Tina Memo No. 2015-001
Internal.

Improving the Stability of MRI Parameter Estimates using Regional Sub-Sampling

S.V. Notley and N.A. Thacker

Last updated
14 / 2 / 2015

Imaging Science and Biomedical Engineering Division,
Medical School, University of Manchester,
Stopford Building, Oxford Road,
Manchester, M13 9PT.

Abstract

In our previous work we have shown empirically how the use of regional subsets of data could be used to improve the errors on estimated tissue density parameters for MRI segmentation. Here a discussion of the current state of the art in the literature and justification for the need for this work is followed with an investigation into the theoretical origins of this improvement. We show that this approach can be related to conventional regression techniques and the performance can be predicted using estimates of the Minimum Variance Bound (MVB). By doing so we can prove that the Fisher information available for estimation of parameters is increased by discarding data (on the basis of spatial information) when the interpretation of this data is otherwise highly ambiguous (i.e. density functions overlap). All mathematical derivations are provided, along with simulation results designed to illustrate the improved parameter accuracy over a range of typical MRI contrast variation.

Introduction

The work presented in this report is concerned with the automatic segmentation of MR images and the associated problem of making quantitative volumetric estimates of tissue quantities. The main motivation is to develop a better understanding of empirical observations made in [5] which showed, that for overlapping intensity distributions, significant accuracy in model parameter estimates can be gained through recovering spatial information via regional sub-sampling. In comparison to other methods, the approach has some similarities to the use of spatial 'priors' during the expectation step of an Expectation Maximisation (EM) algorithm. However, in contrast to other current segmentation algorithms that use spatial priors and/or contextual information, this method is based upon more conventional regression approaches to data analysis and parameter estimates are unbiased.

This method under study works by estimating the parameters for specific tissue distributions from regional subsets of data which have been selected to enhance the relative proportion of this tissue. The idea can be related to more conventional density distribution fitting via a two step modification in approach. Firstly, we argue that as EM (Likelihood) based estimation can be applied to any region of data then there can be no logical objection to partitioning the data into multiple regions and fitting them simultaneously (using additional separate sets of tissue quantities or 'priors'). These regions are selected to maximise the quantity of each tissue without modifying the assumed density functions. Having done this we then observe that the information available for each tissue in the 'non-tissue' regions is likely to be small. This data can therefore be discarded entirely during the estimation of tissue specific parameters (i.e. mean, and covariance) without conflicting with the principles of Likelihood estimation. The original insight for this work was that ambiguous interpretations of data within the 'expectation' step destabilise parameter estimates. Although reducing the quantity of data reduces the information with which to constrain parameters, the effects of ambiguity may be more important.

A critical review of current methods considered to be 'state-of-the-art' is presented that discusses the results concerning segmentation accuracies found in the literature. This review places the current work in context and addresses the accuracy of these state-of-the-art methods in terms of the approaches and models employed. The review section discusses the need for appropriate intensity distribution models and that under those conditions improvements in estimation of model parameters leads to improved volumetric estimates.

In order to confirm the origins of the expected benefits of regional sub-sampling we set out to determine if the observed performance was consistent with a theoretical model of estimation. Crucially, the MVB can be applied to the equivalent Likelihood function which describes the multiple region regression, allowing us to predict the accuracy of parameter estimates prior to the final step of discarding 'non-tissue' regions. This can then be compared with the observed performance of the regional subset parameter estimation algorithm. For mathematical tractability this theoretical analysis is employed within a simplified framework that considers only two Gaussian tissue distributions. Without loss of generality, we can consider this to be analogous to estimation of white and grey matter density parameters, on the assumption that all other parameters of a more general model are already known. If regional sub-sampling can be shown to make better use of the information available for this case, then we can safely assume that this would likely be true for the general model, thereby accounting for our previous findings. We show here that Monte-carlo investigation of the MVB bear out our hypothesis. We conclude that regional sub-sampling of data during the process of parameter estimation can indeed result in an increase in the available Fisher Information.

Remarks Concerning Current Methodologies

Modern medical imaging techniques are a ubiquitous set of tools that are used in both clinical and research environments. From a clinical perspective robust image analysis techniques are generally sought to aid an expert

in making diagnostic assessments and clinical decisions; from a research perspective, in areas such as neuroscience and experimental medicine, such analysis techniques are more likely aimed at providing quantitative measurements [10]. That is not to say that quantitative robust techniques are not useful or sought after for clinical applications but rather the opposite. Research projects generally have the time to perform studies at a group level which can give the statistical power necessary to draw conclusions overcoming methodological short comings. In a clinical setting this is not the case and conclusions need to be drawn from a small number of samples and thus requires high levels of precision.

In this work we are primarily concerned with magnetic resonance imaging (MRI) of brain tissue and the common task of segmenting an image; that is, in an ideal situation with infinite precision, the decomposition of an image into individual components composed entirely of one underlying tissue type. In reality the image formation process has finite resolution and is subject to various forms of error and thus the task becomes an optimisation/estimation problem. In general this requires the selection of an appropriate model that is able to account for the variations seen in the data. A common model that is used to represent the intensity distributions found from MR images of brain tissue is a pure Gaussian mixture model. As will be discussed below this model over simplifies the image formation process and does not fully describe the distributions found in real data.

The finite resolution of the image formation process means that any single voxel is not guaranteed to be composed purely of a single tissue type but of a combination of tissues in varying amounts; the so called partial volume effect (PVE). Segmentation of MRI images often forms the backbone of analysis techniques that then attempt to quantify brain tissue and anatomical structures and thus has increasing importance in studies concerned with brain development, neuro-degeneration, dementia etc and in the assessment of neurological disorders[11]. If partial volume effects are not modelled correctly then algorithms will be restricted to simply assigning the most probable class label given the grey level leading to over- and/or under-estimated volume estimates of tissue quantities. Correct modelling of the partial volume effect not only allows a voxel to be labelled as partial but also allows the most probable tissue proportions to be estimated allowing quantitative volumetric estimates.

Particular attention should be paid to the use of the word *quantitative* in the previous statements. By this we mean the ability to make *unbiased* measurements that reflect real world phenomena and to know the accuracy of such measurements. Klauschen *et al* [11] stated in 2009 that 'In the search for biological causes of brain volume differences between diagnostic groups, or individual changes during time in longitudinal studies, the variations due to MRI measurement technique, data quality and image segmentation procedure should be explored and accommodated in the data analysis.'. They further cite a study performed by Clark *et al* [8] in which it was concluded that the choice of segmentation algorithm had the largest impact on variability over other factors such as the pulse sequence. With respect to these findings they reviewed three widely used brain volumetry methods: FSL, SPM5 and FreeSurfer. They used a combination of Brainweb simulated data (1mm^3 isotropic voxels) and expert segmented real data (0.94 x 0.94 x 1.4mm^3 voxels).

As further confirmation of the results presented by Clark *et al* they also found significant inter- and intra-method variation in segmentation results for both whole brain and total grey/white matter volumes. They also argue and present the conclusion that to reliably detect a 3% change in grey matter volume in a longitudinal study (i.e. intra-subject), the best performing algorithm would require approximately 150 subjects! The estimate of group size was based on having the largest grey matter volume being 24% larger than the smallest over their cohort. In the authors view it would be an interesting aside to note how many studies using such algorithms meet the above criteria to detect reliable changes. The main conclusion from this work was that even with good quality data the best performing algorithm (SPM) had significant volumetric errors of between 10% and 20%.

A review from 2008 by Tsang *et al* [17] also reviewed SPM5 and FSL. They used Brainweb simulated data (unstated voxel dimensions) and real brain data (3T magnet, T1 spin echo with 1x1x2mm^3 voxels and T2 turbo-spin echo with 0.9x0.9x3mm^3 voxels). Their methodology used Dice [12] coefficients as similarity measures and found errors of around 10% with the Brainweb data and errors of around between 20% and 30% for the real brain data.

A more recent review of segmentation algorithms [10] from 2014 comparing SPM8, FSL and Brainsuite performed a similar analysis (using Brainweb data with 1mm^3 isotropic voxels and expert segmented real data, voxel dimensions are unstated) and found similar results. They calculated Dice and Jaccarde [4] coefficients as similarity measures and show class labelling errors of around 10% for Brainweb simulated data and this falls to as low as 20% when used on real datasets.

We would like to refer here to problems found in Brainweb simulated data that are highlighted by Bromiley [5]. This work explains that in order for an evaluation with simulated data to be meaningful the simulation *must* model all of the major effects seen in real MR images. However, it is argued that the modelling of partial volume effects as described in [9] are flawed and not an accurate description of the geometric measurement of MR signal. The simulation model ignores the contribution of noise and biological variation to the spread of a pure tissue distribution and ignores entirely the volumetric (geometrical) contributions to the partial volume effect. As well as

issues related to content management a further problem was observed in the histogram of Brainweb data in that a number of spurious peaks were seen outside of the pure tissue peaks. These peaks were again attributed to a combination of numerical artefacts and poor distribution modelling. These effects can be observed by the simple process of examining histograms of Brainweb data. It is concluded that results found using Brainweb data for algorithms that include partial modelling effects should be treated with caution.

The three independent and corroborating pieces of work cited above also include expert segmented data which to a certain extent circumvents the problems found with Brainweb data. In light of this corroborating evidence and the simple theoretical estimate of possible accuracies stated above, we must ask the question of why such low levels of performance are being found in these algorithms given that they are considered by many as state of the art.

Previous work by Thacker *et al* [16] estimated the grey and white matter accuracies obtainable on a per voxel basis when using a closed form solution (based upon a linear model of partial volume image formation) for a number of MRI sequences. For inversion recovery spin echo (IRSE) and inversion recovery turbo-spin echo (IRTSE) they found **per voxel** accuracies of around 25%. If we assume a given tissue typically has a total number of voxels of the order of 10,000 then for the given accuracy we would expect errors made on the total volume to be of the order of 0.25% ($25/\sqrt{10000}$). This analysis assumes that pure tissue parameters are known exactly and the data otherwise can be considered as random samples. The large discrepancy between this estimate and those observed in commonly used segmentation algorithms is worthy of some comment. It suggests to us that the poor performance of segmentation algorithms is due to systematic errors arising from poor modelling, and is not due to statistical errors arising more directly from random image noise.

There are several factors which might be causing the very poor segmentation accuracy seen in independent evaluations. Firstly, the images themselves may contain artifacts which invalidate the assumed data density and image formation model. The most obvious of these is receiver coil effects which can alter the sensitivity to signal across the bore of the magnet. In addition however, larger field machines (i.e. 3T and above) while having less (easily visible) high frequency noise are likely to have greater (less visible) grey level variations due to secondary physics processes [11]. At higher field strengths we may even expect the *effective* T1 and T2 parameters to vary across a single tissue, making the standard assumption of a tissue being reliably described by a single Gaussian distribution unrealistic. If this is the case then we would need independent information with which to predict the true image contrast at each point in the image. We may therefore never be able to segment 3T data with much better performance than current state of the art (10 %).

If we take steps to avoid (or correct) field inhomogeneity effects and stay with lower field machines (1.5 T) then the main problem associated with obtaining accurate segmentations will be appropriate choice of model and the accuracy of estimation of associated density parameters. Indeed, for significantly overlapping density distributions (such as grey matter and white matter in the brain) the accuracy with which we can determine the separation between tissue means will generate approximately proportional errors on tissue labels. This conclusion can be understood by considering the consequence of errors on mean tissue parameters on the Bayes Error Rate. For approximately equal 'priors' and tissue variance, halving the error on the tissue means has the potential to reduce the segmentation errors by a similar factor. It is this observation which leads us to conclude that to improve segmentation accuracies we must focus on better estimation of tissue model parameters and appropriate data density models. By controlling these effects carefully we may be able to approach much better segmentation accuracies (i.e. 3%) sufficient for the detection of subtle volume change in individual subjects.

Both SPM and FSL use probabilistic approaches that use a finite mixture model of the image intensity distribution to make classifications. The whole process first requires the selection of a suitable approximate model. The problem then becomes a matter of estimating the unknown model parameters of each model component such that the best fit to the total intensity distribution is found. A classification decision is then made based on the most dominant model component (in a probabilistic sense) for a given voxels intensity. The problem of model selection and model fitting invariably requires making assumptions about the underlying tissue properties and the image formation process. For many researchers, the emphasis put on this process focuses on the efficiency of parameter determination, rather than appropriateness of the model approximation, i.e. choices which allow simple solutions. Unfortunately, no matter what model is used it will always be possible to find a 'best' fit to the intensity distribution regardless of how 'good' the fit actually is. Using principled solutions for parameter estimation cannot eliminate the problems generated by having already selected a poor model.

Typically the assumption is made that the underlying tissue samples are piecewise constant and that the measurements made within a particular voxel are subject to additive noise of normal distribution[1]. This leads directly to the intensity model being composed of a set number of Gaussian distributions related to each tissue type; the finite Gaussian Mixture Model. Hidden with this model is the assumption that each voxel is composed of pure

[1]It can also be assumed that tissue samples are piecewise continuous, i.e. drawn from a distribution, and then subject to additive noise. Generally, it is assumed in this case that both tissue and noise distributions are Gaussian leading to the same overall model.

tissue and that the intensity of a voxel is independent of the surrounding voxels. As stated previously, it is well known that these assumptions are approximate and that due to the finite size of each voxel there is a significant proportion of voxels are subject to PVEs. It would be wrong here to assume that PVE's are simply equivalent to a point spread process (or image blurring). Slight image blur will transfer a small percentage of signal across a tissue boundary, and so have no real effect on the assumed density model. Near-linearity of the image formation process also ensures little effect on consequent volume estimation or tissue boundary definition. Geometrical PVE's on the other hand will change the assumed density distributions significantly by generating data values that lie almost uniformly along the line between two pure tissue Gaussians. Depending on the voxel dimensions, PVEs can affect as many as 40% of the voxels in an image [refs]. We would argue that key to understanding the performance of these algorithms is to understand the underlying assumptions, especially those concerning partial volume effects [6].

As previously stated, the amount of PVEs within an image are dependent on (and proportional to) the voxel dimensions i.e. thick slice acquisitions are likely to have a lot more voxels with partial volumes than for smaller isotropic acquisitions [4]. Given that a thick slice acquisition with voxel dimensions 1 X 1 x 5mm^3 (5mm in depth with a 1mm inter-slice gap) can have up 40% PVEs then we can extrapolate that an isotropic acquisition with $1mm^3$ voxels will have around 10% PVEs. We note that this is the same order of error found in the review papers mentioned for data with isotropic 1mm^3 voxels. A simple extrapolation to the T1 and T2 weighted acquisitions for the work of Tsang *et al* would predict errors of between 20% and 30%, again of the same order as the errors found. One explanation for the magnitude of this problem is errors on model parameters (as described earlier) arising due to poor initial model selection.

Another possible cause which is rarely considered follows from the fact that brain image research is primarily concerned with only grey matter (GM), white matter (WM), and cerebrospinal fluid (CSF). However, MR images of the head will contain skull, fat and bone tissue types whose intensity distribution in the image significantly overlaps with the distribution of the tissues of interest. A common method used to eliminate this ambiguity is to perform a skull stripping/brain extraction procedure that intends to leave only the three tissues of interest. If this is not done appropriately then this method will lead to a sculpting of the intensity distribution which can also introduce significant parameter bias during the fitting of the density models.

Figure 1 is taken from [18] showing an intensity distribution from an MR image of a brain and shows the individual distributions that have been fit using a pure Gaussian mixture model. It is quite possible that the skull stripping procedure has erroneously removed CSF voxels and sculpted the histogram in such a way that the mean of the CSF Gaussian is too far to the right in the intensity distribution. Further to this since there is no modelling of partial volume voxels, *all* of the voxels with an intensity to the left of the dotted line will be attributed to pure CSF leading to an over estimation of CSF volume. Indeed, the quantitative consequences due to partial volume effects are expected to be largest for well separated tissues, and also become more significant when using multi-spectral approaches. The easiest way to visualise partial volume effects is in the 2D scattergrams of two different MRI sequences.

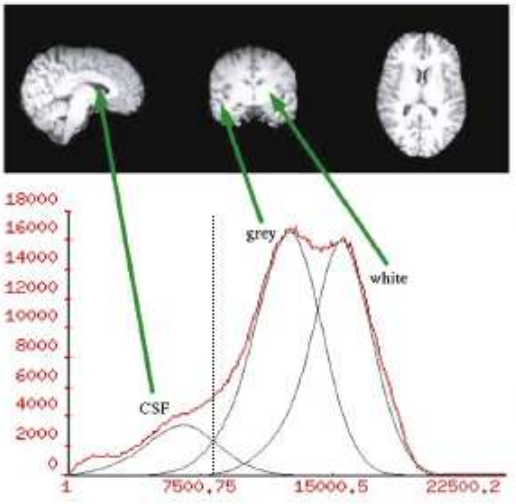
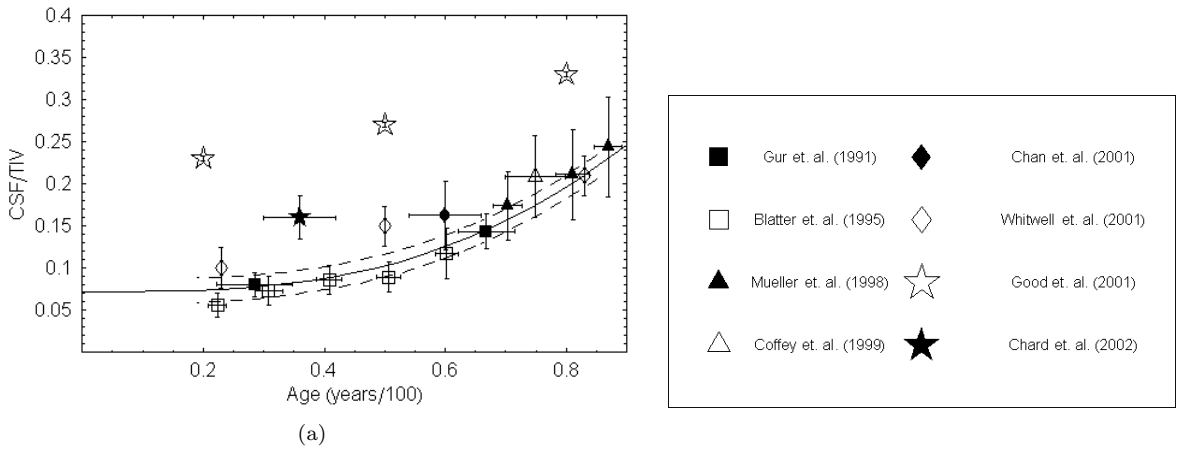


Figure 1: Figure taken from Woolrich *et al* [18] showing bias in pure Gaussian mixture modelling leading to over estimation of CSF

This corroborates a review of results found in the literature [15] showing that pure Gaussian mixture models have



(a)

Figure 2: Previous published TIV normalised CSF volume measurements. The solid curves show the functional fit to the data presented in this paper: the dashed curves represent the 1σ error bounds. The points show data from the literature. Points with error bars in both the x and y directions represent data published numerically, whilst points with error bars only in the y direction represent data read from graphs.

a bias that significantly over estimates CSF volume compared to expert CSF volume estimates. Figure 2 is from [15] and shows manual CSF volume estimates as a function of subject age taken from a number of independent publications. The solid line shows a function fit to volume estimates made using a segmentation algorithm which includes partial volume modelling. The two sets of results found using SPM (Clear star and filled star) are clearly over estimates of the CSF volume.

Another popular algorithm, FSL, is based on an extension of the Markov random field (MRF) theory and termed the Hidden Markov random field (HMRF). In this model the underlying true (hidden) tissue labels are assumed to be a MRF; that is to say that any individual voxel is dependent on its neighbouring voxels. The observed grey levels are 'emissions' of this underlying field and is typically modelled as a mixture of pure Gaussians i.e. the emission field models the effects of noise and/or tissue distribution on the pure underlying labels.

For the finite mixture (FM) model the probability distribution of any voxels grey level, y, is given by:

$$P(y|\theta) = \sum_{t \in T} f(y_i; \theta_t) P(t) \tag{1}$$

where $f(y_i; \theta_l)$ is the model distribution for a tissue t, θ_t are the model parameters and $P(t)$ is the probability of a tissue t occurring. Simply put, each tissue distribution is weighted by the relative amount of tissue found in the image.

In contrast, in the HMRF model the probability distribution of a voxels grey level is dependent on the neighbouring voxels tissue class and not on the the global relative tissue quantity. A given voxels grey level distribution (given the emission functions parameters, θ, and the neighbouring MRF tissue labels, x_{N_i}) is given by:

$$P(y_i|x_{N_i}, \theta) = \sum_{t \in T} f(y_i; \theta_l) P(t|x_{N_i}) \tag{2}$$

where $f(y_i; \theta_l)$ is the emission function for a given tissue type t, θ_l is the emission function parameters for a tissue t and $P(t|x_{N_i})$ is the probability of the voxel being tissue t given the neighbouring voxels tissue labels. In this case each tissue distribution is weighted by the tissue type of the surrounding voxels. Zhang *et al* argue that the FM model can be seen as a degenerative special case of HMRF model where each voxel is independent of surrounding voxels. i.e. $P(t|x_{N_i}) = P(t)$.

It is clear that both models are generalised since no actual distributions for MRI data are specified and considerations/assumptions related to the image formation process must still be made. The work by Zhang *et al* states that the FM model is limited due to the loss of spatial information in moving to the intensity distribution domain and the inherent assumption that all data points are independent samples drawn from a population. Because of this they argue that the FM model only works 'well' in images that have a good signal-to-noise ratio (SNR). They then further argue that MR images do not have good SNR due to artifacts such as partial volume effects and bias field distortion and are thus not suited to the FM model.

The justification for the HMRF model is that spatial information is retained by taking into account neighbouring tissue class labels. We would argue that this is more accurately described as contextual information rather than spatial as directional spatial correlations are ignored. This contextual information is then used to constrain the labelling of a voxel and is a form of smoothing that removes high frequency specular noise. The assumption underlying the method is that there is a true underlying class label map (the MRF) that is corrupted by noise, including measurement error and the quantisation errors leading to PVEs (the emission functions), to produce the final image. The work of Zhang *et al* then assumes the form of the emission functions to be Gaussian and results in what they call the Gaussian Hidden Markov Random Field (GHMRF) model.

To us, this model seems unrelated to the image formation process, i.e. it is not possible to start from a biological model of a region of tissue and use the processes which correspond to the model assumptions to generate image data. Rather it is only an 'effective' (non fundamental) approximation which models empirical intensity distributions and their correlations. In our opinion the correct way to construct a model would require a tissue density (not class label) map (MRF) that includes PV voxels, Gaussian emission functions that model sources of noise related to measurement error, the point spread process, tissue non-uniformity etc, and additional *non-Gaussian* emission functions that model PVEs between pure tissue classes. Of these processes, the one most suitable for modelling by a GHMRF will be the point spread process, which as we mentioned earlier, has probably little effect on quantitative interpretation of data.

As it stands the GHMRF model can only give class labels and thus any volume estimates are, at best, restricted to the resolution of the voxel dimensions but also more likely over-/under estimated due to the incorrect labelling or partial volume voxels. Use of inappropriate models for the modelling of data density distribution will lead to parameter estimation biases as well as a mismatch between the definition of the measurement required (a volume) and the method used to obtain it (class label counts). A later publication by Smith *et al* includes a short paragraph that indicates that the FSL can include PVEs; no technical details are provided and to authors knowledge are not published in any journals. From the documentation provide by the FSL authors on their official website [1], the amount of weighting that the MRF model has on the final segmentation is controlled by a heuristic parameter. This parameter is altered by the user until a *subjective* decision is made on what is visually acceptable. This subjective assessment cannot give a quantitative analysis of the data.

With respect to SPM5 the authors clearly acknowledge that partial volume effects will change the intensity distribution [3]. They further state that more complex models that account for partial volume effects have been developed but no further comments are made. The SPM method uses the simple finite Gaussian mixture model which takes no account of PVEs. The problems caused due to this form the basis of the substantial argument given above for the large errors seen in this model. SPM8 also includes a HMRF method to impose a contextual constraint on the classification of voxels which has been discussed in detail above.

Concluding Remarks

We argue that some of the most popular image segmentation algorithms used, and considered as state-of-the-art, have some short comings in their approach. This is evidenced not only by consideration of the underlying image formation process but also by a number of corroborating reviews which show that at best these algorithms will work to an error of 10%. When the results of these reviews is considered in detail, with respect to the image acquisition parameters, there is evidence of a correlation between algorithm performance and voxel dimension. This correlation seems to be in agreement with what we would expect if partial volume effects are not modelled appropriately.

The segmentation/volumetric measurement process is a problem of finding the correct model and then estimating the parameters of that model so that we may explain the data. If the model does not closely fit the data then there are parts of the data that are not explained and will be incorrectly attributed to model components leading to biases in the estimation of model parameters. Significant problems with CSF volume estimation bias are seen to be removed when PVE terms are included. A competitive evaluation of segmentation algorithms also demonstrated that the best algorithm performance was obtained for the method which contained PVE terms [**?**]. We cannot however conclude that this was the only reason for this result, as many algorithms contained other differences, such as the use of spatial tissue 'priors'. The observation that use of spatial priors is not simply justified by appealing to the authority of 'Bayes Theorem' may come a surprise to some of us. However, the process of gaining increased segmentation stability via the use such 'priors' seems to be suspect in the context of a quantitative analysis (Appendix A).

If we have a model that is capable of fitting the data distribution then it stands to reason that we would like to fit that model as precisely as possible to obtain the most stable segmentation and volumetric measurements. The work presented by Bromiley *et al* derives specific regions of interest to gain improved parameter estimates in a

manner similar to the use of spatial priors, but is careful to avoid introducing bias and sculpting artifacts. The method uses region masks that are de-correlated with tissue boundaries. This ensures that a good (less ambiguous) sample of specific tissue classes are retained within each region without invalidating the general model. It is the proposal to use these samples for improvement of model parameter stability which the following sections of this report address.

Minimum Variance Bound Analysis of Regional Sub-Sampling

The previous section provides the motivation for this work which comes from a need for improved quantitative medical image segmentation methods with respect to magnetic resonance imaging (MRI) of brain tissue. The work presented is based upon a full multi-dimensional model of the intensity distribution of MRI brain image slices developed by Bromiley and Thacker [6]. The model includes a key principled model of partial volume effects and gradient information. They used goodness-of-fit measures that indicated the model was able to fit the data for all but 6% of the voxels. They also demonstrated that biases in SPM that systematically produced significantly higher estimates of CSF volume than is supported by the literature.

In the case of images of brain tissue it may be difficult to assess the robustness and accuracy of an algorithm without a gold standard. However, we believe that such robustness may be achieved by adhering to a methodology that develops models based on sound theoretical principles and makes use of statistical predictions of algorithmic performance. Using such a methodology algorithmic performance on real data may be validated against statistical predictions provided that we *ensure* that any data used adheres to the underlying assumptions of the model. It should be noted that if the assumptions of the model do not match the data (and vice-versa) then all predictions of behaviour are no longer valid and any measurements made cannot be considered as quantitative.

With the previous comments in mind, the work presented in this paper provides an initial theoretical justification for the approach presented in [2], that empirically demonstrated that sub-sampled data may be used to improve the accuracy of the estimation of model parameters over those found from the whole image. This empirical study was careful to address the problems identified with the Brainweb MRI simulator to ensure that results found with the segmentation algorithm used were dependable.

This analytical framework presented here shows that such a technique is statistically valid and demonstrates that the iterative Expectation-Maximisation (EM) algorithm may be used within this framework to provide statistically efficient estimates of the true parameters. One aspect of the approach which may not be immediately obvious is that parameter estimation is improved by making use of spatial information. However, unlike other segmentation approaches the spatial information is not based on an *a-priori* class map; rather that spatial information is derived from the data itself and is thus unbiased.

A common approach to image segmentation is to generate a model of the intensity distribution of the image and to fit the parameters of this model via maximum likelihood (ML) methods. A full model of the intensity distribution of MRI brain image slices, that includes a key principled model of partial volume effects, has been published in *Annals of the BMVA* [6] and presented at a number conferences including Medical Image Understanding and Analysis (MIUA)[7, 14, 13]. The model was developed in such a way that ML estimate of the parameters for this model could be found using a specific implementation of the EM algorithm.

It may be observed that in brain images the different tissue types tend to occur in contiguous blocks and that for some sub-regions of the image the pure tissue peaks in the histogram can be easily identified. It was commented in the previous work that if it were possible to select a-priori all of the voxels containing only one tissue type then this would provide the most efficient estimation of the parameters for that tissue type. The information required to extract such a perfect sub-region is spatial information that relates a particular voxel to its underlying tissue class. In generating a global histogram of an image it is such information that is lost and leads to less efficient parameter estimates than is theoretically possible. We stress that even though spatial information may be contained within image data, it is *not* available a-priori even in normal brain tissue due to inter- and intra-subject variations let alone in subjects with pathological deformations.

The approach taken in [2] is that an *approximation* of the pure tissue regions may be obtained from a first pass segmentation based on the initial global histogram. This first pass segmentation is then used to define sub-regions of the image whose histograms have a prominent peak due to a single tissue type. The sub-regions are defined in such a way that they are still valid in terms of the assumed model since we do not sculpt the histogram using hard thresholds. Since, no histogram sculpting takes place and each sub-region is valid (being boot-strapped from the data) there should be no introduction of bias to the estimation of the parameter and use of the EM algorithm is still a valid approach to finding the ML estimate; the proof of convergence is still valid for each sub-region.

The theoretical approach described in the next section considers an image composed of two tissue types whose intensity distributions are modelled as Gaussian. The distributions model all sources of variation including measurement noise, biological variation and Poisson error on the distribution. The stability of parameter estimates from the data for the intensity distributions is then analysed using the Minimum Variance Bound (MVB) also known as the Cramér-Rao bound. The MVB states that for an unbiased estimator the variance of the estimates is at least as high as the inverse Fisher information matrix. The Fisher information matrix itself is defined as the expectation of the second derivatives of the log-likelihood function with respect to the model parameters.

The following sections derive the analytic equations for the Fisher information matrix (and thus the MVB). This is followed by a section describing and presenting results from a numerical evaluation of these equations and are validated against a monte-carlo simulation that estimates parameters using two versions of the EM algorithm. The first version uses a statistically principled update equation (called Simultaneous Fitting of Region Parameters, SFRP) and is found to have a close match with the derived MVB. The second version uses an empirical update equation that uses only the data available within each region and is shown to be an approximation of the first version under small tissue overlaps. This second version of the EM algorithm (called Combined Fitting, CFRP) is of interest since it is more computationally efficient and also may provide a robust method of handling outlier/pathological tissue data.

Preliminaries

In the following analysis the term pixel is used to denote an individual element of a digital image that in reality has been generated via magnetic resonance imaging techniques. As such it should be kept in mind that the value of each pixel maps to a three dimensional voxel that may be composed of a combination of tissue types. Further to this, even for a set of voxels that are composed purely of one tissue, the density of tissue for any voxel of this set is also varying.

Consider an image that is composed of a set of pixels indexed by a set I such that the grey level of any pixel is given by x_i where $i \in I$. We define an ideal scenario where the image is composed entirely of pixels drawn from two tissue types, T_1 & T_2.

Biological variations are modelled by the grey level of pixels of a given tissue type being drawn from a random variable with a Gaussian distribution of mean, μ_T, and standard deviation, σ_T. Additive measurement noise on all pixels is then modelled as random samples drawn from a random variable with a Gaussian distribution of zero mean and standard deviation, σ_n, and thus models isotropic measurement errors.

A region of an image is defined via an indexing set, R. The quantity of a specific tissue type within a given region is given by $Q_T(R)$. For the analysis that follows we consider that the whole image partitioned into two regions, $R_1 \subset I$ and $R_2 \subset I$. Each region is generated in a way such that it indexes pixels of both tissue types and that the regions are mutually exclusive such that the following conditions are strictly adhered to; $R_1 \cup R_2 = I$ and $R_1 \cap R_2 = \emptyset$. With these restrictions in place the true proportions of the quantities of tissue in each region is given by:

$$Q_{T_1}^{true}(R_1) = Q_{T_1}(I) - \kappa_{ov} \tag{3}$$

$$Q_{T_2}^{true}(R_1) = \kappa_{ov} \tag{4}$$

$$Q_{T_1}^{true}(R_2) = \kappa_{ov} \tag{5}$$

$$Q_{T_2}^{true}(R_2) = Q_{T_2}(I) - \kappa_{ov} \tag{6}$$

where κ_{ov} is a constant that defines the amount of 'overlap' of tissue across the two regions.

For the following analysis we will consider the minimum variance bound (MVB) on the estimation of the mean, μ_{T_1}, of tissue T_1. The log-likelihood over the two regions is given by:

$$lnL(x_i; \boldsymbol{\theta}) = \sum_{n=1}^{N} \sum_{i \in R_n}^{|R_n|} lnP_{R_n}(x_i; \boldsymbol{\theta}) \tag{7}$$

where N is the number of regions (thus for the image as a whole $N = 1$) and:

$$P_{R_n}(x_i; \boldsymbol{\theta}) = \sum_{t \in T}^{|T|} d_t(x_i) Q_t(R_n) \tag{8}$$

T is the set of tissues and $\boldsymbol{\theta}$ is a vector of model parameters; which for a whole image composed of two tissues is defined as:

$$\boldsymbol{\theta} = [\mu_{T_1}, \mu_{T_2}, Q_{T_1}(I), Q_{T_2}(I)]^T \tag{9}$$

and for the two tissue and two region case is defined as:

$$\boldsymbol{\theta} = [\mu_{T_1}, \mu_{T_2}, Q_{T_1}(R_1), Q_{T_2}(R_1), Q_{T_1}(R_2), Q_{T_2}(R_2)]^T \tag{10}$$

The distributions and derivatives for the tissues are given by:

$$d_{T_a}(x_i) = \frac{1}{\sigma_{T_a}\sqrt{2\pi}} e^{-\frac{(x-\mu_{T_a})^2}{2{\sigma_{T_a}}^2}} \tag{11}$$

$$\frac{\partial d_{T_a}(x_i)}{\partial \mu_{T_b}} = \begin{cases} \frac{(x_i-\mu_{T_a})}{{\sigma_{T_a}}^2} d_{T_a}(x_i) & \text{for } a = b \\ 0 & \text{for } a \neq b \end{cases} \tag{12}$$

$$\frac{\partial^2 d_{T_a}(x_i)}{\partial \mu_{T_b}^2} = \begin{cases} \frac{d_{T_a}(x_i)}{\sigma_{T_a}^2}\left(\frac{(x-\mu_{T_a})^2}{\sigma_{T_a}^2} - 1\right) & \text{for } a = b \\ 0 & \text{for } a \neq b \end{cases} \tag{13}$$

Log-Likelihood Derivatives

Derivatives w.r.t mean

First Derivative

The log-likelihood for two regions is given by equation 7 and thus the first derivative of the log-likelihood w.r.t. a tissue mean μ_{T_a} is given by:

$$\frac{\partial lnL(x_i;\boldsymbol{\theta})}{\partial \mu_{T_a}} = \sum_{n=1}^{N} \sum_{i \in R_n}^{|R_n|} \frac{1}{P_{R_n}(x_i;\boldsymbol{\theta})} \cdot \frac{\partial P_{R_n}(x_i;\boldsymbol{\theta})}{\partial \mu_{T_a}} \tag{14}$$

where $\frac{\partial P_{R_n}(x_i;\boldsymbol{\theta})}{\partial \mu_{T_a}}$ is given by (directly from equation 8):

$$\frac{\partial P_{R_n}(x_i;\boldsymbol{\theta})}{\partial \mu_{T_a}} = \frac{(x_i - \mu_{T_a})}{\sigma_{T_a}^2} d_{T_a}(x_i) Q_{T_a}(R_n) \tag{15}$$

Second Derivative

The second derivative of the log-likelihood w.r.t. a tissue μ_{T_b} is given by:

$$\frac{\partial^2 lnL(x_i;\boldsymbol{\theta})}{\partial \mu_{T_a} \partial \mu_{T_b}} = \sum_{n=1}^{N} \sum_{i \in R_n}^{|R_n|} \frac{\partial^2 lnP_{R_n}(x_i;\boldsymbol{\theta})}{\partial \mu_{T_a} \mu_{T_b}} \tag{16}$$

where:

$$\frac{\partial^2 lnP_{R_n}(x_i;\boldsymbol{\theta})}{\partial \mu_{T_a} \partial \mu_{T_b}} = \left[\frac{\partial^2 P_{R_n}(x_i;\boldsymbol{\theta})}{\partial \mu_{T_a} \mu_{T_b}} P_{R_n}(x_i;\boldsymbol{\theta}) - \frac{\partial P_{R_n}(x_i;\boldsymbol{\theta})}{\partial \mu_{T_a}} \cdot \frac{\partial P_{R_n}(x_i;\boldsymbol{\theta})}{\mu_{T_b}}\right] \cdot \frac{1}{P_{R_n}(x_i;\boldsymbol{\theta})^2} \tag{17}$$

$$\frac{\partial^2 P_{R_n}(x_i;\boldsymbol{\theta})}{\partial \mu_{T_a} \partial \mu_{T_b}} = \begin{cases} \frac{d_{T_a}(x_i) Q_{T_a}(R_n)}{\sigma_{T_a}^2}\left(\frac{(x-\mu_{T_a})^2}{\sigma_{T_a}^2} - 1\right) & \text{for } a = b \\ 0 & \text{for } a \neq b \end{cases} \tag{18}$$

Cross-Derivative w.r.t Quantity

$$\frac{\partial^2 lnL(x_i;\boldsymbol{\theta})}{\partial \mu_{T_a} \partial Q_{T_b}(R_c)} = \sum_{n=1}^{N} \sum_{i \in R_n}^{|R_n|} \left[\left(\frac{\partial^2 P_{R_n}(x_i;\boldsymbol{\theta})}{\partial \mu_{T_a} Q_{T_b}(R_c)} P_{R_n}(x_i;\boldsymbol{\theta}) - \frac{\partial P_{R_n}(x_i;\boldsymbol{\theta})}{\partial \mu_{T_a}} \cdot \frac{\partial P_{R_n}(x_i;\boldsymbol{\theta})}{\partial Q_{T_b}(R_c)} \right) \cdot \frac{1}{P_{R_n}(x_i;\boldsymbol{\theta})^2} \right] \tag{19}$$

where:

$$\frac{\partial P_{R_n}(x_i;\boldsymbol{\theta})}{\partial Q_{T_b}(R_c)} = \begin{cases} d_{T_b}(x_i) & \text{for } c = n \\ 0 & \text{for } c \neq n \end{cases} \tag{20}$$

and:

$$\frac{\partial^2 P_{R_n}(x_i;\boldsymbol{\theta})}{\partial \mu_{T_a} \partial Q_{T_b}(R_c)} = \begin{cases} \frac{(x_i - \mu_{T_a})}{\sigma^2_{T_a}} d_{T_a}(x_i) & \text{for } a = b \ \& \ c = n \\ 0 & \text{otherwise} \end{cases} \tag{21}$$

For the cases of only two tissue types, T_a & T_b, we may use the substitution $P_{R_c}(x_i;\boldsymbol{\theta}) - d_{T_a}(x_i)Q_{T_a}(R_c) = d_{T_b}(x_i)Q_{T_b}(R_c)$ to derive the following two equations:
For $a = b$,

$$\frac{\partial^2 lnL(x_i;\boldsymbol{\theta})}{\partial \mu_{T_a} \partial Q_{T_b}(R_c)} = \sum_{i \in R_c} \frac{\frac{(x_i - \mu_{T_a})}{\sigma^2_{T_a}} d_{T_a}(x_i) d_{T_b}(x_i) Q_{T_b}(R_c)}{P_{R_n}(x_i;\boldsymbol{\theta})^2} \tag{22}$$

For $a \neq b$,

$$\frac{\partial^2 lnL(x_i;\boldsymbol{\theta})}{\partial \mu_{T_a} \partial Q_{T_b}(R_c)} = -\sum_{i \in R_c} \frac{\frac{(x_i - \mu_{T_a})}{\sigma^2_{T_a}} d_{T_a}(x_i) d_{T_b}(x_i) Q_{T_a}(R_c)}{P_{R_n}(x_i;\boldsymbol{\theta})^2} \tag{23}$$

Derivatives w.r.t Quantity

First Derivative

From equation 7:

$$\frac{\partial lnL(x_i;\boldsymbol{\theta})}{\partial Q_{T_a}(R_c)} = \sum_{n=1}^{N} \sum_{i \in R_n}^{|R_n|} \frac{1}{P_{R_n}(x_i;\boldsymbol{\theta})} \cdot \frac{\partial P_{R_n}(x_i;\boldsymbol{\theta})}{\partial Q_a(R_c)} \tag{24}$$

where:

$$\frac{\partial P_{R_n}(x_i;\boldsymbol{\theta})}{\partial Q_{T_a}(R_c)} = \begin{cases} d_{T_a}(x_i) & \text{for } n = c \\ 0 & \text{otherwise} \end{cases} \tag{25}$$

thus equation 24 reduces to:

$$\frac{\partial lnL(x_i;\boldsymbol{\theta})}{\partial Q_{T_a}(R_c)} = \sum_{i \in R_c}^{|R_c|} \frac{d_{T_a}(x_i)}{P_{R_c}(x_i;\boldsymbol{\theta})} \tag{26}$$

Second Derivative

$$\frac{\partial^2 lnL(x_i;\boldsymbol{\theta})}{\partial Q_{T_a}(R_c) \partial Q_{T_b}(R_d)} = -\sum_{i \in R_c}^{|R_c|} \left[\frac{\partial P_{R_c}(x_i;\boldsymbol{\theta})}{\partial Q_{T_b}(R_d)}] \cdot \frac{d_{T_a}(x_i)}{P_{R_c}(x_i;\boldsymbol{\theta})^2} \right] \tag{27}$$

Equation 27 is equal to zero if $c \neq d$, and so, even though there is a weak coupling across regions due to the common mean parameters, there is no information related to the data that is contained in the quantity cross-derivatives across regions. Thus, the form of equation ?? reduces to:

$$\frac{\partial^2 lnL(x_i;\boldsymbol{\theta})}{\partial Q_{T_a}(R_c)\partial Q_{T_b}(R_c)} = -\sum_{i\in R_c}^{|R_c|} \frac{d_{T_a}(x_i)d_{T_b}(x_i)}{P_{R_c}(x_i;\boldsymbol{\theta})^2} \tag{28}$$

The MVB and Fisher information matrix

For an unbiased estimated the MVB is given by:

$$cov(\boldsymbol{\theta}) \geq I(\boldsymbol{\theta})^{-1} \tag{29}$$

where $\boldsymbol{\theta}$ is the vector of the model parameters (eqns. 9 & 10) and $I(\boldsymbol{\theta})$ is the Fisher information matrix is given by:

$$I_{m,k}(\boldsymbol{\theta}) = -E\left[\frac{\partial^2}{\partial\theta_m\partial\theta_k} lnL(\boldsymbol{x},\boldsymbol{\theta})\right] \tag{30}$$

Simulations

A monte-carlo simulation of the proposed approach and a numerical evaluation of the MVB have been simulated in Matlab. The following sections provide the technical details of the simulations and the results for the stability of the parameters follows. Both simulations are abstracted away from actual images and evaluated at the histogram level demonstrating the general validity of the approach. Similarly, the units of parameters and data points described in the follow sections are abstracted to grey levels of arbitrary units. However, the distribution mean, spread and quantities have been selected to mimic the typical MR distributions of grey and white matter seen in real images.

Numerical Evaluation of Minimum Variance Bound

The numerical evaluation of the MVB was executed by first defining two 'true' tissue vectors, $\boldsymbol{t}_1$ & $\boldsymbol{t}_2$. The length of each vector was set to $|\boldsymbol{t}_1| = 5000$ & $|\boldsymbol{t}_2| = 10000$ and was composed of data points drawn from Gaussian distributions of standard deviation, $\sigma_{t_n} = 15$, and means $\mu_{t_1} = 100$ & $\mu_{t_2} = [105, 200]$.

To examine the theoretical behaviour of the proposed algorithm the analysis was evaluated are varying amounts of tissue overlap, κ_{ov}, ranging from 100 pixels to 4500 pixels in steps of 100. From the tissue vectors an 'image' vector, $\boldsymbol{I}$, and two region vectors were created, $\boldsymbol{R}_1$ & $\boldsymbol{R}_2$. The image vector was simply the concatenation of the two tissue vectors, $\boldsymbol{I} = \{\boldsymbol{t}_1, \boldsymbol{t}_2\}$. The region vectors were created from the tissue vectors based such that $\boldsymbol{R}_1 = \{t_i \in \boldsymbol{t}_1 | i = [0, n], t_j \in \boldsymbol{t}_2 | j = [0, \kappa_{ov}]\}$ and $\boldsymbol{R}_2 = \{t_i \in \boldsymbol{t}_2 | i = [0, m], t_j \in \boldsymbol{t}_1 | j = [0, \kappa_{ov}]\}$, where $n = |\boldsymbol{t}_1| - \kappa_{ov}$ & $m = |\boldsymbol{t}_2| - \kappa_{ov}$. For consistency with the theoretical framework the region vectors were created without resampling so that the conditions, $R_1 \cup R_2 = I$ and $R_1 \cap R_2 = \emptyset$, are satisfied i.e. all the 'true' tissue data samples are used only once.

Simulations were then run for the whole image vector, $\boldsymbol{I}$, and for the two regions, $\boldsymbol{R}_1$ & $\boldsymbol{R}_2$. The equations are derived on the basis of there being N regions and thus for the image vector $N = 1$. The quantities of tissues $Q_{T_m}(R_n)$ were set initially to the true known values and were then iterated (100 iterations) for consistency with the monte-carlo using:

$$Q_{T_m}(R_n) = \sum_{i\in R_n}^{|R_n|} \frac{P(x_i|T_m)Q_{T_m}(R_n)}{\sum_{m\in T} P(x_i|T_m)Q_{T_m}(R_n)} \tag{31}$$

The Fisher information matrix of equation *** is defined as an expectation. Here the expectation was approximated by taking the average matrices computed over 50 trials. The approximate Fisher information matrices were then inverted using the Matlab pseudo-inverse function (pinv) to give the MVB for each parameter of the model.

Monte-Carlo Evaluation of Parameter Estimation

The framework for the monte-carlo simulation is similar to the simulation for the variance bound in that two tissue vectors, $\boldsymbol{t}_1$ & $\boldsymbol{t}_2$, are defined. In this case the length of the data vectors is subject to a Poisson error so that the vector lengths, $|\boldsymbol{t}_1|$ & $|\boldsymbol{t}_2$ are drawn from Poisson distributions with mean, $\lambda = 5000$ & $\lambda = 10000$ respectively. As previously, the data samples for each tissue are drawn from Gaussian distributions of standard deviation, $\sigma_{t_n} = 15$, and means $\mu_{t_1} = 100$ & $\mu_{t_2} = [105, 200]$.

The whole image vector, $\boldsymbol{I}$, is formed by concatenating the two tissue vectors, $\boldsymbol{I} = \{\boldsymbol{t}_1, \boldsymbol{t}_2\}$. In this case for the creation of the region vectors the actual value of tissue overlap, κ_{ov}, is drawn from a Poisson distribution. For each individual run of the monte-carlo the algorithm is evaluated over range of mean overlap values, λ, ranging from 4500 pixels to 100 pixels. New random tissue vectors are generated for each run and for each value of κ_{ov} to avoid introducing systematic effects into the evaluation.

The maximum likelihood estimation of the parameters was estimated using specific implementations of the Expectation-Maximisation (EM) algorithm. The algorithm was applied in two related but crucially different manners. The first method makes parameter estimates simultaneously for all tissue types using all of the data, and can be related to a theoretical expression for MVB. The second method uses the regional approach to effectively disregard as much data as possible that is not related to the parameters of interest without sculpting the intensity distribution.

In each method the e-step is applied to each region individually using:

$$P(t_m|x_i) = \frac{P(x_i|t_m)Q_{t_m}(R_n)}{\sum_m P(x_i|t_m)Q_{t_m}(R_n)} \tag{32}$$

and:

$$Q_{t_m}(R_n) = \sum_{i \in R_n} P(t_m|x_i) \tag{33}$$

where $P(x_i|t_m)$ is given by the Gaussian distribution for tissue t_m as defined above.

The initial quantities $Q_{t_m}(R_n)$ of each tissue type are derived using the mean values of the Poisson distribution from which they were drawn. The initial mean values for each tissue, mu_{t_m}, were set to a random value drawn from a uniform distribution with mean equal to the true means and a spread of 15 grey levels. The initial values for the mean in each run of the monte-carlo was randomly chose from a uniform distribution around the true mean with a maximum spread of 5 grey levels. The EM algorithm was iterated 150 times at each point in the simulation.

The maximisation step estimates the new mean parameters and it is this area that the two methods differ:

Method 1 - Simultaneous Fitting of Region Parameters (SFRP)

This method uses a statistically valid method of combining the results from each region simultaneously using all of the data. It is this method that is consistent with the MVB derived in the previous section. In this case the estimate of the new parameters is given by:

$$\mu_m = \frac{\sum_{n=1}^{N} \sum_{i \in R_n}^{|R_n|} P(t_m|x_i)x_i}{\sum_{n=1}^{N} \sum_{i \in R_n}^{|R_n|} P(t_m|x_i)} \tag{34}$$

where N is the number of regions. For the case of the whole image, I, the number of regions is 1 and the equations remain valid.

Method 2 - Combined Fitting of Region Parameters (CFRP)

The method of splitting the data into regions is attempting to exploit other information, spatial in the case of images, to create regions that are composed of a majority samples from one specific class. For example in our toy example we have two regions R_A and R_B and two tissue types, T_1, and T_2. Region R_A is constructed such that the majority of samples come from T_1 and a minority from the other tissue. It is this region that is expected to give improved estimates for parameters related to T_1. Likewise region R_B is expected to give improved estimates for parameters related to T_2.

The CFRP method considers each tissue type, T, in turn and only uses the data from the region whose majority of samples are of that tissue type. That data from that region is used to estimate the model parameters for that tissue type θ_T while parameters related to other tissue types are held constant (at the values estimated from the other regions). The full set of model parameters are estimated by considering each tissue type in turn and iterating the whole EM algorithm until convergence.

This method of estimating the tissue parameters is more computationally efficient as it only need to make computations on reduced data sets. The method may also have benefits in the future aim of identifying pathological tissue in MR images as outliers. With this in mind we wish to evaluate the CFRP method with respect to the MVB.

Results

Figure 3 shows the stability of μ_1 found from the theoretical evaluation and from the monte-carlo simulation with respect to differing amounts of tissue overlap using the SFRP. For comparison the results for the whole image and for the region based algorithm are presented. For the whole image the MVB and monte-carlo results are independent of the tissue overlap, κ_{ov}, and this remains constant (to within statistical error). For the analysis by region with overlap values in the range 3000 to 3750 there is no significant improvement over applying the algorithm to the whole image. However, over the rest of the range there is a significant improvement in the stability of μ_1. As the amount of over lapping tissue is reduced the increase in stability improves in an almost linear fashion. This is due to the first region being dominated by tissue, t_1, where at overlaps of 500 and below a more than two fold improvement in stability. Interestingly, improvement in the stability is also seen as the tissue overlap is increased. This can be explained since at the larger values of tissue overlap, the ratio of t_1 to t_2 in region two is tending towards 1 (in the whole image the ratio is 0.5).

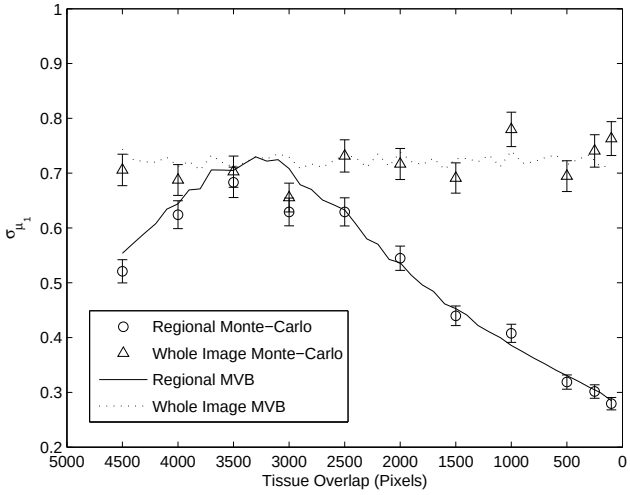


Figure 3: Standard deviation of μ_1 estimate as a function of tissue overlap using SFRP ($\mu_1 = 100$, $\mu_2 = 120$, $\sigma_1 = 15$, $\sigma_2 = 15$).

Figure 4 shows the stability of μ_1 as a function of the second tissue mean, μ_2 i.e. a function of the separation of the two tissues. It can be seen that when the two tissue grey levels are well separated (beyond two standard deviations) the gains made by using the regional approach are minimal. This is expected since each tissue distribution is already well defined and the uncertainty in estimated the parameters is almost entirely due to the inherent characteristics of each tissue and the noise. i.e uncertainty due to overlapping distributions is already minimised.

When the tissue means are within two standard deviations of each other there is a significant improvement in the stability of the μ_1 parameter found with the regional approach over simply using the full image. For the most part the monte-carlo simulation matches the theoretical predictions using the MVB but there is a discrepancy and small amounts of separation less than 10 grey levels. At this point the true tissue means separation is less than the random spread of the initial starting conditions.

Figure 5 shows the same theoretical results found as for those in figure 3 except using the CFRP method. It can be seen that there is a large discrepancy between the theoretical MVB and the CFRP monte-carlo for tissue overlaps greater than 1000 pixels. However, for smaller amount of tissue overlap there is a good match between the two

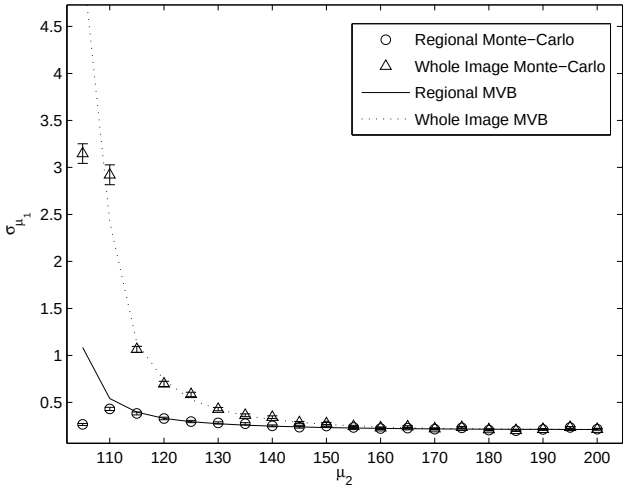


Figure 4: Standard deviation of μ_1 estimate as a function of μ_2 using SFRP ($\mu_1 = 100$, $\sigma_1 = 15$, $\sigma_2 = 15$, tissue overlap = 500).

results and we conclude that at these levels of overlap the CFRP method approximates the statistically principled SFRP method.

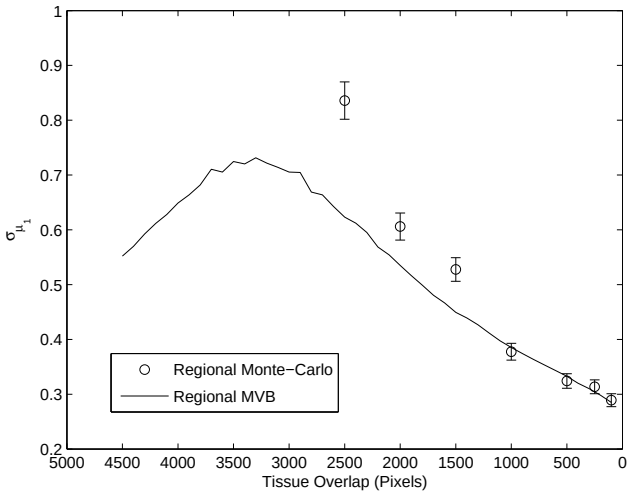


Figure 5: Standard deviation of μ_1 estimate as a function of tissue overlap using CFRP ($\mu_1 = 100$, $\mu_2 = 120$, $\sigma_1 = 15$, $\sigma_2 = 15$).

Conclusions

This work has made an in depth review of the current 'state-of-the-art' image segmentation techniques commonly employed with MR images of brain tissue. It was found that the literature supports evidence that these methods are prone to large errors. In particular, despite providing subjectively acceptable outputs, volumes or regions of interest generated by commonly used software are likely to be contaminated by between 10 and 15% of inappropriately labelled tissue. Results are significantly worse for anisotropic voxels, as often acquired in clinical practice in order to reduce sequence acquisition and clinical review times. This has important consequences for clinical and bio-medical research, whenever these techniques are applied, limiting their applicability.

We believe that once appropriate steps have been taken to ensure good quality data (avoiding 3T machines and avoiding or correcting field inhomogeneity) these large sources of error are due to the use of incorrect models, especially with respect to partial volumes, of the intensity distributions leading to biases and instability in the

estimate of model parameters. If an appropriate model is used, including partial volume distributions, then the problem of making volumetric measurements becomes the problem of making the most accurate estimate of the model parameters; the volume is derived from the model distributions. In this manner estimates of the volumetric error can be derived giving true quantitative measurements.

One common approach for improving parameter stability is the use of spatial 'prior' distributions, which aid in the separation of otherwise ambiguous regions. However, these approaches bias the interpretation of specific datasets in the direction of the assumed average (normal) distribution, and inhibit the measurement of pathological changes. From this point of view we considered empirical work that showed that by using a regional sub-sampling approach, based on an initial segmentation of the MR image, gave an improved stability in the model parameters. This previous work used a simple two tissue Gaussian Monte-Carlo, Brainweb based Monte-Carlo and real data.

In the current work, the MVB bound equations were derived with respect to using an arbitrary number of image regions, the whole image being a special case. For mathematical tractability, the theoretical work was derived within a simplified framework that considers two Gaussian distributed tissue classes. While this model is extremely simplified the main purpose was to provide an initial theoretical justification for regional sub-sampling approach that corroborates the empirical evidence. It was found that the Fisher information available for estimation of parameters can indeed sometimes be increased by discarding ambiguous data, when density functions have significant overlap. A monte-carlo simulation using a more empirical approach (CFRP) was found to approximate the MVB when small amounts of contaminating tissue quantities were used. The specific ranges of parameters which result in improvement are in accordance with typical MRI images of the brain. The improvements in parameter accuracy are predicted to be around a factor of two, and this combined with the additional demonstrated benefits from use of PVE models (as they would have to be to make the more stable estimates meaningful) are expected to make an additional significant improvement over 'state-of-the-art' for MRI brain segmentation.

This approach is of interest for two reasons: 1) it is computationally less expensive than the conventional EM algorithm due to the reduced data size and 2) the exclusion of data not related to known tissues would exclude pathological tissue.

The robust identification of unmodeled/pathological tissue as outliers will be the subject of future investigation.

References

[1] Segmentation and structural analysis. *http://fsl.fmrib.ox.ac.uk/fslcourse*, 2014.

[2] P. a. Bromiley and N. A. Thacker. When Less is More: Improvements in Medical Image Segmentation through Spatial Sub-Sampling. *Tina-Memo*, (2007-005), 2007. Internal Memo.

[3] J. Ashburner and K.J. Friston. Image segmentation. In R.S.J. Frackowiak, K.J. Friston, C. Frith, R. Dolan, K.J. Friston, C.J. Price, S. Zeki, J. Ashburner, and W.D. Penny, editors, *Human Brain Function*. Academic Press, 2nd edition, 2003.

[4] J. Ashburner and K.J. Friston. Unified segmentation. *Neuroimage*, 26:839–851, 2005.

[5] P. A. Bromiley. Problems with the Brainweb MRI Simulator in the Evaluation of Medical Image Segmentation Algorithms, and an Alternative Methodology. *Tina Memo*, (2008-002), 2007. Internal Memo.

[6] P. A. Bromiley and N. A. Thacker. Multi-dimensional Medical Image Segmentation with Partial Volume and Gradient Modelling. *Annals of the BMVA*, 2008(2):1–23, 2008.

[7] P. A. Bromiley, N. A. Thacker, M. L. J. Scott, M. Pokrić, A. J. Lacey, and T. F. Cootes. Bayesian and non-bayesian probabilistic models for medical image analysis. *Image and Vision Computing*, 21(10):851–864, 2003.

[8] K. A. Clark, R. P. Woods, D. A. Rottenburg, A. W. Toga, and J. C. Mazziotta. Impact of Aquisition Protocols and Processing Streams on Tissue Segmentation of T1 Weighted MR Images. *Neuroimage*, 29:185–202, 2006.

[9] D. L. Collins, A. P. Zijdenbos, V. Kollokian, J. G. Sled, N. J. Kabani, C. J. Holmes, and A. C. Evans. Design and Construction of a Realistic Digital Brain Phantom. *IEEE Trans. Med. Eng.*, 17(3):463–468, 1998.

[10] K. Kazemi and N. Noorizadeh. Quantitative Comparison of SPM, FSL, and Brainsuite for Brain MR Image Segmentation. *J. Biomed Phys Eng*, 4(1):13–26, 2014.

[11] F. Klauschen, A. Goldman, V. Barra, A. Meyer-Lindenberg, and A. Lundervold. Evaluation of Automated Brain MR Image Segmentation and Volumetry Methods. *Human Brain Mapping*, 30:1310–1327, 2009.

[12] C. Li, C. Kao, J. C. Gore, and Z. Ding. Minimization of Region-Scalable Fitting Energy for Image Segmentation. *IEEE Trans Imag Process*, 17:1940–1949, 2008.

[13] M. Pokrić, N. A. Thacker, , M. L. J. Scott, and A. Jackson. Multi-dimensional medical image segmentation with partial voluming. In *MIUA*, pages 77–80, 2001.

[14] M. Pokrić, N. A. Thacker, and A. Jackson. The importance of partial voluming in multi-dimensional medical image segmentation. In *MICCAI*, pages 1293–1294, 2001.

[15] N. A. Thacker, P. A. Bromiley, and D. C. Williamson. Multi-dimensional Medical Image Segmentation with Partial Volume and Gradient Modelling. *Tina-Memo*, (2004-009), 2006.

[16] N. A. Thacker, A. Jackson, X. P. Zhu, and K. L. Li. Mathematical Segmentation of Grey Matter, White Matter and Cerebral Spinal Fluid from MR Image Pairs. *Tina-Memo*, (2000-006), 2000. Internal Memo.

[17] O. Tsang, A. Gholipaur, K. Gopinath, R. Briggs, and I. Panahi. Comparison of Tissue Segmentation Algorithms in Neuroimage Analysis Software Tools. In *30th Annual International IEEE EMBS Conference*, pages 3924–3928, 2008.

[18] M. W. Woolrich, S. Jbadi, B. Patenaude, M. Chappell, S. Makni, T. Behrens, C. Beckmann, M. Jenkinson, and S. M. Smith. Bayesian Analysis of Neuroimaging Data in FSL. *Neuroimage*, 45:173–186, 2009.

Appendix A: Bayesian Sensitivity and Bias.

The use of priors in estimation tasks is now widespread across many disciplines. The standard description of Bayesian estimation is the construction of[2]

$$P(\theta|x) \;=\; \frac{P(x|\theta)P(\theta)}{\int P(x|\theta)P(\theta)d\theta}$$

The best value of θ is that which maximises this expression, given the Likelihood distribution $P(x|\theta)$ for measurement x and prior distribution $P(\theta)$.

Bayesian techniques are popular and have many enthusiatic proponents. When challenged on the theoretical basis (and yes there are challenges), most practitioners will resort to "it seems to work" as a justification. The key word here is "seems", and validity of the statement depends on how deeply we are prepared to test this claim. Anyone considering using "priors" to improve their analysis methods should take a some time to understand the consequences. It will be shown below that with a little work it is possible to justify the opposite statement "it doesn't seem to work". With some systems we can considerably simplify the process of incorporating Bayes priors into analysis in order to do this. For Gaussian distributions, the resulting distribution is also a Gaussian and its peak is given by a weighted average of the data and prior mean.

Imagine we have a weighing machine with (Gaussian) accuracy σ, and we use it to measure the weights w_i (our x above) of a group of people with a biological Gaussian distribution (our $P(\theta)$ above) of mean w_0 and standard deviation also σ. Without using Bayes priors to adjust our estimates, the expected distribution of measurements has an S.D. of $\sqrt{\sigma^2 \;+\; \sigma^2} \;=\; \sqrt{2}\sigma$.

If we apply the Bayes prior then each measurement is effectively combined with another 'prior' measurement of equal accuracy. The combination process will yield the average $w_i' \;=\; (w_i + w_0)/2$ for each measurement. Each Bayes estimate will therefore have an apparent reproducibility of $\sigma/2$ while each measurement will also be biased by a factor $(w_i \;-\; w_0)/2$ so than the apparent biological variation (for exact measurements) will be $\sigma/2$. The combined apparent distribution across the entire sample with therefore now be $\sqrt{\sigma^2/4 \;+\; \sigma^2/4} \;=\; \sigma/\sqrt{2}$.

If we are looking at the overall distribution of all sampled data it looks as if we have halved the variance of our system ($\sqrt{2}\sigma \;\rightarrow\; \sigma/\sqrt{2}$). We may be very happy with the result, because our measurements are much better behaved. This is about as far as most people go with their assessment, "it seems to work". But paradoxically, we have also achieved a distribution which, including noise, is narrower than the known biological variation (σ)! This should worry us, we appear to be removing the very thing we seek to measure.

Have we usefully increased the accuracy of our measurement of w_i by replacing it with $(w_i \;+\; w_0)/2$? Imagine that we wish to make a measurement for someone who is 3 S.D. heavier than the average, and wish to say something about the probability that the subject is heavier than the average. (This is analogous to detecting

[2] If this were summarising my own work I would normaly try to explain which if these terms are probabilites and which densities, but I am just summarising here what others do.

pathology in tissue labelling tasks). If we were to take multiple measurements, on average the subject will have a measured weight of $(w_i + w_0)/2 = w_0 + 3\sigma/2$, this is a change of $3\sigma/2$ from the mean with an accuracy of $\sigma/2$ i.e. a 3.D. effect. **This is precicely the same statistical significance we had before using the prior.**

This example is not a specific choice of numbers which provides this behaviour, you can try it for yourself with different starting distributions and values and will see that this is always the case. It is also true for any comparisons between measurements which use the same prior. If we take account of all the changes made to the measurement distribution in our assessment (as we would need to in a scientific summary), any improvement in stability is directly offset by the reduction in signal. The sensitivity to detect change is not improved by the use of priors. Logically we could argue that it would be a very strange result if it did, as any analysis could be then improved by combining with an arbitrary prior. The statistical equivalent of a free lunch.

The use of a prior has mapped our initial variable onto an equivalent one with reduced variation. On average a measurement of $w_0 + 3\sigma$ will be observed to have a biased value of $w_0 + 3\sigma/2$. This bias may be difficult to appreciate. On the face of it, given an appropriate prior mean, an average of measurements from **different** random samples will have no bias. However, repeat measurements of the **same** sample will have a systematic bias towards this mean.

You may feel that the weighing machine example is too simple to really illustrate anything useful[3], but you only need one example of something not working to disprove the theory (Popper). However, you can always construct your own problem and work it through with some numbers. As a suggestion, consider weighing male and female subjects, each with their own prior distribution. Then try to answer the question: "What is the average weight difference observed between two (male and female) subjects who are actually known to be of equal weight?". This problem is equivalent to a great many null hypothesis experiments and you may be suprised at the result. If you keep it simple (assume equal variances on priors and measurements), you will reach a conclusion in less than 1/4 of a page of simple algebra. (The answer is given in the footnote[4]).

The presence of bias will cause problems. Firstly, it will hinder any attempt to combine information from multiple sources. Secondly, it prevents an observation being interpreted as an absolute value. In addition, if the measurement accuracy of the incoming data varies, while the prior distribution is fixed, then the amount of bias will also vary. For engineering applications this may not worry us, and many will point out that we may now have a solution to an ill-posed inverse problem which could not be solved before. But this property has detrimental consequences in quantitative and scientific applications, where **getting a unique estimate is less important than honestly summarising the information content of data**.

The problem has arisen because the co-hort of samples assumed in the probability theory is not applicable to practical use of measurements. For the weighing machine the priors need to be equivalent to unbiased estimates specific to each individual. In the case of image segmentation our x is the grey level image data and θ is now a tissue class label or tissue proportion within a voxel, we must be able to construct a spatial prior so that it is specific to each subject and constitutes an unbiased estimate of what is expected at each voxel location. In contrast, we could have taken a second independent measurement and taken the average. This approach does not introduce bias and would provide a genuine increase in our statistical power by $\sqrt{2}$. For MR image segmentation the use of a different protocol will be more *statistically efficient*.

In conclusion; the standard claim that the noise distribution on measurements is reduced by use of priors, which gives rise to opinions that "it seems to work", does not take account of the accompanying bias (or systematic error). Use of priors in this way converts some of the statistical measurement error (which is directly observable and relatively easy to deal with) into a systematic error (which is not observable and difficult to deal with). In any quantitative application (such as science or medicine), which requires use of; hypothesis tests, parameter covariances, or simply an absolute measurement, "it doesn't seem to work".

We are not claiming here that all uses of Bayes' theorem are flawed, only that it is the responsibility of the researcher to see to it that it does something meaningful. In order to use priors for quantitative estimation the process should not be interpretable as an inappropriate combination of a measurement with the mean of a general distribution (also referred to as MAP estimation). To use such an approach you must expect to have to test it, generally with a suitable Monte-Carlo simulation, in order to check that there is nothing untoward happening. Unfortunately, use of spatial priors during image segmentation has been shown to be analogous to our weighing machine example. While it might superficially increase the visual stability of segmentations, and reduce labelling errors across the total dataset, it will not increase the information available for detecting change. Segmentation

[3] I use this example because I was once told by a peer at a conference that, as "it seems to work", a scale manufacturer could improve his sales figures by using this method, and therefore he (and anyone else) would need to be stupid not to do so.

[4] With the obvious notation $\Delta w = (w_f - w_m)/2.0$

biases are well known in the literature, but what may not be well known is that they will vary in an unknown way across a segmentation map and between subjects, i.e. for quantitative purposes the "measurements" are worse not better.

Some might wonder why this appendix is fixated on the bias induced by use of a prior, after-all the obvious methodological alternative (Likelihood) is regularly described as biased, and we are not making a fuss about that. Some will use this argument as an excuse to ignore any problem with bias. There are two issues here, if we investigate the Likelihood bias issue, its origin is completely different and can be dealt with appropriately (Tina memo 2005-008). The bias is not attributable to Likelihood as a concept but to the "inappropriate" averaging of parameter estimates. When used carefully (making sure the assumed distributions match those in the data, the models are correct and any optimisation found a useful solution), Likelihood is a scientifically valid way to summarise measurements. Multiple estimates should not just be averaged, but combined using the separate Likelihood functions. But more simply, it is impossible to make a solid logical case for using a flawed method on the basis of not knowing how to use another one properly.