UNIVERSITÉ NICE SOPHIA ANTIPOLIS

ÉCOLE DOCTORALE STIC
SCIENCES ET TECHNOLOGIES DE L'INFORMATION ET DE LA COMMUNICATION

THÈSE DOCTORALE

pour obtenir le titre de

Docteur en Sciences

de l'Université Nice Sophia Antipolis
Discipline : AUTOMATIQUE, TRAITEMENT DU SIGNAL ET DES IMAGES

Soutenue par

Loïc LE FOLGOC

Apprentissage Statistique pour la Personnalisation de Modèles Cardiaques à partir de Données d'Imagerie

Superviseur de thèse : Hervé DELINGETTE
Co-Superviseur: Nicholas AYACHE
preparée à Inria Sophia Antipolis, équipe ASCLEPIOS
soutenue le 27 novembre 2015

Jury :

Rapporteurs :	Nikos PARAGIOS	-	École Centrale Paris (CVC)
	Bertrand THIRION	-	Inria (équipe PARIETAL)
Examinateurs:	Antonio CRIMINISI	-	Microsoft Research Cambridge (MLP)
	Sébastien OURSELIN	-	University College London (TIG)
Superviseur:	Hervé DELINGETTE	-	Inria (équipe ASCLEPIOS)
Co-Superviseur:	Nicholas AYACHE	-	Inria (équipe ASCLEPIOS)

UNIVERSITY OF NICE - SOPHIA ANTIPOLIS

DOCTORAL SCHOOL STIC

SCIENCES ET TECHNOLOGIES DE L'INFORMATION ET DE LA COMMUNICATION

P H D T H E S I S

to obtain the title of

PhD of Science

of the University of Nice Sophia Antipolis

Specialty : AUTOMATION, SIGNAL AND IMAGE PROCESSING

Defended by

Loïc LE FOLGOC

Statistical Learning for Image-based Personalization of Cardiac Models

Thesis Advisor: Hervé DELINGETTE

Thesis Co-Advisor: Nicholas AYACHE

prepared at INRIA Sophia Antipolis, ASCLEPIOS Team

defended on November 27, 2015

Jury :

Reviewers:	Nikos PARAGIOS	-	École Centrale Paris (CVC)
	Bertrand THIRION	-	Inria (PARIETAL Research Team)
Examinators:	Antonio CRIMINISI	-	Microsoft Research Cambridge (MLP)
	Sébastien OURSELIN	-	University College London (TIG)
Advisor:	Hervé DELINGETTE	-	Inria (ASCLEPIOS Research Team)
Co-Advisor:	Nicholas AYACHE	-	Inria (ASCLEPIOS Research Team)

This thesis was prepared at Inria, ASCLEPIOS Research Project, Sophia Antipolis, France. It was partly funded by the Microsoft Research – Inria Joint Centre and by the European Research Council through the ERC Advanced Grant MedYMA (2011-291080) on Biophysical Modeling and Analysis of Dynamic Medical Images.

Apprentissage Statistique pour la Personnalisation de Modèles Cardiaques à partir de Données d'Imagerie

Abstract: Cette thèse porte sur un problème de calibration d'un modèle électromécanique de cœur, personnalisé à partir de données d'imagerie médicale $3D + t$; et sur celui — en amont — de suivi du mouvement cardiaque. Les perspectives à long terme de la simulation personnalisée de la fonction cardiaque incluent l'aide au diagnostic et à la planification de thérapie, ainsi que la prévention des risques cardiovasculaires.

A cette fin, nous adoptons une méthodologie fondée sur l'apprentissage statistique. Pour la calibration du modèle mécanique, nous introduisons une méthode efficace mêlant apprentissage automatique et une description statistique originale du mouvement cardiaque utilisant la représentation des courants $3D + t$. Notre approche repose sur la construction d'un modèle statistique réduit reliant l'espace des paramètres mécaniques à celui du mouvement cardiaque.

L'extraction du mouvement à partir d'images médicales avec quantification d'incertitude apparaît essentielle pour cette calibration, et constitue l'objet de la seconde partie de cette thèse. Plus généralement, nous développons un modèle bayésien parcimonieux pour le problème de recalage d'images médicales. Notre contribution est triple et porte sur un modèle étendu de similarité entre images, sur l'ajustement automatique des paramètres du recalage et sur la quantification de l'incertitude. Nous proposons une technique rapide d'inférence gloutonne, applicable à des données cliniques $4D$.

Enfin, nous nous intéressons de plus près à la qualité des estimations d'incertitude fournies par le modèle. Nous comparons les prédictions du schéma d'inférence gloutonne avec celles données par une procédure d'inférence fidèle au modèle, que nous développons sur la base de techniques MCMC. Nous approfondissons les propriétés théoriques et empiriques du modèle bayésien parcimonieux et des deux schémas d'inférence.

Keywords: Problème inverse, suivi de mouvement cardiaque, recalage non-rigide, modélisation bayésienne parcimonieuse structurée, détermination automatique, méthodes de Monte-Carlo par chaînes de Markov

Statistical Learning for Image-Based Personalization of Cardiac Models

Abstract: This thesis focuses on the calibration of an electromechanical model of the heart from patient-specific, image-based data; and on the related task of extracting the cardiac motion from $4D$ images. Long-term perspectives for personalized computer simulation of the cardiac function include aid to the diagnosis, aid to the planning of therapy and prevention of risks.

To this end, we explore tools and possibilities offered by statistical learning. To personalize cardiac mechanics, we introduce an efficient framework coupling machine learning and an original statistical representation of shape & motion based on 3D+t currents. The method relies on a reduced mapping between the space of mechanical parameters and the space of cardiac motion.

The second focus of the thesis is on cardiac motion tracking, a key processing step in the calibration pipeline, with an emphasis on quantification of uncertainty. We develop a generic sparse Bayesian model of image registration with three main contributions: an extended image similarity term, the automated tuning of registration parameters and uncertainty quantification. We propose approximate inference schemes that are tractable on $4D$ clinical data.

Finally, we wish to evaluate the quality of uncertainty estimates returned by the approximate inference scheme. We compare the predictions of the approximate scheme with those of an inference scheme developed on the grounds of reversible jump MCMC. We provide more insight into the theoretical properties of the structured sparse Bayesian model and into the empirical behaviour of both inference schemes.

Keywords: Inverse Problem, Cardiac Motion Tracking, Non-rigid Registration, Structured Sparse Bayesian Learning, Automatic Relevance Determination, Markov Chain Monte Carlo

Acknowledgments

I would like to start by thanking my advisor Hervé Delingette, for your scientific oversight, your guidance and your availability throughout my thesis. I am grateful for all the ideas that you gave me to investigate and for all that I got to learn during my stay. I was both given the chance to keep pushing my ideas until I was happy with them, and given the (occasional regains of) motivation and safeguards to carry them through. The advice and directions that you provided during many helpful discussions shaped this work into better form. To Nicholas Ayache, thank you for welcoming me in the team, for providing such a great work environment and for what I learned from you. I truly enjoyed my stay at Asclepios. To Antonio Criminisi, many thanks for your availability and for your suggestions that often brought a fresh perspective on my research. It was always a pleasure to visit you at Cambridge.

Secondly, I want to express my sincere thanks to my reviewers, Bertrand Thirion and Nikos Paragios. I am grateful for the time you spent reviewing this manuscript, for your sharp comments and your encouragements. I would like to extend my thanks to Sébastien Ourselin for accepting to be member of the jury and for attending my defense. I am very glad to have all of you in my jury.

It is great time I thank all of my co-workers for the friendly atmosphere in the team! Special thanks to the people with whom I shared an office: to Ján Margeta and Stéphanie Marchesseau for your warm welcome when I arrived in the lab, for your support and for (very) patiently answering many of my random questions. To Hervé Lombaert for the helpful ideas, for the friendly discussions at work or around a drink. A shout-out to Hakim Fadil and Loïc Cadour for taking all the not-so-fun engineer jokes with a smile and always helping regardless. Dear Asclepios dinosaurs and youngsters alike, I thank you for the fond memories shared during my stay: whether hiking, climbing, gaming, kayaking, visiting a friend at his marriage, horse-riding or goofing around, you made it a fun few years (yes, the Bollywood movies too).

Big thanks to Isabelle Strobant for all of the help of course, but also for your kindness. I also take this opportunity to thank Maxime Sermesant and Xavier Pennec for the valuable discussions, the shared ideas and the constructive feedback; and for contributing to the friendly atmosphere!

Of course I ought to acknowledge Laurent Massoulié and the Microsoft Research – Inria Joint Center for supporting my research but, just as importantly, I would also like to express my warm thanks to you, to Hélène Bessin-Rousseau and to my colleagues there for giving me a spontaneous and friendly welcome upon each of my visit!

Finally I wish to simply thank my friends, my family, my parents and my brother.

Contents

CHAPTER 1

Introduction

Contents

In the following pages, we briefly review the motivations for this thesis, introducing the necessary background on cardiac function and modelling with emphasis on mechanics and motion. We reposition the problems we will address with respect to the state of the art and give an overview of the manuscript organization.

1.1 Context and Motivations

Cardiac diseases are associated worldwide with high morbidity and mortality. Detecting and preventing risks for such diseases, improving patient health, reducing morbidity has generated immense interest and is the subject of active research in many communities. The continued improvement of imaging modalities brings forward a wealth of opportunities to obtain information about the cardiac function non-invasively, with expected benefits for improved clinical diagnosis and therapy planning.

As a reflection of these challenges, the communities of medical image computing, computer anatomy and physiology have made tremendous progress over the past decades towards simulation of the human biology (and of particular interest to us, the cardiac function) as well as towards the automated analysis and understanding of medical images. Firstly, imaging softwares (MedInria, ParaView, Segment, Cardioviz3D, ITKSnap...) provide a set of standard tools for visualization and low-level processing of images, that already make it possible to extract clinically relevant (albeit simple) indices of the cardiac function (e.g. the volume of cardiac ventricles). Advances have been made towards the automated extraction of the cardiac geometry (also know as segmentation) and that of the cardiac motion (a.k.a. motion tracking or registration) from structural and longitudinal magnetic resonance or echocardiographic data. The geometry and kinematics of the

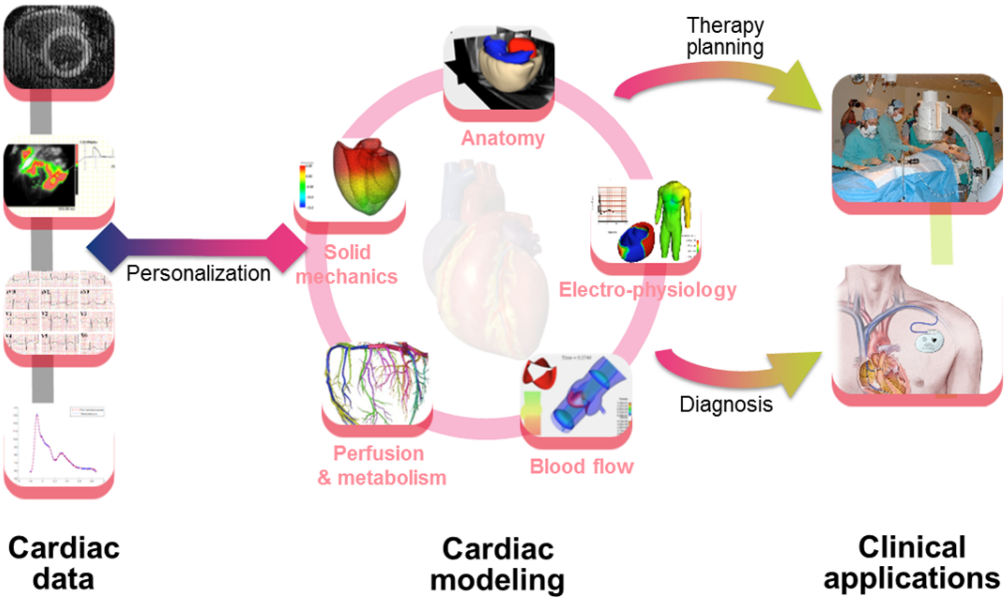


Figure 1.1: The Virtual Patient Heart long-term vision couples knowledge-driven computer models of the cardiac anatomy and physiology with patient-specific data to yield a faithful, personalized simulation that may aid clinical diagnosis as well as help predict successful therapies.

patient's heart potentially yield valuable tokens of their clinical state, either via derived indices (e.g. the ejection fraction) or direct analysis of the motion (dyssynchronous contraction). Of course, open challenges – either intrinsic to the tasks themselves, or brought forward by limitations in the image acquisition protocols – subsist and continue to motivate active research on the topic. Secondly, systematic advances into the biophysical and mathematical modelling of the cardiac physiology have resulted in a variety of computer simulators able to reproduce aspects of the cardiac function with increasingly high degree of faithfulness. This includes the simulation and integration of haemodynamics, electrophysiology, mechanics, metabolism...

This thesis work falls within the join Microsoft Research – Inria Medilearn project. The Medilearn project gathers researchers from the Machine Learning and Perception group (MSR Cambridge), Inria Asclepios and Parietal teams around several research axes, of which the latter directly motivates our work: the development of novel analytic approaches to learn from large-scale datasets of medical images, the automatic analysis and indexation of medical images [Margeta 2011, Lombaert 2014, Margeta 2015, Hoyos-Idrobo 2015, Lombaert 2015] and the development of accurately personalized heart models with the longer-term goal of assisting diagnosis and therapy selection for patients with a heart condition. Naturally patient-specific, personalized simulation of the cardiac behaviour calls for the coupling of computer simulations and of data assimilation techniques that incorporate (typically) image-based information so that the model may reproduce the observed behaviour. Thus the motivation for this work, as we will now see after a brief introduction

of the necessary background, is firstly the exploration of adequate strategies for the personalization of cardiac models – more specifically cardiac mechanics – which then motivated the development of novel tools for the decisive preprocessing step of extracting cardiac kinematics, a.k.a. motion tracking.

1.2 The Cardiac Function

The heart is an involuntary muscle that pumps blood throughout the human body via the circulatory system, in a closed loop. The deoxygenated veinous blood reaches the right atrium before being pushed into the right ventricle. During the cardiac contraction, it is ejected towards the lungs and reoxygenated. Similarly, it then arrives in the left atrium, is pushed down to the left ventricle and ejected towards the aorta and the arterial system during the contraction. Atria and ventricles are separated by atrioventricular valves that open (in a single direction) or close as a result of the pressure differential; similarly arterial valves control the flow between ventricles and arteria. Hence the cardiac contraction is divided into four phases, as summarized in Fig. 1.2. During the *filling* phase, the ventricles fill up with blood, first passively then actively due to the contraction of the atria. In a second phase (*isovolumetric contraction*), the atrioventricular valves close and the ventricles start contracting, resulting in a ventricular pressure rise. Upon reaching and exceeding the arterial pressure, the arterial valves open and the blood is ejected towards the arteria (*ejection* phase). Arterial valves close as the pressure differential reverts back and the *isovolumetric relaxation* starts. As soon as the atrium pressure exceeds the ventricular pressure, atrioventricular valves open and the filling phase starts anew.

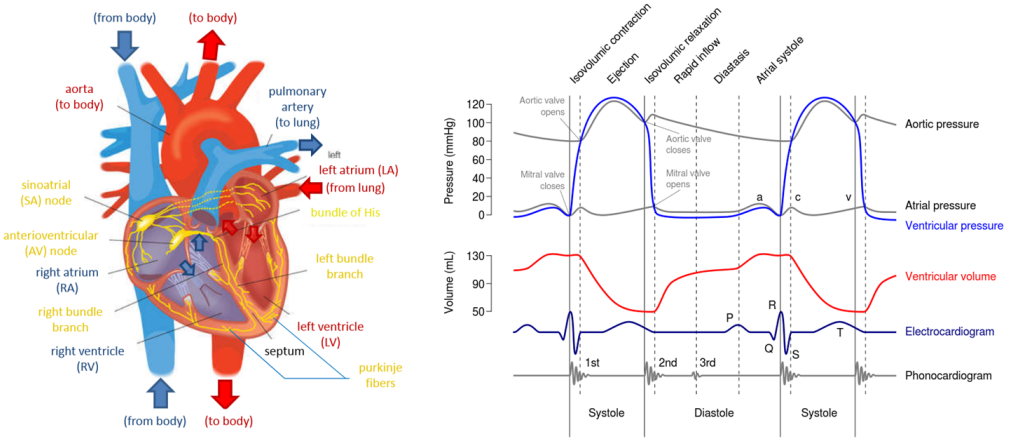
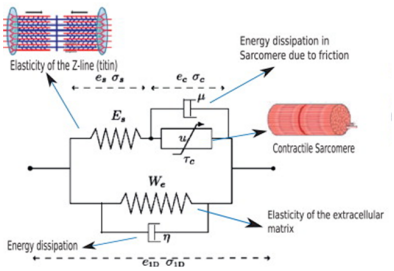


Figure 1.2: (Left) Simplified cardiac anatomy, illustrating the four chambers and their connection to the circulatory system. The blue compartment deals with de-oxygenated blood, the red compartment with oxygenated blood. The conduction system responsible for the synchronous contraction of the heart is overlayed in yellow. *Image from [Prakosa 2013a].* (Right) The Wiggers diagram summarizes the phases of the cardiac cycle. *Image from Wikipedia.*

The contraction of the cardiac muscle occurs preferentially along elongated cardiac fibers, which enroll in a structured manner around the cardiac chambers. The contraction rate is controlled thanks to a natural pacemaker, the sinoatrial node, that generates changes of electrical potential across cardiac cell membranes. The electrical wave propagates through the conduction system and diffuses through the cardiac muscle and is responsible for the (normally) synchronous contraction of the cardiac muscle (so-called myocardium).

1.3 A Macroscopic Multi-Scale Model of Cardiac Mechanics

A variety of models were developed for the simulation of cardiac electrophysiology or mechanics, ranging from cellular models of ionic interactions, of muscle cell structure and activity, to macroscopic models of the electrical wave propagation and of the myocardium mechanics. Our work relies on an implementation of the Bestel-Clement-Sorine (BCS) model [Bestel 2001, Chapelle 2012] described in [Marchesseau 2013a]. We refer the reader to the same manuscript for a discussion of the modelling literature and details on the BCS model, while only providing the necessary summary here. The BCS electromechanical model progressively derives, from a fine description of cellular processes and the recourse to statistical mechanics, a mesoscopic then a macroscopic model of cardiac mechanics. The macroscopic model, as illustrated in Fig. 1.3, is ultimately entirely determined by 14 global parameters, sometimes refined regionally. The model accounts for an element of active contractility (active stress τ_c) controlled by an electrical input u with a simplified behaviour dependent on four main input variables (including the time at which the electrical wave reaches the point of interest and the duration of the action potential), for energy dissipations and for the passive properties of the muscle cells and of the muscle matrix. The amplitude of the electrical input (parameters k_{ATP} and k_{RS}) govern the rates of active contraction and relaxation.



Notation	Parameter Name
σ_0 (MPa)	Max Contraction
k_0 (MPa)	Max Stiffness
k_{ATP} (s^{-1})	Contraction Rate
k_{RS} (s^{-1})	Relaxation Rate
E_s (MPa)	Linear Modulus
α	Cross-bridges Unfasten Rate
μ (MPa.s)	Viscosity
c_1 (kPa)	Mooney Rivlin Modulus
c_2 (kPa)	Mooney Rivlin Modulus
K (MPa)	Bulk Modulus
τ (s)	Wind. Charact. Time
R_p (MPa.m^{-3}.s)	Wind. Periph. Resistance
Z_c (MPa.m^{-3}.s)	Wind. Charact. Resistance
L (kPa.s^2.m^{-3})	Wind. Total Art. Inertance

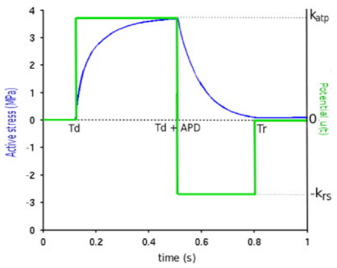


Figure 1.3: The electromechanical model. (Left) The mechanical model. The convention is that stress is additive for modules in parallel and deformations are additive (after linearization) for modules in series. (Middle) The 14 model parameters. (Right) Simplified model of the action potential controlling the contraction. Td: depolarization time (beginning of active contraction). APD: action potential duration. Tr: repolarization time (end of active relaxation, beginning of passive relaxation). Illustration adapted from [Marchesseau 2013a].

1.4 Patient-Specific Model Personalization: an Inverse Problem

Assimilating available data and tuning model parameters for the cardiac model to reproduce the observed behaviour will henceforth be referred to as cardiac model personalization. Data typically consists in $3D$ time series of MR images, which allows to study the motion of the cardiac muscle over the cycle. Several modalities for dynamic MRI exist. Cine SSFP MRI is acquired $2D+t$ slice by $2D+t$ slice across several cardiac cycles then reconstructed by resynchronizing stacks from ECG signals. The acquisition process leads to a lower number of axial slices and a lower axial resolution (as in Fig. 1.4 (Left)). Moreover, parts of the cardiac anatomy such as the basal area typically fall outside of the field of view, introducing higher uncertainty in the tracking of motion. Alternatively, tagged MRI images the cardiac motion via regular, grid-like 'tags' that deform over time, following the displacement of underlying physical points in the myocardium. This makes the tracking of motion at tag intersections easier. On the other hand anatomical structures are not visible in this modality (only tags), and tagged MRI data availability is much less widespread in the clinical practice. Challenges peculiar to the task of extracting the motion from cardiac images will be the basis for the second part of our thesis work.

Motion tracking is a common preprocessing step for model personalization as it makes processed data more amenable to comparison with model-based simulations, in the form of time series of volumetric meshes. Indeed, automatic or interactive tools already exist to create the $3D$ myocardium geometry [Ecabert 2011, Larrabide 2009], which can then be propagated via the output of motion tracking from a reference frame to the rest of the cycle. The task of personalizing cardiac mechanics from cardiac kinematics can then be seen as that of optimizing model parameter values for the simulated sequence of meshes to approximate in some sense the data-driven sequence of meshes. Optimizing parameter values from observed data is an *inverse* problem, by opposition to the *forward* problem of simulating the cardiac motion from given parameter values.

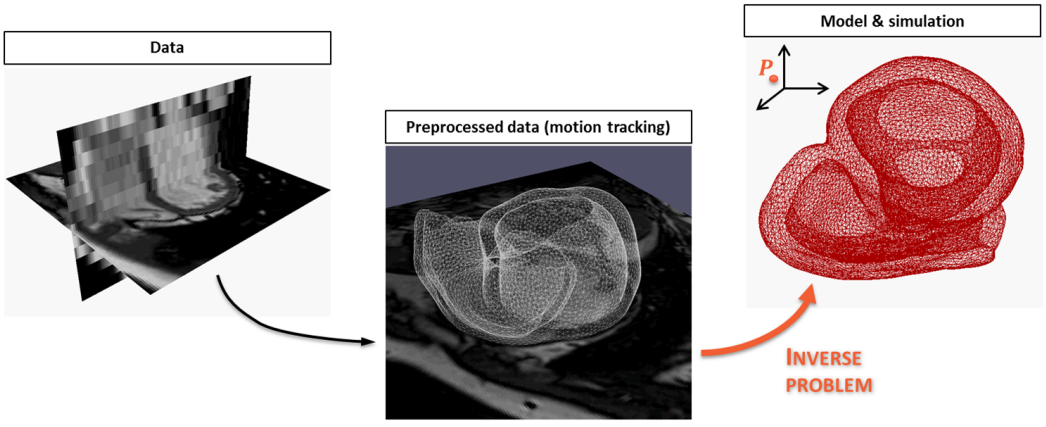


Figure 1.4: From patient-specific images to a personalized simulation: an inverse problem

This inverse problem has been tackled by different authors over the last five years. [Sundar 2009] and [Delingette 2012] propose adjoint variational methods to personalize

contractility parameters from synthetic cardiac motion, or real pathological cases for the latter. [Moireau 2011, Chabiniok 2012, Imperiale 2011] use Reduced Order Unscented Kalman Filtering (ROUKF) to estimate the contractility. [Xi 2011] compared ROUKF and Sequential Quadratic Programming for the estimation of passive material parameters from synthetic data. [Marchesseau 2013b] uses ROUKF to personalize contractilities from the observation of regional volumes on healthy and pathological cases. Variational methods rely on the differentiation of a cost function, which calls for expensive evaluations of the entire forward model and results in a long processing time. Sequential methods typically attempt to reduce the computational load by assimilating and correcting the discrepancy between observations and the predicted state at every time step during the simulation. These methods are intrusive (the code for the forward model has to be modified) and sensitive to initialization. [Marchesseau 2012] proposes to alleviate the dependency on the initial 'guess' with an Unscented (Kalman) Transform based calibration from global indices.

The motivation for our work finds its source in the challenges posed by the computational complexity and the highly non-linear behaviour of the forward model for classical optimization strategies. This context prompted us to explore the possibilities offered by machine learning. Specifically, can we accelerate the personalization procedure by systematic, parallelizable characterization of the relationship between the parameter space and the cardiac motion in an offline phase? Can we make the personalization more reliable and accurate? Anticipating on some of our conclusions, it was found that the estimation of parameters from the cardiac motion was subject to significant residual uncertainty even in well controlled synthetic settings. This uncertainty has several sources, as depicted in Fig. 1.5, partly intrinsic to the task at hand and partly caused by the experimental setting. In particular uncertainty in the motion tracking is crucial and highly conditions the ensuing personalization. Hence the rest of this thesis focuses on the development of reliable registration tools, with emphasis on a framework that opens perspectives for uncertainty quantification. Indeed, accounting for uncertainty on input kinematics in the personalization process is likely to improve its accuracy and robustness. In fact the inverse problem could be given a full probabilistic treatment, returning a faithful characterization of likely combinations of model parameters; preliminary work as recently been conducted in that direction [Neumann 2014].

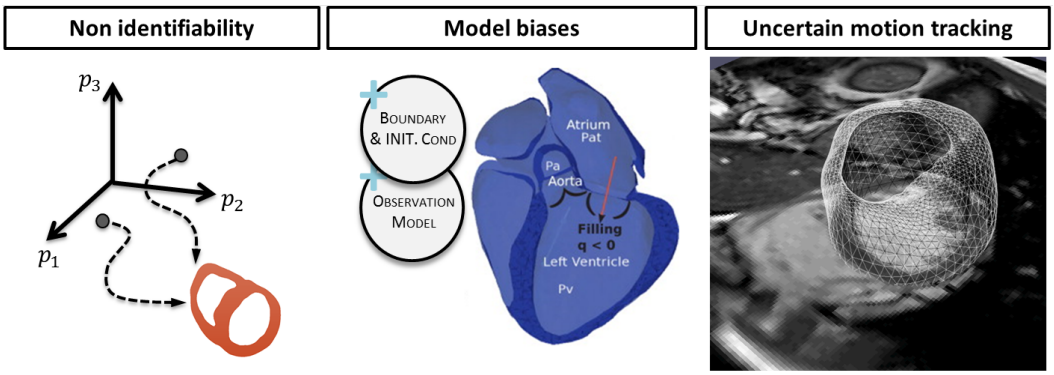


Figure 1.5: Sources of uncertainty in the personalization of cardiac mechanics

As an alternative to the motion tracking preprocessing, the personalization scheme could proceed to directly match parameters from image data but this calls for an adequate simulator of images given a prescribed motion. Such simulators are not yet widespread despite major recent advances, e.g. [Prakosa 2013b, Prakosa 2013a] and for $3D$ echocardiographic images, [Alessandrini 2015].

1.5 Cardiac Motion Tracking & Image Registration

Cardiac motion tracking can be more abstractly apprehended as a particular instance of image registration. Pairwise image registration addresses the problem, given two images I and J representing related objects or organs, of finding a transformation of space that matches homologous features between objects. Thorough reviews of the field can be found for instance in [Goshtasby 2012] and [Sotiras 2013]. Fig. 1.6 provides a graphic overview of the basic methodological review that follows.

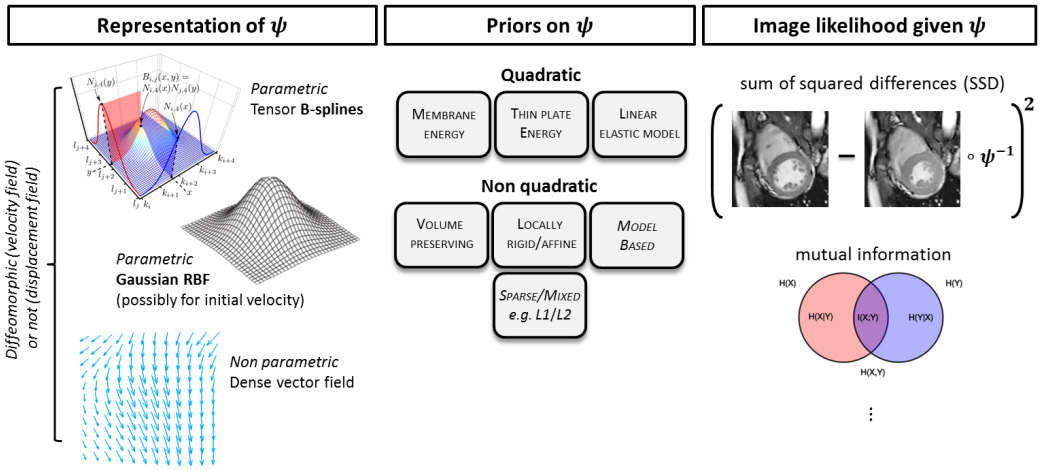


Figure 1.6: Building blocks for an image registration algorithm, relevant both from the classical standpoint of variational optimization and from that of probabilistic registration. Terminology changes according to the context (prior and image likelihood vs. regularizer and image similarity criterion). *B-spline and Gaussian RBF illustrations from [Brunet 2010].*

The quality of match is assessed via a metric of disparity, either between the intensity profiles of registered images, or between key features (landmark- or intensity-based) extracted by some other means. Intensity-based measures of discrepancy include the sum-of-squared differences of voxel intensities (SSD), which can be improved upon by modeling spatially varying noise levels [Simpson 2013] and artefacts [Hachama 2012], or by relaxing assumptions over the intensity mapping between images – *e.g.* to a piecewise constant mapping [Richard 2009], to a locally affine mapping [Cachier 2003, Lorenzi 2013] or to a more complex, non-linear (Parzen-window type) intensity mapping [Janoos 2012]. Other criteria derived from information theory, such as mutual information, allow the registration of images of different modalities [Wells III 1996]. Feature-based strategies are often highly

application dependent, although block matching [Ourselin 2000, Clatz 2005] and computer vision inspired features [Lowe 2004, Dalal 2005, Bay 2008, Calonder 2010, Rublee 2011] count among the most generic and widespread.

In many cases (such as our application of interest), images are not merely related by a rigid transform (neither a similarity or an affine transform) and a fully non-rigid transformation is sought. This has given rise to a variety of parametric or non-parametric representations of the transformation: piecewise-affine [Pitiot 2003], polyaffine [Arsigny 2003, Arsigny 2009], free form deformations parametrized by B-splines [Rueckert 2006], dense displacement fields [Thirion 1998], dense diffeomorphic representations [Dupuis 1998, Beg 2005, Arsigny 2006, Ashburner 2007, Singh 2015], parametric (e.g. Gaussian radial basis functions) diffeomorphic representations [Durrleman 2007, Sommer 2013], etc.

Non-rigid registration is ill-posed (due to the *aperture problem*, to the sheer dimensionality of the space of admissible transformations, etc.) and the incorporation of various soft or hard constraints has been proposed. In the dominant formulation following the work of [Broit 1981] (see also [Gee 1998]), registration attempts to minimize an energy that is the trade-off between the image discrepancy criterion and an $L2$ regularizer inspired from mechanics (e.g. from membrane or thin-plate models) or interpolation theory. Other constraints include incompressibility [Rohlfing 2003], local rigidity [Loeckx 2004] or mixed $L1/L2$ norms inducing sparsity of the parametrization [Shi 2012].

Finally, various gradient-free or gradient-based strategies have been implemented to tackle the resulting optimization problem [Klein 2007], from preconditioned gradient descent to a range of quasi-Newton methods possibly accelerated by full multigrid approaches [Haber 2006]. Alternatively, [Glocker 2007, Glocker 2008a, Glocker 2011] showed tremendous speed improvements by exploiting specific structure of the prior and image similarity energies (via Markov Random Fields), by quantizing the search space for displacements and solving the resulting problem by efficient linear programming. The framework allows for the efficient computation of min-marginals that measure the variations of the model energy under different constraints [Kohli 2008, Glocker 2008b]: a local measure of uncertainty is then derived and used to spatially adapt the displacement quantization, further improving the framework efficiency [Parisot 2014].

Many approaches to cardiac motion tracking have been proposed [Mäkelä 2002]. [Chandrashekara 2004] registers tagged MRI via the use of multi-level FFDs and normalized mutual information (NMI). Alternative grid topologies have been suggested for free form deformations that attempt to more closely match the heart shape [Heyde 2013] with expected gains in compacity and computational efficiency. In [McLeod 2013a, Mcleod 2013b], a cardiac-specific polyaffine model of motion with soft-incompressibility constraint is used instead. [De Craene 2010] addresses the problem of temporal consistency of the FFD by parametrizing the spatio-temporal velocity field with B-Spline kernels that have both a spatial and a temporal extent. The method is applied to both tagged MRI [De Craene 2012] and ($3D$ US) echocardiographic [De Craene 2012] sequences. [Shi 2013] showed an increase in registration accuracy when appending a sparsity-inducing prior ($L1$ norm) to a bending energy regularizer, yielding a sparse multi-level FFD representation. [Osman 1999] develops a dedicated harmonic phase

tracking algorithm for the tagged MR modality, exploiting the peculiar image structure in this modality; recent extensions of the approach showed state-of-the-art result (see [Zhou 2015], including a discussion of other tagged MR-specific methods). Finally, public benchmarks to evaluate the accuracy in terms of motion and strain of cardiac tracking methods were recently proposed [Tobon-Gomez 2013, De Craene 2013] for tagged MRI and $3D$ US data.

Many of the afore-mentioned approaches exploit parametrizations specifically crafted for the purpose of cardiac motion tracking, or third party data such as anatomical segmentations, to help the registration process. Although it is often a reasonable means to achieve better performance, it also comes at an obvious cost in terms of broadness of applicability. We will be interested in maintaining the genericity of the proposed methods while remaining competitive with the state-of-the-art performance-wise.

1.6 Open challenges in image registration: from an optimization standpoint to a probabilistic standpoint

Two main issues are left largely open in the state of the art. Firstly, registration algorithms are heavily dependent on a priori unknown hyperparameters, such as the optimal trade-off between image similarity and regularization. Optimal values of these hyperparameters are affected by a variety of spurious factors such as rescaling of the intensity profiles, resampling of images, coarseness of the parametrization (number of degrees of freedom in the parametrization) and the image acquisition process. Optimal values are also likely to change over the course of the cardiac cycle and from one subject to another (for instance, in the presence of pathologies affecting the cardiac motion). This renders cross-validation-based procedures extremely cumbersome or inefficient at finding optimal parameters. A natural question follows: can we automate the process of finding optimal hyperparameters? Can we find a principled, (somewhat) objective basis on which to define such optimality?

Secondly, confidence in the registration output and its accuracy is hampered by several limiting factors: inadequacy of the parametrization or of the prior assumptions, unknown model hyperparameters, lack of observability of the motion (aperture problem, textureless regions...). We would greatly benefit from estimates of the local confidence in the registration output that integrates these various unknown. In short, can we quantify uncertainty in the registration task?

There is yet little literature on the subject while acknowledging the associated challenges [Taron 2009, Kybic 2010], and we review it with more depth in the corresponding chapters. We note for now the seminal work of [Simpson 2012, Simpson 2013] and [Janoos 2012, Risholm 2013] that reinterpret registration in a Bayesian setting. This theoretical framework brings adequate tools to answer (at least partially) the above questions. Three main challenges have to be tackled to push this framework towards routine use: the first is to improve the modelling of registration (prior models and above all, the observation model that relates registered images), the second is the curse of dimensionality induced by the size of the transformation representation, the third is the development of efficient and

faithful inference schemes to characterize the quantities of interest.

1.7 Manuscript Organization and Objectives

The goals of this thesis led us to consider the following three problems and questions:

1. The personalization of cardiac mechanics poses challenges related to the computational complexity and the highly non-linear behaviour of the forward model [Marchesseau 2012, Marchesseau 2013b]. *Can machine learning help us accelerate and improve the estimation of parameters?*

2. The success of the personalization is crucially dependent on the quality of the preliminary motion tracking. Moreover unavoidable uncertainty in the motion tracking, along with other sources of bias and uncertainty, render the inverse problem fundamentally underdetermined. This legitimates the recourse to a probabilistic personalization of model parameters. In this thesis we focus on the upstream sub-task of developing reliable algorithms for, and quantifying uncertainty in the motion tracking. We focus on algorithms and methods that are generic, broadly applicable for medical image registration. We contribute with improved models of registration and partial answers to the following open issues. *Can we automate the process of tuning the registration parameters or circumvent the issue altogether? Can we quantify uncertainty in the registration process?*

3. *Are the estimates of uncertainty qualitatively satisfactory and useful?*

This thesis is organized around published or submitted work. Chapter 2 presents a machine learning approach for the personalization of cardiac mechanics, based on the work of [Le Folgoc 2012].

Chapter 3 develops a novel sparse Bayesian model of registration and a fast inference scheme. Bayesian modelling allows for self-tuning of model hyperparameters. By forcing the parametrization of the transformation to be sparse, we expect a lower computational complexity of the Bayesian inference and we make the evaluation of the covariance matrix on transformation parameters tractable. This covariance on transformation parameters can be translated into estimates of uncertainty on the displacement. The framework is applied on tasks of motion tracking on real and synthetic $3D$ cardiac data of various modalities. The chapter is based on [Le Folgoc 2015b] (submitted). We note for reference the early counterpart of this work presented in [Le Folgoc 2014], where several aspects of the methodology differed, including the recourse to a block matching strategy.

Chapter 4 is concerned with the soundness of uncertainties predicted by the sparse Bayesian model. The question prompted us to formalize in a fully Bayesian manner the proposed sparse Bayesian model of registration. Assumptions behind the fast inference scheme of chapter 3 are explicited and an alternative, exact inference scheme is proposed. The estimates of the 'optimal' transformation and of uncertainty returned by the two inference schemes are compared and analyzed on theoretical and empirical grounds. The chapter is based on the following article in preparation [Le Folgoc 2015a].

CHAPTER 2

Machine Learning and 4D Currents for the Personalization of a Mechanical Model of the Heart

Contents

2.1 Introduction

Patient-specific models may help better understand the role of biomechanical and electrophysiological factors in cardiovascular pathologies. They may also prove to be useful in predicting the outcome of potential therapeutic interventions for individual patients. In this chapter we focus on the mechanical personalization of the Bestel-Clement-Sorine (BCS) model, as described in [Bestel 2001, Chapelle 2012].

Model personalization aims at optimizing model parameters so that the behaviour of the personalized model matches the acquired patient-specific data (e.g. cine-MR images). Several approaches to the problem of cardiac model personalization have been suggested in the recent years, often formulating the inverse problem via the framework of variational data assimilation [Delingette 2012] or that of optimal filtering theory [Liu 2009b, Imperiale 2011, Chabiniok 2012]. The output of these methods is dependent on the set of parameters used to initialize the algorithm; for this reason calibration procedures are introduced as a preprocessing stage, such as the one developed in

[Marchesseau 2012]. Furthermore these approaches rely on on-line simulations, as an accurate estimation of the effect of parameter changes along several directions in the parameter space is required to drive the parameter estimation. Due to the complexity of the direct simulation these approaches are costly in time and computations.

In this chapter, we explore a novel machine-learning approach, in which the need for initialization and on-line simulation is removed, by moving the analysis of the parameter effects on the kinematics of the model (and thus the bulk of the computations) to an off-line learning phase. In this work we assume the tracking of the heart motion from images to be given (e.g. via [Mansi 2011]) and focus on the mechanical personalization of the cardiac function from meshes. Our work makes use of currents, a mathematical tool which was originally introduced to the medical imaging community in the context of shape registration [Vaillant 2005, Durrleman 2007] and offers a unified, correspondence-free statistical representation of geometrical objects. Our main contributions include the construction of 4D currents to represent, and perform statistics on $3D + t$ beating hearts and the proposal of a machine-learning framework to personalize electromechanical cardiac models.

The remaining of this chapter is organized as follows. In the first part we introduce the background on currents necessary to present the rest of our work. We develop our method in the following section, then present and discuss experimental results in the final sections.

2.2 Currents for Shape Representation

2.2.1 A Statistical Shape Representation Framework

Currents provide a unified representation of geometrical objects of any dimension, embedded in the Euclidean space $\mathbb{R}^n$, that is fit for statistical analysis. The framework of currents makes use of geometrically rich and well-behaved data spaces allowing for the proper definition of classical statistical concepts. Typically the existence of an inner product structure provides a straightforward way to define the mean and principal modes of a data set for instance, as in the Principal Component Analysis (PCA). These comments motivate an approach of currents from the perspective of kernel theory in this section, although currents are formally introduced in a more general way *via* the field of differential topology. The connection to differential topology is particularly relevant to outline the desirable properties of currents when dealing with discrete approximations of continuous shapes, in terms of convergence and consistence of the representation [Durrleman 2010].

A well-known theorem due to Moore and Aronszajn [Aronszajn 1951] states that for any symmetric, positive definite (p.d.) kernel on a set $\mathcal{X}$, there exists a unique Hilbert space $\mathcal{H}_K \subset \mathbb{R}^{\mathcal{X}}$ for which K is a reproducing kernel. This result suggests a straightforward way of doing statistics on $\mathcal{X}$ as long as a p.d. kernel K can be engineered on this set, by mapping any point $\mathrm{x} \in \mathcal{X}$ to a function $K(\mathrm{x}, \cdot) \in \mathcal{H}_K$ and exploiting the Hilbert space structure in $\mathcal{H}_K$. Furthermore, practical computations can be efficiently tracted thanks to

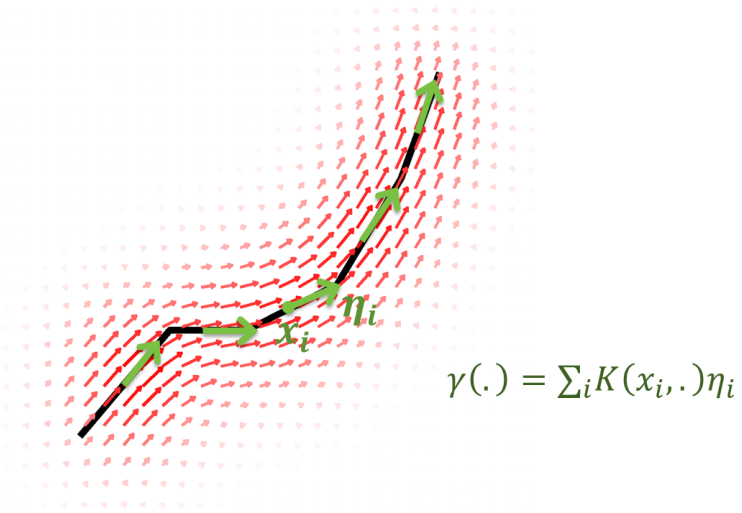


Figure 2.1: Discrete curve (in black) described as a collection of tangents $\{(\mathrm{x}_i, \eta_i)\}_{1\leq i\leq p}$ in green. In red: its associated representation as a smooth vector field (or as a dual smooth differential form), obtained by spatial convolution with a Gaussian kernel.

the *reproducing kernel* property - namely, for any $\mathrm{x}, \mathrm{y} \in \mathcal{X}$, we have

$$(K(\mathrm{x}, \cdot) \mid K(\mathrm{y}, \cdot))_{\mathcal{H}_K} = K(\mathrm{x}, \mathrm{y}) \,, \tag{2.1}$$

and more generally yet, for any $f \in \mathcal{H}_K$, $(K(\mathrm{x}, \cdot)|f)_{\mathcal{H}_K} = f(\mathrm{x})$. Expanding on this, one can compute statistics on pairs of points and m-vectors $(\mathrm{x}, \eta) \in \mathbb{R}^n \times \Lambda_m \mathbb{R}^n$ by mapping them to functions $K(\mathrm{x}, \cdot)\eta$ and making use of the reproducing property

$$(K(\mathrm{x}, \cdot)\eta | K(\mathrm{y}, \cdot)\nu) = \eta^\intercal \nu K(\mathrm{x}, \mathrm{y}) \,. \tag{2.2}$$

Eq. 2.2 simply extends Eq. 2.1 to vector-valued functions, making use of the fact that the tensor product of two kernels is again a kernel over the product space. Expanding the framework even further, we can regard a discrete shape as a finite set $\{(\mathrm{x}_i, \eta_i)\}_{1\leq i\leq p}$, where η_i describes the tangent space at x_i, and associate to it a signature function $\sum_{1\leq i\leq p} K(\mathrm{x}_i, \cdot)\eta_i$. Fig. 2.1 illustrates this construction, which can also be acknowledged as a special case of the convolution kernel on discrete structures described in [Haussler 1999] and [Gärtner 2002]. The correlation between two discrete shapes $\{(\mathrm{x}_i, \eta_i)\}_{1\leq i\leq p}$ and $\{(\mathrm{y}_j, \nu_j)\}_{1\leq j\leq q}$ can then be measured by the inner product

$$(\sum_i K(\mathrm{x}_i, \cdot)\eta_i \mid \sum_j K(\mathrm{y}_j, \cdot)\nu_j) = \sum_{i,j} \eta_i^\intercal \nu_j K(\mathrm{x}_i, \mathrm{y}_j) \,. \tag{2.3}$$

The above inner-product defines a correspondence-free way to measure proximity between shapes, trading hard correspondences for an aggregation of the measures of proximity between each simplex of one shape with every simplex of the other shape in the sense of a kernel $K(\cdot, \cdot)$. We have yet to specify a choice of kernel K. In the following, we will

consider the multivariate Gaussian kernel with variance Σ:

$$K_\Sigma(\mathrm{x},\mathrm{y}) = \frac{1}{\{(2\pi)^n|\Sigma|\}^{1/2}} \exp -\frac{1}{2}(\mathrm{x}-\mathrm{y})^\intercal \Sigma^{-1}(\mathrm{x}-\mathrm{y}) \ .$$

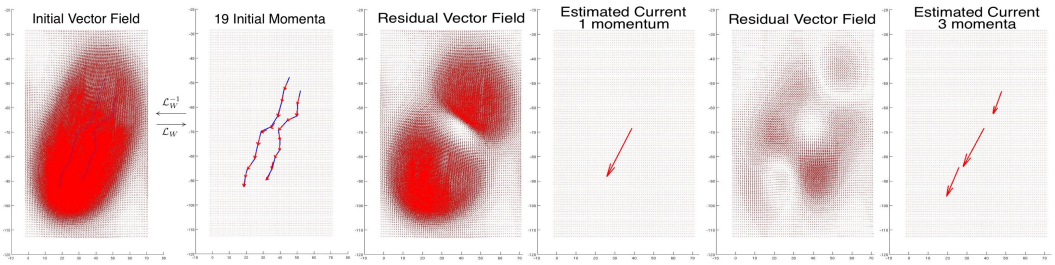
The choice of kernel width Σ can be interpreted as a choice of scale at which the shape of interest is observed: shape variations occurring at a lower scale are likely to be smoothed by the convolution and go unnoticed. This mechanism naturally introduces some level of noise insensitivity in the analysis. This parameter is decided on with regard to the mesh resolution and the level of noise in the data.

Finally, the linear point-wise evaluation functional $\delta_\mathrm{x}^\eta \colon \omega \mapsto \omega(\mathrm{x})(\eta)$ is continuous and dual to $K(\mathrm{x},\cdot)\eta$ by the reproducing kernel property. In the following we will refer to δ_x^η as a *delta-current* or a *moment*. To summarize, the discretized m-manifold $\{(\mathrm{x}_i,\eta_i)\}_{1\leq i\leq p}$ admits equivalent representations as the current $\sum_i \delta_{\mathrm{x}_i}^{\eta_i}$, its dual differential m-form $\sum_{1\leq i\leq p} K(\mathrm{x}_i,\cdot)\eta_i^\intercal$ or its dual vector field $\sum_{1\leq i\leq p} K(\mathrm{x}_i,\cdot)\eta_i$.

2.2.2 Computational Efficiency and Compact Approximate Representations

This framework lends itself to an efficient implementation. Firstly, the inner product between two discrete shapes can be computed in linear time with respect to the number of momenta through the use of a translation invariant kernel. Indeed $\gamma(\cdot) = \sum_i K(\mathrm{x}_i,\cdot)\eta_i$ may then be precomputed at any desired accuracy on a discrete grid by convolution, and rewriting $\sum_{i,j} \eta_i^\intercal \nu_j K(\mathrm{x}_i,\mathrm{y}_j)$ as $\sum_j \gamma(\mathrm{y}_j)^\intercal \nu_j$ exhibits the linear dependency w.r.t. the number of momenta.

Secondly, if the mesh diameter is small with respect to the scale Σ, the initial delta-current representation will be highly redundant. [Durrleman 2008] introduced an iterative method to obtain compact approximations of currents at a chosen scale and with any desired accuracy. We rely on this procedure at training time to fasten computations and reduce the memory load. This algorithm is inspired from the Matching Pursuit method [Davis 1997] and illustrated in Fig. 2.2. A compact current is built from the current S to approximate (of dual field γ) by iteratively adding a single delta current $\delta_{\mathrm{x}_n}^{\eta_n}$ to the previous approximation S_{n-1}, in such a way that the difference $\|S - S_n\|_{\mathcal{H}'_\Sigma}$ steadily decreases. This is achieved by greedily placing the moment at the maximum (in $\|\cdot\|_2$ norm) x_n of the residual field $\gamma(\cdot) - \gamma_{n-1}(\cdot)$, then choosing the optimal η, i.e. the one that minimizes $\|\gamma - \{\gamma_{n-1} + K(\mathrm{x}_n,\cdot)\eta\}\|^2_{\mathcal{H}_\Sigma}$. It is shown in [Durrleman 2008] that this algorithm is greedy in $\|\cdot\|_{\mathcal{H}_\Sigma}$ norm, and converges both in $\|\cdot\|_{\mathcal{H}_\Sigma}$ norm and $\|\cdot\|_\infty$ norm. The stopping criterion is on the residual norm $\|\gamma(\cdot) - \gamma_n(\cdot)\|^2_{\mathcal{H}_\Sigma}$. Our implementation uses a discrete kernel approximation of the Gaussian kernel, rather than an FFT based scheme, for fast local updates of the residual field.



(a) Initial Configuration (b) iteration 1 ($\|R_1\|_\infty = 0.51\sigma$) (c) iteration 3 ($\|R_3\|_\infty = 0.11\sigma$)

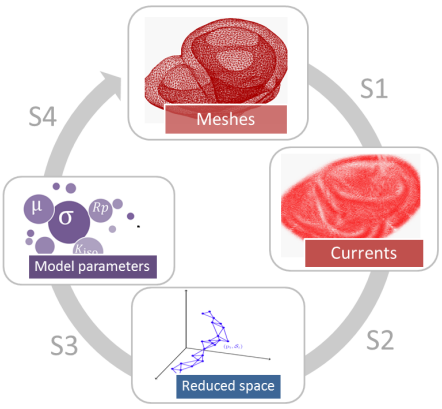
Figure 2.2: Sparse deconvolution scheme for currents: (a) the initial configuration. Right: two discrete curves in blue and their mean in red, as the collection of all tangents (from both curves) seen as momenta in the space of currents. Left: the Gaussian convolution of the initial momenta gives the dual representation of the mean as a dense vector field. (b) (Resp. (c)): first (resp. third) iteration of the matching pursuit algorithm: estimated momenta on the right panel, residual vector field on the left panel. The momenta converge to the true solution while the residual vector field tends to zero. Figure from [Durrleman 2009].

2.3 Method

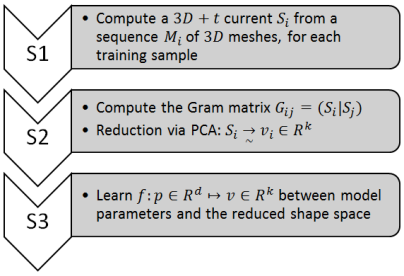
The workflow for the proposed machine-learning based parameter estimation method couples three successive processing steps: the first one aims at generating a current from an input sequence of meshes, so as to obtain a statistically relevant representation; the second one consists in a dimensionality reduction step, so as to derive a reduced shape representation in $\mathbb{R}^k$, which leads to computationally efficient statistical learning; the third step tackles the matter of finding a relationship between the reduced shape space and the (biophysical) model parameters. The three modules are mostly independent and can easily be adjusted in their own respect. As a machine learning based method, our work involves an off-line learning stage and an on-line testing stage: all three modules of the pipeline are involved during each stage. Fig. 2.3 gives a visual overview of our approach. The rest of this section describes the three afore-mentioned processing steps and their use during learning and testing stages.

2.3.1 Current Generation from Mesh Sequences

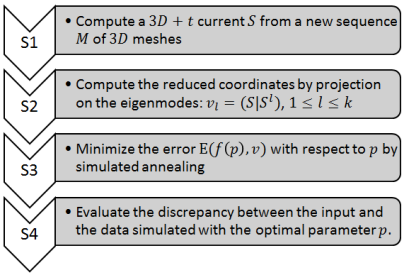
Let us briefly describe the way we build a current from a time sequence of 3D meshes. We first extract surface meshes from the volumetric meshes. This choice derives from the assumption that the displacement of surface points can be recovered more easily than the displacement of all points within the myocardium, given a sequence of images; thus learning from surface meshes may be more relevant for real applications. In this work we assume the trajectory of surface points to be entirely known, as opposed to the displacement in the direction normal to the contour only (aperture problem). Several variants to derive currents for 4D object representation can be discussed (e.g. [Durrleman 2010]), but their relevance largely depends on the application and complete processing work flow from the original data.



(a) Step-by-step process



(b) Learning phase



(c) Testing phase

Figure 2.3: Overview of the learning and testing phases.

In this work, we rely on the remark that the concatenation of smoothly deformed surface meshes can be visualized as a (3D) hyper-surface in 4D (Fig. 2.4). The ith simplex of this hyper-surface generates a current $\delta_{\mathrm{x}_i}^{\eta_i}$, where x_i is its barycenter and η_i is the vector of $\mathbb{R}^4$ normal to its support and of length the volume of the simplex. The current associated to the series of meshes is the aggregation of such delta currents, $\sum_i \delta_{\mathrm{x}_i}^{\eta_i}$. This construction captures both the geometry of the heart and its motion.

2.3.2 Shape Space Reduction

Since learning a direct mapping between the space of model parameters and the space of 3D+t currents is a cumbersome task, we introduce an intermediate step of dimensionality reduction via PCA. During the learning stage, we compute the mean current and principal modes of variation from the learning database of N currents $\{S_i\}_{1\leq i\leq N}$ generated from the N training mesh sequences $\{M_i\}_{1\leq i\leq N}$ as described in §2.3.1. This is achieved efficiently by computing the Gram matrix of the data $\mathbf{G}_{ij} = (S_i|S_j)$ column by column and using the so-called *kernel trick* [Schölkopf 2002]. Each column of $\mathbf{G}$ is computed in $\mathcal{O}(N \cdot P)$, where P is the maximum number of momenta among all currents S_j (cf. §2.2.1). Finally, we compute an approximate compact representation at the scale Σ of the mean current $\bar{T}$ and of the K first modes of variation $\{T_k\}_{1\leq k\leq K}$ to accelerate computations of inner products involving these currents, as in [Durrleman 2008].

At testing time and given a new current S, we derive its coordinates $v = (v_1, \cdots, v_K)$ in the reduced shape space by projection on the principal modes of variation, $v_k = (S - \bar{T}|T_k)$.

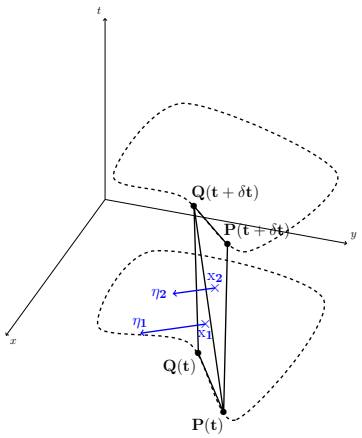
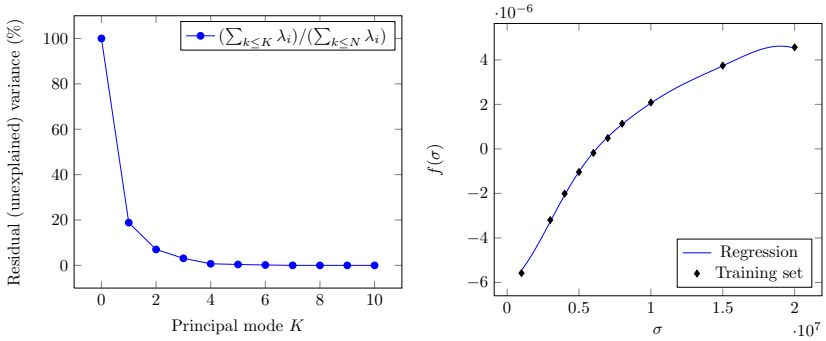


Figure 2.4: Current generation from a mesh element, illustrated on an element of contour in 2D deformed in time. The simplex PQ is followed over two consecutive timesteps, which gives a quad embedded in 3D. The quad is divided into two triangles, from which we get two current deltas, applied at each triangle barycenter, orthogonal to the support of their corresponding triangles and of norm the area of the triangle. For a surface in 3D deformed over time, each element of the triangulation followed over two consecutive timesteps generates a hyper-prism embedded in 4D, which is in turn decomposed in three tetrahedra from which we obtain three momenta.

2.3.3 Regression Problem for Model Parameter Learning

It remains to link the physiological (model) parameters to the reduced shape space. Although we are ultimately interested in finding an optimal set of parameters $\mathrm{p} \in \mathbb{R}^d$ from an observation $v \in \mathbb{R}^K$ we will actually learn a mapping in the other direction, $f\colon \mathrm{p} \in \mathbb{R}^d \mapsto v \in \mathbb{R}^K$. We motivate this choice by three arguments. Firstly, the observation v is a deterministic output of the cardiac model given a parameter set p and thus the mapping f is well-defined; however there may be several parameter sets resulting in the same observable shape and deformation, as parameter identifiability is not *a priori* ensured. Secondly, the parameter space is expected to be of smaller dimensionality than the reduced shape space and therefore easier to sample for combinatorial reasons. Finally, we can also expect that the set of biologically admissible model parameters be relatively well-behaved; on the other hand few points in the shape space may actually relate to anatomically reasonable hearts: thus mapping every $v \in \mathbb{R}^k$ to a parameter set could be impractical.

The regression function f is learned by kernel ridge regression using a Gaussian kernel [Hoerl 1970], and admits a straightforward close-form expression. During the testing phase, given a new observation v, we solve the optimization problem $\arg\min_{\mathrm{p}} \|f(\mathrm{p}) - v\|^2$ by Simulated Annealing [Xiang, Y. 2012]. This optimization problem involves an analytical mapping between low-dimensional spaces, as opposed to optimizing directly over the 4D meshes or currents. Thus it will not constitute a computational bottleneck regardless of



(a) Residual variance

(b) Regression function

Figure 2.5: Dimensionality reduction assessment for the first experiment. (Left) Residual unexplained variance with only the first $K = 0, 1, \ldots$ modes. (Right) Mapping between the parameter space and the first mode of variation, one-to-one over the domain of interest.

the chosen optimization scheme. Naturally, if a prior on the likelihood of a given parameter set $\mathrm{p} \in \mathbb{R}^d$ were known (e.g. via a biophysical argument), it could be integrated in the cost function in the form of a prior energy term $\lambda \cdot R(\mathrm{p})$.

2.4 Experimental Results

In our first experiment we focus on the prediction of the maximum contractility parameter σ_0 of the BCS model, defined globally for the whole cardiac muscle. Building on the sensitivity analysis from [Marchesseau 2012], we consider that σ_0 covers the range of values from 10^6 to $2\ 10^7$ in an anatomically plausible way. We form a training base of ten cases $\{\mathrm{p}_i, \mathcal{M}_i\}$ by sampling this range deterministically and launching simulations with the corresponding parameter sets, for a single heart geometry from the STACOM'2011 dataset. Following the PCA, the first principal mode of variation is found to explain 81% of the variance, thus we set the reduced shape space to be of dimension 1 ($K = 1$); the regression function ($\sigma = 0.3$, $\lambda = 10^{-5}$) bijectively maps the model parameter space and the reduced shape space. In all experiments, the model parameters are affinely mapped to $[-1, 1]$ for convenience, for the regression and optimization stages. We use an isotropic Gaussian kernel of width 1cm in space and 50ms in time.

We evaluate the performance of our approach by cross-validation on an independent test set $\{\mathrm{p}_j, \mathcal{M}_j\}_{0\leq j<N'}$ by randomly choosing parameter sets in the admissible range of parameters and launching the corresponding simulations. We thereafter refer to p_j as the *real* parameter (value) and to the output of our approach p_j^* as the *optimal* parameter (value). Our test set is of size $N' = 100$ samples. The whole personalization pipeline, from the current generation to the parameter optimization phase, takes roughly 2 minutes per sample on a regular laptop. We define the relative error on the parameter value for a given test sample j as $\varepsilon_r \mathrm{p}_j = |\mathrm{p}_j^* - \mathrm{p}_j|/\mathrm{p}_j$. In addition to the relative error, we consider the absolute error over the range of admissible parameters, $\varepsilon_a \mathrm{p}_j = |\mathrm{p}_j^* - \mathrm{p}_j|/|\mathrm{p}_{\max} - \mathrm{p}_{\min}|$.

We refer to ε_ap as an absolute error but express it for convenience as a percentage of the admissible parameter variation. Over the test set, we found a mean relative (resp. absolute) error of 9.2% (resp. 4.5%) and a median relative (resp. absolute) error of 6.8% (resp. 2.3%).

As a preliminary evaluation of the robustness of our approach with respect to geometry changes, ten samples are generated following the same procedure as before but using another heart geometry of the STACOM dataset. The 10 mesh sequences are manually registered (translation, rotation and scale) to the training geometry based on the end-diastole mesh before applying the normal pipeline, as described in Section 2.3. The mean relative (resp. absolute) error on the contractility parameter over our sample is 25% (9.3%), with 15% (resp. 7.5%) median relative (absolute) error.

The second experiment proceeds similarly to the first one, but we simultaneously estimate the contractility σ_0, the relaxation rate k_{rs} and the viscosity μ. For the training phase, the parameter space is sampled on a $7 \times 7 \times 7$ grid with σ_0 in the range $[10^6, 2\ 10^7]$, k_{rs} in $[5, 50]$ and μ in $[10^5, 8\ 10^5]$. The explained variance with 1 eigenmode of the PCA (resp. 2 to 5) out of the $N = 343$ modes equals 63.2% of the total variance (resp. 80.3%, 89.5%, 94.1%, 96.7%). We set the dimension of the reduced shape space to $K = 3$. The performance is tested on $N' = 100$ random samples. Because we can no longer assume the parameter set to be identifiable *a priori*, we introduce another measure of the goodness of fit of our personalization by directly evaluating the error on the observations. Given two surface mesh sequences $\mathcal{M} = \{\mathcal{M}_i\}_{1 \leq i \leq T}$ and $\mathcal{M}' = \{\mathcal{M}'_i\}_{1 \leq i \leq T}$, we define the pseudo-distance $d_{\text{sur}}(\mathcal{M}, \mathcal{M}') = \max_i d_s(\mathcal{M}_i, \mathcal{M}'_i)$ where $d_s(\mathcal{M}_i, \mathcal{M}'_i)^2$ is the mean square distance of the points of the surface $\mathcal{M}_i$ to the surface $\mathcal{M}'_i$. Additionally given one-to-one correspondences between $\mathcal{M}$ and $\mathcal{M}'$, we can define the distance $d_{\text{nod}}(\mathcal{M}, \mathcal{M}') = \max_i d_p(\mathcal{M}_i, \mathcal{M}'_i)$, where $d_p(\mathcal{M}_i, \mathcal{M}'_i)$ is the mean distance between corresponding nodes of $\mathcal{M}_i$ and $\mathcal{M}'_i$. While d_{sur} intuitively relates to an upper bound for the matching between surface meshes at any time step, d_{nod} conveys more information about the quality of the matching of point trajectories. The results for this experiment are reported in Table 2.1 and in Fig. 2.6. As a comparison, two mesh sequences corresponding to extreme values in the parameter set will yield a value for $d_{\text{sur}}(\mathcal{M}, \mathcal{M}')$ (resp. $d_{\text{nod}}(\mathcal{M}, \mathcal{M}')$) of the order of 6mm (resp. 8mm).

In addition we compute the optimal parameters and performance indicators for a different choice of the reduced space dimension K, obtaining quasi-identical statistics for $K = 4$. Finally, we test here again the robustness with respect to changes of the heart geometry. Using the same procedure as before on 10 test samples on a different geometry,

	$\varepsilon_r\sigma_0$ ($\varepsilon_a\sigma_0$)	$\varepsilon_r k_{rs}$ ($\varepsilon_a k_{rs}$)	$\varepsilon_r\mu$ ($\varepsilon_a\mu$)	d_{sur} (mm)	d_{nod} (mm)
Mean	15.2% (8.0%)	48.8% (26.4%)	40.5% (20.0%)	0.92mm	1.42mm
Median	13.2% (6.3%)	44.7% (19.6%)	32.1% (17.5%)	0.80mm	1.32mm

Table 2.1: Experiment 2 - results

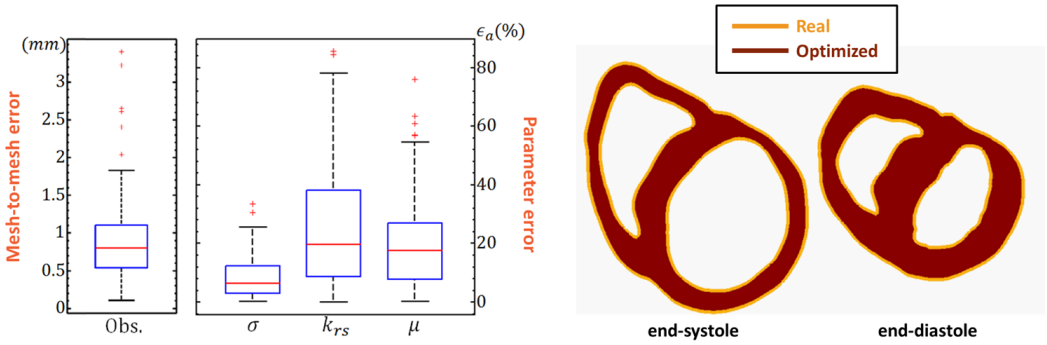


Figure 2.6: Benchmark for the second experiment. (Boxplots) Statistics are computed over the test set. The left boxplot quantifies the quality of fit in the observation space (mesh-to-mesh error). The right boxplot reports errors between estimated and ground truth parameters. (Mesh slices) Visual display of the quality of fit, on an example case, on a 2D slice of the cardiac muscle at two representative frames during the cardiac cycle (respectively end of relaxation phase and end of contraction phase).

we find a mean error of 1.4mm and a median value at 1.3mm for d_{sur} (respectively, 1.8mm and 1.6mm for d_{nod}).

2.5 Discussion

Despite working around the bias and error introduced by the model and image processing in real applications, our synthetic experiments show promising performance for our framework in terms of accuracy, tolerance to non-linear effects of parameters, robustness and computational efficiency. The accuracy of our approach was found to be below the typical voxel dimension (1mm), while a priori optimizing among a very wide range of parameter values at test time, and using a reasonable number of training samples at learning time. Although a single geometry is used for the training phase, the accuracy was of the same order on similar (non-pathological) heart geometries. Naturally, further work should handle geometry variability in a proper way, taking it into account at the training stage, and adding "shape factors" to the model parameter space capturing 3D shape variability. Moreover the addition in the pipeline of a pre-clustering stage with respect to the heart geometry, so as to distinguish very different geometries and treat them separately, should reduce the number of samples required to cover the whole parameter space while achieving better model personalization.

The proposed framework also brings an interesting perspective on the issue of parameter identifiability. It should be noticed that we achieve good results in terms of spatial distance between the matched model and observations while significant differences in the parameter space may still be observed. Parameter identifiability encompasses two distinct aspects. Firstly, small variations of the parameter values may result in changes that are not noticeable at the scale of reference. This sensitivity to parameters partially explains the er-

ror on the retrieved set of parameters. In our approach, the kernel width for currents impacts the ability of the algorithm to discern shape differences. In the future we will experiment with smaller kernel widths and improve algorithms to handle increased computational cost. Secondly in joint parameter estimation, a whole subset in the parameter space may result in identical observations, which also affects parameter identifiability. Such considerations can be analyzed in depth at the regression or optimization steps: several parameter sets with similar costs along with a measure of local sensitivity around these values may be additionally output by the Simulated Annealing algorithm. Biophysical priors may also be introduced at the optimization step by penalizing unlikely parameter sets without adding significant computational cost.

Finally more efficient machine learning algorithms should be tested in lieu of PCA, so as to capture non-linear 4D shapes variation, and to obtain and exploit precise information about the manifold structure of 4D heart shapes. Not only will this be of help with parameter identifiability and to derive efficient representations in the reduced shape space, but it could also provide valuable feedback for "smart" sampling of the parameter space.

2.6 Conclusion

A machine-learning current-based method has been proposed for the personalization of electromechanical models of the heart from patient-specific kinematics. A framework to encapsulate information regarding shape and motion in a way that allows the efficient computation of statistics via 4D currents has been described. This approach has been evaluated on synthetic data using the BCS model, with the joint estimation of the maximum contraction, relaxation rate and viscosity. It is found that the proposed method is accurate, computationally efficient and robust.

CHAPTER 3

Sparse Bayesian Registration of Medical Images for Self-Tuning of Parameters and Spatially Adaptive Parametrization of Displacements

Contents

3.1 Introduction

Non-rigid image registration is the ill-posed task of inferring a deformation Ψ from a pair of observed (albeit typically noisy), related images I and J. Classical approaches propose to

minimize a functional which weighs an image similarity criterion $\mathcal{D}$ against a regularizing (penalty) term $\mathcal{R}$:

$$\arg\min_{\Psi} \mathcal{E}(\Psi) = \beta \cdot \mathcal{D}(I, J, \Psi) + \lambda \cdot \mathcal{R}(\Psi) \tag{3.1}$$

Prior knowledge to precisely model the space of plausible deformations or the regularizing energy is generally unavailable. The optimal trade-off between the image similarity term and the regularization prior is itself difficult to find. Typically the user would manually adjust the ratio λ/β until a qualitatively good fit is achieved, which is time consuming and calls for some degree of expertise. Alternatively if quantitative benchmarks are available on a similar set of images, they can serve as a metric of reference on which to optimize parameters, under the assumption that the value that achieves optimality is constant across the dataset – say, for images of the same modality or among a given sequence. Unfortunately, this assumption generally does not hold. A major feat in the recent literature [Richard 2009, Simpson 2012, Risholm 2013] was to realize that this issue can be tackled automatically by reinterpreting registration in a probabilistic setting. [Gee 1998] first noted that, in a Bayesian paradigm, the two terms in Eq. (3.1) relate respectively to a likelihood and prior on the latent transformation Ψ. In fact the parameters λ and β themselves can be treated as hidden random variables, equipped with broad prior distributions, and jointly inferred with Ψ or integrated out.

In practice, analytical inference is precluded and various strategies are devised for ap-

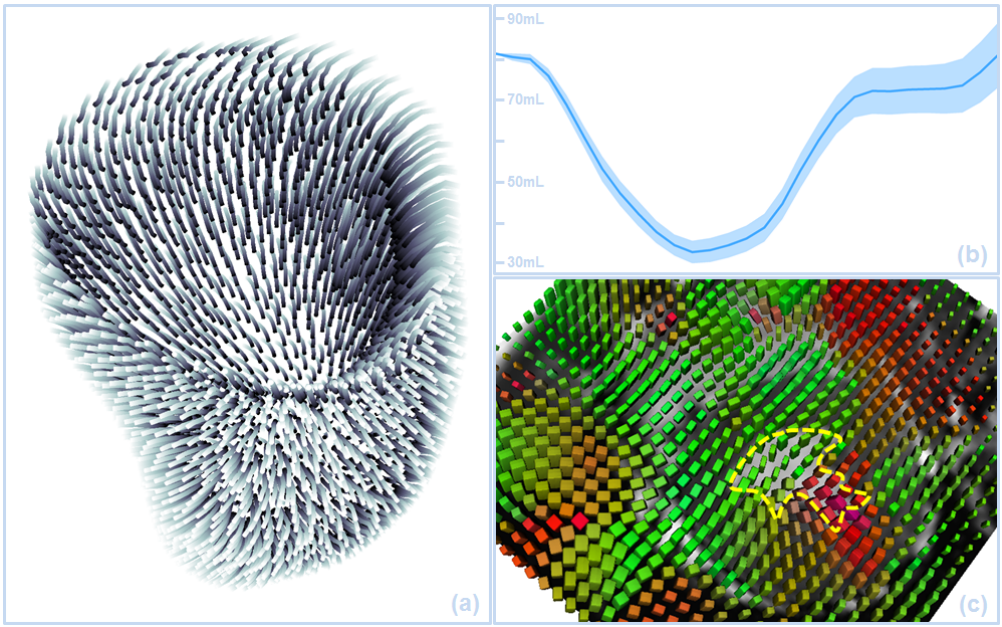


Figure 3.1: (a) Trajectories of points on the endocardium, following the registration of a time series of cardiac MR images by the proposed approach. (b) LV volume over time and 99.7% confidence interval. (c) Tensor visualization of directional uncertainty at end-systole, rasterized at voxel centers of a 2D slice. For a thorough description, please refer to the discussion in section 3.5.

proximate inference. [Risholm 2013] characterize the distributions of interest from MCMC samples. This is a most principled, generic and accurate approach provided that enough samples can be drawn within the available computational budget. Aside from monitoring the progress of the scheme, two difficulties arise: crafting an efficient proposal distribution over Ψ and computing the acceptance probability of the proposed sample. To circumvent this latter issue, the authors sample from an approximate posterior distribution derived in a variational free-energy framework. Alternatively, the *full* inference can be tackled in a variational framework. In this spirit, [Simpson 2012, Simpson 2015] derive a fast Bayesian approach using variational Bayes Expectation Maximization. This offers an appealing compromise between the computational burden and the quality of the estimates, depending on the chosen family of variational (approximate) posterior distributions.

Despite the incurred computational cost, this probabilistic view of registration offers substantial benefits. In this chapter, we demonstrate how such a framework can be extended so as to automatically select the scale and location of bases used to parameterize the transformation Ψ. The complexity of the mapping adapts to the underlying dataset, yielding a reduced set of relevant degrees of freedom: finer bases get introduced only in the presence of coherent image information and motion, while coarser bases ensure better extrapolation of the motion to textureless, uninformative regions. Spatial refinement of the parametrization was previously handled heuristically [Rohde 2003]; or it led to alternative formulations of registration via spatially anisotropic filtering [Stefanescu 2004]. [Stewart 2003] proposed to select a deformation model on the basis of information criteria, in a computationally favorable setting, choosing regionally among a limited pool of models (*e.g.* similarity, affine, quadratic). Here and to our knowledge, for the first time, it is approached on principled grounds within a probabilistic framework. We develop a statistical model of registration in which a reduced parametrization of the transformation is automatically inferred from the data jointly with all model parameters; and propose an efficient algorithmic scheme that renders inference over this model tractable for real scale registration tasks. In particular we extend the scope of a state-of-the-art tool for sparse regression and classification [Tipping 2001, Tipping 2003]. To increase its potency in the context of registration, we lift a strong assumption of independency on hidden variables, so that it now handles generic quadratic regularizing priors at no cost in algorithmic complexity. We also generalize it to multi*valued* regression (regression of vector fields as opposed to scalar fields), so as to preserve the natural invariance to a change of coordinate system.

This chapter expands on earlier work of the authors [Le Folgoc 2014] in several ways. Firstly, we propose a different approximation of the likelihood term, effectively removing a computational bottleneck – specifically, the voxelwise, local optimization of the image similarity *via* dense block-matching. Instead, a step of global optimization of the posterior distribution w.r.t. the reduced parametrization (typically, a hundred degrees of freedom) is performed. Furthermore we introduce a flexible noise model that can account robustly for acquisition noise and artefacts, and seamlessly adapts to a range of image modalities. We demonstrate our approach on tasks of motion tracking on real cardiac data, specifically time sequences of 3D cine or tagged MR images and echocardiographic images.

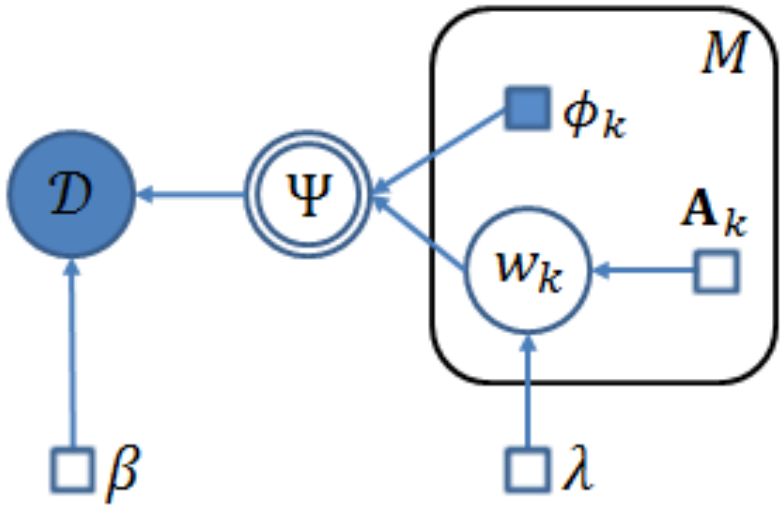


Figure 3.2: Graphical representation of the probabilistic registration model. The data pair D is put in relation via a transformation Ψ of space, up to some noise (β). The transformation is parameterized by weights w_k on a predefined overcomplete set of basis functions $\{\phi_k, k = 1 \cdots M\}$. Priors on the transformation smoothness and on the relevance of individual bases introduce additional parameters (λ and $\mathbf{A}_k$ respectively). Random variables are circled, whereas deterministic parameters are in squares. Arrows capture conditional dependencies. Shaded nodes are observed. The doubly circled node indicates that the transformation Ψ is fully determined by its parent nodes (the ϕ_k and $\boldsymbol{w}_k$). The plate stands for M (replicated) groups of nodes, of which only a single one is shown explicitely.

3.2 Statistical Model of Registration

Registration assumes that the dataset of interest describes objects related *via* some transformation of space. In a medical context, this occurs for instance when the motion of organs is followed over time in a sequence of images. Registration then aims at recovering the underlying transformation of space from the observed data. This can be formally regarded as an *inference* problem and handled as such. We start by expliciting our statistical model of pairwise registration. Fig. 3.2 provides a graphical depiction thereof. We specify a generative model of the data (*e.g.* images, landmarks) $D = \{D_1, D_2\}$ given the transformation Ψ, along with a prior over the admissible set of transformations. The general strategy to infer the parameters of this model is exposed at the end of the section.

The abstract graphical model depicted in Fig. 3.2 bears a strong resemblance to that of the Relevance Voxel Machine of [Sabuncu 2011], developed independently by the authors for regression and classification tasks. In Section 3.3 we propose alternative inference schemes with significant gains in algorithmic complexity, very much in the spirit of how the later work of [Tipping 2003] improved upon the original Relevance Vector Machine [Tipping 2001]. This effectively renders the approach applicable to non-rigid registration.

3.2.1 Data Likelihood

A good transformation Ψ should adequately map the datasets $D = \{D_1, D_2\}$, up to some misalignment and residual error attributable to the data formation process. Our knowledge of this process is captured in a *likelihood* model, which assigns a probability $p(D|\Psi)$ for the data D to be observed under some transformation Ψ. The data likelihood is commonly related to the data-matching energy that appears in the classic optimization framework of Eq. (3.1) by:

$$\mathcal{D}(D, \Psi) = -\ln p(D|\Psi) + \text{const}\,. \tag{3.2}$$

For landmark registration, $\mathcal{D}(D, \Psi)$ is generally chosen to be the sum of squared distances between pairings. Alternatively the quadratic loss can be replaced by robust losses such as the L_1 norm, or other loss functions derived from the heavy tailed family of Student-t distributions [Tipping 2005]. For pairwise registration of images, various data-matching terms were introduced in the literature. The simplest and most common image similarity term is the sum of squared difference (SSD) of voxel intensities, which can be improved upon by modeling spatially varying noise levels [Simpson 2013] and artefacts [Hachama 2012], or by relaxing assumptions over the intensity mapping between images – *e.g.* to a piecewise constant mapping [Richard 2009], to a locally affine mapping [Cachier 2003] or to a more complex, non-linear (Parzen-window type) intensity mapping [Janoos 2012]. Mutual information is another popular image similarity, especially in the context of registering images of different modalities [Wells III 1996], and has been successfully applied to the registration of cardiac images [Chandrashekara 2004].

SSD is a simple yet efficient image similarity term for registration of monomodal cardiac images. It naturally lends itself to a probabilistic interpretation and eases mathematical derivations. The target image J is modeled as the warped source image $I \circ \Psi^{-1}$ further corrupted by additive, independent identically distributed (i.i.d.) noise $e_i \sim \mathcal{N}(0, \beta)$ at each voxel $i = 1 \cdots N$:

$$J = I \circ \Psi^{-1} + \boldsymbol{e} \tag{3.3}$$

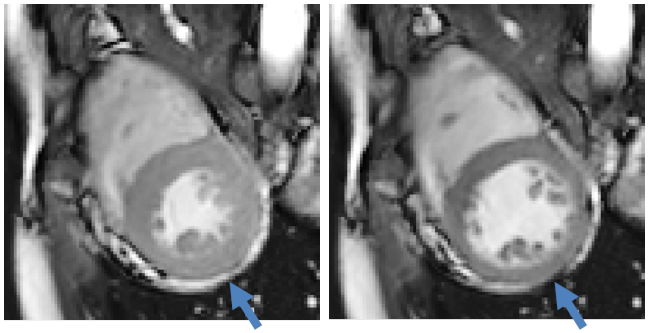
where $\boldsymbol{e} \sim \mathcal{N}(0, \beta\mathbf{I})$, $\mathbf{I}$ the $N \times N$ identity matrix. β is a global scaling parameter: it denotes the precision (or inverse variance) of the noise across the image. The SSD model can be described in a more familiar manner by the energy of Eq. (3.4), where $\{c_i\}_{i=1}^N$ is the list of voxel centers in the fixed image and $C_i = \Psi^{-1}(\mathrm{c_i})$ are the paired coordinates in the moving image.

$$\mathcal{D}_\beta(I, J; \Psi) = \frac{\beta}{2} \sum_{i=1}^{N} \big(J\big[\mathrm{c_i}\big] - \mathrm{I}\big[\mathrm{C_i}\big]\big)^2 \tag{3.4}$$

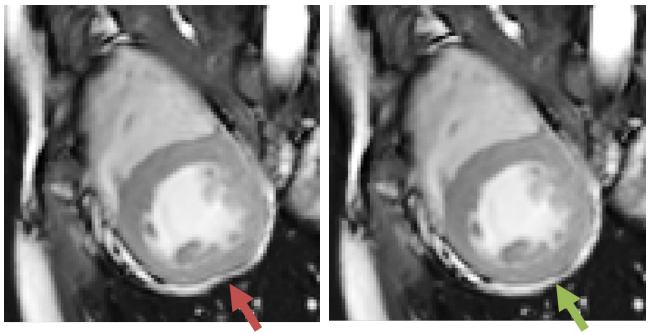
Since the SSD is quadratic w.r.t to intensity differences of paired voxels in the registered images, both the penalty for intensity discrepancies and the *rate* at which it grows can become arbitrarily high. This renders registration particularly vulnerable to strong local intensity biases, introduced for instance by topology changes in the imaged objects or by acquisition artefacts. Fig. 3.3 demonstrates such an occurrence where the shortcomings of SSD result in a qualitatively poor registration. Furthermore residual misalignments between structures of interest in the pair of registered images typically yield higher intensity residuals than those observed at background voxels, as seen in Fig. 3.4a. Sources of model

bias and acquisition noise cannot be captured together in a plausible manner with a single, spatially uniform noise level. In other words, the SSD noise model lacks both the flexibility and robustness required to cope with the various sources of discrepancy in the intensity profiles of the fixed and warped images I and J. To address this limitation we propose to model the noise at each voxel $i = 1 \cdots N$ with a mixture of Gaussian distributions $e_i \sim \sum_{l \leq L} \pi_l \mathcal{N}(0, \beta_l)$. The corresponding data matching energy is given in Eq. (4.2), where we denoted by $Z_l = \sqrt{2\pi/\beta_l}$ the normalizing constant for the Gaussian probability distribution function.

$$\mathcal{D}_{\boldsymbol{\beta},\boldsymbol{\pi},L}(I, J; \Psi) = -\sum_{i=1}^{N} \log \sum_{l=1}^{L} \frac{\pi_l}{Z_l} \exp -\frac{\beta_l}{2} \left(J\big[\mathrm{c_i}\big] - \mathrm{I}\big[\mathrm{C_i}\big] \right)^2 \tag{3.5}$$



a. Moving image b. Fixed image



c. SSD loss d. Robust loss

Figure 3.3: We illustrate the appeal of a robust variant of the SSD image loss based on a mixture-of-Gaussians model (GMM). Images (c,d) display the output warped images obtained after registering images (a) and (b), using respectively the SSD-based likelihood or the GMM-based likelihood (section 3.2.1). The arrows point towards a specific region that highlights the limitations of the SSD: the subset of hypo-intense voxels bordering the myocardium in the fixed image has no evident counterpart in the moving image. The SSD still drives the motion towards the best matches intensity-wise, which induces implausible tangential stretch of the myocardium. The GMM, on the other hand, incorporates a natural mechanism to downweight regions that cannot be reliably paired from image to image based on intensity values. The inferred motion is qualitatively closer to our expectations.

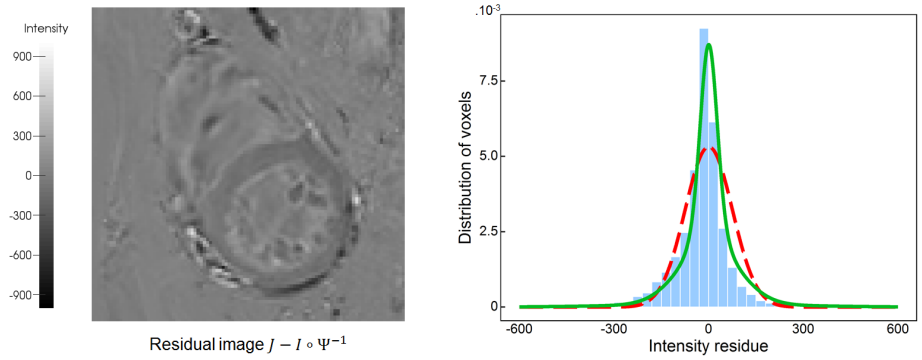


Figure 3.4: (Left) Example residual image after pairwise image registration, as per the example of Fig. 3.3. Artefacts and structures of variable appearance between registered images stand out in a distinct manner, much unlike ambient noise. The intensity of the cardiac muscle itself differed between moving and fixed image. (Right) Histogram of intensity residuals, with SSD and GMM fits overlayed respectively in red and green.

At each voxel, the residue is implicitely associated to a component of the mixture. This naturally yields a spatially varying model of noise that is better suited to render the complexity of noise patterns in medical images. Unlike previous work [Simpson 2013, Le Folgoc 2014] we do not assume that the noise varies smoothly across the image, as patterns arising from misalignment and imaging artefacts are local in nature. Moreover the parameters L, $\boldsymbol{\pi} = \{\pi_l\}_{l \leq L}$ and $\boldsymbol{\beta} = \{\beta_l\}_{l \leq L}$ of the mixture will be jointly inferred from the data – so that it fits the expected distribution of intensity residuals between registered images. Fig. 3.4b shows the histogram of intensity residuals obtained after registering the images depicted in the 2D example of Fig. 3.3, along with the Gaussian mixture fit jointly during the registration process. From the standpoint of energy-based formulations, the procedure effectively learns from the data the most appropriate data matching energy among a prior family of candidates. The respective profiles of the standard SSD loss and the learned Gaussian mixture (GMM) loss are comparatively displayed in Fig 3.5 for the aforementioned example. The characteristic inflexion of the GMM loss, with a reduced growth rate as the intensity residual becomes higher, is responsible for its robustness towards intensity artefacts compared to the standard SSD quadratic loss. From a computational standpoint, Eq. (4.2) fortunately admits variational quadratic upperbounds that serve as mathematically sound proxies for the exact loss and make it as convenient to use as the SSD. It is in fact handled as an iteratively reweighted (voxelwise) SSD when necessary. A similar procedure is described fully by *e.g.* [Archambeau 2007]. For the sake of simplicity and clarity of exposition, most of the derivations are therefore presented for the SSD loss.

Another limitation of SSD shared by all aforementioned variants is to assume that each voxel provides an independent value of intensity. This assumption does not hold in practice however [Simpson 2012]: the residual between the warped image $I \circ \Psi^{-1}$ and its counterpart J exhibits local spatial correlations, either intrinsic to the image acquisition and pre-processing (*e.g.* image pre-smoothing, image upsampling) or introduced as

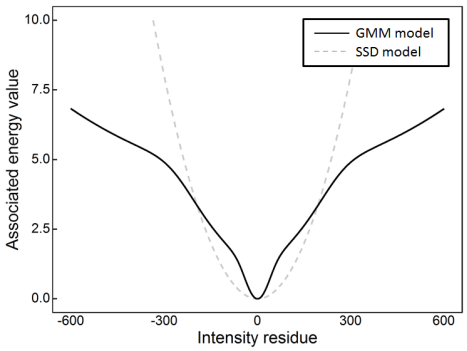


Figure 3.5: Energy profiles for the SSD (grey dashes) and GMM (black line). The penalty is plotted as a function of the difference of intensities in the registered images. The mixture-of-Gaussians model achieves robustness by incorporating concave inflexions that result in a soft threshold on the penalty incurred for large intensity residuals.

a consequence of registration misalignments. Ignoring local correlations in the noise pattern leads to an artificial increase in the number of independent observations and induces over-confidence in the data term. On the other hand, modeling precisely the noise structure would come at a significant computational cost. Instead, we follow [Simpson 2012] in artificially downweighting the SSD term by a factor α that captures redundancies in voxelwise observations, based on a virtual decimation procedure suggested by [Groves 2011].

A final fallacy in our generative model of data stems not from the way residuals are modeled, but from the later assumption that the intensity profile can be evaluated with infinite accuracy at any point in the moving image; whereas we actually rely on interpolation of a discrete scalar field. In regions where strong intensity gradients occur between adjacent voxels (*e.g.* in the case of neighbouring hypo- and hyper-intense regions), ignoring interpolation errors leads to an unreasonably high sub-voxel confidence in the optimally paired coordinates $C_i = \Psi^{-1}(\mathrm{c_i})$. Indeed the region in which the intensity value $I[\mathrm{C_i}]$ in the moving image coincides, up to the noise level, with the value $J[\mathrm{c_i}]$ in the fixed image may then shrink down in the gradient direction unreasonably below sub-voxel dimensions. To account for interpolation uncertainty, [Wachinger 2014] model interpolation as a noisy observation of the hidden, true intensity value and propose an approximate scheme to marginalize over latent variables. We opt instead for a simple scheme that couples efficiently with a Laplace approximation of the data likelihood, as explained in section 3.3.1.

3.2.2 Representation of displacements

In the context of non-rigid registration, an admissible space of transformations should be specified. In this work we restrict ourselves to a small deformation framework, $\Psi^{-1} = \mathrm{Id} + \mathrm{u}$, with a parameterized representation of the displacement field $\mathrm{u}\colon \mathrm{x} \in \mathbb{R}^d \mapsto \mathrm{u}(\mathrm{x}) \in \mathbb{R}^d$. We constrain the displacement field to be expressed over a dictionary $\{\phi_k\}_{k=1}^M$ of radial basis functions, specifically Gaussian kernels $\phi_k(\mathrm{x}) = K_{S_k}(\mathrm{x}_k, \mathrm{x})\,\mathbf{I}$ where $\mathbf{I}$ is the $d \times d$

identity matrix and

$$K_S(x,y) = \exp -\frac{1}{2}(x-y)^\intercal S^{-1}(x-y)\,. \tag{3.6}$$

In other words, the displacement field u is parameterized by a set of weights $\boldsymbol{w}_k \in \mathbb{R}^d$ associated to each basis ϕ_k:

$$\mathrm{u}_{\boldsymbol{w}}(\mathrm{x}) = \sum_{1\leq k\leq M} \phi_k(\mathrm{x})\boldsymbol{w}_k = \boldsymbol{\phi}(\mathrm{x})^\intercal \boldsymbol{w}\,. \tag{3.7}$$

$\boldsymbol{\phi}(\mathrm{x}) = \big(\phi_1(\mathrm{x})\cdots\phi_M(\mathrm{x})\big)^\intercal$ and $\boldsymbol{w}^\intercal = \big(w_1^\intercal\cdots w_M^\intercal\big)$ are respectively the concatenation, for $k = 1\cdots M$, of $\phi_k(\mathrm{x})$ and w_k. The basis centers x_k span a predefined regular grid of points, typically the whole range of voxel centers. The kernel width S_k is also allowed to vary and spans a user-predefined set of values $S_1, S_2, \cdots, S_q$. This allows for a redundant, multiscale representation of displacements. In other words we benefit both from a compact representation *via* larger kernels, and from the ability to capture finer local details *via* smaller kernels.

This dictionary of basis functions can be seen as a finite-dimensional approximation to the space $\mathcal{H}_S$ spanned by Gaussian kernels of a given width $S \leq \min_k\{S_k\}$. Such spaces have attractive properties and are related in the literature to spaces of *currents* [Gori 2013]. In particular, we derive in A.1 a family of *analytical*, computationally efficient probabilistic priors over $\mathcal{H}_S$.

3.2.3 Transformation Prior

In non-rigid registration, the displacement $\mathrm{u}_{\boldsymbol{w}}$ is insufficiently constrained by the data and some regularizing prior has to be imposed over its parameters. This prior distribution encapsulates our knowledge of the deformation and our modeling assumptions (see for instance [Sotiras 2013] for an exhaustive review of deformation priors). We will consider Gaussian priors of the form

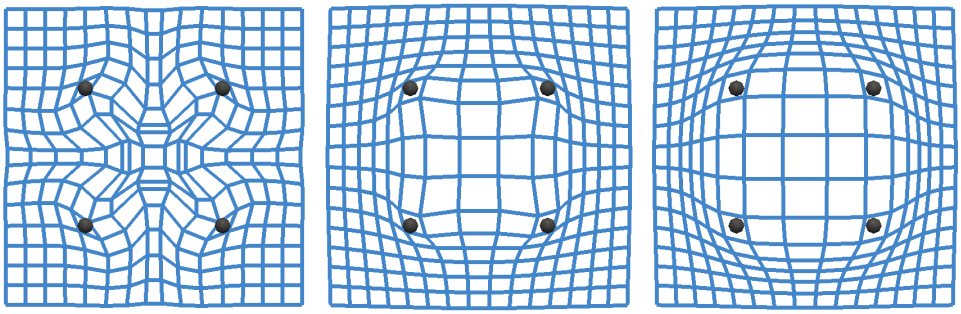
$$p(\boldsymbol{w}|\lambda,\{\mathbf{A}_k\}) \propto \exp -\frac{1}{2}\left\{\lambda\boldsymbol{w}^\intercal\mathbf{R}\boldsymbol{w} + \sum_{k=1}^{M}\boldsymbol{w}_k^\intercal\mathbf{A}_k\boldsymbol{w}_k\right\} \tag{3.8}$$

where λ and $\{\mathbf{A}_k\}_{k=1\cdots M}$ are so called hyperparameters. The motivation for such a prior is two-fold.

Regularity control. Gaussian priors in the form of Eq. (3.9) let us penalize physically implausible deformations. They have been commonly used in the literature starting with [Broit 1981], both because of their natural interpretability and soundness in mechanical terms, and their convenience from an algorithmic and computational standpoint.

$$q(\boldsymbol{w}|\lambda) \propto \exp -\frac{1}{2}\lambda\boldsymbol{w}^\intercal\mathbf{R}\boldsymbol{w} \tag{3.9}$$

The structure of the precision matrix $\mathbf{R}$ can be adjusted to penalize the magnitude of the first derivative [Gee 1998] or higher order derivatives [Rueckert 1999], effectively



Penalizing compressibility ($D = \text{div}$) | Elastic membrane energy ($D = \nabla$) | Bending (thin-plate) energy ($D = \nabla^2$)

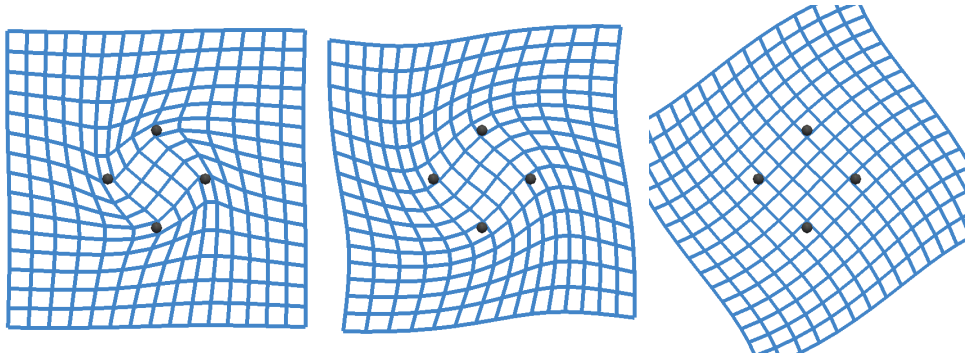
Figure 3.6: Impact of the regularization model. Displacements are parameterized by isotropic Gaussian kernels of set width $\sigma = 0.25$. From left to right, the regularizer varies. The data consists of 4 points regularly sampled on the unit circle, forming an axis aligned square, pulled twice as far away from the origin as they initially were. The warped grid obtained by regression is displayed along with the ground truth displacement.

encoding a wide range of priors [Ashburner 2007, Ashburner 2013]. We specifically consider the subset of quadratic forms that exploit the structure of the space of Gaussian kernels introduced in Section 3.2.2; namely priors of the type $\boldsymbol{w}^\intercal \mathbf{R} \boldsymbol{w} = \|D\mathrm{u}_{\boldsymbol{w}}\|^2_{\mathcal{H}}$, with $D\mathrm{u}$ any differential operator acting on u. Details are reported in A.1. By doing so, we obtain among others an analytical, computationally efficient implementation of a membrane energy with $D = \nabla$, a bending (thin plate) energy with $D = \nabla^2$, and a penalty favoring incompressible, divergence free, behaviours by setting $D = \text{div}$. Fig. 3.6 illustrates the respective impacts of such penalties.

Basis selection. The second factor in our prior, recalled in Eq. (3.10), induces the basis selection mechanism that we exploit in this chapter.

$$q(\boldsymbol{w}|\{\mathbf{A}_k\}) \propto \prod_{k=1}^{m} \exp -\frac{1}{2} \boldsymbol{w}_k^\intercal \mathbf{A}_k \boldsymbol{w}_k \tag{3.10}$$

The additional term $\boldsymbol{w}_k^\intercal \mathbf{A}_k \boldsymbol{w}_k$ for each basis ϕ_k lets us penalize independently the recourse to this basis to capture the displacement, by penalizing high magnitudes of its associated weight $\boldsymbol{w}_k$. The limit case of infinite $\boldsymbol{A}_k$ actually constrains $\boldsymbol{w}_k$ to be null and thus forbids the use of ϕ_k to represent the signal. In section 3.2.6 we propose a principled way to determine optimal values of the set $\{\mathbf{A}_k\}_{k=1\cdots M}$, from which most of them turn out to be infinite: we thus obtain a sparse representation of the displacement from our initial, over-complete dictionary.



Basis scale: $\sigma = 0.25$ | Basis scale: $\sigma = 1.0$ | Basis scale: $\sigma = 4.0$

Figure 3.7: Impact of the basis scale on the inferred transform. From left to right, the displacement field is parameterized by isotropic Gaussian kernels of increasing width. The data consists of 8 points regularly sampled on the unit circle. The underlying motion is a rotation of $\pi/4$ radian. Only four of these eight rotated samples are displayed for readibility. The scale of the bases used to represent the transform strongly affects the area of influence of the data points, as can be seen from the scale at which the regressed transform resembles a global rotation.

3.2.4 Sparse Coding & Registration

Sparsity-inducing priors have a two-fold motivation. The first benefit is in terms of algorithmic complexity. Unless resorting to low parametric models, the shear size of the parametrization makes direct optimization cumbersome without the recourse to sophisticated solvers. The computation of exact covariance matrices that are typically involved in probabilistic approaches also becomes unfeasible, while diagonal approximations used in their stead ignore significant interactions induced by the data and priors, and encoded by off-diagonal terms. Secondly, basis selection mechanisms adaptively constrain the admissible space of deformations, automatically tuning the degrees of freedom to the smallest sufficient set able to capture the observed displacement. Coupled with a multiscale set of basis functions, this yields a data-driven, automatic spatial refinement of the granularity of the displacement field that complements the otherwise scale-blind L_2 regularization. Adaptive, multiscale regularization was shown to yield state-of-art results *e.g.* in denoising natural scenes [Fanello 2014]. It is also critical for medical image registration. The spread of informative image structures lacks spatial homogeneity, yet those structures must correctly drive the whole registration. Moreover, the amount of noise and artefacts may vary across the image in hardly predictable patterns – and the degree of coarseness at which the displacement is modeled should be refined in consequence. Fig. 3.7 illustrates the key impact of scale when relying on limited observations, on a simple toy example where the underlying motion is a rotation.

L_1 priors have been widely and successfully used in all areas of sparse coding, including for registration [Shi 2013]. Other sparsity-inducing norms such as k-support norms and

variants [Argyriou 2012, Belilovsky 2015a], that improve over the performance of the L_1 norm w.r.t. the degree of sparsity in presence of strongly correlated explanatory variables, have recently been proposed. They were shown to be particularly attractive, including on tasks of functional MR imaging [Jenatton 2012, Belilovsky 2015b]. Here, we turn instead towards sparse Bayesian learning, with the prospect of joint estimation of model parameters and that of uncertainty quantification. For an extensive review of sparse methods, we refer the reader to the work of [Bach 2012], and to that of [Mohamed 2012] for a benchmark of L_1 and bayesian sparse learning methods. The prior of Eq. (3.10) was first introduced by [Tipping 2001] for regression and classification tasks with the so called Relevance Vector Machine. The authors demonstrated its relevance for sparse coding when used in conjunction with the framework of Automatic Relevance Determination [MacKay 1992]. [Bishop 2000] offer an alternative sparse Bayesian learning (SBL) view on the Relevance Vector Machine, where they opt for a Variational Bayes treatment. [Wipf 2008] further investigate links between the SBL and ARD frameworks and resulting schemes. Alternatively, Eq. (3.8) can be interpreted as a generalized spike-and-slab prior [Mitchell 1988] despite using a different parametrization, provided that each $\mathbf{A}_k$ is constrained to a binary state – either null or infinite.

3.2.5 Hyper-priors

In this work, we treat hyperparameters $\{\mathbf{A}, \beta, \lambda\}$ as frequentist parameters rather than latent random variables. For a fully Bayesian formulation, prior distributions would be imposed over model parameters. When conjugate priors are available this yields a very elegant model over which, for instance, Variational Bayes approximate inference may be used to derive closed form iterative updates. We refer the reader to the work of [Simpson 2012] for an application to image registration.

However, in absence of strong prior knowledge over the values of λ or β, broad uninformative priors are typically chosen. In the limit this effectively yields identical updates to the ones we derive with point estimates of the parameters while simplifying the exposition. Moreover the uniform prior on $\mathbf{A}_k$ has the appealing benefit of making our inference scheme invariant to rescaling of basis functions, as will be seen from Eq. (3.22), (3.23). This is highly desirable as we rely on a multiscale dictionary of basis functions (section 3.2.2) whose scaling factors may otherwise be hard to relate.

3.2.6 Model Inference

Our probabilistic model is summarized in a graphical manner in Fig. 3.2. The registration task now consists in inferring the displacement from the observed data and the prescribed graphical model. We describe the high-level approach in what follows. Section 3.3 presents the method from an algorithmic standpoint.

Hyperparameter inference. We first estimate the values of the model hyperparameters $\{\{\mathbf{A}_i\}, \lambda, \beta\}$ by maximizing the so called *marginal likelihood* of the data $p(D|\mathbf{A}, \lambda, \beta)$. This framework is known as that of type-II maximum likelihood in the statistical literature.

Note that in computing the marginal likelihood, we integrate over the parameters $\boldsymbol{w}$ (these parameters are *marginalized* out):

$$p(D|\mathbf{A},\lambda,\beta) = \int_{\boldsymbol{w}} p(D|\boldsymbol{w},\beta)p(\boldsymbol{w}|\mathbf{A},\lambda)d\boldsymbol{w} \tag{3.11}$$

where for convenience of notation we introduce the block diagonal matrix $\mathbf{A} = \mathrm{diag}(\mathbf{A}_i)$.

Posterior distribution of model parameters. Given the optimal parameters $\mathbf{A}^*, \beta^*, \lambda^* = \arg\max_{\mathbf{A},\beta,\lambda} p(D|\mathbf{A},\lambda,\beta)$, we can derive statistics of interest on the transformation ψ by first computing the *posterior* distribution of its parametric representation $\boldsymbol{w}$ conditioned on the observed data D. Bayes' rule asserts that:

$$p(\boldsymbol{w}|D,\mathbf{A}^*,\beta^*,\lambda^*) = \frac{p(D|\boldsymbol{w},\beta^*)p(\boldsymbol{w}|\mathbf{A}^*,\lambda^*)}{p(D|\mathbf{A}^*,\lambda^*,\beta^*)}\,. \tag{3.12}$$

In particular, the maximum of Eq. (3.12) minimizes Eq. (3.1). This bridges the gap with the classical framework in which registration is seen as the task of optimizing a functional. In Eq. (3.12) however, the hyperparameters assume *optimal* values $\mathbf{A}^*, \beta^*, \lambda^*$ defined w.r.t. the dataset of interest D. Moreover, the posterior distribution does not merely encode a point estimate of $\boldsymbol{w}$. Its higher-order moments relate to the variability in the inferred parameters. We consider a Gaussian approximation $\mathcal{N}(\boldsymbol{\mu},\Sigma)$ of Eq. (3.12) around its posterior mode to obtain an approximate second moment of the distribution as Σ.

Predictive distribution of displacements. Our model Eq. (4.3) implies a linear relationship between the displacement value $\mathrm{u}(\mathrm{x}) = \boldsymbol{\phi}(\mathrm{x})^\intercal \boldsymbol{w}$ at any $\mathrm{x} \in \mathbb{R}^d$ and the model parameters $\boldsymbol{w}$. Formally, the posterior distribution of $\mathrm{u}(\mathrm{x})$ is then given by Eq. (3.13), where $\boldsymbol{\theta} = \{\mathbf{A},\beta,\lambda\}$ denotes the set of hyperparameters and δ the Dirac distribution.

$$p(\mathrm{u}(\mathrm{x})|D,\boldsymbol{\theta}^*) = \int_{\boldsymbol{w}} \delta_{\boldsymbol{\phi}(\mathrm{x})^\intercal \boldsymbol{w}}[\mathrm{u}(\mathrm{x})]\, p(\boldsymbol{w}|D,\boldsymbol{\theta}^*)d\boldsymbol{w} \tag{3.13}$$

If the posterior distribution $p(\boldsymbol{w}|D,\boldsymbol{\theta}^*)$ of $\boldsymbol{w}$ is normally distributed $\mathcal{N}(\boldsymbol{\mu},\Sigma)$, $\mathrm{u}(\mathrm{x})$ is in turn Gaussian with mean $\boldsymbol{\phi}(\mathrm{x})^\intercal\boldsymbol{\mu}$ and covariance $\boldsymbol{\phi}(\mathrm{x})^\intercal\Sigma\boldsymbol{\phi}(\mathrm{x})$. More generally, $\mathrm{u}|D,\boldsymbol{\theta}^* \colon \mathrm{x} \mapsto \mathrm{u}(\mathrm{x})$ is a Gaussian process with mean $\bar{\mathrm{u}}(\,\cdot\,) = \boldsymbol{\phi}(\,\cdot\,)^\intercal\boldsymbol{\mu}$ and covariance

$$\mathrm{Cov}\,(\mathrm{u}(\mathrm{x}),\mathrm{u}(\mathrm{y})) = \boldsymbol{\phi}(\mathrm{x})^\intercal\Sigma\boldsymbol{\phi}(\mathrm{y})\,. \tag{3.14}$$

Full posterior vs. frequentist posterior. Strictly speaking we would in fact rather estimate the *full* posterior, where hyperparameters have been integrated out as per Eq. (3.15). When using uniform priors over hyperparameters, the posterior probability $p(\boldsymbol{\theta}|D)$ is proportional to the marginal likelihood $p(D|\boldsymbol{\theta})$.

$$p(\mathrm{u}(x)|D) = \int_{\boldsymbol{\theta}} p(\mathrm{u}(x)|D,\boldsymbol{\theta})p(\boldsymbol{\theta}|D)\,d\boldsymbol{\theta} \tag{3.15}$$

The full posterior generally cannot be computed without resorting to MCMC estimates. However if hyperparameters are well determined by the data, $p(\boldsymbol{\theta}|D) \approx \delta_{\boldsymbol{\theta}^*}$ becomes

highly peaked around its mode $\boldsymbol{\theta}^*$. The quality of this assumption is discussed at greater length by [Tipping 2001]. The full posterior of Eq. (3.15) is then approximately equal to the frequentist posterior of Eq. (3.13). This motivates our two-step approach of first looking for the maximum likelihood hyperpameters $\boldsymbol{\theta}^*$ before computing the frequentist posterior.

3.3 Inference schemes

As described in the previous section (3.2.6), our inference strategy is based on the maximization of the marginal likelihood as per Eq. (3.11). Unfortunately, the closed form evaluation of this type of integral is generally intractable, whereas its approximation via MCMC sampling schemes can often be prohibitively costly. [Tipping 2001] notes that when the likelihood and the prior are normally distributed, the data conveniently also follows a Gaussian distribution; furthermore the form of Eq. (3.11) becomes such that tractable maximization schemes can be derived with respect to the hyperparameters. We propose a Gaussian approximation of the likelihood term and extend the inference strategy of [Tipping 2003] to the broader class of priors described in 3.2.3.

3.3.1 Approximation of the Model Likelihood

The assumption of Gaussianity of the data given the model parameters $\boldsymbol{w}$ does not stand for registration purposes and several strategies can be considered to approximate the model likelihood. In earlier work [Le Folgoc 2014], a Taylor expansion of the likelihood around one of its local maxima was applied, resulting in a dense block matching step. Here instead Laplace's method is used around the current estimate of the mode $\boldsymbol{\mu}_{\text{MP}}$ of the posterior distribution, found by quasi Newton (BFGS) optimization on Eq. (3.12). In other words, as in conventional energy-based registration, we numerically solve for the minimizer $\boldsymbol{\mu}_{\text{MP}}$ of Eq. (3.16) using the current estimate of hyperparameters $\mathbf{A}, \lambda, \beta$:

$$\mathcal{E}(\boldsymbol{w}) = \mathcal{D}_\beta(I, J; \Phi\boldsymbol{w}) + \frac{1}{2}\boldsymbol{w}^\intercal(\mathbf{A} + \lambda\mathbf{R})\boldsymbol{w} \tag{3.16}$$

Most notably however, Eq. (3.16) in effect only involves the sparse subset of selected bases ϕ_k for which $\mathbf{A}_k < +\infty$. Proceeding with Laplace's approximation of the data likelihood and dropping the term involving the Hessian of the image[1], we arrive at Eq. (3.17):

$$\mathcal{D}(I, J; \boldsymbol{w}, \beta) \approx \frac{1}{2}\sum_{i=1}^{N} \left(\mathrm{t_i} - \boldsymbol{\phi}(\mathrm{c_i})\boldsymbol{w}\right)^\intercal \beta\mathbf{H}_i \left(\mathrm{t_i} - \boldsymbol{\phi}(\mathrm{c_i})\boldsymbol{w}\right) . \tag{3.17}$$

The virtual observations $\mathrm{t_i} \in \mathbb{R}^{\mathrm{d}}$ and the confidence tensors $\mathbf{H}_i \in \mathcal{M}_{d\times d}$ are given by Eq. (3.18), where $\mathrm{C_i} = \mathrm{c_i} + \mathrm{u_{MP}}(\mathrm{c_i})$ stands for the current (posterior) estimate of the pairing for $\mathrm{c_i}$.

$$\mathrm{t_i} = \mathrm{u_{MP}}(\mathrm{c_i}) - \frac{I(\mathrm{C_i}) - J(\mathrm{c_i})}{\|\nabla I(\mathrm{C_i})\|^2}\nabla I(\mathrm{C_i}) , \quad \mathbf{H}_i = \nabla I(\mathrm{C_i})\nabla I(\mathrm{C_i})^\intercal \tag{3.18}$$

[1] This outer product approximation is justified in *e.g.* [Bishop 2006].

These virtual pairings immediately relate to the optical flow: if we dropped the confidence tensors $\mathbf{H}_i$, the above would yield an approximation of Eq. (3.1),(3.12) much in the spirit of the *demons* algorithm [Thirion 1998, Cachier 2004]. The tensors $\mathbf{H}_i$ vary sharply across the image however, *e.g.* as edges or boundaries are crossed. They assign anisotropic, spatially varying confidence in voxelwise pairings and account for how informative and structured the image is at the point of interest. The local approximation of Eq. (3.17) transforms an image-based criterion into a landmark-based one, and the proximity to formulations in the related literature [Rohr 2003] is indeed striking in this form.

The confidence $\beta\mathbf{H}_i$ in the virtual voxelwise pairing $t_i + c_i$ can grow arbitrarily high for arbitrarily high intensity gradients. These expressions result from linearizing the intensity profile I around the current pairing C_i, and are blind to interpolation uncertainty in evaluating $I(C_i)$ and $\nabla I(\mathrm{C_i})$. To address this shortcoming, we propose to replace $\beta\mathbf{H}_i$ by

$$\left((\beta\mathbf{H}_i)^{-1} + \mathbf{D}_{\text{int}}\right)^{-1} = \frac{1}{1 + \text{tr}[\beta\mathbf{H}_i\mathbf{D}_{\text{int}}]} \cdot \beta\mathbf{H}_i \tag{3.19}$$

which implements a soft upper threshold on the precision, as a heuristic for interpolation uncertainty. $\mathbf{D}_{\text{int}}$ acts as a minimum covariance: it is a diagonal matrix set to the square of –say– half the voxel spacing to prevent unreasonable subvoxel confidence.

3.3.2 Form of the Marginal Likelihood

We now assume that the data $\boldsymbol{t}$ is generated by corrupting the true signal $\boldsymbol{u} = \Phi\boldsymbol{w}$ with Gaussian noise $\boldsymbol{e} \sim \mathcal{N}(0, \beta\mathbf{H})$:

$$\boldsymbol{t} = \Phi\boldsymbol{w} + \boldsymbol{e} \tag{3.20}$$

Note that in block form, Eq. (3.17) yields an approximate likelihood model in the form of Eq. (3.20). Given the hyperparameters of the model, the posterior distribution of $\boldsymbol{w}$ conditioned on the data – obtained by combining the likelihood and prior with Bayes' rule according to Eq. (3.12) – is Gaussian $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$ with

$$\boldsymbol{\mu} = \Sigma\Phi^\intercal\beta\mathbf{H}\boldsymbol{t} \quad \Sigma = (\Phi^\intercal\beta\mathbf{H}\Phi + \lambda\mathbf{R} + \mathbf{A})^{-1} \tag{3.21}$$

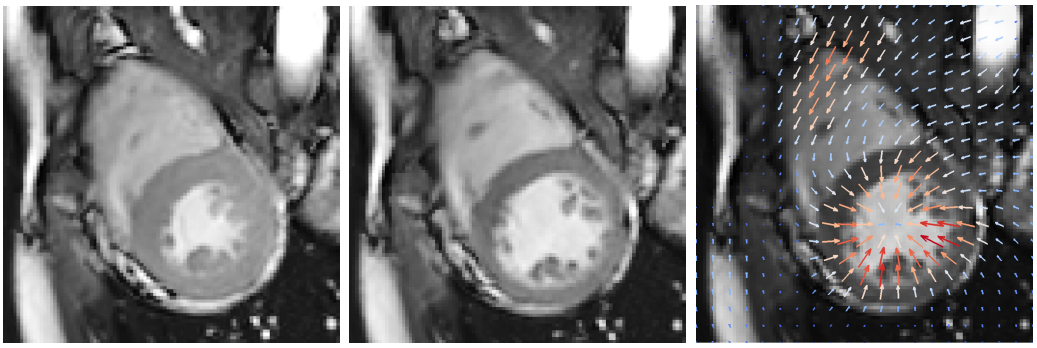
The integrand in Eq. (3.11) involves the product of two Gaussian factors, and by integrating out the weights $\boldsymbol{w}$ we obtain the *evidence* or *marginal likelihood* for the hyperparameters:

$$p(\boldsymbol{t}|\mathbf{A}, \lambda, \beta) = |\, 2\pi\,\mathbf{C}\,|^{-1/2} \cdot \exp\left\{-\frac{1}{2}\boldsymbol{t}^\intercal\mathbf{C}^{-1}\boldsymbol{t}\right\} \tag{3.22}$$

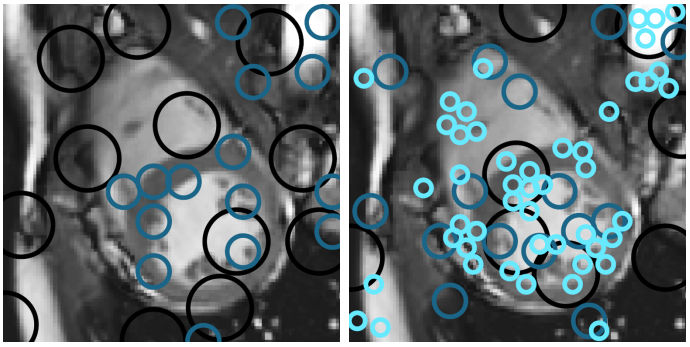
where by identification $\mathbf{C}^{-1} = \beta\mathbf{H} - (\beta\mathbf{H})\Phi\Sigma\Phi^\intercal(\beta\mathbf{H})$. In other words, the distribution of the data $\boldsymbol{t}$ conditioned on the hyperparameters $\{\mathbf{A}, \lambda, \beta\}$ is Gaussian $\mathcal{N}(0, \mathbf{C})$. Furthermore, it follows from the Woodbury matrix identity that

$$\mathbf{C} = (\beta\mathbf{H})^{-1} + \Phi(\mathbf{A} + \lambda\mathbf{R})^{-1}\Phi^\intercal\,. \tag{3.23}$$

Interestingly, the process of setting the hyperparameters can then be seen as fitting a covariance model $\mathbf{C}$ to the data $\boldsymbol{t}$ via a maximum likelihood approach. Note also that the two factors in Eq. (3.22) have antagonistic effects: while the left hand term penalizes



a. Moving image b. Fixed image c. Output displacement field



d. Adaptive parametrization (In the initial steps) e. Adaptive parametrization (At the end)

Figure 3.8: Basis selection mechanism displayed on an example 2D registration between slices of cardiac MR images (cf. sec 3.4.1), respectively at ES (a) and ED (b). (d,e) Bases selected in the initial steps of the algorithm vs. at the end. The locations and scales of the Gaussian RBFs are indicated by circles (isocontour at 1 std). (c) Inverse displacement field output by the algorithm (scale factor: 2), smoothly varying across the whole image.

covariance matrices $\mathbf{C}$ that waste mass (via $|\mathbf{C}|$), the right hand term gives incentive to spend mass to better explain the data $\boldsymbol{t}$. This compromise leads to sparsity, as revealed by a careful look at the form of $\mathbf{C}$ in Eq. (3.23). Indeed, regardless of the value of $\mathbf{A}$, part of the data is explained *for free* by the contribution $(\beta\mathbf{H})^{-1}$ of the noise to $\mathbf{C}$; thus only a few degrees of freedom need be active ($\mathbf{A}_k < \infty$) to fully explain the data.

3.3.3 Hyperparameter inference

Following the strategy discussed in 3.2.6, we wish to maximize the marginal likelihood (Eq. (3.22)), or equivalently its logarithm, with respect to the model hyperparameters. We consider schemes that monotonically converge towards a local maximum by *iteratively* maximizing, or merely increasing, the evidence w.r.t. either one of the hyperparameters $\mathbf{A}$, β or λ.

Maximization w.r.t. $\mathbf{A}$ | a hill-climbing scheme. The procedure relies on a sequence of

additions and deletions of candidate basis functions starting from an empty set $\mathcal{S} = \emptyset$ of active bases (all $\mathbf{A}_k$'s set to ∞). This notably avoids the $\mathcal{O}(M^3)$ cost that would stem from the matrix inversion in Eq. (3.21), if all M bases were included in the active set $\mathcal{S}$ at the start (all $\mathbf{A}_k < \infty$). At each iteration, we take a single action among the addition of a previously inactive basis ($k \notin \mathcal{S}$), or the update or deletion of an active one ($k \in \mathcal{S}$), in a principled way. Specifically, we implement the action that leads to the largest gain in evidence. This is possible because, when all other hyperparameters $\mathbf{A}_{-k}$, λ, β are fixed, the contribution $l(\mathbf{A}_k)$ of a given basis to (the logarithm of) the evidence for any value of its associated hyperparameter $\mathbf{A}_k$ can be singled out in the form of Eq. (3.24). Details about the statistics involved $\boldsymbol{\kappa}_k$, $\boldsymbol{s}_k$ and $\boldsymbol{q}_k$ and the maximization of $l(\mathbf{A}_k)$ are left to A.2.

$$\begin{aligned} l(\mathbf{A}_k) = \log|\mathbf{A}_k + \boldsymbol{\kappa}_k| - \log|\mathbf{A}_k + \boldsymbol{\kappa}_k + \boldsymbol{s}_k| \\ + \boldsymbol{q}_k^\intercal \left\{\mathbf{A}_k + \boldsymbol{\kappa}_k + \boldsymbol{s}_k\right\}^{-1} \boldsymbol{q}_k \end{aligned} \tag{3.24}$$

Naturally, as bases are added or removed from the active set $\mathcal{S}$ via updates of their associated hyperparameter, the potential contribution $\max_{\mathbf{A}_k} l(\mathbf{A}_k)$ of other bases is subject to change. In practice, the statistics $\boldsymbol{\kappa}_k$, $\boldsymbol{s}_k$ and $\boldsymbol{q}_k$ indeed depend on $\mathbf{A}_{-k}$, λ, β via the moments $\boldsymbol{\mu}$ and Σ of the posterior distribution (Eq. (3.21)). Therefore, after updating a given $\mathbf{A}_i$, we exploit rank-one matrix identities to update these moments with a complexity of at most $\mathcal{O}(|\mathcal{S}|^2)$ (A.4). We then update all the $\boldsymbol{\kappa}_k$, $\boldsymbol{s}_k$ and $\boldsymbol{q}_k$ with a complexity of $\mathcal{O}(|\mathcal{S}|)$ per basis. Therefore the worst-case complexity of the updates is of $\mathcal{O}(|\mathcal{S}|^2 + M|\mathcal{S}|)$, as opposed to $\mathcal{O}(M^3)$ for EM updates. The decrease in complexity is rather extreme, as typical respective orders of magnitude of $|\mathcal{S}|$ and M for registration would be 10^2 and 10^6.

To gain a better intuition on how this scheme proceeds, let us look back at the quantities involved in Eq. (3.24). Firstly, $\boldsymbol{q}_k$ captures the relevance of ϕ_k towards providing a better explanation for the data. It is related to the projection of the residual on the basis ϕ_k. When L_2 regularization is used ($\lambda > 0$), $\boldsymbol{q}_k$ also involves the projection of the residual on nearby active bases ϕ_j ($j \in \mathcal{S}$). Secondly $\boldsymbol{s}_k$ captures the amount of overlap between the basis ϕ_k under consideration and those already in the active set $\mathcal{S}$, therefore accounting for competition between strongly correlated bases. $\boldsymbol{\kappa}_k$ arises from the added regularization $\exp -\frac{1}{2}\lambda \boldsymbol{w}^\intercal \mathbf{R} \boldsymbol{w}$, which was absent in the regular RVM ($\lambda = 0$). The sum $\mathbf{A}_k + \boldsymbol{\kappa}_k$ entering Eq. (3.24) highlights the competition between the shrinkage (sparsity-inducing) mechanism due to $\mathbf{A}_k$ and the L_2-norm energy regularization that is accounted for by the quantity $\boldsymbol{\kappa}_k$.

Maximization w.r.t λ and β | EM updates. We estimate the hyperparameters λ, β using Expectation Maximization updates. Instead of directly maximizing our target criterion, the marginal-likelihood $p(\boldsymbol{t}|\mathbf{A}, \lambda, \beta)$, EM proceeds by maximizing a surrogate quantity, the expected complete log-likelihood:

$$\arg\max_{\lambda,\beta} \mathbb{E}_{p(\boldsymbol{w}|D,\mathbf{A}_*,\beta_*,\lambda_*)} \left[\log p(\boldsymbol{t}, \boldsymbol{w}|\mathbf{A}_*, \lambda, \beta)\right] . \tag{3.25}$$

The expectation is taken with respect to the current estimate of the posterior distribution $p(\boldsymbol{w}|D, \mathbf{A}_*, \beta_*, \lambda_*)$ of $\boldsymbol{w}$. β_*, λ_* are quantities fixed at their current values, as opposed

to variables to optimize. In practice, we iterate between reestimation of the noise level β and of the regularization trade-off λ, fixing the other parameter in turn. The EM scheme is appealing as it guarantees an increase in marginal likelihood, despite maximizing a surrogate. Furthermore, when the distributions involved are Gaussian, the expected complete log-likelihood (3.25) can be expressed directly in terms of the mean and covariance matrix of the Gaussian distributions. For the regularizing trade-off λ, this leads to maximizing the smooth, convex function:

$$\lambda^* = \arg\max_{\lambda \geq 0} -\frac{\lambda}{2}\mathrm{tr}(\Sigma \mathbf{R}) + \frac{1}{2}\log|\mathbf{A} + \lambda \mathbf{R}| - \frac{\lambda}{2}\boldsymbol{\mu}^\intercal \mathbf{R}\boldsymbol{\mu}\,. \tag{3.26}$$

The solution can be found numerically in inexpensive $\mathcal{O}(|\mathcal{S}|)$ iterations after a single singular value decomposition in $\mathcal{O}(|\mathcal{S}|^3)$. Alternatively, if we restrict $\mathbf{A}_k$ to be either null or infinite as suggested in section 3.2.4, we obtain the simpler analytical update:

$$|\mathcal{S}| \cdot \lambda^{*-1} = \boldsymbol{\mu}^\intercal \mathbf{R}\boldsymbol{\mu} + \mathrm{tr}(\Sigma \mathbf{R})\,. \tag{3.27}$$

Leaving optimization details aside, we can instead gain some insight into the EM update by examining the solution of the constrained optimization problem Eq. (3.26). Setting aside the case where the constraint is active ($\lambda^* = 0$), the solution λ^* cancels the gradient of the function to maximize. From a careful analysis of the terms involved in the gradient, the following identity holds at λ^*:

$$\mathbb{E}_{p(\boldsymbol{w}|\mathbf{A}_*,\lambda^*)}\left[\boldsymbol{w}^\intercal \mathbf{R}\boldsymbol{w}\right] = \mathbb{E}_{p(\boldsymbol{w}|D,\mathbf{A}_*,\beta_*,\lambda_*)}\left[\boldsymbol{w}^\intercal \mathbf{R}\boldsymbol{w}\right] \tag{3.28}$$

λ^* is chosen such that a sample drawn from the updated prior has the same energy as a sample drawn from the inferred posterior, on average. In other words, the EM update modifies the strength λ of the prior to best reflect what has been inferred from the data. The noise level β is updated according to Eq. (3.29). It accounts for a bias between the estimated and true displacements, and for the variance in the estimated displacement.

$$N \cdot \beta^{*-1} = \|\boldsymbol{t} - \Phi\boldsymbol{\mu}\|^2_{\mathbf{H}} + \mathrm{tr}(\Sigma \Phi^\intercal \mathbf{H} \Phi)\,. \tag{3.29}$$

In other words, β^{*-1} is set to the expected error (averaged over pixels) if sampling from the current estimate of the posterior.

3.3.4 Algorithmic overview

Algorithm 1 summarizes our pipeline for registration. The algorithm mostly works by iterative refinements of the subset of active basis functions $\{\phi_j, j \in \mathcal{S}\}$ with fast updates, based on an *approximate* likelihood model around the current mode of the displacement parameters $\boldsymbol{\mu}_{\mathrm{MP}}$. Every now and then, the noise and regularization parameters λ, β are updated, which makes it necessary to recompute every statistic in full. Before doing so, we re-estimate the posterior mode $\boldsymbol{\mu}_{\mathrm{MP}}$ from the *true* likelihood model and the current (reduced) set $\mathcal{S}$ of active bases and recompute the likelihood approximation around this mode. To further accelerate the pipeline, the scheme can be coupled with a multiresolution pyramidal scheme, starting with downsampled (smoothed) versions of the image I and J and progressively moving through the pyramid of images to the images I and J at full resolution.

Algorithm 1: Sparse Bayesian registration algorithm

Initialize $\mathbf{A}_k = \infty$ for all k ($\mathcal{S} = \emptyset$)
Initialize $\boldsymbol{\mu} = 0$ and $\boldsymbol{\Sigma} = 0$
repeat
  Update β according to Eq. (3.29).
  Find the posterior mode $\boldsymbol{\mu}_{\text{MP}}$ of Eq. (3.12) for the current values of $\mathbf{A}, \lambda, \beta$ by quasi-Newton (BFGS) search.
  Compute a quadratic approximation of the likelihood around $\boldsymbol{\mu}_{\text{MP}}$: Eq. (3.17),(3.18)
  Recompute $\boldsymbol{\mu}$ ($= \boldsymbol{\mu}_{\text{MP}}$) and $\boldsymbol{\Sigma}$ in full from Eq. (3.21)
  Recompute $\boldsymbol{q}_k$, $\boldsymbol{s}_k$ and $\boldsymbol{\kappa}_k$ in full from Eq. (A.31), (A.32) and Eq. (A.33), (A.34), for all k
  for p iterations **do**
    $\forall k \in \mathcal{S}$ (resp. $k \notin \mathcal{S}$), compute the gain $\max_{A_k} \Delta l(\mathrm{A}_k)$ in marginal likelihood obtained by updating or deleting k from $\mathcal{S}$ (resp. adding k to $\mathcal{S}$), using Eq. (3.24) and A.2.
    Select the most favorable action i s.t. $\max_{A_i} \Delta l(\mathrm{A}_i) \geq \max_{A_k} \Delta l(\mathrm{A}_k)$ for all k.
    Set $A_i^* = \arg\max_{A_i} l(\mathrm{A}_i)$ and update $\mathcal{S}$.
    Update $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ via rank-one matrix identities from Eq. (A.22), (A.25), (A.24).
    Update $\boldsymbol{q}_k$, $\boldsymbol{s}_k$ and $\boldsymbol{\kappa}_k$ for all k using rank-one updates *e.g.* Eq. (A.35), (A.36) and Eq. (A.33), (A.34).
  Update λ according to Eq. (3.26).
until no action leads to a significant increase in marginal likelihood.

3.4 Experiments & Results

We experiment with the proposed sparse Bayesian framework on tasks of cardiac motion tracking. The goal in cardiac motion tracking is to accurately recover the hidden motion of the cardiac muscle over the course of the cardiac cycle from a time series of 3D images. The first experimental setup aims at clarifying and analyzing the empirical behaviour of the proposed algorithm and focuses on a simple example of 2D pairwise registration. All other sections involve full $3D + t$ motion tracking on various imaging modalities, namely cine SSFP sequences, tagged sequences and a synthetic ultrasound dataset. Fig. 3.9 displays example 2D slices from frames of each modality.

The experimental setup is identical across all modalities, and indeed we aim to demonstrate that the proposed framework is flexible enough to adapt seamslessly to the peculiarities of each dataset. For all 3D experiments, the multiscale parametrization of the displacement field consists of isotropic Gaussian kernels at two scales, of respective variance $S_1 = 20^2$ mm^2 and $S_2 = 10^2$ mm^2, plus an anisotropic Gaussian kernel of variance 10^2 mm^2 in the short axis plane and 20^2 mm^2 along the long axis. Indeed since our framework imposes no restriction on the parametrization of the displacement field, a natural way to put this advantage to use is to introduce anisotropic bases of potential anatomical relevance. All scales are jointly optimized upon. As explained in section 3.3.4, the multiscale

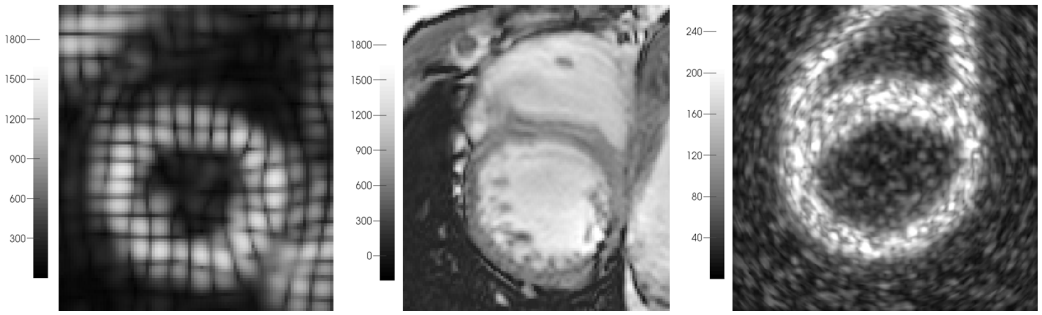


Figure 3.9: Example slices for the cardiac imaging modalities that we experiment on. (Left) 3D tagged MR image (Middle) 3D cine SSFP MR image (Right) 3D echocardiographic image. In all experiments the same flexible model of registration is applied successfully, regardless of the artefacts and patterns peculiar to each modality.

representation of the displacement field is coupled with a classic multiresolution, pyramidal scheme on the images of interest themselves – they are downsampled by a factor 2 (and smoothed) at three different resolutions. In other words the lowest resolution level is subsampled by a factor of 4 compared to the original image. Finally, we use a bending (thin-plate) energy as a regularizer (Section 3.2.3) and we constrain the hyperpriors $\mathbf{A}_k$ to a binary state (sections 3.2.4 and 3.3.3) to prevent competition between two types of regularization: the regularity induced by $\lambda\mathbf{R}$ on the derivatives of the displacement and that induced by $\mathbf{A}_k$ on the magnitude of specific weights w_k. Note also that our registration scheme does not make use of pre-segmentations of regions of interest, which could be challenging or otherwise impractical to obtain.

3.4.1 Self-tuning registration algorithm: an analysis

As explained in the previous sections, the parameters introduced in the model of registration are inferred jointly during the course of the algorithm. We now leverage the example 2D registration of Fig. 3.3 3.8 to provide some insight into our schemes and analyze the convergence of key parameters throughout iterations.

Basis selection & regularization. From an algorithmic standpoint, the algorithm proceeds by iteratively adding dictionary bases into – or deleting them from – the active parametrization of the displacement field. Every few iterations, the noise model and regularization level λ are re-adjusted based on the current estimate of the displacement and its uncertainty. Fig. 3.10 demonstrates how basis selection mechanisms empirically combine with parameter re-estimation throughout iterations to provide a seamless convergence towards a reasonable local minimum.

The regularization level λ is initialized thanks to a heuristic that provides a "very large" value for the registration at hand. Initially this effectively prevents the addition of finer dictionary bases in the model, whose impact on the signal regularity is too high at this stage. Instead coarse bases are favoured, which capture the global trends in the observed displace-

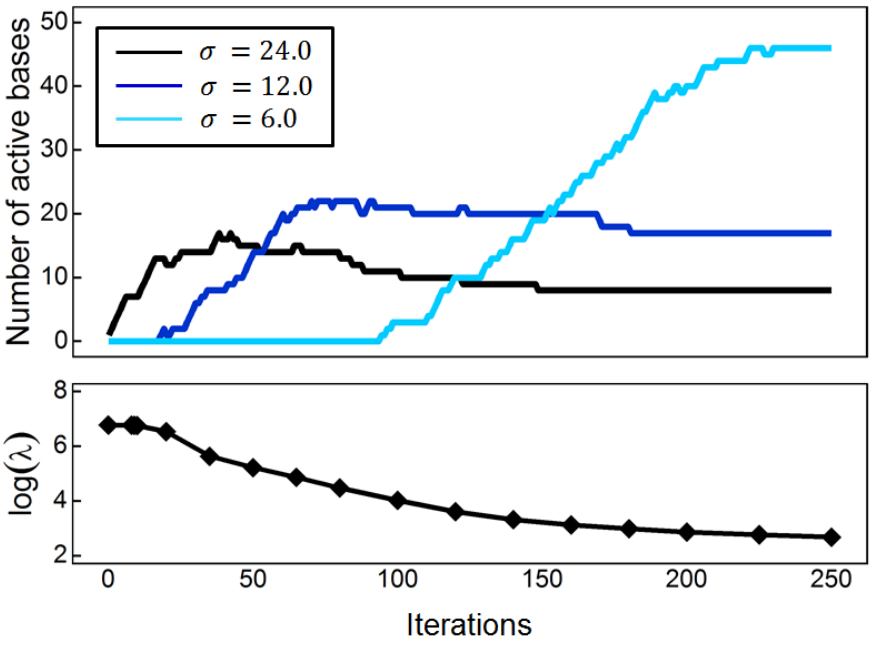


Figure 3.10: Basis selection mechanism and its coupling with the jointly estimated regularization level, across iterations. (Top) Addition, update or deletion of dictionary bases in the active parametrization of the displacement field across iterations. Three distinct scales are used in the representation of displacements (1 curve per scale). (Bottom) Regularization parameter λ, updated every few iterations, plotted against the number of iterations run since the beginning of the registration.

ment. The regularization level is consequently refined to reflect the actual regularity in this inferred displacement. As λ decreases towards a more sensible value, finer bases get incorporated in the active set to capture finer local details of the visible motion, or to ensure that these finer details of the inferred motion blend smoothly with the rest of the displacement field. In case of significant overlap between a subset of fine bases and a coarse basis, the basis at the coarsest scale may automatically be removed – as it no longer contributes towards a better explanation of the data. Towards the last iterations, algorithmic convergence has roughly been reached. This is evidenced by the plateau value of λ and by the fact that most actions consist in small updates of active bases (specifically, their orientation) rather than in the addition or deletion of dictionary bases.

Fig. 3.8 further illustrates this mechanism of basis selection, with the location of bases in the active parametrization being depicted at two points in time – in the initial steps of the algorithm and at the end.

Noise model estimation. The noise model is jointly estimated over the course of the algorithm. In all experiments it displayed a fast convergence towards its final inferred distribution. This is examplified by Fig. 3.11, where the probability density functions corresponding to the Gaussian mixture inferred at an early iteration and at the final

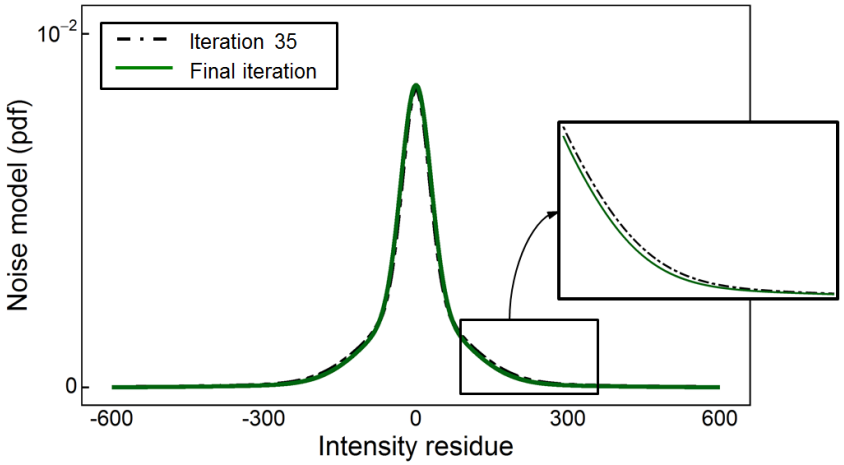


Figure 3.11: Inferred noise model, *i.e.* inferred expected distribution of intensity residuals between the fixed and warped images. The learned distribution is comparatively shown at an early stage in the registration (black dashed line) and at convergence (green solid line). The curves are nearly indistinguishable, hinting towards fast convergence in this example. The zoom on the distribution tails still evidences a slightly lower noise level at the end of registration, as intuitively expected.

iteration are hardly distinguishable. This partly reflects the fact that our registration experiments involve small displacements, with the coarse patterns in the underlying motion being captured early on. The Gaussian mixture is also quick to adapt to changes in the distribution of intensity residuals that arise from our multiresolution pyramidal scheme, when hopping from a smoothed downsampled image to the next level in the pyramid of images, as seen from Fig 3.12. The jumps from a coarser image resolution to a finer image resolution occur at iteration 10 and iteration 20.

Robustness w.r.t. initialization. Beyond the seamless convergence of model parameters over the course of the algorithm, one may also wonder whether their estimated final value displays consistency across a range of possible initializations. Fig. 3.13 provides evidence towards the empirical robustness of the estimated level of regularity λ w.r.t. its initial value. The final value of λ typically varies by significantly less than an order of magnitude when initialized from a range of values covering several orders of magnitude.

3.4.2 Synthetic 3D Ultrasound Cardiac Dataset

We first demonstrate our approach on synthetic sequences of 3D ultrasound data provided as part of a registration challenge organized for the 2012 MICCAI workshop on Statistical Atlases and Computational Models of the Heart (STACOM). Details on the challenge methodology can be found in [De Craene 2013] along with participant results. These synthetic images count approximately 10 million voxels each, at an extremely fine isotropic

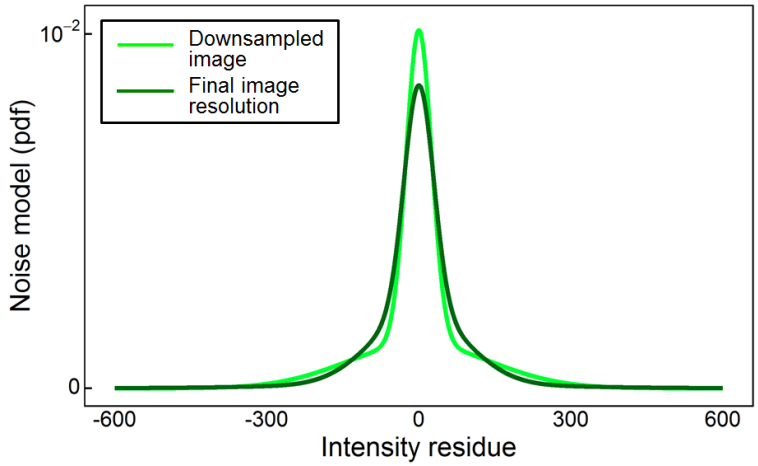


Figure 3.12: Inferred noise model at the beginning of registration and at the end. The registration scheme relies on a multiresolution representation of registered images, jumping from a coarse resolution to a finer resolution at predetermined iterations. The inferred distribution of intensity residuals is shown for the lowest resolution image (light green) and finest resolution image (dark green). As expected the noise model learned on downsampled, smoothed images has a higher probability of low noise (higher peak around 0) but also of high noise (higher tails) due to the increased misalignment at the beginning of registration.

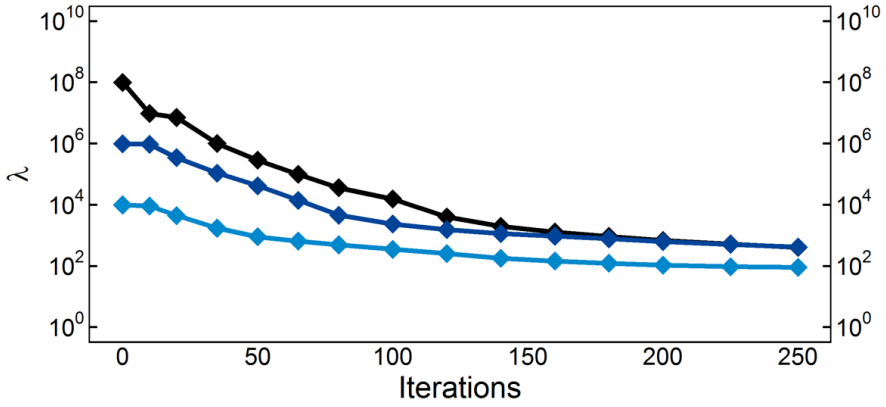
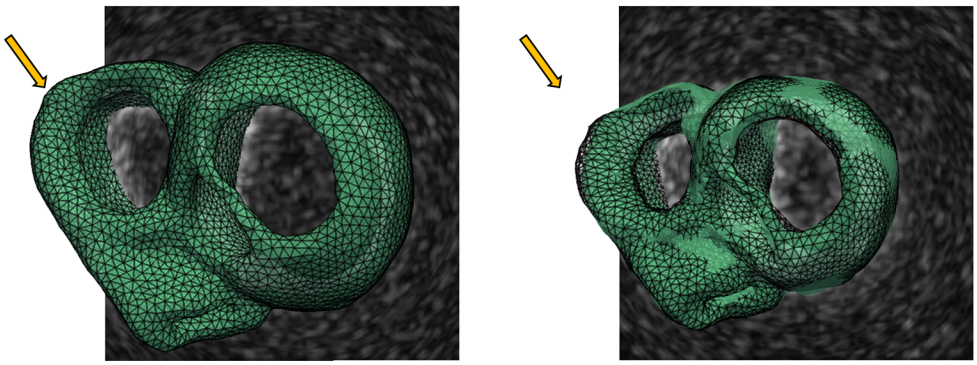


Figure 3.13: Robustness of the inferred regularity level w.r.t. its initial estimation. The 2D registration is run 3 times, and initialized each time with a differing level of regularity λ (respectively 10^4, 10^6, 10^8). Each curve shows the evolution of λ over the course of the associated run. While the initial value of λ spans 4 orders of magnitude, its final estimate varies by at most a factor of 4 across runs.



end-diastole frame end-systole frame

Figure 3.14: Ground truth mesh (green transparent surface) vs. reference mesh transported *via* registration (overlayed black wireframe). The extrapolated motion out of the field of view (where the arrow points) remains close to the ground truth. The maximum error does not exceed 4mm. Best seen by zooming in.

resolution of 0.33mm. Without further optimization of our code w.r.t. RAM management we had to downsample them by a factor of 2 to process them. We thus worked at a resolution of 0.66mm at the finest level.

The appeal of this benchmark is to offer a dense ground truth in terms of motion and strain inside the cardiac muscle, as displacements in the myocardium are directly prescribed from the output of an electromechanical model of the heart as part of the workflow of image generation. For each sequence of images, the ground truth consists of a sequence of meshes of the left and right ventricles deformed over the cardiac cycle. The data extracted from such ground truth meshes can be compared to that obtained by deforming the mesh at a reference time point (namely, end diastole) throughout the cardiac cycle with the transformation output by the proposed registration approach. Prior to any quantitative assessment, we first comment that the visual and qualitative behaviour of the proposed approach was found to be satisfactory, even in terms of extrapolation – as indeed the inferred motion remained consistent in areas of the right ventricle that fell outside of the field of view (Fig. 3.14). This indicates an effective regularization mechanism, despite being automatically tuned.

We evaluate the accuracy of our approach on a first subset of sequences that image the same motion at various Signal-to-Noise Ratios (SNRs). Because the proposed approach infers a consistent motion both inside and outside of the field of view, we find natural to assess its accuracy from statistics based on the whole mesh. This slightly departs from the methodology of [De Craene 2013] where part of the left ventricle only is considered. Fig. 3.15 reports the median point-to-point error in the inferred displacement for each time frame, where the median statistics is computed from every node in the mesh. At the best SNR, the highest error is observed around end systole with a median of 0.83mm, although the spread of error values becomes wider in the last frames. This falls in the same range as that reported for challenge participants by [De Craene 2013] – although slightly higher

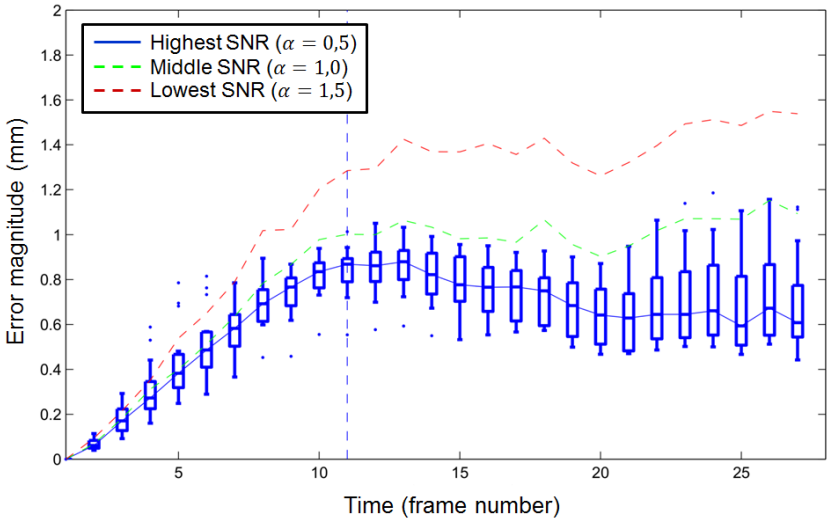


Figure 3.15: Accuracy benchmark on the 3D US STACOM 2012 normal dataset, reporting the median tracking error over time for varying SNRs (blue, green, red curves). For the reference SNR (in blue), overlayed quartiles (boxplots) picture the dispersion of errors.

than the most accurate methodology. Of course part of the error is likely to be attributable to the use of downsampled, smoothed images with a resolution of 0.66mm as opposed to 0.33mm. Besides as the signal to noise ratio degrades, we observe as expected a global trend of increased error magnitude. As seen from Fig. 3.16, the increased SNR impacts the noise model (with a higher prevalence of small intensity residuals at high SNR) learned by the proposed approach, which in turn becomes more conservative in its estimates of displacements.

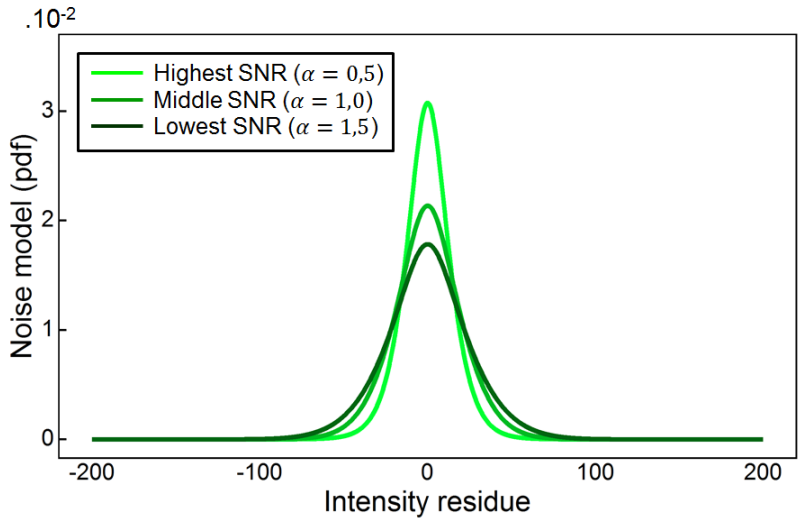


Figure 3.16: Evolution of the inferred noise model for increasing Signal-to-Noise Ratios.

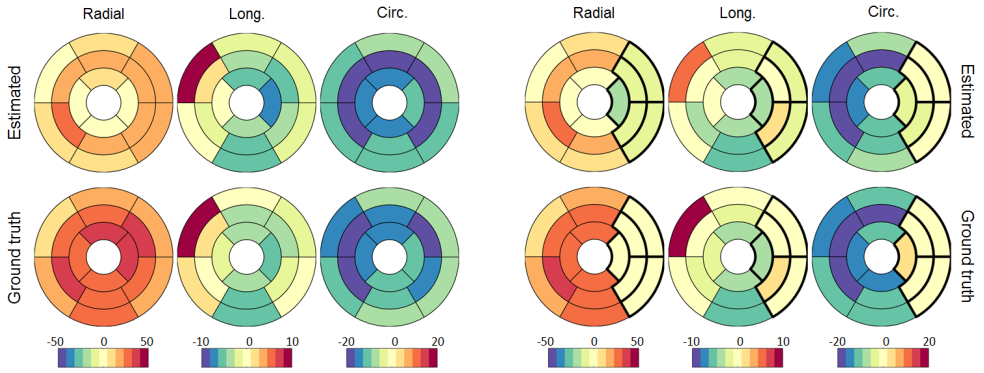


Figure 3.17: Bull's eye plots of the radial, longitudinal and circumferential strain components at end-systole, averaged over AHA segments: estimated (top) and ground truth (bottom). A healthy case (left) and an ischemic case (right) are reported.

Fig. 3.17a reports (Green Lagrangian) strain measures at end systole averaged over AHA segments. This provides indirect evidence of the relevance of the automatically tuned regularity level λ and of the displacement parametrization. Ground truth values of strain obtained from the corresponding ground truth mesh are compared to those estimated from the output of registration. Variations in the strain across segments are generally well captured, even more so for its longitudinal and circumferential components. Similarly to most methodologies however, the radial strain – which captures the thickening of the muscle during the contraction – appears to be globally somewhat underestimated in the left ventricle. Such bias in the radial strain might result from a slight bias in the estimated placement of the endo- and/or epicardium. This might indicate a slightly conservative estimate of displacements due to a coarse parametrization or over-regularized transformation. The following table provides statistics on the number of bases of each scale used for the parametrization of the displacement field, for the normal case at highest SNR. The number of active bases on these sequences is typically smaller than that used in our experiments on cine and tagged data, with a lesser reliance on fine-scale bases. It may evidence increased conservatism in the estimated displacements, as well as indicate greater regularity of the synthetic ground truth motion.

Basis type	Median # (Q1 – Q3)
$\sigma = 20$mm	17.5 (14.25 – 19)
anisotropic σ	15 (11.25 – 20.5)
$\sigma = 10$mm	34 (31.25 – 38)
Total	**64.5 (60 – 71)**

Table 3.1: Number of bases at each scale in the active parametrization of the displacement field (pooled over all frames in the sequence). Median, first and third quartiles are reported.

The benchmark also provides datasets that aim at reproducing pathological cardiac function, including a case where certain AHA segments become quasi akinetic due to ischemy. Fig. 3.17b summarizes estimated regional strains for this case, with qualitative retrieval of the ischemic segments (bolded contours), as emphasized by the comparison with the normal case. The accuracy on the ischemic case is similar to that of the normal case at identical SNR, with a median error at end systole of 0.80mm.

3.4.3 STACOM 2011 tagged MRI benchmark

In 2011 the MICCAI workshop on Statistical Atlases and Computational Models of the Heart (STACOM) proposed a cardiac motion analysis challenge. The challenge datasets, aimed at evaluating the accuracy of motion tracking algorithms, are openly hosted by the Cardiac Atlas Project. The data includes a set of 15 sequences of 3D tagged MR images. Fig. 3.9 (Left) shows an example slice for such an image. The grid-like tags overlayed on the region of interest allow to follow the motion of keypoints on the boundary of, or inside the cardiac muscle. Each sequence in the dataset thus comes with a corresponding set of 12 landmarks, the motion of which was manually tracked over time. The landmarks are typically located where tags intersect, and divided in three groups of 4 points in the basal, mid-ventricular and apical areas of the left ventricle. Details of the experimental setting along with challenger results are provided and analyzed by [Tobon-Gomez 2013].

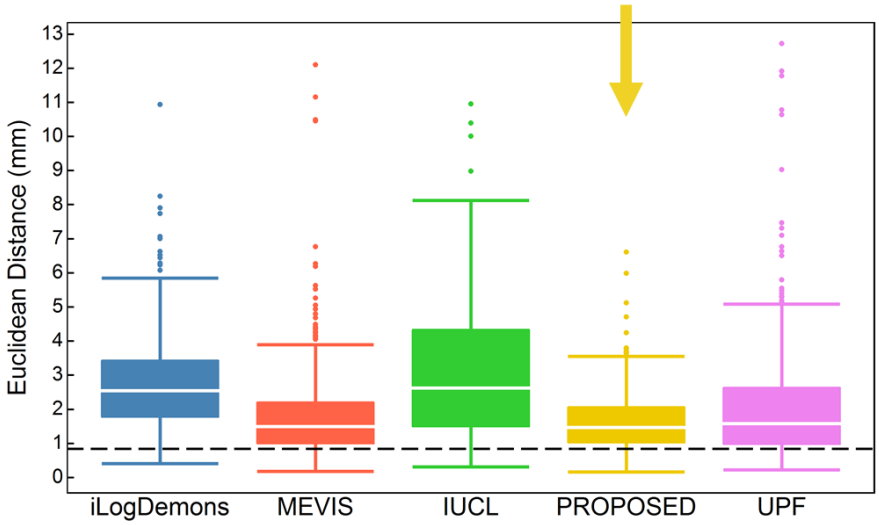


Figure 3.18: Accuracy benchmark on the 3D tag STACOM 2011 dataset, reporting boxplots of tracking errors on all methodologies. The doted black line represents the average inter-observer variability.

We validate our approach on this dataset of real 3D tagged MR sequences. Manual landmarks serve as ground truth from which the accuracy of our methodology at End-Systole (ES) is assessed. Fig. 3.18 summarizes challengers' results along with ours. The proposed approach achieves state-of-art results on this benchmark with a median accuracy of 1.46mm. As a point of comparison, the variability in the landmark tracking was esti-

mated as part of the challenge methodology at 0.84mm. We perform two simple statistical tests to quantify the statistical significance of the increase in accuracy of our methodology compared to the challengers: a pairwise Student-t test and a pairwise Kolmogorov-Smirnov test. The tests are run for each pair of samples involving the proposed approach against a challenge participant's. The Student-t test aims at detecting significant differences in the true mean error of our method versus a challenger's, whereas the Kolmogorov-Smirnov test more generally aims at detecting whether the underlying distribution of errors differ. Figures are reported in Table 3.2 and provide some evidence towards a significant improvement from at least 3 of the 4 methodologies.

Challenger	Student-t p-value	KS p-value
iLogDemons	$< \mathbf{2.2 \cdot 10^{-16}}$	$< \mathbf{2.2 \cdot 10^{-16}}$
MEVIS	$\mathbf{0.0099}$	0.1385
IUCL	$< \mathbf{2.2 \cdot 10^{-16}}$	$< \mathbf{2.2 \cdot 10^{-16}}$
UPF	$\mathbf{2.45 \cdot 10^{-5}}$	$\mathbf{0.00024}$

Table 3.2: Statistical significance of the increase in accuracy on the STACOM 2011 3D motion tracking challenge. We report p-values of pairwise tests for the proposed approach versus each participant's. Bolded values highlight significant improvements at the 5% significance level.

Rather than over-emphasizing the reach of these statistical tests we would like the reader to appreciate the ability of the proposed formulation to achieve in a quasi automatic manner results qualitatively and quantitatively on par with the state of the art, as demonstrated on this benchmark. In particular it should be emphasized that all parameters involved in the proposed formulation – the noise model and regularity level λ, the active parametrization of the displacement field – were automatically determined during registration. All 3D registration experiments reported in this section involve little user interaction, in effect run from the same model and settings.

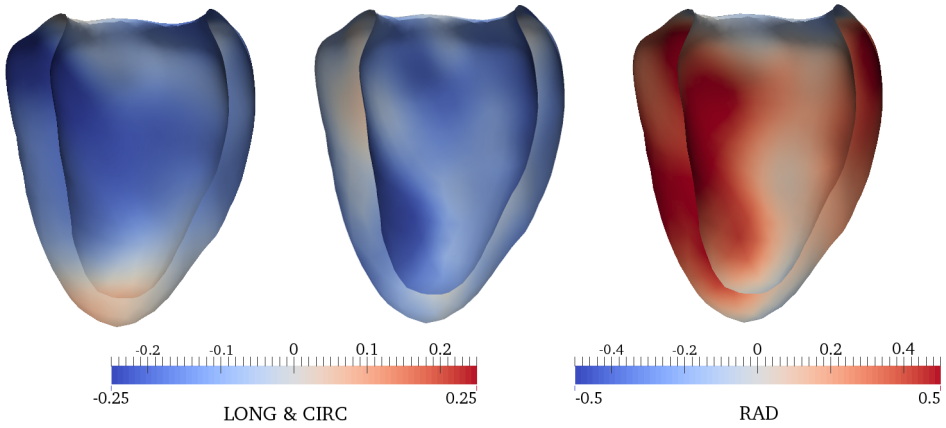


Figure 3.19: Strain at ES, computed from the 3D tag data of volunteer V9.

Interestingly the strain maps and mesh deformations produced by our scheme, as illustrated for instance in Fig. 3.19, also appear to be qualitatively on par with the best challenge results in that respect, and superior to that of the closest competing methodology accuracy wise (please refer to [Tobon-Gomez 2013] for a direct counterpart to Fig. 3.19). This hints towards the fact that the automatically adjusted weights of the data energy versus the regularization energy, beyond their theoretical grounds, are of practical relevance. Finally we report in Table 3.3 the number of bases of each scale used for the parametrization of the displacement field.

Basis type	Median # (Q1 – Q3)
$\sigma = 20$mm	17 (14.75 – 20)
anisotropic σ	30 (25 – 33.25)
$\sigma = 10$mm	100 (90 – 110.25)
Total	**148 (134 – 160)**

Table 3.3: Number of bases at each scale in the active parametrization of the displacement field (pooled over all sequences and all frames). Median, first and third quartiles are reported.

3.4.4 Cine MRI dataset: qualitative results and uncertainty

In addition to tagged MRI sequences, the STACOM 2011 challenge datasets include 15 sequences of cine SSFP MR cardiac images. For each of the 15 volunteers the left and right ventricles are imaged in 3D, over 30 frames covering the cardiac cycle. Original images had a low inter-slice resolution of 8mm compared to the in-plane resolution of 1.25mm, and we upsampled them (typically by a factor of 5) prior to the registration process to prevent a degradation of numerical accuracy. To obtain a ground truth by direct manual tracking of landmarks over time was deemed difficult for this image modality. Instead the accuracy of the proposed algorithm was evaluated by cross-comparison with direct 3D+t segmentation results. Specifically, the endocardium was delineated over time on 2D slices using the freely available software Segment[2] [Heiberg 2005], yielding a 3D point set of discretized contours. A 3D surface was then reconstructed as the zero level set of a signed distance map computed by radial basis interpolation, after estimating the normal to the surface at every point in the set from a local neighborhood[3]. We then assessed the discrepancy between the reference end diastole segmentation transported over time *via* the output of registration, and the surface estimated by direct segmentation of the endocardium at each time step.

Fig. 3.20 summarizes the distribution of errors over time, pooled over all 15 sequences and all contour points, displaying the evolution of key quantile-based statistics. The median error reaches a satisfactory maximum of 1.82mm for frame 10, which roughly coincides with the end systole time for all volunteers. As a point of comparison, the volumes under

[2] http://segment.heiberg.se
[3] http://hdl.handle.net/10380/3149

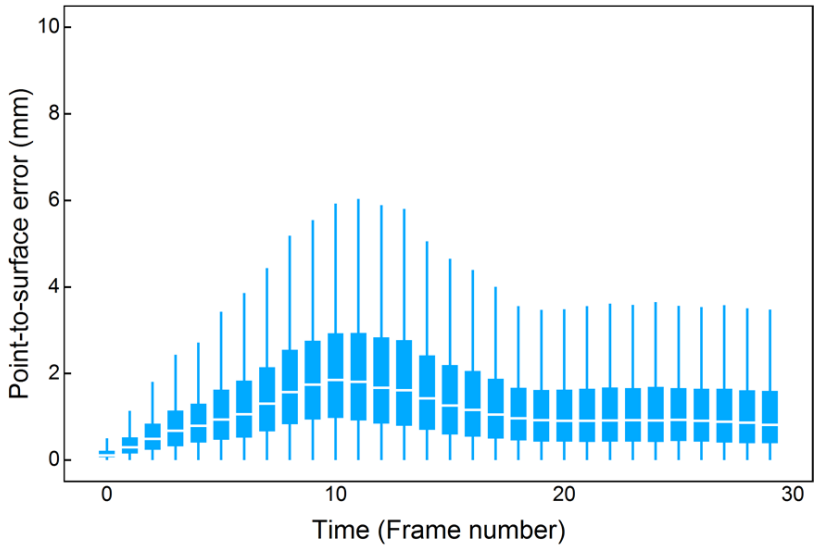


Figure 3.20: Accuracy benchmark on the cine SSFP STACOM 2011 dataset, reporting median error over time along with quartiles. Surfaces reconstructed from slice-by-slice 3D+t segmentation serve as ground truth. Points on the discrete contours delineated at time 0 are transported over time with the registration output and point-to-surface distances are gathered. For each time step, errors over all 15 sequences and all contour points are pooled. Errors at time 0 are induced by the surface reconstruction.

consideration have a spacing of 1.25mm in the short-axis plane (*i.e.* within slices) and 8mm along the long-axis (*i.e.* inter-slice). The wide spread of error values partly reflects the challenge in obtaining a 3D segmentation of the endocardium that remains consistent over time (*e.g.* due to the variable appearance of papillary muscles). Misalignment of short-axis slices in 3D volumes, which may arise from the (slice by slice) image acquisition process, also accounts for some of the largest discrepancies. We observed no evident spatial pattern in the distribution of errors, although the segmentation rarely reached the very tip of the apical region.

The mixture-of-Gaussians noise model proved adequate here again, as distinct components captured variations in the level of noise of an order of magnitude (a factor of 10 between the standard deviations of extreme components). Indeed regions of interest change appearance over time and motion, and tend to be assigned higher noise levels than the baseline acquisition noise level. Voxels in basal slices, with visible outflow tracts and apparent topology changes, also tend to fall in the noisiest components. Finally, Fig. 3.21 attests to the high variability (several orders of magnitude) of the optimal model parameter λ for varying sequences and time steps, which would render its manual estimation via a trial-and-error or cross-validation approach cumbersome. The apparent bimodality of the histogram might reflect the fact that cardiac phases with significant contraction or relaxation, around end systole, alternate with phases of lesser motion around end diastole.

Finally, because of the strong anisotropy in spacing (lower inter-slice resolution) that is characteristic of the image acquisition process, cine MR data examplifies the necessity

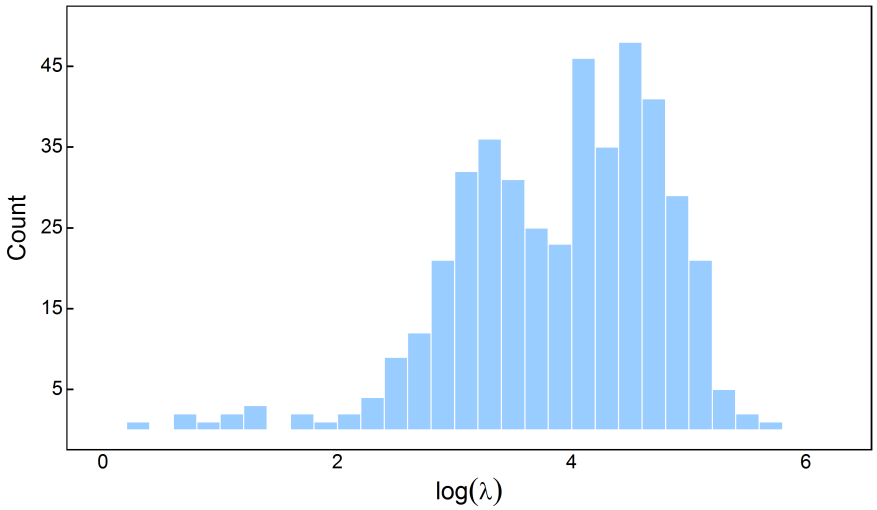


Figure 3.21: Histogram of inferred values for the regularity hyperparameter λ, pooled over all 15 sequences and 30 frames per sequence.

of modeling uncertainty in the interpolation of discrete intensity profiles (cf. sections 3.2.1 and 3.3.1). When not accounting for it the regularity of the inferred transform was systematically found, upon visual inspection, to be of lesser quality in the direction of lower resolution. This behaviour is to be expected if the scheme is blind to its increased reliance on interpolation to find locations of matching intensity values in the moving image. Image upsampling prior to registration also relies on interpolation and was accounted for in an identical manner. Table 3.4 reports statistics on the number of bases of each scale used for the parametrization of the displacement field.

Basis type	Median # (Q1 – Q3)
$\sigma = 20$mm	18 (10 – 28)
anisotropic σ	29 (21 – 38)
$\sigma = 10$mm	44 (29 – 57)
Total	**93 (75 – 111)**

Table 3.4: Number of bases in the active parametrization of the displacement field (pooled over all sequences and all frames), at each scale. Median, first and third quartiles are reported.

3.5 Discussion and conclusions

We proposed a data-driven, spatially adaptative, multiscale parametrization of deformations for registration. It uses larger kernels in regions of high uncertainty due to *e.g.* lack of image gradients or incoherent information in registered images, and uses smaller kernels

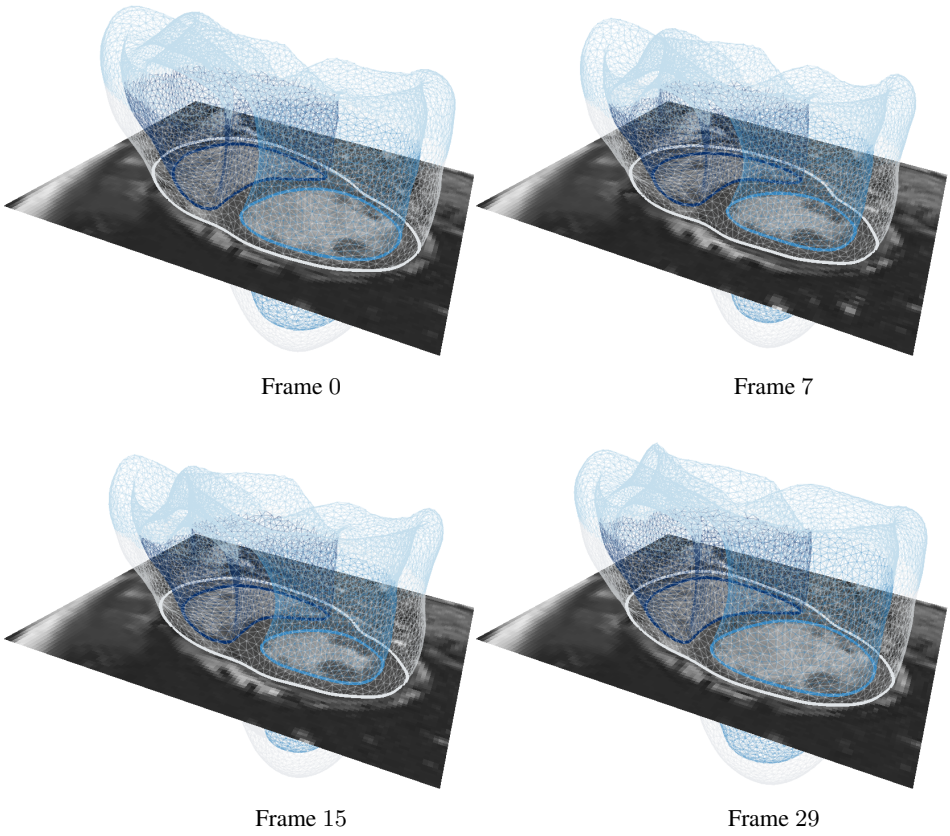


Figure 3.22: Example registration for the cine SSFP dataset of volunteer V5. We propagate the segmentation from the reference frame to the rest of the time-series with the output of the registration. The resulting mesh is overlayed on a 2D slice and visualized at four representative timesteps. The 3D mesh attests to the regularity of the underlying transform, and to its coherence over the cardiac cycle.

where a finer motion can be estimated with confidence from local cues in paired images. This is achieved in a Bayesian framework, so that the approach retains natural advantages of probabilistic formulations such as the joint inference of registration hyperparameters. In effect this yields a self-tuning algorithm of general scope with attractive capabilities in terms of achieved accuracy and regularity. The prime contribution of this work is a procedure for fast marginal likelihood maximisation in sparse Bayesian models, that relaxes the assumptions made by [Tipping 2003] for the fast Relevance Vector Machine at virtually no algorithmic cost. It broadens the scope of the RVM much in the same direction as the mixed $L1$-$L2$ Elastic Net regularization [Zou 2005] extends the $L1$-norm LASSO regularization [Tibshirani 1996]. This scheme applies to the wide range of classification and regression tasks that share the same abstract graphical representation as Fig. 3.2 or variants thereof.

While we left the question of uncertainty quantification mostly unadressed, we note that the proposed framework provides us with a readily computable, compact $|\mathcal{S}| \times |\mathcal{S}|$ covariance matrix $\mathbf{\Sigma}$ that summarizes uncertainty on the degrees of freedom in the active

parametrization. The covariance on transformation parameters can be turned into directional estimates of uncertainty at any point in space by simple linear algebra, or can be sampled from at a marginal $\mathcal{O}(|\mathcal{S}|^2)$ cost to efficiently explore the *joint* variability of the full transformation. Sampling the transformation itself, unlike sampling displacements independently at each point in space, preserves correlations in the displacement of close-by points. This can be instrumental in deriving empirical estimates of uncertainty on integral geometrical quantities. For instance, Fig. 3.1(b) reports estimates of uncertainty in the volume enclosed over time by the endocardium surface, as segmented on the reference frame (at time 0), for a cine MRI sequence (volunteer 5). For the same volunteer, Fig. 3.1(c) summarizes, in the form of a tensor map, the uncertainty in the inferred displacement field at end-systole, accounting for uncertainty in the output of each frame-to-frame registration between end-diastole and end-systole. Tensors are rasterized at the voxel centers of the end-systole frame. Each tensor encodes (the square root of) the 3×3 covariance matrix of the displacement at a given point and is evidently elongated in directions of higher uncertainty. Due to voxel spacing anisotropy in the cine SSFP dataset, the direction of higher uncertainty is, consistently across space, aligned with the long-axis. The color scheme thus encodes the second principal direction of highest uncertainty. Steep intensity gradients in the underlying image typically translate into directions where tensors are least elongated. Tensor magnitude and principal directions vary smoothly across space, as estimates of uncertainty incorporate information of a local (and in fact, global) nature. The yellow dashed line gives a visual cue as to the position of the left ventricle endocardium boundary. Future work will have to assess extensively the quality of the approximate posterior returned by our fast scheme compared to the exact model posterior (as explored *e.g.* via MCMC techniques), and the empirical agreement between the exact model posterior and intuition.

We experimented with the proposed framework of registration on tasks of motion tracking on dynamic cardiac data. A flexible noise model based on mixtures of Gaussian distributions was introduced and performed suitably on all tested modalities, advantageously replacing models of smoothly varying noise. This is likely due to an increased ability to represent intricate spatial patterns of intensity residuals arising from acquisition noise and artefacts, registration misalignment and variable appearance of organs over time. Despite using generic multiscale RBFs, the inferred parametrizations of 3D displacements were highly sparse, typically involving no more than a hundred degrees of freedom. We note that tagged MR images encouraged the recourse to finer bases; beyond the higher resolution of these volumes (compared to cine SSFP data), it is likely that tags were found to be reliable, informative structures along all directions of motion. While good accuracy was achieved on synthetic echocardiographic time series with a reduced number of bases, the synthetic motion from which the sequences were reconstructed is likely to have enjoyed greater regularity as well.

Despite not relying on temporal regularization to help in tracking the motion of the cardiac muscle (as in *e.g.* [De Craene 2012]), the temporal and spatial consistency of deformations was judged satisfactory, as evidenced by Fig. 3.1(a), 3.22. Still, incorporating temporal regularization and moving towards a large deformation framework [Beg 2005, Arsigny 2006] with geodesic-by-part trajectories may be fruitful for this application. It could also constitute a step in moving from a data-driven parametrization of

displacements (beneficial to the quality of registration) towards an anatomically relevant parametrization. Indeed the proposed framework of basis selection may be suitable for learning a parametric atlas of motion from a small dataset of 3D+t images in the spirit of *e.g.* [Allassonnière 2007, Durrleman 2013, Gori 2013].

CHAPTER 4

Quantifying Registration Uncertainty with Sparse Bayesian Modelling: Is Sparsity a Bane?

Contents

4.1 Introduction

Non-rigid image registration is an ill-posed task that supplements limited, noisy data with 'inexact but useful' prior knowledge to infer an optimal deformation between images of interest. As a standard processing step in many pipelines for medical imaging, for computational anatomy & physiology, registration would greatly benefit from the development of principled strategies to analyze its output and to subsequently re-evaluate model assumptions. Bayesian modelling and inference, potentially in conjunction with hypothesis testing, provides a framework to explicitly incorporate prior assumptions and re-assess their relevance in retrospect. We focus on another expected benefit of Bayesian approaches, the opportunity to go beyond point estimates of the optimal deformation towards quantification of uncertainty in the solution.

In a seminal paper Gee and Bajcsy [Gee 1998] lay the groundwork for a Bayesian interpretation of registration, extending the mechanical formulation of Broit [Broit 1981]. Exploiting the Gaussian Markov random field structure inherited from a finite-element discretization of the domain, they characterize the posterior distribution of displacements by

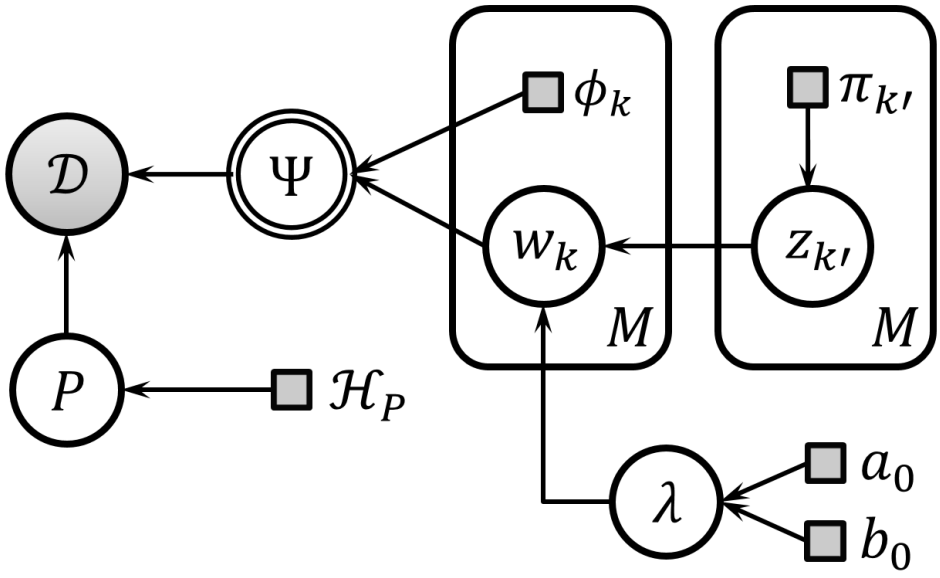


Figure 4.1: Graphical representation of the probabilistic registration model. The process of generating the data D involves a transformation Ψ of space, and some noise whose model is controlled by a set of parameters P. Hyperpriors (with hyperparameters $\mathcal{H}_P$) are in turn imposed over the noise parameters P. The transformation is parameterized by weights w_k on a predefined overcomplete set of basis functions $\{\phi_k,\, k = 1 \cdots M\}$. Priors on the transformation smoothness and on the relevance of individual bases introduce additional parameters (λ and $z_{k'}$ respectively). Random variables are circled, whereas deterministic parameters are in squares. Arrows capture conditional dependencies. Shaded nodes are observed variables or fixed hyperparameters. The doubly circled node indicates that the transformation Ψ is fully determined by its parent nodes (the ϕ_k and $\boldsymbol{w}_k$). The plate stands for M (replicated) groups of nodes, of which only a single one is shown explicitly.

Gibbs sampling. Risholm *et al* [Risholm 2013] extend the approach to the case of unknown confidence on the observed data and on model priors respectively, aiming to address the critical issue of finding an objective trade-off between data fit and regularity-inducing priors. The so-called temperature hyperparameters are treated as latent variables and approximately marginalized over, while an MCMC chain with full dimensional Metropolis-Hastings transitions traverses the space of transformation parameters. Simpson *et al* [Simpson 2012] propose an alternate methodology based on variational Bayes inference. Using a parametric (FFD) representation of the displacement field, the posterior distribution is approximated within a 'convenient' family for which transformation parameters and model hyperparameters factorize. As a result, estimates of uncertainty quantify variability in the displacement field *conditionally* to the inferred hyperparameters, but disregard uncertainty induced by hyperparameter variability. The variational factorization nevertheless proves to be computationally advantageous. This work and ensuing developments [Simpson 2013] also expose crucial limitations of classic generative models of data for registration. Although uncertainty quantification is peripheral to their work,

Richard *et al* [Richard 2009] develop for the related task of atlas building a mixed SAEM and MCMC approach where nodes of the finite-element mesh are updated via Metropolis-Hastings-Within-Gibbs transitions; Zhang *et al* [Zhang 2013] implement a mixed SAEM and Hybrid Monte Carlo approach for a Bayesian MAP estimation of the template and of temperature hyperparameters in a diffeomorphic setting.

Our contribution is twofold. Whist applied to a particular model of registration introduced in earlier work [Le Folgoc 2015b], our inference schemes open broader avenues for efficient and objective comparison between competing registration models. At a high level, the space of transformation parameters is explored by a reversible jump Markov chain. Reversible jump MCMC [Green 1995] provides a principled mechanism to elegantly jump between competing models of varying dimensionalities while sampling the posterior density of interest. In this work, reversible jump transitions allow to seamlessly refine the parametrization of the transformation, adapting the granularity of the parametrization to the granularity of the underlying motion and the local informativeness of the image, all the while exploring the most likely deformations. As it carefully avoids the computation of so-called Bayes factor, the method proceeds significantly faster than state-of-the-art MCMC inference schemes for registration. At a lower level, we capitalize on closed form marginalization of most nuisance variables, and integrate second-order knowledge of the posterior distribution in proposal kernels. This yields an algorithm that reliably and consistently traverses the parameter space towards the most likely deformations in spite of the model intricacies.

The proposed 'sparse Bayesian' model of registration [Le Folgoc 2015b] is adapted from the Relevance Vector Machine (RVM) devised by Tipping *et al* [Tipping 2001, Tipping 2003] for the purpose of classification and regression. In these articles approximate inference is achieved by marginal likelihood maximization (the 'evidence framework'), which provides a point estimate of the sparsity-governing hyperparameters. An approximate Gaussian posterior for the regressor parameters, conditionally to the optimal set of hyperparameters, is then derived. Rasmussen *et al* [Rasmussen 2005] exposed the peculiar behaviour of the RVM in terms of predictive uncertainty, yielding *lower* uncertainty away from data (see *e.g.* [Quiñonero-Candela 2005]). Several ad-hoc methods were subsequently suggested to circumvent this behaviour. We adopt a different stance, showing that while the *approximate* posterior derived by evidence maximization suffers from inconsistent predictive uncertainties, the *exact* posterior resulting from the proposed sparse Bayesian model does not.

The chapter unfolds as follows. In part 4.2 we describe the sparse Bayesian model of registration and devise a principled strategy for exact inference. The proposed design of the Markov chain exploits insight gained into the model to bypass standard impediments of MCMC schemes. Hyperparameter uncertainty is fully accounted for by marginalization of the nuisance variables. In part 4.3 we review breakdown scenarii in which the approximate posterior significantly departs from the true posterior, leading to poor approximate predictive uncertainty. In part 4.4 we conduct preliminary experiments to assess the validity of MCMC uncertainty estimates.

4.2 Statistical Model and Inference

Registration infers, from prior knowledge and limited data $\mathcal{D}$, a transformation of space Ψ that pairs homologous features in objects of interests (*e.g.* organs or vessels, in a medical setting). The section starts with a succinct description of the registration model, developed by the authors in Chapter 3, and offers insight into its mechanisms. Fig. 4.1 provides a graphical representation thereof. An MCMC approach for systematic characterization of the posterior distribution is then devised.

4.2.1 Bayesian Model of Registration

4.2.1.1 Likelihood model

The generative model of data makes explicit the relationship between the data $\mathcal{D}$ and the spatial mapping Ψ. It is specified by a likelihood model $p(\mathcal{D}|\Psi; P)$ (often conditioned on a set of hyperparameters P) that typically assumes the form of a Boltzmann distribution $p(\mathcal{D}|\Psi; P) \propto \exp -\mathcal{E}_{\mathcal{D}}(\mathcal{D}, \Psi; P)$. For landmark registration, a transformation that approximately maps corresponding key points $\{t_i\}$ and $\{T_i\}$, $i = 1 \cdots N$, between a template object and a target object is sought. A standard choice of energy is the sum of squared

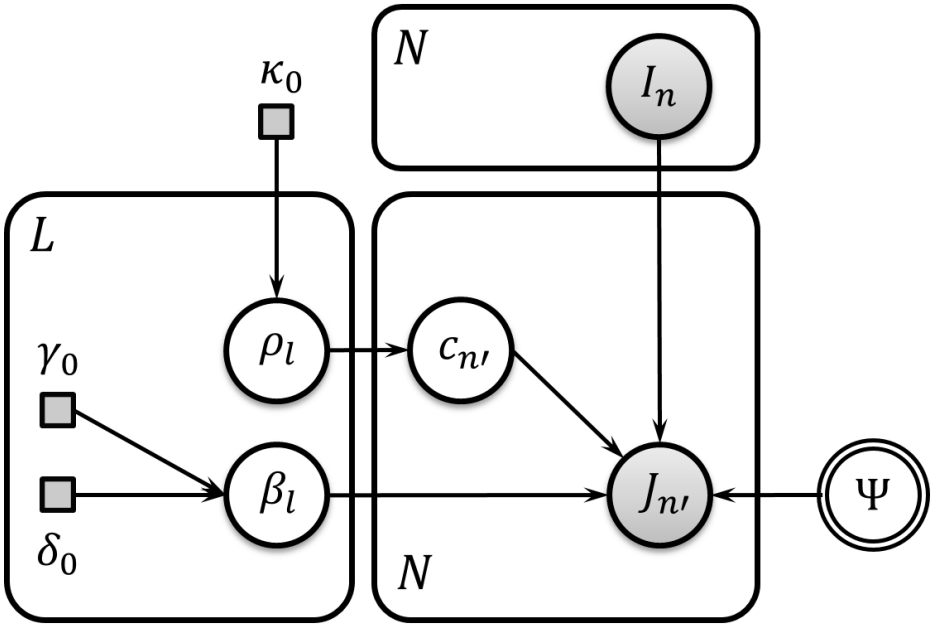


Figure 4.2: Graphical representation of the generative data model. Graphical codes are identical to those of Fig. 4.1. Residuals between the fixed image J and the warped image $I \circ \Psi^{-1}$ are assumed to be distributed according to a mixture of L Gaussian components whose parameters ρ_l (probability of falling in the lth component) and β_l (inverse variance a.k.a. precision parameter for the lth Gaussian component) are regarded as latent variables. $c_{n\prime} \in \{1 \cdots L\}$ assigns the corresponding voxel to one of the L mixture components.

distances between pairings, up to multiplicative factor:

$$\mathcal{E}_{\mathcal{D}}(\mathcal{D}, \Psi; \beta) = \frac{\beta}{2} \sum_{i=1}^{N} \|T_i - \Psi(t_i)\|^2 . \tag{4.1}$$

For pairwise registration of a fixed image J and a moving image I, a mixture-of-Gaussians model of intensity residuals is adopted. At voxel center v_i, the intensity residual $r_i = J(v_i) - I[\Psi^{-1}(v_i)]$ is assigned to the lth component of the mixture, $1 \leq l \leq L$, if the L-way categorical variable $c_i \in \{1 \cdots L\}$ takes value l. If so the residual r_i follows a normal distribution $\mathcal{N}(0, \beta_l^{-1})$. The component assignment c_i follows a categorical distribution and takes value l with probability ρ_l, normalized such that $\sum_{l=1}^{L} \rho_l = 1$. For distinct voxels v_i and v_j, residuals r_i and r_j (resp. component assignments c_i and c_j) are assumed to be independent. The corresponding GMM energy $\mathcal{E}_{\mathcal{D}}(\mathcal{D}, \Psi; \boldsymbol{\beta}, \boldsymbol{\rho})$ is given by Eq. (4.2), with $Z_l = \sqrt{2\pi/\beta_l}$ a normalizing constant:

$$-\sum_{i=1}^{N} \log \sum_{l=1}^{L} \frac{\rho_l}{Z_l} \exp -\frac{\beta_l}{2} \left(J[v_i] - I[\Psi^{-1}(v_i)] \right)^2 \tag{4.2}$$

Fig. 4.2 summarizes this model of data in graphical form. The assumption of independence of voxelwise residuals is known not to hold (see *e.g.* [Simpson 2012, Le Folgoc 2015b]) and to affect the outcome of the probabilistic registration. Since a proper probabilistic account of correlations in intensity residuals is both beyond the scope of this work and irrelevant to the ensuing developments, the *Virtual Decimation* scheme of [Simpson 2012] is reproduced instead for simplicity.

4.2.1.2 Transformation parametrization

A small deformation standpoint is adopted for convenience. The displacement field $\mathrm{u} \colon \mathrm{x} \in \Omega \subset \mathbb{R}^d \mapsto u(x) = \Psi^{-1}(x) - x \in \mathbb{R}^d$ is parametrized by a linear combination of M basis functions $\phi_k(.)$ with associated weight $w_k \in \mathbb{R}^d$:

$$u(x) = \sum_{1 \leq k \leq M} \phi_k(x) w_k = \boldsymbol{\phi}(x)^\intercal \boldsymbol{w} . \tag{4.3}$$

$\boldsymbol{\phi}(\mathrm{x}) = \left(\phi_1(\mathrm{x}) \cdots \phi_M(\mathrm{x})\right)^\intercal$ and $\boldsymbol{w}^\intercal = \left(w_1^\intercal \cdots w_M^\intercal\right)$ are respectively the concatenation, for $k = 1 \cdots M$, of $\phi_k(\mathrm{x})$ and w_k. Arbitrary choices of basis functions ϕ_k are possible. B-splines (*e.g.* [Rueckert 1999]) present desirable properties in terms of smoothness and interpolation. Here the ϕ_k's instead consist of multiscale Gaussian radial basis functions (RBFs) whose centers lie on a regular grid of points (typically, decimated voxel centers). Multiscale Gaussian RBFs possess attractive analytical and computational properties.

4.2.1.3 Transformation priors

The weights $\boldsymbol{w}$ are endowed with a generalized Spike-&-Slab prior that favours both smoothness of the resulting displacement field and sparsity in its parametrization. The properties of this prior are central to the proposed 'sparse Bayesian' modelling and to our

analysis thereof. Each basis ϕ_k is assigned a distinct activation variable z_k that controls its inclusion in the active parametrization (or exclusion therefrom). If $z_k = 0$ the basis ϕ_k is pruned out of the active parametrization. We do so by designing $p(w_k|z_k = 0)$ as a Dirac distribution centered at 0. If $z_k = 1$ the basis ϕ_k is included in the parametrization. The prior on such bases is designed as a joint, structured Gaussian distribution that penalizes lack of smoothness in the induced displacement field. Let us denote by $\mathcal{S}$ the set of such indices k for which $z_k = 1$ and by $\boldsymbol{w}_\mathcal{S}$ the concatenation of the corresponding subset of weights $\{w_k,\, k \in \mathcal{S}\}$. For an arbitrary linear differential operator D, we wish to penalize high values of the quadratic energy $\|D\mathrm{u}\|^2 = \boldsymbol{w}_\mathcal{S}^\intercal \mathbf{R}_\mathcal{S} \boldsymbol{w}_\mathcal{S}$, where $\mathbf{R}_\mathcal{S}$ is the $|\mathcal{S}| \times |\mathcal{S}|$ matrix whose k,l-th coefficient is $\langle D\phi_k | D\phi_l \rangle$. The Gaussian distribution $\mathcal{N}(\boldsymbol{w}_\mathcal{S} \,|\, \mathbf{0}, \{\lambda\, d|\mathcal{S}|\}^{-1} \mathbf{R}_\mathcal{S}^{-1})$ is a natural choice of prior for $p(\boldsymbol{w}_\mathcal{S}|\mathcal{S})$, that we adopt henceforth. Note that the covariance normalization by $d|\mathcal{S}|$, where d is the image dimension, departs from that of [Le Folgoc 2015b]. Under this prior $\lambda\, d|\mathcal{S}| \cdot \boldsymbol{w}_\mathcal{S}^\intercal \mathbf{R}_\mathcal{S} \boldsymbol{w}_\mathcal{S}$ is $\chi^2(d|\mathcal{S}|)$ distributed so that λ immediately relates to the expectation of the energy: $\mathbb{E}_{p(\boldsymbol{w}_\mathcal{S}|\mathcal{S})}(\|D\mathrm{u}\|^2) = \lambda^{-1}$ and $\mathbb{E}_{p(\boldsymbol{w})}(\|D\mathrm{u}\|^2) = \lambda^{-1}$. The prior over all weights $\boldsymbol{w}$ conditioned on the state of the gate variables $\boldsymbol{z} = \begin{pmatrix} z_1 & \cdots & z_M \end{pmatrix}^\intercal$ is best summarized in the form of Eq. (4.4), where $-\mathcal{S}$ is the complement of $\mathcal{S}$:

$$p(\boldsymbol{w}|\boldsymbol{z}, \lambda) = \mathcal{N}(\boldsymbol{w}_\mathcal{S} \,|\, \mathbf{0}, \frac{1}{\lambda\, d|\mathcal{S}|} \mathbf{R}_\mathcal{S}^{-1}) \cdot \mathcal{N}(\boldsymbol{w}_{-\mathcal{S}} \,|\, \mathbf{0}, \mathbf{0})\,. \tag{4.4}$$

4.2.1.4 Hyperpriors

Parameters introduced in the specification of priors are in turn treated as latent variables. λ is endowed with a Gamma prior $\Gamma(\lambda|a_0, b_0)$ that is conjugate to $p(\boldsymbol{w}|\boldsymbol{z}, \lambda)$. The parameters β_l (resp. β) involved in the likelihood model for image (resp. landmark) registration are endowed with independent Gamma priors $\Gamma(\beta_l|\gamma_0, \delta_0)$. The noise mixture proportions $\boldsymbol{\rho} = \{\rho_1 \cdots \rho_L\}$ are assigned a Dirichlet prior $\mathrm{Dir}(\boldsymbol{\rho}|\boldsymbol{\kappa})$, with $\boldsymbol{\kappa} = (\kappa_1 \cdots \kappa_L)$.

Independent Bernoulli priors $\mathcal{B}(z_k|\pi_k)$ on each z_k constitute a natural, conjugate hyperprior specification for the activation variables $\boldsymbol{z}$. The positive mass $1 - \pi_k$ concentrated at $w_k = 0$ as a result explicitly encodes sparsity. Assuming all $\pi_k = \pi_0$ to be equal, all parametrizations using the same number of active bases $|\mathcal{S}|$ are a priori equally probable. In addition the cost of including a new basis in the active parametrization is independent of the current number of active bases. However, we opt instead for a stronger prior, $p(\boldsymbol{z}) \propto \Gamma(\frac{d|\mathcal{S}|}{2})^{-1}$. The Gamma function $\Gamma(\cdot)$ is a natural extension of the (integer) factorial to real values, yielding a prior that increasingly penalizes each new inclusion. This prior was found to perform better w.r.t. sparsity, as can be theoretically argued from the analysis of the marginal prior $p(\boldsymbol{w}|\boldsymbol{z})$.

4.2.2 Model analysis

4.2.2.1 Marginal prior and marginal likelihood

Critical insight into the statistical model can be gained by considering the prior $p(\boldsymbol{w}|\boldsymbol{z}, \mathcal{H})$ and likelihood $p(\mathcal{D}|\boldsymbol{w}, \boldsymbol{c}, \mathcal{H})$ with so called *temperature* parameters λ and $\boldsymbol{\beta}$ marginalized

over, *e.g.*:

$$p(\boldsymbol{w}|\boldsymbol{z},\mathcal{H}) = \int_{\mathbb{R}_+} p(\boldsymbol{w}|\boldsymbol{z},\lambda,\mathcal{H})p(\lambda|\mathcal{H})d\lambda\,. \tag{4.5}$$

The multivariate Student distribution $t_\nu(\cdot|\boldsymbol{\mu},\boldsymbol{\Lambda})$ with location parameter $\boldsymbol{\mu}$, inverse scale matrix $\boldsymbol{\Lambda}$ and ν degrees of freedom naturally appears in analytic derivations (cf. Appendix A.5), yielding the following expressions for the prior and likelihood:

$$p(\boldsymbol{w}|\boldsymbol{z},\mathcal{H}) = \mathcal{N}\left(\boldsymbol{w}_{-\mathcal{S}}\middle|\mathbf{0},\mathbf{0}\right) t_{\nu_\lambda}\left(\boldsymbol{w}_{\mathcal{S}}\middle|\,\mathbf{0},\frac{a_0}{b_0}\,d|\mathcal{S}|\,\mathbf{R}_{\mathcal{S}}\right) \tag{4.6}$$

$$p(\mathcal{D}|\boldsymbol{w},\boldsymbol{c},\mathcal{H}) = \prod_{l=1}^{L} t_{\nu_l}\left(\boldsymbol{I}_l\circ\Psi_{\boldsymbol{w}}^{-1}\middle|\,\boldsymbol{J}_l,\frac{\gamma_0}{\delta_0}\mathbf{I}\right) \tag{4.7}$$

where $\nu_\lambda = 2a_0$, $\nu_l = 2\gamma_0$, $\mathcal{S}$ is the set of active bases and $|\mathcal{S}| = \sum_k z_k$ its cardinal. $\boldsymbol{J}_l = \left(\cdots J[v_i]\cdots\right)^{\intercal}_{i|c_i=l}$ is the vector of voxel values in image J, for those voxels assigned to component l, and $\boldsymbol{I}_l\circ\Psi_{\boldsymbol{w}}^{-1} = \left(\cdots I[\Psi_{\boldsymbol{w}}^{-1}(v_i)]\cdots\right)^{\intercal}_{i|c_i=l}$ is similarly defined for the warped image $I\circ\Psi_{\boldsymbol{w}}^{-1}$. For a fixed choice of active bases $\boldsymbol{z}$, the posterior distribution of the weights $p(\boldsymbol{w}|\boldsymbol{z},\boldsymbol{c},\mathcal{D},\mathcal{H})$ is proportional to the product of the prior Eq. (4.6) and likelihood Eq. (4.7). In the limit of uninformative hyperpriors $a_0,\gamma_0\to 0$, $\beta_0,\delta_0\to 0$ and assuming $L=1$ for the sake of illustration,

$$p(\boldsymbol{w}|\boldsymbol{z},\boldsymbol{c},\mathcal{D},\mathcal{H}) \propto \mathcal{N}\left(\boldsymbol{w}_{-\mathcal{S}}\middle|\mathbf{0},\mathbf{0}\right)\,\frac{1}{\chi_{\text{lik}}[\boldsymbol{w}]^N}\frac{1}{\chi_{\text{pr}}[\boldsymbol{w}]^{d|\mathcal{S}|}}\,. \tag{4.8}$$

where $\chi_{\text{lik}}[\boldsymbol{w}]^2$ is the data error and $\chi_{\text{pr}}[\boldsymbol{w}]^2 = \|D\mathrm{u}_{\boldsymbol{w}}\|^2$ the regularizing energy. In particular the posterior distribution is invariant to rescaling of the data error, and hence to rescaling of the intensity profile, after marginalizing over temperature parameters. Note also that, for a fixed parametrization $\boldsymbol{z}$, the ratio of posterior probabilities of two distinct parameter sets $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$ may become arbitrarily overwhelmed by the prior as the number of bases in the parametrization grows ($|\mathcal{S}|\gg N$). If not for sparsity, this might render MCMC characterization of the posterior unreliable (using *e.g.* Metropolis Hastings transitions), potentially making its outcome dependent on the size of the parametrization. Fortunately the proposed sparse model has a clear mechanism to prevent overparametrization and render overlapping bases largely mutually exclusive, as discussed next.

4.2.2.2 Prior probability of basis inclusion

Interactions between overlapping bases can be better understood by looking at the probability $p(z_k|\boldsymbol{w}_{-k},\boldsymbol{z}_{-k},\mathcal{H})$ of inclusion of a new basis z_k given a known configuration $\boldsymbol{z}_{-k}$ for the other bases and their associated weights $\boldsymbol{w}_{-k}$. The state $\boldsymbol{w}_{-k}$ of other bases informs us about the expected regularity of the signal $\mathrm{u}_{\boldsymbol{w}}$, introducing dependencies between z_k and $\boldsymbol{z}_{-k}$ conditionally to $\boldsymbol{w}_{-k}$. Denoting by $\tilde{\boldsymbol{z}}$ (resp. $\boldsymbol{z}$) the state with $z_k=1$ (resp. $z_k=0$), we see from Bayes' rule that:

$$\frac{p(z_k=1|\boldsymbol{w}_{-k},\boldsymbol{z}_{-k})}{p(z_k=0|\boldsymbol{w}_{-k},\boldsymbol{z}_{-k})} = \frac{p(\boldsymbol{w}_{-k}|\tilde{\boldsymbol{z}})}{p(\boldsymbol{w}_{-k}|\boldsymbol{z})}\frac{p(\tilde{\boldsymbol{z}})}{p(\boldsymbol{z})} \tag{4.9}$$

where the dependence on hyperparameters is made implicit for convenience of notations. Leaving details of derivations aside, we note that in the limit of uninformative values, the ratio of Eq. (4.9) takes the form of

$$\frac{p(\tilde{\boldsymbol{z}})}{p(\boldsymbol{z})}\left(\frac{|\kappa_k|}{|R_{k,k}|}\right)^{1/2}\left(1-\frac{\mu_{\text{pr}}^{k\intercal}\mathrm{R}_{k,k}\mu_{\text{pr}}^{k}}{\boldsymbol{w}_{-k}^{\intercal}\mathbf{R}_{\mathcal{S}}\boldsymbol{w}_{-k}}\right)^{-\frac{d|\mathcal{S}|}{2}} \tag{4.10}$$

where $\mathcal{S}$ is the set of active bases (excluding k), $\mu_{\text{pr}}^{k}=-\mathrm{R}_{k,k}^{-1}\mathbf{R}_{k}^{\intercal}\boldsymbol{w}_{-k}$ and $\kappa_k=\mathrm{R}_{k,k}-\mathbf{R}_{k}^{\intercal}\mathbf{R}_{\mathcal{S}}^{-1}\mathbf{R}_{k}$. The middle factor penalizes the inclusion of basis k if it overlaps with bases in the active set $\mathcal{S}$, in the sense of the metric induced by $\mathbf{R}$. κ_k is a measure of overlap of basis k with all bases in the active set $\mathcal{S}$ and is null if basis k is perfectly collinear to $\mathcal{S}$. The right most factor favors the inclusion of basis k if it is a priori expected to yield a significant increase in regularity.

4.2.3 Posterior Exploration by MCMC Sampling

For any set of points $X=\{x_1\cdots x_n\}$ in the admissible domain Ω, consider the vector of displacements $\boldsymbol{u}_X^{\intercal}=\big(u(x_1)^{\intercal}\cdots u(x_n)^{\intercal}\big)$. We wish to characterize the joint posterior distribution $p(\boldsymbol{u}_X|\mathcal{D},\mathcal{H})$ of any such vector of displacements for any discrete set X. To that aim we merely need to characterize the posterior distribution $p(\boldsymbol{w}|\mathcal{D},\mathcal{H})$ of the weights $\boldsymbol{w}$ involved in the parametrization of the transformation Ψ^{-1} sufficiently well.

4.2.3.1 Related work

MCMC methods are tools of predilection to explore arbitrarily complex distributions in a principled manner. Gibbs sampling [Geman 1984] cycles between latent variables, sampling from their conditional distributions in turn while other model variables remain fixed. It is attractive when conditional distributions are known in closed form whereas the joint distribution is untractable or computationally costly to sample. When the conditional cannot be sampled directly, a component-wise proposal may be used instead within a Metropolis-Hastings (MH) step (Metropolis-Within-Gibbs). Unfortunately, Gibbs sampling of temperature parameters is prone to failure, with the chain drifting away from regions of high probability for the duration of any finite MCMC run. Collapsing temperature parameters λ, $\boldsymbol{\beta}$ when sampling regressor variables $\boldsymbol{w}$ is highly opportune. In the context of registration, Risholm et al. [Risholm 2013] propose a MH scheme where marginalizing over temperature parameters induces the expensive computation of partition functions, for which an intricate procedure based on Laplace approximations is designed. In the proposed model, the computation of partition functions (specifically, marginal likelihoods, a.k.a. *evidences*) may arise as well when sampling gate variables z_k. Selecting a specific configuration $\boldsymbol{z}$ can be interpreted as a choice between competing models of varying complexity and dimensionality. The problem of estimating the evidence for a model is well studied in the statistical literature. A variety of methods exist, ranging from the straightforward Laplace approximation to more principled approaches typically exploiting samples from the (possibly augmented) posterior, including Chib's method [Chib 1995],

importance sampling, bridge sampling, path sampling (see *e.g.* [Gelman 1998]) and reversible jump MCMC [Green 1995]. The latter approach is in fact primarily concerned with sampling from a posterior distribution involving competing models (freely jumping between models in the process) and merely obtains evidence ratios as a byproduct. Reversible jump MCMC is appealing in our setting where competing models $\boldsymbol{z}$ are organized in series of nested models of increasing complexity, rendering its machinery mostly invisible. Reversible jump MCMC proceeds in the general framework of Metropolis-Hastings, hence a sound proposal must be crafted. We derive a sensible family of proposals from a modal analysis of the posterior distribution.

4.2.3.2 Modal analysis of the posterior & proposal

For the model described in 4.2.1, the Laplace approximation of the (conditional) posterior $p(\boldsymbol{w}|\mathcal{D},\boldsymbol{z},\boldsymbol{c},\mathcal{H})$ around its mode $\boldsymbol{w}_* = \arg\max_{\boldsymbol{w}} p(\boldsymbol{w}|\mathcal{D},\boldsymbol{z},\boldsymbol{c},\mathcal{H})$ takes the following form:

$$\begin{aligned} -\log\, p(\boldsymbol{w}|\mathcal{D},\boldsymbol{z},\boldsymbol{c},\mathcal{H}) \approx \\ \frac{1}{2}\frac{\gamma_0 + N/2}{\delta_0 + \chi^2_{\text{lik}}/2}\,(\boldsymbol{T}_* - \boldsymbol{\Phi}\boldsymbol{w})^\intercal \mathbf{H}_*(\boldsymbol{T}_* - \boldsymbol{\Phi}\boldsymbol{w}) + \\ \frac{1}{2}\frac{a_0 + |\mathcal{S}|/2}{b_0 + \chi^2_{\text{pr}}|\mathcal{S}|/2}\, d|\mathcal{S}|\, \boldsymbol{w}^\intercal \mathbf{R}_{\mathcal{S}}\boldsymbol{w} + \text{const}\,. \end{aligned} \tag{4.11}$$

where for the sake of illustration we take a single component mixture ($L = 1$, $c_i = 1$ for all i, $\boldsymbol{\beta} = \beta$). $\chi^2_{\text{pr}} = \boldsymbol{w}_*^\intercal \mathbf{R}_{\mathcal{S}}\boldsymbol{w}_*$ is the energy in the displacement field, χ^2_{lik} is the data error $\chi^2_{\text{lik}} = \sum_{i=1}^{N}(J[v_i] - I[\Psi_*^{-1}(v_i)])^2$ and we discard higher order terms in b_0, δ_0. $\boldsymbol{T}_*^\intercal = \left(T_{1*}^\intercal \cdots T_{N*}^\intercal\right)$ is a set of *virtual* pairings whose value does not depend on $\boldsymbol{\beta}$, λ. $\mathbf{H}_*$ is a block diagonal matrix whose ith diagonal block $\mathbf{H}_i^*$ is the $d \times d$ precision matrix associated to the ith virtual pairing T_{i*}. The factors stemming from the marginalization:

$$\beta_* = \frac{\gamma_0 + N/2}{\delta_0 + \chi^2_{\text{lik}}/2}\,, \quad \lambda_* = \frac{a_0 + |\mathcal{S}|/2}{b_0 + \chi^2_{\text{pr}}|\mathcal{S}|/2} \tag{4.12}$$

are commensurable to temperature parameters. The approximation of the conditional posterior is Gaussian (Eq. (4.11) is quadratic) and admits the more obvious canonical form $\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})$, with $\boldsymbol{\mu} = \boldsymbol{\Sigma}\boldsymbol{\Phi}^\intercal(\beta_*\mathbf{H}_*)\boldsymbol{T}_*$ and $\boldsymbol{\Sigma} = (\boldsymbol{\Phi}^\intercal\beta_*\mathbf{H}_*\boldsymbol{\Phi} + \lambda_*|\mathcal{S}|\mathbf{R}_{\mathcal{S}})^{-1}$. The Laplace approximation provides a reasonable approximation of the posterior and a judicious starting point to design proposals. Component-wise proposals that leave most of the activation variables z_l and the corresponding weights w_l unchanged will be of particular interest to us (cf. section 4.2.3.3). A natural idea is to use the conditionals $\tilde{w}_k \sim \mathcal{N}(\mu^k_{\text{pos}}, \Sigma_k)$ of the Laplace approximation $\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})$ as proposal distributions. Because they neither require the actual computation of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ nor involve inner products $\phi_k^\intercal(\beta_*\mathbf{H}_*)\phi_l$, these 'Gibbs-like' proposals are computationally appealing. As a final tweak to alleviate modal assumptions, we reintroduce dependency on the current value of w_k, yielding the following component-wise proposal instead, with $0 \leq r_{\text{HMALA}} \leq 1$ and $s \geq 1$:

$$q_k(w_k \to \tilde{w}_k) = \mathcal{N}\left(\tilde{w}_k \,|\, m_k(w_k), s\Sigma_k\right) \tag{4.13}$$

$$m_k(w_k) = (1 - r_{\text{HMALA}})\, w_k + r_{\text{HMALA}}\, \mu^k_{\text{pos}} \tag{4.14}$$

If not set to 1, the factor s accounts for potentially fatter tails of the true conditional posterior in the proposal. μ^k_{pos} and Σ_k depend on $\mathbf{H}_*$ and $\boldsymbol{T}_*$, which in the formal reasoning based on the Laplace approximation are computed around $\Psi_*^{-1}(\cdot) = \text{Id} + \boldsymbol{\phi}(\cdot)^\intercal \boldsymbol{w}_*$. In fact $\boldsymbol{T}_*$ and $\mathbf{H}_*$ can be replaced by $\boldsymbol{T}_{\boldsymbol{w}}$ and $\mathbf{H}_{\boldsymbol{w}}$ computed from a (local) quadratic approximation of $p(\boldsymbol{w}|\mathcal{D}, \boldsymbol{z}, \boldsymbol{c}, \lambda_*, \beta_*)$ around the current $\Psi^{-1}(\cdot) = \text{Id} + \boldsymbol{\phi}(\cdot)^\intercal \boldsymbol{w}$. In that case Eq. (4.13), (4.14) exactly coincide with a component-wise Hessian preconditioned Metropolis Adjusted Langevin Algorithm (HMALA) [Roberts 1996, Girolami 2011, Zhang 2011], which exploits first and second order local information about the target distribution for increased efficiency. However the local approximation generates additional computations at each step and offers little gain if we expect the posterior to be unimodal. Given our experimental settings, we use the global approximation with adaptation during the burn-in phase (at that stage λ_*, β_*, $\boldsymbol{T}_*$ and $\mathbf{H}_*$ are recomputed every few iterations from statistics $|\mathcal{S}|$, χ^2_{pr}, χ^2_{lik} averaged with decaying weights over past samples).

4.2.3.3 Reversible jump MCMC scheme

The groundwork for this scheme was laid in sections 4.2.2.1, 4.2.2.2, 4.2.3.2. The reversible jump procedure itself lets us generate samples of the joint posterior $p(\boldsymbol{w}, \boldsymbol{z}, \boldsymbol{c}|\mathcal{D}, \mathcal{H})$ with temperature parameters marginalized over. Dropping irrelevant variables in the generated samples, we obtain samples of the marginals of interest, *e.g.* $p(\boldsymbol{w}|\mathcal{D}, \mathcal{H})$. The reversible jump scheme simply proposes to move from a current state $\boldsymbol{w}, \boldsymbol{z}, \boldsymbol{c}$ to a new state $\tilde{\boldsymbol{w}}, \tilde{\boldsymbol{z}}, \tilde{\boldsymbol{c}}$ and computes a Metropolis-Hastings acceptance ratio for the proposal, leading to acceptance or rejection of the new state. For the sake of simplicity, proposals for a new state of $\boldsymbol{w}, \boldsymbol{z}$ may be made separately from those of $\boldsymbol{c}$. For the latter, the most natural proposal exactly results in collapsed Gibbs sampling of each c_i, see *e.g.* [Murphy 2012][1]. For $\boldsymbol{w}, \boldsymbol{z}$ we design basic moves that – when combined – allow to add, remove or switch active bases as well as update several components of $\boldsymbol{w}$. These basic moves are combined to craft proposal distributions $Q(\boldsymbol{w}, \boldsymbol{z} \to \tilde{\boldsymbol{w}}, \tilde{\boldsymbol{z}})$ for which the probability of a move $\boldsymbol{w}, \boldsymbol{z} \to \tilde{\boldsymbol{w}}, \tilde{\boldsymbol{z}}$ has direct symmetries with that of the reverse move $\tilde{\boldsymbol{w}}, \tilde{\boldsymbol{z}} \to \boldsymbol{w}, \boldsymbol{z}$, so that the acceptance ratio

$$\min\left(1, \frac{p(\tilde{\boldsymbol{w}}, \tilde{\boldsymbol{z}}, \boldsymbol{c}|\mathcal{D}, \mathcal{H})}{p(\boldsymbol{w}, \boldsymbol{z}, \boldsymbol{c}|\mathcal{D}, \mathcal{H})} \frac{Q(\tilde{\boldsymbol{w}}, \tilde{\boldsymbol{z}}, \boldsymbol{c} \to \boldsymbol{w}, \boldsymbol{z}, \boldsymbol{c})}{Q(\boldsymbol{w}, \boldsymbol{z}, \boldsymbol{c} \to \tilde{\boldsymbol{w}}, \tilde{\boldsymbol{z}}, \boldsymbol{c})}\right) \tag{4.15}$$

becomes particularly straightforward to compute. The basic moves are:

a) Basis removal. For a basis k such that $z_k = 1$, set $\tilde{z}_k = 0$ and $\tilde{w}_k = 0$. The symmetric move is the basis addition.

b) Component-wise update. For a basis k such that $z_k = 1$, propose a new $\tilde{w}_k \sim q_k(w_k \to \tilde{w}_k)$ according to Eq. (4.13), (4.14) with a fixed $0 \leq r_{\text{HMALA}} \leq 1$. This move is its own symmetric (using the reverse update).

[1] A complete and concise summary of the relevant derivations and schemes is given in http://www.kamperh.com/notes/kamper_bayesgmm13.pdf

Algorithm 2: Proposal $Q_k(\boldsymbol{w}, \boldsymbol{z}, \text{reverse\_traversal} \to \tilde{\boldsymbol{w}}, \tilde{\boldsymbol{z}}, \text{reverse\_traversal}^*)$.

n_{neighb} is an integer fixed in advance.
Set $\tilde{\boldsymbol{w}} = \boldsymbol{w}$, $\tilde{\boldsymbol{z}} = \boldsymbol{z}$.

Draw one of 3 competing events: ***on-off***, ***exchange***, ***update***.

if ϕ_k **inactive** *and* **update then**
  Exit. No action to implement (as $\tilde{w}_k = 0$).

if exchange *and* ϕ_k **active then**
  Draw an inactive basis ϕ_{k^*} to replace ϕ_k. A proposal that favours well-aligned bases is designed.

else if exchange *and* ϕ_k **inactive then**
  Draw an active basis ϕ_{k^*} to replace ϕ_k. A proposal that favours well-aligned bases is designed.

if ϕ_k **active** *and* **on-off** *or* **exchange then**
  Set $\tilde{z}_k = 0$ and $\tilde{w}_k = 0$.

if on-off *or* **update then**

- For ***update***: set $\mathcal{I} = \{k\}$.
- For ***on-off***: Set $\mathcal{I} \subset \mathcal{S} \backslash \{k\}$ to a list of n_{neighb} active bases, favouring bases well-aligned with ϕ_k.
- If reverse_traversal $= 1$, reverse the ordering of $\mathcal{I}$.

for $l \in \mathcal{I}$ **do**
  $\tilde{w}_l^{\text{new}} \sim q_l(\tilde{\boldsymbol{w}} \to \tilde{\boldsymbol{w}}^{\text{new}})$ and $\tilde{w}_l = \tilde{w}_l^{\text{new}}$

if ϕ_k **inactive** *and* **on-off then**
  Set $\tilde{z}_k = 1$ and $\tilde{w}_k \sim q_k(w_k \to \cdot)$, using $r_{\text{HMALA}} = 1$.

else if ϕ_k **inactive** *and* **exchange then**
  Set $\tilde{z}_{k^*} = 1$ and $\tilde{w}_{k^*} \sim q_{k^*}(w_{k^*} \to \tilde{w}_{k^*})$ ($r_{\text{HMALA}} = 1$).

if on-off *or* **exchange then**
  Switch the state of the binary variable reverse_traversal.

c) Basis addition. For a basis k such that $z_k = 0$, set $\tilde{z}_k = 1$ and propose a new $\tilde{w}_k$ according to Eq. (4.13), (4.14) with $r_{\text{HMALA}} = 1$. The symmetric move is the basis removal.

The family of proposals $Q_k(\cdot)$ that we design combines these basic moves in such a way that when reversed, the sequence of moves induced by the proposal $Q_k(\boldsymbol{w}, \boldsymbol{z}, \boldsymbol{c} \to \tilde{\boldsymbol{w}}, \tilde{\boldsymbol{z}}, \boldsymbol{c})$ coincides exactly with the sequence of moves induced by $Q_k(\tilde{\boldsymbol{w}}, \tilde{\boldsymbol{z}}, \boldsymbol{c} \to \boldsymbol{w}, \boldsymbol{z}, \boldsymbol{c})$. The proposal and reverse proposal travel along the same path in opposite directions, drastically reducing the computational load when evaluating Eq. (4.15). Each proposal Q_k revolves primarily around the corresponding basis ϕ_k and is defined as per Algorithm 2 (where we introduced a binary variable reverse_traversal to address technicalities). Using Q_k, we define a transition kernel P_k conventionally: given the current state $\boldsymbol{w}_t, \boldsymbol{z}_t, \boldsymbol{c}_t$, we propose a new state $\tilde{\boldsymbol{c}} = \boldsymbol{c}_t$, $\tilde{\boldsymbol{w}}, \tilde{\boldsymbol{z}} \sim Q_k(\boldsymbol{w}_t, \boldsymbol{z}_t \to \cdot)$. The state is accepted with probability given by Eq. (4.15), in which case we set $(\boldsymbol{w}_{t+1}, \boldsymbol{z}_{t+1}, \boldsymbol{c}_{t+1}) = (\tilde{\boldsymbol{w}}, \tilde{\boldsymbol{z}}, \tilde{\boldsymbol{c}})$; otherwise we stay at the current state and $(\boldsymbol{w}_{t+1}, \boldsymbol{z}_{t+1}, \boldsymbol{c}_{t+1}) = (\boldsymbol{w}_t, \boldsymbol{z}_t, \boldsymbol{c}_t)$. Computation of the acceptance ratio is relatively straightforward by construction, since the ratio of posterior probabilities

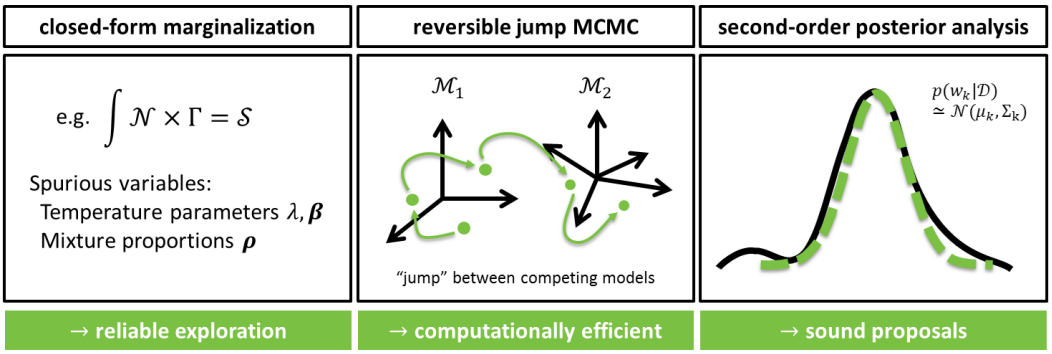


Figure 4.3: Highlight and rationale for the main constituents of the MCMC scheme.

involved in Eq. (4.15) can be rewritten as:

$$\frac{p(\mathcal{D}|\tilde{\boldsymbol{w}}, \tilde{\boldsymbol{z}}, \boldsymbol{c}, \mathcal{H})}{p(\mathcal{D}|\boldsymbol{w}, \boldsymbol{z}, \boldsymbol{c}, \mathcal{H})} \cdot \frac{p(\tilde{\boldsymbol{w}}|\tilde{\boldsymbol{z}}, \boldsymbol{c}, \mathcal{H}) p(\tilde{\boldsymbol{z}}|\mathcal{H})}{p(\boldsymbol{w}|\boldsymbol{z}, \boldsymbol{c}, \mathcal{H}) p(\boldsymbol{z}|\mathcal{H})} \tag{4.16}$$

The leftmost factor is a ratio of likelihoods and need only be evaluated once for a proposed transition. As the denominator is known from the previous iteration, only the numerator need be evaluated. In the context of registration, this part corresponds to the image term and would involve costly computations if evaluated repeatedly. Note also that for basis functions with compact support (or approximately so), only part of the image term need be updated to evaluate the ratio. The ratio on the right-hand side and the ratio of proposals are simply decomposed over the sequence of previously defined basic moves, then efficiently evaluated using Eq. (4.6), (4.13), (4.14) and expressions similar to Eq. (A.39). For the latter, statistics κ_k are kept up to date (for all bases) using efficient rank one updates derived in [Le Folgoc 2015b]. Alternatively, the necessary statistic κ_k can be recomputed from scratch only for the bases under consideration. This is usually much more efficient (cf. algorithmic complexity in 4.2.3.5).

Each transition kernel P_k satisfies a detailed balance condition. In terms of these transition kernels, the MCMC chain proceeds as follows. Random variables $k_1, k_2, \ldots$ taking values in $\{1, 2, \ldots, M\}$ are chosen according to some scheme and the corresponding transition kernel P_{k_t} is used at time t. Conventional schemes include the random-scan, where the $\{k_t\}$ are i.i.d uniform, and the deterministic scan that cycles through $\{1, 2, \ldots, M\}$ in natural order (see *e.g.* [Roberts 2006]). For the random scan, the global transition kernel also satisfies detailed balance conditions. For both schemes, the MCMC chain has stationary distribution $p(\boldsymbol{w}, \boldsymbol{z}, \boldsymbol{c}|\mathcal{D}, \mathcal{H})$ after incorporating collapsed Gibbs updates of $\boldsymbol{c}$. Highlights of the MCMC scheme main constituents are summarized in Fig. 4.3.

4.2.3.4 Markov chain mixing improvement

Similarly to Gibbs sampling of temperature parameters, Gibbs sampling of voxel GMM assignments $\boldsymbol{c}$ within updates separated from those of $\boldsymbol{w}, \boldsymbol{z}$ potentially hampers the mixing of the Markov chain for any finite, practical duration of the MCMC run. If at any point in time, a data point that should be regarded as an outlier (*e.g.* an image artifact), or a group of

such points, is assigned to a 'non-outlier' mixture component, the disjoint sampling generally causes the chain to remain stuck in the vicinity of the corresponding local mode of the posterior $p(\boldsymbol{w}, \boldsymbol{z}, \boldsymbol{c}|\mathcal{D}, \mathcal{H})$: the desired reverse assignment move virtually occurs with probability zero after readjustment of $\boldsymbol{w}, \boldsymbol{z}$. This defect is critical as such failure scenarii happen with overwhelming probability. Fortunately, joint proposals for $\boldsymbol{w}, \boldsymbol{z}, \boldsymbol{c}$ can be designed at little cost, even more so after noting that the component-wise proposals for w_k (Eq. (4.13), (4.14)) and z_k only *indirectly* depend on $\boldsymbol{c}$. The transition $Q_k(\boldsymbol{w}, \boldsymbol{z}, \boldsymbol{c} \to \tilde{\boldsymbol{w}}, \tilde{\boldsymbol{z}}, \tilde{\boldsymbol{c}})$ proceeds in two steps. First, $\tilde{\boldsymbol{w}}, \tilde{\boldsymbol{z}}$ is proposed as per Algorithm 2. Then, $\tilde{\boldsymbol{c}}$ is sampled by component-wise collapsed Gibbs sampling of each $\tilde{c}_i \sim p(\tilde{c}_i \,|\, \tilde{c}_{j<i}, c_{j>i}, \tilde{\boldsymbol{w}}, \mathcal{D}, \mathcal{H})$ in turn. For efficiency, only the subset of voxels in the support of updated basis functions is sampled, and voxel assignments are updated only once in case of overlapping supports. The two-step move is accepted or rejected based on the acceptance ratio (4.15), replacing $\boldsymbol{c}$ by $\tilde{\boldsymbol{c}}$ where necessary. The order of voxel traversal is reversed according to the state of reverse_traversal. Sampling $\tilde{\boldsymbol{c}}$ and computing its contribution to the acceptance ratio exclusively involves the residuals r_i and $\tilde{r}_i$ of updated voxels prior and after the update $\boldsymbol{w}, \boldsymbol{z} \to \tilde{\boldsymbol{w}}, \tilde{\boldsymbol{z}}$, which were already required to compute the likelihood change in Eq. (4.16).

4.2.3.5 Algorithmic complexity

The algorithmic complexity associated to a transition kernel P_k (proposal and acceptance-reject) is $\mathcal{O}(|\mathcal{S}| \cdot |\mathcal{I}_+| + \sum_{l \in \mathcal{I}_+} V_l + L\, C)$, noting $\mathcal{I}_+$ the set of updated bases, V_l the number of voxels in the support of basis ϕ_l and C the number of voxels whose assignments $\tilde{c}_i$ are resampled. The first term includes part of the cost of the proposal $\boldsymbol{w}, \boldsymbol{z} \to \tilde{\boldsymbol{w}}, \tilde{\boldsymbol{z}}$ and its impact on the ratio of prior probabilities. The second term is replicated three times and can be heavily parallelized in each case: once to compute $\mu^l_{\text{pos}}, \Sigma_l$ in Eq. (4.13), (4.14) for $l \in \mathcal{I}_+$, twice to evaluate and store differences in the displacement fields (resp. residual images) over the support of basis functions in $\mathcal{I}_+$ following their update. The last term accounts for all computations related to resampled voxel GMM assignments $\tilde{\boldsymbol{c}}$. When a move that involves the inclusion or removal of a basis function from the active set is accepted, an additional $\mathcal{O}(|\mathcal{S}|^2 + M \cdot |\mathcal{S}|)$ cost is involved to maintain statistics κ_k over all bases in the dictionary, with the right-hand term being parallelizable into M disjoint $\mathcal{O}(|\mathcal{S}|)$ operations. The $\mathcal{O}(M \cdot |\mathcal{S}|)$ cost upon inclusion or deletion of a basis can be replaced by a $\mathcal{O}(|\mathcal{S}|^2)$ cost per proposed move, which is usually more efficient.

4.2.3.6 Initialization

The chain is initialized from the output of the deterministic algorithm presented in [Le Folgoc 2015b] which progresses greedily in the space of parameters $\{\boldsymbol{z}, \lambda, P\}$ towards a local maximum of their joint posterior. We comment, however, that any registration algorithm could reasonably be used to initialize the chain.

4.3 Predictive Uncertainties: Marginal Likelihood Maximization vs. Exact Inference

The 'sparse Bayesian' model presented in Fig. 4.1 is inspired by the Spike-&-Slab model of Mitchell and Beauchamp [Mitchell 1988] and the Relevance Vector Machine (RVM) proposed by Tipping [Tipping 2001] for tasks of regression and classification. In the latter work, the author approaches the problem of inferring an optimal sparse regression function from the standpoint of Automatic Relevance Determination (ARD). Point estimates of the hyperparameters that govern basis selection (and in fact of all hyperparameters) are sought in a first step by maximizing the marginal likelihood or *evidence* as per Eq. (4.17):

$$\begin{aligned} \boldsymbol{\theta}^* &= \arg\max_{\boldsymbol{\theta}} p(\mathcal{D}|\boldsymbol{\theta},\mathcal{H}) \\ &= \arg\max_{\boldsymbol{\theta}} \int_{\boldsymbol{w}} p(\mathcal{D}|\boldsymbol{w},\boldsymbol{\theta},\mathcal{H}) p(\boldsymbol{w}|\boldsymbol{\theta},\mathcal{H}) d\boldsymbol{w} \end{aligned} \tag{4.17}$$

where $\boldsymbol{\theta} = \{\boldsymbol{z}, P, \lambda\}$ using our notations. If non-uniform, proper hyperpriors on $\boldsymbol{\theta}$ are assumed, $\boldsymbol{\theta}^*$ maximizes the posterior $p(\boldsymbol{\theta}|\mathcal{D},\mathcal{H}) \propto p(\mathcal{D}|\boldsymbol{\theta},\mathcal{H})p(\boldsymbol{\theta}|\mathcal{H})$ instead. In a second step, the distribution of weights w_k is characterized conditionally to the selected model,

$$p(\boldsymbol{w}|\mathcal{D},\mathcal{H}) \approx p(\boldsymbol{w}|\boldsymbol{\theta}^*,\mathcal{D},\mathcal{H})\,. \tag{4.18}$$

This strategy is typically successful in reaching strongly sparse solutions with good predictive power but, above all else, is motivated by its computational efficiency. Dedicated schemes relying on linear algebra and rank one updates make it possible to efficiently, iteratively build the set $|\mathcal{S}|$ of relevant basis functions ϕ_k from scratch. See for instance [Tipping 2003], and [Le Folgoc 2015b] for an extension to the wider family of priors required for registration tasks. The approximation of Eq. (4.18) is justified by observing that the full posterior $p(\boldsymbol{w}|\mathcal{D},\mathcal{H})$ is obtained by summing over all conditional posteriors $p(\boldsymbol{w}|\boldsymbol{\theta},\mathcal{D},\mathcal{H})$, conditioned on the value $\boldsymbol{\theta}$, weighted by the posterior probability $p(\boldsymbol{\theta}|\mathcal{D},\mathcal{H})$ for this value:

$$p(\boldsymbol{w}|\mathcal{D},\mathcal{H}) = \int_{\boldsymbol{\theta}} p(\boldsymbol{w}|\boldsymbol{\theta},\mathcal{D},\mathcal{H}) p(\boldsymbol{\theta}|\mathcal{D},\mathcal{H}) d\boldsymbol{\theta} \tag{4.19}$$

Now if the available data $\mathcal{D}$ is informative enough, $p(\boldsymbol{\theta}|\mathcal{D},\mathcal{H})$ will be sharply peaked around its mode(s). In the limit case where $p(\boldsymbol{\theta}|\mathcal{D},\mathcal{H})$ is a Dirac centered at its single mode $\boldsymbol{\theta}^*$, Eq. 4.18 is retrieved exactly, and the two-step scheme outlined in Eq. (4.17), (4.18) is justified. Moreover in the case of sparsity governing parameters $\boldsymbol{z} = \begin{pmatrix} z_1 & \cdots & z_M \end{pmatrix}^\intercal$, Tipping [Tipping 2001] argues that, even if several combinations of parameters are highly probable due to the presence of redundant functions ϕ_k in the dictionary of bases, they should roughly lead to the same optimal solution u^* and an approximate mode (or the expectation) of $p(\mathrm{u}|\mathcal{D},\mathcal{H})$ should still be correctly evaluated. Regardless, we now demonstrate why this evidence-based approximation will typically fail to properly approximate higher order moments of the full posterior, resulting for instance in poor approximation of the real predictive uncertainty. There are two main breakdown situations for the evidence-based approximation of the full posterior assumed in Eq. 4.18.

Firstly in absence of data, the assumption that the posterior distribution $p(\boldsymbol{\theta}|\mathcal{D},\mathcal{H})$ of hyperparameters is well approximated by a Dirac collapses. Indeed the posterior then

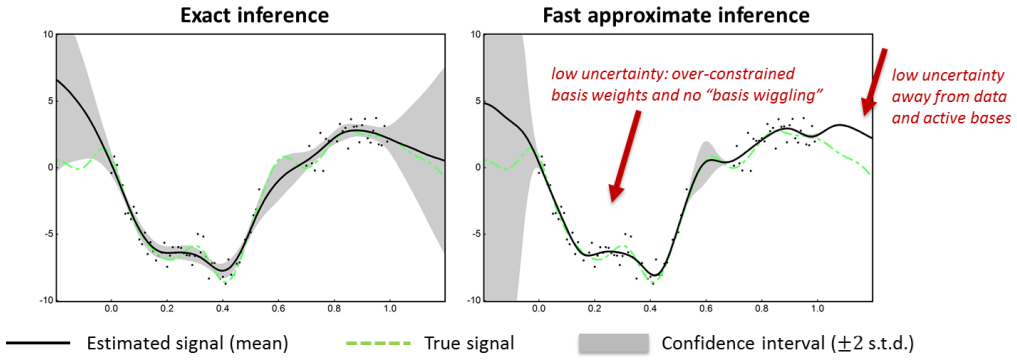


Figure 4.4: Comparison of approximate evidence-based inference and faithful MCMC inference for the sparse Bayesian model, on a $1D$ regression task. Data points (black dots) are sampled with additive $i.i.d.$ Gaussian noise from the true signal (dashed green line). The consistence of the fast and faithful estimates of the regressor function (black lines) is satisfactory (w.r.t. uncertainty levels), even more so in the presence of data. Esimates of uncertainty (grey ribbon), however, can be inconsistent.

resembles the prior distribution $p(\boldsymbol{\theta}|\mathcal{H})$, which is typically flat. This scenario is apropos in the case of basis selection parameters z_k, since associated basis functions ϕ_k have a local support over which reliable data may be missing. Away from data and without strong incentive to include the basis to increase the deformation regularity, the probability of basis inclusion (resp. exclusion) is π_k (resp. $1 - \pi_k$), and for neutral values of π_k, the choice of excluding the basis is arbitrary.

Secondly and even in presence of data, many combinations of active bases could have quasi-identical probability. When using radial basis functions for instance, the location of basis centers can be slightly perturbed without significantly affecting the posterior probability of the new configuration. The optimal value of basis weights $\boldsymbol{w}$ under two such perturbations will slightly differ however, as well as the resulting transformation Ψ. The evidence-based approximation of Eq. (4.18) relies on a single – perhaps only marginally superior – configuration, whereas the true posterior sums over all such configurations, as seen from Eq. (4.19). As it turns out, 'basis wiggling' accounts for a significant part of the uncertainty.

4.4 Preliminary Experiments and Results

The following experiments aim to qualitatively evaluate the consistency of approximate evidence-based and faithful MCMC inference, as well as provide insight into the MCMC inference scheme, starting again from the example $2D$ registration of Chapter 3 as shown on Fig. 4.5.

For the approximate-based inference, the methodology of Chapter 3 is used without change. The multiscale dictionary is identical, using Gaussian RBFs at three different scales (isotropic, $\sigma = 6mm$, $12mm$ and $24mm$). For the MCMC scheme, the chain was

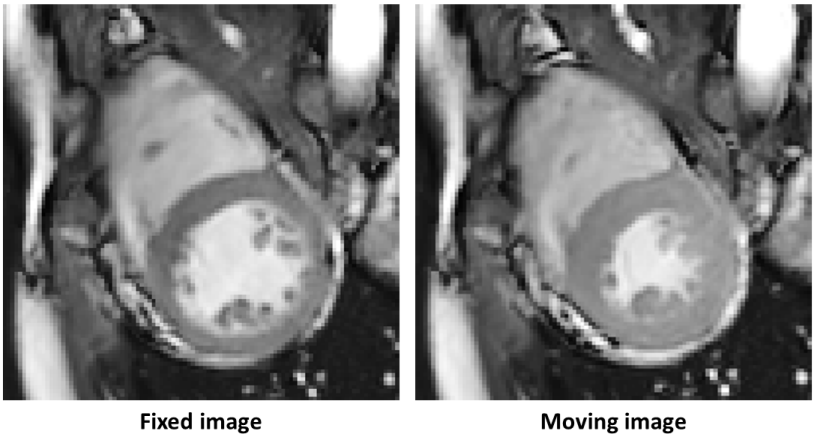


Figure 4.5: $2D$ registration setting: fixed and moving images.

run for roughly $7 \cdot 10^5$ transitions and 500 samples were regularly extracted. Approximately $7 \cdot 10^4$ additional samples were discarded as part of the burn-in phase, during which the parameters of the proposal distribution were fine-tuned (cf. section 4.2.3.2). This tuning relies on a set of sufficient statistics, such as the average energy and the average voxel-wise square intensity residuals per sample. The averages are computed using a scheme that downweights the early samples, typically using a first phase of fixed learning rate before reverting to a classical (inverse linear) weighting, drawing inspiration from the SAEM scheme of *e.g.* [Richard 2009]. The free parameter s controlling the spread of proposals compared to the second-order approximation of the posterior (section 4.2.3.2) was set to 1 (spread unchanged). The observed acceptance rate varies between 20–45% under reasonable variations of the experimental setting, and between 27–34% during the run of interest under the settings described above. Examples of samples are reported in Fig. 4.6. As an order of magnitude, the run takes 20 minutes on a standard laptop with a naive implementation.

Fig. 4.7 reports the mean displacement reported respectively by the evidence-based inference scheme and by the MCMC inference scheme. As anticipated from the discussion of section 4.3, very good agreement between the evidence-based and MCMC-based esti-

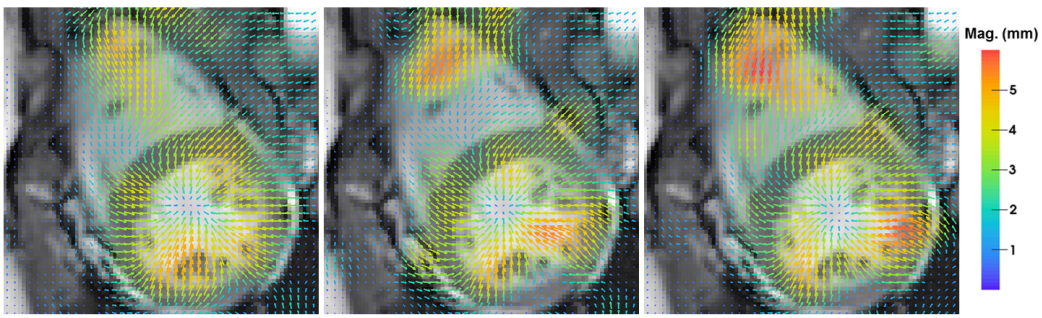


Figure 4.6: Examples of samples returned by the MCMC run.

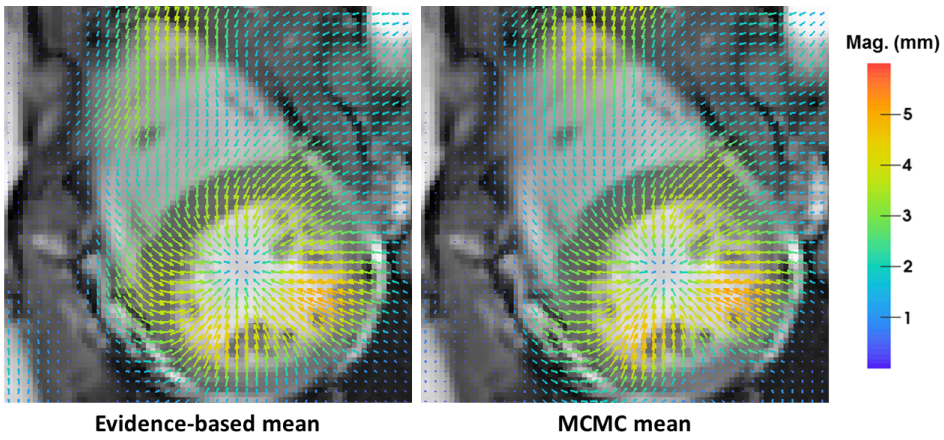


Figure 4.7: Comparison of the approximate evidence-based posterior mean displacement vs. the MCMC-based posterior mean.

mates of the displacement is observed. Upon close inspection, minor differences are noted in some areas with flat intensity profiles or otherwise low confidence (such as that resulting from artefacts, or disagreeing intensities in the fixed and moving image). Their magnitude is lower than the level of uncertainty in the output of registration, as estimated from the MCMC scheme.

Fig. 4.8 reports the estimates of uncertainty obtained from the MCMC characterization of the posterior. Any relevant statistics can be estimated empirically from the set of samples returned by the run. To study the spatial localization of uncertainty, we visualize at each voxel center x_i the 2×2 empirical covariance matrix of the posterior distribution $p(u(x_i)|I, J, \mathcal{H})$ of the corresponding displacement vector $u(x_i)$. This is reasonable under the assumption that the posterior on displacements is approximately mono-modal and Gaussian. The voxelwise empirical covariance matrix, or its square root (homogeneous to a standard deviation), can be visualized as a $2D$ tensor that encodes uncertainty at this point along any direction. Fig. 4.8 displays the resulting tensor map (Left) and a scalar summary (Right).

The order of magnitude of reported uncertainties (typically $\sim 1mm$ for a 95% confidence interval) is consistent with both the magnitude of the underlying motion (no more than $5mm$, see fig. 4.7) and the resolution (voxel dimensions: $1.25mm \times 1.25mm$). As expected, uncertainty is higher in regions with little structured content (no intensity gradients) and in the direction of contours.

Finally, Fig. 4.9 demonstrates the benefit of a careful design of the Markov chain. The left-most figure displays the estimated mean displacement, under the afore-mentioned experimental setting, if moves in the space of transformation parameters are done separately from the resampling of voxelwise assignments to components of the noise mixture instead of jointly (right-most figure). In this example, a local discrepancy in the intensity profiles of the fixed and moving images induces a spurious maximum in the joint posterior distribution of transformation parameters and voxel labels (cf. section 4.2.3.4). A systematic

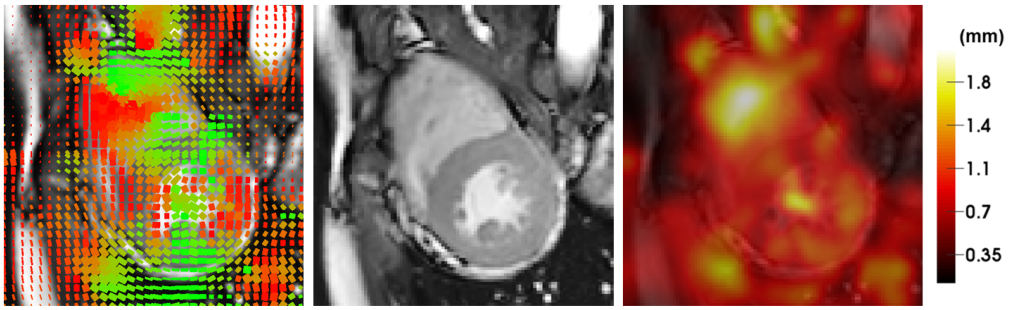


Figure 4.8: Estimates of uncertainty obtained by characterizing the posterior distribution of the sparse Bayesian model by MCMC sampling. (Left) Tensor visualization of the displacement uncertainty: each tensor encodes the square root of the empirical 2×2 covariance displacement matrix at this location. Color scheme: direction of first eigenvector. (Middle) Moving image (Right) Trace of the square root covariance.

and indefinite drift towards this mode was observed in all runs where the sampling was performed in an alternated manner, whereas systematic recovery was observed under the improved scheme. Similar observations were made in preliminary experiments where temperature parameters were treated by Gibbs sampling instead of analytically marginalized over.

4.5 Discussion and Conclusions

In this chapter we explored the properties of the proposed sparse Bayesian model of registration for the purpose of uncertainty quantification. We emphasize the distinction between the Bayesian model itself and inference schemes used to characterize the distributions of interest within this model. In particular we presented in the previous chapter a greedy evidence-based approximate inference scheme. The question underlying the present work, motivated by [Quiñonero-Candela 2005, Rasmussen 2005], is twofold. Does this approximate inference procedure provide consistent estimates of the moments of the true posterior distribution of displacements, specifically of the expectation and variance? And perhaps more fundamentally, are the estimates of displacement and uncertainty provided by the true sparse model useful and sound?

To this aim, we proposed a reversible jump (transdimensional) MCMC scheme for the systematic, quasi-exact characterization of the posterior distribution. Special care was taken in the design of the chain to ensure proper mixing: nuisance variables such as temperature parameters (noise and regularization levels) and mixture parameters were analytically marginalized over, while transformation parameters and voxel assignments to mixture components were jointly sampled. This was shown to prevent the chain from indefinitely drifting towards poor local maxima of the posterior on examples of practical relevance. This work hints at the good empirical correspondence between the optimal estimates of

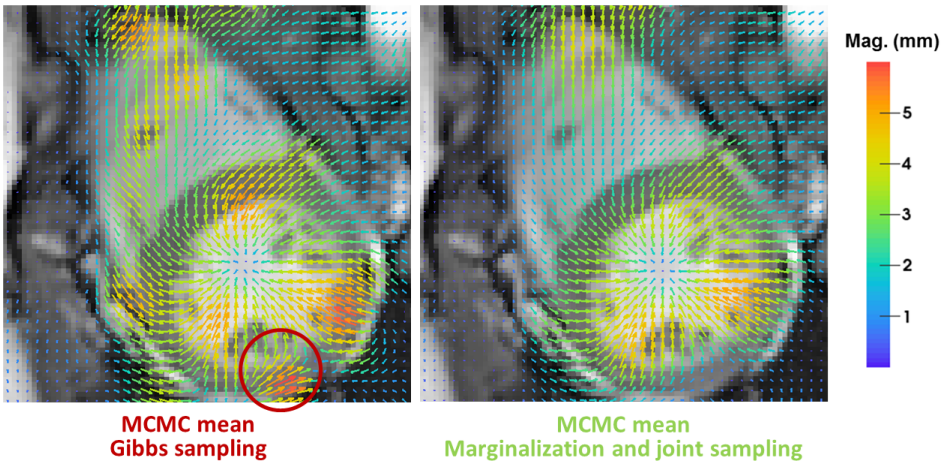


Figure 4.9: Comparison of estimates of the posterior mean returned by MCMC characterization. (Left) Alternated sampling of on the one hand, activation variables $\boldsymbol{z}$ and corresponding weights $\boldsymbol{w}$, on the other hand voxel mixture assignments $\boldsymbol{c}$. (Right) Joint sampling, as per the approach proposed in section 4.2.3.4.

displacement returned by the approximate and exact inference schemes, particularly in the presence of informative data. On the other hand, it evidences limitations of the approximate evidence-based scheme for uncertainty quantification even in simplified experimental settings. This observation is supported by simple insight on the underlying approximations made by the evidence-based scheme, and how it proceeds to select the active parametrization.

Despite limitations of the approximate scheme, preliminary experiments hint at the soundness of the true uncertainty estimates (as characterized from the MCMC analysis). Of course, a major factor affecting the validity of uncertainty estimates is the soundness of the model assumptions. Various model biases can plague the quality of uncertainty estimates in the context of registration, such as the assumed model of interpolation of image intensities and the inadequacy of the noise or prior models. A thorough, application-driven evaluation of the quantified uncertainty, assessing the respective impacts of these factors should be conducted: so far, this work contributes by providing the methodological framework and schemes to do so. As a methodological perspective, the Bayesian framework and the proposed MCMC inference scheme in fact allow to relax even further both the prior assumptions (e.g. allowing the structure of the quadratic prior to be learned) and the observation model.

To address limitations of the evidence-based scheme, approximate inference schemes that correct identified causes of breakdown can be devised, for instance by keeping track of several sets of relevant explanatory variables via tree structures [Schniter 2008]. In fact, we note instead the satisfactory computational complexity of the proposed MCMC approach and the high potential for parallelization, which could make it usable in the research routine even on real clinical data. Moreover classic techniques used to accelerate registration optimization schemes (subsampling of voxels, sensible scanning of the multi-

scale parametrization) remain applicable in the proposed framework, with some care. In the proposed MCMC approach, we did not make use of full-dimensional moves over the space of transformation parameters. Calibrating such transitions calls for the particularly expensive computation of large (non-diagonal) Hessian matrices, which render them inefficient unless e.g., exploiting dedicated procedures inspired from limited memory quasi-Newton methods [Zhang 2011]. Such moves could be incorporated in the proposed scheme in the future, although we found component-wise transitions to be particularly suitable, provided that the set of active bases must be jointly explored.

CHAPTER 5

Conclusion and Perspectives

Contents

This thesis was motivated by the following questions. Can we learn cardiac characteristics, such as parameter values governing constitutive mechanics, by observing the cardiac motion? Can statistical learning make the process faster or more reliable? The question in all its generality remains broad in scope, but this work brings partial answers, insight and methodological contributions to approach the issue.

This thesis started with a preliminary machine learning approach for personalization of cardiac mechanics from image-based $3D + t$ data. The work evidenced high uncertainty in the estimation of mechanical parameters, with sources of uncertainty typically ranging from lack of parameter identifiability when personalizing from purely kinematic data, to model biases, to uncertainty and errors in the preprocessing steps of segmentation and motion tracking. We then focused on the development of novel tools for motion tracking, and more generally image registration, enabling uncertainty quantification. The work explored possibilities offered by Bayesian modelling to address open limitations of classical optimization-based registration schemes. We demonstrated how sparse Bayesian modelling could result in greater automation of the registration task and could render possible the use of adaptative observation and prior models (a.k.a. image similarity and regularizing priors). We provided methodological tools for, and insight into the problem of quantifying uncertainty in registration, and specifically under the sparse Bayesian registration model. The computational tractability of the proposed Bayesian model was demonstrated on a range of modalities on real $3D$ and $4D$ cardiac data.

5.1 Personalization of cardiac mechanics: contribution & perspectives

In [Le Folgoc 2012], we proposed a preliminary machine learning framework for personalization of cardiac mechanics from image-based data. The space of model parameters is systematically sampled at a training stage: simulations are run for each parameter set,

and the resulting training set is used to learn a 'causal' regressor between parameter and observation spaces. The observation space consists in time sequences of $3D$ meshes, encapsulating shape and motion. This is consistent with the output of the mechanical model and, at test time, such sequences can be obtained once the kinematics have been extracted from images by motion tracking. At test time, parameters are optimized so as to minimize the discrepancy between simulation and observation in the shape and motion space. To alleviate the computational cost, an intermediate reduced observation space is introduced: a low dimensional descriptor of the shape and motion serves as surrogate for the mesh sequences, and allows for fast optimization, even with complex optimizers (e.g. simulated annealing). We proposed to derive the low-dimensional descriptor by linear dimensionality reduction in a Hilbert space of smooth $4D$ currents, exploiting the advantageous space structure to guarantee efficient computations and bypass the need for (node-to-node pairwise) mesh correspondences in the training set and at test time. The framework is highly parallelizable during the training stage and fast at test time.

The work evidences the issue of parameter identifiability even in simplified synthetic experimental settings. While state-of-art sequential optimization techniques [Moireau 2011, Imperiale 2011, Xi 2011, Chabiniok 2012, Marchesseau 2013b] give a partial access to uncertainty in the estimation, a dedicated probabilistic framework would let one crucially [Liu 2009a, Brynjarsdóttir 2014] account for multimodality of the distribution of probable parameters, for various sources of biases (e.g. boundary and limit conditions, modelization of mechanics, observation model) and for uncertainty in the preprocessing (segmentation, motion tracking). Bayesian treatment of the inverse problem remains so far marginal in the cardiac electromechanical personalization community except for the recent work of [Neumann 2014] and the seminal paper of [Konukoglu 2011] (applied to the personalization of the electrophysiology), although plentiful inspiration can be found from varied literature in the fields of statistics, inverse problems and computational physics, e.g. [Kennedy 2001, Arendt 2012, Stuart 2010, Kaipio 2011]. Some of the building blocks for a tractable Bayesian analysis are already present in the proposed work: moving from an optimization to a probabilistic standpoint, the simulated annealing based optimization could be replaced with dedicated MCMC schemes for exploration of the posterior distribution; the training phase along with the dimensionality reduction address the more general issue of approximating the costly forward model (the simulation and the observation model) with an inexpensive surrogate. The sampling of the parameter space required to build the approximate model may advantageously be done over sparse grids (e.g. [Ma 2009]) with proper scaling w.r.t. the dimensionality of the parameter space.

5.2 Sparse Bayesian modelling for image registration: contribution & perspectives

5.2.1 Contributions and perspectives on image registration

This work developed a probabilistic framework of non-rigid image registration that enables better automation of the registration process thanks to principled hyperparameter inference.

The strategy is successful in finding a good, objective trade-off between image similarity and regularization as in e.g. [Simpson 2012]; but it also allowed for data-driven adaptivity of both the model of intensity residuals (via a mixture model) and of the representation of deformations, seamlessly coupling a coarse parametrization to ensure proper extrapolation of the structured image content to textureless, flat-intensity regions, and a finer parametrization to capture local patterns of the observed motion in regions of high confidence. To our knowledge, it is the first time spatial adaptivity of the parametrization of transformations was attempted on probabilistic grounds. In our case, this was made easier by the recourse to specific sparsity-inducing priors with straightforward probabilistic interpretations. In experiments on clinical data we retained the benefits of fixed, coarsely parametrized FFD (Free Form Deformation) models w.r.t. regularity of the inferred motion, while typically maintaining state of the art accuracy. In the context of cardiac motion tracking this may give opportunities to consistently capture subtle temporal patterns of asynchronous contraction such as septal flashes. We also note that the approach was completely generic in terms of representation of the transformation, exploiting little prior knowledge on the specific application to cardiac imaging – and specifically no segmentation of the cardiac muscle. This proved to be convenient as cardiac segmentation remains a challenging task, but dedicated prior knowledge or constraints on the parametric representation of transformations could of course be seamslessly integrated into the model.

As a perspective, moving from a small displacement framework towards a model of large diffeomorphic deformation (as in [Arsigny 2006, Ashburner 2007, Arsigny 2003] or [Miller 2002, Beg 2005, Durrleman 2007, Zhang 2013, Sommer 2013]) could be extremely fruitful in the context of intersubject registration, or longitudinal analysis with poor temporal resolution. In addition, the framework for basis selection developed in this work for pairwise registration could be directly extended to derive a common parametric atlas of deformations in group-wise analyses. In the present work, the adaptive parametrization was a means to achieve compacity of representation with gains in terms of computational tractability, in terms of accuracy and smoothness of the transformation. However, if performed at a group scale, the reduced representation could become intrinsically related to the trends and variability in the population, and open avenues for statistical analysis. A final aspect left open by our work regarding image registration is the derivation of a statistically coherent, yet efficient, model of images with particular care taken into the modelling of spatial correlations. Spatial correlations of intensity residuals were ignored in the generative model we proposed, and the over-confidence on data stemming from this approximation was corrected on an ad-hoc basis by introducing a joint scheme for virtual decimation of voxels. While partial solutions to model correlations are brought to the table by Fourier analysis in the case of spatially homogeneous noise, the difficulty lies in accounting at the same time for non-stationarity of the noise patterns in images.

5.2.2 Contributions and perspectives on statistical learning

Although specifically applied to registration problems, the methodological tools that we put in place have a broader scope of application to regression, classification and unsuper-

vised learning tasks. While it is somewhat difficult here to have a thorough vision of the machine learning and statistical literature, a fully generic extension of the fast marginal likelihood computation scheme of [Tipping 2003] that handles the incorporation of generic quadratic priors (in addition to the sparsity-inducing terms) is proposed here for the first time, to our knowledge. We note that the extended prior model itself was also proposed for image-based tasks of classification and regression under the Relevance Voxel Machine of [Sabuncu 2012], where inference followed the earlier guidelines of [Tipping 2001], later accelerated for specific priors exploiting sparsely connected graphs [Ganz 2013]. Alternatively, the scheme developed in [Le Folgoc 2014, Le Folgoc 2015b] can be seen as a fast greedy inference scheme for a Spike-&-Slab prior model with arbitrary structure of dependency between explanatory variables (explanatory variables are typically assumed to be a priori independent, e.g. in [Mitchell 1988]). Relevant applications for the proposed model and inference span a range of probabilistic regression and classification tasks where some benefit can be expected from enforcing a given form of regularity or connectivity.

Secondly we explored the properties of the proposed model for the purpose of uncertainty quantification. In that respect, the sparse Bayesian model of [Tipping 2001, Bishop 2000] was previously reinterpreted in the work of [Candela 2004] from the standpoint of sparse Gaussian processes. Their work evidenced limitations of the RVM approximate evidence-based inference [Quiñonero-Candela 2005, Rasmussen 2005], which questioned its ability to gauge adequate confidence intervals. In [Le Folgoc 2015a], our work clarifies the respective responsibilities of the inference scheme and of the model. We provide theoretical arguments and empirical evidence for the good behaviour of the sparse Bayesian model itself in terms of uncertainty quantification, although indeed revealing limitations of approximate evidence-based inference for that purpose. We contribute by proposing a feasible transdimensional Markov Chain Monte Carlo scheme for exploration of the parameter space, exploiting insight into the model to avoid standard impediments of MCMC implementations. On the applicative side, extensive evaluation of uncertainty estimates still needs to be conducted, with specific targets to judge its usefulness from a clinical standpoint. On the methodological side, the possibilities offered by statistical modelling and analysis could be further explored, moving from Bayesian modelling and inference towards Bayesian (or non-Bayesian) analysis of modelling assumptions and assessment of their soundness. On a shorter term, our Bayesian framework and MCMC inference opens up avenues for tractable model comparison and would allow to relax even further both prior assumptions (e.g. allowing the structure of the quadratic prior to be learned) and the observation model.

APPENDIX A

Sparse Bayesian Registration: Technical Appendices

Contents

A.1 Closed form regularizers for Gaussian reproducing kernel Hilbert spaces

Gaussian reproducing kernel Hilbert space. Given a $d \times d$ symmetric positive definite (s.p.d.) matrix S and the Gaussian kernel $K_S(x, y) = \exp -\frac{1}{2}(x-y)^\intercal S^{-1}(x-y)$, we consider the space $\mathcal{H}^d_S$ of integrable d-vector fields $f : \mathbb{R}^d \to \mathbb{R}^d$ such that $\|f\|_S < +\infty$, where

$$\|f\|^2_S = \frac{1}{(2\pi)^d} \int_\xi \|\widehat{f}(\xi)\|_2^2 \, \widehat{K}_S^{-1}(\xi) d\xi \tag{A.1}$$

involves the Fourier transform $\hat{f} = \mathcal{F}[f]$ of f, defined by $\hat{f}(\xi) = \int_{\mathrm{x}} \exp\{-i\mathrm{x}^\intercal \xi\} f(\mathrm{x}) d\mathrm{x}$. Endowed with the inner product A.2

$$\langle f | g \rangle_S = \frac{1}{(2\pi)^d} \int_\xi \widehat{f}(\xi)^\dagger \widehat{g}(\xi) \, \widehat{K}_S^{-1}(\xi) d\xi \tag{A.2}$$

$(\mathcal{H}^d_S, \langle \, | \, \rangle_S)$ is a reproducing kernel Hilbert space with attractive theoretical and algorithmic properties. Perhaps more intuitively, $\mathcal{H}^d_S$ is the completion of the space spanned by all finite combinations of $K_S(x, \cdot)\boldsymbol{\alpha}$, for $x \in \mathbb{R}^d$ and $\boldsymbol{\alpha} \in \mathbb{R}^d$.

A multiscale space. From A.1 and the properties of Gaussian kernels under Fourier transform, the atom $K_{\tilde{S}}(x, \cdot)\boldsymbol{\alpha}$ lies in $\mathcal{H}^m_S$ if and only if $\tilde{S} > S/2$, in the sense of positive definiteness. In particular, given a sequence $S_1 \leq \cdots \leq S_q$ of $d \times d$ s.p.d. matrices, their associated r.k.h.s. are nested: $\mathcal{H}^d_{S_1} \supseteq \cdots \supseteq \mathcal{H}^d_{S_q}$. This property leads to a principled framework to represent displacements in a multiscale fashion, jointly regularized at all

scales.

Closed form regularizers. The successive partial derivatives $\partial_{x^{i_1}\cdots x^{i_p}} f$ of elements $f \in \mathcal{H}_S^d$ exist and all lie in $\mathcal{H}_S^d$ [Zhou 2008]. As such, we may consider the family of regularizers of the form $\mathcal{R}_D(f) = \|Df\|_S^2$, where D is a differential operator. For any f that can be written over a finite number of atoms $K_{S_k}(x_k, \cdot)\boldsymbol{\alpha}_k$, the properties of Gaussian kernels under Fourier transform, multiplication (by Gaussian kernels) and summation yield closed form expressions for such regularizers. We illustrate this on two classic penalty terms for registration: the membrane and bending energies ($D = \nabla^s$, $s = 1, 2$). Recall that for any p-uplet of integers $i_1 \cdots i_p \in \{1, \cdots, d\}$, $\mathcal{F}[\partial_{x^{i_1}\cdots x^{i_p}} f](\xi) = j^p \xi^{i_1} \cdots \xi^{i_p}\, \mathcal{F}[f](\xi)$. Thus for any $f \in \mathcal{H}_S^d$,

$$\mathcal{F}[\nabla \cdot f](\xi) = j\xi^\intercal\, \mathcal{F}[f](\xi)\,. \tag{A.3}$$

and for any even integer s,

$$\mathcal{F}[\nabla^s f](\xi) = (j\xi)^s\, \mathcal{F}[f](\xi) \tag{A.4}$$

The Fourier transform of a Gaussian kernel is given by $\mathcal{F}[K_{S_k}(x_k, \cdot)](\xi) = |2\pi S_k|^{1/2} e^{-j\xi^\intercal x_k} \exp\{-\frac{1}{2}\xi^\intercal S_k \xi\}$, which is again Gaussian. Regrouping the exponential factors when appropriate, we derive the following expression for inner products of the form $\langle \nabla^s(K_{S_k}\boldsymbol{\alpha}_k) \,|\, \nabla^s(K_{S_l}\boldsymbol{\alpha}_l) \rangle_S$:

$$C_{S_k,S_l,d} \cdot \boldsymbol{\alpha}_k^\intercal \left\{ \int_\xi \|\xi\|^{2s} \mathcal{F}[K_{S_{k,l}}(z_{k,l}, \cdot)](\xi) d\xi \right\} \boldsymbol{\alpha}_l \tag{A.5}$$

where $C_{S_k,S_l,d} = \frac{1}{(2\pi)^d} \left(\frac{|S_k| \cdot |S_l|}{|S| \cdot |S_{k,l}|} \right)^{1/2}$ is a constant, $S_{k,l} = S_k + S_l - S$ and $z_{k,l} = x_l - x_k$. Recognizing the inverse Fourier transform of $\nabla^{2s}[K_{S_{k,l}}(\vec{0}, \cdot)]$ evaluated at $z_{k,l}$, we finally obtain that

$$\left\| \nabla^s \left(\sum\nolimits_k K_{S_k}(x_k, \cdot)\boldsymbol{\alpha}_k \right) \right\|_S^2 = \boldsymbol{\alpha}^\intercal \left[\mathbf{R}_{k,l}^s \right]_{1 \le k,l \le M} \boldsymbol{\alpha} \tag{A.6}$$

$$\mathbf{R}_{k,l}^s = \left(\frac{|S_k| \cdot |S_l|}{|S| \cdot |S_{k,l}|} \right)^{1/2} (-\Delta)^s [K_{S_{k,l}}](z_{k,l})\, \mathbf{I} \tag{A.7}$$

for s integer and $\mathbf{I}$ the $d \times d$ identity matrix. Straightforward analytical expressions of $(-\Delta)^s[K_{S_{k,l}}](z_{k,l})$ are obtained for $s = 1, 2$ by computing the derivatives of the Gaussian kernel.

A.2 Contribution of a basis to the log marginal likelihood

It follows from Eq. (3.22) that the log marginal likelihood $\mathcal{L} = \log p(\boldsymbol{t}|\mathbf{A}, \lambda, \sigma^2)$ is given up to additive constant by

$$\mathcal{L} = -\frac{1}{2} \left\{ \log |\mathbf{C}| + \boldsymbol{t}^\intercal \mathbf{C}^{-1} \boldsymbol{t} \right\} \tag{A.8}$$

with $\mathbf{C} = (\beta\mathbf{H})^{-1} + \Phi\mathbf{L}\Phi^\intercal$, where we define

$$\mathbf{L} \triangleq (\mathbf{A} + \lambda\mathbf{R})^{-1}\,. \tag{A.9}$$

Noting that $\mathbf{C}$ exclusively depends on the basis k via the kth diagonal coefficient of $\mathbf{A}$ and column of Φ, we would like to single out the contribution $l(\mathbf{A}_k)$ of any such basis ϕ_k to the global log marginal likelihood in the form:

$$\mathcal{L} = l(\mathbf{A}_k) + \mathcal{L}_{-k} \tag{A.10}$$

where $\mathcal{L}_{-k}$ does not depend on the basis k. If we denote by $\mathbf{L}_{-k}$ the inverse of the matrix obtained by removing the kth column from $\mathbf{L}^{-1} = \mathbf{A} + \lambda\mathbf{R}$ (or equivalently by setting $\mathbf{A}_k = +\infty$ in $\mathbf{L}$), we see from the Woodbury rank one matrix identity that $\mathbf{L} = \mathbf{L}_{-k} + \mathrm{U}_k\mathrm{L}_{kk}\mathrm{U}_k^\intercal$, with $\mathrm{U}_k^\intercal = \begin{pmatrix}(\lambda\mathbf{L}_{-k}\mathrm{R}_k)^\intercal & \mathrm{I}\end{pmatrix}$ and $\mathrm{L}_{kk} = (\mathbf{A}_k + \boldsymbol{\kappa}_k)^{-1}$, where the $d \times d$ matrix $\boldsymbol{\kappa}_k$ is defined as:

$$\boldsymbol{\kappa}_k \triangleq \lambda\mathbf{R}_{kk} - (\lambda\mathrm{R}_k)^\intercal\mathbf{L}_{-k}(\lambda\mathrm{R}_k) \tag{A.11}$$

By injecting this latter decomposition of $\mathbf{L}$ into the expression of $\mathbf{C}$, we derive a decomposition of $\mathbf{C}$ into the sum of a term that does not depend on the kth basis and of a rank one term:

$$\mathbf{C} = \mathbf{C}_{-k} + (\Phi\mathrm{U}_k)(\mathbf{A}_k + \boldsymbol{\kappa}_k)^{-1}(\Phi\mathrm{U}_k)^\intercal \tag{A.12}$$

Letting $\mathbf{C}_{-k}^{-1} \triangleq (\mathbf{C}_{-k})^{-1}$, a second application of rank one update identities for the determinant and the inverse gives the two following expressions A.13 and A.14 for the two terms in the right-hand side of the log marginal likelihood expression A.8:

$$|\mathbf{C}| = |\mathbf{C}_{-k}| \cdot |\mathbf{A}_k + \boldsymbol{\kappa}_k|^{-1} \cdot |\mathbf{A}_k + \boldsymbol{\kappa}_k + \boldsymbol{s}_k| \tag{A.13}$$

$$\boldsymbol{t}^\intercal\mathbf{C}^{-1}\boldsymbol{t} = \boldsymbol{t}^\intercal\mathbf{C}_{-k}^{-1}\boldsymbol{t} - \boldsymbol{q}_k^\intercal\left(\mathbf{A}_k + \boldsymbol{\kappa}_k + \boldsymbol{s}_k\right)^{-1}\boldsymbol{q}_k \tag{A.14}$$

We introduced the statistics $\boldsymbol{s}_k$ and $\boldsymbol{q}_k$ respectively defined as:

$$\boldsymbol{s}_k \triangleq (\Phi\mathrm{U}_k)^\intercal\mathbf{C}_{-k}^{-1}(\Phi\mathrm{U}_k) \tag{A.15}$$

$$\boldsymbol{q}_k \triangleq (\Phi\mathrm{U}_k)^\intercal\mathbf{C}_{-k}^{-1}\boldsymbol{t} \tag{A.16}$$

We thus retrieve the expression Eq. (3.24) for $l(\mathbf{A}_k)$, which was used without proof in Section 3.3.3. It is of practical significance to the algorithmic complexity of our schemes that the quantities involved ($\mathbf{L}$, $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$) do not actually depend on bases that are not in the active set $\mathcal{S}$ (*i.e.* all bases s.t. $\mathbf{A}_m = +\infty$). Similarly $\boldsymbol{s}_k$, $\boldsymbol{\kappa}_k$ and $\boldsymbol{q}_k$ only involve the set of active bases $\mathcal{S}$ augmented with the kth basis, due of the form of U_k.

The maximization of Eq. (3.24) under constraint that $\mathbf{A}_k$ is a symmetric positive semidefinite $d \times d$ matrix involves the gradient of the (unconstrained) function $l(\mathbf{A}_k)$:

$$\nabla l(\mathbf{A}_k) = -\boldsymbol{\sigma}_k\left\{\boldsymbol{q}_k\boldsymbol{q}_k^\intercal - \boldsymbol{s}_k - \boldsymbol{s}_k(\mathbf{A}_k + \boldsymbol{\kappa}_k)^{-1}\boldsymbol{s}_k\right\}\boldsymbol{\sigma}_k \tag{A.17}$$

where $\boldsymbol{\sigma}_k$ is shorthand for $(\mathbf{A}_k + \boldsymbol{\kappa}_k + \boldsymbol{s}_k)^{-1}$, and $\nabla l(\mathbf{A}_k)$ is a $d \times d$ matrix. Since $\boldsymbol{\sigma}_k$ is symmetric positive definite and $\boldsymbol{q}_k\boldsymbol{q}_k^\intercal$ is of rank one, $\nabla l(\mathbf{A}_k)$ has at most one negative eigenvalue. More precisely, if $\boldsymbol{q}_k\boldsymbol{q}_k^\intercal - \boldsymbol{s}_k$ is negative then $\nabla l(\mathbf{A}_k)$ is positive definite for all $\mathbf{A}_k$ and the improper maximizer of $l(\mathbf{A}_k)$ lies at infinity $\mathbf{A}_k \to +\infty$. Otherwise there is exactly one negative eigenvalue and we look for maximizers of the form $\mathbf{A}_k^{-1} = \alpha_k^{-1}\boldsymbol{\eta}_k\boldsymbol{\eta}_k^\intercal$.

This is consistent with the intuitive comment that $\mathbf{A}_k^{-1} \in \mathcal{M}_{d,d}$ cannot be fully determined from a single "observation" $\boldsymbol{q}_k \in \mathbb{R}^d$ and should be degenerate. Rewriting Eq. (3.24) as a function of α, η leads to maximizing A.18 under constraint that α is positive (dropping the index k for convenience). Note also that A.18 is invariant under reparametrization $\boldsymbol{\eta} \to \nu\boldsymbol{\eta}, \alpha \to \alpha/\nu^2$.

$$l(\alpha, \boldsymbol{\eta}) = -\log\left\{1 + \frac{\boldsymbol{\eta}^\intercal \boldsymbol{s}\boldsymbol{\eta}}{\alpha + \boldsymbol{\eta}^\intercal\boldsymbol{\kappa}\boldsymbol{\eta}}\right\} + \frac{(\boldsymbol{q}^\intercal\boldsymbol{\eta})^2}{\alpha + \boldsymbol{\eta}^\intercal(\boldsymbol{\kappa} + \boldsymbol{s})\boldsymbol{\eta}} \tag{A.18}$$

At a maximizer $\alpha^*, \boldsymbol{\eta}^* = \arg\max_{\alpha,\boldsymbol{\eta}} l(\alpha, \boldsymbol{\eta})$ the constraint is either active ($\alpha^* = 0$) or inactive ($\alpha^* > 0$). If inactive, the solution actually maximizes the unconstrained function A.18 and is given by $\alpha^* = \bar{\alpha}(\bar{\boldsymbol{\eta}})$, $\boldsymbol{\eta}^* = \bar{\boldsymbol{\eta}}$ where

$$\bar{\alpha}(\boldsymbol{\eta}) = \frac{(\boldsymbol{\eta}^\intercal \boldsymbol{s}\boldsymbol{\eta})^2}{(\boldsymbol{q}^\intercal\boldsymbol{\eta})^2 - \boldsymbol{\eta}^\intercal \boldsymbol{s}\boldsymbol{\eta}} - \boldsymbol{\eta}^\intercal\boldsymbol{\kappa}\boldsymbol{\eta}\,, \tag{A.19}$$

$$\bar{\boldsymbol{\eta}} = \boldsymbol{s}^{-1}\boldsymbol{q}\,. \tag{A.20}$$

In this case $l(\alpha^*, \boldsymbol{\eta}^*)$ is simply equal to $\bar{l}(\boldsymbol{\eta}^*)$, where $\bar{l}(\boldsymbol{\eta})$ is defined by A.21 with $\xi(\boldsymbol{\eta}) \triangleq (\boldsymbol{q}^\intercal\boldsymbol{\eta})^2/\boldsymbol{\eta}^\intercal \boldsymbol{s}\boldsymbol{\eta}$.

$$\bar{l}(\boldsymbol{\eta}) \triangleq -\log\xi(\boldsymbol{\eta}) + \xi(\boldsymbol{\eta}) - 1 \tag{A.21}$$

In addition $\bar{l}(\bar{\boldsymbol{\eta}})$ can be seen to always provide an upper bound to the maximum contribution of a basis to the evidence, $\max_{\alpha,\boldsymbol{\eta}} l(\alpha, \boldsymbol{\eta})$. In the case where the constraint is active, $\alpha^* = 0$, we numerically optimize over the unit sphere in $\mathbb{R}^d$ to find $\boldsymbol{\eta}^*$. This case occurs when the l_2-norm regularization is by itself sufficient along the direction $\boldsymbol{\eta}^*$, and no additional shrinkage is deemed necessary. To save on unnecessary computations, we first check that the upper bound $\bar{l}(\bar{\boldsymbol{\eta}})$ to the maximum contribution of the basis k to the evidence is superior to the current best contribution among bases already handled, as this is a necessary condition for $\mathbf{A}_k$ to be updated as this iteration.

A.3 Update of μ, Σ, L

Updates of the moments of the posterior distribution $\boldsymbol{\mu}$, $\boldsymbol{\Sigma} = (\Phi^\intercal(\beta\mathbf{H})\Phi + \mathbf{A} + \lambda\mathbf{R})^{-1}$ and of $\mathbf{L} = (\mathbf{A} + \lambda\mathbf{R})^{-1}$ upon deletion from the model, update or addition to the model of a basis i are done similarly to [Tipping 2003] and follow from Woodbury identities. Denoting updated quantities with a tilde, we get in the case of deletion:

$$\tilde{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma} - \Sigma_i\Sigma_{ii}^{-1}\Sigma_i^\intercal\,, \tag{A.22}$$

$$\tilde{\mathbf{L}} = \mathbf{L} - \mathrm{L}_i\mathrm{L}_{ii}^{-1}\mathrm{L}_i^\intercal \tag{A.23}$$

and

$$\tilde{\boldsymbol{\mu}} = \boldsymbol{\mu} - \Sigma_i(\Sigma_{ii}^{-1}\boldsymbol{\mu}_i)\,. \tag{A.24}$$

These rank one updates carefully avoid matrix-matrix products and have a $\mathcal{O}(|\mathcal{S}|^2)$ complexity. In the case of the addition of a basis, we first compute the new column of $\tilde{\boldsymbol{\Sigma}}$ (resp. $\tilde{\mathbf{L}}$) before updating its full body as:

$$\tilde{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma} + \tilde{\Sigma}_i\tilde{\Sigma}_{ii}^{-1}\tilde{\Sigma}_i^\intercal\,, \tag{A.25}$$

$$\tilde{\mathbf{L}} = \mathbf{L} + \tilde{\mathrm{L}}_i \tilde{\mathrm{L}}_{ii}^{-1} \tilde{\mathrm{L}}_i^\intercal \,, \tag{A.26}$$

where the column $\tilde{\Sigma}_i$ (resp. $\tilde{\mathbf{L}}_i$) is given in $\mathcal{O}(|\mathcal{S}|^2)$ by

$$\tilde{\Sigma}_i = \begin{pmatrix} \boldsymbol{\Sigma} \Pi_i \tilde{\Sigma}_{ii} \\ \tilde{\Sigma}_{ii} \end{pmatrix} \,, \quad \tilde{\mathrm{L}}_i = \begin{pmatrix} \mathbf{L}(\lambda \mathrm{R}_i) \tilde{L}_{ii} \\ \tilde{L}_{ii} \end{pmatrix} \tag{A.27}$$

and

$$\tilde{\Sigma}_{ii} = (\boldsymbol{s}_i + \boldsymbol{\kappa}_i + \mathbf{A}_i)^{-1} \,, \quad \tilde{L}_{ii} = (\boldsymbol{\kappa}_i + \mathbf{A}_i)^{-1} \,. \tag{A.28}$$

Π_i is the column vector of $d \times d$ matrices defined by $\Pi_i = \Phi^\intercal(\beta \mathbf{H}) \phi_i + \lambda \mathbf{R}_i$. The case of the update of a basis i is treated as a deletion followed by an addition, updating $\boldsymbol{s}_i$ and $\boldsymbol{\kappa}_i$ in-between these actions as they are needed in Eq. (A.28).

A.4 Update of s_i, κ_i and q_i

From the resolvent identity, we note that $\Sigma \Phi^\intercal(\beta \mathbf{H}) \Phi \mathbf{L} = \mathbf{L} - \Sigma$. Using such relationships after developing the factors in A.15 and A.16, we derive alternative expressions for $\boldsymbol{s}_m$ and $\boldsymbol{q}_m$:

$$\boldsymbol{s}_m = \phi_m^\intercal(\beta \mathbf{H}) \phi_m - \Pi_m^\intercal \boldsymbol{\Sigma}_{-m} \Pi_m + (\lambda \mathrm{R}_m)^\intercal \mathbf{L}_{-m} (\lambda \mathrm{R}_m) \tag{A.29}$$

$$\boldsymbol{q}_m = \phi_m^\intercal(\beta \mathbf{H}) \boldsymbol{t} - \Pi_m^\intercal \boldsymbol{\Sigma}_{-m} \Phi^\intercal(\beta \mathbf{H}) \boldsymbol{t} \tag{A.30}$$

where Π_m is a column vector of 3×3 matrices defined by $\Pi_m = \Phi^\intercal(\beta \mathbf{H}) \phi_m + \lambda \mathbf{R}_m$. Π_m can be interpreted as the inner product of basis m with all the active bases w.r.t an appropriate metric, in the sense that its jth coefficient is given by: $\Pi_{jm} = \phi_j^\intercal(\beta \mathbf{H}) \phi_m + \lambda \, (D\phi_j | D\phi_m)_{\mathcal{H}}$. In the specific case where $\lambda = 0$, we retrieve the quantities and expressions derived by [Tipping 2003] for the RVM. We found useful to introduce surrogate quantities $\boldsymbol{t}_m$ and $\boldsymbol{r}_m$ respectively defined according to A.31 and A.32:

$$\boldsymbol{t}_m \triangleq \phi_m^\intercal(\beta \mathbf{H}) \phi_m - \Pi_m^\intercal \boldsymbol{\Sigma} \Pi_m + (\lambda \mathrm{R}_m)^\intercal \mathbf{L} (\lambda \mathrm{R}_m) \tag{A.31}$$

$$\boldsymbol{r}_m \triangleq \phi_m^\intercal(\beta \mathbf{H}) \boldsymbol{t} - \Pi_m^\intercal \boldsymbol{\Sigma} \Phi^\intercal(\beta \mathbf{H}) \boldsymbol{t} \tag{A.32}$$

These quantities merely differ from $\boldsymbol{s}_m$ and $\boldsymbol{q}_m$ in that the index $-m$ was dropped from $\boldsymbol{\Sigma}_{-m}$ and $\mathbf{L}_{-m}$. Our underlying motivation is to update simpler quantities $\boldsymbol{t}_m$ and $\boldsymbol{r}_m$ that still retain a straightforward link to the statistics $\boldsymbol{s}_m$ and $\boldsymbol{q}_m$ of interest for the computation of $l(\mathbf{A}_m)$. Indeed, for a basis l that does not lie in the model, $\boldsymbol{\Sigma}_{-l} = \boldsymbol{\Sigma}$ and $\mathbf{L}_{-l} = \mathbf{L}$. Therefore, the quantities under consideration coincide: $\boldsymbol{s}_l = \boldsymbol{t}_l$ and $\boldsymbol{q}_l = \boldsymbol{r}_l$. For a basis j that lies in the model and noting that $\boldsymbol{\Sigma}_{-j} = \boldsymbol{\Sigma} - \Sigma_j \Sigma_{jj}^{-1} \Sigma_j^\intercal$, we obtain the statistics of interest efficiently as:

$$\begin{aligned} \boldsymbol{s}_j = \boldsymbol{t}_j + & \left[\Pi_j^\intercal \Sigma_j \right] \, \Sigma_{jj}^{-1} \, \left[\Pi_j^\intercal \Sigma_j \right]^\intercal \\ & - \, [(\lambda \mathrm{R}_j)^\intercal \mathrm{L}_j] \, L_{jj}^{-1} \, [(\lambda \mathrm{R}_j)^\intercal \mathrm{L}_j]^\intercal \end{aligned} \tag{A.33}$$

$$\boldsymbol{q}_j = \boldsymbol{r}_j + \left[\Pi_j^\intercal \Sigma_j \right] \, \Sigma_{jj}^{-1} \mu_j \tag{A.34}$$

Thus, we always maintain the quantities $\boldsymbol{t}_m$ and $\boldsymbol{r}_m$ (for every basis) and recompute $\boldsymbol{s}_m$ and $\boldsymbol{q}_m$ either at no cost for inactive bases or, for bases in the active set $\mathcal{S}$, in $\mathcal{O}(|\mathcal{S}| \cdot d)$. Updates of $\boldsymbol{t}_m$ and $\boldsymbol{r}_m$ upon deletion from the model, update or addition to the model of a basis i are done similarly to [Tipping 2003], in $\mathcal{O}(|\mathcal{S}| \cdot d)$ per basis. For instance, in the addition case, it follows from Woodbury identities that

$$\begin{aligned} \tilde{\boldsymbol{t}}_m = \boldsymbol{t}_m - & \left[\Pi_m^\intercal \tilde{\Sigma}_i\right] \tilde{\Sigma}_{ii}^{-1} \left[\Pi_m^\intercal \tilde{\Sigma}_i\right]^\intercal \\ + & \left[(\lambda \mathrm{R}_m)^\intercal \tilde{\mathrm{L}}_i\right] \tilde{L}_{ii}^{-1} \left[(\lambda \mathrm{R}_m)^\intercal \tilde{\mathrm{L}}_i\right]^\intercal \end{aligned} \tag{A.35}$$

and

$$\tilde{\boldsymbol{r}}_m = \boldsymbol{r}_m - \left[\Pi_m^\intercal \tilde{\Sigma}_i\right] \boldsymbol{r}_i \tag{A.36}$$

where $\tilde{\boldsymbol{r}}_m$ and $\tilde{\boldsymbol{t}}_m$ denote updated quantities, as opposed to quantities prior to the update $\boldsymbol{r}_m$ and $\boldsymbol{t}_m$. The quantities indexed by i are computed (once for all bases) following A.3.

A.5 Marginal prior & marginal likelihood

We assume a Gaussian prior $p(\boldsymbol{w}|\lambda) = |\frac{\lambda}{2\pi}\mathbf{R}|^{1/2} \exp -\frac{\lambda}{2}\boldsymbol{w}^\intercal \mathbf{R}\boldsymbol{w}$ and a conjugate Gamma prior with support over strictly positive real numbers, $p(\lambda|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\lambda^{\alpha-1}e^{-\beta\lambda}$, where $\Gamma(\alpha) = \int_0^\infty t^{\alpha-1}e^{-t}dt$ is the Gamma function. Then the marginal distribution $p(\boldsymbol{w}) = \int p(\boldsymbol{w}|\lambda)p(\lambda)d\lambda$ is given by:

$$p(\boldsymbol{w}) = \frac{\Gamma(\alpha + \dim(\boldsymbol{w})/2)}{\Gamma(\alpha)} \left|\frac{1}{2\pi\beta}\mathbf{R}\right|^{1/2} \left(1 + \frac{\chi^2}{2\beta}\right)^{-(\alpha+\dim(\boldsymbol{w})/2)} \tag{A.37}$$

with $\chi^2 = \boldsymbol{w}^\intercal \mathbf{R}\boldsymbol{w}$. This is a non-standardized multivariate Student $t_{2\alpha}(\boldsymbol{w}|\,\mathbf{0}, \frac{1}{\beta/\alpha}\mathbf{R})$, denoting by $t_\nu(\cdot|\boldsymbol{\mu},\boldsymbol{\Lambda})$ the multivariate Student distribution with location parameter $\boldsymbol{\mu}$, inverse scale matrix $\boldsymbol{\Lambda}$ and ν degrees of freedom:

$$t_\nu(\boldsymbol{t}|\boldsymbol{\mu},\boldsymbol{\Lambda}) = \frac{\Gamma(\frac{\nu+\dim(\boldsymbol{t})}{2})}{\Gamma(\frac{\nu}{2})} \left|\frac{1}{\nu\pi}\boldsymbol{\Lambda}\right|^{1/2} \left(1 + \frac{1}{\nu}(\boldsymbol{t}-\boldsymbol{\mu})^\intercal \boldsymbol{\Lambda}(\boldsymbol{t}-\boldsymbol{\mu})\right)^{-(\nu+\dim(\boldsymbol{t}))/2} \tag{A.38}$$

In the limit case of $\alpha, \beta \to 0$, $\chi^2 \gg 2\beta$, $p(\boldsymbol{w}) \propto \frac{1}{\chi^{dim(\boldsymbol{w})}}$. Insight into this latter improper prior can be gained by commenting that it also arises naturally from the two following assumptions: scale invariance (a priori) on χ, $p(\chi) \propto 1/\chi$, and rotational invariance of $p(\boldsymbol{w})$ w.r.t. the euclidean metric defined by $\mathbf{R}$.

The form of the marginal prior given by Eq. (4.6) follows from the above. For a fixed set of active bases $\mathcal{S}$ specified by the state of activation variables $\boldsymbol{z} = \begin{pmatrix} z_1 & \cdots & z_M \end{pmatrix}^\intercal$, the marginal prior $p(\boldsymbol{w}|\boldsymbol{z},\mathcal{H})$ is the product of a Dirac $\mathcal{N}(\boldsymbol{w}_{-\mathcal{S}} \,|\, \mathbf{0},\mathbf{0})$ on weights $\boldsymbol{w}_{-\mathcal{S}}$ associated to inactive bases, and of a multivariate Student distribution $t_{\nu_\lambda}(\boldsymbol{w}_\mathcal{S}|\,\mathbf{0}, \frac{a_0}{b_0}\, d|\mathcal{S}|\, \mathbf{R}_\mathcal{S})$ on weights $\boldsymbol{w}_\mathcal{S}$ associated to active bases. Analogous derivations lead to Eq. 4.7 for the marginal likelihood.

In the reversible jump scheme developed in section 4.2.3.3, ratios $\frac{p(\tilde{\boldsymbol{w}}|\tilde{\boldsymbol{z}},\mathcal{H})}{p(\boldsymbol{w}|\boldsymbol{z},\mathcal{H})}$ of marginal priors are involved, where exactly one activation variable and its associated weight differ between both states, *e.g.* ($z_k = 0$, $w_k = 0$), ($\tilde{z}_k = 1$, $\tilde{w}_k$ arbitrary); and for any $l \neq k$, $\tilde{z}_k = z_k$ and $\tilde{w}_k = w_k$. The ratio can be written as:

$$\frac{\Gamma(a_0 + d\frac{|\mathcal{S}|+1}{2})}{\Gamma(a_0 + d\frac{|\mathcal{S}|}{2})} \left(\frac{|\mathcal{S}|+1}{|\mathcal{S}|}\right)^{d\frac{|\mathcal{S}|}{2}} \left(\frac{d(|\mathcal{S}|+1)}{2\pi b_0}\right)^{\frac{d}{2}} \frac{|\mathbf{R}_{\tilde{\mathcal{S}}}|^{\frac{1}{2}}}{|\mathbf{R}_{\mathcal{S}}|^{\frac{1}{2}}} \frac{\left(1 + \frac{d|\mathcal{S}|\chi^2}{2b_0}\right)^{a_0 + d\frac{|\mathcal{S}|}{2}}}{\left(1 + \frac{d(|\mathcal{S}|+1)\tilde{\chi}^2}{2b_0}\right)^{a_0 + d\frac{|\mathcal{S}|+1}{2}}} \tag{A.39}$$

where $\mathcal{S}$ is the set of active bases excluding k, $\tilde{\mathcal{S}}$ includes basis k, $\chi^2 = \boldsymbol{w}_{\mathcal{S}}^\intercal \mathbf{R}_{\mathcal{S}} \boldsymbol{w}_{\mathcal{S}}$ and $\tilde{\chi}^2 = \tilde{\boldsymbol{w}}_{\tilde{\mathcal{S}}}^\intercal \mathbf{R}_{\tilde{\mathcal{S}}} \tilde{\boldsymbol{w}}_{\tilde{\mathcal{S}}} = \chi^2 + 2\tilde{w}_k^\intercal \mathbf{R}_k \boldsymbol{w}_{\mathcal{S}} + \tilde{w}_k^\intercal \mathbf{R}_{kk} \tilde{w}_k$. The ratio $\frac{|\mathbf{R}_{\tilde{\mathcal{S}}}|}{|\mathbf{R}_{\mathcal{S}}|}$ of determinants, using rank-one matrix update identities, is equal to $|\mathbf{R}_{kk} - \mathbf{R}_k^\intercal \mathbf{R}_{\mathcal{S}}^{-1} \mathbf{R}_k|$.

Bibliography

[Alessandrini 2015] Martino Alessandrini, Mathieu De Craene, Olivier Bernard, Sophie Giffard-Roisin, Pascal Allain, Juergen Weese, Eric Saloux, Herve Delingette, Maxime Sermesant and Jan D'Hooge. *A Pipeline for the Generation of Realistic 3D Synthetic Echocardiographic Sequences: Methodology and Open-access Database.* IEEE Transactions on Medical Imaging, page in press, 2015. (Cited on page 7.)

[Allassonnière 2007] Stéphanie Allassonnière, Yali Amit and Alain Trouvé. *Towards a coherent statistical framework for dense deformable template estimation*. Journal of the Royal Statistical Society: Series B (Statistical Methodology), vol. 69, no. 1, pages 3–29, 2007. (Cited on page 56.)

[Archambeau 2007] Cédric Archambeau and Michel Verleysen. *Robust bayesian clustering*. Neural Networks, vol. 20, no. 1, pages 129–138, 2007. (Cited on page 29.)

[Arendt 2012] Paul D Arendt, Daniel W Apley and Wei Chen. *Quantification of model uncertainty: Calibration, model discrepancy, and identifiability*. Journal of Mechanical Design, vol. 134, no. 10, page 100908, 2012. (Cited on page 78.)

[Argyriou 2012] Andreas Argyriou, Rina Foygel and Nathan Srebro. *Sparse Prediction with the k-Support Norm*. In Advances in Neural Information Processing Systems, pages 1457–1465, 2012. (Cited on page 34.)

[Aronszajn 1951] N. Aronszajn. Theory of reproducing kernels. Harvard University, 1951. (Cited on page 12.)

[Arsigny 2003] Vincent Arsigny, Xavier Pennec and Nicholas Ayache. *Polyrigid and polyaffine transformations: A new class of diffeomorphisms for locally rigid or affine registration*. Medical Image Computing and Computer-Assisted Intervention-MICCAI 2003, pages 829–837, 2003. (Cited on pages 8 and 79.)

[Arsigny 2006] Vincent Arsigny, Olivier Commowick, Xavier Pennec and Nicholas Ayache. *A log-euclidean framework for statistics on diffeomorphisms*. In Medical Image Computing and Computer-Assisted Intervention–MICCAI 2006, pages 924–931. Springer, 2006. (Cited on pages 8, 55 and 79.)

[Arsigny 2009] Vincent Arsigny, Olivier Commowick, Nicholas Ayache and Xavier Pennec. *A fast and log-euclidean polyaffine framework for locally linear registration*. Journal of Mathematical Imaging and Vision, vol. 33, no. 2, pages 222–238, 2009. (Cited on page 8.)

[Ashburner 2007] John Ashburner. *A fast diffeomorphic image registration algorithm*. NeuroImage, vol. 38, no. 1, pages 95 – 113, 2007. (Cited on pages 8, 32 and 79.)

[Ashburner 2013] John Ashburner and Gerard R. Ridgway. *Symmetric diffeomorphic modelling of longitudinal structural MRI*. Frontiers in Neuroscience, vol. 6, no. 197, 2013. (Cited on page 32.)

[Bach 2012] Francis Bach, Rodolphe Jenatton, Julien Mairal and Guillaume Obozinski. *Optimization with sparsity-inducing penalties*. Foundations and Trends® in Machine Learning, vol. 4, no. 1, pages 1–106, 2012. (Cited on page 34.)

[Bay 2008] Herbert Bay, Andreas Ess, Tinne Tuytelaars and Luc Van Gool. *Speeded-up robust features (SURF)*. Computer vision and image understanding, vol. 110, no. 3, pages 346–359, 2008. (Cited on page 8.)

[Beg 2005] M.Faisal Beg, MichaelI. Miller, Alain Trouvé and Laurent Younes. *Computing Large Deformation Metric Mappings via Geodesic Flows of Diffeomorphisms*. International Journal of Computer Vision, vol. 61, no. 2, pages 139–157, 2005. (Cited on pages 8, 55 and 79.)

[Belilovsky 2015a] Eugene Belilovsky, Andreas Argyriou, Gaël Varoquaux and Matthew B. Blaschko. *Convex relaxations of penalties for sparse correlated variables with bounded total variation* . Machine Learning, pages 1–21, June 2015. (Cited on page 34.)

[Belilovsky 2015b] Eugene Belilovsky, Katerina Gkirtzou, Michail Misyrlis, Anna Konova, Jean Honorio, Nelly Alia-Klein, Rita Goldstein, Dimitris Samaras and Matthew Blaschko. *Predictive sparse modeling of fMRI data for improved classification, regression, and visualization using the k-support norm*. Computerized Medical Imaging and Graphics, page 1, 2015. (Cited on page 34.)

[Bestel 2001] J. Bestel, F. Clément and M. Sorine. *A biomechanical model of muscle contraction*. In Niessen, W.J., Viergever, M.A. (eds.) MICCAI. LNCS, vol. 2208, pages 1159–1161. Springer, 2001. (Cited on pages 4 and 11.)

[Bishop 2000] Christopher M Bishop and Michael E Tipping. *Variational relevance vector machines*. In Proceedings of the Sixteenth conference on Uncertainty in artificial intelligence, pages 46–53. Morgan Kaufmann Publishers Inc., 2000. (Cited on pages 34 and 80.)

[Bishop 2006] Christopher M Bishop*et al.* Pattern recognition and machine learning, volume 1. springer New York, 2006. (Cited on page 36.)

[Broit 1981] Chaim Broit. *Optimal registration of deformed images*. 1981. (Cited on pages 8, 31 and 57.)

[Brunet 2010] F. Brunet. *Contributions to Parametric Image Registration and 3D Surface Reconstruction*. PhD thesis, Université d'Auvergne, Technische Universität München, 2010. (Cited on page 7.)

[Brynjarsdóttir 2014] Jenný Brynjarsdóttir and Anthony O'Hagan. *Learning about physical parameters: The importance of model discrepancy*. Inverse Problems, vol. 30, no. 11, page 114007, 2014. (Cited on page 78.)

[Cachier 2003] Pascal Cachier, Eric Bardinet, Didier Dormont, Xavier Pennec and Nicholas Ayache. *Iconic feature based nonrigid registration: the PASHA algorithm*. Computer vision and image understanding, vol. 89, no. 2, pages 272–298, 2003. (Cited on pages 7 and 27.)

[Cachier 2004] Pascal Cachier and Nicholas Ayache. *Isotropic energies, filters and splines for vector field regularization*. Journal of Mathematical Imaging and Vision, vol. 20, no. 3, pages 251–265, 2004. (Cited on page 37.)

[Calonder 2010] Michael Calonder, Vincent Lepetit, Christoph Strecha and Pascal Fua. *Brief: Binary robust independent elementary features*. Computer Vision–ECCV 2010, pages 778–792, 2010. (Cited on page 8.)

[Candela 2004] Joaquin Quinonero Candela and Lars Kai Hansen. *Learning with uncertainty-Gaussian processes and relevance vector machines*. PhD thesis, 2004. (Cited on page 80.)

[Chabiniok 2012] R. Chabiniok, P. Moireau, P.-F. Lesault, A. Rahmouni, J.-F. Deux and D. Chapelle. *Estimation of tissue contractility from cardiac cine-MRI using a biomechanical heart model*. Biomechanics and Modeling in Mechanobiology, vol. 11, no. 5, pages 609–630, 2012. (Cited on pages 6, 11 and 78.)

[Chandrashekara 2004] Raghavendra Chandrashekara, Raad H Mohiaddin and Daniel Rueckert. *Analysis of 3-D myocardial motion in tagged MR images using nonrigid image registration*. IEEE Transactions on Medical Imaging, vol. 23, no. 10, pages 1245–1250, 2004. (Cited on pages 8 and 27.)

[Chapelle 2012] D. Chapelle, P. Le Tallec, P. Moireau and M. Sorine. *An energy-preserving muscle tissue model: formulation and compatible discretizations*. IJMCE, 10(2):189-211, 2012. (Cited on pages 4 and 11.)

[Chib 1995] Siddhartha Chib. *Marginal likelihood from the Gibbs output*. Journal of the American Statistical Association, vol. 90, no. 432, pages 1313–1321, 1995. (Cited on page 64.)

[Clatz 2005] Olivier Clatz, Hervé Delingette, Ion-Florin Talos, Alexandra J Golby, Ron Kikinis, Ferenc Jolesz, Nicholas Ayache, Simon K Warfield*et al. Robust nonrigid registration to capture brain shift from intraoperative MRI*. Medical Imaging, IEEE Transactions on, vol. 24, no. 11, pages 1417–1427, 2005. (Cited on page 8.)

[Dalal 2005] Navneet Dalal and Bill Triggs. *Histograms of oriented gradients for human detection*. In Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, volume 1, pages 886–893. IEEE, 2005. (Cited on page 8.)

[Davis 1997] G. Davis, S. Mallat and M. Avellaneda. *Adaptive greedy approximations*. Constructive approximation, vol. 13, no. 1, pages 57–98, 1997. (Cited on page 14.)

[De Craene 2010] Mathieu De Craene, Gemma Piella, Nicolas Duchateau, Etel Silva, Adelina Doltra, Hang Gao, Jan D'hooge, Oscar Camara, Josep Brugada, Marta Sitges*et al.* *Temporal diffeomorphic free-form deformation for strain quantification in 3D-US images*. In Medical Image Computing and Computer-Assisted Intervention–MICCAI 2010, pages 1–8. Springer, 2010. (Cited on page 8.)

[De Craene 2012] Mathieu De Craene, Gemma Piella, Oscar Camara, Nicolas Duchateau, Etelvino Silva, Adelina Doltra, Jan D'hooge, Josep Brugada, Marta Sitges and Alejandro F Frangi. *Temporal diffeomorphic free-form deformation: Application to motion and strain estimation from 3D echocardiography*. Medical Image Analysis, vol. 16, no. 2, pages 427–450, 2012. (Cited on pages 8 and 55.)

[De Craene 2013] Mathieu De Craene, Stephanie Marchesseau, Brecht Heyde, Hang Gao, M Alessandrini, Olivier Bernard, Gemma Piella, AR Porras, E Saloux, L Tautz*et al.* *3D Strain Assessment in Ultrasound (STRAUS): A synthetic comparison of five tracking methodologies*. IEEE Transactions on Medical Imaging, 2013. (Cited on pages 9, 44 and 46.)

[Delingette 2012] H. Delingette, F. Billet, K.C.L. Wong, M. Sermesant, K. Rhode, M. Ginks, CA Rinaldi, R. Razavi, N. Ayache*et al.* *Personalization of Cardiac Motion and Contractility from Images using Variational Data Assimilation*. IEEE Trans. Biomed. Eng., vol. 59, no. 1, page 20, 2012. (Cited on pages 5 and 11.)

[Dupuis 1998] Paul Dupuis, Ulf Grenander and Michael I Miller. *Variational problems on flows of diffeomorphisms for image matching*. Quarterly of applied mathematics, vol. 56, no. 3, page 587, 1998. (Cited on page 8.)

[Durrleman 2007] S. Durrleman, X. Pennec, A. Trouvé and N. Ayache. *Measuring brain variability via sulcal lines registration: a diffeomorphic approach*. In Ayache, N., Ourselin, S., Maeder, A. (eds.) MICCAI 2007, Part I. LNCS, vol. 4791, pages 675–682. Springer, Heidelberg, 2007. (Cited on pages 8, 12 and 79.)

[Durrleman 2008] S. Durrleman, X. Pennec, A. Trouvé and N. Ayache. *Sparse approximation of currents for statistics on curves and surfaces*. In Metaxas, D., Axel, L., Szekely, G., and Fichtinger, G. (eds.) Proceedings MICCAI, Part II, LNCS, vol. 5242, pages 390–398. Springer, 2008. (Cited on pages 14 and 16.)

[Durrleman 2009] Stanley Durrleman, Xavier Pennec, Alain Trouvé and Nicholas Ayache. *Statistical models of sets of curves and surfaces based on currents*. Medical image analysis, vol. 13, no. 5, pages 793–808, 2009. (Cited on page 15.)

[Durrleman 2010] S. Durrleman. *Statistical models of currents for measuring the variability of anatomical curves, surfaces and their evolution*. Ph.D. Thesis, INRIA, March 2010. (Cited on pages 12 and 15.)

[Durrleman 2013] Stanley Durrleman, Stéphanie Allassonnière and Sarang Joshi. *Sparse adaptive parameterization of variability in image ensembles*. International Journal of Computer Vision, vol. 101, no. 1, pages 161–183, 2013. (Cited on page 56.)

[Ecabert 2011] Olivier Ecabert, Jochen Peters, Matthew J Walker, Thomas Ivanc, Cristian Lorenz, Jens von Berg, Jonathan Lessick, Mani Vembar and Jürgen Weese. *Segmentation of the heart and great vessels in CT images using a model-based adaptation framework*. Medical Image Analysis, vol. 15, no. 6, pages 863–876, 2011. (Cited on page 5.)

[Fanello 2014] Sean Ryan Fanello, Cem Keskin, Pushmeet Kohli, Shahram Izadi, Jamie Shotton, Antonio Criminisi, Ugo Pattacini and Tim Paek. *Filter Forests for Learning Data-Dependent Convolutional Kernels*. In Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on, pages 1709–1716. IEEE, 2014. (Cited on page 33.)

[Ganz 2013] Melanie Ganz, Mert R Sabuncu and Koen Van Leemput. *An improved optimization method for the relevance voxel machine*. In Machine Learning in Medical Imaging, pages 147–154. Springer, 2013. (Cited on page 80.)

[Gärtner 2002] T. Gärtner, P.A. Flach, A. Kowalczyk and A.J. Smola. *Multi-instance kernels*. In Proceedings of the 19th International Conference on Machine Learning, pages 179–186, 2002. (Cited on page 13.)

[Gee 1998] James C Gee and Ruzena K Bajcsy. *Elastic matching: Continuum mechanical and probabilistic analysis*. Brain warping, vol. 2, 1998. (Cited on pages 8, 24, 31 and 57.)

[Gelman 1998] Andrew Gelman and Xiao-Li Meng. *Simulating normalizing constants: From importance sampling to bridge sampling to path sampling*. Statistical science, pages 163–185, 1998. (Cited on page 65.)

[Geman 1984] Stuart Geman and Donald Geman. *Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images*. Pattern Analysis and Machine Intelligence, IEEE Transactions on, no. 6, pages 721–741, 1984. (Cited on page 64.)

[Girolami 2011] Mark Girolami and Ben Calderhead. *Riemann manifold langevin and hamiltonian monte carlo methods*. Journal of the Royal Statistical Society: Series B (Statistical Methodology), vol. 73, no. 2, pages 123–214, 2011. (Cited on page 66.)

[Glocker 2007] Ben Glocker, Nikos Komodakis, Nikos Paragios, Georgios Tziritas and Nassir Navab. *Inter and intra-modal deformable registration: Continuous deformations meet efficient optimal linear programming*. In Information Processing in Medical Imaging, pages 408–420. Springer, 2007. (Cited on page 8.)

[Glocker 2008a] Ben Glocker, Nikos Komodakis, Georgios Tziritas, Nassir Navab and Nikos Paragios. *Dense image registration through MRFs and efficient linear programming*. Medical image analysis, vol. 12, no. 6, pages 731–741, 2008. (Cited on page 8.)

[Glocker 2008b] Ben Glocker, Nikos Paragios, Nikos Komodakis, Georgios Tziritas and Nassir Navab. *Optical flow estimation with uncertainties through dynamic MRFs*. In Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on, pages 1–8. IEEE, 2008. (Cited on page 8.)

[Glocker 2011] Ben Glocker, Aristeidis Sotiras, Nikos Komodakis and Nikos Paragios. *Deformable medical image registration: Setting the state of the art with discrete methods**. Annual review of biomedical engineering, vol. 13, pages 219–244, 2011. (Cited on page 8.)

[Gori 2013] Pietro Gori, Olivier Colliot, Yulia Worbe, Linda Marrakchi-Kacem, Sophie Lecomte, Cyril Poupon, Andreas Hartmann, Nicholas Ayache and Stanley Durrleman. *Bayesian Atlas Estimation for the Variability Analysis of Shape Complexes*. In Medical Image Computing and Computer-Assisted Intervention–MICCAI 2013, pages 267–274. Springer, 2013. (Cited on pages 31 and 56.)

[Goshtasby 2012] A Ardeshir Goshtasby. *Image registration methods*. In Image Registration, pages 415–434. Springer, 2012. (Cited on page 7.)

[Green 1995] Peter J Green. *Reversible jump Markov chain Monte Carlo computation and Bayesian model determination*. Biometrika, vol. 82, no. 4, pages 711–732, 1995. (Cited on pages 59 and 65.)

[Groves 2011] Adrian R Groves, Christian F Beckmann, Steve M Smith and Mark W Woolrich. *Linked independent component analysis for multimodal data fusion*. Neuroimage, vol. 54, no. 3, pages 2198–2217, 2011. (Cited on page 30.)

[Haber 2006] Eldad Haber and Jan Modersitzki. *A multilevel method for image registration*. SIAM Journal on Scientific Computing, vol. 27, no. 5, pages 1594–1607, 2006. (Cited on page 8.)

[Hachama 2012] Mohamed Hachama, Agnès Desolneux and Frédéric JP Richard. *Bayesian technique for image classifying registration*. Image Processing, IEEE Transactions on, vol. 21, no. 9, pages 4080–4091, 2012. (Cited on pages 7 and 27.)

[Haussler 1999] D. Haussler. *Convolution kernels on discrete structures*. Rapport technique, Technical report, UC Santa Cruz, 1999. (Cited on page 13.)

[Heiberg 2005] Einar Heiberg, L Wigstrom, Marcus Carlsson, AF Bolger and M Karlsson. *Time resolved three-dimensional automated segmentation of the left ventricle*. In Computers in Cardiology, 2005, pages 599–602. IEEE, 2005. (Cited on page 51.)

[Heyde 2013] Brecht Heyde, Daniel Barbosa, Piet Claus, Frederik Maes and Jan D'hooge. *Influence of the grid topology of free-form deformation models on the performance of 3d strain estimation in echocardiography*. In Functional Imaging and Modeling of the Heart, pages 308–315. Springer, 2013. (Cited on page 8.)

[Hoerl 1970] A.E. Hoerl and R.W. Kennard. *Ridge regression: Biased estimation for nonorthogonal problems*. Technometrics, pages 55–67, 1970. (Cited on page 17.)

[Hoyos-Idrobo 2015] Andrés Hoyos-Idrobo, Yannick Schwartz, Gaël Varoquaux and Bertrand Thirion. *Improving sparse recovery on structured images with bagged clustering*. In International Workshop On Pattern Recognition In Neuroimaging (PRNI), 2015 , Palo alto, United States, June 2015. (Cited on page 2.)

[Imperiale 2011] A. Imperiale, R. Chabiniok, P. Moireau and D. Chapelle. *Constitutive Parameter Estimation Using Tagged-MRI Data*. In Metaxas, D., Axel, L. (eds.) Proceedings of FIMH'11, LNCS 6666, pages 409–417. Springer, 2011. (Cited on pages 6, 11 and 78.)

[Janoos 2012] Firdaus Janoos, Petter Risholm and William Wells III. *Bayesian characterization of uncertainty in multi-modal image registration*. Biomedical Image Registration, pages 50–59, 2012. (Cited on pages 7, 9 and 27.)

[Jenatton 2012] Rodolphe Jenatton, Alexandre Gramfort, Vincent Michel, Guillaume Obozinski, Evelyn Eger, Francis Bach and Bertrand Thirion. *Multi-scale Mining of fMRI data with Hierarchical Structured Sparsity*. SIAM Journal on Imaging Sciences, vol. 5, no. 3, pages 835–856, July 2012. (Cited on page 34.)

[Kaipio 2011] Jari P Kaipio and Colin Fox. *The Bayesian framework for inverse problems in heat transfer*. Heat Transfer Engineering, vol. 32, no. 9, pages 718–753, 2011. (Cited on page 78.)

[Kennedy 2001] Marc C Kennedy and Anthony O'Hagan. *Bayesian calibration of computer models*. Journal of the Royal Statistical Society. Series B, Statistical Methodology, pages 425–464, 2001. (Cited on page 78.)

[Klein 2007] Stefan Klein, Marius Staring and Josien PW Pluim. *Evaluation of optimization methods for nonrigid medical image registration using mutual information and B-splines*. Image Processing, IEEE Transactions on, vol. 16, no. 12, pages 2879–2890, 2007. (Cited on page 8.)

[Kohli 2008] Pushmeet Kohli and Philip HS Torr. *Measuring uncertainty in graph cut solutions*. Computer Vision and Image Understanding, vol. 112, no. 1, pages 30–38, 2008. (Cited on page 8.)

[Konukoglu 2011] Ender Konukoglu, Jatin Relan, Ulas Cilingir, Bjoern H Menze, Phani Chinchapatnam, Amir Jadidi, Hubert Cochet, Mélèze Hocini, Hervé Delingette, Pierre Jaïs*et al.* *Efficient probabilistic model personalization integrating uncertainty on data and parameters: Application to eikonal-diffusion models in cardiac electrophysiology*. Progress in Biophysics and Molecular Biology, vol. 107, no. 1, pages 134–146, 2011. (Cited on page 78.)

[Kybic 2010] Jan Kybic. *Bootstrap resampling for image registration uncertainty estimation without ground truth*. Image Processing, IEEE Transactions on, vol. 19, no. 1, pages 64–73, 2010. (Cited on page 9.)

[Larrabide 2009] Ignacio Larrabide, Pedro Omedas, Yves Martelli, Xavier Planes, Maarten Nieber, Juan A Moya, Constantine Butakoff, Rafael Sebastián, Oscar Camara, Mathieu De Craene*et al.* *GIMIAS: an open source framework for efficient development of research tools and clinical prototypes*. In Functional imaging and modeling of the heart, pages 417–426. Springer, 2009. (Cited on page 5.)

[Le Folgoc 2012] Loic Le Folgoc, Hervé Delingette, Antonio Criminisi and Nicholas Ayache. *Current-based 4D shape analysis for the mechanical personalization of heart models*. In Bjoern Menze, Georg Langs, Albert Montillo, Zhuowen Tu and Antonio Criminisi, editeurs, MCV - MICCAI Workshop on Medical Computer Vision - 2012, volume 7766, pages 283–292, Nice, France, October 2012. Menze, Bjoern and Langs, Georg and Montillo, Albert and Tu, Zhuowen and Criminisi, Antonio, Springer Berlin Heidelberg. (Cited on pages 10 and 77.)

[Le Folgoc 2014] Loic Le Folgoc, Hervé Delingette, Antonio Criminisi and Nicholas Ayache. *Sparse Bayesian Registration*. In Medical Image Computing and Computer-Assisted Intervention–MICCAI 2014, pages 235–242. Springer, 2014. (Cited on pages 10, 25, 29, 36 and 80.)

[Le Folgoc 2015a] Loïc Le Folgoc, Hervé Delingette, Antonio Criminisi and Nicholas Ayache. *Quantifying registration uncertainty with sparse Bayesian modelling: is sparsity a bane?* to be submitted in IEEE Transactions in Medical Imaging, 2015. (Cited on pages 10 and 80.)

[Le Folgoc 2015b] Loïc Le Folgoc, Hervé Delingette, Antonio Criminisi and Nicholas Ayache. *Sparse Bayesian Registration of Medical Images for Self-Tuning of Parameters and Spatially Adaptive Parametrization of Displacements*. submitted to Medical Image Analysis, 2015. (Cited on pages 10, 59, 61, 62, 68, 69, 70 and 80.)

[Liu 2009a] Fei Liu, MJ Bayarri, JO Berger*et al.* *Modularization in Bayesian analysis, with emphasis on analysis of computer models*. Bayesian Analysis, vol. 4, no. 1, pages 119–150, 2009. (Cited on page 78.)

[Liu 2009b] H. Liu and P. Shi. *Maximum a posteriori strategy for the simultaneous motion and material property estimation of the heart*. IEEE Trans. Biomed. Eng., vol. 56, no. 2, pages 378–389, 2009. (Cited on page 11.)

[Loeckx 2004] Dirk Loeckx, Frederik Maes, Dirk Vandermeulen and Paul Suetens. *Nonrigid image registration using free-form deformations with a local rigidity constraint*. In Medical Image Computing and Computer-Assisted Intervention–MICCAI 2004, pages 639–646. Springer, 2004. (Cited on page 8.)

[Lombaert 2014] Herve Lombaert, Darko Zikic, Antonio Criminisi and Nicholas Ayache. *Laplacian Forests: Semantic Image Segmentation by Guided Bagging*. In Medical Image Computing and Computer-Assisted Intervention–MICCAI 2014, pages 496–504. Springer, 2014. (Cited on page 2.)

[Lombaert 2015] Herve Lombaert, Antonio Criminisi and Nicholas Ayache. *Spectral Forests: Learning of Surface Data, Application to Cortical Parcellation*. Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015, 2015. (Cited on page 2.)

[Lorenzi 2013] Marco Lorenzi, Nicholas Ayache, Giovanni B Frisoni, Xavier Pennec, Alzheimer's Disease Neuroimaging Initiative (ADNI*et al. LCC-Demons: a robust and accurate symmetric diffeomorphic registration algorithm*. NeuroImage, vol. 81, pages 470–483, 2013. (Cited on page 7.)

[Lowe 2004] David G Lowe. *Distinctive image features from scale-invariant keypoints*. International journal of computer vision, vol. 60, no. 2, pages 91–110, 2004. (Cited on page 8.)

[Ma 2009] Xiang Ma and Nicholas Zabaras. *An efficient Bayesian inference approach to inverse problems based on an adaptive sparse grid collocation method*. Inverse Problems, vol. 25, no. 3, page 035013, 2009. (Cited on page 78.)

[MacKay 1992] David JC MacKay. *Bayesian interpolation*. Neural computation, vol. 4, no. 3, pages 415–447, 1992. (Cited on page 34.)

[Mäkelä 2002] Timo Mäkelä, Patrick Clarysse, Outi Sipilä, Nicoleta Pauna, Quoc Cuong Pham, Toivo Katila and Isabelle E Magnin. *A review of cardiac image registration methods*. Medical Imaging, IEEE Transactions on, vol. 21, no. 9, pages 1011–1021, 2002. (Cited on page 8.)

[Mansi 2011] T. Mansi, X. Pennec, M. Sermesant, H. Delingette and N. Ayache. *iLogDemons: A Demons-Based registration algorithm for tracking incompressible elastic biological tissues*. International journal of computer vision, vol. 92, no. 1, pages 92–111, 2011. (Cited on page 12.)

[Marchesseau 2012] S. Marchesseau, H. Delingette, M. Sermesant, K. Rhode, S.G. Duckett, C.A. Rinaldi, R. Razavi and N. Ayache. *Cardiac Mechanical Parameter Calibration based on the Unscented Transform*. In Ayache, N., Delingette, H., Golland, P., Mori, K. (eds.) MICCAI, LNCS, volume 7511. Springer, 2012. (Cited on pages 6, 10, 12 and 18.)

[Marchesseau 2013a] Stéphanie Marchesseau. *Simulation of patient-specific cardiac models for therapy planning*. Theses, Ecole Nationale Supérieure des Mines de Paris, January 2013. (Cited on page 4.)

[Marchesseau 2013b] Stéphanie Marchesseau, Hervé Delingette, Maxime Sermesant, Rocio Cabrera Lozoya, Catalina Tobon-Gomez, Philippe Moireau, Rosa Maria Figueras I Ventura, Karim Lekadir, Alfredo Hernandez, Mireille Garreau, Erwan Donal, Christophe Leclercq, Simon G. Duckett, Kawal Rhode, Christopher Aldo

Rinaldi, Alejandro F. Frangi, Reza Razavi, Dominique Chapelle and Nicholas Ayache. *Personalization of a Cardiac Electromechanical Model using Reduced Order Unscented Kalman Filtering from Regional Volumes*. Medical Image Analysis, vol. 17, no. 7, pages 816–829, May 2013. (Cited on pages 6, 10 and 78.)

[Margeta 2011] Ján Margeta, Ezequiel Geremia, Antonio Criminisi and Nicholas Ayache. *Layered Spatio-Temporal Forests for Left Ventricle Segmentation from 4D Cardiac MRI Data*. In MICCAI workshop: Statistical Atlases and Computational Models of the Heart (STACOM), 2011. (Cited on page 2.)

[Margeta 2015] Jan Margeta, Antonio Criminisi, R Cabrera Lozoya, Daniel C Lee and Nicholas Ayache. *Fine-tuned convolutional neural nets for cardiac MRI acquisition plane recognition*. Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization, no. ahead-of-print, pages 1–11, 2015. (Cited on page 2.)

[McLeod 2013a] Kristin McLeod, Christof Seiler, Maxime Sermesant and Xavier Pennec. *A near-incompressible poly-affine motion model for cardiac function analysis*. In Statistical Atlases and Computational Models of the Heart. Imaging and Modelling Challenges, pages 288–297. Springer, 2013. (Cited on page 8.)

[Mcleod 2013b] Kristin Mcleod, Christof Seiler, Nicolas Toussaint, Maxime Sermesant and Xavier Pennec. *Regional analysis of left ventricle function using a cardiac-specific polyaffine motion model*. In Functional Imaging and Modeling of the Heart, pages 483–490. Springer, 2013. (Cited on page 8.)

[Miller 2002] Michael I Miller, Alain Trouvé and Laurent Younes. *On the metrics and Euler-Lagrange equations of computational anatomy*. Annual review of biomedical engineering, vol. 4, no. 1, pages 375–405, 2002. (Cited on page 79.)

[Mitchell 1988] Toby J Mitchell and John J Beauchamp. *Bayesian variable selection in linear regression*. Journal of the American Statistical Association, vol. 83, no. 404, pages 1023–1032, 1988. (Cited on pages 34, 70 and 80.)

[Mohamed 2012] Shakir Mohamed, Katherine A Heller and Zoubin Ghahramani. *Bayesian and L1 Approaches for Sparse Unsupervised Learning*. Proceedings of the 29th International Conference in Machine Learning, 2012. (Cited on page 34.)

[Moireau 2011] P. Moireau and D. Chapelle. *Reduced-order Unscented Kalman Filtering with application to parameter identification in large-dimensional systems*. ESAIM: Control, Optimisation and Calculus of Variations, vol. 17, no. 02, pages 380–405, 2011. (Cited on pages 6 and 78.)

[Murphy 2012] Kevin P Murphy. Machine learning: a probabilistic perspective. MIT press, 2012. (Cited on page 66.)

[Neumann 2014] Dominik Neumann, Tommaso Mansi, Bogdan Georgescu, Ali Kamen, Elham Kayvanpour, Ali Amr, Farbod Sedaghat-Hamedani, Jan Haas, Hugo Katus, Benjamin Meder*et al.* *Robust image-based estimation of cardiac tissue parameters and their uncertainty from noisy data*. In Medical Image Computing and Computer-Assisted Intervention–MICCAI 2014, pages 9–16. Springer International Publishing, 2014. (Cited on pages 6 and 78.)

[Osman 1999] Nael F Osman, William S Kerwin, Elliot R McVeigh and Jerry L Prince. *Cardiac motion tracking using CINE harmonic phase (HARP) magnetic resonance imaging*. Magnetic resonance in medicine: official journal of the Society of Magnetic Resonance in Medicine/Society of Magnetic Resonance in Medicine, vol. 42, no. 6, page 1048, 1999. (Cited on page 8.)

[Ourselin 2000] S. Ourselin, A. Roche, S. Prima and N. Ayache. *Block Matching: A General Framework to Improve Robustness of Rigid Registration of Medical Images*. In ScottL. Delp, AnthonyM. DiGoia and Branislav Jaramaz, editeurs, Medical Image Computing and Computer-Assisted Intervention – MICCAI 2000, volume 1935 of *Lecture Notes in Computer Science*, pages 557–566. Springer Berlin Heidelberg, 2000. (Cited on page 8.)

[Parisot 2014] Sarah Parisot, William Wells, Stéphane Chemouny, Hugues Duffau and Nikos Paragios. *Concurrent tumor segmentation and registration with uncertainty-based sparse non-uniform graphs*. Medical image analysis, vol. 18, no. 4, pages 647–659, 2014. (Cited on page 8.)

[Pitiot 2003] Alain Pitiot, Grégoire Malandain, Eric Bardinet and Paul M Thompson. *Piecewise affine registration of biological images*. In Biomedical Image Registration, pages 91–101. Springer, 2003. (Cited on page 8.)

[Prakosa 2013a] Adityo Prakosa. *Analysis and simulation of multimodal cardiac images to study the heart function*. Theses, Université Nice Sophia Antipolis, January 2013. (Cited on pages 3 and 7.)

[Prakosa 2013b] Adityo Prakosa, Maxime Sermesant, Hervé Delingette, Stéphanie Marchesseau, Eric Saloux, Pascal Allain, Nicolas Villain and Nicholas Ayache. *Generation of synthetic but visually realistic time series of cardiac images combining a biophysical model and clinical images*. Medical Imaging, IEEE Transactions on, vol. 32, no. 1, pages 99–109, 2013. (Cited on page 7.)

[Quiñonero-Candela 2005] Joaquin Quiñonero-Candela and Carl Edward Rasmussen. *A unifying view of sparse approximate Gaussian process regression*. The Journal of Machine Learning Research, vol. 6, pages 1939–1959, 2005. (Cited on pages 59, 74 and 80.)

[Rasmussen 2005] Carl Edward Rasmussen and Joaquin Quinonero-Candela. *Healing the relevance vector machine through augmentation*. In Proceedings of the 22nd international conference on Machine learning, pages 689–696. ACM, 2005. (Cited on pages 59, 74 and 80.)

[Richard 2009] Frédéric JP Richard, Adeline MM Samson and Charles A Cuénod. *A SAEM algorithm for the estimation of template and deformation parameters in medical image sequences*. Stat Comput, vol. 19, no. 4, 2009. (Cited on pages 7, 24, 27, 59 and 72.)

[Risholm 2013] Petter Risholm, Firdaus Janoos, Isaiah Norton, Alex J Golby and William M Wells III. *Bayesian characterization of uncertainty in intra-subject non-rigid registration*. Medical image analysis, vol. 17, no. 5, pages 538–555, 2013. (Cited on pages 9, 24, 25, 58 and 64.)

[Roberts 1996] Gareth O Roberts and Richard L Tweedie. *Exponential convergence of Langevin distributions and their discrete approximations*. Bernoulli, pages 341–363, 1996. (Cited on page 66.)

[Roberts 2006] Gareth O Roberts and Jeffrey S Rosenthal. *Harris recurrence of Metropolis-within-Gibbs and trans-dimensional Markov chains*. The Annals of Applied Probability, pages 2123–2139, 2006. (Cited on page 68.)

[Rohde 2003] Gustavo K Rohde, Akram Aldroubi and Benoit M Dawant. *The adaptive bases algorithm for intensity-based nonrigid image registration*. Medical Imaging, IEEE Transactions on, vol. 22, no. 11, pages 1470–1479, 2003. (Cited on page 25.)

[Rohlfing 2003] Torsten Rohlfing, Calvin R Maurer Jr, David Bluemke, Michael Jacobs*et al. Volume-preserving nonrigid registration of MR breast images using free-form deformation with an incompressibility constraint*. Medical Imaging, IEEE Transactions on, vol. 22, no. 6, pages 730–741, 2003. (Cited on page 8.)

[Rohr 2003] K Rohr, M Fornefett and H.S Stiehl. *Spline-based elastic image registration: integration of landmark errors and orientation attributes*. Computer Vision and Image Understanding, vol. 90, no. 2, pages 153 – 168, 2003. (Cited on page 37.)

[Rublee 2011] Ethan Rublee, Vincent Rabaud, Kurt Konolige and Gary Bradski. *ORB: an efficient alternative to SIFT or SURF*. In Computer Vision (ICCV), 2011 IEEE International Conference on, pages 2564–2571. IEEE, 2011. (Cited on page 8.)

[Rueckert 1999] Daniel Rueckert, Luke I Sonoda, Carmel Hayes, Derek LG Hill, Martin O Leach and David J Hawkes. *Nonrigid registration using free-form deformations: application to breast MR images*. IEEE Transactions on Medical Imaging, vol. 18, no. 8, pages 712–721, 1999. (Cited on pages 31 and 61.)

[Rueckert 2006] Daniel Rueckert, Paul Aljabar, Rolf A Heckemann, Joseph V Hajnal and Alexander Hammers. *Diffeomorphic registration using B-splines*. In Medical Image Computing and Computer-Assisted Intervention–MICCAI 2006, pages 702–709. Springer, 2006. (Cited on page 8.)

[Sabuncu 2011] Mert R Sabuncu and Koen Van Leemput. *The Relevance Voxel Machine (RVoxM): a Bayesian method for image-based prediction*. In Medical Image

Computing and Computer-Assisted Intervention–MICCAI 2011, pages 99–106. Springer, 2011. (Cited on page 26.)

[Sabuncu 2012] Mert R Sabuncu and Koen Van Leemput. *The relevance voxel machine (rvoxm): A self-tuning bayesian model for informative image-based prediction*. Medical Imaging, IEEE Transactions on, vol. 31, no. 12, pages 2290–2306, 2012. (Cited on page 80.)

[Schniter 2008] Philip Schniter, Lee C Potter and Justin Ziniel. *Fast Bayesian matching pursuit*. In Information Theory and Applications Workshop, 2008, pages 326–333. IEEE, 2008. (Cited on page 75.)

[Schölkopf 2002] B. Schölkopf and A.J. Smola. Learning with kernels: Support vector machines, regularization, optimization, and beyond. the MIT Press, 2002. (Cited on page 16.)

[Shi 2012] Wenzhe Shi, Xiahai Zhuang, Luis Pizarro, Wenjia Bai, Haiyan Wang, Kai-Pin Tung, Philip Edwards and Daniel Rueckert. *Registration Using Sparse Free-Form Deformations*. MICCAI, pages 659–666, 2012. (Cited on page 8.)

[Shi 2013] Wenzhe Shi, Martin Jantsch, Paul Aljabar, Luis Pizarro, Wenjia Bai, Haiyan Wang, Declan O'Regan, Xiahai Zhuang and Daniel Rueckert. *Temporal sparse free-form deformations*. Medical Image Analysis, vol. 17, no. 7, pages 779–789, 2013. (Cited on pages 8 and 33.)

[Simpson 2012] Ivor JA Simpson, Julia A Schnabel, Adrian R Groves, Jesper LR Andersson and Mark W Woolrich. *Probabilistic inference of regularisation in non-rigid registration*. NeuroImage, vol. 59, no. 3, pages 2438–2451, 2012. (Cited on pages 9, 24, 25, 29, 30, 34, 58, 61 and 79.)

[Simpson 2013] Ivor JA Simpson, Mark W Woolrich, Manuel Jorge Cardoso, David M Cash, Marc Modat, Julia A Schnabel and Sebastien Ourselin. *A Bayesian Approach for Spatially Adaptive Regularisation in Non-rigid Registration*. Medical Image Computing and Computer-Assisted Intervention–MICCAI 2013, pages 10–18, 2013. (Cited on pages 7, 9, 27, 29 and 58.)

[Simpson 2015] IJA Simpson, MJ Cardoso, M Modat, DM Cash, MW Woolrich, JLR Andersson, JA Schnabel, S Ourselin, Alzheimer's Disease Neuroimaging Initiative*et al. Probabilistic Non-Linear Registration with Spatially Adaptive Regularisation*. Medical image analysis, 2015. (Cited on page 25.)

[Singh 2015] Nikhil Singh, François-Xavier Vialard and Marc Niethammer. *Splines for diffeomorphisms*. Medical image analysis, 2015. (Cited on page 8.)

[Sommer 2013] Stefan Sommer, Mads Nielsen, Sune Darkner and Xavier Pennec. *Higher-order momentum distributions and locally affine LDDMM registration*. SIAM Journal on Imaging Sciences, vol. 6, no. 1, pages 341–367, 2013. (Cited on pages 8 and 79.)

[Sotiras 2013] Aristeidis Sotiras, Christos Davatzikos and Nikos Paragios. *Deformable medical image registration: A survey*. IEEE Transactions on Medical Imaging, vol. 32, no. 7, pages 1153–1190, 2013. (Cited on pages 7 and 31.)

[Stefanescu 2004] Radu Stefanescu, Xavier Pennec and Nicholas Ayache. *Grid powered nonlinear image registration with locally adaptive regularization*. Medical image analysis, vol. 8, no. 3, pages 325–342, 2004. (Cited on page 25.)

[Stewart 2003] Charles V Stewart, Chia-Ling Tsai and Badrinath Roysam. *The dual-bootstrap iterative closest point algorithm with application to retinal image registration*. Medical Imaging, IEEE Transactions on, vol. 22, no. 11, pages 1379–1394, 2003. (Cited on page 25.)

[Stuart 2010] Andrew M Stuart. *Inverse problems: a Bayesian perspective*. Acta Numerica, vol. 19, pages 451–559, 2010. (Cited on page 78.)

[Sundar 2009] Hari Sundar, Christos Davatzikos and George Biros. *Biomechanically-constrained 4D estimation of myocardial motion*. In Medical Image Computing and Computer-Assisted Intervention–MICCAI 2009, pages 257–265. Springer, 2009. (Cited on page 5.)

[Taron 2009] Maxime Taron, Nikos Paragios and M-P Jolly. *Registration with uncertainties and statistical modeling of shapes with variable metric kernels*. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 31, no. 1, pages 99–113, 2009. (Cited on page 9.)

[Thirion 1998] J-P Thirion. *Image matching as a diffusion process: an analogy with Maxwell's demons*. Medical image analysis, vol. 2, no. 3, pages 243–260, 1998. (Cited on pages 8 and 37.)

[Tibshirani 1996] Robert Tibshirani. *Regression shrinkage and selection via the lasso*. Journal of the Royal Statistical Society. Series B (Methodological), pages 267–288, 1996. (Cited on page 54.)

[Tipping 2001] Michael E Tipping. *Sparse Bayesian learning and the relevance vector machine*. The Journal of Machine Learning Research, vol. 1, pages 211–244, 2001. (Cited on pages 25, 26, 34, 36, 59, 70 and 80.)

[Tipping 2003] Michael E Tipping, Anita C Faul*et al.* *Fast marginal likelihood maximisation for sparse Bayesian models*. In Workshop on artificial intelligence and statistics, volume 1. Jan, 2003. (Cited on pages 25, 26, 36, 54, 59, 70, 80, 84, 85 and 86.)

[Tipping 2005] Michael E Tipping and Neil D Lawrence. *Variational inference for Student-t models: Robust Bayesian interpolation and generalised component analysis*. Neurocomputing, vol. 69, no. 1, pages 123–141, 2005. (Cited on page 27.)

[Tobon-Gomez 2013] Catalina Tobon-Gomez, Mathieu De Craene, Kristin McLeod, Lennart Tautz, Wenzhe Shi, Anja Hennemuth, Adityo Prakosa, H Wang, Gerry

Carr-White, Stam Kapetanakis*et al.* *Benchmarking framework for myocardial tracking and deformation algorithms: An open access database*. Medical Image Analysis, 2013. (Cited on pages 9, 49 and 51.)

[Vaillant 2005] M. Vaillant and J. Glaunes. *Surface matching via currents*. In Christensen, G.E., Sonka, M. (eds.) IPMI 2005. LNCS, vol. 3565, pages 381–392. Springer, Heidelberg, 2005. (Cited on page 12.)

[Wachinger 2014] Christian Wachinger, Polina Golland, Martin Reuter and William Wells. *Gaussian Process Interpolation for Uncertainty Estimation in Image Registration*. In Medical Image Computing and Computer-Assisted Intervention–MICCAI 2014, pages 267–274. Springer, 2014. (Cited on page 30.)

[Wells III 1996] William M Wells III, Paul Viola, Hideki Atsumi, Shin Nakajima and Ron Kikinis. *Multi-modal volume registration by maximization of mutual information*. Medical Image Analysis, vol. 1, no. 1, pages 35–51, 1996. (Cited on pages 7 and 27.)

[Wipf 2008] David P Wipf and Srikantan S Nagarajan. *A new view of automatic relevance determination*. In Advances in neural information processing systems, pages 1625–1632, 2008. (Cited on page 34.)

[Xi 2011] J. Xi, P. Lamata, J. Lee, P. Moireau, D. Chapelle and N. Smith. *Myocardial transversely isotropic material parameter estimation from in-silico measurements based on a reduced-order unscented Kalman filter*. Journal of the Mechanical Behavior of Biomedical Materials, 2011. (Cited on pages 6 and 78.)

[Xiang, Y. 2012] Xiang, Y., Gubian, S., Suomela, B. and Hoeng, J. *Generalized Simulated Annealing for Efficient Global Optimization: the GenSA Package for R.* The R Journal, 2012. Forthcoming. (Cited on page 17.)

[Zhang 2011] Yichuan Zhang and Charles A Sutton. *Quasi-Newton methods for Markov chain Monte Carlo*. In Advances in Neural Information Processing Systems, pages 2393–2401, 2011. (Cited on pages 66 and 76.)

[Zhang 2013] Miaomiao Zhang, Nikhil Singh and P Thomas Fletcher. *Bayesian estimation of regularization and atlas building in diffeomorphic image registration*. In Information Processing in Medical Imaging, pages 37–48. Springer, 2013. (Cited on pages 59 and 79.)

[Zhou 2008] Ding-Xuan Zhou. *Derivative reproducing properties for kernel methods in learning theory*. Journal of computational and Applied Mathematics, vol. 220, no. 1, pages 456–463, 2008. (Cited on page 82.)

[Zhou 2015] Yitian Zhou, Olivier Bernard, Eric Saloux, Alain Manrique, Pascal Allain, Sherif Makram-Ebeid and Mathieu De Craene. *3D harmonic phase tracking with anatomical regularization*. Medical Image Analysis, vol. 26, no. 1, pages 70 – 81, 2015. (Cited on page 9.)

[Zou 2005] Hui Zou and Trevor Hastie. *Regularization and variable selection via the elastic net*. Journal of the Royal Statistical Society: Series B (Statistical Methodology), vol. 67, no. 2, pages 301–320, 2005. (Cited on page 54.)