CHALMERS |  UNIVERSITY OF GOTHENBURG

PREPRINT 2014:1

Multivariate latent Gaussian random field mixture models

DAVID BOLIN
JONAS WALLIN
FINN LINDGREN

Department of Mathematical Sciences
Division of Mathematical Statistics
CHALMERS UNIVERSITY OF TECHNOLOGY
UNIVERSITY OF GOTHENBURG
Gothenburg Sweden 2014

Preprint 2014:1

Multivariate latent Gaussian random field mixture models

David Bolin, Jonas Wallin and Finn Lindgren

Department of Mathematical Sciences
Division of Mathematical Statistics
Chalmers University of Technology and University of Gothenburg
SE-412 96 Gothenburg, Sweden
Gothenburg, January 2014

Preprint 2014:1
ISSN 1652-9715

Matematiska vetenskaper
Göteborg 2014

Multivariate latent Gaussian random field mixture models

David Bolin[1], Jonas Wallin[2] and Finn Lindgren[3]

[1]*Mathematical Sciences, Chalmers University of Technology, Gothenburg, Sweden*
[2]*Centre for Mathematical Sciences, Lund University, Lund, Sweden*
[3]*Mathematical Sciences, University of Bath, Bath, United Kingdom*

Abstract: A novel class of models is introduced, with potential areas of application ranging from land-use classification to brain imaging and geostatistics. The model class, denoted latent Gaussian random filed mixture models (LGFM models), combines the Markov random field mixture model with latent Gaussian random field models. The latent model, which is observed under measurement noise, is defined as a mixture of several, possible multivariate, Gaussian random fields. Which of the fields that is observed at each location is modeled using a discrete Markov random field. In order to use the method for massive massive data sets that arises in many possible areas of application, such as brain imaging, a computationally efficient parameter estimation method is required. Such an estimation method, based on a stochastic gradient algorithm, is developed and the model is tested on a magnetic resonance imaging application.

Key words: Gaussian mixture; Markov random fields; Random fields; Stochastic gradients

1 Introduction

Gaussian mixture models (GMMs) have successfully been used for classification in several areas of application ranging from video surveillance [Stauffer and Grimson, 1999] to speaker identification [Reynolds and Rose, 1995]. Also in geostatistics and statistical image analysis, classification and image segmentation is often performed using GMMs in combination with the Expectation Maximization (EM) algorithm [Dempster et al., 1977] for estimation. Let $\mathbf{Y}_1, \ldots, \mathbf{Y}_m$ be observations of some, possibly multivariate, process $\mathbf{Y}(\mathbf{s})$ at locations $\mathbf{s}_1, \ldots, \mathbf{s}_m$. The classical GMM can then be formulated as

$$\pi(\mathbf{Y}_i|\boldsymbol{\theta}) = \sum_{k=1}^{K} w_{ik}\pi_k(\mathbf{Y}_i|\boldsymbol{\theta}_k), \tag{1}$$

independently for all $i = 1, \ldots m$, where K is the number of classes, w_{ik} denotes the prior probability of $\mathbf{Y}_i$ belonging to class k, and $\pi_k(\mathbf{Y}_i|\boldsymbol{\theta}_k)$ denotes the distribution of class k, which is assumed to be Gaussian, $\mathsf{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$.

A drawback with classification based on the classical GMM is that any spatial dependency of the data is ignored. A common strategy to account for spatial dependency in the data is allow for dependency in the allocation variables (w_{ik}), which can be done in several ways. One way is to model the class probabilities, w_{ik}, using a logistic normal model

$$w_{ik} = \frac{\exp(\eta_{ik})}{\sum_j \exp \eta_{ij}}, \tag{2}$$

where $\boldsymbol{\eta}_k$ are assumed to be latent Gaussian fields [Fernández and Green, 2002]. Estimation under this model is difficult, and one generally has to resort to computationally expensive MCMC methods. Furthermore, for classification problems, the model is not ideal as the spatial model forces the posterior weights to be smoothly varying, which often can reduce the predictive power of the model.

Another way to allow for dependency in the mixture weights is to note that in the random variable $\mathbf{Y}_i$ defined in (1) equals, in distribution,

$$\sum_{k=1}^{K} z_{ik}\mathbf{G}_{ik}, \tag{3}$$

where $\mathbf{G}_{ik} \sim \mathsf{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ and $z_{ik} = 1(x_i = k)$ is an indicator function for the event $x_i = k$, where x_i is a multinomial distributed r.v. defined though the probabilities $\mathsf{P}(x_i = k) = w_{ik}$. Using this formulation of the GMM, spatial dependency can be introduced by assuming that $\mathbf{x} = \{x_i\}$ is a discrete MRF [see e.g. Held et al., 1997, Zhang et al., 2001, Van Leemput et al., 1999]. We refer to this model as a MRF mixture model.

Allowing for spatial dependency in the mixture weights is often reasonable and improves the classification for spatial problems. However, from a modeling perspective the MRF mixture models are not ideal since the data within each class is assumed to be independent observations of the same Gaussian distribution, while one would also like to allow for spatial dependency of the data within each class. Consider, for example, land-use classification from satellite images, where the classes in the mixture are assumed to correspond to distinct land types such as forest, fields, water, etc. For a given class, say forest, the measured values will depend on, for example, vegetation density and vegetation composition which makes the assumption of independent measurements within the class unrealistic.

In geostatistics, the most common approach to model spatially dependent data is to use latent Gaussian random fields [see e.g. Cressie, 1991, Cressie and Wikle, 2011]. Collecting all measurements $\{\mathbf{Y}_i\}$ in a vector $\mathbf{Y}$, a latent Gaussian model can be written as

$$\mathbf{Y} = \mathbf{B}\boldsymbol{\beta} + \mathbf{A}\boldsymbol{\xi} + \boldsymbol{\varepsilon}, \tag{4}$$

where $\boldsymbol{\xi}$ is a (multivariate) mean-zero Gaussian random field, $\mathbf{A}$ is a matrix that connects the measurements to the latent field, and ε_i is Gaussian measurement noise. The matrix $\mathbf{B}$ contains covariates for the mean evaluated at the measurement locations, and the latent field evaluated at the measurement locations is given by $\mathbf{X} = \mathbf{B}\boldsymbol{\beta} + \mathbf{A}\boldsymbol{\xi}$. This modeling approach is often preferable if the latent process is smoothly varying and it is highly useful for noise reduction and spatial interpolation in cases of partial observations [Stein, 1999]. However, the latent Gaussian random fields are poorly equipped to deal with the discontinuity of both process and covariance common for data in classification problems.

The aim of this work is twofold. First, we want to provide a new class of models that extends the MRF mixture models and can be used for spatial modeling of data that is usually studied in spatial classification problems. The goal is to provide a model class that can be used for classification but also for noise reduction and spatial interpolation. The model class we propose, which we will refer to as the latent Gaussian random field mixture (LGFM) models, combines the MRF mixture models and the latent Gaussian models, by assuming that the latent field is a MRF mixture of Gaussian random fields. The possible application areas for this model class ranges from geostatistics and land-use classification problems to brain imaging and MRI modeling and estimation.

The second goal of this work is to provide an efficient estimation method for the LGFM and MRF mixture models that simplifies their usage for applications with massive datasets. The main computational bottle neck for likelihood-based estimation methods, for both the LGFM models and the MRF models, is the computation of the normalizing constants of the joint densities. For the MRF models there exists several ways to handle this issue, and the two most common methods are gradient-based estimation and pseudo-likelihood estimation [Guyon, 1995]. Recently, gradient methods for large-scale GRF models have been developed for likelihood estimation that efficiently deals with the normalizing constants [Anitescu et al., 2012, Stein et al., 2013]. We propose a stochastic version of the EM gradient method [Lange, 1995] based on pseudo-likelihoods, which handles the normalizing constant for both the LGFM and the MRF mixture models efficiently.

The structure of this work is as follows. In Section 2, the model class is introduced and connections to other related models are discussed. Section 3 contains an introduction to a particular choice of the model components which is suitable for modeling of large datasets. Section 4 introduces an estimation procedure that is suitable for this model class but also for the standard MRF mixture models and the latent Gaussian models in cases of large datasets. In Section 5, the model class is used for noise reduction in magnetic resonance (MR) imaging. Finally, Section 6 contains a discussion of possible extensions and further work.

2 Latent Gaussian random field mixture models

Let $\mathbf{Y}$ be the vector of, possibly multivariate, observations. The general structure of the LGFM models is then

$$\begin{aligned}\mathbf{Y}_i &= \mathbf{X}(\mathbf{s}_i) + \boldsymbol{\varepsilon},\\ \mathbf{X}(s) &= \sum_{k=1}^{K} z_k(\mathbf{s})\mathbf{X}_k(\mathbf{s}),\\ \mathbf{X}_k(s) &= \sum_{j=1}^{n} \mathbf{B}_{kj}(\mathbf{s})\beta_{kj} + \boldsymbol{\xi}_k(\mathbf{s}).\end{aligned} \tag{5}$$

Here, $\boldsymbol{\varepsilon}$ is mean-zero Gaussian measurement noise and $\mathbf{X}(\mathbf{s})$ is the latent process. The latent process is described as a mixture of K Gaussian random field models, $\mathbf{X}_1, \ldots, \mathbf{X}_K$, and $\mathbf{z}$ is an indicator field that determines which class that is present at each location. Each Gaussian component is modeled using some covariates $\mathbf{B}_{kj}$ for the mean and a mean-zero Gaussian random field $\boldsymbol{\xi}_k$ with some covariance structure, which may be different for the different classes.

This general class contains several interesting models, and some examples of realizations of univariate models with $K = 2$ are shown in Figure 1. In the examples, X_k are independent stationary Gaussian Matérn fields. The indicator field z is obtained as $z_1(s) = \mathbb{I}_{Z(s)>0}(s)$, $z_2(s) = \mathbb{I}_{Z(s)\leq 0}(s)$ where $Z(s)$ is a Gaussian Matérn field, i.e. $z_1(s) = 1$ and $z_2(s) = 0$ for all s where $Z(s) > 0$ and $z_1(s) = 0$ and $z_2(s) = 1$ otherwise. In Panels (a) and (b), $Z(s)$ is independent of X_k. Panel (a) shows an example where X_1 and X_2 have the same covariance function but different mean values and Panel (b) shows an example where X_1 and X_2 have the same mean values but different correlation ranges. One can also imagine that z depends on some of the latent fields. Panels (c) and (d) are the same as Panels (a) and (b) except that $Z = X_1$. Thus, X_1 is only observed if it is positive and otherwise X_2 is observed.

There is a connection to the popular linear coregionalization models (LCM) [Zhang, 2007] in geostatistics. In our notation, an LCM can be written as

$$Y(\mathbf{s}_i) = \boldsymbol{\mu}(\mathbf{s}_i) + \sum_{k=0}^{K} \xi_k(\mathbf{s}_i).$$

Thus, this model would be a special case of the LGFM models if we allowed $z_k(\mathbf{s}) = 1$ for all k and $\mathbf{s}$.

For spatial classification problems, the domain for $\mathbf{s}$ is often discrete, e.g. pixels in satellite images or voxels in MR images. In such situations, the model can be written more compactly as

$$\mathbf{Y} = \sum_{k=1}^{K} \mathbf{z}_k \cdot (\mathbf{B}_k\boldsymbol{\beta}_k + \mathbf{A}\boldsymbol{\xi}_k) + \boldsymbol{\varepsilon}, \tag{6}$$

where $\cdot$ denotes element-wise multiplication, $\mathbf{B}$ is a matrix containing the covariates evaluated at the measurement locations, and $\mathbf{A}$ is a measurement matrix that determines which components in $\boldsymbol{\xi}_k$ that are observed. The latent field evaluated at the measurement locations is now given by $\mathbf{X} = \sum_{k=1}^{K} \mathbf{z}_k \cdot (\mathbf{B}\boldsymbol{\beta}_k + \mathbf{A}\boldsymbol{\xi}_k)$, which is a spatially correlated mixture of Gaussian random fields. Thus, there is a clear connection between this model and the MRF mixture models; a MRF mixture model with spatially dependent components is obtained by choosing $\mathbf{z}$ as the indicator field of a discrete MRF.

For practical applications of the model one is typically interested in estimates of the latent field $\mathbf{X}$ given the data. For spatial prediction and noise reduction, $\mathsf{E}(\mathbf{X}|\mathbf{Y}, \boldsymbol{\Psi})$, where $\boldsymbol{\Psi}$ is an estimate of the model parameters, is used as a point-estimate of the latent field and $\mathsf{V}(\mathbf{X}|\mathbf{Y}, \boldsymbol{\Psi})$ is used as a

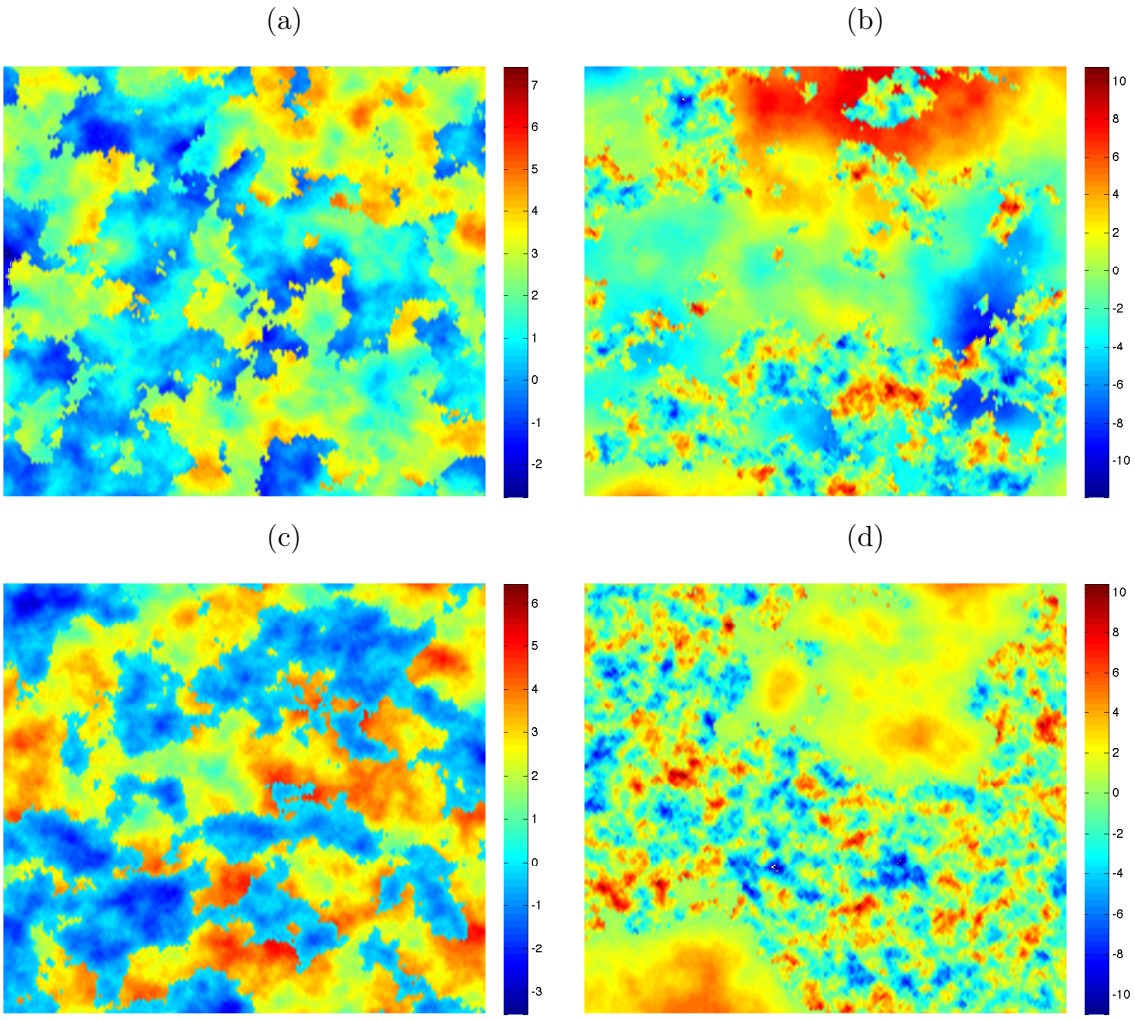


Figure 1: Examples of spatial mixture models with $K = 2$. The latent fields X_1 and X_2 are independent stationary Gaussian Matérn fields and z is obtained as $z_1(s) = Z(s) > 0$, $z_2(s) = Z(s) < 0$ where $Z(s)$ is a Gaussian Matérn field. In Panel (a), X_1 and X_2 have the same covariance function but different mean values and $X(s)$ is independent of X_k. In Panel (b), X_1 and X_2 have the same mean values but different correlation ranges and $X(s)$ is independent of X_k. Panels (c) and (d) are the same as Panels (a) and (b) respectively, except that $Z = X_1$.

measure of the uncertainty in that prediction. To calculate these, we note that

$$\begin{aligned}\mathsf{E}(\mathbf{X}|\mathbf{Y},\boldsymbol{\Psi}) &= \mathsf{E}[\mathsf{E}(\mathbf{X}|\mathbf{Y},\mathbf{z},\boldsymbol{\Psi})|\boldsymbol{\Psi},\mathbf{Y}],\\ \mathsf{V}(\mathbf{X}|\mathbf{Y},\boldsymbol{\Psi}) &= \mathsf{E}[\mathsf{V}(\mathbf{X}|\mathbf{Y},\mathbf{z},\boldsymbol{\Psi})|\boldsymbol{\Psi},\mathbf{Y}] + \mathsf{V}[\mathsf{E}(\mathbf{X}|\mathbf{Y},\mathbf{z},\boldsymbol{\Psi})|\boldsymbol{\Psi},\mathbf{Y}].\end{aligned}$$

Here, $\mathsf{E}(\mathbf{X}|\mathbf{Y},\mathbf{z},\boldsymbol{\Psi})$ and $\mathsf{V}(\mathbf{X}|\mathbf{Y},\mathbf{z},\boldsymbol{\Psi})$ can be calculated analytically since these are posterior means and covariances for Gaussian distributions. The outer expectation and variances, taken over $\mathbf{z}$, are typically not known analytically but can be estimated using Monte Carlo integration by sampling from $\pi(\mathbf{z}|\mathbf{Y},\boldsymbol{\Psi})$. While sampling $\mathbf{z}$, $\mathsf{E}(\mathbf{z}|\mathbf{Y},\boldsymbol{\Psi})$ can be estimated and used to classify the data.

Since the model class is mainly targeted at applications on discrete domains, we choose to study the discrete model in more detail and leave the practical details of the continuous models for further research. In the following section, we outline a reasonable choice for the different components in the model that makes the model applicable to large spatial problems. And in Section 4, an estimation procedure for this particular model is presented.

(a) (b)

Figure 2: A fist order neighborhood structure (a) and corresponding sets of conditionally independent pixels (b).

3 Model components

In this section, we present a particular choice for the model components in (6) which is suitable for modeling of massive multivariate spatial datasets. To increase the computational efficiency of the model, Markov properties are used both for the indicator process $\mathbf{s}$ and for the latent fields $\boldsymbol{\xi}_k$.

3.1 A discrete MRF model for z

A suitable model for the indicator field, $\mathbf{z}$, determining the class belongings for each pixel, is a discrete MRF. We let $\mathbf{x}$ be a discrete MRF taking values in $\{1, \ldots, K\}$ and define $z_{ik} = 1(x_i = k)$. The joint distribution of $\mathbf{x}$ can be formulated using the Gibbs distribution $p(\mathbf{x}) = Z^{-1} \exp(-W(\mathbf{x}))$ where $W(\mathbf{x}) = \sum_{\mathcal{C}} V_{\mathcal{C}}(\mathbf{x})$ is the sum of the potential for all cliques generated by the neighborhood structure and $Z = \sum_{\omega} \exp(-W(\omega))$.

There are many potential choices for the neighborhood structure, but we use a simple first-order neighborhood $\mathcal{N}_\star$, which on a regular lattice in $\mathbb{R}^2$ consists of the four closest nodes, in euclidean distance, and in $\mathbb{R}^3$ consists of the six closest nodes. In $\mathbb{R}^2$, this neighborhood structure is illustrated in Figure 2 (a) where $\bullet$ denotes the neighbors of the pixel $\star$. For this neighborhood structure, there are only first and second-order cliques, and we use the potentials $V_{\{u\}}(\mathbf{x}) = \alpha_k$ when $x_u = k$, and $V_{\{u,v\}}(\mathbf{x}) = \beta_k$ when $x_u = k$ and $x_v = k$.

Hence, the model has parameters $\boldsymbol{\alpha} = \{\alpha_k\}$ and $\boldsymbol{\beta} = \{\beta_k\}$ where $\boldsymbol{\alpha}$ determines the prior probabilities for each class k and $\boldsymbol{\beta}$ are interaction parameters that governs the strength of the spatial dependency. Since only the differences of the elements in $\boldsymbol{\alpha}$ affect the model, we fix α_1 to zero. Simplified models are obtained by either fixing all $\alpha_k = 0$ or by assuming that all β_k are equal to some common parameter β.

3.2 A Gaussian random field model for $\boldsymbol{\xi}$

We assume that $\boldsymbol{\xi}_k \sim \mathsf{N}(\mathbf{0}, \mathbf{Q}_k^{-1})$ is a multivariate spatial Gaussian random field with a covariance structure that is separable with respect to space the dimension of the data. This means that $\mathbf{Q}_k$ can be written as $\mathbf{Q}_k = \mathbf{Q}_{kd} \otimes \mathbf{Q}_{ks}$, where $\mathbf{Q}_{ks}$ is determined by a spatial covariance model and $\mathbf{Q}_{kd}$ is the multivariate part. The motivation behind this particular choice is that if there is no spatial dependence in the data, one can choose $\mathbf{Q}_{ks}$ as the identity matrix and the model reduces to a standard MRF mixture model. Since the precision matrix $\mathbf{Q}_{dk}$ corresponds to the covariance matrix $\boldsymbol{\Sigma}_k$ in the MRF mixture model, we do not assume any special structure of this matrix. It is therefore

parametrized as $\mathbf{Q}_{dk} = \mathbf{R}_{dk}^\top \mathbf{R}_{dk}$ where

$$\mathbf{R}_{dk} = \begin{bmatrix} \exp(\eta_1) & \eta_2 & \eta_4 & \cdots & \eta_x \\ 0 & \exp(\eta_3) & \eta_5 & \cdots & \vdots \\ 0 & 0 & \ddots & & \\ 0 & 0 & 0 & 0 & \exp(\eta_{d(d+1)/2}) \end{bmatrix} \tag{7}$$

is the unique Choesky factor of $\mathbf{Q}_{dk}$ with $d(d+1)/2$ parameters $\boldsymbol{\eta}_k$.

In general, there are no restrictions on the spatial structure of the process, specified through $\mathbf{Q}_s$. However, since we want to use the method for large problems we choose a model so that $\mathbf{Q}_s$ is sparse. For a discrete domain, we can then choose any type of GMRF model, e.g. the popular CAR models [Besag, 1974]. The particular choice we use is a CAR model that corresponds to a Gaussian Matérn field. Constructing the spatial precision matrix using the SPDE connection [Lindgren et al., 2011] between the discrete CAR models and the continuous Matérn fields allows us to use separate discretizations for $\mathbf{z}$ and $\boldsymbol{\xi}$, which is desirable if the data is such that the process $\boldsymbol{\xi}$ is smoothly varying compared to the resolution for $\mathbf{z}$. The basic idea is to use a basis expansion $\xi(\mathbf{s}) = \sum_{i=1}^n \varphi_i(\mathbf{s}) w_i$, where $\{\varphi_i\}$ are known compactly supported piecewise linear basis functions and $\mathbf{w} = \{w_i\}$ is a zero mean multivariate normal distribution with precision matrix $\mathbf{Q}_s = c\mathbf{K}\mathbf{C}^{-1}\mathbf{K}$, where $\mathbf{K} = (\mathbf{G} + \kappa^2\mathbf{C})$ with $G_{ij} = \langle \nabla\varphi_i, \nabla\varphi_j \rangle$, $C_{ii} = \langle \varphi_i, 1 \rangle$ and c as a positive scaling constant. The number of basis functions, n, can be chosen smaller than the number of locations in the domain for $\mathbf{z}$ in order to increase the computational efficiency of the model.

This particular choice of $\mathbf{Q}_s$ corresponds to a Matérn field with shape parameter $\alpha = 2$, which for models in $\mathbb{R}^3$ results in the exponential covariance function. Since the parameter κ^2 needs to be positive, we parametrize it as $\kappa^2 = \exp(\kappa_0)$. The constant c, in the precision matrix, is chosen so that the spatial part have variance one, which achieved for $c = \Gamma(2 - D/2)(4\pi)^{-D/2}\kappa^{D-4}$, where D denotes the dimension of the spatial domain. This way, $\mathbf{Q}_s$ determined the spatial correlation and $\mathbf{Q}_d$ controls the variances.

The particular choice of covariance structure presented here is a so called proportional correlation model [Chiles and Delfiner, 1999] as the resulting stationary covariance function for $\boldsymbol{\xi}$ can be written as $C(\mathbf{h}) = \mathbf{Q}_d^{-1}\rho(\mathbf{h}), \mathbf{h} \in \mathbb{R}^d$. There are several fully parametric alternatives to this model, such multivariate Matérn fields [Hu et al., 2013].

3.3 The measurement noise $\boldsymbol{\varepsilon}$

We assume that the measurement noise $\boldsymbol{\varepsilon}$ is mean-zero Gaussian white noise with a spatially constant variance. One can either assume that the noise is the same for each dimension of the data, $\boldsymbol{\Sigma}_\varepsilon = \sigma^2\mathbf{I}_{nd}$, or one can allow for a separate variance for each dimension of the data, $\boldsymbol{\Sigma}_\varepsilon = \text{diag}(\sigma_1^2, \ldots, \sigma_d^2) \otimes \mathbf{I}_n$. Here, $\mathbf{I}_m$ denotes an $m \times m$ identity matrix. Since the variance parameters σ_i are positive, we parametrize them as $\sigma_i = \exp(\sigma_{i0})$

4 Parameter estimation

Parameter estimation for MRF mixture models is difficult, and allowing for spatial dependency within each class introduces further complications. Furthermore, we want these models to be useful for massive multivariate problems in $\mathbb{R}^3$, which are common in MR imaging, and this makes computational efficiency of the estimation procedure paramount.

The MRF mixture models are typically either estimated with some modified version of the EM algorithm or through Monte Carlo (MC) methods. Both of these procedures are too computationally demanding to be useful for the LGFM models. Instead, we base our estimation on the EM gradient (EMG) algorithm. The main idea behind this method is that if one can easily calculate the gradient

$\nabla_{\boldsymbol{\Psi}} \log L(\boldsymbol{\Psi}; \mathbf{z}, \mathbf{Y})$ of the augmented likelihood, then knowing the posterior $\pi(\mathbf{z}|\mathbf{y}, \boldsymbol{\Psi})$ one can compute the exact gradient of the log likelihood $\log L(\boldsymbol{\Psi}; \mathbf{Y})$ as

$$\begin{aligned}\nabla_{\boldsymbol{\Psi}} \log L(\boldsymbol{\Psi}; \mathbf{Y}) =& \nabla_{\boldsymbol{\Psi}} \log \pi(\mathbf{Y}|\boldsymbol{\Psi}) = \frac{1}{\pi(\mathbf{Y}|\boldsymbol{\Psi})} \nabla_{\boldsymbol{\Psi}} \int \pi(\mathbf{Y}, \mathbf{z}|\boldsymbol{\Psi}) \, \mathrm{d}\mathbf{z} \\ =& \int \frac{\pi(\mathbf{Y}, \mathbf{z}|\boldsymbol{\Psi})}{\pi(\mathbf{Y}|\boldsymbol{\Psi})} \nabla_{\boldsymbol{\Psi}} \log \pi(\mathbf{Y}, \mathbf{z}|\boldsymbol{\Psi}) \, \mathrm{d}\mathbf{z} = \int \pi(\mathbf{z}|\mathbf{Y}, \boldsymbol{\Psi}) \nabla_{\boldsymbol{\Psi}} \log \pi(\mathbf{Y}, \mathbf{z}|\boldsymbol{\Psi}) \, \mathrm{d}\mathbf{z} \\ =& \mathsf{E}_{\mathbf{z}} \left[\nabla_{\boldsymbol{\Psi}} \log \pi(\mathbf{Y}, \mathbf{z}|\boldsymbol{\Psi}) | \mathbf{Y}, \boldsymbol{\Psi} \right].\end{aligned}$$

The idea is then to use the exact gradient in a gradient descent method. At step p in the EMG algorithm, the gradient of the likelihood is calculated and a step

$$\boldsymbol{\Psi}^{(p+1)} = \boldsymbol{\Psi}^{(p)} + \mathbf{S} \nabla_{\boldsymbol{\Psi}} \log L(\boldsymbol{\Psi}; \mathbf{Y})$$

where $\mathbf{S}$ is a matrix determining the step size. Taking $\mathbf{S} = \gamma \mathbf{I}$ where $\mathbf{I}$ is the identity matrix, we obtain an ordinary gradient descent method which has linear convergence. Ideally, we would like to take $\mathbf{S}$ as the inverse of the Hessian matrix $\mathbf{H}$ to obtain a Newton method with quadratic convergence. Often, one cannot compute the true Hessian matrix of the log-likelihood, and Lange [1995] instead proposed using

$$\mathbf{S} = \mathsf{E}_{\mathbf{z}}(H_{\boldsymbol{\Psi}}(\log \pi(\mathbf{Y}, \mathbf{z}|\boldsymbol{\Psi}))|\mathbf{Y}, \boldsymbol{\Psi}), \tag{8}$$

where $H_{\boldsymbol{\Psi}}(f)_{ij} = \frac{\partial^2 f}{\partial \boldsymbol{\Psi}_i \partial \boldsymbol{\Psi}_i}$ is the hessian operator. The motivation behind this choice of scaling matrix is that from dealing with spatial data we have experienced that the two first conditional moments often are little affected by changes in the parameters, which would indicate that $\mathbf{S}$ is a good approximation of the true hessian with the advantage of being readily available in most situations.

In the MRF mixture models, we cannot evaluate the gradient of the likelihood analytically, and one can then use MC sampling to estimate the gradient as

$$\nabla_{\boldsymbol{\Psi}} \log L(\boldsymbol{\Psi}; \mathbf{Y}) = \mathsf{E}_{\mathbf{z}} \left[\nabla_{\boldsymbol{\Psi}} \log \pi(\mathbf{Y}, \mathbf{z}|\boldsymbol{\Psi}) | \mathbf{Y}, \boldsymbol{\Psi} \right] \approx \frac{1}{T} \sum_{t=1}^{T} \nabla_{\boldsymbol{\Psi}} \log \pi(\mathbf{Y}, \mathbf{z}^{(t)}|\boldsymbol{\Psi}),$$

where $\mathbf{z}^{(t)}$ are draws from $\pi(\mathbf{z}|\mathbf{Y}, \boldsymbol{\Psi})$. In a similar fashion, one can use MC sampling to evaluate the approximate Hessian that is used to determine the step size

$$\mathbf{S} \approx \frac{1}{T} \sum_{t=1}^{T} H_{\boldsymbol{\Psi}}(\log \pi(\mathbf{Y}, \mathbf{z}^{(t)}|\boldsymbol{\Psi})).$$

We refer to this estimation procedure as the MCEMG algorithm.

To simplify the presentation, we split this section in three parts. In the first part, we go through the details of the estimation for the MRF mixture model, presenting a version of the method based on pseudo-likelihoods. In the second part we cover estimation for the latent Gaussian model, and one should note here that the estimation method is an attractive alternative for estimation of latent Gaussian models for massive datasets since it avoids all calculations of log-determinants, which is usually the computational bottleneck in maximum-likelihood estimation procedures for such problems. Finally, we combine the results for the MRF mixture models and the latent Gaussian models to and estimation procedure for the full LGFM model.

4.1 Estimation of the MRF mixture model

As a first step towards an estimation method for the LGFM models, we in this section discuss parameter estimation of the MRF mixture models. To make the results of this section more easily applicable to the LGFM model, we parametrize the Gaussian distributions using the mean and cholesky factor

of the precision matrix. Let $\boldsymbol{\theta}_k = \{\boldsymbol{\mu}_k, \mathbf{Q}_k\}$ where $\mathbf{Q}_{dk} = \boldsymbol{\Sigma}_k^{-1}$ is parametrized as $\mathbf{Q}_{dk} = \mathbf{R}_{dk}^\top \mathbf{R}_{dk}$ and $\mathbf{R}_{dk}$ has the form (7). Thus, the model parameters that need to be estimated are $\boldsymbol{\Psi} = \{\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\theta}\}$, where $\boldsymbol{\theta} = \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K\}$ contains all parameters for the Gaussian distributions, $\boldsymbol{\theta}_k = \{\boldsymbol{\mu}_k, \boldsymbol{\eta}_k\}$.

Maximum likelihood estimation for this model is difficult since there is no simple form for the data likelihood. However, if we augment the data with the hidden class belongings, the augmented likelihood has a simpler form, $L(\boldsymbol{\Psi}; \mathbf{z}, \mathbf{Y}) = \pi(\mathbf{z}|\boldsymbol{\alpha}, \boldsymbol{\beta})\pi(\mathbf{Y}|\mathbf{z}, \boldsymbol{\theta})$. This suggest that we could use an EM algorithm [Dempster et al., 1977] where one would iterate calculating the function

$$\mathcal{Q}\left(\boldsymbol{\Psi}, \boldsymbol{\Psi}^{(p)}\right) = E\left[\log L(\boldsymbol{\Psi}; \mathbf{z}, \mathbf{Y}) | \mathbf{Y}, \boldsymbol{\Psi}^{(p)}\right], \tag{9}$$

where $\boldsymbol{\Psi}^{(p)}$ denotes the current estimate of $\boldsymbol{\Psi}$ at the pth iteration of the algorithm and then maximize $\mathcal{Q}\left(\boldsymbol{\Psi}, \boldsymbol{\Psi}^{(p)}\right)$ with respect to $\boldsymbol{\Psi}$ in order to obtain the next estimate of the parameter vector.

Unfortunately, the normalizing constant Z for the MRF distribution depends on the parameters and is intractable for large problems. Thus, we cannot evaluate $\pi(\mathbf{z}|\boldsymbol{\alpha}, \boldsymbol{\beta})$. A solution to this problem is to replace $\pi(\mathbf{z}|\boldsymbol{\alpha}, \boldsymbol{\beta})$ with a pseudo-likelihood, $\pi_p(\mathbf{z}|\boldsymbol{\alpha}, \boldsymbol{\beta})$, which is a product of the full conditionals of $\mathbf{x}$. Let $f_{ik} = \sum_{j \in \mathcal{N}_i} z_{jk}$ denote the sum of the neighboring pixels to z_{ik}, the conditional class probability of a pixel i can then be written as $\mathsf{P}(x_i = k | f_{ik}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \mathsf{E}(z_{ik} | f_{ik}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \propto \exp(\alpha_k + \beta_k f_{ik})$, and the pseudo-likelihood is

$$\pi_p(\mathbf{z}|\boldsymbol{\alpha}, \boldsymbol{\beta}) = \prod_i \pi(x_i | x_j, j \in \mathcal{N}_i, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \prod_i \frac{\exp(\sum_k \alpha_k z_{ik} + \sum_k \beta_k z_{ik} f_{ik})}{\sum_k \exp(\alpha_k + \beta_k f_{ik})}.$$

To avoid bias due to this procedure, only conditionally independent pixels are included in the product simultaneously, and the coding method [Besag, 1974] is used to combine the estimates based on different combinations of conditionally independent sets of pixels. Since the neighborhood structure in Figure 2 (a) is used, two sets of conditionally independent pixels are obtained using the checkerboard pattern shown in Figure 2 (b), where the black nodes are conditionally independent given the white nodes and vice versa.

Hence, the function we need to calculate the expectation of to obtain $\mathcal{Q}\left(\boldsymbol{\theta}, \boldsymbol{\theta}^{(p)}\right)$ is

$$\log PL(\boldsymbol{\Psi}; \mathbf{z}, \mathbf{Y}) = \log \pi(\mathbf{Y}|\mathbf{z}, \boldsymbol{\theta}) + \sum_{k,i} \alpha_k z_{ik} + \sum_{k,i} \beta_k z_{ik} f_{jk} - \sum_i \log\left(\sum_k \exp(\alpha_k + \beta_k f_{ik})\right). \tag{10}$$

Using the pseudo likelihood for the MRF part of $\mathcal{Q}$, the function can be written as

$$\mathcal{Q}(\boldsymbol{\Psi}, \boldsymbol{\Psi}^{(p)}) = \mathsf{E}(\log(\pi_p(\mathbf{z}|\boldsymbol{\alpha}, \boldsymbol{\beta})) + \sum_i \sum_k \mathsf{E}(z_{ik} | \mathbf{Y}, \boldsymbol{\theta}^{(p)}) \log \pi(Y_i | \boldsymbol{\theta}_k).$$

We cannot evaluate the expectation of the pseudo likelihood analytically, and we therefore replace it with an MC approximation, which requires sampling from the posterior distribution. Using Bayes formula and the independence assumption, one has

$$\mathsf{E}\left(z_{ik} \mid f_{ik}, \mathbf{Y}, \boldsymbol{\Psi}\right) \propto p\left(y_i \mid \boldsymbol{\theta}_k\right) \exp\left(\alpha_k + \beta_k f_{ik}\right) = \exp\left(\tilde{\alpha}_{ik} + \beta_k f_{ik}\right)$$

with $\tilde{\alpha}_{ik} = \alpha_k + \log \pi(Y_i|\boldsymbol{\theta}_k)$. Thus, the posterior distribution is simply a non-stationary extension of the original MRF model. We can therefore use Gibbs sampling to draw samples $\mathbf{z}^{(t)}$ from the posterior. Dividing the nodes using the checkerboard pattern in Figure 2 (b), and denoting the black nodes $\mathbf{z}_b$ and the white nodes $\mathbf{z}_w$, Gibbs sampling of $\mathbf{z}$ is performed by iterating sampling $\mathbf{z}_w^{(t)}$ from $\pi(\mathbf{z}_w | \mathbf{z}_b^{(t-1)}, \mathbf{Y}, \boldsymbol{\Psi})$ and sampling $\mathbf{z}_b^{(t)}$ from $\pi(\mathbf{z}_b | \mathbf{z}_w^{(t)}, \mathbf{Y}, \boldsymbol{\Psi})$.

Now, this is about as far as one gets with the EM algorithm since the M step is highly problematic. Versions of the MRF mixture model has been used several times in tissue classification of magnetic resonance images [Held et al., 1997, Zhang et al., 2001, Van Leemput et al., 1999], and in these

situations the model is usually fitted to data using an EM estimator for the Gaussian parameters together with an iterated conditional modes (ICM) estimator for the MRF parameters. Convergence of this mixed estimation procedure is not easy to motivate theoretically, and the method can be computationally demanding.

However, the EM gradient method is straight-forward to implement. The derivatives need to evaluate the gradient are presented in Appendix A. At step p in the EM gradient algorithm, we run the Gibbs sampler to approximate the gradient and the scaling $\mathbf{S}$ and then take a step $\boldsymbol{\Phi}^{(p+1)} = \boldsymbol{\Phi}^{(p)} + \mathbf{S}\nabla_{\boldsymbol{\Psi}} \log PL(\boldsymbol{\Psi}; \mathbf{Y})$. Thus, there is no need for numerical optimization or Taylor approximations to calculate the parameter updates, as is needed if an EM algorithm is used. Note that $\nabla_{\boldsymbol{\Psi}} \log PL(\boldsymbol{\Psi}'; \mathbf{Y}) = \nabla_{\boldsymbol{\Psi}} \mathcal{Q}(\boldsymbol{\Psi}, \boldsymbol{\Psi}')|_{\boldsymbol{\Psi}=\boldsymbol{\Psi}'}$, thus the function maximized in the gradient algorithm is the same function maximized in the EM-algorithm.

4.2 Estimation of the latent Gaussian model

As a second step towards the estimation procedure for the full LGFM models, we in this section discuss the estimation of the latent Gaussian model (4) where $\boldsymbol{\xi}$ is given introduced in Section 3.2 and $\boldsymbol{\varepsilon}$ is introduced in Section 3.3. To simplify the presentation, we assume that the measurement noise has a common variance for all dimensions of the data, and the extension to separate noise variances is trivial.

Let $\boldsymbol{\Psi} = \{\boldsymbol{\mu}, \boldsymbol{\eta}, \sigma, \kappa\}$ be the vector containing all model parameters. Since the model is Gaussian, likelihood estimation of all parameters can be performed by numerical optimization of $\log \pi(\boldsymbol{\Psi}|\mathbf{Y})$, which has a closed form [see e.g. Bolin and Lindgren, 2011]. Even though this procedure is commonly used and theoretically straight-forward, it is computationally demanding. The problem is that one needs to calculate the determinant of $\hat{\mathbf{Q}} = \mathbf{Q} + \frac{1}{\sigma^2}\mathbf{A}^\top\mathbf{A}$ and solve the quadratic form $\mathbf{Y}^\top\mathbf{A}\hat{\mathbf{Q}}^{-1}\mathbf{A}^\top\mathbf{Y}$ each time the optimizer evaluates the likelihood. This is most efficiently done using sparse Cholesky factorization and back-substitution; however, even though one has a separable covariance structure, this does not help when calculating the Cholesky factor, which makes the evaluation of the likelihood highly computationally demanding for large multivariate spatial problems.

The need to calculate the determinant of $\hat{\mathbf{Q}}$ is avoided if the EMG method is used. Hence, the likelihood is augmented with the latent variable $\boldsymbol{\xi}$ and we calculate the gradient and the scaling matrix $\mathbf{S}$ by the procedure described above. The augmented log-likelihood is

$$l = \log \pi(\mathbf{Y}, \boldsymbol{\xi}|\boldsymbol{\Psi}) = -m\sigma_0 - \frac{1}{2e^{2\sigma_0}}(\mathbf{Y} - \mathbf{B}\boldsymbol{\beta} - \mathbf{A}\boldsymbol{\xi})^\top(\mathbf{Y} - \mathbf{B}\boldsymbol{\beta} - \mathbf{A}\boldsymbol{\xi}) + \frac{1}{2}\log|\mathbf{Q}| - \frac{1}{2}\boldsymbol{\xi}^\top\mathbf{Q}\boldsymbol{\xi},$$

and the derivatives needed to evaluate the gradient of the log-likelihood $L(\boldsymbol{\Psi}; \mathbf{Y})$ are presented in Appendix B.

The gradient method replaces computing $|\hat{\mathbf{Q}}|$ with computing various traces and there are two computational issues that have to be solved for the method to be applicable to large data sets. The first is to solve $\hat{\boldsymbol{\xi}} = \hat{\mathbf{Q}}^{-1}\mathbf{b}$ for a vector $\mathbf{b}$, which can be done using sparse cholesky factorizations and back-substitution. However, in order to reduce the computationally complexity we instead use the preconditioned conjugate gradient method (PCG) with a robust incomplete Cholesky preconditioner [Ajiz and Jennings, 1984] to solve the equation.

The second issue is to solve the various traces of inverse matrices present in the expressions for the gradients. Recent work in spatial statistics [Anitescu et al., 2012, Stein et al., 2013] has proposed solving this issue using stochastic programming. The basic idea is to note that $\mathsf{E}[\mathbf{u}^\top\mathbf{Q}\mathbf{u}] = tr(\mathbf{Q})$ for any vector $\mathbf{u}$ of independent random variables u_i with mean zero and variance one [Hutchinson, 1990]. Thus, we can rewrite all the traces in the gradient ∇l as expectations, which can ben be approximated using Monte Carlo integration. For example $\text{tr}\left(\mathbf{Q}_s^{-1}\frac{\partial \mathbf{Q}_s}{\partial \phi_j}\right) = \mathsf{E}\left[\mathbf{u}^\top\frac{\partial \mathbf{Q}_s}{\partial \phi_j}\mathbf{Q}_s^{-1}\mathbf{u}\right]$ is replaced with $k^{-1}\sum_{i=1}^k \mathbf{u}_i^\top\frac{\partial \mathbf{Q}_s}{\partial \phi_j}\mathbf{Q}_s^{-1}\mathbf{u}_i$. The standard choice for $\mathbf{u}_i$ is to use mean-zero Bernoulli random variables but for spatial problems the variance of the estimator can be reduced by for example using the probing vectors proposed by Aune et al. [2012]. The PCG method is used to efficiently calculate $\mathbf{Q}_s^{-1}\mathbf{u}_i$.

The resulting approximation, ∇l_k, of the gradient ∇l is a random function with $\mathsf{E}[\nabla l_k] = \nabla l$. Shapiro et al. [2009] shows that, under mild conditions, the local minimum of ∇l_k converges to a local minimum of ∇l with probability one as $k \to \infty$. Using the iterative methods in combination with the EM gradient method results in a highly computationally efficient method for estimating latent Gaussian models.

4.3 Estimation of the LGFM model

With the estimators for the MRF mixture model and the latent Gaussian model derived, it is now simply a matter of combining these two for making the estimator for the LGFM model. We augment the data-likelihood by both the MRF $\mathbf{z}$ and the GRFs $\boldsymbol{\xi} = \{\boldsymbol{\xi}_1, \dots, \boldsymbol{\xi}_k\}$, and let $l = \log \pi(\mathbf{Y}, \mathbf{z}, \boldsymbol{\xi}|\boldsymbol{\Psi})$ where $\boldsymbol{\Psi}$ now denotes all model parameters. To calculate the required gradient, we note that

$$\begin{aligned}
\nabla_{\boldsymbol{\Psi}} \log L(\boldsymbol{\Psi}; \mathbf{Y}) &= \int \pi(\mathbf{z}, \boldsymbol{\xi}|\mathbf{Y}, \boldsymbol{\Psi}) \nabla_{\boldsymbol{\Psi}} \log \pi(\mathbf{Y}, \mathbf{z}, \boldsymbol{\xi}|\boldsymbol{\Psi}) \,\mathrm{d}\mathbf{z} \,\mathrm{d}\boldsymbol{\xi} \\
&= \int \pi(\mathbf{z}|\mathbf{Y}, \boldsymbol{\Psi}) \int \pi(\boldsymbol{\xi}|\mathbf{z}, \mathbf{Y}, \boldsymbol{\Psi}) \nabla_{\boldsymbol{\Psi}} \log \pi(\mathbf{Y}, \mathbf{z}, \boldsymbol{\xi}|\boldsymbol{\Psi}) \,\mathrm{d}\boldsymbol{\xi} \,\mathrm{d}\mathbf{z} \\
&= \mathsf{E}_{\mathbf{z}} \left(\mathsf{E}_{\boldsymbol{\xi}} \left(\nabla_{\boldsymbol{\Psi}} \log \pi(\mathbf{Y}, \mathbf{z}, \boldsymbol{\xi}|\boldsymbol{\Psi}) \mid \mathbf{z}, \mathbf{Y}, \boldsymbol{\Psi} \right) \mid \mathbf{Y}, \boldsymbol{\Psi} \right) \\
&= \mathsf{E}_{\mathbf{z}} \left(\nabla_{\boldsymbol{\Psi}} \log \pi(\mathbf{z}|\boldsymbol{\alpha}, \boldsymbol{\beta}) + \mathsf{E}_{\boldsymbol{\xi}} \left(\nabla_{\boldsymbol{\Psi}} \log \pi(\mathbf{Y}, \boldsymbol{\xi}|\mathbf{z}, \boldsymbol{\sigma}) \mid \mathbf{z}, \mathbf{Y}, \boldsymbol{\Psi} \right) \mid \mathbf{Y}, \boldsymbol{\Psi} \right).
\end{aligned}$$

As in previous section, the expectation with respect to $\mathbf{z}$ must be approximated using MC sampling. However, since the expectation with respect to $\boldsymbol{\xi}$ is known analytically, see Appendix B, we can use Rao-Blackwellization to calculate gradient as

$$\nabla_{\boldsymbol{\Psi}} \log L(\boldsymbol{\Psi}; \mathbf{Y}) = \frac{1}{N} \sum_{t=1}^{T} \left(\nabla_{\boldsymbol{\Psi}} \log \pi(\mathbf{z}^{(t)}|\boldsymbol{\alpha}, \boldsymbol{\beta}) + \mathsf{E}_{\boldsymbol{\xi}} \left(\nabla_{\boldsymbol{\Psi}} \log \pi(\mathbf{Y}, \boldsymbol{\xi}|\mathbf{z}^{(t)}, \boldsymbol{\sigma}) \mid \mathbf{z}^{(t)}, \mathbf{Y}, \boldsymbol{\Psi} \right) \right).$$

This means that we can use the gradients calculated in the previous sections with two minor changes for implementing the estimation procedure for the full LGFM model.

The first difference is that the Gaussian likelihood (4.2) for each, independent, field $\boldsymbol{\xi}_k$ is replaced with

$$\begin{aligned}
\log \pi(\mathbf{Y}, \boldsymbol{\xi}_k|\mathbf{z}^{(t)}, \boldsymbol{\Psi}) = &- m_k^{(t)} \sigma_0 - \frac{1}{2e^{2\sigma_0}} (\mathbf{Y}^{(t)} - \mathbf{B}_k^{(t)} \boldsymbol{\beta} - \mathbf{A}_k^{(t)} \boldsymbol{\xi}_k)^\top (\mathbf{Y}^{(t)} - \mathbf{B}_k^{(t)} \boldsymbol{\beta} - \mathbf{A}_k^{(t)} \boldsymbol{\xi}_k) \\
&+ \frac{1}{2} \log |\mathbf{Q}_k| - \frac{1}{2} \boldsymbol{\xi}_k^\top \mathbf{Q}_k \boldsymbol{\xi}_k
\end{aligned}$$

where $m_k^{(t)} = d \sum_j z_{kj}$ and $\mathbf{Y}^{(t)}$, $\mathbf{A}_k^{(t)}$ and $\mathbf{B}_k^{(t)}$ are constructed by taking $\mathbf{Y}$, $\mathbf{A}$, and $\mathbf{B}$ and only keeping the rows that corresponds to the pixels with $\mathbf{z}_k^{(t)} = 1$. Thus, m, $\mathbf{A}$ and $\mathbf{B}$ are replaced with $m_k^{(t)}$, $\mathbf{A}_k^{(t)}$ and $\mathbf{B}_k^{(t)}$ respectively in the Gaussian gradients presented in Appendix B.

The second difference is how $\mathbf{z}^{(t)}$ is simulated. Unlike for the regular MRF mixture model, $\mathbf{Y}|\theta_k$ is not a vector independent variables and the sampling method for $\mathbf{z}$ in the MRF mixture model therefore has to be modified. To simulate $\mathbf{z}^{(t)}$, we introduce an extra step in the Gibbs sampler for the MRF mixture model as follows

1. Sample the Gaussian fields $\{\boldsymbol{\xi}_k\}^{(t)}$ from their respective distributions $\pi(\boldsymbol{\xi}_k|\mathbf{Y}, \mathbf{z}^{(t-1)}, \boldsymbol{\Psi})$.
2. Sample $\mathbf{z}_w^{(t)}$ from $\pi(\mathbf{z}_w|\mathbf{z}_b^{(t-1)}, \mathbf{Y}, \{\boldsymbol{\xi}_k\}^{(t)}, \boldsymbol{\Psi})$.
3. Sample $\mathbf{z}_b^{(t)}$ from $\pi(\mathbf{z}_b|\mathbf{z}_w^{(t)}, \mathbf{Y}, \{\boldsymbol{\xi}_k\}^{(t)}\}, \boldsymbol{\Psi})$.

Since $\mathbf{Y}|\{\boldsymbol{\xi}_k\}^{(t)}, \boldsymbol{\Psi}$ is a vector of independent variables, the second and third step of the Gibbs sampler are performed in the same way as for the MRF mixture model. It should also be noted that the sampled fields $\{\boldsymbol{\xi}_k\}^{(t)}$ are not used in the optimization other than to generate $\mathbf{z}^{(t)}$.

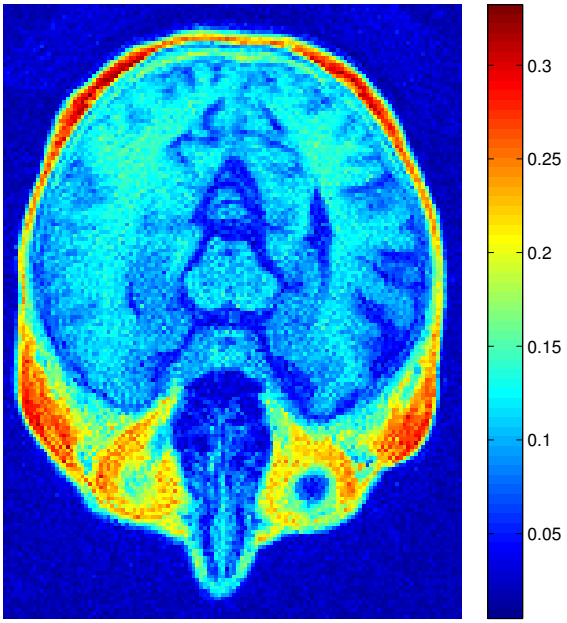


Figure 3: A noisy MR image of size 166×124 pixels.

The simulation from $\pi(\boldsymbol{\xi}_k|\mathbf{Y}, \mathbf{z}^{(t-1)}, \boldsymbol{\Psi})$ is typically solved using Cholesky factorization of $\hat{\mathbf{Q}}_k = \mathbf{Q}_k + \frac{1}{\sigma^2}(\mathbf{A}_k^{(t)})^\top \mathbf{A}_k^{(t)}$; however, this is not possible for large data sets. We instead use the following method, from [Papandreou and Yuille, 2011], which avoids the calculation of Cholesky factors entirely,

1. Generate $\mathbf{x} = \left((\mathbf{K}\mathbf{C}^{-1/2}) \otimes \mathbf{R}_{dk}\right) \mathbf{x}_1 + \frac{1}{\sigma}(\mathbf{A}_k^{(t)})^\top \mathbf{x}_2$ where $\mathbf{x}_1$ and $\mathbf{x}_2$ are vectors of independent $N(0,1)$ random variables.
2. Solve $\hat{\mathbf{Q}}_k \boldsymbol{\xi}_k = \mathbf{x} + \frac{1}{\sigma^2}(\mathbf{A}_k^{(t)})^\top (\mathbf{Y}^{(t)} - \mathbf{B}_k^{(t)} \boldsymbol{\beta})$.

Also here, the PCG method with a robust incomplete Cholesky preconditioner is used to solve the linear equation in the second step.

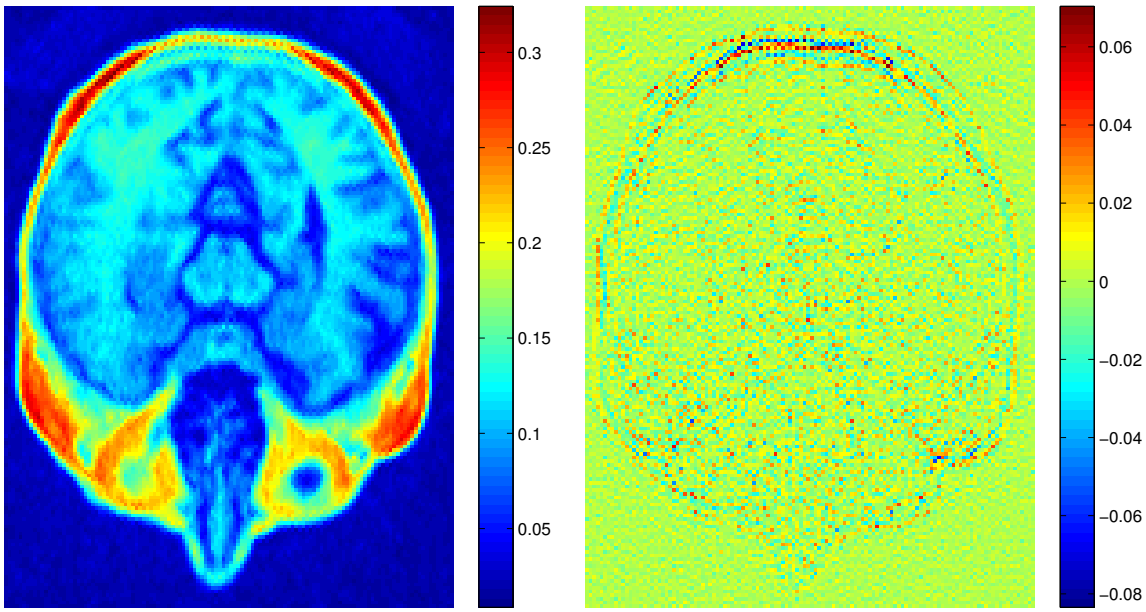
5 An application to magnetic resonance imaging

There are a number of possible applications to brain imaging that could be considered for this model class. However, in this section we only present a simple application to noise reduction. The MR image we analyze is a subset of data that previously has been used for CT substitute generation and is described in detail in Johansson et al. [2011]. The image is taken with a radial UTE sequence with a 10 degree flip angle, a repetition time of 6 ms, and an echo time of 0.07 ms. The UTE images were reconstructed to a matrix with $192 \times 192 \times 192$ voxels with isotropic resolution and a voxel size of 1.33 mm. For simplicity, we analyze only one slice of this data, which is of size 192×192 pixels. After removing parts of the slice that only contains areas outside the head, we obtain the image shown in Figure 3 which is of size 166×124 pixels.

As seen in the figure, the data is somewhat noisy and the goal is therefore use statistical techniques to reduce the noise in the image. As a first method, we use a standard latent Gaussian model, which can be described as the LGFM model in Section 3 with $K = 1$. The resulting estimate, $\hat{X}$, is shown in Figure 4 (a) and the kriging residuals, $\hat{X} - Y$, are shown in Figure 4 (b). If the model was correct, there should be no spatial structure in the residuals. However, we clearly see the contour of the head in the residuals, which means that this simple latent Gaussian model likely is insufficient for doing noise reduction of this image.

	LGM	LGFM_1	LGFM_2	LGFM_3
κ^2	0.0256	0.0550	0.0132	0.0005
σ^2	0.0396	0.0303	0.0303	0.0303
τ	2.8437	4.1980	194.14	0.0081
μ	1.3568	1.5297	0.3857	3.2251

Table 1: Parameter estimates for the latent Gaussian model (LGM) and the three mixture components of the LGFM model. The spatial dependency parameter β for MRF in the LGFM model was assumed to be the same for all classes, and was estimated to 2.73, and the prior parameters α_k were fixed to zero. The estimation was done on data standardized to have variance one.



(a) (b)

Figure 4: Estimated MR image using a latent Gaussian model (a) and the kriging residuals (b).

As an alternative to the latent Gaussian model, we fit a LGFM model with three mixture components. The reason for choosing three components is to keep the model simple while being able to separate the air outside the head and the bone from the other tissue types, as these two classes clearly stands out in the image. In order to keep the model simple, the MRF parameters α_k are fixed to zero and a common β parameter is assumed for all classes. Estimates of the other parameters are shown in Table 1, which also shows the parameter estimates for the latent Gaussian model as a reference.

Starting values for the LGFM estimation are obtained by first doing a classification of the data using a standard Gaussian mixture model and then estimating a latent Gaussian model for each class in the estimated mixture. The classification using the LGFM model is shown in Figure 5, Panel (a) and the corresponding classification is show in Panel (b). One should note that this classification is unsupervised and obtained as a byproduct while fitting the LGFM model, and it clearly finds the desired regions in the image. Panel (c) shows the difference between the LGFM estimate and the LGM estimate in Figure 4 (a), and one sees that the difference is quite large, especially near the tissue boundaries. Finally, Panel (d) shows the kriging residuals of the LGFM model in the same color scale as the residuals of the latent Gaussian model in Figure 4 (b), and although there is still some structure in the residuals, the result is much better.

Thus, the LGFM model performs much better than the latent Gaussian model, and one of the

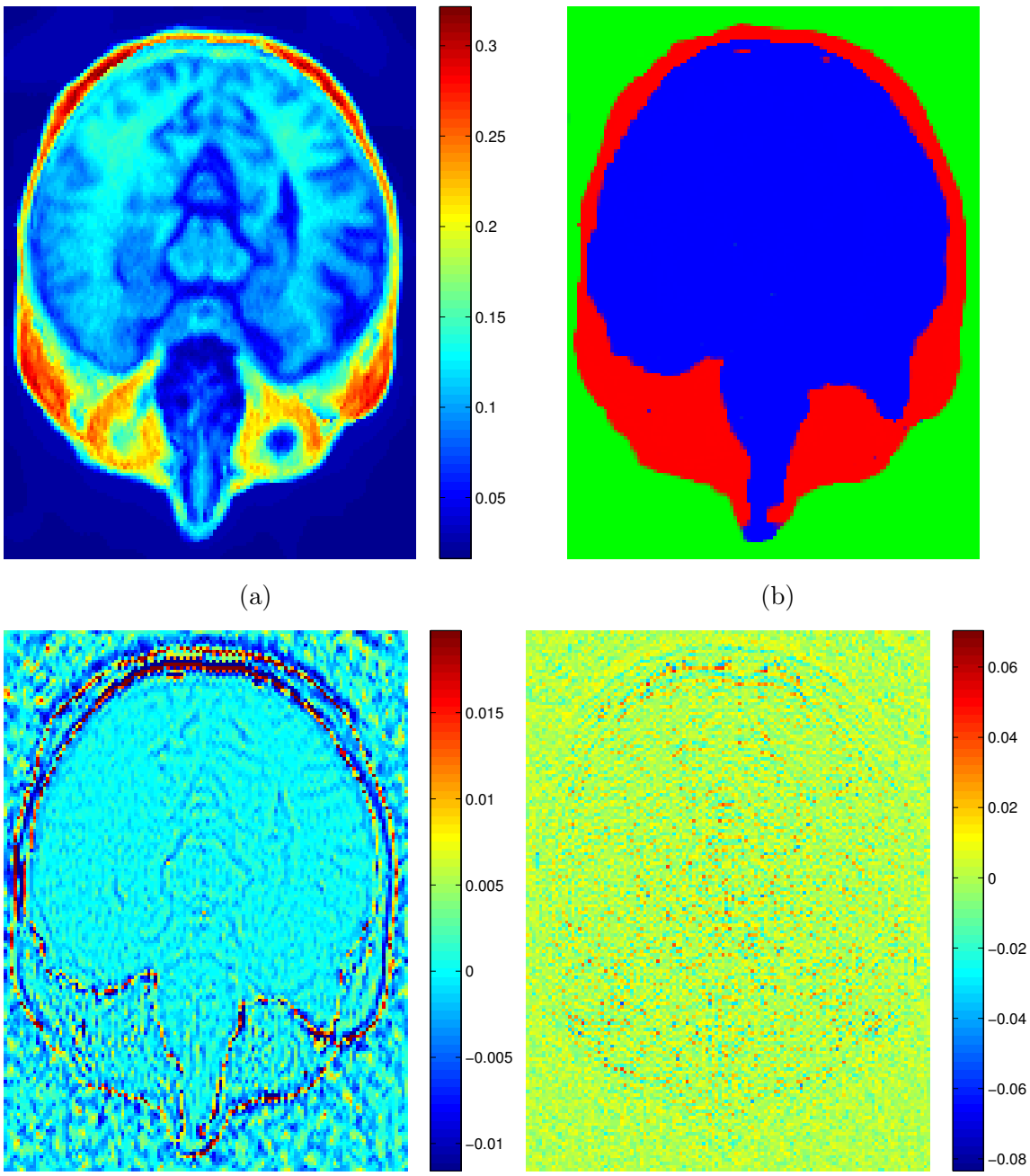


Figure 5: Estimated MR image using a LGFM model with three classes (a) and the corresponding classification (b). Panel (c) shows the difference between this estimate and the estimate using a latent Gaussian model, the color scale has been truncated to the middle 98% of the values to improve the visibility, which means that the largest differences are truncated in the color scale. Panel (d) shows the kriging residuals for the LGFM model, which shows much less spatial structure than the corresponding residuals for the latent Gaussian model. The color scale in Panel (d) has been set to match the color scale in Figure 4 (b).

reasons for this is that the model parameters are allowed to vary between the classes. This behavior could also be obtained by using a non-stationary latent Gaussian model, where the parameters are allowed to vary with space. However, the second important reason for the better behavior of the LGFM model, which is much harder to obtain using a non-stationary latent Gaussian model or an adaptive smoother is that the estimate for each class only uses data that is classified as belonging to that class. This allows for much sharper changes in the resulting estimate, and such behavior cannot be obtained in any simple way using an ordinary latent Gaussian model.

In this example, the main purpose was noise reduction and using the LGFM model we obtained a classification of the image as a byproduct. If the main objective was segmentation, a method worth mentioning is the popular adaptive segmentation method by Wells III et al. [1996]. It is worth noting that this method fits into the general LGFM framework. In our notation, their model that is used for classification can be written as

$$\log(\mathbf{Y}_i) = \boldsymbol{\xi}_i + \sum_{k=1}^{K} z_{ik}\mathbf{G}_{ik} \tag{11}$$

where the field $\boldsymbol{\xi}$ is denoted a bias field and the second part is a standard gaussian mixture model (z is not a MRF in this model). This model can be reformulated as a transformed LGFM model, without measurement noise and with dependent mixture fields $\boldsymbol{\xi}_{ki} = \boldsymbol{\xi}_i + \mathbf{G}_{ki}$. An important difference that should be noted is that Wells III et al. [1996] assumes that the covariance matrix for $\boldsymbol{\xi}$ is a known band matrix and makes no attempts at estimating it, while we estimate the covariance function for each class.

6 Discussion

This work has introduced the class of LGFM models as well as a computationally efficient stochastic gradient parameter estimation method for the model class.

There are a number of directions in which this work can be extended. The methods were tested on a simple noise reduction application in brain imaging and we are working on more applications, such as substitute CT generation and land-use classification. We focused on a particular model here that is suitable for modeling of massive data sets on regular grids, but it would also be interesting to test the model for more typical geostatistical problems in continuous space. This would not require much work though the particular MRF model for the allocation process would have to be modified.

The proposed estimation method is not only useful for the LGFM models but also for regular MRF mixture models and latent Gaussian models. We have not shown any theoretical properties of the estimator here and to the authors knowledge, there are no applicable results available to show consistency of the estimator for the proposed model class. Comets and Gidas [1992] showed consistency for the maximum likelihood estimator for the MRF mixture models, but the consistency of the maximum likelihood estimator for the LGFM models, the pseudo likelihood estimators for the MRF mixture models, and the pseudo likelihood estimators for the LGFM models are to the authors knowledge unknown, and certainly something for further research.

Finally, the basic estimation method is straightforward to implement. However, we used several sophisticated techniques to reduce the computational cost of the estimation, which increases the complexity of the implementation. We are therefore working on a software package that implements these methods and will simplify their practical usage.

Acknowledgements

The authors are grateful to Adam Johansson for providing the MR data used in Section 5.

A MRF gradients

Let $l = \log\left(PL(\mathbf{\Psi}; \mathbf{z}, \mathbf{Y})\right)$, where PL is the pseudo likelihood of the MRF mixture model (10), the derivatives in the gradient are then given by

$$\begin{aligned}
\frac{\partial\, l}{\partial\, \alpha_k} &= \sum_i z_{ik} - \sum_i \frac{\exp(\alpha_k + \beta_k f_{ik})}{\sum_l \exp(\alpha_l + \beta_l f_{il})} \\
\frac{\partial\, l}{\partial\, \beta_k} &= \sum_i z_{ik} f_{ik} - \sum_i \frac{\exp(\alpha_k + \beta_k f_{ik}) f_{ik}}{\sum_l \exp(\alpha_l + \beta_l f_{il})} \\
\nabla_{\boldsymbol{\mu}_k} l &= \sum_i z_{ik} \mathbf{Q}_{dk} (\mathbf{Y}_i - \boldsymbol{\mu}_k) \\
\frac{\partial\, l}{\partial\, \eta_{kj}} &= \sum_i z_{ik} \left(\mathbb{I}_{diag} - \frac{1}{2} \left(Y_i - \boldsymbol{\mu}_k\right)^\top \frac{\partial\, \mathbf{Q}_{dk}}{\partial\, \eta_{kj}} \left(Y_i - \boldsymbol{\mu}_k\right) \right).
\end{aligned}$$

Here $\mathbb{I}_{diag}$ is one if η_{kj} is an element on the main diagonal of $\mathbf{R}_{dk}$ and zero otherwise. We have

$$\frac{\partial\, \mathbf{Q}_{dk}}{\partial\, \eta_{kj}} = \frac{\partial\, \mathbf{R}_{dk}^\top}{\partial\, \eta_{ki}} \mathbf{R}_{dk} + \mathbf{R}_{dk}^\top \frac{\partial\, \mathbf{R}_{dk}}{\partial\, \eta_{ki}} \tag{12}$$

where the derivative $\frac{\partial\, \mathbf{R}_{dk}}{\partial\, \eta_{ki}}$ is a matrix with all elements zero except the element that corresponds to $\boldsymbol{\eta}_{ki}$. These expressions can be obtained with almost no extra cost while running the Gibbs sampler to sample $\mathbf{z}$. The derivatives needed to evaluate the scaling matrix $\mathbf{S}$ are

$$\begin{aligned}
\frac{\partial^2 l}{\partial\, \alpha_{k_1} \partial\, \alpha_{k_2}} &= \sum_i \left(-\mathbb{I}_{k_1 = k_2} \frac{\exp(\alpha_{k_1} + \beta_k f_{ik_1})}{\sum_l \exp(\alpha_l + \beta_l f_{il})} + \frac{\exp(\alpha_{k_1} + \beta_{k_1} f_{ik_1}) \exp(\alpha_{k_2} + \beta_{k_2} f_{ik_2})}{(\sum_l \exp(\alpha_l + \beta_l f_{il}))^2} \right) \\
\frac{\partial^2 l}{\partial\, \beta_{k_1} \partial\, \beta_{k_2}} &= \sum_i \left(-\mathbb{I}_{k_1 = k_2} \frac{\exp(\alpha_{k_1} + \beta_k f_{ik_1}) f_{ik_1}^2}{\sum_l \exp(\alpha_l + \beta_l f_{il})} + \frac{\exp(\alpha_{k_1} + \beta_{k_1} f_{ik_1}) \exp(\alpha_{k_2} + \beta_{k_2} f_{ik_2}) f_{ik_1} f_{ik_2}}{(\sum_l \exp(\alpha_l + \beta_l f_{il}))^2} \right) \\
\frac{\partial^2 l}{\partial\, \alpha_{k_1} \partial\, \beta_{k_2}} &= \sum_i \left(-\mathbb{I}_{k_1 = k_2} \frac{\exp(\alpha_{k_1} + \beta_k f_{ik_1}) f_{ik_1}}{\sum_l \exp(\alpha_l + \beta_l f_{il})} + \frac{\exp(\alpha_{k_1} + \beta_{k_1} f_{ik_1}) \exp(\alpha_{k_2} + \beta_{k_2} f_{ik_2}) f_{ik_1}}{(\sum_l \exp(\alpha_l + \beta_l f_{il}))^2} \right)
\end{aligned}$$

where $\mathbb{I}_{k_1 = k_2}$ controls that that factor is only included when $k_1 = k_2$. We also need the the derivatives of the parameters for the independent Gaussian distributions:

$$\begin{aligned}
\Delta_{\boldsymbol{\mu}_k} l &= -\sum_i z_{ik} \mathbf{Q}_{dk} \\
\frac{\partial^2 l}{\partial\, \eta_{kj_1} \partial\, \eta_{kj_2}} &= -\frac{1}{2} \sum_i z_{ik} \left(\mathbf{Y}_i - \boldsymbol{\mu}_k\right)^\top \frac{\partial^2\, \mathbf{Q}_{dk}}{\partial\, \eta_{jj_1} \partial\, \eta_{kj_2}} \left(\mathbf{Y}_i - \boldsymbol{\mu}_k\right) \\
\frac{\partial}{\partial\, \eta_{kj}} \nabla_{\boldsymbol{\mu}_k} l &= \sum_i z_{ik} \frac{\partial\, \mathbf{Q}_{dk}}{\partial\, \eta_{kj}} (\mathbf{Y}_i - \boldsymbol{\mu}_k)
\end{aligned}$$

where

$$\frac{\partial^2\, \mathbf{Q}_{dk}}{\partial\, \eta_{jj_1} \partial\, \eta_{kj_2}} = \mathbb{I}_{j_1 = j_2} \mathbb{I}_{diag} \frac{\partial\, \mathbf{Q}_{dk}}{\partial\, \eta_{kj}} + \frac{\partial\, \mathbf{R}_{dk}^\top}{\partial\, \eta_{kj_1}} \frac{\partial\, \mathbf{R}_{dk}}{\partial\, \eta_{kj_2}} + \frac{\partial\, \mathbf{R}_{dk}^\top}{\partial\, \eta_{kj_2}} \frac{\partial\, \mathbf{R}_{dk}}{\partial\, \eta_{kj_2}} \tag{13}$$

Except for the derivatives with respect to $\boldsymbol{\mu}_k$, all these derivatives are also need for the estimation of the LGFM model.

B Gaussian gradients

Let $l = \log \pi(\mathbf{Y}, \boldsymbol{\xi}|\boldsymbol{\Psi})$, the expectation of the derivatives then needed to evaluate the gradients of the Gaussian likelihood ($\log \pi(\mathbf{Y}|\boldsymbol{\Psi})$) are

$$\begin{aligned}
\mathsf{E}\left[\frac{\partial l}{\partial \kappa_0}|\mathbf{Y}, \boldsymbol{\Psi}\right] &= nd(D/4-1) + de^{\kappa_0} \operatorname{tr}\left(\mathbf{K}^{-1}\mathbf{C}\right) + \hat{\boldsymbol{\xi}}^\top \mathbf{Q}_d \otimes \tilde{\mathbf{Q}}_s \hat{\boldsymbol{\xi}} + \operatorname{tr}\left(\mathbf{Q}_d \otimes \tilde{\mathbf{Q}}_s \hat{\mathbf{Q}}^{-1}\right), \\
\mathsf{E}\left[\frac{\partial l}{\partial \eta_j}|\mathbf{Y}, \boldsymbol{\Psi}\right] &= \mathbb{I}_{diag} - \frac{1}{2}\hat{\boldsymbol{\xi}}^\top \frac{\partial \mathbf{Q}_d}{\partial \eta_j} \otimes \mathbf{Q}_s \hat{\boldsymbol{\xi}} - \frac{1}{2}\operatorname{tr}\left(\frac{\partial \mathbf{Q}_d}{\partial \eta_j} \otimes \mathbf{Q}_s \hat{\mathbf{Q}}^{-1}\right), \\
\mathsf{E}\left[\frac{\partial}{\partial \boldsymbol{\beta}} l|\mathbf{Y}, \boldsymbol{\Psi}\right] &= e^{-2\sigma_0}\mathbf{B}^\top\left(\mathbf{Y} - \mathbf{B}\boldsymbol{\beta} - \mathbf{A}\hat{\boldsymbol{\xi}}\right), \\
\mathsf{E}\left[\frac{\partial}{\partial \sigma_0} l|\mathbf{Y}, \boldsymbol{\Psi}\right] &= -m + e^{-2\sigma_0}\left(\left(\mathbf{Y} - \mathbf{B}\boldsymbol{\beta} - \mathbf{A}\hat{\boldsymbol{\xi}}\right)^\top\left(\mathbf{Y} - \mathbf{B}\boldsymbol{\beta} - \mathbf{A}\hat{\boldsymbol{\xi}}\right) + \operatorname{tr}\left(\mathbf{A}^\top\mathbf{A}\hat{\mathbf{Q}}^{-1}\right)\right).
\end{aligned}$$

Here, $\frac{\partial \mathbf{Q}_d}{\partial \eta_j}$ is given by (12), $\hat{\boldsymbol{\xi}}$ is the expected value of $\boldsymbol{\xi}$ given the current parameter estimates, $\hat{\boldsymbol{\xi}} = \hat{\mathbf{Q}}^{-1}\frac{1}{\sigma^2}\mathbf{A}^\top(\mathbf{Y} - \mathbf{B}\boldsymbol{\beta})$, and $\tilde{\mathbf{Q}}_s = (1 - D/4)\mathbf{Q}_s - ce^{\kappa_0}\mathbf{K}$. For the scaling $\mathbf{S}$, we also need expectation of the second derivatives:

$$\begin{aligned}
\mathsf{E}\left[\frac{\partial^2}{\partial \kappa_0^2} l|\mathbf{Y}, \boldsymbol{\Psi}\right] &= e^{\kappa_0} d \operatorname{tr}\left(\mathbf{K}^{-1}\mathbf{C}\right) + e^{2\kappa_0} d \operatorname{tr}\left(\mathbf{Q}_s^{-1}\mathbf{C}\right) + \hat{\boldsymbol{\xi}}^\top \mathbf{Q}_d \otimes \frac{\partial \tilde{\mathbf{Q}}_s}{\partial \kappa_0}\hat{\boldsymbol{\xi}} + \operatorname{tr}\left(\mathbf{Q}_d \otimes \frac{\partial \tilde{\mathbf{Q}}_s}{\partial \kappa_0}\hat{\mathbf{Q}}^{-1}\right), \\
\mathsf{E}\left[\frac{\partial^2}{\partial \eta_i \partial \eta_j} l|\mathbf{Y}, \boldsymbol{\Psi}\right] &= -\frac{1}{2}\hat{\boldsymbol{\xi}}^\top \frac{\partial^2 \mathbf{Q}_d}{\partial \eta_i \partial \eta_j} \otimes \mathbf{Q}_s \hat{\boldsymbol{\xi}} - \frac{1}{2}\operatorname{tr}\left(\frac{\partial^2 \mathbf{Q}_d}{\partial \eta_i \partial \eta_j} \otimes \mathbf{Q}_s \hat{\mathbf{Q}}^{-1}\right) \\
\mathsf{E}\left[\frac{\partial^2}{\partial \kappa_0 \partial \eta_j} l|\mathbf{Y}, \boldsymbol{\Psi}\right] &= \hat{\boldsymbol{\xi}}^\top \frac{\partial \mathbf{Q}_d}{\partial \eta_j} \otimes \tilde{\mathbf{Q}}_s \hat{\boldsymbol{\xi}} + \operatorname{tr}\left(\frac{\partial \mathbf{Q}_d}{\partial \theta_j} \otimes \tilde{\mathbf{Q}}_s \hat{\mathbf{Q}}^{-1}\right) \\
\mathsf{E}\left[\frac{\partial^2}{\partial \boldsymbol{\beta}^2} l|\mathbf{Y}, \boldsymbol{\Psi}\right] &= -e^{-2\sigma_0}\mathbf{B}^\top\mathbf{B} \\
\mathsf{E}\left[\frac{\partial^2}{\partial \sigma^2} l|\mathbf{Y}, \boldsymbol{\Psi}\right] &= -2e^{-2\sigma_0}\left(\left(\mathbf{Y} - \mathbf{B}\boldsymbol{\beta} - \mathbf{A}\hat{\boldsymbol{\xi}}\right)^\top\left(\mathbf{Y} - \mathbf{B}\boldsymbol{\beta} - \mathbf{A}\hat{\boldsymbol{\xi}}\right) + \operatorname{tr}\left(\mathbf{A}^\top\mathbf{A}\hat{\mathbf{Q}}^{-1}\right)\right) \\
\mathsf{E}\left[\frac{\partial^2}{\partial \beta \partial \sigma_0} l|\mathbf{Y}, \boldsymbol{\Psi}\right] &= -2\frac{\partial}{\partial \boldsymbol{\beta}} l.
\end{aligned}$$

Here $\frac{\partial^2 \mathbf{Q}_d}{\partial \eta_i \partial \eta_j}$ is given by (13) and

$$\frac{\partial \tilde{\mathbf{Q}}_s}{\partial \kappa_0} = -\frac{(D-4)^2}{8}\mathbf{Q}_s + c(3-D)e^{\kappa_0}\mathbf{K} - ce^{2\kappa_0}\mathbf{C}.$$

References

MA Ajiz and Alan Jennings. A robust incomplete choleski-conjugate gradient algorithm. *International Journal for Numerical Methods in Engineering*, 20(5):949–966, 1984.

Mihai Anitescu, Jie Chen, and Lei Wang. A matrix-free approach for solving the parametric gaussian process maximum likelihood problem. *SIAM Journal on Scientific Computing*, 34(1):A240–A262, 2012.

Erlend Aune, Daniel P Simpson, and Jo Eidsvik. Parameter estimation in high dimensional gaussian distributions. *Statistics and Computing*, pages 1–17, 2012.

J. Besag. Spatial interaction and the statistical analysis of lattice systems (with discussion). *J. Roy. Statist. Soc. Ser. B Stat. Methodol.*, 36:192–225, 1974.

David Bolin and Finn Lindgren. Spatial models generated by nested stochastic partial differential equations, with an application to global ozone mapping. *Ann. Appl. Statist.*, 5(1):523–550, 2011.

J.-P. Chiles and P. Delfiner. *Geostatistics, Modeling Spatial uncertainty*. Wiley Series in Probability and statistics, 1999.

Francis Comets and Basilis Gidas. Parameter estimation for gibbs distributions from partially observed data. *Ann. Appl. Probab.*, 2(1):142–170, 02 1992.

N. Cressie. *Statistics for Spatial Data*. John Wiley & Sons Ltd, New York, NY, USA, 1991.

N. Cressie and C.K. Wikle. *Statistics for Spatio-Temporal Data*. Wiley Series in Probability and Statistics. Wiley, Hoboken, New Jersey, 2011.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B Stat. Methodol.*, 39(1):1–38, 1977.

Carmen Fernández and Peter J. Green. Modelling spatially correlated data via mixtures: a bayesian approach. *J. Roy. Statist. Soc. Ser. B Stat. Methodol.*, 64(4):805–826, 2002. ISSN 1467-9868.

X. Guyon. *Random Fields on a Network: Modeling, Statistics, and Applications*. Graduate Texts in Mathematics. Springer, 1995. ISBN 9780387944289.

Karsten Held, E Rota Kops, Bernd J Krause, William M Wells III, Ron Kikinis, and H-W Muller-Gartner. Markov random field segmentation of brain mr images. *Medical Imaging, IEEE Transactions on*, 16(6):878–886, 1997.

Xiangping Hu, Daniel Simpson Lindgren, Håvard Rue, et al. Multivariate gaussian random fields using systems of stochastic partial differential equations. *arXiv preprint arXiv:1307.1379*, 2013.

MF Hutchinson. A stochastic estimator of the trace of the influence matrix for laplacian smoothing splines. *Communications in Statistics-Simulation and Computation*, 19(2):433–450, 1990.

Adam Johansson, Mikael Karlsson, and Tufve Nyholm. CT substitute derived from MRI sequences with ultrashort echo time. *Med. Phys.*, 38:2708, 2011.

Kenneth Lange. A gradient algorithm locally equivalent to the EM algorithm. *J. Roy. Statist. Soc. Ser. B Stat. Methodol.*, pages 425–437, 1995.

Finn Lindgren, Håvard Rue, and Johan Lindström. An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach (with discussion). *J. Roy. Statist. Soc. Ser. B Stat. Methodol.*, 73:423–498, 2011.

George Papandreou and Alan L Yuille. Efficient variational inference in large-scale Bayesian compressed sensing. In *IEEE Workshop on Information Theory in Computer Vision and Pattern Recognition*, pages 1332–1339. IEEE, 2011.

Douglas A Reynolds and Richard C Rose. Robust text-independent speaker identification using Gaussian mixture speaker models. *IEEE Transactions on Speech and Audio Processing*, 3(1):72–83, 1995.

Alexander Shapiro, Darinka Dentcheva, and Andrzej P Ruszczyński. *Lectures on stochastic programming: modeling and theory*, volume 9. SIAM, 2009.

Chris Stauffer and W Eric L Grimson. Adaptive background mixture models for real-time tracking. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2. IEEE, 1999.

M. L. Stein. *Interpolation of Spatial Data: Some Theory for Kriging.* Springer-Verlag, New York, 1999.

Michael L. Stein, Jie Chen, and Mihai Anitescu. Stochastic approximation of score functions for gaussian processes. *Ann. Appl. Statist.*, 7(2):1162–1191, 06 2013.

Koen Van Leemput, Frederik Maes, Dirk Vandermeulen, and Paul Suetens. Automated model-based tissue classification of MR images of the brain. *IEEE Transactions on Medical Imaging*, 18(10): 897–908, 1999.

Williams M Wells III, W Eric L Grimson, Ron Kikinis, and Ferenc A Jolesz. Adaptive segmentation of MRI data. *IEEE Transactions on Medical Imaging*, 15(4):429–442, 1996.

Hao Zhang. Maximum-likelihood estimation for multivariate spatial linear coregionalization models. *Environmetrics*, 18(2):125–139, 2007.

Yongyue Zhang, Michael Brady, and Stephen Smith. Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. *IEEE Transactions on Medical Imaging*, 20(1):45–57, 2001.