THE UNIVERSITY OF QUEENSLAND

AUSTRALIA

Maximum pseudolikelihood estimation with Markov random fields in the segmentation of brain magnetic resonance images

Amy Chan

Bachelor of Mathematics (Hons I)

A thesis submitted for the degree of Doctor of Philosophy at
The University of Queensland in 2018

School of Mathematics and Physics

Abstract

Having an accurate model of the tissue structure of the brain is useful in studying the development and progression of neurodegenerative diseases like dementia. Brain magnetic resonance images (MRI) can be used to create such models, by detecting the tissue boundaries in the image and classifying each voxel as a particular tissue. This is task known as image segmentation.

Many segmentation methods use a mixture-Markov random field probabilistic model for the image intensities, which can then be used to determine the most likely segmentation of the image. This model consists of normal distribution for the image intensities of each tissue, and a Markov random field (MRF) for the prior distribution of tissue labels. The purpose of the MRF is to incorporate spatial dependence between the labels of neighbouring voxels, adding smoothness to the segmentation to remove noise.

In this thesis, we develop and validate methods to perform automatic tissue segmentation of brain MRI. We specifically focus on the MRF component of the image model. This is used to model spatial dependence between neighbouring tissue labels, which has the effect of spatial regularisation on the segmentations.

This thesis begins with a general introduction to mixture models, Markov random fields, and the combined mixture-MRF model as it applies to image segmentation. Estimation of the tissue intensity parameters and also of the segmentation using Expectation-Maximisation (EM) is explained, as well as effective approximations required to accommodate the MRF's intractable normalising constant.

First, the homogeneous Potts MRF is introduced. It is used ubiquitously for MRI segmentation. The Potts model has one parameter that controls the strength of the MRF compared to the normal intensity probabilities, and hence the smoothness of the resulting segmentation. In the literature and in practice, this parameter is almost always fixed to a value chosen by manual tuning or with the use of training data. When no training data is available, selection of an appropriate parameter value is subjective and can affect the accuracy of the segmentation. We propose use of the maximum pseudolikelihood estimator (Besag, 1986) to automatically determine the value of this parameter and show how to incorporate it into the EM algorithm. The proposed method adaptively determines the amount of spatial regularisation on a per-image basis, without needing training data or an anatomical atlas. The maximum pseudolikelihood estimator (MPLE) is statistically consistent. It is also computationally tractable, involving only univariate maximisation of a concave function, and a straightforward extension of EM. The proposed method is demonstrated on real brain MRI and compared to various existing methods that require manual specification of the smoothing parameter. It is also compared to the least-squares method of Derin and Elliott

(1987) which has previously been used to automatically determine the smoothing parameter by Van Leemput et al. (1999b). The MPLE produces segmentations that are comparable or significantly more accurate than these.

Next, the image model is extended to use the non-homogeneous Potts MRF, which has not been studied in detail for tissue segmentation. While the homogeneous Potts MRF has one parameter that controls global smoothness, the non-homogeneous MRF has multiple pairwise parameters that allow different smoothness constraints depending on the specific neighbouring tissues. The MRF additionally has unary parameters allowing for tissue-specific prior information to be incorporated. The role of each of these parameters is studied in isolation and together. The previously proposed MPLE is applied to this MRF to automatically determine the parameters. The method is applied to real brain images. Model selection using pseudolikelihood information criterion (Forbes and Peyrard, 2003) suggests that the MRF with smoothing parameters but without unary parameters is favoured. However, segmentation accuracy suggests that the non-homogeneous Potts MRF (with various combinations of unary and smoothing parameters) is not more beneficial than the homogeneous Potts MRF. A review of similar MRFs in the literature suggests that the use of prior anatomical knowledge is required to constrain the parameters of the non-homogeneous Potts MRF to tailor it for brain segmentation. Leaving all parameters free to be estimated can lead to oversmoothing, particularly if a given tissue boundary is relatively rare compared to others.

Finally, the image model is extended to consider anisotropic MRFs. Based on the Potts MRF, these allow for smoothing that can incorporate local features of the image in addition to the tissue labels. Drawing from the principles of Perona-Malik diffusion (Perona and Malik, 1990), a model is designed and proposed to smooth the segmentation tangentially along a detected edge but not across it, with strength proportional to the detected edge strength. Similar anisotropic MRFs have been used for tissue segmentation before, but are discriminative models requiring training data and different solution methods. The proposed model is generative and may be estimated using Expectation-Maximisation and maximum pseudolikelihood estimation, thus requiring no training. The model MRF and two variants are applied to brain MRI, and their segmentation accuracy compared to the homogeneous Potts MRF. The two supplementary MRFs underperform the homogeneous Potts MRF but demonstrate that the anisotropy is being appropriately applied. The proposed MRF significantly outperforms the homogeneous Potts MRF and demonstrates anisotropic smoothing as intended. Suggestions are made to further improve the framework and MRF to make better use of the local image structure.

In summary the thesis comprises two main directions of research. First, automatic determination of MRF parameters in the mixture-Markov random field framework may be achieved in a computationally tractable manner using maximum pseudolikelihood, avoiding poor segmentations due to manual specification of the spatial parameter. Second, different MRFs allow for finer control of smoothing on a tissue-specific or even more local neighbourhood-specific level, and when properly specified may improve segmentation accuracy.

Declaration by author

This thesis is composed of my original work, and contains no material previously published or written by another person except where due reference has been made in the text. I have clearly stated the contribution by others to jointly-authored works that I have included in my thesis.

I have clearly stated the contribution of others to my thesis as a whole, including statistical assistance, survey design, data analysis, significant technical procedures, professional editorial advice, financial support and any other original research work used or reported in my thesis. The content of my thesis is the result of work I have carried out since the commencement of my higher degree by research candidature and does not include a substantial part of work that has been submitted to qualify for the award of any other degree or diploma in any university or other tertiary institution. I have clearly stated which parts of my thesis, if any, have been submitted to qualify for another award.

I acknowledge that an electronic copy of my thesis must be lodged with the University Library and, subject to the policy and procedures of The University of Queensland, the thesis be made available for research and study in accordance with the Copyright Act 1968 unless a period of embargo has been approved by the Dean of the Graduate School.

I acknowledge that copyright of all material contained in my thesis resides with the copyright holder(s) of that material. Where appropriate I have obtained copyright permission from the copyright holder to reproduce material in this thesis and have sought permission from co-authors for any jointly authored works included in the thesis.

Publications during candidature

Conference papers

Chan, A., Wood, I. A., and Fripp, J. (2016). Maximum Pseudolikelihood Estimation for Mixture-Markov Random Field Segmentation of the Brain. In *2016 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, pages 1–7. IEEE

Journal papers

Boyd, R., George, J., Fripp, J., Panneck, K., Chan, A., Fiori, S., Guzzetta, A., Ware, R., Rose, S., and Colditz, P. (2015). Relationship between early brain structure on Mri, white matter integrity (diffusion Mri) and neurological function at 30 weeks post menstrual age in infants born very preterm. *Developmental Medicine & Child Neurology*, 57:8–9

Lai, M., D'Acunto, G., Guzzetta, A., Fripp, J., Chan, A., Rose, S., Ngenda, N., Whittingham, K., Colditz, P., and Boyd, R. (2015). Randomised controlled trial of PREMM: Early somatosensory stimulation (massage) in preterm infants. *Developmental Medicine & Child Neurology*, 57:94–95

George, J., Fripp, J., Shen, K., Pannek, K., Chan, A., Ware, R., Rose, S., Colditz, P., and Boyd, R. (2015). Relationship between white matter integrity and neurological function in preterm infants at 30 weeks postmenstrual age. *Developmental Medicine & Child Neurology*, 57:88–89

George, J., Fripp, J., Shen, K., Pannek, K., Chan, A., Ware, R., Rose, S., Colditz, P., and Boyd, R. (2016). Relationship between white matter integrity at 3T Mri and neurological function in preterm infants at 30 weeks postmenstrual age. *Developmental Medicine & Child Neurology*, 58:33–34

Publications included in this thesis

Chan, A., Wood, I. A., and Fripp, J. (2016). Maximum Pseudolikelihood Estimation for Mixture-Markov Random Field Segmentation of the Brain. In *2016 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, pages 1–7. IEEE Incorporated as a part of Chapter 3.

Contributor	Statement of contribution
Amy Chan (Candidate)	Conception and design (70%) Analysis and interpretation (80%) Drafting and production (80%)
Ian Wood	Conception and design (15%) Analysis and interpretation (10%) Drafting and production (10%)
Jurgen Fripp	Conception and design (15%) Analysis and interpretation (10%) Drafting and production (10%)

Contributions by others to the thesis

All chapters were written entirely by the candidate, with editorial advice provided by Dr Ian Wood and Dr Jurgen Fripp. The research in this thesis was developed under the guidance and suggestions of Dr Ian Wood and Dr Jurgen Fripp.

Statement of parts of the thesis submitted to qualify for the award of another degree

None.

Research Involving Human or Animal Subjects

No animal or human subjects were involved in this research.

Acknowledgements

Completing my doctorate has been simultaneously one of the best and worst times of my life. I have in turns enjoyed and lamented the student lifestyle. I felt both completely lost in a sea of knowledge, and the satisfaction of coming up with a reasonable idea and realising you've gained enough knowledge to do so.

This thesis offered an interesting challenge, being at the intersection of statistics and imaging. It has been at times frustrating to undertake a thesis straddling two fields, needing to constantly learn more about each, running back and forth between the two, in order to gain proficiency in the middle ("Jack of all trades, master of none"!). Yet it has been very rewarding to finally reach that middle and discover that it is a mastery in and of itself. Taking techniques from statistics with various theoretical plaudits and adapting them to practical use in image segmentation has taught me much about compromise, and renewed my sense of wonder at the ability of mathematics and statistics to transcend decades and technologies.

This thesis would not have been possible without my supervisors. I thank Dr Ian Wood for many hours of discussion not only on my research, but also on how to develop as a researcher and writer, how to manage the lack of motivation and writer's block that all students experience, and being generally good to talk to. I appreciate the shared laughs from many of our meetings. Thank you for your willingness to learn about imaging and brains, and to stay up to date as I went further and further into narrower and narrower areas of research. Thank you also for helping me through my "mid-thesis crisis", without which this thesis probably would not exist.

I thank Dr Jurgen Fripp for his willingness to learn about the nitty gritty of parameter estimation in Markov random fields, sometimes at the expense of losing sight of the practical application. Your reminders to pull back and think about context within the big picture when I get stuck on tiny small details are much appreciated. I greatly value your vast store of knowledge on the many different imaging techniques out there.

I also thank Dr Geoffrey McLachlan for giving me a good grounding in statistics and offering helpful advice and suggestions in the early days of my doctorate.

One thing I have learned while writing this thesis is the value of community. To my fortnightly Friday board games group, to the folk of the Cactus and Succulent Society QLD, the trivia gang, my fellow students (many now post-docs) and jigsaw crew at the AeHRC, BGSC, QUGS, the Royal Raiders, and #nethack on freenode - you have all been valuable sources of friendship, support and encouragement.

I thank my friends and family – Andrew, Milly, Jess, and too many others to list – for offering support, commiseration and encouragement, and also for your patience with me for missing out on so many occasions particularly in the last year, due to "I have to work on my thesis".

Last but not least thanks to the DevTeam[1], NAO, and Keldon's AI - many, many happy hours squandered.

[1] who think of everything

Financial support

This research was supported by an Australian Government Research Training Program Scholarship and the Commonwealth Scientific Industrial Research Organisation.

Keywords

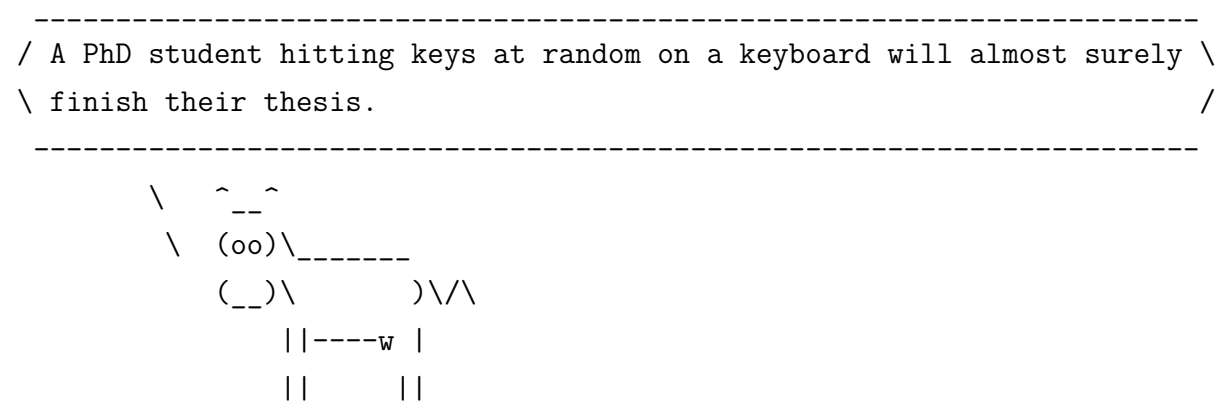
Markov random field, image segmentation, magnetic resonance imaging, pseudolikelihood, Potts, mixture models, expectation-maximisation

Australian and New Zealand Standard Research Classifications (ANZSRC)

ANZSRC code: 010406, Applied Statistics, 50%
ANZSRC code: 080106, Image Processing, 50%

Fields of Research Classification

FoR code: 0104, Statistics, 50%
FoR code: 0801, Artificial Intelligence & Image Processing, 50%


A PhD student hitting keys at random on a keyboard will almost surely
finish their thesis.
^__^
(oo)_______
(__)\ )\/\
||----w |
|| ||

Contents

List of Figures

List of Tables

List of Abbreviations used in the thesis

CSF	cerebrospinal fluid
EM	Expectation-Maximisation
GM	grey matter
ICE, ICM	Iterated Conditional Modes/Expectation
LS, LSE	least-squares (estimator)
MF	mean-field
MRF	Markov random field
MRI	magnetic resonance images
MPL, MPLE	maximum pseudolikelihood (estimator)
pdf	probability density function
PL	pseudolikelihood
WM	white matter

List of Notation used in the thesis

Indices

i	index used for voxel i
m	index used for voxel m, usually a neighbour of i
$\boldsymbol{\partial i}$	indices in the neighbourhood of voxel i
n	total number of voxels
j, k	indices used for tissue labels
g	total number of tissues

Variables

$\boldsymbol{Y}_i, Y_i, \boldsymbol{y}_i, y_i$	random variable (uppercase) and realisation (lowercase) of intensity at voxel i, either vector or scalar
$\boldsymbol{Y}, \boldsymbol{y}$	intensities for all voxels, i.e. $(y_1, y_2, \ldots, y_n)$
$\boldsymbol{Z}_i, \boldsymbol{z}_i$	random variable and realisation of tissue labels; $\boldsymbol{z}_i \in \{0,1\}^g$ such that $\sum_{j=1}^{g} z_{ij} = 1$
z_{ij}	the jth element of $\boldsymbol{z}_i$
$\langle \boldsymbol{z}_i \rangle$	mean-field approximation of $\boldsymbol{z}_i$
$\boldsymbol{Z}, \boldsymbol{z}$	random variable and realisation of intensity at voxel i
$\boldsymbol{e}_j$	an indicator vector of length g that is 0 everywhere except the jth element, which is 1
$\boldsymbol{z}_{\boldsymbol{\partial i}}$	labels in the neighbourhood of voxel i, i.e. $\boldsymbol{z}_m$ such that $m \in \boldsymbol{\partial i}$

Distributions

f	usually a density function involving continuous variables such as $\boldsymbol{y}$ or y_i
ϕ	the normal probability density function
$\boldsymbol{\Theta}$	parameters of a the intensity distribution
μ_j, σ_j^2	mean and standard deviation of the normal distribution corresponding to the jth tissue
p	usually a probability function over discrete variables such as $\boldsymbol{z}$ or $\boldsymbol{z}_i$
$\tilde{p}$	pseudolikelihood or mean-field approximation of p
$U_i(\boldsymbol{z}_i \vert \boldsymbol{z}_{\boldsymbol{\partial i}})$	MRF potential of voxel i
C	normalising constant of $p(\boldsymbol{z})$
u_{ij}	number of neighbours of voxel i with label j, i.e. $\sum_{j=1}^{g} z_{ij}$
δ_{im}	distance between voxels i and m
$\boldsymbol{\Psi}$	parameters of the Markov random field
β	single smoothing parameter of the homogeneous Potts model
β_{jk}	multiple smoothing parameters of the non-homogneous Potts model, applying to the boundary between tissues j and k
$\boldsymbol{B}$	$g \times g$ matrix with element (j,k) being β_{jk} when $j \neq k$; $\beta_{kj} = \beta_{jk}$ and $\beta_{jj} = 0$
$\boldsymbol{\alpha}, \alpha_j$	unary parameters of the non-homogeneous Potts model such that $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_g)$

Miscellaneous

$\mathbb{E}\,[]$	expectation
Q	Q-function in Expectation-Maximisation
τ_{ij}	posterior probability that voxel i belongs to tissue j
$x^{(t)}$	the quantity x on iteration t

Chapter 1

Introduction

As we develop and then age, our brain changes continuously. Various substructures within the brain may change in shape or size as part of healthy aging (Dennis and Thompson, 2013). On the other hand, the presence of neurodegenerative diseases such as dementia can also affect the brain. Having an accurate model of the brain is vital to studying the progression of such diseases and how they differ from the processes of normal aging.

Clinically, qualitative measures are commonly used to diagnose and assess the severity of neurological disorders. These are both time consuming and require and are dependent on rater expertise/experience. With the rapid development and improvement of medical technology, more accurate diagnoses can be found by including biomarkers from cerebrospinal fluid analyses and images obtained by magnetic resonance imaging and positron emission tomography (Dubois et al., 2007). Quantitative measurement of, for example, the volume of the hippocampus (Jack et al., 1997, 2000; Schuff et al., 2009) or the thickness of the cortical wall (Thompson et al., 2003) has the potential to better characterise the nature of dementia.

The challenge is obtaining accurate measurements of the brain and then accurately modelling it. These models are constructed from medical images of the brain, such as magnetic resonance images (MRI). Automated measurement of structures in the brain from MRI saves both time and the need for a fully-trained expert, and can be highly reliable (Han et al., 2006).

Before such measurements can be taken, a reliable reconstruction of the brain from the MRI is needed. In the brain, there are three main tissue types - cerebrospinal fluid (CSF), grey matter (GM) and white matter (WM). Classifying each spatial location of the MRI to the underlying tissue type being imaged there is known as a *tissue segmentation* of the brain. The underlying tissue type at a given location may be inferred from the observed signal of the MRI there, as well as prior anatomical knowledge. The task is made more difficult by the presence of artefacts that can degrade the quality of the image, for example scanner noise, the machine's bias field, or patient motion. An example of an MRI and corresponding segmentation are shown in figure 1.1.

Segmentation of brain MRI is the primary focus of this thesis. In particular, we are interested in

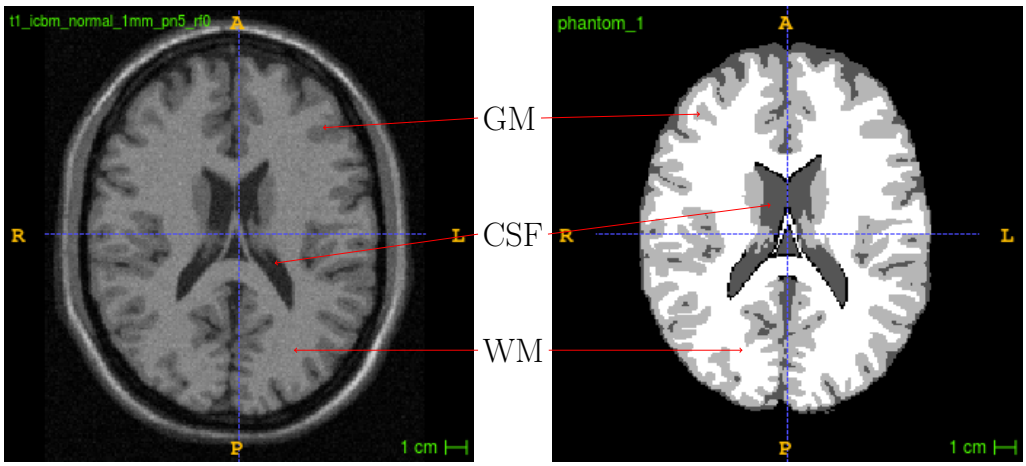


Figure 1.1: Brain tissue is primarily grey matter (GM), white matter (WM) or cerebrospinal fluid (CSF). Left: an MRI. Right: corresponding tissue classification.

the use of probability models of brain MRI for segmentation, and the incorporation of adaptive spatial regularisation into these models.

1.1 Background

1.1.1 Magnetic resonance imaging

An MRI machine has a strong static magnetic field in which the object to be imaged is placed, causing the nuclear spins of the object to become aligned. This mostly corresponds to hydrogen atoms in water present in the body. A radiofrequency field is briefly applied in the transverse plane to the static field, causing the spins to align with it. When this field is removed, the protons precess or relax back to their equilibrium position, producing a signal that is detected. The static magnetic field has a physical gradient in field strength, allowing the physical location of the signal to be inferred. In this way a signal is recorded from a dense grid of spatial locations within and around the object to be imaged.

The image itself may be viewed as a set of measurements on a regular (square or cubic) grid, either in 2D or 3D. Each cell of the grid is known as a *pixel* for a 2D image, or a *voxel* for a 3D image (also called a volume). For example, colour images may have 3 integer values at each pixel, being red, green and blue intensity values. For an MRI, each voxel contains the signal strength at the corresponding point of real space. Different tissues have different relaxation times, allowing them to be distinguished in the image. An MRI is typically a 3D *volume*, consisting of many 2D *slices* (figure 1.2).

Depending on the imaging sequence used in the MRI, different types of image can be produced. For example, a T1-weighted image shows CSF as the darkest tissue, white matter as the brightest, and grey matter intermediate. In a T2-weighted image, CSF is the brightest tissue, grey matter

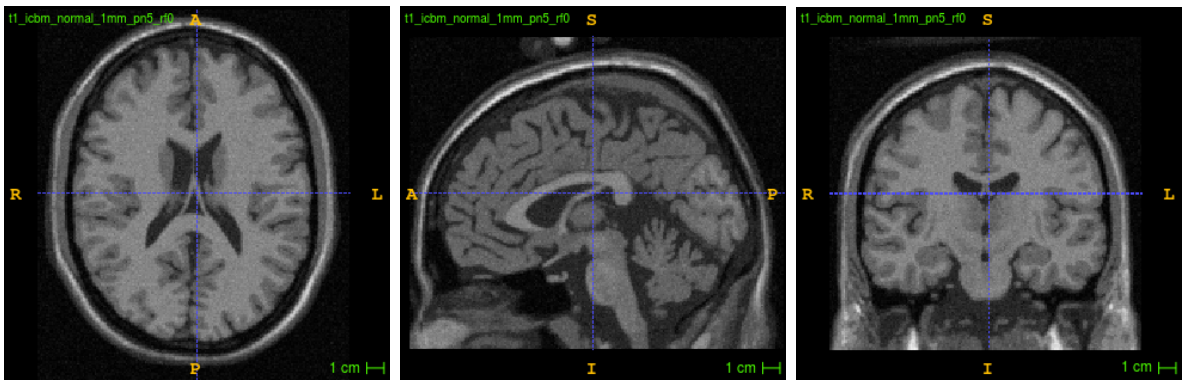


Figure 1.2: Example slices of an MRI of the brain

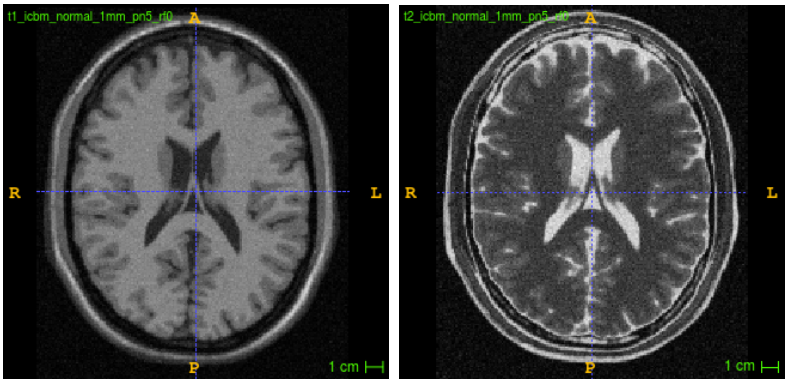


Figure 1.3: T1 image (left) vs T2 image (right). In a T1 image, CSF is the darkest tissue, followed by grey matter with white matter as the brightest. The order is reversed in a T2 image.

grey, and white matter dark (figure 1.3).

1.1.2 Brain MRI segmentation

A segmentation of the brain may mean multiple things:

- a *tissue segmentation* of the brain into major tissues such as cerebrospinal fluid, grey matter, and white matter, dura, glial tissue, and so on.
- an *anatomical segmentation* of the brain into finer anatomical regions such as the hippocampus, ventricles, thalamus, and so on.

This can be done manually, with semi-automatic, and fully-automatic methods. See (Despotović et al., 2015; Balafar et al., 2010; Withey and Koles, 2008) for reviews.

Manual segmentation involves a technician or expert manually delineating the regions of interest on the MRI. This can be extremely time-consuming. While manual segmentations are often used as a 'gold standard', they can be subjective and suffering from poor inter- and intra-rater reliability (Clarke et al., 1995; Collier et al., 2003). For example, Gurleyik and Haacke (2002) reported an inter-observer error of as large as 16% for five experts performing manual segmentation on the caudate nucleus. Even for one well-trained expert, segmentations can

differ significantly depending on what segmentation protocol is used (Boccardi et al., 2011). Additionally, it can take many man-hours to manually delineate tissues in each subject. Using semi-automatic or fully-automatic methods has the advantage of restoring objectivity to the segmentations while also saving manual labour.

Semi-automatic segmentation is that which is mostly automatic, but requires some manual input. For example, a technician could click on regions of the MRI they know to be grey matter, white matter, and CSF, and these are used to initialise a segmentation algorithm. These approaches can greatly increase reliability of the segmentations (Yushkevich et al., 2006). However, methods like this still require manual input, though much less than a full manual segmentation.

Related to semi-automatic segmentation methods are those that are fully automatic to run, but require training data. Examples of these methods include neural networks and deep learning-based approaches (Zhang et al., 2015; Moeskops et al., 2016; Litjens et al., 2017; Shen et al., 2017, and the references therein). These methods can offer very promising results and a good compromise between the accuracy of manual segmentation, and the convenience and objectiveness of automatic segmentation. However, such methods still require training data, typically comprised from images and their matching manual segmentations. They can fail if a test image is presented that is significantly different from the training data (e.g. in image contrast, or brain morphology). There is a need for methods that do not require manual intervention, or extensive training data.

In this thesis, we focus on fully-automatic segmentation methods that do not require training data. Automatic methods may largely be classified into those that use prior anatomical knowledge (an atlas), and those that do not. However, combinations of these are also often used, and many methods that do not require an atlas may still make use of one if available. An atlas is typically a representative MRI, along with a hard or probabilistic labelling of it into tissues or regions. For example, each voxel may have a probability associated with it to be white matter, grey matter, or cerebrospinal fluid. An unlabelled input MRI is registered to the atlas (possibly non-rigidly), aligning the two brains. The labels are propagated from the atlas onto the registered input brain, which may then be transformed back into the original space. The advantage of atlas-based methods is that the atlas may be labelled in finer detail than could be inferred from the MRI alone. For example, two cortical walls pressed so closely as to appear a single contiguous region based on the MRI alone, could be properly distinguished as two separate walls.

However, the registration may easily fail if the unlabelled brain does not match the atlas closely enough. Some examples of this can be seen in figure 1.4. If the brain to be segmented have pathologies or injuries, there may not exist a mapping between it and the atlas. Another example is the neonate brain, which changes significantly over a short period of time (Rutherford, 2002). If the atlas selected for the brain does not match its current developmental state closely enough, registration will fail. For this reason, we focus on fully-automatic brain MR segmentation that

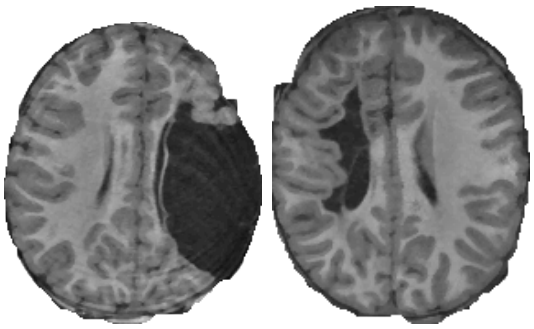


Figure 1.4: Examples of brains with injuries from cerebral palsy

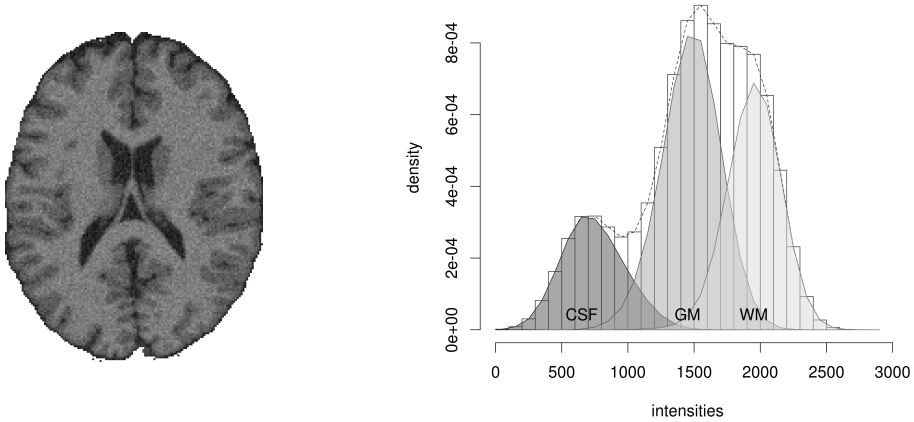


Figure 1.5: MRI and histogram of its voxel intensities - overall and by tissue (from manual segmentation)

does not require an atlas or prior training.

Segmentation methods that do not require an atlas make use of image intensities. Such methods can be edge- and surface-based, for example active contours and level set methods (Tsai et al., 2001; Vese and Chan, 2002). Edges in the image are located with intensity gradient information, and used to define regions of interest. Such methods often also incorporate region-based metrics, identifying regions in the image as having homogeneous intensity within each region (Wang et al., 2009; Huang et al., 2009).

A large number of segmentation methods for the brain focus on the clustering of image intensities. Figure 1.5 shows a T1 MRI and its corresponding intensity histogram, as well as the intensity distribution of each of the three main tissues (determined by a manual segmentation of the image). Since CSF is generally dark, WM is bright, and GM is in between, the most basic approach is simply to threshold the image intensities to determine an image classification. This can be quite subject to noise. More sophisticated clustering methods include k-means (Cocosco et al., 2003; Vrooman et al., 2007) and fuzzy C-means (Ahmed et al., 2002).

However, the most common clustering method for brain tissue segmentation, and the focus of this thesis, employs a Gaussian mixture model of the MRI intensities (we will cover this in further detail in Chapter 2). On examining figure 1.5, the image intensities appear to be well approximated by three overlapping Gaussian distributions, one per tissue type. In fact,

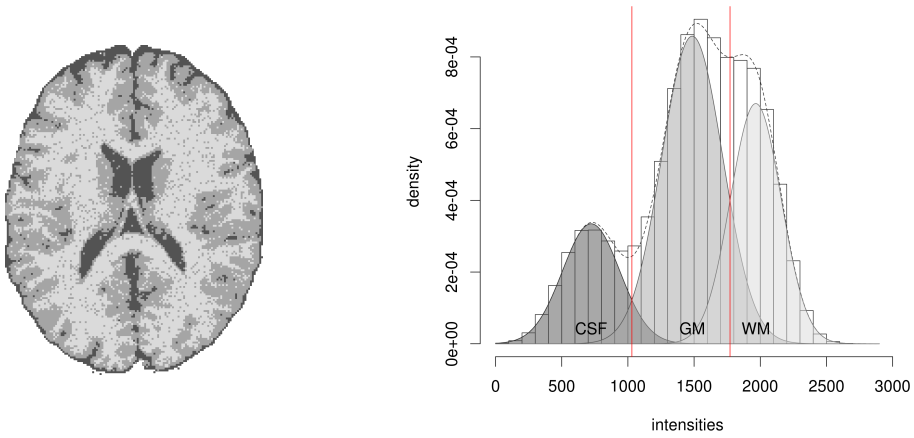


Figure 1.6: EM fit from a Gaussian mixture model and corresponding segmentation

it has been shown that noise in tissue intensities is Rician, but may be approximated by a Gaussian given the signal-to-noise ratio of MRI (Gudbjartsson and Patz, 1995). A mixture of three Gaussians model may be fit to an MRI's intensities, inducing a segmentation of the image by assigning each voxel to the Gaussian it is most likely to belong to.

As can be seen (figure 1.6), this method is susceptible to noise in the image. All voxels with a given intensity will be classified as the same tissue, even if entirely surrounded by a different tissue. To address this, the segmentation can be smoothed. Morphological operators such as openings and closings can be applied to the EM segmentation to remove isolated noise. However, such operators are 'blind' to the image around them, and indiscriminately fill in all features of the same size regardless of the surrounding image data.

An alternative is to incorporate the smoothness constraint into the image model itself, so that smoothing is context-aware of the local intensity information. To do this, it is standard to use a Markov Random Field (MRF) as a prior probability distribution over the tissue labels in the mixture model. The Potts model for atomic spins from statistical mechanics (Potts, 1952) is most commonly used. When considering the probability for a given voxel to be a given tissue conditioned on the tissues of its neighbours, the Potts model prefers the majority tissue in the neighbourhood. The Gaussian distribution of the intensities of each tissue is retained from the standard mixture model. In this way, the smoothing of the Potts model is weighted by the intensity probabilities, so that the smoothing is both intensity- and spatially-dependent. With some modifications, Expectation-Maximisation can be adapted to handle the MRF (we will show these details in Chapter 2).

This model - each tissue distributed according to a Gaussian, and the prior distribution of the tissues with the Potts MRF - is ubiquitous in MR segmentation. Introduced for image segmentation by Besag (1986), it is a component of common segmentation packages such as NiftySeg (Cardoso et al., 2009, 2011), Expectation Maximisation Segmentation (Van Leemput et al., 1999b), Atropos (Avants et al., 2011) and FAST (Zhang et al., 2001). From this basis many extensions may be made; for example, one can add bias-field correction (Van Leemput

(a) 0.1 (b) 0.3 (c) 1 (d) 5 (e) 10

Figure 1.7: Segmentations with various smoothing values β

et al., 1999a; Wells et al., 1996) or partial volume correction (Noe and Gee, 2001; Shattuck et al., 2001; Van Leemput et al., 2003).

The Potts MRF has one non-negative parameter β that controls the strength of the smoothing. This is typically set by manual tuning, or to arbitrarily-chosen values in the literature (for example, $\beta = 1.5$ (Besag, 1986), $\beta = 1$ (Zhang et al., 2001; McLachlan et al., 1996), $\beta = 0.7$ (Owen, 1986; Ripley, 1986), $\beta = 0.3$ (Avants et al., 2011), $\beta = 0.25$ (Cardoso et al., 2009)). Larger values correspond to stronger smoothing, while 0 corresponds to no smoothing. Misspecifying this parameter can lead to not enough smoothing, or too much (figure 1.7). In addition, the value of β that is best for one MRI may not be the same as that for a different MRI. An automatic method to determine the amount of smoothing, i.e. β, on a per-image basis is of value, and could then be used in all methods based on the mixture-MRF formulation.

In addition, the Potts MRF is quite basic, with its single parameter β only allowing for the same uniform smoothness across the entire image. In the brain, it is known that some tissue boundaries are more convoluted than others (e.g. cortical folding vs the ventricle boundary). MRFs that allow smoothing to be applied at different scales, for example on a per-tissue basis, or incorporating further local image features, are worth investigating.

1.2 Aims

This thesis addresses the issues mentioned in the previous section through development of a fully-automatic, three-dimensional brain MR segmentation algorithm that does not require training data. We aim to segment the skull-stripped brain into the three primary tissues, cerebrospinal fluid (CSF), grey matter (GM) and white matter (WM). Our segmentation method is based on constructing a probability model for the MRI, which is then used to classify it into tissues. Specifically, this thesis focuses on the Markov random field (MRF) prior probability distribution over the tissue labels.

The thesis may be viewed as a study of adaptive spatial smoothing of brain MR segmentations

using Markov random fields, where the smoothing is applied on a global, per-tissue, or local level. There are two aspects to this, given below.

1.2.1 Automatic determination of the smoothing parameter

The standard method for probabilistic method uses a Gaussian distribution of intensities for each tissue, and the Potts MRF for the tissue labels. The Potts MRF has one parameter that determines the amount of smoothing to apply to the image, but there is no well-accepted, principled method to determine the value of this parameter automatically, with manual tuning being common.

The first aim of the thesis is to develop methods to automatically and adaptively determine the smoothing parameter for the Potts MRF. The method should be computationally tractable, should not require training data, and should be able to adjust the parameter on a per-image basis.

There have been a number of attempts at automatically setting the smoothing parameter, including Bayesian approaches (Woolrich et al., 2005; Woolrich and Behrens, 2006) and regression-based estimators (Van Leemput et al., 1999b). The Bayesian approaches are computationally intensive and slow, requiring many simulations of the desired Markov random field at each iteration of the algorithm. The regression-based estimator does not have this drawback, but relies on building a neighbourhood histogram of the image which is time-consuming, and suffers additional limitations on its use.

The proposed method utilises the *pseudolikelihood* (Besag, 1975) and *mean-field* (Chandler, 1987) approximations in order to determine a suitable value for the smoothing parameter. The method is computationally tractable and easy to interpret and understand. Additionally, the method is a natural extension of the modified Expectation-Maximisation algorithm already used in existing methods, so does not represent much implementational burden to incorporate into existing methods.

We focus on segmentation of brain tissues only, i.e. we assume that the skull has already been stripped in the images and artefacts such as bias-field already corrected. However, the image model used in this thesis is common to many segmentation algorithms that can also perform these tasks, and the methods developed in this thesis can be readily incorporated into these algorithms.

Fully-automatic estimation of the smoothing parameter in the Potts MRF will be studied in chapter 3. We hypothesise that estimation of the parameter individually for each image will provide more accurate segmentations than setting it to the same fixed value for each image.

1.2.2 Different types of MRF

The second aim of this thesis is to investigate the use of more complex forms of MRF in brain segmentation. As previously mentioned, the Potts MRF can only apply the same smoothing uniformly across the image. It may be advantageous to allow different tissue boundaries to be smoothed to different degrees. For example, the GM-WM boundary of the cortical folds could be permitted to be less smooth than the GM-CSF boundary of the ventricles. Additionally, the standard Potts MRF cannot explicitly account for certain tissues being less prevalent than others. In Chapter 4, we aim to use a more general form of the Potts MRF to enable tissue-specific smoothing and control of tissue proportions. This allows both per-tissue smoothing and relative tissue proportions to be controlled. Additionally, we will use the method developed in the first aim to estimate the parameters of the MRF. We hypothesise that this MRF will allow greater sensitivity to the different tissues when smoothing, and be able to adapt to images where the tissue proportions are very different from each other.

It is also of interest to incorporate MRF smoothing on an even finer scale than the tissue level. For example, local image features such as edge orientation and strength can be used to further adjust the smoothing so as not to smooth away thin features such as the cortical folds. In Chapter 5, we will develop and investigate MRF to achieve anisotropic smoothing, and use the method developed in the first aim to estimate its parameters. We hypothesise use of anisotropic MRFs will prevent thin features from being smoothed away, while still permitting strong smoothing of noise in otherwise homogeneous regions.

1.3 Contributions

The key contributions of the thesis are

1. to demonstrate the effectiveness of MRF parameter estimation (by any method) as opposed to fixing the spatial regularisation parameter to a manually-chosen constant. This results in a fully-automatic intensity-based brain segmentation algorithm with adaptive spatial regularisation. While maximum pseudolikelihood estimation in MRFs has been performed before (e.g. Celeux et al. (2003)), it has not been studied in detail with regards to neighbourhood size, choice of MRF approximation, or the form of the MRF itself when applied to MR segmentation.

2. to specifically demonstrate the suitability of maximum pseudolikelihood estimation for MRF parameter estimation in brain segmentation, as compared to other estimation techniques. Maximum pseudolikelihood is more computationally tractable than existing Bayesian methods (Woolrich et al., 2005; Woolrich and Behrens, 2006). It makes use of quantities already calculated in the the MRF segmentation framework, so is modular and straightforward to implement into methods that already use this framework, unlike

the regression estimator of (Van Leemput et al., 1999b). Finally, results from real brain datasets indicate that using maximum pseudolikelihood to automatically determine the spatial regularisation is superior to the current methods which fix it, especially when no atlas can be used.

3. to explore more complex MRF models (with parameter estimation) and assess their suitability for/tailor them to brain MR segmentation. These MRFs can smooth on finer scales: on a per-tissue basis (as studied in Chapter 4), and secondly, in local neighbourhoods (as studied in Chapter 5). We use maximum pseudolikelihood to adaptively smooth with these MRFs also.

1.4 Overview of the thesis

In chapter 2, we give the necessary mathematical and imaging background that will be used throughout the remaining chapters.

In chapter 3, we consider the most common form of MRF used for brain segmentation, the homogeneous Potts MRF, which requires one global smoothing parameter. We study and compare two methods to automatically determine this parameter, least-squares estimation and maximum pseudolikelihood estimation. We argue that the latter is ideal for brain MR segmentation as it can easily be incorporated into existing methods, as it involves an optimisation that is concave, computationally tractable, and uses quantities already calculated in existing methods. We demonstrate its use on a real-brain dataset and show it has favourable performance compared to existing fixed-parameter methods. Although the method itself is not new, we make a detailed study of how the neighbourhood specification and approximation of the MRF are related to segmentation accuracy. To our knowledge, a study focusing on these aspects has not been presented before. The work of this chapter can be thought of as smoothing on a global (image-wide) level.

In chapter 4, we consider smoothing on a per-tissue level, which may be more realistic for the brain. We do this by considering the non-homogeneous Potts MRF, a generalisation of the homogeneous Potts MRF of Chapter 3. We show how to automatically determine the model parameters, comparing least-squares estimation with maximum pseudolikelihood estimation. We focus on various forms of the non-homogeneous Potts MRF, separating out its unary and pairwise terms and studying their effect on the resulting segmentation in detail. We demonstrate the use of these MRFs on a real-brain dataset. This expands on existing work with a similar per-tissue-smoothing MRF.

In chapter 5, we consider smoothing on a local neighbourhood level, by allowing local features such as the presence, strength and orientation of edges to be incorporated into the MRF. As before we show how to automatically determine the model parameters and demonstrate the

algorithm on a real-brain dataset, showing promising results against the models considered thus far. Incorporation of local features like this into the probability model is novel.

Chapter 2

Mathematical background

2.1 Introduction

Throughout this thesis, the same general probability model is used to describe the intensities of a brain MRI. This comprises of a Gaussian (normal) mixture model on the brain intensities, and a Markov random field as the prior density for the brain tissue classification. The details of the mixture model and MRF vary, but much of the underlying theory remains the same. In this chapter, we cover the common basics of the image model used throughout the thesis.

The first part of the chapter covers the normal mixture model and how it is used for image segmentation. The second part briefly covers Markov random fields, difficulties in working with them and likelihood approximations used to mitigate those difficulties. The last part shows how to incorporate a Markov random field into the normal mixture model for image segmentation.

2.2 Mixture models

Mixture models provide an important tool for statistical modelling and inference. A *mixture model* is a density that is a linear combination of other densities. One advantage of using a mixture model is that quite complex densities may be built up of simpler and well-known component densities, which need not all be the same. Mixtures are particularly useful for clustering applications. The resulting model can be used to calculate the probability that a given observation (intensity) belongs to a particular mixture component (tissue type), providing a soft classification of the data. This may be converted into a segmentation by e.g. assigning each pixel to the mixture component it has the highest posterior probability of belonging to. McLachlan and Peel (2000) provide an extensive treatment of the theory of finite mixture models.

Let $\boldsymbol{Y}_i$ $(i = 1, \ldots, n)$ be a random sample of size n, where $\boldsymbol{Y}_i$ is a p-dimensional random vector and has probability density function (pdf) $f(\boldsymbol{y}_i)$. Let $\boldsymbol{y}_i$ be a realisation of $\boldsymbol{Y}_i$, and let

$\boldsymbol{y} = (\boldsymbol{y}_1^T, \ldots, \boldsymbol{y}_n^T)^T$ denote the entire observed data.

The random vector $\boldsymbol{y}_i$ is said to have come from a g-component mixture if its pdf takes the form

$$f(\boldsymbol{y}_i; \boldsymbol{\Theta}) = \sum_{j=1}^{g} \pi_j f_j(\boldsymbol{y}_i; \boldsymbol{\theta}_j), \tag{2.1}$$

where component j has pdf f_j with parameters $\boldsymbol{\theta}_j$, and $\boldsymbol{\Theta}$ are the elements of $({\boldsymbol{\theta}_1}^T, \ldots, {\boldsymbol{\theta}_g}^T)^T$ known *a priori* to be distinct. It is not necessary that f_j be of the same form (e.g. all Gaussian). The *mixing proportions* π_j $(j = 1, \ldots, g)$ are non-negative and sum to one.

It is assumed that each observation belongs to one component of the mixture, but it is not necessarily known which. It is helpful to introduce latent (unobserved) random variables $\boldsymbol{Z} = (\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n)$ indicating the component of the mixture each observation belongs to, and correspondingly $\boldsymbol{z}_i$ being a realisation of $\boldsymbol{Z}_i$. Each $\boldsymbol{Z}_i$ is a vector of length g consisting of exactly one 1 in the jth position and all other elements 0. Let the jth element of $\boldsymbol{Z}_i$ be Z_{ij}. Then observation i is said to be in component j if and only if $Z_{ij} = 1$. Alternatively, we may write that $\boldsymbol{Z}_i = \boldsymbol{e}_j$, where $\boldsymbol{e}_j$ is the vector that is 1 in the jth position and 0 elsewhere.

In a standard mixture model, we assume $\boldsymbol{Z}_i$ are independently and identically distributed according to $p(\boldsymbol{z}_i; \boldsymbol{\Psi})$. To arrive at the pdf $f(\boldsymbol{y}_i)$ in (2.1), we suppose that $\boldsymbol{Z}_i$ are drawn from the multinomial distribution with probabilities $\boldsymbol{\Psi} = (\pi_1, \ldots, \pi_g)$. That is, $p(\boldsymbol{Z}_i = \boldsymbol{e}_j) = \pi_j$, and the parameters to be estimated are $\boldsymbol{\Psi} = (\pi_1, \ldots, \pi_{g-1})$ (as the proportions sum to 1, this determines π_g). In this formulation, we may identify $f_j(\boldsymbol{y}_i)$ with $f(\boldsymbol{y}_i | \boldsymbol{Z}_i = \boldsymbol{e}_j)$, which we will usually write as $f(\boldsymbol{y}_i | \boldsymbol{e}_j)$ for brevity.

2.2.1 Expectation-maximisation

Observing only $\boldsymbol{y}$, we wish to estimate the mixture parameters $\boldsymbol{\Theta}$ and mixing proportions $\boldsymbol{\Psi}$ For clustering and classification applications we may also wish to retrieve the latent variables $\boldsymbol{z}$. Expectation-Maximisation (EM) (Dempster et al., 1977) can be used to search for maximum likelihood estimates for the tissue means and covariance matrices and mixing proportions. Only the main equations are shown here; for their derivations, see Appendix A.

Since $Z_{ij} = 1$ for exactly one j and is 0 for all others and $\boldsymbol{Z}_i$ are assumed independent, the marginal density of the class labels is

$$p(\boldsymbol{z}; \boldsymbol{\Psi}) = \prod_{i=1}^{n} \prod_{j=1}^{g} \pi_j^{z_{ij}}. \tag{2.2}$$

We also assume that the observed values $\boldsymbol{Y}_i$ are independent given their labels $\boldsymbol{Z}_i$, so that

$$\begin{aligned} f(\boldsymbol{y}|\boldsymbol{z};\boldsymbol{\Theta}) &= \prod_{i=1}^{n} f(\boldsymbol{y}_i|\boldsymbol{z}_i;\boldsymbol{\Theta}) \\ &= \prod_{i=1}^{n}\prod_{j=1}^{g} f(\boldsymbol{y}_i|\boldsymbol{Z}_i=\boldsymbol{e}_j;\boldsymbol{\theta}_j)^{z_{ij}}. \end{aligned} \tag{2.3}$$

Combining (2.3) and (2.2) yields the joint likelihood:

$$\mathcal{L}(\boldsymbol{\Theta},\boldsymbol{\Psi};\boldsymbol{Y},\boldsymbol{Z}) = f(\boldsymbol{y},\boldsymbol{z};\boldsymbol{\Theta},\boldsymbol{\Psi}) = \prod_{i=1}^{n}\prod_{j=1}^{g}(\pi_j f(\boldsymbol{y}_i|\boldsymbol{e}_j;\boldsymbol{\theta}_j))^{z_{ij}}. \tag{2.4}$$

The Q-function, being the expectation of the log-likelihood, is then

$$Q(\boldsymbol{\Theta},\boldsymbol{\Psi}|\boldsymbol{\Theta}^{(t)},\boldsymbol{\Psi}^{(t)}) = \sum_{i=1}^{n}\sum_{j=1}^{g}\mathbb{E}\left[z_{ij}|\boldsymbol{y};\boldsymbol{\Theta}^{(t)},\boldsymbol{\Psi}^{(t)}\right](\log {\pi_j}^{(t)} + \log f(\boldsymbol{y}_i|\boldsymbol{e}_j;\boldsymbol{\Theta}^{(t)}),$$

where (t) indicates values at the t-th iteration.

On the E-step, the expected label values given the data are calculated:

$${\tau_{ij}}^{(t)} = \mathbb{E}\left[z_{ij}|\boldsymbol{y};\boldsymbol{\Theta}^{(t)},\boldsymbol{\Psi}^{(t)}\right] = \frac{{\pi_j}^{(t)} f(\boldsymbol{y}_i|\boldsymbol{e}_j;{\boldsymbol{\theta}_j}^{(t)})}{\sum_{j=1}^{g}{\pi_j}^{(t)} f(\boldsymbol{y}_i|\boldsymbol{e}_j;{\boldsymbol{\theta}_j}^{(t)})}. \tag{2.5}$$

The quantities τ_{ij} also happen to be the posterior probability that observation i belongs to component j of the mixture, $p(\boldsymbol{Z}_i = j|\boldsymbol{y};\boldsymbol{\Theta}^{(t)},\boldsymbol{\Psi}^{(t)})$.

On the M-step, the Q-function is maximised with respect to the parameters $\boldsymbol{\Theta}$ and $\boldsymbol{\Psi}$, yielding:

$$\begin{aligned} {\pi_j}^{(t+1)} &= \frac{\sum_{i=1}^{n}{\tau_{ij}}^{(t)}}{n}. \\ \boldsymbol{\Theta}^{(t+1)} &= \arg\max_{\boldsymbol{\Theta}} \frac{\partial Q}{\partial \boldsymbol{\Theta}} \end{aligned} \tag{2.6}$$

The E and M steps are repeated until a stop condition has been satisfied. Common stop conditions are the convergence of the parameter values or the convergence of the observed-data log-likelihood. The observed-data likelihood is given by

$$f(\boldsymbol{y};\boldsymbol{\Theta},\boldsymbol{\Psi}) = \sum_{\boldsymbol{z}} f(\boldsymbol{y}|\boldsymbol{z};\hat{\boldsymbol{\Theta}})p(\boldsymbol{z};\hat{\boldsymbol{\Psi}}).$$

2.2.2 Normal mixture models

In the context of brain MRI segmentation, $\boldsymbol{Y}_i$ represents the intensity of an n-pixel MRI at voxel i. This is typically scalar, though if a multichannel image is taken it will be a vector (e.g. simultaneous or co-registered PET-MR, or T1 and T2-weighted MRI). Each observation $\boldsymbol{Y}_i$

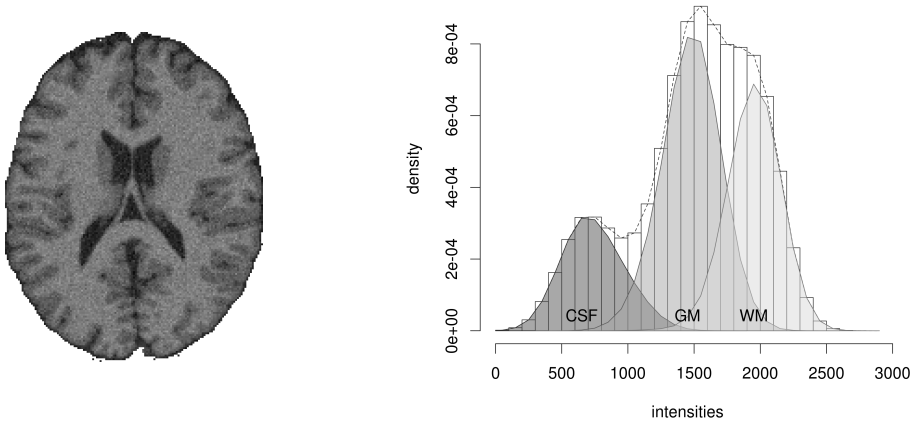


Figure 2.1: Axial MRI slice and intensity histogram, with tissue intensity distributions from a manual segmentation

represents measurements at a single pixel, and the sample $\boldsymbol{Y}$ consists of the measurements at all pixels of a single image (possibly multi-channel), as opposed to $\boldsymbol{Y}_i$ being an entire image of a subject and $\boldsymbol{Y}$ being a images of many subjects.

We are concerned with segmentation of the brain into CSF, GM and WM only (not including bone, background etc) so have $g = 3$. The unobserved variables $\boldsymbol{Z}_i$ give the tissue label at each pixel, and $\boldsymbol{z}_i$ represents a particular segmentation of the image.

Figure 2.1 shows a brain MRI and the intensity histogram of the brain voxel intensities. It also shows the intensity distribution of each tissue, where the tissues are determined by an expert manual segmentation. The distribution is trimodal, with one mode corresponding to each of the tissues. CSF has the lowest average intensity, followed by grey matter, and then white matter. Each tissue's intensity distribution appears to be normally distributed. In fact, use of a 3-component normal mixture model for brain MRI segmentation is standard. The skull may be stripped from the image beforehand using various techniques so that only brain tissue is included. Multiple Gaussians per tissue are also sometimes used (Ashburner and Friston, 1997).

We will assume each mixture component $f(\boldsymbol{y}_i|\boldsymbol{Z}_i = \boldsymbol{e}_j)$ to be Gaussian with parameters mean $\boldsymbol{\mu}_j$ and covariance matrix $\boldsymbol{\Sigma}_j$:

$$f(\boldsymbol{y}_i; \boldsymbol{\Theta}) = \sum_{j=1}^{g} \pi_j \phi(\boldsymbol{y}_i; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j),$$

where ϕ is the Gaussian pdf. For a normal mixture model, the M-step for the mean and covariance matrix is (see Appendix A for the derivation):

$$\begin{aligned} \boldsymbol{\mu}_j{}^{(t+1)} &= \frac{\sum_{i=1}^{n} \tau_{ij}{}^{(t)} \boldsymbol{y}_i}{\sum_{i=1}^{n} \tau_{ij}{}^{(t)}} \\ \boldsymbol{\Sigma}_j{}^{(t+1)} &= \frac{\sum_{i=1}^{n} \tau_{ij}{}^{(t)} (\boldsymbol{y}_i - \boldsymbol{\mu}_j{}^{(t+1)})^T (\boldsymbol{y}_i - \boldsymbol{\mu}_j{}^{(t+1)})}{\sum_{i=1}^{n} \tau_{ij}{}^{(t)}}. \end{aligned} \tag{2.7}$$

A mixture model is only identifiable up to the labels j; for example, switching the parameters and mixing proportions of components 1 and 2 will yield the same density. This is generally restored by imposing some constraint on the parameters. For example, for scalar y_i (as we will deal with in this thesis),

$$\mu_1 \leq \mu_2 \leq .. \leq \mu_3.$$

In this thesis our example datasets consist of T1 MR images; this convention is equivalent to having $j = 1$ for CSF, $j = 2$ for GM and $j = 3$ for WM.

2.2.3 Image segmentation

The aim of image segmentation is not so much to determine the tissue parameters $\boldsymbol{\Theta}$, but rather to determine the underlying segmentation $\hat{\boldsymbol{z}}$. Once estimates for the parameters have been determined, they can be used to calculate the posterior probability that each observation i belongs to a particular class j, i.e. τ_{ij}. The class memberships $\boldsymbol{z}_i$ (i.e. the hard image segmentation into tissue classes) may be estimated for each voxel by

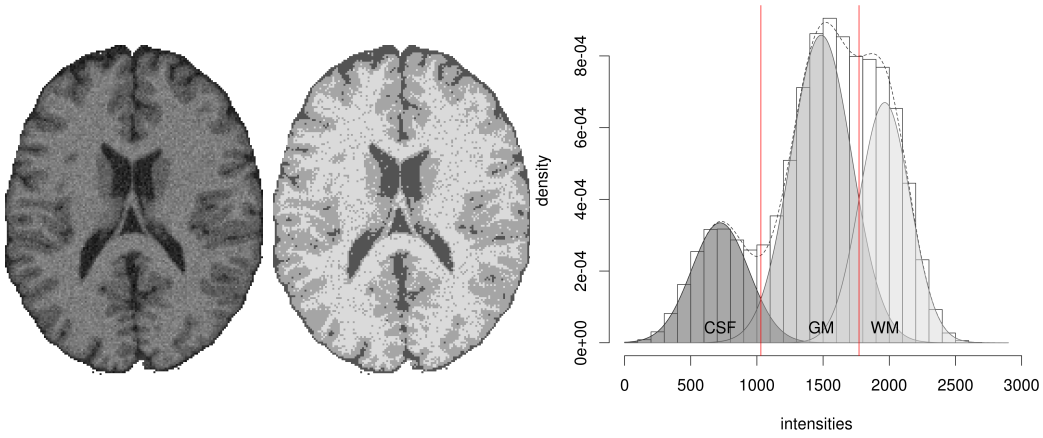
$$\hat{\boldsymbol{z}}_i = \underset{\boldsymbol{e}_j, j=1,\ldots,g}{\arg\max}\, p(\boldsymbol{Z}_i = \boldsymbol{e}_j | \boldsymbol{Y}_i; \boldsymbol{\Theta}^{(t)}, \boldsymbol{\Psi}^{(t)}) = \arg\max_j {\tau_{ij}}^{(t)}. \tag{2.8}$$

2.3 Markov Random Fields

The standard mixture model assumes that all voxel labels are independent. This leads to the classification rule (2.8), which will assign all voxels that have the same intensity, the same tissue label. This can be a problem with noisy images - isolated bright or dark pixels will be classified purely according to their intensity, even if they are located in regions of opposite brightness. This can lead to segmentations that are themselves quite noisy, as can be seen in figure 2.2.

In practice, tissue labels are not independent. Rather, a pixel's label should depend on the labels and intensities of its neighbouring pixels. It is more likely that an isolated bright pixel in a dark image region should belong to the same tissue as its dark neighbours, than a component with bright mean intensity.

There have been various attempts to incorporate spatial smoothness into intensity-based segmentation. The most basic involve convolving the pixel intensities with e.g. a Gaussian kernel *before* fitting a standard Gaussian mixture. This smooths the intensities, reducing noise. The problem with this is that noise and edges are smoothed uniformly, so that sharpness around legitimate boundaries of tissues is lost. Also, the Gaussian mixture still assumes that each voxel is independent of the others. Rather than incorporating the spatial dependence into the model itself, this modifies the observations (image intensities) prior to fitting in order to make the model more applicable.



(a) MRI (b) EM segmentation (c) intensity thresholds on the image histogram

Figure 2.2: Segmentation with a 3-component Gaussian mixture model is susceptible to image noise.

Another alternative to pre-processing the input image $\boldsymbol{y}$, is post-processing the output *segmentation* $\boldsymbol{z}$ instead. After the segmentation is obtained, morphological operations such as dilations and erosions may be used to fill in small holes in the segmentation and smooth the tissue boundaries. This suffers from similar problems to preprocessing - in particular, the sulci and gyri forming the convoluted boundary of the brain can be of sufficiently small size in the image that they are smoothed as well as the noise.

A more elegant option is to incorporate the spatial dependence directly into the probability model itself. The voxel intensity $\boldsymbol{Y}_i$ could be allowed to depend on the intensities and/or labels of its neighbours. This would be a suitable model for image blurring, where the observation at location i is corrupted by observations from nearby locations, or where observations are made on a coarser grid than the underlying location lattice. More suited to our situation is to allow each voxel's *label* (rather than intensity) to depend on the labels and/or intensities of its neighbours. This encodes the statement that locations in close proximity are more likely to be of the same tissue.

We will proceed by allowing each voxel's label $\boldsymbol{Z}_i$ to depend on the labels of its neighbours. We still assume that $\boldsymbol{Y}_i|\boldsymbol{Z}_i$, the intensities given their labels, are conditionally independent, but relax the assumption of independence between $\boldsymbol{Z}_i$. A suitable way to capture the dependence of each pixels on its neighbours is through a Markov random field.

Let us represent a set of variables by an undirected graph: each variable is a vertex, while an edge between variables indicates dependence between these variables. Vertices that are not directly connected by an edge should depend on each other only through intermediate nodes that form a path between the vertices. A Markov random field is a probability distribution that encapsulates the dependencies in the graph. For a more extensive treatment, as well as analogues for directed and hierarchical graphs, see (Koller and Friedman, 2009; Lauritzen, 1996).

For a concise review of statistical inference for MRFs, see (Stoehr, 2017).

More formally, let (V, E) be the vertices and edges of an undirected graph, and $X_i, i \in V$ be random variables, one per vertex. First, we define the notion of *conditional independence* (Dawid, 1980): we say that a variable X_i is *conditionally independent of* X_j *given* X_m, written $X_i \perp\!\!\!\perp X_j | X_m$, if the conditional probability $p(X_i|X_j, X_m)$ is a function of only X_m. For a subset of vertices A, let the notation X_A denote X_i such that $i \in A$. Let $\boldsymbol{\partial i}$ denote the set of *neighbours* of vertex i; that is, all vertices that are connected by an edge to i. Then, $X_{\boldsymbol{\partial i}}$ denotes the variables that depend on X_i. The random variables form a *Markov random field* if the following properties are satisfied:

- Pairwise Markov property: two variables corresponding to vertices that are not connected are conditionally independent given the rest of the variables. For any i and m not connected by an edge,
$$X_i \perp\!\!\!\perp X_m | X_{V \setminus \{i,m\}}.$$
- Local Markov property: a variable (vertex) is conditionally independent of all other variables (vertices) not including its neighbours, given its neighbours. For any i,
$$X_i \perp\!\!\!\perp X_{V \setminus (\{i\} \cup \boldsymbol{\partial i})} | X_{\boldsymbol{\partial i}}.$$
- Global Markov property: disjoint sets of variables are independent giving a separating subset. For any sets of vertices A, B and S such that S separates A from B,
$$X_A \perp\!\!\!\perp X_B | X_S.$$
A subset of vertices S is said to *separate* other sets A and B, if removing S from the graph disconnects A and B into separate connected components. Equivalently, every (if any) path from A to B passes through S.

It can be shown that for an undirected graph, the global property implies the local, which implies the pairwise (Lauritzen, 1996, proposition 3.4). However, they are not in general equivalent. In the context of image segmentation, $X_i = \boldsymbol{Z}_i$ and V is the set of voxels in the image.

2.3.1 Hammersley-Clifford theorem

It is usually more convenient to define dependences between variables locally, i.e. the pdf of each node given its neighbours $p(\boldsymbol{Z}_i = \boldsymbol{z}_i | \boldsymbol{Z}_{\boldsymbol{\partial i}} = \boldsymbol{z}_{\boldsymbol{\partial i}})$ is given. We will shorten this to $p(\boldsymbol{z}_i | \boldsymbol{z}_{\boldsymbol{\partial i}})$ for convenience of notation in the remainder of the thesis. We assume $p(\boldsymbol{z}_i | \boldsymbol{z}_{\boldsymbol{\partial i}})$ belongs to the

exponential family. It is common to write it as

$$\begin{aligned} p(\boldsymbol{z}_i|\boldsymbol{z}_{\partial \boldsymbol{i}};\boldsymbol{\Psi}) &= \frac{\exp(-U_i(\boldsymbol{z}_i|\boldsymbol{z}_{\partial \boldsymbol{i}};\boldsymbol{\Psi}))}{C_i} \\ C_i &= \sum_{k=1}^{g} \exp(-U_i(\boldsymbol{e}_k|\boldsymbol{z}_{\partial \boldsymbol{i}};\boldsymbol{\Psi})), \end{aligned} \tag{2.9}$$

where the negative sign is by convention. The function U_i is often called a *potential*.

Given a set of local conditional pdfs, two questions occur:

- What is the corresponding joint density $p(\boldsymbol{z})$?
- Under what conditions are $p(\boldsymbol{z}_i|\boldsymbol{z}_{\partial \boldsymbol{i}})$ even compatible with each other?
- If they are compatible, does $p(\boldsymbol{z})$ satisfy the Markov properties?

The answers to these questions are addressed in Besag (1974). First, it is assumed that p is *positive*, i.e. all realisations $\boldsymbol{z}$ have positive probability. Then, the joint density may be found by taking the product of conditionals, normalised to sum to one (see (2.2) of Besag (1974)):

$$\begin{aligned} p(\boldsymbol{z};\boldsymbol{\Psi}) &= \frac{1}{C}\prod_{i=1}^{n} p(\boldsymbol{z}_i|\boldsymbol{z}_{\partial \boldsymbol{i}};\boldsymbol{\Psi}), \\ C &= \sum_{\text{all possible } \boldsymbol{z}'} \prod_{i=1}^{n} p(\boldsymbol{z}'_i|\boldsymbol{z}'_{\partial \boldsymbol{i}};\boldsymbol{\Psi}) \end{aligned} \tag{2.10}$$

where $\boldsymbol{z}_{\partial \boldsymbol{i}}$ denotes all $\boldsymbol{z}_m$ that $\boldsymbol{z}_i$ depends on i.e. all the neighbours of i, and $\boldsymbol{\Psi}$ are the parameters of the MRF. We will omit the dependence on $\boldsymbol{\Psi}$ unless relevant for ease of notation.

As to whether a given joint pdf forms a valid Markov random field, the Hammserley-Clifford gives the sufficient and necessary conditions. It was first proven by Hammersley and Clifford in an unpublished manuscript (Hammersley and Clifford, 1971), and later proved more generally and concisely by Besag (Besag, 1974). A positive $p(\boldsymbol{z})$ forms a valid Markov random field (satisfies the Markov properties) if and only if it factorises over the cliques of its underlying graph.

A *clique* of a graph is a fully-connected subset of vertices. All the cliques of a 2D grid/lattice where each node has 4 neighbours, or where each node has 8 neighbours, are shown in Figure 2.3. For a pdf to factorise over cliques of a graph means that it can be written

$$p(\boldsymbol{z}) = \frac{1}{C}\prod_{\text{cliques } c} \psi_c(\boldsymbol{z}_c),$$

where each c is a clique, $\boldsymbol{z}_c$ are all $\boldsymbol{z}_i$ such that i is in c, and C is a normalising constant. The clique potentials ψ_c may depend on $\boldsymbol{z}$ only through those $\boldsymbol{z}_i$ where i is in the clique c. In general, this factorisation is not unique. For example, suppose ψ_c was defined over pairwise neighbours only in the 8-neighbour lattice of figure 2.3. Then $p(\boldsymbol{z})$ could be factorised such that each clique c consisted of two nodes only (pairwise neighbours). Alternatively, the graph could be

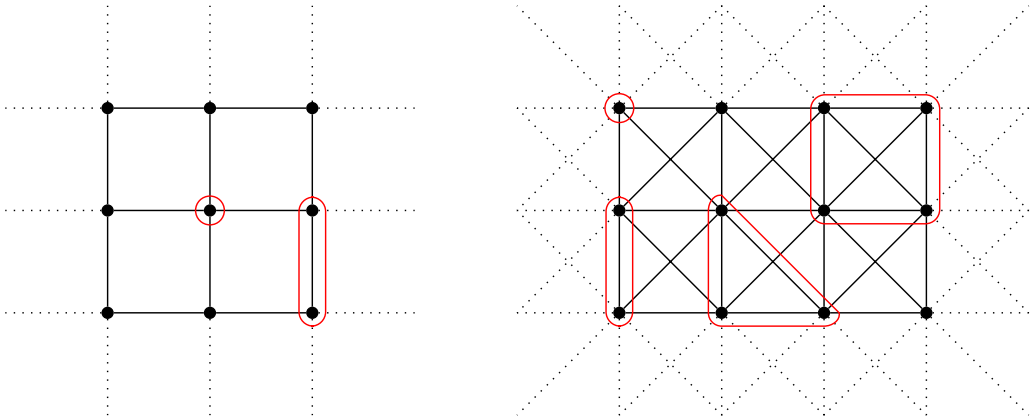


Figure 2.3: All clique shapes for a regular 2D lattice with 4 and 8 neighbours (left and right respectively). The maximal cliques are of size 2 (4 neighbours) or 4 (8 neighbours).

decomposed into cliques of size 4, and then ψ_c would be obtained by multiplying all the pairwise potentials together. What is important is that p can be decomposed into a product of functions over cliques.

In summary, the Hammersley-Clifford theorem allows us to define a potential over the local cliques of a graph which is often more convenient, while guaranteeing that the resulting (normalised) product is a valid joint pdf for an MRF.

As an example, in image segmentation, one can imagine each voxel to be a node of a regular lattice as demonstrated in figure 2.3 for a 2D image. The property that each voxel's label should depend on its neighbours' labels can be encoded by defining a conditional pdf

$$p(\boldsymbol{Z}_i = \boldsymbol{e}_j | \boldsymbol{z}_{\partial \boldsymbol{i}}) = \frac{\exp(\beta u_{ij})}{\sum_{k=1}^{g} \exp(\beta u_{ik})}, \tag{2.11}$$

where $u_{ij} = \sum_{m \in \partial \boldsymbol{i}} z_{mj}$ is the number of neighbours of voxel i that share the same label j. When the parameter β is positive, this pdf encodes that it is more likely for a voxel to take the same label as that of the majority of its neighbours. This is the Potts model of statistical mechanics (or the Ising model, when there are only 2 labels) and is the most common form of MRF used in image segmentation (Besag, 1986); we will explore it in detail in Chapter 3.

By exploiting the binary nature of z_{ij} we may rewrite (2.11) as

$$p(\boldsymbol{z}_i | \boldsymbol{z}_{\partial \boldsymbol{i}}) = \frac{\exp(\beta \sum_{j=1}^{g} z_{ij} u_{ij})}{\sum_{k=1}^{g} \exp(\beta u_{ik})},$$

so that by (2.10) the joint pdf is given by

$$\begin{aligned} p(\boldsymbol{z}) &= \frac{1}{C} \exp(\beta \sum_{i=1}^{n} \sum_{j=1}^{g} z_{ij} u_{ij}) \\ &= \frac{1}{C} \exp(2\beta \sum_{i,m \text{ neighbours}} \sum_{j=1}^{g} z_{ij} z_{mj}), \end{aligned}$$

where the factor of 2 is because the neighbours are double-counted. It can be seen that this decomposes into a product over pairwise cliques $\{i, m\}$:

$$\phi_{\{i,m\}}(\boldsymbol{z}_i, \boldsymbol{z}_m) = \exp(2\beta \boldsymbol{z}_i^T \boldsymbol{z}_m).$$

Hence, $p(\boldsymbol{z})$ forms a valid MRF.

In this thesis we will consider MRFs that have non-zero clique potentials for singletons and adjacent pairs of voxels only as opposed to tuples or higher, as these quickly become intractable. For a thorough review on image applications of MRFs, see (Li, 2009), or (Winkler, 2012) for a more probabilistic approach.

2.3.2 Likelihood approximations

There are generally four different tasks one wishes to achieve with an MRF:

- sampling from it,
- estimating its parameters,
- finding a maximum-likelihood realisation $\boldsymbol{z}$ (possibly given observations $\boldsymbol{y}$).

However, all of these are hampered by calculation of the MRF's normalising constant (also called the partition function) C in (2.10).

$$C = \sum_{\text{all possible } \boldsymbol{z}'} \prod_{i=1}^{n} p(\boldsymbol{z}'_i | \boldsymbol{z}'_{\partial \boldsymbol{i}})$$

The normalising constant involves a sum over all possible states of $\boldsymbol{z}$. For an n-voxel image with g colours, there are g^n possible images. For a typical brain MRI, $g = 3$ and n is in the millions. This makes calculation of C intractable.

When it comes to sampling, the primary methods are based on Gibbs sampling (Geman and Geman, 1984), including the Swendsen-Wang method (Swendsen and Wang, 1987), and Wolff's algorithm (Wolff, 1989). As these are based on Gibbs sampling, they operate on the local probabilities $p(\boldsymbol{z}_i | \boldsymbol{z}_{\partial \boldsymbol{i}})$ rather than the intractable joint density.

We are mainly focused on the problems of parameter estimation and inferring $\boldsymbol{z}$ from observations $\boldsymbol{y}$. Most approaches focus on replacing $p(\boldsymbol{z})$ with an approximation $\tilde{p}(\boldsymbol{z})$ that is more tractable to compute, before proceeding These include the *pseudolikelihood* and *mean-field* approximations, and are a major focus of this thesis.

2.3.2.1 Pseudolikelihood approximation

The *pseudolikelihood* (PL) approximation was introduced in Besag (1975) and applied in the context of imaging in Besag (1986). It is a type of composite likelihood as studied by (Lindsay, 1988; Varin, 2008), whereby a joint likelihood is approximated by a product of more tractable marginal or conditional likelihoods. For a review of the different forms of composite likelihood, see Varin et al. (2011); we focus on Besag's pseudolikelihood. Besag's pseudolikelihood approximates $p(\boldsymbol{z})$ as the product of the conditional probabilities $p(\boldsymbol{z}_i|\boldsymbol{z}_{\partial i})$:

$$\begin{aligned} p(\boldsymbol{z};\boldsymbol{\Psi}) \approx \tilde{p}(\boldsymbol{z};\boldsymbol{\Psi}) &= \prod_{i=1}^{n} p(\boldsymbol{z}_i|\boldsymbol{z}_{\partial i};\boldsymbol{\Psi}) \\ &= \frac{1}{\tilde{C}(\boldsymbol{z};\boldsymbol{\Psi})} \prod_{i=1}^{n} \exp(-U_i(\boldsymbol{z}_i|\boldsymbol{z}_{\partial i};\boldsymbol{\Psi})) \\ \tilde{C}(\boldsymbol{z};\boldsymbol{\Psi}) &= \prod_{i=1}^{n} \sum_{k=1}^{g} \exp(-U_i(\boldsymbol{e}_k|\boldsymbol{z}_{\partial i};\boldsymbol{\Psi})) \end{aligned} \tag{2.12}$$

Comparing the pseudolikelihood to the full likelihood (2.10),

$$\begin{aligned} p(\boldsymbol{z};\boldsymbol{\Psi}) &= \frac{1}{C(\boldsymbol{\Psi})} \prod_{i=1}^{n} \exp(-U_i(\boldsymbol{z}_i|\boldsymbol{z}_{\partial i};\boldsymbol{\Psi})), \\ C(\boldsymbol{\Psi}) &= \sum_{\boldsymbol{z}'} \prod_{i=1}^{n} \exp(-U_i(\boldsymbol{z}'_i|\boldsymbol{z}'_{\partial i};\boldsymbol{\Psi})), \end{aligned}$$

it can be seen that the pseudolikelihood replaces the global normalising constant C (a sum over g^n terms) with a product of local normalising constants C_i. Since each C_i is a sum over only g terms, the pseudolikelihood is computationally tractable. However, while the global constant C depends only on the parameters $\boldsymbol{\Psi}$, the pseudolikelihood constant $\tilde{C}$ also depends on the realisation $\boldsymbol{z}$.

Besag's pseudolikelihood approximates the likelihood as a product of conditionals over the individual voxels. For this reason it was termed *point-pseudolikelihood* by Qian and Titterington (1992), who also studied line- and block-versions. In this thesis, 'pseudolikelihood' refers to the point-pseudolikelihood unless otherwise specified. Rydén and Titterington (1998) showed how the pseudolikelihood could be used in conjunction with Gibbs sampling to sample from an MRF.

2.3.2.2 Mean-field approximation

An alternative though similar approximation to the pseudolikelihood approximation is the mean-field approximation of the likelihood. Mean field theory originated in statistical mechanics as a tool to study phase transitions in interacting systems. For an extensive treatment, see Chandler (1987). The core idea is to replace interaction terms with their expected or mean values in order to decouple them. There are multiple mean-field approximations for a given

MRF, e.g. each neighbour $\boldsymbol{z}_m$ is replaced by its mean value, or where pairwise interactions $\boldsymbol{z}_i^T \boldsymbol{z}_m$ are replaced by their mean values, but we focus on the former only:

$$\begin{aligned} p(\boldsymbol{z}) \approx \tilde{p}(\boldsymbol{z}) &= \prod_{i=1}^{n} p(\boldsymbol{z}_i | \langle \boldsymbol{z}_{\partial \boldsymbol{i}} \rangle) \\ &= \prod_{i=1}^{n} \frac{\exp(-U_i(\boldsymbol{z}_i | \langle \boldsymbol{z}_{\partial \boldsymbol{i}} \rangle))}{\sum_{k=1}^{g} \exp(-U_i(\boldsymbol{e}_k | \langle \boldsymbol{z}_{\partial \boldsymbol{i}} \rangle))}. \end{aligned} \tag{2.13}$$

Here, $\langle \cdot \rangle$ is the expected value with respect to $\tilde{p}_{MF}$, and $\langle \boldsymbol{z}_{\partial \boldsymbol{i}} \rangle$ is shorthand for $\langle \boldsymbol{z}_m \rangle$ such that $m \in \boldsymbol{\partial i}$. Note that while $\boldsymbol{z}_i$ is binary in nature $\in \{0,1\}^g$, $\langle \boldsymbol{z}_{\partial \boldsymbol{i}} \rangle$ is continuous $\in [0,1]^g$ with $\sum_{j=1}^{g} \langle \boldsymbol{z}_i \rangle_j = 1$.

The mean values $\langle \boldsymbol{z}_i \rangle$ may be found by solving the fixed-point or *self-consistency* equations (Zhang, 1992)

$$\langle \boldsymbol{z}_i \rangle = \sum_{j=1}^{g} \boldsymbol{e}_j p(\boldsymbol{Z}_i = \boldsymbol{e}_j | \langle \boldsymbol{z}_{\partial \boldsymbol{i}} \rangle) \tag{2.14}$$

where the right-hand side is exactly the expectation of $\boldsymbol{z}_i$ under $p(\cdot | \langle \boldsymbol{z}_{\partial \boldsymbol{i}} \rangle)$. Brouwer's fixed-point theorem guarantees existence of at least one solution to this equation, and iteration of the above equation typically leads to one of the solutions (Wu and Doerschuk, 1995).

Alternatively, the mean-field approximation may be justified by searching for a factorisable $\tilde{p}(\boldsymbol{z})$ that minimises its Kullback-Leibler divergence to the true pdf $p(\boldsymbol{z})$. Let $\mathcal{F}$ be the class of factorisable densities of the form $\tilde{p}(\boldsymbol{z}) = \prod_{i=1}^{n} p_i(\boldsymbol{z}_i)$. The desired approximation is

$$\arg\min_{\tilde{p} \in \mathcal{F}} \mathbb{E}_{\tilde{p}} \left[\log \left(\frac{\tilde{p}(\boldsymbol{z})}{p(\boldsymbol{z})} \right) \right], \tag{2.15}$$

where the expectation is taken with respect to the candidate pdf $\tilde{p}$. It can be shown that solving this equation yields exactly the previous self-consistency equations (Hofmann and Buhmann, 1997, section 4.1).

When the MRF is *hidden*, there are two choices when trying to approximate the joint likelihood $f(\boldsymbol{y}, \boldsymbol{z})$. Either $p(\boldsymbol{z})$ or $p(\boldsymbol{z}|\boldsymbol{y})$ may be approximated, from which the other can be derived. Approximating the marginal density $p(\boldsymbol{z})$ yields the equations already shown (2.14). However, note that one solution to these equations is the uniform solution $\langle \boldsymbol{z}_i \rangle = \frac{1}{g}$ for all i. Celeux et al. (2003), Example 1 claims this solution is unique for $\beta < \frac{g}{|\boldsymbol{\partial i}|}$ and $g \leq 4$ where $|\boldsymbol{\partial i}|$ is the neighbourhood size (the proof's existence is mentioned but it is not given). This degenerate solution is not of interest as it prohibits estimation of the MRF parameters, being independent of them. Thus, it is preferable to use the mean-field approximation for $p(\boldsymbol{z}|\boldsymbol{y})$, as noted by

(Celeux et al., 2003; Archer and Titterington, 2002). This yields the equations

$$\begin{aligned}\langle \boldsymbol{z}_i \rangle &= \sum_{j=1}^{g} \boldsymbol{e}_j p(\boldsymbol{e}_j | \boldsymbol{y}_i, \langle \boldsymbol{z}_{\partial \boldsymbol{i}} \rangle) \\ &= \sum_{j=1}^{g} \boldsymbol{e}_j \frac{f(y_i|\boldsymbol{e}_j) p(\boldsymbol{e}_j | \langle \boldsymbol{z}_{\partial \boldsymbol{i}} \rangle)}{\sum_{k=1}^{g} f(y_i|\boldsymbol{e}_k) p(\boldsymbol{e}_k | \langle \boldsymbol{z}_{\partial \boldsymbol{i}} \rangle)}\end{aligned} . \tag{2.16}$$

It can be seen that the mean-field (MF) approximation (2.13) is very similar the the pseudolikelihood approximation (2.12), except that the mean-field approximation uses the expected values for the neighbours $\langle \boldsymbol{z}_{\partial \boldsymbol{i}} \rangle$, while the pseudolikelihood approximation uses the discrete values $\boldsymbol{z}_{\partial \boldsymbol{i}}$. Since both approximations are tractable, they simplify the problems of sampling, parameter estimation, and estimation of $\boldsymbol{z}$.

2.4 Expectation-Maximisation for a mixture-MRF model

Our image model now consists of voxel intensities that are normally distributed given the voxel label, and labels that are distributed according to a Markov random field. We wish to retrieve the maximum-likelihood segmentation $\boldsymbol{z}$ while also fitting the unknown parameters, when only the MRI $\boldsymbol{y}$ is observed. As for a standard mixture model, the hidden nature of $\boldsymbol{z}$ makes the problem suitable for Expectation Maximisation. The image model is (exploiting the binary nature of z_{ij} to write $f(\boldsymbol{y}|\boldsymbol{z})$ and $p(\boldsymbol{z})$):

$$\begin{aligned} f(\boldsymbol{y}|\boldsymbol{z}; \boldsymbol{\Theta}) &= \prod_{i=1}^{n} \prod_{j=1}^{g} f(\boldsymbol{y}_i | \boldsymbol{Z}_i = \boldsymbol{e}_j; \boldsymbol{\Theta})^{z_{ij}} \\ f(\boldsymbol{y}_i|\boldsymbol{e}_j; \boldsymbol{\Theta}) &= \phi(\boldsymbol{y}_i; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) \sim \mathcal{N}(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) \\ p(\boldsymbol{z}; \boldsymbol{\Psi}) &= \frac{1}{C} \prod_{i=1}^{n} \prod_{j=1}^{g} \exp(-U_i(\boldsymbol{e}_j | \boldsymbol{z}_{\partial \boldsymbol{i}}; \boldsymbol{\Psi}))^{z_{ij}}. \end{aligned}$$

This model is often termed 'GMM-MRF' or 'GMM-HMRF' (Gaussian mixture model (hidden) Markov random field).

The joint likelihood is

$$\begin{aligned} f(\boldsymbol{y}, \boldsymbol{z}; \boldsymbol{\Theta}, \boldsymbol{\Psi}) &= f(\boldsymbol{y}|\boldsymbol{z}; \boldsymbol{\Theta}) p(\boldsymbol{z}; \boldsymbol{\Psi}) \\ &= \frac{1}{C} \prod_{i=1}^{n} \prod_{j=1}^{g} \left(\phi(\boldsymbol{y}_i; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) \exp(-U_i(\boldsymbol{z}_i | \boldsymbol{z}_{\partial \boldsymbol{i}}; \boldsymbol{\Psi})) \right)^{z_{ij}} . \end{aligned}$$

The Q-function is then

$$Q(\boldsymbol{\Theta}, \boldsymbol{\Psi}|\boldsymbol{\Theta}^{(t)}, \boldsymbol{\Psi}^{(t)}) = \left[\sum_{i=1}^{n}\sum_{j=1}^{g} \mathbb{E}\left[z_{ij}|\boldsymbol{y}; \boldsymbol{\Theta}^{(t)}, \boldsymbol{\Psi}^{(t)}\right] \log f_j(\boldsymbol{y}_i; \mu_j, \boldsymbol{\Sigma}_j) - \mathbb{E}\left[z_{ij} U_i(\boldsymbol{e}_j|\boldsymbol{z}_{\partial \boldsymbol{i}}; \boldsymbol{\Psi})|\boldsymbol{y}; \boldsymbol{\Theta}^{(t)}, \boldsymbol{\Psi}^{(t)}\right]\right] - \log C(\boldsymbol{\Psi}). \tag{2.17}$$

There are two problems to be dealt with here. First, the normalising constant C is not tractable to calculate. The problem can be avoided if the MRF parameters $\boldsymbol{\Psi}$ are held fixed rather than estimated, as then C need not be computed. In fact, it is very common to do this in image segmentation for this reason; we will return to this in Chapter 3. The second problem is that the expectations cannot be computed due to the dependence of each voxel on its neighbours.

The natural solution based on the theory presented thus far is to replace $p(\boldsymbol{z})$ with the pseudolikelihood or mean-field approximations $\tilde{p}(\boldsymbol{z})$:

$$\tilde{p}_{PL}(\boldsymbol{z}) = \prod_{i=1}^{n} \frac{\exp(-U_i(\boldsymbol{z}_i|\boldsymbol{z}_{\partial \boldsymbol{i}}))}{C_i(\boldsymbol{z}_{\partial \boldsymbol{i}})}$$
$$\tilde{p}_{MF}(\boldsymbol{z}) = \prod_{i=1}^{n} \frac{\exp(-U_i(\boldsymbol{z}_i|\langle \boldsymbol{z}_{\partial \boldsymbol{i}} \rangle))}{C_i(\langle \boldsymbol{z}_{\partial \boldsymbol{i}} \rangle)},$$

where the dependence of the pixel-wise normalising constant C_i on its neighbours is written for emphasis. In what follows, we will use the pseudolikelihood approximation. The mean-field approximation is identical except that the neighbours $\boldsymbol{z}_{\partial \boldsymbol{i}}$ are replaced with the mean values $\langle \boldsymbol{z}_{\partial \boldsymbol{i}} \rangle$. For the most part there is no difference in theory; we will point out when there is. Since the approximations $\tilde{p}(\boldsymbol{z})$ replace C by a product of C_i, this resolves the difficulty in computing C in the Q-function.

The second difficulty was computation of the expectations in (2.17). If the mean-field approximation is used, the replacement of $\boldsymbol{Z}_{\partial \boldsymbol{i}}$ with $\langle \boldsymbol{z}_{\partial \boldsymbol{i}} \rangle$, which are constants, uncouples each term in $\tilde{p}(\boldsymbol{z})$ and hence the computation is straightforward:

$$\begin{aligned} \mathbb{E}\left[z_{ij}|\boldsymbol{y}; \boldsymbol{\Theta}^{(t)}, \boldsymbol{\Psi}^{(t)}\right] &= \Pr(\boldsymbol{Z}_i = \boldsymbol{e}_j|\boldsymbol{y}; \boldsymbol{\Theta}^{(t)}, \boldsymbol{\Psi}^{(t)}) \\ &= \frac{\phi(\boldsymbol{y}_i; \boldsymbol{\mu}_j{}^{(t)}, \boldsymbol{\Sigma}_j{}^{(t)}) \tilde{p}(\boldsymbol{e}_j; \boldsymbol{\Psi}^{(t)})}{f(\boldsymbol{y}_i; \boldsymbol{\Theta}^{(t)})} \\ &= \frac{\phi(\boldsymbol{y}_i; \boldsymbol{\mu}_j{}^{(t)}, \boldsymbol{\Sigma}_j{}^{(t)}) p(\boldsymbol{e}_j|\langle \boldsymbol{z}_{\partial \boldsymbol{i}} \rangle; \boldsymbol{\Psi}^{(t)})}{\sum_{k=1}^{g} \phi(\boldsymbol{y}_i; \boldsymbol{\mu}_k{}^{(t)}, \boldsymbol{\Sigma}_k{}^{(t)}) p(\boldsymbol{e}_j|\langle \boldsymbol{z}_{\partial \boldsymbol{i}} \rangle; \boldsymbol{\Psi}^{(t)})} \end{aligned} \tag{2.18}$$

The second expectation in (2.17) simplifies to

$$\mathbb{E}\left[z_{ij} U_i(\boldsymbol{e}_j|\langle \boldsymbol{z}_{\partial \boldsymbol{i}} \rangle; \boldsymbol{\Psi}|\boldsymbol{y}; \boldsymbol{\Theta}^{(t)}, \boldsymbol{\Psi}^{(t)})\right] = \mathbb{E}\left[z_{ij}|\cdot\right] U_i(\boldsymbol{e}_j|\langle \boldsymbol{z}_{\partial \boldsymbol{i}} \rangle; \boldsymbol{\Psi}|\boldsymbol{y}; \boldsymbol{\Theta}^{(t)}, \boldsymbol{\Psi}^{(t)}),$$

and the quantity $\mathbb{E}\left[z_{ij}|\cdot\right]$ is calculated as previously. It is interesting to note that the expectation (2.18) is identical to calculation of the mean-field approximation $\langle \boldsymbol{z}_i \rangle$ (2.16), as both compute the same expectation. The only difference is that the mean-field approximation should be computed

iteratively with the results used to updated the mean-field values. The E-step is computed simultaneously, treating $\langle \boldsymbol{z}_i \rangle$ as the fixed neighbour values.

If the pseudolikelihood approximation is used, then calculation of the marginal probability $\Pr(\boldsymbol{Z}_i = \boldsymbol{e}_j)$ is intractable as each voxel still depends on its neighbours. Following (Kay and Titterington, 1986; Kay, 1986), this can be resolved by replacing the expectation under $p(\boldsymbol{z}_i)$ with $p(\boldsymbol{z}_i|\boldsymbol{z}_{\partial \boldsymbol{i}}{}^{(t)})$ where $\boldsymbol{z}_{\partial \boldsymbol{i}}{}^{(t)}$ is the current best estimate for the neighbours. The expectations then proceed identically to when the mean-field approximation is used, except that the mean-field neighbours $\langle \boldsymbol{z}_{\partial \boldsymbol{i}} \rangle$ are replaced with the current best (discrete) approximation $\boldsymbol{z}_{\partial \boldsymbol{i}}{}^{(t)}$.

With these modifications, the Q-function becomes

$$\begin{aligned} Q(\boldsymbol{\Theta}, \boldsymbol{\Psi}|\boldsymbol{\Theta}^{(t)}, \boldsymbol{\Psi}^{(t)}) = \sum_{i=1}^{n}\sum_{j=1}^{g} \mathbb{E}\left[z_{ij}|\boldsymbol{y}, \boldsymbol{z}_{\partial \boldsymbol{i}}{}^{(t)}; \boldsymbol{\Theta}^{(t)}, \boldsymbol{\Psi}^{(t)}\right]\Big(\log f_j(\boldsymbol{y}_i; \mu_j, \boldsymbol{\Sigma}_j) - U_i(\boldsymbol{e}_j|\boldsymbol{z}_{\partial \boldsymbol{i}}{}^{(t)}; \boldsymbol{\Theta}) \\ - \log C_i(\boldsymbol{\Psi})\Big) \\ C_i(\boldsymbol{\Psi}) = \sum_{k=1}^{g} \exp(-U_i(\boldsymbol{e}_k|\boldsymbol{z}_{\partial \boldsymbol{i}}{}^{(t)}; \boldsymbol{\Psi})). \end{aligned}$$

By replacing $\boldsymbol{z}_{\partial \boldsymbol{i}}$ with $\boldsymbol{z}_{\partial \boldsymbol{i}}{}^{(t)}$ the dependence between voxels is decoupled, so the expectations may be computed:

$$\mathbb{E}\left[z_{ij}|\boldsymbol{y}, \boldsymbol{z}_{\partial \boldsymbol{i}}{}^{(t)}\boldsymbol{\Theta}^{(t)}, \boldsymbol{\Psi}^{(t)}\right] = \tau_{ij}{}^{(t)} = \frac{p(\boldsymbol{e}_j|\boldsymbol{z}_{\partial \boldsymbol{i}}{}^{(t)}, \boldsymbol{\Psi}^{(t)})\phi(\boldsymbol{y}_i; \boldsymbol{\mu}_j{}^{(t)}, \boldsymbol{\Sigma}_j{}^{(t)})}{\sum_{k=1}^{g} p(\boldsymbol{e}_k|\boldsymbol{z}_{\partial \boldsymbol{i}}{}^{(t)}, \boldsymbol{\Psi}^{(t)})\phi(\boldsymbol{y}_i; \boldsymbol{\mu}_k{}^{(t)}, \boldsymbol{\Sigma}_k{}^{(t)})}. \tag{2.19}$$

Comparing this to the E-step of the standard mixture model, the equations are the same, except that $\pi_j = \Pr(\boldsymbol{Z}_i = \boldsymbol{e}_j)$ of the standard mixture model has been replaced with $p(\boldsymbol{e}_j|\boldsymbol{z}_{\partial \boldsymbol{i}}{}^{(t)})$. The M-step then proceeds identically to the standard mixture model, with the new τ_{ij} as defined above.

2.4.1 Approximating $\boldsymbol{z}$

The final piece to the algorithm is how to determine values $\boldsymbol{z}^{(t)}$ to be used for $\boldsymbol{z}_{\partial \boldsymbol{i}}{}^{(t)}$ in computing the expectations. The MAP estimate

$$\hat{\boldsymbol{z}}_{MAP} = \arg\max_{\boldsymbol{z}} p(\boldsymbol{z}|\boldsymbol{y}) = \arg\max_{\boldsymbol{z}} f(\boldsymbol{y}|\boldsymbol{z})p(\boldsymbol{z}).$$

is intractable given the dependence between neighbouring $\boldsymbol{z}_i$ and that the maximisation must be over all $\boldsymbol{z}$ simultaneously. Thus we return to using approximations of $p(\boldsymbol{z})$. These reduce the simultaneous estimate of $\boldsymbol{z}$ to pointwise estimates.

If the mean-field approximation is being used, then the fixed-point equation (2.16) may be iterated to find $\langle \boldsymbol{z}_i \rangle$ from the mean-field approximation of $p(\boldsymbol{z})$. However, as suggested by Celeux et al. (2003), it is more advantageous to use the mean-field approximation of the posterior

$p(\boldsymbol{z}|\boldsymbol{y})$. The self-consistency equations (2.16) become

$$\begin{aligned}\langle \boldsymbol{z}_i \rangle^{(t+1)} &= \sum_{j=1}^{g} \boldsymbol{e}_j \Pr(\boldsymbol{Z}_i = \boldsymbol{e}_j|\boldsymbol{y}) \\ &= \sum_{j=1}^{g} \boldsymbol{e}_j \frac{f(\boldsymbol{y}_i|\boldsymbol{Z}_i = \boldsymbol{e}_j) p(\boldsymbol{Z}_i = \boldsymbol{e}_j|\langle \boldsymbol{z}_{\partial \boldsymbol{i}} \rangle^{(t,t+1)})}{\sum_{k=1}^{g} f(\boldsymbol{y}_i|\boldsymbol{Z}_i = \boldsymbol{e}_k) p(\boldsymbol{Z}_i = \boldsymbol{e}_k|\langle \boldsymbol{z}_{\partial \boldsymbol{i}} \rangle^{(t,t+1)})}.\end{aligned} \tag{2.20}$$

The superscript $(t, t+1)$ means to use the most recent value of $\langle \boldsymbol{z}_m \rangle$, be it from the previous or current iteration (depending on whether that voxel has been updated yet). This is identical to the τ_{ij} calculation (2.19), except that it is performed sequentially with $\langle \boldsymbol{z}_{\partial \boldsymbol{i}} \rangle$ always consisting of the most-recently computed values. The scheme is guaranteed to converge to a fixed point (Wu and Doerschuk, 1995).

Another popular alternative is to use Iterated Conditional Modes (Besag, 1986). Rather than performing the maximisation over all voxels simultaneously, Besag proposed to update each voxel sequentially according to the mode of its conditional likelihood:

$$\begin{aligned}\boldsymbol{z}_i{}^{(t+1)} &= \arg\max_{\boldsymbol{e}_j} \Pr(\boldsymbol{Z}_i = \boldsymbol{e}_j|\boldsymbol{y}, \boldsymbol{z}_{\partial \boldsymbol{i}}{}^{(t,t+1)}) \\ &= \arg\max_{\boldsymbol{e}_j} f(\boldsymbol{y}_i|\boldsymbol{Z}_i = \boldsymbol{e}_j) p(\boldsymbol{Z}_i = \boldsymbol{e}_j|\boldsymbol{z}_{\partial \boldsymbol{i}}{}^{(t,t+1)})\end{aligned} \tag{2.21}$$

ICM will converge to a (possibly local) maximum of the posterior probability, as

$$p(\boldsymbol{z}|\boldsymbol{y}) = p(\boldsymbol{z}_i, \boldsymbol{z}_{-i}|\boldsymbol{y}) = p(\boldsymbol{z}_i|\boldsymbol{y}, \boldsymbol{z}_{-i}) p(\boldsymbol{z}_{-i}|\boldsymbol{y}) = p(\boldsymbol{z}_i|\boldsymbol{y}, \boldsymbol{z}_{\partial \boldsymbol{i}}) p(\boldsymbol{z}_{-i}|\boldsymbol{y}), \tag{2.22}$$

and $p(\boldsymbol{z}_i|\boldsymbol{y}, \boldsymbol{z}_{\partial \boldsymbol{i}})$ is the quantity ICM maximises for each i. Here $\boldsymbol{z}_{-i}$ means all $\boldsymbol{z}_m$ such that $m \neq i$. In practice Besag found the algorithm to converge very quickly (in the sense that the number of voxels changing per iteration decreased rapidly), often in less than 10 cycles. It is of interest to note that ICM is very similar to the mean-field procedure, except that ICM uses the mode as the estimate, while the mean-field procedure uses the mean. In addition, ICM is equivalent to the standard classification rule used in EM (2.8), except that the in (2.8) the τ_{ij} quantities are calculated simultaneously rather than sequentially.

A variation of ICM, termed "Iterated Conditional Expectations" (ICE), was proposed by Owen (Owen, 1986, Owen (1989)). The aim of ICE was to avoid the hard assignation of labels (2.21), since this will treat (for example) a voxel with probability 51% to be a particular label the same as a voxel with probability 99% to be a particular label. ICE uses soft neighbours (expected values) rather than hard thresholding to update the posterior probabilities. This turns out to be exactly the mean-field update.

The ICM and MF updates should be performed sequentially over the voxels. However, a slight computational saving can be made by employing Besag's "coding sets" (Besag, 1974). This method was developed for parameter estimation of MRFs; we will cover that aspect in the next chapter. It can also be used in sequential voxel-update operations. It is applicable to any MRF

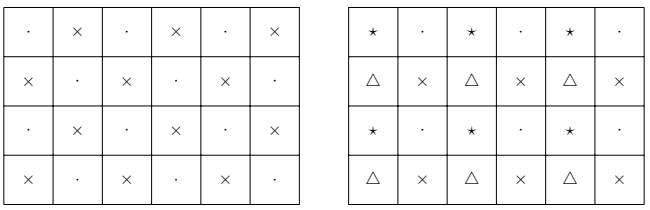


Figure 2.4: Coding sets for a two-dimensional image grid with 4 neighbours (orthogonal only) and 8 (orthogonal and diagonal). No pixels in a given set are neighbours.

whose underlying graph can be partitioned into subsets such that no two elements of the same subset are neighbours. That is, all nodes in a given subset are mutually independent, given their neighbours. Examples are shown in figure 2.4 for the two-dimensional 4- and 8-neighbour cases, where each subset is represented by a different symbol. No two pixels in the same subset are neighbours. Thus, every pixel in a given coding set may be updated simultaneously, which is the same as updating them sequentially since all neighbours remain un-updated. The coding sets are visited sequentially, updating all voxels in each set simultaneously. We show coding schemes for three-dimensional images with various neighbourhood configurations in Appendix B.

2.5 Algorithm

The elements described so far define a family of EM-like algorithms as described in Celeux et al. (2003). They are used to perform image segmentation using a mixture model with an MRF prior. They all fit in a general algorithm consisting of three core steps, with minor variations at each step. On iteration t:

1. **(C-step)** Form an estimate of the current labels $\boldsymbol{z}^{(t)}$ to be used as neighbours; either discrete (for the pseudolikelihood approximation) or continuous (for the mean-field approximation).
2. **(E-step)** Calculate ${\tau_{ij}}^{(t)}$ using (2.19), using $\boldsymbol{z}^{(t)}$ from the C-step where needed for $\boldsymbol{z}_{\partial \boldsymbol{i}}{}^{(t)}$:

$$ {\tau_{ij}}^{(t)} = \frac{p(\boldsymbol{Z}_i = \boldsymbol{e}_j | \boldsymbol{z}_{\partial \boldsymbol{i}}{}^{(t)}, \boldsymbol{\Psi}^{(t-1)}) \phi(\boldsymbol{y}_i; {\boldsymbol{\mu}_j}^{(t-1)}, {\boldsymbol{\Sigma}_j}^{(t-1)})}{\sum_{k=1}^{g} p(\boldsymbol{Z}_i = \boldsymbol{e}_k | \boldsymbol{z}_{\partial \boldsymbol{i}}{}^{(t)}, \boldsymbol{\Psi}^{(t-1)}) \phi(\boldsymbol{y}_i; {\boldsymbol{\mu}_k}^{(t-1)}, {\boldsymbol{\Sigma}_k}^{(t-1)})}. $$

3. **(M-step)** Maximise Q with respect to $\boldsymbol{\Theta}$ to obtain the intensity parameters. The update equations for the mixture parameters remain the same (with the new definition of τ_{ij}):

$$ {\boldsymbol{\mu}_j}^{(t)} = \frac{\sum_{i=1}^{n} {\tau_{ij}}^{(t)} \boldsymbol{y}_i}{\sum_{i=1}^{n} {\tau_{ij}}^{(t)}} $$
$$ {\boldsymbol{\Sigma}_j}^{(t)} = \frac{\sum_{i=1}^{n} {\tau_{ij}}^{(t)} (\boldsymbol{y}_i - {\boldsymbol{\mu}_j}^{(t)})^T (\boldsymbol{y}_i - {\boldsymbol{\mu}_j}^{(t)})}{\sum_{i=1}^{n} {\tau_{ij}}^{(t)}}. $$

These steps are repeated until the parameters or observed log-likelihood stop changing up to some relative tolerance. The observed log-likelihood $f(\boldsymbol{y})$ cannot be found exactly, but can be

approximated by using the pseudolikelihood or mean-field versions of $p(\boldsymbol{z})$:

$$\log \mathcal{L}(\boldsymbol{\Theta}, \boldsymbol{\Psi}; \boldsymbol{y}, \boldsymbol{z}) = \sum_{i=1}^{n} \sum_{j=1}^{g} z_{ij} (\log f_j(\boldsymbol{y}_i; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) + \log p(\boldsymbol{e}_j | \boldsymbol{z}_{\partial \boldsymbol{i}}, \boldsymbol{\Psi})).$$

On the C-step, the ICM or ICE/mean-field algorithms (2.21) and (2.20) are be used to find the current estimate of the tissue labels. Alternatively, each $\boldsymbol{z}_i{}^{(t+1)}$ may be sampled multinomially from the current posterior density i.e. $\tau_{ij}, j = 1, \ldots, g$ for each i, as in Gibbs sampling.

Many other methods instead treat $\boldsymbol{z}^{(t)}$ from the C-step as the true values of $\boldsymbol{z}$ and maximise the joint log-likelihood rather than the Q-function, thus skipping the E-step. This algorithm corresponds to 'C-M' in the framework described, though in the literature is usually referred to as 'restoration-maximisation'. When the C-step uses ICM, this is the procedure described by Besag (1986), though he primarily focuses on the case where the parameters are known and no M-step is required. However, some authors found that treating $\boldsymbol{z}^{(t)}$ as truth and omitting the E-step lead to bias in the intensity parameter estimates and image segmentations/reconstructions (Titterington, 1984; Little and Rubin, 1983; Qian and Titterington, 1991). Rather, $z^{(t)}$ should only be used to approximate the neighbours $\boldsymbol{z}_{\partial \boldsymbol{i}}$ rather than being treated as truth.

When the 'C-M' algorithm uses the mean-field update in the C-step, this is the procedure described by Zhang (1992). Although this method also treats the $z^{(t)}$ from the C-step as truth in the M-step, it is worth noting that the mean-field C-step produces $\langle \boldsymbol{z}_i \rangle$ that are the same as the τ_{ij} produced in the E-step. The only difference is that the former is performed sequentially while the latter is performed simultaneously.

One further set of variations consists of reordering the steps E-C-M. The posterior probabilities τ_{ij} of the E-step are used to generate a realisation $z^{(t)}$ in the C-step. This realisation is then used as truth in the M-step, maximising the joint log-likelihood rather than the Q-function. When ICM is used in the C-step, this is Celeux and Govaert (1992)'s Classification EM algorithm. When one cycle of Gibbs sampling is used in the C-step, this is Celeux and Diebolt (1985)'s Stochastic EM algorithm. When multiple realisations are generated in the C-step with the average parameter estimate over these realisations taken in the M-step, this is the approach favoured by Qian and Titterington (1991).

Celeux et al. (2003) explored many of the variations, finding CEM with Gibbs sampling on the C-step and using the mean-field approximation to be superior to other options.

While these algorithms work and are commonly used in practice, there appear to be no convergence guarantees on the modified E-M algorithm using pseudolikelihood or mean-field theory. As previously established, so long as the C-step is performed sequentially, both the mean-field and ICM versions are guaranteed to converge. However, the monotonic increase and convergence of the observed log-likelihood by maximising Q does not hold in general. Gao and Song (Gao and Song, 2011) showed that these properties are preserved, but only if the

pseudolikelihood consists of a product of *marginal* likelihoods e.g. $\prod_i p(\boldsymbol{z}_i)$, not *conditional* likelihoods $\prod_i p(\boldsymbol{z}_i|\boldsymbol{z}_{\partial i})$. The reliance on and need for an estimate $\boldsymbol{z}^{(t)}$ is the cause of the problem, as the expected values are somewhat reliant on the current realisation.

2.6 Conclusion

In this chapter we have given a brief overview of the following topics, with focus on their use in image segmentation:

- Gaussian mixture models and their solution by Expectation Maximisation,
- Markov random fields and pseudolikelihood/mean-field approximations to improve tractability,
- a combined Gaussian mixture model with MRF prior, and how it may be solved using EM in combination with the pseudolikelihood or mean-field approximation.

Future chapters will draw on these elements, with different specific forms of MRF.

Chapter 3

Homogeneous Potts MRF

3.1 Introduction

Tissue segmentation provides valuable information for brain tissue analysis, enabling study of how the volume and shape of various brain structures are affected by injury, stroke or disease. Fully-automated methods to segment grey matter (GM), white matter (WM), and cerebrospinal fluid (CSF) save the many man-hours required to manually segment MRI, and avoid problems such as inter- and intra-observer bias.

As outlined in Chapter 1, there are many techniques to segment the brain. One very popular method for segmentation of the brain is to use a mixture-MRF to model the MRI. Each tissue's voxel intensities are assumed to have a Gaussian distribution, while the tissues themselves use a Markov random field as a prior to incorporate spatial dependence, smoothing the segmentation and eliminating noise.

Introduced to the field of image segmentation by Besag (1986), the Potts model from statistical mechanics (Potts, 1952) is ubiquitous as the specific choice of MRF in the model. A Gaussian mixture with the Potts MRF as a prior is used in many major segmentation tools and pipelines that are considered gold standards in automatic segmentation. These include FAST (Zhang et al., 2001) and Niftyseg (Cardoso et al., 2009), and Atropos (Avants et al., 2011). Even when an MRF is used only as a post-hoc smoother rather than incorporated into the probability model itself, the Potts MRF is used. This is the case in the leading and commonly-used atlas-based segmentation pipeline FreeSurfer (Ashburner and Friston, 2005).

In the Potts MRF, a parameter β is used to control the spatial regularisation applied by the MRF. It is very common in image segmentation to fix β *a priori* rather than estimating it. However, there is no clear method to determine an appropriate value. Rather, β has been determined by manual tuning, in order to obtain a (subjective) visually suitable result. In the field of MR segmentation, all of the commonly-used segmentation software packages previously mentioned require the user to specify β, but also set a default value that has been configured by

the developers. FAST uses $\beta = 1$, NiftySeg uses $\beta = 0.25$, while Atropos uses $\beta = 0.3$.

If β is set inappropriately, the resulting segmentation could retain too much noise, or could be so smooth as to obscure important fine brain features. Determining an appropriate β value is image dependent and often subjective. In addition, as we will see, the appropriate β value is not necessarily fixed from patient to patient or image to image, particularly in MRI where small changes in machines and acquisition parameters affect the reconstructed image. Automatic estimation of β removes the need to guess appropriate smoothing values, and recognises that different parameter values may be needed for each individual image.

In this chapter, we introduce and validate a method to adaptively determine the level of spatial smoothing in the Potts model on a per-image basis. This is done in a fully-automatic fashion rather than requiring manual tuning. We achieve this by statistical estimation of the underlying smoothing value in the Potts model for each image. This must be achieved in a tractable manner. In this chapter we explore the use of the *maximum pseudolikelihood estimator* (MPLE) (Besag, 1974) for the Potts MRF in MR segmentation. Parameter estimation in this way has been performed before, e.g. (Celeux et al., 2003), but rarely in brain MR segmentation. We compare it to the least-squares estimator implemented in the software 'Expectation Maximisation Segmentation' (EMS) (Van Leemput et al., 1999b) as well as common default fixed values. Additionally, we show that choosing the "wrong" values for the MRF parameters can lead to a severe loss of segmentation accuracy. The primary contributions of this chapter are

- the comparison of estimation with the MPLE against commonly-used fixed parameter values or the least-squares estimator, in the context of brain segmentation.
- the systematic and detailed study of how the MPLE performs under different neighbourhood size and MRF approximations.

3.1.1 Aim

Aim 1 is to develop a method to adaptively determine the amount of spatial regularisation applied in mixture-MRF segmentation with the simplified Potts MRF. The desired properties are:

- Property 1: The method should be able to adapt to the characteristics of each individual image, as some images may require less smoothing than others. Thus, it should operate on a per-image basis.
- Property 2: The method should not require training data, though may be able to make use of it if available. This will allow it to be robust to images dissimilar to those in the training set (particularly pertinent as it is unclear what "dissimilar" means for a Potts MRF). This also removes the need to obtain training data - images with manual segmentations, or segmentations that will be treated as ground truth.

Aim 2 pertains to the practical aspects of the developed method:

- Property 3: The method should be computationally tractable.
- Property 4: The method should be straightforward to incorporate into existing segmentation algorithms. While the mixture-Potts image model is very common throughout MR segmentation, it is often only the basis of a more sophisticated algorithm. For example, it has been extended to incorporate bias-field correction (Van Leemput et al., 1999a; Zhang et al., 2001), partial-volume estimation (Noe and Gee, 2001; Shattuck et al., 2001; Van Leemput et al., 2003), as well as the use of anatomical priors (Van Leemput et al., 1999b; Cardoso et al., 2011). If the parameter estimation method is modular, it can easily be inserted into these methods without much additional work.

In this chapter, we will see that the maximum pseudolikelihood estimator satisfies all desired properties in the first two aims, making it particularly suited to MR segmentation.

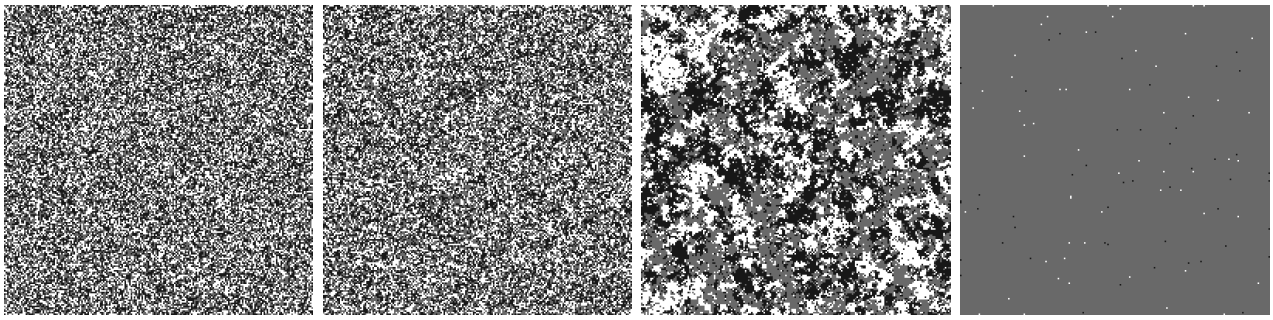
Aim 3 is to validate the method on real data to determine if and in what circumstances automatic determination of β can be of value. We will also compare the method to existing mixture-MRF algorithms: NiftySeg (Cardoso et al., 2009), Atropos (Avants et al., 2011), and FAST (Zhang et al., 2001), which all use fixed β values, and EMS (Van Leemput et al., 1999b), which uses the least-squares estimator.

3.2 Background

3.2.1 Potts MRF

The Potts model (Potts, 1952) is used almost exclusively as a prior for the tissue labels in image segmentation. It originated from the field of statistical mechanics as a many-state extension to the Ising model (Ising, 1925). The Ising model is a model of ferromagnetism, giving the probability of particular configurations of magnetic spins in a lattice. Each site can have one of two states, or spins ($g = 2$). This is akin to each voxel taking one of two labels, a binary image. The Potts model is an extension that allows each site to have $g > 2$ states; for us, an image with more than two tissue labels. Use of this model in image segmentation and restoration was popularised by Besag (1986). The simplified form - the homogeneous Potts field in the absence of an external field - is the one most commonly used in imaging. Recalling that $\boldsymbol{Z}_i$ indicates the label of voxel i of which there are g possibilities, the local conditional form of the Potts MRF is:

$$\begin{aligned} p(\boldsymbol{Z}_i = \boldsymbol{e}_j | \boldsymbol{z}_{\partial \boldsymbol{i}}; \beta) &\propto \exp(\beta u_{ij}), \quad j = 1, \ldots, g \\ u_{ij} &= \sum_{m \in \partial \boldsymbol{i}} \frac{z_{mj}}{\delta_{im}} \\ \beta &\geq 0, \end{aligned} \tag{3.1}$$



(a) 0.1 (b) 0.3 (c) 1.0 (d) 5.0

Figure 3.1: Example Potts MRFs with various smoothing values β and 4 neighbours

where δ_{im} is the distance between voxels i and m. The term u_{ij} is the number of neighbours of voxel i that have label j (scaled by their distance to i). Thus the Potts MRF assigns the log-probability of each voxel's label to be proportional to the number of neighbours matching that label. The parameter β controls the strength of this relationship.

The joint distribution can be written

$$\begin{aligned} p(\boldsymbol{z}) &= \frac{1}{C} \prod_{i=1}^{n} \prod_{j=1}^{g} \exp(\beta u_{ij})^{z_{ij}}, \text{ where} \\ C &= \sum_{\boldsymbol{z}'} \prod_{i=1}^{n} \prod_{j=1}^{g} \exp(\beta u_{ij}))^{z'_{ij}}. \end{aligned} \tag{3.2}$$

It is also common to see the joint distribution written in terms of pairwise potentials:

$$p(\boldsymbol{z}) = \frac{1}{\tilde{C}} \exp(\tilde{\beta} \sum_{\substack{i,m \\ \text{neighbours}}} \frac{\boldsymbol{z}_i^T \boldsymbol{z}_m}{\delta_{im}}),$$

where the sum is over all voxel pairs that are neighbours and $\boldsymbol{z}_i^T \boldsymbol{z}_m$ is 1 if voxel i's label matches its neighbour m's, and 0 otherwise. This form has $\tilde{\beta} = 2\beta$ because each neighbour pair is double-counted in the previous equation.

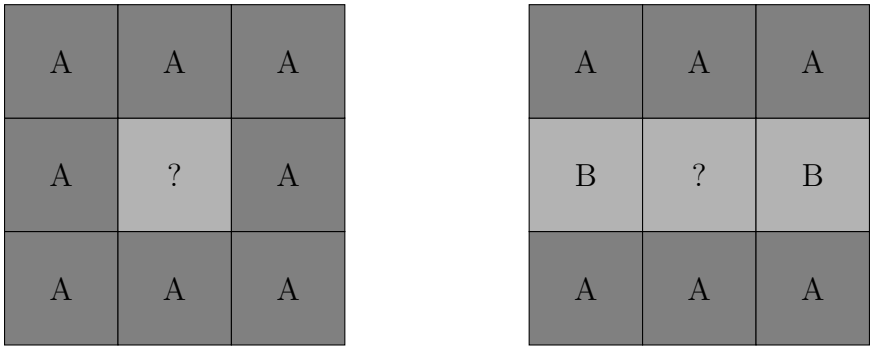
The normalising constant C consists of a sum over all g^n possible label configurations $\boldsymbol{z}'$, which is intractable. The pointwise *pseudolikelihood* and *mean-field* approximations, discussed in the previous chapter, are often used in place of $p(\boldsymbol{z})$.

3.2.2 Spatial regularisation parameter

The primary focus of this chapter is on the spatial regularisation parameter β. In the original context of ferromagnetism models, β is the inverse of the system's thermodynamic temperature, $\beta = \frac{1}{k_B T}$, where k_B is Boltzmann's constant and T is the absolute temperature of the system. When β is small, the system is disordered (not magnetic); the dependence between neighbouring sites is weak. When β is large and positive (as the temperature lowers), the material exhibits

(a) original MRI (b) 0.1 (c) 0.3 (d) 1 (e) 5

Figure 3.2: Segmentations with various smoothing values β



(a) Isolated pixels of noise can be smoothed (b) Thin features may also be smoothed

Figure 3.3: Example pixel configurations with two labels, A and B.

ferromagnetism. In statistical mechanics the focus is typically on studying critical temperatures (β values) at which the material transitions from one state to the other - for magnetism, this is known as the Curie temperature.

In image segmentation, β controls the strength of the spatial dependence between neighbours; it is constrained to be non-negative (a negative β would encourage neighbouring voxels to have *different* labels to each other). Figure 3.1 shows examples of the Potts MRF at various β values. If $\beta = 0$, neighbours are independent; each label is equally likely and independent of its neighbours. When $\beta = \infty$, the model becomes a majority-voting system, where the label of pixel i is purely determined by the label of the majority of its neighbours. Intermediate values trade off between the two.

When the Potts prior is combined with the Gaussian probabilities over each tissue label's intensities, the probability of a voxel having a given label given its intensity and neighbouring labels is

$$p(\boldsymbol{Z}_i = \boldsymbol{e}_j | \boldsymbol{z}_{\partial \boldsymbol{i}}, y_i) \propto \exp(\beta u_{ij}) f(y_i | \boldsymbol{e}_j)$$

It is evident that β acts to control the balance of intensity and spatial information. When $\beta = 0$, only the intensity distribution contributes to the label and not a voxel's neighbours. When As β gets larger, the MRF has higher precedence over the intensities, producing, smoother segmentations than at lower values of β. In this way, β may be thought of as a spatial

regularisation parameter. This can be seen in figure 3.2, which shows segmentations produced by the EM algorithm described in the previous chapter, at various β values. At $\beta = 0$, one retrieves the standard Gaussian mixture model, but with fixed mixing proportions $\Pr(\boldsymbol{Z}_i = \boldsymbol{e}_j) = \frac{1}{g}$.

Choice of β can be critical in order to smooth noise without blurring features of interest. Consider figure 3.3, which shows a voxel marked '?' and the labels of its neighbouring voxels. The question is what label to assign voxel '?'. Suppose there are two labels for an image - 'A' (dark) and 'B' (light). The left-hand neighbourhood shows an isolated voxel of noise that should be smoothed by the MRF. Denoting voxel '?' by the subscript i, its probability to have either label is

$$p(\boldsymbol{Z}_i = \mathrm{A}|y_i) \propto \exp(8\beta) f(y_i|\mathrm{A})$$
$$p(\boldsymbol{Z}_i = \mathrm{B}|y_i) \propto \exp(0\beta) f(y_i|\mathrm{B})$$

Since voxel '?' is light in colour, $f(y_i|\mathrm{B})$ is large compared to $f(y_i|\mathrm{A})$. However, the MRF probability $\exp(\beta u_{ij})$ is higher for label B than A. A sufficiently large β will overcome the intensity pdf to assign voxel '?' to label 'A' with high probability.

On the other hand, consider the right-hand neighbourhood, which shows a thin line. Here,

$$p(\boldsymbol{Z}_i = \mathrm{A}|y_i) \propto \exp(6\beta) f(y_i|\mathrm{A})$$
$$p(\boldsymbol{Z}_i = \mathrm{B}|y_i) \propto \exp(2\beta) f(y_i|\mathrm{B})$$

Depending on the resolution of the MRI, it is likely that we wish to preserve this feature, as it could represent some fine anatomical region in the brain (for example, extrasulcal CSF between cortical folds). If β is too large, the MRF probability will dominate and label 'A' will be assigned, obscuring the feature. Thus β must be selected sufficiently large to remove noise as in the left-hand neighbourhood, but not so large as to oversmooth the right-hand neighbourhood. Clearly, this also depends on the intensity distribution $f(y_i|\boldsymbol{z}_i)$ and its spread compared to that of $p(\boldsymbol{z}_i|\boldsymbol{z}_{\partial i})$; there is no immediately obvious method to select β.

3.2.3 Related work

Here we briefly discuss methods to determine or set β in two different fields - the statistical literature from which much of the theory and methods originated, and the medical imaging literature. The application of methods from the former to the latter requires consideration of ease of implementation as well as computational efficiency, given the large size of medical images.

3.2.3.1 Fixed-parameter segmentation

For the reasons mentioned previously, it is common to have β fixed, with the value chosen manually.

The founding statistical papers on image restoration all used small two-dimensional images to demonstrate the methods. Besag (1986) found $\beta = 1.5$ worked well for 6-label synthetic images. Jubb and Jennison (1991) used $\beta = 4$ on very noisy synthetic binary images but mentioned $\beta = 1$ to be a good general value. Owen (1986) suggested a smaller value of $\beta = 0.7$ for binary images with convoluted edges (the example used was a map of the coastline of northern Scotland). All examples used a 2D neighbourhood consisting of the 8 immediately orthogonal and diagonal neighbours. All of these values were tuned manually.

In medical image segmentation, it is perhaps even more common to use fixed β values. In papers making use of the mixture-Potts formulation use of $\beta = 1$ is popular (McLachlan et al., 1996; Jubb and Jennison, 1991; Zhang et al., 2001). Many of the leading, publicly-available software packages for MRI segmentation implement a mixture-MRF model as the basis of their image segmentation pipelines, with fixed β values.

The 'FAST' tool from the Oxford Centre for Functional MRI of the Brain's software library (Zhang et al., 2001) uses $\beta = 1$ for a neighbourhood consisting of the orthogonal neighbours only (6 for a 3D MRI). Its implementation uses $\beta = 0.1$ for all neighbours in the $3 \times 3 \times 3$ neighbourhood centred on the pixel of interest (26 neighbours for a 3D MRI). Advanced Normalization Tools' 'Atropos' (Avants et al., 2011) uses $\beta = 0.3$ with 26 neighbours.

NiftySeg, developed at University College London, has two main segmentation algorithms. The first (`seg_EM`) uses a default fixed $\beta = 0.25$. The second (`seg_LoAd`) uses a more advanced form of the Potts model. It requires use of an anatomical atlas to further split the tissue classes: CSF into external CSF (between the sulci and gyri) and internal CSF (in the ventricles), GM into deep GM (within the brain) and cortical GM (forming the folds on the outer boundary of the brain). It uses $\beta = 0.5$ for tissues anatomically likely to be near each other - for example, internal CSF with deep GM - and $\beta = 0.3$ for tissues anatomically unlikely to be near each other - for example, internal CSF with cortical GM. We will explore this form of tissue-based smoothing in Chapter 4.

Two other major segmentation tools merit mention. SPM (Ashburner and Friston, 2005, 1997) and FreeSurfer (Fischl et al., 2002) use mixtures to represent the image intensities, but do not use MRFs in the label priors. Rather, these are atlas-driven methods that rely on registering the input brain to that of a brain that is already labelled (an "atlas") in order to propagate tissue labels, using the image intensities to aid the process. Both of these allow an MRF to be used to smooth the segmentation after it has been obtained, but these MRFs are not incorporated into the image model and are not given intensity information.

3.2.3.2 MRF parameter estimation

When it comes to automatic estimation of β, there is a large disconnect between the statistical and medical imaging worlds. While MRF parameter estimation has been extensively investigated

in the statistical field, little of it has been implemented in the corresponding medical imaging segmentation software. This is partly due to the fact that the statistical papers were largely written before modern advances in computing power. These papers tended to focus more on statistical aspects of the estimators (e.g. asymptotic behaviour and efficiency; recovery of parameter values); images used as demonstrations were typically two-dimensional, artificial or simulated, binary (two-colour), and less than 256x256 pixels. On the other hand, medical images are three-dimensional, may have many classes, and typically contain millions of voxels. Additionally, the focus is not so much on asymptotic behaviour or recovery of β as the true underlying value is irrelevant to the application, but on accuracy of the segmentation $\boldsymbol{z}$. Investigating whether a method can be successfully transferred into medical image segmentation requires careful consideration.

Traditional statistical approaches to MRF parameter estimation fall broadly into two classes - stochastic or deterministic. In general, they aim to find maximum-likelihood estimates of the MRF parameters. Stochastic approaches are typically quite slow as they rely on repeated sampling, while deterministic approaches can become stuck in local maxima.

Stochastic approaches typically use Markov Chain Monte Carlo (MCMC) and in particular Gibbs sampling in order to generate samples from the MRF. These are used to calculate expectations empirically, which are used within e.g. a gradient descent algorithm (Younes, 1991; Jalobeanu et al., 2002) or EM (Qian and Titterington, 1991) to find approximate maximum-likelihood estimates. While these can produce estimates that are more accurate than likelihood approximations, they are computationally expensive. This is because they require many samples to be drawn on each iteration.

Deterministic options focus on replacing the MRF probability with a tractable approximation, and maximising that. A precursor to these approaches is the 'coding method' of Besag (1974). Here, the image voxels are divided into 'coding sets' such that no two voxels in a given set are neighbours (see section 2.4.1 and Appendix B for further detail). Within each coding set, the pseudolikelihood is the true likelihood. Thus, the maximum-pseudolikelihood estimate for each coding set is the maximum-likelihood estimate. Since the pseudolikelihood is computationally tractable, an estimate is readily obtained for each coding set. However, as mentioned by Besag, it is unclear how the estimates should be combined. Possolo (1986) showed that the estimate for each coding set is statistically consistent (converges to the true parameter) as the lattice becomes infinitely large. However in practice, with finite-size lattices, the estimates may differ significantly between coding sets, making averaging unsatisfactory (Kashyap and Chellappa, 1983).

The coding method was later extending to maximisation of the point-pseudolikelihood over the entire image (Besag, 1975). Qian and Titterington (1992) extended this to investigate and compare maximisation of the point-, line- and block-pseudolikelihoods, demonstrated on small (64x64 pixel) two-dimensional satellite images. This can be combined with simulated annealing

as described in Lakshmanan and Derin (1989). Maximising the pseudolikelihood and mean-field approximations was also investigated by Dunmur and Titterington (1998) and found to be superior to simple thresholding (which is what the standard mixture model amounts to), though only for binary-coloured, small images. Alternatively there is the *histogram estimator* (also known as the *least-squares estimator*); this was originally developed for binary images and is analogous to logistic regression (Possolo, 1986) and later extended to multi-labelled images (Derin and Elliott, 1987; Gurelli and Onural, 1994; Borges, 1999).

As mentioned, very few of these methods for automatically determining β have been incorporated into medical image segmentation. The few examples include use of MCMC approaches in segmentation of satellite images (Pereyra et al., 2013) and functional MRI (Woolrich et al., 2005). However, MCMC is not generally computationally tractable due to the large number of simulations required, so does not satisfy property 3 defined in the Aims.

In terms of deterministic approaches, Woolrich and Behrens (2006) used approximate variational Bayes for segmentation of functional MRI, which is computationally tractable and also uses the Potts MRF. However, in order to achieve this, the discrete tissue labels must be approximated by continuous versions (a logistic transform of a Gaussian Markov random field). The MRF is no longer the same as the Potts model. This inhibits straightforward incorporation into existing methods (Property 4 defined in the Aims).

The least-squares estimator of Derin and Elliott (1987) was studied for the Potts MRF in brain segmentation by Van Leemput et al. (1999b) and made available in the 'Expectation Maximisation Segmentation' tool. It involves a least-squares regression, but must first construct a neighbourhood histogram of the image at each iteration. As we will see, this restricts its use and potentially introduces bias into the estimates.

On the other hand, methods that make use of likelihood approximations (e.g. mean-field or pseudolikelihood) and simply maximise them with respect to the MRF parameters are computationally tractable, but have not been extensively studied in MR segmentation. Chaari et al. (2013) used maximum pseudolikelihood estimation with the mean-field approximation with application to brain fMRI (each voxel classified into two classes). In this work the brain was first subdivided into 600 anatomical regions before fitting a mixture-MRF to each individual region (a much smaller number of voxels), so the suitedness of MPL to segmentation of large datasets in particular was notaddressed In the remainder of this chapter, we will show that maximum-pseudolikelihood (or mean-field approximation) methods are particularly suited to be adapted into the existing mixture-MRF model for MRI segmentation of the whole brain volume simultaneously.

In this chapter, we focus on maximum-pseudolikelihood estimators (where we use this term to also include the mean-field approximation, making clear when a particular form is meant). The pseudolikelihood estimator is computationally tractable (property 3 defined in the Aims), is applied to individual images separately (property 1), and does not need training data (property

2). It seems particularly suited to large three-dimensional medical images and has not been applied to them before. Additionally, we will see that it is particularly well-suited to be incorporated into the existing EM framework ubiquitously used for mixture-MRF segmentation (property 4). We also choose the least-squares estimator for comparison, as a method that has already been applied to MR segmentation with code made openly available.

3.3 Method

The image model is the same as that presented in the previous chapter. Let $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ be random variables where Y_i is the intensity of voxel i in an n-voxel MR volume. We consider the case of a single-channel MRI, i.e. Y_i is scalar, though the theory is readily applied to a multichannel/multivariate case. Let g be the number of tissue classes. We use $g = 3$: cerebrospinal fluid (CSF), grey matter (GM) and white matter (WM). Let $\boldsymbol{Z} = (\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n)$ be random variables giving the tissue classification or label of each voxel. $\boldsymbol{Z}_i$ indicates the tissue label of voxel i. Let $\boldsymbol{e}_j$ be the indicator vector with a 1 in the jth position and 0 elsewhere. $\boldsymbol{Z}_i = \boldsymbol{e}_j$ if and only if voxel i is tissue j. The set of voxels that neighbour voxel i is denoted $\boldsymbol{\partial i}$, and $\boldsymbol{z}_{\boldsymbol{\partial i}}$ are the labels of all such neighbours. Lowercase letters e.g. y_i and $\boldsymbol{z}_i$ are used to represent realisations of $\boldsymbol{Y}_i$ and $\boldsymbol{Z}_i$.

The distribution of the intensities given their label is written $f(y_i|\boldsymbol{Z}_i = \boldsymbol{e}_j)$, or just $f(y_i|\boldsymbol{e}_j)$. In this thesis, we assume the intensities of voxels to be independently and normally distributed, given their label. The labels are distributed according to the Potts MRF (3.1) and (3.2).

$$\begin{aligned} Y_i|(\boldsymbol{Z}_i = \boldsymbol{e}_j) &\sim \mathcal{N}(\mu_j, \sigma_j^2) \\ \boldsymbol{Z} &\sim \text{Potts}(\beta), \beta \geq 0. \end{aligned}$$

The joint log-likelihood is approximated using the pseudolikelihood or mean-field approximations,

$$f(\boldsymbol{y}, \boldsymbol{z}; \boldsymbol{\Theta}, \beta) = \prod_{i=1}^{n} \prod_{j=1}^{g} \left(\phi(y_i; \mu_j, \sigma_j^2) \frac{\exp(\beta u_{ij})}{\sum_{k=1}^{g} \exp(\beta u_{ik})} \right)^{z_{ij}}$$

where $\boldsymbol{\Theta} = (\mu_1, \ldots, \mu_g, \sigma_1^2, \ldots, \sigma_j^2)$ are the intensity parameters, and ϕ is the normal probability density function.

In fitting the image model to an MRI, we need to determine the optimal mixture and MRF parameters, as well as recover the optimal segmentation. As described in the previous chapter, EM combined with the pseudolikelihood or mean-field approximations is used to determine estimates of the Gaussian intensity parameters and segmentation. We now focus specifically on two methods for estimation of β: maximum pseudolikelihood (including the mean-field variant), and the least-squares estimator. The former has not been studied for medical image segmentation, while the latter is chosen for comparison as it has been implemented and made

openly available in this context (Van Leemput et al., 1999b).

3.3.1 Maximum Pseudolikelihood Estimation

The pseudolikelihood and mean-field approximations are used in the EM algorithm to make computation of the MRF tractable. These approximations are given by:

$$\tilde{p}(\boldsymbol{z}) = \prod_{i=1}^{n} \prod_{j=1}^{g} \left(\frac{\exp(\beta u_{ij})}{\sum_{k=1}^{g} \exp(\beta u_{ik})} \right)^{z_{ij}}.$$

The pseudolikelihood approximation (Besag, 1986) calculates the number of neighbours as

$$u_{ij} = \sum_{m \in \boldsymbol{\partial i}} \frac{z_{mj}}{\delta_{im}}.$$

The mean-field approximation of the likelihood for the Potts MRF (Zhang, 1992) uses the same equation but replaces the discrete z_{mj} with their expected values under the approximation:

$$u_{ij} = \sum_{m \in \boldsymbol{\partial i}} \frac{\langle z_{mj} \rangle}{\delta_{im}},$$

where $\langle z_{ij} \rangle$ satisfies

$$\langle z_{ij} \rangle = \frac{\exp(\beta u_{ij})}{\sum_{k=1}^{g} \exp(\beta u_{ik})}$$

or

$$\langle z_{mj} \rangle = \frac{\exp(\beta u_{ij}) f(y_i | \boldsymbol{Z}_i = \boldsymbol{e}_j)}{\sum_{k=1}^{g} \exp(\beta u_{ik}) f(y_i | \boldsymbol{Z}_i = \boldsymbol{e}_k)}. \tag{3.3}$$

The former version is a mean-field approximation to $p(\boldsymbol{z})$ only. We use the latter, the mean-field approximation for $p(\boldsymbol{z}|\boldsymbol{y})$. Not incorporating the intensity information into the mean-field approximation can lead to biased results or a trivial approximation (Celeux et al., 2003; Archer and Titterington, 2002).

Since these approximations are the same up to choice of the neighbours $\boldsymbol{z}_{mj}$, we refer to both as the "pseudolikelihood approximation" for convenience, and will make clear if a statement applies to only one of the approximations.

Since the pseudolikelihood and mean-field approximations are already used when solving a fixed-β mixture-Potts model with EM (as outlined in Chapter 2), it is natural to consider maximising the pseudolikelihood or mean-field approximation of the likelihood to determine an estimate for β. Here, we will use the term 'maximum pseudolikelihood' (MPL) to mean maximisation of *either* the pseudolikelihood (PL) or mean-field (MF) approximations of the likelihood, as they differ only in the calculation of the neighbours u_{ij}. Where applicable we will make clear if a particular approximation is meant.

The pseudolikelihood for the Potts MRF is

$$\tilde{p}(\boldsymbol{z};\beta) = \prod_{i=1}^{n}\prod_{j=1}^{g}\left(\frac{\exp(\beta u_{ij})}{\sum_{k=1}^{g}\exp(\beta u_{ik})}\right)^{z_{ij}},$$

where calculation of the neighbour counts u_{ij} depends on the particular approximation used.

The *maximum pseudolikelihood estimator* (MPLE), suggested by Besag (1975), estimates β simply by maximising the log-pseudolikelihood.

$$\hat{\beta}_{MPL} = \arg\max_{\beta}\sum_{i=1}^{n}\sum_{j=1}^{g} z_{ij}\left(\beta u_{ij} - \log(\sum_{k=1}^{g}\exp(\beta u_{ik}))\right). \tag{3.4}$$

To extend this to the case of a hidden MRF, the log joint-pseudolikelihood $f(\boldsymbol{y}|\boldsymbol{z})\tilde{p}(\boldsymbol{z})$ may be maximised instead. Since the intensity component does not depend on β, this is equivalent to (3.4). In the context of EM, one may maximise the Q-function. Again, the intensity component is constant with respect to β, so this becomes:

$$\hat{\beta}_{MPLQ} = \arg\max_{\beta}\sum_{i=1}^{n}\sum_{j=1}^{g} {\tau_{ij}}^{(t)}\left(\beta {u_{ij}}^{(t)} - \log(\sum_{k=1}^{g}\exp(\beta {u_{ik}}^{(t)}))\right). \tag{3.5}$$

In the above, ${\tau_{ij}}^{(t)}$ are calculated on the E-step while the neighbours ${u_{ij}}^{(t)}$ are calculated on the C-step.

A proof of consistency for the maximum pseudolikelihood estimate for a *visible* MRF (i.e., $\boldsymbol{z}$ is observed) as the lattice size n tends to infinity was sketched in Besag (1975) and proved rigorously by Geman and Graffigne (1986). It has also been shown that the maximum-*likelihood* estimate for a hidden MRF ($\boldsymbol{z}$ is hidden and only $\boldsymbol{y}$ is observed) is consistent (Comets and Gidas, 1992). However, we are unaware of any similar proofs for the maximum-*pseudolikelihood* estimate for a *hidden* MRF.

An MRF that is log-linear in its parameters has concave log-pseudolikelihood and Q-function with respect to the parameters, and its gradient/Hessian are obtainable in closed form. Thus it is amenable to maximisation by e.g. gradient ascent.

3.3.1.1 Concavity

To see that the Q-function is concave, consider a general MRF with conditional probability

$$p(\boldsymbol{z}_i|\boldsymbol{z}_{\partial i};\boldsymbol{\Psi}) = \frac{\exp(-U_i(\boldsymbol{z}_i|\boldsymbol{z}_{\partial i};\boldsymbol{\Psi}))}{C_i(\boldsymbol{\Psi})}$$

$$C_i(\boldsymbol{\Psi}) = \sum_{k=1}^{g}\exp(-U_i(\boldsymbol{e}_k|\boldsymbol{z}_{\partial i};\boldsymbol{\Psi})).$$

We assume only that U_i is linear in its parameters $\boldsymbol{\Psi}$. The Hessian (denoted $\nabla^2{}_{\boldsymbol{\Psi}}$) is

$$\begin{aligned}\nabla^2{}_{\boldsymbol{\Psi}} \log p(\boldsymbol{z}_i|\boldsymbol{z}_{\partial i};\boldsymbol{\Psi}) &= -\nabla^2_{\boldsymbol{\Psi}} U_i(\boldsymbol{z}_i|\boldsymbol{z}_{\partial i};\boldsymbol{\Psi}) - \nabla^2_{\boldsymbol{\Psi}} \log C_i(\boldsymbol{\Psi}) \\ &= -\nabla^2_{\boldsymbol{\Psi}} \log C_i(\boldsymbol{\Psi})\end{aligned} \tag{3.6}$$

since U_i is linear in $\boldsymbol{\Psi}$.

The (conditional) Fisher information matrix under $p(\boldsymbol{z}_i|\boldsymbol{z}_{\partial i})$ is the negative expectation of the Hessian:

$$\begin{aligned}\mathcal{I}_i(\boldsymbol{\Psi}) &= -\,\mathbb{E}_{\mathrm{Z_i}|\mathrm{Z}_{\partial i}=\mathrm{z}_{\partial i}}\left[\nabla^2 \log p(\boldsymbol{Z}_i|\boldsymbol{z}_{\partial i};\boldsymbol{\Psi})\right] \\ &= \mathbb{E}_{\mathrm{Z_i}|\mathrm{Z}_{\partial i}=\mathrm{z}_{\partial i}}\left[\nabla^2_{\boldsymbol{\Psi}} \log C_i(\boldsymbol{\Psi})\right] \\ &= \nabla^2_{\boldsymbol{\Psi}} \log C_i(\boldsymbol{\Psi}),\end{aligned}$$

where the last line follows because C_i does not depend on $\boldsymbol{z}_i$, only $\boldsymbol{z}_{\partial i}$ and $\boldsymbol{\Psi}$. Thus,

$$\nabla^2_{\boldsymbol{\Psi}} \log p(\boldsymbol{z}_i|\boldsymbol{z}_{\partial i};\boldsymbol{\Psi}) = -\mathcal{I}_i(\boldsymbol{\Psi}). \tag{3.7}$$

In particular, this is independent of the value of $\boldsymbol{z}_i$, only depending on $\boldsymbol{z}_{\partial i}$. Since the Fisher information matrix is always positive semi-definite, the Hessian of the log-conditional probability is negative semi-definite.

The Q-function for a Gaussian mixture using such an MRF is

$$\sum_{i=1}^{n}\sum_{j=1}^{g} {\tau_{ij}}^{(t)}(\log \phi(y_i;\mu_j,\sigma_j^2) + \log p(\boldsymbol{e}_j|{\boldsymbol{z}_{\partial i}}^{(t)};\boldsymbol{\Psi})) \tag{3.8}$$

Using (3.7), the Hessian of the Q-function is:

$$\nabla^2_{\boldsymbol{\Psi}} \sum_{i=1}^{n}\sum_{j=1}^{g} {\tau_{ij}}^{(t)} \log p(\boldsymbol{e}_j|{\boldsymbol{z}_{\partial i}}^{(t)};\boldsymbol{\Psi})) = -\sum_{i=1}^{n}\sum_{j=1}^{g} {\tau_{ij}}^{(t)}\mathcal{I}_i(\boldsymbol{\Psi}) \quad = -\sum_{i=1}^{n}\mathcal{I}_i(\boldsymbol{\Psi}),$$

where the last line follows since $\sum_{j=1}^{g} {\tau_{ij}}^{(t)} = 1$ and $\nabla^2 \log p(\boldsymbol{e}_j|\boldsymbol{z}_{\partial i})$ does not depend on j. The last line is the sum of negative semi-definite matrices and hence is also negative semi-definite. Thus the Q-function is negative semi-definite. Similarly, the log-pseudolikelihood and log-joint likelihood (using the pseudolikelihood) can be shown to be negative semidefinite. This is independent on whether u_{ij} are calculated using the mean-field or pseudolikelihood approximations.

Since the Potts MRF has $U_i = -\beta u_{ij}$ with $\boldsymbol{\Psi} = \beta$, it is log-linear with respect to β. By the above, the Q-function is concave with respect to β, though this is not necessarily strict.

3.3.1.2 Gradient

Furthermore, the gradient of the Q function (or log-pseudolikelihood) with respect to β is obtainable in closed-form, making it suited to gradient descent algorithms. Here we derive an explicit expression for the gradient of the Q-function with respect to β for a log-linear MRF. Starting from the log conditional probability,

$$\begin{aligned}
\nabla_{\boldsymbol{\Psi}} \log p(\boldsymbol{z}_i|\boldsymbol{z}_{\partial \boldsymbol{i}}; \boldsymbol{\Psi}) &= -\nabla_{\boldsymbol{\Psi}} U_i(\boldsymbol{z}_i|\boldsymbol{z}_{\partial \boldsymbol{i}}; \boldsymbol{\Psi}) - \nabla_{\boldsymbol{\Psi}} \log C_i(\boldsymbol{\Psi}) \\
&= -\nabla_{\boldsymbol{\Psi}} U_i(\boldsymbol{z}_i|\boldsymbol{z}_{\partial \boldsymbol{i}}; \boldsymbol{\Psi}) - \frac{\nabla_{\boldsymbol{\Psi}} C_i(\boldsymbol{\Psi})}{C_i(\boldsymbol{\Psi})} \\
&= -\nabla_{\boldsymbol{\Psi}} U_i(\boldsymbol{z}_i|\boldsymbol{z}_{\partial \boldsymbol{i}}; \boldsymbol{\Psi}) + \sum_{k=1}^{g} \frac{\nabla_{\boldsymbol{\Psi}} U_i(\boldsymbol{e}_k|\boldsymbol{z}_{\partial \boldsymbol{i}}; \boldsymbol{\Psi}) \exp(-U_i(\boldsymbol{e}_k|\boldsymbol{z}_{\partial \boldsymbol{i}}; \boldsymbol{\Psi}))}{C_i(\boldsymbol{\Psi})} \\
&= -\nabla_{\boldsymbol{\Psi}} U_i(\boldsymbol{z}_i|\boldsymbol{z}_{\partial \boldsymbol{i}}; \boldsymbol{\Psi}) + \sum_{k=1}^{g} \nabla_{\boldsymbol{\Psi}} U_i(\boldsymbol{e}_k|\boldsymbol{z}_{\partial \boldsymbol{i}}; \boldsymbol{\Psi}) p(\boldsymbol{e}_k|\boldsymbol{z}_{\partial \boldsymbol{i}}; \boldsymbol{\Psi}).
\end{aligned}$$

The gradient of the Q-function (3.8) is then

$$\begin{aligned}
\sum_{i=1}^{n}\sum_{j=1}^{g} {\tau_{ij}}^{(t)} \nabla_{\boldsymbol{\Psi}} \log p(\boldsymbol{e}_j|\boldsymbol{z}_{\partial \boldsymbol{i}}{}^{(t)}; \boldsymbol{\Psi}) &= \sum_{i=1}^{n}\sum_{j=1}^{g} {\tau_{ij}}^{(t)} \left(-\nabla_{\boldsymbol{\Psi}} U_i(\boldsymbol{e}_j|\boldsymbol{z}_{\partial \boldsymbol{i}}{}^{(t)}) + \sum_{k=1}^{g} \nabla_{\boldsymbol{\Psi}} U_i(\boldsymbol{e}_k|\boldsymbol{z}_{\partial \boldsymbol{i}}{}^{(t)}) p(\boldsymbol{e}_k|\boldsymbol{z}_{\partial \boldsymbol{i}}) \right) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{g} \nabla_{\boldsymbol{\Psi}} U_i(\boldsymbol{e}_j|{u_{ij}}^{(t)}) \left(-{\tau_{ij}}^{(t)} + p(\boldsymbol{e}_j|\boldsymbol{z}_{\partial \boldsymbol{i}}{}^{(t)}) \right),
\end{aligned} \tag{3.9}$$

where the last line follows as the sum with respect to k does not depend on j, and $\sum_{j=1}^{g} {\tau_{ij}}^{(t)} = 1$. For the gradient of the joint log-likelihood, ${\tau_{ij}}^{(t)}$ is replaced by z_{ij}. From this last equation, we see that one way to maximise the Q-function is to equate the prior probability $p(\boldsymbol{Z}_i = \boldsymbol{e}_j|\boldsymbol{z}_{\partial \boldsymbol{i}})$ to the posterior probability $p(\boldsymbol{Z}_i = \boldsymbol{e}_j|y_i)$ for each pixel. In the EM algorithm, this is done for the current segmentation by adjusting the parameters. It is interesting to note that the mean-field equations (3.3) are the same, except that these seek to find $\boldsymbol{z}_i$ for the current parameters.

For the Potts MRF, the gradient of the Q-function is

$$\nabla_{\beta} Q(\boldsymbol{y}, \boldsymbol{z}; \boldsymbol{\Theta}, \beta) = \sum_{i=1}^{n}\sum_{j=1}^{g} {\tau_{ij}}^{(t)} ({u_{ij}}^{(t)} - \sum_{k=1}^{g} {u_{ij}}^{(t)} p(\boldsymbol{e}_k|\boldsymbol{z}_{\partial \boldsymbol{i}}{}^{(t)}; \beta)). \tag{3.10}$$

The only terms that need to be recomputed during the β optimisation are the probabilities $p(\boldsymbol{e}_k|\boldsymbol{z}_{\partial \boldsymbol{i}}; \beta)$, as the neighbour terms u_{ik} are fixed.

3.3.2 Least-squares estimate

The least-squares estimator (LSE) is an alternative method to estimate log-linear MRF parameters. It was first suggested for binary MRFs by Possolo (1986), called the "logit estimator" due to its similarity to use of minimum chi-square estimation in logistic regression (Berkson, 1949).

It was independently put forward for multi-valued MRFs by Derin and Elliott (1987), termed the "histogram estimator". Variants of this estimator and their properties were studied further in e.g. (Gurelli and Onural, 1994; Gurelli, 1996; Borges, 1999). The term "least-squares estimator" arose in the field of brain MRI segmentation when the estimator was used by Van Leemput et al. (1999b).

Suppose the MRF can be written

$$\exp(-U_i(\boldsymbol{e}_j|\boldsymbol{z}_{\partial i};\boldsymbol{\Psi})) = \exp(-\boldsymbol{V}_i(\boldsymbol{e}_j|\boldsymbol{z}_{\partial i})^T\boldsymbol{\Psi}), \tag{3.11}$$

where $\boldsymbol{V}_i$ is a $|\boldsymbol{\Psi}| \times 1$ vector of coefficients for each parameter. For the Potts MRF, $\boldsymbol{V}_i$ is the scalar $-u_{ij}$ and $\boldsymbol{\Psi} = \beta$.

In essence, the estimator aims to minimise the difference between empirical and expected neighbourhood ratios. From Bayes' rule,

$$\frac{p(\boldsymbol{z}_i|\boldsymbol{z}_{\partial i})}{p(\boldsymbol{z}_i, \boldsymbol{z}_{\partial i})} = \frac{1}{p(\boldsymbol{z}_{\partial i})},$$

where the dependency of p on $\boldsymbol{\Psi}$ has been dropped for ease of notation. As the right-hand side does not depend on the value of $\boldsymbol{z}_i$ itself, it can be seen that

$$\frac{p(\boldsymbol{e}_j|\boldsymbol{z}_{\partial i})}{p(\boldsymbol{e}_j, \boldsymbol{z}_{\partial i})} = \frac{p(\boldsymbol{e}_k|\boldsymbol{z}_{\partial i})}{p(\boldsymbol{e}_k, \boldsymbol{z}_{\partial i})}$$

for any labels j and k with the same neighbourhood labels $\boldsymbol{z}_{\partial i}$. Substituting $p(\boldsymbol{z}_i|\boldsymbol{z}_{\partial i})$ in the log-linear form (3.11), rearranging, and taking the logarithm yields

$$\begin{aligned} \frac{\exp(-\boldsymbol{V}_i(\boldsymbol{e}_j|\boldsymbol{z}_{\partial i})^T\boldsymbol{\Psi})}{p(\boldsymbol{e}_j, \boldsymbol{z}_{\partial i})} &= \frac{\exp(-\boldsymbol{V}_i(\boldsymbol{e}_k|\boldsymbol{z}_{\partial i})^T\boldsymbol{\Psi})}{p(\boldsymbol{e}_k, \boldsymbol{z}_{\partial i})} \\ (-\boldsymbol{V}_i(\boldsymbol{e}_j|\boldsymbol{z}_{\partial i}) + \boldsymbol{V}_i(\boldsymbol{e}_k|\boldsymbol{z}_{\partial i}))^T\boldsymbol{\Psi} &= \log\left(\frac{p(\boldsymbol{e}_j, \boldsymbol{z}_{\partial i})}{p(\boldsymbol{e}_k, \boldsymbol{z}_{\partial i})}\right). \end{aligned} \tag{3.12}$$

Each neighbourhood $\boldsymbol{z}_{\partial i}$ and pair of tissues (j, k) yields an equation of the above form. The resulting system is generally overdetermined and may be solved using ordinary least-squares regression, provided the right-hand side may be estimated or evaluated.

The right hand side has terms such as $p(\boldsymbol{e}_j, \boldsymbol{z}_{\partial i})$. This is the probability of seeing a voxel with label j and neighbourhood labels $\boldsymbol{z}_{\partial i}$. In Derin and Elliott (1987), the ratio $p(\boldsymbol{e}_j, \boldsymbol{z}_{\partial i})/p(\boldsymbol{e}_k, \boldsymbol{z}_{\partial i})$ is estimated by $N(j, \boldsymbol{z}_{\partial i})/N(k, \boldsymbol{z}_{\partial i})$, where $N(j, \boldsymbol{z}_{\partial i})$ is the number of times the neighbourhood $\boldsymbol{z}_{\partial i}$ occurs in the image with centre pixel of label j. One consequence of this is that if $N(j, \boldsymbol{z}_{\partial i})$ or $N(k, \boldsymbol{z}_{\partial i}) = 0$, the neighbourhood and centre combination cannot be used in the system of equations. A given set of neighbours $\boldsymbol{z}_{\partial i}$ must appear with at least two different centre labels j and k in order to contribute to the system of equations.

For the Potts MRF, the LSE is determined by the least-squares solution to

$$\beta(u_{ij} - u_{ik}) = \log\left(\frac{N(j, \boldsymbol{z}_{\partial i})}{N(k, \boldsymbol{z}_{\partial i})}\right), \tag{3.13}$$

For each distinct neighbourhood $\boldsymbol{z}_{\partial i}$, up to $\binom{g}{2}$ equations may be added to the system (one per unique (j, k) pair). Overall, one equation per unique $(j, k, \boldsymbol{z}_{\partial i})$ tuple where j and k are labels and $\boldsymbol{z}_{\partial i}$ is a neighbourhood configuration may be added.

Consistency for the LSE has been established for the Ising model (Possolo, 1986; Guyon and Künsch, 1992), but not for the Potts model ($g > 2$). Several variants of the LSE exist, though these have only been defined for Ising models ($g = 2$). Possolo (1986) used weighted least-squares to solve the system instead of ordinary least-squares. For an Ising model, there is only one equation per neighbourhood and the right-hand side is simply $\log(q/(1-q))$ where $q = p(\boldsymbol{e}_1|\boldsymbol{z}_{\partial i})$ is the probability of seeing the neighbourhood with one label as opposed to the other as the centre voxel.

Gurelli and Onural (1994) and Gurelli (1996) constructed an estimate for the right-hand side that minimises mean bias assuming $N(1, \boldsymbol{z}_{\partial i})$ is binomially distributed with parameter q, over a range of possible q. However, this requires computation of an estimate for each $N(1, \boldsymbol{z}_{\partial i})$ and $N(0, \boldsymbol{z}_{\partial i})$ combination, which are stored in a lookup table and used throughout the algorithm. For a large image, this is not feasible. In addition, estimates are still not available when $N(j, \boldsymbol{z}_{\partial i}) = 0$ for some j.

Borges (1999) devised an estimate of $\log(q/(1-q))$ that its minimised mean square error under similar assumptions to the minimum-bias version of Gurelli and Onural. Remarkably, the resulting integrals were analytically solvable, resulting in a closed-form estimate that is valid even when $N(j, \boldsymbol{z}_{\partial i}) = 0$. However this estimate has not been derived when $g > 2$.

A further consideration for the LSE is that there is no simple analogue for a hidden MRF. Following the derivation for the visible MRF,

$$\begin{aligned} f(y_i, \boldsymbol{z}_i, \boldsymbol{z}_{\partial i}) &= f(y_i|\boldsymbol{z}_i, \boldsymbol{z}_{\partial i}) p(\boldsymbol{z}_i|\boldsymbol{z}_{\partial i}) p(\boldsymbol{z}_{\partial i}) \\ &= f(y_i|\boldsymbol{z}_i) p(\boldsymbol{z}_i|\boldsymbol{z}_{\partial i}) p(\boldsymbol{z}_{\partial i}), \end{aligned}$$

where the last line follows as y_i depends on $\boldsymbol{z}_i$ only. From this the following relation may be derived for a given fixed neighbourhood $\boldsymbol{z}_{\partial i}$ with any two centre labels $\boldsymbol{e}_j$ or $\boldsymbol{e}_k$, or any two centre intensities y_i or y_i':

$$\frac{p(\boldsymbol{e}_j|\boldsymbol{z}_{\partial i})}{p(\boldsymbol{e}_k|\boldsymbol{z}_{\partial i})} = \frac{f(y_i, \boldsymbol{e}_j, \boldsymbol{z}_{\partial i}) f(y_i'|\boldsymbol{e}_k)}{f(y_i', \boldsymbol{e}_k, \boldsymbol{z}_{\partial i}) f(y_i|\boldsymbol{e}_j)}.$$

From this the corresponding system of equations, analogous to (3.12), is

$$-U_i(\boldsymbol{e}_j|\boldsymbol{z}_{\partial i}) + U_i(\boldsymbol{e}_k|\boldsymbol{z}_{\partial i}) = \log\left(\frac{f(y_i, \boldsymbol{e}_j, \boldsymbol{z}_{\partial i}) f(y_i'|\boldsymbol{e}_k)}{f(y_i', \boldsymbol{e}_k, \boldsymbol{z}_{\partial i}) f(y_i|\boldsymbol{e}_j)}\right). \tag{3.14}$$

As before, the left-hand-side is linear in the MRF parameters, allowing the system to be solved by least-squares. However, an estimate of $f(y_i, \boldsymbol{z}_i, \boldsymbol{z}_{\partial i})$ must be found. In practice, the intensities y_i are discrete and take on a finite number of allowable values (e.g. 0 to 4095 for 12-bit integers), so a histogram-type method can be performed as earlier. However, given the large number of possible $(y_i, \boldsymbol{z}_i)$ combinations, it is unlikely for these to occur enough times for each neighbourhood $\boldsymbol{z}_{\partial i}$ to provide realistic estimates.

3.3.3 Algorithm

The algorithm is the same as that described in section 2.5, which itself follows the presentation of Celeux et al. (2003). An additional step is added to the M-step to estimate β. Namely, on iteration t:

1. **(C-step)** Form an estimate of the current labels $\boldsymbol{z}^{(t)}$ to be used as neighbours; either discrete (for the pseudolikelihood approximation) or continuous (for the mean-field approximation). The pseudolikelihood version uses the Iterated Conditional Modes (ICM) update

$$\boldsymbol{z}_i{}^{(t+1)} = \boldsymbol{e}_j \text{ where } j = \arg\max_k \exp(\beta u_{ij}{}^{(t,t+1)})\phi(y_i; \mu_k{}^{(t-1)}, \sigma_k^{2(t-1)}),$$

   where ϕ is the pdf of the normal distribution. The mean-field version uses the mean-field update

$$\langle \boldsymbol{z}_i \rangle^{(t+1)} = \sum_{j=1}^{g} \boldsymbol{e}_j \frac{\exp(\beta u_{ij}{}^{(t,t-1)})\phi(y_i; \mu_j{}^{(t-1)}, \sigma_j^{2(t-1)})}{\sum_{k=1}^{g} \exp(\beta u_{ik}{}^{(t,t-1)})\phi(y_i; \mu_k{}^{(t-1)}, \sigma_k^{2(t-1)})}$$

   These updates should be performed sequentially. To save time, we divide the voxels into coding sets (see Appendix B) and update each set simultaneously, visiting them sequentially.

2. **(E-step)** Calculate $\tau_{ij}{}^{(t)}$, using $\boldsymbol{z}^{(t)}$ from the C-step to compute the neighbour term u_{ij}:

$$\tau_{ij}{}^{(t)} = \frac{\exp(\beta^{(t-1)} u_{ij}{}^{(t)})\phi(y_i; \mu_j{}^{(t-1)}, \sigma_j^{2(t-1)})}{\sum_{k=1}^{g} \exp(\beta^{(t-1)} u_{ik}{}^{(t)})\phi(y_i; \mu_k{}^{(t-1)}, \sigma_k^{2(t-1)})}.$$

3. **(M-step)** Maximise Q with respect to $\boldsymbol{\Theta}$ to obtain the intensity parameters.

$$\mu_j{}^{(t)} = \frac{\sum_{i=1}^{n} \tau_{ij}{}^{(t)} y_i}{\sum_{i=1}^{n} \tau_{ij}{}^{(t)}}$$
$$\boldsymbol{\Sigma}_j{}^{(t)} = \frac{\sum_{i=1}^{n} \tau_{ij}{}^{(t)} (y_i - \mu_j{}^{(t)})^2}{\sum_{i=1}^{n} \tau_{ij}{}^{(t)}}.$$

   Then update β using either the MPL or LS estimators.

   For the MPLE, numerically maximise the univariate, concave Q-function with respect to

β, using the gradient (3.10) as necessary:

$$\hat{\beta}_{MPL} = \arg\max_{\beta} \sum_{i=1}^{n} \sum_{j=1}^{g} {\tau_{ij}}^{(t)} (\beta {u_{ij}}^{(t)} - \log(\sum_{k=1}^{g} \exp(\beta {u_{ik}}^{(t)}))).$$

For the LSE, first count how many times each combination of centre and surrounding label configurations $(\boldsymbol{z}_i, \boldsymbol{z}_{\partial \boldsymbol{i}})$ occur in order to obtain estimates for $N(\boldsymbol{z}_i, \boldsymbol{z}_{\partial \boldsymbol{i}})$. One way to do this is to map each neighbourhood to an integer by appending the centre label to that of its neighbours in a fixed, predefined order. This corresponds to a base-g integer of length $|\boldsymbol{z}_{\partial \boldsymbol{i}}| + 1$ where $|\boldsymbol{z}_{\partial \boldsymbol{i}}|$ is the number of neighbours.

For example, if $g = 3$ and a voxel has label '2' and its north, east, south, west, top, and bottom neighbours are '1', '1', '2', '3', '3', '2', this can be represented as the base-3 integer '1001221', decreasing each label by 1 to ensure that every integer from 0 to $g^{|\boldsymbol{z}_{\partial \boldsymbol{i}}|+1} - 1$ represents a (centre, neighbourhood) combination and vice-versa. These integers may be used as indexes into a frequency table.

Once the frequency table is computed, it is used to construct the right-hand side of the system of equations (3.13):

$$\beta(u_{ij} - u_{ik}) = \log\left(\frac{N(j, \boldsymbol{z}_{\partial \boldsymbol{i}})}{N(k, \boldsymbol{z}_{\partial \boldsymbol{i}})}\right),$$

where there is one equation for each unique $(j, k, \boldsymbol{z}_{\partial \boldsymbol{i}})$, and discarding all equations for which $N(j, \boldsymbol{z}_{\partial \boldsymbol{i}})$ or $N(k, \boldsymbol{z}_{\partial \boldsymbol{i}})$ is zero. Finally, the above system is solved for β using least-squares.

These steps are repeated until the relative change in approximate observed log-likelihood falls below a pre-specified tolerance (1e-5 in these experiments), or it decreases. The approximate observed-data log-likelihood is given by

$$\log f(\boldsymbol{y}) \approx \sum_{i=1}^{n} \log\left(\sum_{j=1}^{g} \phi(y_i | \boldsymbol{e}_j; {\mu_j}^{(t)}, \sigma_j^{2(t)}) p(\boldsymbol{e}_j | {\boldsymbol{z}_{\partial \boldsymbol{i}}}^{(t)}; \boldsymbol{\Psi}^{(t)})\right).$$

EM on a standard mixture model guarantees an increase in the observed log-likelihood and the Q function. Gao and Song (2011) proved the ascent property holds when the pseudolikelihood is a product of *marginal* likelihoods, but not for conditional likelihoods. However, as far as we are aware, there is no analogous result for pseudolikelihoods that are products of conditionals; thus, it is possible that the Q-function and observed log-likelihood may decrease.

We initialise the algorithm by fitting a standard normal mixture model with 3 components to the image (i.e., without the MRF). This yields an initial segmentation to be used as the neighbours, as well as means and standard deviations. Where β is to be estimated by LS or MPL, it is derived from this initial segmentation.

3.4 Experiments

We perform a number of experiments to investigate the value of estimation of β by maximum pseudolikelihood. We list our aims, hypotheses and corresponding experiments briefly here, and explain them in further detail in the corresponding sections. In general, experiments are performed by segmenting a dataset of real brain MRI with the EM algorithm, with or without β estimation.

Optimal configuration for the MPLE
Aim: determine the best choice (in terms of segmentation accuracy) of MRF approximation and neighbourhood size to be used with the MPLE.
Experiment: We compare segmentations obtained using MPL estimation with all combinations of MRF approximation (mean-field or pseudolikelihood) and various neighbourhood sizes (6, 18 and 26 neighbours).
Hypothesis: The mean-field approximation will allow greater sensitivity to voxels that are not strongly allocated into one particular class, resulting in higher segmentation accuracy. Similarly, larger neighbourhoods should increase the accuracy of the estimation due to more local information being available.

MRF estimation
Aims:

- compare estimation of β to fixing it to commonly-used default values.
- compare the MPLE to the LSE.

Experiment: We compare segmentations with the same configuration as Atropos, NiftySeg, and FAST (which use fixed β) to segmentations using the MPLE and LSE where the LSE is comparable to EMS.
Hypothesis: We expect estimation of β to yield more accurate results to fixed-β methods in general. MPLE segmentations may be more accurate than the LS estimations as they can take into account the fact that the MRF is hidden, while the LSE cannot.

Grid search
Aims:

- Study the dependence of segmentation accuracy on β.
- Determine whether MPLE can recover the (or a) fixed-β value that produces the highest segmentation accuracy.

Experiment: We select some example subjects and perform a grid search over a wide range β values, and compare these to the MPLE segmentations.
Hypothesis: The grid search, by its nature, should yield some segmentations with the same or higher accuracy than MPLE. However, we expect the corresponding fixed-β values to differ by image, and that the MPLE is able to estimate β near these values (or achieve a similar

segmentation accuracy). We expect segmentation accuracy to decrease drastically as β increases due to oversmoothing.

The experiments were evaluated on images from the Internet Brain Segmentation Repository (IBSR) (Rohlfing, 2012).[1] The dataset used consists of T1-weighted coronal MR volumes of 18 normal subjects of ages 7 to 71. Each volume consists of 128 coronal slices spaced at 1.5mm with in-plane resolution varying from 0.84×0.84mm to 1.00×1.00mm. This dataset also contains manual segmentations to compare the automatic segmentations to. The images are already skull-stripped with bias-correction already performed, so no further preprocessing was done. That is, all non-brain voxels (such as skull, fat) are already removed from the image as we wish to concentrate on segmentation of the brain only.

The most relevant way to evaluate performance of the various segmentation algorithms is to measure the accuracy of the resulting MR segmentation against some reference or 'ground truth' image. Maximisation of the observed-data log-likelihood and accuracy of the β estimates are not relevant to the application; in any case, the former cannot be evaluated exactly, and the 'true' parameter estimates for the latter cannot be not known. For brain segmentation, the ground truth image would consist of the underlying true physical boundaries of the tissues in the brain, discretised by the image grid. However, the true underlying tissue boundaries cannot be found non-destructively and without moving the brain from the position in which it was imaged. Rather, the 'ground truth' will generally take the form of the MR image manually segmented by an expert.

By its nature, manual segmentation can be subjective, depending on the experience of the expert as well as the protocols used to segment the brain. For example, imagine a thin pipe-like feature that is thinner than one voxel. Thus, a given voxel may contain the entire pipe-like feature and also some surrounding tissue of a different type. The nature of a hard segmentation is that each voxel may be counted as only one tissue type. One protocol may stipulate that such a voxel should be classified as the surrounding tissue (being the majority), while another may stipulate that it should be classified as the tissue of the pipe-like feature. Thus though we use expert manual segmentations as 'ground truths' from which to evaluate accuracy, it is important to remember that even the expert segmentations may differ slightly between experts.

As reference segmentations are hard (each voxel has exactly one tissue), any segmentations we wish to evaluate must also be hard. EM procedures produce a soft segmentation, giving the posterior probability of each voxel to be each particular tissue. We convert these to hard segmentations by assigning voxel to the tissue with the largest posterior probability.

Performance against the manual segmentations will be evaluated by two metrics, *segmentation accuracy* and *Dice similarity*. Let A and B represent two segmentations, being sets of indices for each tissue. That is, $A_j, j = 1, \ldots, g$ are non-intersecting subsets of the indices $1, \ldots, n$

[1] The MR brain data sets and their manual segmentations were provided by the Center for Morphometric Analysis at the Massachusetts General Hospital and are available at http://www.cma.mgh.harvard.edu/ibsr/.

whose union is the entire brain, where $i \in A_j$ implies that voxel i is assigned to tissue j in segmentation A.

The *segmentation accuracy* is quantified as the overall percentage of voxels correctly classified. Since the reference and test segmentations have the same number of voxels (all of the brain voxels), this is well-defined.

$$\text{accuracy}(A, B) = \frac{|A \cap B|}{|A|}.$$

The *Dice similarity coefficient* (commonly called 'Dice score' or 'Dice index') (Dice, 1945) is used to compare segmentations on a tissue-by-tissue basis. The Dice coefficient for a given tissue between two segmentations A and B is given by the number of correctly-classified voxels divided by the average area classified (of that tissue):

$$\text{Dice}(A_j, B_j) = \frac{2|A_j \cap B_j|}{|A_j| + |B_j|}.$$

It ranges from 0 to 1, with 1 meaning a perfect match between the two segmentations of that tissue. The reason for using Dice coefficient for each tissue rather than accuracy is that the number of voxels classified as a particular tissue may not be equal between the two segmentations, whereas the number of overall voxels in the brain (used for the accuracy) is.

3.4.1 Choice of approximation and neighbourhood size

The aim of this experiment is to determine the optimum configuration (in terms of segmentation accuracy) for the MPL estimation. In particular, we choose between different MRF approximations and different neighbourhood sizes.

Existing algorithms use various combinations of MRF approximation and neighbourhood size. Atropos and FAST use the pseudolikelihood approximation, while NiftySeg uses the mean-field approximation. The mean-field approximation retains the probability that a given voxel belongs to any class, while the pseudolikelihood does not due to the need to threshold these probabilities. This allows voxel states to propagate further, and may help for voxels that are not decidedly in one tissue class or the other (e.g. with intermediate intensity). For this reason we hypothesise that the mean-field approximation to outperform the pseudolikelihood approximation.

We consider different types of neighbourhood in a $3 \times 3 \times 3$-voxel vicinity about voxel i. We compare neighbourhoods with the 6 orthogonal neighbours, 18 in-plane neighbours, and full 26 neighbours figure 3.4. Atropos uses all 26 neighbours of the 3x3x3 cube, while FAST and NiftySeg use 6. Use of all 26 neighbours may yield more accurate results as it permits more local information to be used in distinguishing between different neighbourhoods. However, use of 6 neighbours is far more common in the literature and in practice (e.g. Zhang et al. (2001); Van Leemput et al. (1999a)). It is likely that this for computational efficiency in calculating the

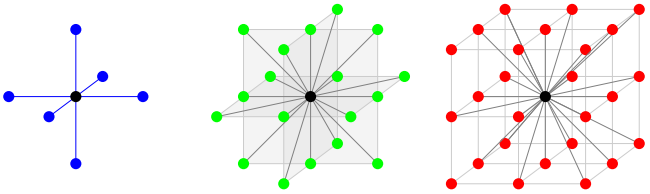


Figure 3.4: Neighbourhoods in a 3x3x3 cube with 6, 18 and 26 neighbours. Each point represents a voxel while lines connecting them indicate these two voxels are neighbours.

neighbourhood statistics of each voxel.

We test all combinations of MRF approximation with all combinations of neighbourhood size; these are shown in table 3.1. The reason every combination of neighbourhood size and MRF approximation is tested together rather than independently is to examine if they interact. Based on the results of this experiment, we select a neighbourhood size and MRF approximation to use with MPL for the remainder of the experiments.

Table 3.1: Experiment summary: MRF neighbourhood size and approximation configurations using MPLE

neighbourhood size	MRF approximation
6	MF
6	PL
18	MF
18	PL
26	MF
26	PL

3.4.2 MRF estimation

Now we compare estimation of β using maximum pseudolikelihood (MPL) and using least-squares (LS) to fixing it to commonly-used default values. Maximum pseudolikelihood estimation has not previously been used in this context. Least-squares estimation is included as it is the method as presented by Van Leemput et al. (1999b) for brain segmentation (though that paper used a more general form of the Potts MRF). The common fixed β values we compare to are 1 as used by FAST (Zhang et al., 2001), 0.25 as used by NiftySeg's `seg_EM` program (Cardoso et al., 2009), and 0.3 as used by Atropos (Avants et al., 2011). We use the neighbourhood size of 6 and the pseudolikelihood approximation for MPL as determined by the results of the previous experiment. For the least-squares approximation, we match the EMS settings as specified in Van Leemput et al. (1999a), using the pseudolikelihood approximation with a neighbourhood size of 6. As well as comparing estimation of β to fixed β, we are interested in comparing the least-squares estimator to the maximum pseudolikelihood estimator. For the MPL estimator we used the PL approximation and 6 neighbours as determined by the previous experiment.

We note that FAST, NiftySeg and Atropos all have the capability to incorporate an atlas prior and tend to be used with one, hence it could be that their default β values were chosen with this in mind. However, there is no way to tell if this is the case and if so, how to adjust the defaults for the case of no atlas. Hence, we use the default values as-is.

Given the vast range of fixed β values used as defaults, we expect that the corresponding segmentations will have differences in accuracy across the dataset. The algorithms that estimate β should be able to adjust to the various images, resulting in higher segmentation accuracy.

Table 3.2: Experiment summary: comparison of various mixture-MRF algorithms.

method	neighbourhood size	MRF approximation	β
Atropos	26	PL	0.3
FAST	6	PL	1
NiftySeg	6	MF	0.25
LS	6	PL	estimated
MPL	6	PL	estimated

3.4.3 Grid search

To further examine how choice of β affects the segmentation, we pick 2 random subjects from the dataset and perform a grid search over fixed β values. From this, the sensitivity of segmentation accuracy with respect to β may be studied. It is anticipated that the response of accuracy to β will be most variable for small β variables, after which the accuracy will degrade due to oversmoothing.

We also wish to see whether maximum pseudolikelihood estimation can recover the most accurate value(s) of β, or whether it can achieve a similar accuracy to the highest obtained in the grid search.

All subjects were segmented with the PL approximation and 6 neighbours (using the same configuration as MPL from the previous experiment for comparability), and fixed β values ranging from 0.00 to 100.00 at various increments (see table 3.3).

Table 3.3: Fixed β values used for grid search

range (inclusive)	increment
0.00 to 1.00	0.05
1.00 to 2.50	0.10
2.50 to 5.00	0.25
5.00 to 10.00	1.00
10.00 to 100.00	10.00

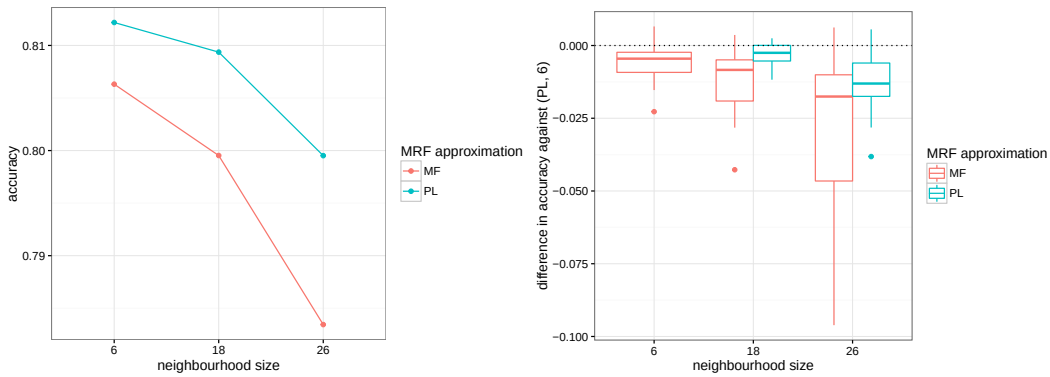


Figure 3.5: Mean accuracy for different MRF approximations under different neighbourhood sizes using MPLE (left); paired differences by subjects against pseudolikelihood with neighbourhood size 6 (right). Neighbourhood size (X) axis is categorical, not numeric.

3.5 Results

3.5.1 Choice of approximation and neighbourhood size

Table 3.4: Comparison of different MRF approximation and neighbourhood size combinations using MPL, ordered by accuracy decreasing. Estimated $\hat{\beta}$, accuracy and Dice similarities are means over all subjects.

approximation	neighbourhood size	$\hat{\beta}$	accuracy	Dice (CSF)	Dice (GM)	Dice (WM)
PL	6	1.88	**0.812**	**0.625**	**0.843**	0.820
PL	18	0.68	0.809	0.614	0.839	0.827
MF	6	1.93	0.806	0.609	0.836	0.825
MF	18	0.68	0.800	0.592	0.829	0.829
PL	26	0.49	0.800	0.581	0.826	**0.841**
MF	26	0.48	0.783	0.554	0.810	0.841

Table 3.5: ANOVA of mixed-effects model of accuracy against neighbourhood size (categorical) and MRF approximation), controlling for subject random effects.

	Sum Sq	Mean Sq	NumDF	DenDF	F	Pr(>F)
neighbourhood size	0.006	0.003	2	87.0	29.410	<0.001*
MRF approximation	0.003	0.003	1	87.0	29.282	<0.001*

Table 3.4 shows the mean segmentation accuracy for various combinations of MRF approximation and neighbourhood size. These are also shown in figure 3.5 (left) to assess the overall trends. However, the data comprises repeated measures as each subject is segmented with all combinations of MRF approximation and neighbourhood size. Hence in figure 3.5 (right) we show the difference in accuracy relative to a baseline of the pseudolikelihood approximation with a neighbourhood size of 6.

From figure 3.5 it can be seen that the mean-field approximation performs worse in terms of accuracy compared the pseudolikelihood approximation for all neighbourhood sizes. It can also

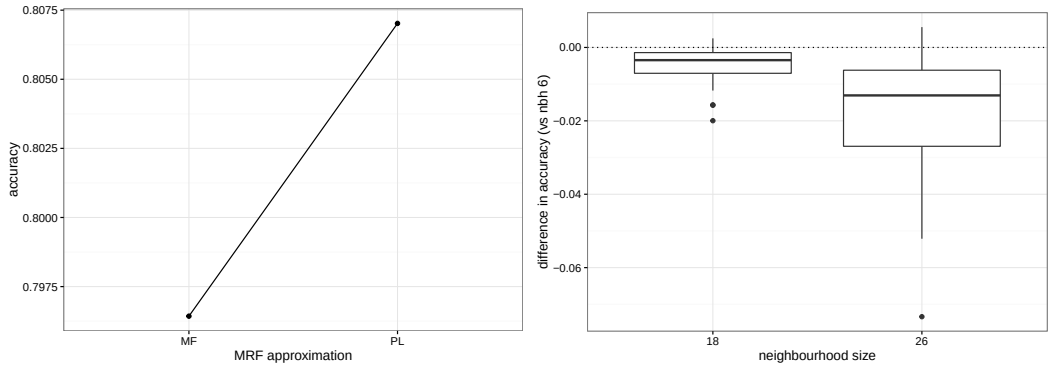


Figure 3.6: Main effects plot - average accuracy for different neighbourhood sizes, and paired difference in accuracy relative to size 6.

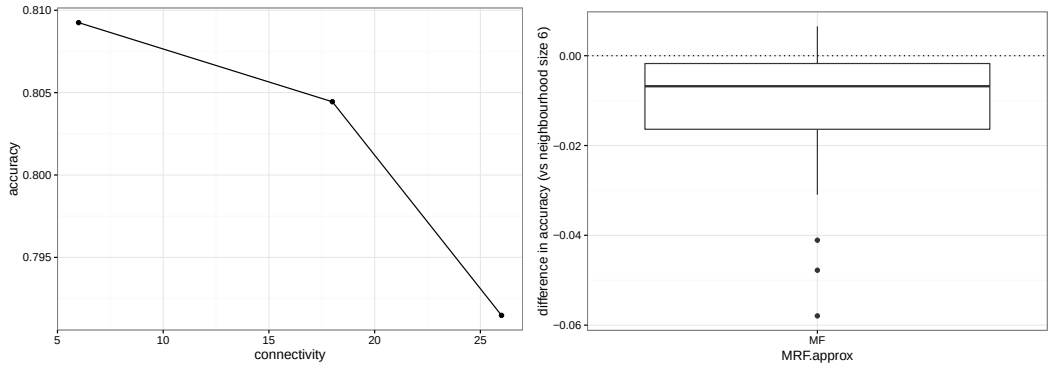


Figure 3.7: Main effects plot - average accuracy for MF and PL approximations, and paired difference in accuracy relative to the PL approximation.

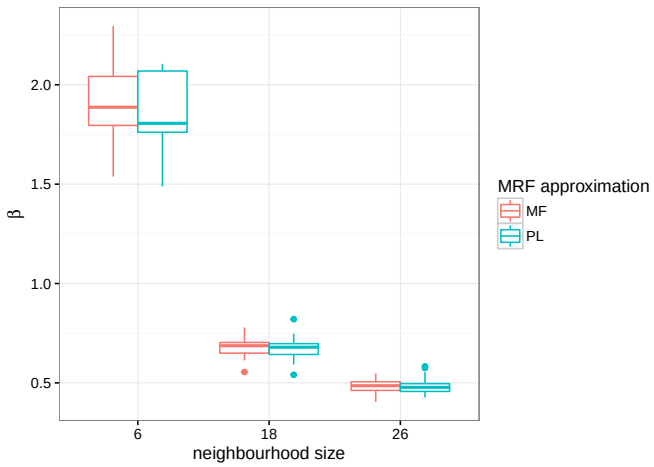


Figure 3.8: Estimated β values for various configurations. Neighbourhood size (X) axis is categorical, not numeric.

Table 3.6: Post-hoc pairwise comparisons of accuracy for different neighbourhood sizes combinations using Tukey's method.

Comparison	Estimate	p
6 - 18	0.005	0.117
6 - 26	0.018	<0.001*
18 - 26	0.013	<0.001*

Table 3.7: Post-hoc pairwise comparison of accuracy for different MRF approximations using Tukey's method.

Comparison	Estimate	p
MF - PL	-0.011	<0.001*

be seen that in general, the accuracy decreases as the neighbourhood size increases. There appears to be no interaction between the two. A mixed-effects model was fitted to test for an effect of neighbourhood size (as a categorical variable, not numeric) and MRF approximation on accuracy, with a random intercept permitted for each subject. No interaction term was included. The model is shown in table 3.5. Segmentation accuracy was found to differ significantly depending on neighbourhood size and MRF approximation ($p < 0.05$); the fitted effects and per-subject paired differences can be seen in figures 3.6, 3.7.

Post-hoc pairwise comparisons were performed between the various neighbourhood sizes and MRF approximations independently, using Tukey's method to adjust for multiple comparisons (tables 3.6, 3.7). Estimated marginal means of the fixed effects are shown (i.e. predictions are made over a grid of the other covariate and averaged to marginalise them). The mean-field approximation performed significantly worse than the pseudolikelihood approximation, and 26 neighbours performed significantly worse than 6 or 18 neighbours. Thus, we use the pseudolikelihood approximation with 6 neighbours for MPL for the remainder of the experiments. We chose 6 neighbours instead of 18 for computational efficiency and because it achieved higher average accuracy than 18 neighbours (though not significant).

In terms of β values, from figure 3.8 it can be seen that the choice of approximation makes very little difference to the estimated values, both in terms of spread and location. However, it is clear that a higher number of neighbours is associated with a lower estimated β value. We will investigate this relationship further in section 3.6.1.2.

3.5.2 MRF estimation

Table 3.8 and figures 3.9, 3.10 show the average accuracy and Dice scores achieved across the dataset by the various algorithms. Estimation of the MRF yielded higher accuracy than any of the fixed-β comparisons, though we acknowledge that as NiftySeg, FAST and Atropos are

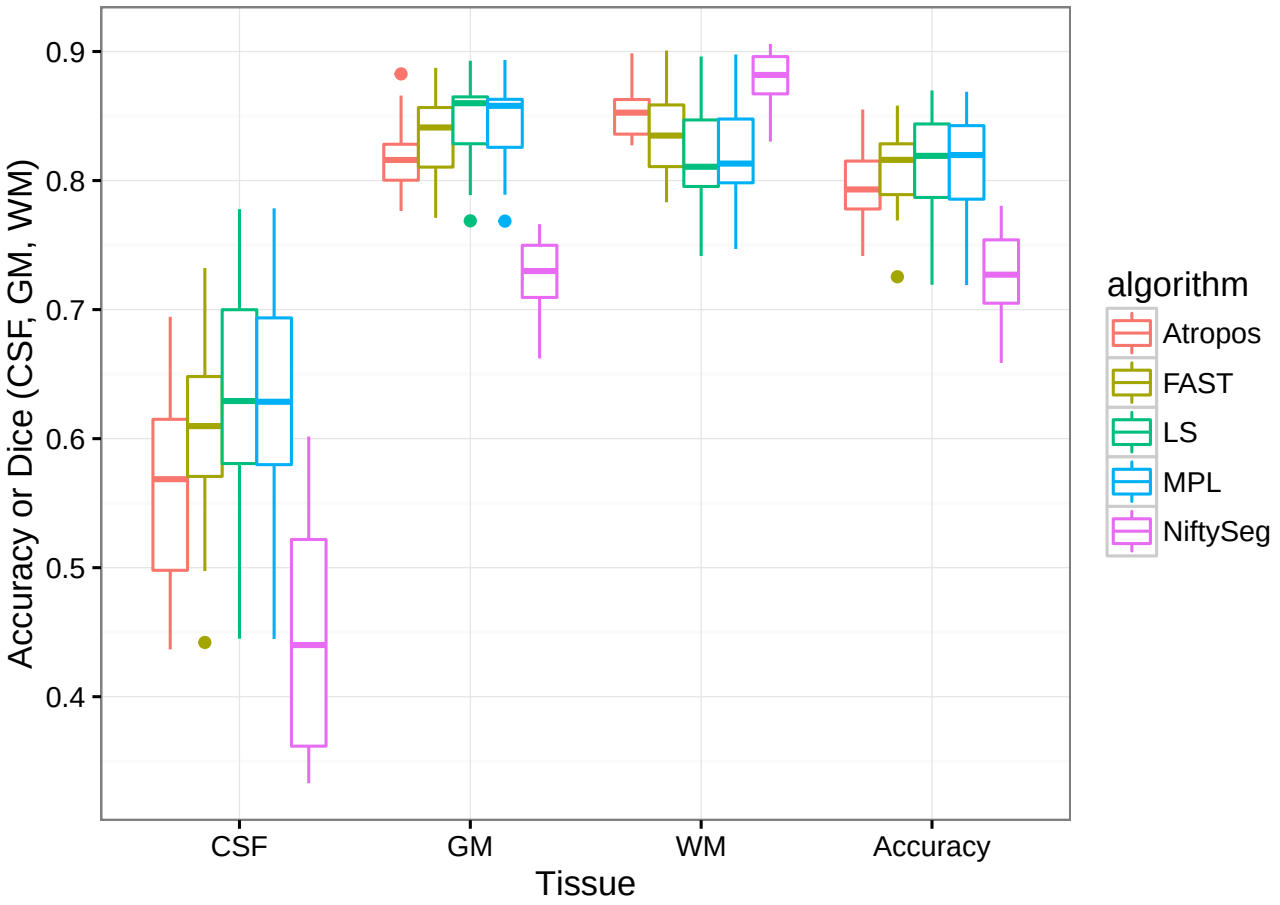


Figure 3.9: Segmentation metrics (accuracy or Dice coefficient) for the various methods.

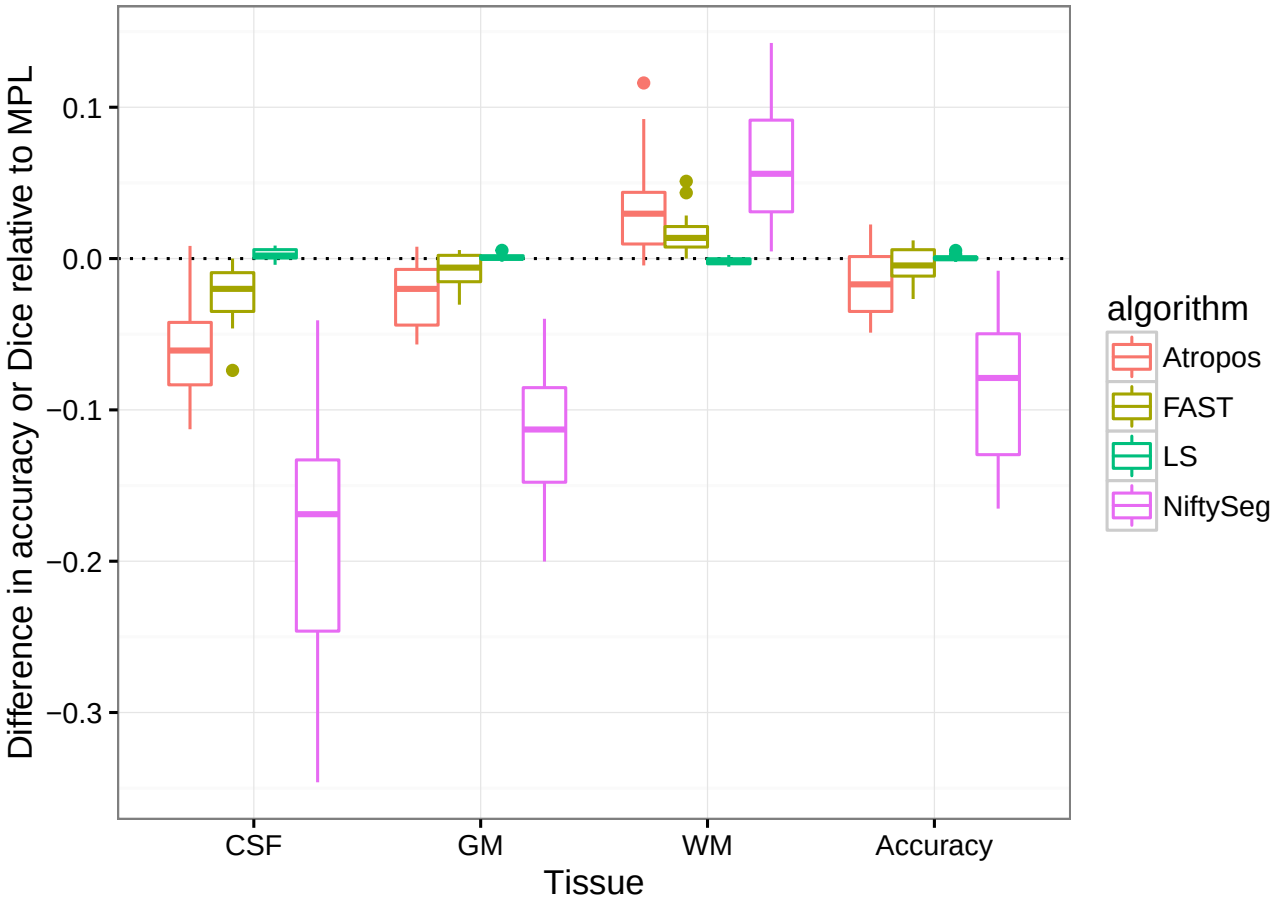


Figure 3.10: Paired differences in accuracy/Dice, relative to MPL.

Table 3.8: Average performance for different algorithms, ordered by accuracy decreasing.

algorithm	neighbourhood size	β	approximation	accuracy	Dice (CSF)	Dice (GM)	Dice (WM)
LS	6	2.14	PL	**0.813**	**0.628**	**0.843**	0.818
MPL	6	1.88	PL	0.812	0.625	0.843	0.820
FAST	6	1.00	PL	0.808	0.602	0.835	0.837
Atropos	26	0.30	PL	0.796	0.566	0.819	0.854
NiftySeg	6	0.25	MF	0.727	0.446	0.727	**0.878**

Table 3.9: Mixed-effects model of segmentation accuracy by algorithm, controlling for subject blocking.

	Sum Sq	Mean Sq	NumDF	DenDF	F	Pr(>F)
algorithm	0.097	0.024	4	68.0	48.670	<0.001*

usually used with atlas priors, their default fixed β values may not have been selected for the case of no atlas. In terms of per-tissue Dice score, it seems that gains in Dice coefficient for WM are generally offset by losses in Dice for GM and CSF and vice-versa. All methods had difficulty accurately segmenting CSF compared to other tissues. All algorithms achieved relatively similar overall accuracy except for NiftySeg, which clearly had the lowest accuracy. A mixed-effects model was fit to accuracy against algorithm controlling for repeated subjects and is shown in table 3.9. Post-hoc comparisons with Tukey's method (table 3.10) showed that NiftySeg was significantly worse than all other methods; all other methods were not significantly different to each other in accuracy.

Comparing both estimated-β methods, least-squares had slightly higher average accuracy than MPL estimation though this was not significant. Both methods had very similar Dice coefficients and overall accuracy. Both algorithms yielded estimated β values larger than one (figure 3.11). The least-squares estimates were more widely spread over the dataset.

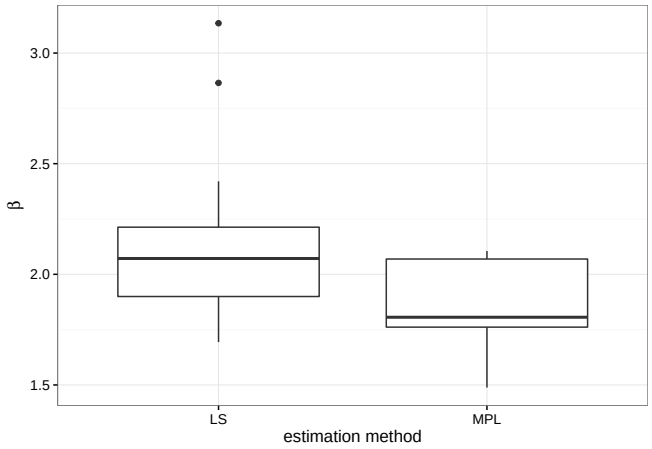


Figure 3.11: Range of estimated β values

Table 3.10: Post-hoc pairwise comparisons for differences in accuracy for different algorithms using Tukey's method. Only significant differences are shown.

Comparison	Estimate	p
Atropos - NiftySeg	0.069	<0.001*
FAST - NiftySeg	0.081	<0.001*
LS - NiftySeg	0.086	<0.001*
MPL - NiftySeg	0.085	<0.001*

Table 3.11: β values for various algorithms, ordered by accuracy decreasing.

IBSR_10			IBSR_17		
algorithm	β	accuracy	algorithm	β	accuracy
grid-search	2.10	0.847	grid-search	0.75	0.795
LS	2.02	0.846	FAST	1.00	0.793
MPL	1.80	0.845	MPL	2.06	0.781
FAST	1.00	0.832	LS	2.18	0.781
Atropos	0.30	0.817	Atropos	0.30	0.777
NiftySeg	0.25	0.767	NiftySeg	0.25	0.701

3.5.3 Grid search

Figure 3.12 shows the accuracy obtained for segmentations performed with fixed β values as defined in table 3.3, for 2 randomly-selected subjects. For small β, the accuracy appears to be lowest when $\beta = 0$, gradually increasing. Maximum accuracy is attained before $\beta = 10$, after which it appears to plateau, even as β becomes very large. Table 3.11 shows the fixed β at which the maximum segmentation accuracy was attained (if there were multiple, it is the smallest such). As we will see later, it is possible to reach a point where changing increasing β further cannot change the label assignation any more; this occurs when every voxel matches the majority label in its neighbourhood.

Figure 3.12 and table 3.11 also show the β and segmentation accuracy of other algorithms of interest. The reason the Atropos and NiftySeg points do not lie on the corresponding fixed-β line is because they use a different neighbourhood size (Atropos has 26 neighbours) or MRF approximation (NiftySeg uses MF) to that used to perform the grid search (6 neighbours, PL approximation). The reason MPL and LS do not have the same accuracy as their corresponding fixed-β counterparts is due to the fact that for MPL and LS, β is not held fixed; it can change on each iteration. This produces a resulting change in segmentation accuracy. As expected, the highest accuracy was obtained by the grid-search. However, the two estimated-β methods (MPL and LS) had close to this accuracy. FAST ($\beta = 1$) was the closest to the maximum accuracy within the fixed-β methods, and also had β closest to the grid-search maximum β of the fixed-β methods.

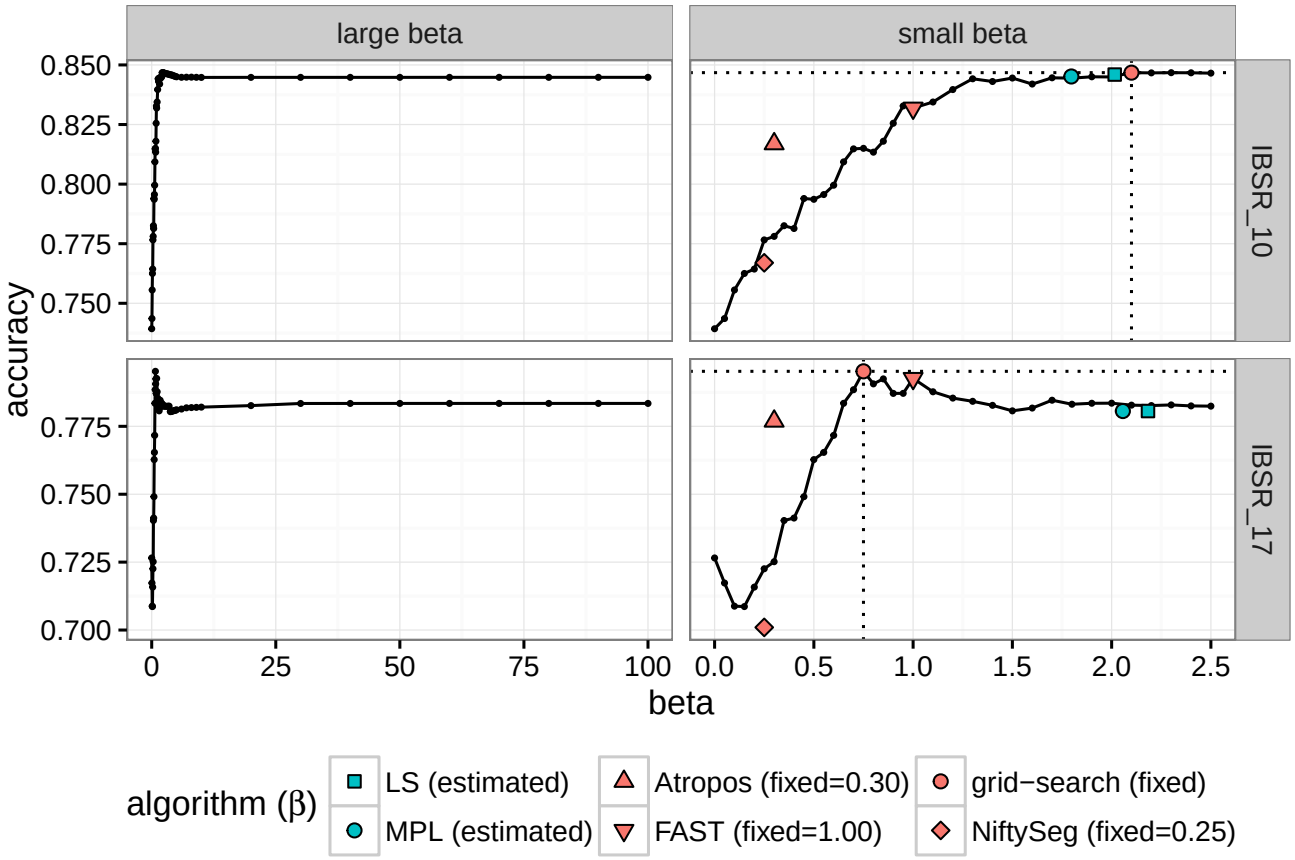


Figure 3.12: Accuracy for various fixed beta values (X axis) and subjects (one per row). Different β limits (columns) are shown to emphasise various behaviours. Several fixed-β algorithms are shown for comparison.

3.6 Discussion

3.6.1 Choice of approximation and neighbourhood size

As the effects of MRF approximation and neighbourhood size were found to be independent of each other, we will address them separately.

3.6.1.1 MRF approximation

It was expected that the mean-field approximation should produce higher accuracy than the pseudolikelihood approximation. This is because the mean-field approximation retains more information about the neighbour states than the pseudolikelihood approximation - that is, it retains posterior label probabilities while the pseudolikelihood approximation thresholds them. The mean-field approximation also has the property of being optimal in the sense of minimising Kullback-Leibler divergence to the true MRF pdf in the class of factorisable functions, which includes pseudolikelihood. It also fits with the ethos of Expectation-Maximisation, which is to use expected values of the class labels rather than actual realisations. Thus, it is surprising to notice that the pseudolikelihood segmentation had significantly higher overall accuracy than the corresponding mean-field segmentation. It is unclear why this is the case.

3.6.1.2 Neighbourhood size

It was expected that additional neighbours would help to distinguish different neighbourhoods that be identical with fewer neighbours, leading to greater sensitivity of the MRF and thus higher accuracy. For example, in figure 3.13 two different image features are shown: one is a flat surface as for an interface between two tissues, while the other has one isolated voxel of a different tissue label to the rest. The 6-neighbour MRF cannot distinguish between these features, but they should be treated differently - the flat surface is more likely a feature to preserve, while the isolated voxel is more likely a feature to smooth. In 26-connectivity, these surfaces would be distinguished from each other. Instead, the accuracy decreases as the neighbourhood size increases, which was unexpected.

Figure 3.14 shows a mid-slice of the PL segmentation for subject IBSR_18 with different neighbourhood sizes. As the neighbourhood size increases, it appears that there is more CSF and WM and less GM in the resulting segmentations. Some small isolated areas of WM in the 6-neighbour segmentation have merged to become smoother in the 26-neighbour segmentation at the expense of grey matter which seems eroded. It is unclear why this is, except that the different numbers of neighbours have shifted the tissue means such that the tissue boundaries have changed.

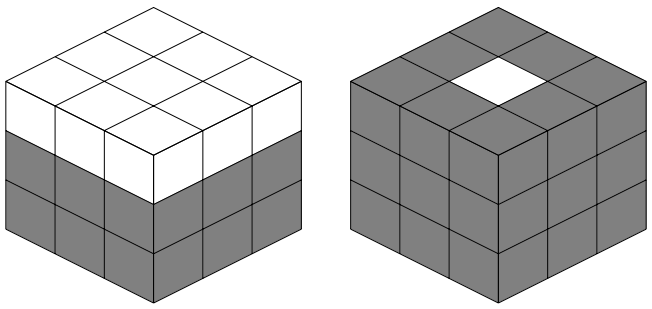
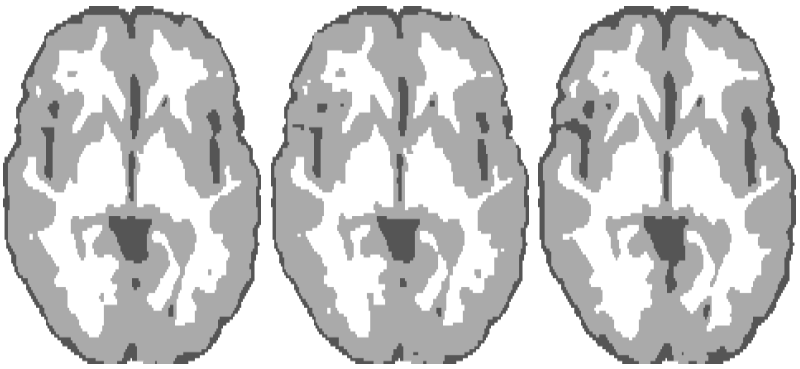


Figure 3.13: Two different 3x3x3 neighbourhoods that the 6-neighbourhood MRF cannot distinguish between.



(a) 6 neighbours (b) 18 neighbours (c) 26 neighbours

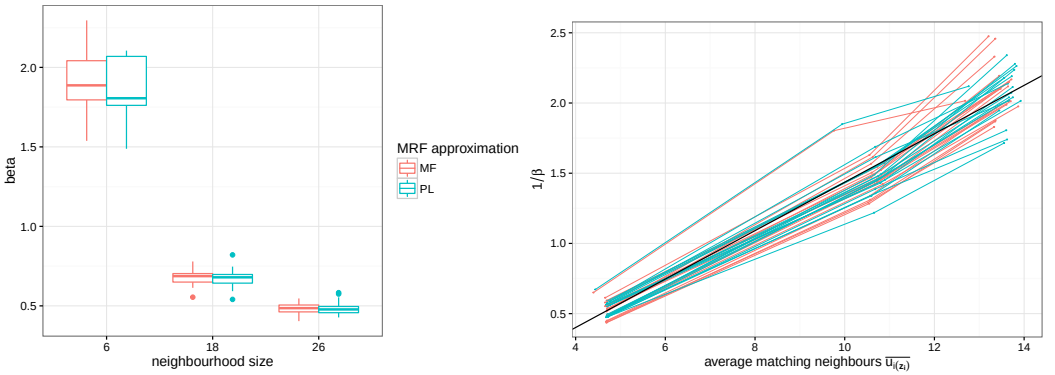
Figure 3.14: Mid-brain slice of segmentations of subject IBSR_18 with the PL approximation.

The estimated β for each neighbourhood size decreases as the number of neighbours increases. It is also much more variable for smaller neighbourhoods. Let $u_{i(\boldsymbol{z}_i)}$ indicate the number of voxels in voxel i's neighbourhood that match its label. The vector of $u_{i(\boldsymbol{z}_i)}$ for all i is a sufficient statistic for the tissue labels $\boldsymbol{z}_i$ for the Potts MRF. As the neighbourhood size increases, so does the $u_{i(\boldsymbol{z}_i)}$, due to there being more voxels in the neighbourhood. In order to maintain a similar influence of the MRF relative to the intensity pdf, the MRF exponent $\beta u_{i(\boldsymbol{z}_i)}$ should remain roughly constant across neighbourhood sizes. Thus, it is unsurprising that as the neighbourhood size increases, β decreases in response.

Table 3.12: Average number of matching-label neighbours in MPL segmentations, relative to the weighted neighbourhood size.

neighbourhood size (weighted)	MRF approximation	average $u_{i(\boldsymbol{z}_i)}$
6 (5.2)	MF	4.7 (89%)
6 (5.2)	PL	4.7 (90%)
18 (12.3)	MF	10.6 (86%)
18 (12.3)	PL	10.7 (87%)
26 (16.0)	MF	13.5 (84%)
26 (16.0)	PL	13.6 (85%)

We compute the mean $u_{i(\boldsymbol{z}_i)}$ for each image, denoted $\overline{u_{i(\boldsymbol{z}_i)}}$, and compare this to the weighted neighbourhood size in table 3.12. The weighted neighbourhood size is the number of neighbours



(a) Estimated β values with MPL.

(b) $1/\beta$ against the average number of matching-label neighbours. The population fit of the mixed-effects model in table 3.13 is shown.

Figure 3.15: Estimated and fitted β values with MPL.

weighted by inverse distance, i.e. $\sum_{m \in \partial \boldsymbol{i}} \frac{1}{\delta_{im}}$; this is the same as the number of matching neighbours if the entire neighbourhood were of one label. The average number of matching neighbours are all quite high relative to the weighted neighbourhood size, reflecting that an MRI is largely homogeneous in nature. As the neighbourhood size increases, the average number of matching neighbours decreases slightly since there are more possible neighbourhoods, reducing the likelihood of the all-homogeneous neighbourhood.

Table 3.13: Mixed linear regression: $1/\beta$ against average number of matching neighbours and MRF approximation with a fixed intercept and per-subject random intercepts. The marginal R^2 is 0.95. The estimate for "MRF approximation" is additive for the PL approximation.

	Coefficient (95% CI)	Test statistic	df	p-value
$\overline{u_{i(\boldsymbol{z}_i)}}$	0.17 (0.2, 0.18)	F = 3739.7	1, 88.1	<0.01*
MRF approximation	-0.01 (-0.1, 0.03)	F = 0.4	1, 88.0	0.51
Intercept	-0.29 (-0.4, -0.21)	T = -7.3	61.8	<0.01*

Figure 3.15 shows $1/\beta$ against the $\overline{u_{i(\boldsymbol{z}_i)}}$ for each image. This shows an approximately linear relationship with no apparent difference between MRF approximations. A linear regression of $1/\beta$ against $u_{i(\boldsymbol{z}_i)}$ controlling for MRF approximation and random per-subject intercepts yielded a linear relationship with no difference between MRF approximations (table 3.13). The marginal R^2 as defined by Nakagawa and Schielzeth (2013) is used to assess model fit, being the proportion of the total variance explained by the fixed effects only. This yields a very high $R^2 = 0.95$, supporting the strength of the linearity. However, the global intercept was non-zero, so it cannot be said that $\beta u_{i(\boldsymbol{z}_i)}$ remains constant across neighbourhood size.

The E-step described to fit the mixture-MRF model is approximate. It does not perform a full marginalisation of the latent variables $\boldsymbol{z}$ due to the large number of possible states, and the intractability of calculating the MRF normalising constant for each of them. Instead, the E-step makes use of a realisation $\boldsymbol{z}^{(t)}$ and computes the expectation of each voxel individually, using the realisation as neighbours. Thus, the current segmentation may have undue influence in the

algorithm; the result is biased due to the need for it in the E-step. More neighbours could mean that the segmentation (i.e.the MRF) is biases the results even more.

Another possible cause for this unexpected result could lie with the use of coding sets to implement the MRF update. Ideally a fully-asynchronous update should be used (e.g. visiting voxels in a random and different order every iteration). Coding sets are commonly used to greatly improve computational efficiency. It could be that the effect of using a coding scheme to update is increased with larger neighbourhoods.

In summary, it appears that adding extra neighbours is negatively associated with segmentation accuracy for the MPL estimator. This turns out to be beneficial, as computing neighbour label counts u_{ij} with only 6 neighbours is more efficient than with 26 neighbours.

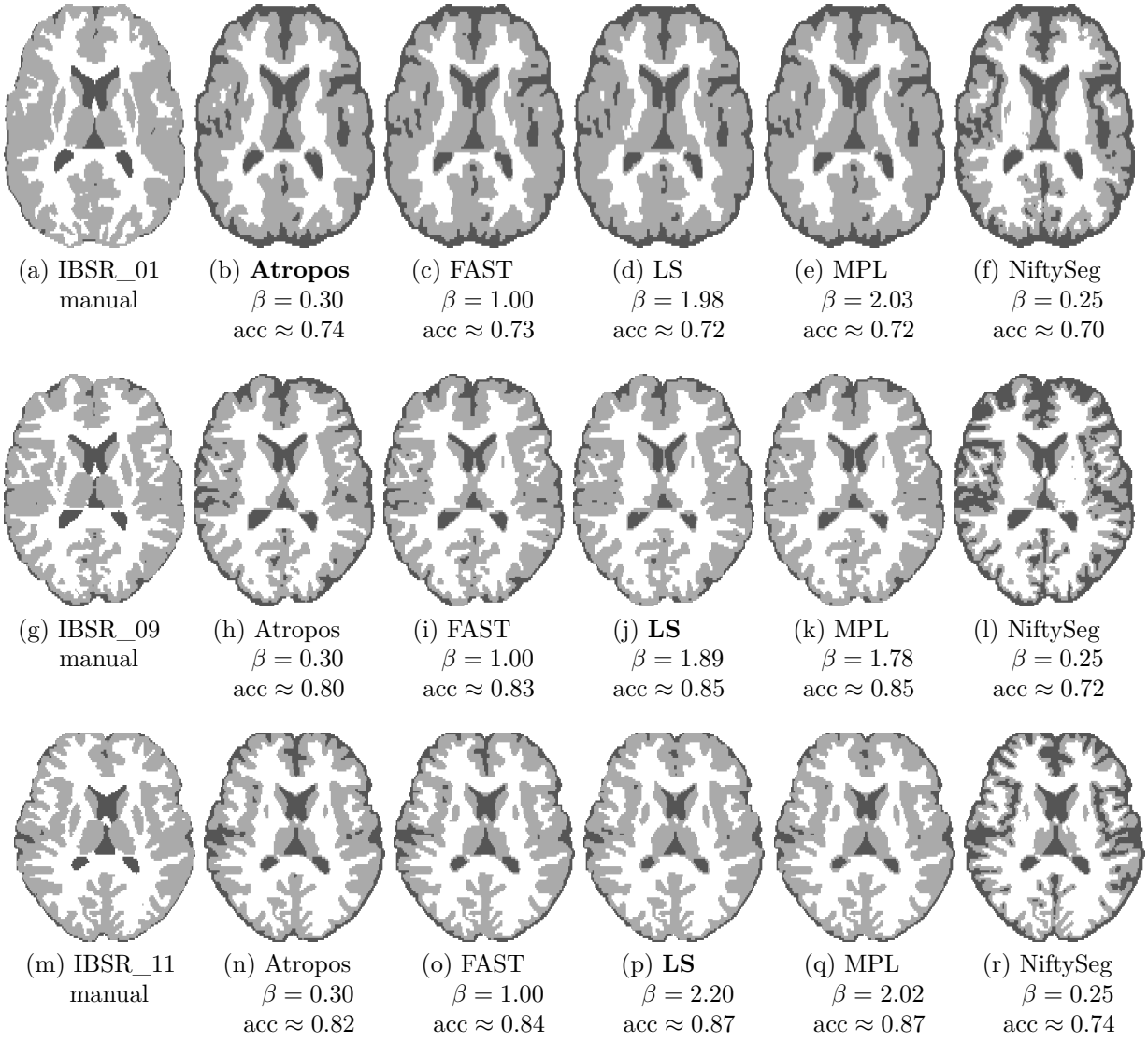
3.6.2 MRF estimation

3.6.2.1 Qualitative comparison of segmentations

With the exception of NiftySeg, all the segmentation accuracies are quite similar at approximately 80%. While the difference in accuracies between methods may seem small so as to be insignificant, one must consider that the majority of the brain is homogeneous in intensity, and easily classified by intensity. It is primarily along the tissue borders that differences will be seen due to slightly different decision boundaries between methods, and in voxels of noise. These make up a relative small proportion of the brain, hence small percentage point differences in accuracy can translate to noticeable qualitative differences in segmentations. In addition in a clinical context, changes due to (for example) aging may only be very small per year and localised to particular regions of the brain, so application-specific metrics may yield more noticeable differences.

Figure 3.16 shows sample segmentations produced by the various methods, alongside the manual segmentation. Both relatively poorly- and well-segmented subjects are shown (IBSR_01 and IBSR_11 respectively). All segmentation methods had difficulty segmenting sub-cortical grey matter such as the thalamus and basal ganglia; for example, they are absent in subject IBSR_09, merged into the insula for IBSR_01, and present but diminished for IBSR_11. Sub-cortical grey matter structures typically have indistinct boundaries and intensity closer to white matter, and cannot be segmented based on intensity alone; atlas-based approaches are needed to properly segment these regions (Pohl et al., 2005). As a result, grey matter volumes were lower those of the corresponding manual segmentation (figure 3.17), with the additional volume being mostly allocated to WM. Splitting the Dice metrics into sub-cortical and cortical GM may elucidate the difference in methods more clearly, but these labellings were not available.

It can also be seen that all segmentations produced significantly more CSF (in percentage volume) than the manual segmentation. This is particularly so for the NiftySeg segmentations; we examine the reason in the next section. Examining the sample segmentations, this additional CSF



(a) IBSR_01 manual | (b) **Atropos** $\beta = 0.30$ acc ≈ 0.74 | (c) FAST $\beta = 1.00$ acc ≈ 0.73 | (d) LS $\beta = 1.98$ acc ≈ 0.72 | (e) MPL $\beta = 2.03$ acc ≈ 0.72 | (f) NiftySeg $\beta = 0.25$ acc ≈ 0.70

(g) IBSR_09 manual | (h) Atropos $\beta = 0.30$ acc ≈ 0.80 | (i) FAST $\beta = 1.00$ acc ≈ 0.83 | (j) **LS** $\beta = 1.89$ acc ≈ 0.85 | (k) MPL $\beta = 1.78$ acc ≈ 0.85 | (l) NiftySeg $\beta = 0.25$ acc ≈ 0.72

(m) IBSR_11 manual | (n) Atropos $\beta = 0.30$ acc ≈ 0.82 | (o) FAST $\beta = 1.00$ acc ≈ 0.84 | (p) **LS** $\beta = 2.20$ acc ≈ 0.87 | (q) MPL $\beta = 2.02$ acc ≈ 0.87 | (r) NiftySeg $\beta = 0.25$ acc ≈ 0.74

Figure 3.16: Example segmentations for subjects (in rows) by various methods (in columns) and the manual segmentation, with β value and segmentation accuracy ('acc'). Low accuracy (IBSR_01) and high-accuracy (IBSR_11) subjects are shown. The highest-accuracy method for each subject is in bold.

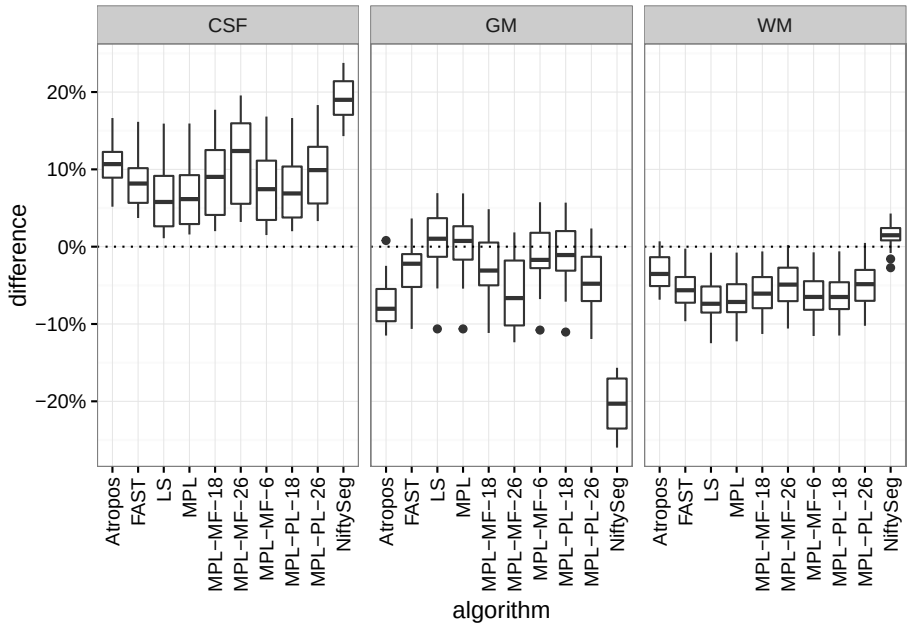


Figure 3.17: Difference in tissue volume (as a percentage of brain volume) relative to manual segmentations

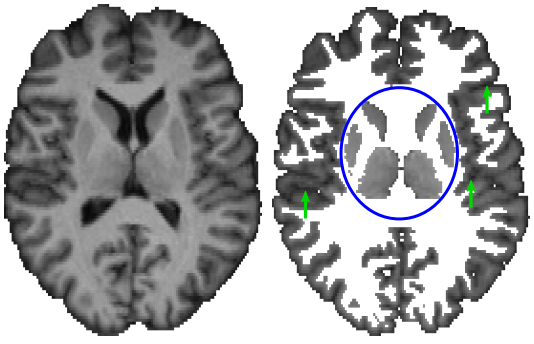


Figure 3.18: MRI of subject IBSR_09 and grey matter region according to the manual segmentation. CSF in the sulci the brain appears to be classified as GM (green arrows). The intensity of deep-GM structures (blue circle) is much lighter than the cortical GM.

appears to be sulcal CSF - between the folds of the cortical surface. The manual segmentations from the IBSR dataset consider sulcal CSF voxels to be classified as grey matter since they typically occupy less than a voxel, which others may consider to be CSF (Valverde et al., 2015). However, all algorithms tested have classified these voxels as CSF, as the intensity is noticeably darker than the remainder of the CSF (figure 3.18). This is reflected in the poor Dice coefficients for CSF. There is a corresponding negative effect on the Dice coefficients for GM due to the shift in tissue boundary. Ideally, the manual segmentations could be adjusted to include sulcal CSF, or such voxels could be excluded from the Dice scores. However, as we lacked a manual accurate classification of sulcal CSF, we could not do this. This speaks somewhat to the need for an automated segmentation algorithm to avoid differences in manual protocol between brains and ensure consistency of segmentations. Alternatively, use of an additional co-registered MRI taken with a different sequence would help (such as a T2 scan, on which CSF is much more visible). Then the intensities would be vectors consisting of the T1 and T2 intensity at each voxel rather than scalars.

3.6.2.2 Value of β estimation

The two estimated-β methods, LS and MPL, had the highest two overall accuracies. This suggests that estimation of β is desirable to fixing it. Though FAST ($\beta = 1$, 6 neighbours) and Atropos ($\beta = 0.3$, 26 neighbours) did not perform significantly worse than LS or MPL, NiftySeg ($\beta = 0.25$, 6 neighbours) did. While fixing β saves the computation of estimation and e.g. $\beta = 1$ as with FAST seems reasonable for this dataset, there is no way to know *a priori* that $\beta = 1$ is suitable for the dataset. The value of estimation is help to avoid the situation of an inappropriate β being chosen (for this dataset, an example is NiftySeg's $\beta = 0.25$).

NiftySeg significantly underperforms the other methods. We note that it has the lowest β of all the fixed-β methods (Atropos has similar $\beta = 0.3$, but a much higher neighbourhood size; as discussed previously, there is evidence for a linear relationship between the two). As β approaches 0, the mixture-MRF model reduces to a standard Gaussian mixture model with a multinomial prior with fixed equal mixing proportions for each tissue. The fixed and equal mixing proportions bias the segmentation to have equal tissue volumes, which is not realistic. This is particularly noticeable in CSF, which takes up a relatively small proportion of the brain: an average of 8% in these examples, based on the manual segmentations. Even if sulcal CSF were added to the manual segmentations, CSF still takes well less than 1/3 of the brain volume. Indeed, when examining the tissue volumes in figure 3.17 it can be seen that NiftySeg has much larger CSF proportion than any of the other segmentations. Upon examining the NiftySeg segmentations in figure 3.16 it can be seen that the sulcal CSF is much thicker than produced by the other algorithms. It is also reflected in the lowest Dice coefficient for CSF of all the methods. In this case, the issue is not so much under-smoothing (through β being too small), but in the forcing of equal tissue proportions in the tissue prior as β approaches 0.

Examining the estimated β values, MPL (1.9) and LS (2.1) produced the highest values, on average, being greater than 1. Figure 3.11 shows the range of estimated β values (for 6 neighbours) over the dataset. These values are all far larger than the comparable 6-neighbour fixed-β values used by FAST ($\beta = 1$) and NiftySeg ($\beta = 0.25$), yet FAST's performance was comparable while NiftySeg's was worse. Atropos' $\beta = 0.3$ is much lower than the corresponding (PL, 26 neighbours) estimated $\beta = 0.49$ found in the MPL configuration experiment (table 3.4), yet attains almost identical segmentation accuracy ($\approx 80\%$). This suggests that the response of segmentation accuracy to β may be relatively flat in a region near the maximum pseudolikelihood value, yet sensitive as β approaches 0. We verify this trend in the grid-search experiment (section 3.6.3).

On the other hand, we also note that the estimated β values might be high because of the need for an approximate E-step in the EM algorithm. Since the latent variables $\boldsymbol{z}$ are not completely marginalised out (we must condition on $\boldsymbol{z}_{\partial \boldsymbol{i}}{}^{(t)}$), the segmentation of the C-step can affect the estimated β value in the M-step, which will in turn affect the segmentation in the next C-step and so on. Since the majority of neighbourhoods in a segmentation are homogeneous, the MPLE can favour a high β. If β increases, ICM (in the C-step) will favour a smoother segmentation, and so on.

The estimated β values are very variable over the dataset. The range in β values are $\Delta\beta = 0.62$ for MPL and $\Delta\beta = 1.44$ for LS, which are relatively large considering the significant difference in accuracy results between e.g. $\beta = 1$ for FAST and $\beta = 0.25$ for NiftySeg which differ by only $\Delta\beta = 0.75$. This further justifies the need for β to be adjusted on an individual per-image basis, as the estimated β values differ greatly.

We must also acknowledge that FAST, NiftySeg and Atropos are typically used with a co-registered anatomical atlas to aid the segmentation. This gives the probability π_{ij} that each voxel i is a given tissue j. It is usually constructed from a number of manually-segmented brain MRI that are co-registered to give a probabilistic map of brain tissue. The atlas is incorporated by multiplying it through the tissue prior and renormalising. Since these tools were designed primarily for use with an atlas (though also work without one), it could be that the default β values were selected based on test cases with atlases. Since the atlas probability is multiplied by the MRF probability, β can have less effect than when there is no atlas, so the algorithm may be more robust to mis-specifying it. Correspondingly, the segmentation accuracy may be much flatter with respect to β when an atlas is used, so that a large range of different β will produce comparable accuracies.

The problem with requiring an atlas is that it must be registered to the image to be segmented in order to be effective, and this is of itself a difficult task. In addition, if the input brain has pathologies or if the atlas brain is quite different to the input brain (e.g. due to age differences), the atlas can misguide the segmentation (see Pagnozzi et al. (2015) for examples of this). Automatic selection of β is thus even more important, since there is no atlas to mitigate "poor" values of β.

3.6.2.3 Comparison of estimators

The two estimators discussed - the MPLE and LSE - were selected for different reasons. The LSE is the only estimator that has been used in tissue segmentation to automatically determine β with no training data (Van Leemput et al., 1999b). It was implemented and made available as the "Expectation Maximisation Segmentation" package (Van Leemput, 2001). The MRF used is a different form of the Potts MRF (we will study it in Chapter 4), but the derivation of the estimator is the same. The MPLE is chosen because it satisfies the properties defined in the Aims well.

We found no significant difference in accuracy between the MPL and LS methods. Given this, we prefer the MPL estimator to the LS estimator. Both estimators satisfy properties 1 and 2: they estimate β on a per-image basis and do not require training data or an atlas (though an atlas can be incorporated by multiplying the MRF probabilities by the atlas and re-normalising). However, the MPLE better fulfils properties 3 (computational tractability) and 4 (straightforwardness of implementation). In addition, it is more interpretable.

For the MPLE, the Q-function is concave with respect to β, has gradient readily available in closed-form, and consists of a univariate maximisation only, so is computationally cheap to optimise. Thus MPLE is computationally tractable. In addition, the Q-function has already been computed as part of the E-M algorithm, regardless of whether β is estimated or not (namely, the terms ${\tau_{ij}}^{(t)}$ and ${u_{ij}}^{(t)}$). Since ${\tau_{ij}}^{(t)}$ and ${u_{ij}}^{(t)}$ do not change with β (they use the fixed $\beta^{(t-1)}$), there is no additional computation required to do the optimisation besides the optimisation itself. This makes the MPLE particularly suited for implementation into existing segmentation algorithms that use the mixture-Potts image model (such as FAST, NiftySeg, and Atropos); only an additional maximisation in the M-step need be added, and all terms required have already been computed.

By contrast, the LSE requires computation of the neighbourhood counts $N(j, \boldsymbol{z}_{\partial \boldsymbol{i}})$ for each label j and neighbourhood $\boldsymbol{z}_{\partial \boldsymbol{i}}$ before the regression is performed, though the regression itself is computationally cheap. As the neighbourhood size increases this can become unviable. Additionally, the LSE has no mean-field analogue like the pseudolikelihood does. Counting neighbourhood frequencies $N(\boldsymbol{z}_i, \boldsymbol{z}_{\partial \boldsymbol{i}})$ requires discrete labels, hence is incompatible with the mean-field approximation. Both these aspects make it difficult to incorporate the LSE into existing algorithms; if it uses the mean-field approximation, this must be thresholded in order to count the neighbourhoods. The neighbourhood counts themselves must also be computed, as they would not normally be if β were fixed.

The LSE may also be very sensitive to the current segmentation. A given neighbourhood $\boldsymbol{z}_{\partial \boldsymbol{i}}$ must appear with at least two different centre labels j and k in order to contribute to the system of equations. This has two implications: first, it is possible that there are insufficient different neighbourhood and centre voxel combinations to provide enough equations for the LSE estimator to be used. However, this is unlikely with only one parameter to be estimated. Second,

neighbourhoods that are rare in the segmentation may have similar influence to those that are common. The majority of neighbourhoods in an MRI have all neighbours as well as the centre voxel of the same tissue label, for example $\boldsymbol{z}_{\partial i}$ are all GM and j is GM. If there are no occurrences of $\boldsymbol{z}_{\partial i}$ all being GM but k being a different tissue (as for an isolated voxel of noise), then the all-GM $\boldsymbol{z}_{\partial i}$ cannot be used to determine β. For an image with no noise, this may be the case for the all-CSF, all-GM, and all-WM neighbourhoods. Then the homogeneous neighbourhoods could not influence the β estimate, despite making up the vast majority of neighbourhoods in the image. This leaves β to be determined by a minority of neighbourhoods. Suppose on the next iteration, a single all-GM neighbourhood is changed such that the neighbours are still all GM, but the centre is WM. Suddenly, this neighbourhood may contribute to the β estimate with $N(\text{GM}, \boldsymbol{z}_{\partial i})$ very large and $N(\text{WM}, \boldsymbol{z}_{\partial i}) = 1$. Thus the addition of even a single neighbourhood may alter the estimate significantly.

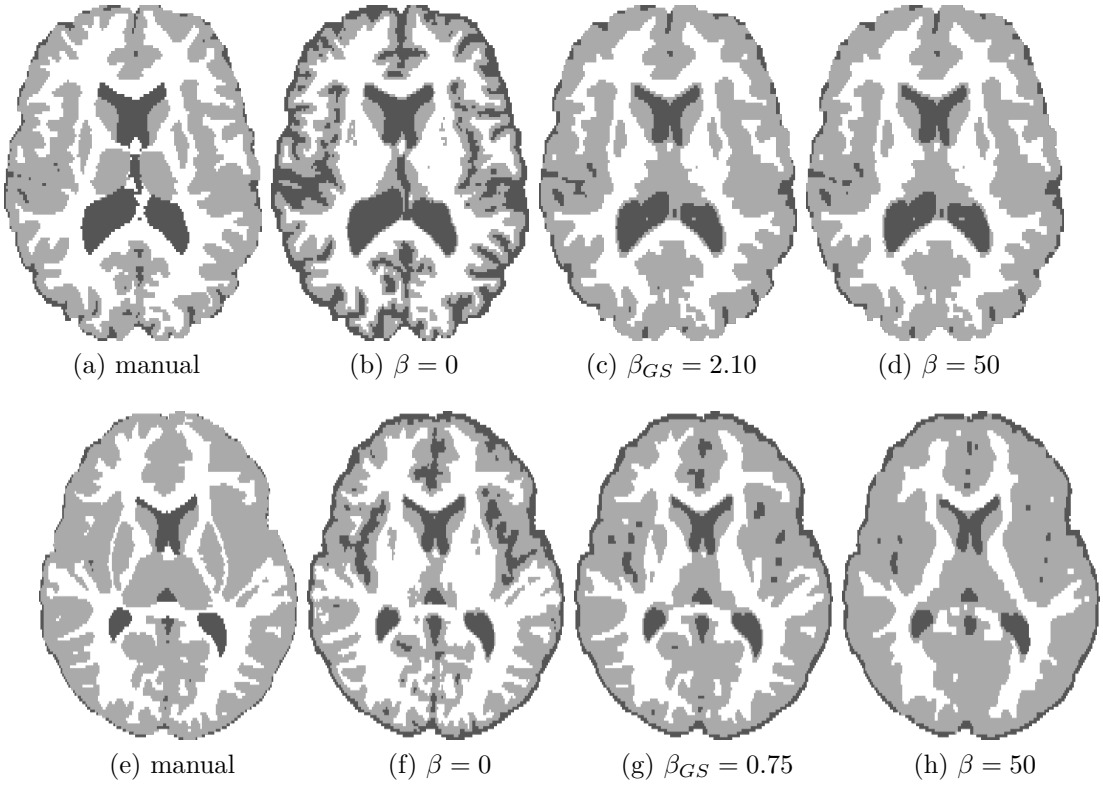
Another consideration is that the LSE can only be found using the prior MRF probability $\tilde{p}(\boldsymbol{z})$. It cannot also incorporate the intensity distribution $f(\boldsymbol{y}|\boldsymbol{z})$ due to the need to empirically estimate $f(y_i, \boldsymbol{e}_j, \boldsymbol{z}_{\partial i})$ in the right-hand side of the system of equations ((3.14)). On the other hand, the MPLE can incorporate this information by maximising the joint log-pseudolikelihood $f(\boldsymbol{y}|\boldsymbol{z})\tilde{p}(\boldsymbol{z})$ or Q-function instead of simply $\tilde{p}(\boldsymbol{z})$. For the MPLE, maximisation of the MRF pseudolikelihood alone is inferior to maximising the joint pseudolikelihood/Q-function (Archer and Titterington, 2002), since it does not use all available information. It could be that the same applies to the LSE.

In terms of interpretability, the MPLE is straightforward and fits with the existing goals of EM: it maximises an approximation to the likelihood (or Q function). This is exactly the M-step for β. If a given neighbourhood appears many times in the image, it has more weight in the log-likelihood (though given that the majority of neighbourhoods in a natural image are homogeneous and therefore are more likely under a higher β, this could lead to oversmoothing).

The LSE does not maximise likelihood, but rather matches observed to expected neighbourhood ratios. Gurelli and Onural (Gurelli and Onural, 1994) proposed that for the Ising MRF ($g = 2$), the LS estimator corresponds to maximising the log-pseudolikelihood of distinct neighbourhood configurations separately, and combining these estimates through least-squares. However this does not generalise to $g > 2$. A given neighbourhood may contribute up to 3 separate equations of the form (3.13) with $g = 3$ labels, and this will not in general each give the same β for that neighbourhood.

In our experiments, we found β values were generally higher for LS than MPL figure 3.11. This does not appear to have resulted in the LS segmentations being particularly oversmooth compared to the MPL segmentations, though the β values do not appear significantly higher. While a strong presence of homogeneous neighbourhoods drives β higher under MPL estimation as just discussed, it is unclear how these affect the LS estimate.

Guyon and Künsch (1992) showed that asymptotically, the MPLE and LS estimators for the



(a) manual (b) $\beta = 0$ (c) $\beta_{GS} = 2.10$ (d) $\beta = 50$

(e) manual (f) $\beta = 0$ (g) $\beta_{GS} = 0.75$ (h) $\beta = 50$

Figure 3.19: Manual segmentation and segmentations produced for various fixed β values. β_{GS} is the fixed β with the highest segmentation accuracy from the grid search. Subjects IBSR_10 (top) and IBSR_17 (bottom).

visible Ising MRF ($g = 2$) are equivalent, if weighted least squares is used for the latter with carefully-chosen weights. As far as we are aware, there is no analogue for the Potts MRF ($g > 2$), nor for a hidden MRF.

3.6.3 Grid search

Figure 3.12 shows the accuracy-vs-β curve of the two subjects examined, and figure 3.19 shows selected segmentations at small, high, and the maximum-accuracy β from the grid search (denoted β_{GS}). For the two subjects examined, accuracy is lowest when β is small and increases quickly, reaching a maximum by $\beta = 2.5$ before decreasing and plateauing. Accuracies above 90% are never achieved, which shows that the mixture-Potts MRF alone cannot perfectly segment the brain; this is not surprising due to difficult regions such as the sub-cortical grey matter, and given that the brain is a complex structure that cannot be fully captured by a Markov random field.

It appears that over-estimating β is preferable to under-estimating it. Rather than continually decreasing, the accuracy asymptotes as β becomes large. The loss in accuracy when β is higher than β_{GS} is not nearly as much as when β is comparably lower.

First we consider behaviour for small β. As discussed for NiftySeg, as β approaches 0 the

model approaches a standard mixture model with fixed mixing proportions of 1/3. As CSF in particular makes up well less than 1/3 of the brain, it is over-emphasised, leading to a loss of accuracy. This can be seen in the segmentations for $\beta = 0$, figure 3.19c and figure 3.19g, where there is considerably more CSF than the manual segmentation.

The fixed β values attaining the highest accuracy, β_{GS}, were 2.10 and 0.75. These are quite variable, suggesting that a single value of β is not appropriate for all images in a dataset.

As expected, the method that used β closest to the grid-search value attained closest to the highest accuracy. For subject IBSR_10 ($\beta_{GS} = 2.10$), this was the two estimated-β methods. For subject IBSR_17 ($\beta_{GS} = 0.75$), this was FAST. Here, the two estimated-β methods both over-estimated β. On the other hand, we note that neither method is maximising segmentation accuracy directly, as this is not typically available. MPL uses log-pseudolikelihood as a proxy for accuracy, while LS matches ratios of neighbourhood statistics. In light of this, it is not surprising that these methods do not always achieve the highest accuracy. The key point is that while estimation of β does not always yield the highest-accuracy β, it yields values that attain near the accuracy maximum, and avoids choosing values that are significantly worse (e.g. NiftySeg's $\beta = 0.25$ for these images).

It is curious to see that the accuracy asymptotes as β becomes larger. Initially we had expected that as β became large, the accuracy would drop dramatically due to oversmoothing (i.e. the entire segmentation becomes uniformly one label, which has high MRF likelihood). Instead, the segmentations are still brain-like, and even have small isolated circles of tissue (for example in IBSR_17's $\beta = 50$ segmentation, figure 3.19h). We infer that past a certain point the value of β does not matter; the same segmentation will be obtained regardless of its value (given a particular set of starting parameters). This segmentation comprises a local maximum in which every voxel is of the same tissue label as its neighbourhood majority. Referring back to IBSR_17's $\beta = 50$ segmentation, although there are isolated dots of tissue, they are convex in shape.

To see this, we fix the intensity parameters and segmentation, and consider β at which a given voxel i changes from one label to the other. In particular, we show that as β becomes large, the voxel must eventually be classified to the label of its neighbourhood majority. Suppose that the neighbourhood majority is label j, i.e. $u_{ij} > u_{ik} \forall j \neq k$ (if $u_{ij} = u_{ik}$, then the MRF probability to be tissue k or j is the same regardless of β, so is decided by intensity alone). Suppose also that voxel i has label k, not the neighbourhood majority. For voxel i to remain label k, its posterior probability to be label k must dominate all other labels, i.e.

$$\exp(\beta u_{ik})\phi_{ik} > \exp(\beta u_{ij})\phi_{ij}, \tag{3.15}$$

where ϕ_{ik} is shorthand for the Normal intensity density $\phi(y_i; \mu_k, \sigma_k^2)$. Rearranging for β,

$$\log \frac{\phi_{ik}}{\phi_{ij}} > \beta(u_{ij} - u_{ik}).$$

Since u_{ij} is the neighbourhood majority, $u_{ij} - u_{ik} > 0$ and so

$$\beta < \log \frac{\phi_{ik}}{\phi_{ij}} / (u_{ij} - u_{ik}).$$

The above must be satisfied in order for voxel i to remain label k. But as β becomes large, it must eventually violate the constraint, as the right-hand-side does not depend on β. Thus eventually all voxels must eventually be classified as their neighbourhood majority.

In the special case that there are more than one labels that are the equal neighbourhood majority, for example 3 neighbours of GM and 3 neighbours of CSF, the discrimination rule (3.15) is independent of β, and the label of that voxel is determined by intensity alone.

This is why the accuracy asymptotes as β becomes large - the segmentation eventually reaches a point where every voxel has been classified to its neighbourhood majority (or determined by intensity in the case of an equal majority), and cannot change any further. An example is seen in the $\beta = 50$ segmentations in figure 3.19. Of course, this 'asymptotic' segmentation is only local, and depends on the initialisation.

3.7 Conclusion

We have studied the Potts MRF as a prior for mixture-MRF brain MRI segmentation, and how it may be solved using Expectation-Maximisation. This forms the basis for most of the current popular intensity-based segmentation tools such as FAST, NiftySeg and Atropos. Atlas-based tool FreeSurfer also makes use of the Potts MRF as a post-hoc segmentation smoother. These tools all require the operator to specify a value for the Potts MRF smoothing parameter β and offer default values, all different to each other. It is not clear which value to use on a given image, and we have demonstrated that segmentation accuracy can varying greatly depending on the choice. In addition, FAST, NiftySeg and Atropos are commonly used with an anatomical atlas where the effect of β may be lessened. It is unclear whether the default values are selected for use with or without an atlas, and how to adjust them. For this reason, β should be determined automatically.

In this chapter we aimed to develop a way to automatically specify β for the Potts MRF on a per-image basis (property 1) that does not require training data (property 2), is computationally tractable (property 3), and straightforward to implement into existing algorithms (property 4). We proposed use of the *maximum pseudolikelihood estimator* of Besag (1974) for this purpose. Use of MPL estimation of β is novel in brain MRI segmentation. The MPLE satisfies the first two properties by design. The Q function and pseudolikelihood are concave with respect to β and the gradient is readily computable in closed-form, making the maximisation itself computationally cheap. Our method simply implements the M step of the EM algorithm for the β parameter as well as the intensity parameters, whereas existing methods do this for

the intensity parameters only. Existing methods already make use of the pseudolikelihood approximation even when β is fixed. This in particular makes the MPLE simple to implement into existing methods, as all quantities required to perform the maximisation have already been computed. Thus, it satisfies properties 3 and 4.

We have investigated the effect of neighbourhood size and likelihood approximation on segmentation accuracy when used with the MPLE. We found the smallest neighbourhood of 6 neighbours to be ideal, as additional neighbours appear to unduly influence the MRF due to the E-step being approximate rather than exact. Use of only 6 neighbours instead of 26 simplifies computation of the MRF. Unexpectedly, the pseudolikelihood approximation was found to be more accurate than the mean-field approximation.

The third aim of the chapter was to validate the method on real data and compare it to existing methods. To demonstrate the value of β estimation, we have compared segmentations using the MPLE to various standard fixed-β values: 0.25 (NiftySeg), 0.3 (Atropos), and 1 (FAST) on a real dataset of brains. We found estimation to produce significantly more accurate segmentations than NiftySeg, and slightly though not significantly more accurate segmentations than FAST or Atropos. The estimated β values ranged from 1.5 to 2.1. These were all larger than the largest fixed value (FAST at $\beta = 1$), suggesting that the common fixed-β values may not be appropriate. Thus estimating β will do no worse, and may do better than, existing fixed-β methods.

To further examine the role of β, we performed a grid search over fixed β values for two subjects. While there exist some fixed-β value(s) that attain higher accuracy than MPL, these are variable; there is no reason to expect the same value to work for all images. Also, there is no way to find these optimal (in terms of segmentation accuracy) values without manual tuning and without a manual segmentation to compare accuracy against. MPL is a reasonable alternative that does not require training data or manual tuning. It always attained close to the maximum accuracy found from the grid search, while avoiding "poor" β values.

The grid search also revealed that choosing β "too low" can result in segmentations with much lower accuracy than choosing it "too high". When β is too low, tissues with low proportions such as CSF are over-emphasised. On the other hand, the accuracy curve seems to be relatively flat with respect to β at and after the maximum-accuracy value. As β becomes large, the segmentation converges to one in which every voxel is classified to the same label as the majority of its neighbours. The particular segmentation will depend on the initialisation. While this oversmoothing does cause a loss of accuracy, it is not nearly as much as when β is too small.

We also compared the MPLE with the least-squares estimator (LSE), as investigated in Van Leemput et al. (1999b) for a different form of the Potts MRF. This estimator, like the MPLE, is statistically consistent as the image size goes to infinity for an observed MRF. It is the only case known to the authors where parameter estimation for the Potts MRF has been implemented and made readily available for brain MRI segmentation. The segmentations produced with the LSE had higher accuracy compared to MPLE, though not significantly different to MPLE. As

both estimators produce similar accuracies, we prefer the MPLE for a number of reasons:

- The LSE matches local neighbourhood frequency ratios to theoretical ratios. This is not easily interpretable in a global sense, as maximum likelihood is.
- The LSE cannot be applied in as many situations as MPLE. It cannot be used with the mean-field approximation, due to the need to approximate $p(j, \boldsymbol{z}_{\partial i})$ by empirical frequencies. It also cannot incorporate the intensity probabilities.
- A given neighbourhood must appear with at least two different centre labels in order to form an equation; otherwise, it is omitted. This is not necessarily guaranteed. The effect of omitting neighbourhoods from the system of equations is not clear.
- Although the least-squares estimation itself is computationally efficient, in order to set up the equations the neighbourhood frequencies $p(j, \boldsymbol{z}_{\partial i})$ must be calculated for the current segmentation. As the neighbourhood size increases, this becomes more cumbersome.

On the other hand, the MPLE does not require calculation of any additional quantities before the maximisation is performed, as the pseudolikelihood is already calculated as part of the existing EM algorithm.

In conclusion, we have presented a method for automatic spatial regularisation in brain MRI segmentation using the MPLE, which has not been considered for this purpose before. We strongly advocate for estimation of β using MPLE:

- It produces segmentations no worse than standard fixed-β methods, and sometimes better if β is fixed to an inappropriate value.
- For a given (visible) segmentation, the MPL estimator is known to be consistent.
- It falls naturally within the EM framework - the Q-function (making use of pseudolikelihood) is maximised with respect to β in the M-step, along with the other parameters.
- The Q-function with a fixed segmentation is concave with respect to β. The gradient is readily available in closed-form and computationally tractable due to use of an MRF approximation. The resulting maximisation is univariate and concave, hence does not add noticeable additional computational burden.
- MPL is very readily incorporated into the existing methods using this mixture-Potts model with β fixed. This is because the Q-function is already calculated in such methods, so does not need to be re-calculated for the maximisation.

Chapter 4

Non-homogeneous Potts MRF

4.1 Introduction

The single-beta Potts MRF is widely used in MRI segmentation to incorporate spatial smoothness into the image model. Its single smoothing parameter β has a straightforward interpretation: larger is smoother; smaller is less smooth. Because of this, it is often tuned manually. However, as we have shown in the previous chapter, it is preferable to automatically determine β; manually specifying it can result in poor segmentations. An additional disadvantage is that this smoothing is applied globally across the segmentation. As a result, the MRF may oversmooth parts of the brain while undersmoothing others. This can be a particular issue in MRI, where signal-to-noise ratio is often non-uniform and tissue composition can be heterogeneous, resulting in variable contrast.

The full or non-homogeneous Potts MRF allows smoothing to be determined on a per-tissue basis, allowing for finer control. The Potts MRF allows a different smoothing β to be set depending on which two tissues are neighbours. For example, the grey matter-white matter cortical boundary along the sulci and gyri of the brain is typically much more convoluted than the CSF-white matter boundary along the ventricles. The Potts MRF can allow for smoothing to be less strong along the former boundary than the latter.

The single-beta Potts MRF consists of a pairwise component only (counting whether a voxel and its neighbour are of the same label). This means that it cannot account for unequal tissue proportions as a standard mixture model can. The full Potts MRF has a unary component akin to the mixing proportion of a standard mixture model, while retaining the smoothing component. Separation of the proportion (unary) and smoothing (pairwise) parameters may improve the sensitivity of the smoothing parameters so that they are not affected by large homogeneous regions of the brain.

Since the full Potts MRF has many more parameters to estimate (one per pair of tissues for the smoothing, and one per tissue for the tissue proportions), it is not reasonable to manually

specify the parameters - they must be determined automatically. Possibly for this reason, the full Potts MRF has not been widely used in brain segmentation. Van Leemput et al. (1999b) used an MRF similar to the full Potts MRF for brain segmentation with estimation via the least-squares estimator. However, it was not identifiable, and had many more parameters than the version we present here. Cardoso et al. (2011) also used a form of the full Potts MRF that was simplified to have only two smoothing parameters, chosen manually. Forbes et al. (2013) introduces the full Potts MRF with maximum pseudolikelihood estimation for spatial disease risk mapping, which is later used in brain lesion segmentation by their students and collaborators (Maggia et al., 2016; Kabir et al., 2007; Menze et al., 2015). However, while all papers mention the full Potts MRF and the potential of its use for brain MRI segmentation, they reduce the model to the single-beta MRF in application.

In this chapter, we make a detailed study of the full Potts MRF, various parametrisations of it, and various methods of parameter estimation, applying all of these to brain MRI segmentation to compare their performance. In particular, we separate out the unary (mixing proportion) and pairwise (tissue-wise smoothing) components of the MRF potential and examine their roles in detail, experimenting with various combinations of these. We show how maximum pseudolikelihood estimation (MPLE) is used to adaptively determine the MRF parameters. We also derive the corresponding least-squares estimator (LSE), similar to Van Leemput et al. (1999b), and compare it to the MPLE. The properties that made MPLE preferable to the LSE for the single-beta MRF are largely preserved for the full Potts MRF:

- The Q function is negative-semidefinite, with gradient available in closed-form.
- The maximisation does not require new quantities (such as neighbourhood frequencies) to be calculated as the least-squares estimator does.
- The MPLE can be used in more scenarios than the LSE and is less prone to problems of missing neighbourhoods or lack of neighbourhood diversity.

We also discuss how the parameters might be constrained using prior knowledge to improve the MRF.

As already mentioned, the use of the full Potts MRF with MPL for parameter estimation is not novel (Forbes et al., 2013). However, while the full Potts MRF has been discussed for brain segmentation before (Maggia et al., 2016; Kabir et al., 2007; Menze et al., 2015), it has not been used (in all such papers the single-beta form was used). This may have been because these papers were focused on the application rather than the method; as we will discover, the full Potts MRF with no prior constraints on or training of the parameters is not necessarily suitable for segmentation. By contrast, we concentrate specifically on the full Potts MRF with a detailed study of its various parametrisations and a focus on model selection. In addition, we focus on the estimation method, comparing the MPLE to the LSE.

4.1.1 Aim

The aims of this chapter are:

- To study the use of the full Potts MRF in a fully-automatic segmentation method for brain MRI.
- To study how the full Potts MRF may be used to more finely tailor the MRF prior to brain segmentation, with regards to controlling for the relative frequency of tissues and neighbouring tissues.
- To study various parameterisations of the full Potts MRF and which is most suited to brain segmentation, with respect to both practical outcomes and standard statistical model selection metrics.
- To compare estimation of the parameters of this MRF using the MPLE to the LSE.

4.2 Background

The image model is still a mixture-MRF model: intensities $y_i,\ i = 1, \ldots, n$ are normally distributed for each tissue, and the tissue labels $\boldsymbol{z}_i,\ i = 1, \ldots, n$ are distributed according to a Markov random field. Recall g is the number of tissue labels, and $\boldsymbol{z}_i$ is a vector of length g indicating which label voxel i takes; $\boldsymbol{z}_i = \boldsymbol{e}_j$ means that voxel i has label j. Here $\boldsymbol{e}_j$ is the indicator vector of 0s everywhere except in the jth position which has a 1. In this chapter, the normal distribution for the intensity conditioned on tissue remains unchanged; we focus rather on different forms of the MRF than the single-beta Potts model.

4.2.1 Non-homogeneous Potts MRF

The full Potts model was introduced and briefly discussed for image restoration by Besag (1986); however, he did not use the full form but the simplified single-beta version. The full Potts MRF has local conditional potential given by:

$$
\begin{aligned}
p(\boldsymbol{Z}_i = \boldsymbol{e}_j | \boldsymbol{z}_{\partial \boldsymbol{i}}) &\propto \exp(\alpha_j - \sum_{\substack{k=1 \\ k \neq j}}^{g} \beta_{jk} u_{ik}) \\
\alpha_1 &= 0 \\
u_{ik} &= \sum_{m \in \partial \boldsymbol{i}} \frac{z_{mk}}{\delta_{im}} \\
\beta_{jk} &= \beta_{kj} \\
\beta_{jk} &> 0
\end{aligned}
\tag{4.1}
$$

As with the single-beta MRF, u_{ik} denotes the number of neighbours of voxel i with the label k, scaled by the distance between the voxels. The model parameters are the unary parameters α_j and the smoothing parameters β_{jk}. For reasons explained below, we constrain $\alpha_1 = 0$ and $\beta_{jk} = \beta_{kj}$, so that the model parameters $\boldsymbol{\Psi}$ consist of α_j where $j = 2, \ldots, g$ with $\alpha_1 = 0$ and β_{jk} where $1 \leq j < k \leq g$, giving a total of $(g-1) + \binom{g}{2}$ parameters.

The unary parameters α_j are related to the relative proportions of the label j in the image, though are not the same unless $\beta = 0$. Then, the model because a standard normal mixture model with spatially independent multinomial prior with probabilities $\pi_j = \exp(\alpha_j)$. These parameters depend only on z_{ij} so do not perform any smoothing. Similar to the standard mixture model, only $g-1$ parameters are required for an MRF with g labels as α_j are only unique up to an additive constant. To see this, set $\tilde{\alpha_j} = \gamma + \alpha_j$ where γ is a constant and $j = 1, \ldots, g$. Then $\exp(\tilde{\alpha_j}) = \exp(\gamma)\exp(\alpha_j)$ and the constant $\exp(\gamma)$ is cancelled out in the denominator. Without loss of generality we set $\alpha_1 = 0$, but retain it in the model equations for ease of notation and interpretation. The reason to include the α_j parameters into the model is that the β parameters control the smoothness of the interface between different tissues, but cannot directly allow for different base proportions of these tissues. Alternatively, the terms $\exp(\alpha_j)$ may be thought of as multiplying prior knowledge of the tissue proportions by the smoothing portion of the MRF.

The pairwise term over the neighbours gives a penalty for neighbouring voxels that do not match voxel i's label (compared to the single-beta MRF, which encourages neighbouring voxels to be of the same label). The parameters β_{jk} control the penalty for labels j and k co-occurring in the same neighbourhood. These may be used to enforce anatomical constraints on tissues unlikely to appear adjacent to each other. For the model to be a valid MRF, we must have $\beta_{jk} = \beta_{kj}$, as pairwise potentials must be symmetric. As with the single-beta MRF, the larger and more positive β_{jk} becomes, the more tissue j and k are penalised from co-occurring in the same neighbourhood.

The single-beta MRF may be recovered by allowing all the smoothing parameters to be the same, i.e. $\beta_{jk} = \beta \; \forall \; j, k$ and setting $\alpha_j = 0 \; \forall \; j$. Then

$$\exp(-\sum_{j \neq k} \beta_{jk} u_{ik}) = \exp(-\beta \sum_{j \neq k} u_{ik}) = \exp(-\beta(|\boldsymbol{\partial i}| - u_{ij})) \propto \exp(\beta u_{ij}),$$

where $|\boldsymbol{\partial i}|$ is the number of neighbours and the constant $\exp(-\beta|\boldsymbol{\partial i}|)$ is cancelled out in the numerator and denominator.

It will often be convenient to write (4.1) in matrix form. Let $\boldsymbol{\alpha}$ be the length-g vector $(0, \ldots, \alpha_g)^T$ (recall that $\alpha_1 = 0$).. Let $\boldsymbol{B}$ be a $g \times g$ symmetric matrix, where $\boldsymbol{B}_{jk} = \beta_{jk}$ for

$j < k$, $\boldsymbol{B}_{jk} = \beta_{jk} = \beta_{kj}$ for $j > k$ and $\boldsymbol{B}_{jj} = 0$. For example, with $g = 3$ tissues,

$$\boldsymbol{B} = \begin{bmatrix} 0 & \beta_{12} & \beta_{13} \\ \beta_{12} & 0 & \beta_{23} \\ \beta_{13} & \beta_{23} & 0 \end{bmatrix}.$$

There are $\binom{g}{2}$ unique β_{jk} parameters. Then, the conditional pdf may be written

$$\begin{aligned} p(\boldsymbol{z}_i|\boldsymbol{z}_{\partial \boldsymbol{i}}; \boldsymbol{\Psi}) &= \frac{\boldsymbol{z}_i^T \boldsymbol{\alpha} - \boldsymbol{z}_i^T \boldsymbol{B} \boldsymbol{u}_i}{C_i}, \text{ where} \\ C_i &= \sum_{j=1}^{g} \exp(\boldsymbol{e}_j^T \boldsymbol{\alpha} - \boldsymbol{e}_j^T \boldsymbol{B} \boldsymbol{u}_i) \end{aligned} \tag{4.2}$$

In the above we have introduced the notation $\boldsymbol{u}_i$ for the vector of neighbour counts $(u_{i1}, u_{i2}, \ldots, u_{ig})^T$.

4.2.2 Related work

4.2.2.1 Similar MRFs

In this chapter, we focus on models where the pairwise parameters depend only on the *class* of the voxel and its neighbours. This reduces the number of pairwise parameters $\binom{g}{2}$, very few compared to the number of voxels. A number of MRFs similar to the non-homogeneous Potts model are used in other application areas, often with pairwise parameters that depend on the *direction* of the neighbour, or the position of the neighbours.

Texture segmentation is the task of identifying regions of different texture in an image - for example, grass, leather, wood. One way to achieve this is by use of a hidden Markov random field with *directional* parameters. For example, Derin and Elliott (1987) use the MRF

$$p(\boldsymbol{z}_i|\boldsymbol{z}_{\partial \boldsymbol{i}}) \propto \exp(\boldsymbol{z}_i^T \alpha + \beta_h \boldsymbol{z}_i^T (\boldsymbol{z}_E + \boldsymbol{z}_W) + \beta_v \boldsymbol{z}_i^T (\boldsymbol{z}_N + \boldsymbol{z}_S) + \beta_{d1} \boldsymbol{z}_i^T (\boldsymbol{z}_{NE} + \boldsymbol{z}_{SW}) + \beta_{d2} \boldsymbol{z}_i^T (\boldsymbol{z}_{NW} + \boldsymbol{z}_{SE})),$$

where $\boldsymbol{z}_E$ is the east neighbour, $\boldsymbol{z}_W$ is the west neighbour, β_h is the horizontal parameter, and so on. The set of parameters can be used to characterise particular textures. For example, a texture consisting of vertical stripes may low or even negative β_h, encouraging a pixel to be a *different* label than its horizontal neighbour, and a high β_v, encouraging a pixel to take the *same* label as its vertical neighbour. The parameters should be learned on various textures in order to distinguish between them. In the mentioned paper, least squares estimation was used. Manjunath and Chellappa (1991) used a similar formulation but with a Gaussian MRF over the image intensities rather than underlying labels. We will consider directional MRFs of this nature in the next chapter.

Another MRF with a similar form is the Boltzmann machine (Ackley et al., 1985). These are a

type of neural network that also form an MRF. Observations i, called "units", are binary-valued. In our notation, $\boldsymbol{z}_{i1} = 1$ corresponds to unit i being off, while $\boldsymbol{z}_{i2} = 1$ corresponds to it being on. The probability of the system is given by

$$p(\boldsymbol{z}) \propto \exp(-(\sum_{i=1}^{n} \alpha_i z_{i2} + \sum_{i<m} \beta_{im} \boldsymbol{z}_i^T \boldsymbol{z}_m))$$

Here the α_i parameters, known as the *bias* of unit i, are specific to each unit (compared to our α_j which are specific only to the *value* at each i). Likewise, there is one parameter β_{im} for each pair of units; these are the *connection strengths*. Unlike our neighbourhood structure, for a Boltzmann machine, every unit can be a neighbour to every other unit. The connection strengths or weights β_{im} are typically learned based on training data, with a value of zero meaning no connection between the ith and mth units.

4.2.2.2 Brain segmentation

The full Potts MRF is not commonly used in brain segmentation, possibly due to the fact that parameter estimation is difficult and often not performed. Manually tuning the parameters is difficult given the number of parameters and their non-intuitive interaction.

Van Leemput et al. (1999b) used the MRF

$$p(\boldsymbol{z}_i | \boldsymbol{z}_{\partial i}; \boldsymbol{\Psi}) \propto \pi_{ij} \exp(-\boldsymbol{z}_i^T \boldsymbol{G} \boldsymbol{u}_i^G - \boldsymbol{z}_i^T \boldsymbol{H} \boldsymbol{u}_i^H), \tag{4.3}$$

where the $g \times g$ matrices $\boldsymbol{G}$ and $\boldsymbol{H}$ are not constrained to be symmetric, nor do they have 0 on the diagonal. For the moment, assume $\pi_{ij} = 1$. Here $\boldsymbol{u}_i^G$ is the neighbour count for the in-plane (east, west, south and north) neighbours while $\boldsymbol{u}_i^H$ is the neighbour count for the top and bottom neighbours. These are *not* distance-weighted. It is broadly similar to the full Potts MRF, except that:

- Rather than a single matrix of $\boldsymbol{B}$ parameters, there is a different one for the within-slice neighbours ($\boldsymbol{G}$) than the between-slice neighbours ($\boldsymbol{H}$), to account for a difference in resolution. We handle this by use of distance scaling in the neighbour counts. Van Leemput et al. found that $\boldsymbol{G}$ and $\boldsymbol{H}$ were roughly the same, due to the isotropic spacing of the images used.
- The matrices $\boldsymbol{G}$ and $\boldsymbol{H}$ are not constrained to be symmetric, so do not form a valid MRF. An MRF must have non-directional relationships between voxels, i.e. the pairwise potential should be the same for voxels (i, m) as (m, i).
- The diagonal elements of $\boldsymbol{G}$ and $\boldsymbol{H}$ fill the role of $\boldsymbol{z}_i^T \boldsymbol{\alpha}$.

While the full Potts MRF uses $g - 1$ α_j parameters and $\binom{g}{2}$ β_{jk} parameters making $\frac{(g-1)(g+2)}{2}$ in total, the version used in Van Leemput et al. (1999b) has strictly more parameters: $2g^2$. The non-symmetry of $\boldsymbol{G}$ and $\boldsymbol{H}$, while causing the prior to no longer be an MRF, do not practically

affect the segmentations. The parameters are only identifiable up to the sum $(G_{jk} + G_{kj})$, but the individual parameter values themselves do not matter to the segmentation. The parameter matrices were estimated using the least-squares estimator.

The elements of $\boldsymbol{G}$ and $\boldsymbol{H}$ were constrained to reflect anatomical prior knowledge. If this was not done, it was found that the MRF oversmoothed the segmentation. For example, the authors wished a voxel surrounded by CSF and GM to have the same (prior) probability to be either, leaving the classification decision up to the intensity pdf rather than the MRF. To do this, the constraints $\boldsymbol{G}_{\text{CSF,CSF}} = \boldsymbol{G}_{\text{CSF,GM}}$ and $\boldsymbol{G}_{\text{GM, CSF}} = \boldsymbol{G}_{\text{GM, GM}}$ were imposed. Similar constraints were imposed for GM and WM. However, given that $\boldsymbol{G}_{jk}$ represents the penalty (or bonus) for tissue j to occur next to tissue k, the diagonal elements $\boldsymbol{G}_{jj}$ should ideally be negative to encourage homogeneous regions of tissue, while the $\boldsymbol{G}_{jk}$ should be positive in order for smoothing to occur along the j-k boundary or 0 to disable smoothing along this boundary. In order to have the CSF-GM boundary decided on intensity alone, one should instead set $\boldsymbol{G}_{\text{CSF,GM}} = \boldsymbol{G}_{\text{GM,CSF}} = 0$.

Finally, a registered anatomical atlas was incorporated such that π_{ij} in the equation above reflects the atlas probability that voxel i is tissue j. We can consider this as the Potts MRF with voxel-specific α_{ij} parameters rather than merely tissue-specific α_j parameters, i.e.

$$\pi_{ij} = \exp(\alpha_{ij}).$$

In this thesis, the primary aim is to focus on applications where no atlas information is available. However, the MRFs we study can all be used with an atlas prior by multiplying them through as above.

Forbes et al. (2013) presents the full Potts MRF with maximum pseudolikelihood estimation, as we intend to in this chapter. The application is to spatial disease risk mapping. In practice, they use the single-beta form of the Potts MRF, though they mention that the full $\boldsymbol{B}$ matrix could be estimated on training data or tuned by an expert (e.g., by constraining some parameters to be multiples of others based on *a priori* knowledge). The full Potts MRF is also mentioned in a number of brain lesion segmentation application papers (Maggia et al., 2016; Kabir et al., 2007; Menze et al., 2015) with reference to this paper. The methods are similar to brain tissue segmentation but include an extra tissue class for lesions. Detail is brief (as the papers focus on the application), but while they mention the full Potts MRF, it is implied that the single-beta MRF is used (with MPL estimation). While we present the same model as Forbes et al. (2013) and also with MPL estimation, we apply the full model rather than (in addition to) the single-beta form and specifically focus on model specification and selection. We also compare MPL estimation with least-squares estimation.

Cardoso et al. (2011) used the Potts MRF with multiple smoothing parameters β_{jk} and no unary parameters ($\alpha_j = 0$). The smoothing parameters were not estimated. Instead the β_{jk} penalties were set to 0.5 if tissues j and k were "anatomically plausible", and 3 if they were

"anatomically implausible". To enable this, an anatomical atlas was used to further subdivide the tissue classes: CSF into internal CSF (e.g. the ventricles) and external CSF (e.g. on the outside boundary of the brain), and GM into deep GM and cortical GM. Then, tissue pairs may be graded on whether they are anatomically likely to appear next to each other or not: for example, deep grey-matter is unlikely to border external CSF, but is likely to border internal CSF. If CSF and GM were not further split in this way, such grading would not be possible, as any two of CSF, GM, or WM are likely to share a boundary in a healthy brain.

Wels et al. (2011) used an MRF with unary parameters α_j and one smoothing parameter β fixed to 1, which was further weighted by the difference in the neighbouring voxel intensities (we will study this form of local anisotropic smoothing in Chapter 4). The α_j parameters were fixed to prior tissue probabilities learned from a probabilistic atlas; however, rather the prior being specific to each voxel and tissue, they were only dependent on the tissue, i.e.

$$\alpha_j = \log(\tilde{\pi}_j),$$

where $\tilde{\pi}_j$ is the atlas-learned probability that any voxel is tissue j. In a way, this is similar to the mixing proportions/multinomial prior probabilities of a standard spatially independent mixture model.

The difference of our work to both of these cases (Cardoso et al. (2011) and Wels et al. (2011)) is that none of our parameters are pre-determined, or require an atlas or training data to determine. Rather, all parameters are estimated from the current best segmentation and observed data (voxel intensities). We wish to see if the full Potts MRF is a better tissue prior than the single-beta MRF of the previous chapter, when both do not have atlas information.

Roche and Forbes (2014) replaced the labels $\boldsymbol{z}_i$ with continuous versions representing tissue concentrations, in order to explicitly handle partial volume effects. For example, $\boldsymbol{z}_i = (1/2, 0, 1/2)$ corresponds to a voxel that is comprised of 50% CSF and 50% WM in the underlying tissue, so that the observed intensity is a mixture of these. On the surface it may seem that this is equivalent to our formulation of a single discrete label per voxel and conditional intensity distribution given that label, but it is not. In our formulation, given $\boldsymbol{z}_i$, the observed intensity is drawn from a *single* normal distribution with parameters corresponding to $\boldsymbol{z}_i$'s single class. In the continuous formulation, the observed intensity at each voxel arises from a *mixture* of normal distributions according to the mixing proportions in $\boldsymbol{z}_i$. Their MRF is (after a change of notation)

$$\exp(-\boldsymbol{z}_i^T \boldsymbol{A} \boldsymbol{z}_i - \beta \sum_{m \in \boldsymbol{\partial i}} |\boldsymbol{z}_i - \boldsymbol{z}_m|^2), \tag{4.4}$$

where $\boldsymbol{A}$ is a $g \times g$ matrix with the same structure as our $\boldsymbol{B}$. The pairwise component is quite different to our $\boldsymbol{z}_i^T \boldsymbol{z}_m$, being more suited to measuring the distance between the continuous $\boldsymbol{z}_i$ and $\boldsymbol{z}_m$. Because of the way the $\boldsymbol{z}_i$ variables are formulated, $\boldsymbol{A}$ may be interpreted similarly to our $\boldsymbol{B}$ matrix. They fixed the $\boldsymbol{A}$ matrix to values learned from training data. Our model is quite different to this (as are the other details of their model and solution method), but we

mention the MRF as $\boldsymbol{A}$ can be interpreted as smoothing between tissue classes, and we also observe an artefact in the segmentations arising from this that those authors did. We will discuss it further in section 4.6.1.2.

4.3 Method

First, we briefly restate the image model used. Then, we give the full Potts MRF and various special cases of it that are used to study the effects of its various parameters. We show how to estimate the parameters with maximum pseudolikelihood and derive the gradient. Then, we show how to form the least-squares estimate for the parameters, for comparison to the method of Van Leemput et al. (1999b). Finally, we outline how to solve the model for the mixture and MRF parameters and segmentation using Expectation-Maximisation.

The image model is the same as that presented in the previous chapter. Let $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ be random variables where Y_i is the intensity of voxel i in an n-voxel MR volume. We consider the case of a single-channel MRI, i.e. Y_i is scalar, though the theory is readily applied to a multichannel/multivariate case. Let g be the number of tissue classes. We use $g = 3$: cerebrospinal fluid (CSF), grey matter (GM) and white matter (WM). Let $\boldsymbol{Z} = (\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n)$ be random variables giving the tissue classification or label of each voxel. $\boldsymbol{Z}_i$ indicates the tissue label of voxel i. Let $\boldsymbol{e}_j$ be the indicator vector with a 1 in the jth position and 0 elsewhere. $\boldsymbol{Z}_i = \boldsymbol{e}_j$ if and only if voxel i is tissue j. The set of voxels that neighbour voxel i is denoted $\boldsymbol{\partial i}$, and $\boldsymbol{z}_{\boldsymbol{\partial i}}$ are the labels of all such neighbours. Lowercase letters e.g. y_i and $\boldsymbol{z}_i$ are used to represent realisations of $\boldsymbol{Y}_i$ and $\boldsymbol{Z}_i$.

The intensities are assumed Normally distributed given their label. The labels are distributed according to the Potts MRF.

$$\boldsymbol{Y}_i | (\boldsymbol{Z}_i = \boldsymbol{e}_j) \sim \mathcal{N}(\mu_j, \sigma_j^2)$$
$$\boldsymbol{Z} \sim \text{Potts}(\boldsymbol{\Psi}).$$

However, we consider a number of forms of the full Potts MRF to examine and dissociate the effects of the various parameters. The EM algorithm is used to solve for the intensity and MRF parameters as described previously. This is described in section 2.5.

4.3.1 Choice of MRF

We consider a number of forms of the Potts MRF to examine and dissociate the effects of the various parameters. They are summarised in table 4.1. Throughout, we use α and "alpha" to refer to the unary parameters of the MRF that are spatially independent, and β, $\boldsymbol{B}$ or "beta" to refer to the pairwise parameters that are spatially dependent and enable smoothing.

First, the full Potts MRF, called the "alpha-multi-beta" model, retains all α and β parameters

for the greatest flexibility of all models considered:

$$\begin{aligned} p(\boldsymbol{Z}_i = \boldsymbol{e}_j | \boldsymbol{z}_{\partial \boldsymbol{i}}) &\propto \exp(\alpha_j - \sum_{\substack{k=1 \\ k \neq j}}^{g} \beta_{jk} u_{ik}) = \exp(\boldsymbol{e}_j^T \boldsymbol{\alpha} - \boldsymbol{e}_j^T B \boldsymbol{u}_i) \\ \alpha_1 &= 0 \\ \beta_{jk} &> 0 \\ \boldsymbol{\Psi} &= (\alpha_2, \ldots, \alpha_g, \beta_{jk} \text{ such that } j < k)^T \end{aligned} \tag{4.5}$$

By disabling the α_j parameters we obtain the "multi-beta" model which is most similar to that used by Van Leemput et al. (1999b) and Cardoso et al. (2011):

$$\begin{aligned} p(\boldsymbol{Z}_i = \boldsymbol{e}_j | \boldsymbol{z}_{\partial \boldsymbol{i}}) &\propto \exp(- \sum_{\substack{k=1 \\ k \neq j}}^{g} \beta_{jk} u_{ik}) = \exp(-\boldsymbol{e}_j^T \boldsymbol{B} \boldsymbol{u}_i) \\ \beta_{jk} &> 0 \\ \boldsymbol{\Psi} &= (\beta_{jk} \text{ such that } j < k)^T \end{aligned} \tag{4.6}$$

By setting all β_{jk} to the same value β and disabling the α_j parameters, we recover the single-beta model of the previous chapter for comparison:

$$\begin{aligned} p(\boldsymbol{Z}_i = \boldsymbol{e}_j | \boldsymbol{z}_{\partial \boldsymbol{i}}) &\propto \exp(\beta u_{ij}) = \exp(\beta \boldsymbol{e}_j^T \boldsymbol{u}_i) \\ \beta &> 0 \\ \boldsymbol{\Psi} &= \beta \end{aligned} \tag{4.7}$$

We also consider the single-beta MRF with additional α_j parameters to see if these can account for different tissue proportions. This yields the "alpha-single-beta" model:

$$\begin{aligned} p(\boldsymbol{Z}_i = \boldsymbol{e}_j | \boldsymbol{z}_{\partial \boldsymbol{i}}) &\propto \exp(\alpha_j + \beta u_{ij}) = \exp(\boldsymbol{e}_j^T \boldsymbol{\alpha} + \beta \boldsymbol{e}_j^T \boldsymbol{u}_i) \\ \beta &> 0 \\ \boldsymbol{\Psi} &= \beta \end{aligned} \tag{4.8}$$

By comparing the models with a single smoothing β to those with multiple β_{jk} we can examine if the finer level of smoothing is beneficial. By comparing the models with and without α_j parameters, we can examine if a unary potential is needed beyond the smoothing portion of the MRF.

For the alpha-multi-beta and alpha-single-beta models, we also investigate setting the $\boldsymbol{\alpha}$ parameters to the log-proportions of each tissue type (normalised so that $\alpha_1 = 0$), i.e.

$$\exp(\alpha_j) = \sum_{i=1}^{n} \tau_{ij} / \sum_{i=1}^{n} \tau_{i1}.$$

The reason to use τ_{ij} rather than z_{ij} is that ideally, τ_{ij} has been marginalised over all realisations $\boldsymbol{z}$ in the E-step (though in practice, the E-step is only approximate). This update is analogous to how the mixing proportion parameters π_j are set in a standard mixture model. However unlike the standard mixture model, this is *not* the same as maximising Q with respect to α_j.

The motivation is to view the unary potential as a prior probability for each tissue class, similar to Wels et al. (2011). In this way we can view the MRF as the product of a spatially independent tissue prior with a spatially-depending smoothing MRF. However, unlike Wels et al. (2011) where α_j are determined from an atlas, we explore determining them from the data so as not to rely on an atlas. Our aim is to test if constraining the α_j parameters to reflect class proportions and leaving the pairwise parameters to smooth the segmentation is preferable to leaving all of them unconstrained.

We call these models "fixed-alpha-single-beta" and "fixed-alpha-multi-beta". The term "fixed" is not in the sense of being fixed to a constant throughout the segmentation, but rather as fixed to the logarithm of the current tissue proportions, which will change in each iteration.

Table 4.1: Summary of MRFs

potential	$p(\boldsymbol{e}_j \vert \boldsymbol{z}_{\partial \boldsymbol{i}}; \boldsymbol{\Psi}) \propto$	$\boldsymbol{\Psi}$	$\vert\boldsymbol{\Psi}\vert$
single-beta	$\exp(-\beta \boldsymbol{e}_j^T \boldsymbol{u}_i)$	β	1
multi-beta	$\exp(-\boldsymbol{e}_j^T B \boldsymbol{u}_i)$	$(\beta_{jk}$ such that $j < k)^T$	$\binom{g}{2}$
alpha-single-beta	$\exp(\boldsymbol{e}_j^T \boldsymbol{\alpha} - \beta \boldsymbol{e}_j^T \boldsymbol{u}_i)$	$(\alpha_2, \ldots, \alpha_g, \beta)^T$	$(g-1)+1$
alpha-multi-beta	$\exp(\boldsymbol{e}_j^T \boldsymbol{\alpha} - \boldsymbol{e}_j^T \boldsymbol{B} \boldsymbol{u}_i)$	$(\alpha_2, \ldots, \alpha_g,$ β_{jk} such that $j < k)^T$	$(g-1)+\binom{g}{2}$
fixed-alpha-single-beta	$\exp(\boldsymbol{e}_j^T \boldsymbol{\alpha} - \beta \boldsymbol{e}_j^T \boldsymbol{u}_i)$ $\alpha_j = \log(\sum_{i=1}^n \tau_{ij} / \sum_i \tau_{i1})$	β	1
fixed-alpha-multi-beta	$\exp(\boldsymbol{e}_j^T \boldsymbol{\alpha} - \boldsymbol{e}_j^T \boldsymbol{B} \boldsymbol{u}_i)$ $\alpha_j = \log(\sum_{i=1}^n \tau_{ij} / \sum_i \tau_{i1})$	$(\beta_{jk}$ such that $j < k)^T$	$\binom{g}{2}$

We determine the parameters $\boldsymbol{\Psi}$ using maximum pseudolikelihood estimation. We also derive the corresponding least-squares estimators and use this for comparison, akin to Van Leemput et al. (1999b).

4.3.1.1 Model selection

The Akaike or Bayesian information criteria (AIC and BIC respectively) cannot be computed for an MRF as the true maximum-likelihood estimates of the parameters are not known. Even if they are, the observed-data likelihood $f(y)$ cannot be computed due to the intractability of the MRF. Ji and Seymour (1996) proposed a pseudolikelihood information criterion (PLIC) for *visible* MRFs that is, in essence, the BIC with $p(\boldsymbol{z})$ replaced by the pseudolikelihood. For a

hidden MRF, Forbes and Peyrard (2003) derived an analogous PLIC:

$$\text{PLIC} = 2\log f(y; \hat{\boldsymbol{\Theta}}, \hat{\boldsymbol{\Psi}}) - |\boldsymbol{\Psi}|\log(n). \tag{4.9}$$

With this sign convention, the model with the *highest* PLIC is selected. The likelihood $f(\boldsymbol{y})$ is computed using the pseudolikelihood or mean-field approximation, at the estimated parameter values $\hat{\boldsymbol{\Psi}}$ and $\hat{\boldsymbol{\Theta}}$ and estimated segmentation $\hat{\boldsymbol{z}}$. The neighbours $\hat{\boldsymbol{z}_{\partial \boldsymbol{i}}}$ are taken from the estimated segmentation $\hat{\boldsymbol{z}}$. When ICM is used to produce these the PLIC of Stanford and Raftery (2002) is recovered.

4.3.2 Maximum pseudolikelihood estimation

In what follows, we use the alpha-multi-beta Potts MRF as it is the most general form. The pseudolikelihood approximation to $p(\boldsymbol{z})$ is given by

$$\tilde{p}(\boldsymbol{z}; \boldsymbol{\Psi}) = \prod_{i=1}^{n} \frac{\exp(\boldsymbol{z}_i^T \boldsymbol{\alpha} - \boldsymbol{z}_i^T \boldsymbol{B} \boldsymbol{u}_i)}{\exp(\sum_{k=1}^{g} \boldsymbol{e}_k^T \boldsymbol{\alpha} - \boldsymbol{e}_k^T \boldsymbol{B} \boldsymbol{u}_i)},$$

where

$$\boldsymbol{u}_i = \sum_{m \in \boldsymbol{\partial i}} \delta_{im}^{-1} \boldsymbol{z}_m,$$

and δ_{im} is the distance between voxels i and m. The mean-field approximation is the same, but the neighbours are calculated via

$$\boldsymbol{u}_i = \sum_{m \in \boldsymbol{\partial i}} \delta_{im}^{-1} \langle \boldsymbol{z}_m \rangle,$$

where $\langle \boldsymbol{z}_m \rangle$ satisfies either

$$\langle \boldsymbol{z}_i \rangle = \frac{p(\boldsymbol{z}_i | \boldsymbol{z}_{\partial \boldsymbol{i}})}{\sum_{k=1}^{g} p(\boldsymbol{e}_k | \boldsymbol{z}_{\partial \boldsymbol{i}})}$$

or

$$\langle \boldsymbol{z}_i \rangle = \frac{p(\boldsymbol{z}_i | \boldsymbol{z}_{\partial \boldsymbol{i}}) f(y_i | \boldsymbol{z}_i)}{\sum_{k=1}^{g} p(\boldsymbol{e}_k | \boldsymbol{z}_{\partial \boldsymbol{i}}) f(y_i | \boldsymbol{e}_k)}$$

depending on if $p(\boldsymbol{z})$ or $p(\boldsymbol{z}|\boldsymbol{y})$ is to be approximated. We use the latter as justified by Celeux et al. (2003) and Archer and Titterington (2002).

Maximum pseudolikelihood estimation determines $\boldsymbol{\Psi}$ by optimising the Q-function with respect to them.

$$\boldsymbol{\Psi} = \arg\max_{\boldsymbol{\Psi}} \sum_{i=1}^{n} \sum_{j=1}^{g} \tau_{ij} (\log p(\boldsymbol{z}_i | \boldsymbol{z}_{\partial \boldsymbol{i}}; \boldsymbol{\Psi}) - \log(\sum_{k=1}^{g} p(\boldsymbol{e}_k | \boldsymbol{z}_{\partial \boldsymbol{i}}; \boldsymbol{\Psi})))$$

where τ_{ij} are the (approximate) posterior probabilities $p(\boldsymbol{z}_i | y_i)$, calculated using (4.15). For

example, for the alpha-multi-beta potential:

$$\boldsymbol{\Psi} = \arg\max_{\boldsymbol{\alpha},\boldsymbol{B}} \sum_{i=1}^{n}\sum_{j=1}^{g} \tau_{ij}(\boldsymbol{z}_i^T\boldsymbol{\alpha} - \boldsymbol{z}_i^T\boldsymbol{B}\boldsymbol{u}_i - \log(\sum_{k=1}^{g}\exp(\boldsymbol{e}_k^T\boldsymbol{\alpha} - \boldsymbol{e}_k^T\boldsymbol{B}\boldsymbol{u}_i)).$$

We will retain our previous convention of using "pseudolikelihood" in the sense of "maximum pseudolikelihood" to mean either approximation, and make it explicit when we wish to reference one or the other.

By the same argument presented in section 3.3.1.1, the Q-function is negative semi-definite with respect to $\boldsymbol{\Psi}$.

4.3.2.1 Gradient

The gradient of the Q-function may be obtained in closed form, making it amenable to gradient descent algorithms. It was derived previously (3.9) as:

$$\nabla_{\boldsymbol{\Psi}} Q(\boldsymbol{y},\boldsymbol{z};\boldsymbol{\Theta},\boldsymbol{\Psi}) = \sum_{i=1}^{n}\sum_{j=1}^{g} \nabla_{\boldsymbol{\Psi}} U_i(\boldsymbol{e}_j|\boldsymbol{z}_{\boldsymbol{\partial i}}{}^{(t)};\boldsymbol{\Psi}) \left(-\tau_{ij}{}^{(t)} + p(\boldsymbol{e}_j|\boldsymbol{z}_{\boldsymbol{\partial i}}{}^{(t)};\boldsymbol{\Psi})\right),$$

where $U_i(\boldsymbol{e}_j|\boldsymbol{z}_{\boldsymbol{\partial i}})$ is $-\boldsymbol{e}_j^T\boldsymbol{\alpha} + \boldsymbol{e}_j^T\boldsymbol{B}\boldsymbol{u}_i$ for the alpha-multi-beta model, and a simplified version for the others (the negative is from the convention of writing the MRF probability as $\exp(-U_i(\cdot))$).

For the $\boldsymbol{\alpha}$ parameters, it is convenient to take derivatives with respect to $\boldsymbol{\alpha}$ even though $\alpha_1 = 0$; the first element of the resulting gradient vector is ignored. Now

$$\nabla_{\boldsymbol{\alpha}} U_i(\boldsymbol{e}_j|\boldsymbol{z}_{\boldsymbol{\partial i}}{}^{(t)}) = -\boldsymbol{e}_j,$$

so that

$$\begin{aligned}\nabla_{\boldsymbol{\alpha}} Q(\boldsymbol{y},\boldsymbol{z};\boldsymbol{\Theta},\boldsymbol{\Psi}) &= \sum_{i=1}^{n}\sum_{j=1}^{g} -\boldsymbol{e}_j \left(-\tau_{ij}{}^{(t)} + p(\boldsymbol{e}_j|\boldsymbol{z}_{\boldsymbol{\partial i}}{}^{(t)};\boldsymbol{\Psi})\right), \\ &= \sum_{i=1}^{n} \boldsymbol{\tau}_i{}^{(t)} - \boldsymbol{p}_i{}^{(t)},\end{aligned}$$

where $\boldsymbol{\tau}_i{}^{(t)} = (\tau_{i1}{}^{(t)}, \tau_{i2}{}^{(t)}, \ldots, \tau_{ig}{}^{(t)})^T$ is voxel i's posterior probability to be each tissue label, and similarly $\boldsymbol{p}_i = (p(\boldsymbol{e}_1|\boldsymbol{z}_{\boldsymbol{\partial i}}{}^{(t)}), p(\boldsymbol{e}_2|\boldsymbol{z}_{\boldsymbol{\partial i}}{}^{(t)}), \ldots, p(\boldsymbol{e}_g|\boldsymbol{z}_{\boldsymbol{\partial i}}{}^{(t)})^T$ is voxel i's prior or MRF probability to take each tissue label. As $\alpha_1 = 0$, we only consider elements 2 to g of the above gradient vector; however, it is convenient to write it with the full $\boldsymbol{\alpha}$.

For $\boldsymbol{B}$ in the alpha-multi-beta and multi-beta potentials, since $\boldsymbol{B}$ is symmetric, we only give the derivatives for β_{jk} where $j < k$. We write

$$\boldsymbol{B} = \boldsymbol{B}' + \boldsymbol{B}'^T,$$

where $\boldsymbol{B}'$ is the upper-diagonal matrix with $(\boldsymbol{B}')_{jk} = \beta_{jk}$ for $j < k$, and all other entries are 0.

Then

$$U_i(\boldsymbol{e}_j|\boldsymbol{z}_{\partial \boldsymbol{i}}) = -\boldsymbol{e}_j^T\boldsymbol{\alpha} + \boldsymbol{e}_j^T(\boldsymbol{B}' + \boldsymbol{B}'^T)\boldsymbol{u}_i$$

and

$$\nabla_{\boldsymbol{B}'} U_i(\boldsymbol{e}_j|\boldsymbol{z}_{\partial \boldsymbol{i}}{}^{(t)}) = \boldsymbol{e}_j \boldsymbol{u}_i{}^{(t)^T} + \boldsymbol{u}_i{}^{(t)}\boldsymbol{e}_j^T$$

The derivative with respect to β_{jk} where $j < k$ is found by the (j,k)th element of

$$\nabla_{\boldsymbol{B}'} Q(\boldsymbol{y}, \boldsymbol{z}; \boldsymbol{\Theta}, \boldsymbol{\Psi}) = \sum_{i=1}^{n}\sum_{j=1}^{g} \left(\boldsymbol{e}_j \boldsymbol{u}_i{}^{(t)^T} + \boldsymbol{u}_i{}^{(t)}\boldsymbol{e}_j^T\right)\left(-\tau_{ij}{}^{(t)} + p(\boldsymbol{e}_j|\boldsymbol{z}_{\partial \boldsymbol{i}}{}^{(t)}; \boldsymbol{\Psi})\right)$$

Through tedious algebra, it can be shown that

$$\nabla_{\boldsymbol{B}'} Q(\boldsymbol{y}, \boldsymbol{z}; \boldsymbol{\Theta}, \boldsymbol{\Psi}) = (-\boldsymbol{\tau}^{(t)} + \boldsymbol{p})^T \boldsymbol{u}^{(t)} + \boldsymbol{u}^{(t)^T}(-\boldsymbol{\tau}^{(t)} + \boldsymbol{p}),$$

where

- $\boldsymbol{\tau}^{(t)}$ is the $n \times g$ matrix with $(\boldsymbol{\tau}^{(t)})_{ij} = \tau_{ij}{}^{(t)}$,
- $\boldsymbol{u}^{(t)}$ is the $n \times g$ matrix with $(\boldsymbol{u}^{(t)})_{ij} = u_{ij}{}^{(t)}$,
- $\boldsymbol{p}$ is the $n \times g$ matrix with $(\boldsymbol{p})_{ij} = p(\boldsymbol{e}_j|\boldsymbol{z}_{\partial \boldsymbol{i}}{}^{(t)}; \boldsymbol{\Psi})$,

and we only consider the (j,k)th element with $j < k$ for the derivative with respect to β_{jk}. For MRFs with only one β parameter, the β part of $U_i(\cdot)$ is $-\beta u_{ij}$, so

$$\nabla_{\beta} Q(\boldsymbol{y}, \boldsymbol{z}; \boldsymbol{\Theta}, \boldsymbol{\Psi}) = \sum_{i=1}^{n}\sum_{j=1}^{g} -u_{ij}{}^{(t)}\left(-\tau_{ij}{}^{(t)} + p(\boldsymbol{e}_j|\boldsymbol{z}_{\partial \boldsymbol{i}}{}^{(t)}; \boldsymbol{\Psi})\right),$$

Finally, the derivatives are:

$$\nabla_{\boldsymbol{\alpha}} Q(\boldsymbol{y}, \boldsymbol{z}; \boldsymbol{\Theta}, \boldsymbol{\Psi}) = \sum_{i=1}^{n} \boldsymbol{\tau}_i{}^{(t)} - \boldsymbol{p}_i{}^{(t)} \qquad \text{(alpha-single-beta,alpha-multi-beta)}$$

$$\nabla_{\boldsymbol{B}'} Q(\boldsymbol{y}, \boldsymbol{z}; \boldsymbol{\Theta}, \boldsymbol{\Psi}) = (-\boldsymbol{\tau}^{(t)} + \boldsymbol{p})^T \boldsymbol{u}^{(t)} + \boldsymbol{u}^{(t)^T}(-\boldsymbol{\tau}^{(t)} + \boldsymbol{p}) \qquad \text{(alpha-multi-beta, multi-beta)}$$

$$\nabla_{\beta} Q(\boldsymbol{y}, \boldsymbol{z}; \boldsymbol{\Theta}, \boldsymbol{\Psi}) = \sum_{i=1}^{n}\sum_{j=1}^{g} -u_{ij}{}^{(t)}\left(-\tau_{ij}{}^{(t)} + p(\boldsymbol{e}_j|\boldsymbol{z}_{\partial \boldsymbol{i}}{}^{(t)}; \boldsymbol{\Psi})\right) \qquad \text{(single-beta, alpha-single-beta)}$$

where the appropriate equations are used depending on the parameters of the model. These have been written in this way to emphasise that they can be achieved by straight-forward matrix multiplications and additions, or in the case of a single β, element-wise matrix multiplication.

4.3.3 Least-squares estimation

Least-squares estimation can be applied to MRFs that are log-linear in their parameters. Suppose the MRF can be written

$$\exp(-U_i(\boldsymbol{e}_j|\boldsymbol{z}_{\partial\boldsymbol{i}};\boldsymbol{\Psi})) = \exp(-\boldsymbol{V}_i(\boldsymbol{e}_j|\boldsymbol{z}_{\partial\boldsymbol{i}})^T\boldsymbol{\Psi}),$$

where $\boldsymbol{V}_i$ is a $|\boldsymbol{\Psi}| \times 1$ vector of coefficients for each parameter. From Bayes' rule,

$$\frac{p(\boldsymbol{z}_i|\boldsymbol{z}_{\partial\boldsymbol{i}})}{p(\boldsymbol{z}_i,\boldsymbol{z}_{\partial\boldsymbol{i}})} = \frac{1}{p(\boldsymbol{z}_{\partial\boldsymbol{i}})},$$

where the dependency of p on $\boldsymbol{\Psi}$ has been dropped for ease of notation. As the right-hand side does not depend on the value of $\boldsymbol{z}_i$ itself, it can be seen that

$$\frac{p(\boldsymbol{Z}_i = \boldsymbol{e}_j|\boldsymbol{z}_{\partial\boldsymbol{i}})}{p(\boldsymbol{e}_j,\boldsymbol{z}_{\partial\boldsymbol{i}})} = \frac{p(\boldsymbol{Z}_i = \boldsymbol{e}_k|\boldsymbol{z}_{\partial\boldsymbol{i}})}{p(\boldsymbol{e}_k,\boldsymbol{z}_{\partial\boldsymbol{i}})}$$

for any distinct labels j and k with the same neighbourhood $\boldsymbol{z}_{\partial\boldsymbol{i}}$. Rearranging and substituting $p(\boldsymbol{z}_i|\boldsymbol{z}_{\partial\boldsymbol{i}})$ yields

$$\begin{aligned}\frac{p(\boldsymbol{e}_j,\boldsymbol{z}_{\partial\boldsymbol{i}})}{p(\boldsymbol{e}_k,\boldsymbol{z}_{\partial\boldsymbol{i}})} &= \frac{p(\boldsymbol{e}_j|\boldsymbol{z}_{\partial\boldsymbol{i}})}{p(\boldsymbol{e}_k|\boldsymbol{z}_{\partial\boldsymbol{i}})} \\ &= \exp(-\boldsymbol{V}_i(\boldsymbol{e}_j|\boldsymbol{z}_{\partial\boldsymbol{i}})^T\boldsymbol{\Psi} + \boldsymbol{V}_i(\boldsymbol{e}_k|\boldsymbol{z}_{\partial\boldsymbol{i}})^T\boldsymbol{\Psi}) \qquad (4.10)\\ (-\boldsymbol{V}_i(\boldsymbol{e}_j|\boldsymbol{z}_{\partial\boldsymbol{i}}) + \boldsymbol{V}_i(\boldsymbol{e}_k|\boldsymbol{z}_{\partial\boldsymbol{i}}))^T\boldsymbol{\Psi} &= \log\left(\frac{p(\boldsymbol{e}_j,\boldsymbol{z}_{\partial\boldsymbol{i}})}{p(\boldsymbol{e}_k,\boldsymbol{z}_{\partial\boldsymbol{i}})}\right)\end{aligned}$$

The right-hand side is estimated by

$$\log\left(\frac{p(\boldsymbol{e}_j,\boldsymbol{z}_{\partial\boldsymbol{i}})}{p(\boldsymbol{e}_k,\boldsymbol{z}_{\partial\boldsymbol{i}})}\right) \approx \log\frac{N(j,\boldsymbol{z}_{\partial\boldsymbol{i}})}{N(k,\boldsymbol{z}_{\partial\boldsymbol{i}})}$$

where $N(j,\boldsymbol{z}_{\partial\boldsymbol{i}})$ is the number of times the neighbourhood $\boldsymbol{z}_{\partial\boldsymbol{i}}$ occurs with centre label j in the current segmentation. Up to $\binom{g}{2}$ equations for each neighbourhood may be added to the equation, which can then be solved for $\boldsymbol{\Psi}$ using least-squares regression.

For the single-beta MRF and the β component of the alpha-single-beta MRF,

$$\boldsymbol{V}_i(\boldsymbol{e}_j|\boldsymbol{z}_{\partial\boldsymbol{i}}) = -u_{ij}. \qquad (4.11)$$

For the α component of the alpha-single-beta and alpha-multi-beta MRFs,

$$\boldsymbol{V}_i(\boldsymbol{e}_j|\boldsymbol{z}_{\partial\boldsymbol{i}}) = -\boldsymbol{e}_j, \qquad (4.12)$$

without the first component since $\alpha_1 = 0$. For the smoothing parameters β_{jk} in the multi-beta and alpha-multi-beta MRFs, we flatten the upper half of $\boldsymbol{B}$ (not including the diagonal) into a

vector:

$$\boldsymbol{\Psi} = (\beta_{12}, \beta_{13}, \ldots, \beta_{1g}, \beta_{23}, \ldots, \beta_{2g}, \ldots, \beta_{(g-1)(g)})^T$$

Then $\boldsymbol{V}_i(\boldsymbol{e}_r|\boldsymbol{z}_{\partial \boldsymbol{i}})$ is a $\binom{g}{2}$ vector whose elements are

$$\begin{cases} u_{ik} & \text{if the corresponding } \beta_{jk} \text{ has } j = r, \\ u_{ij} & \text{if the corresponding } \beta_{jk} \text{ has } k = r, \\ 0 & \text{otherwise.} \end{cases}$$

For example, when $g = 3$ we have

$$\begin{aligned} \boldsymbol{\Psi} &= (\beta_{12}, \quad \beta_{13}, \quad \beta_{23})^T \\ \boldsymbol{V}_i(\boldsymbol{e}_1|\boldsymbol{z}_{\partial \boldsymbol{i}}) &= (u_{i2}, \quad u_{i3}, \quad 0)^T \\ \boldsymbol{V}_i(\boldsymbol{e}_2|\boldsymbol{z}_{\partial \boldsymbol{i}}) &= (u_{i1}, \quad 0, \quad u_{i3})^T \\ \boldsymbol{V}_i(\boldsymbol{e}_3|\boldsymbol{z}_{\partial \boldsymbol{i}}) &= (\ 0, \quad u_{i1}, \quad u_{i2})^T \end{aligned} \tag{4.13}$$

To create the full $\boldsymbol{V}_i$ for a given MRF, the appropriate $\boldsymbol{\alpha}$ and (β or $\boldsymbol{B}$) coefficients are concatenated from (4.11), (4.12) and (4.13), remembering to omit the α_1 coefficient.

4.3.4 Algorithm

The algorithm used is unchanged from the previous chapter; we repeat it here for clarity. We use Expectation-Maximisation to estimate the MRF and intensity parameters, using the pseudolikelihood or mean-field approximation in the Q-function for computational tractability. The Q function is given by

$$Q(\boldsymbol{\Theta}, \boldsymbol{\Psi}|\boldsymbol{\Theta}^{(t)}, \boldsymbol{\Psi}^{(t)}) = \sum_{i=1}^{n} \sum_{j=1}^{g} {\tau_{ij}}^{(t)} (\log \phi(y_i; \mu_j, \sigma_j^2) + \log p(\boldsymbol{e}_j|\boldsymbol{z}_{\partial \boldsymbol{i}}; \boldsymbol{\Psi})) \tag{4.14}$$

On iteration t:

1. **(C-step)** Form an estimate of the current labels $\boldsymbol{z}^{(t)}$ to be used as neighbours; either discrete (for the pseudolikelihood approximation) or continuous (for the mean-field approximation). The pseudolikelihood version uses the Iterated Conditional Modes (ICM) update

$${\boldsymbol{z}_i}^{(t+1)} = \boldsymbol{e}_j \text{ where } j = \arg\max_k p(\boldsymbol{e}_j|{\boldsymbol{z}_{\partial \boldsymbol{i}}}^{(t,t+1)}; \boldsymbol{\Psi}) \phi(y_i; {\mu_k}^{(t-1)}, {\sigma_k^2}^{(t-1)}).$$

   The mean-field version uses the mean-field update

$$\langle \boldsymbol{z}_i \rangle^{(t+1)} = \sum_{j=1}^{g} \boldsymbol{e}_j \frac{p(\boldsymbol{e}_j|{\boldsymbol{z}_{\partial \boldsymbol{i}}}^{(t,t-1)}; \boldsymbol{\Psi}) \phi(y_i; {\mu_j}^{(t-1)}, {\sigma_j^2}^{(t-1)})}{\sum_{k=1}^{g} p(\boldsymbol{e}_k|{\boldsymbol{z}_{\partial \boldsymbol{i}}}^{(t,t-1)}; \boldsymbol{\Psi}) \phi(y_i; {\mu_k}^{(t-1)}, {\sigma_k^2}^{(t-1)})}$$

These updates should be performed sequentially. To save time, we divide the voxels into coding sets (see Appendix B) and update each set simultaneously, visiting them sequentially.

2. **(E-step)** Calculate ${\tau_{ij}}^{(t)}$, using $\boldsymbol{z}^{(t)}$ from the C-step to compute the neighbour term u_{ij}:

$$ {\tau_{ij}}^{(t)} = \frac{p(\boldsymbol{e}_j|\boldsymbol{z}_{\partial \boldsymbol{i}}{}^{(t)}; \boldsymbol{\Psi}^{(t)})\phi(y_i; {\mu_j}^{(t-1)}, {\sigma_j^2}^{(t-1)})}{\sum_{k=1}^{g} p(\boldsymbol{e}_k|\boldsymbol{z}_{\partial \boldsymbol{i}}{}^{(t)}; \boldsymbol{\Psi}^{(t)})\phi(y_i; {\mu_k}^{(t-1)}, {\sigma_k^2}^{(t-1)})}. \tag{4.15} $$

3. **(M-step)** Maximise Q with respect to $\boldsymbol{\Theta}$ to obtain the intensity parameters.

$$ \begin{aligned} {\mu_j}^{(t)} &= \frac{\sum_{i=1}^{n} {\tau_{ij}}^{(t)} y_i}{\sum_{i=1}^{n} {\tau_{ij}}^{(t)}} \\ {\boldsymbol{\Sigma}_j}^{(t)} &= \frac{\sum_{i=1}^{n} {\tau_{ij}}^{(t)} (y_i - {\mu_j}^{(t)})^2}{\sum_{i=1}^{n} {\tau_{ij}}^{(t)}}. \end{aligned} $$

Then, update $\boldsymbol{\Psi}$ using either the MPL or LS estimators.

For the MPLE, numerically maximise the negative semi-definite Q-function with respect to $\boldsymbol{\Psi}$, using the gradient (alpha-single-beta,alpha-multi-beta) as necessary.

For the LSE, first construct the neighbourhood histogram $N(\boldsymbol{z}_i, \boldsymbol{z}_{\partial \boldsymbol{i}})$ as described in section 2.5. Once the frequency table is computed, it is used to construct the right-hand side of the system of equations (4.10). The left-hand side is constructed using (4.11), (4.12) and (4.13). The system can then be solved using linear least-squares regression for $\boldsymbol{\Psi}$.

These steps are repeated until the relative change in approximate observed log-likelihood falls below a pre-specified tolerance (1e-5 in these experiments), or it decreases. This is

$$ \log f(\boldsymbol{y}) \approx \sum_{i=1}^{n} \log \left(\sum_{j=1}^{g} \phi(y_i|\boldsymbol{e}_j; {\mu_j}^{(t)}, {\sigma_j^2}^{(t)}) p(\boldsymbol{e}_j|\boldsymbol{z}_{\partial \boldsymbol{i}}{}^{(t)}; \boldsymbol{\Psi}^{(t)}) \right). $$

EM on a standard mixture model guarantees an increase in the observed log-likelihood and Q; however, since we use an approximate E-step and a likelihood approximation for $p(\boldsymbol{z})$ rather than the true likelihood, we no longer have this guarantee.

We initialise the algorithm by fitting a standard normal mixture model with 3 components to the image (i.e., without the MRF). This yields an initial segmentation to be used as the neighbours, as well as means and standard deviations. The initial MRF parameters $\boldsymbol{\Psi}^{(0)}$ are estimated from this initial segmentation.

4.4 Experiments

Our aim is to compare the full MRF to the commonly-used single-beta MRF to see if the more specific smoothing and addition of unary potentials is advantageous. Overall, we compare each of the single-beta, alpha-single-beta, multi-beta, and full alpha-multi-beta MRFs using both MPL and LS estimators. We also allow the α_j parameters to be free, or constrain them to log-tissue proportions. The reason for using these variants of the full Potts MRF is to separate out the effect of the pairwise (smoothing) and unary (proportion) parameters. The experiments are outlined in table 4.2.

The closest analogue to Van Leemput et al. (1999b) is the alpha-multi-beta model with LS estimation. There is no comparable method that uses the full MRF while fixing the parameters except for Cardoso et al. (2011), but this requires use of an atlas in order to further split GM into cortical- and deep-GM, and CSF into internal- and external-CSF, and we do not use an atlas in this work. Hence, we do not compare to any fixed-parameter methods. For the single-beta MRF, we have already found that estimation yields segmentations not significantly different from popular fixed-β methods, and in some cases significantly better (particularly if the fixed β is mis-specified). The same reasoning applies here.

In our second experiment, we emulate Wels et al. (2011) in setting α_j to the log-proportion of voxels in class j at each iteration relative to class 1:

$$\begin{aligned} {\alpha_j}^{(t)} &= \log(\tilde{\pi}_j/\tilde{\pi}_1) \\ \tilde{\pi}_j &= \sum_{i=1}^{n} {\tau_{ij}}^{(t)}/n. \end{aligned}$$

In Wels et al. (2011), $\tilde{\pi}_j$ were determined through an anatomical atlas. We instead experiment with determining these directly from the current segmentation. The motivation behind this is to the view tissue prior as

$$p(\boldsymbol{e}_j|\boldsymbol{z}_{\partial \boldsymbol{i}}) \propto \tilde{\pi}_j \exp(-\boldsymbol{e}_j^T \boldsymbol{B} \boldsymbol{u}_i),$$

i.e. the product of prior knowledge about the tissue proportions $\tilde{\pi}_j$, and an MRF that is used exclusively for smoothing. These models are termed the "fixed-alpha-single-beta" and "fixed-alpha-multi-beta" MRFs. We estimate their parameters with MPLE only.

For these experiments we use a neighbourhood size of 6 with the pseudolikelihood approximation, as determined in the previous chapter. All images were segmented with both MPL and LS for each MRF.

The algorithm was evaluated on images from the Internet Brain Segmentation Repository (IBSR) (Rohlfing, 2012).[1] The dataset used consists of T1-weighted coronal MR volumes of 18 normal subjects of ages 7 to 71. Each volume consists of 128 coronal slices spaced at 1.5mm with

[1]The MR brain data sets and their manual segmentations were provided by the Center for Morphometric Analysis at the Massachusetts General Hospital and are available at http://www.cma.mgh.harvard.edu/ibsr/.

Table 4.2: Experiment summary. A neighbourhood of size 6 was used with the pseudolikelihood approximation. All images were segmented with both MPL and LS for each MRF.

MRF	Parameters ($g = 3$)	
single-beta	$\mathbf{\Psi} = \beta$	
alpha-single-beta	$\mathbf{\Psi} = (\alpha_2, \alpha_3, \beta)^T$	
multi-beta	$\mathbf{\Psi} = (\beta_{12}, \beta_{13}, \beta_{23})^T$	
alpha-multi-beta	$\mathbf{\Psi} = (\alpha_2, \alpha_3, \beta_{12}, \beta_{13}, \beta_{23})^T$	
fixed-alpha-single-beta	$\mathbf{\Psi} = \beta$	$\alpha_j = \log(\sum_{i=1}^n \tau_{ij} / \sum_{i=1}^n \tau_{i1})$ $(j > 2)$
fixed-alpha-multi-beta	$\mathbf{\Psi} = (\beta_{12}, \beta_{13}, \beta_{23})^T$	$\alpha_j = \log(\sum_{i=1}^n \tau_{ij} / \sum_{i=1}^n \tau_{i1})$ $(j > 2)$

in-plane resolution varying from 0.84×0.84mm to 1.00×1.00mm. This dataset also contains manual segmentations to compare the automatic segmentations to. The images are already skull-stripped with bias-correction already performed, so no further preprocessing was done. That is, all non-brain voxels (such as skull, fat) are already removed from the image as we wish to concentrate on segmentation of the brain only.

Performance against the manual segmentations is evaluated by two metrics, *segmentation accuracy* and *Dice similarity*. Let A and B represent two segmentations, being sets of indices for each tissue. That is, $A_j, j = 1, \ldots, g$ are non-intersecting subsets of the indices $1, \ldots, n$ whose union is the entire brain, where $i \in A_j$ implies that voxel i is assigned to tissue j in segmentation A.

The *segmentation accuracy* is quantified as the overall percentage of voxels correctly classified. Since the reference and test segmentations have the same number of voxels (all of the brain voxels), this is well-defined.

$$\text{accuracy}(A, B) = \frac{|A \cap B|}{|A|}.$$

The *Dice similarity coefficient* (commonly called 'Dice score' or 'Dice index') (Dice, 1945) is used to compare segmentations on a tissue-by-tissue basis. The Dice coefficient for a given tissue between two segmentations A and B is given by the number of correctly-classified voxels divided by the average area classified (of that tissue):

$$\text{Dice}(A_j, B_j) = \frac{2|A_j \cap B_j|}{|A_j| + |B_j|}.$$

It ranges from 0 to 1, with 1 meaning a perfect match between the two segmentations of that tissue. The reason for using Dice coefficient for each tissue rather than accuracy is that the number of voxels classified as a particular tissue may not be equal between the two segmentations, whereas the number of overall voxels in the brain (used for the accuracy) is.

4.5 Results

A large number of least-squares segmentations did not complete due to errors in the regression. This occurred exclusively in MRFs with multiple smoothing parameters and was due to the coefficient matrix of the system of equations being singular. These are shown in table 4.3. For this reason, we only include potentials estimated with least-squares if all subjects were successfully segmented in future results. All MPL segmentations completed successfully. This might be another reason that the full Potts MRF is not often used for tissue segmentation; to our knowledge, only the least-squares estimator has been proposed for MRF parameter estimation in tissue segmentation (without requiring extensive sampling as with MCMC approaches, or training data).

Table 4.3: Number of subjects (out of 18) successfully segmented using the least-squares estimator.

MRF	number successfully segmented
alpha-single-beta	18
single-beta	18
alpha-multi-beta	6
multi-beta	9

4.5.1 Model selection

We consider the value of the various MRFs using two metrics (table 4.4):

- pseudolikelihood information criterion (PLIC), and
- accuracy/Dice coefficient.

While the former is more appropriate for model selection, the latter is of more practical interest. The *highest* PLIC is desirable according to the definition in (4.9). We only consider the MPL-estimated methods, as not all LS methods completed successfully. In addition, the PLIC should be evaluated at the maximum-(pseudo)likelihood estimates, and the LS method does not attempt to maximise pseudolikelihood. For the same reason, the fixed-α_j MRFs were not included.

From table 4.4 it can be seen that the multi-beta model has the highest PLIC, followed by single-beta. It appears that addition of the α_j parameters is not called for, nor is fixing them to the log-class proportions (which reduces the number of parameters). Also, multiple smoothing parameters β_{jk} has higher PLIC than just one β.

On the other hand, if only the segmentation accuracy is considered (figure 4.1 and table 4.4), fixing the α_j to the log-class proportions produces the most accurate segmentations. Having the α_j parameters with no constraints is not beneficial. Additionally, a single smoothing parameter gives higher accuracy than multiple. In terms of specific tissue Dice scores, the same trend

Table 4.4: Model selection. Average PLIC evaluated at MPL estimates (left); average accuracy (right), decreasing.

MRF	PLIC
multi-beta	-8306583
single-beta	-8315326
alpha-multi-beta	-8342722
alpha-single-beta	-8364843
fixed-alpha-single-beta	-8618450
fixed-alpha-multi-beta	-8618656

MRF	accuracy
fixed-alpha-single-beta	0.819
fixed-alpha-multi-beta	0.819
single-beta	0.812
multi-beta	0.810
alpha-single-beta	0.807
alpha-multi-beta	0.802

noted in the previous chapter is also present here: gains in CSF and GM come at the expense of WM, and vice-versa.

Figure 4.2 shows the paired difference in accuracy for each MRF potential compared to the single-beta MRF. Overall, it appears that the only model worth using over the single-beta MRF are those with $\boldsymbol{\alpha}$ fixed to the tissue proportions. However, these two models also have extremely variable performance relative to the single-beta MRF in CSF in particular, ranging from 0.08 worse to almost 0.15 better (again, largely offset by the opposite trend in WM so that the overall gain in accuracy is much more moderate by comparison).

A mixed-effects model was fit to accuracy against MRF potential controlling for repeated subjects and is shown in table 4.5, showing significant differences in accuracy depending on the MRF used. Post-hoc comparisons with Tukey's method (table 4.6 shows the significant differences) showed that the fixed-$\boldsymbol{\alpha}$ methods performed significantly better than their counterparts with unconstrained $\boldsymbol{\alpha}$, while the unconstrained alpha-multi-beta performed significantly worse than single-beta.

Table 4.5: Mixed-effects model of segmentation accuracy for different MRFs using MPL, controlling for subject blocking.

	Sum Sq	Mean Sq	NumDF	DenDF	F	Pr(>F)
MRF	0.004	0.001	5	85.0	7.953	<0.001*

Table 4.6: Post-hoc pairwise comparisons for differences in accuracy using Tukey's method. Only significant differences are shown.

Comparison	Estimate	p
amb - fixed-amb	-0.016	<0.001*
amb - fixed-asb	-0.017	<0.001*
amb - sb	-0.010	0.032*
asb - fixed-amb	-0.012	0.008*
asb - fixed-asb	-0.012	0.007*

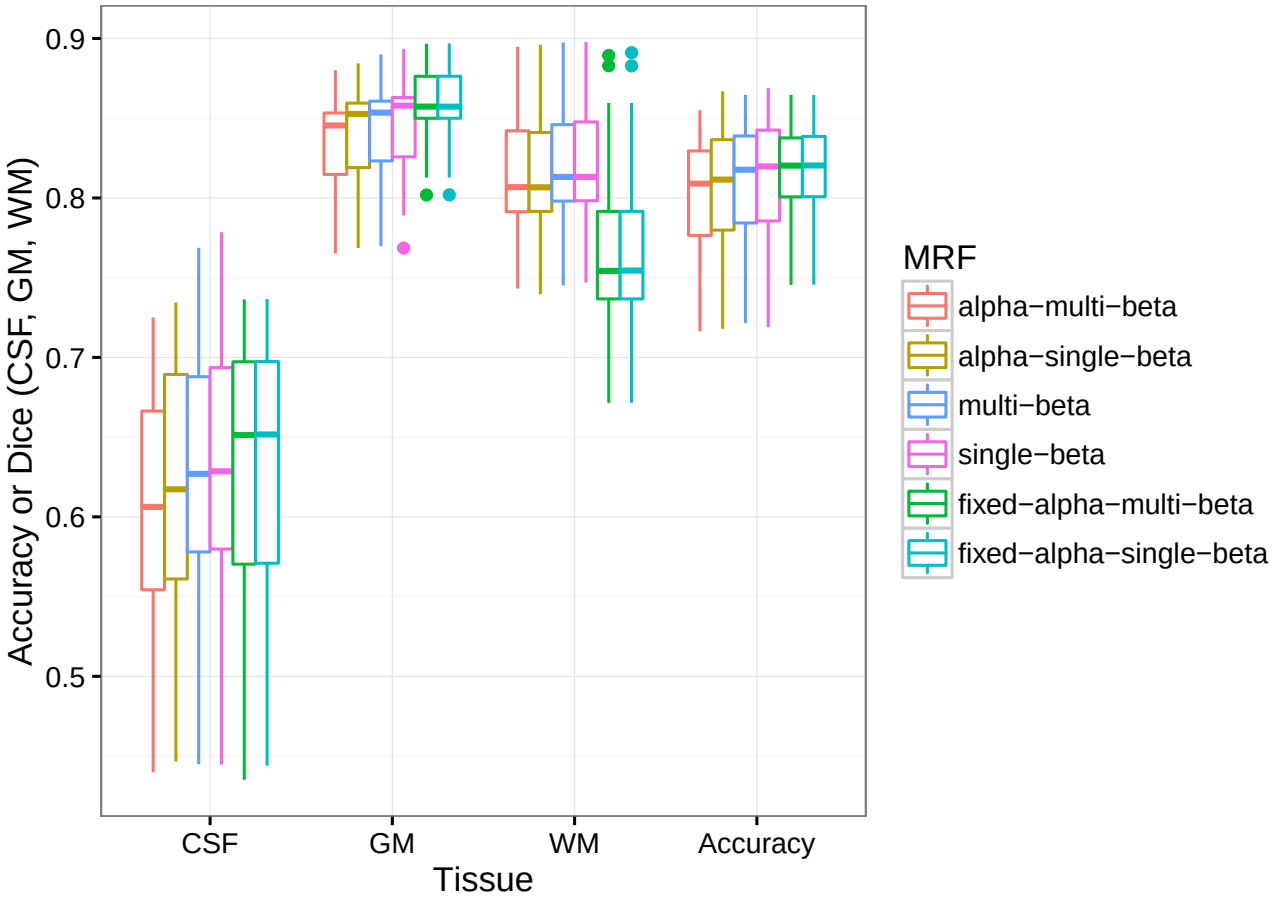


Figure 4.1: Segmentation metrics (accuracy or Dice coefficient) for the various MRFs using MPL.

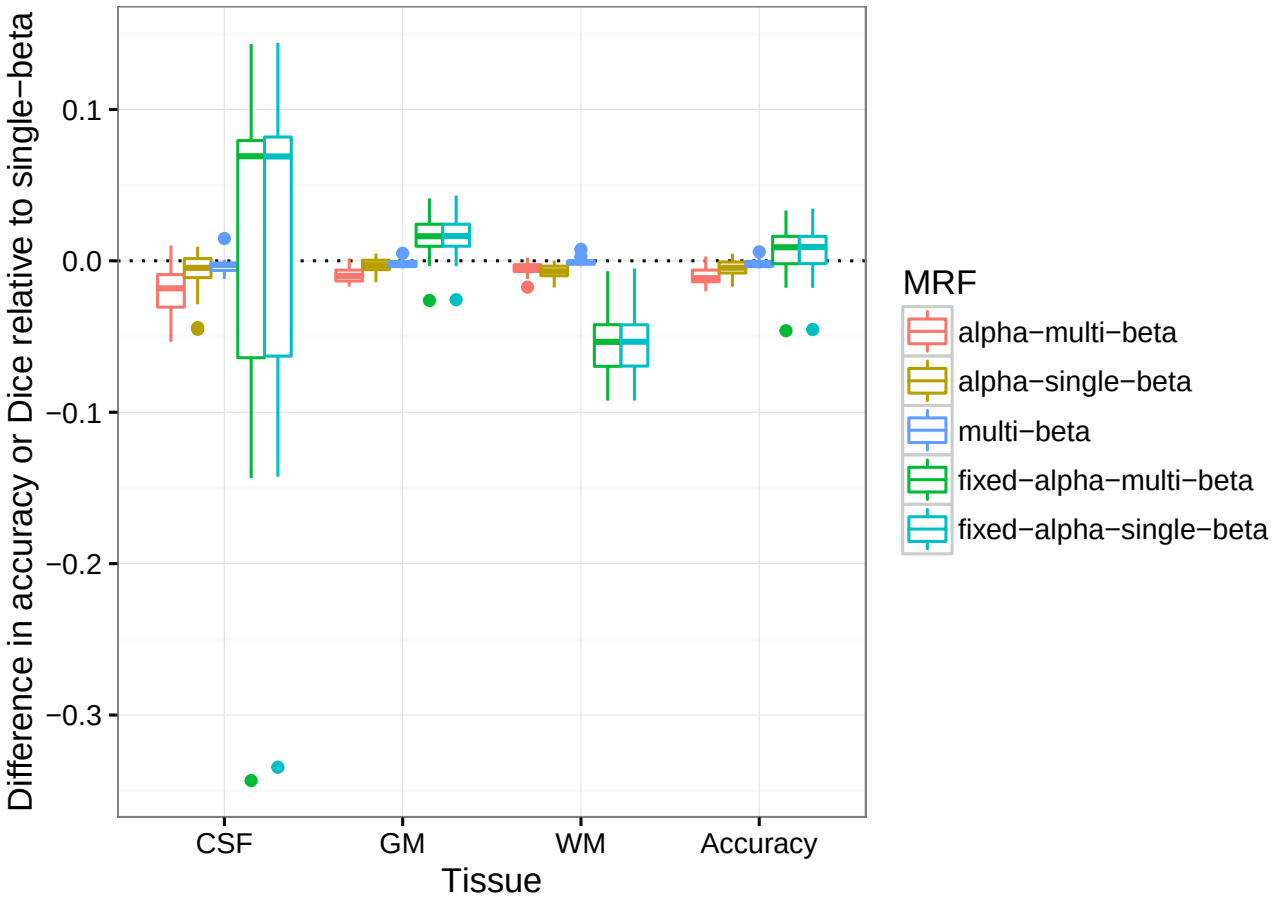


Figure 4.2: Paired differences in accuracy/Dice, relative to the single-beta MRF.

Table 4.7: Comparison of accuracy and Dice coefficient between LS and MPL for those MRF potentials where LS succeeded on every subject.

MRF	accuracy		Dice (CSF)		Dice (GM)		Dice (WM)	
	LS	MPL	LS	MPL	LS	MPL	LS	MPL
alpha-single-beta	**0.816**	0.807	**0.672**	0.616	**0.853**	0.839	0.779	**0.813**
single-beta	**0.813**	0.812	**0.628**	0.625	**0.843**	0.843	0.818	**0.820**

4.5.2 Comparison of estimators

It is difficult to compare the MPLE to the LSE due to the LSE often failing to find estimates for β_{jk}. Only comparisons on the single-beta and alpha-single-beta models may be made. Table 4.7 compares the accuracy and Dice coefficients for these MRFs. Overall, using the LSE produces segmentations with higher accuracy than with the MPLE, except in white matter. Interestingly, using MPL the single-beta model has the higher accuracy, while with the LSE the alpha-single-beta model has the higher accuracy.

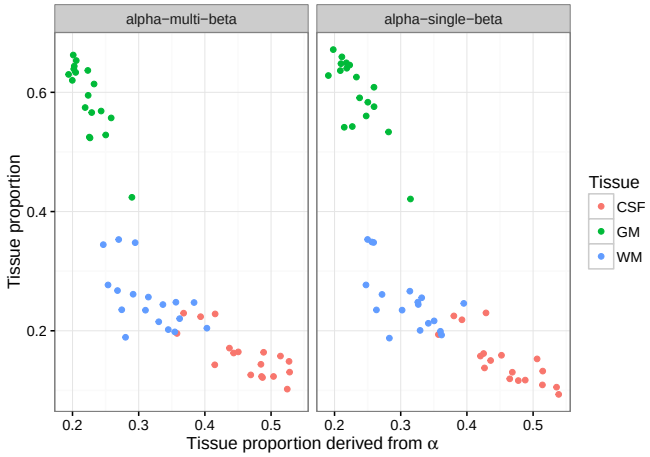


Figure 4.3: Tissue proportions compared to $\exp(\alpha_j)$ (normalised to sum to 1).

4.5.3 Parameter values

While the main aim is to explore the efficacy of the different MRFs with regards to segmentation accuracy, it is also of interest to examine the estimated parameter values.

First, we explore the validity of fixing α_j to the tissue proportions by examining whether the unconstrained models had a correspondence between $\exp(\alpha_j)$ and the tissue proportions. The tissue proportions from the final segmentation are plotted against the proportions derived from the final $\boldsymbol{\alpha}$ parameters in figure 4.3. The derived proportions are given by

$$\pi_j^* = \frac{\exp(\alpha_j)}{(\sum_{k=1}^{g} \exp(\alpha_k))},$$

that is, $\exp(\alpha_j)$ normalised to sum to 1 to account for $\alpha_1 = 0$. The tissue proportions from the segmentation are calculated by

$$\tilde{\pi}_j = \frac{\sum_{i=1}^{n} z_{ij}}{n}.$$

It appears that tissue proportion may have an inverse relationship with $\tilde{\pi}_j$. If the β parameters were omitted (set to 0), then the model would be a standard normal mixture no spatial dependence, and we would have $\tilde{\pi}_j = \pi_j^*$. Introduction of smoothing parameter(s) clearly breaks this correspondence.

Figure 4.4 show the β values for the different MRF potentials with multiple β_{jk}. In this figure β_{jk} have been converted to indicate the associated tissues (1=CSF; 2=GM; 3=WM), e.g. β_{12} applies to CSF and GM. The graph also displays the β values for the same model fitted with only a single β value. The MPL values only are shown, since the LS estimator could not always determine estimates for β_{jk}.

When α_j fixed to log-tissue proportions, the β and β_{jk} estimates are both higher and much more variable than when they are unconstrained. For all multi-beta models, β_{13} for the CSF-WM boundary is much higher than the other two β_{jk}, which remain close to the corresponding

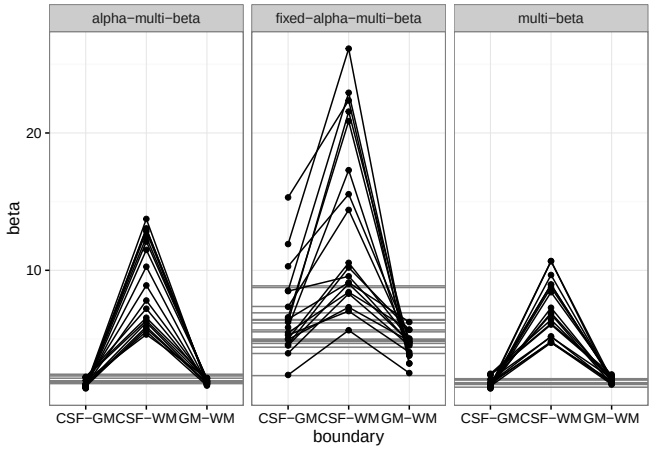


Figure 4.4: β_{jk} values estimated by MPL for various potentials; one line per subject. The β values for the corresponding model with only a single β are shown as horizontal lines.

single-β value. It appears that allowing for tissue-specific β_{jk} has allowed greater sensitivity for CSF-WM in particular.

4.6 Discussion

4.6.1 Model selection

Since the LSE often failed to find estimates for β_{jk}, we only consider the MPL segmentations in model selection. Figure 4.5 shows sample segmentations using MPL for various subjects. For all brains, we found relatively little difference between the MRFs within each subject, being mostly present in small isolated regions of tissue or thin features being smoothed or preserved. As in the previous chapter, the deep-grey matter regions were poorly segmented in all subjects due to lack of an anatomical atlas to inform the algorithm. GM was also oversmoothed in the cortical folds in most segmentations. Segmentations with α_j fixed to log-tissue proportions seem to have less CSF and more GM, which is oversmoothed, regardless of whether one β or multiple β_{jk} are used.

All segmentations suffer from the same partial-volume problem noted in the previous chapter: on the boundary of dark CSF and bright WM, a thin border of GM is retained. The intensity of such voxels is intermediate to CSF and WM, matching GM. The smoothing portion of the MRF should give these voxels much higher probability to be CSF or WM than GM due to few neighbours being GM. It appears that the corresponding smoothing parameters are not strong enough for the MRF probability to overcome the difference in intensity probability, and if they were stronger they would oversmooth the rest of the brain. Alternatively, the neighbourhood may not be large enough to capture a majority of CSF and WM neighbours as opposed to GM, particularly if the shell of GM surrounding the ventricle extends vertically (giving the top and bottom neighbours of the 6 available as GM).

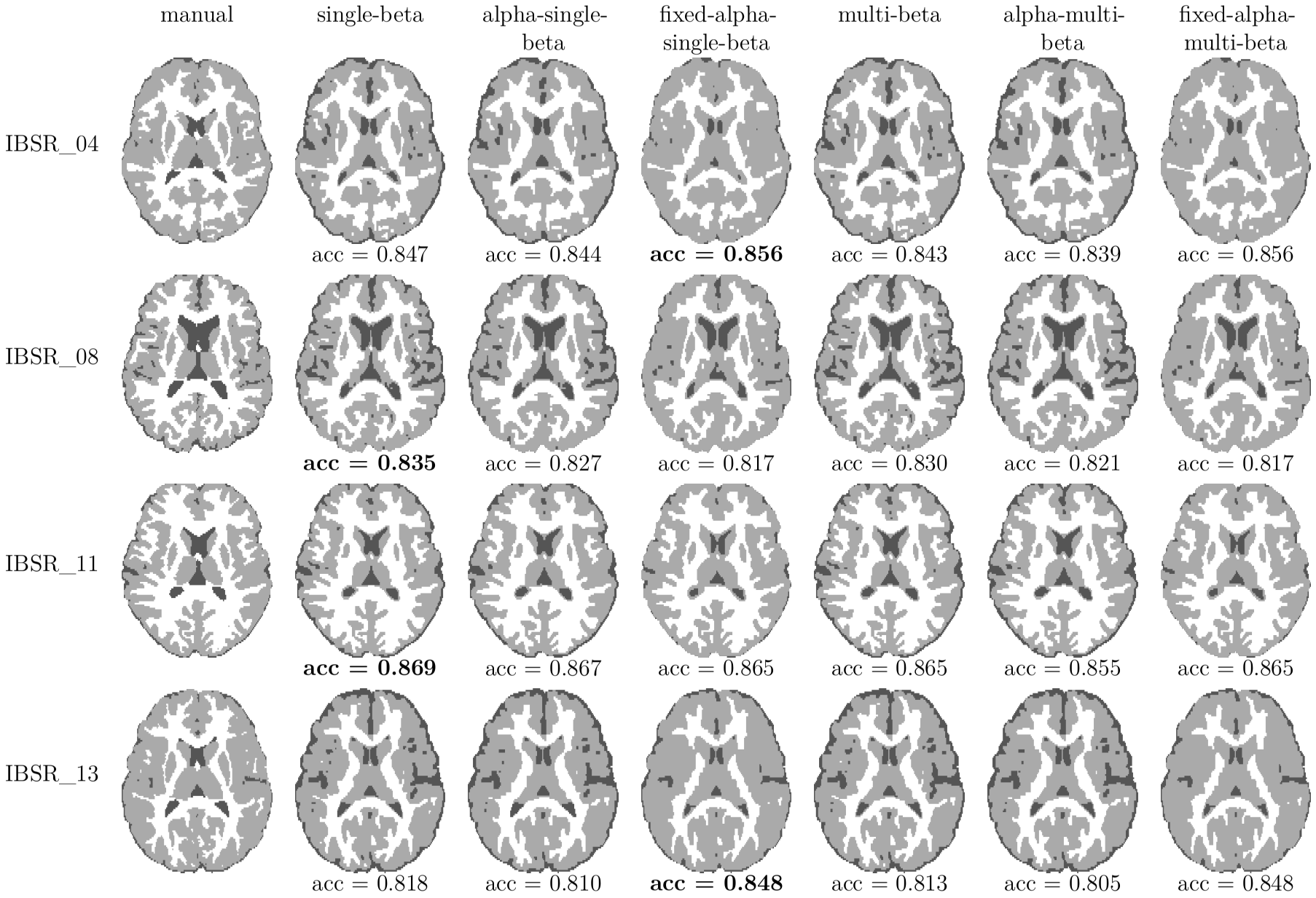


Figure 4.5: Example segmentations for different MRFs and subjects. Differences are mainly observed with less CSF for fixed-α_j MRFs, and in thin strips of CSF/WM for the others. Segmentation accuracy is displayed and in bold for the highest-accuracy MRF per subject.

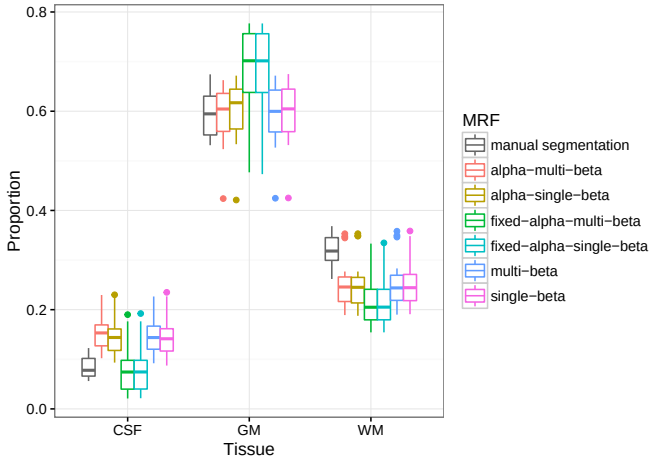


Figure 4.6: Tissue proportions for various MRFs compared to the manual segmentation.

4.6.1.1 Unary parameters

When selecting a model to use, a decision must be made as to the most important objective of the segmentation task: to achieve highest segmentation accuracy, or to favour models that are statistically justifiable.

PLIC favours a model without the unary parameters. In addition, if α_j must be used, they should not be fixed to the log-tissue proportions. This last is unsurprising as the log-tissue proportions are not the maximum pseudolikelihood estimates for α_j unless there is no smoothing component to the MRF, so the base pseudolikelihood of these models suffers and is not sufficiently offset by having fewer parameters. In fact, as evidenced by figure 4.3, unconstrained $\exp(\alpha_j)$ are not at all equivalent to tissue proportions and may even be inversely proportional to them. This is the opposite trend than if α_j were constrained or if there were no smoothing parameters.

Despite PLIC placing the fixed-α_j models as the worst options, these achieve the highest overall accuracy, though not significantly more than the single-beta model of the previous chapter. This highlights that pseudolikelihood and accuracy are different objective functions. Although the fixed-α_j models may not be statistically desirable, they may be a closer match to the underlying *physical* model.

Figure 4.6 shows the tissue proportions of the automatic segmentations compared to the manual segmentation. Models with constrained α_j have tissue proportions that are quite different to the other models. While all the MRFs with unconstrained parameters had too much CSF, the two models with constrained-$\boldsymbol{\alpha}$ were able to match the manual segmentation more closely. On the other hand, the GM and WM proportions were respectively much higher and lower than those of the manual segmentations and unconstrained models. In particular, the GM seems to be very oversmoothed in these segmentations as can be seen in figure 4.5. The fact that the fixed-α_j MRFs have higher overall accuracy overall compared to the other MRFs could be an artefact of the manual segmentations for the IBSR dataset labelling extra-sulcal CSF as GM (Valverde et al., 2015). Though fixing α_j achieved higher segmentation accuracy, too many fine

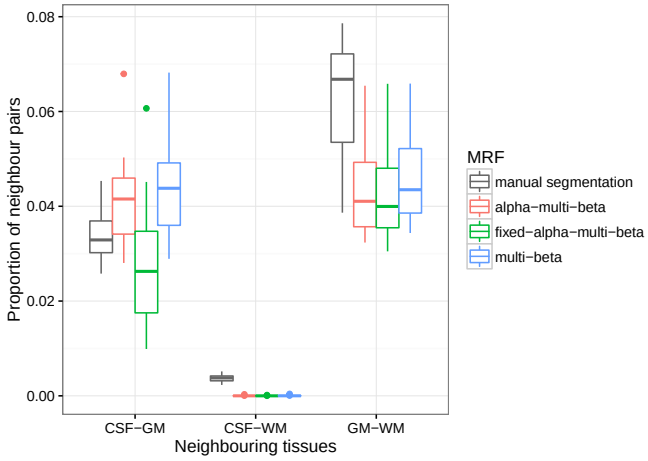


Figure 4.7: Proportion of neighbouring voxel pairs with different tissues for MRFs with multiple β_{jk} parameters.

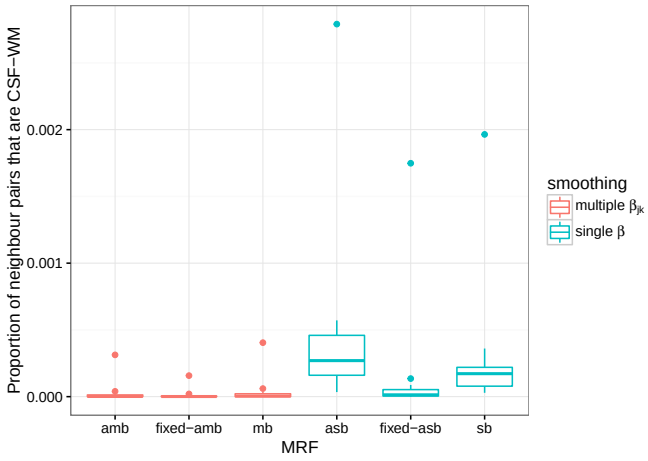


Figure 4.8: Proportion of (CSF, WM) neighbouring voxel pairs for each MRF potential

features are lost in the GM.

Overall, it does not appear worth including α_j parameters compared to the standard single-beta MRF. When unconstrained, these models achieve lower accuracy than the standard single-beta MRF, with the difference being significant for the alpha-multi-beta MRF. When constrained, the models achieve higher accuracy but not significantly, and visual inspection of the resulting segmentations reveals that GM is undesirably oversmoothed. While fixing α_j to log-class proportions can significantly change the resulting tissue proportions, it appears that they must be constrained using external anatomical knowledge (e.g. derived from an atlas as in Wels et al. (2011)) to be useful.

4.6.1.2 Smoothing parameters

In terms of whether to use a single smoothing parameter β or multiple β_{jk}, again PLIC suggests to use multiple, while segmentation accuracy suggests to use only one. We deduce that while having tissue-specific smoothing parameters induces a higher probability between the MRF

prior and the data (i.e. higher PLIC), this doesn't match the underlying anatomical model, resulting in a loss of accuracy.

When talking about β_{jk} for specific j and k, we remind the reader that we use the convention of 1=CSF, 2=GM, 3=WM (in increasing order of intensity). Examining the β_{jk} values in figure 4.4, we see that β_{12} (CSF-GM) and β_{23} (GM-WM) do not change much in value compared to if only a single β value had been used. However, β_{13} (CSF-WM) becomes much larger. It is also of interest to note that the proportion of neighbour pairs that are CSF and WM, is very low in all the resulting segmentations (figure 4.7). The likelihood for the full Potts MRF can be written

$$p(\boldsymbol{z}) \propto \exp(\sum_{j=1}^{g} \alpha_j n_j - \sum_{j \neq k} \beta_{jk} n_{jk}),$$

where $n_j = \sum_{i=1}^{n} z_{ij}$ is the number of voxels with label j, and $n_{jk} = \sum_{i=1}^{n} z_{ij} u_{ik}$ is the number of (distance-weighted) neighbouring voxel pairs with labels j and k.

It is unclear why a low n_{13} and high β_{13} should be associated with each other; a low (or zero) β_{13} would maximise $p(\boldsymbol{z})$ given it must be non-negative, in the absence of an intensity pdf. The inclusion of the intensity pdf must constrain the parameter values such that this does not occur.

Another possibility is that due to n_{13} being very small, the gradient of β_{13} on the Q-function is much lower than the other β_{jk} parameters, so its individual value operates on a different scale. Figure 4.7 shows the proportion of neighbour pairs in the image that consist of two different tissues. The proportion of CSF neighbouring WM (n_{13} normalised) is significantly lower than all the other pairs. The MRF probability (and Q function) will thus be relatively insensitive to the value of β_{13}, so the fact that it is so much larger than the other β_{jk} may not be practically meaningful.

Another consequence of a high β_{13} is that CSF and WM are prohibited from being in the same neighbourhood due to the large penalty. However, it is unclear whether a high β_{13} causes n_{13} to be low, is caused by n_{13} being low, or both. It is possible that since the E-step is approximate and very dependent on the current segmentation, there is an undesirable feedback loop between the C-step and M-steps. That is, since the current segmentation has a low occurrence of CSF and WM neighbouring each other, the estimated β_{13} is higher; but since β_{13} is higher, even fewer instances of CSF and WM occur in the next segmentation, which feeds into the next β_{13} and so on. This can be seen in figure 4.7, where the proportion of CSF-WM voxel pairs in the manual segmentation is much higher than in any of the automatic segmentations.

Recall the MRF of Roche and Forbes (2014) (4.4), which has a unary potential

$$\exp(-\boldsymbol{z}_i^T \boldsymbol{A} \boldsymbol{z}_i),$$

that can be largely interpreted in the same way as our pairwise potential, due to the way their $\boldsymbol{z}_i$ are formulated. The $\boldsymbol{A}$ matrix serves the same purpose as our $\boldsymbol{B}$ matrix. They fixed the

(a) Manual segmentation (b) LS segmentation

Figure 4.9: Example segmentation from which β_{jk} cannot be estimated using LS. CSF and WM never appear in the same orthogonal/6-neighbour neighbourhood, except when GM is in the centre.

$\boldsymbol{A}$ matrix to values learned from training data. They also found that the parameter in $\boldsymbol{A}$ corresponding to CSF and WM was very large. For them, this caused an artificial shell of GM around the ventricles, as we observe also. They mentioned that this might be addressed by introduction of a spatially-varying prior, presumably meaning an anatomical atlas.

When only one β parameter is used, the feedback problem is somewhat abated as β cannot be so specific. This can be seen in figure 4.8, which shows the proportion of (CSF, WM) voxel pairs by MRF potential. It can be seen that all MRFs with a single β parameter have more (CSF, WM) voxel pairs than those with multiple β_{jk} parameters. This is likely to be due to the high β_{13} in multi-beta models.

4.6.2 Comparison of estimators

The LSE often failed to find an estimate for the MRF parameters. This happened exclusively in models with multiple β_{jk} parameters. On closer examination of such a case, it was noted that the matrix of coefficients in the system of equations (4.10) was singular.

In particular, it was common for the column corresponding to β_{13} to be entirely 0, leaving this parameter unable to be estimated. This parameter corresponds to the CSF-WM boundary, which is not as common as the other boundaries in a human brain. Figure 4.9 shows the manual segmentation for one subject for which the multi-beta LS estimation failed, alongside the LS segmentation at that point. At this point in the segmentation procedure, there were very few neighbourhoods that had both CSF and WM present in the same neighbourhood, and all of these had GM in the centre. A typical example is the mis-classified thin strip of GM surrounding the ventricles between CSF and WM, arising as a consequence of partial volume effects.

To see how a scarcity of (CSF, WM) neighbours can cause its corresponding column of coefficients to be zero, consider the LS equations corresponding to the β_{jk} parameters for a neighbourhood

$\boldsymbol{z_{\partial i}}$ with $g = 3$:

$$\begin{pmatrix} -u_{i2} + u_{i1} & -u_{i3} & u_{i3} \\ -u_{i2} & -u_{i3} + u_{i1} & u_{i2} \\ -u_{i1} & u_{i1} & -u_{i3} + u_{i2} \end{pmatrix} \begin{pmatrix} \beta_{12} \\ \beta_{13} \\ \beta_{23} \end{pmatrix} = \begin{pmatrix} \log N(1, \boldsymbol{z_{\partial i}})/N(2, \boldsymbol{z_{\partial i}}) \\ \log N(1, \boldsymbol{z_{\partial i}})/N(3, \boldsymbol{z_{\partial i}}) \\ \log N(2, \boldsymbol{z_{\partial i}})/N(3, \boldsymbol{z_{\partial i}}) \end{pmatrix}.$$

Recall that 1=CSF, 2=GM and 3=WM. Since a given neighbourhood must occur with two *different* centre labels in order to be included in the system of equations, and all neighbourhoods with both CSF and WM only had GM in the centre, no such neighbourhoods were represented in the system of equations. Assume a neighbourhood is present in the system of equations. The first equation of the three possible requires that neighbourhood to appear at least once with CSF in the centre, and at least once with GM in the centre. The coefficient for β_{13} is $-u_{i3}$, i.e. the number of WM neighbours. However, since CSF does not appear with WM in the same neighbourhood unless GM is in the centre, we must have no WM neighbours ($u_{i3} = 0$) in order to have $N(1, \boldsymbol{z_{\partial i}}) = N(\text{CSF}, \boldsymbol{z_{\partial i}}) > 0$. This leaves a coefficient of 0 for β_{13}.

Likewise, the second equation requires the neighbourhood to appear with both CSF and WM in the centre, and the coefficient is $-u_{i3} + u_{i1}$. In order to have $N(\text{CSF}, \boldsymbol{z_{\partial i}}) > 0$ we must have no WM neighbours ($u_{i(\text{WM})} = u_{i3} = 0$) and in order to have $N(\text{WM}, \boldsymbol{z_{\partial i}}) > 0$ we must have no CSF neighbours ($u_{i(\text{CSF})} = u_{i1} = 0$), thus the coefficient of β_{13} is again 0. Similar reasoning shows that the coefficient of β_{13} in the third equation is also 0.

As a result, the matrix of equations is singular and β_{13} cannot be estimated. This could be avoided if the mean-field approximation were used to compute the u_{ij} since these are rarely 0; however, then it becomes unclear how to count $N(j, \boldsymbol{z_{\partial i}})$. It might also be avoided if the neighbourhood size was increased so that given tissue combinations occurred more frequently in the same neighbourhood, but this would also necessarily reduce the probability of a given $\boldsymbol{z_{\partial i}}$ occurring at least once with two different centre voxels.

Another alternative is to use the minimum mean-square error version of the LSE, derived by Borges (1999) for a binary MRF. This variant derives an estimate for the right-hand side of the LS equations, being (for a binary MRF with labels 0 and 1) $\log p/(1-p)$ where p is $p(1, \boldsymbol{z_{\partial i}})$ and $1 - p = p(0, \boldsymbol{z_{\partial i}})$. It is derived by minimising the square error of the estimate assuming that $N(1, \boldsymbol{z_{\partial i}})$ is binomially distributed for a given $\boldsymbol{z_{\partial i}}$ with parameter p. Since p is unknown, this error is integrated over the range of p which is assumed uniformly distributed on $[0, 1]$. The entire quantity is then minimised with respect to the estimate. Remarkably, the resulting equation has a closed-form solution and is computable even when the neighbourhood does not appear with a given centre tissue, or indeed when the neighbourhood is not present in the image at all. However, the estimator has not been extended to an MRF with more than two labels.

Overall, even though the LSE produced higher accuracy segmentations on the single-beta and alpha-single-beta models than MPL, we cannot recommend it due to this inherent instability. The MPLE does not suffer this problem, as it only requires a single *pair* of neighbours to be

CSF and WM in order for β_{13} to be present in the likelihood.

4.7 Conclusion

In this chapter we have studied different forms of the full Potts MRF for use as a prior in mixture-MRF brain MRI segmentation, as an alternative to the simplified single-beta Potts MRF of the previous chapter. The aim was to enable finer control of the smoothing applied by the MRF, by changing the single β parameter to multiple β_{jk} parameters that relate to the boundary between tissues j and k. Additionally, incorporation of spatially-independent unary parameters α_j that depend only a voxel's own label was thought to give the MRF flexibility to adjust for imbalanced tissue proportions. Alternatively, by constraining α_j to match the current log-tissue proportions, the MRF can be viewed as a tissue prior multiplied by a pairwise MRF for spatial regularity.

To separate the effects of the various parameters, we compared the single-beta MRF of the previous chapter to various forms of the full Potts MRF: with α_j enabled, disabled, or fixed to log-tissue proportions, and with a single smoothing parameter β or multiple tissue-specific smoothing parameters β_{jk}.

We extended the work of the previous chapter to show how maximum pseudolikelihood estimation could be applied to such models. The MPLE retains its desirable features: its Hessian is negative semi-definite, so any local maximum in $\boldsymbol{\Psi}$ for a given segmentation $\boldsymbol{z}$ is also a global maximum (though possibly not unique). Its gradient is available in closed form for ease of use with numerical optimisers. The objective function is the Q-function which has already been calculated as part of EM; it does not require additional computation to set up the optimisation. We also showed how to use the least-squares estimator to estimate the MRF parameters. This approach the same as that implemented by Van Leemput et al. (1999b), though with a slightly different MRF.

When comparing estimators, we found that the LSE often fails to form a solvable system of equations when multiple β_{jk} parameters are used. This was due to the very low occurrence of CSF and WM as neighbours with different centre voxel labels, so that the coefficient of the corresponding β_{13} was never non-zero, and the system was under-determined. By contrast, the MPLE only requires at least one occurrence of CSF and WM in the entire image in order to find an estimate for β_{13}. This extreme dependence of the LSE on the segmentation's neighbourhood profile makes less preferable to the MPLE. In future work, one option may be to use the minimum mean-square error variant of the LSE developed by Borges (1999), extending it to more than 2 tissue labels. Even then, the neighbourhood counts $N(j, \boldsymbol{z}_{\partial i})$ must be calculated which represents an extra computational burden over the MPLE. However, when the LSE *was* successful (for the single-beta and alpha-single-beta MRFs), it produced segmentations more accurate than MPLE, though not significantly so.

PLIC and segmentation accuracy were used to aid in model selection. PLIC favoured multiple β_{jk} over a single β and not to use any α_j parameters, particularly not fixed to log-tissue proportions. Segmentation accuracy recommended the opposite: constraining α_j and using a *single* β gave the highest accuracy. However, no MRF was significantly more accurate than the standard single-beta MRF.

Use of unconstrained α_j does not benefit the segmentation accuracy at all. Constraining α_j to log-tissue proportions produced segmentations with tissue proportions that were quite different to the other MRFs, with CSF more closely matching the manual segmentation, but much more GM. The resulting loss of definition of the cortical folds is such that use of constrained α_j is not worth the higher accuracy (which may be an artefact of the IBSR manual segmentations being oversmooth). Constraining α_j may still be of worth, but these should incorporate external prior knowledge of tissue proportions as in Wels et al. (2011) rather than being determined directly from the current segmentation.

A similar effect was noticed with the smoothing parameters. Although use of multiple β_{jk} parameters may be beneficial in allowing more specific smoothing, the reliance of the E-step on the current segmentation and subsequent feedback loop with the M-step bias them too much. In particular it appears that if a given tissue combination is relatively rare, this drives the corresponding ${\beta_{jk}}^{(t)}$ parameter higher, which further inhibits that tissue combination from appearing, which feeds back into the next ${\beta_{jk}}^{(t+1)}$ estimate. This was noticed along the CSF-WM boundary with β_{13} being very high compared to the other β_{jk} and preventing partial-volume voxels on this boundary from being classified as anything other than GM.

Overall, we believe that while multiple α_j and β_{jk} parameters can be beneficial, they must be constrained using prior and external anatomical knowledge in order to be most effective. That is, they should be used to impose constraints on the segmentation rather than allowing themselves to be driven by it. This may be one reason that papers such as (Maggia et al., 2016; Kabir et al., 2007; Menze et al., 2015) mention the full Potts MRF but only use the single-beta MRF. Incorporating such knowledge requires use of an anatomical atlas. For example, α_j can fixed to constant expected log-tissue proportions rather than being allowed to vary throughout as in Wels et al. (2011). Alternatively voxel-specific α_{ij} may be used to incorporate an anatomical atlas, being equivalent to multiplying the MRF by the atlas. Cardoso et al. (2011) demonstrated how β_{jk} constraints may be imposed on tissue pairs anatomically unlikely to neighbour each other by splitting each tissue into more specific sub-classes using an atlas.

In the absence of prior anatomical knowledge, or when an atlas cannot be used due to e.g. brain injury, the single-beta MRF is preferred for its simplicity and interpretability, and given no other MRF produced significantly more accurate segmentations.

Chapter 5

Anisotropic MRFs

5.1 Introduction

Using the Potts MRF prior for the tissue labels acts to smooth the segmentation. The forms of MRF considered up to now have a pairwise term that treats each neighbour of the same tissue label identically. For example, in the single-beta MRF, all neighbours of the same colour as the centre voxel contribute β/δ_{im} to that pixel's MRF potential. In this way, the MRF is *directionally isotropic* (up to a distance rescaling).

In previous chapters, it was found that the Potts MRF can sometimes oversmooth the image, with narrow sulci and gyri on the cortical surface being filled in. This happened regardless of whether a single or multiple smoothing parameters were used, but was exacerbated with the latter. While multiple smoothing parameters can enable finer control over specific tissue boundaries, they must be constrained using e.g. an anatomical atlas. When the parameters are unconstrained, the model is almost entirely driven by the data rather than imposing a constraint on it due to the use of an approximate E-step in the EM algorithm. We seek to compromise between these extremes, avoiding the problem of many unconstrained parameters while still allowing the smoothing to be variable across the image.

One solution is to only use a single smoothing parameter β, but weight it for each voxel pair to modulate the amount of smoothing. Rather than the weight depending only on the tissue classification of the neighbours as in the previous chapter, we allow it to incorporate of neighbour direction and other local image information. If the dependence of these weights on the image characteristics is defined generatively, no extra parameters need to be estimated. This avoids the extreme dependence of parameter values on the current segmentation noted with the multi-beta MRFs, while still allowing β to vary across the image.

In this chapter, we present a framework for incorporating local image characteristics into the MRF prior. In particular, we incorporate information about edge strength and orientation in the local neighbourhood in order to smooth along edges, but not across them. Our choice

of MRF potential is inspired by diffusion-based image smoothing and Perona-Malik diffusion (Perona and Malik, 1990).

We begin by stating the problem to motivate the desired properties of the MRF potential. A brief overview of anisotropic and Perona-Malik-based diffusion is given. This along with the desired properties lead naturally to a number of choices for anisotropic MRF prior. These MRFs are then utilised within the EM framework already presented to segment brain MRI. We compare the properties of the different anisotropic MRFs, as well as their performance against the single-beta MRF. We also incorporate parameter estimation.

The advantages of the work presented here are:

- we provide a generic framework to incorporate edge orientation and strength into an anisotropic MRF potential;
- the anisotropy is incorporated into the MRF itself rather than as a pre- or post-processing step;
- the smoothing parameter β is determined automatically through maximum pseudolikelihood estimation rather than fixed and tuned manually.

5.1.1 Aim

We seek to design an MRF potential that incorporates information about the strength and orientation of local image features (in this case, tissue boundaries or edges) and retains them rather than smoothing across them.

Recall that the image model consists of a mixture model, with each voxel's intensity y_i normally distributed given its label $\boldsymbol{z}_i$. The label $\boldsymbol{z}_i$ is distributed according to a Markov random field, dependent on the labels of its neighbours, $\boldsymbol{z}_{\partial \boldsymbol{i}}$.

$$\begin{aligned} Y_i|\boldsymbol{Z}_i = j &\sim \mathcal{N}(\mu_j, \sigma_j^2) \\ \boldsymbol{Z}_i|\boldsymbol{Z}_{\partial \boldsymbol{i}} &\sim \mathrm{MRF}(\boldsymbol{\Psi}) \end{aligned}$$

The conditional probability for a voxel to be a particular tissue is

$$p(\boldsymbol{z}_i|\boldsymbol{z}_{\partial \boldsymbol{i}}; \boldsymbol{\Psi}) \propto \exp(-U_i(\boldsymbol{z}_i|\boldsymbol{z}_{\partial \boldsymbol{i}}; \boldsymbol{\Psi})) \tag{5.1}$$

The potential U_i can consist of a unary term which depends only on $\boldsymbol{z}_i$ and is spatially independent, and a pairwise term depending on voxel pairs $(\boldsymbol{z}_i, \boldsymbol{z}_m)$, $m \in \boldsymbol{\partial i}$. The pairwise term is used to smooth the segmentation, while the unary term can be used to incorporate spatially-independent prior knowledge with the smoothing.

We will write the MRF potential as

$$U_i(\boldsymbol{z}_i|\boldsymbol{z}_{\partial \boldsymbol{i}}) = W_i(\boldsymbol{z}_i) + \sum_{m \in \partial \boldsymbol{i}} W_{im}(\boldsymbol{z}_i, \boldsymbol{z}_m),$$

where W_i is the unary term and W_{im} is the pairwise term. The single-beta MRF has

$$W_i(\boldsymbol{z}_i) = 0$$
$$W_{im}(\boldsymbol{z}_i, \boldsymbol{z}_m) = -\frac{1}{\delta_{im}} \beta \boldsymbol{z}_i^T \boldsymbol{z}_m.$$

The full Potts MRF has

$$W_i(\boldsymbol{z}_i) = -\boldsymbol{z}_i^T \alpha$$
$$W_{im}(\boldsymbol{z}_i, \boldsymbol{z}_m) = \frac{1}{\delta_{im}} \boldsymbol{z}_i^T \boldsymbol{B} \boldsymbol{z}_m.$$

In this chapter we aim to extend the pairwise term W_{im} to also incorporate local image edge strength and orientation. We will use the pairwise term of the single-beta model as a base due to its simplicity, modifying it to weight β for each neighbour to adjust the amount of smoothing applied. Because of this weighting, there is no need for tissue-specific parameters β_{jk}.

To this end, we allow the pairwise term to access image intensity, writing it as

$$W_{im}(\boldsymbol{z}_i, \boldsymbol{z}_m, y_i, y_m) = -\frac{1}{\delta_{im}} \beta w_{im}(\boldsymbol{z}_i, \boldsymbol{z}_m, y_i, y_m) \boldsymbol{z}_i^T \boldsymbol{z}_m, \tag{5.2}$$

where the function $w_{im}(\cdot) \in [0, 1]$ determines how strongly voxel m contributes to the conditional probability. If $w_{im} = 0$, voxel m does not influence the label of voxel i. If $w_{im} = 1$, voxel m contributes the same amount as the original single-beta model. Hence w_{im} should range from 1 for neighbours that are within the feature we wish to retain, to 0 for neighbours without.

We concern ourselves with the detection of edges in the brain (surfaces in 3D). A voxel deemed to be "across" an edge should not contribute to the local tissue majority, while one "along" or tangent to an edge should contribute as it originally did. This should help the MRF to distinguish whether a thin strip of tissue should be smoothed as noise, or preserved as a feature.

Suppose at a given pixel, an edge is detected with orientation $\boldsymbol{v} \in \mathbb{R}^d$, where d is the dimension of the image. Additionally, suppose we have some measure of the strength of the edge $\lambda \in [0, \infty)$, where $\lambda = 0$ means "no edge", and larger positive values represent how "strong" an edge is. Let $\boldsymbol{im}$ denote the vector from voxel i to voxel m. Also, given two vectors $\boldsymbol{a}$ and $\boldsymbol{b}$, let $\boldsymbol{a} \perp \boldsymbol{b}$ denote that $\boldsymbol{a}$ and $\boldsymbol{b}$ are perpendicular and $\boldsymbol{a} \parallel \boldsymbol{b}$ that they are parallel. The criteria stated so far translate to the following desired properties of w_{im}:

- If there is a strong edge, voxels along the edge should have weight 1 while those across

the edge should have weight 0:

$$\lambda \text{ “large”} \implies w_{im} \approx \begin{cases} 1, \boldsymbol{im} \parallel \boldsymbol{v} \\ 0, \boldsymbol{im} \perp \boldsymbol{v} \end{cases} \tag{P1}$$

- If there is no or a weak detected edge, standard isotropic smoothing should occur:

$$\lambda \text{ “small”} \implies w_{im} \approx 1 \ \forall m \in \boldsymbol{\partial i} \tag{P2}$$

The first property ensures that in regions where an edge is detected, only neighbours along the edge, as opposed to across it, count towards the neighbourhood tissue majority. The second property states that if there is no detected edge in the neighbourhood, then the standard single-beta MRF can be used.

5.2 Background

5.2.1 Image-based diffusion

In order to find w_{im} that weights neighbours anisotropically, downweighting those across edges, we draw inspiration from the field of diffusion-based image smoothing. We give a brief overview of Perona-Malik diffusion, only touching on aspects that will directly motivate the choice of w_{im}. For further details, see Perona and Malik (1990) and Weickert (1998).

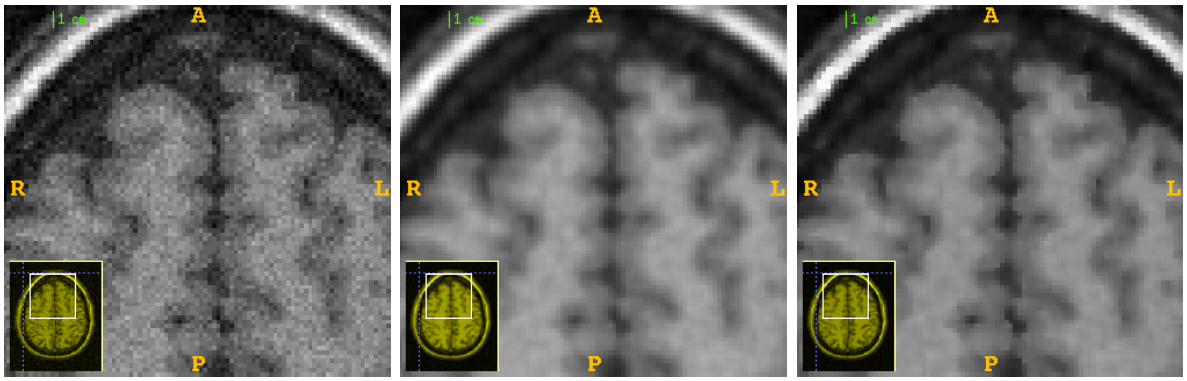
One way to smooth a noisy image is to convolve it with a Gaussian kernel. That is, the intensity at each pixel is replaced by the weighted sum of its own intensity and intensities in its neighbourhood, with the weights corresponding to a Gaussian or discretised version of it. Smoothing in this way is directionally isotropic since the Gaussian kernel is symmetric. This effectively smooths noise from images, but also smooths across image boundaries, blurring them.

Gaussian smoothing can be understood as a diffusion process on the image intensities. Let $I(\boldsymbol{x}, t)$ represent the image at time t and at spatial location $\boldsymbol{x}$. The MRI is observed at time 0, and we only observe the value of I at spatial locations corresponding to the pixel grid. That is, our vector of observed intensities $\boldsymbol{y}$ is such that $y_i = I(\boldsymbol{x}, 0)$ where $\boldsymbol{x}$ corresponds to the coordinates of voxel i. The following equation is the heat equation with the initial value equal to the observed image:

$$\begin{aligned} \partial_t I &= \operatorname{div}\left(\boldsymbol{D}\boldsymbol{\nabla} I\right), \\ I(\boldsymbol{x}, t = 0) &= \text{original image}, \end{aligned} \tag{5.3}$$

where the *diffusion tensor* $\boldsymbol{D}$ is a d-dimensional symmetric positive-definite matrix, and d is the dimension of the image. $\boldsymbol{D}$ itself may be a function of the pixel location and time.

When $\boldsymbol{D}$ is the identity matrix, (5.3) becomes the standard heat equation. The solution to



(a) Original image (b) Gaussian (c) Perona-Malik

Figure 5.1: Examples of isotropic (Gaussian) and anisotropic (Perona-Malik) image diffusion.

this equation at time t is the convolution of I with a Gaussian with diagonal covariance matrix $2t\mathbf{Id}$ where $\mathbf{Id}$ is the identity matrix. This smooths equally at all pixels and isotropically in all directions, including across edges.

5.2.2 Perona-Malik diffusion

The incorporation of anisotropic smoothing into the above framework was pioneered by Perona and Malik (Perona and Malik, 1990). They allowed $\boldsymbol{D}$ to vary with the pixel position $\boldsymbol{x}$, considering $\boldsymbol{D} = w(\boldsymbol{x}, t)\mathbf{Id}$. The *diffusivity function* $w(\boldsymbol{x}, t)$ provides the strength of smoothing to be performed at location $\boldsymbol{x}$. The idea is that w should approach 0 at or near image edges, and 1 away from edges. This corresponds to no smoothing at and near image edges, and isotropic smoothing away from edges, preventing them from being blurred. Figure 5.1 shows the difference between Gaussian (isotropic) smoothing and Perona-Malik (anisotropic) smoothing. Gaussian diffusion blurs all features including the edges (quite noticeable on the GM-CSF interface), while anisotropic diffusion preserves these while still blurring the homogeneous regions.

Perona and Malik proposed that w be a function of the magnitude of the image gradient at location $\boldsymbol{x}$, that is $w(\boldsymbol{x}, t) = w(|\boldsymbol{\nabla} I_t(\boldsymbol{x})|)$. Here $\boldsymbol{\nabla} I_t(\boldsymbol{x})$ is the gradient of the image after diffusion at time t, at location $\boldsymbol{x}$. We omit the dependence on t for clarity in what follows. The reason for using the image gradient is that sharp edges are typically recognised as a sudden change in image intensity at the edge. This corresponds to a large image gradient magnitude near an edge. The two functions proposed were:

$$\begin{aligned} w_{PM1}(|\boldsymbol{\nabla} I|) &= \exp(-(|\boldsymbol{\nabla} I|/\kappa)^2) \\ w_{PM2}(|\boldsymbol{\nabla} I|) &= \left(1 + (|\boldsymbol{\nabla} I|/\kappa)^2\right)^{-1}. \end{aligned} \tag{5.4}$$

The difference between the two is in the rate at which the diffusion coefficient decays to 0; it takes longer in the inverse version (see figure 5.2). In a homogeneous patch of the image, $|\boldsymbol{\nabla} I|$ approaches 0, as would be expected away from an edge. Both w_{PM1} and w_{PM2} approach 1 so

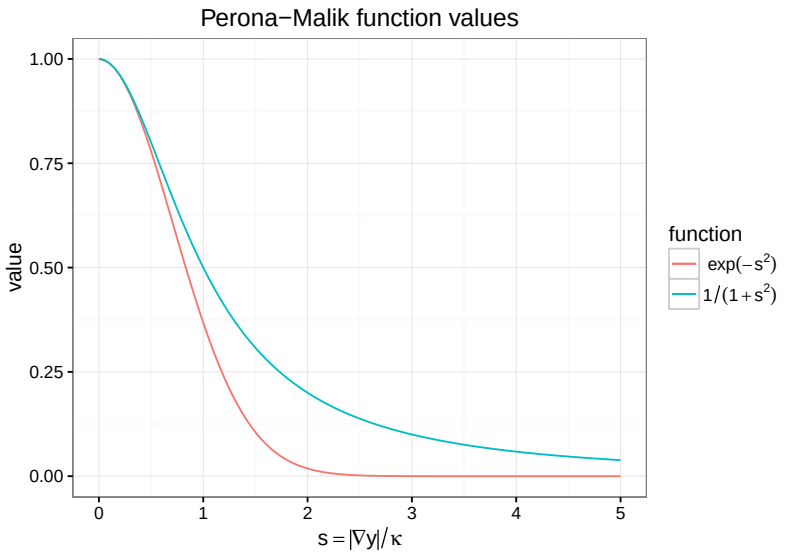


Figure 5.2: Comparison of Perona-Malik functions.

that isotropic smoothing is performed. Near an edge $|\boldsymbol{\nabla} I|$ is large and both functions approach 0, preventing smoothing.

It can be shown that Perona-Malik diffusion smooths edges with $|\boldsymbol{\nabla} I| < \kappa$, and *sharpens* edges with $|\boldsymbol{\nabla} I| > \kappa$. The parameter κ acts as a soft threshold of sorts, distinguishing gradient jumps due to noise from those due to true edges. The larger it is, the higher the image gradient must be in order to be preserved. Perona and Malik (1990) suggested to use the 90% quantile of $|\boldsymbol{\nabla} I|$ for κ. Black et al. (1998) suggested to use the median absolute deviation (MAD) of the gradient magnitude, which is a more robust, consistent estimator of the standard deviation (Rousseeuw and Leroy, 2005) (much like the median is to the mean). Boykov and Funka-Lea (2006) and Wels (2010) set it to σ_j, being the (current estimate of the) standard deviation of the intensities of tissue j.

Perona-Malik diffusion achieves the type of anisotropic smoothing outlined in the aims. A natural way to proceed is to set w_{im} to one of the Perona-Malik functions, where the gradient is along the direction $\boldsymbol{im}$.

5.2.3 Related work

There are a number of existing related works using MRFs that incorporate local image features. In general, these either apply the MRF separately after a standard mixture has been applied rather than incorporating it directly (Ward et al., 2017, Bériault et al. (2013)), or can be seen as examples of specific w_{im} but do not consider the design of the MRF in the same detail as presented here (Pagnozzi et al., 2015, Wels (2010)). Additionally, these works either omit β (Pagnozzi et al., 2015), or manually tune it (Wels, 2010), or set it by optimising segmentation accuracy on training data (Ward et al., 2017). In our case, we aim to develop methods that can be used in situations when no training data is available. There is also no reason to believe the

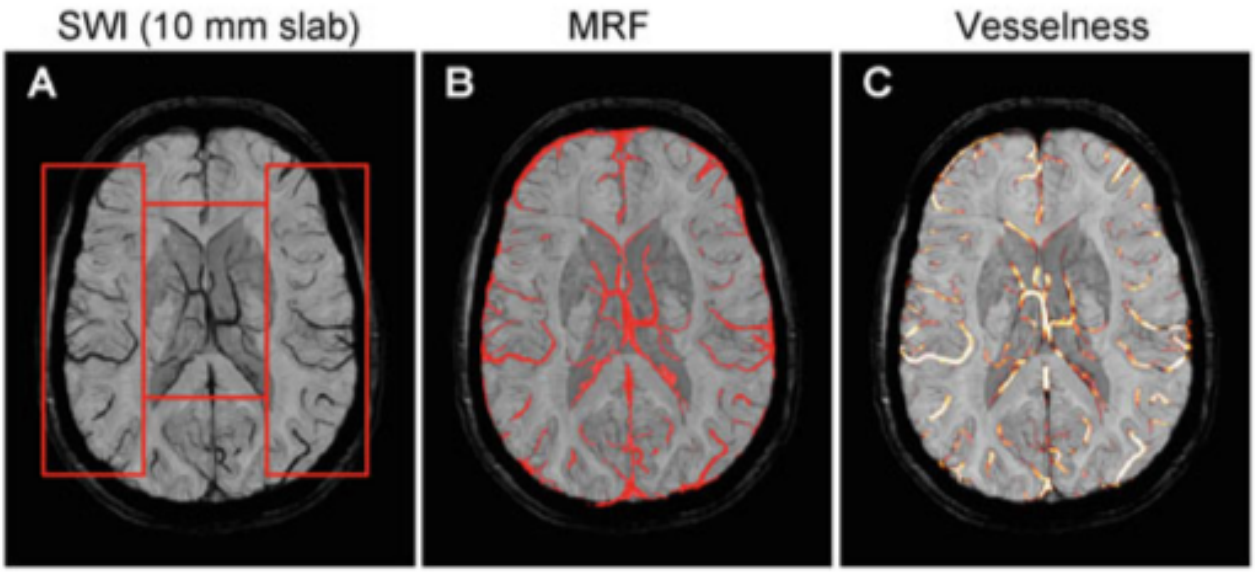


Figure 5.3: a) Example of a Susceptibility Weighted Image showing veins (dark intensity). b) Anisotropic MRF segmentation incorporating Frangi filter. c) Frangi's "vesselness" filter. From (Bériault et al., 2013).

same fixed value for β will work optimally for all images, as found in Chapter 3, nor what this value should be if so; hence, we estimate β.

5.2.3.1 Vein segmentation

An anisotropic MRF has been previously used in vein segmentation from brain MRI (Ward et al., 2017). Specific MRI sequences are used to obtain images showing blood vessels in the brain (figure 5.3). A standard (i.e. not with MRF) 2-component Gaussian mixture with components "vein" and "not vein" was fit to the intensities. After this, an anisotropic MRF was applied using Iterated Conditional Modes (Besag, 1986) on the resulting segmentation to refine it. The MRF made use of the Frangi filter (Frangi et al., 1998) which itself uses the smoothed image Hessian, effective at detecting tubular structures in 3D, lines in 2D. The Frangi filter produces a "vesselness" measure between 0 and 1, as well the orientation of the proposed vessel at each voxel. Ward et al. incorporated it into the MRF by setting w_{im} to be the dot product of neighbour m's direction from i with the vein direction. The "vesselness" strength itself was not used. Parameters of the model were estimated using cross-validation against segmentation accuracy on training data. The difference of this work to ours is that we incorporate the MRF directly into the mixture model. We also concentrate on the case where no training data is available, so the parameters must be estimated directly from the image being segmented.

Anisotropic MRFs have also been used to segment connective tissue in the optic nerve (Grau et al., 2006). Like brain vein segmentation, the MRF is designed to detect tubular features. It makes use of the image structure matrix (derived from eigen-analysis of the image gradient's outer product with itself) rather than the Frangi filter. This work is closely related to the work

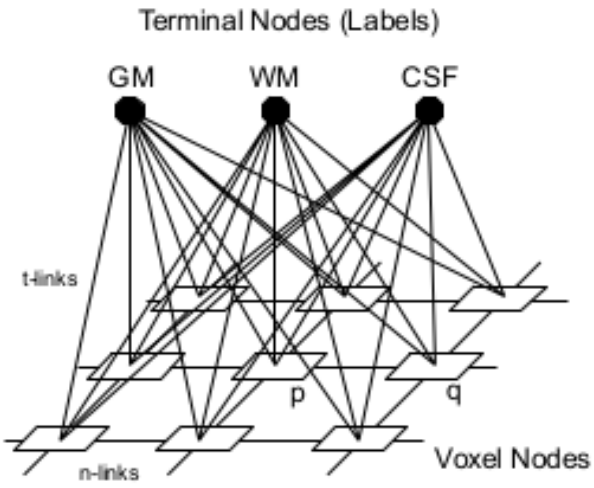


Figure 5.4: Example of graph-cuts formulation, showing voxels connected to their neighbours and 3 terminal nodes for the tissues. From (Song et al., 2006).

presented in this chapter, but is specific to detection of narrow tubular structures; our goal is instead to preserve edges. Additionally, the MRF parameter β is tuned manually.

5.2.3.2 Brain segmentation

In the field of brain MRI segmentation, Pagnozzi et al. (2015) used an MRF that used local image intensities as well as tissue labels. The MRF was designed to encourage neighbours of the same tissue to have similar intensities. A neighbouring voxel with a different tissue and different intensity to that being considered was not penalised. Conversely, a penalty would be applied for a neighbour of the *same* tissue but different intensity. This proved better than the standard single-beta MRF, though could unintentionally penalise noise voxels as the images were not pre-smoothed. The parameter β was not explicitly included in the MRF, so is equivalent to $\beta = \frac{1}{2}$ in the formulation given here.[1] A similar anisotropic MRF has also been used in Wels (2010) and Wels et al. (2011), with the segmentation found using graph-cut segmentation.

Cardoso et al. (2011) used the multi-beta Potts MRF with constant $\beta = 0.5$ or 3 depending on the particular tissues being considered. This β was further weighted by a per-voxel value between 0 and 1 corresponding to whether the voxel was in the sulci and gyri of the brain or not. The result was that MRF smoothing was weakened in these regions of the brain, which are prone to oversmoothing. While this weighting is anisotropic across the image - it varies at each voxel location, at a given voxel - the weight does not vary for each neighbour of a given voxel.

5.2.3.3 Graph-cut segmentation

The idea of anisotropically weighting neighbours has been used in the graph-cuts approach to image segmentation. Graph cuts minimises an energy over the image, rather than an a probability model. However, the two are closely linked. The voxels of the image are represented as nodes on a weighted graph, width edges between the nodes of neighbouring voxels. In addition, three "hidden" terminal nodes are added - one for each tissue (GM, WM, CSF). All voxels are connected by an edge to each of these (see figure 5.4). The weight on the edges connecting the voxel nodes to the tissue nodes ("t-links" in figure 5.4) is an intensity fidelity term, for example the conditional intensity probability $-\log f(y_i|\boldsymbol{z}_i)$. The weight on the edges between voxels ("n-links") is associated with the log-MRF probability. Graph cut segmentation aims to 'cut' edges such the total cost of cut edge weights is minimised, while ensuring that every voxel node is connected to exactly one tissue node, and that no edges exist between voxel nodes connected to different tissues. The segmentation is retrieved by associating each voxel with the tissue node it is connected to.

Incorporation of anisotropic edge weights was done in (Song et al., 2006; Boykov and Funka-Lea, 2006). In Boykov and Funka-Lea (2006) the edge weights comprised a penalty on same-labelled neighbours with very different intensities. In Song et al. (2006) the edge weights had a similar intensity penalty based on the Lorentzian error norm. Additionally, they incorporated an edge probability based on the image gradient between the neighbours (Malik et al., 2001). These weights may be used in an MRF by setting the MRF potential to the edge weights. In particular our derivation from the perspective of anisotropic smoothing yields the same pairwise term as Boykov and Funka-Lea (2006) as one of its implementations.

One advantage of using a probability model rather than graph cuts is that the former is generative, and provides posterior probabilities rather than just hard classifications. The corresponding β or smoothing parameter in a graph cuts model weights the t-link and n-link terms against each other. It must typically be set manually or with use of training data, whereas the maximum-likelihood provides a natural framework for parameter estimation and can be used when training data is not available. On the other hand, as graph-cut segmentation focuses on minimising an energy, there are no intractable normalising constants; the MRF potential is used as-is as an n-link weight.

Wels (Wels, 2010, chapter 2; Wels et al., 2011) used the same pairwise cost as Boykov and Funka-Lea (2006) and incorporated it into an MRF, though did not explore it in the setting of anisotropic diffusion, or consider related anisotropy-motivated MRFs. This pairwise cost was chosen to enforce intensity homogeneity within tissue classes; we will derive the same term in one of our examples by considering anisotropic smoothing. They also had a per-voxel term in the MRF potential (akin to the α term in the multi-beta MRF previously presented), but

[1]Omitting β is equivalent to setting $\beta = 1$, but the pairwise potential in (Pagnozzi et al., 2015) is $\frac{1}{2}W_{im}(\boldsymbol{z}_i, \boldsymbol{z}_m, y_i, y_m)$ with the 1/2 to counteract neighbours being double-counted; we have absorbed the 1/2 into our β.

these were trained from a number of manually-labelled images or atlases. They did not explore estimation of β but rather set it to $\beta = 0.6$ in (5.2).

To re-emphasise, the difference between our method and these is our estimation of β, along with demonstration of how to modify the Potts MRF to incorporate neighbour-specific weights designed to preserve edges.

5.3 Method

The MRF takes the form

$$p(\boldsymbol{z}_i|\boldsymbol{z}_{\partial \boldsymbol{i}}; \boldsymbol{\Psi}) \propto \exp(W_i(\boldsymbol{z}_i) + \beta \sum_{m \in \partial \boldsymbol{i}} \frac{1}{\delta_{im}} w_{im} \boldsymbol{z}_i^T \boldsymbol{z}_m),$$

where w_{im} is the anisotropic weight function. In the previous chapter, we found that including unary potentials of the form $W_i(\boldsymbol{z}_i) = \boldsymbol{z}_i^T \alpha$ was not helpful unless α were constrained using external information. For example, Wels et al. (2011) set α_j to log-tissue proportions learned from an external atlas. We found that doing the same but using log-tissue proportions from the current segmentation had slightly higher accuracy than when they were omitted, but on visual inspection of the algorithms was seen to cause severe oversmoothing of grey matter. Therefore in this chapter, we omit unary potentials and set $W_i(\boldsymbol{z}_i) = 0$.

5.3.1 Choice of weight function

We incorporate ideas from anisotropic diffusion into the design of an anisotropic MRF potential by using the Perona-Malik w_{im} functions to weight each neighbour's contribution to the smoothing. Suppose at voxel i an edge is detected with direction $\boldsymbol{v} \in \mathbb{R}^d$ and strength $\lambda \in [0, \infty)$, where $\lambda = 0$ means "no edge", and larger positive values represent how "strong" an edge is. We seek a weight between neighbours i and m, $w_{im} \in [0, 1]$, such that

$$\lambda \text{ "large"} \implies w_{im} \approx \begin{cases} 1, \boldsymbol{im} \parallel \boldsymbol{v} \\ 0, \boldsymbol{im} \perp \boldsymbol{v} \end{cases} \tag{P1}$$

$$\lambda \text{ "small"} \implies w_{im} \approx 1 \; \forall m \in \partial \boldsymbol{i} \tag{P2}$$

A natural choice for λ is one of the Perona-Malik functions w_{PM1} or w_{PM2}, which we denote w_{PM}. The Perona-Malik functions determine the edge direction to be orthogonal to the local image gradient, i.e. $\boldsymbol{v}$ is such that $\boldsymbol{v} \cdot \boldsymbol{\nabla} I = 0$. We will do likewise, except that we use $\boldsymbol{\nabla} I^s$ in place of $\boldsymbol{\nabla} I$, being the gradient of the image after pre-smoothing with a Gaussian of standard deviation s. Pre-smoothing is required to regularise the image gradient, which is otherwise ill-posed; for further information, see (Weickert, 1998, Section 1.3). A small s is chosen, in order

to perform a weak smoothing: enough to soften isolated voxels of noise while not affecting edge sharpness overly much. We use $s = 1$ (a 1-voxel radius).

The Perona-Malik functions both approach 1 as the gradient (edge strength) approaches 0, satisfying property (P2). To achieve directional anisotropy in (P1), the edge strength, edge direction, and neighbour direction must be combined. Three MRF potentials are presented below.

Isotropic Perona-Malik ($\mathbf{PM_{iso}}$):

$$w_{im} = w_{PM}(|\boldsymbol{\nabla} I_i^s|), \tag{5.5}$$

where $|\boldsymbol{\nabla} I_i^s|$ indicates the magnitude of the image gradient evaluated at voxel i. This is natural approach of setting $w_{im} = w_{PM}$ directly and is most like the original Perona-Malik diffusion. w_{im} varies with voxels i, but for a given i is the same for all neighbours m. As with Perona-Malik diffusion, this process is anisotropic in that the diffusion coefficients vary at each voxel location, but directionally isotropic as each neighbour at that location is treated equally. It is only the *strength* of diffusion that varies at each voxel. Thus, this MRF does not satisfy (P1).

In effect, at each voxel the MRF becomes

$$p(\boldsymbol{z}_i|\boldsymbol{z}_{\partial \boldsymbol{i}}; \boldsymbol{y}_{\partial \boldsymbol{i}}) \propto \exp(\tilde{\beta}_i \sum_{m \in \partial \boldsymbol{i}} \boldsymbol{z}_i^T \boldsymbol{z}_m),$$

where $\tilde{\beta}_i = \beta w_{PM}(|\boldsymbol{\nabla} I_i^s|)$. $\mathrm{PM}_{\mathrm{iso}}$ is akin to the single-beta MRF, where the β may change at each voxel but not for each neighbour within that voxel's neighbourhood. Although this MRF does not satisfy (P1), we will use it to compare the difference between incorporating directional anisotropy and not. To satisfy (P2), it is desirable to incorporate "true" directional anisotropy into the diffusion.

Anisotropic Perona-Malik using the directional derivative ($\mathbf{PM}_{|\boldsymbol{\nabla} I^s \cdot \hat{\boldsymbol{im}}|}$):

$$w_{im} = w_{PM}(|\boldsymbol{\nabla} I^s \cdot \hat{\boldsymbol{im}}|), \tag{5.6}$$

where $\hat{\boldsymbol{im}}$ is the unit vector pointing from voxel i to voxel m. This attempts to incorporate edge orientation by replacing the image gradient with the directional image gradient. If $\boldsymbol{\nabla} I^s \cdot \hat{\boldsymbol{im}} = 0$, then m lies along an edge, and $w_{im} = 1$ as requested by (P1). If m happens to lie directly in the direction of the gradient, then it lies across the edge ($\boldsymbol{im} \perp \boldsymbol{v}$), and $w_{im} = w_{PM}(|\boldsymbol{\nabla} I^s|)$. This is approximately 1 when the gradient/edge is weak ($|\boldsymbol{\nabla} I^s| \approx 0$), satisfying (P2). If the edge is strong, $|\boldsymbol{\nabla} I^s|$ is large and w_{PM} approaches 0. This prohibits that neighbour from counting towards the neighbourhood majority tissue, satisfying (P1).

One disadvantage of $\mathrm{PM}_{|\boldsymbol{\nabla} I^s \cdot \hat{\boldsymbol{im}}|}$ is that it is symmetric. Suppose voxel i lies on an edge between GM and WM, and its colour indicates it is likely to be GM (see figure 5.5). Neighbour m_1, nominally WM, has very little contribution to the MRF due to it being almost perpendicular

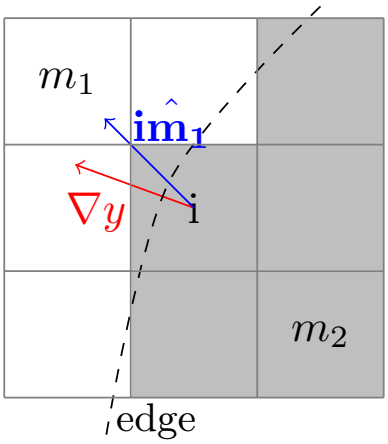


Figure 5.5: Neighbourhood of voxel i showing an intensity edge and the orientation of the gradient ($\boldsymbol{\nabla} I^s$) with two of the neighbours labelled.

to the edge direction, as desired. On the other hand, neither does neighbour m_2 (nominally GM) despite it being on the "same side" of the edge as pixel i, since its direction is also nearly perpendicular to the edge.

Anisotropic Perona-Malik using intensity differences ($\mathbf{PM}_{|y^s_m - y^s_i|}$):

$$w_{im} = w_{PM}(|y^s_m - y^s_i|), \tag{5.7}$$

where y^s_i denotes the ith pixel of $\boldsymbol{y}^s$ being the image pre-smoothed by a Gaussian kernel of size s (again, this is for gradient regularisation). This alternative approximates the directional image gradient by a forward difference. Perona and Malik (1990) suggested this form as a natural discretisation of their diffusion problem. Since this is a forward difference, it is not symmetric. In figure 5.5, $|y^s_{m_1} - y^s_i|$ is large, so its contribution to the neighbourhood tissue label majority is close to 0. On the other hand, $|y^s_{m_2} - y^s_i|$ is small, so it retains close to its usual contribution as the single-beta MRF.

General remarks

We choose the exponential form w_{PM1}

$$w_{PM1} = \exp(-\left(\frac{|y^s_m - y^s_i|}{\kappa}\right)^2),$$

as this is the same pairwise edge weight used in the graph-cut formulation of (Boykov and Funka-Lea, 2006, Wels (2010)). The potential used in Pagnozzi et al. (2015) is similar, but the argument is not squared. Their motivation was to encourage homogeneity of intensities within each tissue class. It is interesting to see that the different motivations - to smooth anisotropically along edges, and to ensure intensity homogeneity within tissues - lead to the same MRF potential. The exponential and inverse forms of the Perona-Malik functions are quite similar in behaviour; either could be used, and the same comments apply to both.

It is also worth noting that in Perona and Malik's original paper (Perona and Malik, 1990, Section VI-B), the authors relate their diffusion model to finding maxima with respect to the image intensities, of a Markov random field over the image intensities, with fixed parameters (namely, only κ). Perona and Malik show that there exists an energy function such that minimising it by gradient descent with respect to I is equivalent to anisotropic diffusion, though the form of the energy function itself is unknown. Their model does not have tissue labels as in our segmentation problem, as the aim is to adjust image intensities, rather than use the intensities to reduce the image to a very small number of tissue classes.

Finally, we do not give these choices of w_{im} as the *only* options, or even 'optimal' ones; they are natural suggestions arising from consideration Perona-Malik diffusion. A strong alternative would be to use concepts from anisotropic diffusion in the style of Weickert (Weickert, 1998), where the diffusion matrix $\boldsymbol{D}$ in (5.3) is not a scalar multiple of the identity matrix. Instead, it is derived from the eigenvalues and eigenvectors of the *image structure tensor* $\boldsymbol{\nabla} I^T \boldsymbol{\nabla} I$. The structure tensor is a more sophisticated feature detector than the image gradient, as it is able to distinguish between (e.g.) tubular structures vs. surfaces in three dimensions. For an example of use of the structure matrix with Weickert's coherence-enhancing filter in an anisotropic MRF to segment tubular structures in the eye, see (Grau et al., 2006). Alternatively, other measures of edge strength could be used. For example, the edge probability of (Malik et al., 2001) could be incorporated as in (Song et al., 2006).

5.3.2 Parameter estimation

The MRF has two parameters: κ, and β. They are identifiable and serve different purposes: κ is an intensity normalisation parameter that applies to $\boldsymbol{y}$, while β is a spatial regularisation parameter corresponding to the labels $\boldsymbol{z}$ and balancing the intensity probabilities with the spatial probabilities.

As previously mentioned, existing literature fixes the value of β, either by using training data, or arbitrarily (Pagnozzi et al., 2015, Wels (2010), Ward et al. (2017)). Alternatively, omitting it, replacing it entirely by w_{im}, sets $\beta = 1$. Instead, we estimate β, as there is no principled or well-reasoned value to fix it to in absence of training data. For the single-beta MRF, fixing $\beta = 1$ proved less effective than estimating it, and estimated values were found to be generally greater than 1, implying a larger amount of smoothing is needed. Once the weights w_{im} are incorporated, the average neighbour counts $\tilde{u_{ij}}$ will generally be smaller than the corresponding single-beta u_{ij}, since the weights are bounded by 0 and 1 and usually less than 1 unless the image gradient is exactly 0. Therefore, if $\beta = 1$ was too small for the single-beta MRF, we presume it will be here also. This justifies our choice to estimate β.

We use maximum pseudolikelihood to estimate β. As the MRF is still log-linear in β, it is concave with respect to β for fixed $\hat{\boldsymbol{z}}$ and mixture parameters. We do not consider the least-squares

estimator as we have in previous chapters, as it cannot be adapted for the weights w_{im}.

The Perona-Malik functions both require a parameter κ to be specified. In this framework, the parameter κ may be interpreted as an intensity normalisation parameter. It also determines the rate at which the weights decay to 0 (see figure 5.2), with a larger value corresponding to a slower decay. Boykov and Funka-Lea (2006) and Wels (2010) set it to $\hat{\sigma}_j$, being the current estimate of the standard deviation of tissue j's intensities. Perona and Malik (1990) suggested to set κ to the 90% quantile of $|\boldsymbol{\nabla} I|$. Black et al. (1998) suggested the median absolute deviation, a robust estimate of the standard deviation of the gradient magnitude. Let ∂_i be the gradient magnitude of the image $|\boldsymbol{\nabla} I|$ at voxel i, and $\boldsymbol{\partial} = (\partial_1, \ldots, \partial_n)$ be the length-n vector of gradient magnitudes evaluated at each voxel. The median absolution deviation (MAD) is

$$\text{MAD}(\boldsymbol{\partial}) = 1.4826\text{median}(|\partial_i - \text{median}(\boldsymbol{\partial})|), \tag{5.8}$$

where the factor of 1.4826 is derived from the quantile function of the standard normal distribution and is such that the MAD is a consistent estimator for the standard deviation. We followed Black et al. (1998) and used the MAD as κ. We found that using the 90% quantile of $|\boldsymbol{\nabla} I|$ was too high, such that $w_{PM}(|\boldsymbol{\nabla} I|/\kappa)$ (and more so the other weight functions) was too insensitive, varying very little across the image and thus acting as an isotropic weight.

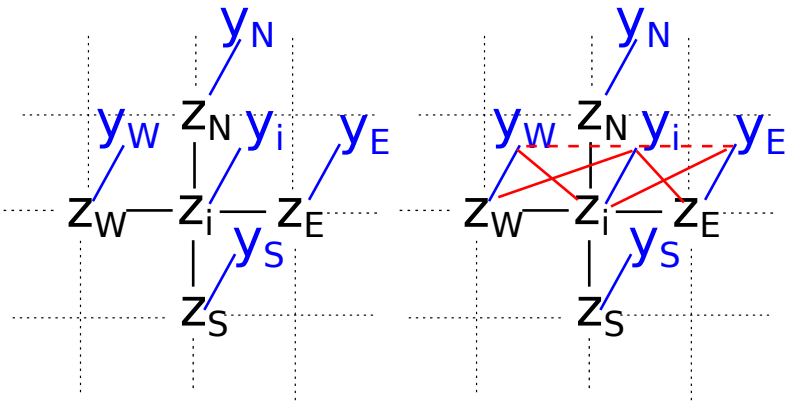
5.3.3 Limitations

There are two limitations to this work that must be acknowledged. First, the parameter κ forms part of the MRF parameters $\boldsymbol{\Psi}$, yet we do not estimate it; rather, we fix it. Second, allowing w_{im} and hence the MRF to depend on the intensities y_i and y_m means we can no longer view the MRF as $p(\boldsymbol{z}_i|\boldsymbol{z}_{\partial i})$.

5.3.3.1 Intensity normalisation parameter κ

For these experiments, the intensity normalisation parameter κ was chosen to be the MAD of the gradient magnitude in the image, as per (Black et al., 1998). Lower values correspond to a faster decay of the neighbour weights towards zero as the intensity difference increases; higher values slow the decay. Thus a lower value of κ penalises the difference in intensity more heavily than a higher value, and causes the MRF to be more sensitive to the image gradient. On the other hand, if it is too low, too much noise will be preserved.

It could be argued that κ should be estimated, as β is, using e.g. maximum pseudolikelihood. The two parameters are identifiable: while they both control smoothing, κ smooths image intensities (gradients) while β acts on tissue labels. However, due to the nonlinearity of κ in the MRF potential, the problem is no longer concave should κ be estimated. Also, it was thought that κ should not be allowed to vary throughout the iterations, as the intensities do not vary



(a) Using the single-beta MRF (b) Using the anisotropic MRF

Figure 5.6: The underlying dependence graph for mixture-MRF segmentation. In the anisotropic MRF, the additional solid red edges are added (only shown for the east and west neighbours). See the text for further details.

throughout the iterations. Use of MPL would allow κ to change as the current segmentation changed, and possibly feed back into the next iteration's segmentation undesirably. Rather, we think it should be viewed as an intensity normalisation constant than a parameter of the MRF. Nevertheless, estimation of κ alongside β is worth further investigation.

5.3.3.2 Introduction of intensities into the MRF prior

All our choices of w_{im} depend on y_i and $\boldsymbol{y}_{\partial \boldsymbol{i}}$, where $\boldsymbol{y}_{\partial \boldsymbol{i}}$ are the intensities at voxels neighbouring voxel i, i.e. y_m such that $m \in \boldsymbol{\partial i}$. These are used to calculate the various gradient approximations. For example, in $\text{PM}_{|y_m^s - y_i^s|}$ w_{im} depends on y_i and y_m, while in PM_{iso} it depends on all of y_i and $\boldsymbol{y}_{\partial \boldsymbol{i}}$. By introducing the intensities into w_{im}, we can no longer write

$$p(\boldsymbol{z}_i | \boldsymbol{z}_{\partial \boldsymbol{i}}; \boldsymbol{\Psi}) \propto \exp(\beta \sum_{m \in \boldsymbol{\partial i}} \frac{1}{\delta_{im}} w_{im}(\boldsymbol{z}_i, \boldsymbol{z}_m) \boldsymbol{z}_i^T \boldsymbol{z}_m).$$

Rather, we have

$$p(\boldsymbol{z}_i | y_i \boldsymbol{z}_{\partial \boldsymbol{i}}, \boldsymbol{y}_{\partial \boldsymbol{i}}; \boldsymbol{\Psi}) \propto \exp(\beta \sum_{m \in \boldsymbol{\partial i}} \frac{1}{\delta_{im}} w_{im}(y_i, \boldsymbol{y}_{\partial \boldsymbol{i}}) \boldsymbol{z}_i^T \boldsymbol{z}_m). \tag{5.9}$$

For the intensity pdf, we have

$$f(y_i | \boldsymbol{z}_i; \boldsymbol{\Theta}) = \sum_{j=1}^{g} z_{ij} \phi(y_i; \mu_j, \sigma_j^2). \tag{5.10}$$

First, it is not clear that these two conditional distributions are compatible. Second, $p(\boldsymbol{z}_i | \boldsymbol{y}_i, \boldsymbol{z}_{\partial \boldsymbol{i}}, \boldsymbol{y}_{\partial \boldsymbol{i}}; \boldsymbol{\Psi})$ does not necessarily form an MRF.

To see this, consider Figure 5.6a which shows the undirected graphical model corresponding to mixture-MRF segmentation with the Potts model. Every voxel has two nodes, one with an

associated intensity y_i and the other with associated label $\boldsymbol{z}_i$. An edge is drawn between nodes that are conditionally dependent, e.g. for y_i and $\boldsymbol{z}_i$

$$f(y_i, \boldsymbol{z}_i | \boldsymbol{y}_{-i}, \boldsymbol{z}_{-i}) \neq f(y_i | \boldsymbol{y}_{-i}, \boldsymbol{z}_{-i}) p(\boldsymbol{z}_i | \boldsymbol{y}_{-i}, \boldsymbol{z}_{-i}),$$

where $\boldsymbol{y}_{-i}$ means all intensities y_m except for $m = i$, and similarly for $\boldsymbol{z}_{-i}$. Now for the mixture-MRF model with the Potts MRF,

$$p(\boldsymbol{z}_i | \boldsymbol{y}_{-i}, \boldsymbol{z}_{-i}) = p(\boldsymbol{z}_i | \boldsymbol{z}_{\partial \boldsymbol{i}}).$$

Separately, we note that

$$\begin{aligned} f(y_i, \boldsymbol{z}_i | \boldsymbol{y}_{-i}, \boldsymbol{z}_{-i}) &= f(y_i | \boldsymbol{z}_i, \boldsymbol{y}_{-i}, \boldsymbol{z}_{-i}) p(\boldsymbol{z}_i | \boldsymbol{y}_{-i}, \boldsymbol{z}_{-i}) \\ &= f(y_i | \boldsymbol{z}_i) p(\boldsymbol{z}_i | \boldsymbol{z}_{\partial \boldsymbol{i}}). \end{aligned}$$

The two variables Y_i and $\boldsymbol{Z}_i$ would be conditionally independent if $f(y_i|\boldsymbol{z}_i)$ were equal to $f(y_i|\boldsymbol{y}_{-i}, \boldsymbol{z}_{-i})$. However, since $f(y_i|\boldsymbol{z}_i)$ depends on the value of $\boldsymbol{z}_i$, this is not the case; thus, there is an edge connecting y_i and $\boldsymbol{z}_i$ in the graph. Similarly, there are edges between $\boldsymbol{z}_i$ and its neighbours $\boldsymbol{z}_m$ where $m \in \boldsymbol{\partial i}$.

By the Hammersley-Clifford theorem this forms a valid MRF, because the joint distribution $f(\boldsymbol{y}, \boldsymbol{z})$ may be decomposed (up to a normalising constant) into terms between pairwise cliques $(y_i, \boldsymbol{z}_i)$ and $\boldsymbol{z}_i, \boldsymbol{z}_m$. The clique potential associated with the $(y_i, \boldsymbol{z}_i)$ is

$$\prod_{j=1}^{g} f(y_i | \boldsymbol{Z}_i = \boldsymbol{e}_j)^{z_{ij}}$$

where $f(y_i|\boldsymbol{Z}_i = \boldsymbol{e}_j)$ is the normal probability density function with tissue j's parameters. The clique potential for $(\boldsymbol{z}_i, \boldsymbol{z}_m)$ is

$$\exp(\beta \boldsymbol{z}_i^T \boldsymbol{z}_m).$$

Figure 5.6b shows the undirected graphical model corresponding to the $\text{PM}_{|y_m^s - y_i^s|}$ "MRF" presented in this chapter (the solid lines only). There are additional edges between $\boldsymbol{z}_i$ and the neighbouring y_m with $m \in \boldsymbol{\partial i}$ due to the presence of $w_{im}(y_i, y_m)$ in the MRF potential. For $\text{PM}_{|\boldsymbol{\nabla} I^s \cdot \hat{\boldsymbol{im}}|}$ and PM_{iso} this graph has even more edges due to the gradient approximation requiring all neighbouring intensity values. Since it is unclear if the conditional densities (5.9) and (5.10) are compatible with each other, it is also unclear if an edge may be drawn between y_i and y_m (dotted line).

If there is no edge, then the joint pdf cannot be decomposed into a product of terms over cliques. This is because the smallest decomposition of the joint pdf must be over sets consisting of four nodes $\{y_i, y_m, \boldsymbol{z}_i, \boldsymbol{z}_m\}$; smaller sets such as pairs and triples cannot be isolated in (5.9). However, if there is no edge between y_i and y_m, the four nodes do not form a clique, which must

be fully-connected. Hence, there exists no decomposition of the joint pdf into a product over cliques, and by the Hammersley-Clifford theorem, $f(\boldsymbol{y}, \boldsymbol{z})$ does not form a valid MRF.

Previously mentioned existing work does not appear to recognise this inconsistency, with the exception of Grau et al. (2006). They noted the difficulty of the coupling between $\boldsymbol{z}$ and $\boldsymbol{y}$, but viewed $w_{im}(y_i, \boldsymbol{y}_{\partial \boldsymbol{i}})$ as constants (one per voxel pair) rather than realisations of a random variable. In all cases, they proceeded by using the EM algorithm with pseudolikelihood as has been discussed in previous chapters, treating $p(\boldsymbol{z}_i|\boldsymbol{z}_{\partial \boldsymbol{i}}, \boldsymbol{y}_{\partial \boldsymbol{i}}; \boldsymbol{\Psi})$ as if it were simply $p(\boldsymbol{z}_i|\boldsymbol{z}_{\partial \boldsymbol{i}}; \boldsymbol{\Psi})$.

One possible solution is to discard the current conditional formulation and instead build up a new MRF from clique potentials to satisfy the Hammersley-Clifford theorem. First, induce an explicit dependency between voxels i and m. For example, one could explicitly incorporate a blurring of intensities between the neighbouring voxels (in addition to the dependence between neighbouring labels); (Besag, 1986, section 5.3) suggested an autonormal model might be used for this purpose. Then, *define* the clique potential $\{y_i, y_m, \boldsymbol{z}_i, \boldsymbol{z}_m\}$ as

$$\psi(y_i, y_m, \boldsymbol{z}_i, \boldsymbol{z}_m) = \exp(\beta \frac{1}{\delta_{im}} w_{im}(y_i, y_m) \boldsymbol{z}_i^T \boldsymbol{z}_m + \log(f(y_i|\boldsymbol{z}_i)) + \log(f(y_m|\boldsymbol{z}_i)) + b(y_i, y_m)),$$

where $b(y_i, y_m)$; represents the relationship (e.g. intensity blur) between y_i and y_m. Then, form the pdf (up to a normalising constant)

$$f(\boldsymbol{y}, \boldsymbol{z}) \propto \prod_{\text{neighbours } i,m} \psi(y_i, y_m, \boldsymbol{z}_i, \boldsymbol{z}_m).$$

This is a valid MRF, since it is a product of clique potentials with cliques of size 4.

From here, $f(\boldsymbol{y}, \boldsymbol{z})$ may be approximated by a pseudolikelihood or mean-field approximation. There is some freedom in the choice of conditional pdf to be used here. For example,

$$f_{PL}(\boldsymbol{y}, \boldsymbol{z}) \approx \prod_i p(\boldsymbol{z}_i|\boldsymbol{y}, \boldsymbol{z}_{\partial \boldsymbol{i}}) f(y_i|\boldsymbol{y}_{\partial \boldsymbol{i}}, \boldsymbol{z})$$

or

$$f_{PL}(\boldsymbol{y}, \boldsymbol{z}) \approx \prod_i f(\boldsymbol{y}_i, \boldsymbol{z}_i|\boldsymbol{y}_{\partial \boldsymbol{i}}, \boldsymbol{z}_{\partial \boldsymbol{i}})$$

are possibilities. The conditional probabilities must be derived. For example, to find $p(\boldsymbol{z}_i|\boldsymbol{y}, \boldsymbol{z}_{\partial \boldsymbol{i}}) = p(\boldsymbol{z}_i|\boldsymbol{y}, \boldsymbol{z}_{-i})$, one can consider the difference in the joint pdf by change of just voxel i from label j to k, and renormalising:

$$\frac{p(\boldsymbol{Z}_i = \boldsymbol{e}_j|\boldsymbol{y}, \boldsymbol{z}_{\partial \boldsymbol{i}})}{p(\boldsymbol{Z}_i = \boldsymbol{e}_k|\boldsymbol{y}, \boldsymbol{z}_{\partial \boldsymbol{i}})} = \frac{f(\boldsymbol{y}, \boldsymbol{z}_{-i}, \boldsymbol{Z}_i = \boldsymbol{e}_j)}{f(\boldsymbol{y}, \boldsymbol{z}_{-i}, \boldsymbol{Z}_i = \boldsymbol{e}_k)}$$

$$\sum_{j=1}^{g} p(\boldsymbol{Z}_i = \boldsymbol{e}_j|\boldsymbol{y}, \boldsymbol{z}_{\partial \boldsymbol{i}}) = 1$$

These may not take the same form as the conditional densities (5.9) and (5.10).

This is an approach that can be explored in future work. For now, we acknowledge this limitation of the model and proceed with the EM algorithm as an approximation.

5.3.3.3 Discriminate random fields

Finally, we note the similarity between (5.9) and a conditional random field. A *conditional random field*, introduced in (Lafferty et al., 2001), is an undirected graphical model of which the MRF is a special case. They have since been extended to multi-dimensional lattice structures where they are known as *discriminative random fields* (DRFs) (Kumar and Hebert, 2003, 2006). In image analysis, they appear to be primarily used in foreground-background segmentation, i.e. $g = 2$ (Blake et al., 2004; Boykov and Jolly, 2001). A DRF is the discriminative counterpart to MRFs, which are generative. Other examples of discriminative models are support vector machines, or neural networks. That is, a DRF constructs only $p(\boldsymbol{z}|\boldsymbol{y})$ without considering the marginal or joint distributions; further, the functional dependence of $\boldsymbol{z}$ on $\boldsymbol{y}$ is learned rather than specified. By contrast, the mixture-MRF approach specifies $p(\boldsymbol{z})$ and $p(\boldsymbol{y}|\boldsymbol{z})$ to construct the joint and posterior distributions. A DRF containing unary and pairwise potentials (as studied in this thesis) takes the (local) form

$$p(\boldsymbol{z}_i|\boldsymbol{z}_{\partial \boldsymbol{i}}, \boldsymbol{y}) \propto \exp(U_i(\boldsymbol{z}_i, \boldsymbol{y}) + \sum_{m \in \partial \boldsymbol{i}} U_{im}(\boldsymbol{z}_i, \boldsymbol{z}_m, \boldsymbol{y})).$$

The difference between a DRF and MRF is that the DRF allows the clique potentials to depend on the observed data, not just the neighbours $\boldsymbol{z}_{\partial \boldsymbol{i}}$. A pseudolikelihood for $p(\boldsymbol{z}|\boldsymbol{y})$ may be constructed analogously to an MRF by multiplying together the individual conditional probabilities.

At first glance, it may appear that a DRF should be used for the work in this chapter due to the similarities with (5.9). However, since a DRF is discriminative, the functional forms of the potentials are not specified *a priori* and must be learned. For example, the unary potential is written:

$$U_i(\boldsymbol{z}_i, \boldsymbol{y}) = \psi(\boldsymbol{w}^T \boldsymbol{f}_i(\boldsymbol{y})),$$

where

- $\boldsymbol{f}_i(\boldsymbol{y})$ is a *feature vector* computed at voxel i, such as the intensity, the image gradient, information from an anatomical atlas. Whatever features thought to be relevant to the application may be included.
- $\boldsymbol{w}$ is an unknown vector of weights for each feature that must be learned.
- ψ is a link function (similar to a generalised linear model), for example logit.

In this way the specific contribution of each feature to the pdf (i.e. the weights $\boldsymbol{w}$) need not be modelled explicitly, but is instead learned. Once learned (usually through maximum pseudolikelihood), the weights $\boldsymbol{w}$ are fixed regardless of image, unlike our $\boldsymbol{\Psi}$ which may vary

with *each* image. Then when the model is presented with a new image, the segmentation is usually determined with graph-cuts or ICM.

Despite the striking similarities between the DRF and MRF, the key practical difference is that the DRF requires labelled training data to learn the parameters. In contrast, our method must estimate the parameters and segmentation at the same time from the single image to be segmented. However, it is possible the work could be reformulated as a DRF were training data available.

5.3.4 Algorithm

With this in mind, the algorithm used is almost the same as the previous chapter. We repeat it here for clarity, explaining what changes are needed to incorporate the weights w_{im}. Where we previously had the MRF probability $p(\boldsymbol{e}_j|\boldsymbol{z}_{\partial \boldsymbol{i}};\boldsymbol{\Psi})$, we replace with the new probability $p(\boldsymbol{e}_j|\boldsymbol{z}_{\partial \boldsymbol{i}}, y_i, \boldsymbol{y}_{\partial \boldsymbol{i}};\boldsymbol{\Psi})$, though this is not strictly a valid MRF, nor is it a density over $\boldsymbol{z}$ only. We use Expectation-Maximisation to estimate the MRF and intensity parameters, using the pseudolikelihood or mean-field approximation in the Q-function for computational tractability. The Q function is given by

$$Q(\boldsymbol{\Theta},\boldsymbol{\Psi}|\boldsymbol{\Theta}^{(t)},\boldsymbol{\Psi}^{(t)}) = \sum_{i=1}^{n}\sum_{j=1}^{g} {\tau_{ij}}^{(t)}(\log\phi(y_i;\mu_j,\sigma_j^2) + \log p(\boldsymbol{e}_j|\boldsymbol{z}_{\partial \boldsymbol{i}}, y_i, \boldsymbol{y}_{\partial \boldsymbol{i}};\boldsymbol{\Psi})) \tag{5.11}$$

where

$$p(\boldsymbol{z}_i|\boldsymbol{z}_{\partial \boldsymbol{i}};\boldsymbol{\Psi}) = \frac{\exp(\beta\sum_{m\in\partial \boldsymbol{i}}\frac{1}{\delta_{im}}w_{im}(y_i,\boldsymbol{y}_{\partial \boldsymbol{i}};\kappa)\boldsymbol{z}_i^T\boldsymbol{z}_m)}{C_i}$$

$$C_i = \sum_{k=1}^{g}\exp(\beta\sum_{m\in\partial \boldsymbol{i}}\frac{1}{\delta_{im}}w_{im}(y_i,\boldsymbol{y}_{\partial \boldsymbol{i}};\kappa)\boldsymbol{e}_k^T\boldsymbol{z}_m)$$

and the weights function is one of $\mathrm{PM}_{|y_m^s - y_i^s|}$ (5.7), $\mathrm{PM}_{|\boldsymbol{\nabla} I^s \cdot \hat{\boldsymbol{im}}|}$ (5.6), or $\mathrm{PM}_{\mathrm{iso}}$ (5.5), repeated below.

$$\mathrm{PM}_{\mathrm{iso}}: \qquad w_{im}(y_i,\boldsymbol{y}_{\partial \boldsymbol{i}}) = \exp(-\left(\frac{|\boldsymbol{\nabla} I^s|}{\kappa}\right)^2)$$

$$\mathrm{PM}_{|\boldsymbol{\nabla} I^s \cdot \hat{\boldsymbol{im}}|}: \qquad w_{im}(y_i,\boldsymbol{y}_{\partial \boldsymbol{i}}) = \exp(-\left(\frac{|\boldsymbol{\nabla} I^s \cdot \hat{\boldsymbol{im}}|}{\kappa}\right)^2)$$

$$\mathrm{PM}_{|y_m^s - y_i^s|}: \qquad w_{im}(y_i,y_m) = \exp(-\left(\frac{|y_m^s - y_i^s|}{\kappa}\right)^2)$$

compared to

$$\text{single-beta}: w_{im} = 1.$$

The weight κ is pre-set as the median absolute deviation of the image gradient (5.8).

Pre-smoothing of the image to obtain y^s (or image gradient $\boldsymbol{\nabla} I^s$) was performed with $s = 1$. For $\mathrm{PM}_{\mathrm{iso}}$ and $\mathrm{PM}_{|\boldsymbol{\nabla} I^s \cdot \hat{\boldsymbol{im}}|}$, the gradient of the image is computed by convolution of the image with an approximation to the first derivative of a Gaussian; this is equivalent to finding the

gradient of the Gaussian-smoothed image. This process yields an approximation to the intensity gradient vector at each voxel. For specific details, we refer the reader to Deriche (1993); the Insight ToolKit software (Yoo et al., 2002) was used to perform this.

We note that as these weighting functions w_{im} only depend on intensity gradient (which does not change) and the fixed parameter κ, w_{im} only needs to be calculated once at the start of the algorithm for each pair of voxels. One may then compute the number of neighbours of voxel i with label j as

$$\tilde{u_{ij}} = \sum_{m \in \boldsymbol{\partial i}} \frac{1}{\delta_{im}} w_{im}(y_i, \boldsymbol{y_{\partial i}}) \boldsymbol{e}_j^T \boldsymbol{z}_m$$

and then compute $p(\boldsymbol{e}_j | \boldsymbol{z_{\partial i}}, y_i, \boldsymbol{y_{\partial i}}; \boldsymbol{\Psi})$ as

$$\frac{\exp(\beta \tilde{u_{ij}})}{\sum_{k=1}^{g} \exp(\beta \tilde{u_{ik}})}.$$

This is similar to simplified Potts MRF (i.e., the single-beta MRF) except for the neighbour count having additional weights per neighbour (the single-beta MRF has $w_{im} = 1$).

The rest of the algorithm proceeds as follows. On iteration t:

1. **(C-step)** Form an estimate of the current labels $\boldsymbol{z}^{(t)}$ to be used as neighbours; either discrete (for the pseudolikelihood approximation) or continuous (for the mean-field approximation). The pseudolikelihood version uses the Iterated Conditional Modes (ICM) update

   $$\boldsymbol{z}_i{}^{(t+1)} = \boldsymbol{e}_j \text{ where } j = \arg\max_k p(\boldsymbol{e}_j | \boldsymbol{z_{\partial i}}{}^{(t,t+1)}, y_i, \boldsymbol{y_{\partial i}}; \boldsymbol{\Psi}) \phi(y_i; \mu_k{}^{(t-1)}, \sigma_k^{2(t-1)}).$$

   The mean-field version uses the mean-field update

   $$\langle \boldsymbol{z}_i \rangle^{(t+1)} = \sum_{j=1}^{g} \boldsymbol{e}_j \frac{p(\boldsymbol{e}_j | \boldsymbol{z_{\partial i}}{}^{(t,t-1)}, y_i, \boldsymbol{y_{\partial i}}; \boldsymbol{\Psi}) \phi(y_i; \mu_j{}^{(t-1)}, \sigma_j^{2(t-1)})}{\sum_{k=1}^{g} p(\boldsymbol{e}_k | \boldsymbol{z_{\partial i}}{}^{(t,t-1)}, y_i, \boldsymbol{y_{\partial i}}; \boldsymbol{\Psi}) \phi(y_i; \mu_k{}^{(t-1)}, \sigma_k^{2(t-1)})}$$

   These updates should be performed sequentially. To save time, we divide the voxels into coding sets (see Appendix B) and update each set simultaneously, visiting them sequentially.
2. **(E-step)** Calculate $\tau_{ij}{}^{(t)}$, using $\boldsymbol{z}^{(t)}$ from the C-step to compute the neighbour term u_{ij}:

   $$\tau_{ij}{}^{(t)} = \frac{p(\boldsymbol{e}_j | \boldsymbol{z_{\partial i}}{}^{(t)}, y_i, \boldsymbol{y_{\partial i}}; \boldsymbol{\Psi}^{(t)}) \phi(y_i; \mu_j{}^{(t-1)}, \sigma_j^{2(t-1)})}{\sum_{k=1}^{g} p(\boldsymbol{e}_k | \boldsymbol{z_{\partial i}}{}^{(t)}, y_i, \boldsymbol{y_{\partial i}}; \boldsymbol{\Psi}^{(t)}) \phi(y_i; \mu_k{}^{(t-1)}, \sigma_k^{2(t-1)})}. \tag{5.12}$$

3. **(M-step)** Maximise Q with respect to $\boldsymbol{\Theta}$ to obtain the intensity parameters.

$$\mu_j{}^{(t)} = \frac{\sum_{i=1}^n \tau_{ij}{}^{(t)} y_i}{\sum_{i=1}^n \tau_{ij}{}^{(t)}}$$
$$\boldsymbol{\Sigma}_j{}^{(t)} = \frac{\sum_{i=1}^n \tau_{ij}{}^{(t)} (y_i - \mu_j{}^{(t)})^2}{\sum_{i=1}^n \tau_{ij}{}^{(t)}}.$$

   Then, update $\boldsymbol{\Psi}$ using maximum pseudolikelihood estimation. This amounts to numerically maximising the concave Q-function with respect to β, using the gradient (3.9) as necessary. The gradient equation for β is the same as that for the original single-beta MRF, with $\tilde{u_{ij}}$ in place of u_{ij}.

These steps are repeated until the relative change in approximate observed log-likelihood falls below a pre-specified tolerance (1e-5 in these experiments), or it decreases. This is

$$\log f(\boldsymbol{y}) \approx \prod_{i=1}^n \log \left(\sum_{j=1}^g \phi(y_i | \boldsymbol{e}_j; \mu_j{}^{(t)}, \sigma_j^{2(t)}) p(\boldsymbol{e}_j | \boldsymbol{z}_{\partial \boldsymbol{i}}{}^{(t)}, y_i, \boldsymbol{y}_{\partial \boldsymbol{i}}; \boldsymbol{\Psi}^{(t)}) \right).$$

EM on a standard mixture model guarantees an increase in the observed log-likelihood and Q; however, we no longer have this guarantee.

We initialise the algorithm by fitting a standard normal mixture model with 3 components to the image (i.e., without the MRF). This yields an initial segmentation to be used as the neighbours, as well as means and standard deviations. The initial $\beta^{(0)}$ is estimated from this initial segmentation.

5.4 Experiments

Our aim is to see if any of the anisotropic MRF potentials PM_{iso}, $\text{PM}_{|y_m^s - y_i^s|}$ and $\text{PM}_{|\boldsymbol{\nabla} I^s \cdot \hat{\boldsymbol{im}}|}$ are effective at preserving edges that would otherwise be smoothed with the single-beta MRF. We compare the segmentations arising from using each different potential, with parameter β estimated using maximum pseudolikelihood estimation. We additionally compared estimating β to setting it fixed to 1 (i.e. omitting it) to study the value of parameter estimation in the anisotropic MRF setting. In these experiments we used the pseudolikelihood approximation with a size-6 neighbourhood.

The images segmented were from the Internet Brain Segmentation Repository (IBSR) (Rohlfing, 2012).[2] The dataset used consists of T1-weighted coronal MR volumes of 18 normal subjects of ages 7 to 71. Each volume consists of 128 coronal slices spaced at 1.5mm with in-plane resolution varying from 0.84×0.84mm to 1.00×1.00mm. This dataset also contains manual segmentations

[2]The MR brain data sets and their manual segmentations were provided by the Center for Morphometric Analysis at the Massachusetts General Hospital and are available at http://www.cma.mgh.harvard.edu/ibsr/.

to compare the automatic segmentations to. The images are already skull-stripped with bias-correction already performed, so no additional preprocessing was done. That is, all non-brain voxels (such as skull, fat) are already removed from the image as we wish to concentrate on segmentation of the brain only.

Performance against the manual segmentations was evaluated using segmentation accuracy (for overall accuracy) and Dice similarity (for per-tissue accuracy). Let A and B represent two segmentations, being sets of indices for each tissue. That is, $A_j, j = 1, \ldots, g$ are non-intersecting subsets of the indices $1, \ldots, n$ whose union is the entire brain, where $i \in A_j$ implies that voxel i is assigned to tissue j in segmentation A.

The segmentation accuracy is the overall percentage of voxels correctly classified. Since the reference and test segmentations have the same number of voxels (all of the brain voxels), this is well-defined.

$$\text{accuracy}(A, B) = \frac{|A \cap B|}{|A|}.$$

The Dice similarity coefficient (commonly called 'Dice score' or 'Dice index') (Dice, 1945) is used to compare segmentations on a tissue-by-tissue basis. The Dice coefficient for a given tissue between two segmentations A and B is given by the number of correctly-classified voxels divided by the average area classified (of that tissue):

$$\text{Dice}(A_j, B_j) = \frac{2|A_j \cap B_j|}{|A_j| + |B_j|}.$$

Both measures range from 0 to 1, with 1 meaning a perfect match between the two segmentations of that tissue and 0 meaning no match. The reason for using Dice coefficient for each tissue rather than accuracy is that the number of voxels classified as a particular tissue may not be equal between the two segmentations, whereas the number of overall voxels in the brain (used for the accuracy) is.

5.5 Results

Table 5.1: Average accuracy and Dice for different MRF potentials

MRF	$\hat{\beta}$ (average)	accuracy	Dice (CSF)	Dice (GM)	Dice (WM)
$\text{PM}_{\lvert y_m^s - y_i^s \rvert}$	4.38	**0.818**	**0.634**	**0.846**	**0.829**
single-beta	1.88	0.812	0.625	0.843	0.820
$\text{PM}_{\lvert \boldsymbol{\nabla} I^s \cdot \hat{\boldsymbol{im}} \rvert}$	15.63	0.808	0.621	0.839	0.817
PM_{iso}	135.15	0.803	0.600	0.832	0.827

Table 5.1 and figure 5.7 show the average accuracy and per-tissue Dice scores across the dataset for each MRF potential. The $\text{PM}_{|y_m^s - y_i^s|}$ potential gave the best performance in all metrics. Use

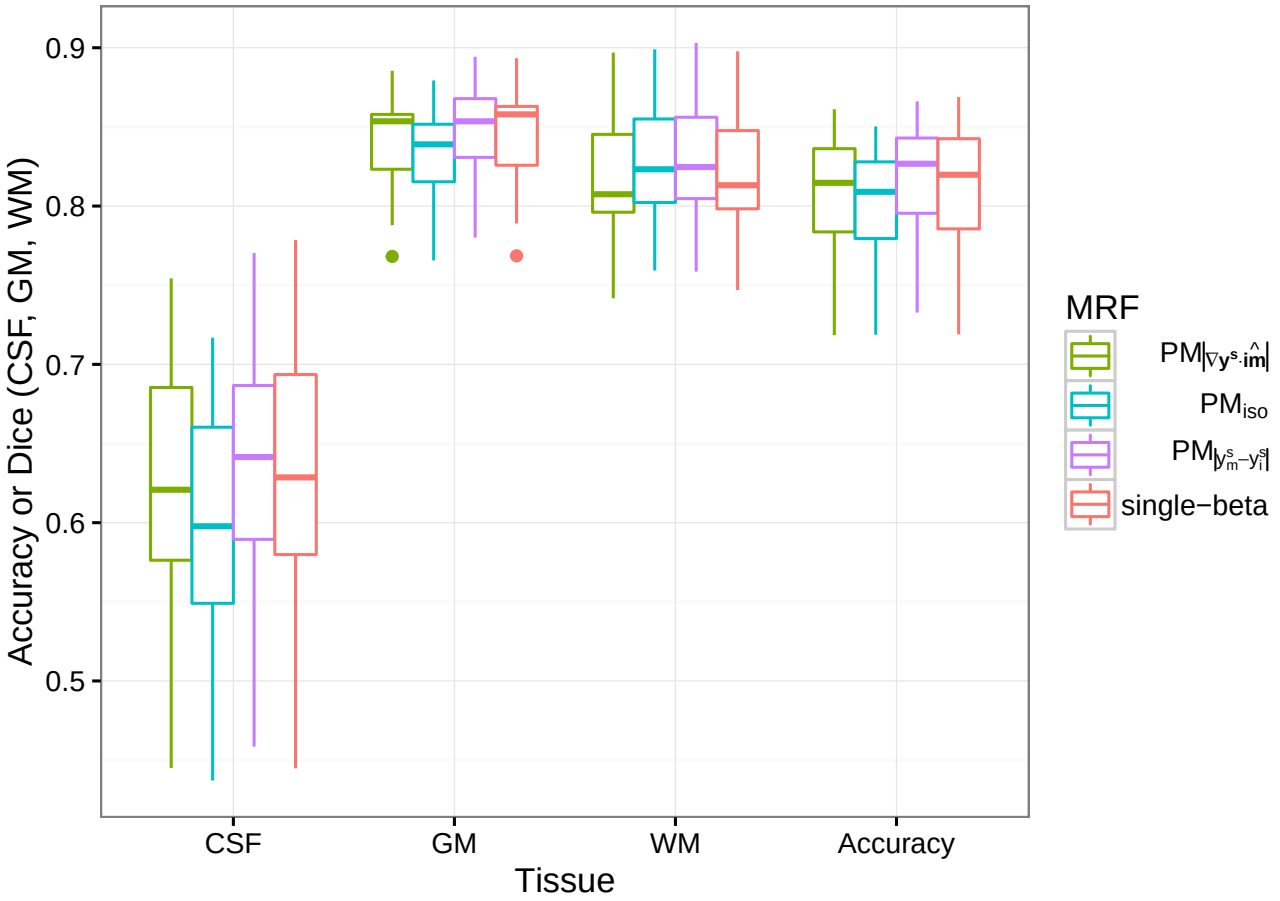


Figure 5.7: Segmentation metrics (accuracy or Dice coefficient) for the various MRFs using MPL

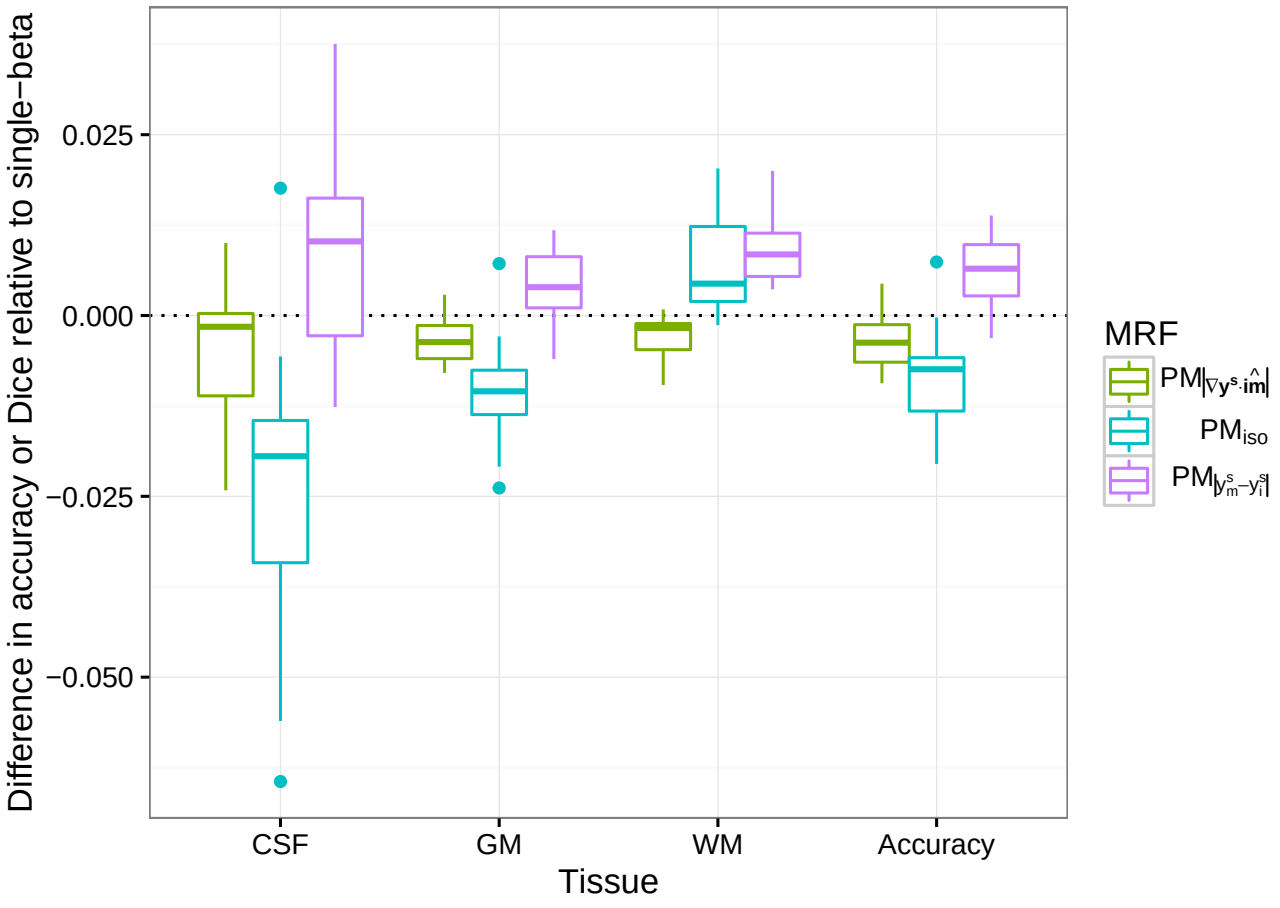


Figure 5.8: Paired difference in accuracy and Dice score, relative to the single-beta MRF

Table 5.2: Mixed-effects model of segmentation accuracy for different models using MPL, controlling for subject blocking.

	Sum Sq	Mean Sq	NumDF	DenDF	F	Pr(>F)
MRF	0.002	0.001	3	51.1	52.321	<0.001*

Table 5.3: Post-hoc pairwise comparisons for differences in accuracy using Tukey's method. Only significant differences are shown.

Comparison	Estimate	p
$\text{PM}_{\lvert\boldsymbol{\nabla} I^s \cdot \hat{\boldsymbol{im}}\rvert}$ - $\text{PM}_{\lvert y_m^s - y_i^s\rvert}$	-0.010	<0.001*
$\text{PM}_{\lvert\boldsymbol{\nabla} I^s \cdot \hat{\boldsymbol{im}}\rvert}$ - PM_{iso}	0.005	<0.001*
$\text{PM}_{\lvert\boldsymbol{\nabla} I^s \cdot \hat{\boldsymbol{im}}\rvert}$ - single-beta	-0.004	0.016*
$\text{PM}_{\lvert y_m^s - y_i^s\rvert}$ - PM_{iso}	0.015	<0.001*
$\text{PM}_{\lvert y_m^s - y_i^s\rvert}$ - single-beta	0.006	<0.001*
PM_{iso} - single-beta	-0.009	<0.001*

of an anisotropic MRF yielded a large improvement in CSF and GM, though at the cost of WM segmentation (with the exception of the $\text{PM}_{\lvert y_m^s - y_i^s\rvert}$ potential).

Figure 5.8 shows the improvement in accuracy or Dice score of each anisotropic MRF relative to the single-beta MRF. It appears that use of the $\text{PM}_{\lvert y_m^s - y_i^s\rvert}$ MRF is warranted over the single-beta MRF, while the other anisotropic MRFs are worse. This was confirmed by a mixed-effects model of segmentation accuracy against MRF, controlling for repeated subjects, which showed a significant effect of choice of MRF (table 5.2). Pairwise comparisons were performed using Tukey's method and significant differences are shown in table 5.3. We are only interested in comparison of the anisotropic MRFs to the single-beta MRF. We see that the $\text{PM}_{\lvert y_m^s - y_i^s\rvert}$ MRF has significantly more accurate segmentations than the single-beta MRF, while the others are significantly less accurate.

The estimated β values can be seen in table 5.1 and figure 5.9. All of the anisotropic MRFs had higher β than the single-beta MRF, and the PM_{iso} MRF very much so.

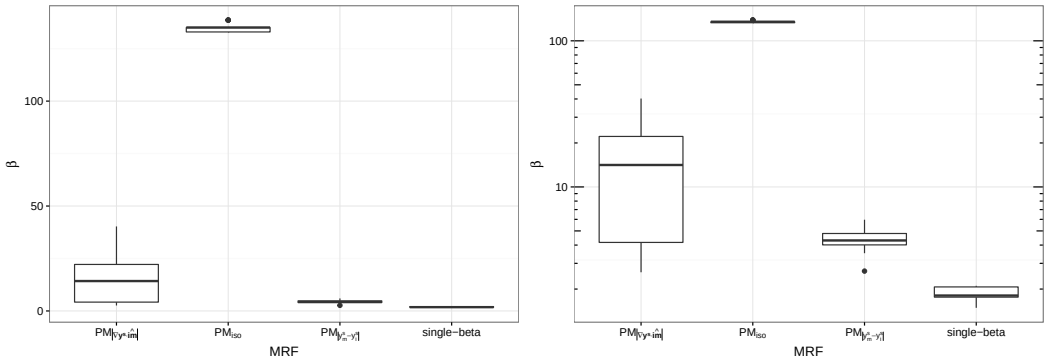


Figure 5.9: β values for various MRFs (see also Table 5.1). The same plot on the right-hand side has a log10 scale on the Y axis.

5.6 Discussion

Figure 5.10 shows some sample segmentations under the various potentials. As with all previous chapters, there is difficulty with deep grey-matter structures, which can be resolved by incorporating an anatomical atlas into the tissue prior. It is likely that the accuracy for all segmentations suffers because the IBSR manual segmentations systematically label sulcal CSF as WM (Valverde et al., 2015). However, as this mislabelling is consistent, we are still able to compare the fine features of our segmentations.

5.6.1 Comparison of anisotropic potentials

We consider whether or not to use an anisotropic MRF potential, and if so, which one to use. From the results seen in tbl:ch5.results.whichPotential, the $\mathrm{PM}_{|y_m^s - y_i^s|}$ MRF has potential, obtaining significantly higher segmentation accuracy overall, and higher Dice coefficient in each tissue to the single-beta MRF. The other anisotropic potentials had significantly lower accuracy than the single-beta MRF, and also had lower Dice coefficient in all tissues except in WM for $\mathrm{PM}_{\mathrm{iso}}$.

Figure 5.10 shows sample segmentations produced by the various potentials. The circles point out some of the differences, in particular where the $\mathrm{PM}_{|y_m^s - y_i^s|}$ potential has preserved features that the single-beta potential has not. These features are typically thin and narrow, exactly those features that the anisotropic MRFs were designed to preserve. For example, there are many cases of WM in the GM cortical folds not being smoothed over or having the ends filled in. The other anisotropic MRFs also generally preserve these features to varying extents. In some cases the feature that has been preserved is not in the manual segmentation, for example the extrasulcal CSF in subject IBSR_06. However, this is more an artefact of the intensity parameters placing the CSF-GM boundary too high rather than the MRF. The anisotropic MRFs merely attempt to ensure that long thin features are preserved, regardless of label.

As a specific example, Figure 5.11 shows the image region in the circle drawn on subject 1 in figure 5.10. A thin strip of grey matter between the ventricles (CSF) has been preserved by the $\mathrm{PM}_{|y_m^s - y_i^s|}$ potential but not by the others. Again, we note that in the manual segmentation, there is no strip of GM between the ventricles and suppose this to be a consequence of the mixture model component estimating the tissue parameters such that the CSF-GM boundary is too high in intensity. Nevertheless, figure 5.11 serves as a demonstrative example as to how the different MRFs act anisotropically.

The figure shows the image intensities in the region, the directional gradient at each voxel (the lengths of the arrows proportional to the gradient magnitude and the arrows themselves indicating the direction), and the segmentations produced by the various potentials. The numbers on the segmentations show the weights assigned to the neighbouring voxels when

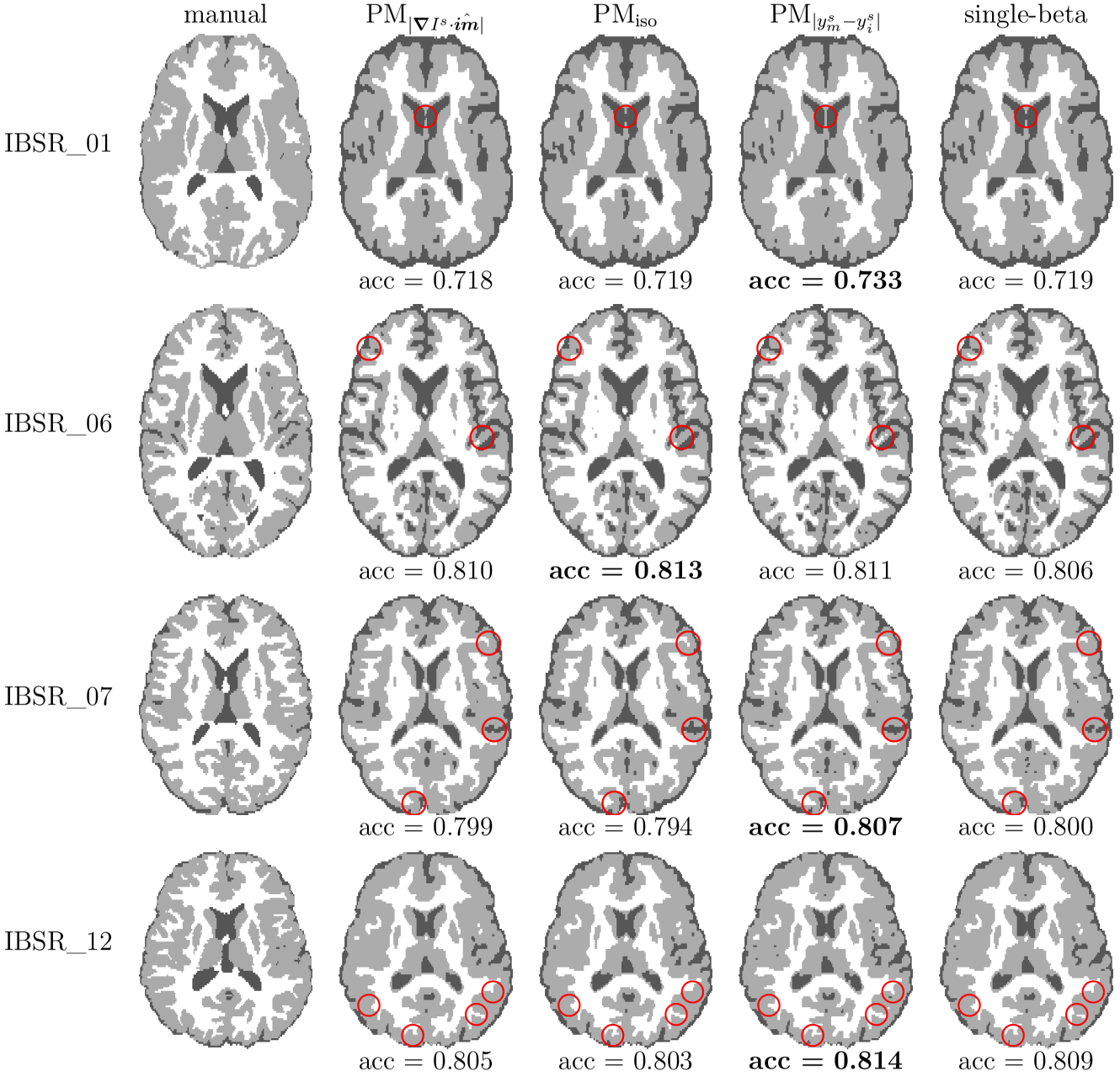


Figure 5.10: Example segmentations for different anisotropic MRFs. The red circles indicate some points of difference between the potentials, particularly where thin sulci have been preserved to various degrees by using an anisotropic MRF. Segmentation accuracy is displayed and in bold for the highest-accuracy MRF per subject.

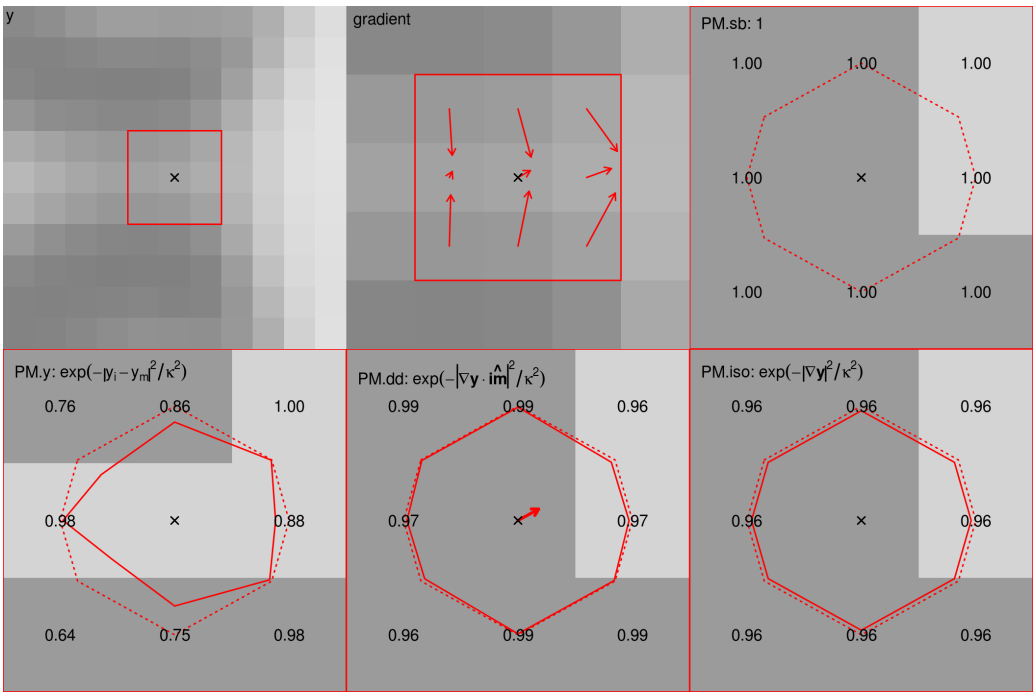


Figure 5.11: Zoomed-in portion of the circled area of subject 1, showing the raw intensities ("y"), the directional gradient, and the segmentations produced by the various MRF potentials. The red lines and numbers show the weights w_{im} assigned to each neighbour relative to the single-beta MRF (dotted red line). See the discussion for explanation.

considering what label to assign the centre voxel. The red lines also visualise the weights: the dotted red line represents the weights for the single-beta MRF. For example, the single-beta MRF has weights of 1 for each neighbour, so the dotted red line shows the discretisation of a circle with radius 1. The solid red lines for the anisotropic MRFs show the weights for those MRFs relative to the single-beta weights.

The $\text{PM}_{|y_m^s - y_i^s|}$ MRF has preserved the thin strip of grey matter between the ventricles (which are darker CSF), while the other MRFs have not. The intensity of the centre voxel is intermediate between GM and CSF due to the partial volume effect, but the normal probabilities with the respective estimated tissue parameters still favour GM for this voxel. On the other hand, the single-beta MRF (with any positive β) will prefer to label this voxel CSF rather than GM, being the majority label in the neighbourhood. The value of β estimated for the single-beta MRF is sufficiently large that CSF is chosen for the final voxel label upon combining the intensity and MRF probabilities under the single-beta MRF.

As the intensity gradient magnitude is fairly weak in this region (compared to the magnitudes over the entire image), PM_{iso} also produces weights that are almost 1. Since these are isotropic, the PM_{iso} potential acts very much like the single-beta potential in this neighbourhood, and likewise chooses CSF for the centre voxel.

One might expect the gradient-based potential $\text{PM}_{|\boldsymbol{\nabla} I^s \cdot \hat{\boldsymbol{im}}|}$ to succeed in detecting this feature. However as can be seen from the gradient image, since the thin strip of GM is only one voxel wide, it forms a narrow ridge running along the horizontal axis. As this represents a local

maximum in vertical dimension, the gradient cannot detect it, and hence points horizontally with a small magnitude. The small magnitude of the gradient causes the $\text{PM}_{|\boldsymbol{\nabla} I^s \cdot \hat{i\boldsymbol{m}}|}$ potential to be isotropic (as designed), so that again CSF is chosen for the centre voxel. Even if the gradient magnitude were strong, the edge would incorrectly be detected as running vertically due to the gradient direction. The $\text{PM}_{|\boldsymbol{\nabla} I^s \cdot \hat{i\boldsymbol{m}}|}$ potential can detect edges, but not thin lines. This symmetry may explain why it performs worse than the single-beta MRF overall.

Finally, we consider the $\text{PM}_{|y^s_m - y^s_i|}$ potential. As this is a one-sided finite difference approximation to the gradient (as opposed to a central approximation), it does not suffer from the problem of the vertical gradients cancelling each other out like the $\text{PM}_{|\boldsymbol{\nabla} I^s \cdot \hat{i\boldsymbol{m}}|}$ potential did. As can be seen from the figure, the north and south neighbours are downweighted due to their differing intensity, while the east and west neighbours retain a higher weight (as does the south-east neighbour). This MRF downweights the north and south neighbours sufficiently that GM is correctly chosen for the centre voxel. We note that at this point of the algorithm, the west neighbour in the $\text{PM}_{|y^s_m - y^s_i|}$ potential is already GM while it is CSF for all the other potentials, which further increases the MRF probability for GM over CSF. However, this west voxel has itself been classified as GM rather than CSF for the same reasons just explained. It is difficult to isolate an instance where only a single voxel has been affected by the choice of MRF, due to the spatial dependence of the MRF. If one voxel is affected differently by the different MRFs, this flows on to the neighbours.

In summary, the $\text{PM}_{|y^s_m - y^s_i|}$ potential is able to detect edges by approximating the directional gradient with a forward difference. Using this, it can successfully preserve thin features that the single-beta MRF might otherwise smooth. The $\text{PM}_{|\boldsymbol{\nabla} I^s \cdot \hat{i\boldsymbol{m}}|}$ potential performs poorly on thin features that are local maxima in one direction. The PM_{iso} potential is not directionally isotropic, and can only disable smoothing entirely in regions of high intensity change, rather tangent to the direction of that change.

5.6.2 Parameter values

From table 5.1 it can be seen that the anisotropic MRFs have higher estimated β than the single-beta MRF. The weights w_{im} are identically 1 for the single-beta MRF, and in general less than 1 for the anisotropic MRFs. The "effective" number of neighbours for each voxel is thus lower for an anisotropic MRF than for the single-beta MRF. On average, the estimated β is increased to compensate (compared to the single-beta MRF).

The estimated β for the PM_{iso} potential is extremely high. This is because this MRF essentially "turns off" all neighbours in a neighbourhoods where the gradient magnitude is large ($w_{im} \approx 0 \; \forall m$ for such i). The only remaining neighbourhoods are those where the gradient magnitude is locally small. By definition, this means that the intensities are locally very similar, and hence the tissue labels are likely to be all of one tissue in such neighbourhoods. The maximum-

likelihood β for a uniform (one-label) image is ∞. This explains why the estimated β is so much higher for the PM_{iso} potential - it is presented an almost uniform image over which to maximise β, with non-uniform neighbourhoods being disabled by the weights. On the other hand, the other anisotropic MRFs treat each neighbour differently and hence do not disable entire neighbourhoods, avoiding this. However, the high β value for PM_{iso} does not mean that the resulting segmentations are oversmoothed as they would be on a single-beta MRF, due to the low weights w_{im} balancing out high values of β in neighbourhoods with strong edges. This can be seen in figure 5.10 - the PM_{iso} segmentations are not drastically smoother than the others.

5.6.3 The intensity normalisation parameter κ

We briefly add to our previous discussion of κ. In the example shown in figure 5.11, the weights were all close to 1, so the potentials (PM_{iso} and $\text{PM}_{|\boldsymbol{\nabla} I^s \cdot \hat{\boldsymbol{im}}|}$) did not act very differently from the single-beta potential. We know that the lower κ is, the more sensitive the weight function is to the image gradient, and the more harshly it will penalise "large" gradients. It could be that the κ chosen, the median absolute deviation of the image gradient magnitude, was too large for the PM_{iso} and $\text{PM}_{|\boldsymbol{\nabla} I^s \cdot \hat{\boldsymbol{im}}|}$ potentials.

Anecdotally, it was noticed that decreasing κ (which decreases w_{im}) led to an increase in the estimated β values. By similar reasoning to why the anisotropic MRFs generally have higher β than the single-beta MRF, it is likely that this was to preserve the average potential value. As κ decreases, so do the neighbour weights (for a fixed intensity difference); β may increase so that the value of W_{im} remains roughly constant. This does not necessarily mean that the accuracy is unchanged by adjusting κ; κ controls the strength of the anisotropy in the neighbourhood, while β controls the strength of the smoothing once the smoothing directions have been determined. The parameter values affect how the MRF is applied locally, even if the global energy remains similar.

5.6.4 Alternate anisotropic schemes

The potentials presented so far are all based off the functions used in Perona-Malik diffusion. An alternative is to make use of anisotropic diffusion as developed by Weickert (Weickert, 1998). This makes use of the image *structure matrix*, a discretised and regularised version of $\boldsymbol{\nabla} I^T \boldsymbol{\nabla} I$. The eigenvalues and eigenvectors of this matrix define an ellipsoid that summarise the local gradient distribution of the image. They can be used to detect intensity jumps as with the image gradient. However, the eigenvalues can also be used to distinguish between thin tubular structures, edges, plates, and structureless noise. This provides much more information than the image gradient. Frangi's vesselness filter (Frangi et al., 1998) previously mentioned makes use of the structure matrix to detect veins, being thin tubular structures.

Weickert transforms the structure matrix eigenvalues to adjust the strength of diffusion in each of the eigenvector directions anisotropically, creating a *diffusion matrix*. The PDE (5.3) is solved with this diffusion matrix as $\boldsymbol{D}$. How to incorporate this matrix in to the MRF with the local tissue labels, image intensities and neighbour direction is not immediately clear. One possibility is given by Grau et al. (2006) who uses

$$w_{im} = \hat{\boldsymbol{im}}^T \boldsymbol{D} \hat{\boldsymbol{im}},$$

where $\boldsymbol{D}$ has been constructed to detect thin cylinders (veins) and smooth along but not across them.

5.7 Conclusion

We have presented a framework for replacing the standard single-beta MRF with one that uses local image characteristics to smooth anisotropically rather than isotropically. We believe that this method could be useful in preserving thin features that would otherwise be smoothed, for example CSF or WM between the cortical folds.

Our primary contribution is to thoroughly define the specific properties that must be satisfied by the MRF potential in order to achieve these aims (properties 1 and 2 described previously), within that framework. We investigated 3 different options for choice of anisotropic potential based on the Perona-Malik diffusion function, that allow these properties to be achieved. These use the image gradient to detect the presence and direction of edges locally and incorporate this into the model to smooth along image edges and not across them. The $\mathrm{PM}_{|y_m^s - y_i^s|}$ MRF potential was the most successful in this goal and produced more accurate segmentations than the standard single-beta MRF. It achieved the aim of retaining only a single β parameter, but allowing its strength to vary across the image without being heavily driven by the current segmentation as we found in the multi-beta MRFs of Chapter 4.

Although the $\mathrm{PM}_{|y_m^s - y_i^s|}$ MRF has been seen in similar works (Wels, 2010; Boykov and Funka-Lea, 2006), these fix the MRF spatial regularisation parameter β arbitrarily, or omit it from the MRF potential. The second contribution of our work is to incorporate estimation of β. The advantage of this is to avoid problems with under- or over-specifying β as was found in Chapter 3.

Although this work shows promising improvements in segmentation accuracy, there are still a number of areas in which it could be improved. First, the "MRF" as defined does not form a valid MRF nor prior probability over $\boldsymbol{z}$ (see section 5.3.3.2); we have outlined how this may be rectified by incorporation of a dependence between intensities of neighbouring voxels.

Second, the intensity normalisation parameter κ can be adjusted to increase the sensitivity of the MRF to the image features. We set it to the median absolute deviation of the gradient magnitude following Black et al. (1998), but perhaps it could instead be estimated along with β.

Third, the MRF itself does not need to take the form of the Perona-Malik functions, and may be further tailored. For example, if an anatomical atlas were available, it could be incorporated into the unary potential $W_i(\boldsymbol{z}_i)$ which we omitted. Another option is to use the image structure matrix in place of the image gradient to detect edges with greater specificity, e.g. surface that form a tube in three dimensions as opposed to a plane; see section 5.6.4.

Chapter 6

Conclusion

6.1 Summary and findings

This thesis has focused on fully-automatic segmentation of brain MRI using a mixture-MRF model of the image intensities and spatial distribution. In particular, we have studied the MRF component of the model, which can be used to incorporate spatial dependence between neighbouring voxels, having the effect of smoothing the segmentation. We have used maximum pseudolikelihood estimation as a means to automatically determine the MRF parameters rather than needing to manually specify them as is standard. We began with the simplified Potts model as it is commonly used as the MRF component for tissue segmentation and showed how to use MPL estimation to determine its spatial regularisation parameter β. We then studied more advanced MRFs that allowed for more specific control of the smoothing, and applied MPLE to these.

In Chapter 2 we covered the mixture-MRF image model and showed how it may be solved for the tissue intensity parameters and optimal segmentation using Expectation-Maximisation, with fixed MRF parameters.

6.1.1 Homogeneous Potts MRF

In Chapter 3, we focused on the homogeneous Potts MRF (also called the "single-beta MRF") which is very commonly used for tissue segmentation in several software tools used in neuroimaging (FAST (Zhang et al., 2001), Atropos (Avants et al., 2011), NiftySeg (Cardoso et al., 2009), EMS (Van Leemput et al., 1999b)). Its smoothing parameter β is typically fixed to a value that has been chosen by the developers of the software. It is very rare for β to be determined in a full-automatic manner, and methods for this usually require appropriate training data. We proposed use of maximum pseudolikelihood estimation to automatically determine the smoothing parameter β based on the input image data only. This has only rarely been done

before in brain MRI segmentation (Forbes et al. (2013) with related application papers (Maggia et al., 2016; Kabir et al., 2007; Menze et al., 2015)), and in none of these was the suitability of MPLE over other methods of estimation studied in detail. The least-squares estimator (LSE) was also studied, being the only other previously published method we are aware of in which β was automatically estimated on a per-image basis without requiring training data. We showed that estimating β using MPLE requires only univariate maximisation of a concave function, which itself was already computed even when β is fixed. This means that implementation of MPLE into existing methods is only a small added computational and coding burden to existing software.

A detailed study was performed using several variants of MPLE, comprising choice of approximation (mean-field or pseudolikelihood) with neighbourhood size (6, 18, or 26). As far as we are aware, a study specifically on MRF configuration choices such as these for brain segmentation has not been presented before. Through experiments on real brain MRI it was found that the pseudolikelihood approximation was surprisingly better than the mean-field approximation. This was unexpected, as the pseudolikelihood loses probabilistic information by thresholding probabilities to obtain labels. It was also found that 6 neighbours was sufficient, and that extra neighbours sometimes caused over-smoothing in the MRF, possibly due to the approximate E-step.

The relationship between segmentation accuracy and fixed β was explored to determine the value of estimation. By performing a grid-search over fixed β values, it was shown that choosing β too low can cause significant losses in accuracy. Setting β too high can result in an over-smooth segmentation, though there is an upper bound to β beyond which the segmentation will not change, having reached a state in which all voxels labels are equal to their majority neighbourhood label. The range of "acceptable" β values and the value attaining the highest segmentation accuracy differed for each image, demonstrating that a single fixed β is not generally appropriate for all images. By contrast, estimation of β (by either MPLE or LSE) was able to select a reasonable value for each image, near the maximum accuracy.

Segmentations produced using MPLE to determine β were compared to segmentations using popular fixed β values and also those with estimation using LS estimation. Estimation produced more accurate segmentations than fixed β, sometimes significantly so. Comparison of MPLE with LSE showed that neither estimator produced segmentations significantly more accurate than the other. However, LSE requires computation of neighbourhood frequencies, which can be prohibitively expensive as neighbourhood size increases. It is also less applicable than MPLE, as it cannot take the intensity distribution into account and cannot be used with the mean-field approximation. It relies heavily on the nature of the current segmentation in its construction, and may counterintuitively prevent common neighbourhoods (e.g. homogeneous ones) from contributing to the β estimate if they do not occur with multiple different labels in the centre. For these reasons, we find that MPLE is better suited for MRF estimation in MRI segmentation.

6.1.2 Non-homogeneous Potts MRF

In chapter 4, we studied different forms of the full Potts MRF for use as a prior in mixture-MRF brain MRI segmentation, as an alternative to the simplified single-beta Potts MRF. This enabled finer control of the smoothing applied by the MRF, on a tissue-specific level. Additionally, incorporation of unary parameters was thought to adjust for imbalanced tissue proportions allowing interpretation of the MRF as a tissue prior multiplied by a pairwise MRF for spatial regularity.

Three new MRFs were proposed to replace the single-beta Potts MRF: alpha-multi-beta with multiple unary and pairwise (smoothing) parameters, multi-beta with just multiple smoothing parameters, and alpha-single-beta, with multiple unary and only a single smoothing parameter. It was also proposed to constrain the unary parameters such they matched current tissue portions rather than being free.

We showed how maximum pseudolikelihood estimation could be applied to these more complex MRFs. The MPL estimator retains its desirable features: its Hessian is negative semi-definite, so any local maximum in $\boldsymbol{\Psi}$ for a given segmentation $\boldsymbol{z}$ is also a global maximum (though possibly not unique). The gradient of the Q-function was computed and shown to be simply calculable by a small number of matrix multiplications and additions. We also derived the corresponding least-squares estimates, building on existing work of Van Leemput et al. (1999b).

We found that LSE was not suitable for use with multiple smoothing parameters, as it was too dependent on the occurrence of specific neighbourhoods in order for its underlying system of equations to be defined. In particular, the rarity of different neighbourhoods containing CSF and WM neighbouring each other meant that LSE often could not find an estimate for the corresponding CSF-WM smoothing parameter.

The proposed MRFs — with no or multiple unary parameters, either free or constrained to tissue proportions, and with single or multiple smoothing parameters — were used to segment real brain MRI using maximum pseudolikelihood estimation to automatically determine the parameters. The alpha-multi-beta and alpha-single-beta MRFs with unary potentials constrained to tissue proportions produced the most accurate segmentations, and the same MRFs with unconstrained unary potentials produced the least accurate segmentations. However, none of the MRFs achieved significantly more accurate segmentations than the previously-used single-beta MRF.

We investigated this further by calculating the tissue proportions from the segmentations with unary potentials constrained to the tissue proportions, to see if they were effective in controlling these proportions. It was found that segmentations with unary potentials fixed to tissue proportions had different tissue proportions to the other MRFs with no or unconstrained unary potentials, more closely matching the true CSF proportion. However, they produced too much grey matter. On visual inspection it was seen that the grey matter was over-smooth. Hence

despite achieving slightly higher accuracy than the single-beta MRF, they are not recommended. Rather, they have shown that constraints on the unary potentials can affect the segmentations, but may require external anatomical information for the constraints to be useful.

A similar conclusion was drawn when multiple smoothing parameters were used. Although they may be beneficial in allowing more specific smoothing, the reliance of the E-step on the current segmentation and subsequent feedback loop with the M-step biases them too much when unconstrained. In particular, it was found that the generally low occurrence of CSF and WM in the brain drove the corresponding parameter estimate higher, which in turn further prohibited CSF and WM from occurring (though this is plausible in a healthy brain). This resulted in an artificial shell of GM separating CSF from WM in the segmentations.

In conclusion, we found that the fully-parameterised Potts MRF may be used to further tailor the tissue prior to brain segmentation, but requires use of specific anatomical knowledge to impose constraints on the parameters. In the absence of these, the single-beta MRF should be used. This conclusion has been hinted at in other publications (e.g. Maggia et al. (2016) presents the full Potts MRF but only uses the single-beta MRF), but has not previously been explicitly demonstrated.

6.1.3 Locally anisotropic models

In chapter 5, we presented a framework for incorporating local image features into the single-beta MRF such that the MRF was anisotropic. We proposed three models, based on Perona-Malik anisotropic diffusion, to demonstrate this framework that use various approximations to the local image gradient in order to smooth along edges but not across them. The $\mathrm{PM}_{|y_m^s - y_i^s|}$ model used a forward difference approximation for the directional derivative, while the $\mathrm{PM}_{|\boldsymbol{\nabla} I^s \cdot \hat{\boldsymbol{im}}|}$ model used the directional derivative itself, and the $\mathrm{PM}_{\mathrm{iso}}$ model used the gradient magnitude. These models have a single β parameter, avoiding the problems of the previous chapter, yet still enable local smoothing, unlike the single-beta MRF. The parameter β is estimated with maximum pseudolikelihood. It was expected that the $\mathrm{PM}_{|y_m^s - y_i^s|}$ model would be most sensitive to presence of local edges as it did not suffer from the symmetry problems of the $\mathrm{PM}_{|\boldsymbol{\nabla} I^s \cdot \hat{\boldsymbol{im}}|}$ and $\mathrm{PM}_{\mathrm{iso}}$ models. We found the $\mathrm{PM}_{|y_m^s - y_i^s|}$ model to produce significantly more accurate segmentations than the standard isotropic single-beta MRF, particularly when preserving thin features. The $\mathrm{PM}_{|\boldsymbol{\nabla} I^s \cdot \hat{\boldsymbol{im}}|}$ and $\mathrm{PM}_{\mathrm{iso}}$ models produced significantly *less* accurate segmentations than the single-beta MRF, showing that the anisotropic potential does indeed have an effect, and must be carefully designed.

6.2 Contributions

The novel contributions of this thesis are

- to undertake a study of parameter estimation in Markov random fields for image segmentation,
- to provide a recipe for use of the maximum pseudolikelihood estimator in the EM algorithm and demonstrating how it is particularly suitable for incorporation into existing fixed-parameter algorithms,
- to perform an explicit comparison of MRF design choices (neighbourhood size and MRF approximation) of the standard homogeneous Potts MRF with regards to segmentation accuracy,
- to make a detailed comparison of various forms of the non-homogeneous Potts MRF with MPL estimation (comparing it to the homogeneous Potts MRF) and apply this to brain MRI segmentation,
- to develop anisotropic MRFs suited to preserving fine structures with parameter estimation via maximum pseudolikelihood estimation to them, and apply these to brain MRI segmentation.

We acknowledge that the thesis does not aim to provide a full pipeline for brain segmentation (e.g. including brain extraction, bias field correction or registration), but rather to provide modifications and improvements to the core mixture-MRF model that is used as a basis in many of these pipelines. Specifically, we have shown how maximum pseudolikelihood estimation may readily be added to any existing method that uses the mixture-Potts MRF. We have also made numerous arguments as to why this should be done. We have studied the effects of various configuration details (neighbourhood size, MRF approximation) specifically with respect to brain segmentation which has not been detailed elsewhere.

The full Potts MRF has additional parameters that need to be specified compared to the single-beta MRF. Since parameter estimation is not often done in segmentation, this means that the full Potts MRF is rarely used; it is sometimes mentioned but typically reduced to the single-beta MRF when applied. We have demonstrated how the full Potts MRF may be used with maximum pseudolikelihood to specify the parameters and undertaken a detailed study of how the various unary and pairwise parameters apply to brain segmentation. We found that the full Potts MRF is not useful with unconstrained parameters compared to the homogeneous Potts MRF. Though this has been hinted at in other papers (e.g. Maggia et al. (2016)), our work provides an explicit demonstration supporting the claim. We have also given suggestions as to how the parameters may be constrained using anatomical information and using maximum pseudolikelihood to determine the specific values.

We have shown how to construct anisotropic MRF potentials using local image features and how to estimate their parameters with maximum pseudolikelihood. While anisotropic MRFs have been used previously for tissue segmentation, these are discriminative models that require training to learn the dependence of the MRF on the image features. By drawing on Perona-Malik diffusion we have provided a generative model whose parameters may be estimated without the use of training data.

6.3 Future work

There are a number of avenues to take to improve the accuracy of the image model used in the thesis; the work of a researcher is never done. The model consists of two components: an MRF over the tissue labels, and an intensity distribution given those labels. This thesis has very much focused on the MRF component. There is also much scope to study how the intensity distribution might be adjusted to better model the image.

6.3.1 Markov random field

The main limitation of maximum pseudolikelihood estimation was that the E-step of the EM algorithm required the current estimate of the segmentation in order to approximate the expectations. Thus the segmentation was not fully marginalised out, leading to an undesirable dependence and feedback loop between C and M steps. The MRF parameters in particular were strongly influenced by the current segmentation, which allowed them to feed back into the next segmentation. This was observed in the multi-beta MRFs of chapter 4 in the CSF-WM boundary. It also tends to result in slightly over-smooth segmentations, since segmentations are mostly homogeneous which favours high β estimates. This dependence also means that the algorithm requires a good initial starting segmentation.

Future work could focus on removing or lessening this dependence in the E-step. Celeux et al. (2003) favoured a 'simulated random fields' variant of the EM algorithm. Here, the C-step did not use ICM or mean-field updates to determine $\boldsymbol{z}$, but instead simulated $\boldsymbol{z}$ using one cycle of Gibbs sampling from the current distribution. This allowed the $\boldsymbol{z}$ flexibility to escape local minima. We experimented with this but found the posterior distribution did not allow the simulated $\boldsymbol{z}$ to change much from what would have been obtained with ICM. However, perhaps multiple cycles of Gibbs sampling or use of MRF-specific algorithms such as those by Swendsen and Wang (1987) or Wolff (1989) could be used here. The caveat is that the more sophisticated the simulation method, the greater the computation time.

In terms of the MRFs themselves, in chapter 4 we studied the fully-parameterised Potts MRF to allow finer control of tissue smoothing. These showed promise, but ultimately it was concluded that the parameters need to be constrained with context-specific knowledge in order to have greatest effect. We attempted to incorporate such constraints in the form of setting the unary parameters to match log-tissue proportions as for a standard mixture model, but while this helped to address over-segmentation of CSF, it compensated by over-segmenting GM instead. Future work could study the form of the constraints in greater detail, and whether an anatomical atlas is absolutely required to impose them or if they may be driven by the data. It could be that setting some β_{jk} to be equal or bounded by each other and likewise with α_j could prove beneficial.

In chapter 5 we discovered our most promising MRF, the $\mathrm{PM}_{|y_m^s - y_i^s|}$ MRF. However as we noted, this is not truly a MRF $p(\boldsymbol{z}_i|\boldsymbol{z}_{\partial i})$ but rather a distribution described by $p(\boldsymbol{z}_i|\boldsymbol{z}_{\partial i}, \boldsymbol{y}_{\partial i})$, and it is not clear that this leads to a valid joint distribution of $\boldsymbol{z}$ and $\boldsymbol{y}$. The basis of the EM algorithm is the Q function being the expectation of the joint distribution. Use of EM for the image model of chapter 5 may not have been valid. If the probabilities could be rectified to form a valid joint distribution, the corresponding pseudolikelihood and EM algorithm could then be applied appropriately. We have given a few suggestions in that chapter as to how this might be achieved. However, treating this model and corresponding algorithms as approximate, it led to better practical results than any other method considered.

Also, the anisotropic MRFs studied in chapter 5 were fairly basic; classical Perona-Malik diffusion differs in strength at each voxel but is isotropic at each voxel. It is only because we introduced a forward approximation to the image gradient that we obtained directional anisotropy in the $\mathrm{PM}_{|y_m^s - y_i^s|}$ potential. An extension of these MRFs would be to use "true" anisotropic diffusion in the style of Weickert (1998) seems very promising as it would make use of the image structure matrix rather than the image gradient.

6.3.2 Intensity distribution

The intensity component assumes that the tissue intensities are normally distributed given the tissue label. However, this is only an approximation. The intensity distribution for a given distribution is not the same across the brain. For example, it is known that sub-cortical grey-matter has intensity closely matching that of white-matter (Pohl et al., 2005); for this reason, none of our segmentations were able to accurately segment these regions. In this case, an atlas may be used to guide the segmentation. Other works have addressed this by using a mixture of multiple normal distributions per tissue (Ashburner and Friston, 2005). One further possibility is to allow the means of the normal distributions to themselves vary slowly across the image, explicitly incorporating the inhomogeneity into the image model (Pohl et al., 2004).

The single-beta Potts MRF is limited in that in only allows for gross control of the smoothness in the segmentation, offering only one parameter to adjust this. An example of this was the difficulty in classifying partial volume voxels, especially those that fell on the boundary of CSF and WM thus having the intensity of GM. The use of a mixture model in combination with an MRF should have enabled these voxels to be correctly classified. Given the neighbours of such a voxel, the intensity distribution is a standard normal mixture with mixing proportions determined by the MRF probabilities evaluated a that voxel. These mixing proportions should favour CSF and WM over GM as they occur in the neighbourhood while GM does not. However, it appears they were not strong enough to overcome the difference in intensity probability. Having β higher can increase the importance of the spatial information to override the intensity density, but would then cause over-smoothing in the rest of the segmentation. An alternative way to deal with this problem could be to explicitly incorporate partial volume effects into the

image model itself. Shattuck et al. (2001) used additional tissue classes for combined tissues, e.g. "CSF/GM", with probability derived by integrating normal densities over linear combinations of the corresponding means:

$$f(y_i|\text{CSF/GM}) = \int_0^1 \phi(y_i; s\mu_{\text{CSF}} + (1-s)\mu_{\text{GM}}, \sigma^2)\,\mathrm{d}s.$$

Another options is to allow the tissue labels $\boldsymbol{z}_i$ to be continuous rather than discrete. Rather than consisting of $g-1$ 0s and one 1, they are permitted to be in the range $[0,1]$ and sum to 1. In a standard mixture, each voxel may have only one label. With this continuous version, each voxel may be fractionally composed of different labels. The difference is that in the former, the marginal $f(y_i)$ is a linear combination of normal densities but $f(y_i|\boldsymbol{z}_i)$ is a single normal density, while in the latter, $f(y_i|\boldsymbol{z}_i)$ is also a linear combination of normal densities. The MRF is then also be modified, typically changing the $\boldsymbol{z}_i^T\boldsymbol{z}_m$ term to $|\boldsymbol{z}_i - \boldsymbol{z}_m|^2$. This approach is demonstrated in (Choi et al., 1991; Nocera and Gee, 1997; Roche and Forbes, 2014), amongst others.

Throughout this thesis we have assumed that the image intensity at each voxel y_i is scalar. However, the existence of many sequences to highlight various tissues in MRI means that there are often multiple images available. For example, T2 provides good contrast for CSF and when used together with T1, may improve overall segmentation accuracy. The techniques developed in this these are all valid for vector $\boldsymbol{y}_i$, and this is an avenue worth investigating. The only change required is to replace the univariate Gaussian intensity distributions with multivariate ones; the corresponding update equations for EM are in Appendix A. However, additional challenges are encountered as the images must be well registered.

Finally, we have assumed that the voxel intensities are conditionally independent given their labels, which are allowed to be dependent. This cannot account for intensity blurring, where the signal recorded at voxel i is contaminated by the signals at neighbouring tissues. Incorporating conditional dependence of observed intensities can be achieved using an MRF, for example an autoregressive model. This is quite common in the field of satellite image classification, and techniques used in that field might be adapted for MRI segmentation.

6.4 Final remarks

Ultimately, I have shown the importance of automatically determining parameter values in MRFs for image segmentation. I have championed the use of maximum pseudolikelihood estimation for its simplicity and ease of implementation into EM. The ability to automatically determine parameter values has enabled use of more sophisticated MRFs (with more parameters) to improve the specificity of the MRF to brain tissue segmentation. I have conducted a detailed study into which forms of MRF work better than others when no training data or anatomical atlases are available, and of modelling choices such as neighbourhood structure and MRF approximation.

Some proposed MRFs performed better than others (notably the anisotropic MRFs), but I hope the work in this thesis will be used as a detailed record of successes, failures and suggestions to aid future research in the area.

Bibliography

Ackley, D. H., Hinton, G. E., and Sejnowski, T. J. (1985). A learning algorithm for Boltzmann machines. *Cognitive science*, 9(1):147–169.

Ahmed, M. N., Yamany, S. M., Mohamed, N., Farag, A. A., and Moriarty, T. (2002). A modified fuzzy c-means algorithm for bias field estimation and segmentation of MRI data. *IEEE Transactions on Medical Imaging*, 21(3):193–199.

Archer, G. E. B. and Titterington, D. M. (2002). Parameter estimation for hidden Markov chains. *Journal of Statistical Planning and Inference*, 108(1):365–390.

Ashburner, J. and Friston, K. (1997). Multimodal Image Coregistration and Partitioning—A Unified Framework. *NeuroImage*, 6(3):209–217.

Ashburner, J. and Friston, K. J. (2005). Unified segmentation. *NeuroImage*, 26(3):839–851.

Avants, B. B., Tustison, N. J., Wu, J., Cook, P. A., and Gee, J. C. (2011). An Open Source Multivariate Framework for n-Tissue Segmentation with Evaluation on Public Data. *Neuroinformatics*, 9(4):381–400.

Balafar, M. A., Ramli, A. R., Saripan, M. I., and Mashohor, S. (2010). Review of brain MRI image segmentation methods. *Artificial Intelligence Review*, 33(3):261–274.

Bériault, S., Archambault-Wallenburg, M., Sadikot, A. F., Collins, D. L., and Pike, G. B. (2013). Automatic Markov Random Field Segmentation of Susceptibility-Weighted MR Venography. In *Clinical Image-Based Procedures. Translational Research in Medical Imaging*, pages 39–47. Springer, Cham.

Berkson, J. B. (1949). Minimum X2 and Maximum Likelihood Solution in Terms of a Linear Transform, with Particular Reference to Bio-Assay. *Journal of the American Statistical Association*, 44(246):273–278.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society B*, pages 192–236.

Besag, J. (1975). Statistical Analysis of Non-Lattice Data. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 24(3):179–195.

Besag, J. (1986). On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society B*, 48:259–302.

Black, M. J., Sapiro, G., Marimont, D. H., and Heeger, D. (1998). Robust anisotropic diffusion. *IEEE Transactions on Image Processing*, 7(3):421–432.

Blake, A., Rother, C., Brown, M., Pérez, P., and Torr, P. (2004). Interactive Image Segmentation Using an Adaptive GMMRF Model. In *Eur. Conf. Comput. Vis.*, volume 3, pages 428–441.

Boccardi, M., Ganzola, R., Bocchetta, M., Pievani, M., Redolfi, A., Bartzokis, G., Camicioli, R., Csernansky, J. G., de Leon, M. J., deToledo-Morrell, L., Killiany, R. J., Lehéricy, S., Pantel, J., Pruessner, J. C., Soininen, H., Watson, C., Duchesne, S., Jack, C. R., and Frisoni, G. B. (2011). Survey of protocols for the manual segmentation of the hippocampus: Preparatory steps towards a joint EADC-ADNI harmonized protocol. *Journal of Alzheimer's disease: JAD*, 26 Suppl 3:61–75.

Borges, C. F. (1999). On the estimation of Markov random field parameters. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(3):216–224.

Boyd, R., George, J., Fripp, J., Panneck, K., Chan, A., Fiori, S., Guzzetta, A., Ware, R., Rose, S., and Colditz, P. (2015). Relationship between early brain structure on Mri, white matter integrity (diffusion Mri) and neurological function at 30 weeks post menstrual age in infants born very preterm. *Developmental Medicine & Child Neurology*, 57:8–9.

Boykov, Y. and Funka-Lea, G. (2006). Graph Cuts and Efficient N-D Image Segmentation. *International Journal of Computer Vision*, 70(2):109–131.

Boykov, Y. Y. and Jolly, M. P. (2001). Interactive graph cuts for optimal boundary amp; region segmentation of objects in N-D images. In *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, volume 1, pages 105–112 vol.1.

Cardoso, M. J., Clarkson, M. J., Ridgway, G. R., Modat, M., Fox, N. C., and Ourselin, S. (2011). LoAd: A locally adaptive cortical segmentation algorithm. *NeuroImage*, 56(3):1386–1397.

Cardoso, M. J. et al. (2009). NiftySeg (version 0.9.4). University College London. `seg_EM` program. Available at https://sourceforge.net/projects/niftyseg/.

Celeux, G. and Diebolt, J. (1985). The SEM algorithm: A probabilistic teacher algorithm derived from the EM algorithm for the mixture problem. *Computational statistics quarterly*, 2(1):73–82.

Celeux, G., Forbes, F., and Peyrard, N. (2003). EM procedures using mean field-like approximations for Markov model-based image segmentation. *Pattern Recognition*, 36(1):131–144.

Celeux, G. and Govaert, G. (1992). A classification EM algorithm for clustering and two stochastic versions. *Computational statistics & Data analysis*, 14(3):315–332.

Chaari, L., Vincent, T., Forbes, F., Dojat, M., and Ciuciu, P. (2013). Fast Joint Detection-Estimation of Evoked Brain Activity in Event-Related fMRI Using a Variational Approach. *IEEE Transactions on Medical Imaging*, 32(5):821–837.

Chan, A., Wood, I. A., and Fripp, J. (2016). Maximum Pseudolikelihood Estimation for Mixture-Markov Random Field Segmentation of the Brain. In *2016 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, pages 1–7. IEEE.

Chandler, D. (1987). *Introduction to Modern Statistical Mechanics*. Oxford University Press, New York.

Choi, H. S., Haynor, D. R., and Kim, Y. (1991). Partial volume tissue classification of multichannel magnetic resonance images-a mixel model. *IEEE Transactions on Medical Imaging*, 10(3):395–407.

Clarke, L. P., Velthuizen, R. P., Camacho, M. A., Heine, J. J., Vaidyanathan, M., Hall, L. O., Thatcher, R. W., and Silbiger, M. L. (1995). MRI segmentation: Methods and applications. *Magnetic Resonance Imaging*, 13(3):343–368.

Cocosco, C. A., Zijdenbos, A. P., and Evans, A. C. (2003). A fully automatic and robust brain MRI tissue classification method. *Medical Image Analysis*, 7(4):513–527.

Collier, D. C., Burnett, S. S. C., Amin, M., Bilton, S., Brooks, C., Ryan, A., Roniger, D., Tran, D., and Starkschall, G. (2003). Assessment of consistency in contouring of normal-tissue anatomic structures. *Journal of Applied Clinical Medical Physics*, 4(1):17–24.

Comets, F. and Gidas, B. (1992). Parameter Estimation for Gibbs Distributions from Partially Observed Data. *The Annals of Applied Probability*, 2(1):142–170.

Dawid, A. P. (1980). Conditional Independence for Statistical Operations. *The Annals of Statistics*, 8(3):598–617.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B*, pages 1–38.

Dennis, E. L. and Thompson, P. M. (2013). Typical and atypical brain development: A review of neuroimaging studies. *Dialogues in Clinical Neuroscience*, 15(3):359–384.

Deriche, R. (1993). Recursively implementating the Gaussian and its derivatives. report, INRIA.

Derin, H. and Elliott, H. (1987). Modeling and Segmentation of Noisy and Textured Images Using Gibbs Random Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-9(1):39–55.

Despotović, I., Goossens, B., and Philips, W. (2015). MRI segmentation of the human brain: Challenges, methods, and applications. *Computational and mathematical methods in medicine*, 2015.

Dice, L. R. (1945). Measures of the Amount of Ecologic Association Between Species. *Ecology*, 26(3):297–302.

Dubois, B., Feldman, H. H., Jacova, C., Dekosky, S. T., Barberger-Gateau, P., Cummings, J., Delacourte, A., Galasko, D., Gauthier, S., Jicha, G., Meguro, K., O'brien, J., Pasquier, F., Robert, P., Rossor, M., Salloway, S., Stern, Y., Visser, P. J., and Scheltens, P. (2007). Research criteria for the diagnosis of Alzheimer's disease: Revising the NINCDS-ADRDA criteria. *The Lancet. Neurology*, 6(8):734–746.

Dunmur, A. P. and Titterington, D. M. (1998). Mean fields and two-dimensional Markov random fields in image analysis. *Pattern Analysis and Applications*, 1(4):248–260.

Fischl, B., Salat, D. H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., van der Kouwe, A., Killiany, R., Kennedy, D., Klaveness, S., Montillo, A., Makris, N., Rosen, B., and Dale, A. M. (2002). Whole Brain Segmentation: Automated Labeling of Neuroanatomical Structures in the Human Brain. *Neuron*, 33(3):341–355.

Forbes, F., Charras-Garrido, M., Azizi, L., Doyle, S., and Abrial, D. (2013). SPATIAL RISK MAPPING FOR RARE DISEASE WITH HIDDEN MARKOV FIELDS AND VARIATIONAL EM. *The Annals of Applied Statistics*, 7(2):1192–1216.

Forbes, F. and Peyrard, N. (2003). Hidden Markov random field model selection criteria based on mean field-like approximations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(9):1089–1101.

Frangi, A. F., Niessen, W. J., Vincken, K. L., and Viergever, M. A. (1998). Multiscale vessel enhancement filtering. In *Medical Image Computing and Computer-Assisted Intervention — MICCAI'98*, pages 130–137. Springer, Berlin, Heidelberg.

Gao, X. and Song, P. X.-K. (2011). Composite likelihood EM algorithm with applications to multivariate hidden Markov model. *Statistica Sinica*, pages 165–185.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:721–741.

Geman, S. and Graffigne, C. (1986). Markov random field image models and their applications to computer vision. In *Proceedings of the International Congress of Mathematicians*, volume 1, pages 1496–1517.

George, J., Fripp, J., Shen, K., Pannek, K., Chan, A., Ware, R., Rose, S., Colditz, P., and Boyd, R. (2015). Relationship between white matter integrity and neurological function in preterm infants at 30 weeks postmenstrual age. *Developmental Medicine & Child Neurology*, 57:88–89.

George, J., Fripp, J., Shen, K., Pannek, K., Chan, A., Ware, R., Rose, S., Colditz, P., and Boyd, R. (2016). Relationship between white matter integrity at 3T Mri and neurological

function in preterm infants at 30 weeks postmenstrual age. *Developmental Medicine & Child Neurology*, 58:33–34.

Grau, V., Downs, J. C., and Burgoyne, C. F. (2006). Segmentation of trabeculated structures using an anisotropic Markov random field: Application to the study of the optic nerve head in glaucoma. *IEEE transactions on medical imaging*, 25(3):245–255.

Gudbjartsson, H. and Patz, S. (1995). The Rician Distribution of Noisy MRI Data. *Magnetic resonance in medicine*, 34(6):910–914.

Gurelli, M. I. (1996). Extension of the modified-histogramming method for multilevel Markov random fields. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 26(1):180–187.

Gurelli, M. I. and Onural, L. (1994). On a parameter estimation method for Gibbs-Markov random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(4):424–430.

Gurleyik, K. and Haacke, E. M. (2002). Quantification of errors in volume measurements of the caudate nucleus using magnetic resonance imaging. *Journal of magnetic resonance imaging: JMRI*, 15(4):353–363.

Guyon, X. and Künsch, H. R. (1992). Asymptotic Comparison of Estimators in the Ising Model. In *SpringerLink*, pages 177–198. Springer New York.

Hammersley, J. and Clifford, P. (1971). Markov fields on finite graphs and lattices.

Han, X., Jovicich, J., Salat, D., van der Kouwe, A., Quinn, B., Czanner, S., Busa, E., Pacheco, J., Albert, M., Killiany, R., Maguire, P., Rosas, D., Makris, N., Dale, A., Dickerson, B., and Fischl, B. (2006). Reliability of MRI-derived measurements of human cerebral cortical thickness: The effects of field strength, scanner upgrade and manufacturer. *NeuroImage*, 32(1):180–194.

Hofmann, T. and Buhmann, J. M. (1997). Pairwise data clustering by deterministic annealing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(1):1–14.

Huang, A., Abugharbieh, R., and Tam, R. (2009). A Hybrid Geometric–Statistical Deformable Model for Automated 3-D Segmentation in Brain MRI. *IEEE transactions on bio-medical engineering*, 56(7):1838–1848.

Ising, E. (1925). Beitrag zur Theorie des Ferromagnetismus. *Zeitschrift für Physik*, 31(1):253–258.

Jack, C. R., Petersen, R. C., Xu, Y., O'brien, P., Smith, G. E., Ivnik, R. J., Boeve, B. F., Tangalos, E. G., and Kokmen, E. (2000). Rates of hippocampal atrophy correlate with change in clinical status in aging and AD. *Neurology*, 55(4):484–490.

Jack, C. R., Petersen, R. C., Xu, Y. C., Waring, S. C., O'brien, P. C., Tangalos, E. G., Smith, G. E., Ivnik, R. J., and Kokmen, E. (1997). Medial temporal atrophy on MRI in normal aging and very mild Alzheimer's disease. *Neurology*, 49(3):786–794.

Jalobeanu, A., Blanc-Féraud, L., and Zerubia, J. (2002). Hyperparameter estimation for satellite image restoration using a MCMC maximum-likelihood method. *Pattern Recognition*, 35(2):341–352.

Ji, C. and Seymour, L. (1996). A consistent model selection procedure for Markov random fields based on penalized pseudolikelihood. *The Annals of Applied Probability*, 6(2):423–443.

Jubb, M. and Jennison, C. (1991). Aggregation and refinement in binary image restoration. In *Institute of Mathematical Statistics Lecture Notes - Monograph Series*, pages 150–162. Institute of Mathematical Statistics, Hayward, CA.

Kabir, Y., Dojat, M., Scherrer, B., Forbes, F., and Garbay, C. (2007). Multimodal MRI segmentation of ischemic stroke lesions. In *2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 1595–1598.

Kashyap, R. and Chellappa, R. (1983). Estimation and choice of neighbors in spatial-interaction models of images. *IEEE Transactions on Information Theory*, 29(1):60–72.

Kay, J. (1986). Contribution to the discussion of paper by J. Besag. *Journal of the Royal Statistical Society B*, 48:293.

Kay, J. and Titterington, D. (1986). Image labelling and the statistical analysis of incomplete data. In *Proc. 2nd Int. Conf. Image Processing and Applications*, pages 44–48.

Koller, D. and Friedman, N. (2009). *Probabilistic Graphical Models: Principles and Techniques*. MIT press.

Kumar, S. and Hebert, M. (2003). Discriminative random fields: A discriminative framework for contextual interaction in classification. In *Proceedings Ninth IEEE International Conference on Computer Vision*, pages 1150–1157 vol.2.

Kumar, S. and Hebert, M. (2006). Discriminative Random Fields. *International Journal of Computer Vision*, 68(2):179–201.

Lafferty, J., McCallum, A., and Pereira, F. C. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

Lai, M., D'Acunto, G., Guzzetta, A., Fripp, J., Chan, A., Rose, S., Ngenda, N., Whittingham, K., Colditz, P., and Boyd, R. (2015). Randomised controlled trial of PREMM: Early somatosensory stimulation (massage) in preterm infants. *Developmental Medicine & Child Neurology*, 57:94–95.

Lakshmanan, S. and Derin, H. (1989). Simultaneous parameter estimation and segmentation of Gibbs random fields using simulated annealing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(8):799–813.

Lauritzen, S. L. (1996). *Graphical Models*, volume 17. Clarendon Press.

Li, S. Z. (2009). *Markov Random Field Modeling in Image Analysis*. Advances in Pattern Recognition. Springer London, London.

Lindsay, B. G. (1988). Composite likelihood methods. *Contemporary mathematics*, 80(1):221–39.

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., van der Laak, J. A. W. M., van Ginneken, B., and Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60–88.

Little, R. J. A. and Rubin, D. B. (1983). On Jointly Estimating Parameters and Missing Data by Maximizing the Complete-Data Likelihood. *The American Statistician*, 37(3):218–220.

Maggia, C., Doyle, S., Forbes, F., Heck, O., Troprès, I., Berthet, C., Teyssier, Y., Velly, L., Payen, J.-F., and Dojat, M. (2016). Assessment of Tissue Injury in Severe Brain Trauma. In Crimi, A., Menze, B., Maier, O., Reyes, M., and Handels, H., editors, *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, Lecture Notes in Computer Science, pages 57–68. Springer International Publishing.

Malik, J., Belongie, S., Leung, T., and Shi, J. (2001). Contour and Texture Analysis for Image Segmentation. *International Journal of Computer Vision*, 43(1):7–27.

Manjunath, B. S. and Chellappa, R. (1991). Unsupervised Texture Segmentation Using Markov Random Field Models. *IEEE Trans. Pattern Anal. Mach. Intell.*, 13(5):478–482.

McLachlan, G. J., Ng, S. K., Galloway, G. J., and Wang, D. (1996). Clustering of magnetic resonance images. *American Statistical Association - 1996 Proceedings of the Statistical Computing Section*, pages 12–17.

McLachlan, G. J. and Peel, D. (2000). *Finite Mixture Models*. Wiley, New York.

Menze, B. H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., Lanczi, L., Gerstner, E., Weber, M., Arbel, T., Avants, B. B., Ayache, N., Buendia, P., Collins, D. L., Cordier, N., Corso, J. J., Criminisi, A., Das, T., Delingette, H., Demiralp, Ç., Durst, C. R., Dojat, M., Doyle, S., Festa, J., Forbes, F., Geremia, E., Glocker, B., Golland, P., Guo, X., Hamamci, A., Iftekharuddin, K. M., Jena, R., John, N. M., Konukoglu, E., Lashkari, D., Mariz, J. A., Meier, R., Pereira, S., Precup, D., Price, S. J., Raviv, T. R., Reza, S. M. S., Ryan, M., Sarikaya, D., Schwartz, L., Shin, H., Shotton, J., Silva, C. A., Sousa, N., Subbanna, N. K., Szekely, G., Taylor, T. J., Thomas, O. M., Tustison, N. J., Unal, G., Vasseur, F., Wintermark, M., Ye, D. H., Zhao, L., Zhao, B., Zikic, D., Prastawa, M., Reyes, M., and Leemput, K. V. (2015). The Multimodal Brain

Tumor Image Segmentation Benchmark (BRATS). *IEEE Transactions on Medical Imaging*, 34(10):1993–2024.

Moeskops, P., Viergever, M. A., Mendrik, A. M., de Vries, L. S., Benders, M. J. N. L., and Išgum, I. (2016). Automatic Segmentation of MR Brain Images With a Convolutional Neural Network. *IEEE Transactions on Medical Imaging*, 35(5):1252–1261.

Nakagawa, S. and Schielzeth, H. (2013). A general and simple method for obtaining R2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, 4(2):133–142.

Nocera, L. and Gee, J. C. (1997). Robust partial-volume tissue classification of cerebral MRI scans. volume 3034, pages 312–322.

Noe, A. and Gee, J. C. (2001). Partial Volume Segmentation of Cerebral MRI Scans with Mixture Model Clustering. In Insana, M. F. and Leahy, R. M., editors, *Information Processing in Medical Imaging*, number 2082 in Lecture Notes in Computer Science, pages 423–430. Springer Berlin Heidelberg.

Owen, A. (1986). Discussion of Ripley's "Statistics, images, and pattern recognition". *Canadian Journal of Statistics*, 14:106–110.

Owen, A. (1989). Image segmentation via iterated conditional expectations. Technical report, Department of Statistics, University of Chicago.

Pagnozzi, A. M., Dowson, N., Bradley, A. P., Boyd, R. N., Bourgeat, P., and Rose, S. (2015). Expectation-Maximization with Image-Weighted Markov Random Fields to Handle Severe Pathology. In *2015 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, pages 1–6.

Pereyra, M., Dobigeon, N., Batatia, H., and Tourneret, J.-Y. (2013). Estimating the granularity coefficient of a Potts-Markov random field within a Markov chain Monte Carlo algorithm. *IEEE transactions on image processing: a publication of the IEEE Signal Processing Society*, 22(6):2385–2397.

Perona, P. and Malik, J. (1990). Scale-space and edge detection using anisotropic diffusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(7):629–639.

Petersen, K. B. and Pedersen, M. S. (2012). The matrix cookbook. http://matrixcookbook.com.

Pohl, K. M., Bouix, S., Kikinis, R., and Grimson, W. E. L. (2004). Anatomical guided segmentation with non-stationary tissue class distributions in an expectation-maximization framework. In *IEEE International Symposium on Biomedical Imaging: Nano to Macro, 2004*, pages 81–84 Vol. 1.

Pohl, K. M., Fisher, J., Levitt, J. J., Shenton, M. E., Kikinis, R., Grimson, W. E. L., and Wells, W. M. (2005). A Unifying Approach to Registration, Segmentation, and Intensity Correction.

In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2005*, Lecture Notes in Computer Science, pages 310–318. Springer, Berlin, Heidelberg.

Possolo, A. (1986). Estimation of binary Markov random fields. *Unpublished manuscript*.

Potts, R. B. (1952). Some generalized order-disorder transformations. *Mathematical Proceedings of the Cambridge Philosophical Society*, 48(1):106–109.

Qian, W. and Titterington, D. (1991). Estimation of parameters in hidden Markov models. *Philosophical Transactions of the Royal Society A*, 337(1647):407–428.

Qian, W. and Titterington, D. M. (1992). Stochastic relaxations and EM algorithms for markov random fields. *Journal of Statistical Computation and Simulation*, 40(1-2):55–69.

Ripley, B. D. (1986). Statistics, images, and pattern recognition. *Canadian Journal of Statistics*, 14(2):83–102.

Roche, A. and Forbes, F. (2014). Partial Volume Estimation in Brain MRI Revisited. In Golland, P., Hata, N., Barillot, C., Hornegger, J., and Howe, R., editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2014*, number 8673 in Lecture Notes in Computer Science, pages 771–778. Springer International Publishing.

Rohlfing, T. (2012). Image Similarity and Tissue Overlaps as Surrogates for Image Registration Accuracy: Widely Used but Unreliable. *IEEE Transactions on Medical Imaging*, 31(2):153–163.

Rousseeuw, P. J. and Leroy, A. M. (2005). *Robust Regression and Outlier Detection*, volume 589. John Wiley & Sons.

Rutherford, M. A. (2002). *MRI of the Neonatal Brain*. W.B. Saunders, London; New York.

Rydén, T. and Titterington, D. M. (1998). Computational Bayesian Analysis of Hidden Markov Models. *Journal of Computational and Graphical Statistics*, 7(2):194–211.

Schuff, N., Woerner, N., Boreta, L., Kornfield, T., Shaw, L., Trojanowski, J., Thompson, P., Jack Jr, C., Weiner, M., and Initiative, A. D. N. (2009). MRI of hippocampal volume loss in early Alzheimer's disease in relation to ApoE genotype and biomarkers. *Brain*, 132(4):1067–1077.

Shattuck, D. W., Sandor-Leahy, S. R., Schaper, K. A., Rottenberg, D. A., and Leahy, R. M. (2001). Magnetic Resonance Image Tissue Classification Using a Partial Volume Model. *NeuroImage*, 13(5):856–876.

Shen, D., Wu, G., and Suk, H.-I. (2017). Deep Learning in Medical Image Analysis. *Annual Review of Biomedical Engineering*, 19(1):221–248.

Song, Z., Tustison, N., Avants, B., and Gee, J. C. (2006). Integrated Graph Cuts for Brain MRI Segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2006*, pages 831–838. Springer, Berlin, Heidelberg.

Stanford, D. C. and Raftery, A. E. (2002). Approximate Bayes factors for image segmentation: The Pseudolikelihood Information Criterion (PLIC). *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(11):1517–1520.

Stoehr, J. (2017). A review on statistical inference methods for discrete Markov random fields. *arXiv:1704.03331 [stat]*.

Swendsen, R. H. and Wang, J.-S. (1987). Nonuniversal critical dynamics in Monte Carlo simulations. *Physical Review Letters*, 58(2):86–88.

The DevTeam (1987). Nethack (versions 3.4.3 and 3.6.0). https://www.nethack.org/.

Thompson, P. M., Hayashi, K. M., De Zubicaray, G., Janke, A. L., Rose, S. E., Semple, J., Herman, D., Hong, M. S., Dittmer, S. S., Doddrell, D. M., and others (2003). Dynamics of gray matter loss in Alzheimer's disease. *Journal of neuroscience*, 23(3):994–1005.

Titterington, D. (1984). Comments on "Application of the Conditional Population-Mixture Model to Image Segmentation". *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, (5):656–658.

Tsai, A., Yezzi, A., and Willsky, A. S. (2001). Curve evolution implementation of the Mumford-Shah functional for image segmentation, denoising, interpolation, and magnification. *IEEE transactions on Image Processing*, 10(8):1169–1186.

Valverde, S., Oliver, A., Cabezas, M., Roura, E., and Lladó, X. (2015). Comparison of 10 brain tissue segmentation methods using revisited IBSR annotations. *Journal of magnetic resonance imaging: JMRI*, 41(1):93–101.

Van Leemput, K. (2001). Expectation-maximization segmentation. Medical Imaging Computing. Available at https://mirc.uzleuven.be/MedicalImageComputing/downloads/ems.php.

Van Leemput, K., Maes, F., Vandermeulen, D., and Suetens, P. (1999a). Automated model-based bias field correction of MR images of the brain. *IEEE Transactions on Medical Imaging*, 18(10):885–896.

Van Leemput, K., Maes, F., Vandermeulen, D., and Suetens, P. (1999b). Automated model-based tissue classification of MR images of the brain. *IEEE Transactions on Medical Imaging*, 18(10):897–908.

Van Leemput, K., Maes, F., Vandermeulen, D., and Suetens, P. (2003). A unifying framework for partial volume segmentation of brain MR images. *IEEE Transactions on Medical Imaging*, 22(1):105–119.

Varin, C. (2008). On composite marginal likelihoods. *AStA Advances in Statistical Analysis*, 92(1):1.

Varin, C., Reid, N. M., and Firth, D. (2011). An overview of composite likelihood methods. *Statistica Sinica*, Vol.21(No.1):5–42.

Vese, L. A. and Chan, T. F. (2002). A multiphase level set framework for image segmentation using the Mumford and Shah model. *International journal of computer vision*, 50(3):271–293.

Vrooman, H. A., Cocosco, C. A., van der Lijn, F., Stokking, R., Ikram, M. A., Vernooij, M. W., Breteler, M. M., and Niessen, W. J. (2007). Multi-spectral brain tissue segmentation using automatically trained k-Nearest-Neighbor classification. *Neuroimage*, 37(1):71–81.

Wang, L., Li, C., Sun, Q., Xia, D., and Kao, C.-Y. (2009). Active contours driven by local and global intensity fitting energy with application to brain MR image segmentation. *Computerized medical imaging and graphics*, 33(7):520–531.

Ward, P. G. D., Ferris, N. J., Raniga, P., Ng, A. C. L., Barnes, D. G., Dowe, D. L., and Egan, G. F. (2017). Vein segmentation using shape-based Markov random fields. In *IEEE International Symposium on Biomedical Imaging*.

Weickert, J. (1998). *Anisotropic Diffusion in Image Processing*, volume 1. Teubner Stuttgart.

Wells, W. M., Grimson, W. L., Kikinis, R., and Jolesz, F. A. (1996). Adaptive segmentation of MRI data. *IEEE Transactions on Medical Imaging*, 15(4):429–442.

Wels, M. (2010). *Probabilistic Modeling for Segmentation in Magnetic Resonance Images of the Human Brain*, volume 33. Logos Verlag Berlin GmbH.

Wels, M., Zheng, Y., Huber, M., Hornegger, J., and Comaniciu, D. (2011). A discriminative model-constrained EM approach to 3D MRI brain tissue classification and intensity non-uniformity correction. *Physics in Medicine & Biology*, 56(11):3269.

Winkler, G. (2012). *Image Analysis, Random Fields and Markov Chain Monte Carlo Methods: A Mathematical Introduction*, volume 27 of *Stochastic Modelling and Applied Probability*. Springer Science & Business Media.

Withey, D. and Koles, Z. (2008). A review of medical image segmentation: Methods and available software. *International Journal of Bioelectromagnetism*, 10(3):125–148.

Wolff, U. (1989). Collective Monte Carlo updating for spin systems. *Physical Review Letters*, 62(4):361.

Woolrich, M. and Behrens, T. (2006). Variational Bayes inference of spatial mixture models for segmentation. *IEEE Transactions on Medical Imaging*, 25(10):1380–1391.

Woolrich, M., Behrens, T., Beckmann, C., and Smith, S. (2005). Mixture models with adaptive spatial regularization for segmentation with an application to fMRI data. *IEEE Transactions on Medical Imaging*, 24(1):1–11.

Wu, C.-H. and Doerschuk, P. C. (1995). Cluster expansions for the deterministic computation of Bayesian estimators based on Markov random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(3):275–293.

Yoo, T. S., Ackerman, M. J., Lorensen, W. E., Schroeder, W., Chalana, V., Aylward, S., Metaxas, D., and Whitaker, R. (2002). Engineering and algorithm design for an image processing API: A technical report on ITK-the insight toolkit. *Studies in Health Technology and Informatics*, pages 586–592.

Younes, L. (1991). Maximum Likelihood Estimation for Gibbsian Fields. *Lecture Notes-Monograph Series*, 20:403–426.

Yushkevich, P. A., Piven, J., Hazlett, H. C., Smith, R. G., Ho, S., Gee, J. C., and Gerig, G. (2006). User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability. *NeuroImage*, 31(3):1116–1128.

Zhang, J. (1992). The mean field theory in EM procedures for Markov random fields. *IEEE Transactions on Signal Processing*, 40(10):2570–2583.

Zhang, W., Li, R., Deng, H., Wang, L., Lin, W., Ji, S., and Shen, D. (2015). Deep convolutional neural networks for multi-modality isointense infant brain image segmentation. *NeuroImage*, 108:214–224.

Zhang, Y., Brady, M., and Smith, S. (2001). Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. *IEEE Transactions on Medical Imaging*, 20(1):45–57.

Appendix A

Derivation of update equations for a normal mixture model

We derive the update equations for the Expectation-Maximisation algorithm for a finite mixture model with spatially independent labels. First we derive the update for the mixing proportions with arbitrary component distributions. Then we give the update equations where the component distributions are multivariate normal. The update equations are all given in (McLachlan and Peel, 2000, Chapter3) without derivation.

As given in Chapter 2, let $\boldsymbol{Y}_i, i = 1, \ldots, n$ be n observations over a p-dimensional sample space. These are assumed distributed according to a g-component mixture. Let component j of the mixture ($j = 1, \ldots, g$) be distributed according to the pdf

$$f(\boldsymbol{y}_i|\boldsymbol{Z}_i = \boldsymbol{e}_j; \boldsymbol{\theta}_j) = f_j(\boldsymbol{y}_i; \boldsymbol{\theta}_j).$$

Each component density f_j need not be the same. Let $\boldsymbol{\Theta}$ be the parameters of $\boldsymbol{Y}_i$ known *a priori* to be distinct, $(\boldsymbol{\theta}_1^T, \ldots, \boldsymbol{\theta}_g^T)^T$.

Let $\boldsymbol{Z}_i, i = 1, \ldots, n$ indicate which component of the mixture each observation is from, where $\boldsymbol{z}_i$ is a vector of length g, with element j being 1 if and only if observation i is in class j and 0 otherwise. These are assumed independently and identically distributed according to the multinomial distribution with probabilities $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_g)$.

A.1 Joint distribution

Due to the binary nature of Z_{ij}, its pdf may be written

$$f(\boldsymbol{z}_i; \boldsymbol{\pi}) = \prod_{j=1}^{g} \pi_j^{z_{ij}}.$$

Since $\boldsymbol{Z}_i$ are independently distributed,

$$f(\boldsymbol{z};\boldsymbol{\pi}) = \prod_{i=1}^{n}\prod_{j=1}^{g}\pi_j^{z_{ij}}.$$

Similarly, the conditional distribution of the observed variables can be written

$$f(\boldsymbol{y}_i|\boldsymbol{z}_i;\boldsymbol{\Theta}) = \prod_{j=1}^{g} f_j(\boldsymbol{y}_i;\boldsymbol{\theta}_j)^{z_{ij}}.$$

Since $\boldsymbol{Y}_i$ are assumed independent given $\boldsymbol{Z}_i$,

$$f(\boldsymbol{y}|\boldsymbol{z};\boldsymbol{\Theta}) = \prod_{i=1}^{n}\prod_{j=1}^{g} f_j(\boldsymbol{y}_i;\boldsymbol{\theta}_j)^{z_{ij}}.$$

The joint distribution is then

$$f(\boldsymbol{y},\boldsymbol{z};\boldsymbol{\Theta},\boldsymbol{\pi}) = \prod_{i=1}^{n}\prod_{j=1}^{g}(\pi_j f_j(\boldsymbol{y}_i;\boldsymbol{\theta}_j))^{z_{ij}},$$

and log-likelihood function is

$$\log\mathcal{L}(\boldsymbol{\Theta},\boldsymbol{\pi};\boldsymbol{Y},\boldsymbol{Z}) = \sum_{i=1}^{n}\sum_{j=1}^{g} z_{ij}(\log\pi_j + \log f_j(\boldsymbol{y}_i;\boldsymbol{\theta}_j)).$$

A.2 E-step

The Q function is obtained by taking the expectation of the log-likelihood with respect to $\boldsymbol{Z}$ given $\boldsymbol{Y}, \boldsymbol{\Theta}^{(t)}, \boldsymbol{\pi}^{(t)}$:

$$\begin{aligned} Q(\boldsymbol{\Theta},\boldsymbol{\pi}|\boldsymbol{\Theta}^{(t)},\boldsymbol{\pi}^{(t)}) &= \mathbb{E}_{\boldsymbol{Z}|\boldsymbol{Y},\boldsymbol{\Theta}^{(t)},\boldsymbol{\pi}^{(t)}}[\log\mathcal{L}] \\ &= \sum_{i=1}^{n}\sum_{j=1}^{g}\mathbb{E}\left[z_{ij}|\boldsymbol{Y},\boldsymbol{\Theta}^{(t)},\boldsymbol{\pi}^{(t)}\right](\log\pi_j + \log f_j(\boldsymbol{y}_i;\boldsymbol{\Theta})). \end{aligned}$$

Now z_{ij} can take the values 1 or 0, so

$$\begin{aligned} \mathbb{E}\left[z_{ij}|\boldsymbol{Y}=\boldsymbol{y};\boldsymbol{\Theta}^{(t)},\boldsymbol{\pi}^{(t)}\right] &= 1\cdot\Pr(\boldsymbol{Z}_i=\boldsymbol{e}_j|\boldsymbol{y};\boldsymbol{\Theta}^{(t)},\boldsymbol{\pi}^{(t)}) + 0\cdot\Pr(\boldsymbol{Z}_i\neq\boldsymbol{e}_j|\boldsymbol{y};\boldsymbol{\Theta}^{(t)},\boldsymbol{\pi}^{(t)}) \\ &= \Pr(\boldsymbol{Z}_i=\boldsymbol{e}_j|\boldsymbol{y}_i;\boldsymbol{\Theta}^{(t)},\boldsymbol{\pi}^{(t)}) \text{ since } \boldsymbol{Z}_i \text{ depends on } \boldsymbol{Y}_i \text{ only} \\ &= \frac{f(\boldsymbol{y}_i|\boldsymbol{Z}_i=\boldsymbol{e}_j;\boldsymbol{\Theta}^{(t)})\Pr(\boldsymbol{Z}_i=\boldsymbol{e}_j;\boldsymbol{\pi}^{(t)})}{f(\boldsymbol{y}_i;\boldsymbol{\Theta}^{(t)},\boldsymbol{\pi}^{(t)})}. \end{aligned}$$

In the numerator, $f(\boldsymbol{y}_i|\boldsymbol{Z}_i=\boldsymbol{e}_j,\boldsymbol{\Theta}^{(t)})$ is the pdf $f_j(\boldsymbol{y}_i;\boldsymbol{\Theta}^{(t)})$, while $\Pr(\boldsymbol{Z}_i=\boldsymbol{e}_j;\boldsymbol{\pi}^{(t)}) = {\pi_j}^{(t)}$.

This yields:

$$\mathbb{E}\left[z_{ij}|\boldsymbol{Y}=\boldsymbol{y};\boldsymbol{\Theta}^{(t)},\boldsymbol{\pi}^{(t)}\right]=\frac{{\pi_j}^{(t)}f_j(\boldsymbol{y}_i;\boldsymbol{\Theta}^{(t)})}{\sum_{j=1}^{g}{\pi_j}^{(t)}f_j(\boldsymbol{y}_i;\boldsymbol{\Theta}^{(t)})}.$$

We will write this value as ${\tau_{ij}}^{(t)}$; note it does not depend on the parameters $\boldsymbol{\Theta}$ and $\boldsymbol{\pi}$, but only on their values at iteration t. Hence the Q-function is

$$\begin{aligned}Q(\boldsymbol{\Theta},\boldsymbol{\pi}|\boldsymbol{\Theta}^{(t)},\boldsymbol{\pi}^{(t)})&=\sum_{i=1}^{n}\sum_{j=1}^{g}{\tau_{ij}}^{(t)}(\log\pi_j+\log f_j(\boldsymbol{y}_i;\boldsymbol{\Theta}))\\ {\tau_{ij}}^{(t)}&=\frac{{\pi_j}^{(t)}f_j(\boldsymbol{y}_i;\boldsymbol{\Theta}^{(t)})}{\sum_{j=1}^{g}{\pi_j}^{(t)}f_j(\boldsymbol{y}_i;\boldsymbol{\Theta}^{(t)})}.\end{aligned}\tag{A.1}$$

A.3 M-step

To perform the M-step we maximise Q with respect to $\boldsymbol{\pi}$ and $\boldsymbol{\Theta}$.

A.3.1 Mixing proportions

We wish to maximise Q subject to $\sum_j \pi_j = 1$. This can be achieved using a Lagrange multiplier. Let

$$L=Q(\boldsymbol{\Theta},\boldsymbol{\pi}|\boldsymbol{\Theta}^{(t)},\boldsymbol{\pi}^{(t)})+\lambda(\sum_{j=1}^{g}\pi_j-1).$$

Then taking the derivative and setting it equal to zero,

$$\begin{aligned}\frac{\partial L}{\partial\pi_k}&=\frac{\partial}{\partial\pi_k}\sum_{i=1}^{n}\sum_{j=1}^{g}{\tau_{ij}}^{(t)}(\log\pi_j+\log f_j(\boldsymbol{y}_i;\boldsymbol{\theta}_j))+\lambda\frac{\partial}{\partial\pi_k}(\sum_j\pi_j-1)\\ &=\sum_{i=1}^{n}\frac{{\tau_{ik}}^{(t)}}{\pi_k}+\lambda\\ &=0.\end{aligned}$$

Rearranging yields

$$\pi_k=-\frac{\sum_{i=1}^{n}{\tau_{ik}}^{(t)}}{\lambda},\quad\forall k\in\{1,\ldots,g\}\tag{A.2}$$

Recalling that $\sum_{j=1}^{g} \pi_j = 1$, we have that

$$\begin{aligned} 1 &= -\sum_{j=1}^{g} \frac{\sum_{i=1}^{n} {\tau_{ij}}^{(t)}}{\lambda} \\ -\lambda &= \sum_{j=1}^{g}\sum_{i=1}^{n} {\tau_{ij}}^{(t)} \\ &= \sum_{i=1}^{n}\sum_{j=1}^{g} {\tau_{ij}}^{(t)} \\ &= \sum_{i=1}^{n} 1 \\ &= n, \end{aligned}$$

since

$$\sum_{j=1}^{g} {\tau_{ij}}^{(t)} = \sum_{j=1}^{g} \frac{{\pi_j}^{(t)} f_j(\boldsymbol{y}_i; {\boldsymbol{\theta}_j}^{(t)})}{\sum_{h=1}^{g} {\pi_h}^{(t)} f_h(\boldsymbol{y}_i; {\boldsymbol{\theta}_h}^{(t)})} = \frac{\sum_{j=1}^{g} {\pi_j}^{(t)} f_j(\boldsymbol{y}_i; {\boldsymbol{\theta}_j}^{(t)})}{\sum_{h=1}^{g} {\pi_h}^{(t)} f_h(\boldsymbol{y}_i; {\boldsymbol{\theta}_h}^{(t)})} = 1.$$

Substituting $\lambda = -n$ into (A.2) yields

$${\pi_j}^{(t+1)} = \frac{\sum_{i=1}^{n} {\tau_{ij}}^{(t)}}{n}, \tag{A.3}$$

with ${\tau_{ij}}^{(t)}$ given by (A.1). Note that this is independent of the component distributions in the mixture, f_j.

A.3.2 Gaussian components

The update equations for $\boldsymbol{\theta}_j$ are obtained by maximising Q with respect to them. We now assume that the component distributions f_j are multivariate Gaussian with mean $\boldsymbol{\mu}_j$ and covariance matrix $\boldsymbol{\Sigma}_j$ and derive the update equations. The pdf of component j is:

$$f_j(\boldsymbol{y}_i; \boldsymbol{\theta}_j) = \det(2\pi\boldsymbol{\Sigma}_j)^{-\frac{1}{2}} \exp(-\frac{1}{2}(\boldsymbol{y}_i - \boldsymbol{\mu}_j)^T \boldsymbol{\Sigma}_j^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu}_j)),$$

where $\boldsymbol{\theta}_j = (\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$. The Q function becomes (dropping constants that do not depend on the parameters):

$$Q(\boldsymbol{\Theta}, \boldsymbol{\pi} | \boldsymbol{\Theta}^{(t)}, \boldsymbol{\pi}^{(t)}) = \sum_{i=1}^{n}\sum_{j=1}^{g} {\tau_{ij}}^{(t)} (\log \pi_j - \frac{1}{2}\log(\det \boldsymbol{\Sigma}_j) - \frac{1}{2}(\boldsymbol{y}_i - \boldsymbol{\mu}_j)^T \boldsymbol{\Sigma}_j^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu}_j))$$

A.3.2.1 Mean

Taking the gradient of Q with respect to the mean $\boldsymbol{\mu}_k$ yields

$$\nabla_{\boldsymbol{\mu}_k} Q = -\frac{1}{2}\sum_{i=1}^{n} {\tau_{ik}}^{(t)} \nabla_{\boldsymbol{\mu}_k} (\boldsymbol{y}_i - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{y}_i - \boldsymbol{\mu}_k),$$

where the the sum over j only has non-zero derivative for the kth term. Now[1]

$$\begin{aligned}\nabla_{\boldsymbol{\mu}_k} (\boldsymbol{y}_i - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{y}_i - \boldsymbol{\mu}_k) &= (\boldsymbol{\Sigma}_k^{-1} + (\boldsymbol{\Sigma}_k^{-1})^T)(\boldsymbol{y}_i - \boldsymbol{\mu}_k) \\ &= 2\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu}_k) \cdot (-1),\end{aligned}$$

since $\boldsymbol{\Sigma}_k$ is symmetric, and hence so is $\boldsymbol{\Sigma}^{-1}$. Then

$$\begin{aligned}\nabla_{\boldsymbol{\mu}_k} Q &= \sum_{i=1}^{n} {\tau_{ik}}^{(t)} \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{y}_i - \boldsymbol{\mu}_k) \\ &= \boldsymbol{\Sigma}_k^{-1} (\sum_{i=1}^{n} {\tau_{ik}}^{(t)} \boldsymbol{y}_i - {\tau_{ik}}^{(t)} \boldsymbol{\mu}_k) \\ &= 0.\end{aligned}$$

Rearranging yields

$${\boldsymbol{\mu}_k}^{(t+1)} = \frac{\sum_{i=1}^{n} {\tau_{ik}}^{(t)} \boldsymbol{y}_i}{\sum_{i=1}^{n} {\tau_{ik}}^{(t)}}. \tag{A.4}$$

A.3.2.2 Covariance

Taking the gradient of Q with respect to $\boldsymbol{\Sigma}_k$ yields

$$\nabla_{\boldsymbol{\Sigma}_k} Q = -\frac{1}{2}\sum_{i=1}^{n} {\tau_{ik}}^{(t)} \nabla_{\boldsymbol{\Sigma}_k} Q(\log(\det \boldsymbol{\Sigma}_k) + (\boldsymbol{y}_i - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{y}_i - \boldsymbol{\mu}_k))$$

Now

$$\frac{\partial}{\partial \boldsymbol{\Sigma}_k} \log\det(\boldsymbol{\Sigma}_k) = (\boldsymbol{\Sigma}_k^T)^{-1} = \boldsymbol{\Sigma}_k^{-1} \text{ as } \boldsymbol{\Sigma}_k = \boldsymbol{\Sigma}_k^T$$

and

$$\begin{aligned}\frac{\partial}{\partial \boldsymbol{\Sigma}_k} (\boldsymbol{y}_i - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{y}_i - \boldsymbol{\mu}_k)) &= -\boldsymbol{\Sigma}_k^{-T} (\boldsymbol{y}_i - \boldsymbol{\mu}_k)^T (\boldsymbol{y}_i - \boldsymbol{\mu}_k) \boldsymbol{\Sigma}_k^{-T} \\ &= -\boldsymbol{\Sigma}_k^{-1} (\boldsymbol{y}_i - \boldsymbol{\mu}_k)^T (\boldsymbol{y}_i - \boldsymbol{\mu}_k) \boldsymbol{\Sigma}_k^{-1},\end{aligned}$$

so the derivative is

$$\frac{\partial Q}{\partial \boldsymbol{\Sigma}_k} = -\frac{1}{2}\sum_{i=1}^{n} {\tau_{ik}}^{(t)} (\boldsymbol{\Sigma}_k^{-1} - \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{y}_i - \boldsymbol{\mu}_k)^T (\boldsymbol{y}_i - \boldsymbol{\mu}_k) \boldsymbol{\Sigma}_k^{-1}).$$

[1] The matrix cookboock (Petersen and Pedersen, 2012) is immensely useful here

Setting the gradient equal to 0 and multiplying by $\boldsymbol{\Sigma}_k$ on the left and right yields

$$\begin{aligned} 0 &= \sum_{i=1}^{n} {\tau_{ik}}^{(t)} (\boldsymbol{\Sigma}_k - (\boldsymbol{y}_i - \boldsymbol{\mu}_k)^T (\boldsymbol{y}_i - \boldsymbol{\mu}_k)) \\ &= \boldsymbol{\Sigma}_k \sum_{i=1}^{n} {\tau_{ik}}^{(t)} - \sum_{i=1}^{n} {\tau_{ik}}^{(t)} (\boldsymbol{y}_i - \boldsymbol{\mu}_k)^T (\boldsymbol{y}_i - \boldsymbol{\mu}_k) \\ \boldsymbol{\Sigma}_k &= \frac{\sum_{i=1}^{n} {\tau_{ik}}^{(t)} (\boldsymbol{y}_i - \boldsymbol{\mu}_k)^T (\boldsymbol{y}_i - \boldsymbol{\mu}_k)}{\sum_{i=1}^{n} {\tau_{ik}}^{(t)}}. \end{aligned} \tag{A.5}$$

A.4 Summary

Restating (A.1), (A.3), (A.4) and (A.5), the update equations for a Gaussian mixture model are:

$$\begin{aligned} {\tau_{ij}}^{(t)} &= \frac{{\pi_j}^{(t)} f_j(\boldsymbol{y}_i; \boldsymbol{\Theta}^{(t)})}{\sum_{j=1}^{g} {\pi_j}^{(t)} f_j(\boldsymbol{y}_i; \boldsymbol{\Theta}^{(t)})} \\ {\pi_j}^{(t+1)} &= \frac{\sum_{i=1}^{n} {\tau_{ij}}^{(t)}}{n} \\ {\boldsymbol{\mu}_j}^{(t+1)} &= \frac{\sum_{i=1}^{n} {\tau_{ij}}^{(t)} \boldsymbol{y}_i}{\sum_{i=1}^{n} {\tau_{ij}}^{(t)}} \\ {\boldsymbol{\Sigma}_j}^{(t+1)} &= \frac{\sum_{i=1}^{n} {\tau_{ij}}^{(t)} (\boldsymbol{y}_i - {\boldsymbol{\mu}_j}^{(t+1)})^T (\boldsymbol{y}_i - {\boldsymbol{\mu}_j}^{(t+1)})}{\sum_{i=1}^{n} {\tau_{ij}}^{(t)}}. \end{aligned} \tag{A.6}$$

It should be noted that the update for the mixing proportions does not require f_j to be Gaussian.

Following McLachlan and Peel (2000) (section 3.2), it is computationally convient to write:

$$\begin{aligned} T_{j1} &= \sum_{i=1}^{n} {\tau_{ij}}^{(t)} \\ \boldsymbol{T}_{j2} &= \sum_{i=1}^{n} {\tau_{ij}}^{(t)} \boldsymbol{y}_i \\ \boldsymbol{T}_{j3} &= \sum_{i=1}^{n} {\tau_{ij}}^{(t)} \boldsymbol{y}_i \boldsymbol{y}_i^T, \end{aligned} \tag{A.7}$$

giving

$$\begin{aligned} {\pi_j}^{(t+1)} &= \frac{T_{j1}}{n} \\ {\boldsymbol{\mu}_j}^{(t+1)} &= \frac{\boldsymbol{T}_{j2}}{T_{j1}} \\ {\boldsymbol{\Sigma}_k}^{(t+1)} &= \frac{\boldsymbol{T}_{j3} - \boldsymbol{T}_{j2} \boldsymbol{T}_{j2}^T / T_{j1}}{T_{j1}}. \end{aligned} \tag{A.8}$$

To see the last line,

$$
\begin{aligned}
\boldsymbol{\Sigma}_j{}^{(t+1)} &= \frac{\sum_{i=1}^n \tau_{ij}{}^{(t)} (\boldsymbol{y}_i - \boldsymbol{\mu}_j{}^{(t+1)})^T (\boldsymbol{y}_i - \boldsymbol{\mu}_j{}^{(t+1)})}{\sum_{i=1}^n \tau_{ij}{}^{(t)}} \\
&= \frac{\sum_{i=1}^n \tau_{ij}{}^{(t)} \boldsymbol{y}_i^T \boldsymbol{y}_i - \tau_{ij}{}^{(t)} \boldsymbol{y}_i^T \boldsymbol{\mu}_j{}^{(t+1)} - \tau_{ij}{}^{(t)} (\boldsymbol{\mu}_j{}^{(t+1)})^T \boldsymbol{y}_i + \tau_{ij}{}^{(t)} (\boldsymbol{\mu}_j{}^{(t+1)})^T \boldsymbol{\mu}_j{}^{(t+1)}}{T_{j1}} \\
&= \frac{\boldsymbol{T}_{j3} - 2(\boldsymbol{\mu}_j{}^{(t+1)})^T \sum_{i=1}^n \tau_{ij}{}^{(t)} \boldsymbol{y}_i + \boldsymbol{T}_{j2}^T \boldsymbol{T}_{j2} / T_{j1}^2 \sum_{i=1}^n \tau_{ij}{}^{(t)}}{T_{j1}} \\
&= \frac{\boldsymbol{T}_{j3} - 2(\boldsymbol{T}_{j2}^T / T_{j1}) \boldsymbol{T}_{j2} + \boldsymbol{T}_{j2}^T \boldsymbol{T}_{j2} / T_{j1}^2 \cdot T_{j1}}{T_{j1}} \\
&= \frac{\boldsymbol{T}_{j3} - 2\boldsymbol{T}_{j2}^T \boldsymbol{T}_{j2} / T_{j1} + \boldsymbol{T}_{j2}^T \boldsymbol{T}_{j2} / T_{j1}}{T_{j1}} \\
&= \frac{\boldsymbol{T}_{j3} - \boldsymbol{T}_{j2}^T \boldsymbol{T}_{j2} / T_{j1}}{T_{j1}}.
\end{aligned}
$$

Appendix B

Coding schemes for three dimensional images

Here we give the coding schemes used for three-dimensional neighbourhoods with 6 neighbours, 18, and 26 neighbours in a $3 \times 3 \times 3$ cube. A coding scheme (Besag, 1974) is a partition of the lattice (in our case, the voxel grid) into sets of vertices into sets such that no two elements in the same set are neighbours. For quantities that should be calculated sequentially using the updated values for each voxel's neighbours at all times, coding schemes allow all voxels in a given set to be updated simultaneously, with the sets visited sequentially. This is useful in calculating Iterated Conditional Modes or mean-field updates. Note that coding schemes are not necessarily unique. We believe the ones shown use the minimum number of voxel partitions (though their arrangement is also not necessarily unique).

Visualising the three-dimensional schemes in two dimensions is difficult. To aid in this, we have drawn a portion of the voxel grids in 2D slices, thowing three slices (since the neighbourhood is $3 \times 3 \times 3$). Each coding set is associated with a particular symbol; every voxel in that set is marked with that symbol. The schemes are periodic. To verify that the coding sets are valid, pick a particular voxel and determine its neighbours; none of them should match the symbol of that voxel. Repeat this for each symbol.

To aid in this, we outline the voxel (symbol) being considered in bold with a blue background. In each slice displayed, we indicate its neighbours by shading them grey. One should verify that none of the grey-shaded neighbours share the same symbol as the blue-shaded voxel. We repeat this for each symbol/coding set.

B.1 6 neighbours

The 6-neighbourhood contains the orthogonal neighbours: north, east, south, west, top and bottom. A coding scheme may be achieved by partitioning the voxels into two sets in a chess-

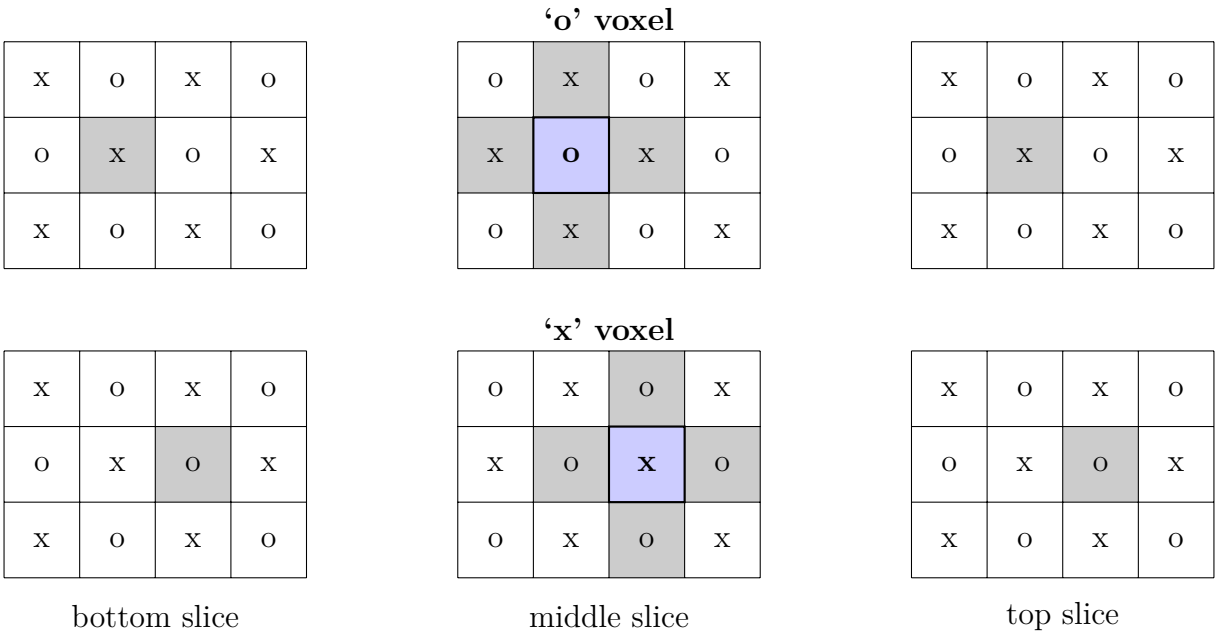


Figure B.1: Coding scheme into 2 sets for 6-neighbourhood.

board pattern. This scheme is periodic within each slice, and should be shifted by 1 row (or column) between slices. It is shown in Figure B.1.

B.2 18 neighbours

The 18-neighbourhood contains the orthogonal neighbours as well as the in-plane neighbours. A coding scheme may be achieved by partitioning the voxels into four sets. The scheme is periodic within each slice (a 2×2 grid). Between slices, it should be offset by one row and one column. The scheme is periodic over a $2 \times 2 \times 2$ cube. It is shown in Figure B.2.

B.3 26 neighbours

The 26-neighbourhood contains all neighbours of the $3 \times 3 \times 3$ cube. A coding scheme may be achieved by partitioning the voxels into eight sets. This scheme uses 4 symbols per slice, and is periodic in a 2×2 grid within each slice. The symbols use alternate every slice. We show the neighbourhoods first for "odd" slices (symbols !, #, ^, @) in Figure B.3 and then for "even" slices (symbols x, o, ., *) in Figure B.4. These slices are alternated as-is.

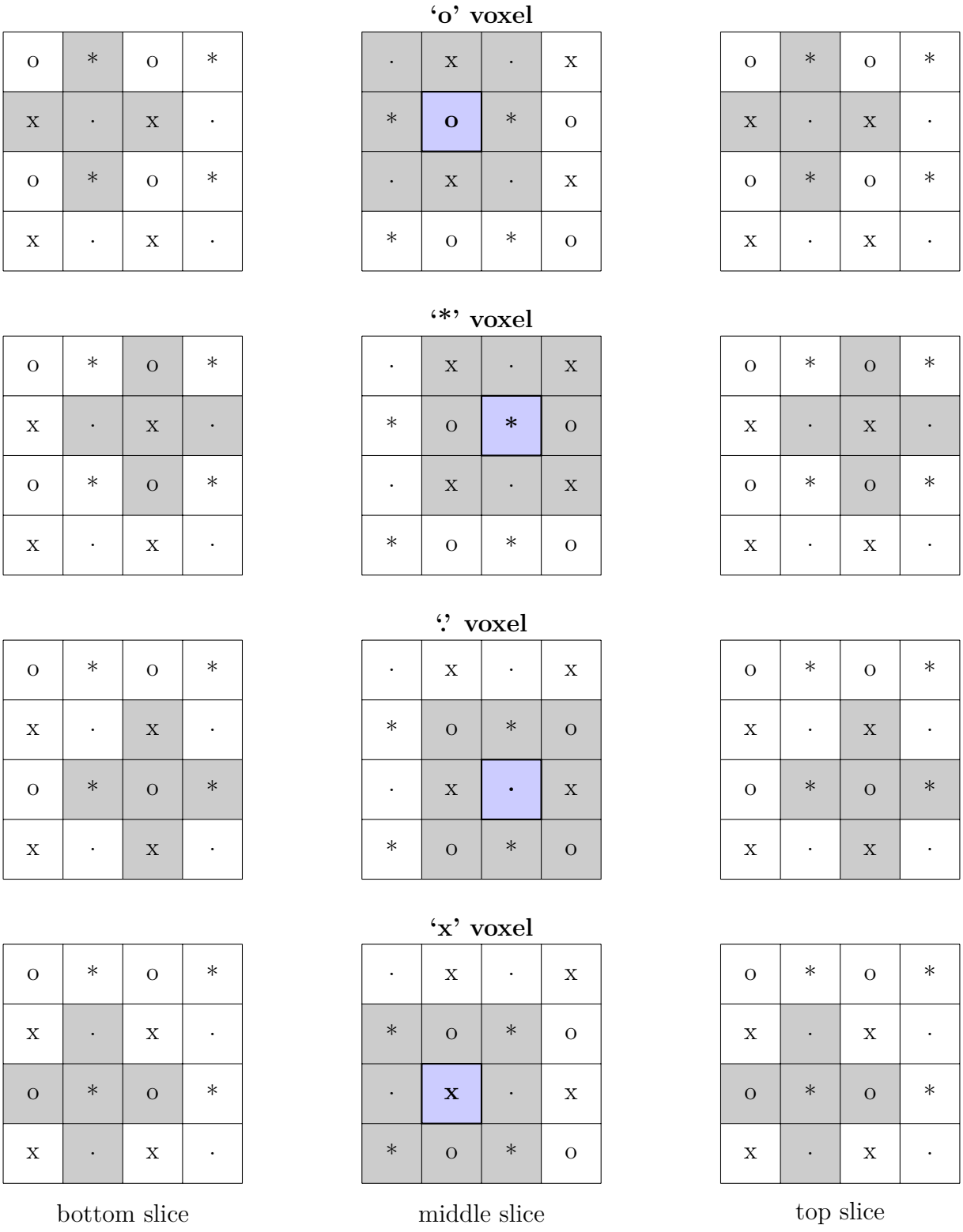


Figure B.2: Coding scheme into 4 sets for 18-neighbourhood.

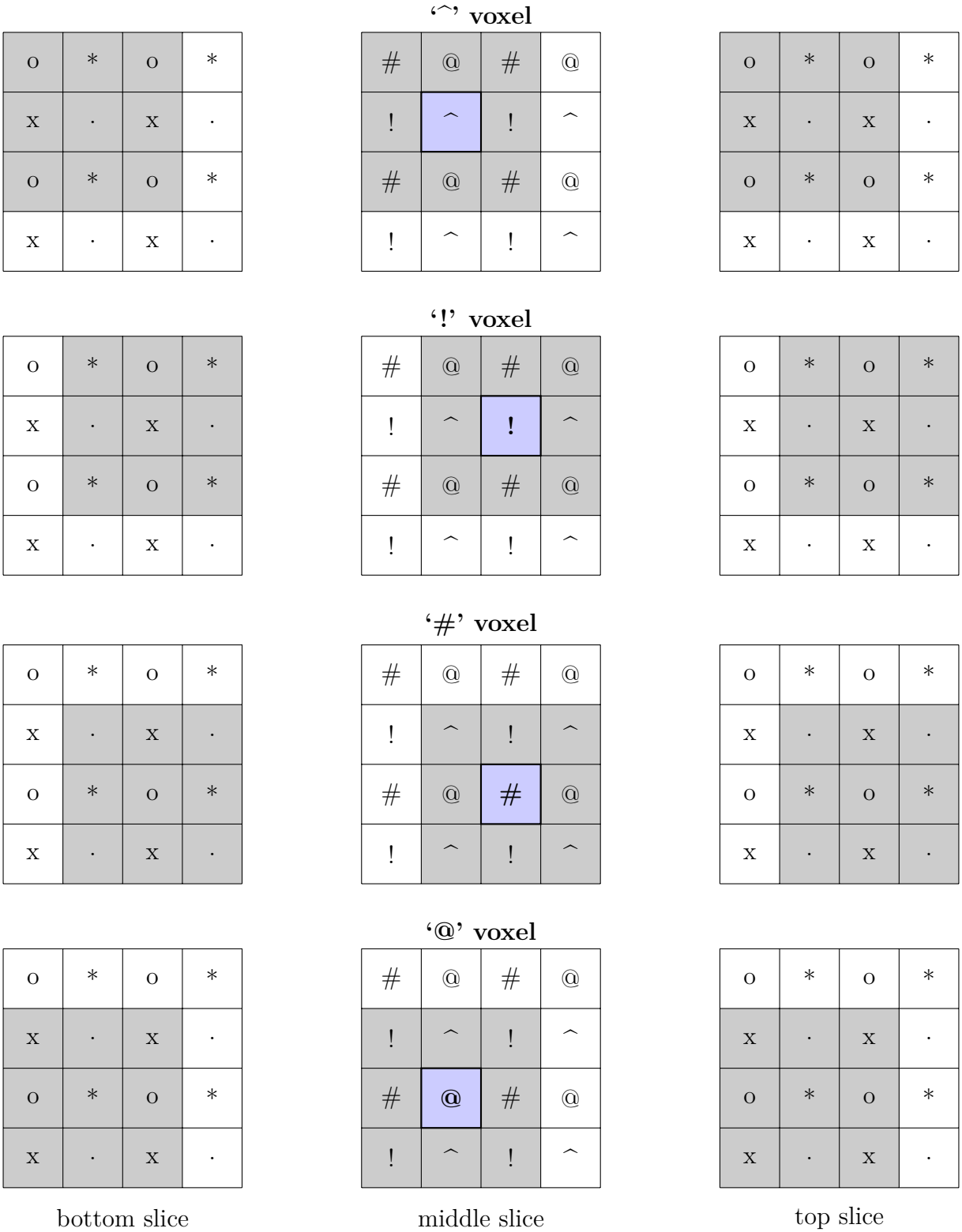


Figure B.3: Coding scheme into 8 sets for the 26-neighbourhood. The “odd” slice is in the centre (symbols !, #, ^, @).

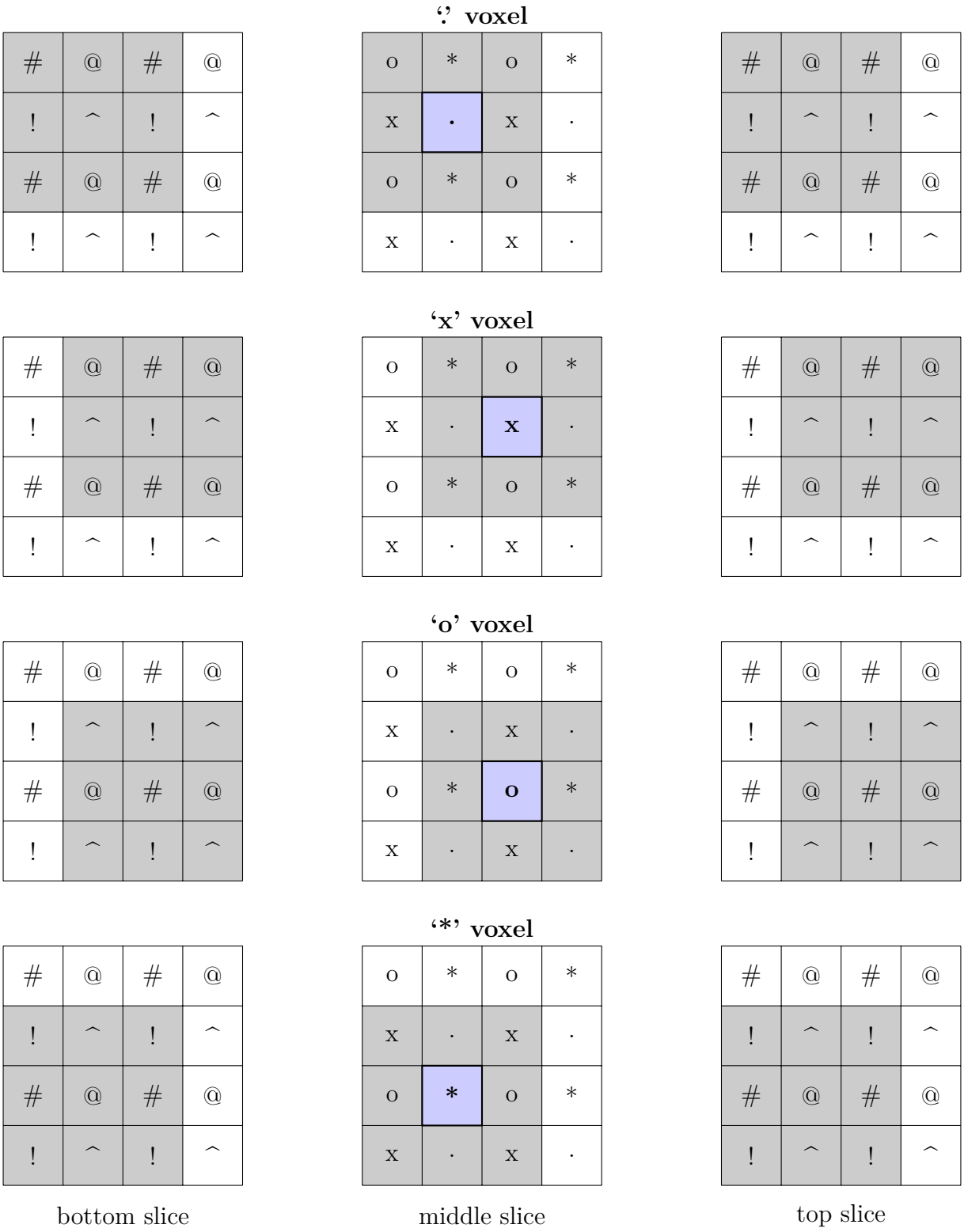


Figure B.4: Coding scheme into 8 sets for the 26-neighbourhood. The "even" slice is in the centre (symbols x, o, ., *).