Model-Based Image Segmentation via Monte Carlo EM, with Application to DCE-MRI

Dawn B. Woodard, Cornell University
Roseline Bilina Falafala, Columbia University
Ciprian Crainiceanu, Johns Hopkins University

Abstract

We introduce an improved method for spatial model-based clustering, and use it to segment three-dimensional Dynamic Contrast Enhanced Magnetic Resonance (DCE-MR) images. Our approach extends an existing Monte Carlo Expectation-Maximization method for Markov random field mixture models, and is guaranteed to converge to a local maximum of the likelihood. Our first extension is to show how to incorporate cluster weight parameters in a computationally tractable way; these parameters are needed to accurately capture small features in the image. Secondly, we incorporate a covariance decomposition to allow control over geometric characteristics of the segmentation. Thirdly, we give a consistent approximation to the observed-data likelihood.

We apply our method to segment DCE-MR images of a subject with multiple sclerosis. In DCE-MR imaging the concentration of a contrast agent in the brain is monitored over time; we segment the brain according to this concentration trajectory. Unlike existing methods, ours yields medically informative segmentations for this application, for instance accurately identifying a particular type of multiple sclerosis lesion.

Keywords: Markov random field, Magnetic Resonance Imaging, model-based clustering, Expectation-Maximization.

1 Introduction

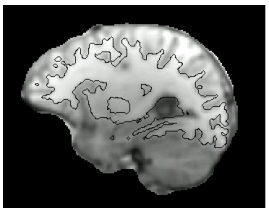
Image segmentation consists of partitioning an image into possibly non-contiguous regions, within which the measurement values are relatively homogeneous (Figure 1(e)). It can fa-

cilitate interpretation or further analysis of the image, and has numerous applications in medicine (Pham et al., 2000), video and image compression (Bosch et al., 2011), and remote sensing (Deng and Clausi, 2004). Model-based clustering provides a formal statistical approach to image segmentation, allowing natural characterization of the regions and extension to other contexts such as time-course imaging, multiple subjects, non-Euclidean measurements, and longitudinal data (Robinson et al., 2010; Zhang et al., 2010b).

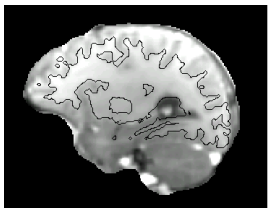
The images are typically two- or three-dimensional, and consist of measurement(s) at each of a grid of voxels (the general term for a pixel in ≥ 2 dimensions). A popular approach to segmentation uses a mixture model based on a Markov random field; this captures the spatial association of the voxels, but is computationally challenging, due in part to an unknown normalizing constant in the likelihood.

We extend an existing Monte Carlo Expectation-Maximization (MCEM) method for image segmentation based on Markov random fields, and apply it to three-dimensional Dynamic Contrast Enhanced Magnetic Resonance (DCE-MR) images of a subject with multiple sclerosis (MS). First, we show how to incorporate cluster weight parameters in a computationally tractable way. Second, we incorporate a covariance decomposition into the model for the mixture components. Third, we give a consistent approximation to the observed-data likelihood (an approximation that converges to the true value in the limit of the number of Monte Carlo samples). Our method, unlike existing approaches, accurately distinguishes small features like MS brain lesions, and yields medically informative segmentations in our application. The computational and statistical challenges in the context of DCE-MRI include: 1. handling three-dimensional images, instead of a two-dimensional slice as in most previous work using Markov random fields; 2. the number of voxels, which is an order of magnitude larger than in most previous work; and 3. the time-series nature of the data.

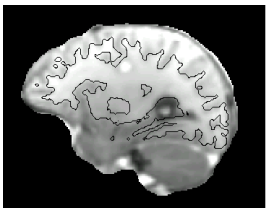
A MCEM method for approximate maximum likelihood estimation in Markov random field mixture models was given in Forbes and Fort (2007; Section 5) and Zhang et al. (2008). Unlike nearly all existing approaches, this method is guaranteed to converge to a local maximum of the likelihood function (Forbes and Fort, 2007). By contrast, the pseudolikelihood, mean-field, and related approximations used in other computational methods (Celeux et al., 2003; Van Leemput et al., 2003; Alfó et al., 2008) either introduce systematic bias (in the



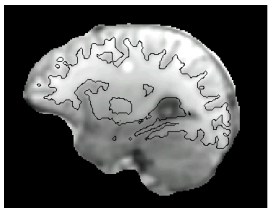
(a) Pre-injection



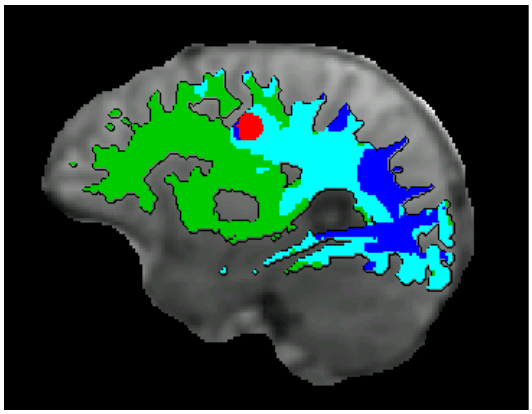
(b) 2 min. post-injection



(c) 30 min. post-inj.



(d) 153 min. post-injection



(e) Segmentation slice

Figure 1: (a)-(d): A two-dimensional slice of a DCE-MRI study at four time points. The black boundary delineates white matter; the bright spot inside the white matter in (c)-(d) corresponds to a multiple sclerosis lesion. (e): The same slice of a three-dimensional segmentation of the white matter, obtained using our method.

case of the pseudolikelihood; Geyer and Thompson 1992), are not guaranteed to converge, or converge to a fixed point that does not correspond to a local maximum of the likelihood (in the case of the mean-field variant in Forbes and Fort 2007).

Our first extension is to include cluster weight parameters; these are critical for capturing small but potentially important features in the image, such as tumors or lesions (Celeux et al., 2004). Many previous Markov random field approaches do not use cluster weight parameters, or fix them at pre-determined values (Besag, 1986; Celeux et al., 2003; McGrory et al., 2009; Zhang et al., 2010b). This is due to the fact that the unknown normalizing constant is a function of these additional parameters, complicating estimation. Our approach addresses this by using approximations to the gradient and Hessian of the normalizing constant, which are both more useful numerically and easier to obtain accurately than the normalizing constant itself.

The weight parameters have been previously used with other computational techniques, including those utilizing the mean-field approximation (Celeux et al., 2004), pseudolikelihood (Van Leemput et al., 2003) and dependency relaxations (Friel et al., 2009). However,

our method appears to be the first to incorporate these parameters while guaranteeing convergence to a local maximum of the likelihood function, or to another estimator of the parameters that has a clear statistical interpretation.

Our second extension is to incorporate the covariance decomposition of Banfield and Raftery (1993) and Celeux and Govaert (1995) into the mixture model. This allows control over geometric characteristics of the segmentation. Our third extension is to give a consistent approximation to the observed-data likelihood, by using thermodynamic integration. This is useful in part because MCEM should be applied using multiple restarts, keeping the parameter values with the highest observed-data likelihood. While an approximation was described in Zhang et al. (2008), it was based on an incorrect expression for the observed-data likelihood, and is not consistent (Section 2).

One could also consider Bayesian estimation for our model; Bayesian approaches for other spatial clustering models are introduced by Johnson and Piert (2009) and Zhang et al. (2010a,b). However, due to the multimodality of the posterior distribution for mixture models, this would require both a sophisticated Markov chain method and a large number of iterations (Woodard and Goldszmidt, 2011). The high computational cost of likelihood evaluation and normalizing constant approximation, together with the large scale of our application, makes such an approach very challenging, so we focus here on MCEM.

We show in a simulation study that our method, unlike the existing MCEM method, is able to accurately capture regions of unequal sizes. We further illustrate this advantage in an application to DCE-MR images of a MS subject. In DCE-MR imaging the subject is injected with a contrast agent, and the concentration of the agent in the tissue is monitored over time using a series of MR images. This contrast enhancement improves viewing of some features of interest. For instance, MS subjects have brain lesions, which are associated with increased permeability of the blood-brain barrier; a primary method for characterizing these lesions is by observing the diffusion of a contrast agent into the brain tissue (Grossman et al., 1988). A two-dimensional slice of a DCE-MRI study at four time points is shown in Figure 1 (a)-(d); the region of primary interest is the white matter, delineated by the black boundary. A lesion is visible in (c)-(d) as a bright spot inside the white matter, where the contrast agent concentration is high; such lesions are called "enhancing lesions." Enhancement is a hallmark

of newly forming lesions, so the characteristics of enhancing lesions are a critical component of diagnostic criteria and outcome measures in clinical trials (Polman et al., 2011).

Current practice is for a specialist to manually identify enhancing lesions from a DCE-MRI study (Shinohara et al., 2012). Additionally, Shinohara et al. (2011) develop context-specific methods for automatic identification of enhancing lesions. Here we instead focus on whole-image segmentation techniques based on the trajectory of MR intensities over time. A two-dimensional slice of an example segmentation using our approach is shown in Figure 1(e). The segmentation automatically distinguishes the enhancing lesions, and identifies other distinct regions within the white matter, which may have relevance in terms of the disease process.

In Section 2 we introduce the clustering model for images and in Section 3 we describe our computational methods. The simulation study is given in Section 4, and Section 5 addresses the DCE-MRI application. We draw conclusions in Section 6.

2 Modeling

An image consists of vector-valued observations $\mathbf{Y}_i$, where $i = 1, \ldots, n$ indexes over a grid that is typically two- or three-dimensional. The vectors $\mathbf{Y}_i$ can represent for instance red, green, blue values in a color image (Panjwani and Healey, 1995) or multiple channels in an MRI image (Choi et al., 1991). The goal of segmentation is to assign to each voxel i a group membership $Z_i \in \{1, \ldots, K\}$; the number of groups K can either be fixed or estimated from the data. The assigned Z_i values ideally should achieve two goals: 1. voxels in the same cluster should have similar values of the observed vector $\mathbf{Y}_i$; and 2. the resulting segmentation of the image should not appear noisy.

Model-based approaches to clustering typically utilize a mixture model, most commonly

$$\mathbf{Y}_i | \mathbf{p}, \boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_K \overset{\text{iid}}{\sim} \sum_{k=1}^{K} p_k f(\mathbf{y}; \boldsymbol{\gamma}_k) \qquad i = 1, \ldots, n \tag{1}$$

where $f(\cdot; \cdot)$ is a density function parameterized by its second argument, $\boldsymbol{\gamma}_k$ is the parameter vector associated with mixture component k, and $\mathbf{p} = (p_1, \ldots, p_K)$ are unknown mixing proportions satisfying $p_k > 0$ and $\sum_{k=1}^{K} p_k = 1$. In a typical context $\mathbf{Y}_i \in \mathbb{R}^d$ is d-dimensional,

in which case a useful choice of f is given by the multivariate Gaussian distribution $f(\mathbf{y};\boldsymbol{\gamma}_k) = N_d(\mathbf{y};\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)$, having unknown mean vector $\boldsymbol{\mu}_k$ and covariance matrix $\boldsymbol{\Sigma}_k$.

Model (1) can be used to cluster the voxels of an image, by: 1. obtaining approximate maximum likelihood estimators of the model parameters $(\mathbf{p},\gamma_1,\ldots,\gamma_K)$ by Expectation-Maximization (EM), then 2. assigning each observation i to the highest-probability cluster k, given the parameter estimates and the observation $\mathbf{y}_i$. However, the segmentation is typically noisy; an example is given in the bottom-left image of Figure 2. This is due to the inaccurate assumption in (1) of independence between adjacent voxels, which is addressed in the following alternative model.

Conditional on the cluster membership $Z_i \in \{1,\ldots,K\}$, we assume that the observation $\mathbf{Y}_i$ at voxel i has a multivariate Gaussian distribution:

$$\mathbf{Y}_i|Z_i,\boldsymbol{\mu},\boldsymbol{\Sigma} \sim N_d(\boldsymbol{\mu}_{Z_i},\boldsymbol{\Sigma}_{Z_i}) \qquad\qquad i=1,\ldots,n \tag{2}$$

independently across i, where $\boldsymbol{\mu} = (\boldsymbol{\mu}_1,\ldots,\boldsymbol{\mu}_K)$ and $\boldsymbol{\Sigma} = (\boldsymbol{\Sigma}_1,\ldots,\boldsymbol{\Sigma}_K)$ are unknown. All methods described here also extend to the case of multivariate t-distributions, by utilizing ideas from Peel and McLachlan (2000), and to other families of distributions. We model the unobserved cluster memberships $\mathbf{Z} = \{Z_i\}_{i=1}^n$ using a *Markov random field*, where $i \sim j$ indicates that voxel j is a spatial neighbor of voxel i and $\sum_{i\sim j}$ indicates a sum over all pairs of voxels that are neighbors:

$$\Pr(\mathbf{Z}|\phi,\mathbf{p}) = g(\phi,\mathbf{p})^{-1}\exp\left\{\phi\sum_{i\sim j}1_{\{Z_i=Z_j\}} + \sum_{i=1}^n \log p_{Z_i}\right\}. \tag{3}$$

The parameters p_k still have the restrictions $p_k > 0$ and $\sum_{k=1}^K p_k = 1$, and the normalizing constant is $g(\phi,\mathbf{p}) \equiv \sum_{\mathbf{z}}\exp\left\{\phi\sum_{i\sim j}1_{\{z_i=z_j\}} + \sum_i \log p_{z_i}\right\}$. Here we use the first-order neighbors, meaning that $i \sim j$ indicates that j is one of the (at most six) voxels that share a face with voxel i. The unknown parameter $\phi \geq 0$ measures the tendency of neighboring voxels to belong to the same cluster. When $\phi = 0$ the model (2)-(3) reduces to the mixture model (1) (since in this case $\Pr(\mathbf{Z}|\phi,\mathbf{p}) = \prod_{i=1}^n p_{Z_i}$, so the cluster memberships Z_i are independent across i, and marginalizing over Z_i yields $\mathbf{Y}_i|\mathbf{p},\boldsymbol{\mu},\boldsymbol{\Sigma} \overset{\text{iid}}{\sim} \sum_{k=1}^K p_k N_d(\mathbf{y};\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)$).

Model (2)-(3) captures distinct groups of voxels while taking into account spatial association. While the term $\sum_{i=1}^n \log p_{Z_i}$ is necessary for (2)-(3) to form an extension of the mixture

model (1), it is typically omitted from Markov random field models for image analysis, or the value of $\mathbf{p}$ is taken to be known. However, without the term $\sum_{i=1}^{n} \log p_{Z_i}$ there is a strong tendency for the model to yield clusters of similar sizes, which is undesirable in cases where there is a distinct subgroup in the data that is either much larger or smaller than n/K (Celeux et al., 2004). The term $\sum_{i=1}^{n} \log p_{Z_i}$ has been previously incorporated by Celeux et al. (2004), where a mean-field approximation is used. However, their method does not have convergence guarantees, and diverged when applied to our three-dimensional images; the value of ϕ increased without bound. This divergence issue also occurred when applying the approach of Besag (1986), which uses a pseudolikelihood approach. To understand this issue, we show in Web Appendix A that the maximum pseudolikelihood estimate of ϕ can be infinite when the maximum likehood estimate is finite, in the context of a simplified version of model (3) in which $\mathbf{Z}$ is directly observed. The mean-field approximation is related to the pseudolikelihood in the sense that both replace the joint distribution (3) with an approximation factorized over i, which may explain why the two approaches are diverging in the same situation.

The geometric features of the clusters are controlled by the covariance matrices $\mathbf{\Sigma}_k$, and by making restrictions on these matrices one can control which features are assumed to be the same across clusters, and which are allowed to vary. For a Gaussian distribution the contours of constant density in $\mathbb{R}^d$ are ellipses, and the volume, shape, and orientation of the ellipses can be separately controlled. To do this we use the eigenvalue decomposition

$$\mathbf{\Sigma}_k = \lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}_k^T \tag{4}$$

(Celeux and Govaert, 1995), where $\lambda_k = |\mathbf{\Sigma}_k|^{1/d}$, $\mathbf{D}_k$ is the matrix of eigenvectors of $\mathbf{\Sigma}_k$, and $\mathbf{A}_k$ is a diagonal matrix such that $|\mathbf{A}_k| = 1$. The quantity λ_k controls the volume (spread) of the kth cluster, $\mathbf{A}_k$ its shape (circular to strongly ellipsoidal) and $\mathbf{D}_k$ its orientation (the direction of the ellipse). For instance, it is common to restrict to spherical clusters, taking $\mathbf{A}_k = \mathbf{D}_k = \mathbf{I}$ but allowing the volume λ_k to vary across clusters. The model where the clusters are spherical ($\mathbf{A}_k = \mathbf{D}_k = \mathbf{I}$) and have the same volume ($\lambda_k = \lambda$) is closely related to the K-means clustering method (Fraley and Raftery, 2002).

We report results for a range of values of K; in our medical imaging context the in-

terpretability of the results is the most relevant criterion for selecting K. Standard model selection criteria, such as information criteria or the posterior probability of K, can yield impractically large values of K in the context of model-based clustering with large datasets. This is because in real data the true cluster distributions are not precisely Gaussian, and their small deviations from Gaussians are better represented in the data as n grows, so additional mixture components are included to capture these small deviations (Biernacki et al., 2000; Forbes et al., 2006; Baudry et al., 2010). This effect is particularly pronounced when the observations $\mathbf{Y}_i$ are multivariate (Wehrens et al., 2004), as in our application. An alternative model selection criterion that addresses this issue was proposed by Cucala and Marin (2012), and could potentially be adapted to our model if an numerical criterion is desired.

The normalizing constant $g(\phi, \mathbf{p})$ in (3) is intractable to compute directly because it is a sum over K^n terms. The numerical methods we introduce use an approximation to the gradient and Hessian of g. It will also be helpful to have an approximation to the observed-data likelihood $L_{obs}(\phi, \mathbf{p}, \boldsymbol{\mu}, \boldsymbol{\Sigma}|\mathbf{y}) = \sum_{\mathbf{z}} \Pr(\mathbf{Z} = \mathbf{z}|\phi, \mathbf{p}) \prod_{i=1}^{n} N_d(\mathbf{y}_i; \boldsymbol{\mu}_{z_i}, \boldsymbol{\Sigma}_{z_i})$, which poses a similar computational challenge. We obtain one using thermodynamic integration; see Section 3.1. A different approximation to the observed-data likelihood in a model related to ours is suggested by Zhang et al. (2008). However, this approximation appears to be based on an incorrect expression for the observed-data likelihood (given on their p.741); in particular, they appear to have provided an approximation to the expected complete-data log-likelihood, and used it as an approximation to the observed-data log-likelihood.

3 Computational Methods

To obtain a segmented image using model (2)-(3), we first find approximate maximum likelihood estimates (MLEs) $\hat{\boldsymbol{\theta}}$ of the parameters $\boldsymbol{\theta} \equiv (\phi, \mathbf{p}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$. Then we obtain estimated cluster memberships $\mathbf{Z}$, by maximizing the posterior probability of $\mathbf{Z}$ conditional on $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$.

Monte Carlo EM. We introduce a MCEM method to approximate the MLE of $\boldsymbol{\theta}$. This method is an extension of the MCEM algorithm given in Zhang et al. (2008) and Forbes and Fort (2007) for a model that is similar to (2)-(3) but does not include the parameters $\mathbf{p}$ or

the decomposition $\boldsymbol{\Sigma}_k = \lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}_k^T$. Our algorithm is guaranteed to converge to a local maximum of the likelihood, by applying the same argument as in Forbes and Fort (2007).

Since the likelihood function for mixture models can be multimodal (even after handling the non-identifiability of the cluster labels; Woodard and Goldszmidt 2011), we run MCEM twice from different and carefully chosen starting positions and use the final parameter values that have the highest observed-data likelihood. We did not find any sensitivity of our method to the initialization on real data, but there was some in the simulated data due to wide spacing between the true mixture components.

Our first initialization takes $\phi = 0$ and sets $\mathbf{p}$, $\boldsymbol{\mu}$, and $\boldsymbol{\Sigma}$ equal to their estimates from the non-spatial mixture model (1) with normal densities $f(\mathbf{y}; \boldsymbol{\gamma}_k) = N_d(\mathbf{y}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$. If there are restrictions on the covariance matrices in the spatial model, we use the same restrictions in the non-spatial model. The second initialization obtains estimates of $\mathbf{p}$, $\boldsymbol{\mu}$, and $\boldsymbol{\Sigma}$ from the same non-spatial model but with equal and diagonal covariance matrices, $\boldsymbol{\Sigma}_k = \lambda \mathbf{I}$. This is the mixture-model analogue of K-means, as discussed in Section 2. The first initialization closely matches the spatial model of interest, while the second has clusters that are spherical and have equal volume in the observation space $\mathbb{R}^d$. The non-spatial mixture models are fit using EM and initialized using hierarchical agglomerative clustering (Fraley and Raftery, 2002) applied to a subsample of the voxels.

In our context, the complete-data vector is $(\mathbf{y}_1, \ldots, \mathbf{y}_n, z_1, \ldots, z_n)$, so the complete-data log-likelihood is $\ell_{\text{comp}}(\boldsymbol{\theta}|\mathbf{y}, \mathbf{Z} = \mathbf{z}) = \ell_1(\boldsymbol{\mu}, \boldsymbol{\Sigma}) + \ell_2(\phi, \mathbf{p})$ where

$$\ell_1(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \equiv \sum_{k=1}^{K} \sum_{i=1}^{n} 1_{\{z_i = k\}} \left[-\frac{1}{2} \log |\boldsymbol{\Sigma}_k| - \frac{1}{2} (\mathbf{y}_i - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_k) \right]$$

$$\ell_2(\phi, \mathbf{p}) \equiv \phi \sum_{i \sim j} 1_{\{z_i = z_j\}} + \sum_{k=1}^{K} (\log p_k) \big(\sum_{i=1}^{n} 1_{\{z_i = k\}} \big) - \log g(\phi, \mathbf{p}).$$

The E step in the tth iteration of EM calculates the conditional expectation of $\ell_{\text{comp}}(\boldsymbol{\theta}|\mathbf{y}, \mathbf{Z} = \mathbf{z})$ with respect to $\Pr(\mathbf{Z}|\mathbf{y}, \boldsymbol{\theta}^{(t)}) \propto \Pr(\mathbf{Z}|\phi^{(t)}, \mathbf{p}^{(t)}) \prod_{i=1}^{n} N_d(\mathbf{y}_i; \boldsymbol{\mu}_{Z_i}^{(t)}, \boldsymbol{\Sigma}_{Z_i}^{(t)})$, where $\boldsymbol{\theta}^{(t)} = (\phi^{(t)}, \mathbf{p}^{(t)}, \boldsymbol{\mu}^{(t)}, \boldsymbol{\Sigma}^{(t)})$ is the current value of the parameter vector. We sample some number S_t of configuration vectors $\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(S_t)} \sim \Pr(\mathbf{Z}|\mathbf{y}, \boldsymbol{\theta}^{(t)})$ using the Swendsen-Wang algorithm, which is an efficient Markov chain method designed for Markov random field models (Swendsen and Wang, 1987). Letting $a_{ik} \equiv \Pr(Z_i = k|\mathbf{y}, \boldsymbol{\theta}^{(t)})$, we use these samples to obtain

Monte Carlo estimates of $\sum_{i=1}^n a_{ik}$, $\sum_{i=1}^n a_{ik}\mathbf{y}_i$, $\sum_{i=1}^n a_{ik}\mathbf{y}_i\mathbf{y}_i^T$, and $E\left[\sum_{i\sim j} 1_{\{Z_i=Z_j\}} \;\middle|\; \mathbf{y}, \boldsymbol{\theta}^{(t)}\right]$, respectively:

$$T_{1k}^{(t)} = S_t^{-1}\sum_{s=1}^{S_t}\sum_{i:z_i^{(s)}=k} 1 \qquad \mathbf{T}_{2k}^{(t)} = S_t^{-1}\sum_{s=1}^{S_t}\sum_{i:z_i^{(s)}=k} \mathbf{y}_i$$

$$\mathbf{T}_{3k}^{(t)} = S_t^{-1}\sum_{s=1}^{S_t}\sum_{i:z_i^{(s)}=k} \mathbf{y}_i\mathbf{y}_i^T \qquad T_4^{(t)} = S_t^{-1}\sum_{s=1}^{S_t}\sum_{i\sim j} 1_{\{z_i^{(s)}=z_j^{(s)}\}}. \tag{5}$$

These yield estimates for the terms of the expected complete-data log-likelihood:

$$\hat{E}\left(\ell_1(\boldsymbol{\mu},\boldsymbol{\Sigma}) \mid \mathbf{y}, \boldsymbol{\theta}^{(t)}\right) = \sum_{k=1}^K -\frac{T_{1k}^{(t)}}{2}\log|\boldsymbol{\Sigma}_k| - \frac{1}{2}\sum_{k=1}^K \mathrm{tr}\left[\boldsymbol{\Sigma}_k^{-1}\left(T_{3k}^{(t)} - 2\boldsymbol{\mu}_k T_{2k}^{(t)T} + \boldsymbol{\mu}_k\boldsymbol{\mu}_k^T T_{1k}^{(t)}\right)\right]$$

$$\hat{E}\left(\ell_2(\phi,\mathbf{p}) \mid y, \boldsymbol{\theta}^{(t)}\right) = \phi T_4^{(t)} + \sum_{k=1}^K (\log p_k) T_{1k}^{(t)} - \log g(\phi,\mathbf{p}). \tag{6}$$

In the M step these functions are maximized over the parameter vector $\boldsymbol{\theta}$. The function $\hat{E}\left(\ell_1(\boldsymbol{\mu},\boldsymbol{\Sigma}) \mid y, \boldsymbol{\theta}^{(t)}\right)$ is largest when $\boldsymbol{\mu}_k^{(t+1)} \equiv \frac{T_{2k}^{(t)}}{T_{1k}^{(t)}}$. Plugging this in and maximizing over $\boldsymbol{\Sigma}$ we find that for $\boldsymbol{\Sigma}_k$ unrestricted, $\boldsymbol{\Sigma}_k^{(t+1)} = \frac{1}{T_{1k}^{(t)}}\left(T_{3k}^{(t)} - \frac{T_{2k}^{(t)}T_{2k}^{(t)T}}{T_{1k}^{(t)}}\right)$. Methods to maximize $\hat{E}\left(\ell_1(\boldsymbol{\mu},\boldsymbol{\Sigma}) \mid y, \boldsymbol{\theta}^{(t)}\right)$ over $\boldsymbol{\Sigma}$ using the decomposition (4) are given in Celeux and Govaert (1995). For the DCE-MRI application we will require $\lambda_k = \lambda$, in which case $\boldsymbol{\Sigma}_k^{(t+1)} = \lambda C_k$ where $\lambda = \frac{\sum_{k=1}^K |W_k|^{1/d}}{n}$, $C_k = \frac{W_k}{|W_k|^{1/d}}$, and $W_k = \left(T_{3k}^{(t)} - \frac{T_{2k}^{(t)}T_{2k}^{(t)T}}{T_{1k}^{(t)}}\right)$.

We maximize $\hat{E}\left(\ell_2(\phi,\mathbf{p}) \mid y, \boldsymbol{\theta}^{(t)}\right)$ numerically in $(\phi, \log p_1, \ldots, \log p_K)$, by approximating the gradient and Hessian and using Newton-Raphson. This yields the global maximum because $\hat{E}\left(\ell_2(\phi,\mathbf{p}) \mid y, \boldsymbol{\theta}^{(t)}\right)$ is concave in $(\phi, \log p_1, \ldots, \log p_K)$ (since (3) is an exponential family distribution, $-\log g(\phi,\mathbf{p})$ is concave in the natural parameter vector). The maximum is only unique up to an additive constant for $\log p_1, \ldots, \log p_K$, so we first do maximization without the restriction that $\sum_{k=1}^K p_k = 1$, then normalize the resulting vector $\mathbf{p}$. The gradient is given by

$$\frac{\partial}{\partial\phi}\hat{E}\left(\ell_2(\phi,\mathbf{p}) \mid y, \boldsymbol{\theta}^{(t)}\right) = T_4^{(t)} - \frac{\partial}{\partial\phi}\log g(\phi,\mathbf{p})$$

$$\frac{\partial}{\partial \log p_k}\hat{E}\left(\ell_2(\phi,\mathbf{p}) \mid y, \boldsymbol{\theta}^{(t)}\right) = T_{1k}^{(t)} - \frac{\partial}{\partial \log p_k}\log g(\phi,\mathbf{p}) \qquad k \in \{1,\ldots,K\}$$

where

$$\frac{\partial \log g(\phi, \mathbf{p})}{\partial \phi} = \frac{\sum_{\mathbf{z}} \exp\{\phi \sum_{i\sim j} 1_{\{z_i=z_j\}} + \sum_i \log p_{z_i}\} \sum_{i\sim j} 1_{\{z_i=z_j\}}}{\sum_{\mathbf{z}} \exp\{\phi \sum_{i\sim j} 1_{\{z_i=z_j\}} + \sum_i \log p_{z_i}\}} = E_{\mathbf{Z}}\left[\sum_{i\sim j} 1_{\{Z_i=Z_j\}}\right]$$

$$\frac{\partial \log g(\phi, \mathbf{p})}{\partial \log p_k} = \frac{\sum_{\mathbf{z}} \exp\{\phi \sum_{i\sim j} 1_{\{z_i=z_j\}} + \sum_i \log p_{z_i}\} \sum_i 1_{\{Z_i=k\}}}{\sum_{\mathbf{z}} \exp\{\phi \sum_{i\sim j} 1_{\{z_i=z_j\}} + \sum_i \log p_{z_i}\}} = E_{\mathbf{Z}}\left[\sum_i 1_{\{Z_i=k\}}\right] \quad (7)$$

and the expectation is with respect to $\Pr(\mathbf{Z}|\phi, \mathbf{p})$. Similarly, the Hessian of $\hat{E}\left(\ell_2(\phi, \mathbf{p}) \mid y, \boldsymbol{\theta}^{(t)}\right)$ is equal to the negative of the covariance matrix of $(\sum_{i\sim j} 1_{\{Z_i=Z_j\}}, \sum_i 1_{\{Z_i=1\}}, \ldots, \sum_i 1_{\{Z_i=K\}})$ under $\Pr(\mathbf{Z}|\phi, \mathbf{p})$. We approximate the expectations and covariance matrix via (Swendsen-Wang) Markov chain Monte Carlo.

Each iteration t of the proposed MCEM method has computation time on the order $O(S_t(nKd^2 + Kd^{2.4}))$, where the term $Kd^{2.4}$ comes from inversion of the $d \times d$ covariance matrices $\boldsymbol{\Sigma}_k^{(t)}$. Most critically, this is linear in the number of voxels n. Both the number of MCEM iterations required for convergence, and the number of Swendsen-Wang iterations required to obtain accurate Monte Carlo estimates, depend on n in a way that is difficult to quantify. The means that the overall running time of the algorithm may grow more quickly than linear in n. However, for the experiments reported here, no more than 300 MCEM iterations were required for convergence to within a tolerance of 5/1000 times the estimated parameter value. In each MCEM iteration we use 8 parallel Swendsen-Wang chains, each simulated for $S_t = 63$ iterations after a burn-in period of 20 iterations (yielding a total of 504 Monte Carlo samples). While this is a low number of Markov chain iterations, there is a strong averaging effect in the Monte Carlo estimates (5), due to the large number of voxels. Because of this, the Monte Carlo standard error estimated using batch means (Flegal et al., 2008) is nearly always less than 2/1000 of the Monte Carlo estimate. We experimented with increasing the number of Swendsen-Wang iterations by a factor of 10 in some simulation and DCE-MRI experiments, and found no substantive change in the final results. With the current settings, our model takes about 6 hours to fit on the simulated data, and 6-48 hours to fit on the full DCE-MRI dataset (depending on the value of K), on a dual quad-core processor with 2.0 GHz speed and 4 GB of memory. Running times are only slightly longer than the method of Zhang et al. (2008); for instance, with $K = 3$ on the DCE-MRI data, our method takes 13.9 hours and theirs takes 11.5 hours.

Estimation of Z. Our estimate of $\mathbf{Z}$ is obtained by maximizing the posterior probability $\Pr(\mathbf{Z}|\mathbf{y}, \hat{\boldsymbol{\theta}})$ using simulated annealing (Bertsimas and Tsitsiklis, 1993). We use $T = 200$ iterations and temperature schedule $\frac{.7}{\ln(t+1)}$ where t is the simulated annealing iteration. This computation requires about 3 minutes on a single processor.

3.1 Approximating the Observed-Data Likelihood

Here we obtain an approximation to the observed-data likelihood $L_{obs}(\boldsymbol{\theta}|\mathbf{y})$. Since

$$L_{obs}(\boldsymbol{\theta}|\mathbf{y}) = \sum_{\mathbf{z}} \Pr(\mathbf{Z} = \mathbf{z}|\phi, \mathbf{p}) \prod_{i=1}^{n} N_d(\mathbf{y}_i; \boldsymbol{\mu}_{z_i}, \boldsymbol{\Sigma}_{z_i}) = g^{-1}(\phi, \mathbf{p}) h(\mathbf{y}, \boldsymbol{\theta}) \qquad \text{where}$$

$$h(\mathbf{y}, \boldsymbol{\theta}) \equiv \sum_{\mathbf{z}} \exp\left\{ \phi \sum_{i \sim j} 1_{\{z_i = z_j\}} + \sum_{i} \left[\log p_{z_i} + \log N_d(\mathbf{y}_i; \boldsymbol{\mu}_{z_i}, \boldsymbol{\Sigma}_{z_i}) \right] \right\},$$

it is sufficient to approximate $g(\phi, \mathbf{p})$ and $h(\mathbf{y}, \boldsymbol{\theta})$, which we do by thermodynamic integration (Ogata, 1989; Gelman and Meng, 1998). For any $\mathbf{p}$ we have $g(0, \mathbf{p}) = \sum_{\mathbf{z}} \prod_{i=1}^{n} p_{z_i} = 1$, so

$$\log g(\phi, \mathbf{p}) = \log \frac{g(\phi, \mathbf{p})}{g(0, \mathbf{p})} = \int_0^{\phi} \frac{\partial \log g(\tilde{\phi}, \mathbf{p})}{\partial \tilde{\phi}} \partial \tilde{\phi} \tag{8}$$

where $\frac{\partial \log g(\phi, \mathbf{p})}{\partial \phi} = E_{\mathbf{Z}} \left[\sum_{i \sim j} 1_{\{Z_i = Z_j\}} \right]$ as shown in (7). We first approximate $E_{\mathbf{Z}} \left[\sum_{i \sim j} 1_{\{Z_i = Z_j\}} \right]$ by Markov chain Monte Carlo for each $\tilde{\phi}$ on a grid of values between 0 and ϕ. Then we combine these values to obtain an approximation to (8) using the trapezoidal rule.

To approximate $h(\mathbf{y}, \boldsymbol{\theta})$, note that $\log h(\mathbf{y}, 0, \mathbf{p}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \log \left[\prod_{i=1}^{n} \sum_{k=1}^{K} p_k N_d(\mathbf{y}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right]$ is simple to compute. Also, analogously to (7), $\frac{\partial \log h(\mathbf{y}, \boldsymbol{\theta})}{\partial \phi} = E_{\mathbf{Z}|\mathbf{y}} \left[\sum_{i \sim j} 1_{\{Z_i = Z_j\}} \right]$ where the expectation is with respect to $\Pr(\mathbf{Z}|\mathbf{y}, \boldsymbol{\theta})$. Computation proceeds as for $g(\phi, \mathbf{p})$. The quantities $g(\phi, \mathbf{p})$ and $h(\mathbf{y}, \boldsymbol{\theta})$, and thus $L_{obs}(\boldsymbol{\theta}|\mathbf{y})$, can be approximated to arbitrary accuracy by increasing both the number of Monte Carlo samples and the resolution of the $\tilde{\phi}$ grid.

4 Simulation Study

We simulate data on a three-dimensional grid $\{1, \ldots, 50\}^3$, which contains 125,000 voxels (grid points). This is about one-third as many voxels as in the DCE-MRI data, chosen for computational ease in repeated experiments. We define two large true clusters (having

many voxels) and one to two small true clusters. The small clusters contain roughly the same percentage of the voxels as does the smallest estimated cluster for the DCE-MRI data. The first small cluster (cluster 1) contains the voxels within the contiguous central $7 \times 7 \times 7$ subgrid $\{22, \ldots, 28\}^3$, i.e. 343 voxels. For the experiments having a second small cluster (cluster 4), it corresponds to the subgrid $(\{1, \ldots, 7\} \cup \{44, \ldots, 50\}) \times \{1, \ldots, 5\}^2$, i.e. two $7 \times 5 \times 5$ regions in corners of the grid with a total of 350 voxels. One of the large clusters (cluster 3) includes the voxels $(\{1, \ldots, 17\} \cup \{34, \ldots, 50\}) \times \{9, \ldots, 50\}^2$, i.e. 59976 voxels in two $17 \times 42 \times 42$ regions adjacent to the grid boundaries that do not overlap with the small clusters. The second large cluster (cluster 2) contains all voxels that do not belong to other clusters, and is a contiguous region.

For each voxel i, Y_i is sampled from the normal distribution associated with its true cluster indicator Z_i, as in (2). The dimension is $d = 1$, and the variance parameters for all the clusters are taken to be $\sigma_k^2 = 4$. The mean parameters for the two large clusters are taken to be $\mu_2 = 0$ and $\mu_3 = -3$. The simulation scenarios are:

1. One small cluster, for a total of $K = 3$ clusters with sizes 343, 64681, and 59976, respectively. We take $\mu_1 = 7$.

2. Two small clusters, for a total of $K = 4$ clusters with sizes 343, 64331, 59976, and 350, respectively. We take $\mu_1 = 7$ and $\mu_4 = -10$.

We apply our method in two ways: first, with unrestricted covariance matrices $\mathbf{\Sigma}_k$; second, assuming a common cluster volume $\lambda_k = \lambda$ under the decomposition (4), as done in our DCE-MRI analysis (see Section 5). In the simulation study $d = 1$, so this corresponds to assuming a common variance σ_k^2 for all clusters. We also compare our method to three other approaches: 1. the method of Zhang et al. (2008) (see Section 3); 2. a non-spatial Gaussian mixture model with unrestricted variances, in the form (1); and 3. the same non-spatial model, applied after smoothing the observations y_i using a moving average. Such pre-smoothing is common in other types of image analysis (Sweeney et al., 2012, 2013) and has been used with image segmentation (Cai et al., 2010). However, it is not clear how to choose the amount of smoothing; with low levels of smoothing the resulting segmentations are noisy, and with more smoothing the small clusters (1 and 4) are not detected at all. For

	Misclass. cluster 1	Misclass. cluster 3	cluster size 1	cluster size 3	$\hat{\mu}_1$	$\hat{\mu}_3$	$\hat{\sigma}_1^2$	$\hat{\sigma}_3^2$	$\hat{\phi}$
Scenario 1									
True value	-	-	343	59976	7	-3	4	4	-
Non-spatial	177 (12)	28916 (458)	238 (18)	57756 (4100)	7.2 (.47)	-3.0 (.08)	3.8 (1.0)	3.9 (.09)	-
Smoothed nonsp.	13.8 (4.0)	11192 (178)	338 (3.7)	59811 (291)	6.5 (.07)	-2.9 (.01)	1.3 (.12)	1.2 (.01)	-
Zhang et al.	386 (30)	2587 (112)	727 (29)	59928 (55)	.54 (.03)	-3.1 (.01)	11 (.30)	3.6 (.03)	.67 (.00)
Our method	86.2 (11)	2880 (105)	403 (9.9)	60117 (78)	6.3 (.15)	-3.1 (.01)	4.5 (.38)	3.6 (.03)	.59 (.00)
Ours, $\lambda_k = \lambda$	84.6 (13)	2876 (106)	394 (10)	60063 (67)	6.5 (.13)	-3.1 (.01)	3.6 (.02)	3.6 (.02)	.59 (.00)
Scenario 2									
True value	-	-	343	59976	7	-3	4	4	-
Non-spatial	173 (9.6)	28935 (208)	238 (27)	59105 (2572)	7.0 (.55)	-3.0 (.06)	4.2 (1.1)	4.0 (.09)	-
Smoothed nonsp.	10.9 (3.2)	11199 (97)	339 (2.9)	59867 (559)	6.5 (.09)	-2.9 (.02)	1.4 (.16)	1.2 (.02)	-
Zhang et al.	332 (14)	2002 (58)	674 (14)	59667 (78)	1.3 (.05)	-3.1 (.01)	13 (.34)	3.6 (.02)	.75 (.00)
Our method	80.0 (17)	2652 (61)	401 (16)	59971 (57)	6.2 (.13)	-3.1 (.01)	4.8 (.38)	3.6 (.02)	.61 (.00)
Ours, $\lambda_k = \lambda$	84.1 (13)	2664 (72)	399 (15)	59965 (56)	6.5 (.12)	-3.1 (.04)	3.6 (.01)	3.6 (.01)	.61 (.00)

Table 1: Results from the simulation study, for five methods and two simulation scenarios. The mean over 10 simulations is given, with standard deviations in parentheses; results for clusters 2 & 4 (omitted) are similar to 1 & 3.

illustration we give results when the smoothed value for each voxel i is taken to be to be $y_i^* = \frac{1}{2}y_i + \frac{\sum_{j:i\sim j} y_j}{2\sum_{j:i\sim j} 1}$, which is a relatively low level of smoothing.

For all methods we first use EM or MCEM to obtain approximate MLEs of the parameters, then fix those parameter values and use simulated annealing to obtain an approximate maximum a posteriori estimate of $\mathbf{Z}$. The estimated clusters are then re-labeled in order of decreasing values of the estimated mean parameter $\hat{\mu}_k$, so that the estimated clusters correspond to the true clusters. To ensure a fair comparison, we initialize each method at two distinct locations, analogous to those used for our method (Section 3), and use the parameter estimates that have the highest observed-data likelihood.

We simulated 10 datasets for each scenario; the results are shown in Table 1. The columns give the number of misclassifications associated with each of clusters 1 and 3 (the number of voxels incorrectly assigned to that cluster, plus the number of voxels that belong to that cluster but are assigned to other clusters), the estimated number of voxels in each of clusters 1 and 3, and some of the parameter estimates. Results for clusters 2 and 4 are qualitatively similar to those of clusters 3 and 1, respectively.

Scenario 1 has similar results to Scenario 2. The non-spatial method underestimates the size of cluster 1 by more than 30% in both scenarios, and underestimates the size of cluster 3 by 3.7% in Scenario 1 and 1.5% in Scenario 2. The smoothed non-spatial method does considerably better in this respect, estimating the sizes of the clusters accurately. The

smoothed method also has 60% fewer misclassifications associated with cluster 3 and 90% fewer misclassifications associated with cluster 1 than the non-spatial method. However, the number of cluster 3 misclassifications is still large, with about 9% of all voxels having this particular error. Both the non-spatial and smoothed non-spatial methods estimate the μ_k parameters relatively accurately (although there is necessarily more estimation error associated with μ_1 than μ_3 since cluster 1 has fewer observations). The non-spatial method estimates the variance parameters σ_k^2 accurately but the smoothed non-spatial method dramatically underestimates them, due to the averaging effect of the smoothing.

The method of Zhang et al. (2008) has fewer than one-quarter as many cluster 3 misclassifications than either of the non-spatial methods. It does as well on average as the smoothed non-spatial method in estimating the size of cluster 3, and has considerably less variability in this estimate between simulations. It estimates μ_3 and σ_3^2 relatively accurately. However, it dramatically underestimates μ_1, and dramatically overestimates σ_1^2. The estimated size of cluster 1 is roughly twice the true size. Because of this, the number of misclassifications associated with cluster 1 is roughly twice as high as even the non-spatial method. This inaccuracy associated with small clusters is due to the fact that this method does not have cluster weight parameters p_k, so it assigns too many voxels to the small cluster, and thus inaccurately estimates the parameters of that cluster.

The two versions of our method perform similarly to each other. Like Zhang et al. (2008) we estimate μ_3, σ_3^2, and the size of cluster 3 accurately, and have far fewer cluster 3 misclassifications than the non-spatial methods. However, our methods estimate cluster 1 more accurately. They have fewer than one-quarter as many cluster 1 misclassifications as Zhang et al. (2008), and fewer than half as many as the non-spatial method. Our methods estimate the size of cluster 1, and the parameters μ_1 and σ_1^2, accurately and with low variability.

The two versions of our method do have 11-33% more cluster 3 misclassifications than the method of Zhang et al. (2008). This is due to the fact that their method estimates the spatial dependence parameter ϕ to be higher for these data, and so does more smoothing within the two large clusters. This occurs because most of the voxels in the simulated data come from two clusters, so the data suggest a higher value of $\sum_{i\sim j} 1_{\{Z_i=Z_j\}}$ than would be the case if all four clusters were similar sizes. Our method accounts for this effect directly,

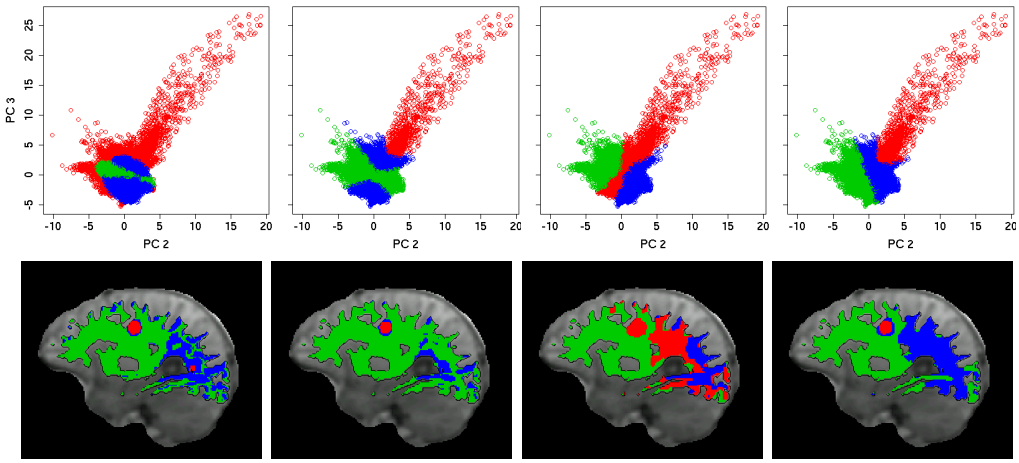


Figure 2: Top row: principal component scores, colored by the cluster assignment, for four methods with $K = 3$ clusters. Bottom row: a two-dimensional slice of the corresponding brain segmentations. The methods are (from left to right) a non-spatial mixture model without and with $\lambda_k = \lambda$; the method of Zhang et al. (2008); and our method.

by estimating different values for the p_k parameters. The method of Zhang et al. (2008) does not have the parameters p_k, so it estimates the spatial dependence ϕ to be higher to account for the high value of $\sum_{i \sim j} 1_{\{Z_i = Z_j\}}$.

5 Application to DCE-MRI

We apply our clustering method to segment the entire white matter volume of a subject with MS, based on a DCE-MRI study consisting of images obtained at 60 time points over 155 minutes, starting two minutes before the contrast agent was injected. These data are also analyzed (as Subject 1) in Shinohara et al. (2011), where a full description can be found. The white matter region for this subject consists of 384,185 voxels.

Figure 3 shows the time series of MR intensities for some voxels that are or are not inside enhancing lesions; since the enhancing lesions are not defined a priori, for this illustration we use a rough definition based on the shape of the intensity time series. The intensities are unnormalized since we analyze a single DCE-MRI study. The sampled time points are unequally spaced, and there appears to be considerable measurement error.

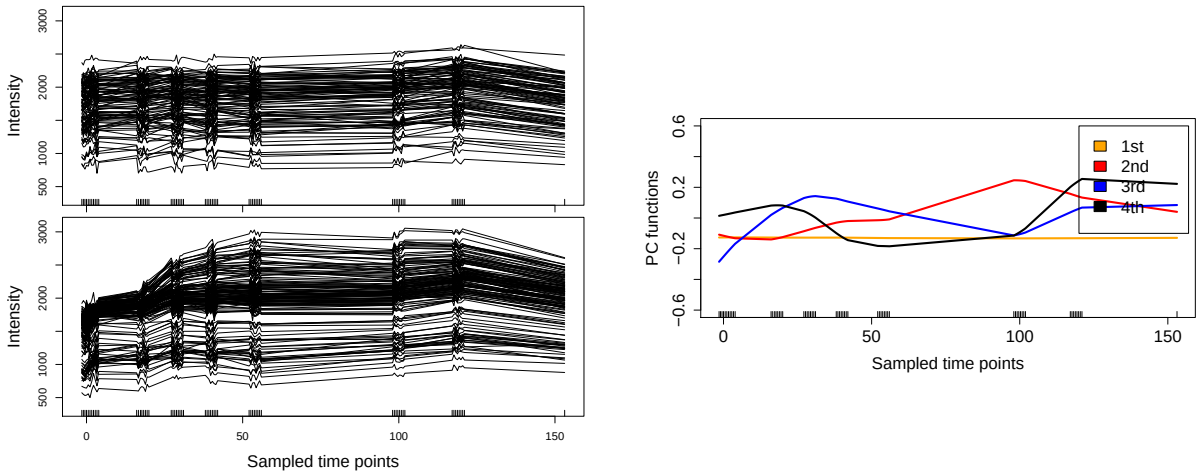


Figure 3: Top left: Time series of MR intensities for .075% of the voxels not inside enhancing lesions. Bottom left: Intensity time series for 10% of the voxels inside enhancing lesions. Right: Regularized principal component functions for all voxel intensities.

Dimension Reduction. We first reduce the time series associated with each voxel to a low-dimensional summary using regularized functional principal components analysis; the resulting PC functions are shown in Figure 3. The first PC function is roughly constant in time, so represents the baseline contrast intensity of the voxels. The second PC represents a gradual increase in contrast intensity up to time 100, followed by a gradual decrease. The third captures a rapid increase in contrast intensity in the first thirty minutes. In the context of MS we are interested in the change in contrast intensity over time, so we omit the first PC from further analysis. The second and third PCs capture 73.3% of the remaining variability, and the second through fourth capture 87.6%. We use the second and third PC scores, since this case is easier to visualize.

Results. Scatterplots of the PC score vectors $\mathbf{y}_i$ are shown in Figure 2. The most prominent feature of the scatterplot (initially ignoring the colors) is a long tail. Voxels in this tail correspond to the enhancing lesions (Shinohara et al., 2011). There also appears to be some structure in the main cloud of points. Fitting a non-spatial (independent) Gaussian mixture model with, for instance, $K = 3$ mixture components and unrestricted covariance matrices results in the cluster assignments shown in the top-left plot of Figure 2.

One cluster captures all of the outlying points, regardless of direction, and the two other

clusters capture features in the central cloud of points. This model has failed to isolate the most obvious feature of the data, namely the tail, from the rest of the data. So the enhancing lesions, which have high values of both PC scores, have been grouped together with white matter that has very different characteristics, including voxels with low values of both scores. Additionally, it is difficult to separately interpret the blue and green clusters, because their estimated mean vectors $\hat{\boldsymbol{\mu}}_k$ are very close. A two-dimensional slice of the three-dimensional white matter segmentation is given in the bottom-left image of Figure 2. An area known to be an enhancing lesion (the circular red area) has been given the same cluster assignment as an area that is known to not be an enhancing lesion (the other red area).

One way to create a more meaningful segmentation is to put a restriction on the covariance matrices $\boldsymbol{\Sigma}_k$ of the mixture components. We restrict all the $\boldsymbol{\Sigma}_k$ to have the same determinant, i.e. restrict $\lambda_k = \lambda$ in (4). Recalling that λ_k is interpreted as the volume of the kth cluster in the space of PC scores, this corresponds to restricting the volumes of the clusters to be the same. This rules out a configuration like the top-left one in Figure 2, where one mixture component has much higher variance in all directions than the other mixture components. We do not place any other restrictions on the covariance matrices, so the orientations $\mathbf{D}_k$ and the shapes $\mathbf{A}_k$ of the clusters are allowed to vary. All the remaining results reported for the DCE-MRI data will use the restriction $\lambda_k = \lambda$, since the results are poor without it.

The results from the non-spatial mixture model with this restriction are shown in the second column of Figure 2. The "tail" in the scatterplot is now identified as a distinct group. The central region of the scatterplot is cut into two clusters, but one of the clusters is noncontiguous, making interpretation more difficult. Also, the segmentation slice in Figure 2 appears somewhat noisy.

The results from the spatial mixture model of Zhang et al. (2008) are shown in the third column of Figure 2. The model is unable to distinguish the enhancing lesions (tail in the scatterplot) as a separate group, because the enhancing lesions contain a tiny proportion of the voxels (roughly 0.3%).

The results of our method with $K = 3$ are given in the last column of Figure 2. The scatterplot shows that each cluster is contiguous in the space of PC scores, and that the

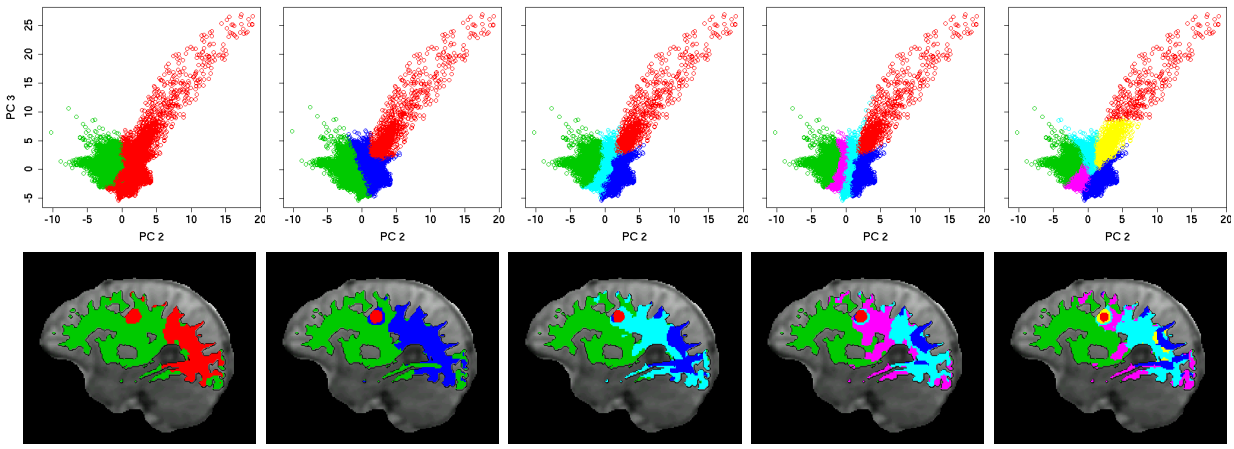


Figure 4: PC score scatterplots and brain image slices, colored by the cluster assignment from our method, for $K = 2$ (leftmost column) through $K = 6$ (rightmost column).

enhancing lesions have been correctly distinguished from the rest of the white matter. Additionally, the brain image shows spatially smooth cluster assignments. The spatial model has effectively de-noised the image, creating contiguous blocks where the tissue is estimated to be of the same type.

In Figure 4 we illustrate the results of our method with $K = 2$ through $K = 6$ clusters. Two clusters ($K = 2$) are clearly insufficient to distinguish critical features. Three and four clusters accurately distinguish the enhancing lesions, and make reasonable distinctions between different parts of the remaining white matter. Larger numbers of clusters (such as $K = 5, 6$) are difficult to justify, because they lead to minor distinctions between the PC scores, as seen on the score scatterplots in Figure 4. Based on these observations, we focus on $K = 4$.

Time series of MR intensities for each cluster are shown in Figure 5, for our method with $K = 4$. The variability within each cluster is large relative to the mean differences between the clusters. Also, there is a jump in intensities for all clusters around time 120 minutes; this also exists in the raw data and appears to be an imaging artifact. Focusing then on times up to 100, we see that the average intensity for the cluster corresponding to the enhancing lesions (solid red line) starts low, increases quickly before time 50, then gradually increases until time 100. The average intensity of each of the other three clusters has fairly constant

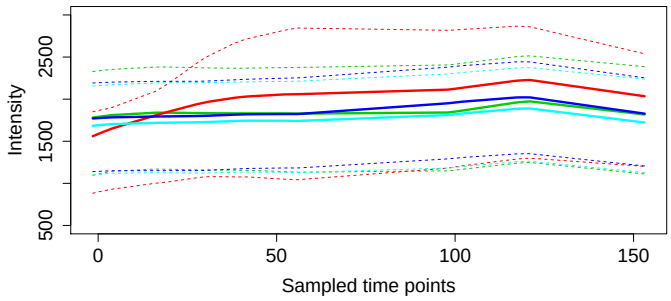


Figure 5: Time series of MR intensities for the clusters, for our method with $K = 4$. Colors correspond to those in the third column of Figure 4; solid lines show average cluster intensities, and dashed lines show pointwise 95% intervals.

slope up to time 100, and the three clusters differ mainly in the steepness of that slope. The dark blue cluster has the steepest slope, while the green cluster has the flattest slope of the three. While these three clusters exhibit little of the early enhancement shown by the red cluster, they have differing degrees of this slower enhancement.

As seen in the third column of Figure 4, the dark blue cluster is mainly located near the back of the brain, while the green cluster is close to the front and the light blue cluster is in between. The differences between the regions could either reflect spatial heterogeneity in the brain tissue, or heterogeneity that is an artifact of the imaging process.

For the DCE-MRI data we did not find any sensitivity of our method to its initialization. The non-spatial mixture model, however, often had different results according to the initialization (the ones reported here are those that had the highest observed-data likelihood). This difference between the two methods may be due to the spatial information incorporated by our model, which has the effect of ruling out parameter values that correspond to spatially nonsmooth segmentations.

6 Conclusions

We introduced a method for image segmentation based on spatial model-based clustering, that can accurately capture small-scale features in addition to large-scale ones. We also proposed a computational method that is efficient enough to handle even high-resolution

three-dimensional images.

We applied our method to DCE-MR imaging of multiple sclerosis subjects, who have brain lesions. The image segmentations provided by a competing non-spatial method accurately distinguish the enhancing lesions from other white matter, but the other identified regions have poor interpretability and the segmentations appear noisy. The image segmentations from the method of Zhang et al. (2008) group the enhancing lesions together with a large portion of the non-enhancing white matter, despite the distinctive signal associated with enhancing lesions. By contrast, the segmentations from our method accurately distinguish the enhancing lesions, and identify other regions that have reasonable interpretation.

A population-level analysis of MS subjects, based on the same principles, might provide additional clinical insights regarding MS. This could be done by applying our clustering methods to multiple subjects simultaneously, fitting our model to the voxels in all subjects after appropriate normalization. Then the resulting clusters would be interpreted as population-level tissue types instead of subject-specific tissue types. Alternatively, the model-based approach that we have taken would allow natural extension to multiple subjects, by incorporation of random effects.

ACKNOWLEDGEMENTS. The authors thank Stuart Geman for his helpful suggestions. This project was supported by grants CMMI-0926814 and DMS-1209103 from the U.S. National Science Foundation, R01EB012547 from the National Institute of Biomedical Imaging and Bioengineering, and R01NS060910 from the National Institute of Neurological Disorders and Stroke (NINDS). This work represents the opinions of the researchers and not necessarily that of the granting organizations. The authors also thank Daniel Reich and the members of the Translational Neuroradiology Unit at NINDS/NIH who collected these data.

References

Alfó, M., Nieddu, L., and Vicari, D. (2008). A finite mixture model for image segmentation. *Statistics and Computing*, 18:137–150.

Banfield, J. D. and Raftery, A. E. (1993). Model-based Gaussian and non-Gaussian cluster-

ing. *Biometrics*, 49:803–821.

Baudry, J.-P., Raftery, A. E., Celeux, G., Lo, K., and Gottardo, R. (2010). Combining mixture components for clustering. *J. of Computational and Graphical Statistics*, 19:332–353.

Bertsimas, D. and Tsitsiklis, J. (1993). Simulated annealing. *Statistical Science*, 8:10–15.

Besag, J. E. (1986). On the statistical analysis of dirty pictures. *J. of the Royal Statistical Society, Series B*, 48:259–302.

Biernacki, C., Celeux, G., and Govaert, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22:719–725.

Bosch, M., Zhu, F., and Delp, E. J. (2011). Segmentation-based video compression using texture and motion models. *IEEE Journal of Selected Topics in Signal Processing*, 5:1366–1377.

Cai, W., Lei, L., and Yang, M. (2010). A Gaussian mixture model-based clustering algorithm for image segmentation using dependable spatial constraints. In Sun, L., editor, *Proc. of the 3rd International Congress on Image and Signal Processing*, pages 1268–1272. IEEE.

Celeux, G., Forbes, F., and Peyrard, N. (2003). EM procedures using mean field-like approximations for Markov model-based image segmentation. *Pattern Recognition*, 36:131–144.

Celeux, G., Forbes, F., and Peyrard, N. (2004). EM-based image segmentation using Potts models with external field. In *Proceedings of Reconnaissance des Formes et Intelligence Artificielle*. Available at `http://mistis.inrialpes.fr/~forbes/publications.html`.

Celeux, G. and Govaert, G. (1995). Gaussian parsimonious clustering models. *Pattern Recognition*, 28:781–793.

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*, 10:395–407.

Cucala, L. and Marin, J.-M. (2012). Bayesian inference on a mixture model with spatial dependence. Submitted; available at `http://www.math.univ-montp2.fr/~cucala/menteith.pdf`.

Deng, H. and Clausi, D. A. (2004). Unsupervised image segmentation using a simple MRF model with a new implementation scheme. *Pattern Recognition*, 37:2323–2335.

Flegal, J. M., Haran, M., and Jones, G. L. (2008). Markov chain Monte Carlo: can we trust the third significant figure? *Statistical Science*, 23:250–260.

Forbes, F. and Fort, G. (2007). Combining Monte Carlo and mean-field-like methods for inference in hidden Markov random fields. *IEEE Transactions on Image Processing*, 16:824–837.

Forbes, F., Peyrard, N., Fraley, C., Georgian-Smith, D., Goldhaber, D. M., and Raftery, A. E. (2006). Model-based region-of-interest selection in dynamic breast MRI. *J. of Computer Assisted Tomography*, 30:675–687.

Fraley, C. and Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *J. of the American Statistical Association*, 97:611–631.

Friel, N., Pettitt, A. N., Reeves, R., and Wit, E. (2009). Bayesian inference in hidden Markov random fields for binary data defined on large lattices. *J. of Computational and Graphical Statistics*, 18:243–261.

Gelman, A. and Meng, X.-L. (1998). Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statistical Science*, 13:163–185.

Geyer, C. J. and Thompson, E. A. (1992). Constrained Monte Carlo maximum likelihood for dependent data. *J. of the Royal Statistical Society, Series B*, 54:657–699.

Grossman, R. I., Braffman, B. H., Brorson, J. R., Goldberg, H. I., Silberberg, D. H., and Gonzalez-Scarano, F. (1988). Multiple sclerosis: serial study of gadolinium-enhanced MR imaging. *Radiology*, 169:117–122.

Johnson, T. D. and Piert, M. (2009). A Bayesian analysis of dual autoradiographic images. *Computational Statistics and Data Analysis*, 53:4570–4583.

McGrory, C. A., Titterington, D. M., Reeves, R., and Pettitt, A. N. (2009). Variational Bayes for estimating the parameters of a hidden Potts model. *Statistics and Computing*, 19:329–340.

Ogata, Y. (1989). A Monte Carlo method for high dimensional integration. *Numerische Mathematik*, 55:137–157.

Panjwani, D. K. and Healey, G. (1995). Markov random field models for unsupervised segmentation of textured color images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17:939–954.

Peel, D. and McLachlan, G. J. (2000). Robust mixture modelling using the t distribution. *Statistics and Computing*, 10:339–348.

Pham, D. L., Xu, C., and Prince, J. L. (2000). Current methods in medical image segmentation. *Annual Review of Biomedical Engineering*, 2:315–337.

Polman, C. H., Reingold, S. C., Banwell, B., et al. (2011). Diagnostic criteria for multiple sclerosis: 2010 revisions to the McDonald criteria. *Annals of Neurology*, 69:292–302.

Robinson, L. F., Wager, T. D., and Lindquist, M. A. (2010). Change point estimation in multi-subject fMRI studies. *NeuroImage*, 49:1581–1592.

Shinohara, R. T., Crainiceanu, C. M., Caffo, B. S., Gaitán, M. I., and Reich, D. S. (2011). Population-wide principal component-based quantification of blood-brain-barrier dynamics in multiple sclerosis. *NeuroImage*, 57:1430–1446.

Shinohara, R. T., Goldsmith, J., Mateen, F. J., Crainiceanu, C., and Reich, D. S. (2012). Predicting breakdown of the blood-brain barrier in multiple sclerosis without contrast agents. *American J. of Neuroradiology*, 33:1586–1590.

Sweeney, E. M., Shinohara, R. T., Shea, C. D., Reich, D. S., and Crainiceanu, C. M. (2012). Automatic lesion incidence estimation and detection in multiple sclerosis using multisequence longitudinal MRI. *American J. of Neuroradiology*, 34:68–73.

Sweeney, E. M., Shinohara, R. T., Shiee, N., Mateen, F. J., Chudgar, A. A., Cuzzocreo, J. L., Calabresi, P. A., Pham, D. L., Reich, D. S., and Crainiceanu, C. M. (2013). OASIS is automated statistical inference for segmentation, with applications to multiple sclerosis lesion segmentation in MRI. *NeuroImage: Clinical*, 2:402–413.

Swendsen, R. H. and Wang, J. S. (1987). Nonuniversal critical dynamics in Monte Carlo simulations. *Physical Review Letters*, 58:86–88.

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*, 22:105–119.

Wehrens, R., Buydens, L. M. C., Fraley, C., and Raftery, A. E. (2004). Model-based clustering for image segmentation and large datasets via sampling. *J. of Classification*, 21:231–253.

Woodard, D. B. and Goldszmidt, M. (2011). Online model-based clustering for crisis identification in distributed computing. *J. of the American Statistical Association*, 106:49–60.

Zhang, X., Johnson, T. D., Little, R. J. A., and Cao, Y. (2008). Quantitative Magnetic Resonance Image analysis via the EM algorithm with stochastic variation. *Annals of Applied Statistics*, 2:736–755.

Zhang, X., Johnson, T. D., Little, R. J. A., and Cao, Y. (2010a). A Bayesian image analysis of radiation induced changes in tumor vascular permeability. *Bayesian Analysis*, 5:189–212.

Zhang, X., Johnson, T. D., Little, R. J. A., and Cao, Y. (2010b). Longitudinal image analysis of tumor/healthy brain change in contrast uptake induced by radiation. *J. of the Royal Statistical Society, Series C*, 59:821–838.