals are defined in terms of the PDFs/ JPDFs, the visualization of the latter is essential to obtain intuition regarding the behaviour of the functionals. The second row presents the marginal $\hat{p}(1)$ of image 1 - present in all five image pairs (first column) as well as the joint $\hat{p}(1,Z)$, $\forall Z$ (remaining columns). Row three shows the $\log\left(\hat{p}(1,Z)\right)$, which reveals the full extent of the clusters[1]. Finally, the fourth row shows the marginal $\hat{p}(Z)$ which varies among pairs. It is important to note that the same binning range, binning width and kernel standard deviation were used for the KDE in all cases. The axes of the PDFs/ JPDF are common in all PDFs/JPDFs. All PDFs and JPDFs are displayed under a common scale to enable consistent visual comparison.

The JE and - MI scores for each of the pairs, together with the marginal $\hat{h}(Z)$ involved in the MI computation, are provided in Table 7.1. We will refer to these values when we consider the formed pairs individually.

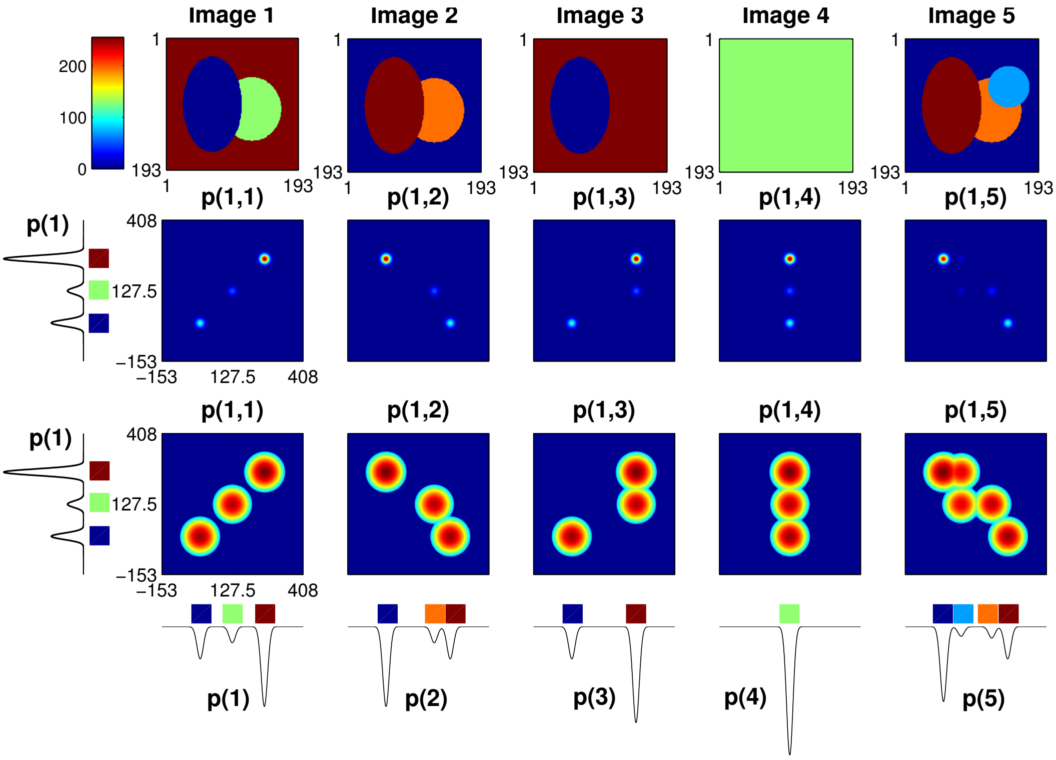


Figure 7.1: Test images and corresponding PDFs/JPDFs for comparing JE and MI. **Top row:** Test images. **Second row:** Marginal PDF of image 1 and JPDFs between image pairs $\{1,Z\}$ where $Z = 1, 2, \ldots, 5$. **Third row:** Same as second row but JPDFs are now in log scale. Bottom row: Marginal PDFs for each of the test images.

We should note that Image 1 is present in all image pairs and remains structurally unchanged. In order to relate the discussion with the image reconstruction framework, image 1 is interpreted as x_{ref} whereas the variable image Z completing the pair corresponds to $x^{\star}$ - except if stated otherwise.

7.3.1 JE vs MI: Image pair {1,1}

Image pair $\{1,1\}$ corresponds to the case of a perfect x_{ref} as it is identical to $x^{\star}$. $\hat{p}(1,1)$ displays the well known example - from image registration studies (for example see [Hill et al., 2001]), diagonal cluster arrangement which is characteristic of this case. This case was discussed in Sec. 4.5.5. The three clusters

[1] we use a KDE with kernel of finite support

Table 7.1: Joint entropy, negative mutual information and marginal entropy values for the image pairs in Figure 7.1

Image pair	$\{1,1\}$	$\{1,2\}$	$\{1,3\}$	$\{1,4\}$	$\{1,5\}$
$\hat{h}(1,Z)$	8.834	8.834	8.834	8.834	9.052 nats
$-MI(1,Z)$	-0.904	-0.897	-0.564	0	-0.845 nats
Image	1	2	3	4	5
$\hat{h}(Z)$	4.869	4.862	4.529	3.965	5.028 nats

in the JPDFs is the minimum number of clusters which can be formed given the fact that both images have three distinct intensities and all structural features are spatially registered. $\hat{h}(1,1)$ and $-MI(1,1)$ attain their theoretical global minimum in this case from all other possible $Z = 1$. For example, because image 1 and image Z are identical, knowing image 1 maximally reduces the uncertainty of the solution (image Z) - which is the definition of MI. By using image 1 as a prior, the solution image $Z = 1$ would be minimally penalized by the regularization functionals -as they attain their minimum values - and consequently the solution image Z would be correctly favoured over the majority of other solutions.

7.3.2 JE vs MI: Image pair {1,2}

Image pair $\{1,2\}$ corresponds again to the case of a structurally correct prior. However, the gray values of image 2 have been transformed according to the arbitrarily chosen non-linear function discussed in Sec. 4.5.5 and subsequently rescaled to the range of image 1. Table 7.1 shows that JE is invariant to the gray value transformation as $\hat{h}(1,2) = \hat{h}(1,1)$ up to the third decimal point. Regarding MI it is apparent that $MI(1,2) \neq MI(1,1)$ as it carries the error from $\hat{h}(1) < \hat{h}(2)$. We remind we seek to minimize -MI. This is due to the partial overlap of the two modes in $\hat{p}(2)$. One would expect that in minimization framework involving $-MI$, the functional would favour a solution Z where $\hat{h}(Z)$ is maximized. This corresponds to a highly spread $\hat{p}(Z)$, which in this case is translated as three completely non-overlaping modes in $\hat{p}(Z)$. Due to the partial overlap however this case does not correspond to the global minimum $-MI$ solution. Hence, if image 1 was used as a prior in a reconstruction scheme and the regularization parameter τ was very high, -MI would bias the gray values of the obtained solution in order to form an optimal $\hat{p}(Z)$ (with three distinct modes). In other words, it would attempt to increase the contrast among the formed features. The above effect will be demonstrated in practice in the case study 7.5 and specifically to Figs. 7.12b & 7.12d, which depict reconstructions regularized by MI for variable τ.

A similar cluster overlap takes place in $\hat{p}(1,2)$ but JE is less affected, as the overlap area in the $\hat{p}(1,2)$ is a smaller fraction of the total area of the 2D plane, whereas the overlap area in $\hat{p}(2)$ is more significant considering the support of the entire $\hat{p}(2)$. In addition, because the movement of clusters in the JPDF has an additional degree of freedom (vertical direction), there can be configurations where overlap of modes appears only in the marginal PDF. Such a case would manifest if the circular (green) feature in image 1 was re-coloured to a value closer to the one of the background (brown). This would result in the corresponding cluster (green/orange) to move up in the vertical direction and the overlap in $\hat{p}(1,2)$ ($\log$-version) would fully vanish. However, the above change would not result in any change in $\hat{p}(2)$ and the overlap would be still present.

To conclude, the maximization of $\hat{h}(Z)$ in MI favours highly varying solutions which can be different from the true solution Z. In this case JE is superior to MI, as the former does not depend on $\hat{h}(Z)$ and hence does not bias the gray values in the pre-described manner.

7.3.3 JE vs MI: Image pair {1,3}

Image pair $\{1,3\}$ corresponds to the case where x_{ref} contains features not present in $x^{\star}$. In this case we show that JE displays a level of structural invariance which can prove both beneficial and disadvantageous in a reconstruction regime. Apparently the removal of the smaller feature in image 3 does not lead to a removal of a cluster in $\hat{p}(1,3)$. As a rule of thumb for these trivial cases, the number of clusters in a JPDF is always equal or greater to number of different gray values in the more varying of the two images - in this case image 1. It becomes greater when there is partial overlap between features (see image pair $\{1,5\}$). The above only holds in these non realistic cases, as the difference between the gray values in the most varying image, is large enough to ensure that they don't contribute to the same cluster, thus there is one distinct cluster for every gray value.

In $\{1,3\}$ there are now two features in image 1 which overlap with the background of image 3. This results in a characteristic alignment of the corresponding clusters in $\hat{p}(1,3)$, with respect to the horizontal direction. We should emphasize that as image 1 is constant, clusters can only move horizontally. Because the number and amplitude of clusters in $\hat{p}(1,3)$ do not change compared to the previous cases and because cluster overlap is minimal, then $\hat{h}(1,3) = \hat{h}(1,1)$ (see Table 7.1). Thus, while using the incorrect image 1 as a prior, the correct solution $Z = 3$ could still be obtained (as a feasible solution) - along with the incorrect $Z = 1$ and $Z = 2$ which return the same JE. Hence, as all solutions $Z = 1|2|3$ correspond to equivalent levels of JE, then when considering an image reconstruction scheme it is up to the data fit term (data likelihood) to select one of the solutions Z. That solution would be the one which maximally satisfies the measured data. Apparently, as JE regularizes the likelihood term by reducing the set of feasible solutions, so does the likelihood term to JE.

On the contrary, MI is greatly altered due to the large change in $\hat{p}(3)$, which reflects the removal of the feature. Two of the modes have now fully merged, creating a dominant mode with high probability. The reduction in uncertainty in image 3 results in a reduced $\hat{h}(3)$ which increases -MI. Hence, MI does not exhibit any invariance in cases such as this.

We mentioned that the structural invariance of JE can also be disadvantageous. Consider the case where image 1 reflects both $x^{\star}$ and prior x_{ref} - this is again the case of a correct prior. Assume also that due to ill-posedness $x^{\star}$ cannot be correctly reconstructed and that the best retrieved estimate $\hat{x}$ resembles image 3. In this case, the extra feature in image 1 is wrongly missing from image 3. However, we have already established that, if image 1 is acting as a prior, theoretically it cannot enforce the reconstruction of the extra feature in image 3 due to the aforementioned structural invariance. This is an important limitation of the JE.

On the contrary, MI is not prone to this problematic case as all features in the prior are strongly enforced to the solution, due to maximization of the marginal entropy term. There is no structural invariance to MI. Unfortunately the latter is very ill-behaved and as we said it can induce further variation in the reconstructed image, such as extra variance or the emphasis of artefacts which appear due to noise. Both behaviours are once again demonstrated in practice in the case study of Sec. 7.5 and specifically in Fig. 7.12. We will revisit these topics later.

Finally, we should emphasize that the structural invariance of JE is only partial. For example see Fig. 7.2 and the considered image pairs $\{1,1\}$, $\{1,2\}$ as well as $\{3,3\}$ and $\{3,4\}$. JE is the same for image pairs $\{1,1\}$, $\{1,2\}$ for the reasons already discussed (corresponds to $\{1,1\}$-$\{1,3\}$ of Fig. 7.1). Consider now $\{3,3\}$-$\{3,4\}$. Image 4 is created by removing a feature from image 3. However, the gray value of the removed feature is also assigned to another part of the image which is not affected by the removal. Apparently the removal of the rightmost feature creates an extra cluster in the JPDF[1] and increases the joint entropy. Information theoretic functionals - especially under the i.i.d. assumption - treat features with equal or similar gray values as one, even when these features are located in different parts of the image. Although such features are not proximal in a spatial sense, they contribute to the same clusters of the JPDF. Hence, by changing one part of the image, it is possible to bias spatially distant parts just because they are populated with values similar to the altered part. Once again, we have observed this in practice and it is discussed in Case III of Sec. 7.5.2 as well as in Fig. 7.8 where the corresponding JPDF are depicted.

7.3.4 JE vs MI: Image pair {1,4}

Image pair $\{1,4\}$ highlights the above findings using the extreme case of a homogeneous solution. As now the solution Z is homogeneous, all gray values pairs formed between the images share the common gray value of Z, which causes all clusters in the $\hat{p}(1,4)$ to vertically align. Again, $\hat{p}(1,4)$ has three clusters, minimally overlapping hence $\hat{h}(1,4) = \hat{h}(1,1)$.

Another important detail regarding JE is revealed. The vertical alignment of the clusters maximally reduces their in-between distance and consequently their overlap. This is more noticeable when kernels of substantial width are used in the KDE. Two clusters with substantial overlap can resemble a single cluster, hence JE is reduced. Thus there is the possibility that the JE between a x_{ref} and an incorrect 'homogeneous' solution x can score a smaller value than the correct solution which resembles the prior. We note that this behaviour is due to approximation of the gray value pairs as 2D Gaussians of substantial width. Theoretically this does not reflect reality as the Gaussians should resemble 2D Dirac delta functions.

The preference of homogeneous solutions by JE has also been observed in practice. Over-regularization with JE can lead to the removal of features from the solution in order to favour the formation of dominant clusters in the JPDF. The removed image features are usually the ones not strongly supported by the data. High values of τ can break any resistance posed by the likelihood regarding

[1] In $\{3,3\}$ we had two clusters corresponding to the superimposed features (circle/circle),(background/background). In $\{3,4\}$ we still have those but also (circle-background)

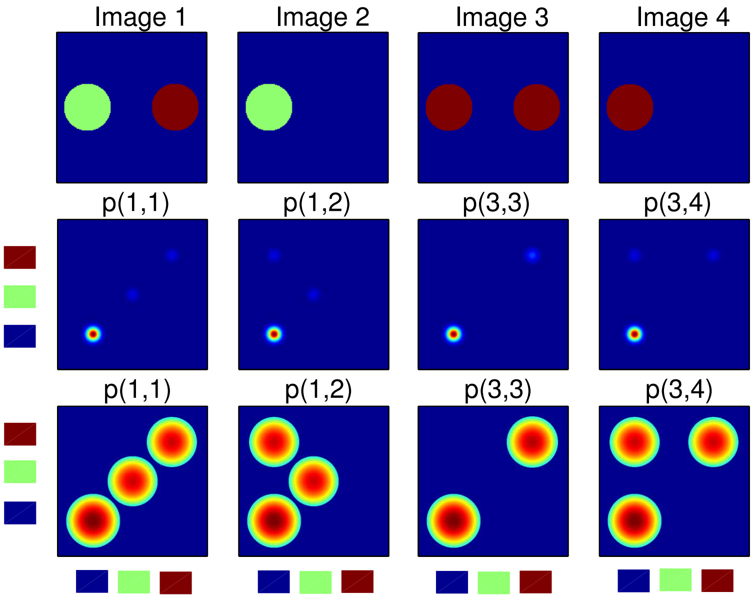


Figure 7.2: Demonstration of JE partial structural invariance. The removal of pair in $\{1, 2\}$ is similar to the $\{1, 3\}$ in Fig. 7.1 and preserves the number of clusters in the JPDF. However, the removal of a non-unique feature in $\{3, 4\}$ does alter the JPDF. The y-axis in the JPDFs correspond to the first (unchanged) image in each pair.

the removal of features, hence the effect is observed. One again, the above effect will be demonstrated in practice in the case study 7.5 and specifically to Figs. 7.12a & 7.12c, which depict reconstructions regularized by JE for variable τ.

Regarding -MI, it is evident (Table 7.1) that its marginal term $\hat{h}(4)$ attains its global minimum for this case ($\hat{p}(4)$ now contains a single mode, although a narrower mode would result in an even lower value) which results in a great reduction in MI. In this case the global minimum of MI=0 is attained. Another minimum proximal to the global one is realized when every pixel in Z is assigned with a different gray value or equivalently when $\hat{p}(Z)$ tends to be uniform (the corresponding image is not depicted).

7.3.5 JE vs MI: Image pair {1,5}

Finally image pair $\{1, 5\}$ corresponds to the case where the solution Z has features not present in the prior space. The minimal number of clusters in $\hat{p}(1, 5)$ is now determined by Z, however an additional cluster manifests due to the partial overlap present in this case. By examining the values in Table 7.1 it is apparent that this case does not constitute an optimum for any of the functionals.

7.4 KDE binning range/ kernel standard deviation

7.4.1 Binning range

The binning range of the KDE was discussed in Sec. 6.1.1.1. In the context of that discussion, we seek a density estimate from a known sample - for example one retrieved from a Gaussian density. In the reconstruction context and in the case of a joint functional - i.e. the JE or MI, we have two samples. The fixed prior x_{ref} and the solution x which evolves during the iterative optimization. The binning range regarding the gray values of x_{ref} is estimated in a manner similar to Sec. 6.1.1.1. In contrast, the range of x is not known *a priori*, as x is subjected to constant improvement.

We approach this problem by preceding the IT regularized reconstruction with a fixed number of iterations based on a more generic reconstruction scheme utilizing regularization functionals such as first-order Tikhonov (TK$_1$) or total variation (TV). The solution estimate $\hat{x}_{\text{init}}$ obtained by the latter is used as an initial solution estimate for the IT enabled reconstruction. We assume that a significant part of the range $\mathcal{R}\left(\hat{x}\right)$ - where $\hat{x}$ is the solution ultimately obtained by the proposed regularization scheme, is captured by $\mathcal{R}\left(\hat{x}_{\text{init}}\right)$. In order to be able to process potential values outside this range, the final range $\mathcal{R}(\breve{x})$ of the equispaced bins $\breve{x}$ utilized by the KDE, is computed by extending $\mathcal{R}\left(\hat{x}_{\text{init}}\right)$ by a fixed amount c at both its ends. Thus $\mathcal{R}(\breve{x}) = \Big[\min\left(\mathcal{R}\left(\hat{x}_{\text{init}}\right)\right) - c, \max\left(\mathcal{R}\left(\hat{x}_{\text{init}}\right)\right) + c\Big]$. Throughout this work we choose a $c = 1.5\mathcal{R}\left(\hat{x}_{\text{init}}\right)$. We note that the extension of $\mathcal{R}\left(\hat{x}_{\text{init}}\right)$ is required in both directions as the $\hat{x}$ is not strictly positive as x_+, but it is expressed in a logarithmic scale so it is unbounded at both ends.

Finally, in the case that some reconstructed values still fall outside $\mathcal{R}(\breve{x})$, we have facilitated a scheme of dynamic extension of $\mathcal{R}(\breve{x})$ which appends sufficient amount of bins in any needed direction. The newly appended bins have the same size $\Delta\breve{x}$ of the initial $\breve{x}$. The magnitude of the extension is large enough to bring any $x_{\text{out}} \notin \mathcal{R}(\breve{x})$ inside $\mathcal{R}(\breve{x})$, as well as a further $0.5supp(K_u)$ from the outermost x_{out}, for the reasons discussed in Sec. 6.1.1.1.

7.4.2 Kernel standard deviation

The kernel standard deviation u is not considered as an optimized quantity in this work, but it is fixed during initialization. Its value is naturally a function of the gray values in x and x_{ref}, from which we seek to retrieve density estimates. In this work we consider u as a function of the range $\mathcal{R}\left(\hat{x}_{\text{init}}\right)$, specifically as its percentage. Thus $u(c) = c\mathcal{R}\left(\hat{x}_{\text{init}}\right)$ with $c \in (0, 1)$. We have identified an empirical optimum c in a pilot reconstruction - where the true solution is considered known, and we maintain the same choice for all subsequent data sets. To identify the optimum we perform multiple reconstructions from the same data set, for multiple values of c. The obtained reconstructions for each $u(c)$ are compared against the true solution and the value c which returns the most accurate reconstructions is nominated as the optimum and employed in all subsequent studies. We explicitly perform this process in 7.5.2.

7.5 Case study 1: 2D numerical simulation

7.5.1 Simulation description

We test the method on a 2D simulated case. We seek the recovery of the optical absorption and reduced scattering coefficient parameter distributions, denoted by μ_a and μ_s' respectively, in a circular object of 25mm diameter. As a light transport model we employ the DA of the RTE which is expressed in terms of μ_a and the diffusion coefficient κ where κ is computed from Eq. 3.24. We simulate the probed anatomy by mapping the target optical distributions on the nodes of an unstructured mesh, consisting of 7261 10-noded triangles and a total of 32,971 nodes. The target optical quantities are depicted in Figure 7.3.

The trans-illumination of the medium and the data acquisition process is simulated from 32 sources and 32 detectors arranged on the boundary at equidistant angular spacing, with sources and detectors being interlaced. The sources are amplitude modulated at a frequency of 100MHz. The sources are activated sequentially. For each source, we solve the frequency domain DA using the TOAST FEM implementation. For each source, the exitance at the boundary is measured by all 32 detectors. Considering the full set of 32 sources, this leads to a total of 1024 detected signals. The full set of collected data constitutes the simulated measured data $\acute{y}$. The measurable quantity $\acute{y}$ at the boundary consists of the logarithmically transformed amplitude as well as the phase of the radiation exiting the medium. The acquired $\acute{y}$ is finally contaminated with 1% of multiplicative Gaussian noise.

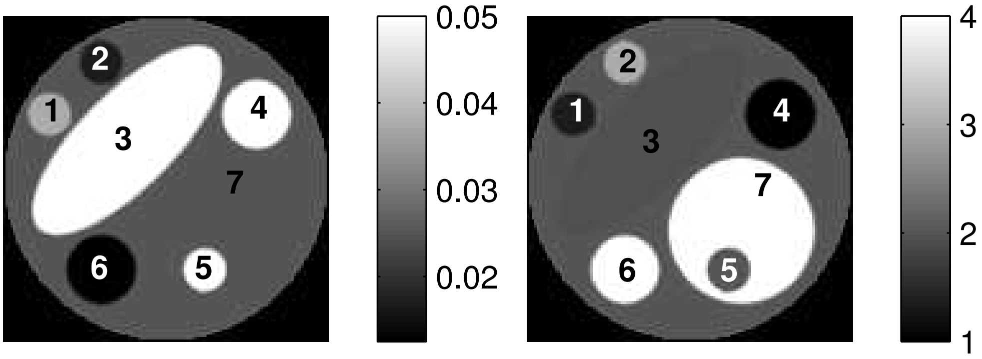


Figure 7.3: Target μ_a and μ_s' images for 2D simulations. The optical coefficients for the labeled regions are:
μ_a - **1**: 0.037, **2**: 0.0167, **[3,4,5]**: 0.05, **6**: 0.0125, **7**: 0.025 mm^{-1}
μ_s' - **1**: 1.33, **2**: 3, **3**: 1.975, **4**: 1, **5**: 2, **[6,7]**:2 mm^{-1}

The reconstruction process is based on the iterative non-linear CG scheme method, which starts by some initial guess $x^{(0)}$ and is then successively improved until pre-defined convergence criteria are met. In this case, the initial guess $x^{(0)} = [{\mu_a}^{(0)}, {\mu_s'}^{(0)}]$ is the result of 10 iterations of CG minimization of Eq. 7.2 with $\Psi(x)$ being a TK$_1$ penalty with $\tau = 1e-4$. The initial guess for the initial TK$_1$ reconstruction is homogeneous for both optical quantities, populated with the target background parameters $\mu_a = 0.025$ and $\mu_s = 2mm^{-1}$.

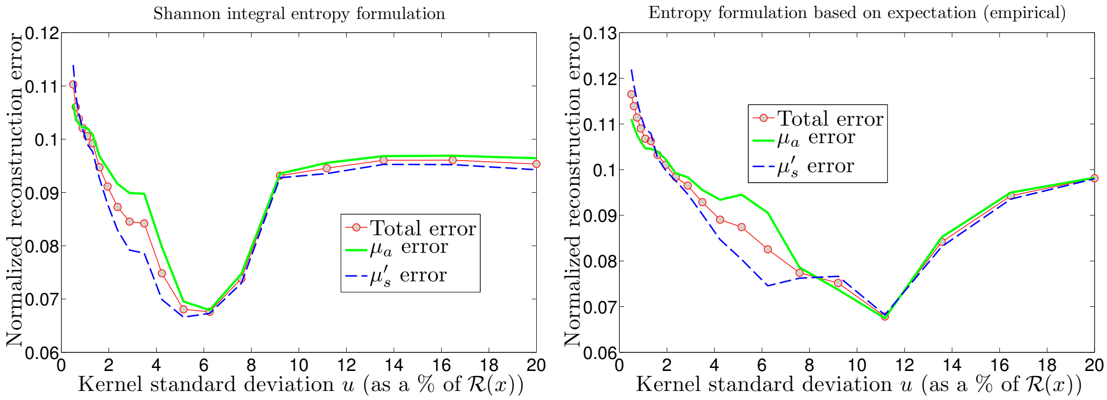
The FEM mesh used for solving the DA during the reconstruction process is different from the one used in the simulation of the data acquisition process. It consists of 3511 nodes and 6840 three-noded triangles, using linear shape functions. Adopting different meshes for the generation of the initial data and during the reconstruction process ensures that the simulation does not involve an inverse crime. The reconstruction is performed on a 128x128 solution basis with square pixels.

7.5.2 Comparison between Shannon and empirical entropy implementation

Before we proceed with testing the capacity of JE and MI to regularize DOT with different priors, we have to choose between the Shannon and empirical entropy formulations introduced in Chapter 6. We execute the forward problem on the software phantom, in order to simulate the data acquisition process. We subsequently create 20 different data sets by contaminating the measured data with 20 different realizations of 1% Gaussian multiplicative noise. For each data set we perform 20 JE-based reconstructions based on the Shannon and empirical entropy formulations for variable kernel standard deviations u. The x_{ref} introduced by the JE for this test is shown in the third column of Fig. 7.5. We compare the accuracy of the reconstructions with the ground truth. Executing the comparison over a wide range of u is essential, as the two formulations perform optimally at different values (see Sec. 6.1.4). A common $\tau = 0.02$ was employed for both formulations. The error between reconstructions $\hat{x}$ and true solution $x^\star$ was measured by the normalized L_2 metric.

$$L_2(\hat{x}, x^\star) = \frac{1}{2}\left(\frac{\|\mu_a - \mu_a^\star\|}{\|\mu_a^\star\|} + \frac{\|\mu_s' - \mu_s'^\star\|}{\|\mu_s'^\star\|}\right) \tag{7.17}$$

Figure 7.4 shows the error plots for the Shannon and empirical entropy formulation. The Shannon formulation *marginally* outperforms the empirical formulation when compared at their absolute minima of 0.6761nats and 0.06783 nats respectively. Although the the number of reconstructions (20) might not be large enough to credit statistical significance to the study, it is a strong indicator regarding the validity of the findings of Sec. 6.1.4, that is, at their optimum u both formulations perform comparably. Henceforth, we employ the Shannon formulation due to its better run-time performance in evaluating the JE and its derivative between images (see Sec 6.4.4).



(a) **(b)**

Figure 7.4: Comparison between Shannon & empirical entropy estimators for variable kernel width u^{opt}. Normalized reconstruction error for **Subfig.** 7.4a Shannon formulation **Subfig.** 7.4b Empirical (expectation based) formulation. Minima at: Shannon 0.6761nats Empirical: 0.06783 nats

Case study 1 - continued Figure 7.5 introduces the $x^\star$ and four different pairs of x_{ref} (columns 2-4) one for μ_a and one for μ_s', all with incommensurate gray values in relation to the $x^\star$. The reconstructed region consists only of the circular domain and the sources/detectors have been placed on its boundary. The remaining region is masked out and serves only for visualization purposes.

It is important to comment on the structure of the x_{ref} as they are designed to assess the capacity of both JE and MI for various cases.

- Case I. The first set consists of the perfect x_{ref} in structural terms as it shares the exact number of features with $x^\star$ and with the exact number of distinct gray values. It should be emphasized that when multiple features in $x^\star$ share the same gray value, their corresponding features in x_{ref} share also a common gray value.

- Case II. The x_{ref} of the second set have been created by summing the $x_{\text{ref}}^{\mu_a}$ and $x_{\text{ref}}^{\mu_s'}$ of Case I, adding an additional feature on $x_{\text{ref}}^{\mu_a}$ and re-scaling the gray values arbitrarily between 1 and 3.8 for both images. The significance of this reference set is as follows. Firstly the prior space has features which do not appear in the target space. Second, due to the way that $x_{\text{ref}}^{\mu_a}$ and $x_{\text{ref}}^{\mu_s'}$ are created, more regions share similar gray values [1,2,3,5,6], hence they are less distinguishable by the KDE. This is the main reason that a continuous KDE is needed, so that the entropic estimates and their derivatives are continuous and enables them to resolve regions with similar values.

- Case III. The third x_{ref} is missing features which are found in $x^\star$. The purpose of this test is to access if the inaccurate x_{ref} suppress the reconstruction of the true features.

- Case IV The final set consists of non-piecewise constant x_{ref}. The are replaced by x_{ref} with a generic structure based on the ones of Case I, but with gray values changing according to a radial gradient. To complicate things further we add 5% of multiplicative Gaussian noise to the prior image. The noise was simulated via

$$x_{\text{ref}} = (x_{\text{ref}})_{\not\eta} + 5\% \times |x_{\text{ref}}| \times \eta, \tag{7.18}$$

where $(x_{\text{ref}})_{\not\eta}$ is the reference image, represented as a $N \times 1$ vector, prior to noise-contamination; and η is a $N \times 1$ vector consisting of random trials drawn from a Gaussian density $\mathcal{N}(\mu, \sigma^2)$, with mean $\mu = 0$ and standard deviation $\sigma = 0.05$.

Figures 7.6 and 7.7 present the obtained reconstructions for both μ_a and μ_s' respectively. Each figure in columns 2-4 show the reconstructed images obtained using both JE (top row) and MI (bottom row), using the corresponding image pairs. We have used the same kernel width u (optimum from Sec. 7.5.2) for both JE and MI as they both computed using the Shannon entropy implementation. In the first column we present the initial guess used for the IT reconstruction and also a converged reconstruction using TK$_1$ regularization to be used for comparison purposes. Figures 7.9 and 7.10 depict the profiles of the images at y-axis coordinates 40 or 100 depending on which is more revealing, taking into account the features in the x_{ref} and $x^\star$ spaces. The formed JPDFs are depicted in Fig. 7.8. Note for example that the added noise in the fourth prior is reflected in the profile (gray/dashed line). A discussion of the obtained results now follows.

Regarding Case I, it is apparent that both JE and MI perform well. Consistent to the our discussion in Sec. 7.3, we emphasize that the MI reconstruction of μ_s' exhibits higher variation in the background region - which should be homogeneous, possibly due to the maximization of $\hat{h}(x)$ term.

Case II is far more revealing about the superiority of JE as it reflects on the theory in Sec. 7.3. We observe that the extra features in the reference images leave a distinct footprint in the MI based reconstructions. The fact that the extra features in x_{ref} appear with distinct gray values in $\hat{x}$ is due to the effect of the marginal entropy term $\hat{h}x$ discussed in Sec. 7.3. Apparently, MI based priors wrongly bias the solution in the case of extra features in the prior space. In contrast, in the case of JE, the enforced blob can attain the gray value of the background so effectively it disappears (vertical alignment in the JPDF as discussed in Fig. 7.1). This is a significant advantage of JE over MI. The effects of the extra feature in x_{ref} can be observed in the JPDFs in Fig. 7.8.

In Case III we observe that the features in $x^\star$ which do not have a corresponding feature in the reference image are successfully reconstructed, although not as resolved as in cases I and II. However, the accuracy of the reconstruction regarding these features is better than the corresponding one in the TK_1 reconstructions, which should be expected as TK_1 smooths out the region borders corresponding to high gradients. Another possible explanation is that by providing accurate prior support to the rest of the domain, the overall reconstruction is improved. Thus the modelling of the overall light propagation becomes more accurate and as a by-product, it leads to the improved recovery of the features which miss explicit prior support. Hence, their improved spatial resolution compared to TK_1 can be considered as a by-product of the overall improvement of the reconstruction. It should be noted that the IT functionals are not easily predictable in their behavior. By adding or removing a blob in the prior space, the effect in the reconstruction is not necessarily constrained in the region of the missing/extra blob. These changes propagate through the PDF/JPDF and affect spatially distant regions in the reconstruction, although their gray values are proximal in the PDF/JPDF space. One can easily see that the removal of blob [4] affects the contrast of the reconstructed blob [1]. This is a potential drawback of the IT functionals. This effect is observable in the JPDFs in Fig. 7.8.

Finally, case IV shows that even with the added complexity induced by the gradient/noise in the prior space, both functionals display increased invariance to the complex incommensurate gray value relationship between x and x_{ref}, producing acceptable reconstructions.

Tables 7.2 and 7.3 show the normalized Sobolev norm distance [Terzopoulos, 1986]

$$\|x - x^\star\|^2_{S,2} = 0.5\frac{\|x - x^\star\|^2_2}{\|x^\star\|^2_2} + 0.5\frac{\|\nabla(x - x^\star)\|^2_2}{\|\nabla x^\star\|^2_2}, \tag{7.19}$$

evaluated between the target and reconstructed images for μ_a and μ'_s respectively. It is apparent that in all cases JE is better in quantitative and qualitative terms. Its superiority is also evident in the profiles of Figs. 7.9 & 7.10.

Table 7.2: Sobolev norm distance between target and reconstructed absorption distributions

Recon.	TK1	1	2	3	4
JE	0.44	0.15	0.22	0.29	0.22
MI	0.44	0.20	0.29	0.31	0.30

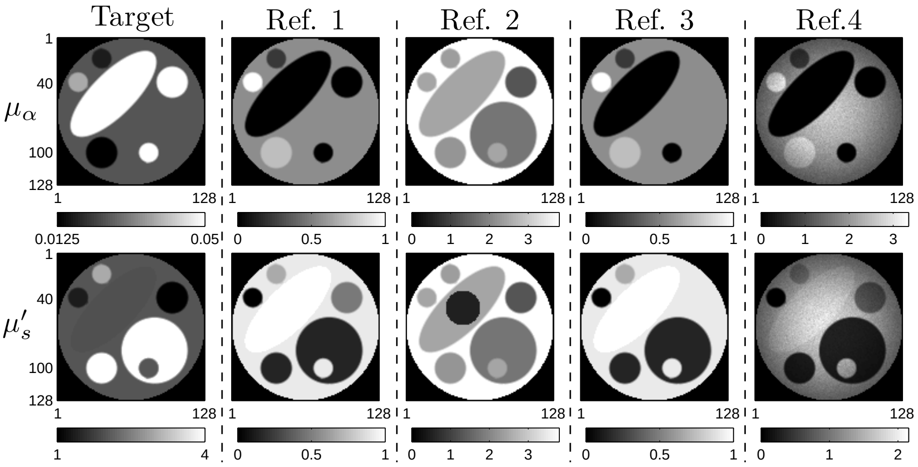


Figure 7.5: Target distributions and the 4 reference images pairs, incommensurately related to the target gray values. Ref 1 displays full correspondence between its features and the ones in the target distributions. Ref. 2 contains features not existing in the target space. Ref. 3 is missing features. The gradient in Ref. 4 is enforced by centring a 2D Gaussian (σ: 50 pixels) on top of Ref. 1 and multiplying the pixels values underneath. We also add 5% Gaussian multiplicative noise.

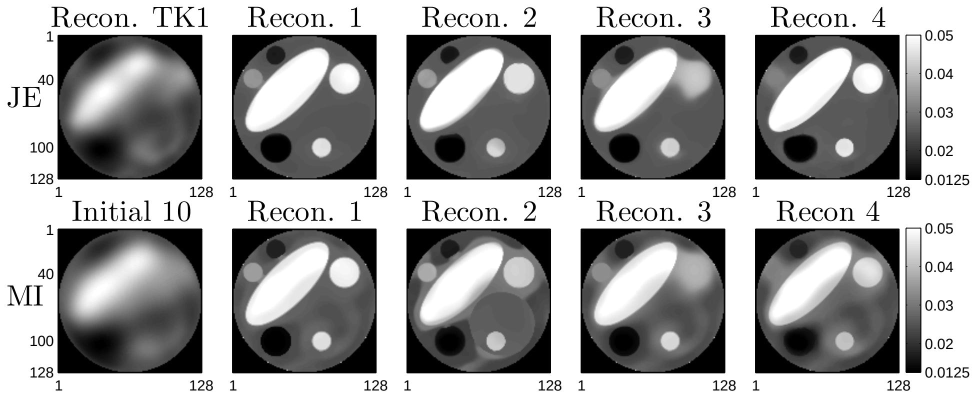


Figure 7.6: μ_a reconstructions by introducing the available reference image pairs with joint entropy or mutual information. The converged TK1 reconstructions are provided for comparison along with the initialization guess.

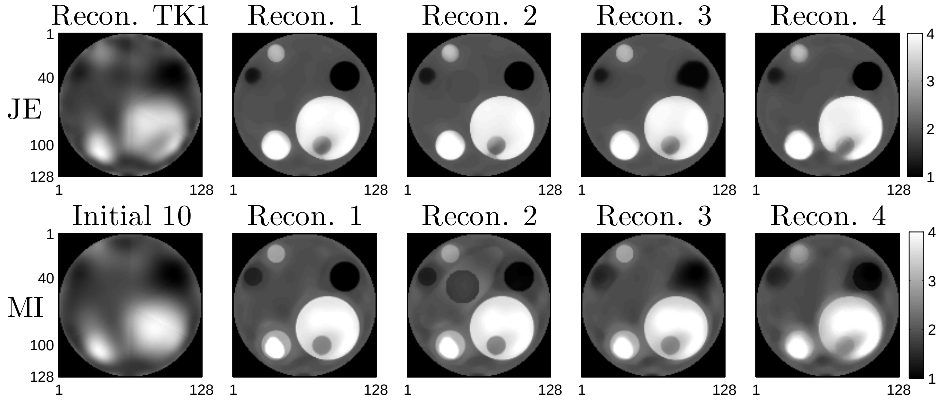
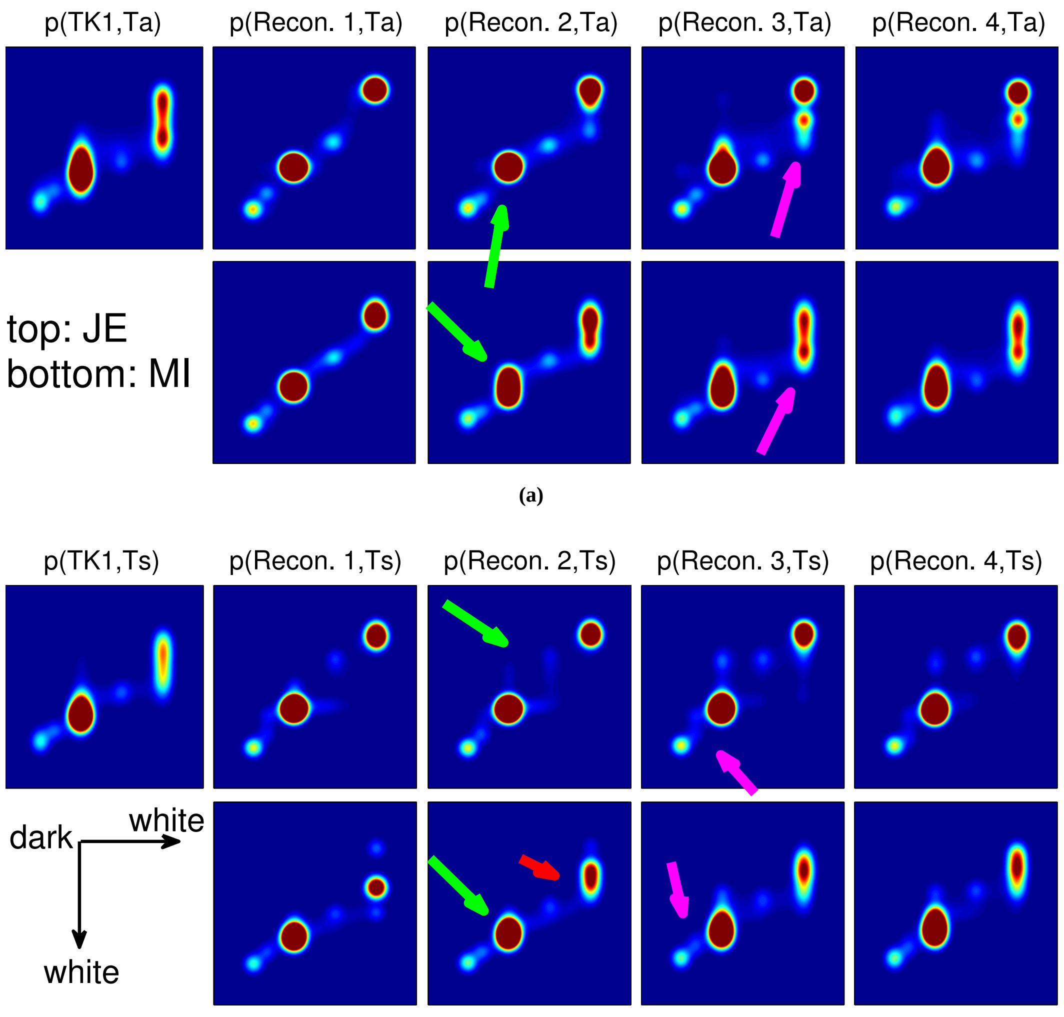


Figure 7.7: μ_s' reconstructions by introducing the available reference image pairs with joint entropy and mutual information.



(a)

(b)

Figure 7.8: Effect of information theoretic regularization on JPDFs. **Subfig. 7.8a:** μ_a. Depiction of the JPDFs for the case of μ_a. Visualization scale cut-off threshold t applied to reveal all clusters. All values $> t$ are shown as brown. For all JPDFs; x-axis: prior, y-axis: Recons. 3^{rd} *col.*: Clusters indicate value of region [7]. JE cluster unaffected. MI cluster shows dispersion in y-axis (disordered background). This increases $\hat{h}$(Recon.2) and reduces $-MI$. 4^{rd} *col.*: Arrows point to the cluster of white regions [3,4,5]. Tail corresponds to the reconstructed blob 4 lucking explicit prior support. JE preserves dominant top cluster corresponding to [3,5]. MI does not, hence [3,5] are affected - see increased variance in [3]. **Subfig. 7.8b:** μ_s'. 3^{rd}: JE mostly unaffected. Few new values created due to minimal bias from extra prior feature [7]. MI - dispersion (overall disorder in [3 + gray background]) and significant change due to strong bias from extra feature (small arrow) 4^{rd} *col.*: Arrows point to value of the black blob lucking support. IMPORTANT: notice that the removal of the blob [4] from the prior, affects its reconstruction but also the reconstruction of blob [1] in JE and more in MI. REASON: Blobs: [1,4] are spatially distant but close in JPDFs. **Overall** usage of priors cluster JPDFs compared to TK1 case, which is interpreted as improved contrast and resolution.

Table 7.3: Sobolev norm distance between target and reconstructed scattering distributions

Recon.	TK1	1	2	3	4
JE	0.42	0.23	0.23	0.24	0.27
MI	0.42	0.296	0.28	0.30	0.34

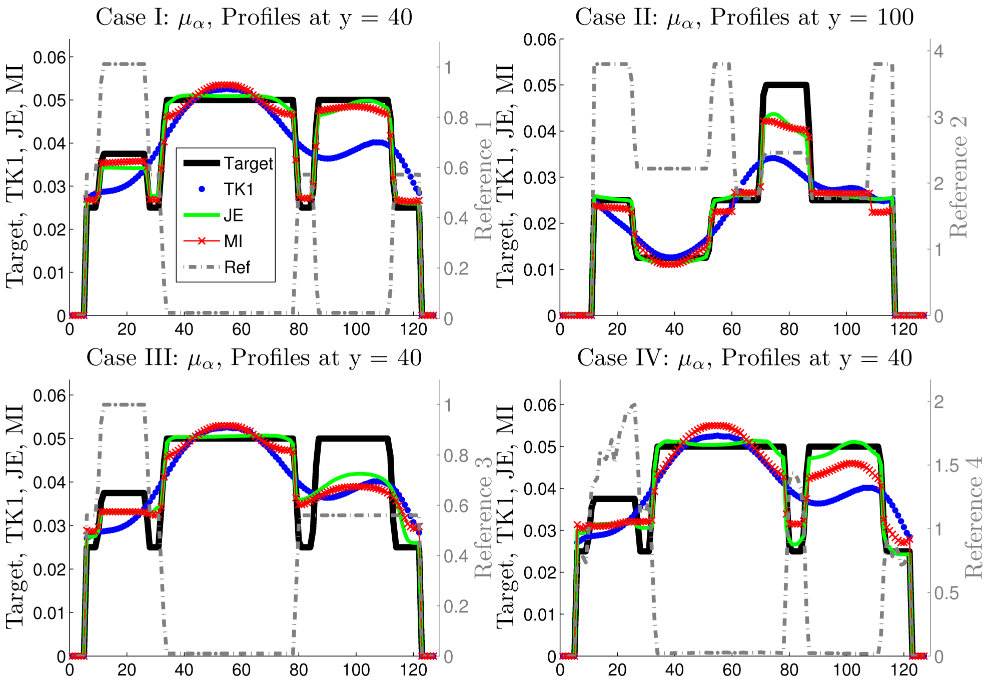


Figure 7.9: Profiles for μ_a at $y = 40$ or $y = 100$ for reconstructions using TK1, JE and MI. We also provide the profiles for $x^\star$ and x_{ref}. The choice of y targets the profiling of different features and was made taking into account the feature correspondence between $x^\star$ and x_{ref} space for each case.

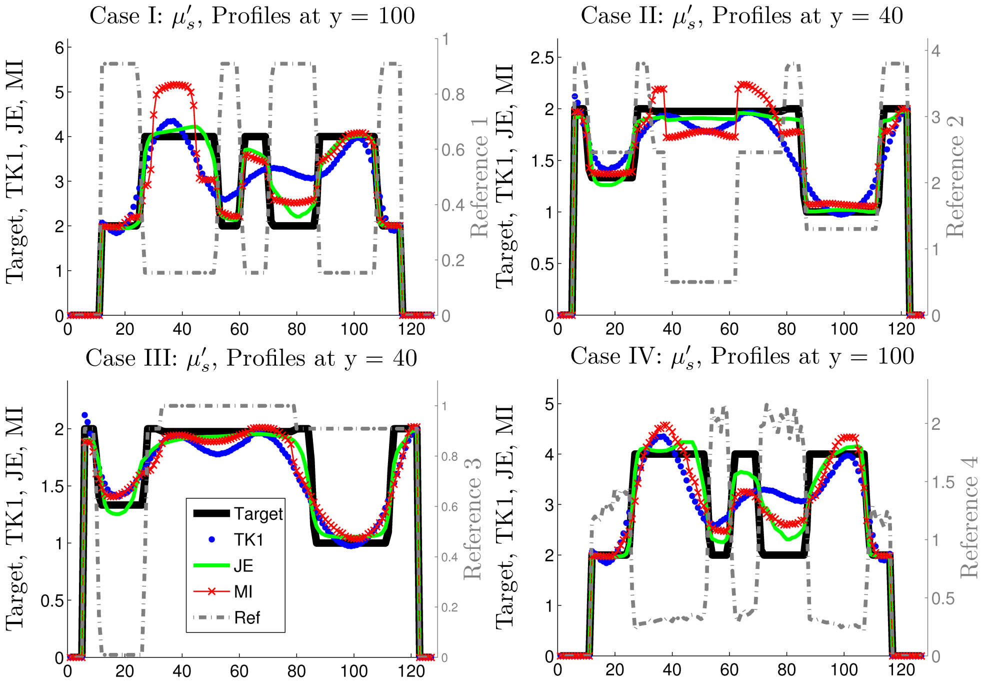
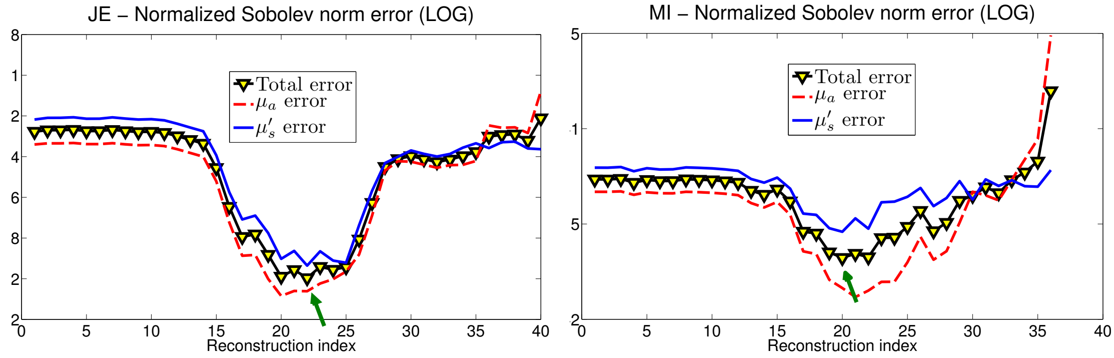


Figure 7.10: Profiles for μ_s' at $y = 40$ or $y = 100$.

In order to fairly compare MI and JE, each was weighted by their individually computed optimum τ, in order to ensure that the observed biasing tendencies were not a product of unfair weighting of one of the methods. This ensures that the results reflect the theoretical capacity of each method and no unfair under- or over- regularization of one of the methods is taking place. For both methods and for each different reference image pair, we performed multiple reconstructions using the same data but variable $\tau \in \left[10^{-4}, 10\right]$, ranging in 40 logarithmically spaced intervals and the τ which returned the best match for each case was adopted.

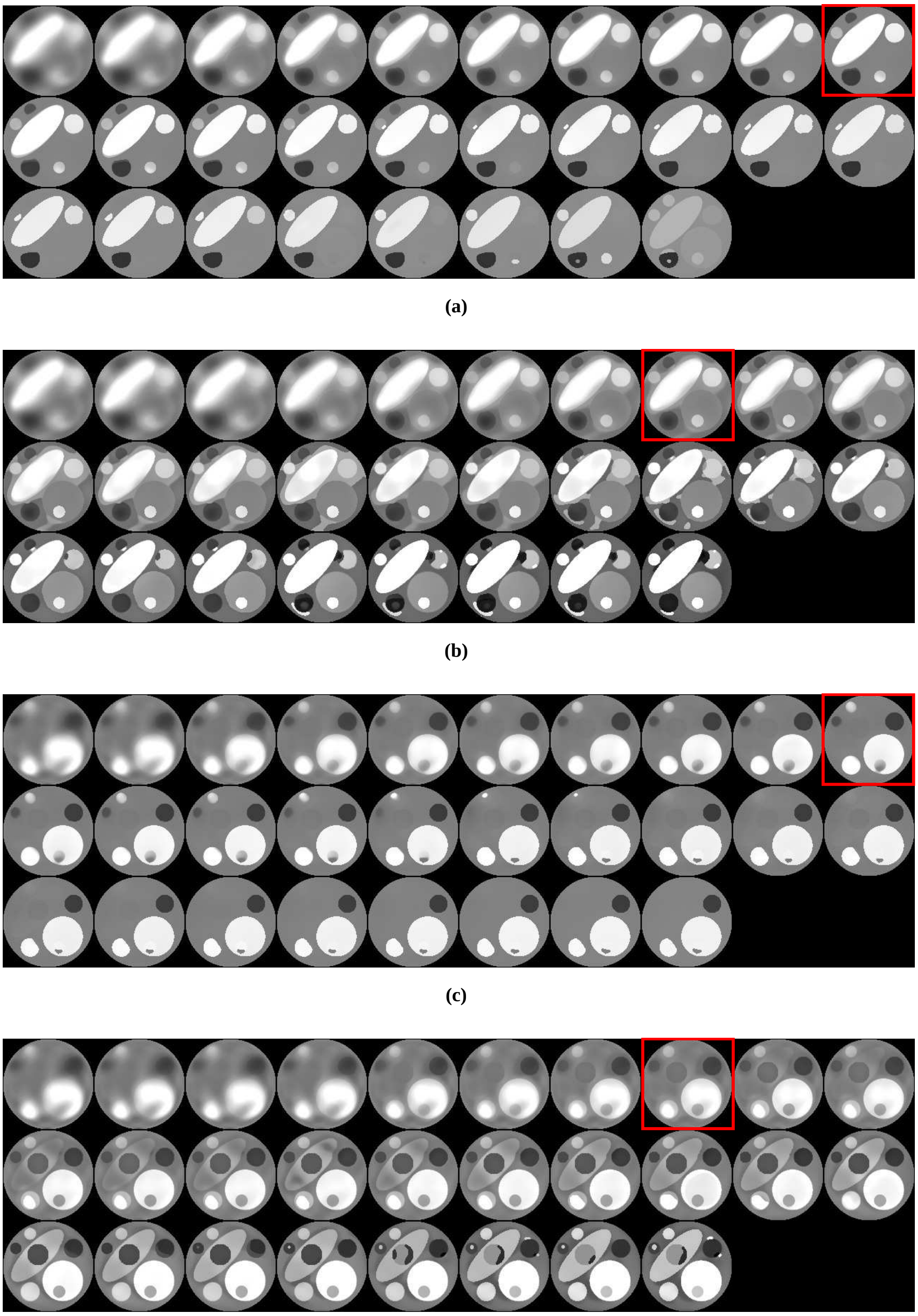
Figure 7.12 shows the $\hat{\mu}_a$ and $\hat{\mu}'_s$ reconstructions obtained using prior pair #2 of Fig. 7.5, for the variable τ and for both JE and MI. The corresponding error plots for the same prior case, are shown in Fig 7.11 as an indicator. The reconstructions corresponding to the minima of the plotted errors are highlighted in Fig. 7.12.

One can observe that JE regularization removes features as τ increases whereas MI fully enforces the prior x_{ref}. This is due to the absence and presence of the marginal term in the two methods. In the JE case, the removal of features leads to vertical alignment of the clusters in the JPDFs (not depicted), their in-between distance is minimized and JE decreases. MI enforces all features (maximum amount of clusters in the PDF/JPDF) and induces higher contrast among features to increase the distance among clusters. The final JE reconstruction in Fig. 7.12a is consistent to the theory. Although all features appear, the global contrast is significantly reduced. The reduction in the contrast among the depicted features - except feature 6 (black circle) - significantly clusters the JPDF (not depicted for this case) thus reducing JE. We have observed in other studies which we have performed (not included here) that for $\tau \gg 0$, the reconstructions approach the homogeneous image.



(a) **(b)**

Figure 7.11: Reconstruction errors (against true solution) for the images depicted in Fig. 7.12. MI errors corresponding to indices $36 \longrightarrow 40$ have been truncated as the correspond to very high values and compromise the clarity of the rest of the plot. Minima are explicitly indicated



(a)

(b)

(c)

(d)

Figure 7.12: Reconstructions using prior pair #2 for 40 log-spaced τ values. Recons. 1 → 12 have been removed (under-regularized, solely data-driven). They are similar to the first two depicted recons. Minimum error recons. have been highlighted. **Subfig.** 7.12a JE-μ_a **Subfig.** 7.12b MI-μ_a **Subfig.** 7.12c JE-μ_s' **Subfig.** 7.12d MI-μ_s'.

7.6 Case study 2: 2D MRI-derived target, MRI prior

7.6.1 Simulation description

This study corresponds to a more realistic case. The structure of the simulated target anatomy is provided by a coronal slice of a 3D MRI scan of the breast. A simple segmentation process classified the tissue to two different types - adipose and glandular. The identified regions were re-assigned with optically meaningful values - although the choice was arbitrary and does not reflect the literature reported optical values of the considered tissue types. In addition, an artificial (non-MRI derived) elliptical perturbation was added to the $\mu_a^\star$ to simulate a tumour and was assigned with a value $6\times$ higher that the background adipose tissue and $1.5\times$ higher than the glandular tissue. Tumours have higher absorption levels due to extra vascularization. The resulting optical image plays the role of the probed medium. The target optical values are summarized in Table 7.4.

Table 7.4: Case study 2: Target optical values. The values are in mm^{-1}. The have been arbitrarily selected and they do not reflect literature reported optical values of the considered tissue types.

Tissue type	μ_a	μ'_s
glandular	0.04	2
adipose	0.01	1
tumour	0.06	-

The MRI providing the structure of $x^\star$ will be used as the prior x_{ref} image. This case is more complicated compared to the previously presented case study for the following reasons. The MRI is not piecewise constant - whereas the target optical image is - and it contains micro-features and noise which is also absent from the target. In addition, the explicitly added feature in $x^\star$ does not have a corresponding entry in x_{ref}, hence it does not benefit from explicit prior support. We thus test the flexibility of the methods to reconstruct it. To complicate things even further we add a circular feature on the MRI x_{ref}. We saw in Sec. 7.3 (discussion regarding image pair $\{1, 3\}$) that partial overlaps such as the ones induced by the added feature, compromise the JE structural invariance. Finally, the regions comprising the target image are non-convex, thus creating small structural details which their retrieval is an added challenge for the relatively low resolution of DOT.

7.6.2 Meshes and reconstruction initialization

We utilize the same meshes of the Case study 1 (see Sec. 7.5). Similar to that case, we ensure not to commit an inverse crime by using different meshes for forward data generation and also for the reconstruction process. Finally, 1% of Gaussian distributed random noise was added to the simulated data.

The initialization guess $\hat{x}_{\text{init}}$ used in the IT reconstruction scheme was provided by a converged TV-regularised solution $\hat{x}_{TV}$. This reconstruction will also be used in order to evaluate the potential achieved improvement by the informational theoretic regularisation scheme. Figure 7.13 presents the true solution $x^\star$, prior x_{ref}, the initialization estimate $\hat{x}_{\text{init}}$ - obtained by a converged TV reconstruction - as well as the formed PDFs $\hat{p}(\hat{x}_{\text{init}}, x_{\text{ref}})$. The above quantities are presented for both μ_a and μ_s'. All reconstructions presented in this case are visualized using the colour range of the target $x^\star$. The magnitude of the challenge is evident. The initial estimate, especially in the case of μ_a fails completely to reproduce structural detail. The simulated tumour is reconstructed with lower contrast and an additional feature - not corresponding to the true solution also appears. The normalized L_2 error between the retrieved $\hat{\mu}_a$ and the true $\mu_a^\star$ is 22.01%. The scattering reconstruction is marginally better resulting to an L_2 error of 20.75%.

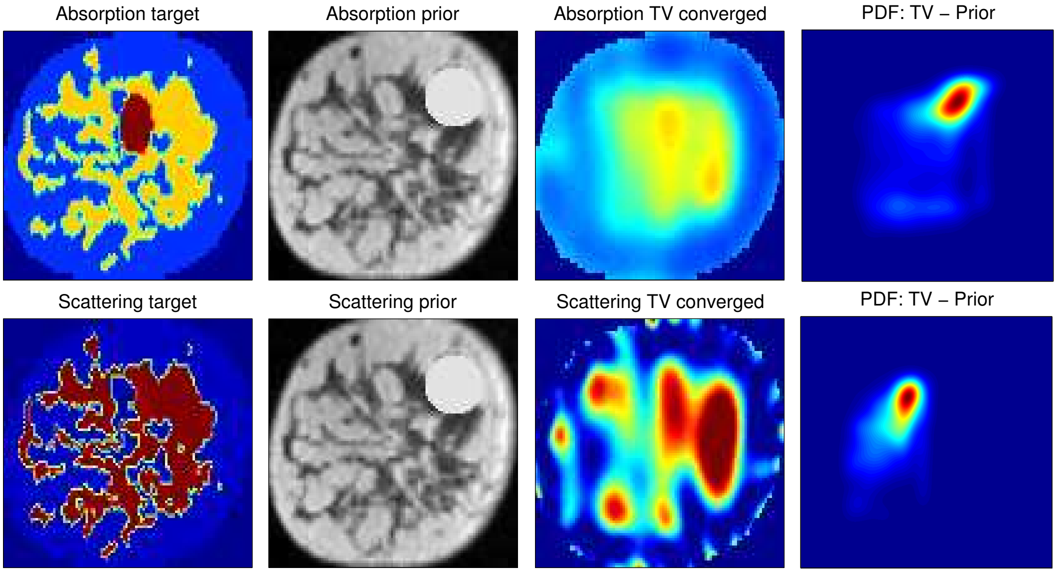


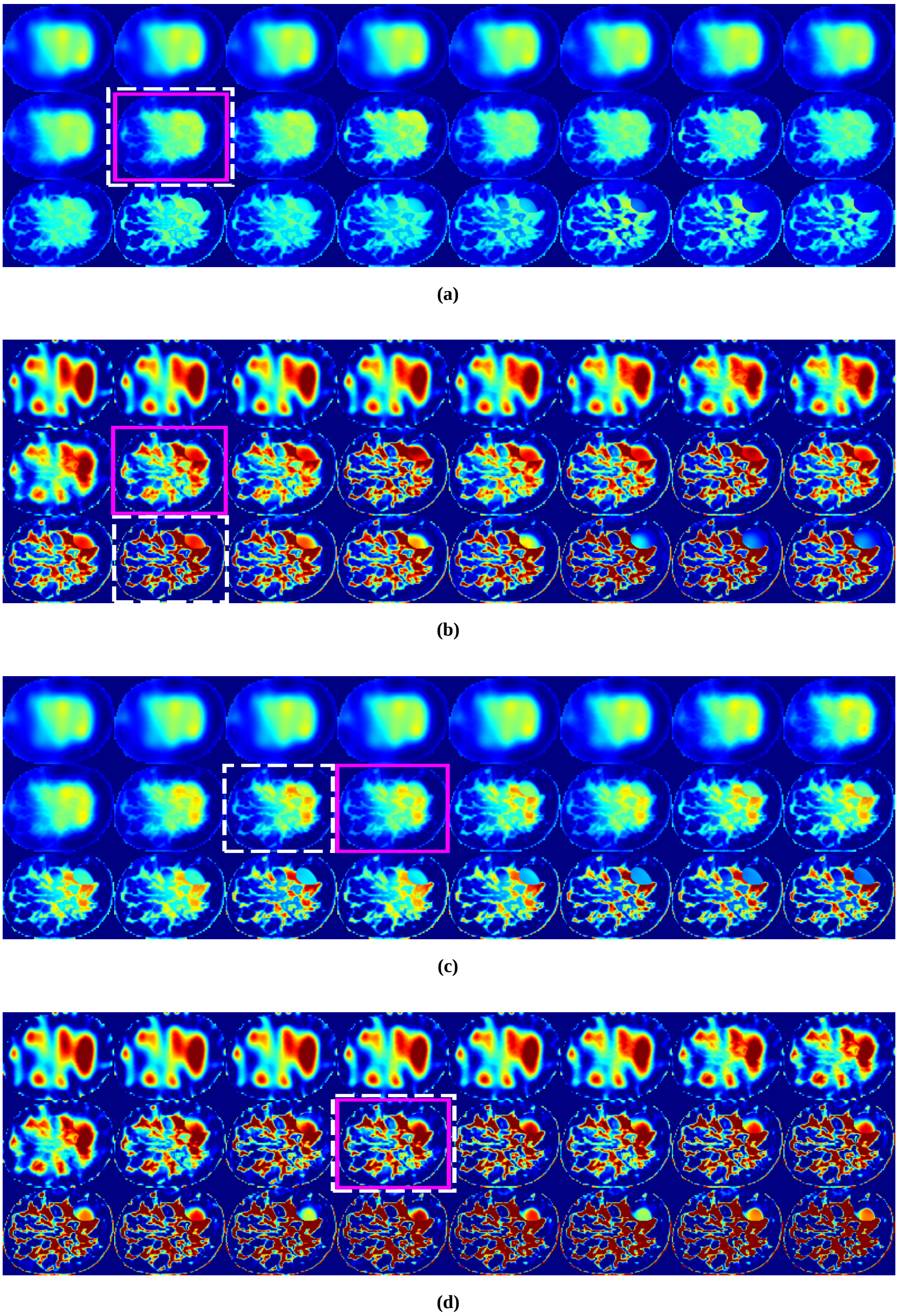
Figure 7.13: Case study 2: Initial images. **Top row:** Absorption **Bottom row:** Scattering **1st col. :** Targets **2nd col. :** MRI-priors **3rd col. :** TV-converged reconstruction $\hat{x}_{TV}$ - used also as the initial estimate $\hat{x}_{\text{init}}$ **4th col. :** JPDFs between $\hat{x}_{TV}$ and x_{ref}

7.6.3 Information theoretic regularization

Similar to the previous case, we have performed multiple JE and MI reconstructions with variable log-spaced τ, similarly to Case 1. These reconstructions are depicted in Fig. 7.14. The top two subfigures correspond to μ_a and μ_s' for the case of JE, whereas the bottom two show μ_a and μ_s' for the case of MI. In every image, the dashed box indicates the reconstruction which matches best (in L_2 norm terms) the target solution for the depicted optical quantity. Thus the optimum μ_a and μ_s' reconstruction can be retrieved for different values of τ. The continuous box indicate the reconstruction which minimizes the combined μ_a and μ_s' error and it is this which is ultimately chosen.

In Subfig.7.14a (JE-μ_a) we once again observe that JE reduces contrast as τ increases. This is not observed in the scattering reconstructions. A potential explanation is that the initial μ_s' obtained from the TV reconstruction is more accurate than the corresponding μ_a initialization. In the case of MI contrast progressively increases as expected from the theory.

The optimal reconstructions corresponding to the minimum combined L_2 error are depicted in Fig. 7.15. There is an evident improvement in the reconstructions, especially in the case of μ_s'. We mentioned



(a)

(b)

(c)

(d)

Figure 7.14: Reconstructions using 40 log-spaced τ values. **Top to bottom:** JE-Absorption, JE-Scattering, MI-Absorption, MI-Scattering. **Continuous boxes** indicate reconstruction corresponding to minimum total L_2 error whereas **dashed boxes** indicate recons. with minimum error for that case (JE/MI, μ_a/μ_s').

that this can be the case due to the better initial μ'_s guess. Another possible reason for the superiority of the retrieved μ'_s over μ_a is inadequate normalization between the two, during the optimization. Thus, μ'_s can potentially be subjected to higher regularization than μ_a. Such problems do not occur in alternative modalities which retrieve a single image.

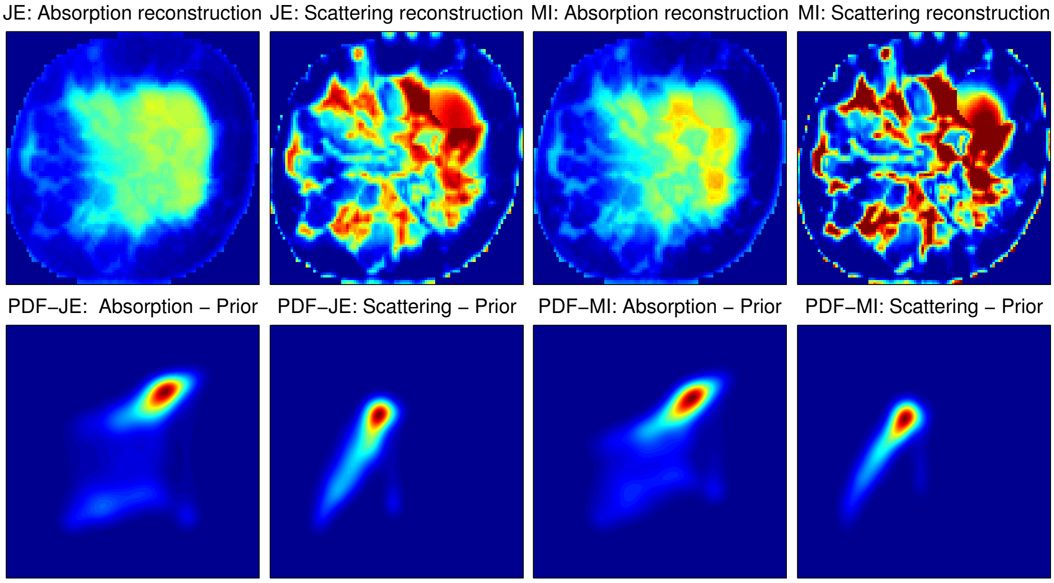

JE: Absorption reconstruction
JE: Scattering reconstruction
MI: Absorption reconstruction
MI: Scattering reconstruction
PDF–JE: Absorption – Prior
PDF–JE: Scattering – Prior
PDF–MI: Absorption – Prior
PDF–MI: Scattering – Prior

Figure 7.15: Case study 2: Reconstructed images. **Top row:** Reconstructed images. **From left to right:** JE-Absorption, JE-Scattering, MI-Absorption, MI-Scattering **Bottom row:** Final JPDFs. **From left to right:** The reconstructed images (image above each PDF) vs the prior x_{ref}.

Regarding the comparison between JE and MI, it is apparent that MI promotes the formation of structure more strongly due to its marginal term, which is optimized when the obtained image exhibits higher variation. The same term promotes higher contrast between the reconstructed values (see discussion in 7.3.2), thus emphasizing all regions which differ from their surroundings. Hence, we observe higher absorption in the tumour region, but for the same reason inaccurate values (centre-right) in the reconstructed image are also emphasized. JE on the other hand, due to the lack of the marginal term does not emphasize neither the artefacts as well as the tumour. Both methods however improve on the initial TV reconstruction.

The normalized L_2 errors between the retrieved μ_a, μ'_s reconstructions against their targets and also the total error are provided in Table 7.5. In this case MI outperforms JE. It should be noted that the best JE-μ'_s reconstruction is not the one indicated in Fig. 7.15 but it occurs for higher τ value (see dashed boxes in 7.14). It results in an error of 7.26% and achieves a reduction of 35% from the initial error - which is higher than the one achieved by MI. However, for the same high τ, the error of the JE enabled μ_a reconstruction is far higher, increasing the total error.

Table 7.5: Case study 2: Normalized L_2 errors

Images considered	Normalized L2 error	Reduction
μ_a: $x^\star$ - $\hat{x}_{TV}$	22.01%	-
μ_a: $x^\star$ - $\hat{x}_{TV}$	11.18%	-
Total initial error	16.60%	-
μ_s': $x^\star$ - $\hat{x}_{JE}$	20.75%	-5.72%
μ_s': $x^\star$ - $\hat{x}_{JE}$	8.11%	-27.47%
Total JE final error	14.43%	-13.07%
μ_s': $x^\star$ - $\hat{x}_{MI}$	19.29%	-12.36%
μ_s': $x^\star$ - $\hat{x}_{MI}$	7.34%	-34.3%
Total MI final error	14.43%	-19.9%

7.7 Case study 3: 3D reconstruction of a software phantom

7.7.1 Simulation description

In this study we test the method in the 3D DOT problem. The 3D case is usually more ill-posed than the 2D setting as the ratio of the number of unknowns (equalling the voxels in the 3D volume) to the acquired data increases. In addition, a significant potion of the light that exits the domain goes undetected, as the detectors are now arranged over a 3D surface and not a 2D boundary.

In this case, the test domain consists of a cylindrical object with radius of $25mm$, and height of $50mm$. Sources and detector sites were placed in 5 rings around the mantle of the cylinder, at elevations -20, 10, 0, 10, 20 from the central plane. 16 sources and 16 detectors were arranged in each ring, totalling a set of 80 sources and 80 detectors. Data was acquired for all source/detector combinations resulting to 6400 acquisition events The domain was represented by a FEM mesh which consisted of 83142 nodes and 444278 4-noded tetrahedral elements. The optical background parameters were set to $\mu_a = 0.01\,mm^{-1}$ and $\mu_s' = 1\,mm^{-1}$. Spherical and elliptical perturbations have also been defined, with increased absorption and scattering coefficients. The arrangement of target objects is shown in Figure 7.16. The measurements consisted of logarithmic amplitude and phase at a source modulation frequency of $f = 100\,$MHz, contaminated with 0.5% of multiplicative Gaussian-distributed random noise.

7.7.2 Reconstruction and initialization

The reconstruction is performed on a $32 \times 32 \times 32$ grid of bilinear voxels. Figure 7.17 shows a converged reconstruction using TV regularization which used a regularization parameter value $\tau = 1e - 5$. The display range in the corresponding images is optimally computed to enhance visual clarity. However, the solutions obtained using the JE and MI priors are displayed in the same range of the true solution. Hence, we also present Figure 7.18 which shows the same reconstruction, but now scaled according to the range of the true solution. The reconstruction is depicted in order to enable visual comparison with the IT reconstructions in order to showcase the magnitude of the improvement. The same reconstruction was used as the initialization guess for the IT reconstructions.

7.7.3 Information theoretic regularization: correct prior

In this case we utilize a prior image x_{ref} which shares the same structure with the true optical solution. However, its gray values are incommensurately related to the true optical parameters. The prior image is depicted in Fig. 7.19. Figures 7.20 & 7.21 present selected reconstructions using JE and MI regularization. The display range of the presented results is that of the true solution for both μ_a and μ_s'. It has to be noted that the selection of the presented images over alternatives obtained for different regularization weighting τ, was based on qualitative criteria (visual inspection). It is evident that both methods improve on the initial TV solution as the correctly retrieve the underlying features of the true solution.

7.7.4 Information theoretic regularization: incorrect prior

In this case we utilize a prior image x_{ref} is not an accurate structural representation of the true solution. It contains extra features with respect to both true $\mu_a^\star$ and $\mu_s'^\star$. As in the previous case, its gray values are incommensurately related to the true optical parameters. The prior image is depicted in Fig. 7.22.

Figures 7.23 & 7.24 present selected reconstructions using JE and MI regularization. The display range of the presented results is that of the true solution for both μ_a and μ_s'. Once again the reconstructions were selected based on visual inspection.

As expected, the JE reconstruction induces less bias due to the extra features in the obtained solution. Neither the ellipsoidal nor the spherical perturbation have a significant impact on the μ_s' and μ_a solutions respectively, from which they should be absent. However, we observe that the small features which do not exist in the $\mu_s'^{\star}$, do appear in the obtained μ_s'. A possible explanation for this result is that these features share similar gray values. According to the discussion of in Sec. 7.3.3, this can compromise the structural invariance of JE.

In the case of MI reconstruction, bias from the extra features is apparent. We note that all the reconstruction which we have examined had noticeable bias - except the ones which were under-regularized and did not significantly improve from the TV reconstruction. For higher values of τ the spatial resolution increased along with the bias.

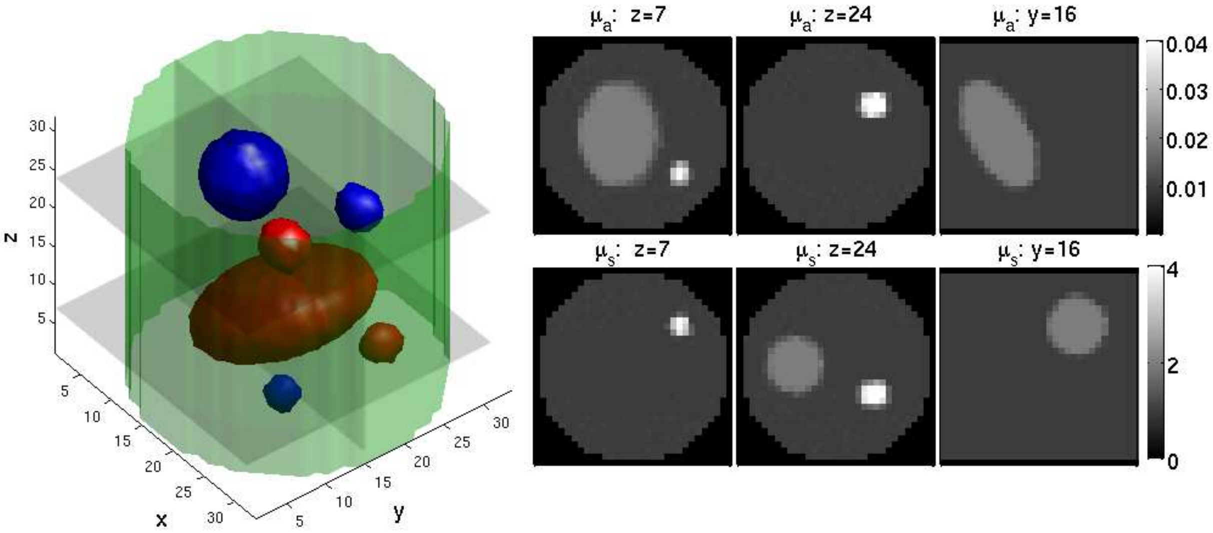


Figure 7.16: Left: Cylindrical software phantom with embedded absorption (red) and scattering (blue) perturbations. The position of the cross-sectional planes used for displaying the reconstruction results is indicated in gray. **Right:** cross sections $z = 7$, $z = 24$ and $y = 16$ through the absorption (top and scattering target. The top three images show μ_a and the bottom three μ_s'.

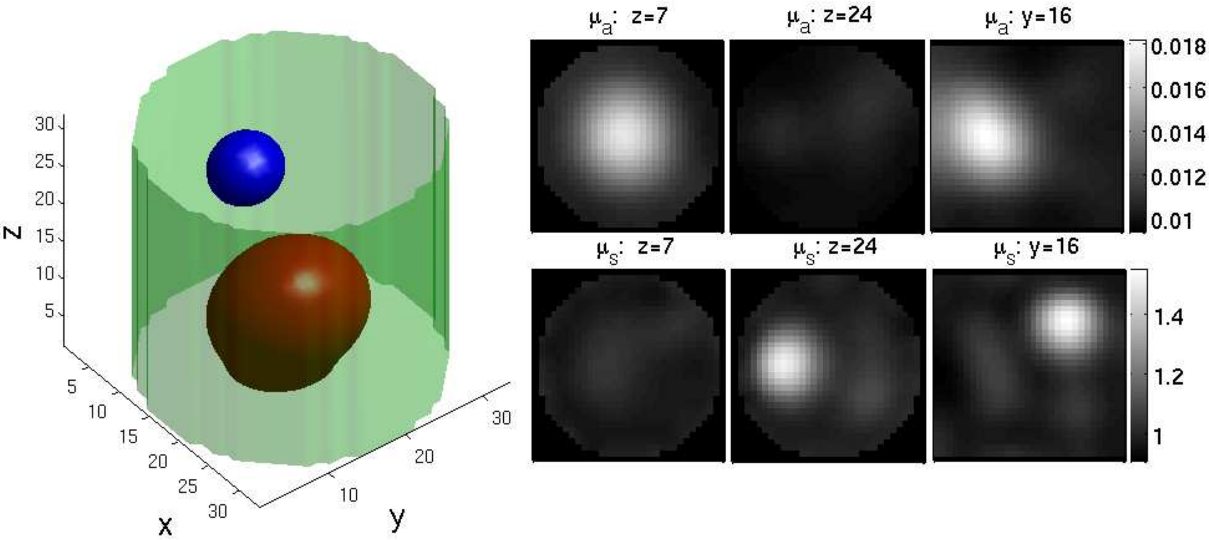


Figure 7.17: Converged TV reconstruction: Visualized using optimal display range to enhance features

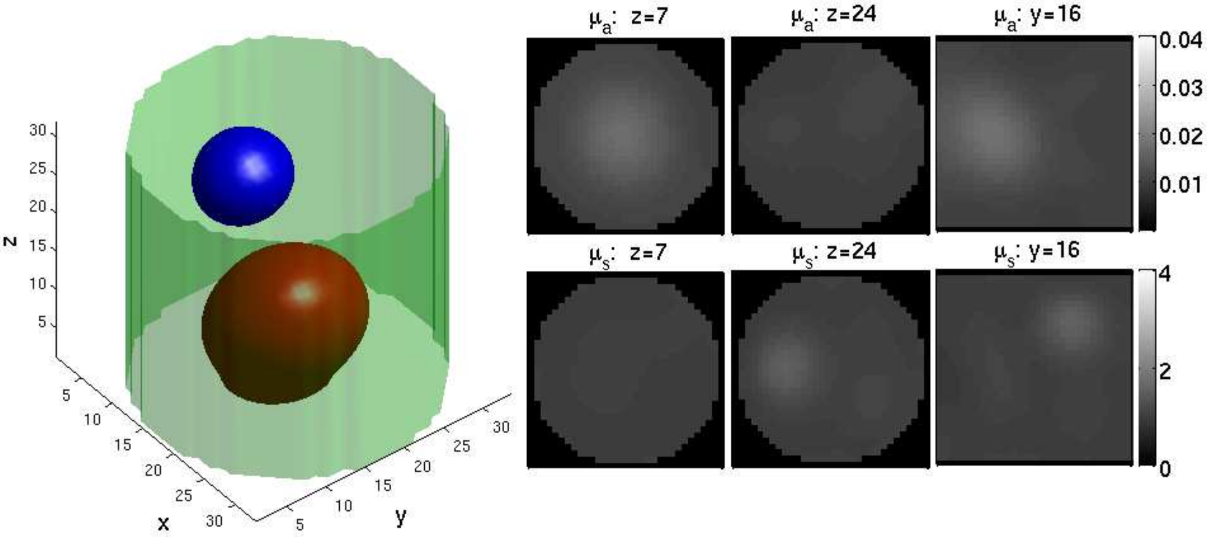


Figure 7.18: Converged TV reconstruction: Visualized using the same display range as the true solution

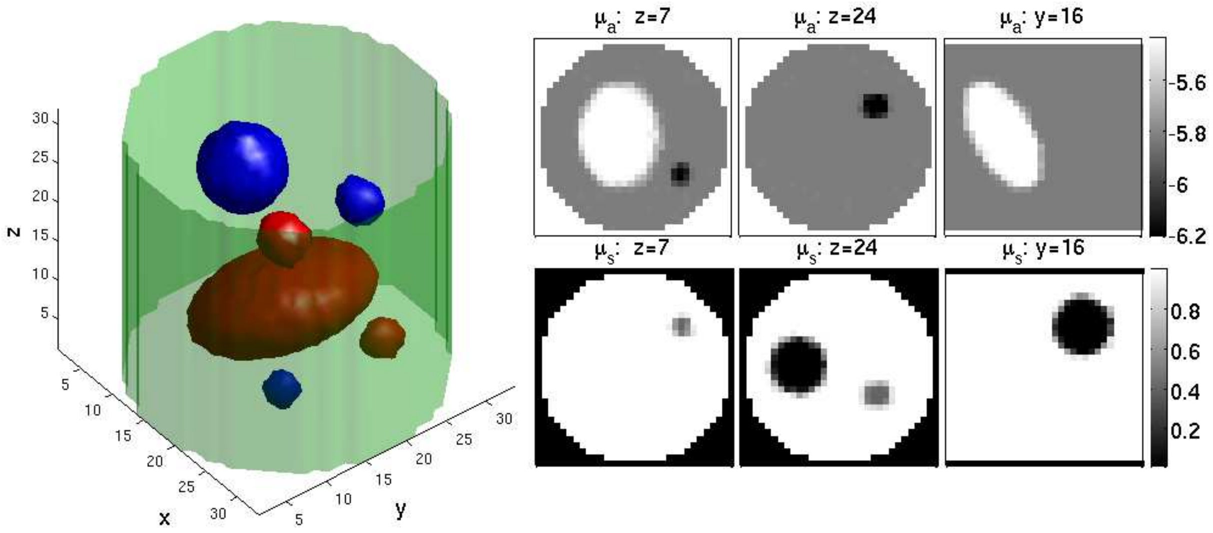


Figure 7.19: Prior image 1: The prior has one-to-one feature correspondence with the true solution. Gray values are incommensurately related to the true solution.

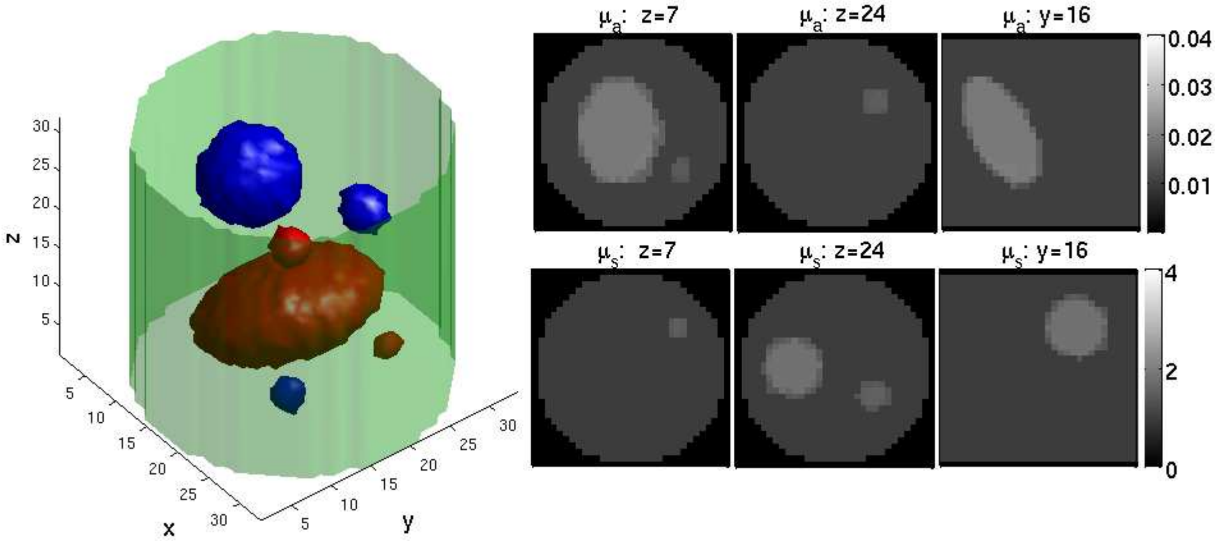


Figure 7.20: Converged JE reconstruction using the structurally correct prior 1

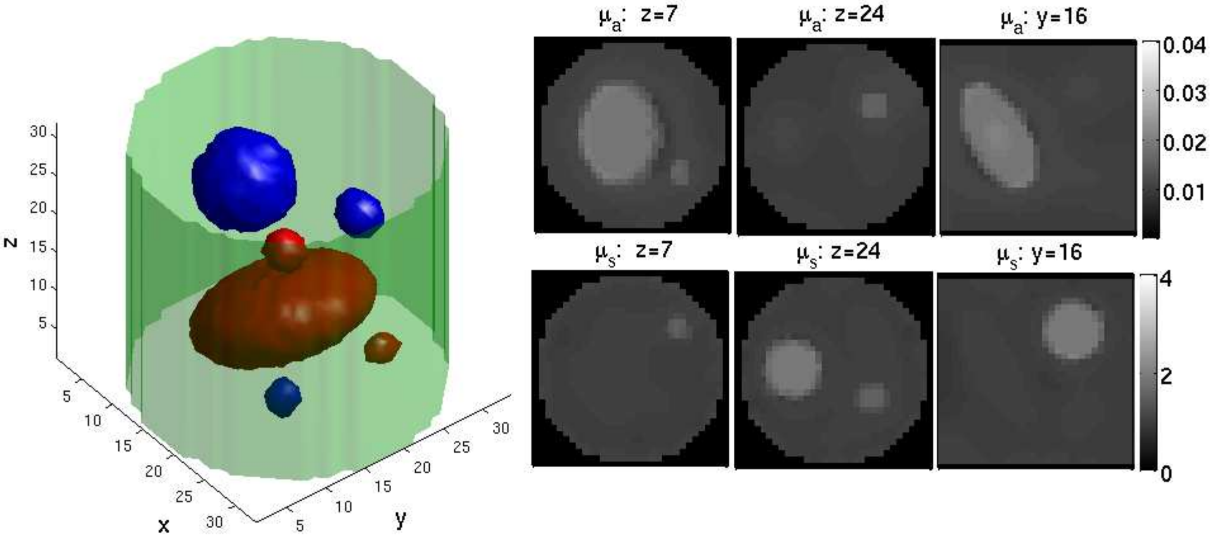


Figure 7.21: Converged MI reconstruction using the structurally correct prior 1

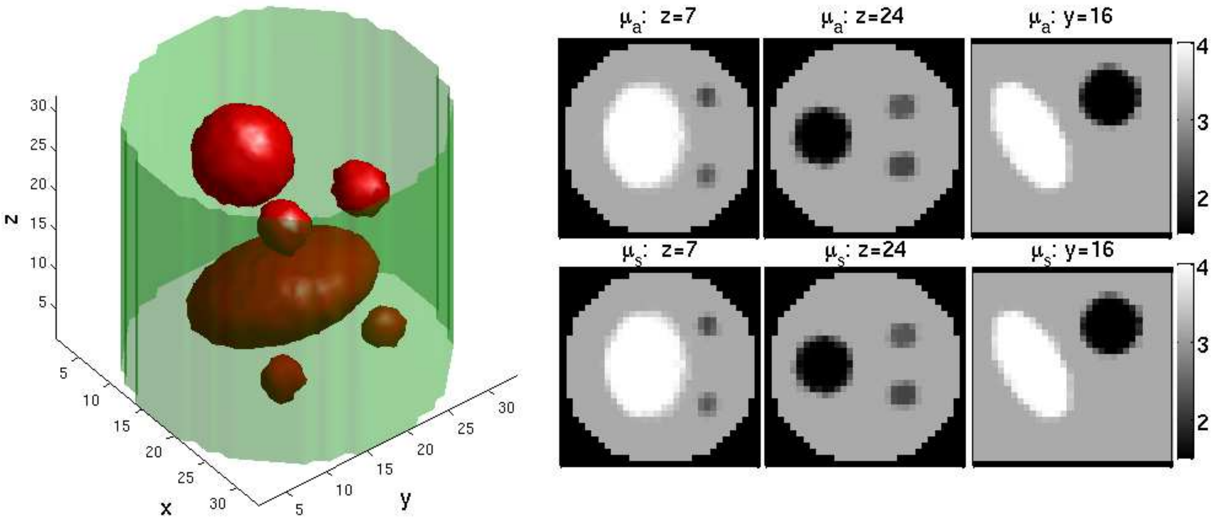


Figure 7.22: Prior image 2: The prior has extra features when considering the true μ_a and true μ_s' solution. Gray values are incommensurately related to the true solution. The displayed isosurface image is not representative of the gray values as the high background value dominates the isosurface rendering. It is solely displayed as an indicator of the locations of the various features.

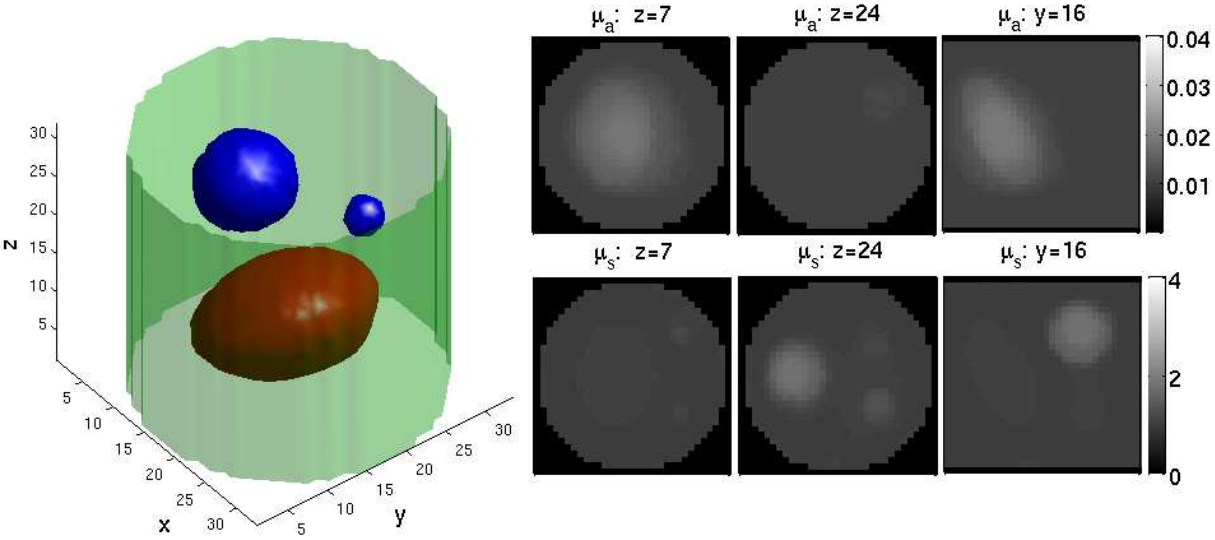


Figure 7.23: Converged JE reconstruction using the structurally incorrect prior 2

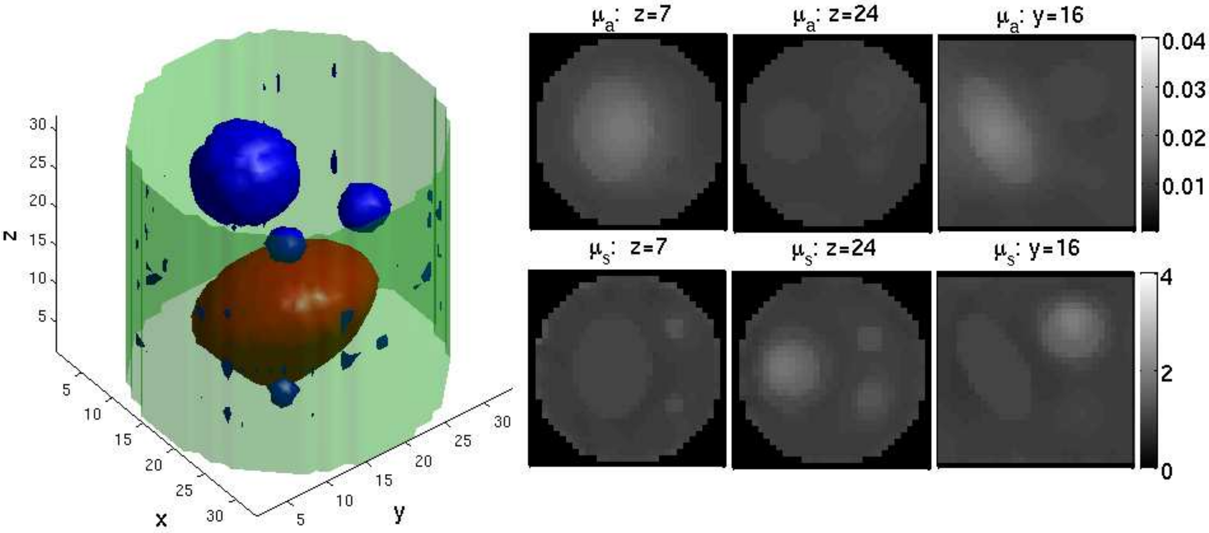


Figure 7.24: Converged MI reconstruction using the structurally incorrect prior 2

7.8 Case study 4: DOT/MRI experimental phantom multi-modality study

In this study we test the IT regularization scheme on experimental data. We seek to retrieve the optical properties of a cylindrical phantom. The phantom is probed by an optical tomography system and is also subjected to an MRI scan. The aim is to reconstruct the optical properties x of the phantom using its MRI representation as the structural prior x_{ref}.

7.8.1 Phantom description

The phantom was constructed[1] by a two-component, room temperature vulcanizing silicone (ELASTOSIL RT 601, Wacker Chemie AG). The cylinder's height was 120mm, whereas its inner diameter was 69mm. Alternative building materials for the phantom include, polysaccharide gels, gelatin, polynivyl alcholol [Mazzara et al., 1996; Ohno et al., 2008; Surry et al., 2004]. Silicone was the material of choice due to its minimal deterioration over time, as well as due to the ease which one can construct regions with different magnetic contrast within it.

Silicone is clear and colourless and its optical properties can be modified by adding scattering and absorbing ingredients. The scatterer used in this instance was TiO_2 powder whereas the absorbing properties were assigned by adding a near infrared (NIR) dye (Pro Jet 900 NP, Avegia Biologics Ltd.) to the silicone. Similar ingredients have been previously used in resin based optical phantoms [Firbank and Delpy, 1993; Firbank et al., 1995]. The dye was first mixed with a small amount of ethanol with the help of an ultrasound cleaner. Then the result was mixed with the silicone in a plastic stirring pot. TiO_2 was mixed with the hardener, again with the help of ultrasonic vibration. The hardener was then added to the stirring pot and mixed carefully to minimise bubble formation. The concentrations of ingredients to achieve approximately desired optical properties were 230.4 mg/dl for TiO_2 and 0.5 mg/dl for the NIR dye. The stirred silicone was poured in a mould made of nylon, which was first treated with silicone spray. The mould consisted of a hollow cylinder, bottom plate and three cylindrical pegs, which leave holes for the perturbations in the cast. The parts were attached to each other with screws.

The cured phantom was removed from the mould. Material for three different perturbations were poured in the holes of the phantom. The first perturbation had double the concentration of TiO_2 compared with the background. The second had double the concentration of the near-infrared dye and a MRI contrast. The third perturbation had only MRI contrast. Silicone itself can be seen in an MRI image and its magnetic properties can be altered with different paramagnetic substances. In this phantom, the MRI contrast was created with Dotarem (Guerbet S.A.), which contains gadoteric acid. Its paramagnetic nature is based on gadolinium. It is used as an MRI contrast agent and normally given by injection. After the perturbations were cured, the holes were filled with silicone with the same optical properties as the background material. The phantom can be seen in Fig. 7.25.

The background optical properties of the phantom were approximately set to $\mu_a = 0.01$ and μ_s' mm^{-1} at a wavelength of 785nm. The base of the three cylindrical cavities were formed on a plane

[1] The phantom was constructed by Atte Lajunen, Department of Biomedical Engineering and Computational Science, Aalto University

parallel to the base of the cylinder and at an elevation of y =56.25mm from the base. The cavities were filled with dyes of different concentration of the background in order to provide optical contrast. Cavity A was filled with double the concentration of the NIR dye aiming to achieve twice the absorption of the background. The aimed optical quantities of the background are $\{2\mu_a, \mu_s\}$. Cavity B had solely MRI contrast, resulting in optical properties $\{\mu_a, \mu_s\}$ same as the background. Finally, cavity C was filled with the scatterer TiO_2 in double concentration of that used for the background, resulting in optical parameters $\{\mu_a, 2\mu_s'\}$ relative to the background. The magnetic properties of cavity C were aimed to be similar to the background ones, thus C should be virtually absent in the corresponding MRI reconstruction.

The internal structure of the phantom as well as its absorption/scattering/magnetic properties are graphically depicted in the schematic of Fig. 7.26. In addition, we provide two sets of MRI images corresponding to *a)* planes parallel to the base of the cylinder, located at elevations y =33.7mm, y =63.7mm & y =78.7mm from the base *b)* planes vertical to the base of the cylinder and in the front-to-back direction, which depict the phantom at various depths, specifically in distances z =19.4mm, z =36.7mm & z =49.6mm from the front.

It should be noted that provided distances are estimates obtained by the information depicted in the full set of slices obtained by the MRI, given that the first slice corresponds to the base and and the last to the top of the cylinder and by utilizing the known height and diameter of the cylinder. For example, we utilize the down-sampled MRI image volume of $32 \times 32 \times 32$ $(\mathsf{x}, \mathsf{y}, \mathsf{z})$ resolution and we then identify the first image slice in the 3D volume stack (starting from the cylinder's base), where the cavities firstly appear. Let the index of that slice be k. We simply compute its distance from the base in mm by $\mathsf{y} = (120/32) * k$. We have applied a similar process for all provided distances. The utilized MRI was down-sampled from its original $256 \times 256 \times 40$ resolution to match the resolution of the optical solution discussed earlier. The reported positions of the considered planes are computed from the down-sampled MRI - as the prior and resulting optical solutions are also represented in this lower resolution.

7.8.2 Data acquisition process

The optical probing of the phantom and the acquisition of the relevant data[1] was obtained using a frequency domain optical tomography system described in [Nissilä et al., 2002; Nissilä et al., 2005]. The wavelength of the incident radiation was set to 785nm. The optical probing was performed by 15 NIR sources and 16 detectors arranged over a two-ring geometry (see Fig. 7.26), with sources and detectors being interlaced and arranged with equidistant angular spacing. The corresponding planes of the two rings were located at elevations of y =65mm and y =70mm from the base of the cylinder. These distances were communicated to us by the scientific staff responsible for the actual measurements.

7.8.3 Reconstruction setting and initialization

The image reconstruction process utilizes an unstructured triangular mesh with 70218 nodes and 26878 elements. The obtained optical solutions are retrieved in the already reported $32 \times 32 \times 32$ grid - which

[1] 1 The data acquisition process was performed by Dr. Ilkka Nissilä, Department of Biomedical Engineering and Computational Science, Aalto University

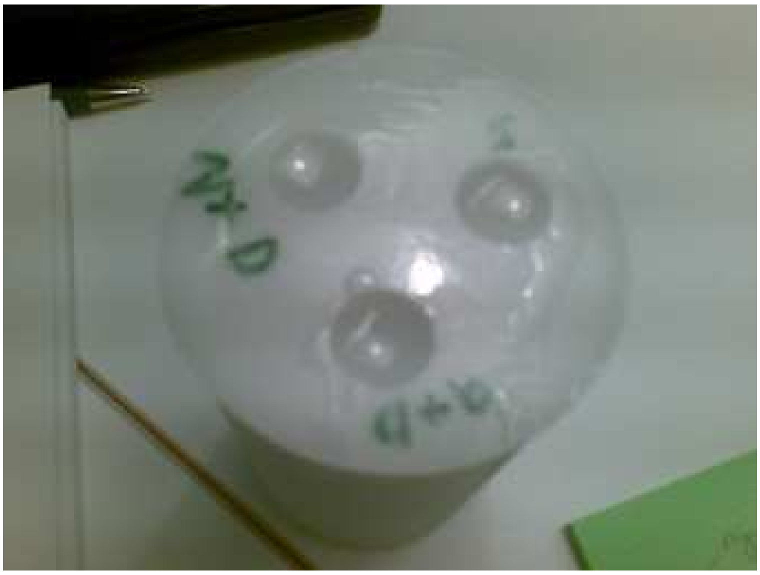


Figure 7.25: Case study 3: Photograph of probed phantom

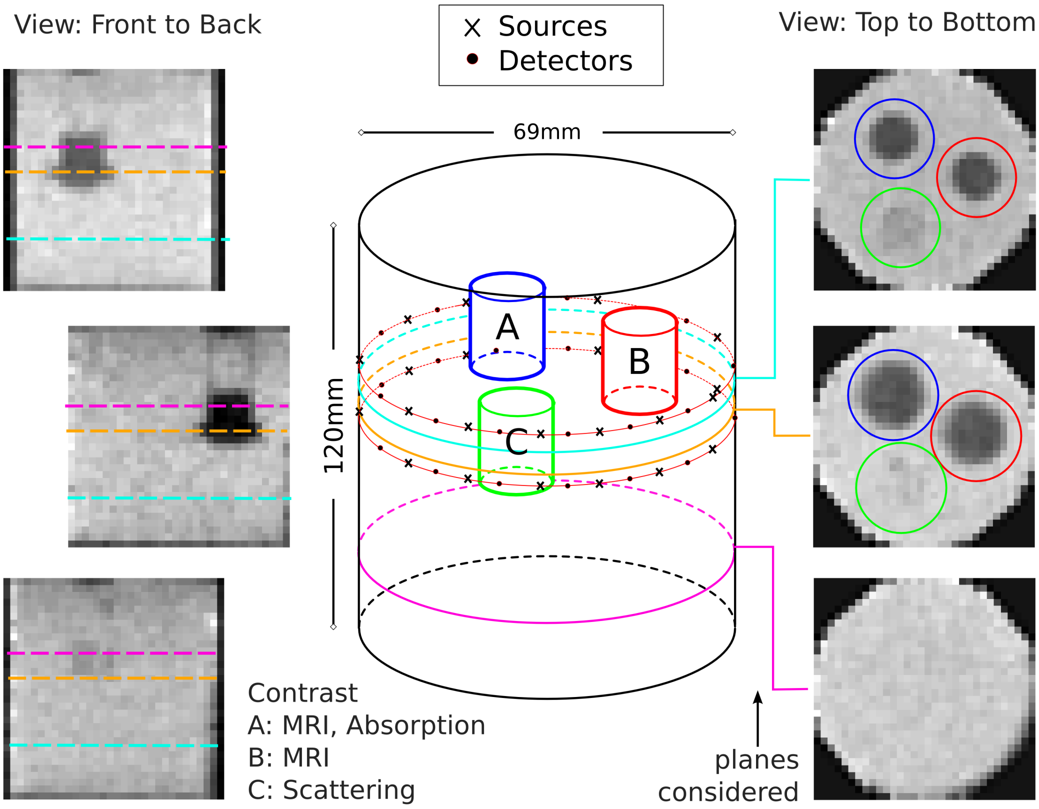


Figure 7.26: Case study 3: Structure of MRI/DOT phantom. **Perturbation A:** i) MR contrast ii) Absorption twice as the background. **Perturbation B:** MR contrast only. **Perturbation C:** i) MR contrast ii) Scattering twice as the background.

also determines the dimensionality of the inverse problem. We firstly perform an initial reconstruction using a converged TV regularization. We then use this retrieved optical solution as the initialization estimate $\hat{x}_{\text{init}}$ for the reconstruction which introduces the MRI prior x_{ref} using the IT functionals. We were given an estimate of the true background optical quantities of the phantom, by the scientific staff responsible for its creation, of $\mu_a = 0.008mm^{-1}$ and $\mu_s' = 0.75mm^{-1}$. These values are employed as the homogeneous initial guess for the TV enabled reconstruction.

We now discuss the obtained results with the considered methods. For each method, all μ_a images are displayed in common scale. Similarly for μ_s'. We note that we show multiple images for each optical quantity, corresponding to the views considered in Fig. 7.26. We also need to note that for some μ_s' reconstructions, there are few pixels in the boundary - close to source positions - which attain significantly higher values than the rest of the image. Specifically these pixels have a value of $1.78mm^{-1}$, whereas the rest of the reconstruction is limited to $1.2mm^{-1}$. We thus impose an upper limit of $1.2mm^{-1}$ to the display range to enhance visual clarity. When scaling has been applied, it is explicitly stated in the caption.

Figure 7.27 showcases the converged TV reconstruction. Regarding the μ_a part, the reconstruction of perturbation A (see Fig. 7.26) is correctly reconstructed - however its spatial resolution is compromised by overestimating its original size. In addition, it is apparent that although perturbations B & C were aimed to be invisible in absorptions terms, this is clearly not the case. Regarding its μ_s' part, the obtained $\hat{x}_{TV}$ is clearly erroneous. Perturbation C which contains a scatterer in $2\times$ the concentration of the background, attains the higher values however its spatial resolution is significantly compromised. In addition, perturbations A & B which were designed to be invisible in μ_s' terms, have a clear footprint in the obtained $\hat{x}_{TV}$.

It is understandable that quantitative analysis of the obtained reconstructions is not trivial in this case as the true solution images are not available.

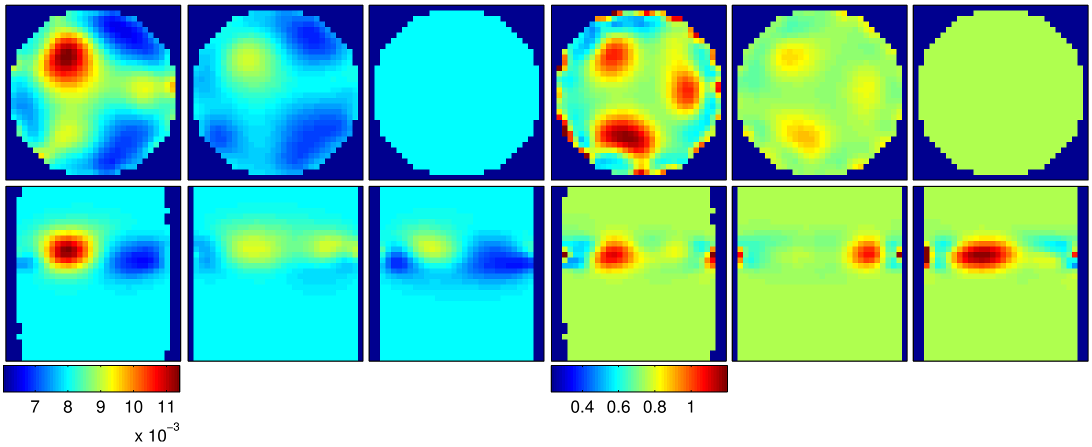


Figure 7.27: Case study 3: TV-reconstructed volume at selected views. Display range cut-off has been applied. Scattering**Columns 1-3:** Absorption **Columns 4-6:** Scattering **Top row:** Top-to-Bottom view **Bottom row:** Front-to-back view

7.8.4 Information theoretic regularization: original prior

The MRI volume is now employed to act as the prior x_{ref} for the IT enabled reconstructions. It is depicted in Fig. 7.28. The non-trivialness of the task at hand becomes evident. The structure between $\mu_a^\star$ and $\mu_s'^\star$ differs[1], however they will receive explicit regularizing support from a single prior. Hence one-to-one correspondence between x_{ref} and either $\mu_a^\star$ of $\mu_s'^\star$ is unavoidably compromised.

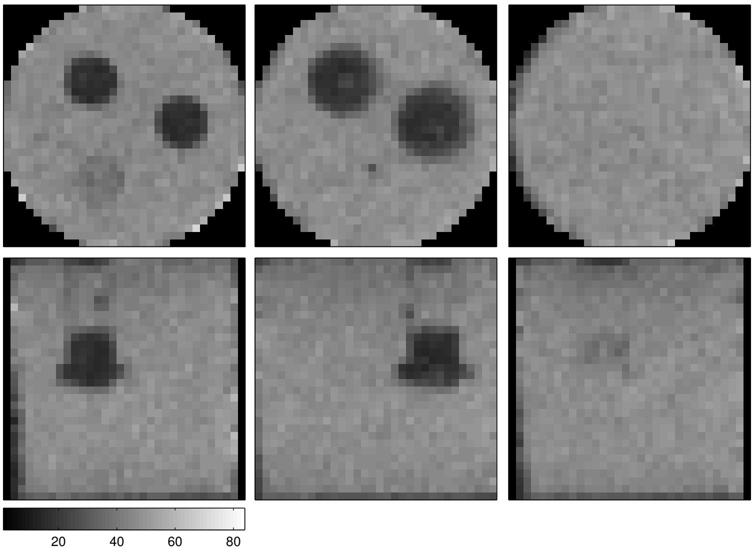


Figure 7.28: Case study 3: MRI prior volume at selected views. Display range cut-off has been applied. **Top row:** Top-to-Bottom view **Bottom row:** Front-to-back view

7.8.4.1 Joint entropy reconstruction

We firstly test the JE functional. We have performed multiple reconstructions for various regularizing weights τ and qualitatively have selected the best. Alternative ways to select the best τ include methods such as the L-curve [Hansen, 1998]. Figure 7.29 shows the obtained results. By visual comparison means only, the structural *a priori* information improves the spatial accuracy of the reconstructed absorption perturbation A. Perturbation B exists in x_{ref} but not in $x^\star$ and has a noticeable impact in both μ_a and μ_s' reconstructions. One would expect that JE due to its structural invariance discussed in Sec.. 7.3.3 would induce bias to the $x^\star$ from the extra prior feature. However, because all MRI visible features have similar gray value representations, the JE invariance is compromised due to the reasons discussed in Sec. 7.3.3. Finally, we can observe that perturbation C is now underestimated, as due to the high τ used for this case, JE tends to reduce the contrast in the image as it results in clustering in the JPDF and hence reduction in the entropy.

7.8.4.2 Mutual information reconstruction

The reconstructions utilizing the MI functional are shown in Fig. 7.30. Regarding the μ_a part of the reconstruction, MI achieves an accurate reconstruction of the perturbation A, in both spatial and quantitative terms, as its attained μ_a value is approximately $2\times$ the background absorption. The strength with

[1] Although a picture of $\mu_a^\star$ and $\mu_s'^\star$ cannot be provided, their expected structure is deduced by Fig. 7.26

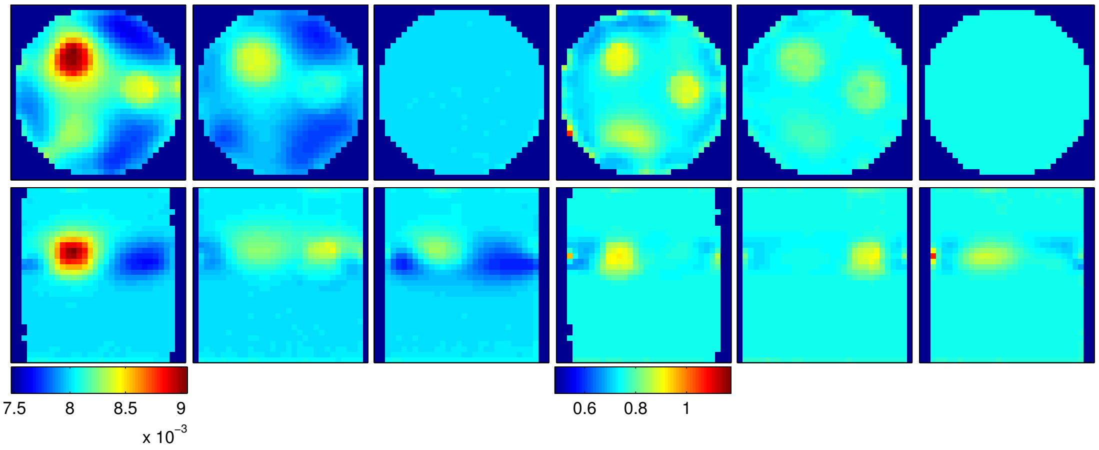


Figure 7.29: Case study 3: JE-reconstructed volume at selected views. Scattering **Columns 1-3:** Absorption **Columns 4-6:** Scattering **Top row:** Top-to-Bottom view **Bottom row:** Front-to-back view

which MI enforces the prior features - due to its marginal entropy term - is evident in the 2nd and 3rd column bottom images, where the structure of the cavity is clearly depicted. Unfortunately, for the same reason we observe significant bias in the μ_s' part of the reconstruction from perturbations A & B. Feature C is practically unchanged - structure wise - due to the lack of explicit prior support, however its contrast increased due to the marginal term in MI.

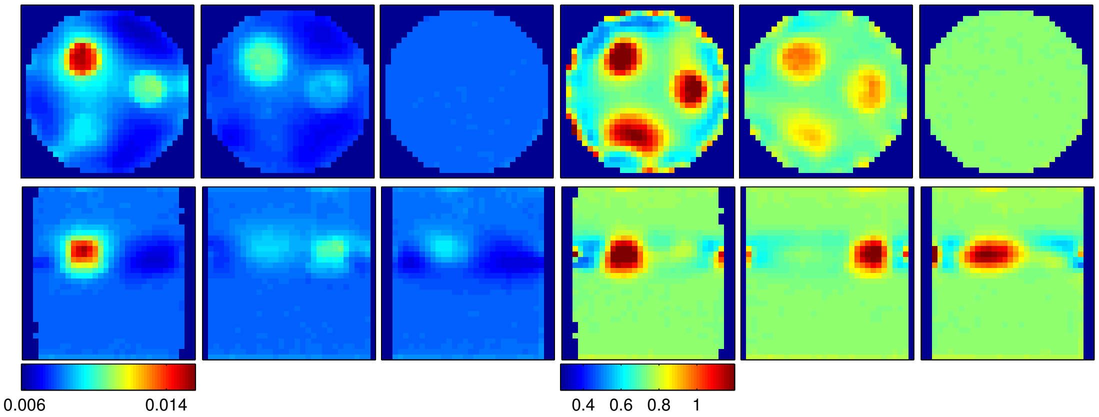


Figure 7.30: Case study 3: MI-reconstructed volume at selected views. Scattering **Columns 1-3:** Absorption **Columns 4-6:** Scattering **Top row:** Top-to-Bottom view **Bottom row:** Front-to-back view

7.8.5 Information theoretic regularization: alternative prior

In this case, we tamper with the prior image by performing a manual rotation - no automatic registration involved - around its axis passing from the centres of its two bases. The magnitude of the rotation was qualitatively selected so that perturbation B would be aligned with the centre of the ‘reconstructed’ perturbation C depicted in the converged TV reconstruction. This simple alteration enables the scattering perturbation C to receive explicit support from the prior image, which now has a distinctive feature in the corresponding location. However, even in this case the prior contains one extra feature when compared to the $\mu_a^\star$ and $\mu_s'{}^\star$. The altered prior image is depicted in Fig. 7.31.

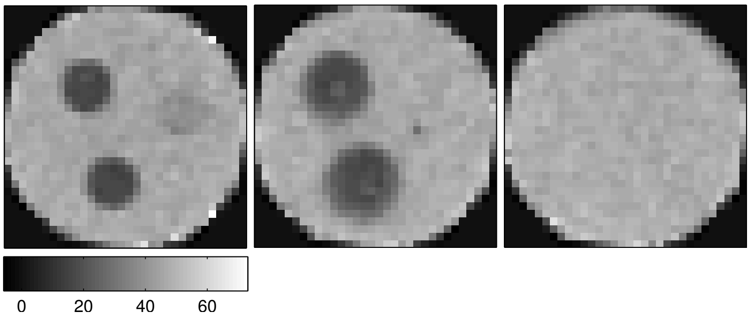


Figure 7.31: Case study 3: MRI prior volume at selected views. **Top row:** Top-to-Bottom view **Bottom row:** Front-to-back view.

7.8.5.1 Joint entropy reconstruction

We re-examine the performance of the JE for the new prior. We employ the same regularizing weight used in the case of the original prior. The major difference in this case regards the improved reconstruction of the scattering perturbation B. This is a direct result of the explicit support by the corresponding feature depicted in the new x_{ref}. However, the contrast of the reconstructed C decreases compared to the one initially retrieved by TV. One should notice that the perturbation C in the MRI image is visible. However, it minimally affects the obtained reconstruction.

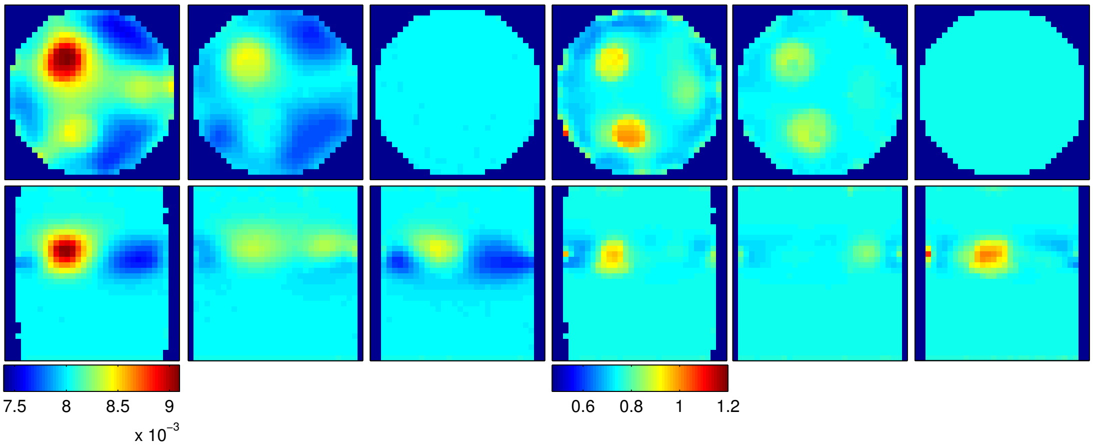


Figure 7.32: Case study 3: JE-reconstructed volume at selected views for the case of rotated prior. Display range cut-off has been applied. Scattering **Columns 1-3:** Absorption **Columns 4-6:** Scattering **Top row:** Top-to-Bottom view **Bottom row:** Front-to-back view

7.8.5.2 Mutual information reconstruction

Finally, we apply the new prior image using MI. The bias in the absorption reconstruction simply changes position following the perturbed - due to rotation - MRI feature. It should be noted that the third and less visible feature in the MRI has a notable impact in both μ_a and μ_s' reconstructions compared to the JE case and this is attributed to the marginal entropy term in the MI functional. The reconstruction of the scattering perturbation C in this case is not as accurate - in terms of spatial accuracy - compared to the JE case and this is an unexpected result as MI enforces structure more strongly than JE. A potential explanation for this behaviour is that the third feature appearing in the reconstruction - corresponding to perturbation B - is strongly opposed by the data results in the entrapment of the optimization in local minima.

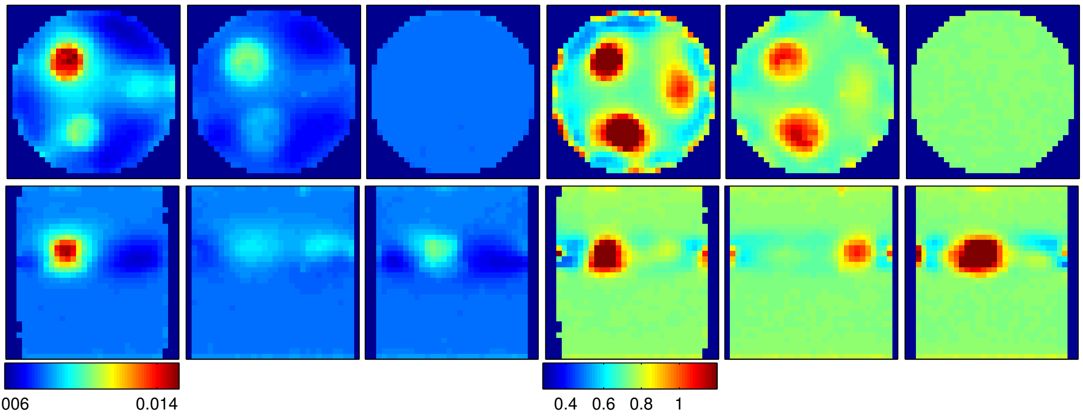


Figure 7.33: Case study 3: MI-reconstructed volume at selected views for the case of rotated prior. Display range cut-off has been applied. Scattering **Columns 1-3:** Absorption **Columns 4-6:** Scattering **Top row:** Top-to-Bottom view **Bottom row:** Front-to-back view

We should emphasize on the complexity of the above experimental study. The magnetic properties of the phantom were specifically designed to produce an MRI image which did not have an one-to-one feature correspondence with the target optical solution. This alone tests the ability of the functionals to introduce partially correct structural priors while minimizing the bias inflicted in regions of the optical solution, which receive inaccurate prior support. One expects that the data would oppose the bias in these regions and if a functional exhibits a level of flexibility, bias would be minimized.

However, in the TV enabled reconstruction of Fig. 7.27 it is evident that the data itself promoted the creation of features in parts of the solution - more specifically in the μ_s' part - which by design were not attributed with optical contrast. The location of these ‘artefacts’ coincided with the cavities in the phantom, which by design should have been invisible to μ_s'. A possible explanation for this result is that the interaction of the propagating light with the cavities, as well as at the interface which they shared with the main body of the phantom, ultimately resulted to scattering. In this case the IT regularization functional are asked to produce an impossible result. That is the incorporation of regionally *incorrect* prior information which promotes the formation of features that are supposedly not a part of the true solution, however the formation of the same features is ultimately favoured by the data itself.

We conclude from this study that the JE absorption reconstruction of Fig 7.29 produced a more spatially accurate result while the inflicted bias from the extra features in the prior was not substantial. The result by MI in Fig. 7.30 was spatially and quantitatively accurate however bias was higher. In addition, the scattering result in the case of the JE rotated prior is encouraging (see Fig 7.31). The accuracy of the reconstructed features which received correct prior support, is a positive indicator about the performance of the method when the entire prior is an accurate representation of the underlying true optical solution.

Before we summarize on the findings of this chapter, as a concluding remark we need to comment on both JE and MI. The inability of JE to strongly enforce structure is simultaneously an advantage (less bias) and disadvantage (weak enforcement of the prior structure). The recent studies which take into account spatial intra-image, inter-pixel dependence (see the reviewed IT regularization literature in Sec. 7.1) suggest that JE can be modified to enforce prior information more strongly. Such alterations will possibly reduce any flexibility of the functional in inducing less bias and also it will suppress data-driven features not supported by the prior. However for all features which are in one-to-one correspondence with the prior image, it is safe to assume that a JE method with intra-image spatial dependency explicitly modeled, would perform better than the version of JE based on the i.i.d. assumption. Our feeling is that if one requires a regularization method which strongly enforces prior information, JE is probably more suitable than the unpredictable and bias-inducing MI. We aim to test structurally dependent JE in the severely ill-posed inverse problem of DOT in the near future.

7.9 Summary and discussion

We have presented an IT regularization framework for DOT, which utilizes the functionals of JE and MI. We assessed the regularizing capacity of the considered functionals and highlighted their in-between differences from a theoretical perspective. JE was found to display partial invariance to potential structural differences between the reconstructed solution and the prior image, which enables it to inflict less bias in the case where the prior contains features not existing in the true optical solution. However, for the same reason JE cannot strongly enforce the prior's structure on the reconstructed solution. In addition, we found that due to the nature of the KDE employed for the purpose of PDF estimation, JE can lead to solutions which are characterized by low contrast between the depicted features, as this corresponds to increased clustering in the JPDF and reduces JE. The latter was demonstrated in practice, with over-regularized JE reconstructions which progressively removed features weakly supported by the data. In contrast, MI was shown to strongly enforce its structure due to the inclusion of the marginal entropy of the optical image, in its formulation. MI does not exhibit any structural invariance, thus features in the prior space will most probably bias the reconstructed image, unless if they are un-regularized. In addition, the marginal term in MI promotes highly varying optical solutions, hence it can emphasize artefacts and induce added variation in the solution.

Chapter 8

Information theoretic regularization in diffuse optical tomography with unregistered structural priors

8.1 Introduction

In this chapter we propose an extension to the information theoretic regularization scheme proposed in Chapter 7, in order to enable the incorporation of spatially unregistered prior information in diffuse optical tomography (DOT). The scheme addresses the second aim in Chapter 1.

Consider the three spaces central to this discussion. These are the true and unknown solution $x^\star$; the space of the optical solution undergoing reconstruction x - with the final estimates explicitly being denoted by $\hat{x}$; and the space of the supplied prior x_{ref} - which by definition is considered spatially *unregistered* with respect to $x^\star$.

In the context of the discussion held in Chapter 7, x_{ref} was characterized as accurate if it contained all the features present in $x^\star$ (one-to-one feature correspondence). In that context x_{ref} accuracy reflected on its contents. Even in the case where an accurate x_{ref} is available, potential for error still exists. One needs to know exactly *where* the *a priori* supplied features should appear in $\hat{x}$, otherwise by forcing them to appear at incorrect locations, we bias the reconstruction and compromise the accuracy of the obtained result. In effect, we seek a x_{ref} which is in accurate registration with $x^\star$. This prior is $x^\star_{\text{ref}}$.

When the above registration condition is not guaranteed *a priori*, we need to explicitly compensate for it. Only when registration is established, x_{ref} can accurately regularize the inverse problem and lead to the accurate retrieval of a solution estimate $\hat{x}$ which resembles $x^\star$. The task at hand was described in Sec. 1.2 with Fig. 1.3 schematically showcasing the task at hand.

Figure 8.1 shows the potential bias which can affect a reconstruction, when x_{ref} are blindly trusted, without having established its accurate spatially registration with $x^\star$. Rows correspond to absorption and scattering. The first column depicts the true optical solutions corresponding to μ_a and μ'_s. Column two depicts a correct prior $x^\star_{\text{ref}}$ in terms of content and alignment. The prior in DOT can be comprised by two different images, one for μ_a and one for μ'_s. Column three depicts a spatially mis-registered x_{ref}.

Column four depicts the difference images formed by subtracting x_{ref} from the $x^{\star}_{\text{ref}}$, revealing the extend of the misalignment. Columns four and five depict $\hat{x}$, obtained by a converged reconstruction via mutual information (MI) and joint entropy (JE) regularization respectively. The *a priori* structural information is enforced to the reconstruction in a manner similar to 'carbon copying'. The bias in the reconstructions is apparent, especially at the regions where the information depicted in x_{ref} does not correspond to the true structure of $x^{\star}$ at the same region, due to misalignment. We should comment on the fact that, although both reconstructions are not accurate qualitatively, one can observe that the bias is greater in the case of MI. According to the discussion in Chapter 7, this is expected as MI priors enforce all their features due to the marginal entropy term in MI. On the contrary, JE induces reduced bias compared to MI at the locations where x_{ref} and $x^{\star}$ are not in agreement. We specifically indicate a sample of areas in the JE reconstruction $\hat{x}_{JE}$ which have been affected by inaccurate *a priori* information, but are not highly biased. Increased bias is observed in the vicinity of the dominant elliptical and circular features. It should be noted that different regularization weights were chosen for the JE and MI reconstructions as the dynamic ranges of the functionals differ. This ensured that the amount of bias induced by MI is not a product of over-regularization but it reflects the natural behavior of the functional. The same approach ensured that the decreased bias induced by JE is not a product of under-regularization

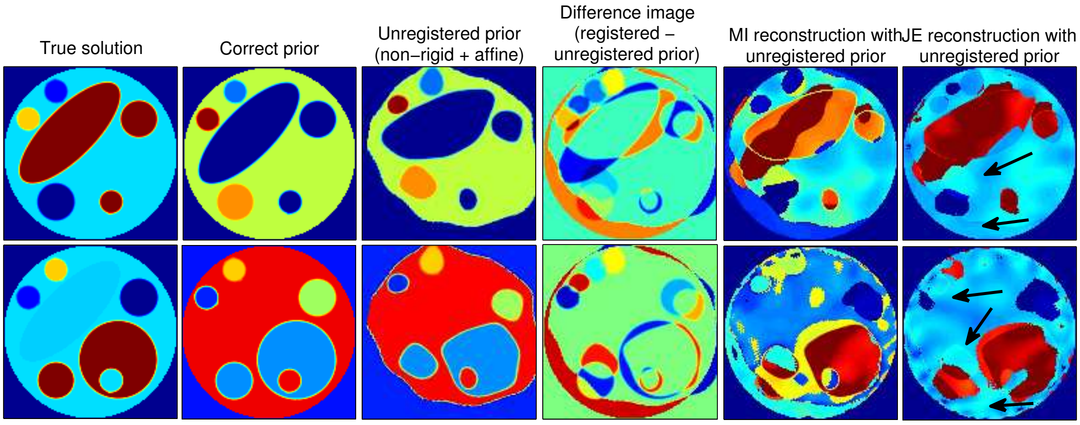

True solution
Correct prior
Unregistered prior (non−rigid + affine)
Difference image (registered – unregistered prior)
MI reconstruction with unregistered prior
JE reconstruction with unregistered prior

Figure 8.1: Depiction of bias in reconstructions induced by the introduction of spatially unregistered priors. **Top row:** absorption **Bottom row:** Scattering **First col.:** Target optical images $x^{\star}$ **Second col.:** Registered priors $x^{\star}_{\text{ref}}$ **Third col:** Unregistered prior x_{ref} (non-rigid + affine) **Fourth col.:** Difference images $(x^{\star}_{\text{ref}} - x_{\text{ref}})$ **Fifth and sixth col.:** Reconstruction via *JE* and *MI* regularization $(\hat{x}_{JE}, \hat{x}_{MI})$

The obvious problem which rises in the attempt of establishing registration between x_{ref} and $x^{\star}$, is that the latter is unknown prior to the reconstruction process. Effectively one is asked to register x_{ref} against an unknown quantity. Firstly, not all information regarding $x^{\star}$ is completely unknown. We have found two categories of methods in the imaging literature regarding the establishment registration between x_{ref} and $x^{\star}$.

The first involves simultaneous probing of the probed anatomy by two (or more) imaging modalities. This approach does not require any knowledge of $x^{\star}$. Spatial registration involves the alignment of the images given a common coordinate system. This alignment however, is not necessary driven by the actual information depicted in the two images - as we already noted one of them is unknown. One simply needs to ensure that the involved coordinate systems are in alignment, under the assumption of course that inter-image corresponding features occupy spatially corresponding areas in both co-ordinate systems. The alignment of the two coordinate systems can be guaranteed by simultaneous probing of the anatomy by the involved modalities - the one which corresponds to $x^\star$ and the ones which provide x_{ref}. The latter can be more easily demonstrated by considering an analogy with photography. One can place a camera on a fixed tripod and take a picture of a scene. Then, the camera is replaced by a second camera - assumingly infra-red in order to maintain the condition of multi-modality. Given that calibration has taken place, i.e. both cameras use lenses of equivalent focal lengths, the resulting pictures should depict exactly the same scene. Differences in the depiction of the features of the true scene might exist between the obtained images, due to the different nature of the two cameras. Some features might be more blurry and some features may be absent in one of the images, but this is not a product of mis-registration. Spatial alignment between corresponding features is guaranteed by the imaging setting and our best estimate regarding the true location of the features is depicted in the image obtained by the higher-resolution camera. In general, the highest-resolution images are employed as x_{ref}. There are reported studies in the literature which introduce prior information in DOT by secondary modalities, where registration is guaranteed by concurrent probing. For example see the DOT-ultrasound imaging (UI) scheme in [Zhu et al., 2005] or a multi-wavelength DOT-magnetic resonance imaging (MRI) coupling [Brooksby et al., 2006]. The concurrent imaging approach enables the establishment of correspondence only in intra-subject studies. If the actual probing is serial, meaning that the data acquisition by both modalities is not exactly simultaneous but one precedes the other, the subject must remain still and the imaged anatomy should not be subjected to deformations, otherwise the underlying physiological information scanned by both modalities, at different times, is not in registration by definition. For example, the accurate registration in serial scanning of organs such as the heart or lungs, cannot be guaranteed as such organs are constantly subjected to deformation due to their natural function.

The second category of methods explicitly address the potential misalignment between $x^\star$ and x_{ref} as a part of the overall task. All these methods are based on the same concept. The global location/shape as well as the local form of x_{ref} is parametrized with respect to some spatial transformation parameters θ. These transformation parameters are included in the set of the optimized quantities. Effectively, the overall task involves a simultaneous reconstruction/registration (SRR) approach. This method has been used in super-resolution imaging where a number of low-resolution images are combined to form a high-resolution image. In order for the images to be combined, they should be registered. In [He et al., 2007; Tom and Katsaggelos, 1995; Zhi et al., 2008] an SRR scheme is adopted. In medical imaging the same method has been adopted for positron emission tomography (PET) and limited view tomography [Bowsher et al., 2006; Van de Sompel and Sir., Brady, M., 2009c]. The study by Bowsher et al. [2006] assumed an intra-modal case and x_{ref} which was rigidly mis-registered with $x^\star$. Van de Sompel and Sir., Brady, M. [2009c] considered non-rigidly deformed, multi-modal priors, similar to what we propose in this work. Our proposal for DOT differs as we reconstruct two images and not one. In addition, we propose a multi-resolution scheme in terms of the B-Spline free form deformation (FFD) control point grid. Finally, we proposed the use of a sole functional for both reconstruction and registration, which guarantees the consistency of the objective function, throughout the process. To the best of our knowledge no such scheme has been proposed for DOT.

The contributions of the work presented in this chapter include

I) the first application of information theoretic regularization with unregistered priors in a severely ill-posed, non-linear inverse problem such as DOT, which requires the registration of the prior against two reconstructed images,

II) the proposition of conditional entropy (CE) as the functional of choice, for the purpose of driving the SRR scheme and finally

III) a scheme of elaborate solution resets which enables the retrieval of improved solutions.

The structure of this chapter is as follows: Section 8.2 re-formulates the inverse problem in DOT in order to enable information theoretic regularization with unregistered anatomical priors. The section contains a discussion regarding the performance of the information theoretic functionals to perform on both solution regularization and prior registration tasks. Section 8.3 introduces the proposed elaborate, alternating optimization scheme comprised by the a B-Spline grid refinement as well as a solution reset scheme. The chapter continues in Sec. 8.4, which describes a series of test cases based on simulated data, designed to test the validity of the proposed scheme as well as the obtained results. Section 8.5 presents a comparison of the results, obtained from the presented test cases. Finally, Sec. 8.6 briefly discusses the presented topics.

8.2 Formulation of the inverse problem

The SRR scheme involves the retrieval of estimates regarding two quantities: the optical solution $\hat{x}$ and the optimal spatial transformation parameters θ which bring x_{ref} in registration with $x^{\star}$, leading to $\hat{x}^{\mathfrak{T}}_{\text{ref}} \rightarrow x^{\star}_{\text{ref}}$. The objective function of the proposed SRR scheme - here defined directly in terms of a minimization scheme - modifies Eq. 7.2 to the following form

$$\left[\hat{x}, \hat{\theta}\right] = \underset{x,\theta}{\text{arg min}} \left[\mathcal{E}(x,\theta) = \left\| \frac{\acute{y} - \mathcal{F}\left(\mathcal{S}^{-1}(x)\right)}{c_1} \right\|^2 + \tau \Psi\Big(x(r), x^{\mathfrak{T}}_{\text{ref}}\left(\mathcal{T}(r';\theta)\right)\Big) \right]. \tag{8.1}$$

where $x^{\mathfrak{T}}_{\text{ref}}\left(\mathcal{T}(r';\theta)\right) = x^{\mathfrak{T}}_{\text{ref}}(r)$ is the transformed prior image x_{ref}, originally defined in r' . The superscript $\mathfrak{T}$ explicitly emphasizes that interpolation is applied to define the transformed image over the coordinate system of $x(r)$ (see Sec. 5.3 for registration specific notation). We should remind that x and x_{ref} - as well as its transformed analogue $x^{\mathfrak{T}}_{\text{ref}}$, are comprised by two distributions, either the optical absorption μ_a and μ'_s images in the case of x or the corresponding reference images $x^{\mu_a}_{\text{ref}}$ and $x^{\mu'_s}_{\text{ref}}$ in the case of x_{ref}. Regarding $\Psi\Big(x(r), x^{\mathfrak{T}}_{\text{ref}}\left(\mathcal{T}(r';\theta)\right)\Big)$, it should be noted that regularization is computed over the locations r which are defined inside the overlap domain $\Omega_{x;x^{\mathfrak{T}}_{\text{ref}}}$ of x and $x^{\mathfrak{T}}_{\text{ref}}$. Finally, it should be emphasized that both reference images $x^{\mu_a}_{\text{ref}}$ and $x^{\mu'_s}_{\text{ref}}$, are subjected to the same spatial transformation, defined by a common transformation parameter vector θ.

8.2.1 Regularization functional

In a combined reconstruction registration scheme, the functional $\Psi\Big(x(r), x_{\text{ref}}^{\mathfrak{T}}\big(\mathcal{T}(r';\theta)\big)\Big)$ is required to perform two tasks. These are

- The regularization of the optical solution x. This is accomplished by penalizing all solutions x which are not 'similar' to the *a priori* supplied x_{ref} and
- the assessment of similarity between $x(r)$ and $x_{\text{ref}}^{\mathfrak{T}}(r)$, with $r \in \Omega_{x;x_{\text{ref}}^{\mathfrak{T}}}$, in order to drive the spatial registration process.

Section 5.6.4 discussed the use of information theoretic functionals for the purposes of image registration. It conclusively demonstrated the inferiority of JE compared to MI, as well as the inferiority of MI compared to normalized mutual information (NMI), for the purpose of assessing similarity in the variable overlap domain and ultimately driving a registration scheme.

On the contrary, the discussion in Chapter 7 regarding the regularization with registered priors, showed that the inclusion of the marginal entropy term $h(x)$ in the MI can introduce bias to the image, rendering JE as the preferable choice for the regularization function. This is also demonstrated in the reconstructions provided at the introductory section of this chapter (see Fig. 8.1). It should be noted that $h(x)$ also exists in the NMI term which was not tested for regularization purposes, however its effects should be similar to its un-normalized variant, the MI.

We are now faced with the choice of a functional which would perform best for both cases. The only other study employed a SRR scheme based on information theoretic priors employed as well an alternating approach was by Van de Sompel and Sir., Brady, M. [2009c]. In that study $\Psi\big(x(r), x_{\text{ref}}^{\mathfrak{T}}(r)\big)$ was set to JE during the reconstruction step and was then switched to negative NMI during the registration step. Such an approach can work in practice if one ensures that both JE and NMI are subjected to minimization and return an improved estimate, at every step of the alternating approach. Such an approach however effectively changes the objective function (Eq. 8.1) between the two steps and raises questions regarding its pure mathematical validity. By doing so, one simply discards the JE derivative with respect to θ and similarly the NMI derivative with respect to x.

In our approach we propose to use a functional which performs better than JE in registration (although it is inferior to both MI and NMI) and as well as JE in reconstruction. By using the same functional for both steps, we keep the global objective function unchanged. The proposed functional is formulated by dropping the marginal entropy term $h\big(x(r)\big)$ of the reconstructed image from the negative MI, however it retains the marginal term of the transformed image $h\big(x_{\text{ref}}^{\mathfrak{T}}(r)\big)$. Consider the MI formulations of Eqs. 4.47-4.49. By dropping the marginal term on the RHS, effectively the resulting function is the CE $h(x \mid x_{\text{ref}}^{\mathfrak{T}})$. In practice we drop the term from the *negative* MI (as we don't maximize MI but minimize its negative), hence the employed functional is defined as

$$\Psi\big(x(r), x_{\text{ref}}^{\mathfrak{T}}(r)\big) = -\Big(h\big(x_{\text{ref}}^{\mathfrak{T}}(r)\big) - h\big(x(r), x_{\text{ref}}^{\mathfrak{T}}(r)\big)\Big) \tag{8.2}$$

$$= h\big(x(r) \mid x_{\text{ref}}^{\mathfrak{T}}(r)\big). \tag{8.3}$$

We now examine the behavior of the above functional (Eq. 8.3) during the reconstruction and registration steps of the alternating optimization scheme.

1. **Reconstruction step :** The optimization of x is driven by the derivative of Eq. 8.2 with respect to x. The derivative $\partial h\big(x_{\text{ref}}^{\mathfrak{T}}(r)\big)/\partial x$ is zero as $x_{\text{ref}}^{\mathfrak{T}}$ does not depend on x. Hence, the derivative of $\Psi\big(x(r), x_{\text{ref}}^{\mathfrak{T}}(r)\big)$ with respect to x is equivalent to the one of JE and consequently it results in the same image update in the context of an iterative image reconstruction scheme. The term $h\big(x_{\text{ref}}^{\mathfrak{T}}(r)\big)$ is constant during the reconstruction step. One has to remember that $r \in \Omega_{x;x_{\text{ref}}^{\mathfrak{T}}}$. For all locations $r \in \Omega_{\mathbf{x}} \backslash \Omega_{x;x_{\text{ref}}^{\mathfrak{T}}}$, the derivative $\partial\Psi\big(x(r), x_{\text{ref}}^{\mathfrak{T}}(r)\big)/\partial x$ is simply set to zero. This means that all parts of the solution x which are not overlapping with $x_{\text{ref}}^{\mathfrak{T}}$, are not subjected to regularization until the registration process advances and they become a part of $\Omega_{x;x_{\text{ref}}^{\mathfrak{T}}}$. In theory, we can employ a second, solely data-driven regularization functional such as $totalvariation(TV)$ to provide regularization for the parts of $\hat{x}$ which are not in $\Omega_{x;x_{\text{ref}}^{\mathfrak{T}}}$. Such a modification is not however considered in the current implementation.

2. **Registration step :** Considering the registration step and by recalling that $x_{\text{ref}}^{\mathfrak{T}}\big(\mathcal{T}(r';\theta)\big) = x_{\text{ref}}^{\mathfrak{T}}(r)$, both terms in Eq. 8.2 depend on θ. We need to examine how $\Psi\big(x(r), x_{\text{ref}}^{\mathfrak{T}}(r)\big)$ behave in a registration framework. Firstly, it is a dissimilarity measure so it attains its minimum when the images are similar. This involves the minimization of JE in Eq. 8.2, which we know from Sec. 5.6.4 attains a minimum at the correct alignment, however this minimum is not global. Recall that the global minimum in JE registration is achieved when both $x(r)$ and $x_{\text{ref}}^{\mathfrak{T}}(r)$ are populated by uniform values in the area of overlap $\Omega_{x;x_{\text{ref}}^{\mathfrak{T}}}$ [1]. The retained marginal term $-h\big(x_{\text{ref}}^{\mathfrak{T}}(r)\big)$ in Eq. 8.2 partially alleviates this effect as it attains its global maximum for uniform images, opposite to what is needed in the minimization of $\Psi\big(x(r), x_{\text{ref}}^{\mathfrak{T}}(r)\big)$.

We repeat the tests of Sec. 5.6.4 to test the behavior of CE in a registration framework. Firstly we repeat the test of Subsec. 5.6.4.1 (Fig. 5.8) where one image is horizontally translated over itself and the functionals are computed for the various overlap configurations. Figure 8.2 shows the plots of the various information theoretic functionals, including CE, the new functional proposed in this work. CE attains a stronger minimum at the correct registration compared to JE, however it is still the global one. One can only expect that registration can be recovered if the initial misalignment falls inside the basin of attraction of the desired minimum - hence it depends on good initialization. It should be emphasized that both MI and NMI are also plagued by local optima, hence they too depend on accurate initialization - although they attain their global one at the correct registration. Local minima can be alleviated by multi-resolution strategies.

We also repeat the test of Subsec. 5.6.4.2, which was used to compare NMI to MI. The functionals are tested for various alignment configurations formed by a target image against itself, where the latter is rotated between $-30° \rightarrow 30°$. In addition we considered three different sizes of its background - varying in the horizontal direction, which we called field of view (FOV) scales. It should be noted that

[1] This results in a joint probability density function (JPDF) of a single entry of maximum probability and corresponds to the global minimum of *JE*

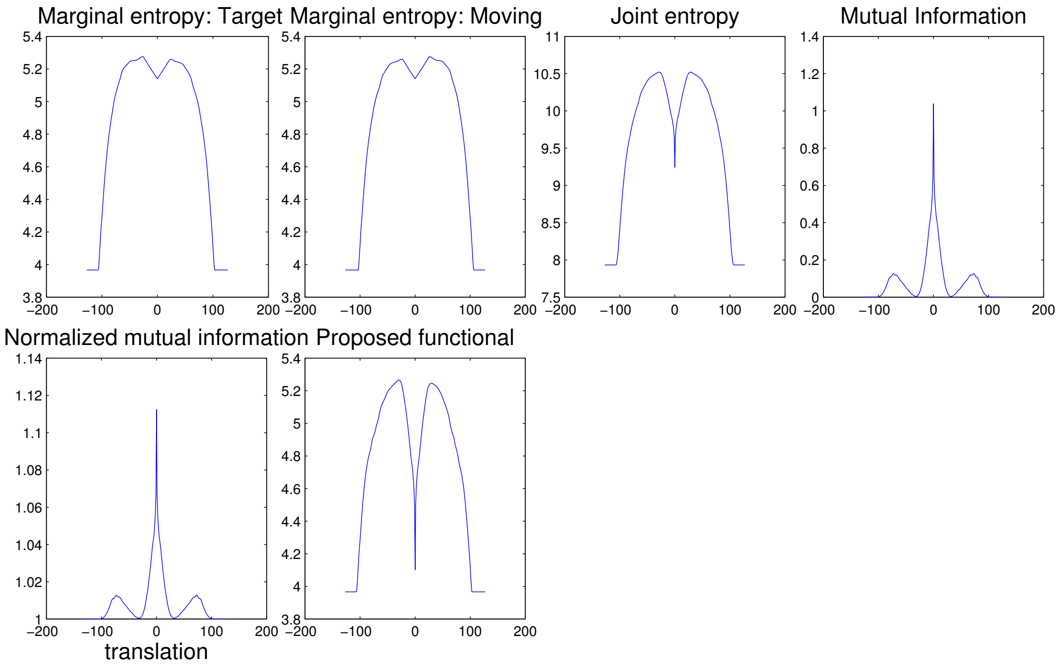


Figure 8.2: Conditional entropy in translational transformations. We repeat the test of Fig. 5.8 and plot all information theoretic functionals in order to compare them against conditional entropy.

the tests of Sec. 5.6.4.2 which resulted in the plots of Fig. 5.10 were performed using the discrete entropy formulation. Figures 8.3 and 8.4 show similar plots - including the proposed functional of CE (negated so it can be compared with NMI), for the discrete and continuous case respectively. It is apparent that in the discrete case, the CE - and similar to NMI, does not have the problem which manifests in MI with the maxima becoming minima. Hence, for each FOV scale the CE attains a global optimum for the correct alignment.

An interesting finding which has not been reported in the rather extensive medical image registration literature, is that in the continuous case, the NMI does not maintain the beneficial capacity of its discrete analogue, that is being invariant to the size of the background included in the $\Omega_{x;x_{\text{ref}}^{\mathcal{T}}}$. It rather exhibits the same problematic behavior as MI. The same result was obtained by both empirical and Shannon implementations. On the contrary, CE exhibits comparable behavior in both discrete and continuous formulations. It has to be noted that this is simply an observation for this case and we have not extensively tested its general validity. A recent publication by Cahill *et al.* Cahill et al. [2008] has brought attention to the non full invariance of NMI to the size of background. Their proposed solution requires known background/foreground statistics which is not a method we want to adopt in this case. In practice, CE performs more than adequately its registration duties, especially in cases where the largest part of both x and x_{ref} are inside the common overlap domain.

8.3 Objective function minimization scheme

We employ a gradient based minimisation scheme to retrieve a solution estimate from Eq. 8.1. The scheme requires the derivatives of its comprising terms with respect to x and θ.

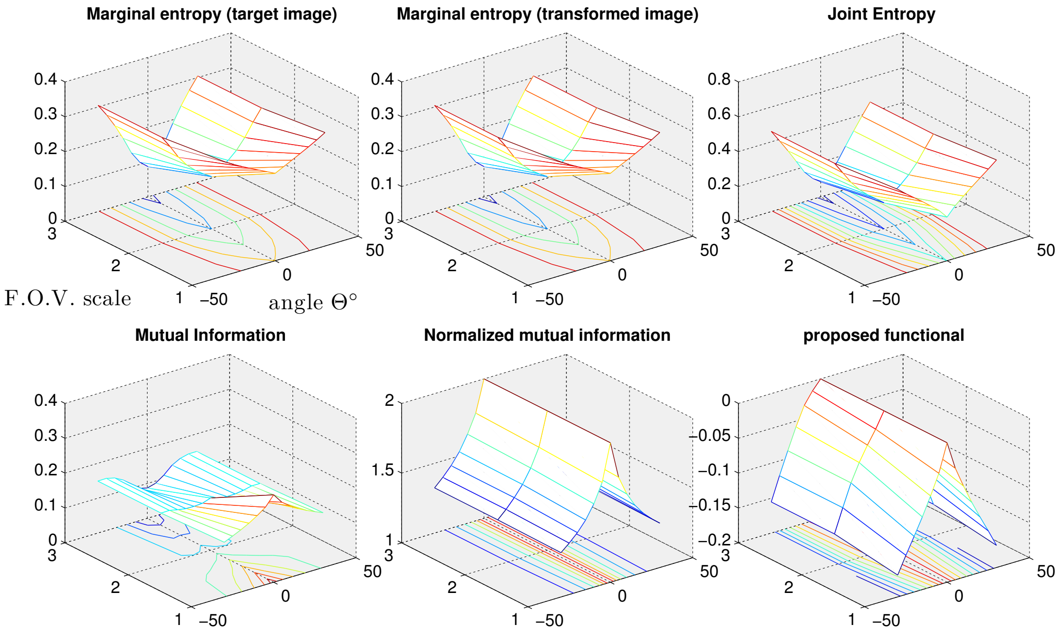


Figure 8.3: Information theoretic functionals against different overlap configurations manifesting from rotational misalignment and for various sizes of field of view: Discrete entropy case.

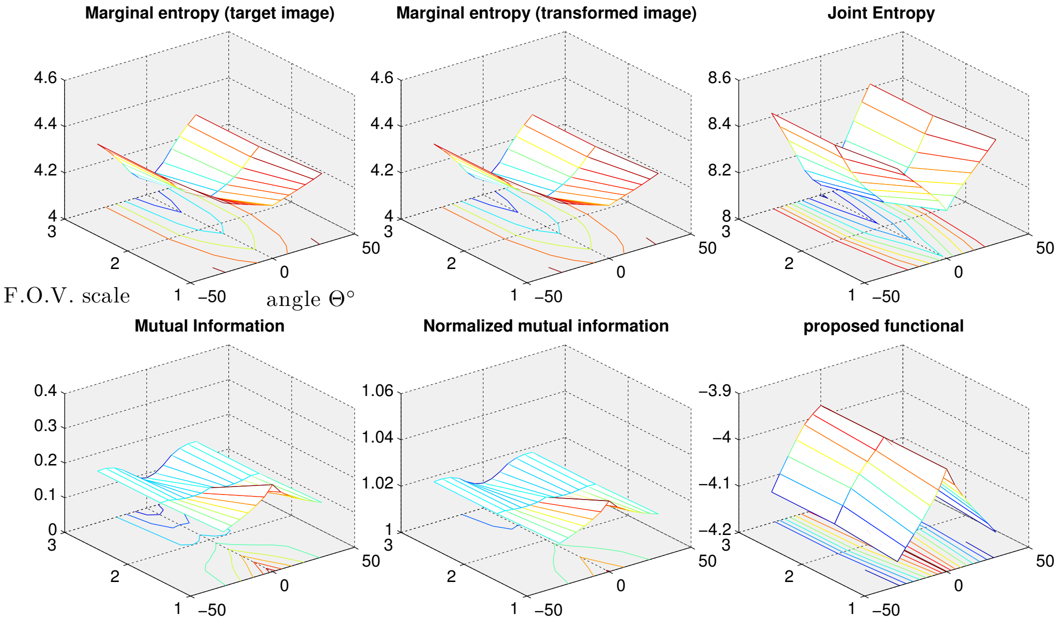


Figure 8.4: Information theoretic functionals against different overlap configurations manifesting from rotational misalignment and for various sizes of field of view: Continuous entropy case.

8.3.1 Derivatives with respect to optical solution

The gradient of the data-fit term $\left(g^{(k)} = \partial\mathcal{E}(x^{(k)})/\partial x^{(k)}\right)$ was discussed in Sec. 7.2.1. In the same section we discussed the marginal and joint entropy derivatives with respect to x. Regarding the regularization functional, as it was noted in Sec. 8.2.1, its derivative $\partial\Psi\left(x(r), x^{\mathfrak{T}}_{\text{ref}}(r)\right)/\partial x$ is equivalent to the derivative of JE (see Eq. 7.14).

8.3.2 Derivatives with respect to transformation parameters

The gradient of the data-fit term with respect to θ is zero, as none of its terms depend on it. Regarding the regularization functional, we employ numerical derivatives. Consider transformation parameters $\theta = \{\theta_1, \theta_2, \ldots, \theta_k, \ldots, \theta_M\}$. Perturbing a single parameter θ_k by some $\Delta\theta_k \to 0$, gives rise to $\theta_k^+ = \{\theta_1, \theta_2, \ldots, \theta_k + \Delta\theta, \ldots, \theta_M\}$ and $\theta_k^- = \{\theta_1, \theta_2, \ldots, \theta_k - \Delta\theta, \ldots, \theta_M\}$. Then the central numerical derivatives of the regularization functional are defined as

$$\frac{\partial\Psi\Big(x(r), x_{\text{ref}}^{\mathfrak{T}}\big(\mathcal{T}(r';\theta)\big)\Big)}{\partial\theta_k} = \frac{\Psi\Big(x(r), x_{\text{ref}}^{\mathfrak{T}}\big(\mathcal{T}(r';\theta_k^+)\big)\Big) - \Psi\Big(x(r), x_{\text{ref}}^{\mathfrak{T}}\big(\mathcal{T}(r';\theta_k^-)\big)\Big)}{2\Delta\theta_k}, \quad \forall k. \tag{8.4}$$

We also note that

$$\frac{\partial\Psi\Big(x(r), x_{\text{ref}}^{\mathfrak{T}}\big(\mathcal{T}(r';\theta)\big)\Big)}{\partial\theta_k} = \frac{\partial\Psi\Big(\mu_a(r), x_{\text{ref}}^{\mu_a\,\mathfrak{T}}\big(\mathcal{T}(r';\theta)\big)\Big)}{\partial\theta_k} + \frac{\partial\Psi\Big(\mu_s'(r), x_{\text{ref}}^{\mu_s'\,\mathfrak{T}}\big(\mathcal{T}(r';\theta)\big)\Big)}{\partial\theta_k}, \quad \forall k. \tag{8.5}$$

One can decompose the transformation parameters to $\theta = \{\theta_{\text{Affine}}, \theta_{FFD}\}$. θ_{FFD} is comprised by the perturbations of FFD grid nodes $\varphi_{i,j}$ (see Eq. 5.15). Each nodal location is perturbed by a fixed amount - common for all nodes, in both x and y directions. Choosing the perturbation amount for θ_{Affine} is not a trivial task, especially if θ_{Affine} is explicitly comprised by parameters expressing translations, rotations, scaling and shearing and not by the general affine parameters of Eq. 5.6. In the former case, the perturbations of the different transformations composing the combined global affine, have to be scaled as they have dis-analogous effect on the optimized functional. For example a rotational perturbation by 0.1 radian results in smaller transformation of the moving image compared to a scaling transformation of 0.1. One can perceive the 'amount' of transformation as a function of the number of displaced pixels as well as the displacement magnitude. The employed package uses empirically derived constants to scale the different types of individual global transformations. This is a common approach employed by many prominent registration packages[1]. A more elaborate, geometry based approach was proposed by Studholme *et al.* [Studholme et al., 1996]. Finally, it should be mentioned that probably the best option - from a computational complexity perspective, is to employ *analytic* derivatives with respect to θ. These can be obtained via chain rule according to

$$\frac{\partial\Psi\Big(x(r), x_{\text{ref}}^{\mathfrak{T}}\big(\mathcal{T}(r';\theta)\big)\Big)}{\partial\theta_k} = \frac{\partial\Psi\Big(x(r), x_{\text{ref}}^{\mathfrak{T}}\big(\mathcal{T}(r';\theta)\big)\Big)}{\partial x_{\text{ref}}^{\mathfrak{T}}\big(\mathcal{T}(r';\theta)\big)} \frac{\partial x_{\text{ref}}^{\mathfrak{T}}\big(\mathcal{T}(r';\theta)\big)}{\partial\mathcal{T}(r';\theta)} \frac{\partial\mathcal{T}(r';\theta)}{\partial\theta} \tag{8.6}$$

The first term in the RHS involves the derivative of the entropy (marginal/joint) with respect to the image gray values, which we discussed in Chapter 6. The second term is simply the gradient of

[1] see for example Insight's ITK `http://www.itk.org/` or
Imperial's IRTK `http://www.doc.ic.ac.uk/~dr/software/`

the image $\nabla_r x^{\mathfrak{T}}_{\text{ref}}(r) = \nabla_{\text{x}} x^{\mathfrak{T}}_{\text{ref}}(r) + \nabla_{\text{y}} x^{\mathfrak{T}}_{\text{ref}}(r)$. The latter term depends on the actual transformation employed. Analytic affine derivatives for the affine case were used by Viola [1995] whereas analytic B-Spline FFD derivatives were employed by Modat et al. [2009], who also presented a GPU enabled implementation in the same study.

8.3.3 Simultaneous reconstruction/registration (SRR)

We minimize the objective function of Eq. 8.1 using a multi-resolution approach, where the variable resolution regards the spacing of the B-Spline FFD control point grid and not the resolution of the involved images. The control point grid refinement is absolutely necessary as it is the only way to recover both large as well as local spatial misalignments (see Sec. 5.5.2.3). During the SRR, it is assumed that global affine misalignments have already been recovered, thus a good initialization estimate is essential. Facilitating the affine registration as a part of the SRR is possible, but not realized in the implemented scheme.

8.3.3.1 Initialization

Prior to the SRR scheme we obtain initial estimates $\hat{x}_{\text{init}}$ and $\hat{\theta}_{\text{init}}$ of the optical solution and transformation parameters respectively. The former is obtained by a reconstruction using a regularization functional such as TV or first-order Tikhonov (TK$_1$). In this work we employ the former. Specifically for the reconstruction initialization we minimize

$$\hat{x}_{\text{init}} = \arg\min_{x} \left[\mathcal{E}_{rec}(x) = \|\acute{y} - \mathcal{F}(x)\|^2 + \tau_1 \text{TV}(x)\right] \tag{8.7}$$

The second estimate $\hat{\theta}_{\text{init}}$ is obtained by registering x_{ref} against the now available $\hat{x}_{\text{init}}$. The process involves an initial affine registration scheme which recovers global misalignments

$$\hat{\theta}^{\text{affine}}_{\text{init}} = \arg\min_{\theta^{\text{affine}}} \left[\mathcal{E}_{reg}(x) = \Psi\big(x(r), x^{\mathfrak{T}}_{\text{ref}}(r'; \theta^{\text{affine}})\big)\right] \tag{8.8}$$

followed by the B-Spline FFD non-rigid registration

$$\hat{\theta}^{\text{non-rigid}}_{\text{init}} = \arg\min_{\theta^{\text{non-rigid}}} \left[\mathcal{E}_{reg}(x) = \Psi\big(x(r), x^{\mathfrak{T}}_{\text{ref}}(r'; \theta^{\text{non-rigid}})\big) + \tau_2 J\big(\mathcal{T}(r'; \theta)\big)\right] \tag{8.9}$$

where $J\big(\mathcal{T}(r'; \theta)\big)$ is the thin plate spline (TPS) penalty function introduced in Sec. 5.5.2, which penalizes for non-smooth transformations. Both schemes utilize a multi-resolution approach, the former in terms of the resolution of the images, whereas the latter in terms of both image resolution and B-Spline control point grid spacing. We will refer to the combined (global+affine) initial estimate as $\hat{\theta}_{\text{init}}$.

8.3.3.2 Iteration in simultaneous reconstruction/registration

Regarding the SRR scheme, we seek to minimize Eq. 8.1. In practice we employ an alternating approach, where each iteration k is split in two steps. The two steps are performed in succession and update the optical and registration part of the solution respectively.

Step $SRR_{1/2}$ The first step updates the optical solution via a minimization of limited iterations

$$x^{(k+1)} = \arg\min_{x} \Big[\mathcal{E}_{SRR_{1/2}}(x^{(k)}) = \|\acute{y} - \mathcal{F}(x^{(k)})\|^2 + \tau\Psi\big(x^{(k)}(r), x^{\mathfrak{T}}_{\text{ref}}(r'; \theta^{(k)})\big)\Big] \tag{8.10}$$

where $x^{(k)}$, $\theta^{(k)}$ are the estimates at the start of the iteration. In the first iteration these are replaced by $\hat{x}_{\text{init}}$ and $\hat{\theta}_{\text{init}}$. In this step, both data-fit term and regularization depend on x and hence both terms are evaluated. In the current implementation, the minimization in this step involves five iterations of conjugate gradients (CG) utilizing a line-search strategy to compute the update step λ_x.

Step $SRR_{2/2}$ The second step updates solely the non-rigid transformation parameters. It assumes that global mis-registration has been recovered. In addition, the involved images are considered in their full resolution without pre-applying blurring, in order to maintain consistency of the overall $\mathcal{E}(x, \theta)$, by assessing the similarity between the same images in both SRR steps. The update is obtained via the minimization scheme

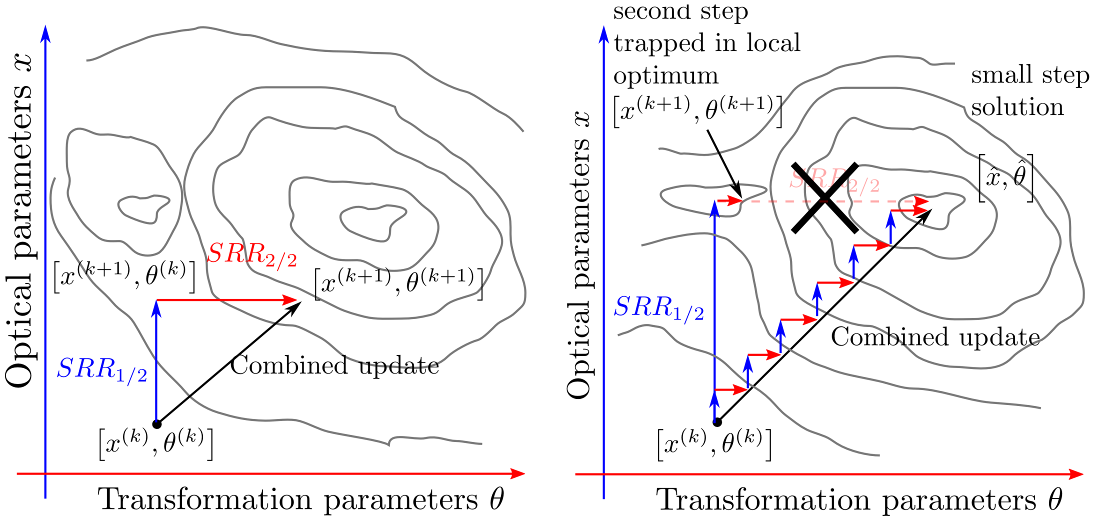
$$\theta^{(k+1)} = \arg\min_{\theta} \Big[\mathcal{E}_{SRR_{2/2}}(\theta^{(k)}) = \Psi\big(x^{(k+1)}(r), x^{\mathfrak{T}}_{\text{ref}}(r'; \theta^{(k)}\big) + \tau_2 J\big(\mathcal{T}(r'; \theta^{(k)})\big)\Big] \tag{8.11}$$

The scheme utilizes $x^{(k+1)}$ which was computed in $SRR_{1/2}$ and is now considered as the current best estimate and a fixed quantity. This step involves the evaluation of only the regularization term (as the data-fit term does not depend on θ), which now acts as an image registration similarity measure. The minimization in this step is performed via five iterations of limited memory BFGS (L-BFGS) (see Sec. 2.6.3), utilizing a line-search strategy to compute the update step λ_θ. The final estimates obtained by SRR at convergence are denoted as $\hat{x}_{SRR}$ and $\hat{\theta}_{SRR}$.

Figure 8.5 schematically shows the geometrical principle of the alternating approach. Successive individual updates in orthogonal axial directions can lead to a solution similar to the one that would have been obtained by a *truly*-simultaneous approach (in reality the orthogonal axes are orthogonal N- and M-dimensional spaces, with N and M being the dimensionality of the individual spaces). However, the magnitude of the individual updates should be kept relatively small, in order for the alternating optimization path to be comparable with the combined one. Otherwise, one could perform a full optical reconstruction followed by a full registration and claim a simultaneous scheme. Such an approach could compromise convergence, as excessive individual steps can be trapped in local minima (see Subfig.8.5b). This is the reason we are forced to limit the number of iterations performed in each of the SRR steps. It should be noted that the choice of five iterations was rather arbitrary. We have tried schemes with less iterations in each of the SRR steps, however the method was prohibitively slow. Identifying the optimum number of iterations is a topic open for further research.

8.3.3.3 B-Spline grid refinements and solution resets

To guarantee global convergence of the B-Spline registration, successive control point grid refinements are necessary. Let G_i denote the grid resolution, where increasing i corresponds to increasing resolution.


Optical parameters x
$\left[x^{(k+1)}, \theta^{(k)}\right]$
$SRR_{2/2}$
$\left[x^{(k+1)}, \theta^{(k+1)}\right]$
$SRR_{1/2}$
Combined update
$\left[x^{(k)}, \theta^{(k)}\right]$
Transformation parameters θ
second step
trapped in local
optimum
$\left[x^{(k+1)}, \theta^{(k+1)}\right]$
small step
solution
$\left[\hat{x}, \hat{\theta}\right]$
$SRR_{2/2}$
Optical parameters x
$SRR_{1/2}$
Combined update
$\left[x^{(k)}, \theta^{(k)}\right]$
Transformation parameters θ

(a) **(b)**

Figure 8.5: Simultaneous reconstruction/registration: Alternating optimization scheme. **Subfig.** 8.5a: Small successive updates can lead to solution estimates similar to a the one retrieved by a *truly*-simultaneous scheme. **Subfig.** 8.5b: If very big steps are allowed, local variations in the solution space can compromise one step and jeopardize the combined convergence.

Refinements: SRR at G_1 is initialized with the aforementioned $\left[\hat{x}_{\text{init}}, \hat{\theta}_{\text{init}}\right]$. Each full SRR results in estimates $\left[\hat{x}_{G_i}, \hat{\theta}_{G_i}\right]$. The next SRR level $i+1$ is initialized by $\left[\hat{x}_{\text{init}}, \hat{\theta}_{G_i}\right]$. Effectively, all the intermediate SRR resolution levels prior to the last, solely update θ.

Solution resets: We have observed that by performing SRR more than once on the same grid level G_i, the retrieved optical solutions are improved. In each repetition we reset the optical solution to the initial $\hat{x}_{\text{init}}$, however we do use the $\hat{x}^{\mathfrak{T}}_{\text{ref}}$ - corresponding to last $\hat{\theta}$ estimate - computed from the previous SRR realization. Let G^z_i denote the z performance of the SRR for G_i. We initialize G^1_i with $\left[\hat{x}_{\text{init}}, \hat{\theta}_{G_{i-1}}\right]$, where $\hat{\theta}_{G_{i-1}}$ corresponds to the converged estimate from the previous refinement level $i-1$. For $i = 1$ we use $\left[\hat{x}_{\text{init}}, \hat{\theta}_{\text{init}}\right]$. Then for all G^z_i, $z \neq 1$, we use $\left[\hat{x}_{\text{init}}, \hat{\theta}_{G^{z-1}_i}\right]$.
The data-flow diagram of the proposed implementation is depicted in Fig. 8.6.

8.3.3.4 Discussion

The reason we employ an alternating approach is the following. If we try to optimize both x and θ simultaneously, we need to normalize the effects of the perturbations of x and θ (this is during the derivative evaluation) on the combined objective function (Eq. 8.1) and consequently the overall optimization. The normalization is required in order to avoid having the descent direction being dominated by one set of parameters. However such a task is non-trivial as x and θ represent very different quantities in physical terms. In addition, the gradient $\nabla_\theta \mathcal{E}\left(x, \theta\right)$ is subjected to the scaling by the regularization parameter τ which is applied to $\Psi\left(x(r), x^{\mathfrak{T}}_{\text{ref}}\left(\mathcal{T}\left(r'; \theta\right)\right)\right)$ in Eq. 8.1. The usually small values of τ vastly minimize the effect of $\nabla_\theta \mathcal{E}\left(x, \theta\right)$ in the optimization compared to $\nabla_x \mathcal{E}\left(x, \theta\right)$ and practically render the moving image $x^{\mathfrak{T}}_{\text{ref}}$ spatially stationary during the full optimization process. This is unacceptable as, the longer

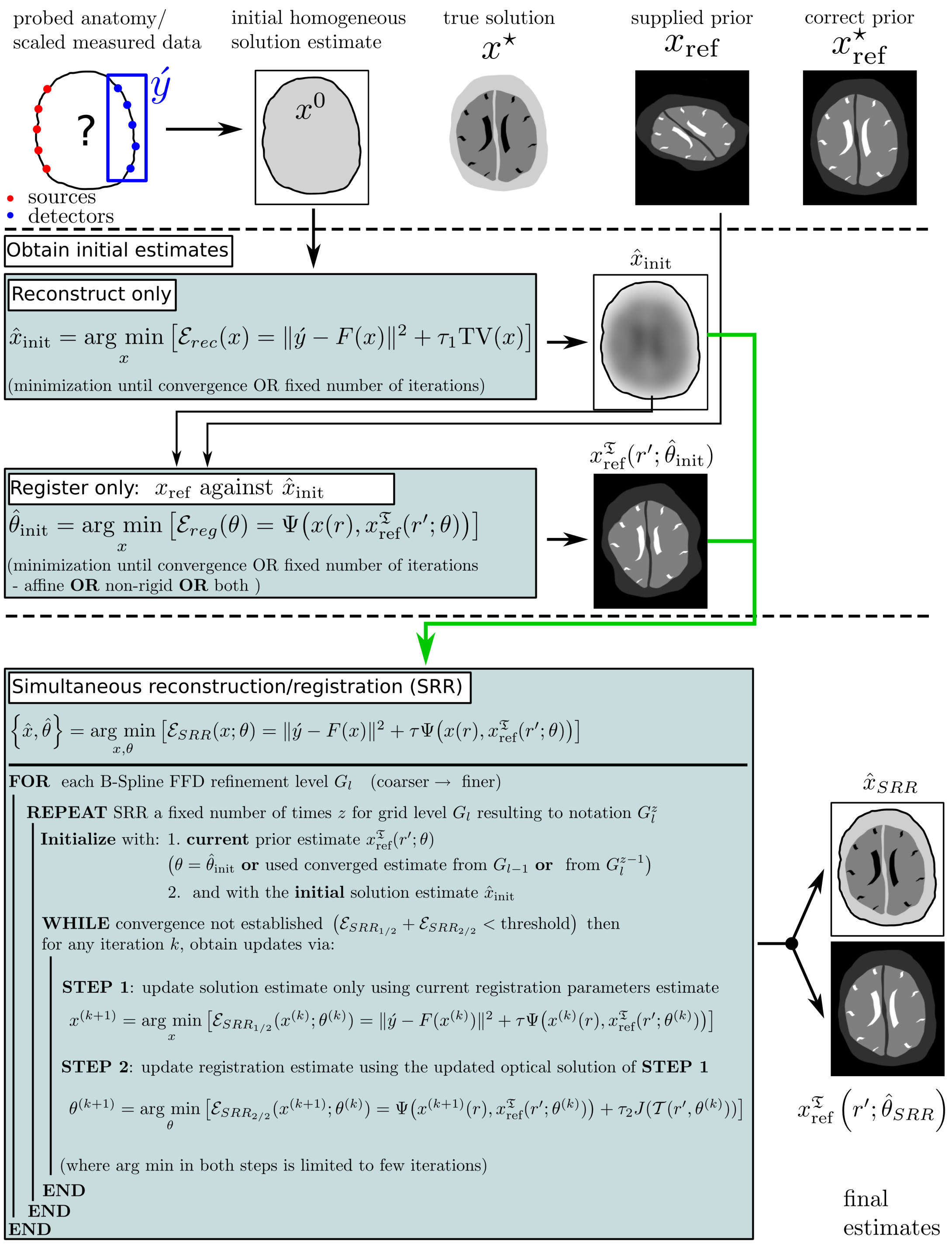

probed anatomy/
scaled measured data
initial homogeneous
solution estimate
true solution
$x^\star$
supplied prior
x_{ref}
correct prior
$x^\star_{\text{ref}}$
?
$\acute{y}$
x^0
sources
detectors
Obtain initial estimates
Reconstruct only
$\hat{x}_{\text{init}} = \arg\min_x \left[\mathcal{E}_{rec}(x) = \|\acute{y} - F(x)\|^2 + \tau_1 \text{TV}(x)\right]$
(minimization until convergence OR fixed number of iterations)
$\hat{x}_{\text{init}}$
Register only: x_{ref} against $\hat{x}_{\text{init}}$
$\hat{\theta}_{\text{init}} = \arg\min_x \left[\mathcal{E}_{reg}(\theta) = \Psi\big(x(r), x^{\mathfrak{T}}_{\text{ref}}(r';\theta)\big)\right]$
(minimization until convergence OR fixed number of iterations
- affine **OR** non-rigid **OR** both)
$x^{\mathfrak{T}}_{\text{ref}}(r';\hat{\theta}_{\text{init}})$
Simultaneous reconstruction/registration (SRR)
$\left\{\hat{x},\hat{\theta}\right\} = \arg\min_{x,\theta} \left[\mathcal{E}_{SRR}(x;\theta) = \|\acute{y} - F(x)\|^2 + \tau\Psi\big(x(r), x^{\mathfrak{T}}_{\text{ref}}(r';\theta)\big)\right]$
FOR each B-Spline FFD refinement level G_l (coarser → finer)
REPEAT SRR a fixed number of times z for grid level G_l resulting to notation G_l^z
Initialize with: 1. **current** prior estimate $x^{\mathfrak{T}}_{\text{ref}}(r';\theta)$
$(\theta = \hat{\theta}_{\text{init}}$ **or** used converged estimate from G_{l-1} **or** from $G_l^{z-1})$
2. and with the **initial** solution estimate $\hat{x}_{\text{init}}$
WHILE convergence not established $\left(\mathcal{E}_{SRR_{1/2}} + \mathcal{E}_{SRR_{2/2}} < \text{threshold}\right)$ then
for any iteration k, obtain updates via:
STEP 1: update solution estimate only using current registration parameters estimate
$x^{(k+1)} = \arg\min_x \left[\mathcal{E}_{SRR_{1/2}}(x^{(k)};\theta^{(k)}) = \|\acute{y} - F(x^{(k)})\|^2 + \tau\Psi\big(x^{(k)}(r), x^{\mathfrak{T}}_{\text{ref}}(r';\theta^{(k)})\big)\right]$
STEP 2: update registration estimate using the updated optical solution of **STEP 1**
$\theta^{(k+1)} = \arg\min_\theta \left[\mathcal{E}_{SRR_{2/2}}(x^{(k+1)};\theta^{(k)}) = \Psi\big(x^{(k+1)}(r), x^{\mathfrak{T}}_{\text{ref}}(r';\theta^{(k)})\big) + \tau_2 J(\mathcal{T}(r',\theta^{(k)}))\right]$
(where arg min in both steps is limited to few iterations)
END
END
END
$\hat{x}_{SRR}$
$x^{\mathfrak{T}}_{\text{ref}}\left(r';\hat{\theta}_{SRR}\right)$
final
estimates

Figure 8.6: Data flow diagram in the proposed simultaneous registration/reconstruction scheme

$x^{\mathfrak{T}}_{\text{ref}}$ remains stationary at a wrong location, the more it biases the optical solution. In turn, the biased optical solution satisfies the registration similarity measure leading to a vicious circle and $x^{\mathfrak{T}}_{\text{ref}}$ is eventually stuck in a local minimum.

In order to consider a fully simultaneous optimization of $\mathcal{E}(x,\theta)$ one needs to normalize the effect of $\nabla_x \mathcal{E}(x,\theta)$ and the scaled transformation derivatives $\tau\nabla_\theta \mathcal{E}(x,\theta)$. This would enable the employment of truly simultaneous optimization schemes which would exhibit faster convergence than the alternating approach. Unfortunately, the exact mathematics of achieving such normalization and keeping the derivatives consistent with the objective function are currently not clear.

Finally, it is essential to emphasize a compromise which had to be made. We stressed earlier that the same functional must be used for both regularization and registration purposes, otherwise the objective function changes during its minimization. Although we establish this conceptually by proposing the use of CE which can adequately perform in both tasks, in practice the implementation of the functional differs between the two steps. Although the proposed efficient entropy evaluation scheme is fast for reconstruction purposes, it is not fast enough to facilitate the computation of the B-Spline FFD transformation derivatives via numerical differences. The number of required evaluations is incapacitating, given the current MATLAB based implementation. The current SRR scheme employs a discrete CE implementation for the registration task. This compromise can be lifted in the future by either employing implementation in more efficient programming environments or simply by utilizing the analytic derivatives of the information theoretic similarity measures with respect to transformation parameters.

8.4 Testing framework and results

We test the validity of the proposed SRR scheme by obtaining preliminary results from case studies based on numerical simulated data. Figure 8.7 introduces the images involved in the simulation. The images depicted in the top row - apart from the third column - correspond to absorption whereas the bottom row corresponds to scattering.

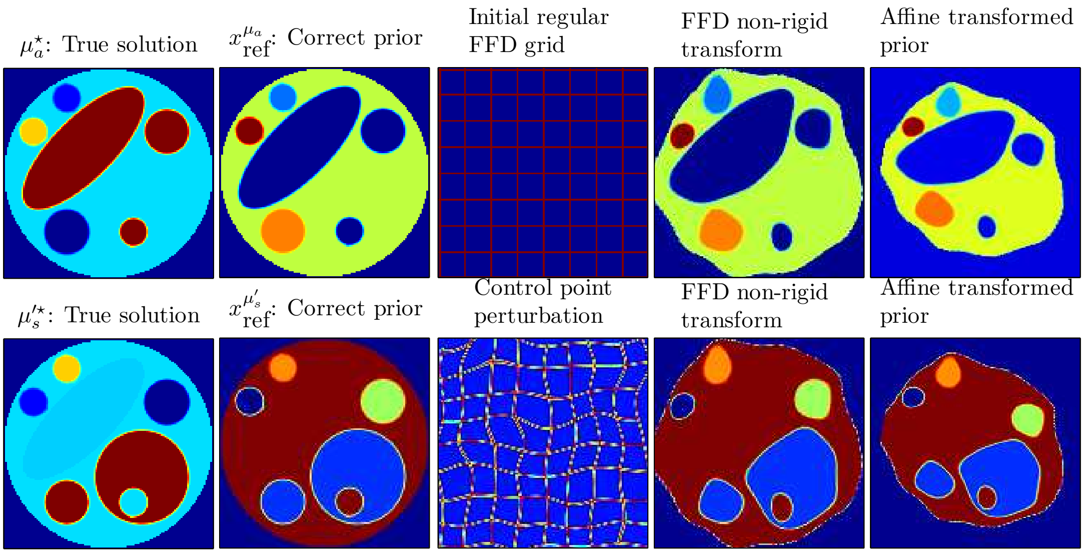

$\mu_a^\star$: True solution
$x_{\text{ref}}^{\mu_a}$: Correct prior
Initial regular FFD grid
FFD non-rigid transform
Affine transformed prior
$\mu_s'^\star$: True solution
$x_{\text{ref}}^{\mu_s'}$: Correct prior
Control point perturbation
FFD non-rigid transform
Affine transformed prior

Figure 8.7: Simultaneous registration/reconstruction test images. **Top row:** Absorption **Bottom row:** Scattering **First col.:** Target optical images $x^\star$ **Second col.:** True, registered priors $x^\star_{\text{ref}}$ **Third col. - top:** Initial FFD B-spline control point arrangement (overlaid on $x^\star_{\text{ref}}$) **Third col. - bottom:** Perturbed FFD B-spline control points **Fourth col.:** Non-rigidly transformed priors via control point perturbation **Fifth col.:** Subsequent affine transformation of the non-rigidly transformed priors. This is the prior x_{ref} which is considered as the supplied unregistered reference image.

The first column depicts the true optical solution and the target reconstruction image. The second column depicts the correct reference images $(x^{\mu_a}_{\text{ref}})^{\star}$ and $(x^{\mu'_s}_{\text{ref}})^{\star}$ comprising $x^{\star}_{\text{ref}}$, which are in accurate spatial registration with $x^{\star}$. The top image in the third column depicts a regularly spaced FFD B-spline grid defined over the domain $\Omega_{x_{\text{ref}}}$ of the true $x^{\star}_{\text{ref}}$. It should be noted that two extra rows and columns of control points are defined outside the domain of the $x^{\star}_{\text{ref}}$, in a manner similar to Fig. 5.4, but they are not visualized. The grid spacing is 16 pixels in both axial directions. The bottom image of the third column shows the non-rigidly deformed control point grid which is used to create the non-rigidly transformed prior images. The perturbation magnitude and direction of every control point, has been *randomly* generated individually across each of the axial directions x and y. More specifically, for any control point φ located at $[\varphi_{\mathsf{x}}, \varphi_{\mathsf{y}}]^{\mathrm{T}}$, its new horizontal location was generated according to

$$\varphi_{\mathsf{x}}^{\text{transformed}} = \varphi_{\mathsf{x}} + 0.03 \times h \times n_1 \tag{8.12}$$

$$\varphi_{\mathsf{x}}^{\text{transformed}} = \varphi_{\mathsf{y}} + 0.03 \times v \times n_2 \tag{8.13}$$

where h corresponds to the maximum horizontal distance between two control points (rightmost-leftmost), similarly v denotes the maximum vertical distance between any two control points on the grid and n_1, n_2 are random trials realized from $\mathcal{N}(0,1)$. In the depicted case, the maximum realized control point perturbation $\sqrt{\Delta\varphi_{\mathsf{x}}^2 + \Delta\varphi_{\mathsf{y}}^2}$ was 17.78 pixels and the minimum 0.52 pixels.

Finally, the fifth column further transforms the non-rigid priors via an affine transformation with arbitrarily selected translations $\mathsf{T}_{\mathsf{x}} = 5,\ \mathsf{T}_{\mathsf{y}} = -5$; rotation $\Theta = -15°$; scaling $\mathsf{S}_{\mathsf{x}} = 0.93$;$\mathsf{S}_{\mathsf{y}} = 0.8$ and shearing $\mathsf{Sh}_{\mathsf{x}} = -0.5$, $\mathsf{Sh}_{\mathsf{y}} = 0.5$. The resulting images are considered to be the prior image x_{ref} used in the following studies.

8.4.1 Case study 1

The setting of the case study is equivalent to the one described in Sec. 7.5. We use the same meshes and the same data-set contaminated with 1% of random, normally distributed noise. Case 1 tests SRR in the case of non-optimal initialization and its ability to recover large prior misalignments as well as a significant portion of the optical solution. Finally, in this case we further compromise the consistency of the objective functional by using different overlap domains for the regularization and registration similarity measures. The former is evaluated by definition in the circular region which is populated by x. The latter however considers that the boundary is known, hence we also consider the dark blue background values surrounding x. In many applications of DOT, the surface of the domain is unequivocally known, for example in the case of brain imaging where the surface can be obtained via photogrammetry [de Souza et al., 2006]. As we will see in the next case study (Sec. 8.4.2), the inclusion of the background of x_{ref} (which constitutes the initial estimate of $x^{\mathfrak{T}}_{\text{ref}}$) in $\Omega_{x;x^{\mathfrak{T}}_{\text{ref}}}$ does not compromise the effort towards the validation of the proposed SRR scheme. In Case study 3 (see Sec. 8.4.3) the above compromise is lifted.

8.4.1.1 Case 1: Initialization

Let $x^{\star}_{\text{ref}}$, $(x^{\mathfrak{T}}_{\text{ref}})_{\text{init}}$ and $(x^{\mathfrak{T}}_{\text{ref}})_{SRR}$ denote the correct prior image $x^{\mathfrak{T}}_{\text{ref}}\big(\mathcal{T}(r',\theta)\big)$ for $\theta = \theta^{\star}$, as well as the pre- and post-SRR priors corresponding to $\hat{\theta}_{\text{init}}$ and $\hat{\theta}_{SRR}$. Prior to the SRR, we obtain the initial esti-

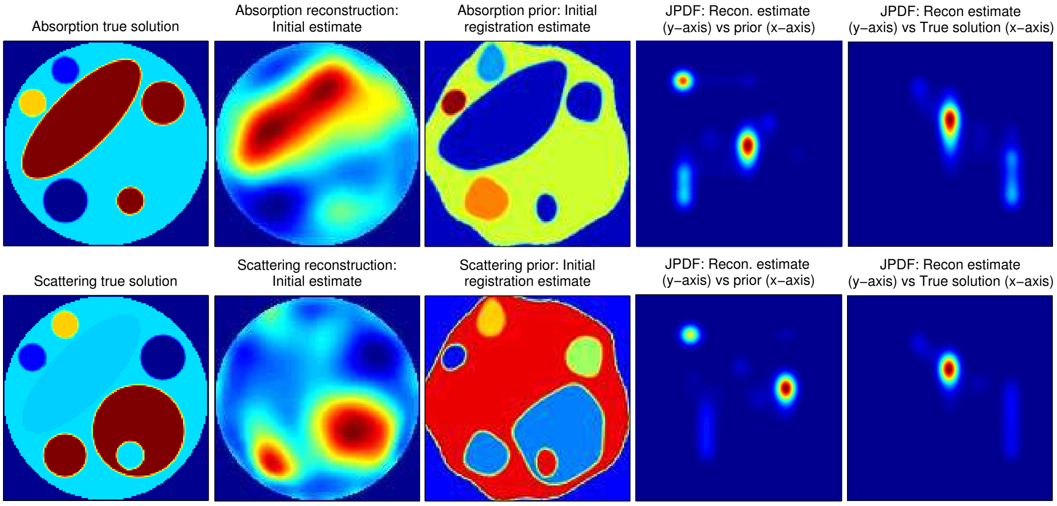


Figure 8.8: Case 1: limited initialization for the SRR scheme. **First col.:** True solutions **Second col.:** Initial reconstruction estimates $\hat{x}_{\text{init}}$ obtained from 10 iterations using TV regularization. **Third col.:** Initial prior $(x^{\mathfrak{T}}_{\text{ref}})_{\text{init}}$ alignment estimate. The corresponding $\hat{\theta}_{\text{init}}$ is obtained by a full affine registration of the x_{ref} against $\hat{x}_{\text{init}}$. **Fourth col.:** JPDF: Initial solution estimates $\hat{x}_{\text{init}}$ (y-axis) vs initial priors estimate $(x^{\mathfrak{T}}_{\text{ref}})_{\text{init}}$ (x-axis) **Fifth col.:** JPDF: Init solution estimates $\hat{x}_{\text{init}}$ (y-axis) vs True solutions $x^{\star}$ (x-axis).

mates $\left[\hat{x}_{\text{init}}, \hat{\theta}_{\text{init}}\right]$. For this case, the former is obtained by a reconstruction utilizing TV regularization, limited to 10 iterations only. Regarding $\hat{\theta}_{\text{init}}$, it is initialized by registering x_{ref} against the obtained $\hat{x}_{\text{init}}$ with *affine* registration only - giving rise to $(x^{\mathfrak{T}}_{\text{ref}})_{\text{init}}$. The first three columns of Fig. 8.8 depict $x^{\star}$; $\hat{x}_{\text{init}}$ and $(x^{\mathfrak{T}}_{\text{ref}})_{\text{init}}$, for both absorption and scattering.

The depicted JPDFs in columns four and five correspond to $\hat{p}\left(\hat{x}_{\text{init}}, (x^{\mathfrak{T}}_{\text{ref}})_{\text{init}}\right)$ and $\hat{p}\left(\hat{x}_{\text{init}}, x^{\star}\right)$. JPDF $\hat{p}\left(\hat{x}_{\text{init}}, x^{\star}\right)$ exhibits increased disorder, as $\hat{x}_{\text{init}}$ is an inaccurate estimate of $x^{\star}$ in terms of quantization and spatial resolution. The increased disorder in $\hat{p}\left(\hat{x}_{\text{init}}, (x^{\mathfrak{T}}_{\text{ref}})_{\text{init}}\right)$ is due to the same reasons, and due to the spatial misalignment between the two images. The additional cluster corresponds to the background, which is solely considered during the registration similarity evaluation

8.4.1.2 Case 1: SRR results

At this point, the SRR algorithm commences. We use CE for both regularization and reconstruction. We employ 400 bins for the JPDF evaluation of both images, however as previously noted the registration functional uses the discrete entropic formulations to render the problem computationally tractable. We note that we choose to visualize the continuous version of $\hat{p}\left(\hat{x}_{\text{init}}, x^{\mathfrak{T}}_{\text{ref}}\right)$ and not the discrete histogram used for registration purposes as the continuous images exhibits more clarity. The continuous JPDF is explicitly computed at initialization (Fig. 8.8) as well as when the scheme converges, solely for visualization purposes.

Regarding the SRR, we use three resolution levels for the B-Spline grid with respective control point spacings $G_1 = [32, 32]$, $G_2 = [16, 16]$ and $G_3 = [8, 8]$. We employ three solution resets for the first level and two for the remaining levels.

It should be mentioned that the last two columns of the introductory Fig. 8.1, depict the reconstruc-

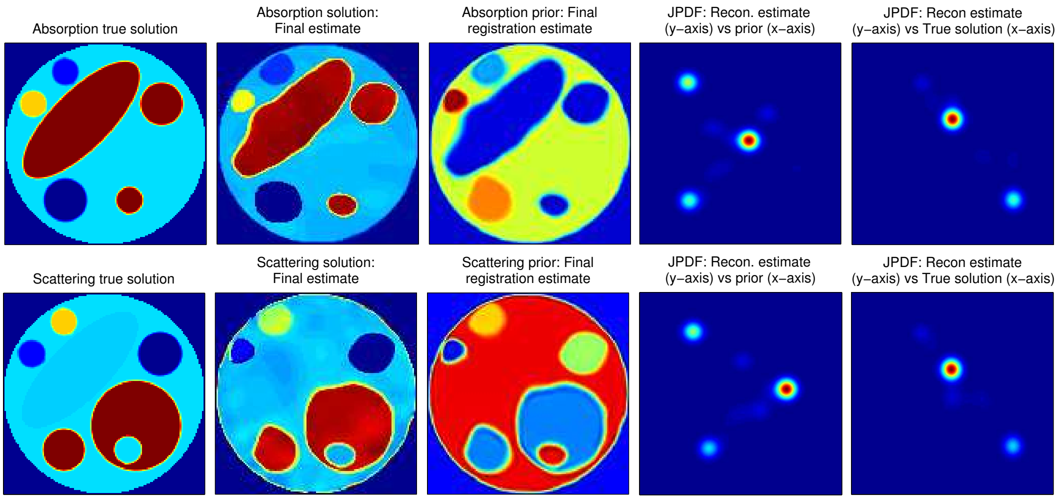


Figure 8.9: Case 1: results obtained by the SRR scheme. Image description similar to Fig. 8.8. Reconstructed images are evidently close to the true solution, as priors have successfully recovered the true solution structure. The JPDFs of the third column ($\hat{x}_{SRR} - x^{\star}$) are clustered due to the increased resolution and accurate contrast/minimized variance of the recovered solution. The JPDFs of the fourth column ($\hat{x}_{SRR} - (x^{\mathfrak{T}}_{\text{ref}})_{SRR}$ are clustered due to the same reasons, as well as due to the improved registration.

tions obtained by blindly applying the unregistered reference images of this case. Figure 8.9 presents the obtained results from the SRR scheme. The reconstructed images are evidently improved from the reconstructions depicted in Fig. 8.1. This is mainly due to the fact that the SRR scheme has recovered a substantial level of the initial misalignment between the unregistered $(x^{\mathfrak{T}}_{\text{ref}})_{\text{init}}$ and the $x^{\star}$. The JPDFs $\hat{p}\big(\hat{x}_{SRR}, x^{\star}\big)$ and $\hat{p}\left(\hat{x}_{SRR}, (x^{\mathfrak{T}}_{\text{ref}})_{SRR}\right)$ exhibit the expected increased clustering compared to their state at initialization (Fig. 8.8). The clustering is interpreted as a reduction in JE - which is the dominant part of CE and results in the minimization of both the registration and reconstruction parts of the algorithm. It is apparent in the registered images, that the current transformation implementation induces new values in $(x^{\mathfrak{T}}_{\text{ref}})_{SRR}$. Possibly these would be eliminated by the use of higher weighting τ_2 on the functional $J\big(\mathcal{T}(r'; \theta^{(k)})\big)$ which penalizes non-smooth transformations. These new values result in spreading of the JPDFs (similar effect to blurring) and a subsequent increase in entropy. This is a problematic issue which can jeopardize convergence. Its resolution is of primary importance, compared to the other potential future improvements.

Table 8.1 reports normalized L_2 errors between the image pairs formed by: the converged $\hat{x}_{TV}$ (see Sec. 8.4.2 for the fully converged TV solution) and $\hat{x}_{SRR}$ against $x^{\star}$; as well as the pre-SRR $(x^{\mathfrak{T}}_{\text{ref}})_{\text{init}}$ and the post-SRR $(x^{\mathfrak{T}}_{\text{ref}})_{SRR}$ against the true $x^{\star}_{\text{ref}}$. The SRR scheme utilizing the unregistered prior, achieves a reduction of 36.3% compared to the full TV reconstruction error. The relevant registration error is reduced by SRR by a significant 43.1%. It should be noted, that the errors would be expected to be reduced even further, if the transformation induced artefacts affecting the $x^{\mathfrak{T}}_{\text{ref}}$ were to be alleviated.

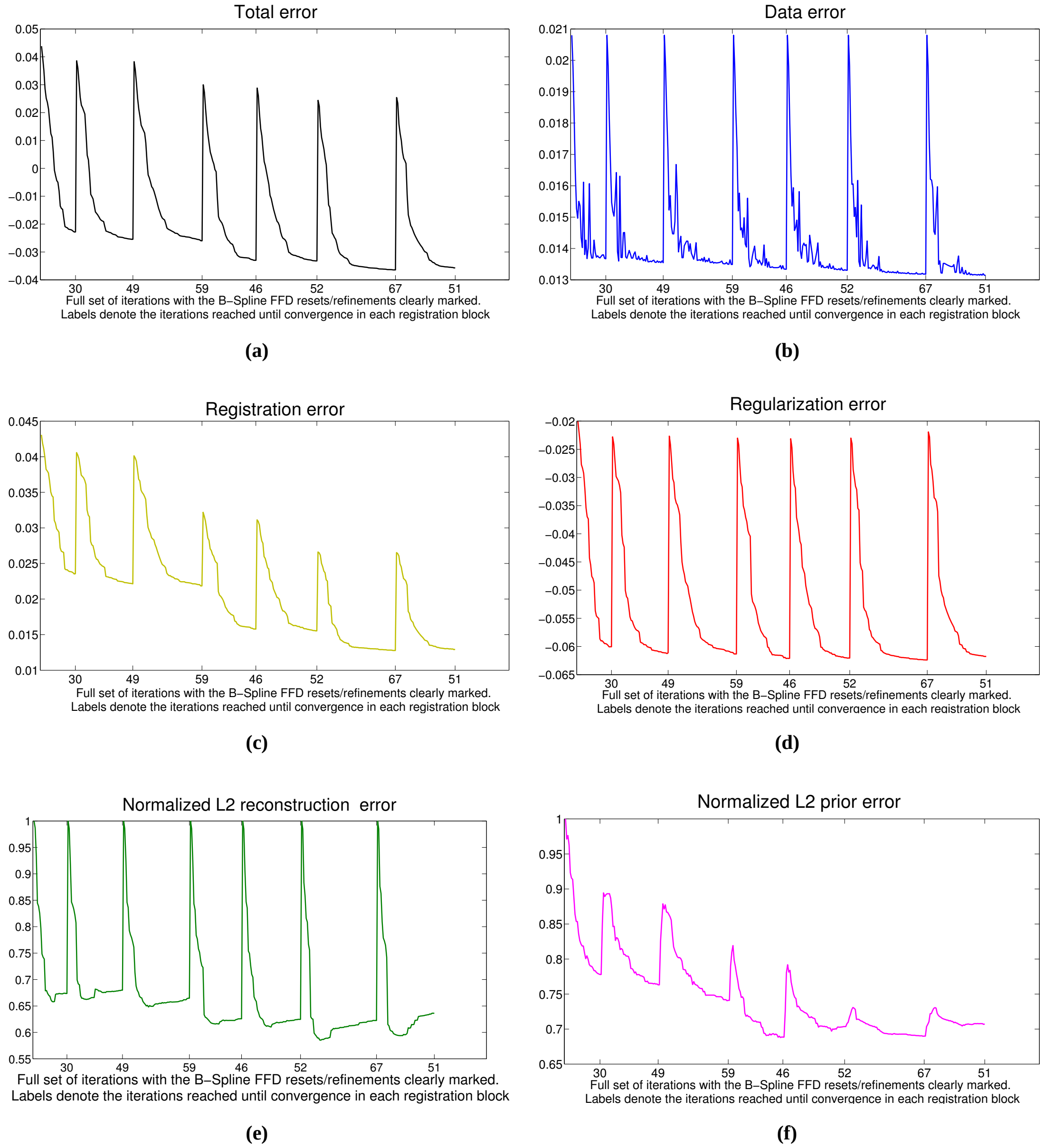
Figure 8.10 presents the corresponding error plots recorded during the SRR optimization. The plotted values correspond to the state of the various quantities at the end of each SRR iteration, in other

Table 8.1: Case 1: Normalized L_2 errors: 1) $x^\star$ vs converged TV-based $\hat{x}_{\textbf{init}}$ 2) $x^\star$ vs $\hat{x}_{SRR}$ 3) $x^\star_{\textbf{ref}}$ vs $(x^{\mathfrak{T}}_{\textbf{ref}})_{\textbf{init}}$ 4) $x^\star_{\textbf{ref}}$ vs $(x^{\mathfrak{T}}_{\textbf{ref}})_{SRR}$

Images considered	Normalized L2 error
$x^\star$, $\hat{x}_{\text{init}}$	19.8%
$x^\star$, $\hat{x}_{SRR}$	12.6%
$x^\star_{\text{ref}}$, $(x^{\mathfrak{T}}_{\text{ref}})_{\text{init}}$	17.6%
$x^\star_{\text{ref}}$, $(x^{\mathfrak{T}}_{\text{ref}})_{SRR}$	10.1%

words just after the registration step $SRR_{2/2}$. All plots are characterized by spikes which correspond to the grid refinements as well as to the solution resets. We also include the L_2 errors between the x undergoing reconstruction as well as the prior $x^{\mathfrak{T}}_{\text{ref}}$ subjected to transformation, against $x^\star$ and $x^\star_{\text{ref}}$ respectively. It should be noted that although the data regularization and registrations errors decrease during resets/refinements, the L_2 errors can exhibit increase. A possible explanation for this behavior is once again the presence of interpolation artefacts induced by registration. More specifically, the regularization functional is minimized while x matches the $x^{\mathfrak{T}}_{\text{ref}}$ - whether the latter contains artefacts or not. There is no feedback mechanism between the L_2 error and the $\psi(x, x_{\text{ref}})$ as L_2 errors are simply based on the unknown $x^\star$ and are computed for validation purposes only when SRR has been completed. The decrease in the regularization, data & registration errors cannot fully guarantee reduction in the reconstructed L_2 error if these artefacts are not treated. It should be noted that the above is the current best explanation we can provide. Other reasons might contribute to the increase in the L_2 errors. Finally, the registration and regularization error differ due to the discrete and continuous implementations respectively as well as due to the different overlap region considered between the two. We note that all errors exhibit significant decrease between their initial and final values.

Finally, Figures 8.11 and 8.12 present the converged optical solutions $\hat{x}_{G^z_i}$ and the priors $(\hat{x}^{\mathfrak{T}}_{\text{ref}})_{G^z_i}$, at the end of each SRR performance for the corresponding G^j_i. Difference images are also provided (see caption for details). Note the further registration which takes place between solution resets.



(a)

(b)

(c)

(d)

(e)

(f)

Figure 8.10: SRR error plots. **First Row: Left:** Total error $\big(\mathcal{D}\big(\acute{y},\mathcal{F}(x)\big)+\tau\Psi(x,x^{\mathfrak{T}}_{\text{ref}})\big)$ (in this case $\Psi(x,x^{\mathfrak{T}}_{\text{ref}})$ corresponds to the continuous implementation used in $SRR_{1/2}$) **Right:** Data error $\mathcal{D}\big(\acute{y},\mathcal{F}(x)\big)$ **Second Row: Left:** Registration error $\Psi(x,x^{\mathfrak{T}}_{\text{ref}})$ (discrete CE) **Right:** $\Psi(x,x^{\mathfrak{T}}_{\text{ref}})$ Continuous CE **Third Row: Left:** Normalized L_2 reconstruction error $(\|(\mu_a-\mu_a^\star)/\mu_a^\star\|+\|(\mu_s'-\mu_s'^\star)/\mu_s'^\star\|)\,/2$. The spikes indicate B-Spline grid refinements/resets. We remind that the algorithm uses grids $\big[G_1^1,\ G_1^2,\ G_1^2,\ G_2^1,\ G_2^1,\ G_3^1,\ G_3^1\big]$. The x-axis values denote the number of iterations in each grid level.

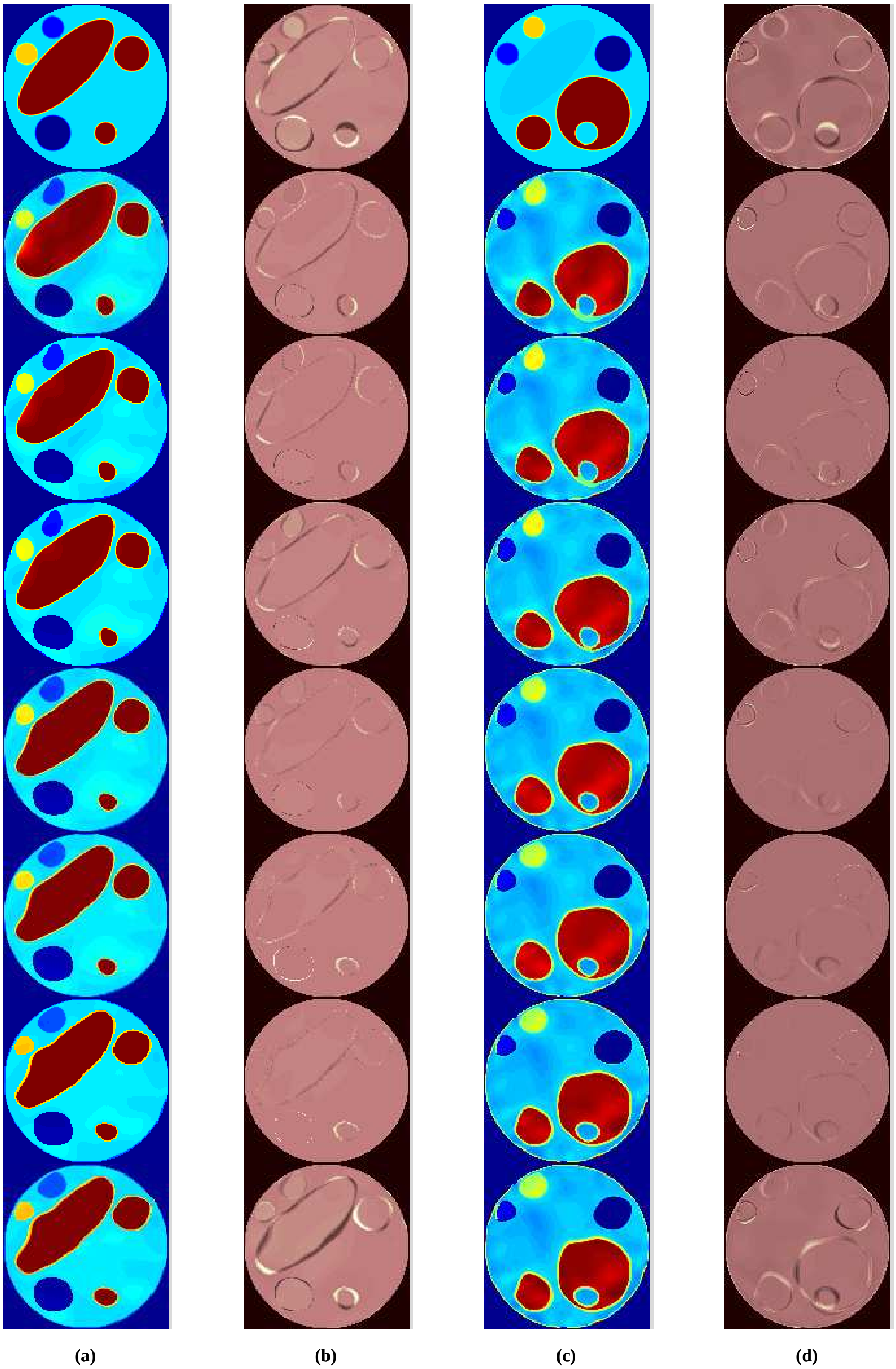


Figure 8.11: Case 1: SRR output at the various FFD grid refinement/reset levels. **Cols. (a),(c):** μ_a, μ_s' **Row 1:** $x^\star$ **Row 2-4:** SRR $G_1 = [32, 32]$. Resets G_1^1, G_1^2, G_1^3. **Rows 5-6:** SRR $G_2 = [16, 16]$. Resets G_2^1, G_2^2. **Rows 7-8:** SRR $G_3 = [8, 8]$. Resets G_3^1, G_3^2. **Cols. (b),(d):** Intra-row differences in cols. (a),(b). **Row 1:** $\hat{x}_{SRR}$-$x^\star$ [8-1] (also interpreted as registration accuracy as $\hat{x}_{SRR} \rightarrow (x^{\mathfrak{T}}_{\text{ref}})_{SRR}$ in structural terms) **Rows 2-7:** Intra-SRR diff/s. [2-1], [3-2] etc. **Row 8:** $\hat{x}_{SRR} - \hat{x}_{\text{init}}$.

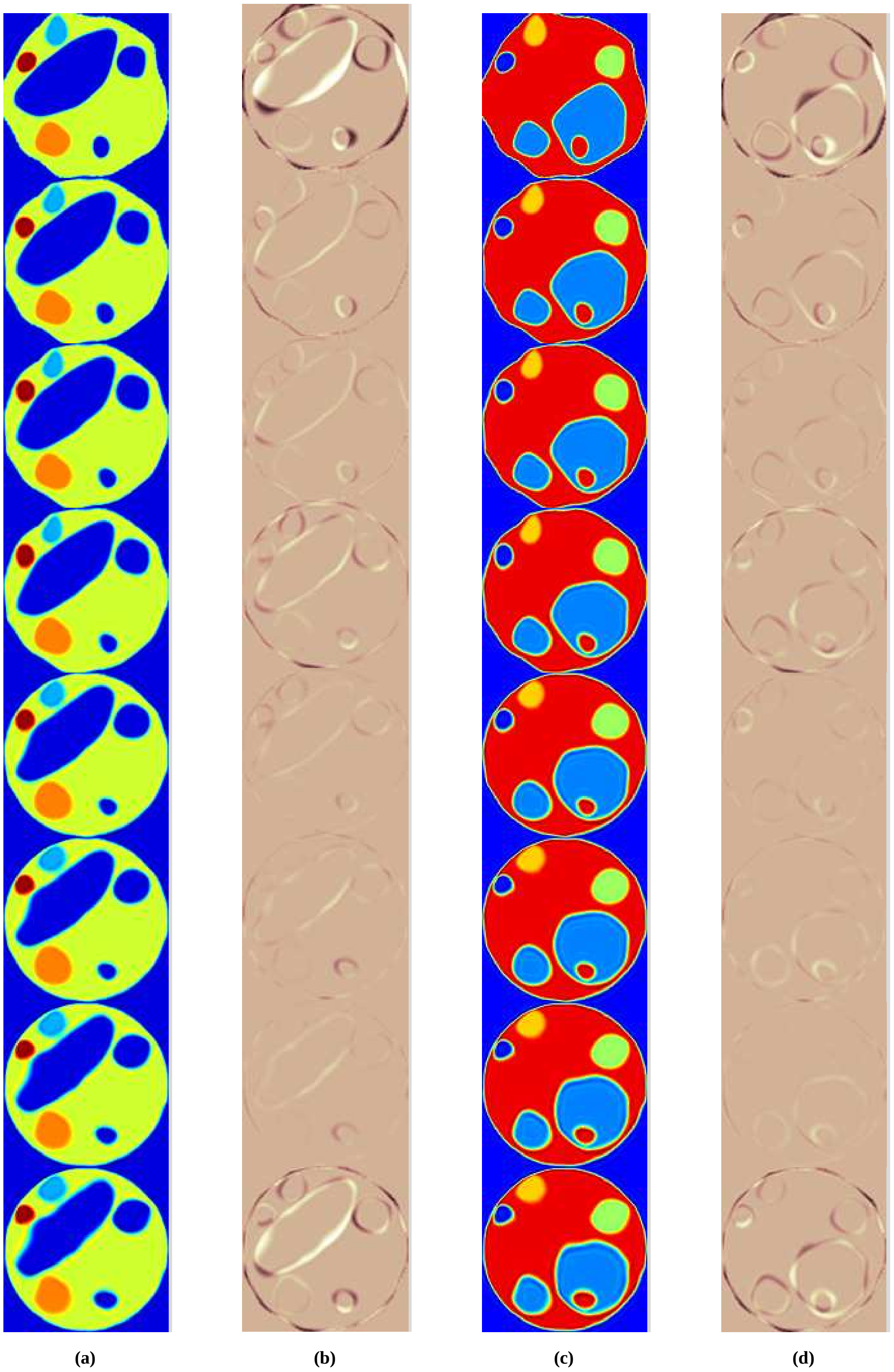


Figure 8.12: Case 1: SRR output at the various FFD grid refinement/reset levels. **Cols. (a),(c):** $x_{\text{ref}}^{\mu_a}$, $x_{\text{ref}}^{\mu_s}$ **Row 1:** $(x_{\text{ref}}^{\mathfrak{T}})_{\text{init}}$ **Row 2-4:** SRR $G_1 = [32, 32]$. Resets G_1^1, G_1^2, G_1^3. **Rows 5-6:** SRR $G_2 = [16, 16]$. Resets G_2^1, G_2^2. **Rows 7-8:** SRR $G_3 = [8, 8]$. Resets G_3^1, G_3^2. **Cols. (b),(d):** Intra-row differences in cols. (a),(b). **Row 1:** $(x_{\text{ref}}^{\mathfrak{T}})_{SRR}$-$(x_{\text{ref}}^{\mathfrak{T}})_{\text{init}}$ (magnitude of total recovered alignment) [8-1] **Rows 2-7:** Intra-SRR diff/s. [2-1], [3-2] etc. **Row 8:** $(x_{\text{ref}}^{\mathfrak{T}})_{SRR} - (x_{\text{ref}}^{\mathfrak{T}})_{\text{init}}$.

8.4.2 Case study 2

The setting of this case is identical to the one of the previous section in terms of data used, overlap regions considered and supplied prior image (see Fig. 8.7), however this time the scheme is initialized by the best possible estimate $\left[\hat{x}_{\text{init}}, \hat{\theta}_{\text{init}}\right]$. The optical initialization estimate $\hat{x}_{\text{init}}$ is obtained by a converged TV reconstruction. The transformation parameters initial estimate $\hat{\theta}_{\text{init}}$ is obtained by registering $x_{\text{ref}}^{\mathfrak{T}}$ to $\hat{x}_{\text{init}}$ until convergence. The registration process involves an affine registration scheme using a multi-resolution pyramid approach, followed by a - new to this case - non-rigid registration using a multi-resolution pyramid approach as well as successive B-spline grid refinement.

We should also note that the registration non-rigid initial registration estimate could be improved by either employing more elaborate registration schemes of by further optimizing secondary parameters, such as smoothing penalties, grid resolutions etc. The important point is that the same registration scheme is employed during the SRR approach hence the relative improvement between initialization and post-SRR is of importance. We note that the non-rigid registration part employs the same smoothness penalty utilized during $SRR_{2/2}$. The initialized optical solution as well as the prior transformed by the initial $\hat{\theta}_{\text{init}}$ are depicted in Fig 8.13.

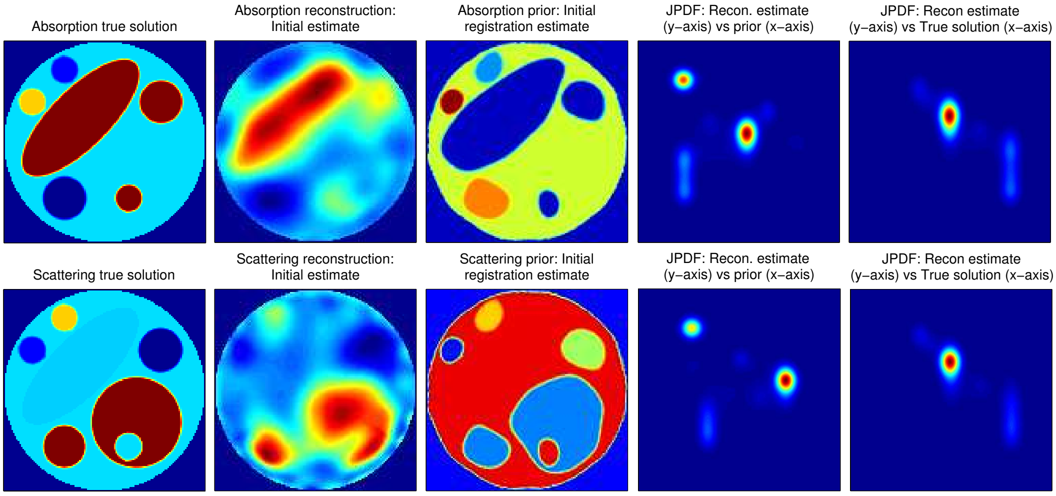


Figure 8.13: Case 2: limited initialization for the SRR scheme. Image description similar to Fig. 8.8. Solution is initialized by a full TV reconstruction. Prior alignment is initialized by a full affine registration against the initial solution estimate, followed by a full non-rigid registration.

The significance of this case is *paramount* towards the validation of the proposed concept and we now explain the reasons. Case 1 utilized less accurate initialization guesses for both prior alignment and solution estimate. The effect of the regularizing functional $\Psi(x, x_{\text{ref}}^{\mathfrak{T}})$ is weighted by usually low values of τ to avoid extreme bias from the introduced prior - except in the cases where one consciously chooses to do so, due to high trust in the supplied prior. For low τ, the regularizing effects of Ψ (that is the enforcement of the prior's structure on x) become more noticeable during the last iterations of the reconstruction, when the gradient of the data-fit term $\nabla_x \mathcal{D}\big(\acute{y}, \mathcal{F}(x)\big)$ (in this case the L_2 norm between data and modeled data) is reduced - due to proximity to the minimum, to a level comparable with the $\tau \nabla_x \Psi(x, \theta)$. At this level the next updates are influenced by both data and regularization, at a comparable level.

We need to ensure that the level of registration achieved in the previous case, is not a result of a largely data-driven reconstruction which strongly opposed or even disregarded the *incorrect* bias from the unregistered and possibly under-regularized prior during the early iterations, during which it holds $\nabla_x \mathcal{D}\big(\acute{y}, \mathcal{F}(x)\big) \gg \tau \nabla_x \Psi(x, x_{\text{ref}}^{\mathfrak{T}})$. A data driven initial reconstruction would avoid local minima due to incorrect bias from the prior and proceed mostly under the influence of $\nabla_x \mathcal{D}\big(\acute{y}, \mathcal{F}(x)\big)$, until near-convergence of $\mathcal{D}\big(\acute{y}, \mathcal{F}(x)\big)$. At that point, the recovered data-driven solution x would be more accurate than the TV-based (10-iterations) $\hat{x}_{\text{init}}$ used in Case 1 and would expectedly be able to further drive a better registration of the prior from that point on. At that final stage, the now more accurately registered prior would have a non-biasing but rather positive regularizing effect (as its structure would be close to registration with the unknown true solution), all objectives would further descend in unison and the final high resolution solution would be obtained. In a more simple terms, we need to ensure that the encouraging results obtained in Case 1, were not simply a product of the registration of x_{ref} against a more accurate, data-driven $\hat{x}$.

For this reason, it is crucial that we can further improve the registration error between prior and $x^{\star}$, compared to the one formed by the *best-possible* initialization estimate $(x_{\text{ref}}^{\mathfrak{T}})_{\text{init}}$ and $x^{\star}$. This analysis has not been performed by any other publication involving SRR. Figure 8.14 shows the obtained results. It is clear that the reconstructions are qualitatively better than the ones obtained by TV. The reported errors in Table 8.2 quantitatively show a 19% improvement of $\hat{x}_{SRR}$ over $\hat{x}_{\text{init}}$ and a highly significant reduction in the error of 30.1%. This result is a strong indicator regarding the validity of the method. In Case 1 we compromised by considering the boundary of $x_{\text{ref}}^{\mathfrak{T}}$ during registration similarity evaluation. In this case, although the boundary is still a part of the overlap region during registration, the boundary mismatch between $(x_{\text{ref}}^{\mathfrak{T}})_{\text{init}}$ and $(x_{\text{ref}}^{\mathfrak{T}})_{SRR}$ is minimal. Thus, the reported decrease in the errors is mainly due to local improvement in the registration of $x_{\text{ref}}^{\mathfrak{T}}$ against x (and not its boundary) and the subsequent improvement of the latter due to more accurate regularization. Finally, Figs. 8.15 & 8.16 show the state of the reconstruction and the transformed priors during the various SRR B-Spline control grid refinement and solution reset stages.

Table 8.2: Case 2: Normalized L_2 errors: 1) $x^{\star}$ vs converged TV-based $\hat{x}_{\text{init}}$ 2) $x^{\star}$ vs $\hat{x}_{SRR}$ 3) $x_{\text{ref}}^{\star}$ vs $(x_{\text{ref}}^{\mathfrak{T}})_{\text{init}}$ 4) $x_{\text{ref}}^{\star}$ vs $(x_{\text{ref}}^{\mathfrak{T}})_{SRR}$

Images considered	Normalized L2 error
$x^{\star}$, $\hat{x}_{\text{init}}$	16.8%
$x^{\star}$, $\hat{x}_{SRR}$	13.6%
$x_{\text{ref}}^{\star}$, $(x_{\text{ref}}^{\mathfrak{T}})_{\text{init}}$	12.6%
$x_{\text{ref}}^{\star}$, $(x_{\text{ref}}^{\mathfrak{T}})_{SRR}$	8.8%

The intra-SRR output images are presented in Figs. 8.15-8.16. SRR error plots are not provided for this case.

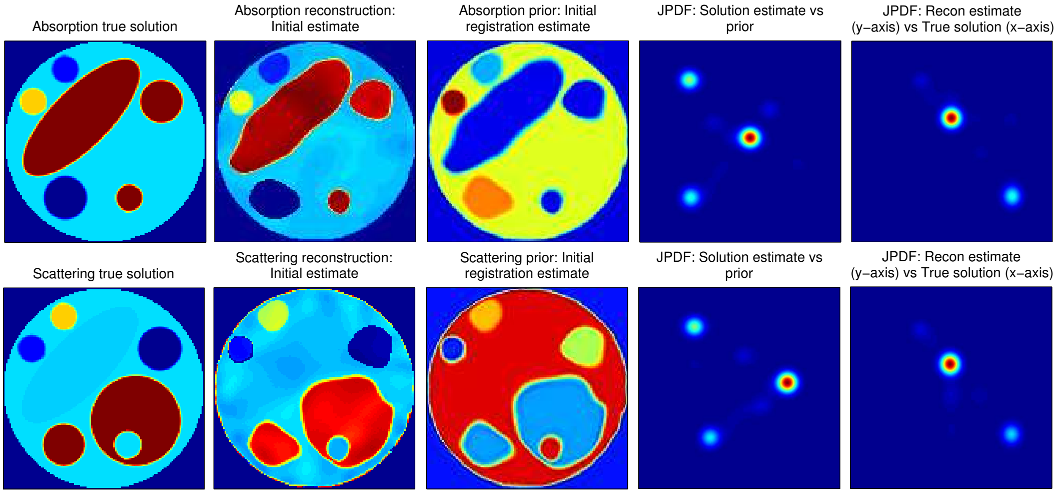

Absorption true solution
Absorption reconstruction: Initial estimate
Absorption prior: Initial registration estimate
JPDF: Solution estimate vs prior
JPDF: Recon estimate (y–axis) vs True solution (x–axis)
Scattering true solution
Scattering reconstruction: Initial estimate
Scattering prior: Initial registration estimate
JPDF: Solution estimate vs prior
JPDF: Recon estimate (y–axis) vs True solution (x–axis)

Figure 8.14: Case 2: results obtained by the simultaneous reconstruction/registration scheme. It is evident that the registration has further improved from the best possible initialization estimate. All JPDFs exhibit increased clustering which indicates reduction in *JE*.

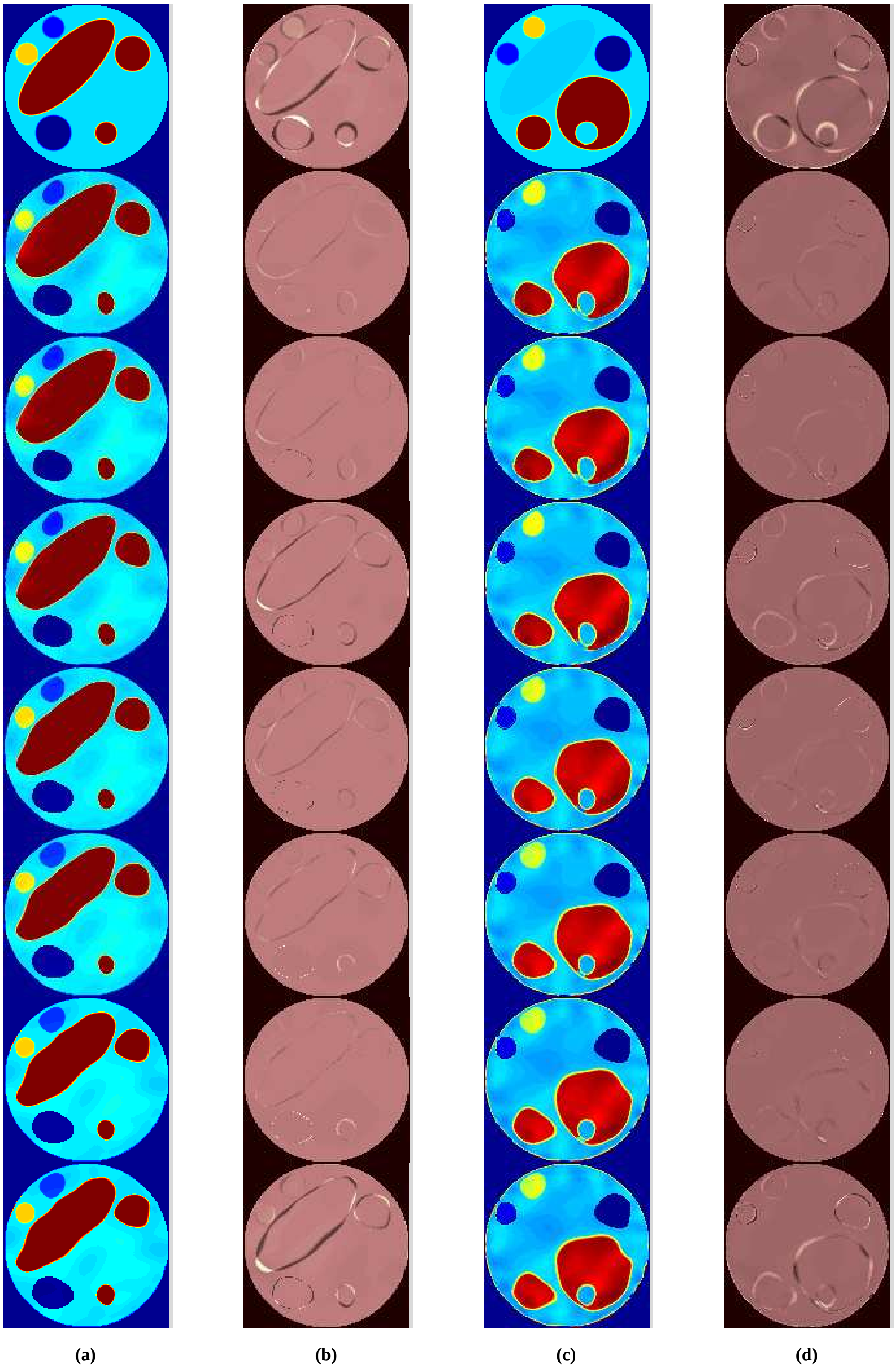


Figure 8.15: Case 2: SRR output at the various FFD grid refinement/reset levels. **Cols. (a),(c):** μ_a, μ_s' **Row 1:** $x^\star$ **Row 2-4:** SRR $G_1 = [32, 32]$. Resets $G_1^1,\ G_1^2,\ G_1^3$. **Rows 5-6:** SRR $G_2 = [16, 16]$. Resets $G_2^1,\ G_2^2$. **Rows 7-8:** SRR $G_3 = [8, 8]$. Resets $G_3^1,\ G_3^2$. **Cols. (b),(d):** Intra-row differences in cols. (a),(b). **Row 1:** $\hat{x}_{SRR}$-$x^\star$ [8-1] (also interpreted as registration accuracy as $\hat{x}_{SRR} \to (x_{\text{ref}}^{\mathfrak{T}})_{SRR}$ in structural terms) **Rows 2-7:** Intra-SRR diff/s. [2-1], [3-2] etc. **Row 8:** $\hat{x}_{SRR} - \hat{x}_{\text{init}}$.

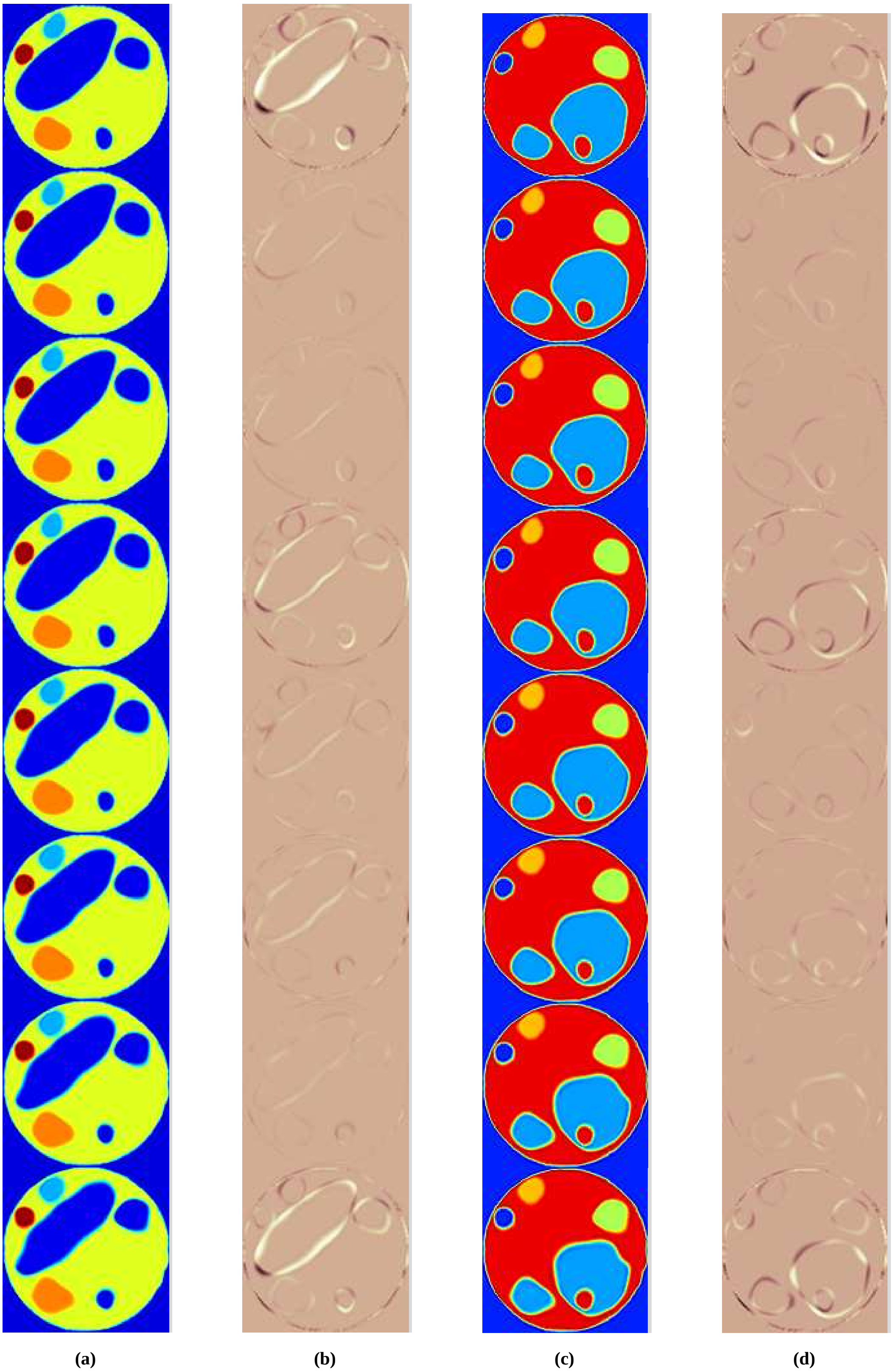


Figure 8.16: Case 2: SRR output at the various FFD grid refinement/reset levels. **Cols. (a),(c):** $x_{\text{ref}}^{\mu_a}$, $x_{\text{ref}}^{\mu_s}$ **Row 1:** $(x_{\text{ref}}^{\mathfrak{T}})_{\text{init}}$ **Row 2-4:** SRR $G_1 = [32, 32]$. Resets $G_1^1,\ G_1^2,\ G_1^3$. **Rows 5-6:** SRR $G_2 = [16, 16]$. Resets $G_2^1,\ G_2^2$. **Rows 7-8:** SRR $G_3 = [8, 8]$. Resets $G_3^1,\ G_3^2$. **Cols. (b),(d):** Intra-row differences in cols. (a),(b). **Row 1:** $(x_{\text{ref}}^{\mathfrak{T}})_{SRR}$-$(x_{\text{ref}}^{\mathfrak{T}})_{\text{init}}$ (magnitude of total recovered alignment) [8-1] **Rows 2-7:** Intra-SRR diff/s. [2-1], [3-2] etc. **Row 8:** $(x_{\text{ref}}^{\mathfrak{T}})_{SRR} - (x_{\text{ref}}^{\mathfrak{T}})_{\text{init}}$.

8.4.3 Case study 3

Finally, in this case we do not consider the boundary of x as a known quantity. The same circular domain of image x is now solely considered as $\Omega_{\mathbf{x}}$, where both regularisation and registration similarity measures are evaluated. Of course, in case of partial overlap between the full (square) $\Omega_{x^{\mathfrak{T}}_{\text{ref}}}$ and $\Omega_{\mathbf{x}}$, it holds that $\Omega_{x^{\mathfrak{T}}_{\text{ref}}} = \Omega_{\mathbf{x}} \cap \Omega_{x^{\mathfrak{T}}_{\text{ref}}}$. We use the same initialization used in Case 1 (see Fig. 8.8) (10 TV iterations/ full affine registration) as well as the same smoothness penalty weighting. The obtained results from this case are presented in Fig. 8.17. It is evident that, as in the previous cases, that $\hat{x}_{SRR}$ improves on the converged $\hat{x}_{TV}$ (see Fig. 8.13). Table 8.3 reports the corresponding errors. Reconstruction improves by 19.5% whereas registration by 14.7%, according to the employed L_2 norm.

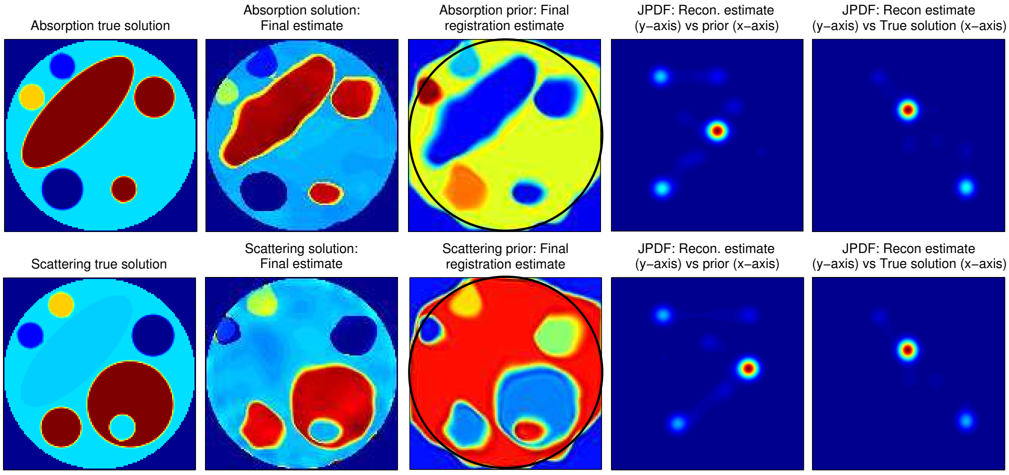


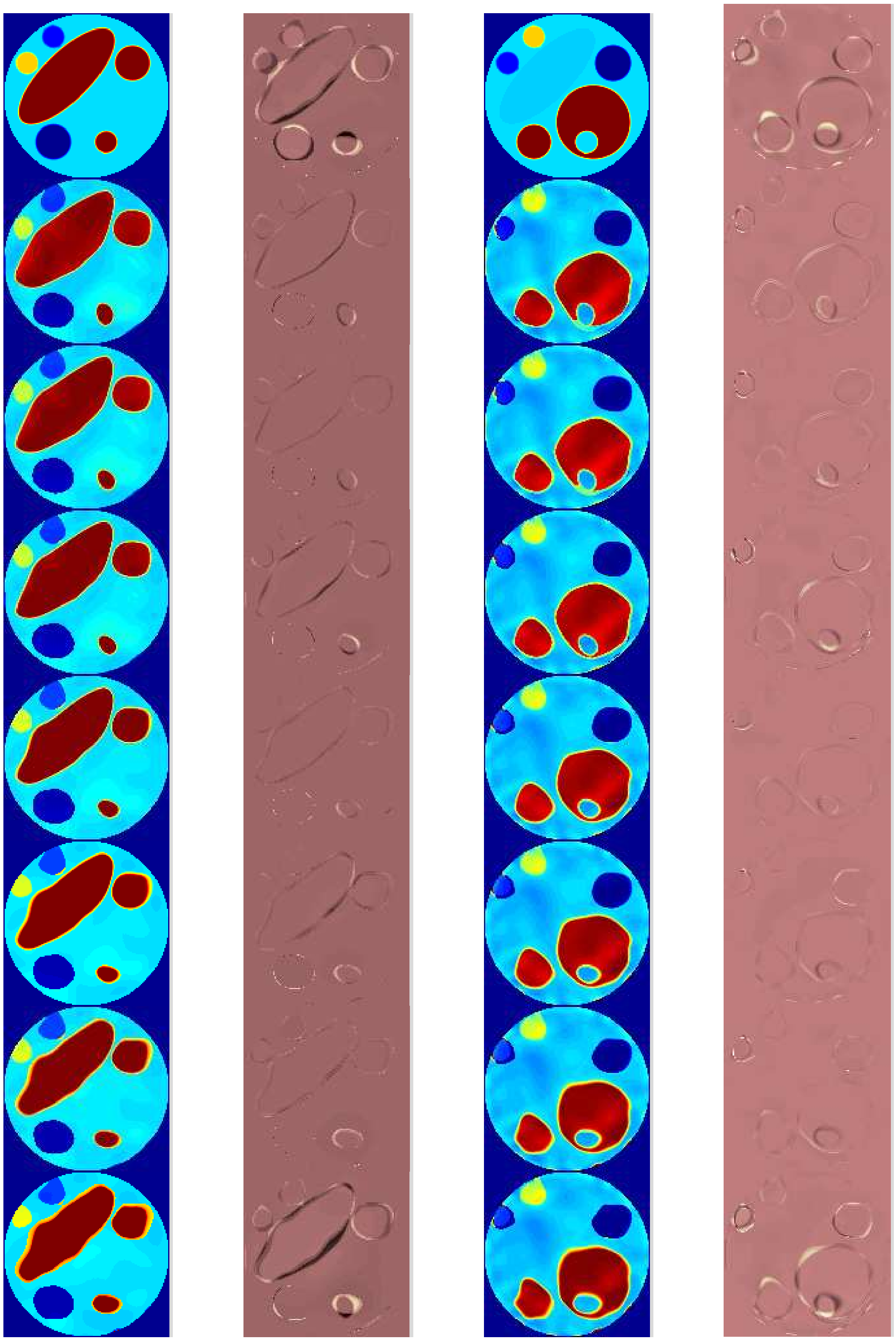
Figure 8.17: Case 3: results obtained by the Simultaneous reconstruction/registration scheme. Registration between corresponding features has been largely established. Local non-rigid transformations have significantly compromised the resolution of the prior at feature boundaries. The obtained All JPDFs exhibit increased clustering which indicates reduction in *JE*.

Table 8.3: Case 3: Normalized L_2 errors: 1) $x^{\star}$ vs converged TV-based $\hat{x}_{\textbf{init}}$ 2) $x^{\star}$ vs $\hat{x}_{SRR}$ 3) $x^{\star}_{\textbf{ref}}$ vs $(x^{\mathfrak{T}}_{\textbf{ref}})_{\textbf{init}}$ 4) $x^{\star}_{\textbf{ref}}$ vs $(x^{\mathfrak{T}}_{\textbf{ref}})_{SRR}$

Images considered	Normalized L2 error
$x^{\star}$, $\hat{x}_{\text{init}}$	16.8%
$x^{\star}$, $\hat{x}_{SRR}$	12.9%
$x^{\star}_{\text{ref}}$, $(x^{\mathfrak{T}}_{\text{ref}})_{\text{init}}$	28.6%
$x^{\star}_{\text{ref}}$, $(x^{\mathfrak{T}}_{\text{ref}})_{SRR}$	20.8%

As boundaries are not considered by the similarity measure, the initial circular boundary of $x^{\mathfrak{T}}_{\text{ref}}$ is now largely deformed compared to the $x^{\star}_{\text{ref}}$. The latter is artificially added on the depicted final $(x^{\mathfrak{T}}_{\text{ref}})_{SRR}$ to assist visual comparison. Achieving accurate registration in this case is not trivial. The two features at the top-left of the circular domain of x are in close proximity with the boundary. This means that if the initial information in $\hat{x}_{\text{init}}$ is such that it drives them outside of the boundary, this case is very difficult to recover from. The reason is that they are no longer included in the registration similarity evaluation.

In addition, we should comment on the effects of the non-rigid transformations on the boundary of the depicted features in $(x_{\text{ref}}^{\mathfrak{T}})_{SRR}$, which is clearly erroneous. The registration process tends to promote the expansion of the distinctive (3-4 pixels wide) boundaries enclosing features in $(x_{\text{ref}}^{\mathfrak{T}})_{\text{init}}$. The same boundary is more evident in $x^{\star}$ (white line encompassing most features). A possible explanation of why these effects are more noticeable in this case is due to the exclusion of the boundary from the similarity evaluation. The alignment of the boundary corresponds to a strong minimizer in the solution space as it is unequivocally known in both images. Thus, the boundary registration is greatly favoured by the registration. As the moving boundary becomes fixed on the target one, then all the pixels in the interior subjected to deformations which push them against the boundary, are penalized by the smoothness penalty. A higher smoothness penalty could possibly alleviate this problem. This is still a topic open for research. Finally, the intra-SRR output images are presented in Figs. 8.18-8.19. SRR error plots are not provided for this case.



(a) (b) (c) (d)

Figure 8.18: Case 3: SRR output at the various FFD grid refinement/reset levels. Output in $\Omega_{\mathbf{x};\mathbf{y}^{\mathfrak{T}}}$ only. **Cols. (a),(c):** μ_a, μ_s' **Row 1:** $x^\star$ **Row 2-4:** SRR $G_1 = [32, 32]$. Resets G_1^1, G_1^2, G_1^3. **Rows 5-6:** SRR $G_2 = [16, 16]$. Resets G_2^1, G_2^2. **Rows 7-8:** SRR $G_3 = [8, 8]$. Resets G_3^1, G_3^2. **Cols. (b),(d):** Intra-row differences in cols. (a),(b). **Row 1:** $\hat{x}_{SRR}$-$x^\star$ [8-1] (also interpreted as registration accuracy as $\hat{x}_{SRR} \rightarrow (x^{\mathfrak{T}}_{\text{ref}})_{SRR}$ in structural terms) **Rows 2-7:** Intra-SRR diff/s. [2-1], [3-2] etc. **Row 8:** $\hat{x}_{SRR} - \hat{x}_{\text{init}}$.

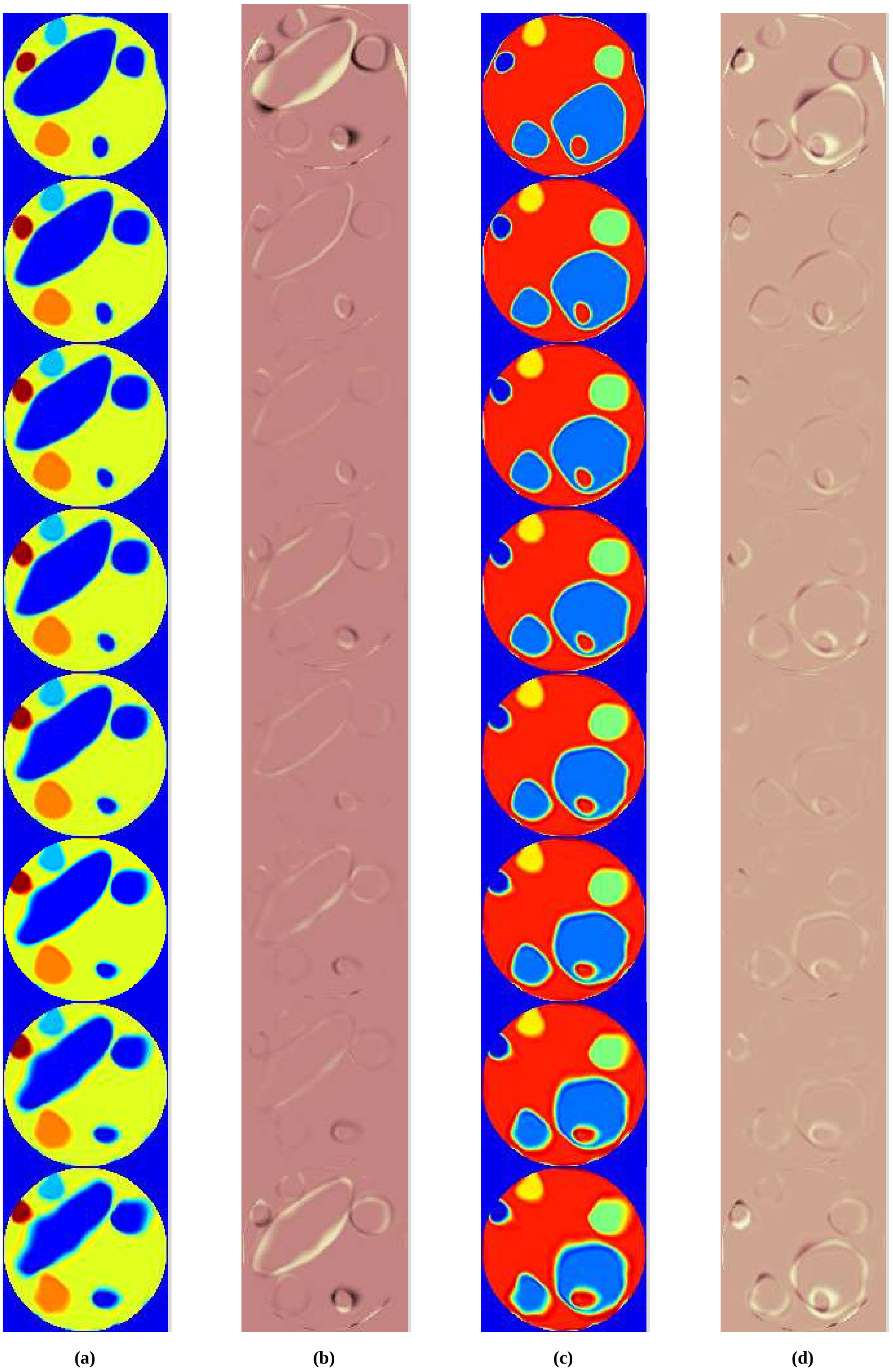


Figure 8.19: Case 3: SRR output at the various FFD grid refinement/reset levels. Output in $\Omega_{\mathbf{x};\mathbf{y}^{\mathfrak{T}}}$ only. **Cols. (a),(c):** $x_{\text{ref}}^{\mu_a}$, $x_{\text{ref}}^{\mu_s}$ **Row 1:** $(x_{\text{ref}}^{\mathfrak{T}})_{\text{init}}$ **Row 2-4:** SRR $G_1 = [32, 32]$. Resets G_1^1, G_1^2, G_1^3. **Rows 5-6:** SRR $G_2 = [16, 16]$. Resets G_2^1, G_2^2. **Rows 7-8:** SRR $G_3 = [8, 8]$. Resets G_3^1, G_3^2. **Cols. (b),(d):** Intra-row differences in cols. (a),(b). **Row 1:** $(x_{\text{ref}}^{\mathfrak{T}})_{SRR}$-$(x_{\text{ref}}^{\mathfrak{T}})_{\text{init}}$ (magnitude of total recovered alignment) [8-1] **Rows 2-7:** Intra-SRR diff/s. [2-1], [3-2] etc. **Row 8:** $(x_{\text{ref}}^{\mathfrak{T}})_{SRR} - (x_{\text{ref}}^{\mathfrak{T}})_{\text{init}}$.

8.5 Comparison among cases

In this section we compare the results obtained from the three discussed cases. The columns of Fig. 8.20 show the regions of interest from (i) $x^{\star}$, (ii) $\hat{x}_{JE}$ with a registered prior $\big($we use the reconstructions obtained by the application of the prior with extra features (iii-v) shows the $\hat{x}_{SRR}$ for the three discussed cases, in the order which they were presented and **6)** a converged $\hat{x}_{TV}$.

Figure 8.21 present errors among the corresponding region of interests (ROIs) of Fig. 8.20. Subfigures 8.21a-8.21b show the bias error for each region defined as $(\bar{z} - \bar{z}^{\star})/\bar{z}^{\star})$, where z denotes the reconstructions from the various schemes mentioned above and $z^{\star}$ the true optical solutions. Negative values correspond to underestimation of a region's mean compared to $x^{\star}$.

Similarly, Figs. 8.21c-8.21d show the variance error, defined as $(Var\,(z) - Var\,(z^{\star}))/Var\,(z^{\star})$. We do not plot the variance of each region but rather its difference from the true variance. The reason for this is that although $x^{\star}$ is seemingly piecewise constant, its ROIs have the distinctive high-contrast ring surrounding them, therefore the variance of these regions is not zero. However, they are are still small enough to produce large discrepancies when compared against the variances of the corresponding reconstructed ROI, where the latter are not homogeneous. Thus in this setting, variance or even the variance difference which is shown, is not the most suitable indication for evaluating the accuracy of reconstructions. However, we have chosen to include the plots for completeness.

Finally, Figs. 8.21e-8.21f show the normalized L_2 error among the corresponding pixels of corresponding regions, defined as $\|z(r) - z^{\star}(r)\| \,/\, \|z^{\star}(r)\|$, $\forall r$ pixels in a the considered region.

Table 8.4 simply plots the mean value from the corresponding plots of all regions, for a single method. For example, in the case of bias, the provided value is the average bias in the whole image, computed as the mean of the regionally reported biasses for that method, for both μ_a and μ_s'. All errors are shown as percentages. Once again the large variance errors are not strong indicators of the reconstructions' accuracy as the reconstructed images are compared to nearly piecewise constant targets. They are provided however for completeness. All SRR methods outperform the TV. This hold for bias and L_2, however not for variance as the TV reconstruction is smoother and does not reconstruct the erroneous high-contrast boundary around features.

Table 8.4: Inter-regional mean errors

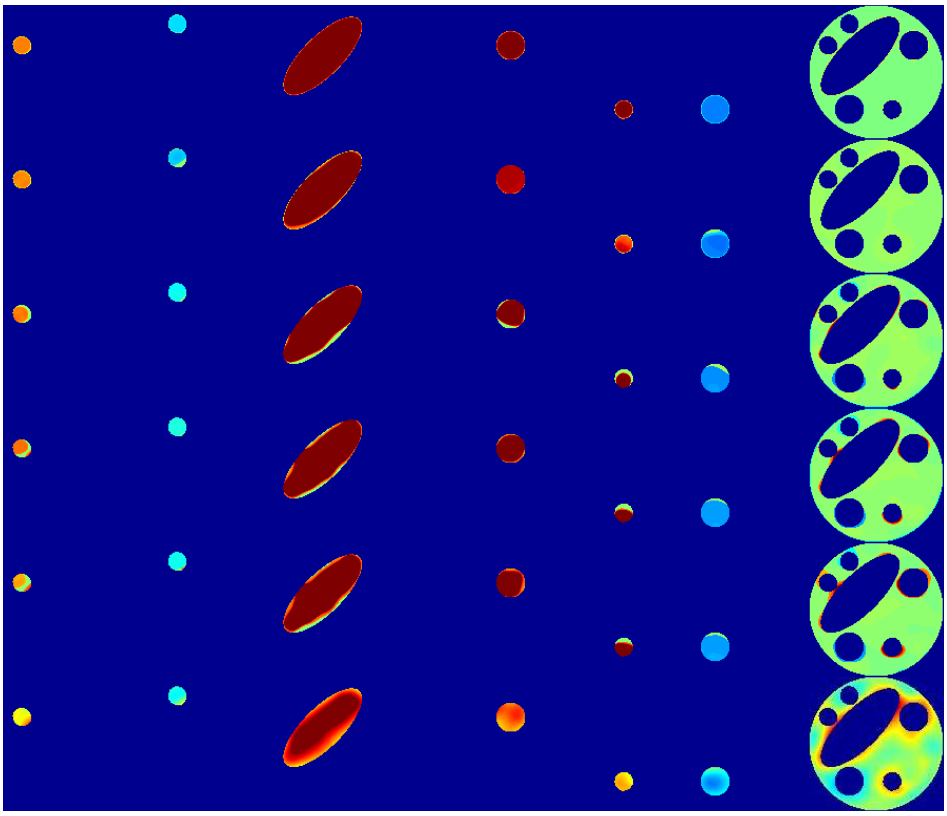
Recon. scheme	Δ-Bias	Δ-Variance	L_2
$\hat{x}_{JE}$	5.92%	276.9 %	12.19%
$\hat{x}_{SRR}$ (case 1)	5.99 %	1230.5%	17.14%
$\hat{x}_{SRR}$ (case 2)	6.09 %	1049.9%	16.23%
$\hat{x}_{SRR}$ (case 3)	6.23 %	846.87%	16.40%
$\hat{x}_{TV}$ (case 3)	15.69 %	712.39%	20.51%

8.6 Discussion

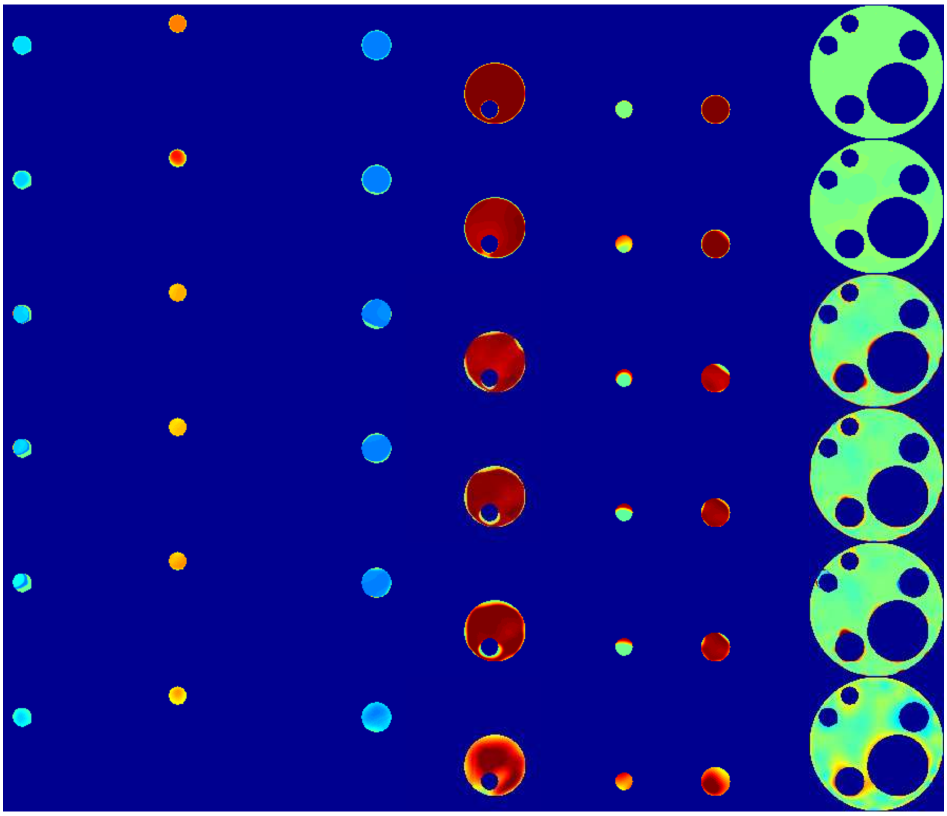
All presented cases used the same randomly deformed reference image. Different priors need to be tested in the future, such as non-piecewise constant priors, priors with no one-to-one feature correspondence (incorrect content) or highly mis-registered priors. Although we would have preferred to show statistical results with multiple randomly generated priors or target images, the current MATLAB based registration scheme still results in prohibitive run-times for such large-scale analysis. It should be noted however that *any* mis-registered prior which can be adequately registered against the $\hat{x}_{\text{init}}$ should produce similar results such as the ones presented. A more interesting case would be to test more complex DOT target images, such as the 2D breast simulation in Chapter 7. The important finding in this chapter - product of the comparison of Case studies 1 & 2 - is that there is a strong indication that SRR schemes further improve registration from its best possible initialization estimate. This indication alone renders SRR schemes worthy of further research.

8.7 Summary

In this chapter we have proposed a simultaneous registration/reconstruction scheme toward the incorporation of unregistered priors in DOT. We have presented preliminary results which act as indicators towards the validity of the principle of SRR. Further analysis has to take place to characterize the capacity of the scheme to perform in cases of prior with inaccurate content; increased initial mis-registration; acquired optical data with higher percentages of contamination and finally more complex target optical solutions. A fully functional scheme would enable the incorporation of *a priori* information from generic population based probabilistic atlases. This would render the latter as potential priors in atlas-to-subject multi-modality imaging. In the case of intra-subject multi-modality imaging, a single high-resolution image of the probed anatomy, could be used repetitively as a prior for subsequent DOT studies.

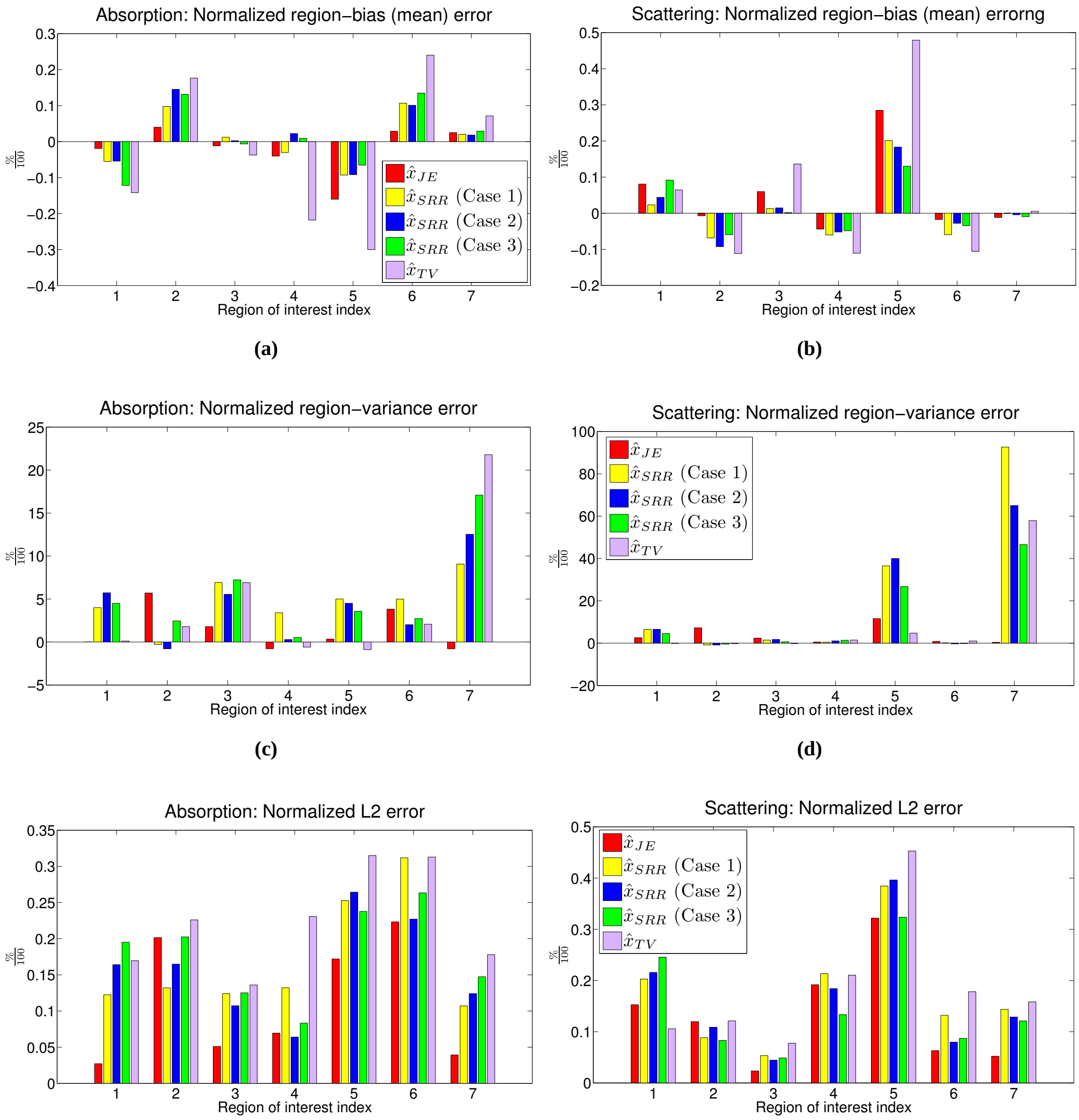


(a)



(b)

Figure 8.20: Regions of interest: Cases 1 & 2. **Subfig.** 8.20a: absorption, **Subfig.** 8.20b: scattering. **Rows**: $x^\star$, $\hat{x}_{JE}$ (reconstruction of 7.5 - prior with extra features), $\hat{x}_{SRR}$ (cases 1 | 2 | 3), $\hat{x}_{TV}$ **Columns** Regions of interest



(a) **(b)**

(c) **(d)**

(e) **(f)**

Figure 8.21: Region of interest: Bias and Variance error. Errors corresponds to the regions depicted in Fig. 8.20. The y-axis corresponds to error $(\%)/100$. The bias error is defined as $(\bar{z} - \bar{z}^\star)/\bar{z}^\star)$, where z denotes the reconstruction from the various schemes (see legend) and $z^\star$ the true optical solutions. Negative values correspond to underestimation. Similarly the variance error is defined as $(Var\,(z) - Var\,(z^\star))/Var\,(z^\star)$.

Part IV

Conclusion

Chapter 9

Summary and future directions

This thesis introduced information theoretic (IT) regularization in the context of diffuse optical tomography (DOT). The scheme enables the incorporation of structural *a priori* information from reference images, with gray values incommensurately related to the optical solution. In addition, the scheme was extended in order to enable the incorporation of spatially unregistered reference images, without using any *a priori* knowledge regarding their correct location. The proposed scheme was developed with emphasis on computational efficiency.

Chapters 2 to 5 introduced the underlying theoretical categories on which the propositions in this thesis have been based, namely *inverse problems* & regularization, DOT, IT and *medical image registration*.

In Chapter 6 we proposed a scheme for the efficient computation of joint entropy (JE) and its derivative in order to enable IT regularization in tractable run times. The proposed scheme extended a method initially proposed by Shwartz et al. [2005], enabling the efficient evaluation and derivative computation of the entropy of a single random variable from its samples. In addition, we characterized two possible implementations of entropy, namely the standard integral formulation by Shannon as well as its formulation as an expectation - termed empirical entropy. Finally we evaluated the accuracy of the obtained derivatives and validated the current implementation.

In Chapter 7 we presented the proposed IT scheme of DOT. The functionals of JE and mutual information (MI) were considered as candidates. A detailed analysis on their theoretical capacity to perform as regularizing functionals was presented. The findings were consistent with the outcomes of numerical simulations specifically designed to test IT regularization in complex cases. The proposed scheme in all cases managed to improve the solutions obtained using generic regularization schemes such as total variation (TV) or first-order Tikhonov (TK_1) regularization. In addition to numerical simulations, the method was tested on experimental data. The task involved the reconstruction of the optical properties of a phantom with both optical and magnetic contrast. A magnetic resonance imaging (MRI) scan depicting the magnetic structure of the phantom was used as the structural *a priori* information to be introduced by the proposed scheme. The magnetic properties of the phantom were specifically designed to produce an MRI image which did not have a one-to-one feature correspondence with the target optical solution. Considering the quality of the optical data itself, the task exhibited an increased level of difficulty. The functionals managed to improve the reconstruction at regions where the prior support was consistent with the underlying optical solution, however unwanted bias was also observed.

Finally, in Chapter 8 we proposed a scheme towards the incorporation of unregistered prior information in DOT. The majority of methods in the literature assume that the spatial co-registration of the prior image and the underlying optical solution is guaranteed at initialization, usually through concurrent probing of the target anatomy from the modality which provides the high-resolution reference images, as well as the modality which seeks to benefit from them. This condition however cannot always be guaranteed and it is disabling when one considers tasks such as the incorporation of *a priori* information from probabilistic atlases, where the notion of concurrent image does not exist. The proposed scheme involves a simultaneous reconstruction/registration (SRR) approach which compensates for potential misalignments of the prior image with respect to the optical solution, in real time and without preconditions. For the purpose of this task, we examined viable choices for similarity measures and we found that conditional entropy can adequately perform for the given task. It behaves as joint entropy in the image reconstruction setting while it improves on JE with respect to its capacity as a similarity measure in the image registration context. It has to be said however that the functional is less capable than the MI or normalized mutual information (NMI) as it is more prone to be affected by local minima in the solution space. We have tested the scheme in a series of numerical simulations and the obtained preliminary results are a positive indicator regarding the validity of the approach. Further research is however necessary in order to test the extent of misalignments which can be compensated by the scheme as well as its robustness to consistently perform in the severely ill-posed setting of DOT.

We have presented an extensive study, both in terms of theory coverage and testing of IT regularization in DOT. It is important to outline a number of identified limitations of the presented study. In terms of the scope of the work undertaken, the study of entropic regularization has been conducted under the popular independent and identically distributed (i.i.d.) assumption regarding the random variables (RVs) considered in the problem. No spatial inter-pixel dependence within a single image has been modeled. In addition, kernel density estimation (KDE) estimation was conducted with kernels of fixed width, where the latter was estimated via exhaustive search in initial pilot studies. Given the issues which were covered, there are a number of potential sources of error which have not been investigated, or at least not in full. Regarding the entropic regularization, it would be useful to conduct a dedicated study in cases of target distributions which are not piece-wise constant. More complex target distributions can potentially compromise the structural invariance of JE, which appears when the prior image has features which do not exist in the solution. Regarding the SRR scheme, we have not tested the framework in the presence of mismatches between the prior and the target solution, in terms of the features present in both distributions. The absence of one-to-one feature correspondence between the two distributions, can introduce local optima in the solution space and compromise both registration and regularization accuracy. The case can become even more complicated if one considers global mismatches between the distributions, such as the presence of gradient fields in one of the image or noise contamination of the prior images[1]

[1] We should note that the presented studies included data noise artificially added to the optical data, but the prior images were piecewise constant

undergoing registration against the optical images. Finally, we have not quantified the possible extend of mis-registration which can be recovered by the scheme, although such task might be impossible given the complexity of the parameters which affect its performance, such as the very structure of images, mis-registration level, ill-posedness levels, noise - all considered simultaneously. Finally, we have to refer to potential limitations which may arise in the practical/clinical application of the method. Clinical multi-modal images exhibit large variations and the level of mismatch between prior and optical images is expected to be high. In the case of simultaneous imaging (registration is guaranteed) one can force the structure of the prior to the solution and expect the optical system to populate it with the best possible optical parameters. We have seen however that as ill-posedness increases, JE ability to enforce structure decreases - due to the i.i.d. assumption. The clinical setting can indeed be severely ill-posed (see Sec. 3.8). Finally, regarding the application of the SRR scheme in a clinical setting, its performance has to be quantified. Some mis-registration can be indeed too high for the scheme to be recovered. For example, breast X-Ray mammography compresses the breast during imaging whereas DOT does not. Such high levels of mis-registration might require special treatment or specific effort to establish the best possible initialization estimate prior to SRR. In the case of using probabilistic atlases of the neonatal brain as priors, mis-registrations can be expected to be more localized and we are more hopeful and excited to test the method in this context.

9.1 Potential improvements

9.1.1 Kernel density estimation

The trivial KDE employed in this work for the purpose of entropy estimation, centres Gaussian kernels of equal bandwidth, over each data point - in this context the gray values of the considered images. More accurate KDE methods can be used where the global width of all Gaussian kernels is considered as an optimized quantity. Such a scheme was recently proposed by [Kazantsev et al., 2010]. An even better approach would be to consider locally adaptive KDEs [Silverman and Green, 1986]. These methods place kernels of variable width at the different data points. This added flexibility allows the accurate modeling of problematic long-tailed densities, by reducing the width of the kernels at data points corresponding to low density regions in the probability density function (PDF), whereas broader kernels are used at high density areas. However, we have not yet established if such methods can be accelerated by the usage of fast Fourier transform (FFT) which is crucial to enable the optimization of entropy in tractable times.

9.1.2 Modeling of intra-image spatial dependency among gray values

The employed KDE conveniently treats the gray values of an image as i.i.d.. This assumption does not reflect reality, as distinct anatomical regions are populated by similar gray values. Hence, one would expect that the reconstructed gray value of a pixel is conditioned by the anatomical region which it resides on. There are a number of recently developed methods in information theoretic regularization of other modalities, which consider intra-image spatial gray value dependence and which we can potentially employ in the future in the DOT scheme. These include the implicit modeling of intra-image spatial dependency by Somayajula et al. [2010] in positron emission tomography (PET) ; the class-conditional entropic regularization proposed by Pedemonte et al. [2010a] and the modeling of spatial dependence proposed by Van de Sompel and Sir., Brady, M. [2009a] via the incorporation of a smoothness prior. We intend to incorporate the explicit modeling of intra-image spatial dependency in the current information theoretic regularization scheme of DOT in the near future. In case such methods are adopted, the efficient entropy and derivative evaluation scheme has to be re-visited.

9.1.3 Optimization

An area of potential improvement concerns the employed optimization schemes. In this work we have employed the first-order, gradient based, iterative optimization scheme of conjugate gradients (CG). The selection of this particular approach over potential alternatives was made after considering a number of factors. Firstly, the objective function in the inverse problem of DOT includes the forward problem (see Eq. 3.60). The computational complexity of the latter can be significantly high, especially in the 3D case. Therefore the optimization scheme has to make as few forward problem evaluations per iteration as possible. The non-gradient methods described in Sec. 2.6.1 require multiple objective function evaluations, rendering their choice inefficient in computational terms. Gradient based schemes successively improve a single initial solution estimate, resulting to a single forward evaluation per iteration. The employment of more advanced gradient-based optimization schemes such as the second-order methods described in Chapter 2 are known to establish convergence in less iterations and higher accuracy, given that the solution space in the vicinity of the optimum resembles a quadratic basin. However, to incorporate such change, we are required to derive the analytic second derivatives of the entropic functionals with respect to the image's gray values. The application of the FFT for the purpose of reducing computational complexity would need to be re-established.

Regarding the simultaneous reconstruction/registration scheme one can pursue the employment of analytic derivatives of the information theoretic similarity measures with respect to the transformation parameters. Analytic derivatives have been used in medical image registration in various studies [Modat et al., 2009; Viola, 1995].

9.1.4 Application to other modalities

In addition to the information theoretic regularization of DOT [Panagiotou et al., 2009b] and PET [Kazantsev et al., 2010; Pedemonte et al., 2010a; Somayajula et al., 2010], we have also obtained preliminary unpublished results from the application of information theoretic regularization in the linear inverse problem of *fluorescence diffuse optical tomography*. We have tested the method on studies based on numerical simulations. This setting introduces additional challenges as the obtained optical reconstructions are usually sparse. The implication of this is that potential anatomical images to be used as priors, would contain dense structural information for the entirety of the probed domain and not specifically for the non-zero regions in fluorescence reconstructions. Such a large scale lack of one-to-one feature correspondence between solution and prior could result in increased bias compared to the optical solution. Developing methods that minimize the bias due to prior/solution structural disagreement is a challenging task.

Information theoretic regularization its still in its infancy. In recent years however, the interest in the method is rapidly increasing. The concepts of JE and MI have proven their worthiness in medical image science due to their inherent ability of measuring similarity between images while by-passing the multi-modality barrier of incommensurately related gray values. It is our feeling that the increased attention on information theoretic regularization will continue and the scheme will evolve to become one of the dominants choices for regularization of inverse problems in imaging. —

Bibliography

Alerstam, E., Svensson, T., and Andersson-Engels, S. (2008). Parallel computing with graphics processing units for high-speed Monte Carlo simulation of photon migration. *J. Biomed. Opt.*, 13(6):060504. 73

Alpert, N. M., Bradshaw, J. F., Kennedy, D., and Correia, J. A. (1990). The principal axes transformation–a method for image registration. *J. Nucl. Med.*, 31(10):1717–1722. 117, 122

Arbel, T., Morandi, X., Comeau, R., and Louis Collins, D. (2001). Automatic non-linear MRI-ultrasound registration for the correction of intra-operative brain deformations. In Niessen, W. and Viergever, M., editors, *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2001*, volume 2208 of *Lecture Notes in Computer Science*, pages 913–922. Springer Berlin / Heidelberg. 10.1007/3-540-45468-3 109. 128

Ardekani, B. A., Braun, M., Hutton, B. F., Kanno, I., and Iida, H. (1996). Minimum cross-entropy reconstruction of PET images using prior anatomical information. *Phys. Med. Biol.*, 41:2497–2517. 85

Aronson, R. (1993). Extrapolation distance for diffusion of light. In Chance, B. and Alfano, R. R., editors, *Photon Migration and Imaging in Random Media and Tissues*, volume 1888, pages 297–305. Proc. SPIE. 67

Arridge, S. R. (1993). The forward and inverse problems in time-resolved infrared imaging. In Muller, G., Chance, B., Alfano, R., Arridge, S., Beuthan, J., Gratton, E., Kaschke, M., Masters, B., Svanberg, S., and van der Zee, P., editors, *Medical Optical Tomography: Functional Imaging and Monitoring*, pages 35–64. SPIE, Bellingham, WA. 80

Arridge, S. R. (1995). Photon measurement density functions. Part 1: Analytical forms. *Appl. Opt.*, 34(31):7395–7409. 76

Arridge, S. R. (1999). Optical tomography in medical imaging. *Inverse Problems*, 15(2):R41–R93. 26, 48, 62, 64, 65, 67, 69, 72, 73, 75, 76, 78, 85, 88

Arridge, S. R., Cope, M., and Delpy, D. T. (1992). Theoretical basis for the determination of optical pathlengths in tissue: Temporal and frequency analysis. *Phys. Med. Biol.*, 37:1531–1560. 73

Arridge, S. R., Dorn, O., Kolehmainen, V., Schweiger, M., and Zacharopoulos, A. (2008a). Parameter and structure reconstruction in optical tomography. In *J. Phys.: Conference Series 125,012001*. 81

Arridge, S. R. and Hebden, J. C. (1997). Optical imaging in medicine: II. Modelling and reconstruction. *Phys. Med. Biol.*, 42:841–853. 73

Arridge, S. R. and Lionheart, W. R. B. (1998). Non-uniqueness in diffusion-based optical tomography. *Opt. Lett.*, 23:882–884. 86

Arridge, S. R., Panagiotou, C., Schweiger, M., and Kolehmainen, V. (2008b). Multimodal diffuse optical tomography: Theory. In *Translational multimodality optical imaging*, chapter 5, pages 101–123. Artech Press. 31, 175

Arridge, S. R. and Schotland, J. C. (2009). Optical tomography: forward and inverse problems. *Inverse Problems*, 25(12):123010. 75, 88

Arridge, S. R. and Schweiger, M. (1993a). Inverse methods for optical tomography. In *Information Processing in Medical Imaging (IPMI'93 Proceedings), Lecture Notes in Computer Science*, volume 687, pages 259–277. Springer-Verlag, Berlin. 79

Arridge, S. R. and Schweiger, M. (1993b). The use of multiple data types in time-resolved optical absorption and scattering tomography (TOAST). In Wilson, J. N. and Wilson, D. C., editors, *Mathematical Methods in Medical Imaging II*, volume 2035, pages 218–229. Proc. SPIE. 69

Arridge, S. R. and Schweiger, M. (1995a). Photon measurement density functions. Part 2: Finite element calculations. *Appl. Opt.*, 34(34):8026–8037. 76

Arridge, S. R. and Schweiger, M. (1995b). Sensitivity to prior knowledge in optical tomographic reconstruction. In Chance, B. and Alfano, R. R., editors, *Optical Tomography, Photon Migration, and Spectroscopy of Tissue and Model Media: Theory, Human Studies, and Instrumentation*, volume 2389, pages 378–388. Proc. SPIE. 79, 80

Arridge, S. R. and Schweiger, M. (1998). A gradient-based optimisation scheme for optical tomography. *Opt. Express*, 2(6):213–226. 76, 178

Arridge, S. R., Schweiger, M., Hiraoka, M., and Delpy, D. T. (1993). A finite element approach for modeling photon transport in tissue. *Med. Phys.*, 20(2):299–309. 69, 71, 72

Arun, K. S., Huang, T. S., and Blostein, S. D. (1987). Least-squares fitting of two 3-D point sets. *IEEE Trans. Patt. Anal. Mach. Intell.*, 9(5):698–700. 128

Ash, R. B. (1990). *Information Theory*. Dover Publications, Inc. 90, 106

Audette, M. A., Ferrie, F. P., and Peters, T. M. (2000). An algorithmic overview of surface registration techniques for medical imaging. *Med. Im. Anal.*, 4(3):201 – 217. 117

Bajcsy, R. and Kovačič, S. (1989). Multiresolution elastic matching. *Comput. Vis. Graph. and Image Process.*, 46(1):1–21. 122

Bazaraa, M. S., Sherali, H. D., and Shetty, C. M. (1993). *Nonlinear Programming: Theory and Algorithms*. Wiley, New York, second edition. 54

Besl, P. J. and McKay, N. D. (1992). A method for registration of 3-D shapes. *IEEE Trans. Patt. Anal. Mach. Intell.*, 14(2):239–256. 128

Binzoni, T., Leung, T. S., Giust, R., Rüfenacht, D., and Gandjbakhche, A. H. (2008). Light transport in tissue by 3D Monte Carlo: Influence of boundary voxelization. *Comput Methods Programs Biomed.*, 89(1):14–23. 73

Björck, Å. (1996). *Numerical Methods for Least Squares Problems*. SIAM, Philadelphia. 46, 47

Boas, D., Culver, J., Stott, J., and Dunn, A. (2002). Three dimensional Monte Carlo code for photon migration through complex heterogeneous media including the adult human head. *Opt. Express*, 10(3):159–170. 74

Boas, D. A. (1996). *Diffuse Photon Probes of Structural and Dynamical Properties of Turbid Media : Theory and Biomedical Applications*. PhD thesis, University of Pennsylvania. 62

Boas, D. A., Brooks, D. H., Miller, E. L., DiMarzio, C. A., Kilmer, M., Gaudette, R. J., and Zhang, Q. (2001). Imaging the body with diffuse optical tomography. *IEEE Sig. Proc. Magazine.*, 18(6):57–75. 26, 28, 86, 88

Bohren, C. F. and Huffman, D. R. (1983). *Absorption and Scattering of Light by Small Particles*. Wiley, New York. 58

Bookstein, F. L. (1989). Principal warps: thin-plate splines and the decomposition of deformations. *IEEE Trans. Pattern Anal. and Mach. Intel.*, 11(6):567–585. 123

Boverman, G., Miller, E. L., Li, A., Zhang, Q., Chaves, T., Brooks, D. H., and Boas, D. A. (2005). Quantitative spectroscopic diffuse optical tomography of the breast guided by imperfect a priori structural information. *Phys. Med. Biol.*, 50:3941–3956. 79, 85

Bowsher, J., DeLong, D., Turkington, T., and Jaszczak, R. (2006). Aligning emission tomography and MRI images by optimizing the emission-tomography image reconstruction objective function. *IEEE Trans. Nucl. Sci.*, 53(3):1248 – 1258. 218

Bracewell, R. (1999). *The Fourier Transform & Its Applications*. McGraw-Hill, 3rd edition. 126, 141, 144, 155

Breeuwer, M., Wadley, J. P., de Bliek, H. L., Buurman, J., Desmedt, P. A., Gieles, P., Gerritsen, F. A., Dorward, N. L., Kitchen, N. D., Velani, B., Thomas, D. G., Wink, O., Blankensteijn, J. D., Eikelboom, B. C., Mali, W. P., Viergever, M. A., Penney, G. P., Gaston, R., Hill, D. L., Maurer, C. R., Hawkes,

D. J., Maes, F., Vandermeulen, D., Verbeeck, R., and Kuhn, M. H. (1998). The easi project–improving the effectiveness and quality of image-guided surgery. *IEEE Trans. Inf. Technol. Biomed.*, 2(3):156–168. 117

Brooksby, B., Jiang, S., Dehghani, H., Pogue, B. W., and Paulsen, K. D. (2005a). Combining near infrared tomography and magnetic resonance imaging to study in vivo breast tissue: implementation of a Laplacian-type regularization to incorporate MR structure. *J. Biomed. Opt.*, 10(5):0515041–10. 80, 85

Brooksby, B., Jiang, S., Dehghani, H., Pogue, B. W., Paulsen, K. D., Kogel, C., Doyley, M., Weaver, J. B., and Poplack, S. P. (2004). Magnetic resonance-guided near-infrared tomography of the breast. *Review of Sci. Inst.*, 75(12):5262–5270. 85

Brooksby, B., Pogue, B., Jiang, S., Dehghani, H., Srinivasan, S., Kogel, C., Tosteson, T. D., Weaver, J., Poplack, S., P., S., and Paulsen, K. D. (2006). Imaging breast adipose and fibroglandular tissue molecular signatures by using hybrid MRI-guided near-infrared spectral tomography. *Proc. of the National Academy of Science*, 103:8828–8833. 218

Brooksby, B., Srinivasan, S., Jiang, S., Dehghani, H., Pogue, B. W., Paulsen, K. D., Weaver, J., Kogel, C., and Poplack, S. P. (2005b). Spectral priors improve near-infrared diffuse tomography more than spatial priors. *Opt. Lett.*, 30(15):1968–1970. 84

Brown, L. G. (1992). A survey of image registration techniques. *ACM Comput. Surv.*, 24(4):325–376. 115

Burch, S. F., Gull, S. F., and Skilling, J. (1983). Image Restoration by a Powerful Maximum Entropy Method. *Comput. Vis. Graph. and Image Process.*, 23:113–128. 45, 107

Cahill, N. D., Schnabel, J. A., Noble, J. A., and Hawkes, D. J. (2008). Revisiting overlap invariance in medical image alignment. In *Proc. IEEE Computer Society Conf. Computer Vision and Pattern Recognition Workshops CVPRW '08*, pages 1–8. 222

Camarillo, D. B., Krummel, T. M., and Salisbury, J. K. (2004). Robotic technology in surgery: past, present, and future. *Am. J. Surg.*, 188(4A Suppl):2S–15S. 115

Candès, E. J., Wakin, M., and Boyd, S. (2007). Enhancing sparsity by reweighted l_1 minimization. *J. of Fourier Analysis and Applications*, 14:877–905. 80

Cao, N., Nehorai, A., and Jacobs, M. (2007). Image reconstruction for diffuse optical tomography using sparsityregularization and expectation-maximization algorithm. *Opt. Express*, 15(21):13695–13708. 80

Chandrasekhar, S. (1950). *Radiative Transfer*. Oxford University Press, London. 62, 63

Cheong, W., Prahl, S. A., and Welch, A. J. (1990). A review of the optical properties of biological tissues. *IEEE J. Quantum Electron.*, 26(12):2166–2185. 85

Christensen, G. E., Rabbitt, R. D., and Miller, M. I. (1996). Deformable templates using large deformation kinematics. *IEEE Trans. Im. Proc.*, 5(10):1435–1447. 122

Chui, E. (2008). *Discrete and continuous Fourier Transforms: Analysis applications and fast algorithms*. Chapman & Hall/CRC Press). 140

Chui, H. and Rangarajan, A. (2003). A new point matching algorithm for non-rigid registration. *Comput. Vis. and Image Understanding*, 89(2-3):114 – 141. Nonrigid Image Registration. 117

Ciarlet, P. G. and Lions, J. L., editors (1991). *Finite Element Methods (Part 1)*, volume 2 of *Handbook of numerical analysis*. Elsevier Science Publishers B. V. 69

Collignon, A., Maes, F., Delaere, D., Vandermeulen, D., Suetens, P., and Marchal, G. (1995). Automated multi-modality image registration based on information theory. *Inf. Proc. in Med. Im.*, pages 263–274. 129

Commowick, O., Arsigny, V., Costa, J., Ayache, N., and Malandain, G. (2006). An efficient locally affine framework for the registration of anatomical structures. In *Biomedical Imaging: Nano to Macro, 2006. 3rd IEEE International Symposium on*, pages 478 –481. 118

Comtat, C., Kinahan, P., Fessler, J. A., Beyer, T., Townsend, D. W., Defrise, M., and Michel, C. (2002). Clinically feasible reconstruction of 3D whole body PET/CT data using blurred anatomical labels. *Phys. Med. Biol.*, 47:1–20. 85

Cooley, J. W. and Tukey, J. W. (1965). An algorithm for the machine calculation of complex fourier series. *Mathematics of Computation*, 19(90):297–301. 141

Corlu, A., Choe, R., Durduran, T., Lee, K., Schweiger, M., Arridge, S. R., Hillman, E. M. C., and Yodh, A. G. (2005). Diffuse optical tomography with spectral constraints and wavelength optimisation. *Appl. Opt.*, 44(11):2082–2093. 84

Corlu, A., Durduran, T., Choe, R., Schweiger, M., Hillman, E., Arridge, S. R., and Yodh, A. G. (2003). Uniqueness and wavelength optimization in continuous-wave multispectral diffuse optical tomography. *Opt. Lett.*, 28:23. 84

Correia, T. M. M. (2010). *Assessment and optimisation of 3D Optical Tomography for brain imaging*. PhD thesis, University College London. 27, 28, 88

Coselmon, M. M., Balter, J. M., McShan, D. L., and Kessler, M. L. (2004). Mutual information based CT registration of the lung at exhale and inhale breathing states using thin-plate splines. *Med. Phys.*, 31(11):2942–2948. 115

Cover, T. M. and Thomas, J. A. (1991). *Elements of information theory*. Wiley-Interscience, New York, NY, USA. 45, 90, 103, 104, 105, 106, 146, 166

Crum, W. R., Griffin, L. D., Hill, D. L. G., and Hawkes, D. J. (2003). Zen and the art of medical image registration: correspondence, homology, and quality. *Neuroimage*, 20(3):1425–1437. 118

Crum, W. R., Hartkens, T., and Hill, D. L. G. (2004). Non-rigid image registration: theory and practice. *Br. J. Radiol.*, 77 Spec No 2:S140–S153. 117, 122

Culver, J., Akers, W., and Achilefu, S. (2008). Multimodality molecular imaging with combined optical and SPECT/PET modalities. *J. Nucl. Med.*, 49(2):169–172. 85

D'Agostino, E., Maes, F., Vandermeulen, D., and Suetens, P. (2003). A viscous fluid model for multimodal non-rigid image registration using mutual information. *Med. Im. Anal.*, 7(4):565 – 575. Medical Image Computing and Computer Assisted Intervention. 122

D'Agostino, E., Maes, F., Vandermeulen, D., and Suetens, P. (2007). Atlas-to-image non-rigid registration by minimization of conditional local entropy. *Inf. Proc. in Med. Im.*, 20:320–332. 117

de Souza, M. A., Hebden, J. C., Gibson, A. P., Sauret, V., and Robson, S. (2006). Developing photogrammetric methods for acquiring realistic head surface models of newborn infants for optical tomography. In *J. Biomed. Opt.*, page SH28. Optical Society of America. 230

Dehghani, H., Carpenter, C. M., Yalavarthy, P. K., Pogue, B. W., and Culver, J. P. (2007). Structural a priori information in near-infrared optical tomography. In Azar, F. S., editor, *Proc. SPIE*, volume 6431, page 64310B. 80

Dehghani, H., Delpy, D. T., and Arridge, S. R. (1999). Photon migration in non-scattering tissue and the effects on image reconstruction. *Phys. Med. Biol.*, 44:2897–2906. 66

Dehghani, H., Pogue, B. W., Shudong, J., Brooksby, B., and Paulsen, K. D. (2003). Three-dimensional optical tomography: Resolution in small-object imaging. *Appl. Opt.*, 42(16):3117–3128. 83

Delpy, D. T., Cope, M., van der Zee, P., Arridge, S. R., Wray, S., and Wyatt, J. (1988). Estimation of optical pathlength through tissue from direct time of flight measurement. *Phys. Med. Biol.*, 33:1433–1442. 61, 69

Diezma-Iglesias, B., Ruiz-Altisent, M., and Barreiro, P. (2004). Detection of internal quality in seedless watermelon by acoustic impulse response. *Biosystems Engineering*, 88(2):221 – 230. 36

Dobson, D. C. and Santosa, F. (1996). Recovery of blocky images from noisy and blurred data. *SIAM J. Appl. Math.*, 56(4):1181–1198. 80

Douiri, A., Schweiger, M., Riley, J., and Arridge, S. R. (2005a). Adaptive diffusion regularization method of inverse problem for diffuse optical tomography. In Licha, K. and Cubeddu, R., editors, *Photon Migration and Diffuse-Light Imaging II*, volume 5859, pages (585916)1–11. Proc. SPIE. 81

Douiri, A., Schweiger, M., Riley, J., and Arridge, S. R. (2005b). Local diffusion regularization method for optical tomography reconstruction by using robust statistics. *Opt. Lett.*, 30:2439–2441. 81

Douiri, A., Schweiger, M., Riley, J., and Arridge, S. R. (2007). Anisotropic diffusion regularisation methods for diffuse optical tomography using edge prior information. *Meas. Sci. Tech.*, 18:87–95. 82

Duda, R. O., Hart, P. E., and Stork., D. G. (2001). *Pattern Classification*. John Wiley & Sons, Inc. 96, 97, 99, 100

Durduran, T., Choe, R., Culver, J. P., Zubkov, L., Holboke, M. J., Giammarco, J., Chance, B., and Yodh, A. G. (2002). Bulk optical properties of healthy female breast tissue. *Phys. Med. Biol.*, 47(16):2847–2861. 85

Egan, W. G. and Hilgeman, T. W. (1979). *Optical Properties of Inhomogeneous Materials*. Academic, New York. 67

Enfield, L., Gibson, A., Everdell, N., Delpy, D., Schweiger, M., Arridge, S., Richardson, C., Keshtgar, M., Douek, M., and Hebden, J. (2007). Three-dimensional time-resolved optical mammography of the uncompressed breast. *Appl. Opt.*, 46:3628–3638. 60

Eriksson, A. P. and Astrom, K. (2006). Bijective image registration using thin-plate splines. *Pattern Recognition, International Conference on*, 3:798–801. 123

Everdell, N., Gibson, A., Tullis, I., Vaithianathan, T., Hebden, J., and Delpy, D. (2004). A frequency multiplexed near infra-red topography system for imaging functional activation in the brain. *Biomedical Topical Meeting*, page WF33. 75

Faber, T. and Stokely, E. (1988). Orientation of 3-D structures in medical images. *IEEE Trans. Patt. Anal. Mach. Intell.*, 10(5):626 –633. 117

Farenick, R. D. (2000). *Algebras of Linear Transformations*. Springer. 39

Ferwerda, H. A. (1999). The radiative transfer equation for scattering media with a spatially varying refractive index. *J. of Optics A: Pure and Applied Optics*, 1(3):L1–L2. 62

Firbank, M., Arridge, S. R., Schweiger, M., and Delpy, D. T. (1996). An investigation of light transport through scattering bodies with non-scattering regions. *Phys. Med. Biol.*, 41:767–783. 66

Firbank, M. and Delpy, D. T. (1993). A design for a stable and reproducible phantom for use in near infrared imaging and spectroscopy. *Phys. Med. Biol.*, 38:847–853. 207

Firbank, M., Oda, M., and Delpy, D. T. (1995). An improved design for a stable and reproducible phantom material for use in near-infrared spectroscopy and imaging. *Phys. Med. Biol.*, 40(5):955–961. 207

Fischer, B. and Modersitzki, J. (2008). Ill-posed medicine&mdash;an introduction to image registration. *Inverse Problems*, 24(3):034008 (16pp). 113

Fitzpatrick, J., Hill, D., and Maurer, Jr., C.R. (2000). *Image Registration*, volume 2, chapter 8, pages 447–513. SPIE Press, San Diego. 115, 128

Fitzpatrick, J., West, J., and Maurer, Jr., C.R. (1998). Predicting error in rigid-body point-based registration. *IEEE Trans. Med. Im.*, 17(5):694 –702. 115, 119, 120, 128

Fletcher, R. and Reeves, C. M. (1964). Function minimization by conjugate gradients. *The Computer Journal*, 7(2):149–154. 51

Flock, S. T., Patterson, M. S., Wilson, B. C., and Wyman, D. R. (1989). Monte Carlo modelling of light propagation in highly scattering tissues. 1: Model predictions and comparison with diffusion theory. *IEEE Trans. Biomed. Eng.*, 36(12):1162–1168. 73

Forsey, D. R. and Bartels, R. H. (1988). Hierarchical b-spline refinement. In *SIGGRAPH '88: Proceedings of the 15th annual conference on Computer graphics and interactive techniques*, pages 205–212, New York, NY, USA. ACM. 124

Fox, N. C., Crum, W. R., Scahill, R. I., Stevens, J. M., Janssen, J. C., and Rossor, M. N. (2001). Imaging of onset and progression of alzheimer's disease with voxel-compression mapping of serial magnetic resonance images. *Lancet*, 358(9277):201–205. 113

Fox, N. C., Freeborough, P. A., and Rossor, M. N. (1996). Visualisation and quantification of rates of atrophy in alzheimer's disease. *Lancet*, 348(9020):94 – 97. 113

Freeborough, P. A. and Fox, N. C. (1998). Modeling brain deformations in alzheimer disease by fluid registration of serial 3D MR images. *J. Comput. Assist. Tomogr.*, 22(5):838–843. 113

Früwald, L., Kettenbach, J., Figl, M., Hummel, J., Bergmann, H., and Birkfellner, W. (2009). A comparative study on manual and automatic slice-to-volume registration of CT images. *European Radiology*, 19:2647–2653. 10.1007/s00330-009-1452-0. 115

Gaudette, R. J., Brooks, D. H., DiMarzio, C. A., Kilmer, M. E., Miller, E. L., Gaudette, T., and Boas, D. A. (2000). A comparison study of linear reconstruction techniques for diffuse optical tomographic imaging of absorption coefficient. *Phys. Med. Biol.*, 45(4):1051–1070. 75

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2003). *Bayesian Data Analysis, Second Edition (Texts in Statistical Science)*. Chapman & Hall/CRC, 2 edition. 100

Gering, D. T., Nabavi, A., Kikinis, R., Hata, N., O'Donnell, L. J., Grimson, W. E., Jolesz, F. A., Black, P. M., and Wells, W. M. (2001). An integrated visualization system for surgical planning and guidance using image fusion and an open MR. *J. Magn. Reson. Imaging*, 13(6):967–975. 115, 117, 118

Gibson, A., Hebden, J., and Arridge, S. R. (2005a). Recent advances in diffuse optical tomography. *Phys. Med. Biol.*, 50:R1–R43. 26, 61, 62, 75, 86, 88

Gibson, A. P., Hebden, J. C., Riley, J., Everdell, N., Schweiger, M., Arridge, S. R., and Delpy, D. T. (2005b). Linear and nonlinear reconstruction for optical tomography of phantoms with nonscattering regions. *Appl. Opt.*, 44(19):3925–3936. 66, 75

Gindi, G., Lee, M., Rangarajan, A., and Zubal, I. G. (1993). Bayesian reconstruction of functional images using anatomical information as priors. *IEEE Trans. Med. Im.*, 12(4):670–680. 85

Girolami, M. and He, C. (2003). Probability density estimation from optimally condensed data samples. *IEEE Trans. Patt. Anal. Mach. Intell.*, 25:1253–1264. 101

Goerres, G. W., Kamel, E., Heidelberg, T.-N. H., Schwitter, M. R., Burger, C., and von Schulthess, G. K. (2002). PET-CT image co-registration in the thorax: influence of respiration. *Eur. J. Nucl. Med. Mol. Imaging.*, 29(3):351–360. 118

Golub, G. H. and Van Loan, C. F. (1996). *Matrix Computations (Johns Hopkins Studies in Mathematical Sciences)*. The Johns Hopkins University Press. 39, 43

Gratton, G., Fabiani, M., Corballis, P. M., Hood, D. C., Goodman-Wood, M. R., Hirsch, J., Friedman, D., and Gratton, E. (1997). Fast and localized event-related optical signals (eros) in the human occipital cortex: comparisons with the visual evoked potential and fMRI. *Neuroimage*, 6:168–180. 26, 85

Groenhuis, R. A. J., Ferwerda, H. A., and Bosch, J. J. T. (1983). Scattering and absorption of turbid materials determined from reflection measurements (parts 1 and 2). *Appl. Opt.*, 22(16):2456–2467. 66, 67

Gudbjartsson, H. and Patz, S. (1995). The rician distribution of noisy MRI data. *Magnetic Resonance in Medicine*, 34(6):910–914. 128

Gull, S. and Daniell, G. (1978). Image reconstruction from incomplete and noisy data. *Nature*, 272:686–690. 107

Guven, M., Yazici, B., Intes, X., and Chance, B. (2005). Diffuse optical tomography with a priori anatomical information. *Phys. Med. Biol.*, 50:2837–2858. 75, 83, 84

Hadamard, J. (1902). Sur les problemes aux derivees partielles et leur signification physique. *Bulletin Princeton University*, 13:49–52. 38

Hajnal, J. V. (2001). *Medical Image Registration*. CRC Press, Cambridge. 115, 128

Hall, P., Sheather, S. J., Jones, M. C., and Marron, J. S. (1991). On optimal data-based bandwidth selection in kernel density estimation. *Biometrika*, 78:263–269. 100

Hall, P. and Wand, M. (1996). On the accuracy of binned kernel density estimators. *J. Multivariate Anal.*, 56:165–184. 143, 150

Hansen, P. C. (1987). The truncated SVD as a method for regularization. *BIT Numerical Mathematics*, 27:534–553. 45

Hansen, P. C. (1990a). The discrete picard condition of discrete ill-posed problems. *BIT Numerical Mathematics*, 30(4):658–672. 45

Hansen, P. C. (1990b). Truncated singular value decomposition solutions to discrete ill-posed problems with ill-determined numerical rank. *SIAM J. on Sci. and Stat. Comput.*, 11(3):503–518. 44, 45

Hansen, P. C. (1992a). Analysis of discrete ill-posed problems by means of the l-curve. *SIAM Review*, 34(4):561–580. 100

Hansen, P. C. (1992b). Numerical tools for analysis and solution of fredholm integral equations of the first kind. *Inverse Problems*, 8(6):849–872. 45

Hansen, P. C. (1998). *Rank-deficient and Discrete Ill-Posed Problems : Numerical Aspects of Linear Inversion*. SIAM, Philadelphia. 37, 38, 39, 44, 45, 46, 47, 100, 211

Harrach, B. (2009). On uniqueness in diffuse optical tomography. *Inverse Problems*, 25(5):055010 (14pp). 86

He, Y., Yap, K.-H., Chen, L., and Chau, L.-P. (2007). A nonlinear least square technique for simultaneous image registration and super-resolution. *IEEE Trans. Im. Proc.*, 16(11):2830 –2841. 218

Hebden, J., Gibson, A., Austin, T., Yusof, R., Everdell, N., Delpy, D., Arridge, S., Meek, J., and Wyatt, J. (2004). Imaging changes in blood volume and oxygenation in the newborn infant brain using three-dimensional optical tomography,. *Phys. Med. Biol.*, 49:1117–1130. 60

Hebden, J., Gibson, A., Yusof, R., Everdell, N., E.M.C.Hillman, Delpy, D., Arridge, S., Austin, T., Meek, J., and Wyatt, J. (2002). Three-dimensional optical tomography of the premature infant brain. *Phys. Med. Biol.*, 47:4155–4166. 26

Hebden, J. C., Arridge, S. R., and Schweiger, M. (1998). Investigation of alternative data types for time resolved optical tomography. In Alfano, R. R. and Fujimoto, J. G., editors, *Adv. in Optical Imaging and Photon Mig.*, volume 21 of *Trends in Optics and Photonics*, pages 162–167, Washington D.C. OSA, Opt. Soc. Am. 69, 86

Hebden, J. C., Schmidt, F. E. W., Fry, M. E., Schweiger, M., Hillman, E. M. C., Delpy, D. T., and Arridge, S. R. (1999). Simultaneous reconstruction of absorption and scattering images by multichannel measurement of purely temporal data. *Opt. Lett.*, 24:534–536. 61

Hecht, E. and Zajac, A. (2002). *Optics*. Addison-Wesley, Reading, MA, fourth edition. 58, 59

Heino, J., Arridge, S., Sikora, J., and Somersalo, E. (2003). Anisotropic effects in highly scattering media. *Phys. Rev. E*, 68:Article number 31908. 74

Heiskala, J., Pollari, M., Metsäranta, M., Grant, P. E., and Nissilä, I. (2009). Probabilistic atlas can improve reconstruction from optical imaging of the neonatal brain. *Opt. Express*, 17(17):14977–14992. 80

Hielscher, A., Klose, A., and Beuthan, J. (2000). Evolution strategies for optical tomographic characterization of homogeneous media. *Opt. Express*, 7(13):507–518. 75

Hielscher, A. H., Klose, A. D., and Hanson, K. M. (1999). Gradient-based iterative image reconstruction scheme for time-resolved optical tomography. *IEEE Trans. Med. Im.*, 18(3):262–271. 76

Hill, D. L. G., Batchelor, P. G., Holden, M., and Hawkes, D. J. (2001). Medical image registration. *Phys. Med. Biol.*, 46(3):R1–R45. 115, 128, 132, 134, 180

Hillman, E. M. C. (2002). Experimental and theoretical investigations of near infrared tomographic imaging methods and clinical applications. Phd thesis, University College London. 65, 86

Hillman, E. M. C., Hebden, J. C., Schmidt, F. E. W., Arridge, S. R., Schweiger, M., and Delpy, D. T. (2000). Calibration techniques and datatype extraction for time-resolved optical tomography. *Review of Sci. Inst.*, 71(9):3415–3427. 61

Hillman, E. M. C., Hebden, J. C., Schweiger, M., Dehgahni, H., Schmidt, F. E. W., Arridge, S. R., and Delpy, D. T. (2001). Time resolved optical tomography of the human forearm. *Phys. Med. Biol.*, 46(4):1117–1130. 60

Hiltunen, P., Calvetti, D., and Somersalo, E. (2008). An adaptive smoothness regularization algorithm for optical tomography. *Opt. Express*, 16(24):19957–19977. 82

Holden, M., Hill, D., Denton, E., Jarosz, J., Cox, T., Rohlfing, T., Goodey, J., and Hawkes, D. (2000). Voxel similarity measures for 3-D serial MR brain image registration. *IEEE Trans. Med. Im.*, 19(2):94 –102. 115

Huang, X., Moore, J., Guiraudon, G., Jones, D., Bainbridge, D., Ren, J., and Peters, T. (2009). Dynamic 2D ultrasound and 3D CT image registration of the beating heart. *IEEE Trans. Med. Im.*, 28(8):1179 –1189. 117

Ino, F., Ooyama, K., and Hagihara, K. (2005). A data distributed parallel algorithm for nonrigid image registration. *Parallel Computing*, 31(1):19 – 43. 124

Intes, X., Maloux, C., Guven, M., Yazici, B., and Chance, N. (2004). Diffuse optical tomography with physiological and spatial a priori constraints. *Phys. Med. Biol.*, 49(12):N155–63. 84

Ishimaru, A. (1978). *Wave Propagation and Scattering in Random Media*, volume 1. Academic, New York. 58, 62, 64, 66, 67

Izenman, A. J. (1991). Recent developments in nonparametric density estimation. *J. of the American Stat. Assoc.*, 86(413):205–224. 100, 101

Jacques, S. L. and Wang, L. (1995). *Optical-Thermal response of laser-irradiated tissue*, chapter Monte Carlo modeling of light transport in tissues, pages 73–100. Plenum, New York. 73

Jacquet, W., Nyssen, E., Bottenberg, P., Truyen, B., and de Groen, P. (2009). 2d image registration using focused mutual information for application in dentistry. *Comput. in Biol. and Med.*, 39(6):545–553. 117

Jaynes, E. (1982). On the rationale of maximum entropy methods. *Proc. IEEE*, 70(9):939–952. 46, 107

Jaynes, E. T. (1957a). Information theory and statistical mechanics. *Phys. Rev.*, 106:620–630. 107

Jaynes, E. T. (1957b). Information theory and statistical mechanics. II. *Phys. Rev.*, 108:171–190. 107

Jaynes, E. T. (1984). Monkeys, kangaroos and n. In Justice, J. H., editor, *In Maximum Entropy and Bayesian Methods in Applied Statistics: Proceedings of the 4th Maximum Entropy Workshop*, pages 26–58. 107

Jenkinson, M. and Smith, S. (2001). A global optimisation method for robust affine registration of brain images. *Med. Im. Anal.*, 5(2):143 – 156. 117

Jiang, Z., Piao, D., Xu, G., Ritchey, J. W., Holyoak, G. R., Bartels, K. E., Bunting, C. F., Slobodov, G., and Krasinski, J. S. (2008). Trans-rectal ultrasound-coupled near-infrared optical tomography of the prostate, part ii: Experimental demonstration. *Opt. Express*, 16(22):17505–17520. 83

Johnson, C. (1987). *Numerical solution of partial differential equations by the finite element method*. Cambridge University Press. 69

Jones, M. C., Marron, J. S., and Sheather, S. J. (1996). A brief survey of bandwidth selection for density estimation. *J. of the American Stat. Assoc.*, 91(433). 100

Judenhofer, M. S., Wehrl, H. F., Newport, D. F., Catana, C., Siegel, S. B., Becker, M., Thielscher, A., Kneilling, M., Lichy, M. P., Eichner, M., Klingel, K., Reischl, G., Widmaier, S., Röcken, M., Nutt, R. E., Machulla, H.-J., Uludag, K., Cherry, S. R., Claussen, C. D., and Pichler, B. J. (2008). Simultaneous PET-MRI: a new approach for functional and morphological imaging. *Nature Medicine*, 14(4):459–465. 85

Junck, L., Moen, J. G., Hutchins, G. D., Brown, M. B., and Kuhl, D. E. (1990). Correlation methods for the centering, rotation, and alignment of functional brain images. *J. Nucl. Med.*, 31(7):1220–1226. 128

Kaipio, J. and Somersalo, E. (2005). *Statistical and Computational Inverse Problems*. Springer, New York. 37, 38, 39, 79, 81, 91, 100

Kaipio, J. P., Kolehmainen, V., Vauhkonen, M., and Somersalo, E. (1999). Inverse problems with structural prior information. *Inverse Problems*, 15:713–729. 82

Kastanis, I. (2007). *Dynamic image and shape reconstruction in undersampled MRI*. PhD thesis, University College London. 52

Kaufman, L. (1987). Implementing and accelerating the em algorithm for positron emission tomography. *IEEE Trans. Med. Im.*, 6(1):37–51. 78

Kaufman, L. (1993). Maximum likelihood, least squares, and penalized least squares for PET. *IEEE Trans. Med. Im.*, 12(2):200–214. 78

Kazantsev, D., Pedemonte, S., Bousse, A., Panagiotou, C., Arridge, S., Hutton, B. F., and Ourselin, S. (2010). PET bayesian reconstruction using automatic bandwidth selection for joint entropy optimization. In *proceedings of IEEE Nuclear Science Symposium and Medical Imaging Conference*, volume M18-299. 31, 101, 107, 138, 176, 254, 255

Keijzer, M., Star, W. M., and Storchi, P. R. M. (1988). Optical diffusion in layered media. *Appl. Opt.*, 27(9):1820–1824. 67

Klose, A. D. and Hielscher, A. H. (2002). Optical tomography using the time-independent equation of radiative transfer–part 2: inverse model. *J. of Quant. Spect. and Radiative Transfer*, 72(5):715 – 732. 76

Klose, A. D. and Hielscher, A. H. (2003). Quasi-Newton methods in optical tomographic imaging. *Inverse Problems*, 19:387. 76

Kolehmainen, V. (2001). *Novel Approaches to Image Reconstruction in Diffusion Tomography*. PhD thesis, University of Kuopio. 37, 38, 39, 46, 47, 63, 64, 68, 69, 79

Kolehmainen, V., Vauhkonen, M., Kaipio, J. P., and Arridge, S. R. (2000). Recovery of piecewise constant coefficients in optical diffusion tomography. *Opt. Express*, 7(13):468–480. 81

Koyama, T., Iwasaki, A., Ogoshi, Y., and Okada, E. (2005). Practical and adequate approach to modeling light propagation in an adult head with low-scattering regions by use of diffusion theory. *Appl. Opt.*, 44(11):2094–2103. 67

Kullback, S. (1959). *Information theory and statistics*. John Wiley and Sons., New York. 90

Lancaster, J. L., Woldorff, M. G., Parsons, L. M., Liotti, M., Freitas, C. S., Rainey, L., Kochunov, P. V., Nickerson, D., Mikiten, S. A., and Fox, P. T. (2000). Automated talairach atlas labels for functional brain mapping. *Hum. Brain Mapp.*, 10(3):120–131. 117

Lee, S., Woberg, G., Chwa, K.-Y., and Shin, S. Y. (1996). Image metamorphosis with scattered feature constraints. *IEEE Trans. Vis. and Comp. Graph.*, 2(4):337–354. 123

Lemieux, L. and Barker, G. J. (1998). Measurement of small inter-scan fluctuations in voxel dimensions in magnetic resonance images using registration. *Med. Phys.*, 25(6):1049–1054. 117, 128

Lemieux, L., Jagoe, R., Fish, D. R., Kitchen, N. D., and Thomas, D. G. (1994). A patient-to-computed-tomography image registration method based on digitally reconstructed radiographs. *Med. Phys.*, 21(11):1749–1760. 128

Lester, H. and Arridge, S. R. (1999). A survey of hierarchical non-linear medical image registration. *Pattern Recognition*, 32(1):129 – 149. 117, 122

Levenberg, K. (1944). A method for the solution of certain non-linear problems in least squares. *Quarterly of Applied Mathematics*, 2(2):164–168. 54

Li, A., Boverman, G., Zhang, Y., Brooks, D., Miller, E. L., Kilmer, M. E., Zhang, Q., Hillman, E. M. C., and Boas, D. A. (2005). Optimal linear inverse solution with multiple priors in diffuse optical tomography. *Appl. Opt.*, 44:1948–1956. 84

Li, A., Miller, E., Kilmer, M. E., Brukilacchio, T. J., Chaves, T., Stott, J., Zhang, Q., Wu, T., Chorlton, M., Moore, R. H., Kopans, D. B., and A., D. A. B. D. (2003). Tomographic optical breast imaging guided by three-dimensional mammography. *Appl. Opt.*, 42:5181–5190. 79, 80, 85

Li, X., Dawant, B. M., Welch, E. B., Chakravarthy, A. B., Freehardt, D., Mayer, I., Kelley, M., Meszoely, I., Gore, J. C., and Yankeelov, T. E. (2009). A nonrigid registration algorithm for longitudinal breast MR images and the analysis of breast tumor response. *J. Magn. Reson. Imaging*, 27(9):1258 – 1270. 113

Likar, B. and Pernuš, F. (1999). Automatic extraction of corresponding points for the registration of medical images. *Med. Phys.*, 26(8):1678–1686. 117

Mackay, D. J. C. (2002). *Information Theory, Inference & Learning Algorithms*. Cambridge University Press. 90

Maes, F., Collignon, A., Vandermeulen, D., Marchal, G., and Suetens, P. J. (1997). Multimodality image registration by maximization of mutual information. *IEEE Trans. Med. Im.*, 16(2):187–198. 117, 129

Maes, F., Vandermeulen, D., and Suetens, P. (1999). Comparative evaluation of multiresolution optimization strategies for multimodality image registration by maximization of mutual information. *Med Image Anal*, 3(4):373–386. 133

Mahanand, B. S. and Kumar, M. A. (2009). Analysis of alzheimer's disease progression in structural magnetic resonance images. *WSEAS Trans. on Computers*, 8(4):579–588. 113

Maintz, J. B. A., Elsen, P. A. v. d., and Viergewer, M. A. (1995). Comparison of feature-based matching of CT and MR brain images. *Comp. Vis., Virt. Real. and Robot. in Med.*, pages 219–228. 128

Maintz, J. B. A. and Viergever, M. A. (1998). A survey of medical image registration. *Med. Im. Anal.*, 2(1):1–36. 115, 116, 122

Makela, T., Clarysse, P., Sipila, O., Pauna, N., Pham, Q. C., Katila, T., and Magnin, I. (2002). A review of cardiac image registration methods. *IEEE Trans. Med. Im.*, 21(9):1011 –1021. 118

Marquardt, D. W. (1963). An algorithm for least-squares estimation of nonlinear parameters. *J. of the Soc. for Indust. and App. Math.*, 11(2):431–441. 54

Martí-López, L., Bouza-Domínguez, J., Hebden, J. C., Arridge, S. R., and Martínez-Celorio, R. A. (2003). Validity conditions for the radiative transfer equation. *J. Opt. Soc. Am. A*, 20(11):2046–2056. 62

Maurer, Jr., C.R. and Fitzpatrick, J. (1993). A review of medical image registration. In *Interactive image–guided neurosurgery*, pages 17–44. 115

Maurer, Jr., C.R., Fitzpatrick, J., Wang, M., Galloway,Jr., R.L., Maciunas, R., and Allen, G. (1997). Registration of head volume images using implantable fiducial markers. *IEEE Trans. Med. Im.*, 16(4):447 –462. 117

Maurer, Jr., C.R., Maciunas, R., and Fitzpatrick, J. (1998). Registration of head CT images to physical space using a weighted combination of points and surfaces [image-guided surgery]. *IEEE Trans. Med. Im.*, 17(5):753 –761. 128

Mazzara, G. P., Briggs, R. W., Wu, Z., and Steinbach, B. G. (1996). Use of a modified polysaccharide gel in developing a realistic breast phantom for MRI. *J. Magn. Reson. Imaging*, 14(6):639–648. 207

Mazziotta, J. C., Toga, A. W., Evans, A., Fox, P., and Lancaster, J. (1995). A probabilistic atlas of the human brain: Theory and rationale for its development : The international consortium for brain mapping (icbm). *Neuroimage*, 2(2, Part 1):89 – 101. 117

Melbourne, A., Hawkes, D., and Atkinson, D. (2009). Image registration using uncertainty coefficients. In *Proceedings of the Sixth IEEE international conference on Symposium on Biomedical Imaging: From Nano to Macro*, pages 951–954. 132

Milstein, A. B., Oh, S., Webb, K. J., Bouman, C. A., Zhanga, Q., Boas, D. A., and Millane, R. P. (2003). Fluorescence optical diffusion tomography. *Appl. Opt.*, 42(16). 80

Modat, M., Ridgway, G. ., Taylor, Z. ., Lehmann, M., Barnes, J., Hawkes, D. ., Fox, N. ., and Ourselin, S. (2009). Fast free-form deformation using graphics processing units. *Computer Methods and Programs in Biomedicine*. 225, 255

Model, R., Orlt, M., Walzel, M., and Hünlich, R. (1997). Reconstruction algorithm for near-infrared imaging in turbid media by means of time-domain data. *J. Opt. Soc. Am. A*, 14(1):313–324. 79

Mohajerani, P., Eftekhar, A. A., Huang, J., and Adibi, A. (2007). Optimal sparse solution for fluorescent diffuse optical tomography: theory and phantom experimental results. *Appl. Opt.*, 46(10):1679–1685. 80

Moore, E. H. (1920). On the reciprocal of the general algebraic matrix. *Bulletin of the Am. Math. Soc.*, 26:394–395. Abstract. 42

Mumcuoglu, E. and Leahy, R. (1994). A gradient projection conjugate gradient algorithm for bayesian PET reconstruction. *Nuclear Science Symposium and Medical Imaging Conference, 1994., 1994 IEEE Conference Record*, 3:1212–1216 vol.3. 78

Mumcuoglu, E. U., Leahy, R., Cherry, S. R., and Zhou, Z. (1994). Fast gradient-based methods for bayesian reconstruction of transmission and emission PET images. *IEEE Trans. Med. Im.*, 13(4):687–701. 78

Murphy, M. J. (2004). Tracking moving organs in real time. *Semin. Radiat. Oncol*, 14(1):91–100. 115

Natarajan, B. K. (1995). Sparse approximate solutions to linear systems. *SIAM J. Comp.*, 24(2):227–234. 80

Nelder, J. A. and Mead, R. (1965). A simplex method for function minimization. *The Computer Journal*, 7(4):308–313. 49

Nielsen, M. and Lillholm, M. (2006). What do features tell about images? In Kerckhove, M., editor, *Scale-Space and Morphology in Computer Vision*, volume 2106 of *Lecture Notes in Computer Science*, pages 39–50. Springer Berlin / Heidelberg. 10.1007/3-540-47778-0 4. 30

Nissilä, I., Kotilahti, K., Fallström, K., and Katila, T. (2002). Instrumentation for the accurate measurement of phase and amplitude in optical tomography. *Review of Sci. Inst.*, 73(9):2206–3312. 208

Nissilä, I., Noponen, T., Kotilahti, K., Tarvainen, T., Schweiger, M., Lipiänen, L., Arridge, S. R., and Katila, T. (2005). Instrumentation and calibration methods for the multichannel measurement of phase and amplitude in optical tomography. *Review of Sci. Inst.*, 76(4):004302. 208

Nocedal, J. and Wright, S. J. (1999). *Numerical Optimization*. Springer Verlag, New York. 50, 53, 54, 78

Ntziachristos, V., Hielscher, A. H., Yodh, A. G., and Chance, B. (2001). Diffuse optical tomography of highly heterogeneous media. *IEEE Trans. Med. Im.*, 20(6):470–478. 88

Ntziachristos, V., Yodh, A. G., Schnall, M., and Chance, B. (2000). Concurrent MRI and diffuse optical tomography of breast after indocyanine green enhancement. *Proc. Nat. Acad. Sci. USA*, 97:2767–2772. 85

Ntziachristos, V., Yodh, A. G., Schnall, M. D., and Chance, B. (2002). MRI-guided diffuse optical spectroscopy of malignant and benign breast lesions. *Neoplasia*, 4(4):347–54. 83, 85

Nuyts, J. (2007). The use of mutual information and joint entropy for anatomical priors in emission tomography. *Nuclear Science Symposium Conference Record*, 6:4149–4154. 176

Oh, S., Milstein, A. B., Millane, R. P., Bouman, C. A., and Webb, K. J. (2002). Source-detector calibration in three-dimensional bayesian optical diffusion tomography. *J. Opt. Soc. Am. A*, 19(10):1983–1993. 75

Ohno, S., Kato, H., Harimoto, T., Ikemoto, Y., Yoshitomi, K., Kadohisa, S., Kuroda, M., and Kanazawa, S. (2008). Production of a human-tissue-equivalent MRI phantom: optimization of material heating. *Magn. Reson. Med. Sci.*, 7(3):131–140. 207

Okada, E. and Delpy, D. T. (2003). Near-infrared light propagation in an adult head model. i. modeling of low-level scattering in the cerebrospinal fluid layer. *Appl. Opt.*, 42(16):2906–2914. 66, 74, 85

Oki, Y., Kawaguchi, H., and Okada, E. (2009). Validation of practical diffusion approximation for virtual near infrared spectroscopy using a digital head phantom. *Opt. Review*, 16(2):153–159. 67

Panagiotou, C., Somayajula, S., Gibson, A. P., Schweiger, M., Leahy, R. M., and Arridge, S. R. (2009a). Diffusion optical tomography using entropic priors. In *Proceedings of the Sixth IEEE International Symposium on Biomedical Imagting (ISBI'09)*, pages 165–168. 31, 175

Panagiotou, C., Somayajula, S., Gibson, A. P., Schweiger, M., Leahy, R. M., and Arridge, S. R. (2009b). Information theoretic regularization in diffuse optical tomography. *J. Opt. Soc. Am. A*, 26(5):1277–1290. 31, 76, 138, 175, 255

Papoulis, A. and Pillai, U. S. (2001). *Probability, Random Variables and Stochastic Processes*. McGraw-Hill Science/Engineering/Math. 90, 91, 92, 95, 96, 97, 98, 99, 100

Parzen, E. (1962). On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33(3):1065–1076. 100, 147

Paulsen, K. D. and Jiang, H. (1996). Enhanced frequency-domain optical image reconstruction in tissues through total variation minimization. *Appl. Opt.*, 35:3447–3458. 81

Pedemonte, S., Cardoso, M. J., Bousse, A., Panagiotou, C., Kazantsev, D., , Arridge, S., Hutton, B. F., and Ourselin, S. (2010a). Class conditional entropic prior for MRI enhanced SPECT reconstruction. In *proceedings of IEEE Nuclear Science Symposium and Medical Imaging Conference*, volume M18-294. 31, 138, 176, 255

Pedemonte, S., Cardoso, M. J., Bousse, A., Panagiotou, C., Kazantsev, D., Arridge, S., Hutton, B. F., and Ourselin, S. (to be submitted, 2010b). Class conditional entropic prior for MRI enhanced SPECT reconstruction. *IEEE Trans. Med. Im.* 31, 138

Pelizzari, C. A., Chen, G. T., Spelbring, D. R., Weichselbaum, R. R., and Chen, C. T. (1989). Accurate three-dimensional registration of CT, PET, and/or MR images of the brain. *J. Comput. Assist. Tomogr.*, 13(1):20–6. 128

Pennec, X., Cachier, P., and Ayache, N. (1999). Understanding the "demon's algorithm": 3D non-rigid registration by gradient descent. In *MICCAI '99: Proceedings of the Second International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 597–605, London, UK. Springer-Verlag. 122

Penrose, R. (1955). A generalized inverse for matrices. *Proc. of the Cambridge Phil.Soc.*, 51:406–413. 42

Phillips, D. L. (1962). A technique for the numerical solution of certain integral equations of the first kind. *J. ACM*, 9(1):84–97. 45

Pineda, A. R., Schweiger, M., Arridge, S., and Barrett, H. H. (2006). Information content of data types in time-domain optical tomography. *J. Opt. Soc. Am. A*, 12:2989–2996. 69

Pitiot, A., Bardinet, E., Thompson, P. M., and Malandain, G. (2006). Piecewise affine registration of biological images for volume reconstruction. *Med. Im. Anal.*, 10(3):465 – 483. Special Issue on The Second International Workshop on Biomedical Image Registration (WBIR'03). 118

Pluim, J. P. W., Maintz, J. B. A., and Viergever, M. A. (2003). Mutual-information-based registration of medical images: a survey. *IEEE Trans. Med. Im.*, 22(8):986–1004. 129

Pogue, B., McBride, T., Osterberg, U., and Paulsen, K. (1999a). Comparison of imaging geometries for diffuse optical tomography of tissue. *Opt. Express*, 4(8):270–286. 60

Pogue, B. and Paulsen, K. D. (1998). High-resolution near-infrared tomographic imaging simulations of the rat cranium by use of a priori magnetic resonance imaging structural information. *Opt. Lett.*, 23:1716–1718. 83

Pogue, B. W., McBride, T. O., Prewitt, J., Osterberg, U. L., and Paulsen, K. D. (1999b). Spatially variant regularization improves diffuse optical tomography. *Appl. Opt.*, 38(13):2950–2961. 80

Pogue, B. W., Patterson, M. S., Jiang, H., and Paulsen, K. D. (1995). Initial assessment of a simple system for frequency domain diffuse optical tomography. *Phys. Med. Biol.*, 40:1709–1729. 79

Polak, E. and Ribière, G. (1969). Note sur la convergence de directions conjugï£¡es. *Rev. Francaise Informat Recherche Operationelle*, 3e Annï£¡e, 16:35–43. 52

Pott, P. and Schwarz, M. (2002). Robots, navigation, telesurgery: state of the art and market overview. *Z Orthop Ihre Grenzgeb*, 140(2):218–231. 115

Prahl, S. A., Keijzer, M., Jacques, S. L., and Welch, A. J. (1989). A Monte Carlo model of light propagation in tissue. In *SPIE Proc. of Dosimetry of Laser Rad. in Med.and Biol.*, number IS 5, pages 102–111. 73

Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P. (1992a). *Numerical Recipes in C*. Cambridge University, Cambridge, England, 2nd edition. 78

Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P. (1992b). *Numerical recipes in C (2nd ed.): the art of scientific computing*. Cambridge University Press, New York, NY, USA. 39, 43, 49, 50, 53, 54, 126

Price, K. V., Storn, R. M., Lampinen, J. A., Salomon, M., Perrin, G.-R., Heitz, F., and Armspach, J.-P. (2005). Parallel differential evolution: Application to 3-D medical image registration. In *Differential Evolution*, Natural Computing Series, pages 353–411. Springer Berlin Heidelberg. 133

Raabe, A., Krishnan, R., Wolff, R., Hermann, E., Zimmermann, M., and Seifert, V. (2002). Laser surface scanning for patient registration in intracranial image-guided surgery. *Neurosurgery*, 50(4):797–801; discussion 802–3. 117

Rangarajan, A., Hsiao, I. T., and Gindi, G. (2000). A Bayesian joint mixture framework for the integration of anatomical information in functional image reconstruction. *J. Math. Imag. Vision*, 12:119–217. 85

Raykar, V. C. and Duraiswami, R. (2006). Fast optimal bandwidth selection for kernel density estimation. *Proceedings of the sixth SIAM International Conference on Data Mining*, pages 524–528. 100

Reza, F. M. (1994). *An Introduction to Information Theory*. Dover Publications. 106

Rice, J. A. (2001). *Mathematical Statistics and Data Analysis*. Duxbury Press. 97

Ries, M., Senneville, B. D. D., Roujol, S., Hey, S., Maclair, G., Kohler, M. O., Quesson, B., and Moonen, C. (2010). Three dimensional motion compensation for real-time MRI guided focused ultrasound treatment of abdominal organs. In Hynynen, K. and Souquet, J., editors, *AIP Conf. Proc.*, volume 1215, pages 239–242. 115

Roche, A., Malandain, G., Pennec, X., and Ayache, N. (1998a). The correlation ratio as a new similarity measure for multimodal image registration. *MICCAI '98 Proc. of the First International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 1115–1124. 129

Roche, A., Roche, A., Malandain, G., Mal, G., Pennec, X., Ayache, N., and Epidaure, P. (1998b). Multimodal image registration by maximization of the correlation ratio. Technical report, INRIA. 129

Rouet, J. M., Jacq, J. J., and Roux, C. (2000). Genetic algorithms for a robust 3-D MR-CT registration. *IEEE Trans. Inf. Technol. Biomed.*, 4(2):126–136. 133

Roy, R. and Sevick-Muraca, E. (1999). Truncated Newton's optimisation sceme for absorption and fluorescence optical tomography : Part I theory and formulation. *Opt. Express*, 4(10):353–371. 76

Roy, R. and Sevick-Muraca, E. (2001). A numerical study of gradient-based nonlinear optimization methods for contrast enhanced optical tomography. *Opt. Express*, 9(1):49–65. 76

Rudin, L. I., Osher, S., and Fatemi, E. (1992). Nonlinear total variation based noise removal algorithm. *Physica D*, 60:259–268. 81

Rueckert, D., Clarkson, M., Hill, D., and Hawkes, D. (2000). Non-rigid registration using higher order mutual information. In *Proc. SPIE*, volume 3979, pages 438–447. 129

Rueckert, D., Sonoda, L. I., Hayes, C., Hill, D. L. G., Leach, M. O., and Hawkes, D. J. (1999). Nonrigid registration using free-form deformations: application to breast MR images. *IEEE Trans. Med. Im.*, 18(8):712–721. 113, 114, 117, 118, 123, 124, 125, 133

Saad, Y. and Schultz, M. H. (1986). GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems. *SIAM J. on Sci. and Stat. Comput.*, 7(3):856–869. 43

Satava, R. M. (1999). Emerging Technologies for Surgery in the 21st Century. *Arch. Surg.*, 134(11):1197–1202. 115

Schmidt, F. E. W. (1999). Development of a time-resolved optical tomography system for neonatal brain imaging. Phd thesis, Universtiy College London. 59, 86

Schmidt, F. E. W., Fry, M. E., Hillman, E. M. C., Hebden, J. C., and Delpy, D. T. (2000). A 32-channel time-resolved instrument for medical optical tomography. *Review of Sci. Inst.*, 71(1):256–265. 61

Schneider, T. D. (1995). Information therory primer. 90, 103

Schweiger, M. (1994). *Application of the Finite Element Method in Infrared Image Reconstruction of Scattering Media*. PhD thesis, University of London. 60, 64, 68, 69, 71, 72

Schweiger, M. and Arridge, S. R. (1997). The finite element model for the propagation of light in scattering media: Frequency domain case. *Med. Phys.*, 24(6):895–902. 65, 68, 69

Schweiger, M. and Arridge, S. R. (1999a). Application of temporal filters to time-resolved data in optical tomography. *Phys. Med. Biol.*, 44:1699–1717. 78, 79

Schweiger, M. and Arridge, S. R. (1999b). Optical tomographic reconstruction in a complex head model using *a priori* region boundary information. *Phys. Med. Biol.*, 44:2703–2721. 82, 83

Schweiger, M. and Arridge, S. R. (2003). Optical tomography with local basis functions. *J. of Electronic Imaging*, 12(4):583–593. 74

Schweiger, M., Arridge, S. R., and Delpy, D. T. (1993). Application of the finite element method for the forward and inverse models in optical tomography. *J. Math. Imag. Vision*, 3:263–283. 69

Schweiger, M., Arridge, S. R., Hiraoka, M., and Delpy, D. T. (1992). Application of the finite element method for the forward model in infrared absorption imaging. In Wilson, D. C. and Wilson, J. N., editors, *Mathematical Methods in Medical Imaging*, volume 1768, pages 97–108. Proc. SPIE. 69

Schweiger, M., Arridge, S. R., Hiraoka, M., and Delpy, D. T. (1995). The finite element model for the propagation of light in scattering media: Boundary and source conditions. *Med. Phys.*, 22(11):1779–1792. 66, 67, 68, 69, 72, 74

Schweiger, M., Arridge, S. R., and Nissilä, I. (2005). Gauss-Newton method for image reconstruction in diffuse optical tomography. *Phys. Med. Biol.*, 50:2365–2386. 53, 72, 75, 76, 78, 79, 80, 178

Schweiger, M., Gibson, A., and Arridge, S. R. (2003). Computational aspects of diffuse optical tomography. *Computing in Science and Engineering*, 5(6):33–41. 87

Sederberg, T. W. and Parry, S. R. (1986). Free-form deformation of solid geometric models. *SIGGRAPH Comput. Graph.*, 20(4):151–160. 123

Segerlind, L. J. (1984). *Applied finite element analysis*. New York: Wiley. 69

Shannon, C. (1948). A mathematical theory of communication. Technical report. 89, 90, 102, 104, 105, 106

Shewchuk, J. R. (1994). An introduction to the conjugate gradient method without the agonizing pain. Technical report, Pittsburgh, PA, USA. 43, 49, 50, 51, 52

Shwartz, S., Zibulevsky, M., and Yoav, Y. S. (2005). Fast kernel entropy estimation and optimization. *Signal Process.*, 85(5):1045–1058. 138, 142, 144, 152, 154, 156, 157, 172, 252

Sikora, J., J.Riley, Arridge, S., Zacharopoulos, A., and Ripoll, J. (2004). Light propagation in diffusive media with non-scattering regions using 3D BEM. In Wilde, S. C., editor, *Proceedings of Third International Conference on Boundary Integral Methods: Theory and Applications*. 74

Silverman, B. W. (1982). AS 176: Kernel density estimation using the fast Fourier transform. *Appl. Stat.*, 31(1):93–99. 140, 141, 143, 144

Silverman, B. W. and Green, P. J. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London. 100, 140, 141, 143, 144, 147, 166, 254

Simonoff, J. (1996). *Smoothing Methods in Statistics*. Statistics. Springer. 100

Skilling, J. and Bryan, R. K. (1984). Maximum Entropy Image Reconstruction - General Algorithm. *Mon. Not. R. Astron. Soc.*, 211:111–+. 107, 108

Som, S., Hutton, B. H., and Braun, M. (1998). Properties of minimum cross-entropy reconstruction of emission tomography with anatomically based prior. *IEEE Trans. Nucl. Sci.*, 46:3014–3021. 85

Somayajula, S., Asma, E., and Leahy, R. M. (2005). PET image reconstruction using anatomical information through mutual information based priors. In *Conf. Rec : IEEE Nucl. Sci. Symp. and Med. Imag. Conf.*, pages 2722–2726. 176

Somayajula, S., Joshi, A., and Leahy, R. (2008). Mutual information based non-rigidmouse registration using a scale-space approach. In *Biomedical Imaging: From Nano to Macro, 2008. ISBI 2008. 5th IEEE International Symposium on*, pages 1147 –1150. 129

Somayajula, S., Panagiotou, C., Rangarajan, A., Quanzheng, L., Arridge, S. R., and Leahy, R. M. (2010). PET image reconstruction using information theoretic anatomical priors. *IEEE Trans. Med. Im.*, 30(3):537–49. 31, 138, 254, 255

Somayajula, S., Rangarajan, A., and Leahy, R. M. (2007). PET image reconstruction using anatomical information through mutual information based priors : A scale space approach. In *Proc. ISBI 2007*, pages 165–168. 176

Spiegel, M. R. and Stephens, L. J. (2008). *Schaum's outline of theory and problems of statistics; 4th ed.* Schaum's outline. McGraw-Hill, New York, NY. 95

Stepnoski, R. A., LaPorta, A., Raccuia-Behling, F., Blonder, G. E., Slusher, R. E., and Kleinfeld, D. (1991). Noninvasive detection of changes in membrane potential in cultured neurons by light scattering. *Proc. Nat. Acad. Sci. USA*, 88(21):9382–9386. 85

Stoll, J. and Dupont, P. (2005). Passive markers for ultrasound tracking of surgical instruments. *Med. Image. Comput. Comput. Assist. Interv.*, 8(Pt 2):41–48. 115

Storn, R. and Price, K. (1997). Differential evolution – a simple and efficient heuristic for global optimization over continuous spaces. *J. of Global Optimization*, 11(4):341–359. 48

Strang, G. (1988). *Linear Algebra and Its Applications*. Brooks Cole. 39, 40, 43

Studholme, C., Hill, D., and Hawkes, D. (1996). Automated 3-D registration of MR and CT images of the head. *Med. Im. Anal.*, 1(2):163 – 175. 118, 224

Studholme, C., Hill, D. L. G., and Hawkes, D. J. (1995). Multiresolution voxel similarity measures for MR-PET registration. In Bizais, Y., Barillot, C., and Paola, R. D., editors, *Inf. Proc. in Med. Im.*, pages 287–298. Dordrecht: Kluwer Academic. 129

Studholme, C., Hill, D. L. G., and Hawkes, D. J. (1999). An overlap invariant entropy measure of 3D medical image alignment. *Pattern Recognition*, 32(1):71 – 86. 132

Surry, K. J. M., Austin, H. J. B., Fenster, A., and Peters, T. M. (2004). Poly(vinyl alcohol) cryogel phantoms for use in ultrasound and MR imaging. *Phys. Med. Biol.*, 49(24):5529. 207

Tang, J., Kuwabara, H., Wong, D. F., and Rahmim, A. (2010). Direct 4d reconstruction of parametric images incorporating anato-functional joint entropy. *Phys. Med. Biol.*, 55(15):4261. 176

Tarantola, A. (2004). *Inverse Problem Theory and Methods for Model Parameter Estimation*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA. 36, 37

Tarvainen, T., Kolehmainen, V., Vauhkonen, M., Vanne, M., Gibson, A. P., Schweiger, M., Arridge, S. R., and Kaipio, J. P. (2005a). Computational calibration method for optical tomography. *Appl. Opt.*, 44(10):1879–1888. 81

Tarvainen, T., Vauhkonen, M., Kolehmainen, V., and Kaipio, J. P. (2005b). A hybrid radiative transfer - diffusion model for optical tomography. *Appl. Opt.*, 44(6):876–886. 67

Tarvainen, T., Vauhkonen, M., Kolehmainen, V., Kaipio, J. R., and Arridge, S. R. (2008). Utilizing the radiative transfer equation in optical tomography. *PIERS Online*, 4(6):655–660. 64, 66

Terzopoulos, D. (1986). Regularization of inverse visual problems involving discontinuities. *IEEE Trans. Pattern Anal. and Mach. Intel.*, 8:413–242. 189

Thevenaz, P., Ruttimann, U. E., and Unser, M. (1998). A pyramid approach to subpixel registration based on intensity. *IEEE Trans. Im. Proc.*, 7(1):27–41. 133

Thevenaz, P. and Unser, M. (2000). Optimization of mutual information for multiresolution image registration. *IEEE Trans. Im. Proc.*, 9(12):2083–2099. 133

Thirion, J.-P. (1996). New feature points based on geometric invariants for 3D image registration. *Int. J. of Comput. Vision*, 18(2):121–137. 117

Thirion, J.-P. (1998). Image matching as a diffusion process: an analogy with maxwell's demons. *Med. Im. Anal.*, 2(3):243 – 260. 122

Tikhonov, A. (1963). Regularization of incorrectly posed problems. *Soviet Math. Dokl.*, 4:1624–1627. 45

Tom, B. and Katsaggelos, A. (1995). Reconstruction of a high-resolution image by simultaneous registration, restoration, and interpolation of low-resolution images. In *Proc. of Intern. Conf. on Image Proc.*, volume 2, pages 539 –542 vol.2. 218

Tribus, M. (1961). *Thermostatics and Thermodynamics*. D. van Nostrand Company, Inc., Princeton, N. J. 103

Troy, T. L., Page, D. L., and Sevick-Muraca, E. M. (1996). Optical properties of normal and diseased breast tissues: prognosis for optical mammography. *J. Biomed. Opt.*, 1(3):342–355. 85

Turgeon, G.-A., Lehmann, G., Guiraudon, G., Drangova, M., Holdsworth, D., and Peters, T. (2005). 2D-3D registration of coronary angiograms for cardiac procedure planning and guidance. *Med. Phys.*, 32(12):3737–3749. 117

Turlach, B. A. (1993). Bandwidth selection in kernel density estimation: a review. Technical report, Univ. Cahtolique de Louvain. 100

Van de Sompel, D. and Sir., Brady, M. (2009a). Robust incorporation of anatomical priors into limited view tomography using multiple cluster modelling of the joint histogram. In *ISBI'09: Proc. of the Sixth IEEE international conference on Symposium on Biomedical Imaging*, pages 1279–1282, Piscataway, NJ, USA. IEEE Press. 255

Van de Sompel, D. and Sir., Brady, M. (2009b). Robust joint entropy regularization of limited view transmission tomography using gaussian approximations to the joint histogram. *Lecture Notes in Computer Science*, 5636:638–650. 176

Van de Sompel, D. and Sir., Brady, M. (2009c). Simultaneous reconstruction and registration algorithm for limited view transmission tomography using a multiple cluster approximation to the joint histogram with an anatomical prior. In *Engineering in Medicine and Biology Society, Annual International Conference of the IEEE*, pages 5733 –5736. 218, 220

Van den Elsen, P., Pol, E.-J., and Viergever, M. (1993). Medical image matching-a review with classification. *Engineering in Medicine and Biology Magazine, IEEE*, 12(1):26 –39. 115, 116

Varah, J. M. (1979). A practical examination of some numerical methods for linear discrete ill-posed problems. *SIAM Review*, 21(1):100–111. 45, 47

Viola, P. A. (1995). Alignment by maximization of mutual information. Technical Report AITR-1548, M.I.T. 48, 75, 95, 98, 101, 104, 105, 106, 128, 129, 133, 146, 225, 255

Vogel, C. R. (2002). *Computational Methods for Inverse Problems*. SIAM. 37, 38, 39, 45, 46, 47, 81, 100

Vogel, C. R. and Oman, M. E. (1998). Fast, robust total variation-based reconstruction of noisy, blurred images. *IEEE Trans. Im. Proc.*, 7:813–824. 81

Walvoord, D., Baum, K., Helguera, M., Krol, A., and Easton, R. (2008). Localization of fiducial skin markers in MR images using correlation pattern recognition for PET/MRI nonrigid breast image registration. In *37th IEEE Applied Imagery Pattern Recognition Workshop*, pages 1 –4. 117

Wand, M. (1994). Fast computation of multivariate kernel estimators. *J.Comp. Graph. Stat.*, 3:433–445. 143, 150

Wand, M. and Jones, M. (1995). *Kernel Smoothing*, volume 60 of *Monographs on Statistics and Applied Probability*. Chapman & Hall. 100

Wang, Z., Bovik, A., Sheikh, H., and Simoncelli, E. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Im. Proc.*, 13(4):600 –612. 30

Webster, J. G., editor (1990). *Electrical Impedance Tomography*. Adam Hilger, Bristol. 82

Wells, W. M., Viola, P., Atsumi, H., Nakajima, S., and Kikinis, R. (1996). Multi-modal volume registration by maximization of mutual information. *Med. Im. Anal.*, 1(1):35–51. 129

Wilson, B. C. and Adam, G. (1983). A Monte Carlo model for the absorption and flux distribution of light in tissue. *Med. Phys.*, 10:824–830. 73

Woods, R. P., Cherry, S. R., and Mazziotta, J. C. (1992). Rapid automated algorithm for aligning and reslicing PET images. *J. Comput. Assist. Tomogr.*, 16:620–633. 129

Woods, R. P., Mazziotta, J. C., and Cherry, S. R. (1993). MRI-PET registration with automated algorithm. *J. Comput. Assist. Tomogr.*, 17(4):536–546. 117, 129

Xu, G., Piao, D., Musgrove, C. H., Bunting, C. F., and Dehghani, H. (2008). Trans-rectal ultrasound-coupled near-infrared optical tomography of the prostate, part i: Simulation. *Opt. Express*, 16(22):17484–17504. 83

Xu, M. and Wang, L. (2006). Photoacoustic imaging in biomedicine. *Review of Sci. Inst.*, 77(4). cited By (since 1996) 236. 27

Yalavarthy, P. K., Pogue, B. W., Dehghani, H., Carpenter, C. M., Jiang, S., and Paulsen, K. D. (2007a). Structural information within regularization matrices improves near infrared diffuse optical tomography. *Opt. Express*, 15(13):8043–8058. 80

Yalavarthy, P. K., Pogue, B. W., Dehghani, H., and Paulsen, K. D. (2007b). Weight-matrix structured regularization provides optimal generalized least-squares estimate in diffuse optical tomography. *Med. Phys.*, 34(6):2085–2098. 80

Yates, T., Hebden, J., Gibson, A., Everdell, N., Arridge, S., and Douek, M. (2005). Optical tomography of the breast using a multi–channel time–resolved imager. *Phys. Med. Biol.*, 50(11):2503–2518. 26

Ye, J. C., Bouman, C. A., Webb, K. J., and Millane, R. P. (2001). Nonlinear multigrid algorithms for Bayesian optical diffusion tomography. *IEEE Trans. Im. Proc.*, 10(5):909–922. 75

Zaidi, H., Montandon, M., and Slosman, D. (2003). Magnetic resonance image-guided attenuation correction in 3D brain positron emission tomography. *Med. Phys.*, 30:937–948. 85

Zhang, Q., Brukilacchio, T. J., Li, A., Stott, J. J., Chaves, T., Hillman, E., Wu, T., Chorlton, M., Rafferty, E., Moore, R., Kopans, D. B., and Boas, D. A. (2005a). Coregistered tomographic X-ray and optical breast imaging: initial results. *J. Biomed. Opt.*, 10(2):024033:1–9. 80

Zhang, X., Toronov, V., and Webb, A. (2005b). Simultaneous integrated diffuse optical tomography and functional magnetic resonance imaging of the human brain. *Opt. Express*, 13(14):5513–5521. 85

Zhi, Y., Peimin, Y., Sheng, L., Juxia, S., and Yuhui, H. (2008). Super resolution based on simultaneous registration and reconstruction. In *Control, Automation, Robotics and Vision, 2008. ICARCV 2008. 10th International Conference on*, pages 1089 –1092. 218

Zhu, Q. (1999). Imager that combines near-infrared diffusive light and ultrasound. *Opt. Lett.*, 24(15):1050–1052. 84

Zhu, Q., Chen, N., and Kurtzman, S. H. (2003). Imaging tumour angiogenesis by use of combined near-infrared diffusive light and ultrasound. *Opt. Lett.*, 25(5):337–339. 84

Zhu, Q., Kurtzma, S. H., Hegde, P., Tannenbaum, S., Kane, M., Huang, M., Chen, N. G., Jagjivan, B., and Zarfos, K. (2005). Utilizing optical tomography with ultrasound localization to image heterogeneous hemoglobin distribution in large breast cancers. *Neoplasia*, 7(3):263–70. 84, 85, 218

Zienkiewicz, O. C. and Taylor, R. L. (1987). *The Finite Element Method*. McGraw-Hill, London, 4th edition. 71

Zienkiewicz, O. C. and Taylor, R. L. (2000). *The Finite Element Method*, volume 1. McGraw-Hill, London, 5th edition. 69, 71

Zitova, B. (2003). Image registration methods: a survey. *Image and Vision Computing*, 21(11):977–1000. 115

Zuo, C. S., Jiang, A., Buff, B. L., Mahon, T. G., and Wong, T. Z. (1996). Automatic motion correction for breast MR imaging. *Radiology*, 198(3):903–906. 117