ARTICLE IN PRESS

Computational Statistics and Data Analysis xx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Latent Gaussian random field mixture models

David Bolin [a], Jonas Wallin [b,*], Finn Lindgren [c]

[a] Chalmers University of Technology and the University of Gothenburg, Gothenburg, Sweden
[b] Lund University, Tycho Brahes väg 1, 220 07 Lund, Sweden
[c] University of Edinburgh, Edinburgh, UK

ARTICLE INFO

Article history:
Received 9 November 2017
Received in revised form 5 August 2018
Accepted 10 August 2018
Available online xxxx

Keywords:
Random field
Spatial statistics
Gaussian mixture
Stochastic gradient
Geostatistics
Gaussian process

ABSTRACT

For many problems in geostatistics, land cover classification, and brain imaging the classical Gaussian process models are unsuitable due to sudden, discontinuous, changes in the data. To handle data of this type, we introduce a new model class that combines discrete Markov random fields (MRFs) with Gaussian Markov random fields. The model is defined as a mixture of several, possibly multivariate, Gaussian Markov random fields. For each spatial location, the discrete MRF determines which of the Gaussian fields in the mixture that is observed. This allows for the desired discontinuous changes of the latent processes, and also gives a probabilistic representation of where the changes occur spatially. By combining stochastic gradient minimization with sparse matrix techniques we obtain computationally efficient methods for both likelihood-based parameter estimation and spatial interpolation. The model is compared to Gaussian models and standard MRF models using simulated data and in application to upscaling of soil permeability data.

1. Introduction

In spatial statistics, data are often linked to spatially varying discrete covariates such as land cover categories in vegetation models (Bolin et al., 2009), geology in soil permeability models (Kim et al., 2005), or brain tissue type in brain imaging applications (Hildeman et al., 2017b). These covariates cause discontinuities in the data that easily can be accounted for *if* one has access to the covariates. Unfortunately, these covariates are often unknown. In this scenario a standard Gaussian random field model is not suitable, due to its inability of handling discontinuities. Here we introduce a class of models, based on a combination of mixture models and Gaussian random fields, to handle this type of data.

One way to analyze data with missing discrete covariates is to first classify the data into the different distinct spatial regions and then model each region separately. In many applications ranging from video surveillance to speaker identification and image analysis (Reynolds and Rose, 1995; Stauffer and Grimson, 1999), Gaussian mixture models (GMMs) are used for the classification problem. A GMM assumes independence between the observations and that the distribution of each observation is $\pi(\mathbf{y}) = \sum_{k=1}^{K} w_k \pi_k(\mathbf{y})$, where K is the number of classes, w_k is the probability of class k, and π_k a multivariate normal density. The assumption of independence between the observations is a clear drawback with GMM-based classification for spatial data. A strategy to account for spatial dependency is to allow for dependency in the allocation variables, which can be done in several ways. One way is to model the class probabilities using a logistic regression model based on Gaussian fields (Fernández and Green, 2002). Another way is to note that a random variable $\mathbf{Y}$ with a GMM distribution can be written as $\mathbf{Y} = \sum_{k=1}^{K} z_k \mathbf{X}_k$. Here $\mathbf{X}_k$ is a Gaussian random variable with density π_k, and $z_k = \mathbb{I}(\tilde{z} = k)$ where $\tilde{z}$ is a discrete random variable with $\mathsf{P}(\tilde{z} = k) = w_k$. Spatial dependency can be introduced by modeling the collection

* Corresponding author.
E-mail address: jonas.wallin@stat.lu.se (J. Wallin).

https://doi.org/10.1016/j.csda.2018.08.007

of the random variables $\tilde{z}$ for all observations as a discrete Markov random field (MRF) (see e.g. Held et al., 1997; Zhang et al., 2001; Van Leemput et al., 1999), which we refer to as an MRF mixture model.

Allowing for spatial dependency in the mixture weights often improves the classification for spatial problems. Yet, it is often not sufficient since it cannot capture the dependence between observations within each class. To account for this, we replace the independent Gaussian variables $\mathbf{X}_k$ for each class by a spatially dependent Gaussian random field (see e.g. Cressie, 1991; Cressie and Wikle, 2011). This allows us to use the model for classification, but also for noise reduction and spatial interpolation in cases where the data consist of noisy partial observations of fields with discontinuities. We refer to models of this type, which are introduced in more detail in Section 2, as latent Gaussian random field mixture (LGFM) models.

The proposed model could be viewed as a non-stationary Gaussian random field, with a specific prior on spatially varying parameters. There is an extensive literature on non-stationary Gaussian fields, see for example Paciorek and Schervish (2006), Fuglstad et al. (2015), Higdon (2001) and Bolin and Lindgren (2011). A non-stationary Gaussian field that resembles the LGFM model is that of Fuentes and Smith (2001), where a process is created as a spatially varying average of stationary Gaussian processes. Other similar modeling approaches are those of Kim et al. (2005), where a tessellation of the spatial domain is used to define a mixture process, and the Bayesian treed Gaussian process models by Gramacy and Lee (2008). However, all these methods either lack the sharp and flexible discontinuities, or the computational efficiency, of the LGFM model.

Since spatial problems often have massive amounts of data, a computationally efficient estimation method is needed in order to fit the LGFM model to data. Further, likelihood estimation for discrete MRFs is problematic due intractable normalizing constants. Two common methods for dealing with this issue are gradient-based minimization and pseudo-likelihood methods (Guyon, 1995; Hildeman et al., 2017b). Recently, gradient-based methods have also been developed for large-scale Gaussian random field models (Anitescu et al., 2012; Stein et al., 2013). We combine these two approaches into a computationally efficient estimation method for LGFM models. The method is a stochastic version of the EM gradient method (Lange, 1995), and is introduced further in Section 3. The model is tested on two simulated data sets in Section 4, and on an application to upscaling soil permeability data in Section 5. Finally, Section 6 contains a discussion of possible extensions and further work. The code used to obtain the results in the article is available at https://bitbucket.org/davidbolin/lgfm/.

2. Latent Gaussian random field mixture models

Let $\mathbf{Y}_1, \ldots, \mathbf{Y}_M$ be d-dimensional observations at locations $\mathbf{s}_1, \ldots, \mathbf{s}_M$ on a regular lattice with n nodes. The structure of a LGFM model for this data is

$$\begin{aligned} \mathbf{X}_k(\mathbf{s}) &= \mathbf{B}_k(\mathbf{s})\boldsymbol{\beta}_k + \boldsymbol{\xi}_k(\mathbf{s}), \quad k = 1, \ldots, K, \\ \mathbf{X}(\mathbf{s}) &= \sum_{k=1}^{K} z_k(\mathbf{s})\mathbf{X}_k(\mathbf{s}), \\ \mathbf{Y}_m &= \mathbf{X}(\mathbf{s}_m) + \boldsymbol{\varepsilon}_m, \quad m = 1, \ldots, M. \end{aligned} \tag{1}$$

Here $\boldsymbol{\varepsilon}_i$ are independent $\mathsf{N}(\mathbf{0}, \boldsymbol{\Sigma}_\varepsilon)$ random variables representing measurement noise for each dimension, with $\boldsymbol{\Sigma}_\varepsilon = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_d^2)$, and the latent process $\mathbf{X}(\mathbf{s})$ is modeled as a mixture of K independent Gaussian random fields $\mathbf{X}_k(\mathbf{s})$. These Gaussian fields are specified using the mean-zero Gaussian fields $\boldsymbol{\xi}_k(\mathbf{s})$ as well as regressions $\mathbf{B}_k(\mathbf{s})\boldsymbol{\beta}_k = \sum_{p=1}^{P}\mathbf{B}_{kp}(\mathbf{s})\beta_{kp}$ on fixed-effects $\mathbf{B}_{kp}(\mathbf{s})$ for the mean values. Finally, $z_k(\mathbf{s}) = \mathbb{I}(\tilde{z}(\mathbf{s}) = k)$ where $\tilde{z}(\mathbf{s})$ is a discrete MRF. In the following two sections, we introduce the statistical models for the discrete MRF and the Gaussian fields $\boldsymbol{\xi}_k(\mathbf{s})$ in more detail, and then discuss properties of the model.

2.1. A model for the Gaussian fields $\boldsymbol{\xi}_k(\mathbf{s})$

In the case of multivariate data, we assume that the Gaussian fields $\boldsymbol{\xi}_k(\mathbf{s})$ have proportional correlation models (Chiles and Delfiner, 1999), which means that their covariance functions can be written as $C(\boldsymbol{\xi}_k(\mathbf{s}_1), \boldsymbol{\xi}_k(\mathbf{s}_2)) = \boldsymbol{\Sigma}_k\rho_k(\|\mathbf{s}_1 - \mathbf{s}_2\|)$, where $\boldsymbol{\Sigma}_k$ is a $d \times d$ covariance matrix and $\rho_k(\cdot)$ is a spatial correlation function. The reason for this particular choice is that it makes the model a natural extension of the regular Gaussian mixture models, which have covariances $C(\boldsymbol{\xi}_k(\mathbf{s}_1), \boldsymbol{\xi}_k(\mathbf{s}_2)) = \boldsymbol{\Sigma}_k\delta_{\mathbf{0}}(\mathbf{s}_1 - \mathbf{s}_2)$, where $\delta_{\mathbf{0}}$ is a regular Dirac distribution.

What remains is to decide on a model for the spatial correlation function. A popular choice is the Matérn correlation function, $\rho(\mathbf{h}) = 2^{1-\nu}\Gamma(\nu)^{-1}(\kappa\|\mathbf{h}\|)^\nu K_\nu(\kappa\|\mathbf{h}\|)$, where Γ is the gamma function and K_ν is a modified Bessel function of the second kind. The positive parameters κ and ν determine the practical correlation range and the differentiability of the process, respectively. An advantage with this covariance function is that one then can use the stochastic partial differential equation (SPDE) connection (Lindgren et al., 2011) between Gaussian Matérn fields and Gaussian Markov random field models (Besag, 1974) to construct a model for $\boldsymbol{\xi}_k(\mathbf{s})$ that has important computational advantages.

Since we assume that the data are on a lattice, we do not need the full generality of the SPDE approach. We can instead use that a conditional autoregressive model of order $p \in \mathbb{N}$, a CAR(p) model, on a lattice in $\mathbb{R}^2$ can be viewed as an approximation of a Gaussian field with a Matérn covariance function with $\nu = p - 1$. The CAR(1) model (which could be viewed as an