LUND UNIVERSITY

Models and Methods for Random Fields in Spatial Statistics with Computational Efficiency from Markov Properties

Bolin, David

2012

Link to publication

Citation for published version (APA):
Bolin, D. (2012). *Models and Methods for Random Fields in Spatial Statistics with Computational Efficiency from Markov Properties*. [Doctoral Thesis (compilation), Mathematical Statistics]. Faculty of Engineering, Centre for Mathematical Sciences, Mathematical Statistics, Lund University.

Total number of authors:
1

General rights
Unless other specific re-use rights are stated the following general rights apply:
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.
• Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying the publication in the public portal

Read more about Creative commons licenses: https://creativecommons.org/licenses/

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

LUND UNIVERSITY

PO Box 117
221 00 Lund
+46 46-222 00 00

Models and Methods for Random Fields in Spatial Statistics

with Computational Efficiency from Markov Properties

David Bolin

Lund University

Faculty of Engineering
Centre for Mathematical Sciences
Mathematical Statistics

Mathematical Statistics
Centre for Mathematical Sciences
Lund University
Box 118
SE-221 00 Lund
Sweden

http://www.maths.lth.se/

Doctoral Theses in Mathematical Sciences 2012:2
ISSN 1404-0034

ISBN 978-91-7473-336-5
LUTFMS-1040-2012

© David Bolin, 2012

Printed in Sweden by Media-Tryck AB, Lund 2012

Abstract

The focus of this work is on the development of new random field models and methods suitable for the analysis of large environmental data sets.

A large part is devoted to a number of extensions to the newly proposed Stochastic Partial Differential Equation (SPDE) approach for representing Gaussian fields using Gaussian Markov Random Fields (GMRFs). The method is based on that Gaussian Matérn field can be viewed as solutions to a certain SPDE, and is useful for large spatial problems where traditional methods are too computationally intensive to use. A variation of the method using wavelet basis functions is proposed and using a simulation-based study, the wavelet approximations are compared with two of the most popular methods for efficient approximations of Gaussian fields. A new class of spatial models, including the Gaussian Matérn fields and a wide family of fields with oscillating covariance functions, is also constructed using nested SPDEs. The SPDE method is extended to this model class and it is shown that all desirable properties are preserved, such as computational efficiency, applicability to data on general smooth manifolds, and simple non-stationary extensions. Finally, the SPDE method is extended to a larger class of non-Gaussian random fields with Matérn covariance functions, including certain Laplace Moving Average (LMA) models. In particular it is shown how the SPDE formulation can be used to obtain an efficient simulation method and an accurate parameter estimation technique for a LMA model.

A method for estimating spatially dependent temporal trends is also developed. The method is based on using a space-varying regression model, accounting for spatial dependency in the data, and it is used to analyze temporal trends in vegetation data from the African Sahel in order to find regions that have experienced significant changes in the vegetation cover over the studied time period. The problem of estimating such regions is investigated further in the final part of the thesis where a method for estimating excursion sets, and the related problem of finding uncertainty regions for contour curves, for latent Gaussian fields is proposed. The method is based on using a parametric family for the excursion sets in combination with Integrated Nested Laplace Approximations (INLA) and an importance sampling-based algorithm for estimating joint probabilities.

Contents

Abstract **i**

List of papers **v**

Acknowledgements **vii**

Introduction and summary **1**

1 Background 1
2 Modeling spatial data 3
3 Hierarchical models, inference, and kriging 7
4 Lattice data and Gaussian Markov random fields 10
5 Estimating excursion sets for latent Gaussian models 16
6 Efficient representations of continuous Gaussian fields 18
7 The SPDE approach 22
8 Extensions of the SPDE method 24
9 Comments on the papers 31

A Fast estimation of spatially dependent temporal vegetation trends using Gaussian Markov random fields **41**

1 Introduction 41
2 Statistical model 43
3 EM parameter estimation 47
4 Testing for significant trends 50
5 Data 51
6 Results 53
7 Extensions 58
8 Concluding remarks 59
A Proof of Proposition 2.2 61

B How do Markov approximations compare with other methods for large spatial data sets? 69
1 Introduction 69
2 Spatial prediction and computational cost 72
3 Wavelet approximations 74
4 Comparison 82
5 Conclusions 96

C Spatial models generated by nested stochastic partial differential equations, with an application to global ozone mapping 103
1 Introduction 103
2 Stationary nested SPDE models 105
3 Computationally efficient representations 112
4 Parameter estimation 118
5 Application: Ozone data 122
6 Concluding remarks 131
A Vector spherical harmonics 132

D Spatial Matérn fields driven by non-Gaussian noise 139
1 Introduction 139
2 Gaussian Matérn fields 141
3 Non-Gaussian SPDE-based models 142
4 Hilbert space approximations 149
5 Sampling from the model 154
6 Parameter estimation 156
7 A simulation study 164
8 Discussion and extensions 166

E Excursion and contour uncertainty regions for latent Gaussian models 175
1 Introduction 175
2 Problem formulation 177
3 Computations 181
4 Tests on simulated data 193
5 Applications 200
6 Discussion 208
A Notes on the MCMC algorithm used in Example 3 209

List of papers

This thesis is based on the following papers, referred to in the text with the capital letters A, B, C, D, and E:

A David Bolin, Johan Lindström, Lars Eklundh and Finn Lindgren (2009). Fast Estimation of Spatially Dependent Temporal Vegetation Trends using Gaussian Markov Random Fields, *Computational Statistics and Data Analysis* 53, 2885-2896

B David Bolin and Finn Lindgren (2009). How do Markov approximations compare with other methods for large spatial data sets? (Submitted)

C David Bolin and Finn Lindgren (2011). Spatial models generated by nested stochastic partial differential equations, with an application to global ozone mapping, *Annals of Applied Statistics* 5:1, 523-550

D David Bolin (2011). Spatial Matérn fields driven by non-Gaussian noise, *Preprints in Mathematical Sciences* 2011:4 , Lund University (Submitted)

E David Bolin and Finn Lindgren (2012). Excursion and contour uncertainty regions for latent Gaussian models

Additional papers not included in the thesis:

1 Georg Lindgren, David Bolin, and Finn Lindgren (2010). Non-traditional stochastic models for ocean waves, *The European Physical Journal - Special Topics* 185:1, 209-224

Acknowledgements

First, I would like thank my co-advisor Finn Lindgren for all his great ideas, enthusiasm, and help along the way of making this thesis. I am also grateful to Johan Lindström for introducing me to the field of spatial statistics a long time ago and for all the help and many interesting discussions, especially in connection with our joint work on Paper A.

A few years ago, I visited the National Center for Atmospheric Research in Boulder, Colorado, and I wish to thank Doug Nychka for the brief meetings we had during that time, which helped me to come up with the ideas for Paper B and Paper C. My thanks also goes to Georg Lindgren for numerous comments on all parts of this thesis; to my advisor Krzysztof Podgórski and fellow Ph.D. student Jonas Wallin for interesting and helpful discussions in connection with my work on Paper D; and to Anders Widd for proofreading and giving comments on the introduction.

I would also like to thank all colleagues at the division of Mathematical Statistics at Lund University that have helped in one way or the other during the last five years, especially Aurelia, James, Joakim, and Mona for the help with various practical matters.

Finally, I am sincerely grateful to my friends and family for always being there.

Lund, April 2012 *David Bolin*

Financial support

This work was partially funded by the multidisciplinary research program FRIVA (Framework Programme for Risk and Vulnerability Analysis) financed by the Swedish Emergency Management Agency, and the Swedish Foundation for Strategic Research (SSF) under grant A3 02:125, Spatial statistics and image analysis for environment and medicine.

Introduction and summary

1 Background

Spatial statistics is the scientific discipline of statistical modeling and analysis of spatially structured phenomena. One of the key features in spatial statistics is the autocorrelation of data; observations at locations in close spatial proximity often tend to be more similar than observations at locations far apart[1]. One of the earliest works involving spatially correlated data was done by R.A. Fisher in the 1920's (Fisher, 1926). Fisher studied design-based inference of agricultural field trials, and noted that plots (rectangular experimental units in the field) close to each other were more similar than plots farther apart, which violated the assumption that the studied plots were mutually independent. Instead of modeling the spatial dependence, Fisher proposed using randomization and blocking of the plots so that larger blocks of plots where approximately independent.

Later, important work by Krige (1951) and Matheron (1963) laid the ground for the field of *geostatistics*; a hybrid field of statistics, mathematics, mining engineering, geology, and other subject matter areas. In their works, some of the first methods for modeling spatial dependence were proposed, many of which now are fundamental in spatial data analysis. Similar methods for modeling spatial dependence were independently derived by the Swedish forestry statistician Bertil Matérn (Matérn, 1960). The doctoral thesis by Bertil Matérn is one of the most important contributions to the field of spatial statistics, and especially it introduced a new class of spatial covariance functions, now bearing his name. The Matérn covariance function is to this day the most popular model of spatial dependence, and is the focus of a large part of this thesis.

Another important branch of spatial statistics is the study of discrete spatial variation, where the models are defined on discrete domains, such as regular grids or lattices. The most popular models in this area are the so-called simultaneous autoregressive (SAR) models, introduced by Whittle (1954), and the conditional

[1] Tobler's first law of geography: "Everything is related to everything else, but near things are more related than distant things." Tobler (1970)

autoregressive (CAR) models, first introduced in the seminal paper by Julian Besag (1974). These models have later been extended and there is now a large theory of Markov Random Fields (MRF) and Gaussian MRFs (GMRFs) (Rue and Held, 2005) which is frequently used in this thesis.

Parallel to the development of spatial data analysis, there has also been much progress in the theory of random fields (see for example Adler, 1981). In many of the classical applications of random field theory in environmental sciences, the cost for obtaining measurements limited the size of the data sets to ranges where computational cost was not an issue. Therefore, the methods were typically developed without any considerations of computational efficiency. Today, however, with the increasing use of remote sensing satellites, producing many large climate data sets, computational efficiency is often a crucial property. Thus, in many applications there is a need to balance the modeling desires with computational limits. This has created an entire new area within random field methods where the computational aspects are put at the center of interest.

It is this area where the common thread of the work constituting this thesis belongs. Hence, the thesis contents lies somewhere in between the applied topics in spatial data analysis and the theoretical studies of random fields, with the emphasis on efficiency of computational methods. A large part of the work is focused on the development of new random field models and methods, but all of these are created with applications in mind and are intended to be applicable to large environmental data sets. Because of this, there are a number of applications to spatial data analysis presented throughout the thesis which serve as examples to show that the developed method indeed can be used in practice.

1.1 Outline

The structure of this thesis summary is as follows. Section 2 gives an introduction to continuous spatial modeling and random fields. Section 3 introduces hierarchical models and the problem of spatial prediction. Section 4 introduces discrete spatial modeling and GMRFs. Computational aspects are also discussed in this section, and an application of a GMRF model taken from Paper A is presented. Section 5 discusses the problem of estimating excursion sets for latent Gaussian fields, which is the topic of Paper E. Section 6 presents some popular computationally efficient representations of continuous Gaussian fields, and Section 7 introduces the stochastic partial differential equation (SPDE) approach which is the main focus of this thesis. Extensions of the SPDE approach are given in Section 8

and, in particular, the nested SPDE approach of Paper C is discussed in Section 8.1 and the non-Gaussian extensions of Paper D are discussed in Section 8.2. Finally, Section 9 concludes with comments on the five appended papers.

2 Modeling spatial data

A statistical model can be seen as a mathematical abstraction of a data-generating mechanism; may it be a physical process, the financial market, or something else. For spatial phenomena, the model is usually a random field.

Definition 2.1. A random field (or stochastic field), $X(\mathbf{s}, \omega)$, $\mathbf{s} \in \mathcal{D}$, $\omega \in \Omega$, is a random function specified by its finite-dimensional joint distributions

$$F(y_1, \ldots, y_n; \mathbf{s}_1, \ldots, \mathbf{s}_n) = \mathsf{P}(X(\mathbf{s}_1) \leq y_1, \ldots, X(\mathbf{s}_n) \leq y_n)$$

for every finite n and every collection $\mathbf{s}_1, \ldots, \mathbf{s}_n$ of locations in $\mathcal{D}$.

The set $\mathcal{D}$ is usually a subset of $\mathbb{R}^d$, and for the special case $d = 1$, $X(\mathbf{s}, \omega)$ is called a random process (or stochastic process). At every location $\mathbf{s} \in \mathcal{D}$, $X(\mathbf{s}, \omega)$ is a random variable where the event ω lies in some abstract sample space Ω. It is important to ensure that the random field has a valid mathematical specification. In general, this is done using the Kolmogorov existence theorem which states that the collection of finite-dimensional distributions defines a valid random field if it is consistent under permutations and marginalization (see e.g. Billingsley, 1986, for details). To simplify the notation, one often writes $X(\mathbf{s})$, removing the dependency on ω from the notation.

An important special case is when the random field is Gaussian. The statistical properties of a Gaussian random field are completely specified by its mean value function, $\mu(\mathbf{s}) = \mathsf{E}(X(\mathbf{s}))$, and covariance function, $C(\mathbf{s}, \mathbf{t}) = \mathrm{Cov}(X(\mathbf{s}), X(\mathbf{t}))$. For existence of a Gaussian field with a prescribed mean and covariance it is enough to ensure that the latter is positive definite.

Definition 2.2. A function $C(\mathbf{s}, \mathbf{t})$ is positive definite if for any finite set of locations $\mathbf{s}_1, \ldots, \mathbf{s}_n$ in $\mathcal{D}$, the covariance matrix

$$\begin{pmatrix} C(\mathbf{s}_1, \mathbf{s}_1) & C(\mathbf{s}_1, \mathbf{s}_2) & \cdots & C(\mathbf{s}_1, \mathbf{s}_n) \\ C(\mathbf{s}_2, \mathbf{s}_1) & C(\mathbf{s}_2, \mathbf{s}_2) & \cdots & C(\mathbf{s}_2, \mathbf{s}_n) \\ \vdots & \vdots & \ddots & \vdots \\ C(\mathbf{s}_n, \mathbf{s}_1) & C(\mathbf{s}_n, \mathbf{s}_2) & \cdots & C(\mathbf{s}_n, \mathbf{s}_n) \end{pmatrix}$$

is non-negative definite.

The mean value function captures the large scale trends in the random field, and is often modeled using a regression basis of some known functions of the spatial coordinates. In accordance with Tobler's first law, spatial covariance functions often decrease with increasing spatial separation of the points $\mathbf{s}$ and $\mathbf{t}$. The exact form of the covariance function, however, has to be chosen to fit the data set at hand. A common simplifying assumption is that the random field is stationary.

Definition 2.3. A random field $X(\mathbf{s})$ is called *strictly stationary* if for any vector $\mathbf{h}$ and for every collection $\mathbf{s}_1, \ldots, \mathbf{s}_n$ of locations in $\mathcal{D}$

$$F(y_1, \ldots, y_n; \mathbf{s}_1 + \mathbf{h}, \ldots, \mathbf{s}_n + \mathbf{h}) = F(y_1, \ldots, y_n; \mathbf{s}_1, \ldots, \mathbf{s}_n).$$

Hence, a random field is strictly stationary if its finite dimensional distributions are shift invariant. In particular this implies that the mean value function and the covariance function are shift invariant if they exist.

Definition 2.4. A random field $X(\mathbf{s})$ is called *weakly stationary* if for any vector $\mathbf{h}$ and any locations $\mathbf{s}, \mathbf{t} \in \mathcal{D}$

$$\mu(\mathbf{s} + \mathbf{h}) = \mu(\mathbf{s}), \qquad \text{and} \qquad C(\mathbf{s} + \mathbf{h}, \mathbf{t} + \mathbf{h}) = C(\mathbf{s}, \mathbf{t}) = C(\mathbf{s} - \mathbf{t}).$$

If the variance, $\mathsf{V}(X(\mathbf{s}))$, of a strictly stationary field is finite, it follows that it is also weakly stationary. A weakly stationary field is in general not strictly stationary; however, if the random field is Gaussian and weakly stationary, it is also strictly stationary. Therefore, as there is no distinction between the two concepts of stationarity in the Gaussian case, one simply writes that the field is stationary.

An important subclass of the weakly stationary fields are the *isotropic* fields. These have covariance functions that depend only on distance, and not direction, between points, i.e. $C(\mathbf{s}_1, \mathbf{s}_2) = C(\|\mathbf{s}_1 - \mathbf{s}_2\|)$. Among the isotropic covariance functions, the Matérn covariance function (Matérn, 1960) is one of the most popular choices for modeling spatial data.

Definition 2.5. Let $K_\nu(x)$ denote the modified Bessel function of the second kind of order ν, and let $\nu, \kappa, \phi > 0$. The Matérn covariance function is then given by

$$C(\mathbf{h}) = \frac{2^{1-\nu}\phi^2}{(4\pi)^{\frac{d}{2}}\Gamma(\nu + \frac{d}{2})\kappa^{2\nu}}(\kappa\|\mathbf{h}\|)^\nu K_\nu(\kappa\|\mathbf{h}\|), \quad \mathbf{h} \in \mathbb{R}^d\,. \tag{1}$$

Here ν is a shape parameter for the covariance function, κ a spatial scale parameter, ϕ^2 a variance parameter, and Γ is the gamma function.

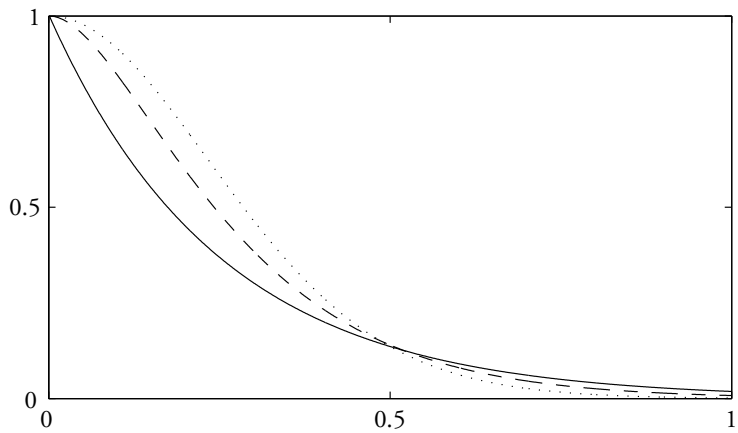

Figure 1: Matérn correlation functions shown for $\nu = 0.5$ (solid line), $\nu = 1.5$ (dashed line), and $\nu = 10$ (dotted line), with $\kappa = \sqrt{8\nu}/0.5$.

There are a few different parameterizations of the Matérn covariance function in the literature, but the one used in the definition above is most suitable in our context. With this parametrization, the variance of a field with this covariance is

$$C(\mathbf{0}) = \frac{\phi^2 \Gamma(\nu)}{(4\pi)^{\frac{d}{2}} \Gamma(\nu + \frac{d}{2}) \kappa^{2\nu}}.$$

In Figure 1, the Matérn covariance function is shown for three different values of ν. An important special case is the exponential covariance function; the solid line in the figure obtained when $\nu = 0.5$. The smoothness of the field increases with ν, and another important special case is the Gaussian covariance function, sometimes also called squared exponential, obtained in the limit as $\nu \to \infty$ if κ is scaled accordingly. For more properties of the Matérn covariance function, see e.g. Stein (1999).

Specifying a Gaussian random field through its covariance function is the most popular method in spatial statistics; however, there are other representations that, in some situations, are more convenient. One alternative is to specify the random field in the frequency domain. By Bochner's theorem (Bochner, 1955), a function C is a valid covariance function if and only if it can be written as

$$C(\mathbf{h}) = \int \exp(i\mathbf{h}^\top \mathbf{k}) \, \mathrm{d}\Lambda(\mathbf{k}) \tag{2}$$

for some non-negative and symmetric measure Λ. Equation (2) is called the spectral representation of the covariance function, and if the measure Λ has a Lebesgue density S, this is called the spectral density. The spectral density associated with the Matérn covariance function (1) is

$$S(\mathbf{k}) = \frac{\phi^2}{(2\pi)^d} \frac{1}{(\kappa^2 + \|\mathbf{k}\|^2)^{\nu+\frac{d}{2}}}.$$

Another popular representation, first proposed by Matheron (1971), that can be used instead of the covariance function is the *(semi)variogram* $\gamma(\mathbf{h})$, that for a stationary process is defined as

$$\gamma(\mathbf{h}) = \frac{1}{2}\mathsf{V}(X(\mathbf{s}+\mathbf{h}) - X(\mathbf{s})).$$

If a process has a constant mean value function (if it exists) and a variogram only depending on $\mathbf{h}$, and not on the location $\mathbf{s}$, it is called *intrinsically stationary*. If a random field is weakly stationary with covariance function $C(\mathbf{h})$, one has that $\gamma(\mathbf{h}) = C(\mathbf{0}) - C(\mathbf{h})$. Hence, all weakly stationary processes are intrinsically stationary. The converse is, however, not true and the class of intrinsically stationary processes is therefore larger than the class of weakly stationary processes. An example of an intrinsically stationary process that is not weakly stationary is the *Brownian motion*. The Brownian motion and its counterpart on $\mathbb{R}^d$, the *Brownian sheet*, can be defined using *white noise*. The following general definition of white noise is taken from Walsh (1986).

Definition 2.6. Let $(\mathcal{D}, \mathcal{M}, \lambda)$ be a σ-finite measure space. A Gaussian white noise based on λ is a random set function $\mathcal{W}$ on the sets $A \in \mathcal{M}$ of finite λ-measure such that $\mathcal{W}(A)$ is a $\mathsf{N}(0, \lambda(A))$ random variable and if $A \cap B = \emptyset$, then $\mathcal{W}(A)$ and $\mathcal{W}(B)$ are independent and $\mathcal{W}(A \cup B) = \mathcal{W}(A) + \mathcal{W}(B)$.

In most cases, $\mathcal{D}$ is an Euclidian space such as $\mathbb{R}^d$ and λ is the Lebesgue measure. The white noise is a Gaussian process defined on the sets in $\mathcal{M}$ with covariance function $\mathrm{Cov}(A, B) = \mathsf{E}(\mathcal{W}(A)\,\mathcal{W}(B)) = \lambda(A \cap B)$. It is straightforward to check that this covariance function is positive definite, and it is therefore a well-defined Gaussian process. There are other ways of defining white noise, and it is often thought of as the derivative of a Brownian sheet. If we set $\mathcal{D} = \mathbb{R}^d_+$ and take λ as the Lebesgue measure, the Brownian sheet is a process $\mathcal{B}(\mathbf{s})$, $\mathbf{s} \in \mathbb{R}^d_+$ defined by $\mathcal{B}(\mathbf{s}) = \mathcal{W}((\mathbf{0}, \mathbf{s}])$. If $\mathbf{s} = (s_1, \ldots, s_d)$ and $\mathbf{t} = (t_1, \ldots, t_d)$, its covariance

function is given by $\mathrm{Cov}(\mathcal{B}(\mathbf{s}), \mathcal{B}(\mathbf{t})) = \prod_{i=1}^{d} \min(s_i, t_i)$. Thus, the covariance function is not stationary and the Brownian sheet is therefore not weakly stationary. For $d = 1$, the Brownian motion is intrinsically stationary since $\gamma(h) = |h|$, but the Brownian sheet is not intrinsically stationary for $d > 1$.

Stationary and isotropic models are easy to work with, but may not be sufficient in certain applications. Introducing non-stationary mean value functions is in general not a problem since one can define the process as $X(\mathbf{s}) = \mu(\mathbf{s}) + Z(\mathbf{s})$, where $Z(\mathbf{s})$ is a stationary process. Introducing non-stationarity in the covariance function is, on the other hand, more problematic since it is difficult to extend stationary and isotropic covariance functions to non-stationary and non-isotropic versions while preserving the crucial property of positive definiteness. One alternative is the spatial deformation method by Sampson and Guttorp (1992) where a stationary covariance model is defined on a transformed domain $h(\mathcal{D})$ which results in a non-stationary covariance function on the original domain $\mathcal{D}$. Most other methods for specifying non-stationary models do not use the covariance representation directly but instead some other representation that induces a certain valid covariance function. We will get back to such alternative representations in Section 6.

3 Hierarchical models, inference, and kriging

Geostatistical measurements are often sampled under measurement noise, and a statistical model for the data thus has to take this into account. Another problem with geostatistical data is that the spatial domain, $\mathcal{D}$, often is a continuous region. Hence, even if one could measure the latent field $X(\mathbf{s})$ exactly, it cannot be sampled exhaustively. One of the most important problems in geostatistics is, therefore, spatial reconstruction of $X(\mathbf{s})$ given a finite number of observations $\mathbf{Y} = (Y_1, \ldots, Y_n)^\top$ of the latent field at locations $\mathbf{s}_1, \ldots, \mathbf{s}_n$ taken under measurement noise.

The most popular method for spatial reconstruction in geostatistics was developed by Georges Matheron. Matheron based his work on the Master's thesis of Daniel G. Krige, and therefore named the method kriging. Depending on the assumptions on the mean value function $\mu(\mathbf{s})$ for the latent field, linear kriging is usually divided into three cases. The method is called *Simple kriging* if $\mu(\mathbf{s})$ is known; *Ordinary kriging* if $\mu(\mathbf{s}) = \mu$ and μ is unknown; and *Universal kriging* if $\mu(\mathbf{s}) = \sum_{k=1}^{m} \beta_k b_k(\mathbf{s})$ where b_k are known basis functions and the parameters β_k

are unknown. The kriging estimator of $X(\mathbf{s})$ at some location $\mathbf{s}_0$ is derived as the minimum mean squared error linear predictor (for extensive details on kriging, see Stein, 1999, Schabenberger and Gotway, 2005). There is, however, a close connection between kriging and estimation in *hierarchical models* which we use.

A hierarchical model is constructed as a hierarchy of conditional probability models that, when multiplied together, yield the joint distribution for all quantities in the model. Let $\mathbf{X}$ be a vector containing $X(\mathbf{s})$ evaluated at the measurement locations and any additional locations where the kriging prediction should be calculated, and let $\boldsymbol{\gamma}$ be a vector containing all model parameters. At the top level of the hierarchical model is the *data model* $\pi(\mathbf{Y}|\mathbf{X}, \boldsymbol{\gamma})$, which specifies the distribution of the measurements given the latent process. Here $\pi(\cdot)$ denotes a density function. This level is sometimes also called the measurement equation since it specifies how the data is generated as a function of the latent process. A typical situation is when the latent field is measured under additive noise,

$$Y_i = X(\mathbf{s}_i) + \epsilon_i \,.$$

A common assumption is that $\epsilon_1, \ldots, \epsilon_n$ are independent identically distributed with some variance σ^2, uncorrelated with the latent process[2]. At the next level is the *process model* $\pi(\mathbf{X}|\boldsymbol{\gamma})$, which typically is given by the model for the continuous latent field of interest. The process model can in itself be written as a hierarchical model, specified by a number of conditional sub-models. The final part of the hierarchical model is the *parameter model* $\pi(\boldsymbol{\gamma})$ which is the joint prior distribution of the parameters. If a parameter model is used, the model is called a Bayesian hierarchical model. An alternative approach is to assume that the parameters are fixed but unknown, in a frequentist setting, and estimate the parameters from data. The model is then sometimes referred to as an empirical-Bayesian model, or empirical hierarchical model (Cressie and Wikle, 2011).

Inference in hierarchical models is performed using the *posterior distribution*

$$\pi(\mathbf{X}, \boldsymbol{\gamma}|\mathbf{Y}) \propto \pi(\mathbf{Y}|\mathbf{X}, \boldsymbol{\gamma})\pi(\mathbf{X}|\boldsymbol{\gamma})\pi(\boldsymbol{\gamma}).$$

Kriging predictions are calculated from the marginal posterior distribution

$$\pi(\mathbf{X}|\mathbf{Y}) \propto \int \pi(\mathbf{X}|\mathbf{Y}, \boldsymbol{\gamma})\pi(\boldsymbol{\gamma}|\mathbf{Y})\, \mathrm{d}\boldsymbol{\gamma},$$

[2] In geostatistics, the variance σ^2 is called the *nugget effect*, a somewhat curious name that has its origin in mining applications where nuggets literally exist and varies on scales so small that they cannot be distinguished from measurement noise.

and one typically reports the posterior mean $\mathsf{E}(\mathbf{X}|\mathbf{Y})$ as a point estimator and the posterior variance $\mathsf{V}(\mathbf{X}|\mathbf{Y})$ as a measure of the uncertainty in the predictor. The posterior distribution for $\mathbf{X}$ and $\boldsymbol{\gamma}$ generally have to be estimated using Markov Chain Monte Carlo (MCMC) methods (see e.g. Robert and Casella, 2004). A better alternative when the process model is Gaussian is to use the Integrated Nested Laplace Approximations (INLA) introduced by Rue et al. (2009), which allows for precise inference in a fraction of the computation time required by MCMC inference for latent Gaussian models.

In an empirical hierarchical model, inference is instead performed using the conditional posterior $\pi(\mathbf{X}|\mathbf{Y}, \hat{\boldsymbol{\gamma}})$. Here $\hat{\boldsymbol{\gamma}}$ is an estimate of $\boldsymbol{\gamma}$ obtained using for example maximum likelihood estimation, or maximum a posteriori estimation in the Bayesian setting[3]. The parameter model $\pi(\boldsymbol{\gamma})$ can often be chosen so that the posterior mean and variance of $\mathbf{X}$ agree with the classical kriging predictions (see e.g. Omre and Halvørsen, 1989). Even if this is not done, we will throughout this thesis refer to the conditional mean of the posterior distribution as the kriging predictor.

Example 1. As an example, consider a purely Gaussian model with known parameters $\boldsymbol{\gamma}$. Let $\mathbf{X}_1$ be a vector containing $X(\mathbf{s})$ evaluated at the measurement locations and let $\mathbf{X}_2$ contain $X(\mathbf{s})$ at the locations, $\mathbf{t}_1, \ldots, \mathbf{t}_m$, for which the kriging predictor should be calculated. With $\mathbf{X} = (\mathbf{X}_1^\top, \mathbf{X}_2^\top)^\top$, one has $\mathbf{X}_1 = \mathbf{A}_1\mathbf{X}$, and $\mathbf{X}_2 = \mathbf{A}_2\mathbf{X}$ for two diagonal matrices $\mathbf{A}_1$ and $\mathbf{A}_2$, and a Gaussian hierarchical model can be written as

$$\mathbf{Y}|\mathbf{X}, \boldsymbol{\gamma} \sim \mathsf{N}(\mathbf{A}_1\mathbf{X}, \boldsymbol{\Sigma}_{\mathcal{E}}), \qquad \text{and} \qquad \mathbf{X}|\boldsymbol{\gamma} \sim \mathsf{N}(\mathbf{0}, \boldsymbol{\Sigma}_X),$$

where $\boldsymbol{\Sigma}_{\mathcal{E}}$ is the covariance matrix for the measurement noise $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^\top$ and $\boldsymbol{\Sigma}_X$ is determined by the covariance function $r(\mathbf{s}, \mathbf{t})$ for the latent field. It is straightforward to show that the posterior is $\mathbf{X}|\mathbf{Y}, \boldsymbol{\gamma} \sim \mathsf{N}(\hat{\boldsymbol{\Sigma}}\mathbf{A}_1\boldsymbol{\Sigma}_{\mathcal{E}}^{-1}\mathbf{Y}, \hat{\boldsymbol{\Sigma}})$, where $\hat{\boldsymbol{\Sigma}} = (\boldsymbol{\Sigma}_X^{-1} + \mathbf{A}_1^\top\boldsymbol{\Sigma}_{\mathcal{E}}^{-1}\mathbf{A}_1)^{-1}$, and the well-known expression for the kriging predictor is given by the conditional mean

$$\begin{aligned}\mathsf{E}(\mathbf{X}_2|\mathbf{Y}, \boldsymbol{\gamma}) &= \mathbf{A}_2\hat{\boldsymbol{\Sigma}}\mathbf{A}_1\boldsymbol{\Sigma}_{\mathcal{E}}^{-1}\mathbf{Y} = \mathbf{A}_2\boldsymbol{\Sigma}_X\mathbf{A}_1^\top(\mathbf{A}_1\boldsymbol{\Sigma}_X\mathbf{A}_1^\top + \boldsymbol{\Sigma}_{\mathcal{E}})^{-1}\mathbf{Y} \\ &= \boldsymbol{\Sigma}_{X_2X_1}(\boldsymbol{\Sigma}_{X_1} + \boldsymbol{\Sigma}_{\mathcal{E}})^{-1}\mathbf{Y} = \boldsymbol{\Sigma}_{X_2X_1}\boldsymbol{\Sigma}_Y^{-1}\mathbf{Y},\end{aligned} \tag{3}$$

[3]Traditionally in geostatistics, the estimation is done in two steps using the variogram. In the first step, the *empirical variogram* is estimated from data using a non-parametric estimator, and in the second step, a parametric model is fitted to the empirical variogram.

where the elements on row i and column j in $\mathbf{\Sigma}_{X_2X_1}$ and $\mathbf{\Sigma}_Y$ are given by the covariances $r(\mathbf{t}_i, \mathbf{s}_j)$ and $r(\mathbf{s}_i, \mathbf{s}_j) + \mathrm{Cov}(\epsilon_i, \epsilon_j)$ respectively. The variance of the kriging predictor is found using the Woodbury identity (Woodbury, 1950) on the matrix $\hat{\mathbf{\Sigma}}$:

$$\begin{aligned}\mathsf{V}(\mathbf{X}_2|\mathbf{Y}, \boldsymbol{\gamma}) = \mathbf{A}_2\hat{\mathbf{\Sigma}}\mathbf{A}_2^\top &= \mathbf{A}_2\left(\mathbf{\Sigma}_X - \mathbf{\Sigma}_X\mathbf{A}_1^\top(\mathbf{\Sigma}_{X_1} + \mathbf{\Sigma}_{\mathcal{E}})\mathbf{A}_1\mathbf{\Sigma}_X\right)\mathbf{A}_2^\top \\ &= \mathbf{\Sigma}_{X_2} - \mathbf{\Sigma}_{X_2X_1}\mathbf{\Sigma}_Y^{-1}\mathbf{\Sigma}_{X_2X_1}^\top.\end{aligned}$$

4 Lattice data and Gaussian Markov random fields

In Section 2, the domain $\mathcal{D}$ was a continuous region, typically a subset of $\mathbb{R}^d$. Another important branch of spatial statistics is the analysis of data on discrete domains, such as lattices (regular grids) or more generally on any collection of countably many spatial locations.

Models on discrete domains are naturally specified differently than those on continuous domains. On discrete domains, a popular choice is to use GMRFs. A random variable

$$\mathbf{x} = (x_1, \ldots, x_n)^\top \sim \mathsf{N}(\boldsymbol{\mu}, \mathbf{Q}^{-1})$$

is called a GMRF with respect to a given neighborhood system if the joint distribution for $\mathbf{x}$ satisfies $\pi(x_i|\mathbf{x}_{-i}) = \pi(x_i|\mathbf{x}_{\mathcal{N}_i})\ \forall\ i$, where $\mathcal{N}_i$ is the neighborhood of i and $\mathbf{x}_{-i}$ denotes all elements in $\mathbf{x}$ except x_i. The neighborhood of i typically consists of all points that, in some sense, are close to i. In theory, there are no restrictions on the size of the neighborhood, and one could for example have $\mathbf{x}_{\mathcal{N}_i} = \mathbf{x}_{-i}$ which shows that any multivariate normal distribution with a symmetric positive definite covariance matrix is a GMRF and vice versa. However, the advantages of the Markov assumption naturally occurs when the neighborhood is small, and on a regular lattice, the neighborhood can for example be chosen as the four closest nodes.

Note that GMRFs are typically parametrized using the *precision matrix* $\mathbf{Q}$, which is the inverse of the covariance matrix. One of the reasons for this is the following important implication of the GMRF condition. If $i \neq j$ then:

$$x_i \perp x_j|\mathbf{x}_{-\{i,j\}} \iff Q_{ij} = 0 \iff j \notin \mathcal{N}_i.$$

This means that the following properties are equivalent: *1)* x_i and x_j are conditionally independent; *2)* the corresponding element in the precision matrix, Q_{ij}, is zero; and *3)* i and j are not neighbors.

Simultaneous model specifications of GMRFs, or so-called SAR models, date back to the work by Whittle (1954). A SAR model can be written as

$$x_i - \sum_{j:j\neq i} \beta_{ij} x_j = e_i, \quad i = 1, \ldots, n,$$

where $e_i \sim \mathsf{N}(0, \kappa_i^{-1})$, $e_i \perp e_j$ for $i \neq j$, and $\boldsymbol{\beta} = [\beta_{ij}]$ is a matrix determined by the neighborhood structure. Following the work of Besag (1974), it is today far more common to implicitly define GMRFs by specifying each of the conditional distributions

$$x_i | \mathbf{x}_{-i} \sim \mathsf{N}\left(\sum_{j:j\neq i} \beta_{ij} x_j, \kappa_i^{-1} \right), \quad i = 1, \ldots, n.$$

These models are known as CAR models, and the conditional distributions must satisfy certain regularity conditions to ensure that a joint model exists with the specified conditional distributions (Rue and Held, 2005). One reason for the popularity of the CAR models is that conditional specifications can be preferable for estimation and model interpretation, and in fact, any simultaneously specified model can be expressed as a conditionally specified model but not vice versa (see Cressie, 1991, for details).

A disadvantage with the CAR models of Besag (1974) and Besag et al. (1991) was that while they have been used for both lattices and spatially motivated graphs, they were only spatially consistent for regular lattices, which limited their applicability. However, recently Lindgren et al. (2011) derived the CAR models in a new way that both removes the lattice constraint and allows for the construction of spatially consistent non-stationary CAR-like models. The method is based on re-formulating the problem in terms of SPDEs in combination with the finite element method from numerical analysis. We return to this method in Section 7.

4.1 Computational details

Since $Q_{ij} \neq 0$ only if i and j are neighbors, most GMRFs have sparse precision matrices. The sparsity of the precision matrix facilitates the use of computationally efficient techniques for sparse matrix operations when working with GMRFs. This fact is used in all papers of this thesis, and in this section we will briefly discuss the computational advantages and some techniques that are used extensively throughout the thesis.

Consider, for example, a regular lattice in $\mathbb{R}^2$ consisting of n locations. Simulating a general Gaussian field $\mathbf{x} \sim \mathsf{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ on the lattice is done by first simulating a vector $\mathbf{v}$ of n independent identically distributed standard Gaussian variables and then calculating

$$\mathbf{x} = \boldsymbol{\mu} + \mathbf{L}\mathbf{v}, \tag{4}$$

where $\mathbf{L}$ is the *Cholesky factor* of $\boldsymbol{\Sigma}$. For a positive definite matrix $\boldsymbol{\Sigma}$, the Cholesky factor is the unique lower triangular matrix with strictly positive diagonal elements satisfying $\boldsymbol{\Sigma} = \mathbf{L}\mathbf{L}^\top$. A simple algorithm for calculating the Cholesky factor of a matrix is given in the following pseudocode, taken from Algorithm 2.8 in Rue and Held (2005). To simplify the presentation, vector notation is used in the algorithm, so $v_{j:n} \leftarrow Q_{j:n,j}$ is short for setting $v_k = Q_{kj}$ for $k = j, \ldots, n$, and so on.

Algorithm 4.1. *Cholesky factorization of a positive definite matrix* $\mathbf{Q}$.

```
for j = 1 → n do
    v_{j:n} ← Q_{j:n,j}
    for k = 1 → j − 1 do
        v_{j:n} ← v_{j:n} − L_{j:n,k} L_{jk}
    end for
    L_{j:n,j} ← v_{j:n,j} / √v_j
end for
return L
```

Several alternative algorithms exist for calculating the Cholesky factor, but as for Algorithm 4.1, they all require $n^3/3$ floating point operations in the overall process. Simulating $\mathbf{x}$ using (4) requires one Cholesky factorization, a matrix-vector multiplication, and a vector addition. The most expensive part is the Cholesky factorization, so the cost for performing the simulation is thus $\mathcal{O}(n^3)$.

To instead simulate a GMRF $\mathbf{x} \sim \mathsf{N}(\boldsymbol{\mu}, \mathbf{Q}^{-1})$ on the lattice, one has to solve

$$\mathbf{x} = \boldsymbol{\mu} + \mathbf{L}^{-\top}\mathbf{v}, \tag{5}$$

where $\mathbf{L}$ now denotes the Cholesky factor of $\mathbf{Q}$. Since $\mathbf{L}$ is lower triangular, one should not calculate its inverse but instead solve the system $\mathbf{L}^\top\mathbf{u} = \mathbf{v}$ in order to evaluate $\mathbf{L}^{-\top}\mathbf{v}$. This method is called *back substitution* because the solution, $\mathbf{u}$, is

calculated in a backward loop

$$u_i = \frac{1}{L_{ii}} \left(v_i - \sum_{j=i+1}^{n} L_{ji} u_j \right), \quad i = n, \ldots, 1.$$

Similarly if one were to calculate $\mathbf{L}^{-1}\mathbf{v}$, this is done using a forward loop and is therefore called *forward substitution*. Performing back substitution (and forward substitution) in general requires n^2 floating point operations, and if $\mathbf{L}$ is sparse this number can be decreased by only evaluating the non-zero terms.

The computational cost for simulating a GMRF is thus determined by the cost for calculating the Cholesky factor $\mathbf{L}$. If $\mathbf{Q}$ is a sparse matrix, Algorithm 4.1 can be modified to take this into account. An easy example is if $\mathbf{Q}$ is a band matrix with band width p. In this case, the bandwidth of $\mathbf{L}$ is p and the elements below the pth diagonal in $\mathbf{L}$ do not have to be evaluated in the algorithm, resulting in a reduction of the number of floating point operations to $n(p^2 + 3p)$, assuming that $p \ll n$. In this situation, it is also easy to show that the number of floating point operations required for performing the back substitution is $2np$. Thus, both the Cholesky factorization and the back substitution are linear in n.

For general sparse matrices $\mathbf{Q}$, one can reorder the nodes so that the reordered matrix has a small bandwidth and then use the methods for band matrices. This technique is called bandwidth reduction, and is intuitively easy to understand and rather easy to implement with simple changes to Algorithm 4.1. However, in general there exist other methods, such as nested dissection or minimum degree methods, that can reduce the number of non-zero elements in $\mathbf{L}$, and therefore the number of required floating point operations, further. For details, see Rue and Held (2005) and the references therein. An example can be seen in Figure 2. In Panel (a), a sparse 100×100 precision matrix $\mathbf{Q}$ is shown. The matrix reordered using a bandwidth reduction method and a minimum degree method are shown in Panels (b) and (c) respectively. The Cholesky factors of the matrices in Panels (a-c) are shown in Panels (d-f). The matrix $\mathbf{Q}$ has 1608 non-zero elements, and the Cholesky factors in Panels (d-f) have 3452, 1836, and 1616 non-zero elements respectively.

Using an efficient Cholesky factorization method, the computational cost for simulating a GMRF using (5) can be reduced from $\mathcal{O}(n^3)$ to $\mathcal{O}(n^{3/2})$ given that the neighborhood size is $m \ll n$ (Rue and Held, 2005), a substantial reduction if n is large. Given the Cholesky factor, many other operations, such as kriging and

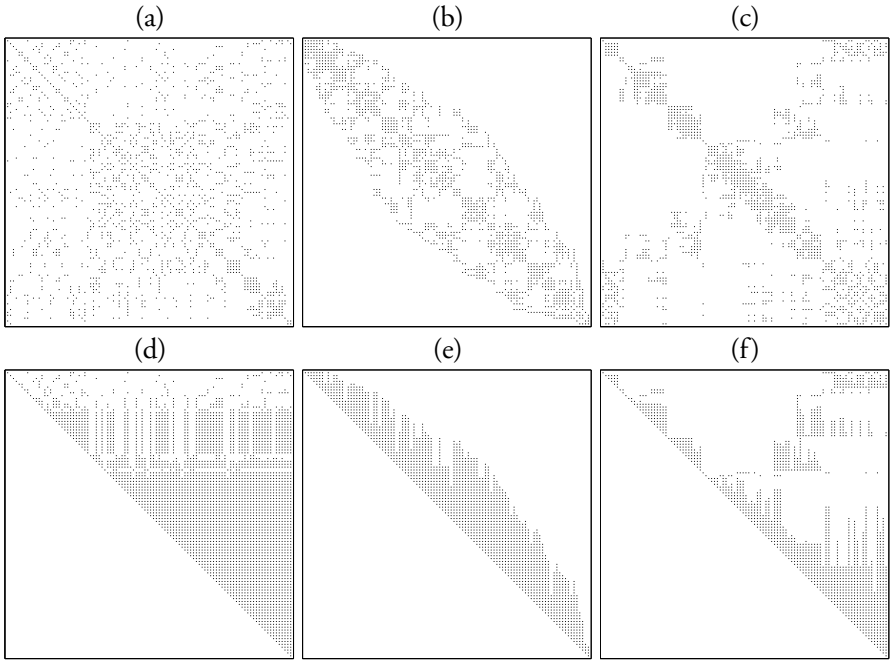

Figure 2: (a) A sparse precision matrix **Q**, and its corresponding Cholesky factor in (d). Only nonzero elements are shown and these are indicated by a dot. (b) The matrix **Q** reordered using a bandwidth reduction method, and its corresponding Cholesky factor in (e). (c) The matrix **Q** reordered using a minimum degree algorithm, and its corresponding Cholesky factor in (f).

likelihood evaluations, can also be calculated efficiently using back substitutions and sparse matrix multiplications, and the computational gain is therefore substantial for most steps of any statistical analysis if sparsity properties can be used. A practical application of a GMRF model is given in the following section.

4.2 Estimation of spatially dependent temporal trends using GMRFs

The African Sahel is a region in northern Africa that has received much attention regarding desertification and climatic variations (Olsson, 1993, Nicholson, 2000, Lamb, 1982), and several authors have used satellite imagery to study vegetation in the region (Tucker et al., 1985, Justice and Hiernaux, 1986, Seaquist et al.,

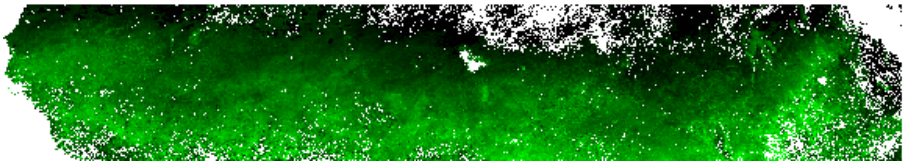

Figure 3: Vegetation index data from the Sahel region 1983. Black denotes a low index value, green a high value, and white missing data or water. Each pixel in the image is of size 8 km $\times$ 8 km.

2003). Notably, Eklundh and Olsson (2003) observed a strong increase in seasonal vegetation index over parts of the Sahel for the period 1982-1999, using Advanced Very High Resolution Radiometer (AVHRR) data from the NOAA/NASA Pathfinder AVHRR Land (PAL) database (Agbu and James, 1994, James and Kalluri, 1994). The PAL dataset consists of calibrated vegetation index measurements which are mapped to 8 km $\times$ 8 km pixels on a regular lattice consisting of approximately 117 000 nodes over the Sahel region. The measurements for 1983 can be seen in Figure 3. The study in Eklundh and Olsson (2003) was based on ordinary least squares (OLS) linear regression on individual time series extracted for each pixel in the satellite images, and a drawback with this method is that the spatial dependencies in the vegetation cover are neglected in the estimation. Hence, improved estimates can be obtained by correctly modeling the spatial dependency. This is the focus of Paper A and a brief summary of the method is given in this section as an example of an application where the ability to use efficient methods for sparse matrices is crucial because of the size of the dataset.

Let $\mathbf{X}_t$ denote the latent vegetation field at time t and assume that the measurements $\mathbf{Y}_t$ at time t are generated as $\mathbf{Y}_t|\mathbf{X}_t, \boldsymbol{\Sigma}_{\epsilon_t} \sim \mathsf{N}(\mathbf{A}_t\mathbf{X}_t, \boldsymbol{\Sigma}_{\epsilon_t})$. Here $\mathbf{A}_t$ is a diagonal matrix determined by which pixels are observed, $\mathbf{X}_t$ is restricted to follow a spatially varying linear regression, $\mathbf{X}_t = \mathbf{K}_1 \cdot t + \mathbf{K}_2$, and $\boldsymbol{\Sigma}_{\epsilon_t}$ is assumed to be diagonal. As a process model for the yearly vegetation $\mathbf{X}_t$ at times $t = 0, 1, \ldots, T-1$, a second-order polynomial intrinsic GMRF (Gamerman et al., 2003, Rue and Held, 2005, Section 3.4.2) model is used. Assuming that $\mathbf{X}_{t_1} \perp \mathbf{X}_{t_2}$ if $t_1 \neq t_2$, the corresponding distributions for the field of regression coefficients $\mathbf{K} = [\mathbf{K}_1^\top, \mathbf{K}_2^\top]$ is $\mathbf{K} \sim \mathsf{N}(\mathbf{0}, (\kappa\mathbf{Q})^{-1})$. Here, κ determines the strength of the spatial dependency and the precision matrix $\mathbf{Q}$ is sparse and determined by the intrinsic GMRF prior and the regression basis.

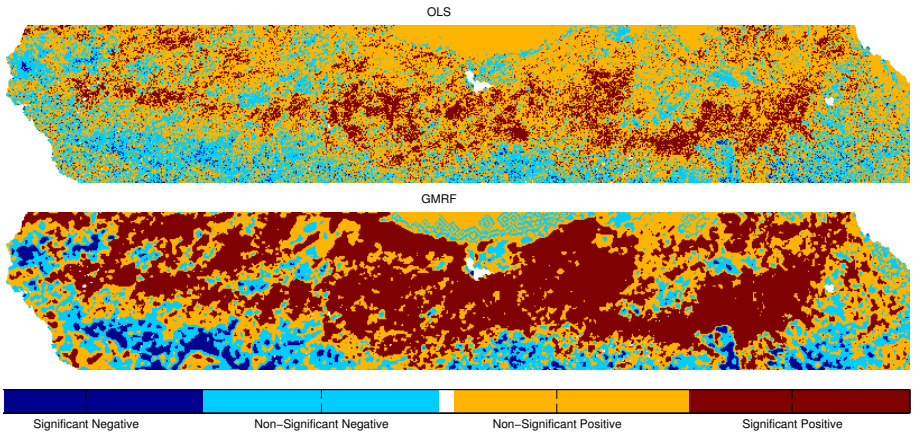

Figure 4: Significance estimates for the slope of the linear trends using the OLS model (upper figure) and the GMRF model (lower figure). Note the large number of significant positive linear trends in the GMRF estimate.

The parameter κ and the measurement noise variance at each pixel are estimated using the Expectation Maximization (EM) algorithm (Dempster et al., 1977), and given the estimated parameters, the posterior distribution of $\mathbf{K}$ is $\mathbf{K}|\mathbf{Y}, \boldsymbol{\Sigma}_\varepsilon, \kappa \in \mathsf{N}(\boldsymbol{\mu}_{K|Y}, \mathbf{Q}_{K|Y}^{-1})$, with

$$\boldsymbol{\mu}_{K|Y} = \mathbf{Q}_{K|Y}^{-1}\mathbf{A}^\top\boldsymbol{\Sigma}_\varepsilon^{-1}\mathbf{Y} \qquad \text{and} \qquad \mathbf{Q}_{K|Y} = \kappa\mathbf{Q} + \mathbf{A}^\top\boldsymbol{\Sigma}_\varepsilon^{-1}\mathbf{A}.$$

Based on the posterior distribution, standard hypothesis testing is used for each pixel to find areas that have experienced a significant change in the vegetation cover over the studied time period. The result can be seen in Figure 4. The GMRF estimates (bottom panel) are smoother and exhibit larger contiguous regions with significant trends than a comparable analysis using OLS (top panel). The larger contiguous regions and smoother estimates will most likely aid interpretation of the data, and make it easier to identify underlying reasons for the detected changes in vegetation.

5 Estimating excursion sets for latent Gaussian models

In certain applications, such as the Sahel example from the previous section, one is not only interested in point estimates of the latent field, but also wants to find

regions where the field exceeds some given level. In Figure 4, the area that represents regions that have experienced a significant increase in vegetation is calculated as $D_m = \{\mathbf{s} : \mathsf{P}(K_1(\mathbf{s}) > 0) \geq 0.95\}$, where the probabilities are calculated under the posterior distribution of K_1. The problem with this definition of D_m is that of multiple hypothesis testing; the confidence level ($\alpha = 0.05$ in this case) does not give us any information about the family-wise error rate. That is, the probability $\mathsf{P}(K_1(\mathbf{s}) > 0, \mathbf{s} \in D_m)$ is, in general, not equal to $1 - \alpha$. How to construct such *excursion sets* with a specified family-wise error is a difficult problem with applications in a wide range of scientific fields, including brain imaging (Marchini and Presanis, 2003) and astrophysics (Beaky et al., 1992), and this is the focus of Paper E.

Formally, we define the positive u excursion set, A_u^+, for a function $f(\mathbf{s})$, $\mathbf{s} \in \mathcal{D}$, by $A_u^+(f) = \{\mathbf{s} \in \mathcal{D}; f(\mathbf{s}) > u\}$. For a stochastic process $X(\mathbf{s})$, the excursion set $E_{u,\alpha}^+(X)$ is defined as the largest set D such that $X(\mathbf{s})$ exceeds the level u with a certain probability $1 - \alpha$ for all $\mathbf{s} \in D$,

$$E_{u,\alpha}^+(X) = \arg\max_D \{|D| : \mathsf{P}(D \subseteq A_u^+(X)) \geq 1 - \alpha\}.$$

There are several ways these sets can be estimated for latent Gaussian models. One method is to simulate from the posterior distribution using MCMC and then find the largest region satisfying the probability constraint based on the simulations. Another method is to use a shape optimization method in combination with a numerical integration method. Notably the quasi Monte Carlo methods by Genz and Bretz (2009) can be used to estimate high-dimensional Gaussian integrals, and therefore Gaussian probabilities. Both these methods are computationally intensive, and much more so than for example calculating kriging predictions. To be able to estimate these sets for large spatial problems, a different strategy, based on a combination of a parametric family for the possible excursion sets, sequential Monte Carlo integration, and Integrated Nested Laplace Approximations is proposed in Paper E. The method is especially efficient if Markov properties of the latent field can be used, and it can also be used for the related problem of finding uncertainty regions for contour curves.

Using the method for the Sahel example, we are able to find regions that have experienced an increase in vegetation while controlling the joint error. The result for the western part of the Sahel region (a subset of the original domain) can be seen in Figure 5. The estimated excursion set $E_{0,0.05}^+(\mathbf{K}_2)$ is shown in red and the point-wise positive significant trends in green. The interpretation of the

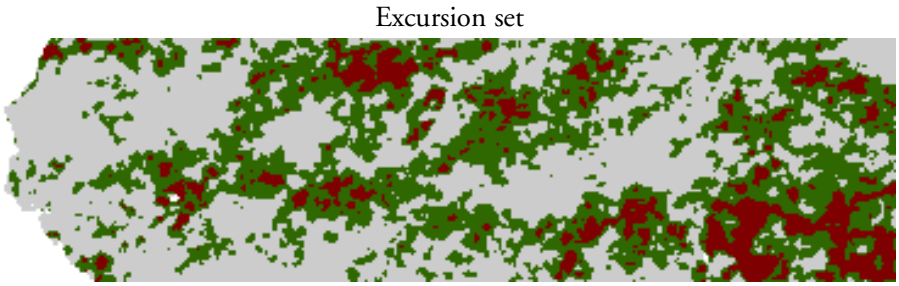

Figure 5: Results from the Sahel vegetation data. The estimated excursion set $E^{+}_{0,0.05}(\mathbf{K}_1)$ is shown in red and the point-wise positive significant trends in green.

result is that one with high certainty can conclude that the areas indicated in red have experienced an increase in vegetation over the studied time period. Hence, conclusions drawn by Eklundh and Olsson (2003) seem valid, also when taking the spatial dependency of the vegetation measurements into account and when estimating the excursion sets controlling the family-wise error.

6 Efficient representations of continuous Gaussian fields

As mentioned in Section 2, continuous Gaussian models are traditionally specified through the mean value function and the covariance function. This method for specifying the models is difficult to use if non-stationary models are needed, and inference using covariance-based models is in general computationally expensive. One example was given in Section 4.1 where simulating Gaussian models was discussed, and another is spatial prediction. For example, to calculate the kriging prediction (3) in Example 1 requires inverting the $n \times n$ covariance matrix $\mathbf{\Sigma}_Y$, which is not computationally feasible if the number of observations, n, is large. The desire of using complicated non-stationary models under computational restrictions has led to a large number of new statistical methods, and a few of these are introduced in this section.

6.1 Low-rank methods

In many of the techniques for building computationally efficient models, the main assumption is that a latent zero-mean Gaussian process $X(\mathbf{s})$ can be ex-

pressed, or at least approximated, through some finite basis expansion

$$X(\mathbf{s}) = \sum_{j=1}^{m} w_j \varphi_j(\mathbf{s}), \tag{6}$$

where w_j are Gaussian random variables and $\{\varphi_j\}_{j=1}^m$ are some pre-defined basis functions. To understand why this increases the computational efficiency, consider the Gaussian hierarchical model in Example 1. Using the approximation (6), $\mathbf{X}$ can be written as $\mathbf{X} = \mathbf{B}\mathbf{w} \sim \mathsf{N}(\mathbf{0}, \mathbf{B}\boldsymbol{\Sigma}_w\mathbf{B}^\top)$, where column i in the matrix $\mathbf{B}$ contains the basis function $\varphi_i(\mathbf{s})$ evaluated at all measurement locations and all locations where the kriging prediction is to be calculated. With $\mathbf{B}_1 = \mathbf{A}_1\mathbf{B}$ and $\mathbf{B}_2 = \mathbf{A}_2\mathbf{B}$, the kriging predictor can be written as

$$\mathsf{E}(\mathbf{X}_2|\mathbf{Y}, \boldsymbol{\gamma}) = \mathbf{B}_2(\boldsymbol{\Sigma}_w^{-1} + \mathbf{B}_1^\top \boldsymbol{\Sigma}_{\mathcal{E}}^{-1}\mathbf{B}_1)^{-1}\mathbf{B}_1\boldsymbol{\Sigma}_{\mathcal{E}}^{-1}\mathbf{Y}. \tag{7}$$

If the measurement noise is uncorrelated, $\boldsymbol{\Sigma}_{\mathcal{E}}$ is diagonal and easy to invert. If $\boldsymbol{\Sigma}_w^{-1}$ is either known, or easy to calculate, the most expensive calculation in (7) is to invert the $m \times m$ matrix $(\boldsymbol{\Sigma}_w^{-1} + \mathbf{B}_1^\top \boldsymbol{\Sigma}_{\mathcal{E}}^{-1}\mathbf{B}_1)$. This requires $\mathcal{O}(m^3)$ floating point operations, and by choosing $m \ll n$, the computational cost for calculating the kriging prediction can thus be substantially decreased.

If the basis expansion (6) is used as a model approximation, a natural question is how the basis functions $\{\varphi_j\}_{j=1}^m$ should be chosen. One way to obtain an, in some sense optimal, expansion of the form (6) is to use the eigenfunctions of the covariance function for the latent field $X(\mathbf{s})$ as a basis, which is usually called the Karhunen-Loève transform. This is, however, seldom used in practice since analytic expressions for the eigenfunctions are only known in a few simple cases. In certain situations, numerical approximations of the eigenfunctions can be obtained by performing principal component analysis on the empirical covariance matrix which is estimated from data. These discrete approximations, sometimes referred to as empirical orthogonal functions (EOFs), can however be poor approximations of the true eigenfunctions since the empirical covariance matrix is affected by the measurement noise. Also, because the EOFs are discrete, they have to be interpolated in some way to produce a continuous model approximation.

A popular method that fit in to the general construct of low-rank approximations is the fixed-rank kriging method by Cressie and Johannesson (2008). Their recommendation is to use multiresolutional basis functions, such as wavelets, and a related method for constructing non-stationary covariance models using wavelets is given in Nychka et al. (2002). There are many other methods that can be

viewed as low-rank approximations, e.g. the predictive process method by Banerjee et al. (2008) and the process convolution method presented in the next section. For additional details on the low-rank methods, see Gelfand et al. (2010).

6.2 Process convolutions

In the process convolution method, the Gaussian random field $X(\mathbf{s})$ on $\mathbb{R}^d$ is specified as a process convolution

$$X(\mathbf{s}) = \int k(\mathbf{s}, \mathbf{u})\mathcal{B}(\,\mathrm{d}\mathbf{u}), \tag{8}$$

where k is some deterministic kernel function and $\mathcal{B}$ is a Brownian sheet. This representation of stationary Gaussian processes is not new, but has become popular lately because it provides an easy method for producing non-stationary models by allowing the convolution kernel to be dependent on location (Barry and Ver Hoef, 1996, Higdon, 2001, Cressie and Ravlicová, 2002, Rodrigues and Diggle, 2010).

If, however, the process is stationary one must have $k(\mathbf{s}, \mathbf{u}) = k(\mathbf{s} - \mathbf{u})$ and the covariance function for X is expressed in terms of the convolution kernel through

$$C(\mathbf{h}) = \int k(\mathbf{u} - \mathbf{h})k(\mathbf{u})\,\mathrm{d}\mathbf{u}. \tag{9}$$

Thus, the covariance function C, the spectrum S, and the kernel k are related through $(2\pi)^d|\mathcal{F}(k)|^2 = \mathcal{F}(C) = S$, where $\mathcal{F}(\cdot)$ denotes the Fourier transform. The method can also be used to define non-Gaussian models by replacing $\mathcal{B}$ with some non-Gaussian process (see e.g. Åberg et al., 2009), and this is discussed further in Paper D.

An approximation that is commonly used in practice for process convolution models is to approximate the integral (8) by a sum

$$X(\mathbf{s}) \approx \sum_{j=1}^{m} k(\mathbf{s} - \mathbf{u}_j)w_j,$$

where $\mathbf{u}_1, \ldots, \mathbf{u}_m$ are some fixed locations in the domain, and w_j are independent zero-mean Gaussian variables with variances equal to the area associated with each location $\mathbf{u}_j$. Thus, with this approximation, the method can be used to obtain a low rank approximation of the form (6), with basis functions $\varphi_j(\mathbf{s}) = k(\mathbf{s} - \mathbf{u}_j)$, of a Gaussian field.

6.3 Covariance tapering

The covariance tapering method, introduced by Furrer et al. (2006), is not a method for constructing covariance models, but a method for approximating a given covariance model to increase the computational efficiency. The idea is to taper the true covariance, $C(\mathbf{h})$, to zero beyond a certain range, θ, by multiplying the covariance function with some compactly supported, positive definite, taper function $r_\theta(\mathbf{h})$. Using the tapered covariance, $C_{tap}(\mathbf{h}) = r_\theta(\mathbf{h})C(\mathbf{h})$, the matrix $\mathbf{\Sigma}_Y$ in the expression for the kriging predictor is sparse, which facilitates the use of sparse matrix techniques.

The sparsity of $\mathbf{\Sigma}_Y$ increases as θ is decreased, but the taper function and the taper range should, of course, also be chosen such that the basic shape of the true covariance function is preserved, and of especial importance for asymptotic considerations is that the smoothness at the origin is preserved. A popular choice for taper functions are the Wendland functions (Wendland, 1995),

$$\text{Wendland}_1\text{:}\ \ r_\theta(\mathbf{h}) = \left(\max\left[1 - \frac{\|\mathbf{h}\|}{\theta}, 0\right]\right)^4 \left(1 + 4\frac{\|\mathbf{h}\|}{\theta}\right),$$

$$\text{Wendland}_2\text{:}\ \ r_\theta(\mathbf{h}) = \left(\max\left[1 - \frac{\|\mathbf{h}\|}{\theta}, 0\right]\right)^6 \left(1 + 6\frac{\|\mathbf{h}\|}{\theta} + \frac{35\|\mathbf{h}\|^2}{2\theta^2}\right).$$

These were for example used by Furrer et al. (2006) to study the accuracy and numerical efficiency of tapered Matérn covariance functions.

6.4 Markov approximations

Although GMRFs are computationally efficient, they are seldom the most natural model choices. One reason is that it is hard to specify the precision matrix such that the corresponding covariance function is similar to some commonly used covariance function for a given data set. However, for regular lattices in $\mathbb{R}^2$, Rue and Tjelmeland (2002) showed that, for a large family of covariance functions, Gaussian fields can be well approximated by GMRFs with small neighborhood structures. The second, more serious problem, is that spatial data is seldom located on regular lattices. Several approaches for using lattice-based GMRFs for non-lattice data have been suggested in the literature; notably, nearest neighbor mapping of the data locations to the grid locations (Hrafnkelsson and Cressie, 2003), assuming that the GMRFs values for non-lattice points are equal to the closest lattice

point (Wikle et al., 1998), or using linear interpolation of the GMRFs lattice values to assign values to non-lattice locations (Werner Hartman, 2006). Although these approaches are simple, a more general approach would be to use an efficient *continuous* representation of the latent Gaussian field where Markov properties could be used, and such a representation is the topic of the next section.

7 The SPDE approach

Recently, Lindgren et al. (2011) derived a method for explicit, and computationally efficient, continuous Markov representations of Gaussian Matérn fields. As previously mentioned, the method can be used to re-define the CAR models to remove the lattice constraint and to allow for non-stationary extensions. However, the most important implication of the work is that it provides a spatially consistent method for approximating continuous Gaussian fields using GMRFs which thus enables the use of sparse matrix techniques for GMRFs when doing inference for continuous Gaussian fields.

The method is based on the fact that a Gaussian Matérn field on $\mathbb{R}^d$ can be viewed as a solution to the SPDE

$$(\kappa^2 - \Delta)^{\frac{\alpha}{2}} X = \phi \mathcal{W}, \tag{10}$$

where $\mathcal{W}$ is Gaussian white noise, Δ is the Laplacian, and $\alpha = \nu + d/2$, this was first noted by Whittle (1963). An approximation of $X(\mathbf{s})$ of the form (6) is then constructed through Hilbert space approximations of the solution to the SPDE with respect to the basis $\{\varphi_k\}$. The procedure can be viewed as a finite element approximation of the solution to the SPDE, where the stochastic weights in (6) are calculated by requiring the stochastic weak formulation of the SPDE to hold for only a specific set of test functions $\{\psi_i, i = 1, \ldots, n\}$,

$$\left\{(\kappa^2 - \Delta)^{\frac{\alpha}{2}} X(\psi_i), i = 1, \ldots, n\right\} \stackrel{d}{=} \{\phi \mathcal{W}(\psi_i), i = 1, \ldots, n\}. \tag{11}$$

To simplify the presentation, we only outline the procedure for the fundamental case $\alpha = 2$. Lindgren et al. (2011) then use $\psi_i = \varphi_i$, and using the basis expansion (6) for X one has

$$(\kappa^2 - \Delta) X(\varphi_i) = \sum_{j=1}^{n} w_j \left\langle \varphi_i, (\kappa^2 - \Delta)\varphi_j \right\rangle,$$

where $\langle f, g\rangle = \int f(\mathbf{s})g(\mathbf{s})\,\mathrm{d}\mathbf{s}$ is an inner product and the integral is taken over the region of interest. The left hand side of (11) can then be written as $\mathbf{Kw}$ where $\mathbf{K}$ is a matrix with elements $\mathbf{K}_{ij} = \left\langle \varphi_i, (\kappa^2 - \Delta)\varphi_j \right\rangle$ and $\mathbf{w} = (w_1, \ldots, w_n)^\top$. Under mild conditions on the basis functions, one has

$$\left\langle \varphi_i, (\kappa^2 - \Delta)\varphi_j \right\rangle = \kappa^2 \left\langle \varphi_i, \varphi_j \right\rangle + \left\langle \nabla\varphi_i, \nabla\varphi_j \right\rangle.$$

Hence, the matrix $\mathbf{K}$ can be written as the sum $\mathbf{K} = \kappa^2\mathbf{C} + \mathbf{G}$ where $\mathbf{C}$ and $\mathbf{G}$ are matrices with elements $\mathbf{C}_{ij} = \left\langle \varphi_i, \varphi_j \right\rangle$ and $\mathbf{G}_{ij} = \left\langle \nabla\varphi_i, \nabla\varphi_j \right\rangle$ respectively. The right hand side of (11) is a Gaussian vector with mean zero and covariance matrix $\phi^2\mathbf{C}$, so one has $\mathbf{w} \sim \mathsf{N}\left(\mathbf{0}, \phi^2\mathbf{K}^{-1}\mathbf{C}\mathbf{K}^{-1}\right)$. The final step is to approximate $\mathbf{C}$ with a diagonal matrix $\tilde{\mathbf{C}}$ with diagonal elements $\tilde{\mathbf{C}}_{ii} = \sum_{k=1}^{n} \mathbf{C}_{ik} = \int \varphi_i(\mathbf{s})\,\mathrm{d}\mathbf{s}$, which makes the precision matrix $\mathbf{Q} = \phi^{-2}\mathbf{K}\mathbf{C}^{-1}\mathbf{K}$ sparse if $\mathbf{G}$ is sparse.

A similar procedure is used for the case $\alpha = 1$, and for higher order $\alpha/2 \in \mathbb{N}$, the approximation is obtained by recursively using the two cases $\alpha = 1$ and $\alpha = 2$.

7.1 Wavelet basis functions and a comparison

A natural question is what type of basis functions one should use in the method. Lindgren et al. (2011) use the standard finite element basis of piecewise linear basis functions induced by some triangulation of the domain $\mathcal{D}$. An example of such a basis, taken from Paper C, on the sphere can be seen in Figure 6. Using these basis functions produces a piecewise linear approximation of the continuous field; however, there might be other types of basis functions that give better approximations of the continuous field.

To investigate this, the procedure is extended using wavelet basis functions in Paper B. Among the most widely used constructions in multiresolution analysis are the B-spline wavelets (Chui and Wang, 1992) and the Daubechies wavelets (Daubechies, 1992), that both have several desirable computational properties. Explicit expression for the matrices $\mathbf{C}$ and $\mathbf{G}$ are derived for the Daubechies wavelets and the B-spline wavelets and the procedure is extended to the corresponding wavelet bases on $\mathbb{R}^d$, using tensor product functions generated by d one-dimensional wavelet bases. By considering the covariance error when the method is used for approximating Gaussian Matérn fields, it is shown that there is no gain in using any more complicated basis functions than the piecewise linear first order B-splines, unless the range of the covariance function is very large.

As opposed to the methods introduced in Section 6.1, the SPDE method can be seen as a *high-rank* model approximation since the sparsity of the preci-

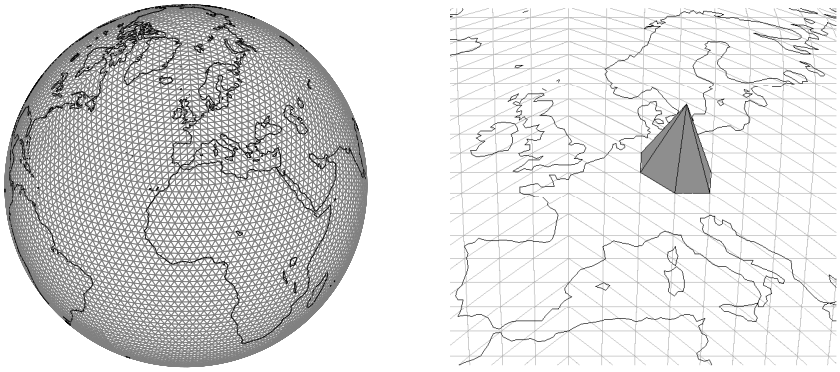

Figure 6: The left part shows a triangulation of the Earth used to define a set of piecewise linear basis functions. Each basis function is one at a node in the triangulation, and decreases linearly to zero at the neighboring nodes. The right part of the figure shows one of these functions.

sion matrix can be used to reduce the computational cost for inference even if as many basis functions are used as there are measurements. To demonstrate the accuracy of the method, a detailed comparison between the SPDE method using different types of wavelet basis functions, the process convolution method, and the tapering method is also performed in Paper B. The computational aspects of the spatial prediction problem are studied, and the results show that the SPDE method generally is more efficient and accurate than both the process convolution approach and the covariance tapering method when used for approximating Gaussian Matérn fields.

8 Extensions of the SPDE method

Instead of defining Matérn fields through the covariance function (1), Lindgren et al. (2011) used the solution to the SPDE (10) as a definition. Besides allowing for the efficient Matérn approximations, the representation has several other advantages: The definition is valid not only on $\mathbb{R}^d$ but on general smooth manifolds, such as the sphere, and facilitates non-stationary extensions by allowing the SPDE parameters κ^2 and ϕ to vary with space. For example, $\log \kappa$ can be

expanded using weighted smooth basis functions,

$$\log \kappa(\mathbf{s}) = \sum_i \beta_i b_i(\mathbf{s}),$$

and similar expansions can be used for ϕ. This extension requires only minimal changes to the method used in the stationary case, see Paper C for a detailed explanation of this case. Spatially varying anisotropy can also be incorporated by replacing the Laplacian Δ in (10) with the more general operator $\nabla \cdot (\mathbf{D}(\mathbf{s})\nabla)$. The model can be extended further by including a drift term $\mathbf{b}(\mathbf{s}) \cdot \nabla$ and temporal dependence, which leads to the general model

$$\left(\frac{\partial}{\partial t} + \kappa(\mathbf{s}, t)^2 + \mathbf{b}(\mathbf{s}, t) \cdot \nabla - \nabla \cdot (\mathbf{D}(\mathbf{s}, t)\nabla) \right) X(\mathbf{s}, t) = \phi(\mathbf{s}, t)\, \mathcal{W}(\mathbf{s}, t), \quad (12)$$

where t is the time variable and $\mathbf{b}$ is a vector field describing the direction of the drift. Further extensions are derived in Paper C and Paper D, and these are discussed in Section 8.1 and Section 8.2 respectively.

8.1 Nested SPDE models

A limitation of the popular Matérn covariance family is that it does not contain any covariance functions with negative values, such as oscillating covariance functions. One way of constructing a larger class of stochastic fields is to consider a generalization of the SPDE (10) of the form $\mathcal{L}_1 X = \mathcal{L}_2 \mathcal{W}$ for some linear differential operators $\mathcal{L}_1$ and $\mathcal{L}_2$. In Paper C, the class of nested SPDE models is introduced. The nested SPDE models are stochastic fields generated by the SPDE

$$\left(\prod_{i=1}^{n_1} (\kappa_i^2 - \Delta)^{\frac{\alpha_i}{2}} \right) X = \left(\prod_{i=1}^{n_2} (b_i + \mathbf{B}_i^\top \nabla) \right) \mathcal{W}, \quad (13)$$

for some parameters $\alpha_i \in \mathbb{N}$ and $\kappa_i > 0$, $b_i \in \mathbb{R}$ and $\mathbf{B}_i \in \mathbb{R}^d$. The class of models contains a wide family of stochastic fields, including both the Gaussian Matérn fields and fields with oscillating covariance functions. Some examples of possible covariance functions that can be obtained with this model are shown in Figure 7.

As for the standard Matérn SPDE, the class of models is easily extended to non-stationary versions on general smooth manifolds by allowing the parameters

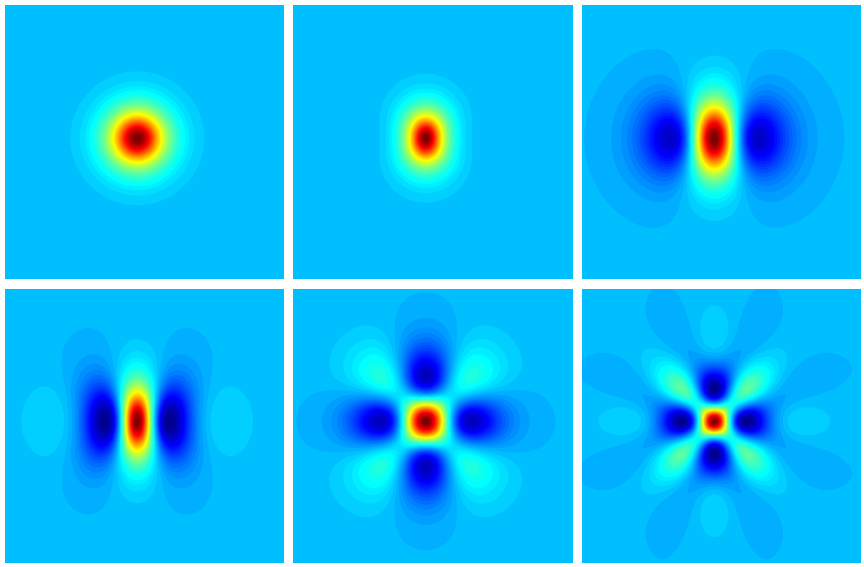

Figure 7: Covariance functions of random fields obtained from the nested SPDE model (13) with different parameters.

in the SPDE to vary with space. Using the Hilbert space approximation technique for the nested SPDE models does not produce GMRF weights directly; however, it is shown that Markov properties still can be used so that all computational advantages are preserved when extending the method to the nested SPDE models.

8.1.1 An application to random linear wave theory

In Paper C, the nested SPDE models are used to analyze a large data set of global total column ozone measurements. Another interesting application of this model class, presented in Lindgren et al. (2010), is its use in random linear wave theory. This is a theory for ocean surface waves that is widely used in, for example, naval architecture and coastal engineering. The condition of the ocean surface is, in this theory, modeled as a stochastic field (Holthuijsen, 2007). The most important concept in random linear wave theory is the wave spectrum, which defines the sea state, i.e. the most relevant properties of the ocean surface. The wave spectrum is most often defined through the directional spectrum $E(\omega, \theta) = S(\omega)D(\theta, \omega)$.

The function $S(\omega)$ is called the wave frequency spectrum, and does not contain any information about the direction of the waves. The directional information is contained in the angularly dependent part $D(\theta, \omega)$, which is normalized such that

$$\int_0^{2\pi} D(\theta, \omega)\, \mathrm{d}\theta = 1, \quad \forall\, \omega \geq 0. \tag{14}$$

The most well-known parametric form of $S(\omega)$ is the so-called Pierson-Moskowitz spectrum $S_{PM}(\omega) = A_{PM}\, \omega^{-5} \exp\left(-B_{PM}\, \omega^{-4}\right)$, where A_{PM} and B_{PM} are parameters related to the main sea state parameters[4]. As for the angularly dependent part, an often used form is the cos-2s-distribution[5] which can be written as $D(\theta, \omega) = c(s) \cos^{2s}\left(2^{-1}(\theta - \theta_1)\right)$, where the constant $c(s)$ is a normalization factor such that the condition (14) is satisfied.

Lindgren et al. (2010) proved the following theorem, showing that the standard wave model with a Pierson-Moskowitz wave-frequency spectrum and a cos-2s angular distribution can be obtained as a limiting case from a subclass of nested SPDE models.

Proposition 8.1. *The spectral density for $X(\mathbf{s})$ given by the nested SPDE model*

$$\left(\frac{B_{PM} g^{-2}}{s+2} - \Delta\right)^{\frac{s+2}{2}} X(\mathbf{s}) = \left(\mathbf{B}^\top \nabla\right)^s \mathcal{W}(\mathbf{s})$$

converges to a random wave model with a Pierson-Moskowitz wave frequency spectrum and a cos-2s angular distribution $D(\theta) = c(s) \cos^{2s}(\theta - \theta_1)$ as $s \to \infty$. For a fixed s, the nested SPDE model has an exact cos-2s angular distribution and an error of $\frac{A_{PM}}{5 B_{PM}(s+1)}$ in the wave frequency spectrum approximation measured in the L_1-norm.

The nested SPDE representation of the wave model has two main advantages. Firstly, since it is a local representation, non-stationary extensions are easy to obtain by spatially varying the parameters of the SPDE. Secondly, the nested SPDE formulation is valid on other domains than $\mathbb{R}^2$, and one can therefore use it for modeling waves on, for example, the globe. Another advantage with is that Hilbert space approximations can be used to obtain computationally efficient model representations.

[4]Usually, one has $A_{PM} = \alpha_{PM} g^2 (2\pi)^4$ and $B_{PM} = 5 f_{PM}^4 4^{-1}$, where α_{PM} and f_{PM} are the energy scale and peak frequency respectively.

[5]Note that $\mathbf{s}$ denotes the space variable in $\mathcal{D}$, while s denotes the parameter in the cos-2s-distribution. Also note that ω and θ in this section are used as parameters in the wave frequency spectrum, and not as an event in the abstract sample space Ω and a tapering range respectively.

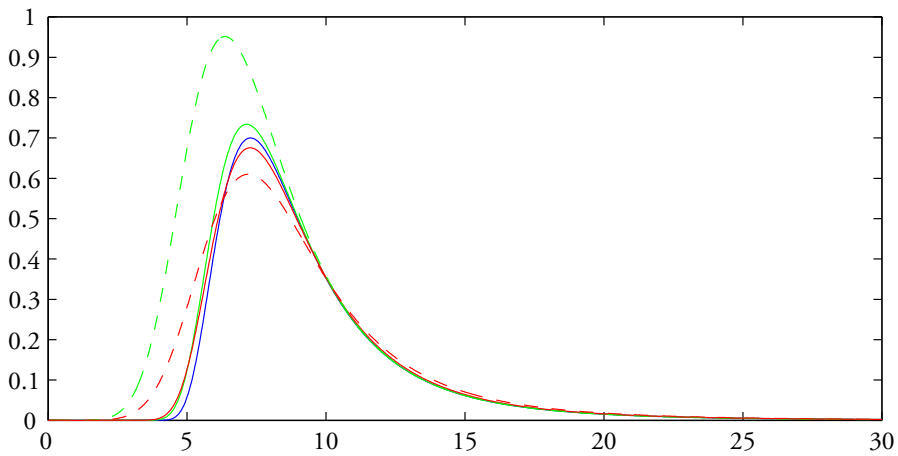

Figure 8: A Pierson-Moskowitz wave frequency spectrum (blue curve) and approximations for the worst case $s = 1$ (dashed curves) and the frequently used case $s = 15$ (solid curves). The green curves are the approximations from Proposition 8.1 and the red curves are L_1-optimal approximations for the given operator orders.

The limitation is that the parameters in the cos-2s-distribution are connected to the order of the approximation, since $\alpha = s+2$ is used. This should not be a big problem in practice since the convergence is reasonably fast. As shown in Figure 8, the wave frequency spectrum is similar to the Pierson-Moskowitz spectrum even for the worst case $s = 1$, and the difference for the, in practice popular, case $s = 15$ will likely be smaller than the parameter estimation error if the model is estimated from data. If one is not satisfied with the approximation, a better approximation can be obtained using $\mathcal{L}_1 = \prod_{i=1}^{s+2}(\kappa_i^2 - \Delta)^{\frac{1}{2}}$. Figure 8 also shows the L_1-optimal approximations assuming that $\mathcal{L}_1$ is on the product form. A last thing to note is that, since the nested SPDE models are purely spatial models, the angularly dependent part is always symmetric in the sense that $D(\theta + \pi) = D(\theta)$. That is, one can obtain cos-2s distributions of the form $\cos^{2s}(\theta - \theta_0)$ but not of the form $\cos^{2s}(\frac{\theta-\theta_0}{2})$. To avoid this problem, one would have to include time in the generating SPDE, which would require a spatio-temporal extension of the nested SPDE models. This is likely possible using an extension similar to (12) but is a subject for further research.

8.2 Spatial Matérn fields driven by non-Gaussian noise

The fact that a Gaussian Matérn field on $\mathbb{R}^d$ can be viewed as a solution to the SPDE (10) seems like a fairly straightforward statement; however, the formal presentation of this connection is far from obvious and requires a detailed analysis. In Paper E, this connection is studied for a more general class of models

$$(\kappa^2 - \Delta)^{\frac{\alpha}{2}} X = \dot{M}, \tag{15}$$

where M is an arbitrary independently scattered L_2-valued random measure with $\mathsf{E}(|M(\,\mathrm{d}\mathbf{x})|^2) = C\,\mathrm{d}\mathbf{x}$ for some $C < \infty$. Examples of such measures are Laplace measures and standard Brownian sheets. The most obvious problem with the statement is that equation (15) has no point-wise meaning. Thus, (15) is viewed as an equation for two random (tempered) distributions so the equation has to be interpreted in the weak sense

$$\mathcal{T} X(\varphi) = \dot{M}(\varphi), \tag{16}$$

where φ is in some appropriate space of test functions and $\mathcal{T} = (\kappa^2 - \Delta)^{\frac{\alpha}{2}}$. To describe the solutions of (15), the Sobolev spaces H_t of fractional order t are needed. These can be seen as an extension of the classical Sobolev space, of L_2 functions with all partial derivatives of order n or less in L_2, to fractional values of n. The following characterization of the solutions to (15) is then shown.

Proposition 8.2. *Assume that M is an independently scattered L_2-valued random measure with $\mathsf{E}(|M(\,\mathrm{d}\mathbf{x})|^2) = C\,\mathrm{d}\mathbf{x}$. Then for $\kappa > 0$, $\alpha > 0$, there exists a random functional $X : H_n \times \Omega \to \mathbb{R}$ such that for a certain set Ω_0, $P(\Omega_0) = 1$ and for all $\omega \in \Omega_0$ and all $\varphi \in H_n$*

$$X(\varphi, \omega) = \int G^\alpha \varphi(\mathbf{x}) M(\,\mathrm{d}\mathbf{x}, \omega),$$

where $G^\alpha \varphi(\mathbf{x}) = \int G_\alpha(\mathbf{s}, \mathbf{x}) \varphi(\mathbf{s})\,\mathrm{d}\mathbf{s}$ and

$$G_\alpha(\mathbf{s}, \mathbf{x}) = \frac{2^{1-\frac{\alpha-d}{2}}}{(4\pi)^{\frac{d}{2}} \Gamma(\frac{\alpha}{2}) \kappa^{\alpha-d}} (\kappa \|\mathbf{s} - \mathbf{x}\|)^{\frac{\alpha-d}{2}} K_{\frac{\alpha-d}{2}}(\kappa \|\mathbf{s} - \mathbf{x}\|).$$

This is the unique H_n-solution to (15) *if $n > d/2$, and moreover we have $X \in H_m$ almost surely for $m < \alpha - d/2$.*

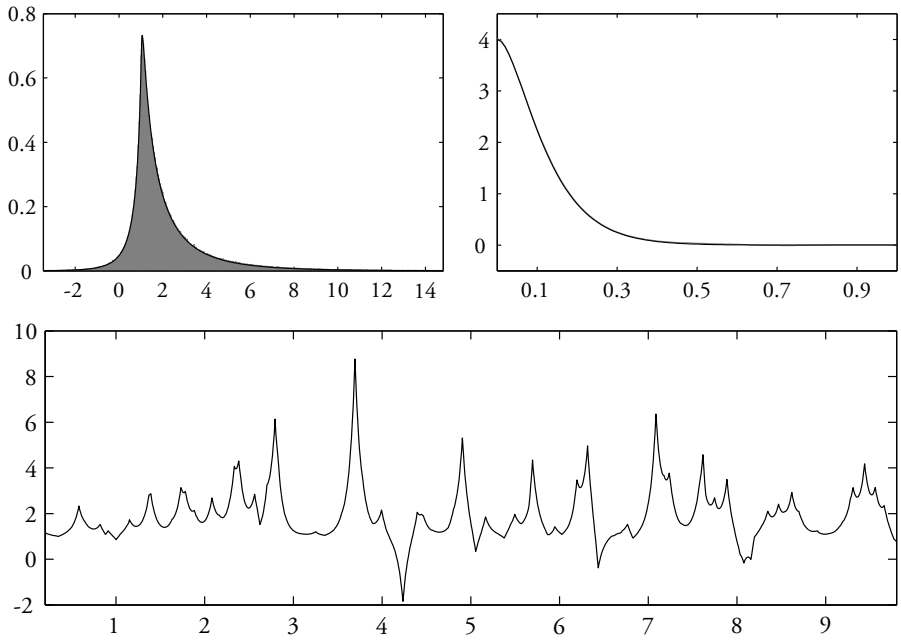

Figure 9: A simulation of a stationary Laplace-driven SPDE (15) on $\mathbb{R}$ (lower panel), its marginal density (upper left) and covariance function (upper right).

Using this proposition in combination with the Sobolev embedding theorem (see e.g. Adams, 1975), one finds the solution can be identified with a random function with a Matérn covariance function if $\alpha > d/2$. This fact is then used to extend the SPDE method by Lindgren et al. (2011) to a more general class of non-Gaussian models with Matérn covariances. An example of such a non-Gaussian process can be seen in Figure 9. As an application, the method is used for the Laplace moving average (LMA) models with Matérn covariances by Åberg and Podgórski (2011). The LMA model is specified as a process convolution where a Matérn kernel function is used and the Brownian sheet is replaced with a Laplace field. Parameter estimation for this model class has previously been performed using method of moments estimation (see Podgórski and Wegener, 2011). However, using the SPDE representation of the LMA model, a maximum likelihood parameter estimation method based on the EM algorithm is derived, as well as an efficient sampling method for the model class.

9 Comments on the papers

Paper A
Fast Estimation of Spatially Dependent Temporal Vegetation Trends using Gaussian Markov Random Fields

Bolin, D., Lindström, J., Eklundh, L., and Lindgren, F.

A summary of the paper is given in Section 4.2. The main contribution of the paper is the formulation of the spatially dependent regression model and the derivation of the estimation procedure for the model. The paper also contributes to the remote sensing community by using the spatially dependent regression model to analyze vegetation data from the African Sahel.

The method was developed in collaboration with J. Lindström, who also did parts of the writing. L. Eklundh wrote parts of the introduction. All other work was done by me, including all implementations and the main part of the writing.

Paper B
How do Markov approximations compare with other methods for large spatial data sets?

Bolin, D. and Lindgren, F.

A summary of the paper is given in Section 7.1. The paper contributes by extending the SPDE method using wavelet basis functions and provides a detailed comparison between the SPDE method, the process convolution method, and the covariance tapering method with respect to their ability to approximate Gaussian Matérn models.

The setup for the comparison was designed in collaboration with the co-author, who also did proofreading of the article and made several improvements to the manuscript. All other work was done by me, including all implementations.

Paper C
Spatial models generated by nested stochastic partial differential equations, with an application to global ozone mapping

Bolin, D. and Lindgren, F.

A summary of the paper is given in Section 8.1. The main contribution of the paper is the introduction of the class of nested SPDE models and the derivation

of its properties. Another contribution is a new representation of a basis that can be used for modeling vector fields on the sphere. The paper also addresses remote sensing problems by using the nested SPDE models to analyze total column ozone data.

Parts of the model formulations and the data analysis were done in collaboration with the co-author, who also did proofreading of the writing and made several improvements and clarifications to the text. All other work was done by me, including all implementations and proofs.

Paper D
Spatial Matérn fields driven by non-Gaussian noise

Bolin, D.

A summary of the paper is given in Section 8.2. The paper contributes by extending the SPDE method to non-Gaussian models using a careful analysis of a general class of SPDEs. It is also shown that the SPDE representation can be used to derive a maximum likelihood estimator and an efficient simulation method for Laplace moving average models.

Paper E
Excursion and contour uncertainty regions for latent Gaussian models

Bolin, D. and Lindgren, F.

A summary of the paper is given in Section 5. The main contributions of the paper are the definitions of the various excursion sets and uncertainty sets for level curves and the introduction of the excursion functions as a visual tool that can be used for illustrating uncertainty in these quantities. A new method for calculating these sets in practice is also introduced and the method is used to analyze the Sahel vegetation data and air pollution data from the Piemonte region in northern Italy.

The problem formulation was done in collaboration with the co-author, who also provided many of the various definitions, and made several improvements and clarifications to the text. All other work was done by me, including the construction of the method used for calculating the sets and all implementations.

References

Åberg, S. and Podgórski, K. (2011). A class of non-Gaussian second order random fields. *Extremes*, 14:187–222.

Åberg, S., Podgórski, K., and Rychlik, I. (2009). Fatigue damage assessment for a spectral model of non-Gaussian random loads. *Probab. Eng. Mech.*, 24:608–617.

Adams, R. A. (1975). *Sobolev Spaces*. Academic Press.

Adler, R. J. (1981). *The Geometry of Random Fields*. Wiley, New York.

Agbu, P. and James, M. (1994). *The NOAA/NASA Pathfinder AVHRR Land Data Set User's Manual*. Goddard Distributed Active Archive Center, NASA, Goddard Space Flight Center, Greenbelt.

Banerjee, S., Gelfand, A. E., Finley, A. O., and Sang, H. (2008). Gaussian predictive process models for large spatial data sets. *J. Roy. Statist. Soc. Ser. B Stat. Methodol.*, 70(4):825–848.

Barry, R. P. and Ver Hoef, J. M. (1996). Blackbox kriging: Spatial prediction without specifying variogram models. *J. Agr. Biol. Environ. Statist.*, 1(3):297–322.

Beaky, M. M., Scherrer, R. J., and Villumsen, J. V. (1992). Topology of large-scale structure in seeded hot dark matter models. *Astrophys. J.*, 387:443–448.

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

Besag, J., York, J., and Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics (with discussion). *Ann. Inst. of Statist. Math.*, 43:1–59.

Billingsley, P. (1986). *Probability and Measure*. John Wiley & Sons Ltd, 2nd edition.

Bochner, S. (1955). *Harmonic analysis and the theory of probability*. University of California press, Berkeley, CA.

Chui, C. K. and Wang, J.-Z. (1992). On compactly supported spline wavelets and a duality principle. *T. Am. Math. Soc.*, 330:903–915.

Cressie, N. (1991). *Statistics for Spatial Data*. John Wiley & Sons Ltd.

Cressie, N. and Johannesson, G. (2008). Fixed rank kriging for very large spatial data sets. *J. Roy. Statist. Soc. Ser. B Stat. Methodol.*, 70(1):209–226.

Cressie, N. and Ravlicová, M. (2002). Calibrated spatial moving average simulations. *Statist. Model.*, 2:267–279.

Cressie, N. and Wikle, C. (2011). *Statistics for Spatio-Temporal Data*. Wiley Series in Probability and Statistics. Wiley.

Daubechies, I. (1992). *Ten Lectures on Wavelets (CBMS-NSF Regional Conference Series in Applied Mathematics)*. Soc for Industrial & Applied Math.

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

Eklundh, L. and Olsson, L. (2003). Vegetation index trends for the African Sahel 1982–1999. *J. Geophys. Res.*, 30:1430–1433.

Fisher, R. A. (1926). The arrangement of field experiments. *J. Min. Agric. G. Br.*, 33:503–513.

Furrer, R., Genton, M. G., and Nychka, D. (2006). Covariance tapering for interpolation of large spatial datasets. *J. Comput. Graph. Statist.*, 15:502–523.

Gamerman, D., Moreira, A., and Rue, H. (2003). Space-varying regression models: specifications and simulation. *Comput. Statist. and Data Anal.*, 42:513–533.

Gelfand, A., Diggle, P., and Guttorp, P. (2010). *Handbook of spatial statistics*. Chapman & Hall/CRC handbooks of modern statistical methods. Taylor & Francis Group.

Genz, A. and Bretz, F. (2009). *Computation of Multivariate Normal and t Probabilities*, volume 195 of *Lecture Notes in Statistics*. Springer.

Higdon, D. (2001). Space and space-time modeling using process convolutions. Technical report.

Holthuijsen, L. H. (2007). *Waves in Oceanic and Coastal Waters*. Cambridge University Press.

Hrafnkelsson, B. and Cressie, N. (2003). Hierarchical modeling of count data with application to nuclear fall-out. *Environ. and Ecological Statist.*, 10(2):179–200.

James, M. and Kalluri, S. (1994). The Pathfinder AVHRR Land data set: An improved coarse resolution data set for terrestrial monitoring. *Internat. J. Remote Sensing*, 15:3347–3363.

Justice, C. and Hiernaux, P. (1986). Monitoring the grasslands of the Sahel using NOAA AVHRR data: Niger 1983. *Internat. J. Remote Sensing*, 7:1475–1497.

Krige, D. G. (1951). A statistical approach to some basic mine valuation problems on the Witwatersrand. *J. Chem., Metal. and Mining Soc. South Africa*, 52(6):119–139.

Lamb, P. (1982). Persistence of Subsaharan drought. *Nature*, 299:46–47.

Lindgren, F., Rue, H., and Lindström, J. (2011). 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.

Lindgren, G., Bolin, D., and Lindgren, F. (2010). Non-traditional stochastic models for ocean waves. *European Phys. J. - Special Topics*, 185:209–224.

Marchini, J. and Presanis, A. (2003). Comparing methods of analyzing fMRI statistical parametric maps. *NeuroImage*, 22:1203–1213.

Matérn, B. (1960). Spatial variation. *Meddelanden från statens skogsforskningsinstitut*, 49(5).

Matheron, G. (1963). Principles of geostatistics. *Econom. Geol.*, 58:1246–1266.

Matheron, G. (1971). The theory of regionalized variables, and its applications. *Centre de Geostatistique, Fontainebleau, Paris*.

Nicholson, S. (2000). Land surface process and Sahel climate. *Rev. of Geophys.*, 38:117–140.

Nychka, D., Wikle, C., and Royle, J. A. (2002). Multiresolution models for nonstationary spatial covariance functions. *Statist. Model.*, 2:315–331.

Olsson, L. (1993). On the causes of famine – drought, desertification and market failure in the Sudan. *Ambio*, 22:395–403.

Omre, H. and Halvørsen, K. (1989). The bayesian bridge between simple and universal kriging. *Mathematical Geology*, 21:767–786.

Podgórski, K. and Wegener, J. (2011). Estimation for stochastic models driven by Laplace motion. *Commun. Statist.- Theory Methods*, 40:3281–3302.

Robert, C. P. and Casella, G. (2004). *Monte Carlo Statistical Methods*. Springer, second edition.

Rodrigues, A. and Diggle, P. J. (2010). A class of convolution-based models for spatio-temporal processes with non-separable covariance structure. *Scand. J. Statist.*, 37:553–567.

Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields; Theory and Applications*, volume 104 of *Monographs on Statistics and Applied Probability*. Chapman & Hall/CRC.

Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations (with discussion). *J. Roy. Statist. Soc. Ser. B Stat. Methodol.*, 71(2):319–392.

Rue, H. and Tjelmeland, H. (2002). Fitting Gaussian Markov random fields to Gaussian fields. *Scand. J. Statist.*, 29(1):31–49.

Sampson, P. D. and Guttorp, P. (1992). Nonparametric estimation of nonstationary spatial covariance structure. *J. Amer. Statist. Assoc.*, 87(417):108–119.

Schabenberger, O. and Gotway, C. (2005). *Statistical methods for spatial data analysis*. Texts in statistical science. Chapman & Hall/CRC.

Seaquist, J., Olsson, L., and Ardö, J. (2003). A remote sensing-based primary production model for grassland biomes. *Ecological Model.*, 169:131–155.

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

Tobler, W. R. (1970). A computer movie simulating urban growth in the detroit region. *Econom. Geogr.*, 46:234–240.

Tucker, C., Vanpraet, C., Sharman, M., and Van Ittersum, G. (1985). Satellite remote sensing of total herbaceous biomass production in the Sengalese Sahel: 1980–1984. *Remote Sensing of Environ.*, 17:233–249.

Walsh, J. (1986). An introduction to stochastic partial differential equations. In *École d'Été de Probabilités de Saint Flour XIV - 1984*, volume 1180 of *Lecture Notes in Mathematics*, chapter 3, pages 265–439. Springer Berlin / Heidelberg.

Wendland, H. (1995). Piecewise polynomial, positive definite and compactly supported radial functions of minimal degree. *Adv. Comput. Math.*, 4:389–396.

Werner Hartman, L. (2006). Bayesian modelling of spatial data using Markov random fields, with application to elemental composition of forest soil. *Math. Geol.*, 38(2):113–133.

Whittle, P. (1954). On stationary processes in the plane. *Biometrika*, 41:434–449.

Whittle, P. (1963). Stochastic processes in several dimensions. *Bull. Internat. Statist. Inst.*, 40:974–994.

Wikle, C. K., Berliner, L. M., and Cressie, N. (1998). Hierarchical Bayesian space-time models. *Environ. and Ecological Statist.*, 5(2):117–154.

Woodbury, M. A. (1950). Inverting modified matrices. *S.R.G. Princeton*, Memo. Rep. n.o. 42.

A

Fast estimation of spatially dependent temporal vegetation trends using Gaussian Markov random fields

DAVID BOLIN[1], JOHAN LINDSTRÖM[1], LARS EKLUNDH[2], AND FINN LINDGREN[1]

[1] *Centre for Mathematical Sciences, Lund University, Lund, Sweden*
[2] *GeoBiosphere Science Centre, Lund University, Lund, Sweden*

Abstract: There is a need for efficient methods for estimating trends in spatio-temporal Earth Observation data. A suitable model for such data is a space-varying regression model, where the regression coefficients for the spatial locations are dependent. A second order intrinsic Gaussian Markov Random Field prior is used to specify the spatial covariance structure. Model parameters are estimated using the Expectation Maximisation (EM) algorithm, which allows for feasible computation times for relatively large data sets. Results are illustrated with simulated data sets and real vegetation data from the Sahel area in northern Africa. The results indicate a substantial gain in accuracy compared with methods based on independent ordinary least squares regressions for the individual pixels in the data set. Use of the EM algorithm also gives a substantial performance gain over Markov Chain Monte Carlo-based estimation approaches.

Key words: Gaussian Markov random fields; temporal trends; vegetation data; EM algorithm; spatio-temporal modeling; African Sahel

1 Introduction

Current awareness of global warming has directed scientific interest towards effective monitoring of land surface changes. Trends in vegetation cover are related to changes in climatic drivers, feedback mechanisms between the atmosphere and land surface, and human interaction. A region with rapid recent changes is the African Sahel. This zone has received much attention regarding desertification and climatic variations (Olsson, 1993, Nicholson, 2000, Lamb, 1982). The region saw an increase in rainfall in the 1960's, but since then, rainfall over the Sahel

has decreased severely (Hulme, 2001). Starting in the late 1960's, the area suffered droughts for over twenty years, culminating with a severe drought and famine in 1983-1984. Since the 1983-1984 famine, and because of the growing archive of meteorological satellite observations, several authors have used satellite imagery to study vegetation in the Sahel (Tucker et al., 1985, Justice and Hiernaux, 1986, Seaquist et al., 2003).

Recently, Eklundh and Olsson (2003) observed a strong increase in seasonal vegetation index over parts of the Sahel using Advanced Very High Resolution Radiometer (AVHRR) data from the NOAA/NASA Pathfinder AVHRR Land (PAL) database (Agbu and James, 1994, James and Kalluri, 1994), for the period 1982-1999. The study was based on ordinary least squares (OLS) linear regression on individual time series extracted for each pixel in the satellite images. The changes were interpreted as a vegetation recovery from the earlier drought periods (Olsson et al., 2005), and was found to be related to an increase in rainfall (Hickler et al., 2005).

However, the time period of space observations is relatively short and the data affected by numerous disturbances, e.g. inter-sensor calibration and satellite drift (Lindström et al., 2006), atmospheric water vapour (Tanré et al., 1992), aerosol variations (Hanan et al., 1995), geometric errors, clouds, and effects of anisotropic surface reflectance (Holben, 1986, Prince and Goward, 1996). The noisy data affect significance levels of the derived trends, reducing the possibility of generating unambiguous trend images. A drawback with the OLS method, is that spatial dependencies in the vegetation cover are neglected. The model presented in this work aims at generating a more efficient analysis by including these dependencies. For comparison, it is tested on the same data as were used in Eklundh and Olsson (2003).

There are several ways of incorporating spatial context into the analysis. A simple approach is to use smoothing, either on the calculated regression coefficients, or directly on the spatio-temporal data using three-dimensional smoothing (Fan and Gijbels, 1996, Di Giacinto et al., 2005). Another approach is to construct a Bayesian Hierarchical model (Wikle et al., 1998, Gelman et al., 2004), where the spatial dependence is introduced either by letting the regression errors be spatially dependent (Cressie, 1991, Chapters 2-5) or by utilising a spatially dependent prior on the regression coefficients (Besag et al., 1991, Gamerman et al., 2003).

Here, we formulate the regression based on a Bayesian hierarchical model.

The data is considered to be noisy observations of underlying fields. A prior on the regression coefficients is then introduced by assuming a spatial dependence structure for each of the underlying fields and further assuming that the time evolutions of the pixels in the fields are restricted to lie in the space spanned by the regression functions. This restriction introduces a spatial prior on the regression coefficients. The parameters of the model can, potentially, be estimated using a Markov Chain Monte Carlo (MCMC) based approach, but the large data-set makes this computationally infeasible (Bolin, 2007). Instead we estimate the model parameters using the Expectation Maximisation (EM) algorithm (Dempster et al., 1977), which allows for reasonably fast computations even when the studied data-set is large.

2 Statistical model

Assume that we have observations in an area of size $n_1 \times n_2$ pixels at the times $t = 0, 1, \ldots, T-1$. We will, from now on, assume that all images are column vectorised, so that an image is represented by a vector of size $n = n_1 n_2$. Denote an observation at time $t \in [0, T-1]$ by $Y_t \in \mathbb{R}^{n-n_t}$, where n_t is the number of missing data points at time t. We assume that each observation is generated as,

$$\mathbf{Y}_t | \mathbf{X}_t, \boldsymbol{\Sigma}_{\varepsilon_t} \in \mathsf{N}(\mathbf{A}_t \mathbf{X}_t, \boldsymbol{\Sigma}_{\varepsilon_t}),$$

where $\mathbf{X}_t$ is an underlying field with prior distribution $\pi(\mathbf{X}_t)$, $\boldsymbol{\Sigma}_{\varepsilon_t}$ is a measurement noise covariance matrix, and $\mathbf{A}_t$ is an observation matrix determining the observed pixels. That is, if the field is measured at the locations $i_1, \ldots, i_{n-n_t}$, the observation matrix, $\mathbf{A}_t$, will be of size $(n - n_t) \times n$, and have all elements equal to zero except for

$$A_{1,i_1} = \ldots = A_{n-n_t, i_{n-n_t}} = 1.$$

Assuming that $\mathbf{X}_{t_1}$ and $\mathbf{X}_{t_2}$ are independent for $t_1 \neq t_2$, the probability density for $\mathbf{X}^\top = (\mathbf{X}_0^\top, \ldots, \mathbf{X}_{T-1}^\top)^\top$ would be

$$\pi(\mathbf{X}) = \prod_{t=0}^{T-1} \pi(\mathbf{X}_t). \tag{1}$$

To estimate time varying trends in the observations, we introduce a $T \times m$-matrix $\mathbf{F} = [\mathbf{f}_1, \ldots, \mathbf{f}_m]$ containing basis functions for the trends. These basis functions

are chosen to be linearly independent, i.e. $\text{rank}(\mathbf{F}) = m$. Restricting $\mathbf{X}$ to follow these trends, we get

$$\mathbf{X} = (\mathbf{F} \otimes \mathbf{I})\mathbf{K}, \tag{2}$$

where $\mathbf{K} = [\mathbf{K}_1^\top, \ldots, \mathbf{K}_m^\top]^\top$ contains the coefficients for the trends and $\mathbf{I}$ is an $n \times n$ identity-matrix. The prior distribution for $\mathbf{K}$ is obtained by evaluating (1) conditionally on the restriction (2).

2.1 Gaussian Markov random fields

A suitable prior describing the spatial dependencies in $\mathbf{X}_t$ is a Gaussian Markov Random Field (GMRF) (see Rue and Held, 2005, for extensive details). A random variable $\mathbf{x} = (x_1, \ldots, x_n)^\top \in \mathsf{N}(\boldsymbol{\mu}, \mathbf{Q}^{-1})$ is called a GMRF if the joint distribution for $\mathbf{x}$ satisfies $\pi(x_i|x_{-i}) = \pi(x_i|x_{\mathcal{N}_i})\ \forall\ i$. Here $\mathcal{N}_i$ is the neighbourhood of i and x_{-i} denotes all elements in $\mathbf{x}$ except x_i. An important implication of this is that if $i \neq j$ then:

$$x_i \perp x_j | x_{-\{i,j\}} \iff Q_{i,j} = 0 \iff j \notin \mathcal{N}_i. \tag{3}$$

This means that the following properties are equivalent: 1) x_i and x_j are conditionally independent 2) the corresponding element in the precision matrix, $Q_{i,j}$ is zero and 3) i and j are not neighbours. Since $Q_{i,j} \neq 0$ only if i and j are neighbours, most GMRFs will have sparse precision matrices. The sparse precision matrix is the main reason for using GMRFs in this work; none of the following computations would be feasible for full matrices because of the large number of observations and nodes in the fields.

In the analysis, we will use a special type of GMRFs called Intrinsic Gaussian Markov Random Fields (IGMRFs). An IGMRF is improper, that is, its precision matrix does not have full rank.

Definition 2.1. Let $\mathbf{x} = (x_1, \ldots, x_n)^\top$, and let $\mathbf{Q}$ be an $n \times n$ symmetric positive semi-definite matrix with rank $n - k$, such that

$$Q_{i,j} \neq 0 \iff j \in \mathcal{N}_i.$$

Then $\mathbf{x}$ is an Intrinsic GMRF, with parameters $\boldsymbol{\mu}$ and $\mathbf{Q}$ if its density is

$$\pi(\mathbf{x}) = \frac{|\mathbf{Q}|_*^{\frac{1}{2}}}{(2\pi)^{\frac{n-k}{2}}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \mathbf{Q}(\mathbf{x} - \boldsymbol{\mu})\right). \tag{4}$$

Here, $|\mathbf{Q}|_*$ denotes the product of the $n-k$ non-zero eigenvalues of $\mathbf{Q}$, and following Rue and Held (2005), we will call this the generalised determinant of $\mathbf{Q}$. It should be noted that the density of an IGMRF is not a true density, since it is invariant to additions of vectors from the null-space of $\mathbf{Q}$. Hence, it cannot be normalised to integrate to one. The parameters, $\boldsymbol{\mu}$ and $\mathbf{Q}$, no longer represent the mean and precision, since these formally do not exist; however, we will for convenience continue denoting them as mean and precision.

We choose a second-order polynomial IGMRF (Gamerman et al., 2003, Rue and Held, 2005, Section 3.4.2) as a smooth prior. The field is invariant to the addition of an arbitrary plane, $p(i,j) = a + bi + cj$, and the precision matrix, $\kappa\mathbf{Q}$ has a rank deficiency of 3. Here κ is a scaling parameter that governs the strength of the dependence in the IGMRF. The second order IGMRF can be seen as a set of penalties on

$$(\Delta^2_{(1,0)} + \Delta^2_{(0,1)})x_{i,j}, \tag{5}$$

where $\Delta_{(1,0)}$ and $\Delta_{(0,1)}$ are forward-differences in the directions $(1,0)$ and $(0,1)$ respectively, and $\Delta^2_{(1,0)} + \Delta^2_{(0,1)}$ is a discrete approximation of the Laplace operator. Due to (5), this IGMRF can be interpreted as penalties on the second derivatives if the field is used to model an underlying continuous field.

2.2 Regression model

Using the second order IGMRF, described above, a distribution for each $\mathbf{X}_t$ in (1), with an unknown precision parameter κ, we get that $\mathbf{X}|\kappa \in \mathsf{N}(0, (\kappa\overline{\mathbf{Q}})^{-1})$, where $\overline{\mathbf{Q}} = \mathbf{I} \otimes \mathbf{Q}_X$ and $\mathbf{Q}_X$ is the precision matrix for a second order IGMRF. Given the restriction in (2), we can now derive the distribution of the parameter field, $\mathbf{K}$.

Proposition 2.2. $\mathbf{K}|\kappa$ *is an IGMRF of rank* $m(n-3)$ *with zero mean and covariance matrix* $(\kappa\mathbf{Q})^{-1}$*, where* $\mathbf{Q} = (\mathbf{F}^\top\mathbf{F}) \otimes \mathbf{Q}_X$.

For proof see Appendix A.

Note that the sparsity structure of $\mathbf{Q}$ is partially determined by the orthogonality of the regression basis, $\mathbf{F}$, see Figure 1.

Using the distribution for $\mathbf{K}|\kappa$ from Proposition 2.2, we can now write the observations, $\mathbf{Y}$, as $\mathbf{Y}|\mathbf{K}, \boldsymbol{\Sigma}_\varepsilon \in \mathsf{N}(\mathbf{AK}, \boldsymbol{\Sigma}_\varepsilon)$. Here, $\boldsymbol{\Sigma}_\varepsilon$ is a block diagonal matrix

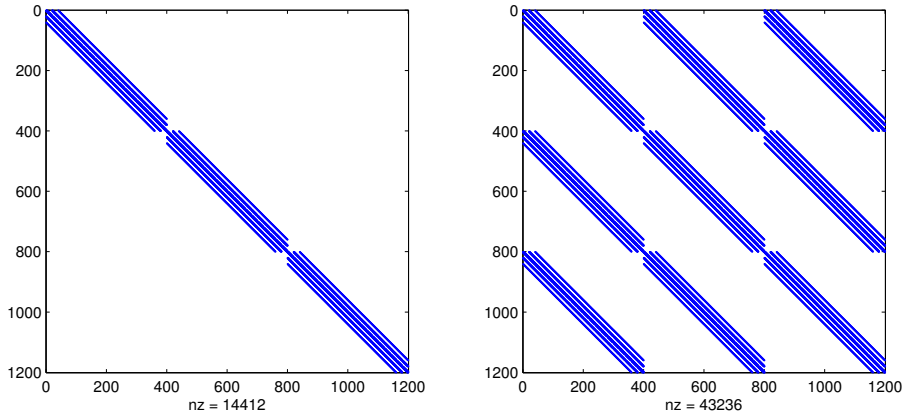

Figure 1: The figure depicts the sparsity structure in $\mathbf{Q}$, Proposition 2.2, for a field with 400 nodes and three regression basis vectors for orthogonal regression basis (left) and non-orthogonal regression basis (right) respectively. The number of non-zero elements in each of the matrices is denoted *nz*.

such that

$$\boldsymbol{\Sigma}_\varepsilon = \begin{bmatrix} \boldsymbol{\Sigma}_{\varepsilon_0} & & 0 \\ & \ddots & \\ 0 & & \boldsymbol{\Sigma}_{\varepsilon_{T-1}} \end{bmatrix},$$

and $\mathbf{A} = \text{diag}(\mathbf{A}_1, \ldots, \mathbf{A}_{T-1})(\mathbf{F} \otimes \mathbf{I})$, where $\mathbf{I}$ is an $n \times n$ identity matrix. The posterior distribution for $\mathbf{K}$ given data, $\boldsymbol{\Sigma}_\varepsilon$, and κ becomes

$$\begin{aligned} \pi(\mathbf{K}|\mathbf{Y}, \boldsymbol{\Sigma}_\varepsilon, \kappa) &\propto \pi(\mathbf{Y}|\mathbf{K}, \boldsymbol{\Sigma}_\varepsilon)\pi(\mathbf{K}|\kappa)\pi(\kappa)\pi(\boldsymbol{\Sigma}_\varepsilon) \\ &\propto \pi(\mathbf{Y}|\mathbf{K}, \boldsymbol{\Sigma}_\varepsilon)\pi(\mathbf{K}|\kappa). \end{aligned}$$

This follows since we choose to use flat priors for κ and $\boldsymbol{\Sigma}_\varepsilon$, i.e. $\pi(\kappa) \propto 1$ and $\pi(\boldsymbol{\Sigma}_\varepsilon) \propto 1$. It can be shown (Rue and Held, 2005, p. 39) that the posterior distribution of $\mathbf{K}$ is $\mathbf{K}|\mathbf{Y}, \boldsymbol{\Sigma}_\varepsilon, \kappa \in \mathsf{N}(\boldsymbol{\mu}_{K|Y}, \mathbf{Q}_{K|Y}^{-1})$, with

$$\boldsymbol{\mu}_{K|Y} = \mathbf{Q}_{K|Y}^{-1}\mathbf{A}^\top\boldsymbol{\Sigma}_\varepsilon^{-1}\mathbf{Y} \qquad \text{and} \qquad \mathbf{Q}_{K|Y} = \kappa\mathbf{Q} + \mathbf{A}^\top\boldsymbol{\Sigma}_\varepsilon^{-1}\mathbf{A}. \tag{6}$$

2.3 Noise model

To complete the model we need a noise model, i.e. we need to determine the structure of $\mathbf{\Sigma}_\varepsilon$. Many of the factors that the measurement noise should model are local phenomena, such as aerosol and cloud cover. Since it seems unreasonable that the scale of these disturbances would be the same over the entire, large, region, we assume a noise model with a different noise variance at each pixel. Further, for simplicity, we take the noise to be spatially uncorrelated. Thus $\mathbf{\Sigma}_\varepsilon$ is a $(nT - \sum_t n_t) \times (nT - \sum_t n_t)$ diagonal matrix with n unique diagonal elements, $\sigma_1^2, \ldots, \sigma_n^2$, which represent the noise variances of the n different pixels. Note that $\mathbf{\Sigma}_\varepsilon$ can be constructed by starting from the matrix $\mathbf{I} \otimes \text{diag}(\sigma_1^2 \ \ldots \ \sigma_n^2)$, where $\mathbf{I}$ is a $T \times T$ identity-matrix, and then removing all rows and columns that correspond to missing observations.

3 EM parameter estimation

By construction, the structure of the precision matrix, $\mathbf{Q}$, in (6) is known a priori, whereas the parameters κ and $\mathbf{\Sigma}_\varepsilon$ have to be estimated. These parameters can, potentially, be estimated using an MCMC based approach. However, two difficulties arise when attempting to estimate this model with MCMC. Firstly, given the large number of variance parameters in $\mathbf{\Sigma}_\varepsilon$, parameters that are likely to be correlated, we find it very challenging to construct an efficient proposal distribution that ensures good mixing of the Markov chain. Secondly, even if this issue is avoided by assuming a simpler measurement noise model with one common variance parameter for all the pixels in the field, the large data-set makes an MCMC based estimation approach computationally infeasible (see Bolin, 2007, for details).

A better alternative is to interpret the problem as a missing data problem. This interpretation facilitates use of the EM algorithm, allowing for much faster computations of maximum likelihood parameter estimates (a comparison of computational times for simulated data is given in the last paragraph of Section 6.1).

Augmenting the observed data, $\mathbf{Y}$, with the unknown regression coefficients, $\mathbf{K}$, we obtain the complete data, $(\mathbf{Y}, \mathbf{K})$, and the augmented likelihood becomes

$$L(\kappa, \mathbf{\Sigma}_\varepsilon | \mathbf{Y}, \mathbf{K}) = \pi(\mathbf{Y}, \mathbf{K} | \kappa, \mathbf{\Sigma}_\varepsilon) = \pi(\mathbf{Y} | \mathbf{K}, \mathbf{\Sigma}_\varepsilon)\pi(\mathbf{K} | \kappa).$$

Taking the parameters as, $\theta = (\kappa, \mathbf{\Sigma}_\varepsilon)$, the loss-function is

$$\mathcal{Q}(\theta, \theta^{(i)}) = \mathsf{E}(\log(L(\theta | \mathbf{K}, \mathbf{Y})) | \mathbf{Y}, \theta^{(i)}), \tag{7}$$

where $\theta^{(i)}$ is an estimate of θ at iteration i, and the expectation is taken over $\mathbf{K}$.

The likelihood consists of two parts, $\pi(\mathbf{Y}|\mathbf{K}, \mathbf{\Sigma}_\varepsilon)$ and $\pi(\mathbf{K}|\kappa)$. For the first part we have that $(\mathbf{Y}|\mathbf{K}, \mathbf{\Sigma}_\varepsilon) \in \mathsf{N}(\mathbf{AK}, \mathbf{\Sigma}_\varepsilon)$, hence

$$\begin{aligned}\log(\pi(\mathbf{Y}|\mathbf{K}, \mathbf{\Sigma}_\varepsilon)) &= \log\left(\frac{1}{(2\pi)^{\frac{n}{2}}|\mathbf{\Sigma}_\varepsilon|^{\frac{1}{2}}}\right) - \frac{1}{2}(\mathbf{Y}-\mathbf{AK})^\top \mathbf{\Sigma}_\varepsilon^{-1}(\mathbf{Y}-\mathbf{AK}) \\ &= -\frac{1}{2}\sum_{j=1}^{n}\left(n_j \log(\sigma_j^2) + \frac{1}{\sigma_j^2}\sum_{k=1}^{n_j}(\mathbf{Y}-\mathbf{AK})_{j_k}^2\right) + \text{const.}\end{aligned}$$

Here, n_j is the number of observations at pixel j (not to be confused with n_t, the number of missing observations at time t), and the last sum runs over these observations, with $(\mathbf{Y}-\mathbf{AK})_{j_k}$ referring to the deviation of the kth observation of pixel j, from the regression line. Similarly, since $\mathbf{K}|\kappa \in \mathsf{N}(0, (\kappa\mathbf{Q})^{-1})$, we get

$$\log(\pi(\mathbf{K}|\kappa)) = \frac{1}{2}\log(|\kappa\mathbf{Q}|_*) - \frac{\kappa}{2}\mathbf{K}^\top\mathbf{QK} + \text{const.} \tag{8}$$

By Proposition 2.2, the rank of $\mathbf{Q}$ is $m(n-3)$, where n is the number of pixels in the image, giving

$$\log(\pi(\mathbf{K}|\kappa)) = \frac{m(n-3)}{2}\log(\kappa) - \frac{\kappa}{2}\mathbf{K}^\top\mathbf{QK} + \text{const.} \tag{9}$$

Thus, the loss-function becomes

$$\begin{aligned}\mathcal{Q}(\theta, \theta^{(i)}) = &\frac{m(n-3)}{2}\log(\kappa) - \frac{\kappa}{2}\mathsf{E}(\mathbf{K}^\top\mathbf{QK}|*) \\ &- \frac{1}{2}\sum_{j=1}^{n}\left(n_j\log(\sigma_j^2) + \frac{1}{\sigma_j^2}\sum_{k=1}^{n_j}\mathsf{E}((\mathbf{Y}-\mathbf{AK})_{j_k}^2|*)\right) + \text{const.},\end{aligned} \tag{10}$$

where the notation $\mathsf{E}(\ldots|*)$ means $\mathsf{E}(\ldots|\mathbf{Y}, \kappa^{(i)}, \mathbf{\Sigma}_\varepsilon^{(i)})$.

To calculate $\arg\max_\theta \mathcal{Q}(\theta, \theta^{(i)})$, we differentiate (10) with respect to the parameters κ and $\sigma_1^2, \ldots, \sigma_n^2$ and set these derivatives equal to zero, yielding

$$\begin{aligned}\sigma_j^{2(i+1)} &= \frac{1}{n_j}\sum_{k=1}^{n_j}\mathsf{E}((\mathbf{Y}-\mathbf{AK})_{j_k}^2|*), \text{ for } 1 \le j \le n, \text{ and} \\ \kappa^{(i+1)} &= \frac{m(n-3)}{\mathsf{E}(\mathbf{K}^\top\mathbf{QK}|*)}.\end{aligned} \tag{11}$$

Recall that these updating equations are derived under the assumption of flat priors for κ and $\sigma_1^2, \cdots, \sigma_n^2$. However it can easily be shown that Gamma and independent inverse-Gamma priors for κ and $\sigma_1^2, \cdots, \sigma_n^2$ respectively will yield tractable updating expressions in (11).

In general, the expectations in (11) can now be found by simulating from the posterior density, $\pi(\mathbf{K}|\mathbf{Y}, \theta^{(i)})$, using a Monte Carlo-approach. However, in this case the observations given the underlying field, $\pi(\mathbf{Y}|\mathbf{K}, \mathbf{\Sigma}_\varepsilon^{(i)})$, are Gaussian which in turn implies that the posterior density, $\pi(\mathbf{K}|\mathbf{Y}, \theta^{(i)})$, is Gaussian and we can calculate the updating rules in (11) analytically. The first expectation in (11) becomes

$$\begin{aligned}\mathsf{E}((\mathbf{Y}-\mathbf{AK})_{j_k}^2|*) &= Y_{j_k}^2 - 2Y_{j_k}\mathbf{A}_{(j_k,\bullet)}\mathsf{E}(\mathbf{K}|*) + \mathsf{E}((\mathbf{A}_{(j_k,\bullet)}\mathbf{K})^2|*) \\ &= (Y_{j_k} - \mathbf{A}_{(j_k,\bullet)}\mathsf{E}(\mathbf{K}|*))^2 + \mathbf{A}_{(j_k,\bullet)}\mathsf{V}_{K|*}\mathbf{A}_{(j_k,\bullet)}^\top. \end{aligned} \tag{12}$$

Here, $\mathbf{A}_{(j_k,\bullet)}$ denotes the row in $\mathbf{A}$ corresponding to the kth observation of pixel j, and $\mathsf{E}(\mathbf{K}|*) = \mu_{K|Y}$ and $\mathsf{V}_{K|*} = \mathbf{Q}_{K|Y}^{-1}$ are given by (6).

Using that both the expected value and the trace of a matrix are linear operators, and the cyclic property of the trace, $\text{tr}(\mathbf{ABC}) = \text{tr}(\mathbf{CAB})$, we can calculate $\mathsf{E}(\mathbf{K}^\top\mathbf{QK}|*)$ as

$$\begin{aligned}\mathsf{E}(\mathbf{K}^\top\mathbf{QK}|*) &= \text{tr}(\mathsf{E}(\mathbf{K}^\top\mathbf{QK}|*)) = \text{tr}(\mathbf{Q}\mathsf{E}(\mathbf{KK}^\top|*)) \\ &= \text{tr}(\mathbf{Q}(\mathbf{Q}_{K|Y}^{-1} + \mathsf{E}(\mathbf{K}|*)\mathsf{E}(\mathbf{K}|*)^\top)) \\ &= \text{tr}(\mathbf{QQ}_{K|Y}^{-1}) + \mathsf{E}(\mathbf{K}|*)^\top\mathbf{Q}\mathsf{E}(\mathbf{K}|*). \end{aligned} \tag{13}$$

Calculating $\text{tr}(\mathbf{QQ}_{K|Y}^{-1})$ and $\mathbf{A}_{(j_k,\bullet)}\mathbf{Q}_{K|Y}^{-1}\mathbf{A}_{(j_k,\bullet)}^\top$ in (12) and (13) might not seem feasible for large fields. However, three things should be noted: Firstly

$$\mathbf{A}_{(j_k,\bullet)}\mathbf{Q}_{K|Y}^{-1}\mathbf{A}_{(j_k,\bullet)}^\top = \text{tr}(\mathbf{A}_{(j_k,\bullet)}^\top\mathbf{A}_{(j_k,\bullet)}\mathbf{Q}_{K|Y}^{-1}).$$

Secondly, due to the trace operator, only the diagonal elements of the products $(\bullet\mathbf{Q}_{K|Y}^{-1})$ need to be calculated. Thirdly $\mathbf{Q}$ and $\mathbf{A}_{(j_k,\bullet)}^\top\mathbf{A}_{(j_k,\bullet)}$ are at least as sparse as $\mathbf{Q}_{K|Y}$, thus to calculate the traces we will *at most* need the elements of $\mathbf{Q}_{K|Y}^{-1}$ that correspond to neighbouring points in the GMRF, given the Cholesky factor these elements can be calculated *without calculating the entire inverse* (Rue and Martino, 2007), making (12) and (13) computationally feasible.

4 Testing for significant trends

A relevant question to ask is where changes in the vegetation have occurred over the course of the studied time period. A method which has been used to answer this question for the Sahel data set, is to find the significant trends in the data (Eklundh and Olsson, 2003). For an OLS regression this is relatively straightforward, although one has to keep in mind that the obtained significant trends describe whether trends at each pixel are individually significant, and not if the trends for all pixels in an entire region are significant. To obtain the latter, one has to consider techniques for multiple hypothesis testing, one option would be to do hypothesis testing on the contour lines of the K_i estimates. How to do such tests is, however, far from trivial, and will not be investigated here.

Using the GMRF model and the parameter estimation algorithm described above, maximum-likelihood estimates of κ and $\sigma_1^2, \cdots, \sigma_n^2$ are obtained. Given these, the conditional posterior for the **K**-field is

$$\mathbf{K}|\mathbf{Y}, \kappa, \sigma_1^2, \cdots, \sigma_n^2 \in \mathsf{N}\left(\boldsymbol{\mu}_{K|Y}, \mathbf{Q}_{K|Y}^{-1}\right),$$

with $\boldsymbol{\mu}_{K|Y}$ and $\mathbf{Q}_{K|Y}$ defined in (6).

The conditional variances of the **K**-estimates are given by the diagonal elements in $\mathbf{Q}_{K|Y}^{-1}$. Let k_i^j be the coefficient corresponding to the trend $\mathbf{f}_j$ at pixel x_i, and let $\hat{\sigma}_i^j$ denote the corresponding standard deviation given by the square root of the relevant diagonal element in $\mathbf{Q}_{K|Y}^{-1}$. A simple hypothesis test would now be to reject the null hypothesis $H_0 : k_i^j = 0$ against $H_1 : k_i^j \neq 0$, at a 95% confidence level if

$$\left|\frac{k_i^j}{\hat{\sigma}_i^j}\right| > \lambda_{0.025},$$

where $\lambda_{0.025}$ is the 2.5%-quantile of the $\mathsf{N}(0, 1)$-distribution.

Simulation studies (see Table 1) indicate that this test rejects the null hypothesis too often, i.e. the corresponding confidence interval for k_i^j is too small. The problem is that the uncertainty in the estimated parameters κ and $\sigma_1^2, \cdots, \sigma_n^2$ are ignored. A solution to this problem is to instead, as in OLS regression, reject the null hypothesis if

$$\left|\frac{k_i^j}{\hat{\sigma}_i^j}\right| > t_{0.025}(f), \tag{14}$$

where $t_{0.025}(f)$ is the 2.5%-quantile of the Student's t-distribution with f degrees of freedom. To perform this test, we have to determine f, which in OLS depends on the number of data points used to estimate k_i^j, and thus $\hat{\sigma}_i^j$. Since κ determines the smoothness in the parameter field, a large κ-value would imply that the regression parameters are highly correlated. In this case, the estimate of a parameter at a specific pixel is influenced, not only by the observations at that pixel, but also by observation of surrounding pixels. On the other hand, if $\kappa = 0$, the model breaks down to n independent OLS regressions, and each parameter is estimated using only the observations of the corresponding pixel. Thus, it is reasonable to assume that f will be dependent on κ. It is, however, not easy to determine an analytic expression for f, and we therefore use a simulation based algorithm for estimating the degrees of freedom: Given estimated values of κ and $\mathbf{\Sigma}_\varepsilon$, we simulate new $\mathbf{K}$-fields and corresponding data sets. From these data, we use the EM algorithm to calculate new ML-estimates, $\kappa^*, \Sigma_\varepsilon^*$, and new posterior means $\mathbf{K}^* = \mathsf{E}(\mathbf{K}|\mathbf{Y}, \kappa^*, \mathbf{\Sigma}_\varepsilon^*)$. The residuals, $\boldsymbol{\epsilon} = \sum_j \mathbf{f}_j^\top \otimes (\mathbf{K}_j - \mathbf{K}_j^*)$, are calculated and each residual, $\epsilon_{i,t}$, is normalised by division with its corresponding standard deviation $\mathsf{V}(\sum_j f_j(t) K_j^*(i))^{1/2}$. In standard OLS regression, the residuals would now follow a Student's-t distribution with unknown degrees of freedom. Thus we estimate f by maximising the log-likelihood, of the t-distribution $\sum_{j,t} \log t_f(\epsilon_{j,t})$, with respect to f, using a simplex search method (Lagarias et al., 1998). Simulation studies, Table 1, indicate that (14) with f estimated as described above gives correct significance levels.

5 Data

Many Earth Observation studies use Normalised Difference Vegetation Index (NDVI) data (Rouse et al., 1973) as a remotely sensed measure of ground vegetation. The principle behind the NDVI is that leaf chlorophyll absorbs red light, whereas the cell structure of leaves reflects light in the near-infrared spectra. With these properties in mind, NDVI is calculated as

$$\text{NDVI} = \frac{R_{\text{NIR}} - R_{\text{RED}}}{R_{\text{NIR}} + R_{\text{RED}}}, \tag{15}$$

where R_{NIR} and R_{RED} are the measured reflectancies in the near-infrared and red spectral bands, respectively. Thus, dense vegetation tends to have positive NDVI values, while soil and other areas with little vegetation tend to have NDVI

values close to zero. Many studies have confirmed that NDVI correlates well with important vegetation variables, e.g. biomass, and vegetation greenness (Myneni et al., 1997, Sellers et al., 1997).

The Sahel is a semi-arid region, and it should be noted that NDVI data is well suited for studies of such areas; because of the low humidity, the atmospheric contamination of the data is fairly low (Chappell et al., 2001), and because of the sparse vegetation, saturated NDVI values are less of a problem than for areas with denser vegetation cover (Prince, 1991).

5.1 Data preprocessing

The PAL database is derived from measurements by the AVHRRs on-board the NOAA series of meteorological satellites (Kidwell, 1998). Before the NDVI values are calculated, the AVHRR data undergoes a number of calibration steps (see Agbu and James, 1994, for details). The data is then mapped to 8 km $\times$ 8 km pixels and 10-day maximum value composites are generated to compensate for negative bias due to clouds (Holben, 1986). In spite of these corrections, the data still contains unwanted noise.

The end result of the above processing is a set of 10-day maximum value composites, for a total of 36 values per year. However we are interested in assessing the year-to-year change in vegetation growth. Therefore the seasonal NDVI integral which gives a measure of the total vegetation growth during each year is of interest.

In order to generate annual data, we use the Savitzky and Golay (1964) algorithm of the TIMESAT processing scheme (Jönsson and Eklundh, 2002, 2004). The method begins by estimating the number of annual growing seasons and the corresponding season lengths. When the approximate onsets and ends of the growing seasons have been found, the time series is smoothed using a robust locally weighted least-squares fit of a quadratic polynomial to the upper envelope of the NDVI data (Press et al., 1992). The fit is adapted to the envelope of the data to account for negatively biased noise, and the weights in the least-squares fitting are determined using cloud cover data from the NOAA/NASA Cloud AVHRR (CLAVR) data set (Stowe et al., 1991). To obtain annual data, the seasonal integrals are calculated from the fitted functions and used in the following analysis. To further reduce the amount of noise in the data, all seasons which have more than one missing observation are removed. Also, to avoid artificial growing seasons in desert areas, all seasons which have less than two NDVI values above 0.1 are set to zero.

6 Results

Recall that the model presented in Section 2.2 (from now on referred to as the GMRF model) is to be used for estimating trends in vegetation data. In this section, the model is compared to simple independent OLS regression for each pixel (from now on referred to as the OLS model), using simulated data and real NDVI data.

6.1 Simulated data

The two models are compared, using 17 years of simulated data. For these tests, we use two orthogonal trend basis functions, one constant and one linearly increasing. Hence, $\mathbf{K}_1$ will contain the intercepts, and $\mathbf{K}_2$ the slopes for the linear trends.

To create a spatially dependent data set, $\mathbf{Y} = (\mathbf{Y}_0, \ldots, \mathbf{Y}_{16})$, we generate dependent $\mathbf{K}_1$- and $\mathbf{K}_2$-fields from the distribution in Proposition 2.2, with $\kappa = 0.5$. To create a data set without spatial dependencies, we instead generate $\mathbf{K}_1$ and $\mathbf{K}_2$ by independently drawing values from $\mathsf{N}(0, 1)$. In both cases, $\mathbf{Y}_t$, $t \in [0, 16]$, is created as $\mathbf{Y}_t = f_1(t)\mathbf{K}_1 + f_2(t)\mathbf{K}_2 + \varepsilon$, where ε is a draw from $\mathsf{N}(0, \mathbf{\Sigma}_{\varepsilon_t})$ and $\mathbf{\Sigma}_{\varepsilon_t}$ is defined as in Section 2.3, with pixel variances σ_i^2 drawn independently from a uniform distribution on $[0, \alpha]$. We use images of size 40×40 pixels in four different cases:

D1	Spatially dependent data with $\kappa = 0.5$ and $\alpha = 1$.
D10	Spatially dependent data with $\kappa = 0.5$ and $\alpha = 10$.
R1	Random data ($\kappa = 0$) with $\alpha = 1$.
R10	Random data with $\alpha = 10$.

The accuracy of the models are compared using the following tests:

1. Calculate the percentage of pixels correctly labelled as significant trends, i.e. the number of significant trends with correct sign divided by the total number of pixels.
2. Estimate the number of times that the confidence intervals for the estimated trend coefficients cover the actual value. This number should be close to 95%.
3. Compare the estimated $\mathbf{K}$-fields with the actual $\mathbf{K}$-fields using the Frobenius norm: $K_i^{\epsilon} = \|\mathbf{K}_i - \mathbf{K}_i^*\|_F$.

		K_1^ϵ	K_2^ϵ	Σ^ϵ	Cov (%)	Sig (%)
D1	OLS	6.84	1.39	8.40	95.01	94.73
	GMRF	3.63	0.74	8.09	94.87	97.12
D10	OLS	21.66	4.42	84.43	95.06	84.40
	GMRF	6.27	1.28	79.86	94.85	95.41
R1	OLS	6.83	1.40	8.38	94.96	94.43
	GMRF	6.81	1.43	11.92	94.88	94.42
R10	OLS	21.63	4.45	83.75	94.90	82.78
	GMRF	21.02	4.57	84.57	94.44	82.58

Table 1: Results for simulated data. Each value is the average over 100 different runs. Here, K_1^ϵ, K_2^ϵ, and Σ^ϵ show the difference, measured in the Frobenius norm, of the true and estimated $\mathbf{K}_1$- , $\mathbf{K}_2$- and Σ_ε-fields, respectively. Cov is the estimated coverage percentage for the $\mathbf{K}_2$ confidence intervals, which should be close to the nominal 95%. Sig is the percentage of correctly labeled significant trends.

4. Compare the $\mathbf{\Sigma}_\varepsilon$ estimates with the actual measurement noise variances: $\Sigma^\epsilon = \|\mathbf{\Sigma}_\varepsilon - \mathbf{\Sigma}_\varepsilon^*\|_F$.

Each of the tests are done on 100 different data sets for each of the four cases described above. The results for the tests are summarised in Table 1. As seen in the table, the two models are fairly similar for random data whereas the GMRF-estimations of the **K**-fields are more accurate for spatially dependent data. This can be seen in Figure 2 where generated **K**-fields are shown together with the two estimates. The lower variance of the estimates also results in a higher percentage of correctly labelled significant trends for spatially dependent data. The GMRF algorithm is implemented in C/C++ using GMRFLib (Rue, 2007). For these simulations, the average time for each step in the EM algorithm was 0.26 seconds, which, for example, resulted in an average computation time of 9.80 seconds for the **D1** simulations. The computations were performed on a dual CPU (2×2.1GHz) personal computer. For comparison, an MCMC based estimation algorithm was also implemented in C/C++, assuming a simpler measurement noise model with one common variance parameter for all the pixels in the field (see Bolin, 2007, for details). The average computation time for one iteration in this algorithm was 0.11 seconds, which resulted in a computation time of 1050 seconds for parameter estimates based on 10^4 iterations.

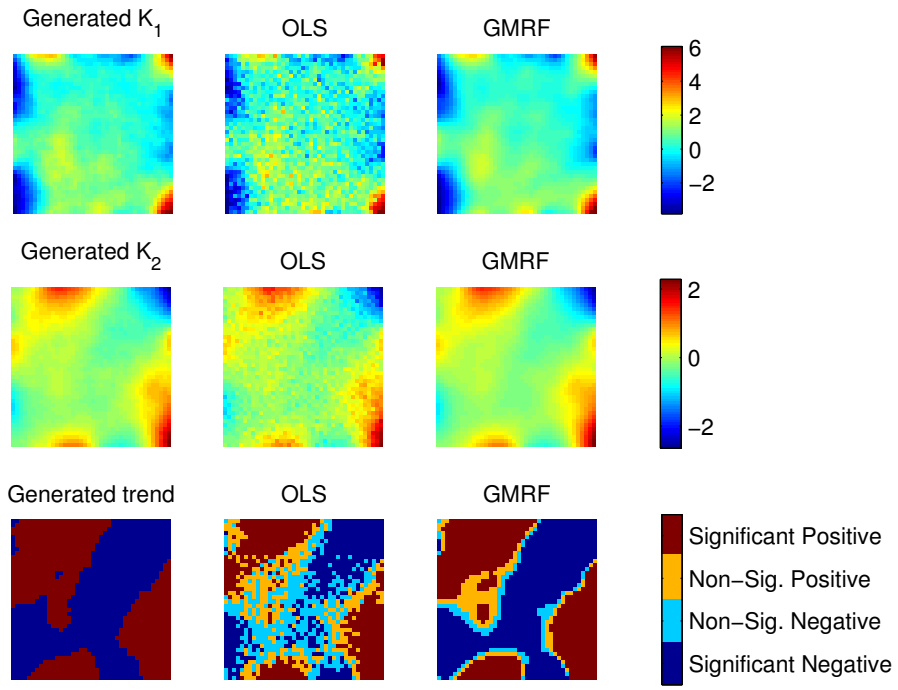

Figure 2: Results from a simulated data set with $\kappa = 0.5$ and $\alpha = 10$. The two upper rows show the true $\mathbf{K}_1$- and $\mathbf{K}_2$-fields and the estimates generated by the OLS model and the GMRF model. Notice that the GMRF estimates are smoother and more accurate. On the bottom row, the left figure indicates where the true $\mathbf{K}_2$-field is positive (red) and negative (blue). The other two figures show the significance estimates for the $\mathbf{K}_2$-trends. Notice that more correct significant trends are found by the GMRF model.

6.2 Sahel data

We base the analysis on the seasonal NDVI integrals of the 18 years of data. The results from the previous section indicate that more accurate vegetation trend estimates can be obtained using the GMRF-model instead of independent OLS regression for each pixel.

The GMRF model assumes as common precision parameter for the entire field, which is probably not valid for the entire Sahel. Therefore, the following results were obtained by dividing the area into smaller subareas, with a (large)

overlap between the subareas, and applying the GMRF model to each of these subareas. This way, κ is assumed to be constant only on the smaller subareas, and the overlap between neighbouring subareas ensure that no discontinuities appear in the joints between the subareas.

Using two orthogonal trend basis functions, one constant, and one linearly increasing, we obtain estimates shown in Figures 3 and 4. In the figures, we see that the estimates are somewhat similar, but that the GMRF estimates are much smoother than the OLS estimates. A big difference in the estimates can be seen in Figure 5. Here, the areas with significant linear trends are shown, and we see that, compared to the OLS, the GMRF estimate has more significant trends and larger contiguous regions with significant trends. This result is expected since the GMRF estimate takes advantage of the spatial dependencies in the data, and therefore uses more observations for each point estimate. As indicated in Section 4, contour plots of the $\mathbf{K}_i$ estimates might also be of interest for visualisation and hypothesis testing. In Figure 6, a contour plot of the GMRF K_2 estimate is shown. In this figure, it is easy to see where large increases in vegetation has occurred. This is a big advantage with the GMRF method, since similar contour plots for the OLS estimate are too noisy to easily interpret. The larger contiguous regions and smoother estimates will most likely aid future interpretation of the data and make it easier to detect underlying reasons for the detected changes in vegetation.

Besides detecting purely linear trends, it is also of interest to find areas which experience a large increase or decrease in seasonal NDVI during the first half of the time period, and a lower increase or decrease during the second half. To do this, we add a third trend basis function, $f_3 = [\frac{T}{2}, \frac{T}{2} - 1, \ldots, 0, 1, \ldots, \frac{T}{2}]$, normalised to have length one. Contour plots of the resulting field, $\mathbf{K}_3$, can be seen in Figure 7. The figure contains two different contour plots, the first one shows the contours of the $\mathbf{K}_3$ field for pixels where the $\mathbf{K}_2$ field is positive. A positive value in this field means that there was a higher increase in vegetation growth during the second part of the time period, whereas a negative value corresponds to a higher increase in vegetation growth during the first part of the time period. The second plot shows the contours for the pixels where the $\mathbf{K}_2$ field is negative. Here, a positive value means that there was a larger decrease in vegetation during the first part, whereas a negative value corresponds to a larger decrease in vegetation during the second half.

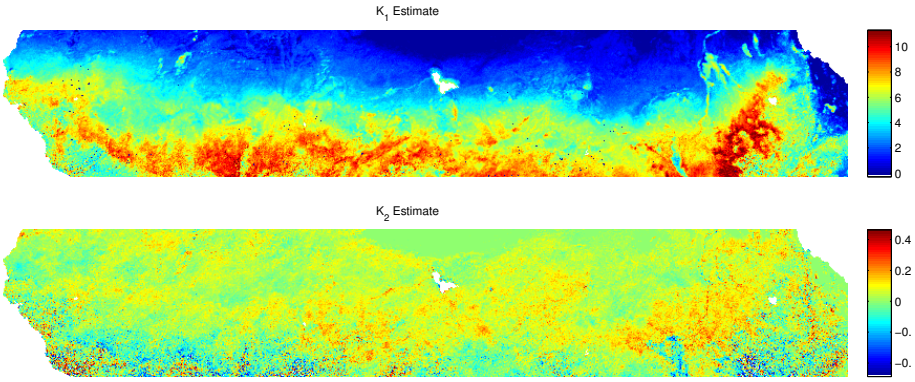

Figure 3: OLS estimates of the intercepts (upper figure) and slopes (lower figure) for the Sahel data. Notice that, especially, the $\mathbf{K}_2$ estimate seems to contain a large amount of noise.

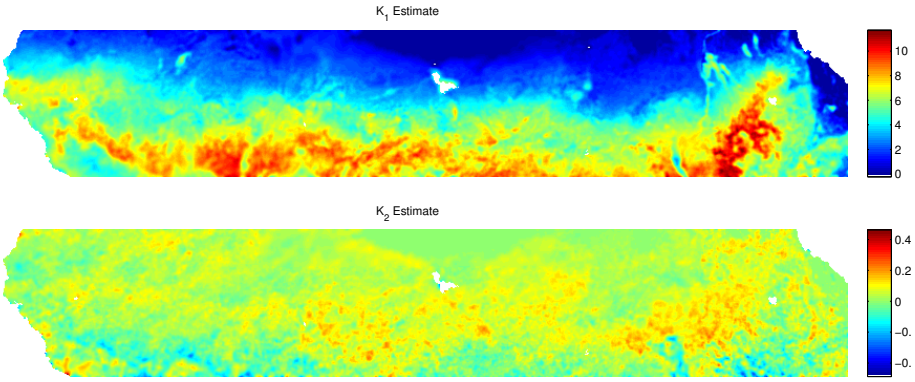

Figure 4: GMRF estimates of the intercepts (upper figure) and slopes (lower figure) for the Sahel data. Notice that both these estimates are much smoother than the corresponding estimates in Figure 3.

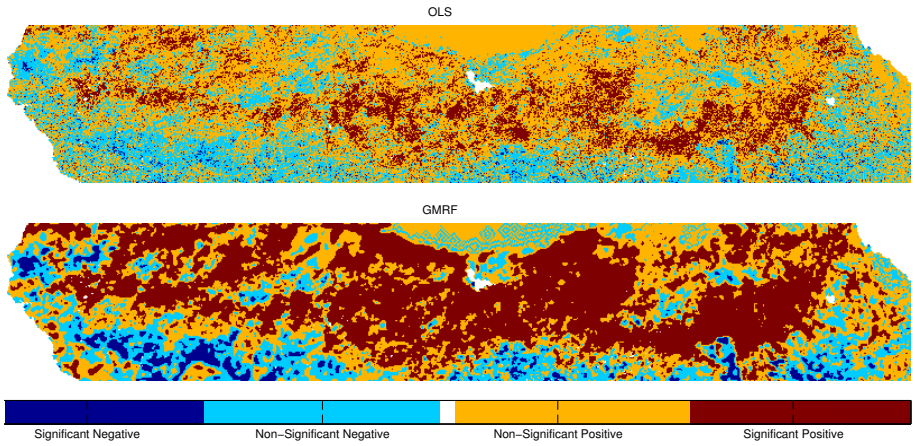

Figure 5: Significance estimates for the slope of the linear trends using the OLS model (upper figure) and the GMRF model (lower figure). Notice the large number of significant positive linear trends in the GMRF estimate.

7 Extensions

In this section, we will discuss three possible extensions of the model, and the increase in complexity and potential difficulties that they would incur. A first extension is to allow for spatial dependencies in the variance of the measurement noise. Since the measurement noise includes phenomena, such as cloud cover and aerosol, the model should, reasonably, allow for varying noise variances. These variances will, however, vary slowly across the image, and the measurement noise estimation could therefore be improved by adding a spatially dependent prior for the noise. There are, in principle, no problems doing this. The only difference in the model will be that explicit updating rules in the EM algorithm are hard to find, most likely necessitating numerical optimisation schemes.

Another interesting extension is to allow for varying strength in the spatial dependencies. That is, instead of using a single precision factor, κ, in the GMRF prior, we let κ vary across the field. This is probably a more accurate model for describing vegetation since the strength of the spatial dependencies most likely varies with the character of the land. Because of the problem structure, the different weights will not separate in the likelihood expression, and the resulting optimisation problem becomes highly non-linear and computationally demanding.

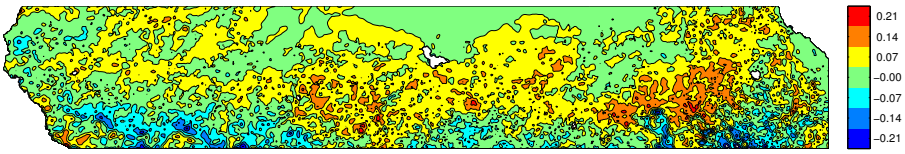

Figure 6: Contour plot of the GMRF estimate of the $\mathbf{K}_2$ field. Here, colours corresponding to positive values correspond to areas with a positive linear increase in vegetation. Notice that a large number of these areas coincides with areas with significant positive trends in Figure 5. Also observe that this figure has a different colour scale than Figure 4.

Finally, introducing non-Gaussian observations to handle heavy tailed residuals, or allowing generalised linear regression models, would also improve the model. However, for non-Gaussian observations, the expectations in (11) will most likely have to be found using simulation techniques, which increases the computational burden of the algorithm.

8 Concluding remarks

There is a need for efficient methods for estimating trends in Earth Observation data. The spatio-temporal regression model constructed in this work shows great promise for utilising the spatial dependencies in satellite-derived NDVI data. Tests on real and simulated data sets indicate that there is a substantial gain in precision, compared with using independent ordinary least squares regressions for the individual pixels. By estimating the model parameters using the EM algorithm, we also obtain a substantial gain in computational cost compared with a full MCMC-based approach.

The GMRF estimates are smoother and exhibit larger contiguous regions with significant trends than a comparative analysis using OLS. The larger contiguous regions and smoother estimates will most likely aid interpretation of the data and make it easier to identify underlying reasons for the detected changes in vegetation.

The model we have used assumes as common precision parameter for the entire field, which is probably not valid for the entire Sahel. Therefore, the results presented in Section 6.2 are obtained by first dividing the area into smaller subareas. This way, κ is assumed to be constant only on the smaller subareas. To

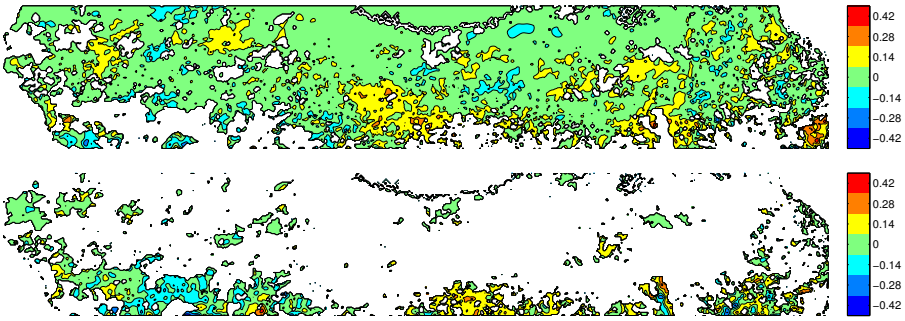

Figure 7: Upper figure: Contour plot of the $\mathbf{K}_3$ field for the pixels where the $\mathbf{K}_2$ field is positive. A positive value in this field means that there was a higher increase in vegetation growth during the second part of the time period, whereas a negative value corresponds to a higher increase in vegetation growth during the first part of the time period. Lower figure: Contour plot of the $\mathbf{K}_3$ field for the pixels where the $\mathbf{K}_2$ field is negative. Here, a positive value means that there was a larger decrease in vegetation during the first part, whereas a negative value corresponds to a larger decrease in vegetation during the second half.

avoid discontinuities in the joints between the areas, extra pixels, overlapping the neighbouring subareas, are added in the computations. Although this procedure will allow for some variation in the strength of the spatial dependencies, it does not solve the problem entirely. Hence, some improvements are anticipated to further improve the results for the Sahel region.

Acknowledgements

D. Bolin and J. Lindström have been partially funded by the Swedish Foundation for Strategic Research (SSF) under grant A3 02:125, Spatial statistics and image analysis for environment and medicine.

Data used by the authors in this study include data produced through funding from the Earth Observing System Pathfinder Program of NASA's Mission to Planet Earth in cooperation with National Oceanic and Atmospheric Administration. The data were provided by the Earth Observing System Data and Information System (EOSDIS), Distributed Active Archive Center at Goddard Space Flight Center which archives, manages, and distributes this data set.

A Proof of Proposition 2.2

This follows from a change of variables in the distribution $\pi(\mathbf{X}|\kappa)$:

$$\pi(\mathbf{X}|\kappa) \propto \exp(-\frac{1}{2}\mathbf{X}^\top \kappa \overline{\mathbf{Q}}\mathbf{X}) = \exp\left(-\frac{\kappa}{2}\mathbf{K}^\top \underbrace{(\mathbf{F}\otimes\mathbf{I})^\top(\mathbf{I}\otimes\mathbf{Q}_X)(\mathbf{F}\otimes\mathbf{I})}_{Q}\mathbf{K}\right).$$

Using common calculation rules (mixed product) for the Kronecker product, the expression for $\mathbf{Q}$ simplifies to

$$\mathbf{Q} = (\mathbf{F}\otimes\mathbf{I})^\top(\mathbf{I}\otimes\mathbf{Q}_X)(\mathbf{F}\otimes\mathbf{I}) = (\mathbf{F}^\top\mathbf{F})\otimes\mathbf{Q}_X. \tag{16}$$

Hence, $\mathbf{K}|\kappa \in \mathsf{N}(0,(\kappa\mathbf{Q})^{-1})$, and the corresponding graph can be determined from the non-zero pattern of $\mathbf{Q}$. To determine the rank of the field, we have to determine the rank of $\mathbf{Q}$. Since $\mathbf{X}_t$ is a second-order IGMRF, $\mathbf{Q}_X$ has rank $n-3$, and $\mathbf{F}^\top\mathbf{F}$ has, by construction, full rank. A basic property of the Kronecker product is that $\text{rank}(\mathbf{A}\otimes\mathbf{B}) = \text{rank}(\mathbf{A})\,\text{rank}(\mathbf{B})$, and hence, $\mathbf{Q}$ has rank $m(n-3)$. □

References

Agbu, P. and James, M. (1994). *The NOAA/NASA Pathfinder AVHRR Land Data Set User's Manual*. Goddard Distributed Active Archive Center, NASA, Goddard Space Flight Center, Greenbelt.

Besag, J., York, J., and Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics (with discussion). *Ann. Inst. of Statist. Math.*, 43:1–59.

Bolin, D. (2007). Estimating vegetation trends in the African Sahel using Gaussian Markov random fields. Master's thesis, Lund University, Lund. LUTFMS–3090–2007.

Chappell, A., Seaquist, J., and Eklundh, L. (2001). Improving the estimation of noise from NOAA AVHRR NDVI for Africa using geostatistics. *Internat. J. Remote Sensing*, 22:1067–1080.

Cressie, N. (1991). *Statistics for Spatial Data*. John Wiley & Sons Ltd.

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

Di Giacinto, V., Dryden, I., Ippoliti, L., and Romagnoli, L. (2005). Linear smoothing of noisy spatial temporal series. *J. Math. and Statist.*, 1(4):300–312.

Eklundh, L. and Olsson, L. (2003). Vegetation index trends for the African Sahel 1982–1999. *J. Geophys. Res.*, 30:1430–1433.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications*. Chapman & Hall.

Gamerman, D., Moreira, A., and Rue, H. (2003). Space-varying regression models: specifications and simulation. *Comput. Statist. and Data Anal.*, 42:513–533.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004). *Bayesian Data Analysis*. Chapman & Hall/CRC, second edition.

Hanan, N., Prince, S., and Holben, B. (1995). Atmospheric correction of AVHRR data for biophysical remote sensing of the Sahel. *Remote Sensing of Environ.*, 51:306–316.

Hickler, T., Eklundh, L., Seaquist, J., Smith, B., Ardö, J., Olsson, L., Sykes, M., and Sjöström, M. (2005). Precipitation controls Sahel greening trend. *Geophys. Res. Lett.*, 32. L21415, doi:10.1029/2005GL024370.

Holben, B. N. (1986). Characteristics of maximum-value composite images from temporal AVHRR data. *Internat. J. Remote Sensing*, 7(11):1417–1434.

Hulme, M. (2001). Climatic perspectives on Sahelian desiccation: 1973–1998. *Global Environ. Change*, 11:19–29.

James, M. and Kalluri, S. (1994). The Pathfinder AVHRR Land data set: An improved coarse resolution data set for terrestrial monitoring. *Internat. J. Remote Sensing*, 15:3347–3363.

Jönsson, P. and Eklundh, L. (2002). Seasonality extraction by function fitting to time-series of satellite sensor data. *IEEE T. Geosci. Remote Sensing*, 40(8):1824–1832.

Jönsson, P. and Eklundh, L. (2004). TIMESAT - a program for analyzing time-series of satellite sensor data. *Comput. and Geosci.*, 30:833–845.

Justice, C. and Hiernaux, P. (1986). Monitoring the grasslands of the Sahel using NOAA AVHRR data: Niger 1983. *Internat. J. Remote Sensing*, 7:1475–1497.

Kidwell, K. B., editor (1998). *NOAA Polar Orbiter Data User's Guide (TIROS–N, NOAA–6, NOAA–7, NOAA–8, NOAA–9, NOAA–10, NOAA–11, NOAA–12, NOAA–13 and NOAA–14) November 1998 Revision*. U.S. Department of Commerce, National Oceanic and Atmospheric Administration.

Lagarias, J. C., Reeds, J. A., Wright, M. H., and Wright, P. E. (1998). Convergence properties of the Nelder-Mead simplex method in low dimensions. *SIAM J. Optim.*, 9:112–147.

Lamb, P. (1982). Persistence of Subsaharan drought. *Nature*, 299:46–47.

Lindström, J., Eklundh, L., Holst, J., and Holst, U. (2006). Influence of solar zenith angles on observed trends in the noaa/nasa 8 km pathfinder normalized difference vegetation index over the african sahel. *Internat. J. Remote Sensing*, 27(10):1973–1991.

Myneni, R., Keeling, C., Tucker, C., Asrar, G., and Nemani, R. (1997). Increased plant growth in the northern high latitudes from 1981 to 1991. *Nature*, 386:698–702.

Nicholson, S. (2000). Land surface process and Sahel climate. *Rev. of Geophys.*, 38:117–140.

Olsson, L. (1993). On the causes of famine – drought, desertification and market failure in the Sudan. *Ambio*, 22:395–403.

Olsson, L., Eklundh, L., and Ardö, J. (2005). A recent greening of the Sahel–trends, patterns and potential causes. *J. Arid Environ.*, 63:556–566.

Press, W., Teukolsky, W., Vetterling, W., and Flannery, B. (1992). *Numerical recipes in FORTRAN: The art of scientific computing*. Cambridge Univ. Press.

Prince, S. (1991). Satellite remote sensing of primary production: comparison of results for Sahelian grasslands 1981–1988. *Internat. J. Remote Sensing*, 12:1301–1312.

Prince, S. and Goward, S. (1996). Evaluation of the NOAA/NASA pathfinder AVHRR land data set for global primary production modelling. *Internat. J. Remote Sensing*, 17:217–221.

Rouse, J., Haas, R., Schell, J., and Deering, D. (1973). Monitoring vegetation systems in the great plains with erts. *Third ERTS Symposium, NASA, SP-351, NASA, Washington, DC*, 1:309–317.

Rue, H. (2007). GMRFLib: Fast and exact simulation of Gaussian Markov random fields on graphs (ver. 3.0-0).

Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields; Theory and Applications*, volume 104 of *Monographs on Statistics and Applied Probability*. Chapman & Hall/CRC.

Rue, H. and Martino, S. (2007). Approximate Bayesian inference for hierarchical Gaussian Markov random field models. *J. Statist. Plann. and Inference*, 137(10):3177–3192.

Savitzky, A. and Golay, M. J. E. (1964). Smoothing and differentiation of data by simplified least squares procedures. *Anal. Chem.*, 36:1627–1639.

Seaquist, J., Olsson, L., and Ardö, J. (2003). A remote sensing-based primary production model for grassland biomes. *Ecological Model.*, 169:131–155.

Sellers, P. J., Dickinson, R. E., Randall, D. A., Betts, A. K., Hall, F. G., Berry, J. A., Collatz, G. J., Denning, A. S., Mooney, H. A., Nobre, C. A., Sato, N., Field, C. B., and Henderson-Sellers, A. (1997). Modeling the exchange of energy, water, and carbon between continents and atmosphere. *Science*, 275:602–609.

Stowe, L., McClain, E., Carey, R., Pellegrino, P., Gutman, G., Davis, P., Long, C., and Hart, S. (1991). Global disribution of cloud cover derived from NOAA/AVHRR operational satellite data. *Adv. in Space Res.*, 3:51–54.

Tanré, D., Holben, B., and Kaufman, Y. (1992). Atmospheric correction algorithm for NOAA AVHRR products: Theory and application. *IEEE T. Geosci. Remote Sensing*, 30:231–248.

Tucker, C., Vanpraet, C., Sharman, M., and Van Ittersum, G. (1985). Satellite remote sensing of total herbaceous biomass production in the Sengalese Sahel: 1980–1984. *Remote Sensing of Environ.*, 17:233–249.

Wikle, C. K., Berliner, L. M., and Cressie, N. (1998). Hierarchical Bayesian space-time models. *Environ. and Ecological Statist.*, 5(2):117–154.

B

How do Markov approximations compare with other methods for large spatial data sets?

DAVID BOLIN[1] AND FINN LINDGREN[2]

[1] *Centre for Mathematical Sciences, Lund University, Lund, Sweden*
[2] *Department of Mathematical Sciences, Norwegian University of Science and Technology, Trondheim, Norway*

Abstract: The Matérn covariance function is a popular choice for modeling dependence in spatial environmental data. Standard Matérn covariance models are, however, often computationally infeasible for large data sets. In this work, recent results for Markov approximations of Gaussian Matérn fields based on Hilbert space approximations are extended using wavelet basis functions. These Markov approximations are compared with two of the most popular methods for efficient covariance approximations; covariance tapering and the process convolution method. The results show that, for a given computational cost, the Markov methods have a substantial gain in accuracy compared with the other methods.

Key words: Matérn covariances; kriging; wavelets; Markov random fields; covariance tapering; process convolutions; Computational efficiency

1 Introduction

The traditional methods in spatial statistics were typically developed without any considerations of computational efficiency. In many of the classical applications of spatial statistics in environmental sciences, the cost for obtaining measurements limited the size of the data sets to ranges where computational cost was not an issue. Today, however, with the increasing use of remote sensing satellites, producing many large climate data sets, computational efficiency is often a crucial property.

In recent decades, several techniques for building computationally efficient models have been suggested. In many of these techniques, the main assumption is that a latent, zero-mean Gaussian process $X(\mathbf{s})$ can be expressed, or at least

approximated, through some finite basis expansion

$$X(\mathbf{s}) = \sum_{j=1}^{n} w_j \xi_j(\mathbf{s}), \tag{1}$$

where w_j are Gaussian random variables and $\{\xi_j\}_{j=1}^{n}$ are some pre-defined basis functions. The justification for using these basis expansions is usually that they converge to the true spatial model as n tends to infinity. However, for a finite n, the choice of the weights and basis functions will greatly affect the approximation error and the computational efficiency of the model. Hence, if one wants an accurate model for a given computational cost, asymptotic arguments are insufficient.

If the process $X(\mathbf{s})$ has a discrete spectral density, one can obtain an approximation of the form (1) by truncating the spectral expansion of the process. Another way to obtain an, in some sense optimal, expansion of the form (1) is to use the eigenfunctions of the covariance function for the latent field $X(\mathbf{s})$ as a basis, which is usually called the Karhunen-Loève (KL) transform. The problem with the KL transform is that analytic expressions for the eigenfunctions are only known in a few simple cases, which are often insufficient to represent the covariance structure in real data sets. Numerical approximations of the eigenfunctions can be obtained for a given covariance function; however, the covariance function is in most cases not known, but has to be estimated from data. In these cases, it is infeasible to use the KL expansion in the parameter estimation, which is often the most computationally demanding part of the analysis. The spectral representation has a similar problem since the computationally efficient methods are usually restricted to stationary models with gridded data, and are not applicable in more general situations. Thus, to be useful for a broad range of practical applications, the methods should be applicable to a wide family of stationary covariance functions, and be extendable to nonstationary covariance structures.

One method that fulfills these requirements is the process convolution approach (Barry and Ver Hoef, 1996, Higdon, 2001, Cressie and Ravlicová, 2002, Rodrigues and Diggle, 2010). In this method, the stochastic field, $X(\mathbf{s})$, is defined as the convolution of a Gaussian white noise process with some convolution kernel $K(\mathbf{s})$. This convolution is then approximated with a sum of the form (1) to get a discrete model representation. Process convolution approximations are computationally efficient if a small number of basis functions can be used, but in practice, this will often give a poor approximation of the continuous convolution model.

A popular method for creating computationally efficient approximations is covariance tapering (Furrer et al., 2006). This method can not be written as an approximation of the form (1), but the idea is instead to taper the true covariance to zero beyond a certain range by multiplying the covariance function with some compactly supported taper function (Gneiting, 2002). This facilitates the use of sparse matrix techniques that increases the computational efficiency, at the cost of replacing the original model with a different model, which can lead to problems depending on the spatial structure of the data locations. However, the method is applicable to both stationary and nonstationary covariance models, and instead of choosing the set of basis functions in (1), the taper range and the taper function have to be chosen.

Nychka et al. (2002) used a wavelet basis in the expansion (1), and showed that by allowing for some correlation among the random variables w_j, one gets a flexible model that can be used for estimating nonstationary covariance structures. As a motivating example, they showed that using a wavelet basis, computationally efficient approximations to the popular Matérn covariance functions can be obtained using only a few nonzero correlations for the weights w_j. The approximations were, however, obtained numerically, and no explicit representations were derived.

Rue and Tjelmeland (2002) showed that general stationary covariance models can be closely approximated by Markov random fields, by numerically minimizing the error in the resulting covariances. Song et al. (2008) extended the method by applying different loss criteria, such as minimizing the spectral error or the Kullback-Leibler divergence. A drawback of the methods is that, just as for the KL and wavelet approaches, the numerical optimisation must in general be performed for each distinct parameter configuration.

Recently, Lindgren and Rue (2007) derived an explicit method for producing computationally efficient approximations to the Matérn covariance family. The method uses the fact that a random process on $\mathbb{R}^d$ with a Matérn covariance function is a solution to a certain stochastic partial differential equation (SPDE). By considering weak solutions to this SPDE with respect to some set of local basis functions $\{\xi_j\}_{j=1}^n$, an approximation of the form (1) is obtained, where the stochastic weights have a sparse precision matrix (inverse covariance matrix), that can be written directly as a function of the parameters, without any need for costly numerical calculations. The method is also extendable to more general stationary and nonstationary models by extending the generating SPDE (Lindgren et al.,

2011, Bolin and Lindgren, 2011).

In this paper, we use methods from Lindgren and Rue (2007) and Lindgren et al. (2011) to algebraically compute the weights w_j for wavelet-based approximations to Gaussian Matérn fields (Section 3). For certain wavelet bases, the weights form a Gaussian Markov Random Field (GMRF), which greatly increases the computational efficiency of the approximation. For other wavelet bases, such as the one used in Nychka et al. (2002), the weights can be well approximated with a GMRF.

In order to evaluate the practical usefulness of the different approaches, a detailed analysis of the computational aspects of the spatial prediction problem is performed (Section 2 and Section 4). The results show that the GMRF methods are more efficient and accurate than both the process convolution approach and the covariance tapering method.

2 Spatial prediction and computational cost

As a motivating example why computational efficiency is important, let us consider spatial prediction. The most widely used method for spatial prediction is commonly known as linear kriging in geostatistics. Let $Y(\mathbf{s})$ be an observation of a latent Gaussian field, $X(\mathbf{s})$, under mean zero Gaussian measurement noise, $\mathcal{E}(\mathbf{s})$, uncorrelated with X and with some covariance function $r_{\mathcal{E}}(\mathbf{s}, \mathbf{t})$,

$$Y(\mathbf{s}) = X(\mathbf{s}) + \mathcal{E}(\mathbf{s}), \tag{2}$$

and let $\mu(\mathbf{s})$ and $r(\mathbf{s}, \mathbf{t})$ be the mean value function and covariance function for $X(\mathbf{s})$ respectively. Depending on the assumptions on $\mu(\mathbf{s})$, linear kriging is usually divided into simple kriging (if μ is known), ordinary kriging (if μ is unknown but independent of $\mathbf{s}$), and universal kriging (if μ is unknown and can be expressed as a linear combination of some deterministic basis functions). To limit the scope of this article, parameter estimation will not be considered, and to simplify the notations, we let $\mu(\mathbf{s}) \equiv 0$. It should, however, be noted that all results in later sections regarding computational efficiency also hold in the cases of ordinary kriging and universal kriging. For more details on kriging, see e.g. Stein (1999) or Schabenberger and Gotway (2005).

Let $r(\mathbf{s}, \mathbf{t})$ have some parametric structure, and let the vector $\boldsymbol{\gamma}$ contain all covariance parameters. Let $\mathbf{Y}$ be a vector containing the observations, $\mathbf{X}_1$ be a vector containing $X(\mathbf{s})$ evaluated at the measurement locations, $\mathbf{s}_1, \ldots, \mathbf{s}_m$, and let

$\mathbf{X}_2$ be a vector containing $X(\mathbf{s})$ at the locations, $\hat{\mathbf{s}}_1, \ldots, \hat{\mathbf{s}}_{\hat{m}}$, for which the kriging predictor should be calculated. With $\mathbf{X} = (\mathbf{X}_1^\top, \mathbf{X}_2^\top)^\top$, one has $\mathbf{X}_1 = \mathbf{A}_1\mathbf{X}$, and $\mathbf{X}_2 = \mathbf{A}_2\mathbf{X}$ for two diagonal matrices $\mathbf{A}_1$ and $\mathbf{A}_2$, and the model can now be written as

$$\begin{aligned}\mathbf{X}|\boldsymbol{\gamma} &\sim \mathsf{N}(\mathbf{0}, \boldsymbol{\Sigma}_X),\\ \mathbf{Y}|\mathbf{X}, \boldsymbol{\gamma} &\sim \mathsf{N}(\mathbf{A}_1\mathbf{X}, \boldsymbol{\Sigma}_{\mathcal{E}}),\end{aligned}$$

where $\boldsymbol{\Sigma}_X$ is the covariance matrix for $\mathbf{X}$ and $\boldsymbol{\Sigma}_{\mathcal{E}}$ contains the covariances $r_{\mathcal{E}}(\mathbf{s}_i, \mathbf{s}_j)$. It is straightforward to show that the posterior is $\mathbf{X}|\mathbf{Y}, \boldsymbol{\gamma} \sim \mathsf{N}(\hat{\boldsymbol{\Sigma}}\mathbf{A}_1\boldsymbol{\Sigma}_{\mathcal{E}}^{-1}\mathbf{Y}, \hat{\boldsymbol{\Sigma}})$, where $\hat{\boldsymbol{\Sigma}} = (\boldsymbol{\Sigma}_X^{-1} + \mathbf{A}_1^\top\boldsymbol{\Sigma}_{\mathcal{E}}^{-1}\mathbf{A}_1)^{-1}$, and the well known expression for the kriging predictor is now given by the conditional mean

$$\begin{aligned}\mathsf{E}(\mathbf{X}_2|\mathbf{Y}, \boldsymbol{\gamma}) &= \mathbf{A}_2\hat{\boldsymbol{\Sigma}}\mathbf{A}_1\boldsymbol{\Sigma}_{\mathcal{E}}^{-1}\mathbf{Y} = \mathbf{A}_2\boldsymbol{\Sigma}_X\mathbf{A}_1^\top(\mathbf{A}_1\boldsymbol{\Sigma}_X\mathbf{A}_1^\top + \boldsymbol{\Sigma}_{\mathcal{E}})^{-1}\mathbf{Y}\\ &= \boldsymbol{\Sigma}_{X_2X_1}(\boldsymbol{\Sigma}_{X_1} + \boldsymbol{\Sigma}_{\mathcal{E}})^{-1}\mathbf{Y} = \boldsymbol{\Sigma}_{X_2X_1}\boldsymbol{\Sigma}_Y^{-1}\mathbf{Y},\end{aligned} \tag{3}$$

where the elements on row i and column j in $\boldsymbol{\Sigma}_{X_2X_1}$ and $\boldsymbol{\Sigma}_Y$ are given by the covariances $r(\hat{\mathbf{s}}_i, \mathbf{s}_j)$ and $r(\mathbf{s}_i, \mathbf{s}_j) + r_{\mathcal{E}}(\mathbf{s}_i, \mathbf{s}_j)$ respectively. To get the standard expression for the variance of the kriging predictor, the Woodbury identity is used on $\hat{\boldsymbol{\Sigma}}$:

$$\begin{aligned}\mathsf{V}(\mathbf{X}_2|\mathbf{Y}, \boldsymbol{\gamma}) &= \mathbf{A}_2(\boldsymbol{\Sigma}_X^{-1} + \mathbf{A}_1^\top\boldsymbol{\Sigma}_{\mathcal{E}}^{-1}\mathbf{A}_1)^{-1}\mathbf{A}_2^\top\\ &= \mathbf{A}_2\boldsymbol{\Sigma}_X\mathbf{A}_2 - \mathbf{A}_2\boldsymbol{\Sigma}_X\mathbf{A}_1^\top(\mathbf{A}_1\boldsymbol{\Sigma}_X\mathbf{A}_1^\top + \boldsymbol{\Sigma}_{\mathcal{E}})\mathbf{A}_1\boldsymbol{\Sigma}_X\mathbf{A}_2^\top\\ &= \boldsymbol{\Sigma}_{X_2} - \boldsymbol{\Sigma}_{X_2X_1}\boldsymbol{\Sigma}_Y^{-1}\boldsymbol{\Sigma}_{X_2X_1}^\top.\end{aligned}$$

If there are no simplifying assumptions on $\boldsymbol{\Sigma}_X$, the computational cost for calculating the kriging predictor is $\mathcal{O}(\hat{m}m + m^3)$, and the cost for calculating the variance is even higher. This means that with 1000 measurements, the number of operations needed for the kriging prediction for a single location is on the order of 10^9. These computations are thus not feasible for a large data set where one might have more than 10^6 measurements.

The methods described in Section 1 all make different approximations in order to reduce the computational cost for calculating the kriging predictor and its variance. These different approximations, and their impact on the computational cost, are described in more detail in Section 4; however, to get a general idea of how the computational efficiency can be increased, consider the kriging predictor for a model of the form (1). The field $\mathbf{X}$ can then be written as

$\mathbf{X} = \mathbf{B}\mathbf{w} \sim \mathsf{N}(\mathbf{0}, \mathbf{B}\boldsymbol{\Sigma}_w\mathbf{B}^\top)$, where column i in the matrix $\mathbf{B}$ contains the basis function $\xi_i(\mathbf{s})$ evaluated at all measurement locations and all locations where the kriging prediction is to be calculated and $\mathbf{w} = (w_1, \ldots, w_n)^\top$. Let $\mathbf{B}_1 = \mathbf{A}_1\mathbf{B}$ and $\mathbf{B}_2 = \mathbf{A}_2\mathbf{B}$ be the matrices containing the basis functions evaluated at the measurement locations and the kriging locations respectively. The kriging predictor is then

$$\mathsf{E}(\mathbf{X}_2|\mathbf{Y}, \boldsymbol{\gamma}) = \mathbf{B}_2(\boldsymbol{\Sigma}_w^{-1} + \mathbf{B}_1^\top\boldsymbol{\Sigma}_{\mathcal{E}}^{-1}\mathbf{B}_1)^{-1}\mathbf{B}_1\boldsymbol{\Sigma}_{\mathcal{E}}^{-1}\mathbf{Y}. \tag{4}$$

If the measurement noise is Gaussian white noise, $\boldsymbol{\Sigma}_{\mathcal{E}}$ is diagonal and easy to invert. If $\boldsymbol{\Sigma}_w^{-1}$ is either known, or easy to calculate, the most expensive calculation in (4) is to solve $\mathbf{u} = (\boldsymbol{\Sigma}_w^{-1} + \mathbf{B}_1^\top\boldsymbol{\Sigma}_{\mathcal{E}}^{-1}\mathbf{B}_1)^{-1}\mathbf{B}_1\boldsymbol{\Sigma}_{\mathcal{E}}^{-1}\mathbf{Y}$. This is a linear system of n equations, where n is the number of basis functions used in the approximation. Thus, the easiest way of reducing the computational cost is to choose $n \ll m$, which is what is done in the convolution approach. Another approach is to ensure that $(\boldsymbol{\Sigma}_w^{-1} + \mathbf{B}_1^\top\boldsymbol{\Sigma}_{\mathcal{E}}^{-1}\mathbf{B}_1)$ is a sparse matrix. Sparse matrix techniques can then be used to calculate the kriging predictor, and the computational cost can be reduced without reducing the number of basis functions in the approximation. If a wavelet basis is used, $\mathbf{B}_1^\top\boldsymbol{\Sigma}_{\mathcal{E}}^{-1}\mathbf{B}_1$ will be sparse, and in Section 3, it is shown that the precision matrix $\mathbf{Q}_w = \boldsymbol{\Sigma}_w^{-1}$ can also be chosen as a sparse matrix by using the Hilbert space approximation technique by Lindgren et al. (2011).

3 Wavelet approximations

In the remainder of this paper, the focus is on the family of Matérn covariance functions (Matérn, 1960) and the computational efficiency of some different techniques for approximating Gaussian Matérn fields. This section shows how wavelet bases can be used in the Hilbert space approximation technique by Lindgren et al. (2011) to obtain computationally efficient Matérn approximations.

3.1 The Matérn covariance family

Because of its versatility, the Matérn covariance family is one of the most popular choices for modeling spatial data. There are a few different parameterizations of the Matérn covariance function in the literature, and the one most suitable in our context is

$$r(\boldsymbol{\tau}) = \frac{2^{1-\nu}\phi^2}{(4\pi)^{\frac{d}{2}}\Gamma(\nu + \frac{d}{2})\kappa^{2\nu}}(\kappa\|\boldsymbol{\tau}\|)^\nu K_\nu(\kappa\|\boldsymbol{\tau}\|), \tag{5}$$

where ν is a shape parameter, κ^2 a scale parameter, ϕ^2 a variance parameter, and K_ν is a modified Bessel function of the second kind of order $\nu > 0$. With this parametrization, the variance is $r(\mathbf{0}) = \phi^2 \Gamma(\nu)(4\pi)^{-\frac{d}{2}} \Gamma(\nu + \frac{d}{2})^{-1} \kappa^{-2\nu}$, and the associated spectral density is

$$S(\boldsymbol{\omega}) = \frac{\phi^2}{(2\pi)^d} \frac{1}{(\kappa^2 + \|\boldsymbol{\omega}\|^2)^{\nu + \frac{d}{2}}}. \tag{6}$$

For the special case $\nu = 0.5$, the Matérn covariance function is the exponential covariance function. The smoothness of the field increases with ν, and in the limit as $\nu \to \infty$, the covariance function is a Gaussian covariance function if κ is also scaled accordingly, which gives an infinitely differentiable field.

3.2 Hilbert space approximations

As noted by Whittle (1963), a random process with the covariance (5) is a solution to the SPDE

$$(\kappa^2 - \Delta)^{\frac{\alpha}{2}} X(\mathbf{s}) = \phi \mathcal{W}(\mathbf{s}), \tag{7}$$

where $\mathcal{W}(\mathbf{s})$ is Gaussian white noise, Δ is the Laplacian, and $\alpha = \nu + d/2$. The key idea in Lindgren et al. (2011) is to approximate the solution to the SPDE using a basis expansion of the form (1). The starting point of the approximation is to consider the stochastic weak formulation of the SPDE

$$\left\{ \left\langle b_i,\, (\kappa^2 - \Delta)^{\frac{\alpha}{2}} X \right\rangle, i = 1, \ldots, n_b \right\} \stackrel{d}{=} \left\{ \langle b_i,\, \phi \mathcal{W} \rangle, i = 1, \ldots, n_b \right\}. \tag{8}$$

Here $\stackrel{d}{=}$ denotes equality in distribution, $\langle f, g \rangle = \int f(\mathbf{s}) g(\mathbf{s})\, \mathrm{d}\mathbf{s}$, and equality should hold for every finite set of test functions $\{b_i, i = 1, \ldots, n_b\}$ from some appropriate space. A finite element approximation of the solution X is then obtained by representing it as a finite basis expansion of the form (1), where the stochastic weights are calculated by requiring (8) to hold for only a specific set of test functions $\{b_i, i = 1, \ldots, n\}$ and $\{\xi_i\}$ is a set of predetermined basis functions. We illustrate the more general results from Lindgren et al. (2011) with the special case $\alpha = 2$, where one uses $b_i = \xi_i$ and then has

$$\left\langle \xi_i,\, (\kappa^2 - \Delta) X \right\rangle = \sum_{j=1}^{n} w_j \left\langle \xi_i,\, (\kappa^2 - \Delta) \xi_j \right\rangle. \tag{9}$$

By introducing the matrix $\mathbf{K}$ with elements $\mathbf{K}_{i,j} = \left\langle \xi_i, (\kappa^2 - \Delta)\xi_j \right\rangle$ and the vector $\mathbf{w} = (w_1, \ldots, w_n)^\top$, the left hand side of (8) can be written as $\mathbf{Kw}$. Since, by Lemma 1 in Lindgren et al. (2011)

$$\left\langle \xi_i, (\kappa^2 - \Delta)\xi_j \right\rangle = \kappa^2 \left\langle \xi_i, \xi_j \right\rangle - \left\langle \xi_i, \Delta\xi_j \right\rangle = \kappa^2 \left\langle \xi_i, \xi_j \right\rangle + \left\langle \nabla\xi_i, \nabla\xi_j \right\rangle,$$

the matrix $\mathbf{K}$ can be written as the sum $\mathbf{K} = \kappa^2\mathbf{C} + \mathbf{G}$ where $\mathbf{C}_{i,j} = \left\langle \xi_i, \xi_j \right\rangle$ and $\mathbf{G}_{i,j} = \left\langle \nabla\xi_i, \nabla\xi_j \right\rangle$. The right hand side of (8) can be shown to be Gaussian with mean zero and covariance $\phi^2\mathbf{C}$ and one thus have that $\mathbf{w} \sim \mathsf{N}(0, \phi^2\mathbf{K}^{-1}\mathbf{C}\mathbf{K}^{-1})$.

For the second fundamental case, $\alpha = 1$, Lindgren et al. (2011) show that $\mathbf{w} \sim \mathsf{N}(\mathbf{0}, \phi^2\mathbf{K}^{-1})$ and for higher order $\alpha \in \mathbb{N}$, the weak solution is obtained recursively using these two fundamental cases. For example, if $\alpha = 4$ the solution to $(\kappa^2 - \Delta)^2 X_0(\mathbf{s}) = \phi\,\mathcal{W}(\mathbf{s})$ is obtained by solving $(\kappa^2 - \Delta)X_0(\mathbf{s}) = \tilde{X}(\mathbf{s})$, where $\tilde{X}$ is the solution for the case $\alpha = 2$. This results in a precision matrix for the weights $\mathbf{Q}_\alpha$ defined recursively as

$$\mathbf{Q}_\alpha = \mathbf{K}\mathbf{C}^{-1}\mathbf{Q}_{\alpha-2}\mathbf{C}^{-1}\mathbf{K}, \quad \alpha = 3, 4, \ldots \tag{10}$$

where $\mathbf{Q}_1 = \phi^{-2}\mathbf{K}$ and $\mathbf{Q}_2 = \phi^{-2}\mathbf{K}^\top\mathbf{C}^{-1}\mathbf{K}$. Thus, all Matérn fields with $\nu + d/2 \in \mathbb{N}$ can be approximated through this procedure. For more details, see Lindgren and Rue (2007) and Lindgren et al. (2011). The results from Rue and Tjelmeland (2002) show that accurate Markov approximations exist also for other ν-values, and one approximate approach to finding explicit expressions for such models was given in the authors' response in Lindgren et al. (2011). However, in many practical applications ν cannot be estimated reliably (Zhang, 2004), and using only a discrete set of ν-values is not necessarily a significant restriction.

3.3 Wavelet basis functions

In the previous section, nothing was said about how the basis functions $\{\xi_i\}$ should be chosen. The following sections, however, shows that wavelet bases have many desirable properties which makes them suitable to use in the Hilbert space approximations on $\mathbb{R}^d$. In this section, a brief introduction to multiresolution analysis and wavelets is given.

A multiresolution analysis on $\mathbb{R}$ is a sequence of closed approximation subspaces $\{V_j\}_{j\in\mathbb{Z}}$ of functions in $L^2(\mathbb{R})$ such that $V_j \subset V_{j+1}$, $\mathrm{cl}\bigcup_{j\in\mathbb{Z}} V_j = L^2(\mathbb{R})$, and $\bigcap_{j\in\mathbb{Z}} V_j = \{0\}$, where cl is the closure, and $f(s) \in V_j$ if and only if $f(2^{-j}s) \in V_0$. This last requirement is the multiresolution requirement because

this implies that all the approximation spaces V_j are scaled versions of the space V_0. A multiresolution analysis is generated starting with a function usually called a father function or a scaling function. The function $\varphi \in L^2(\mathbb{R})$ is called a scaling function for $\{V_j\}_{j\in\mathbb{Z}}$ if it satisfies the two-scale relation

$$\varphi(s) = \sum_{k\in\mathbb{Z}} p_k \varphi(2s - k), \tag{11}$$

for some square-summable sequence $\{p_k\}_{k\in\mathbb{Z}}$ and the translates $\{\varphi(s-k)\}_{k\in\mathbb{Z}}$ form an orthonormal basis for V_0. Given the multiresolution analysis $\{V_j\}_{j\in\mathbb{Z}}$, the wavelet spaces $\{W_j\}_{j\in\mathbb{Z}}$ are then defined as the orthogonal complements of V_j in V_{j+1} for each j, and one can show that W_j is the span of $\{\psi(2^j s - k)\}_{k\in\mathbb{Z}}$, where the wavelet ψ is defined as $\psi(s) = \sum_{k\in\mathbb{Z}}(-1)^k \overline{p_{1-k}}\varphi(2s-k)$.

Given the spaces W_j, V_j can be decomposed as the direct sum

$$V_j = V_0 \oplus W_0 \oplus W_1 \oplus \ldots \oplus W_{j-1}. \tag{12}$$

Several choices of scaling functions have been presented in the literature. Among the most widely used constructions are the B-spline wavelets (Chui and Wang, 1992) and the Daubechies wavelets (Daubechies, 1992) that both have several desirable properties for our purposes.

The scaling function of B-spline wavelets are mth order B-splines with knots at the integers. Because of this, there exists closed form expressions for the corresponding wavelets, and the wavelets have compact support since the mth order scaling function has support on $(0, m+1)$. The wavelets are orthogonal at different scales, but translates at the same scale are not orthogonal. This property is usually referred to as semi-orthogonality.

The Daubechies wavelets form a hierarchy of compactly supported orthogonal wavelets that are constructed to have the highest number of vanishing moments for a given support width. This generates a family of wavelets with an increasing degree of smoothness. Except for the first Daubechies wavelet, there are no closed form expressions for these wavelets; however, for practical purposes, this is not a problem because the exact values for the wavelets at dyadic points can be obtained very fast using the Cascade algorithm (Burrus et al., 1988). In this work, the DB3 wavelet is used because it is the first wavelet in the family that has one continuous derivative. The DB3 wavelet and its scaling function are shown in Figure 1.

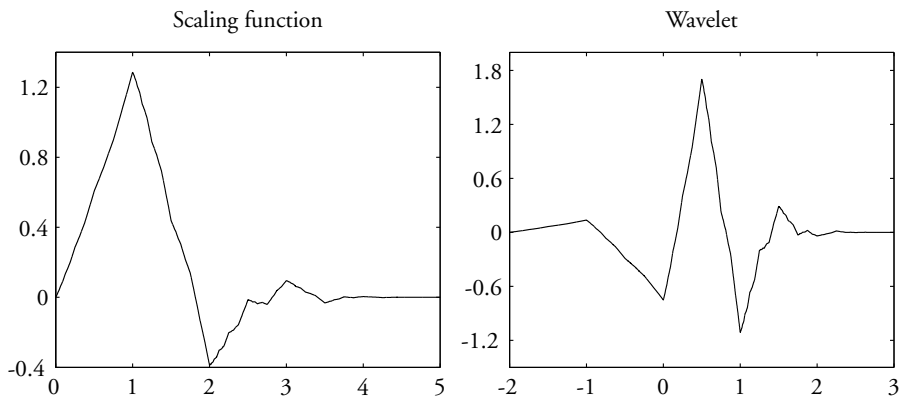

Figure 1: The DB3 scaling function and wavelet.

3.4 Explicit wavelet Hilbert space approximations

To use the Hilbert space approximation for a given basis, the precision matrix, $\mathbf{Q}_\alpha$, for the weights has to be calculated. By (10), we only have to be able to calculate the matrices $\mathbf{C}$ and $\mathbf{G}$ to build the precision matrix for any $\alpha \in \mathbb{N}$. The elements in these matrices are inner products between the basis functions:

$$\mathbf{C}_{i,j} = \int \xi_i(\mathbf{s})\xi_j(\mathbf{s})\,\mathrm{d}\mathbf{s}, \qquad \mathbf{G}_{i,j} = \int (\nabla \xi_i(\mathbf{s}))^\top \nabla \xi_j(\mathbf{s})\,\mathrm{d}\mathbf{s}. \tag{13}$$

This section shows how these elements can be calculated for the DB3 wavelets and the B-spline wavelets. When using a wavelet basis in practice, one always have to choose a finest scale, J, to work with. Given that the subspace V_J is used as an approximation of $L^2(\mathbb{R})$, one can use two different bases. Either one works with the direct basis for V_J, that consists of scaled and translated versions of the father function $\varphi(s)$, or one can use the multiresolution decomposition (12). In what follows, both cases are considered.

3.4.1 Daubechies wavelets on $\mathbb{R}$

For the Daubechies wavelets, the matrix $\mathbf{C}$ is the identity matrix since these wavelets form an orthonormal basis for $L^2(\mathbb{R})$. Thus, only the matrix $\mathbf{G}$ has to be calculated. If the direct basis for V_J is used, the matrix $\mathbf{G}$ contains inner products

of the form

$$\left\langle \nabla\varphi(2^J s - k),\, \nabla\varphi(2^J s - l) \right\rangle = 2^J \left\langle \nabla\varphi(s),\, \nabla\varphi(s - l + k) \right\rangle \equiv 2^J \Lambda(k-l). \quad (14)$$

Because the scaling function has compact support on $[0, 2N-1]$, these inner products are only non-zero if $k - l \in [-(2N-2), 2N-2]$. Thus, the matrix $\mathbf{G}$ is sparse, which implies that the weights $\mathbf{w}$ in (1) form a GMRF. Since there are no closed form expressions for the Daubechies wavelets, there is no hope in finding a closed form expression for the non-zero inner products (14). Furthermore, standard numerical quadrature for calculating the inner products is too inaccurate due to the highly oscillating nature of the gradients. However, utilizing properties of the wavelets, one can calculate an approximation of the inner product of arbitrary precision by solving a system of linear equations. It is outside the scope of this paper to present the full method, but the basic principle is to construct a system of linear equations by using the scaling- and moment equations for the wavelets. This system is then solved using, for example, LU factorization. For details, see Latto et al. (1991).

Using this technique for the DB3 wavelets, the following nonzero values for $\Lambda(\eta)$ are obtained

$$\begin{aligned} \Lambda(0) &= 5.267, & \Lambda(\pm 1) &= -3.390, & \Lambda(\pm 2) &= 0.876, \\ \Lambda(\pm 3) &= -0.114, & \Lambda(\pm 4) &= -0.00535. \end{aligned}$$

These values are calculated once and tabulated for constructing the $\mathbf{G}$ matrix, which is a band matrix with $2^J\Lambda(0)$ on the main diagonal, $2^J\Lambda(1)$ on the first off diagonals, et cetera.

If the multiresolution decomposition (12) is used as a basis for V_J, one also needs the inner products $\left\langle \nabla\psi(2^i s - k),\, \nabla\psi(2^i s - l) \right\rangle$, $i, j \in \mathbb{Z}$. Because of the two-scale relation (11), every wavelet $\psi(2^j s - k)$ can be written as a finite sum of translated scaling functions at scale J. Using this property, the $\mathbf{G}$ matrix can be constructed efficiently using only the already computed Λ values. Figure 2 shows the structure of the $\mathbf{G}$ matrices for a multiresolution DB3 basis with five layers of wavelets and the corresponding direct basis. Note that there are fewer non-zero elements in the precision matrix for the direct basis. Hence, it is more computationally efficient to use the direct basis instead of the multiresolution basis.

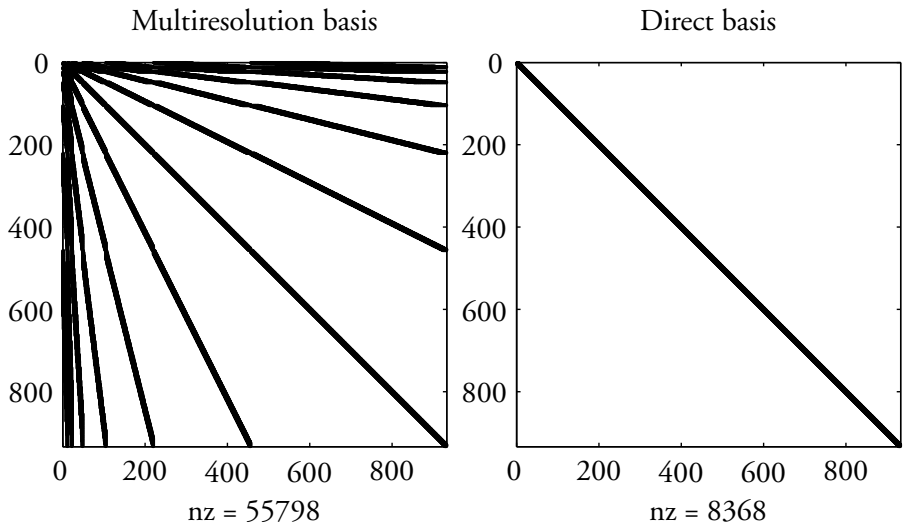

Figure 2: The non-zero elements in the **G** matrices for a multiresolution DB3 basis with five layer of wavelets and the corresponding direct basis. 6.4% of the elements are non-zero for the multiresolution basis whereas only 0.96% of the elements are non-zero for the direct basis.

3.4.2 B-spline wavelets on $\mathbb{R}$

For the B-spline wavelets, the matrices **C** and **G** can be calculated directly using the closed form expressions for the basis functions and their derivatives. When a direct basis is used on $\mathbb{R}$, **C** is a band matrix with bandwidth $m+1$, if the mth order spline wavelet is used. For example, for $m = 1$, calculating (13) gives

$$\mathbf{C}_{i,j} = 2^{-J} \cdot \begin{cases} 2/3, & i = j, \\ 1/6, & |i-j| = 1, \\ 0 & \text{otherwise}, \end{cases} \qquad \mathbf{G}_{i,j} = 2^{J} \cdot \begin{cases} 2, & i = j, \\ -1, & |i-j| = 1, \\ 0 & \text{otherwise}. \end{cases}$$

Since the expression for the precision matrix for the weights **w** contains the inverse of **C**, it is a dense matrix. Hence, $\mathbf{C}^{-1}$ has to be approximated with a sparse matrix if **Q** should be sparse. This issue is addressed in Lindgren et al. (2011) by lowering the integration order of $\langle \xi_i, \xi_j \rangle$, which results in an approximate, diagonal **C** matrix, $\tilde{\mathbf{C}}$, with diagonal elements $\tilde{\mathbf{C}}_{i,i} = \sum_{k=1}^{n} \mathbf{C}_{i,k}$. In Section 4,

the effect of this approximation on the covariance approximation for the basis expansion is studied in some detail. For the multiresolution basis, the matrices are block diagonal, and this approximation is not applicable.

3.4.3 Wavelets on $\mathbb{R}^d$

The easiest way of constructing a wavelet basis for $L^2(\mathbb{R}^d)$ is to use the tensor product functions generated by d one-dimensional wavelet bases. Let φ be the scaling function for a multiresolution on $\mathbb{R}$, the father function can be written as $\bar{\varphi}(x_1, \ldots, x_d) = \prod_{i=1}^{d} \varphi(x_i)$. The scalar product $\langle \nabla \bar{\varphi}(\mathbf{x}),\ \nabla \bar{\varphi}(\mathbf{x} + \boldsymbol{\eta}) \rangle$, where $\boldsymbol{\eta}$ now is a multi-integer shift in d dimensions, can then be written as

$$
\begin{aligned}
\langle \nabla \bar{\varphi}(\mathbf{x}),\ \nabla \bar{\varphi}(\mathbf{x} + \boldsymbol{\eta}) \rangle &= \left\langle \nabla \prod_{i=1}^{d} \varphi(x_i),\ \nabla \prod_{i=1}^{d} \varphi(x_i + \eta_i) \right\rangle \\
&= \sum_{i=1}^{d} \int_{\mathbb{R}^d} \frac{\partial\, \varphi(x_i)}{\partial\, x_i} \frac{\partial\, \varphi(x_i + \eta_i)}{\partial\, x_i} \prod_{j \neq i} \varphi(x_j) \varphi(x_j + \eta_j)\, \mathrm{d}\mathbf{x} \\
&= \sum_{i=1}^{d} \Lambda(\eta_i) \prod_{j \neq i} \int_{\mathbb{R}} \varphi(x_j) \varphi(x_j + \eta_j)\, \mathrm{d}x_j.
\end{aligned}
$$

This expression looks rather complicated, but it implies a very simple Kronecker structure for $\mathbf{G}_d$, the $\mathbf{G}$ matrix in $\mathbb{R}^d$. For example, in $\mathbb{R}^2$ and $\mathbb{R}^3$,

$$
\begin{aligned}
\mathbf{G}_2 &= \mathbf{G}_1 \otimes \mathbf{C}_1 + \mathbf{C}_1 \otimes \mathbf{G}_1 \\
\mathbf{G}_3 &= \mathbf{G}_1 \otimes \mathbf{C}_1 \otimes \mathbf{C}_1 + \mathbf{C}_1 \otimes \mathbf{G}_1 \otimes \mathbf{C}_1 + \mathbf{C}_1 \otimes \mathbf{C}_1 \otimes \mathbf{G}_1,
\end{aligned}
$$

where $\mathbf{G}_1$ and $\mathbf{C}_1$ are the $\mathbf{G}$ and $\mathbf{C}$ matrices for the corresponding one-dimensional basis and $\otimes$ denotes the Kronecker product. Similarly, $\mathbf{C}_2 = \mathbf{C}_1 \otimes \mathbf{C}_1$, and $\mathbf{C}_3 = \mathbf{C}_1 \otimes \mathbf{C}_1 \otimes \mathbf{C}_1$. These expressions hold both if the direct basis for V_J if used or if the multiresolution construction (12) is used for the one-dimensional spaces. For Daubechies wavelets, the $\mathbf{C}$ matrix is the identity matrix for all $d \geq 1$. This also holds for the direct B-spline basis if the diagonal approximation is used for $\mathbf{C}_1$.

4 Comparison

As discussed in Section 2, computational efficiency is often an important aspect in practical applications. However, the computation time for obtaining, for example, an approximate kriging prediction is in itself not that interesting unless one also knows how accurate it is. We will, therefore, in this section compare the wavelet Markov approximations with two other popular methods, covariance tapering and process convolutions, with respect to their accuracy and computationally efficiency when used for kriging.

Before the comparison, we give a brief introduction to the process convolution method and the covariance tapering method and discuss the methods' computational properties. As mentioned in Section 2, the computational cost for the kriging prediction for a single location based on m observations is $\mathcal{O}(m^3)$. In what follows, the corresponding computational costs for the three different approximation methods are presented. We start with the wavelet Markov approximations and then look at the process convolutions and the covariance tapering method. After this, an initial comparison of the different wavelet approximations is performed in Section 4.4 and then the full kriging comparison is presented in Section 4.5-4.6.

4.1 Wavelet approximations

When using a wavelet basis, one can either work with the direct basis for the approximation space V_J or do the wavelet decomposition into the direct sum of $J-1$ wavelet spaces and V_0. If one only is interested in the approximation error, the decomposition into wavelet spaces is not necessary and it is more efficient to work in the direct basis for V_J since this will result in a precision matrix with fewer nonzero elements. Therefore we only use the direct bases for V_J in the comparisons in this section.

The wavelet approximations are of the form (1), so Equation (4) is used to calculate the kriging predictor. However, since an explicit expression for the precision matrix for the weights $\mathbf{w}$ exists for this method, we rewrite the equation as

$$\mathsf{E}(\mathbf{X}_2|\mathbf{Y},\boldsymbol{\gamma}) = \mathbf{B}_2(\mathbf{Q}_w + \mathbf{B}_1^\top \mathbf{Q}_{\mathcal{E}} \mathbf{B}_1)^{-1} \mathbf{B}_1 \mathbf{Q}_{\mathcal{E}} \mathbf{Y},$$

where $\mathbf{Q}_{\mathcal{E}} = \boldsymbol{\Sigma}_{\mathcal{E}}^{-1}$ is diagonal if $\mathcal{E}$ is Gaussian white noise. If the number of kriging locations is small, the computationally demanding step is again to solve a

system of the form

$$\mathbf{u} = (\mathbf{Q}_w + \mathbf{B}_1^\top \mathbf{Q}_\mathcal{E} \mathbf{B}_1)^{-1} \mathbf{v}.$$

Now, if the Daubechies wavelets or the Markov approximated spline wavelets are used, both $\mathbf{Q}_w$ and $\mathbf{B}_1^\top \mathbf{Q}_\mathcal{E} \mathbf{B}_1$ are sparse and positive definite matrices. The system is therefore most efficiently solved using Cholesky factorization, forward substitution, and back substitution. The forward substitution and back substitution are much faster than calculating the Cholesky triangle $\mathbf{L}$, so the computational complexity for the kriging predictor is determined by the calculation of $\mathbf{L}$. Because of the sparsity structure, this computational cost is in general $\mathcal{O}(n)$, $\mathcal{O}(n^{3/2})$, and $\mathcal{O}(n^2)$ for problems in one, two, and three dimensions respectively (see Rue and Held, 2005). If the spline bases are used without the markov approximation, the computational cost instead is $O(n^3)$ since $\mathbf{Q}_w$ then is dense. It should be noted here that any basis could be used in the SPDE approximation, but in order to get good computational properties we need both $\mathbf{Q}_w$ and $\mathbf{B}_1^\top \mathbf{Q}_\mathcal{E} \mathbf{B}_1$ to be sparse. This is the reason for why for example Fourier bases are not appropriate to use in the SPDE formulation since $\mathbf{B}_1$ in this case always is a dense matrix.

4.2 Process convolutions

In the process convolution method, the Gaussian random field $X(\mathbf{s})$ on $\mathbb{R}^d$ is specified as a process convolution

$$X(\mathbf{s}) = \int K(\mathbf{s}, \mathbf{u}) \mathcal{B}(\,\mathrm{d}\mathbf{u}), \tag{15}$$

where K is some deterministic kernel function and $\mathcal{B}$ is a Brownian sheet. One of the advantages with this construction is that nonstationary fields easily are constructed by allowing the convolution kernel to be dependent on location. If, however, the process is stationary we have $K(\mathbf{s}, \mathbf{u}) = K(\mathbf{s} - \mathbf{u})$ and the covariance function for X is $r(\boldsymbol{\tau}) = \int K(\mathbf{u} - \boldsymbol{\tau}) K(\mathbf{u})\,\mathrm{d}\mathbf{u}$. Thus, the covariance function and the kernel K are related through

$$K = \mathcal{F}^{-1}\left(\frac{1}{(2\pi)^{\frac{d}{2}}}\sqrt{\mathcal{F}(r)}\right) = \mathcal{F}^{-1}\left(\frac{1}{(2\pi)^{\frac{d}{2}}}\sqrt{S}\right),$$

where S is the spectral density for $X(\mathbf{s})$ and $\mathcal{F}$ denotes the Fourier transform (Higdon, 2001). Since the spectral density for a Matérn covariance function in

dimension d with parameters ν, ϕ^2, and κ is given by (6), one finds that the corresponding kernel is a Matérn covariance function with parameters $\nu_k = \frac{\nu}{2} - \frac{d}{4}$, $\phi_k^2 = \phi$, and $\kappa_k = \kappa$.

An approximation of (15) which is commonly used in convolution-based modeling is

$$X(\mathbf{s}) \approx \sum_{j=1}^{n} K(\mathbf{s} - \mathbf{u}_j) w_j,$$

where $\mathbf{u}_1, \ldots, \mathbf{u}_n$ are some fixed locations in the domain, and w_j are independent zero-mean Gaussian variables with variances equal to the area associated with each $\mathbf{u}_j$. Thus, this approximation is of the form (1), with basis functions $\xi_j(\mathbf{s}) = K(\mathbf{s} - \mathbf{u}_j)$. When this approximation is used, Equation (4) is used to calculate the kriging predictor. Because the basis functions are Matérn covariance functions, the matrices $\mathbf{B}_1$ and $\mathbf{B}_2$ are dense. Thus, even though both $\mathbf{\Sigma}_{\mathcal{E}}$ and $\mathbf{\Sigma}_w^{-1}$ are diagonal matrices, one still have to solve a system of the form

$$\mathbf{u} = (\mathbf{\Sigma}_w^{-1} + \mathbf{B}_1^\top \mathbf{\Sigma}_{\mathcal{E}}^{-1} \mathbf{B}_1)^{-1} \mathbf{v},$$

where $(\mathbf{\Sigma}_w^{-1} + \mathbf{B}_1^\top \mathbf{\Sigma}_{\mathcal{E}}^{-1} \mathbf{B}_1)$ is a dense $n \times n$ matrix and n is the number of basis functions used in the basis expansion. The computational cost for both constructing and inverting the matrix is $\mathcal{O}(mn^2 + n^3)$. For kriging prediction of $\hat{m}$ locations, the total computational complexity is $\mathcal{O}(\hat{m}n + mn^2 + n^3)$.

4.3 Covariance tapering

Covariance tapering is not a method for constructing covariance models, but a method for approximating a given covariance model to increase the computational efficiency. The idea is to taper the true covariance, $r(\boldsymbol{\tau})$, to zero beyond a certain range, θ, by multiplying the covariance function with some compactly supported positive definite taper function $r_\theta(\boldsymbol{\tau})$. Using the tapered covariance,

$$r_{tap}(\boldsymbol{\tau}) = r_\theta(\boldsymbol{\tau}) r(\boldsymbol{\tau}),$$

the matrix $\mathbf{\Sigma}_Y$ in the expression for the kriging predictor (3) is sparse, which facilitates the use of sparse matrix techniques that increases the computational efficiency. The taper function should, of course, also be chosen such that the basic shape of the true covariance function is preserved, and of especial importance for asymptotic considerations is that the smoothness at the origin is preserved.

Furrer et al. (2006) studied the accuracy and numerical efficiency of tapered Matérn covariance functions, and to be able to compare their results to Matérn approximations obtained by the wavelet Hilbert space approximations and the process convolution method, we use their choice of taper functions:

$$\text{Wendland}_1\text{:}\ r_\theta(\boldsymbol{\tau}) = \left(\max\left[1 - \frac{\|\boldsymbol{\tau}\|}{\theta}, 0\right]\right)^4 \left(1 + 4\frac{\|\boldsymbol{\tau}\|}{\theta}\right),$$

$$\text{Wendland}_2\text{:}\ r_\theta(\boldsymbol{\tau}) = \left(\max\left[1 - \frac{\|\boldsymbol{\tau}\|}{\theta}, 0\right]\right)^6 \left(1 + 6\frac{\|\boldsymbol{\tau}\|}{\theta} + \frac{35\|\boldsymbol{\tau}\|^2}{2\theta^2}\right).$$

These taper functions were first introduced by Wendland (1995). For dimension $d \leq 3$, the Wendland$_1$ function is a valid taper function for the Matérn covariance function if $\nu \leq 1.5$, and the Wendland$_2$ function is a valid taper function if $\nu \leq 2.5$. Furrer et al. (2006) found that Wendland$_1$ was slightly better than Wendland$_2$ for a given ν, so we use Wendland$_1$ for all cases when $\nu \leq 1.5$ and Wendland$_2$ if $1.5 < \nu \leq 2.5$.

If a tapered Matérn covariance is used, the kriging predictor can be written as

$$\mathsf{E}(\mathbf{X}_2|\mathbf{Y}, \boldsymbol{\gamma}) = \boldsymbol{\Sigma}^{tap}_{X_2X_1}(\boldsymbol{\Sigma}^{tap}_{X_1} + \boldsymbol{\Sigma}_{\mathcal{E}})^{-1}\mathbf{Y},$$

where the element on row i and column j in $\boldsymbol{\Sigma}^{tap}_{X_2X_1}$ and $\boldsymbol{\Sigma}^{tap}_{X_1}$ are given by $r_{tap}(\hat{\mathbf{s}}_i, \mathbf{s}_j)$ and $r_{tap}(\mathbf{s}_i, \mathbf{s}_j)$ respectively. Since the tapered covariance is zero for lags larger than the taper range, θ, many of the elements in $\boldsymbol{\Sigma}^{tap}_{X_1}$ will be zero. Thus, the three step approach used for the wavelet Markov approximations can be used to solve the system $\mathbf{u} = (\boldsymbol{\Sigma}^{tap}_{X_1} + \boldsymbol{\Sigma}_{\mathcal{E}})^{-1}\mathbf{Y}$ efficiently. Since the number of non-zero elements for row i in $\boldsymbol{\Sigma}^{tap}_{X_1}$ is determined by the number of measurement locations at a distance smaller than θ from location $\mathbf{s}_i$, the computational cost is determined both by the taper range and the spacing of the observations. Thus, if the measurements are irregularly spaced, it is hard to get a precise estimate of the computational cost. However, for given measurement locations, the taper range can be chosen such that the average number of neighbors to the measurement locations is some fixed number k_θ. The cost for the Cholesky factorization is then similar to the cost for a GMRF with m nodes and a neighborhood size k_θ.

4.4 Covariance approximation

For practical applications of any of the approximation methods discussed here, one is often mostly interested in producing kriging predictions which are close

to the optimal predictions. The error one makes in the kriging prediction is closely related to the method's ability to reproduce the true Matérn covariance function. There are many different wavelet bases one could consider using in the Markov approximation method, and before we consider the kriging problem we will in this section compare some of these bases with respect to their ability to reproduce the Matérn covariance function so that we can choose only a few of the best methods to compare in the next section. As a reference, we also include the process convolution approximation in this comparison.

A natural measure of the error in the covariance approximation is a standardized L^2 norm of the difference between the true-, and approximate covariance functions,

$$\varepsilon_r(\mathbf{s}) = \frac{\int (r(\mathbf{s},\mathbf{u}) - \hat{r}(\mathbf{s},\mathbf{u}))^2 \, \mathrm{d}\mathbf{u}}{\int r(\mathbf{s},\mathbf{u})^2 \, \mathrm{d}\mathbf{u}}. \tag{16}$$

Note here that the true covariance function $r(\mathbf{s},\mathbf{u})$ is stationary and isotropic, while the approximate covariance function $\hat{r}(\mathbf{s},\mathbf{u})$, for the basis expansion (1), generally is not. For the wavelet approximations and the process convolutions, ε_r is periodic in $\mathbf{s}$ since the approximation error in general is smaller where the basis functions are centered, and we therefore use the mean value of $\varepsilon_r(\mathbf{s})$ over the studied region as a measure of the covariance error.

We use the different methods to approximate the covariance function for a Matérn field on the square $[0,10] \times [0,10]$ in $\mathbb{R}^2$. The computational complexity for the kriging predictions depend on the number of basis functions, n, used in the approximations. For the Markov approximated spline bases and the Daubechies 3 basis, this complexity is $O(n^{3/2})$ whereas it is $O(n^3)$ for the spline bases if the Markov approximation is not used and for the process convolution method. We therefore use 100^2 basis functions for the $O(n^{3/2})$ methods and 100 basis functions for the other methods to get the covariance error for the methods when the computational cost is approximately equal.

Figure 3 shows the covariance error for the different methods as functions of the approximate range, $\kappa^{-1}\sqrt{8\nu}$, of the true covariance function for three different values of ν. There are several things to note in this figure:

1. The covariance error decreases for all methods as the range of the true covariance function increases. This is not surprising since the error will be small if the distance between the basis functions (which is kept fixed) is small compared to the true range.

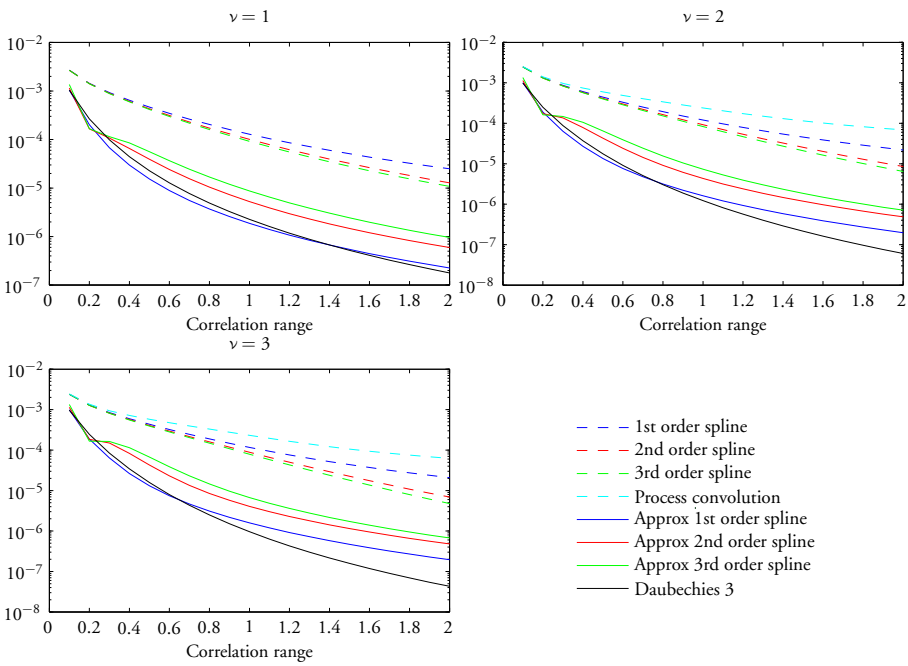

Figure 3: Numeric approximations of the L^2-norm (16) shown as a function of approximate range for different values of ν and different bases in $\mathbb{R}^2$. In this figure, 100^2 basis functions are used for the bases with Markov structure (solid lines), and 100 basis functions are used for the other bases (dashed lines). This gives approximately the same computational complexity for kriging prediction.

2. The solid lines correspond to Markov approximations, which have computational complexity $\mathcal{O}(n^{3/2})$ for calculating the kriging predictor, and the approximations with computational complexity $\mathcal{O}(n^3)$ have dashed lines in the figure.

3. There is no convolution kernel estimate for $\nu = 1$ since the convolution kernel has a singularity at the origin in this case. For the other cases, the locations $\{u_j\}$ for the kernel basis functions were placed on a regular 10×10 lattice in the region.

4. The error one makes by the Markov approximation of the spline bases becomes larger for increasing order of the splines. Note that the third order

spline basis is best without the approximation whereas the first order spline basis is best if the Markov approximation is used.

It is clear from the figure that the Markov approximations have a much lower covariance error for the same computational complexity. Among these, the Daubechies 3 basis is best for large ranges whereas the Markov approximated first order spline basis is best for short ranges. The higher order spline bases have larger covariance errors so, from now on, we focus on the first order spline basis and the Daubechies 3 basis.

4.5 Spatial prediction

In the previous section, several bases were compared with respect to their ability to approximate the true covariance function when used in an approximation of the form (1) of a Gaussian Matérn field. The comparison showed that the Daubechies 3 (DB3) basis and the Markov approximated linear spline (S1) basis are most accurate for a given computational complexity. In this section, the spatial prediction errors for these two wavelet Markov approximations are compared with the process convolution method and the covariance tapering method. In the comparisons, note that the S1 basis is essentially of the same type of piecewise linear basis as used in Lindgren et al. (2011), so the results here also apply to that paper.

Simulation setup

Let $X(\mathbf{s})$ be a Matérn field with shape parameter ν (chosen later as 1, 2, or 3) and approximate correlation range r (later varied between 0.1 and 4). The range r determines κ through the relation $\kappa = \sqrt{8\nu}r^{-1}$ and the variance parameter $\phi = 4\pi\Gamma(\nu+1)\kappa^{\nu}\Gamma(\nu)^{-1}$ is chosen such that the variance of $X(\mathbf{s})$ is 1. We measure X at 5000 measurement locations chosen at random from a uniform distribution on the square $[0,5] \times [0,5]$ in $\mathbb{R}^2$ using the measurement equation (2), where $\mathcal{E}(\mathbf{s})$ is Gaussian white noise uncorrelated with X with standard deviation $\sigma = 0.01$.

Given the measurements, spatial prediction of X to all locations on a 70×70 lattice of equally spaced points in the square is performed using the optimal kriging predictor, the wavelet Markov approximations, the process convolution method, and the covariance tapering method. For each approximate method, the sum of squared differences between the optimal kriging prediction and the approximate method's kriging prediction is used as a measure of kriging error.

We compare the methods for $\nu = 1, 2, 3$, and for each ν we test 40 different ranges varied between 0.1 and 4 in steps of 0.1. For a given ν and a given range, 20 data sets are simulated and the average kriging error is calculated for each method based on these data sets.

Choosing the number of basis functions

To obtain a fair comparison between the different methods, the number of basis functions for each method should be chosen such that the computation time for the kriging prediction is equal for the different methods. The computations needed for calculating the prediction can be divided into three main steps as follows

Step 1. Build all matrices except $\mathbf{M}$ in Step 3 necessary to calculate the kriging predictor.

Step 2. Solve the matrix inverse problem for the given method:

$$\begin{aligned}
&\text{S1, DB3 and Conv.:} && \mathbf{u} = (\boldsymbol{\Sigma}_w^{-1} + \mathbf{B}_1^\top \boldsymbol{\Sigma}_{\mathcal{E}}^{-1}\mathbf{B}_1)^{-1}\mathbf{B}_1\boldsymbol{\Sigma}_{\mathcal{E}}^{-1}\mathbf{Y},\\
&\text{Tapering:} && \mathbf{u} = \left(\boldsymbol{\Sigma}_{X_1}^{tap} + \boldsymbol{\Sigma}_{\mathcal{E}}\right)^{-1}\mathbf{Y},\\
&\text{Optimal kriging:} && \mathbf{u} = \left(\boldsymbol{\Sigma}_{X_1} + \boldsymbol{\Sigma}_{\mathcal{E}}\right)^{-1}\mathbf{Y}.
\end{aligned}$$

Step 3. Depending on which method that is used, build $\mathbf{M} = \mathbf{B}_2$, $\mathbf{M} = \boldsymbol{\Sigma}_{X_2X_1}^{tap}$, or $\mathbf{M} = \boldsymbol{\Sigma}_{X_2X_1}$ and calculate the kriging predictor $\hat{\mathbf{X}} = \mathbf{M}\mathbf{u}$.

For the optimal kriging predictor, and in some cases for the tapering method, the matrix $\mathbf{M}$ cannot be calculated and stored at once due to memory constraints if the number of measurements is large. Each element in $\hat{\mathbf{X}}$ is then constructed separately as $\hat{\mathbf{X}}_i = \mathbf{M}_i\mathbf{u}$, where $\mathbf{M}_i$ is a row in $\mathbf{M}$. It is then natural to include the time it takes to build the rows in $\mathbf{M}$ in the time it takes to calculate $\hat{\mathbf{X}}$, which is the reason for including the time it takes to build $\mathbf{M}$ in Step 3 instead of Step 1.

The computation time for the first step is highly dependent on the actual implementation, and we will therefore focus on the computation time for the last two steps when choosing the number of basis functions. If one only does kriging prediction to a few locations, the second step will dominate the computation time whereas the third step can dominate if kriging is done to many locations. To get results that are easier to interpret, we choose the number of basis functions such

that the time for the matrix inverse problem in Step 2 is similar for the different methods.

Now since the computational complexity for Step 2 is $O(n^3)$ for the convolution method and $O(n^{3/2})$ for the Markov methods, one would think that if n basis functions are used in the convolution method and n^2 basis functions are used for the Markov methods, the computation time would be equal. Unfortunately it is not that simple. If two different methods have computational complexity $O(n^3)$, this means that the computation time scales as n^3 when n is increased for both methods; however, the actual computation time for a *fixed* n can be quite different for the two methods. For example, DB3 is approximately 6 times more computationally demanding than S1 for the same number of basis functions. The reason being that the DB3 basis functions have larger support than the S1 basis functions and this causes the matrices $\mathbf{B}_1$ and $\mathbf{\Sigma}_w^{-1}$ for DB3 to contain approximately 6 times as many nonzero elements compared to S1 for the same number of basis functions. However, the relative computation time will scale as $n_1^{3/2}$ if n_1 is increased for both methods.

To get approximately the same computation time for Step 2 for the different approximation methods, the number of basis functions for S1 is fixed to 100^2. Since DB3 is approximately six times more computationally demanding, the number of basis functions for this method is set to 1600. As mentioned in Lindgren et al. (2011), one should extend the area somewhat to avoid boundary effects from the SPDE formulation used in the Markov methods. We therefore expand the area with two times the range in each direction which results in a slightly higher number of basis functions used in the computations.

The computation time for S1 and DB3 increases if ν is increased since the precision matrix for the weights contain more nonzero elements for larger values of ν. Therefore we use 625 basis functions placed on a regular 25×25 lattice in the kriging area for the convolution method when $\nu = 2$ and use 841 basis functions placed on a regular 29×29 lattice when $\nu = 3$. For the tapering method we chose the tapering range θ such that the expected number of measurements within a circle with radius θ to each kriging location is similar to the number of neighbors to the weights in the S1 method. For $\nu = 1$, $\nu = 2$, and $\nu = 3$ this gives a tapering ranges of 0.4, 0.55, and 0.7 respectively and results in approximately the same number of nonzero elements in the tapered covariance matrix as in the precision matrix Q for the S1 basis.

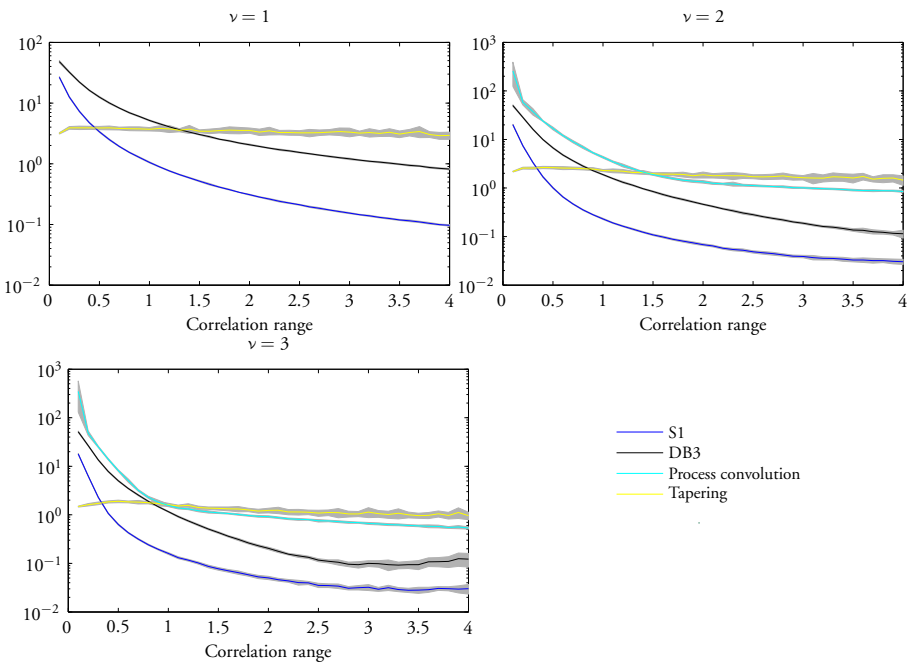

Figure 4: Kriging errors for the different methods as functions of the true covariance function's range. For each range, the values are calculated as the mean of 20 simulations. The lower limit of the bands around the curves is the estimate minus the standard deviation of the samples, and the upper limit is the estimate plus the standard deviation.

Results

Figure 4 shows the average kriging errors for the different methods as functions of the true covariance function's approximate range r. The values for a given ν and r is an average of 20 simulations. The convolution kernels are singular if $\nu = 1$, so there is no convolution estimate for this case. The tapering estimate is best for short ranges, which is not surprising since the covariance matrix for the measurements not is changed much by the tapering if the true range then is shorter than the tapering range. For larger ranges, however, the tapering method has a larger error than the other methods. One reason for this is that the tapered covariance function is very different from the true covariance function if the true range is much larger than the tapering range. Another reason is that the prediction

$\nu = 1$

	Step 1		Step 2		Step 3		Total	
Optimal	37.68	(6.357)	5.074	(0.277)	36.48	(6.231)	79.23	(8.906)
DB3	0.490	(0.049)	0.113	(0.014)	0.293	(0.026)	0.896	(0.057)
S1	0.423	(0.027)	0.088	(0.007)	0.248	(0.018)	0.759	(0.033)
Conv.	–	–	–	–	–	–	–	–
Taper	2.771	(0.191)	0.117	(0.010)	2.051	(0.127)	4.939	(0.229)

$\nu = 2$

Optimal	36.19	(6.965)	5.327	(0.529)	34.94	(6.695)	76.45	(9.675)
DB3	0.600	(0.090)	0.228	(0.039)	0.310	(0.049)	1.138	(0.110)
S1	0.489	(0.055)	0.203	(0.025)	0.260	(0.036)	0.951	(0.070)
Conv.	0.961	(0.027)	0.217	(0.019)	0.942	(0.027)	2.120	(0.043)
Taper	4.184	(1.523)	0.247	(0.028)	3.319	(0.251)	7.750	(1.543)

$\nu = 3$

Optimal	42.75	(6.572)	5.468	(0.380)	41.36	(6.440)	89.58	(9.210)
DB3	0.759	(0.091)	0.394	(0.051)	0.315	(0.033)	1.468	(0.110)
S1	0.569	(0.042)	0.377	(0.035)	0.266	(0.025)	1.213	(0.060)
Conv.	5.656	(1.094)	0.390	(0.024)	5.522	(1.078)	11.57	(1.537)
Taper	6.413	(1.051)	0.421	(0.035)	5.460	(0.402)	12.30	(1.126)

Table 1: Average computation times in seconds for the results in Figure 4. The values are based on the 800 simulations for each value of ν. The standard deviations are shown in the parentheses.

for all locations that do not have any measurements closer than the tapering range is zero in the tapering method. The convolution method has a similar problem if the effective range of the basis functions is smaller than the distance between the basis functions. In this case, the estimates for all locations that are not close to the center of some basis function have a large bias towards zero. These problems can clearly be seen in Figure 5, where the optimal kriging prediction, and the predictions for S1, the tapering method, and the convolution method, are shown for an example where $\nu = 2$ and the range is 1.

The computation times for the different methods are shown in Table 1. These computation times are obtained using an implementation in Matlab on a com-

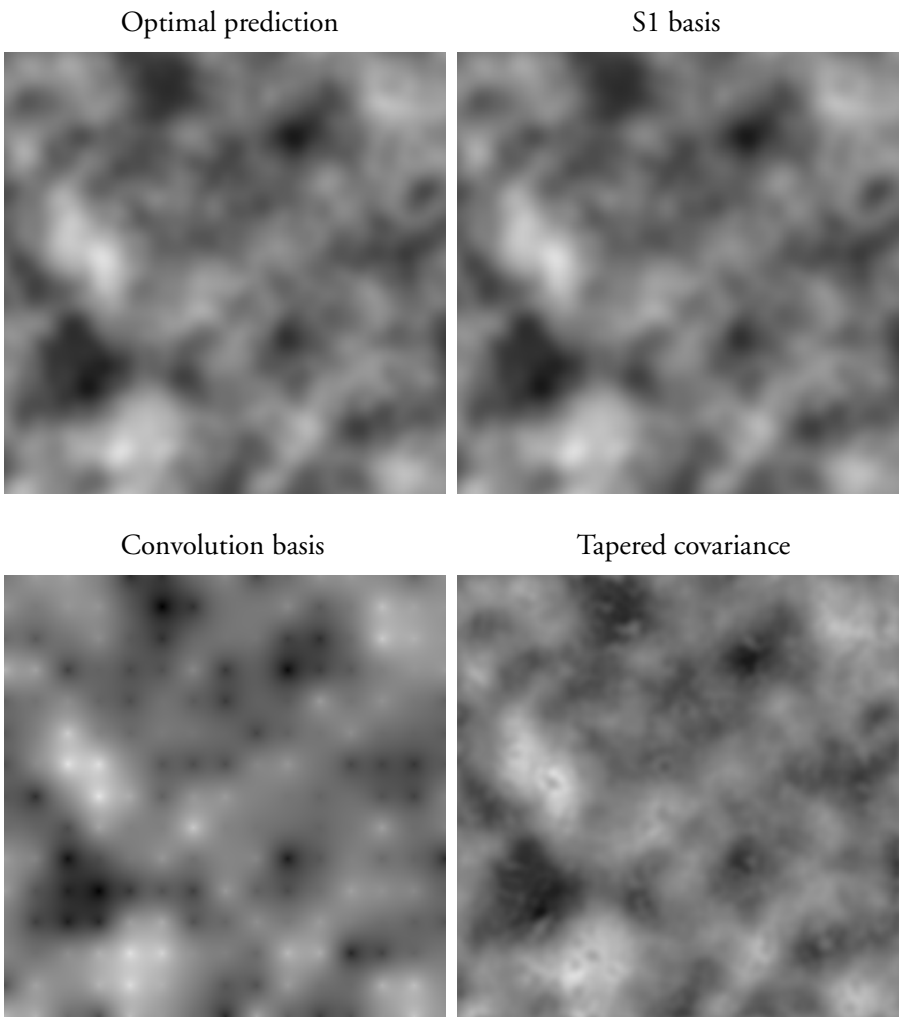

Figure 5: An example of an optimal kriging prediction and predictions using the S1 basis, the convolution basis, and a tapered covariance when $\nu = 2$ and the covariance range is 1. The predictions are based on 5000 observations and are calculated for a 200×200 grid in the square $[0, 5] \times [0, 5]$. The number of basis functions and the tapering range are chosen such that the total time for Step 2 and Step 3 is approximately equal for the different methods.

puter with a 3.33GHz Intel Xeon X5680 processor[1]. As intended, the time for Step 2 is similar for the different methods whereas there is a larger difference between the computation times for Step 3 because the computation time for the kriging prediction scales differently with the number of kriging locations for the different methods. Note that the wavelet methods are less computationally demanding than the tapering method and the convolution method when doing kriging to many locations. The reason being that the matrix $\mathbf{M}$ in Step 3 can be constructed without having to do costly covariance function evaluations.

As mentioned previously is the computation time for Step 1 highly dependent on the actual implementation. However, as for Step 3 can the Markov method's matrices be constructed without doing any covariance function evaluations which is the reason for the faster computation time. One thing to note here is that if the parameters are changed (for example when doing parameter estimation), one does not have to construct all matrices again in the Markov methods as one has to do for the other two methods.

In conclusion we see that S1 is both faster and has a smaller kriging error for all ranges when compared to DB3 and the convolution method and compared to the tapering method it has a smaller kriging error for all but very short ranges. Since the tapering method's computational cost varies with the tapering range, we conclude this section with a study of how changing the tapering range changes the results in order to get a better understanding of which method is to prefer when comparing S1 and the tapering method.

4.6 A study of varying the tapering range

As shown above is the S1 method to prefer over the DB3 method and the convolution method in all our test cases whereas the tapering method had a smaller kriging error for very short ranges. Since this was done using a fixed tapering range, chosen such that the computation time for Step 2 would be similar to the other methods, we now look at what happens if the tapering range is varied when keeping the true range fixed.

The setup is the same as in the previous comparison, a Matérn field with $\nu = 2$, variance 1, and an approximate range r is measured at 5000 randomly chosen locations in a square in $\mathbb{R}^2$. The difference is that we now keep these parameters fixed but instead vary the tapering range from 0.05 to 2 in steps of

[1] implementation available at `http://www.maths.lth.se/matstat/staff/bolin/`

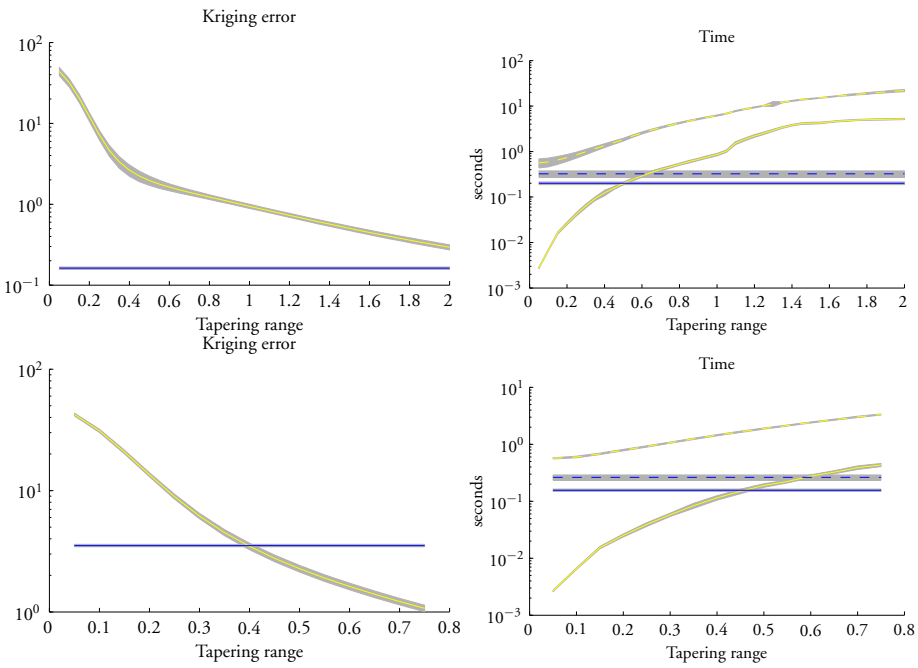

Figure 6: The computation times (right) and kriging errors (left) for the covariance tapering method (yellow lines) as functions of the taper range. Values for the S1 basis (blue lines) are shown for comparison. The range of the true covariance function is 1 (upper panels) and 0.25 (lower panels). The results are averages of 100 simulations, and the grey bands indicate the standard deviation of these samples. The solid lines in the right panels show the computation time for Step 2 and the dashed lines show the total computation time for Step 2 and Step 3.

0.05. We generate 100 data sets and calculate the kriging predictions for the S1 method and the tapering method for all values of the tapering range. Based on these 100 estimates, the average kriging error is calculated for S1 and for each tapering estimate.

The results can be seen in Figure 6. The kriging errors are shown in the left panels and the computation times are shown in the right panels. The blue lines represent the S1 method, which obviously does not depend on the tapering range, and the yellow lines represent the tapering method. In the left panels, the solid lines show the time for Step 2 in the computations and the dashed lines show the

total time for Step 2 and Step 3. In the upper two panels, the true range r is 1, and 100^2 S1 basis functions are used. In this case, S1 is more accurate than the tapering method for all tapering ranges tested, which is not surprising considering the previous results. In the bottom panels of the figure, the true range r is 0.25 and 100^2 S1 basis functions are used. This is a situation where the tapering method was more accurate than S1 in the previous study and we see here that the tapering method is more accurate for tapering ranges larger than 0.4 and that the time for Step 2 is smaller for all tapering ranges smaller than 0.46. Thus, by choosing the tapering range between 0.4 and 0.46, the tapering method is more accurate and has a smaller computation time for Step 2.

The accuracy of the tapering method increases if the ratio between the tapering range and the true range is increased, and the computation time depends on what the distance between the measurements is compared to the tapering range. If the distance between the measurements is large, the tapering method is fast, whereas it is slower if the distance is small. Thus, the situation where the tapering method performs best is when the true covariance range is short compared to the distance between the measurements. However, also for the case when the true range is small, the total time it takes to calculate the tapering prediction is larger than the time it takes to calculate the S1 prediction unless the number of kriging locations is small.

In this work, the taper functions that Furrer et al. (2006) found to be best for each value of ν are used, but the results may be improved by using other taper functions. Changing the taper function will, however, not change the fact that the prediction for all locations that do not have any measurements closer than the tapering range is zero in the tapering method and that the tapered covariance function is very different from the true covariance function if the tapering range is short compared to the true range. Finally, the results for all methods could be improved by finding optimal parameters for the approximate models instead of using the parameters for the true Matérn covariance. For the tapering method, however, Furrer et al. (2006) found that this only changed the relative accuracy by one or two percent.

5 Conclusions

Because of the increasing number of large environmental data sets, there is a need for computationally efficient statistical models. To be useful for a broad range

of practical applications, the models should contain a wide family of stationary covariance functions, and be extendable to nonstationary covariance structures, while still allowing efficient calculations for large problems.

The SPDE formulation of the Matérn family of covariance functions has these properties, as it can be extended to more general nonstationary spatial models (see Bolin and Lindgren, 2011, Lindgren et al., 2011, for details on how this can be done), and allows for efficient and accurate Markov model representations. In addition, as shown by the simulation comparisons, these Markov methods are more efficient and accurate than both the process convolution approach and the covariance tapering method for approximating Matérn fields.

Depending on the context in which a model is used, different aspects are important to make it computationally efficient. If, for example, the model is used in MCMC simulations, one should be able to generate samples from the model given the parameters efficiently, or if the parameters are estimated in a numerical maximum likelihood procedure, one must be able to evaluate the likelihood efficiently. To limit the scope of this article, only the computational aspects of kriging was considered. However, for practical applications, parameter estimation is likely the most computationally demanding part of the analysis. If maximum likelihood estimation is performed using numerical optimization of the likelihood, matrix inverses similar to the one in Step 2 in Table 1 have to be performed in each iteration of the optimization, and it is therefore important that these inverses can be calculated efficiently. We have not discussed estimation here, but the Markov methods are likely most efficient in this situation as well because these do not require costly Bessel function evaluations when calculating the likelihood. However, this is left for future research to investigate in more detail. An introduction to maximum likelihood estimation using the SPDE formulation can be found in Bolin and Lindgren (2011) and Lindgren et al. (2011).

Finally, some relevant methods, such as Cressie and Johannesson (2008) and Banerjee et al. (2008), were not included in the comparison in order to keep it relatively short and also because they are difficult to compare with the methods discussed here. It would be interesting to include more methods in the comparison, but we leave this for future work.

References

Banerjee, S., Gelfand, A. E., Finley, A. O., and Sang, H. (2008). Gaussian predictive process models for large spatial data sets. *J. Roy. Statist. Soc. Ser. B Stat. Methodol.*, 70(4):825–848.

Barry, R. P. and Ver Hoef, J. M. (1996). Blackbox kriging: Spatial prediction without specifying variogram models. *J. Agr. Biol. Environ. Statist.*, 1(3):297–322.

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

Burrus, C., Gopinath, R., and Guo, H. (1988). *Introduction to Wavelets and Wavelet Transforms: A Primer*. Prentice-Hall, New York.

Chui, C. K. and Wang, J.-Z. (1992). On compactly supported spline wavelets and a duality principle. *T. Am. Math. Soc.*, 330:903–915.

Cressie, N. and Johannesson, G. (2008). Fixed rank kriging for very large spatial data sets. *J. Roy. Statist. Soc. Ser. B Stat. Methodol.*, 70(1):209–226.

Cressie, N. and Ravlicová, M. (2002). Calibrated spatial moving average simulations. *Statist. Model.*, 2:267–279.

Daubechies, I. (1992). *Ten Lectures on Wavelets (CBMS-NSF Regional Conference Series in Applied Mathematics)*. Soc for Industrial & Applied Math.

Furrer, R., Genton, M. G., and Nychka, D. (2006). Covariance tapering for interpolation of large spatial datasets. *J. Comput. Graph. Statist.*, 15:502–523.

Gneiting, T. (2002). Compactly supported correlation functions. 83:493–508.

Higdon, D. (2001). Space and space-time modeling using process convolutions. Technical report.

Latto, A., Resnikoff, H. L., and Tenenbaum, E. (1991). The evaluation of connection coefficients of compactly supported wavelets. In *Proceedings of the French-USA Workshop on Wavelets and Turbulence*. Springer-Verlag.

Lindgren, F. and Rue, H. (2007). Explicit construction of GMRF approximations to generalised Matérn fields on irregular grids. *Preprints in Math. Sci. Lund University*, 2007:12.

Lindgren, F., Rue, H., and Lindström, J. (2011). 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.

Matérn, B. (1960). Spatial variation. *Meddelanden från statens skogsforskningsinstitut*, 49(5).

Nychka, D., Wikle, C., and Royle, J. A. (2002). Multiresolution models for nonstationary spatial covariance functions. *Statist. Model.*, 2:315–331.

Rodrigues, A. and Diggle, P. J. (2010). A class of convolution-based models for spatio-temporal processes with non-separable covariance structure. *Scand. J. Statist.*, 37:553–567.

Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields; Theory and Applications*, volume 104 of *Monographs on Statistics and Applied Probability*. Chapman & Hall/CRC.

Rue, H. and Tjelmeland, H. (2002). Fitting Gaussian Markov random fields to Gaussian fields. *Scand. J. Statist.*, 29(1):31–49.

Schabenberger, O. and Gotway, C. (2005). *Statistical methods for spatial data analysis*. Texts in statistical science. Chapman & Hall/CRC.

Song, H., Fuentes, M., and Gosh, S. (2008). A comparative study of Gaussian geostatistical models and Gaussian Markov random field models. *J. Multivariate Anal.*, 99:1681–1697.

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

Wendland, H. (1995). Piecewise polynomial, positive definite and compactly supported radial functions of minimal degree. *Adv. Comput. Math.*, 4:389–396.

Whittle, P. (1963). Stochastic processes in several dimensions. *Bull. Internat. Statist. Inst.*, 40:974–994.

Zhang, H. (2004). Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics. *J. Amer. Statist. Assoc.*, 99(465):250–261.

C

Spatial models generated by nested stochastic partial differential equations, with an application to global ozone mapping

DAVID BOLIN AND FINN LINDGREN

Centre for Mathematical Sciences, Lund University, Lund, Sweden

Abstract: A new class of stochastic field models is constructed using nested stochastic partial differential equations (SPDEs). The model class is computationally efficient, applicable to data on general smooth manifolds, and includes both the Gaussian Matérn fields and a wide family of fields with oscillating covariance functions. Nonstationary covariance models are obtained by spatially varying the parameters in the SPDEs, and the model parameters are estimated using direct numerical optimization, which is more efficient than standard Markov Chain Monte Carlo procedures. The model class is used to estimate daily ozone maps using a large data set of spatially irregular global total column ozone data.

Key words: nested SPDEs; Matérn covariances; non-stationary covariances; total column ozone data

1 Introduction

Building models for spatial environmental data is a challenging problem that has received much attention over the past years. Nonstationary covariance models are often needed since the traditional stationary assumption is too restrictive for capturing the covariance structure in many problems. Also, many environmental data sets today contain massive amounts of measurements, which makes computational efficiency another increasingly important model property. One such data set, which will be analyzed in this work, is the Total Ozone Mapping Spectrometer (TOMS) atmospheric ozone data (McPeters et al., 1996). The data was collected by a TOMS instrument onboard the near-polar, Sun-synchronous or-

biting satellite Nimbus-7, launched by NASA on October 24, 1978. During the sunlit portions of the satellite's orbit, the instrument collected data in scans perpendicular to the orbital plane. A new scan was started every eight seconds as the spacecraft moved from south to north. A number of recent papers in the statistical literature (Cressie and Johannesson, 2008, Jun and Stein, 2008, Stein, 2007) have studied the data, and it requires nonstationary covariance structures as well as efficient computational techniques due to the large number of observations.

A covariance model that is popular in environmental statistics is the Matérn family of covariance functions (Matérn, 1960). The Matérn covariance function has a shape parameter, ν, a scale parameter, κ, and a variance[1] parameter, ϕ^2, and can be parametrized as

$$C(\mathbf{h}) = \frac{2^{1-\nu}\phi^2}{(4\pi)^{d/2}\Gamma(\nu + d/2)\kappa^{2\nu}}(\kappa\|\mathbf{h}\|)^{\nu}K_{\nu}(\kappa\|\mathbf{h}\|), \qquad \mathbf{h} \in \mathbb{R}^d, \tag{1}$$

where K_ν is a modified Bessel function of the second kind of order $\nu > 0$. One drawback with defining the model directly through a covariance function, such as (1), is that it makes nonstationary extensions difficult. Another drawback is that, unless the covariance function has compact support, the computational complexity for calculating the Kriging predictor based on n measurements is $\mathcal{O}(n^3)$. This makes the Matérn covariance model computationally infeasible for many environmental data sets.

Recently, Lindgren et al. (2011) derived a method for explicit, and computationally efficient, Markov representations of the Matérn covariance family. The method uses the fact that a random process on $\mathbb{R}^d$ with a Matérn covariance function is a solution to the stochastic partial differential equation (SPDE)

$$(\kappa^2 - \Delta)^{\alpha/2}X(\mathbf{s}) = \phi\mathcal{W}(\mathbf{s}), \tag{2}$$

where $\mathcal{W}(\mathbf{s})$ is Gaussian white noise, Δ is the Laplace operator, and $\alpha = \nu + d/2$ (Whittle, 1963). Instead of defining Matérn fields through the covariance functions (1), Lindgren et al. (2011) used the solution to the SPDE (2) as a definition. This definition is valid not only on $\mathbb{R}^d$ but also on general smooth manifolds, such as the sphere, and facilitates nonstationary extensions by allowing the SPDE parameters κ^2 and ϕ to vary with space. The Markov representations were obtained by considering approximate stochastic weak solutions to the SPDE; see Section 3 for details.

[1]With this parametrization, the variance $C(\mathbf{0})$ is $\phi^2(4\pi)^{-d/2}\Gamma(\nu)\Gamma(\nu + d/2)^{-1}\kappa^{-2\nu}$.

In this paper we extend the work by Lindgren et al. (2011) and construct a new flexible class of spatial models by considering a generalization of (2). This model class contains a wide family of covariance functions, including both the Matérn family and oscillating covariance functions, and it maintains all desirable properties of the Markov approximated Matérn model, such as computational efficiency, easy nonstationary extensions and applicability to data on general smooth manifolds.

The model class is introduced in Section 2, with derivations of some basic properties, examples of covariance functions that can be obtained from these models and a discussion on nonstationary extensions. Section 3 gives a review of the Hilbert space approximation technique and shows how it can be extended to give computationally efficient representations also for this new model class. In Section 4 a numerical parameter estimation procedure for the nested SPDE models is presented, and the section concludes with a discussion on computational complexity for parameter estimation and Kriging prediction. In Section 5 the model class is used to analyze the TOMS ozone data. In particular, all measurements available from October 1st, 1988 in the spatially and temporally irregular "Level 2" version of the data set are used. This data set contains approximately 180,000 measurements, and the nonstationary version of the model class is used to construct estimates of the ozone field for that particular day. Finally, Section 6 contains some concluding remarks and suggestions for further work.

2 Stationary nested SPDE models

A limitation with the Matérn covariance family is that it does not contain any covariance functions with negative values, such as oscillating covariance functions. One way of constructing a larger class of stochastic fields is to consider a generalization of the SPDE (2):

$$\mathcal{L}_1 X(\mathbf{s}) = \mathcal{L}_2 \mathcal{W}(\mathbf{s}), \tag{3}$$

for some linear operators $\mathcal{L}_1$ and $\mathcal{L}_2$. If $\mathcal{L}_1$ and $\mathcal{L}_2$ are commutative operators, (3) is equivalent to the following system of nested SPDEs:

$$\begin{aligned} \mathcal{L}_1 X_0(\mathbf{s}) &= \mathcal{W}(\mathbf{s}), \\ X(\mathbf{s}) &= \mathcal{L}_2 X_0(\mathbf{s}). \end{aligned} \tag{4}$$

This representation gives us an interpretation of the consequence of the additional differential operator $\mathcal{L}_2$: $X(\mathbf{s})$ is simply $\mathcal{L}_2$ applied to the solution one would get to (3) if $\mathcal{L}_2$ was the identity operator. Equation (3) generates a large class of random fields, even if the operators $\mathcal{L}_1$ and $\mathcal{L}_2$ are restricted to operators closely related to (2). One of the simplest extensions of the Matérn model is to let $\mathcal{L}_1$ be the same as in (2) and use $\mathcal{L}_2 = (b + \mathbf{B}^\top \nabla)$, where ∇ is the gradient, $b \in \mathbb{R}$, and $\mathbf{B} \in \mathbb{R}^d$. The equation then is

$$(\kappa^2 - \Delta)^{\alpha/2} X(\mathbf{s}) = (b + \mathbf{B}^\top \nabla)\mathcal{W}(\mathbf{s}), \tag{5}$$

and $X(\mathbf{s})$ is a weighted sum of a Matérn field and its directional derivative in the direction determined by the vector $\mathbf{B}$. This model is closely related to the models introduced in Jun and Stein (2007) and Jun and Stein (2008), and the connection is discussed later in Section 5. To get a larger class of models, the orders of the operators $\mathcal{L}_1$ and $\mathcal{L}_2$ can be increased, and to get a class of stochastic fields that is easy to work with, the operators are written as products, where each factor in the product is equal to one of the operators in (5). Thus, let

$$\mathcal{L}_1 = \prod_{i=1}^{n_1} (\kappa_i^2 - \Delta)^{\alpha_i/2} \tag{6}$$

for $\alpha_i \in \mathbb{N}$ and $\kappa_i^2 > 0$, and use

$$\mathcal{L}_2 = \prod_{i=1}^{n_2} (b_i + \mathbf{B}_i^\top \nabla) \tag{7}$$

for $b_i \in \mathbb{R}$ and $\mathbf{B}_i \in \mathbb{R}^d$. Hence, the SPDE generating the class of nested SPDE models is

$$\left(\prod_{i=1}^{n_1} (\kappa^2 - \Delta)^{\alpha_i/2} \right) X(\mathbf{s}) = \left(\prod_{i=1}^{n_2} (b_i + \mathbf{B}_i^\top \nabla) \right) \mathcal{W}(\mathbf{s}). \tag{8}$$

There are several alternative equations one might consider; one could, for example, let $\mathcal{L}_2$ be on the same form as $\mathcal{L}_1$, or allow for anisotropic operators on the form $(1 - \nabla^\top \mathbf{A} \nabla)$ for some positive definite matrix $\mathbf{A}$. However, to limit our scope, we will from now on only consider model (8).

2.1 Properties in $\mathbb{R}^d$

In this section some basic properties of random fields generated by (8), when $\mathbf{s} \in \mathbb{R}^d$, are given. First note that all Matérn fields with shape parameters satisfying $\nu + d/2 \in \mathbb{N}$ are contained in the class of stochastic fields generated by (8) since $(\kappa^2 - \Delta)^{\alpha/2}$ can be written on the form (6) for these values of ν. Also note that the order of the operator $\mathcal{L}_2$ cannot be larger than the order of $\mathcal{L}_1$ if $X(\mathbf{s})$ should be at least as "well behaved" as white noise; hence, one must have $\sum_{i=1}^{n_1} \alpha_i \geq n_2$. The smoothness of $X(\mathbf{s})$ is determined by the difference of the orders of the operators $\mathcal{L}_1$ and $\mathcal{L}_2$. In order to make a precise statement about this, the spectral density of $X(\mathbf{s})$ is needed.

Proposition 2.1. *The spectral density for $X(\mathbf{s})$ defined by* (8) *is given by*

$$S(\mathbf{k}) = \frac{\phi^2}{(2\pi)^d} \frac{\prod_{j=1}^{n_2}(b_j^2 + \mathbf{k}^\top \mathbf{B}_j \mathbf{B}_j^\top \mathbf{k})}{\prod_{j=1}^{n_1}(\kappa_j^2 + \|\mathbf{k}\|^2)^{\alpha_j}}.$$

This proposition is easily proved using linear filtering theory (see, for example Yaglom, 1987). Given the spectral density of $X(\mathbf{s})$, the following proposition regarding the sample function regularity can be proved using Theorem 3.4.3 in Adler (1981).

Proposition 2.2. *$X(\mathbf{s})$ defined by* (8) *has almost surely continuous sample functions if $2\sum_{i=1}^{n_1} \alpha_i - 2n_2 > d$.*

Because the stochastic field $X(\mathbf{s})$ is generated by the SPDE (8), the following corollary regarding sample path differentiability is also easily proved using the fact that the directional derivative of $X(\mathbf{s})$ is in the class of nested SPDE models.

Corollary 2.3. *Given that $2\sum_{i=1}^{n_1} \alpha_i - 2n_2 - d > m$, the mth order directional derivative of $X(\mathbf{s})$, $(\mathbf{B}^\top \nabla)^m X(\mathbf{s})$, has almost surely continuous sample functions.*

Hence, as $2\sum_{i=1}^{n_1} \alpha_i - 2n_2$ increases, the sample paths become smoother, and eventually become differentiable, twice differentiable, and so on. One could also give a more precise characterization of the sample path regularity using the notion of Hölder continuity. This is (more or less) straightforward using properties of index-β random fields (Adler, 1981), but outside the scope of this article.

A closed-form expression for the covariance function is not that interesting since none of the methods that are later presented for parameter estimation, spatial prediction or model validation require an expression for the covariance function; however, if one were to use some technique that requires the covariance

function, it can be derived. An expression for the general case is quite complicated, and will not be presented here. Instead we present a recipe for calculating the covariance function for given parameters of the SPDE, with explicit results for a few examples.

To calculate the covariance function of $X(\mathbf{s})$, first calculate the covariance function, $C_{X_0}(\mathbf{h})$, of $X_0(\mathbf{s})$, given by (4). Given this covariance function, the covariance function for $X(\mathbf{s})$ is obtained as

$$C(\mathbf{h}) = \left(\prod_{i=1}^{n_2}(b_i - \nabla^\top \mathbf{B}_i \mathbf{B}_i^\top \nabla)\right) C_{X_0}(\mathbf{h}).$$

The motivation for this expression is again directly from linear filter theory, and the d-dimensional equivalent of the formula for the covariance function for a differentiated stochastic process, $r_{X'}(\tau) = -r''_X(\tau)$. To get an expression for $C_{X_0}(\mathbf{h})$, first use Proposition 2.1 with $\mathcal{L}_2 = I$ to get the spectral density of $X_0(\mathbf{s})$. Using a partial fraction decomposition, the spectral density can be written as

$$S_{X_0}(\mathbf{k}) = \frac{\phi^2}{(2\pi)^d} \sum_{i=1}^{n} \sum_{j=1}^{\alpha_i} \frac{p_{i,j}}{(\kappa_i^2 + \|\mathbf{k}\|^2)^j}, \tag{9}$$

where $p_{i,j}$ is a real constant which can be found using the Heaviside cover-up method (see, for example Thomas and Finney, 1995, page 523). Now, by taking the inverse Fourier transform of (9), the covariance function for $X_0(\mathbf{s})$ is

$$C_{X_0}(\mathbf{h}) = \sum_{i=1}^{n} \sum_{j=1}^{\alpha_i} p_{i,j} C^j_{\kappa_i}(\mathbf{h}),$$

where $C^\nu_\kappa(\mathbf{h})$ denotes a Matérn covariance function with shape parameter ν, scale parameter κ and variance parameter ϕ^2. The final step is to use that the derivative of a Matérn covariance function can be expressed using a Matérn covariance with another shape parameter. More precisely, one has

$$\frac{\partial}{\partial h_i} C^\nu_\kappa(\mathbf{h}) = -\frac{h_i}{2\nu} C^{\nu-1}_\kappa(\mathbf{h}),$$

where h_i denotes the ith component of the vector $\mathbf{h}$. Using these calculations, one can obtain the covariance function for any field given by (8). We conclude this section by showing the covariance function for some simple cases in $\mathbb{R}^2$. The covariance functions for these examples are shown in Figure 1, and realizations of Gaussian processes with these covariance functions are shown in Figure 2.

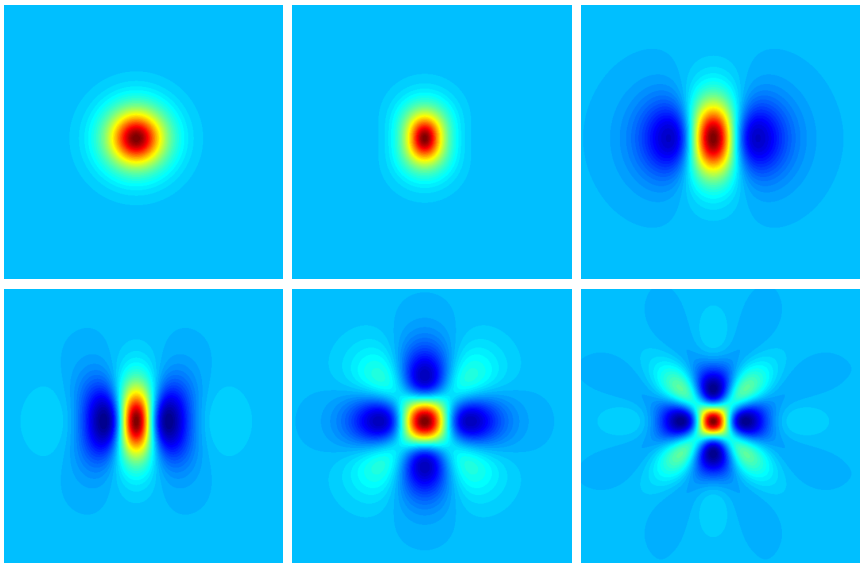

Figure 1: Covariance functions of random fields obtained from model (8) with parameters from Example 1 (top left), Example 2 (top middle and right), Example 3 (bottom left and middle) and Example 4 (bottom right).

Example 1. With $\mathcal{L}_1 = (\kappa^2 - \Delta)^{\alpha/2}$ and $\mathcal{L}_2$ as the identity operator, the standard Matérn covariance function (1) is obtained, shown in the top left panel of Figure 1.

Example 2. The simplest nested SPDE model (5) has the covariance function

$$C(\mathbf{h}) = bC_\kappa^\nu(\mathbf{h}) + \frac{\mathbf{B}^\top\mathbf{B}}{2\nu}C_\kappa^{\nu-1}(\mathbf{h}) - \frac{\mathbf{h}^\top\mathbf{B}\mathbf{B}^\top\mathbf{h}}{4\nu(\nu-1)}C_\kappa^{\nu-2}(\mathbf{h}).$$

A stochastic field with this covariance function is obtained as a weighted sum of a Matérn field $X_0(\mathbf{s})$ and its directional derivative in the direction of $\mathbf{B}$. The field therefore has a Matérn-like behavior in the direction perpendicular to $\mathbf{B}$ and an oscillating behavior in the direction of $\mathbf{B}$. In the upper middle panel of Figure 1, this covariance function is shown for $\mathbf{B} = (1,0)^\top$, $\nu = 3$, and $b = 5$. In the upper right panel of Figure 1, it is shown for $\mathbf{B} = (1,0)^\top$, $\nu = 3$, and $b = 0$.

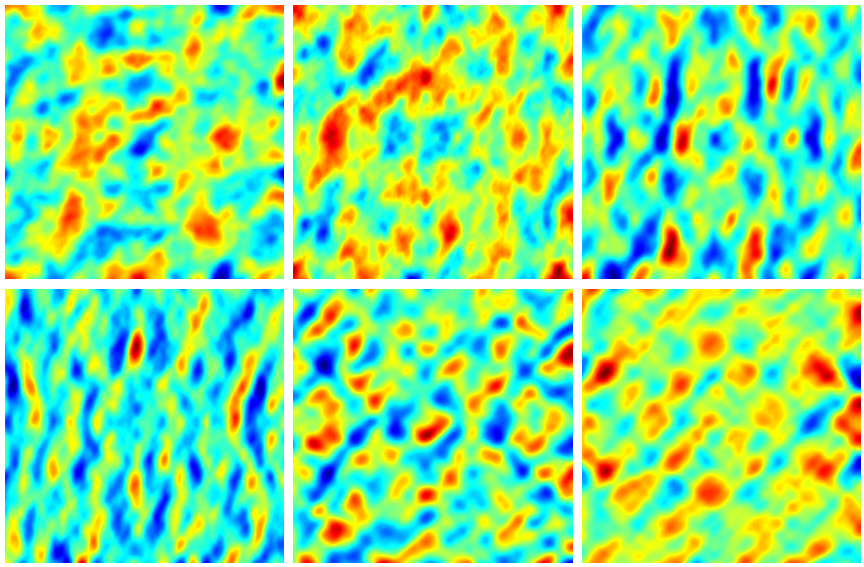

Figure 2: Realizations of random fields obtained from model (8) with different parameters. The realization in each panel corresponds to a stochastic field with the covariance function shown in the corresponding panel in Figure 1.

Example 3. The number of zero crossings of the covariance function in the direction of $\mathbf{B}$ is at most n_2. In the previous example we had $n_2 = 1$, and to obtain a more oscillating covariance function, the order of $\mathcal{L}_2$ can be increased by one:

$$(\kappa^2 - \Delta)^{\alpha/2} X(\mathbf{s}) = (b_1 + \mathbf{B}_1^\top \nabla)(b_2 + \mathbf{B}_2^\top \nabla)\mathcal{W}(\mathbf{s}).$$

This model has the covariance function

$$\begin{aligned} C(\mathbf{h}) \quad = \quad & b_1 b_2 C_\kappa^\nu(\mathbf{h}) + \frac{b_2 \mathbf{B}_1^\top \mathbf{B}_1 + b_1 \mathbf{B}_2^\top \mathbf{B}_2}{2\nu} C_\kappa^{\nu-1}(\mathbf{h}) \\ & + \frac{2(\mathbf{B}_2^\top \mathbf{B}_1)^2 + \mathbf{B}_1^\top \mathbf{B}_1 \mathbf{B}_2^\top \mathbf{B}_2 - \mathbf{h}^\top (b_1 \mathbf{B}_2 \mathbf{B}_2^\top + b_2 \mathbf{B}_1 \mathbf{B}_1^\top)\mathbf{h}}{2^2 \nu(\nu-1)} C_\kappa^{\nu-2}(\mathbf{h}) \\ & - \frac{\mathbf{h}^\top (\mathbf{B}_1 \mathbf{B}_2^\top \mathbf{B}_2 \mathbf{B}_1^\top + 4\mathbf{B}_1 \mathbf{B}_1^\top \mathbf{B}_2 \mathbf{B}_2^\top + \mathbf{B}_2 \mathbf{B}_1^\top \mathbf{B}_1 \mathbf{B}_2^\top)\mathbf{h}}{2^3 \nu(\nu-1)(\nu-2)} C_\kappa^{\nu-3}(\mathbf{h}) \\ & + \frac{(\mathbf{B}_1^\top \mathbf{h}\mathbf{h}^\top \mathbf{B}_2)^2}{2^4 \nu(\nu-1)(\nu-2)(\nu-3)} C_\kappa^{\nu-4}(\mathbf{h}). \end{aligned}$$

In the bottom left panel of Figure 1 this covariance function is shown for $\nu = 5$, $b_1 = b_2 = 0$ and $\mathbf{B}_1 = \mathbf{B}_2 = (1,0)^\top$. With these parameters, the covariance function is similar to the covariance function in the previous example, but with one more zero crossing in the direction of $\mathbf{B}$. For this specific choice of parameters, the expression for the covariance function can be simplified to

$$C(\mathbf{h}) = 3\gamma_2 C_\kappa^{\nu-2}(\mathbf{h}) - 6\gamma_3 h_1^2 C_\kappa^{\nu-3}(\mathbf{h}) + \gamma_4 h_1^4 C_\kappa^{\nu-4}(\mathbf{h}),$$

where $\gamma_k = (2^k \Pi_{i=0}^{k-1}(\nu - k))^{-1}$. In the bottom middle panel of Figure 1 the covariance function is shown for $\nu = 5$, $b_1 = b_2 = 0$, $\mathbf{B}_1 = (1,0)^\top$, and $\mathbf{B}_2 = (0,1)^\top$. Thus, the field $X_0(\mathbf{s})$ is differentiated in two different directions, and the covariance function for $X(\mathbf{s})$ therefore is oscillating in two directions. For these parameters, the covariance function can be written as

$$C(\mathbf{h}) = \gamma_2 C_\kappa^{\nu-2}(\mathbf{h}) - \gamma_3 \mathbf{h}^\top \mathbf{h} C_\kappa^{\nu-3}(\mathbf{h}) + \gamma_4 h_1 h_2 C_\kappa^{\nu-4}(\mathbf{h}).$$

Example 4. The bottom right panel of Figure 1 shows a covariance function for the nested SPDE

$$(\kappa^2 - \Delta)^{\alpha/2} X(\mathbf{s}) = (b_1 + \mathbf{B}_1^\top \nabla)^2 (b_2 + \mathbf{B}_2^\top \nabla)^2 \mathcal{W}(\mathbf{s}).$$

As in the previous examples, the covariance function for a stochastic field generated by this SPDE can be calculated and written on the form

$$C(\mathbf{h}) = \sum_{k=0}^{8} \gamma_k f_k(\mathbf{h}) C_\kappa^{\nu-k}(\mathbf{h}),$$

where $f_k(\mathbf{h}), k = 0, \ldots, 8$, are functions depending on $\mathbf{h}$ and the parameters in the SPDE. Without any restrictions on the parameters, it is a rather tedious exercise to calculate the functions $f_k(\mathbf{h})$, and we therefore only show them for the specific set of parameters that are used in Figure 1: $\nu = 7$, $b_1 = b_2 = 0$, $\mathbf{B}_1 = (1,0)^\top$ and $\mathbf{B}_2 = (0,1)^\top$. In this case $f_0(\mathbf{h}) = f_1(\mathbf{h}) = f_2(\mathbf{h}) = 0$, and the covariance function is

$$\begin{aligned} C(\mathbf{h}) \quad = \quad & 9\gamma_4 C_\kappa^{\nu-4}(\mathbf{h}) - 18\gamma_5 \mathbf{h}^\top \mathbf{h} C_\kappa^{\nu-5}(\mathbf{h}) + 3\gamma_6 (h_1^4 + h_2^4 + 12 h_1^2 h_2^2) C_\kappa^{\nu-6}(\mathbf{h}) \\ & - 6\gamma_7 h_1^2 h_2^2 \mathbf{h}^\top \mathbf{h} C_\kappa^{\nu-7}(\mathbf{h}) + \gamma_8 h_1^4 h_2^4 C_\kappa^{\nu-8}(\mathbf{h}). \end{aligned}$$

2.2 Nonstationary nested SPDE models

Nonstationarity can be introduced in the nested SPDE models by allowing the parameters κ_i, b_i and $\mathbf{B}_i$ to be spatially varying:

$$\begin{aligned}\left(\prod_{i=1}^{n_1}\left(\kappa_i^2(\mathbf{s})-\Delta\right)^{\alpha_i/2}\right)X_0(\mathbf{s}) &= \mathcal{W}(\mathbf{s}),\\ X(\mathbf{s}) &= \left(\prod_{i=1}^{n_2}\left(b_i(\mathbf{s})+\mathbf{B}_i(\mathbf{s})^\top\nabla\right)\right)X_0(\mathbf{s}).\end{aligned} \tag{10}$$

If the parameters are spatially varying, the two operators are no longer commutative, and the solution to (10) is not necessarily equal to the solution of

$$\left(\prod_{i=1}^{n_1}\left(\kappa_i^2(\mathbf{s})-\Delta\right)^{\alpha_i/2}\right)X(\mathbf{s}) = \left(\prod_{i=1}^{n_2}\left(b_i(\mathbf{s})+\mathbf{B}_i(\mathbf{s})^\top\nabla\right)\right)\mathcal{W}(\mathbf{s}). \tag{11}$$

For nonstationary models, we will from now on only study the system of nested SPDEs (10), though it should be noted that the methods presented in the next sections can be applied to (11) as well.

One could potentially use an approach where the spatially varying parameters also are modeled as stochastic fields, but to be able to estimate the parameters efficiently, it is easier to assume that each parameter can be written as a weighted sum of some known regression functions. In Section 5 this approach is used for a nested SPDE model on the sphere. In this case, one needs a regression basis $\{\psi_j(\mathbf{s})\}$ for the vector fields $\mathbf{B}_i(\mathbf{s})$ on the sphere. Explicit expressions for such a basis are given in Appendix A.

3 Computationally efficient representations

In the previous section covariance functions for some examples of nested SPDE models were derived. Given the covariance function, standard spatial statistics techniques can be used for parameter estimation, spatial prediction and model simulation. Many of these techniques are, however, computationally infeasible for large data sets. Thus, in order to use the model for large environmental data sets, such as the ozone data studied in Section 5, a more computationally efficient representation of the model class is needed. In this section the Hilbert space approximation technique by Lindgren et al. (2011) is used to derive such a representation.

The key idea in Lindgren et al. (2011) is to approximate the solution to the SPDE $\mathcal{L}_1 X_0(\mathbf{s}) = \mathcal{W}(\mathbf{s})$ in some approximation space spanned by basis functions $\varphi_1(\mathbf{s}), \ldots, \varphi_n(\mathbf{s})$. The method is most efficient if these basis functions have compact support, so, from now on, it is assumed that $\{\varphi_i\}$ are local basis functions. The weak solution of the SPDE with respect to the approximation space can be written as $\tilde{x}(\mathbf{s}) = \sum_{i=1}^n w_i \varphi_i(\mathbf{s})$, where the stochastic weights $\{w_i\}_{i=1}^n$ are chosen such that the weak formulation of the SPDE is satisfied:

$$[\langle \varphi_i, \mathcal{L}_1 \tilde{x} \rangle_\Omega]_{i=1,\ldots,n} \stackrel{D}{=} [\langle \varphi_i, \mathcal{W} \rangle_\Omega]_{i=1,\ldots,n}. \tag{12}$$

Here $\stackrel{D}{=}$ denotes equality in distribution, Ω is the manifold on which $\mathbf{s}$ is defined, and $\langle f, g \rangle_\Omega = \int_\Omega f(\mathbf{s}) g(\mathbf{s})\, \mathrm{d}\mathbf{s}$ is the scalar product on Ω. As an illustrative example, consider the first fundamental case $\mathcal{L}_1 = \kappa^2 - \Delta$. One has

$$\langle \varphi_i, \mathcal{L}_1 \tilde{x} \rangle_\Omega = \sum_{j=1}^n w_j \langle \varphi_i, \mathcal{L}_1 \varphi_j \rangle_\Omega,$$

so by introducing a matrix $\mathbf{K}$, with elements $\mathbf{K}_{i,j} = \langle \varphi_i, \mathcal{L}_1 \varphi_j \rangle_\Omega$, and the vector $\mathbf{w} = (w_1, \ldots, w_n)^\top$, the left-hand side of (12) can be written as $\mathbf{Kw}$. Since

$$\begin{aligned} \langle \varphi_i, \mathcal{L}_1 \varphi_j \rangle_\Omega &= \kappa^2 \langle \varphi_i, \varphi_j \rangle_\Omega - \langle \varphi_i, \Delta \varphi_j \rangle_\Omega \\ &= \kappa^2 \langle \varphi_i, \varphi_j \rangle_\Omega + \langle \nabla \varphi_i, \nabla \varphi_j \rangle_\Omega, \end{aligned}$$

the matrix $\mathbf{K}$ can be written as $\mathbf{K} = \kappa^2 \mathbf{C} + \mathbf{G}$, where $\mathbf{C}_{i,j} = \langle \varphi_i, \varphi_j \rangle_\Omega$ and $\mathbf{G}_{i,j} = \langle \nabla \varphi_i, \nabla \varphi_j \rangle_\Omega$. The right-hand side of (12) can be shown to be Gaussian with mean zero and covariance $\mathbf{C}$. For the Hilbert space approximations, it is natural to work with the canonical representation, $\mathbf{x} \sim \mathsf{N}_C(\mathbf{b}, \mathbf{Q})$, of the Gaussian distribution. Here, the precision matrix $\mathbf{Q}$ is the inverse of the covariance matrix, and the vector $\mathbf{b}$ is connected to the mean, $\boldsymbol{\mu}$, of the Gaussian distribution through the relation $\boldsymbol{\mu} = \mathbf{Q}^{-1}\mathbf{b}$. Thus, if $\mathbf{K}$ is invertible, one has

$$\mathbf{Kw} \sim \mathsf{N}_C(\mathbf{0}, \mathbf{C}^{-1}) \quad \Longleftrightarrow \quad \mathbf{w} \sim \mathsf{N}_C(\mathbf{0}, \mathbf{KC}^{-1}\mathbf{K}).$$

For the second fundamental case, $\mathcal{L}_1 = (\kappa^2 - \Delta)^{1/2}$, Lindgren et al. (2011) show that $\mathbf{w} \sim \mathsf{N}_C(\mathbf{0}, \mathbf{K})$. Given these two fundamental cases, the weak solution to $\mathcal{L}_1 X_0(\mathbf{s}) = \mathcal{W}(\mathbf{s})$, for any operator on the form (6), can be obtained recursively. If, for example, $\mathcal{L}_1 = (\kappa^2 - \Delta)^2$, the solution is obtained by solving $(\kappa^2 - \Delta) X_0(\mathbf{s}) = \tilde{x}(\mathbf{s})$, where $\tilde{x}$ is the weak solution to the first fundamental case.

The iterative way of constructing solutions can be extended to calculate weak solutions to (8) as well. Let $\tilde{x}_0 = \sum_{i=1}^{n} w_i^0 \varphi_i(\mathbf{s})$ be a weak solution to $\mathcal{L}_1 X_0(\mathbf{s}) = \mathcal{W}(\mathbf{s})$, and let $\mathbf{Q}_{X_0}$ denote the precision for the weights $\mathbf{w}_0 = (w_1^0, \ldots, w_n^0)^\top$. Substituting X_0 with $\tilde{x}_0$ in the second equation of (3), the weak formulation of the equation is

$$\begin{aligned} [\langle \varphi_i, \tilde{x} \rangle_\Omega]_{i=1,\ldots,n} &\stackrel{D}{=} [\langle \varphi_i, \mathcal{L}_2 \tilde{x}_0 \rangle_\Omega]_{i=1,\ldots,n} \\ &= \left[\sum_{j=1}^{n} w_j^0 \langle \varphi_i, \mathcal{L}_2 \varphi_j \rangle_\Omega \right]_{i=1,\ldots,n}. \end{aligned} \tag{13}$$

First consider the case of an order-one operator $\mathcal{L}_2 = b_1 + \mathbf{B}_1^\top \nabla$. By introducing the matrix $\mathbf{H}_1$ with elements $\mathbf{H}_{1i,j} = \langle \varphi_i, \mathcal{L}_2 \varphi_j \rangle_\Omega$, the right-hand side of (13) can be written as $\mathbf{H}_1 \mathbf{w}_0$. Introducing the vector $\mathbf{w} = (w_1, \ldots, w_n)^\top$, the left-hand side of (13) can be written as $\mathbf{C}\mathbf{w}$, and one has

$$\mathbf{w} = \mathbf{C}^{-1} \mathbf{H}_1 \mathbf{w}_0 \quad \Longrightarrow \quad \mathbf{w} \sim \mathsf{N}_C(\mathbf{0}, \mathbf{C}\mathbf{H}_1^{-\top} \mathbf{Q}_{X_0} \mathbf{H}_1^{-1} \mathbf{C}).$$

Now, if $\mathcal{L}_2$ is on the form (7), the procedure can be used recursively, in the same way as when producing higher order Matérn fields. For example, if

$$\mathcal{L}_2 = (b_1 + \mathbf{B}_1^\top \nabla)(b_2 + \mathbf{B}_2^\top \nabla),$$

the solution is obtained by solving $X(\mathbf{s}) = (b_2 + \mathbf{B}_2^\top \nabla)\tilde{x}(\mathbf{s})$, where $\tilde{x}$ is the weak solution to the previous example. Thus, when $\mathcal{L}_2$ is on the form (7), one has

$$\mathbf{w} \sim \mathsf{N}_C(\mathbf{0}, \mathbf{H}^{-\top} \mathbf{Q}_{X_0} \mathbf{H}^{-1}), \qquad \mathbf{H} = \mathbf{C}^{-1} \mathbf{H}_{n_2} \mathbf{C}^{-1} \mathbf{H}_{n_2 - 1} \cdots \mathbf{C}^{-1} \mathbf{H}_1,$$

where each factor $\mathbf{H}_i$ corresponds to the $\mathbf{H}$-matrix obtained in the ith step in the recursion.

3.1 Nonstationary fields

As mentioned in Lindgren et al. (2011), the Hilbert space approximation technique can also be used for nonstationary models, and the technique extends to the nested SPDE models as well. One again begins by finding the weak solution of the first part of the system, $\mathcal{L}_1(\mathbf{s}) X_0(\mathbf{s}) = \mathcal{W}(\mathbf{s})$. The iterative procedure is used for obtaining approximations of high-order operators, so the fundamental step is to find the weak solution to the equation when $\mathcal{L}_1 = (\kappa^2(\mathbf{s}) - \Delta)$. Consider the weak formulation

$$\left[\langle \varphi_i, \left(\kappa^2(\mathbf{s}) - \Delta\right) \tilde{x} \rangle_\Omega \right]_{i=1,\ldots,n} \stackrel{D}{=} [\langle \varphi_i, \mathcal{W} \rangle_\Omega]_{i=1,\ldots,n}, \tag{14}$$

and note that the right-hand side of the equation is the same as in the stationary case, $\mathsf{N}_C(\mathbf{0}, \mathbf{C}^{-1})$. Now, using that

$$\begin{aligned}\langle \varphi_i, \big(\kappa^2(\mathbf{s}) - \Delta\big)\tilde{x}\rangle_\Omega &= \langle \varphi_i, \kappa^2(\mathbf{s})\tilde{x}\rangle_\Omega - \langle \varphi_i, \Delta\tilde{x}\rangle_\Omega \\ &= \langle \varphi_i, \kappa^2(\mathbf{s})\tilde{x}\rangle_\Omega + \langle \nabla\varphi_i, \nabla\tilde{x}\rangle_\Omega,\end{aligned}$$

the left-hand side of (14) can be written as $(\tilde{\mathbf{C}} + \mathbf{G})\mathbf{w}_0$, where $\mathbf{G}$ and $\mathbf{w}_0$ are the same as in the stationary case and $\tilde{\mathbf{C}}$ is a matrix with elements

$$\begin{aligned}\tilde{\mathbf{C}}_{i,j} &= \langle \varphi_i, \kappa^2(\mathbf{s})\varphi_j\rangle_\Omega = \int_\Omega \kappa^2(\mathbf{s})\varphi_i(\mathbf{s})\varphi_j(\mathbf{s})\,\mathrm{d}\mathbf{s} \\ &\approx \kappa^2(\mathbf{s}_j)\int_\Omega \varphi_i(\mathbf{s})\varphi_j(\mathbf{s})\,\mathrm{d}\mathbf{s} = \kappa^2(\mathbf{s}_j)\mathbf{C}_{i,j}.\end{aligned} \tag{15}$$

Since $\{\varphi_i\}$ is assumed to be a local basis, such as B-spline wavelets or some other functions with compact support, the locations $\mathbf{s}_j$ can, for example, be chosen as the centers of the basis functions $\varphi_j(\mathbf{s})$. The error in the approximation of $\tilde{\mathbf{C}}$ is then small if $\kappa^2(\mathbf{s})$ varies slowly compared to the spacing of the basis functions φ_j. From equation (15), one has $\tilde{\mathbf{C}} = \mathbf{C}\kappa$, where κ is a diagonal matrix with elements $\kappa_{j,j} = \kappa^2(\mathbf{s}_j)$. Finally, with $\mathbf{K} = \kappa\mathbf{C} + \mathbf{G}$, one has

$$\mathbf{K}\mathbf{w}_0 \sim \mathsf{N}_C(\mathbf{0}, \mathbf{C}^{-1}) \quad\Longrightarrow\quad \mathbf{w}_0 \sim \mathsf{N}_C(\mathbf{0}, \mathbf{K}\mathbf{C}^{-1}\mathbf{K}).$$

Now given the weak solution, $\tilde{x}_0$, to $\mathcal{L}_1(\mathbf{s})X_0(\mathbf{s}) = \mathcal{W}(\mathbf{s})$, substitute X_0 with $\tilde{x}_0$ in the second equation of (4) and consider the weak formulation of the equation. Since the solution to the full operator again can be found recursively, only the fundamental case $\mathcal{L}_2 = b(\mathbf{s}) + \mathbf{B}(\mathbf{s})^\top\nabla$ is considered. The weak formulation is the same as (13), and one has

$$\begin{aligned}\langle \varphi_i, \tilde{x}\rangle_\Omega &\overset{D}{=} \langle \varphi_i, \mathcal{L}_2\tilde{x}_0\rangle_\Omega = \langle \varphi_i, \big(b(\mathbf{s}) + \mathbf{B}(\mathbf{s})^\top\nabla\big)\tilde{x}_0\rangle_\Omega \\ &= \langle \varphi_i, b(\mathbf{s})\tilde{x}_0\rangle_\Omega + \langle \varphi_i, \mathbf{B}(\mathbf{s})^\top\nabla\tilde{x}_0\rangle_\Omega.\end{aligned}$$

Thus, the right-hand side of (13) can be written as $(\hat{\mathbf{C}} + \hat{\mathbf{H}})\mathbf{w}_0$, where

$$\begin{aligned}\hat{\mathbf{C}}_{i,j} &= \langle \varphi_i, b(\mathbf{s})\varphi_j\rangle_\Omega = \int_\Omega b(\mathbf{s})\varphi_i(\mathbf{s})\varphi_j(\mathbf{s})\,\mathrm{d}\mathbf{s} \approx b(\mathbf{s}_j)\mathbf{C}_{i,j}, \\ \hat{\mathbf{H}}_{i,j} &= \langle \varphi_i, \mathbf{B}(\mathbf{s})^\top\nabla\varphi_j\rangle_\Omega = \int_\Omega \varphi_i(\mathbf{s})\mathbf{B}(\mathbf{s})^\top\nabla\varphi_j(\mathbf{s})\,\mathrm{d}\mathbf{s} \\ &\approx \mathbf{B}(\tilde{\mathbf{s}}_j)^\top\int_\Omega \varphi_i(\mathbf{s})\nabla\varphi_j(\mathbf{s})\,\mathrm{d}\mathbf{s}.\end{aligned}$$

Here, similar approximations as in equation (15) are used, so the expressions are accurate if the coefficients vary slowly compared to the spacing of the basis functions φ_j. The left-hand side of (13) can again be written as $\mathbf{Cw}$, so with $\mathbf{H}_1 = \hat{\mathbf{C}} + \hat{\mathbf{H}}$, one has $\mathbf{w} \sim \mathsf{N}_C(\mathbf{0}, \mathbf{C}\mathbf{H}_1^{-\top}\mathbf{Q}_{X_0}\mathbf{H}_1^{-1}\mathbf{C})$.

3.2 Practical considerations

The integrals that must be calculated to get explicit expressions for the matrices $\mathbf{C}$, $\mathbf{G}$ and $\mathbf{H}$ are

$$\int_\Omega \varphi_i(\mathbf{s})\varphi_j(\mathbf{s})\,\mathrm{d}\mathbf{s}, \qquad \int_\Omega (\nabla\varphi_i(\mathbf{s}))^\top \nabla\varphi_j(\mathbf{s})\,\mathrm{d}\mathbf{s} \quad \text{and} \quad \int_\Omega \varphi_i(\mathbf{s})\nabla\varphi_j(\mathbf{s})\,\mathrm{d}\mathbf{s}.$$

In Section 5 a basis of piecewise linear functions induced by a triangulation of the Earth is used; see Figure 4. In this case, $\varphi_i(\mathbf{s})$ is a linear function on each triangle, and $\nabla\varphi_i(\mathbf{s})$ is constant on each triangle. The integrals, therefore, have simple analytic expressions in this case, and more generally for all piecewise linear bases induced by triangulated 2-manifolds.

Bases induced by triangulations have many desirable properties, such as the simple analytic expression for the integrals and compact support. They are, however, not orthogonal, which causes $\mathbf{C}^{-1}$ to be dense. The weights $\mathbf{w}$, therefore, have a dense precision matrix, unless $\mathbf{C}^{-1}$ is approximated with some sparse matrix. This issue is addressed in Lindgren et al. (2011) by lowering the integration order of $\langle\varphi_i, \varphi_j\rangle$, which results in an approximate, diagonal $\mathbf{C}$ matrix, $\bar{\mathbf{C}}$, with diagonal elements $\bar{\mathbf{C}}_{ii} = \sum_{k=1}^n \mathbf{C}_{ik}$. Bolin and Lindgren (2009) perform numerical studies on how this approximation affects the resulting covariance function of the process, and it is shown that the error is small if the approximation is used for piecewise linear bases. We will, therefore, from now on use the approximate $\mathbf{C}$ matrix in all places where $\mathbf{C}$ is used.

A natural question is how many basis functions one should use in order to get a good approximation of the solution. The answer will depend on the chosen basis, and, more importantly, on the specific parameters of the SPDE model. Bolin and Lindgren (2009) study the approximation error in the Matérn case in $\mathbb{R}$ and $\mathbb{R}^2$ for different bases, and in this case the spacing of the basis functions compared to the range of the covariance function for $X(\mathbf{s})$ determines the approximation error: For a process with long range, fewer basis functions have to be used than for a process with short range to obtain the same approximation error. For more complicated, possibly nonstationary, nested SPDE models, there is no

easy answer to how the number of basis functions should be chosen. Increasing the number of basis functions will decrease the approximation error but increase the computational complexity for the approximate model, so there is a trade-off between accuracy and computational cost. However, as long as the parameters vary slowly compared to the spacing of the basis functions, the approximation error will likely be much smaller than the error obtained from using a model that does not fit the data perfectly and from estimating the parameters from the data. Thus, for practical applications, the error in covariance induced by the Hilbert space approximation technique will likely not matter much. A more important consequence for practical applications when the piecewise linear basis is used is that the Kriging estimation of the field between two nodes in the triangulation is a linear interpolation of the values at the nodes. Thus, variations on a scale smaller than the spacing between the basis functions will not be captured correctly in the Kriging prediction. For practical applications, it is therefore often best to choose the number of basis functions depending on the scale one is interested in the Kriging prediction on.

For the ozone data in Section 5, the goal is to estimate daily maps of global ozone. As we are not interested in modeling small scale variations, we choose the number of basis functions so that the mean distance between basis functions is about 258 km. For this basis, the smallest distance between two basis functions is 222 km, and the largest distance is about 342 km.

Estimating the model parameters using different numbers of basis functions will give different estimates, as the parameters are estimated to maximize the likelihood for the approximate model instead of the exact SPDE. An example of this can be seen in Figure 3 where the estimates of the covariance parameters for model F' (see Section 5 for a model description) for the ozone data are shown for varying numbers of basis functions. Instead of showing the actual parameter estimates, the figure shows the differences between the estimates and the estimate when using the basis shown in Figure 4, which has 9002 basis functions. Increasing the number of basis functions further, the estimates will finally converge to the estimates one would get using the exact SPDE representation. The curve that has not converged corresponds to the dominating parameter in the vector field. Together with κ, this parameter controls the correlation range of the ozone field.

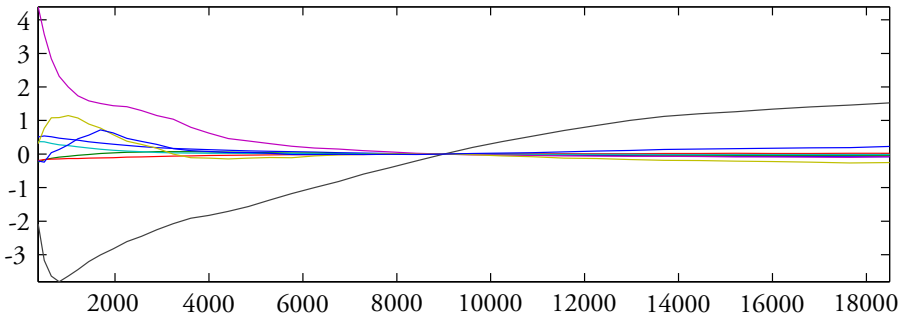

Figure 3: Parameter estimates for the covariance parameters in model F′ for the ozone data as functions of the number of basis functions in the Hilbert space approximations.

4 Parameter estimation

In this section a parameter estimation procedure for the nested SPDE models is presented. One alternative would be to use a Metropolis–Hastings algorithm, which is easy to implement, but computationally inefficient. A better alternative is to use direct numerical optimization to estimate the parameters.

Let $Y(\mathbf{s})$ be an observation of the latent field, $X(\mathbf{s})$, given by (8) or (10), under mean zero Gaussian measurement noise, $\mathcal{E}(\mathbf{s})$, with variance σ^2:

$$Y(\mathbf{s}) = X(\mathbf{s}) + \mathcal{E}(\mathbf{s}). \tag{16}$$

Using the approximation procedure from Section 3, and assuming a regression model for the latent field's mean value function, $\mu(\mathbf{s})$, the measurement equation can then be written as

$$\mathbf{Y} = \mathbf{M}\boldsymbol{\mu} + \boldsymbol{\Phi}\mathbf{w} + \boldsymbol{\epsilon},$$

where $\mathbf{M}$ is a matrix with the regression basis functions evaluated at the measurement locations, and $\boldsymbol{\mu}$ is a vector containing the regression coefficients that have to be estimated. The matrix $\boldsymbol{\Phi}$ contains the basis functions for the Hilbert space approximation procedure evaluated at the measurement locations, and $\mathbf{w}$ is the vector with the stochastic weights. In Section 3 it was shown that the vector $\mathbf{w}$ is Gaussian with mean zero and covariance matrix $\mathbf{H}\mathbf{Q}_{X_0}^{-1}\mathbf{H}^\top$. Both $\mathbf{Q}_{X_0}$ and $\mathbf{H}$ are sparse matrices, but neither the covariance matrix nor the precision matrix for

$\mathbf{w}$ is sparse. Thus, it would seem as if one had to work with a dense covariance matrix, which would make maximum likelihood parameter estimation computationally infeasible for large data sets. However, because of the product form of the covariance matrix, one has that $\mathbf{w} = \mathbf{H}\mathbf{w}_0$, where $\mathbf{w}_0 \sim \mathsf{N}_C(\mathbf{0}, \mathbf{Q}_{X_0})$. Hence, the observation equation can be rewritten as

$$\mathbf{Y} = \mathbf{M}\boldsymbol{\mu} + \boldsymbol{\Phi}\mathbf{H}\mathbf{w}_0 + \boldsymbol{\epsilon}. \tag{17}$$

Interpreting $\boldsymbol{\Lambda} = \boldsymbol{\Phi}\mathbf{H}$ as an observation matrix that depends on some of the parameters in the model, $\mathbf{Y} - \mathbf{M}\boldsymbol{\mu}$ can now be seen as noisy observations of $\mathbf{w}_0$, which has a sparse precision matrix. The advantage with using (17) is that one then is in the setting of having observations of a latent Gaussian Markov random field, which facilitates the usage of sparse matrix techniques in the parameter estimation.

Let $\boldsymbol{\psi}$ denote all parameters in the model except for $\boldsymbol{\mu}$. Assuming that $\boldsymbol{\mu}$ and $\boldsymbol{\psi}$ are a priori independent, the posterior density can be written as

$$\pi(\mathbf{w}_0, \boldsymbol{\mu}, \boldsymbol{\psi}|\mathbf{Y}) \propto \pi(\mathbf{Y}|\mathbf{w}_0, \sigma^2)\pi(\mathbf{w}_0|\boldsymbol{\mu}, \boldsymbol{\psi})\pi(\boldsymbol{\mu})\pi(\boldsymbol{\psi}).$$

Using a Gaussian prior distribution with mean $\boldsymbol{\mu}$ and precision $\mathbf{Q}_\mu$ for the mean parameters, the posterior distribution can be reformulated as

$$\pi(\mathbf{w}_0, \boldsymbol{\mu}, \boldsymbol{\psi}|\mathbf{Y}) \propto \pi(\mathbf{w}_0|\boldsymbol{\mu}, \boldsymbol{\psi}, \mathbf{Y})\pi(\boldsymbol{\mu}|\boldsymbol{\psi}, \mathbf{Y})\pi(\boldsymbol{\psi}|\mathbf{Y}), \tag{18}$$

where $\mathbf{w}_0|\boldsymbol{\mu}, \boldsymbol{\psi}, \mathbf{Y} \sim \mathsf{N}_C(\mathbf{b}, \hat{\mathbf{Q}})$, $\boldsymbol{\mu}|\boldsymbol{\psi}, \mathbf{Y} \sim \mathsf{N}_C(\mathbf{b}_\mu, \hat{\mathbf{Q}}_\mu)$, and

$$\mathbf{b} = \frac{1}{\sigma^2}\boldsymbol{\Lambda}^\top(\mathbf{Y} - \mathbf{M}\boldsymbol{\mu}), \qquad \mathbf{b}_\mu = \mathbf{Q}_\mu\mathbf{m}_\mu + \frac{\mathbf{M}^\top\mathbf{Y}}{\sigma^2} - \frac{\mathbf{M}^\top\boldsymbol{\Lambda}\hat{\mathbf{Q}}^{-1}\boldsymbol{\Lambda}^\top\mathbf{Y}}{\sigma^4},$$
$$\hat{\mathbf{Q}} = \mathbf{Q}_{w_0} + \frac{1}{\sigma^2}\boldsymbol{\Lambda}^\top\boldsymbol{\Lambda}, \qquad \hat{\mathbf{Q}}_\mu = \mathbf{Q}_\mu + \frac{\mathbf{M}^\top\mathbf{M}}{\sigma^2} - \frac{\mathbf{M}^\top\boldsymbol{\Lambda}\hat{\mathbf{Q}}^{-1}\boldsymbol{\Lambda}^\top\mathbf{M}}{\sigma^4}.$$

The calculations are omitted here since these expressions are calculated similarly to the posterior reformulation in Lindström and Lindgren (2008), which gives more computational details. Finally, the marginal posterior density $\pi(\boldsymbol{\psi}|\mathbf{Y})$ can be shown to be

$$\pi(\boldsymbol{\psi}|\mathbf{Y}) \propto \frac{|\mathbf{Q}_{w_0}|^{1/2}\pi(\boldsymbol{\psi})}{|\hat{\mathbf{Q}}|^{1/2}|\hat{\mathbf{Q}}_\mu|^{1/2}|\sigma\mathbf{I}|}\exp\left(\frac{1}{2\sigma^2}\mathbf{Y}^\top\left(\frac{\boldsymbol{\Lambda}\hat{\mathbf{Q}}^{-1}\boldsymbol{\Lambda}^\top}{\sigma^2} - \mathbf{I}\right)\mathbf{Y} + \frac{\mathbf{b}_\mu^\top\hat{\mathbf{Q}}_\mu^{-1}\mathbf{b}_\mu}{2}\right).$$

By rewriting the posterior as (18), it can be integrated with respect to $\mathbf{w}_0$ and $\boldsymbol{\mu}$, and instead of optimizing the full posterior with respect to $\mathbf{w}_0$, $\boldsymbol{\mu}$ and $\boldsymbol{\psi}$, only the marginal posterior $\pi(\boldsymbol{\psi}|\mathbf{Y})$ has to be optimized with respect to $\boldsymbol{\psi}$. This is a lower dimensional optimization problem, which substantially decreases the computational complexity. Given the optimum, $\boldsymbol{\psi}_{\text{opt}} = \text{argmax}_{\boldsymbol{\psi}}\, \pi(\boldsymbol{\psi}|\mathbf{Y})$, $\boldsymbol{\mu}_{\text{opt}}$ is then given by $\boldsymbol{\mu}_{\text{opt}} = \hat{\mathbf{Q}}_{\mu}^{-1}\mathbf{b}_{\mu}$. In practice, the numerical optimization is carried out on $\log \pi(\boldsymbol{\psi}|\mathbf{Y})$.

4.1 Estimating the parameter uncertainty

There are several ways one could estimate the uncertainty in the parameter estimates obtained by the parameter estimation procedure above. The simplest estimate of the uncertainty is obtained by numerically estimating the Hessian of the marginal posterior evaluated at the estimated parameters. The diagonal elements of the inverse of the Hessian can then be seen as estimates of the variance for the parameter estimates.

Another method for obtaining more reliable uncertainty estimates is to use a Metropolis–Hastings based MCMC algorithm with proposal kernel similar to the one used in Lindström and Lindgren (2008). A quite efficient algorithm is obtained by using random walk proposals for the parameters, where the correlation matrix for the proposal distribution is taken as a rescaled version of the inverse of the Hessian matrix (Gelman et al., 1996).

Finally, a third method for estimating the uncertainties is to use the INLA framework (Rue et al., 2009), available as an R package[2]. In settings with latent Gaussian Markov random fields, integrated nested Laplace approximations (INLA) provide close approximations to posterior densities for a fraction of the cost of MCMC. For models with Gaussian data, the calculated densities are for practical purposes exact. In the current implementation of the INLA package, handling the full nested SPDE structure is cumbersome, so further enhancements are needed before one can take full advantage of the INLA method for these models.

4.2 Computational complexity

In this section some details on the computational complexity for the parameter estimation and Kriging estimation are given.

[2] `http://www.r-inla.org/`

The most widely used method for spatial prediction is linear Kriging. In the Bayesian setting, the Kriging predictor simply is the posterior expectation of the latent field X given data and the estimated parameters. This expectation can be written as

$$\mathsf{E}(X|\psi, \mu, \mathbf{Y}) = \mathbf{M}\mu + \mathbf{\Phi}\mathbf{H}\mathsf{E}(\mathbf{w}_0) = \mathbf{M}\mu + \mathbf{\Phi}\mathbf{H}\hat{\mathbf{Q}}^{-1}\mathbf{b}.$$

The computationally demanding part of this expression is to calculate $\hat{\mathbf{Q}}^{-1}\mathbf{b}$. Since the $n \times n$ matrix $\mathbf{Q}$ is positive definite, this is most efficiently done using Cholesky factorization, forward substitution and back substitution: Calculate the Cholesky triangle $\mathbf{L}$ such that $\hat{\mathbf{Q}} = \mathbf{L}\mathbf{L}^\top$, and given $\mathbf{L}$, solve the linear system $\mathbf{L}\mathbf{x} = \mathbf{b}$. Finally, given $\mathbf{x}$, solve $\mathbf{L}^\top\mathbf{y} = \mathbf{x}$, where now $\mathbf{y}$ satisfies $\mathbf{y} = \hat{\mathbf{Q}}^{-1}\mathbf{b}$. Solving the forward substitution and back substitution are much less computationally demanding than calculating the Cholesky triangle. Hence, the computational cost for calculating the Kriging prediction is determined by the cost for calculating $\mathbf{L}$.

The computational complexity for the parameter estimation is determined by the optimization method that is used and the computational complexity for evaluating the marginal log-posterior $\log \pi(\psi|\mathbf{Y})$. The most computationally demanding terms in $\log \pi(\psi|\mathbf{Y})$ are the two log-determinants $\log|\mathbf{Q}_{w_0}|$ and $\log|\hat{\mathbf{Q}}|$ and the quadratic form $\mathbf{Y}^\top \mathbf{\Lambda}\hat{\mathbf{Q}}^{-1}\mathbf{\Lambda}\mathbf{Y}$, which are also most efficiently calculated using Cholesky factorization. Given the Cholesky triangle $\mathbf{L}$, the quadratic form can be obtained as $\mathbf{x}^\top\mathbf{x}$, where $\mathbf{x}$ is the solution to $\mathbf{L}\mathbf{x} = \mathbf{\Lambda}\mathbf{Y}$, and the log-determinant $\log|\hat{\mathbf{Q}}|$ is simply the sum[3] $2\sum_{i=1}^n \log \mathbf{L}_{ii}$. Thus, the computational cost for one evaluation of the marginal posterior is also determined by the cost for calculating $\mathbf{L}$. Because of the sparsity structure of $\hat{\mathbf{Q}}$, this computational cost is $\mathcal{O}(n)$, $\mathcal{O}(n^{3/2})$ and $\mathcal{O}(n^2)$ for problems in one, two and three dimensions respectively (see Rue and Held, 2005, for more details).

The computational complexity for the parameter estimation is highly dependent on the optimization method. If a Broyden–Fletcher–Goldfarb-Shanno (BFGS) procedure is used without an analytic expression for the gradients, the marginal posterior has to be evaluated p times for each step in the optimization, where p is the number of covariance parameters in the model. Thus, if p is large and the initial value for the optimization is chosen far from the optimal value,

[3] Since only the difference between the log-determinants is needed, one should implement the calculation as $2\sum_{i=1}^n(\log L_{(i)}^{w_0} - \log \hat{L}_{(i)})$, where $L_{(i)}^{w_0}$ and $\hat{L}_{(i)}$ are the diagonal elements of the Cholesky factors, sorted in ascending order, and the sum is ordered by increasing absolute values of the differences. This reduces numerical issues.

many thousand evaluations of the marginal posterior may be needed in the optimization.

5 Application: Ozone data

On October 24, 1978, NASA launched the near-polar, Sun-synchronous orbiting satellite Nimbus-7. The satellite carried a TOMS instrument with the purpose of obtaining high-resolution global maps of atmospheric ozone (McPeters et al., 1996). The instrument measured backscattered solar ultraviolet radiation at 35 sample points along a line perpendicular to the orbital plane at 3-degree intervals from 51 degrees on the right side of spacecraft to 51 degrees on the left. A new scan was started every eight seconds, and as the measurements required sunlight, the measurements were made during the sunlit portions of the orbit as the spacecraft moved from south to north. The data measured by the satellite has been calibrated and preprocessed into a "Level 2" data set of spatially and temporally irregular Total Column Ozone (TCO) measurements following the satellite orbit. There is also a daily "Level 3" data set with values processed into a regular latitude-longitude grid. Both Level 2 and Level 3 data have been analyzed in recent papers in the statistical literature (Cressie and Johannesson, 2008, Jun and Stein, 2008, Stein, 2007).

In what follows, the nested SPDE models are used to obtain statistical estimates of a daily ozone map using a part of the Level 2 data. In particular, all data available for October 1st, 1988 is used, which is the same data set that was used by Cressie and Johannesson (2008).

5.1 Statistical model

The measurement model (16) is used for the ozone data. That is, the measurements, $Y(\mathbf{s})$, are assumed to be observations of a latent field of TCO ozone, $X(\mathbf{s})$, under Gaussian measurement noise $\mathcal{E}(\mathbf{s})$ with a constant variance σ^2. We let $X(\mathbf{s})$ have some mean value function, $\mu(\mathbf{s})$, and let the covariance structure be determined by a nested SPDE model. Inspired by Jun and Stein (2008), who proposed using differentiated Matérn fields for modeling TCO ozone, we use the simplest nested SPDE model. Thus, $Z(\mathbf{s}) = X(\mathbf{s}) - \mu(\mathbf{s})$ is generated by the system

$$\begin{aligned} \big(\kappa^2(\mathbf{s}) - \Delta\big) Z_0(\mathbf{s}) &= \mathcal{W}(\mathbf{s}) \\ Z(\mathbf{s}) &= \big(b(\mathbf{s}) + \mathbf{B}(\mathbf{s})^\top \nabla\big) Z_0(\mathbf{s}), \end{aligned}$$

	A	B	C	D	E	F	G	H	I	J	K	L	M
$\kappa^2(\mathbf{s})$	0	1	0	1	2	0	3	2	0	4	3	0	4
$b(\mathbf{s})$	0	1	1	1	2	2	3	2	3	4	3	4	4
$\mathbf{B}(\mathbf{s})$	0	0	1	1	0	2	0	2	3	0	3	4	4
Total	2	8	11	14	18	26	32	34	47	50	62	75	98

Table 1: Maximal orders of the spherical harmonics used in the bases for the different parameters and total number of covariance parameters in the different models for $X(\mathbf{s})$. The actual number of basis functions for $\kappa^2(\mathbf{s})$ and $b(\mathbf{s})$ are given by $(ord + 1)^2$, and for $\mathbf{B}(\mathbf{s})$, the actual number is $2(ord + 1)^2 - 2$, where ord is the maximal order indicated in the table.

where $\mathcal{W}(\mathbf{s})$ is Gaussian white noise on the sphere. If $\kappa(\mathbf{s})$ is assumed to be constant, the ozone is modeled as a Gaussian field with a covariance structure that is obtained by applying the differential operator $(b(\mathbf{s}) + \mathbf{B}(\mathbf{s})^\top \nabla)$ to a stationary Matérn field, which is similar to the model by Jun and Stein (2008). If, on the other hand, κ is spatially varying, the range of the Matérn-like covariance function can vary with location. As in Stein (2007) and Jun and Stein (2008), the mean can be modeled using a regression basis of spherical harmonics; however, since the data set only contains measurements from one specific day, it is not possible to identify which part of the variation in the data that comes from a varying mean and which part that can be explained by the variance–covariance structure of the latent field. To avoid this identifiability problem, $\mu(\mathbf{s})$ is assumed to be unknown but constant. The parameter $\kappa^2(\mathbf{s})$ has to be positive, and for identifiability reasons, we also require $b(\mathbf{s})$ to be positive. We, therefore, let $\log \kappa^2(\mathbf{s}) = \sum_{k,m} \kappa_{k,m} Y_{k,m}(\mathbf{s})$ and $\log b(\mathbf{s}) = \sum_{k,m} b_{k,m} Y_{k,m}(\mathbf{s})$, where $Y_{k,m}$ is the spherical harmonic of order k and mode m. Finally, the vector field $\mathbf{B}(\mathbf{s})$ is modeled using the vector spherical harmonics basis functions $\Upsilon^1_{k,m}$ and $\Upsilon^2_{k,m}$, presented in Appendix A:

$$\mathbf{B}(\mathbf{s}) = \sum_{k,m} \left(B^1_{k,m} \Upsilon^1_{k,m}(\mathbf{s}) + B^2_{k,m} \Upsilon^2_{k,m}(\mathbf{s}) \right).$$

To choose the number of basis functions for the parameters $\kappa^2(\mathbf{s})$, $b(\mathbf{s})$ and $\mathbf{B}(\mathbf{s})$, some model selection technique has to be used. Model selection for this model class is difficult since the models can have both nonstationary mean value functions and nonstationary covariance structures. This makes standard variogram techniques inadequate in general, and we instead base the model selection

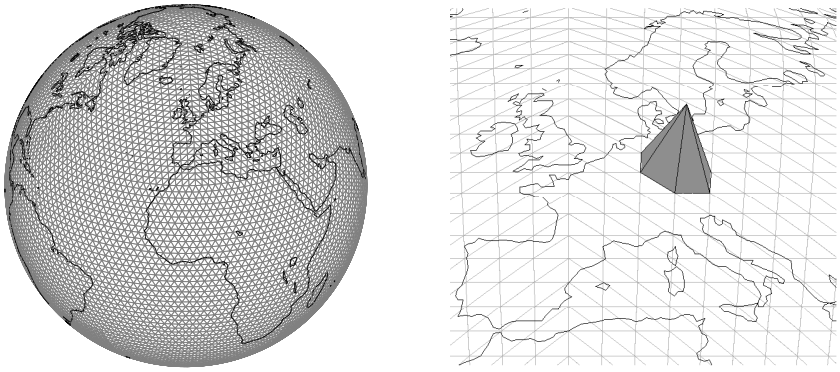

Figure 4: The left part shows the triangulation of the Earth used to define the piecewise linear basis functions in the Hilbert space approximation for ozone data. Each basis function is one at a node in the triangulation, and decreases linearly to zero at the neighboring nodes. The right part of the figure shows one of these functions.

on Akaike's Information Criterion (AIC) and the Bayesian Information Criterion (BIC) (Hastie et al., 2003), which are suitable model selection tools for the nested SPDE models since the likelihood for the data can be evaluated efficiently.

We estimate 13 models with different numbers of covariance parameters, presented in Table 1. The simplest model is a stationary Matérn model, with four parameters to estimate, and the most complicated model has 100 parameters to estimate, including the mean and the measurement noise variance. There are three different types of models in Table 1: In the first type (models B, E, G and J), κ^2 and b are spatially varying and the vector field $\mathbf{B}$ is assumed to be zero. In the second type (models C, F, I and L), b and $\mathbf{B}$ are spatially varying and κ^2 is assumed to be constant. Finally, in the third type (model D, H, K and M), all parameters are spatially varying.

A basis of 9002 piecewise linear functions induced by a triangulation of the Earth (see Figure 4) is used in the approximation procedure from Section 3 to get efficient representations of each model, and the parameters are estimated using the procedure from Section 4. The computational cost for the parameter estimation only depends on the number of basis functions in the Hilbert space approximation, and not on the number of data points, which makes inference efficient even

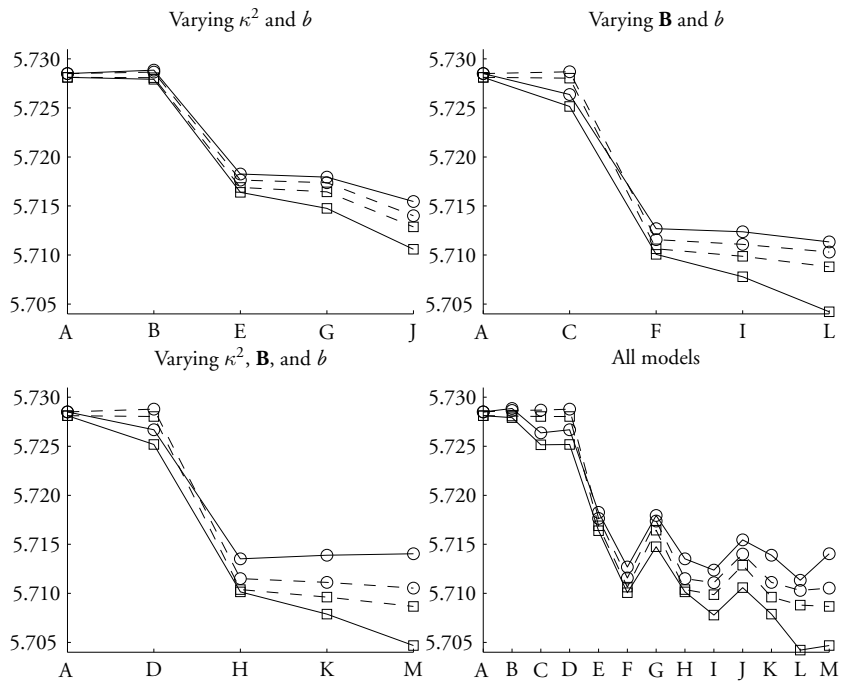

Figure 5: AIC (squares) and BIC (circles) for the models A–M (solid lines) and the axially symmetric models A′–M′ (dashed lines), scaled by a factor 10^{-5}. Note that the major improvement in AIC and BIC occurs when the orders of the basis functions are increased from one to two, and that the model type with spatially varying b and **B** seems to be most appropriate for this data. Also note that the axially symmetric model F′ is surprisingly good considering that it only has 8 covariance parameters.

for this large data set.

AIC and BIC for each of the fitted models can be seen in Figure 5. The figure contains one panel for each of the three model types and one panel where AIC and BIC are shown for all models at once. The major improvement in AIC and BIC occurs when the orders of the basis functions are increased from one to two. For the first model type, with spatially varying κ^2 and b, the figure indicates that the results could be improved by increasing the orders of the basis functions further. However, for a given order of the basis functions, the other two model

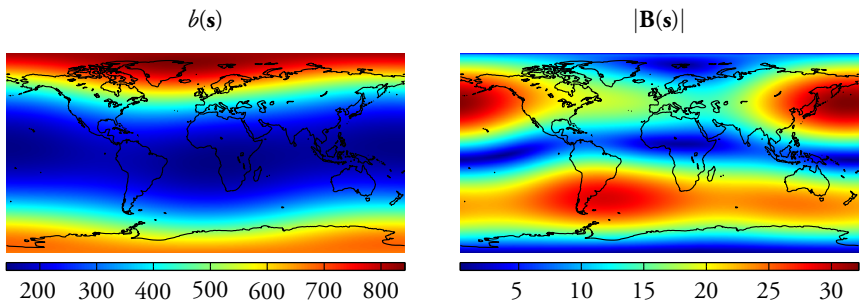

Figure 6: Estimated variance-scaling parameter, $b(\mathbf{s})$, and the norm of the vectors in the estimated vector field $\mathbf{B}(\mathbf{s})$ for model F. Note that the estimates are fairly constant with respect to longitude, which indicates that the latent field could be axially symmetric.

types have much lower AIC and BIC. Also, by comparing AIC and BIC for the second and third model types, one finds that there is not much gain in letting κ^2 be spatially varying. We therefore conclude that a model with spatially varying b and $\mathbf{B}$ is most appropriate for this data.

The estimated parameters $b(\mathbf{s})$ and the length of the vectors $\mathbf{B}(\mathbf{s})$ for model F are shown in Figure 6. One thing to note in this figure is that the two parameters are fairly constant with respect to longitude, which indicates that the latent field could be axially symmetric, an assumption that was made by both Stein (2007) and Jun and Stein (2008). If the latent field indeed was axially symmetric, one would only need the basis functions that are constant with respect to longitude in the parameter bases. Since there is only one axially symmetric spherical harmonic for each order, this assumption drastically reduces the number of parameters for the models in Table 1. Let A′–M′ denote the axially symmetric versions of models A–M. For these models, the number of basis functions for both $\kappa^2(\mathbf{s})$ and $b(\mathbf{s})$ is $ord + 1$, and the number of basis functions for $\mathbf{B}(\mathbf{s})$ is $2(ord + 1) - 2$, where ord is the maximal order indicated in Table 1. The dashed lines in Figure 5 show AIC and BIC calculated for these models. Among the axially symmetric models, model F′ is surprisingly good considering that it only has 8 covariance parameters.

The Kriging estimate and its standard error for model F′ are shown in Figures 7 and 8 respectively. The oscillating behavior near the equator for the standard error is explained by the fact that the satellite tracks are furthest apart there, which results in sparser measurements between the different tracks. Because the

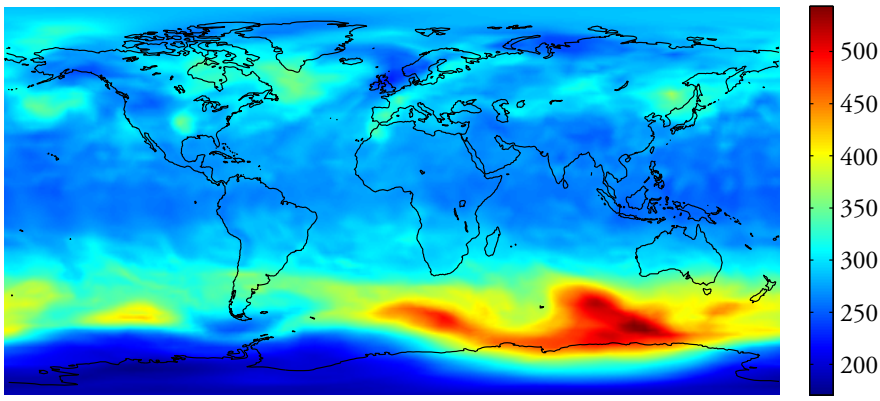

Figure 7: Kriging estimate of TCO ozone in Dobson units using model F′.

measurements are collected using backscattered sunlight, the variance close to the north pole is high, as there are no measurements there. As seen in Figure 9, there is not much spatial correlation in the residuals $\hat{\mathbf{X}} - \mathbf{Y}$, which indicates a good model fit. In Figure 10, estimates of the local mean and variance of the residuals are shown. The mean is fairly constant across the globe, but there is a slight tendency for higher variance closer to the poles. This is due to the fact that the data really is space–time data, as the measurements are collected during a 24-hour period. Since the different satellite tracks are closest near the poles, the temporal variation of the data is most prominent here, and especially near the international date line where data is collected both at the first satellite track of the day and at the last track, 24 hours later. The area with high residual variance is one of those places where measurements are taken both at the beginning and the end of the time period, and where the ozone concentration has changed during the time period between the measurements. One could include this effect by allowing the variance of the measurement noise to be spatially varying; however, one should really use a spatio-temporal model for the data to correctly account for the effect, which is outside the scope of this article.

To see how much the temporal structure near the international date line influences the model fit, the parameters in model F′ are re-estimated without using the first satellite track of the day and without using the last track of the day. The estimated parameters can be seen in Table 2 and, as expected, the estimate of the

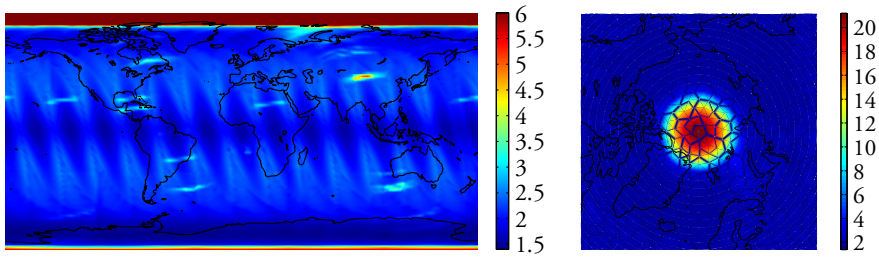

Figure 8: Standard error in Dobson units for the Kriging estimate. The color bar in the left part of the figure has been truncated at 6 Dobson units. The behavior near the north pole can be seen in the right part of the figure.

	κ	σ	b_1	b_2	b_3	B_1	B_2	B_3	B_4
Y_f	0.74	25.60	5.85	0.045	0.34	1.05	2.59	−6.84	−0.84
Y_l	0.73	25.56	5.82	0.033	0.34	0.90	2.38	−7.01	−0.82
Y	0.67	34.09	5.75	0.054	0.36	0.70	2.48	−7.10	−0.68

Table 2: Estimates of the covariance parameters in model F′ using all data but the first track (Y_f), all data but the last track (Y_l), and all data (Y)

measurement noise variance is much lower when not using all date line data. The estimates of the covariance parameters for the latent field also change somewhat, but the large scale structure of the nonstationarity is preserved.

To study how sensitive the Kriging estimates are to the model choice, the ratio between the Kriging estimates for the simple model F′ and the large model M, and the ratio between the corresponding Kriging standard errors, are shown in Figure 11. There is not much difference between the two Kriging estimates, whereas there is a clear difference between the corresponding standard errors. Thus, if one only is interested in the Kriging estimate, it does not matter much which model is used, but if one also is interested in the standard error of the estimate, the model choice greatly influences the results.

5.2 Discussion

Before the nested SPDE models were used on the ozone data, several tests were performed on simulated data to verify that the model parameters in fact could be

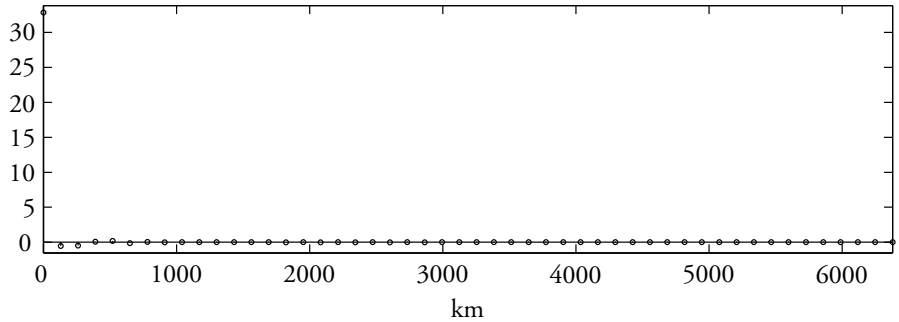

Figure 9: Estimated covariance function for the Kriging residuals using model F′.

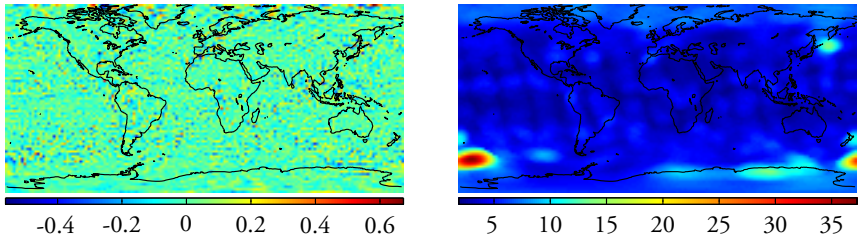

Figure 10: Estimates of the local mean (left) and standard deviation (right) for the Kriging residuals using model F′. The mean is fairly constant across the globe, whereas the standard deviation is higher close to the poles and at the international date line because of the temporal structure in the data.

estimated using the estimation procedure in Section 4. These tests showed that the estimation procedure is robust given that the initial values for the parameters are not chosen too far from the true values. However, for nonstationary models with many covariance parameters, it is not easy to choose the initial values. To reduce this problem, the optimization is done in several steps. A stationary Matérn model (model A) is estimated to get initial values for $\kappa_{0,0}$, $b_{0,0}$ and σ^2. To estimate model B, all parameters are set to zero initially, except for the parameters that were estimated in model A. Another layer of spherical harmonics is added to the bases for $\kappa^2(\mathbf{s})$ and $b(\mathbf{s})$ for estimating model E using the model B parameters as initial values. This step-wise procedure of adding layers of spherical harmonics to the bases is then repeated to estimate the larger models. Numerical studies showed

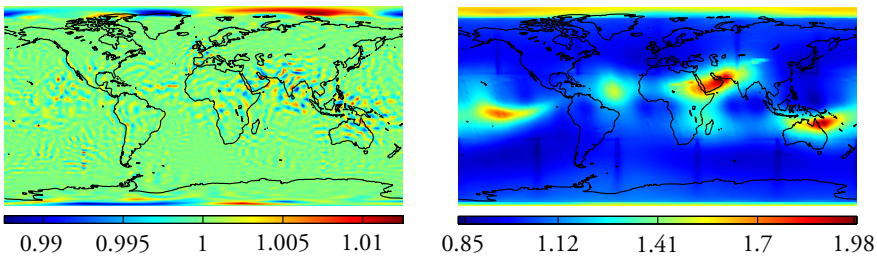

Figure 11: The ratio between the kriging estimates using model F′ and model M (left), and the ratio between the corresponding kriging standard errors (right). Note that there is not much difference between the Kriging estimates, whereas there is a clear difference between the corresponding standard errors.

that this optimization procedure is quite robust even for large models; however, as in most other numerical optimization problems, there are no guarantees that the true optimal values have been found for all models for the ozone data.

The application of the nested SPDE models to ozone data was inspired by Jun and Stein (2008), who proposed using differentiated Matérn fields for modeling TCO ozone, and we conclude this section with some remarks on the similarities and differences between the nested SPDEs and their models. The most general model in Jun and Stein (2008) is on the form

$$\begin{aligned} X(\mathbf{s}) &= P_1(l_2)X_0(\mathbf{s}) + \left(P_2(l_2)\frac{\partial}{\partial l_2} + P_3(l_2)\frac{\partial}{\partial l_1}\right)X_1(\mathbf{s}) \\ &\quad + P_4(l_2)\frac{\partial}{\partial l_1}X_2(\mathbf{s}), \end{aligned} \tag{19}$$

where $X_i, i = 0, 1, 2$, are i.i.d. Matérn fields in $\mathbb{R}^3$, $P_i, i = 1, 2, 3, 4$, are non-random functions depending on latitude, l_1 denoted longitude and l_2 denoted latitude. This model is similar to the model used here, but there are some important differences. First of all, (19) contains a sum of three independent fields, which we cannot represent since the approximation procedure in Section 3 in this case loses its computational benefits. To get a model more similar to the nested SPDE model, one would have to let $P_4(l_2) \equiv 0$, and $X_0(\mathbf{s}) = X_1(\mathbf{s})$. Using $X_0 = X_1$ or X_0 and X_1 as i.i.d. copies of a Matérn field gives different covariance functions, and without testing both cases it is hard to determine what is more appropriate for ozone data.

Another important conceptual difference is how the methods deal with the spherical topology. The Matérn fields in Jun and Stein (2008) are stochastic fields on $\mathbb{R}^3$, evaluated on the embedded sphere, which is equivalent to using chordal distance as the metric in a regular Matérn covariance function. One might instead attempt to evaluate the covariance function using the arc-length distance, which is a more natural metric on the sphere. However, Theorem 2 from Gneiting (1998) shows that for Matérn covariances with $\nu \geq 1$, this procedure does not generate positive definite covariance functions. This means that the arc-length method cannot be used for any differentiable Matérn fields. On the other hand, the nested SPDEs are directly defined on the sphere, and therefore inherently use the arc-length distance.

There is, in theory, no difference between writing the directional derivative of $X(\mathbf{s})$ as $(P_2(l_2)\frac{\partial}{\partial l_2} + P_3(l_2)\frac{\partial}{\partial l_1})X_1(\mathbf{s})$ or $\mathbf{B}(\mathbf{s})^\top \nabla X(\mathbf{s})$, but the latter is easier to work with in practice. If a vector field basis is used to model $\mathbf{B}(s)$, the process will not have any singularities as long as the basis functions are nonsingular, which is the case for the basis used in this paper. If, on the other hand, $P_2(l_2)$ and $P_3(l_2)$ are modeled separately, the process will be singular at the poles unless certain restrictions on the two functions are met. This fact is indeed noted by Jun and Stein (2008), but the authors do not seem to take the restrictions into account in the parameter estimation, which causes all their estimated models to have singularities at the poles.

Finally, the nested SPDE models are computationally efficient also for spatially irregular data, which allowed us to work with the TOMS Level 2 data instead of the gridded Level 3 data.

6 Concluding remarks

There is a need for computationally efficient stochastic models for environmental data. Lindgren et al. (2011) introduced an efficient procedure for obtaining Markov approximations of, possibly nonstationary, Matérn fields by considering Hilbert space approximations of the SPDE

$$\left(\kappa(\mathbf{s})^2 - \Delta\right)^{\alpha/2} X(\mathbf{s}) = \phi(\mathbf{s})\mathcal{W}(\mathbf{s}).$$

In this work, the class of nonstationary nested SPDE models generated by (10) was introduced, and it was shown how the approximation methods in Lindgren

et al. (2011) can be extended to this larger class of models. This model class contains a wider family of covariance models, including both Matérn-like covariance functions and various oscillating covariance functions. Because of the additional differential operator $\mathcal{L}_2$, the Hilbert space approximations for the nested SPDE models do not have the Markov structure the model in Lindgren et al. (2011) has, but all computational benefits from the Markov properties are preserved for the nested SPDE models using the procedure in Section 4. This allows us to fit complicated models with over 100 parameters to data sets with several hundred thousand measurements using only a standard personal computer.

By choosing $\mathcal{L}_2 = b + \mathbf{B}^\top \nabla$, one obtains a model similar to what Jun and Stein (2008) used to analyze TOMS Level 3 ozone data, and we used this restricted nested SPDE model to analyze the global spatially irregular TOMS Level 2 data. This application illustrates the ability to use the model class to produce nonstationary covariance models on general smooth manifolds which efficiently can be used to study large spatially irregular data sets.

The most important next step in this work is to make a spatio-temporal extension of the model class. This would allow us to produce not only spatial but also spatio-temporal ozone models and increase the applicability of the model class to other environmental modeling problems where time dependence is a necessary model component.

A Vector spherical harmonics

When using the nonstationary model (10) in practice, we assume that the parameters in the model can be expressed in terms of some basis functions. If working on the sphere, spherical harmonics is a convenient basis for the parameters taking values in $\mathbb{R}$. On real form, the spherical harmonic $Y_{k,m}(\mathbf{s})$ of order $k \in \mathbb{N}_0$ and mode $m = -k, \ldots, k$ is defined as

$$Y_{k,m}(\mathbf{s}) = \sqrt{\frac{2k+1}{4\pi} \cdot \frac{(k-|m|)!}{(k+|m|)!}} \cdot \begin{cases} \sqrt{2}\sin(ml_1)P_{k,-m}(\sin l_2), & -k \leq m < 0, \\ P_{k,0}(\sin l_2), & m = 0, \\ \sqrt{2}\cos(ml_1)P_{k,m}(\sin l_2), & 0 < m \leq k, \end{cases}$$

where l_2 is the latitude, l_1 is the longitude, and $P_{k,m}(\cdot)$ are associated Legendre functions. We, however, also need a basis for the vector fields $\mathbf{B}_i(\mathbf{s})$, determining the direction and magnitude of differentiation. Since the vector fields in each

point on the sphere must lie in the tangent space of $\mathbb{S}^2$, the basis functions also must satisfy this. A basis with this property is obtained by using a subset of the vector spherical harmonics (Hill, 1954). For each spherical harmonic $Y_{k,m}(\mathbf{s})$, $k > 0$, define the two vector spherical harmonics

$$\begin{aligned} \Upsilon^1_{k,m}(\mathbf{s}) &= \nabla_{\mathbb{S}^2} Y_{k,m}(\mathbf{s}), \\ \Upsilon^2_{k,m}(\mathbf{s}) &= \nabla_{\mathbb{S}^2} Y_{k,m(\mathbf{s})} \times \mathbf{s}. \end{aligned}$$

Here $\times$ denotes the cross product in $\mathbb{R}^3$ and $\nabla_{\mathbb{S}^2}$ is the gradient on $\mathbb{S}^2$. By defining the basis in this way, all basis functions in $\Upsilon^1 = \{\Upsilon^1_{k,m}\}$ and $\Upsilon^2 = \{\Upsilon^2_{k,m}\}$ will obviously lie in the tangent space of $\mathbb{S}^2$. It is also easy to see that the basis is orthogonal in the sense that for any $k, l > 0$, $-k \leq m \leq k$, and $-l \leq n \leq l$, one has

$$\begin{aligned} \langle \Upsilon^1_{k,m}, \Upsilon^2_{l,n} \rangle_{\mathbb{S}^2} &= 0, \\ \langle \Upsilon^1_{k,m}, \Upsilon^1_{l,n} \rangle_{\mathbb{S}^2} &= k(k+1)\delta_{k-l}\delta_{m-n}, \\ \langle \Upsilon^2_{k,m}, \Upsilon^2_{l,n} \rangle_{\mathbb{S}^2} &= k(k+1)\delta_{k-l}\delta_{m-n}. \end{aligned}$$

These are indeed desirable properties for a vector field basis, but for the basis to be of any use in practice, a method for calculating the basis functions explicitly is needed. Such explicit expressions are given in the following proposition.

Proposition A.1. *With* $\mathbf{s} = (x, y, z)^\top$, $\Upsilon^1_{k,m}(\mathbf{s})$ *and* $\Upsilon^2_{k,m}(\mathbf{s})$ *can be written as*

$$\begin{aligned} \Upsilon^1_{k,m}(\mathbf{s}) &= \frac{1}{1-z^2} \begin{bmatrix} -myY_{k,-m}(\mathbf{s}) - c_{k,m}xzY_{k-1,m}(\mathbf{s}) + kxz^2 Y_{k,m}(\mathbf{s}) \\ mxY_{k,-m}(\mathbf{s}) - c_{k,m}yzY_{k-1,m}(\mathbf{s}) + kyz^2 Y_{k,m}(\mathbf{s}) \\ c_{k,m}(1-z^2)Y_{k-1,m}(\mathbf{s}) - (1-z^2)kzY_{k,m}(\mathbf{s}) \end{bmatrix}, \\ \Upsilon^2_{k,m}(\mathbf{s}) &= \frac{1}{1-z^2} \begin{bmatrix} kzyY_{k,m}(\mathbf{s}) - c_{k,m}yY_{k-1,m}(\mathbf{s}) + mzxY_{k,-m}(\mathbf{s}) \\ -kxzY_{k,m}(\mathbf{s}) + c_{k,m}xY_{k-1,m}(\mathbf{s}) + myzY_{k,-m}(\mathbf{s}) \\ -m(1-z^2)Y_{k,-m}(\mathbf{s}) \end{bmatrix}, \end{aligned}$$

where

$$c_{k,m} = \sqrt{\frac{(2k+1)(k^2 - |m|^2)}{2k-1}}.$$

Proof. One has that $\nabla_{\mathbb{S}^2} Y_{k,m} = P_{\mathbb{S}^2}(\nabla_{\mathbb{R}^3} Y_{k,m})$, that is, the gradient on $\mathbb{S}^2$ can be obtained by first calculating the gradient in $\mathbb{R}^3$ and then projecting the result onto

$\mathbb{S}^2$. If c_k^m denotes the normalization constant for the spherical harmonic $Y_{k,m}(\mathbf{s})$, and the recursive relation

$$(1-z^2)\frac{\partial}{\partial z}P_{k,m}(z) = kzP_{k,m}(z) - (k+m)P_{k-1,m}(z)$$

is used, one has that

$$\frac{\partial}{\partial z}Y_{k,m}(\mathbf{s}) = \frac{1}{1-z^2}\left(kzY_{k,m}(\mathbf{s}) - (k+|m|)\frac{c_k^m}{c_{k-1}^m}Y_{k-1,m}(\mathbf{s})\right).$$

Now, using that $\tan(l_1) = x^{-1}y$, one has

$$\begin{aligned}\frac{\partial l_1}{\partial x} &= -\cos^2(l_1)\frac{y}{x^2} = -\frac{y}{1-z^2},\\ \frac{\partial l_1}{\partial y} &= \cos^2(l_1)\frac{1}{x} = \frac{x}{1-z^2},\end{aligned}$$

where the last equalities hold on $\mathbb{S}^2$. Using these relations gives

$$\frac{\partial}{\partial x}Y_{k,m}(\mathbf{s}) = -\frac{my}{1-z^2}Y_{k,-m}(\mathbf{s}), \qquad \frac{\partial}{\partial y}Y_{k,m}(\mathbf{s}) = \frac{mx}{1-z^2}Y_{k,-m}(\mathbf{s}).$$

Thus, with

$$c_{k,m} \triangleq (k+|m|)\frac{c_k^m}{c_{k-1}^m} = \sqrt{\frac{(2k+1)(k^2-|m|^2)}{2k-1}},$$

one has that

$$\nabla_{\mathbb{R}^3}Y_{k,m}(\mathbf{s}) = \frac{1}{1-z^2}\begin{bmatrix} -myY_{k,-m}(\mathbf{s}) \\ mxY_{k,-m}(\mathbf{s}) \\ kzY_{k,m}(\mathbf{s}) - c_{k,m}Y_{k-1,m}(\mathbf{s})\end{bmatrix}.$$

Finally, the desired result is obtained by calculating

$$\begin{aligned}\Upsilon_{k,m}^1 &= \nabla_{\mathbb{S}^2}Y_{k,m} = \mathbf{P}_{\mathbb{S}^2}\nabla_{\mathbb{R}^3}Y_{k,m},\\ \Upsilon_{k,m}^2 &= \Upsilon_{k,m}^1 \times \mathbf{s} = \mathbf{S}_\times \Upsilon_{k,m}^1,\end{aligned}$$

where

$$\mathbf{P}_{\mathbb{S}^2} = (I - \mathbf{s}\mathbf{s}^\top) = \begin{bmatrix} 1-x^2 & -xy & -xz \\ -xy & 1-y^2 & -yz \\ -xz & -yz & 1-z^2\end{bmatrix}, \quad \mathbf{S}_\times = \begin{bmatrix} 0 & -z & y \\ z & 0 & -x \\ -y & x & 0\end{bmatrix}.$$

□

References

Adler, R. J. (1981). *The Geometry of Random Fields*. Wiley, New York.

Bolin, D. and Lindgren, F. (2009). Wavelet Markov approximations as efficient alternatives to tapering and convolution fields (submitted). *Preprints in Math. Sci. Lund University*, 2009:13.

Cressie, N. and Johannesson, G. (2008). Fixed rank kriging for very large spatial data sets. *J. Roy. Statist. Soc. Ser. B Stat. Methodol.*, 70(1):209–226.

Gelman, A., Roberts, G. O., and Gilks, W. R. (1996). Efficient Metropolis jumping rules. *Bayesian Statist.*, 5:599–607.

Gneiting, T. (1998). Simple tests for the validity of correlation function models on the circle. *Statist. Probab. Lett.*, 39:119–122.

Hastie, T., Tibshirani, R., and Friedman, J. H. (2003). *The Elements of Statistical Learning*. Springer, New York.

Hill, E. L. (1954). The theory of vector spherical harmonics. *Am. J. Phys.*, 22:211–214.

Jun, M. and Stein, M. L. (2007). An approach to producing space-time covariance functions on spheres. *Technometrics*, 49(4):468–479.

Jun, M. and Stein, M. L. (2008). Nonstationary covariance models for global data. *Ann. Appl. Statist.*, 2(4):1271–1289.

Lindgren, F., Rue, H., and Lindström, J. (2011). 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.

Lindström, J. and Lindgren, F. (2008). A Gaussian Markov random field model for total yearly precipitation over the African Sahel. *Preprints in Math. Sci. Lund University*, 2008:8.

Matérn, B. (1960). Spatial variation. *Meddelanden från statens skogsforskningsinstitut*, 49(5).

McPeters, R. D., Bhartia, P. K., Krueger, A. J., Herman, J. R., Schlesinger, B., Wellemeyer, C. G., Seftor, C. J., Jaross, G., Taylor, S. L., Swissler, T., Torres, O., Labow, G., Byerly, W., and Cebula, R. P. (1996). *Nimbus-7 Total Ozone Mapping Spectrometer (TOMS) Data Products User's Guide*. NASA Reference Publication 1384.

Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields; Theory and Applications*, volume 104 of *Monographs on Statistics and Applied Probability*. Chapman & Hall/CRC.

Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations (with discussion). *J. Roy. Statist. Soc. Ser. B Stat. Methodol.*, 71(2):319–392.

Stein, M. L. (2007). Spatial variation of total column ozone on a global scale. *Ann. Appl. Statist.*, 1(1):191–210.

Thomas, G. B. and Finney, R. L. (1995). *Calculus and Analytic Geometry*. Addison Wesley, New York, 9 edition.

Whittle, P. (1963). Stochastic processes in several dimensions. *Bull. Internat. Statist. Inst.*, 40:974–994.

Yaglom, A. M. (1987). *Correlation Theory of Stationary and Related Random Functions*, volume 1. Springer-Verlag, New York.

D

Spatial Matérn fields driven by non-Gaussian noise

DAVID BOLIN

Centre for Mathematical Sciences, Lund University, Lund, Sweden

Abstract: The article studies non-Gaussian extensions of a recently discovered link between certain Gaussian random fields, expressed as solutions to stochastic partial differential equations (SPDEs), and Gaussian Markov random fields (GMRFs). The focus is on non-Gaussian random fields with Matérn covariance functions, and in particular we show how the SPDE formulation of a Laplace moving average (LMA) model can be used to obtain an efficient simulation method as well as an accurate parameter estimation technique for the model. These methods are based on an extension of the Hilbert space approximation technique by Lindgren et al. (2011) to SPDEs driven by Laplace noise, and although interesting in itself, the results for the LMA model should be seen as a demonstration of how these techniques can be used, and generalizations to more general SPDEs driven by more general noise processes are readily available.

Key words: Matérn covariances; SPDE; Laplace moving averages; Markov random fields; process convolutions; EM algorithm

1 Introduction

Recently, Lindgren et al. (2011) derived a link between certain Gaussian fields, that can be represented as solutions to stochastic partial differential equations (SPDEs), and Gaussian Markov random fields (GMRFs). The main idea is to approximate these Gaussian fields using basis expansions $\sum_i w_i \varphi_i(s)$ where the stochastic weights $\{w_i\}$ are calculated using the stochastic weak formulation of the corresponding SPDE. For certain choices of the basis functions $\{\varphi_i\}$, especially compactly supported functions, the weights form GMRFs. Because of the Markov property of the weights, fast numerical techniques for sparse matrices can be used when estimating parameters and doing spatial prediction in these models. This greatly improves the applicability to problems involving large data sets, where traditional methods in statistics fail due to computational issues. How-

ever, the advantages of representing Gaussian fields as solutions to SPDEs are not only computational. Using the SPDE representation, non-stationary extensions are easily obtained by allowing spatially varying parameters in the SPDE (Lindgren et al., 2011), and the model class can be generalized to include more general covariance structures by generalizing the class of generating SPDEs (Bolin and Lindgren, 2011). These are indeed useful features from an applied point of view as many applications require complicated non-stationary models to accurately capture the covariance structure of the data.

So far these methods have only been used in Gaussian settings, and it has not been clear whether they are applicable when the Gaussianity assumption cannot be justified. Therefore, this work will focus on extending the SPDE methods beyond Gaussianity. A new type of non-Gaussian models that has proved to be useful in practical applications is the Laplace moving average models (Åberg et al., 2009, Åberg and Podgórski, 2011). These are processes obtained by convolving some deterministic kernel function with stochastic Laplace noise. The models share many good properties with the Gaussian models while allowing for heavier tails and asymmetry in the data, making them interesting alternatives in practical applications (see e.g. Bogsjö et al., 2012). One of the motivating examples in Åberg and Podgórski (2011) is a Laplace moving average model with Matérn covariances. This model can be seen as the solution to the same SPDE that generates Gaussian Matérn field but where the Gaussian white noise forcing is replaced with Laplace noise. It has previously been shown that the SPDE model formulation of Gaussian Matérn fields has many computational advantages compared with the process convolution formulation (Bolin and Lindgren, 2009, Simpson et al., 2010). We demonstrate here that for the Laplace moving average models, the SPDE formulation can also be used to derive a new likelihood-based parameter estimation technique as well as an efficient simulation procedure.

The structure of the paper is as follows. Section 2 contains an introduction to the Matérn covariance family and the SPDE formulation in the Gaussian case. In Section 3, stochastic Laplace fields are introduced, and some properties of the Laplace-driven SPDE model are derived. Subsequently, in Section 4, the Markov approximation technique by Lindgren et al. (2011) is extended to the Laplace model, and its sampling is discussed in Section 5. A parameter estimation technique based on the EM algorithm is derived in Section 6, and Section 7 contains a simulation study showing that it gives reliable parameter estimates. Finally, Section 8 contains a summary and discussion of future work and possible extensions.

2 Gaussian Matérn fields

The Matérn covariance family (Matérn, 1960) is often used when modeling spatial data. There are a few different parameterizations of the Matérn covariance function in the literature, and the one most suitable in our context is

$$C(\mathbf{h}) = \frac{2^{1-\nu}\phi^2}{(4\pi)^{\frac{d}{2}}\Gamma(\nu+\frac{d}{2})\kappa^{2\nu}}(\kappa\|\mathbf{h}\|)^{\nu}K_{\nu}(\kappa\|\mathbf{h}\|), \quad \mathbf{h}\in\mathbb{R}^d, \tag{1}$$

where d is the dimension of the domain, ν is a shape parameter, κ^2 a scale parameter, ϕ^2 a variance parameter, and K_ν is a modified Bessel function of the second kind of order $\nu > 0$. The associated spectrum is

$$S(\mathbf{k}) = \frac{\phi^2}{(2\pi)^d}\frac{1}{(\kappa^2+\mathbf{k}^\top\mathbf{k})^{\nu+\frac{d}{2}}}. \tag{2}$$

As the properties of Gaussian fields are given by their first two moments, the standard way of specifying Gaussian Matérn fields is to chose the mean value, $\mu(\mathbf{s})$ possibly spatially varying, and then let the covariance function be of the form (1). An alternative way of specifying a Gaussian field on $\mathbb{R}^d$ is to view it as a process convolution

$$X(\mathbf{s}) = \int_{\mathbb{R}^d} k(\mathbf{s},\mathbf{u})\mathcal{B}(\,\mathrm{d}\mathbf{u}), \tag{3}$$

where k is some deterministic kernel function and $\mathcal{B}$ is a Brownian sheet (Higdon, 2001). One of the advantages with this construction is that non-stationary extensions are easily constructed by allowing the convolution kernel to be dependent on the location $\mathbf{s}$. If, however, the process is stationary, the kernel k depends only on $\mathbf{s}-\mathbf{u}$ and the covariance function for X is

$$C(\mathbf{h}) = \int_{\mathbb{R}^d} k(\mathbf{u}-\mathbf{h})k(\mathbf{u})\,\mathrm{d}\mathbf{u}.$$

Thus, the covariance function C, the spectrum S, and the kernel k are related through

$$(2\pi)^d|\mathcal{F}(k)|^2 = \mathcal{F}(C) = S,$$

where $\mathcal{F}(\cdot)$ denotes the Fourier transform. Since the spectral density for a Matérn field in dimension d with parameters ν, ϕ^2, and κ is given by (2), one finds that

the corresponding symmetric non-negative kernel is a Matérn covariance function with parameters $\nu_k = \frac{\nu}{2} - \frac{d}{4}$, $\phi_k = \sqrt{\phi}$, and $\kappa_k = \kappa$.

In yet another setting, Gaussian Matérn fields can be viewed as the solution to the SPDE

$$(\kappa^2 - \Delta)^{\frac{\alpha}{2}} X(\mathbf{s}) = \phi\, \mathcal{W}(\mathbf{s}), \tag{4}$$

where $\mathcal{W}(\mathbf{s})$ is Gaussian white noise, $\Delta = \sum_{i=1}^{d} \frac{\partial^2}{\partial \mathbf{s}_i^2}$ is the Laplace operator, and $\alpha = \nu + d/2$ (Whittle, 1963). As discussed in Lindgren et al. (2011), there is an implicit assumption of appropriate boundary conditions needed if one wants the solutions to be stationary Matérn fields.

The connection between (3) and (4) is through the Green's function of the differential operator in (4)

$$G_\alpha(\mathbf{s}, \mathbf{t}) = \frac{2^{1-\frac{\alpha-d}{2}}}{(4\pi)^{\frac{d}{2}} \Gamma(\frac{\alpha}{2}) \kappa^{\alpha-d}} (\kappa \|\mathbf{s} - \mathbf{t}\|)^{\frac{\alpha-d}{2}} K_{\frac{\alpha-d}{2}}(\kappa \|\mathbf{s} - \mathbf{t}\|), \tag{5}$$

that serves as a kernel in (3). It is straightforward to show that $G_\alpha \in L_p(\mathbb{R}^d)$ if and only if $\alpha > \frac{(p-1)d}{p}$ (see for example Samko et al. (1992) p. 538), and in particular $\alpha > d/2$ guarantees that $G_\alpha \in L_2(\mathbb{R}^d)$.

A non-Gaussian model with Matérn covariances could be constructed either using the process convolution formulation (3) where the Brownian sheet is replaced by some non-Gaussian process, or through the SPDE formulation (4) with non-Gaussian noise. Such non-Gaussian extensions are discussed next.

3 Non-Gaussian SPDE-based models

A simple way of moving beyond Gaussianity in the SPDE model (4) is to allow for a stochastic variance parameter ϕ. By choosing ϕ as an inverse-gamma distributed random variable, the resulting field has t-distributed marginal distributions and is therefore sometimes referred to as a t-distributed random field (Røislien and Omre, 2006). In a Bayesian setting, this extension can be interpreted simply as choosing a certain prior distribution for the variance, and one can of course come up with many other non-Gaussian models by changing this distribution. However, models constructed in this way are non-Gaussian only in a very limited sense. Namely, every realization of them behaves exactly as a Gaussian field

with a globally re-scaled variance, and because of this, they are all non-ergodic as the parameters in the prior distribution cannot be estimated from a single realization of the field. One would prefer a non-Gaussian model where the actual sample paths behave differently from a stationary Gaussian field, and one way of achieving this is to let the variance parameter be spatially and stochastically varying. Both Lindgren et al. (2011) and Bolin and Lindgren (2011) explores this option by expressing $\log \phi(\mathbf{s})$ as a regression on a few known basis functions where the stochastic weights are estimated from data. This was interpreted as a non-stationary Gaussian model, but could also be viewed as a, somewhat limited, non-Gaussian model with a slowly spatially varying variance parameter $\phi(\mathbf{s})$. To obtain a model which is intrinsically non-Gaussian also within realizations, one can draw $\phi(\mathbf{s})$ at random independently for each $\mathbf{s}$. The right-hand side of (4) is then a product of two independent noise fields. The following non-Gaussian models essentially can be interpreted as a formal realization of this idea.

One interesting type of distributions, obtained by taking a random variance and mean in an otherwise Gaussian random variable, are the generalized asymmetric Laplace distributions (Åberg et al., 2009). The Laplace distribution is defined through the characteristic function with parameters $\mu, \gamma \in \mathbb{R}$ and $\sigma, \tau > 0$

$$\varphi(u) = e^{i\gamma u}\left(1 - i\mu u + \frac{\sigma^2}{2}u^2\right)^{-\tau}.$$

The distribution is symmetric if $\mu = 0$ and asymmetric otherwise. The shape of the distribution is governed by τ and the scale by σ. The distribution is infinitely divisible, and a useful characterization is that if Z is a standard normal variable and Γ is an independent gamma variable with shape τ, then $\gamma + \mu\Gamma + \sigma\sqrt{\Gamma}Z$ has an asymmetric Laplace distribution.

Stochastic Laplace noise can now be obtained from an independently scattered random measure Λ, defined for a Borel set B in $\mathbb{R}^d$ by the characteristic function

$$\varphi_{\Lambda(B)}(u) = e^{i\gamma m(B)u}\left(1 - i\mu u + \frac{\sigma^2}{2}u^2\right)^{-m(B)},$$

where the measure m is referred to as the control measure of Λ. This does not define Laplace noise in a direct manner, but similarly to how Gaussian white noise can be seen as a differentiated Brownian sheet (Walsh, 1986), Laplace noise can be viewed in the sense of distributions (generalized functions) as a differentiated

Laplace field. The most transparent characterization is through the following series representation of the Laplace field $\Lambda(\mathbf{s})$ on a compact $D \in \mathbb{R}^d$:

$$\Lambda(\mathbf{s}) = \gamma\mathbf{s} + \sum_{k=1}^{\infty} \left(\Gamma_k + G_k\sqrt{\Gamma_k}\right) \mathbf{1}(\mathbf{s} \geq \mathbf{s}_k), \quad \mathbf{s} \in D, \tag{6}$$

where G_k are iid $\mathsf{N}(0,1)$ random variables, $\mathbf{s}_k$ are iid uniform random variables on D, and

$$\mathbf{1}(\mathbf{s} \geq \mathbf{s}_k) = \begin{cases} 1 & \text{if } s_i \geq s_{k,i} \text{ for all } i \leq d, \\ 0 & \text{otherwise.} \end{cases}$$

The random variables Γ_k can be written as $\Gamma_k = e^{-\nu\gamma_k} W_k$ where W_k are iid standard exponential variables and γ_k are the arrival times of a Poisson process with intensity 1. Thus, Laplace noise can be expressed as a distribution (generalized function)

$$\dot{\Lambda} = \gamma + \sum_{i=k}^{\infty} \left(\Gamma_k + G_k\sqrt{\Gamma_k}\right) \delta_{\mathbf{s}_k}, \tag{7}$$

where $\delta_{\mathbf{s}_k}$ is the Dirac delta distribution centered at $\mathbf{s}_k$.

The model of interest is the solution X to the Laplace-driven SPDE

$$(\kappa^2 - \Delta)^{\frac{\alpha}{2}} X = \dot{\Lambda}, \tag{8}$$

where both X and $\dot{\Lambda}$ are viewed as random variables valued in the space of tempered distributions. To clarify in what way the solution to this equation exists, we look at a general SPDE

$$(\kappa^2 - \Delta)^{\frac{\alpha}{2}} X = \dot{M}, \tag{9}$$

where M is an arbitrary independently scattered L_2-valued random measure with $\mathsf{E}(|M(\mathrm{d}\mathbf{x})|^2) = C\,\mathrm{d}\mathbf{x}$ for some constant $C < \infty$. Examples of such measures are the Laplace measures of interest here but also standard Brownian sheets. As usual for fractional Laplacian operators (Samko et al., 1992), $\mathcal{T} = (\kappa^2 - \Delta)^{\frac{\alpha}{2}}$ is defined using the Fourier transform through $\mathcal{F}(\mathcal{T}f) = \mathcal{P}\hat{f}$, where $\hat{f}$ is the Fourier transform of the function f, $(\mathcal{P}\hat{f})(\mathbf{k}) = (\kappa^2 + \mathbf{k}^\top\mathbf{k})^{\frac{\alpha}{2}}\hat{f}(\mathbf{k})$, and the operator $\mathcal{T}$ is well-defined for example for all $f \in L_p(\mathbb{R}^d)$ for $1 \leq p \leq \infty$. The definition

applies also when f is a distribution or, more specifically, a tempered distribution. Thus, (9) is viewed as an equation for two random (tempered) distributions so the equation has to be interpreted in the weak sense

$$\mathcal{T}X(\varphi) = \dot{M}(\varphi), \tag{10}$$

where φ is in some appropriate space of test functions. Now, the action of the self-adjoint operator $\mathcal{T}$ can be moved to the test function on the left-hand side and (10) can be rewritten in a more explicit fashion as

$$X(\mathcal{T}\varphi, \omega) = \int \varphi(\mathbf{s}) M(\,\mathrm{d}\mathbf{s}, \omega). \tag{11}$$

Here, we have included the second argument $\omega \in \Omega$ to highlight that the sought functional X is random, and the equation should hold for ω in a certain full probability set $\Omega_0 \in \Omega$ and universally for each φ.

To describe the solutions of (9), we need the Sobolev spaces H_t of fractional order t. These are usually defined using the Fourier transform in the following way. Let E be the Schwartz space of rapidly decreasing functions on $\mathbb{R}^d$, for $u \in E'$ (the dual of E, also referred to as the space of tempered distributions), define the Fourier transform of u as $\hat{u}(\varphi) = u(\hat{\varphi})$, where $\hat{\varphi}$ is the usual Fourier transform on $\mathbb{R}^d$ of $\varphi \in E$. Define a norm on E by

$$\|u\|_t = \int_{\mathbb{R}^d} (1 + |\mathbf{k}|^2)^t |\hat{u}(\mathbf{k})|^2 \,\mathrm{d}\mathbf{k}$$

and let H_t be the completion of E in this norm. By Plancherel's theorem, one has that $H_0 = L_2(\mathbb{R}^d)$ and one can show that for the special case $t = n \in \mathbb{N}$, H_n is identical to the classical Sobolev space of L_2 functions with all partial derivatives of order n or less in L_2. The space H_{-t} is the dual space of H_t and does in general contain distributions.

Let us note that the right hand side of (11) in principle may not be defined on a full probability set uniformly for all φ. However, one can regularize M so that $\varphi \to M(\varphi)$ is in fact a random distribution. Indeed, since

$$\mathsf{E}(|M(\varphi)|^2) = C \int \varphi(\mathbf{s})^2 \,\mathrm{d}\mathbf{s} = C\|\varphi\|_0^2,$$

the random linear functional $\varphi \to M(\varphi)$ is continuous in probability on H_n for any $n \geq 0$, and by Theorem 4.1 in Walsh (1986) there exists a version of M

which is almost surely in H_{-n} for $n > d/2$. From now on we always assume that we deal with such a version.

Following Walsh (1986), we say that $X(\cdot, \omega)$ is an H_n-solution of (9) if for a.e. ω, $X(\cdot, \omega)$ is an element of H_{-n} and (11) holds for every $\varphi \in H_n$. In other words, we aim at finding a random functional X that almost surely is a distribution and satisfies (9) as a continuous functional on H_n. The proof of the following proposition is similar to the proof of Proposition 9.1 in Walsh (1986) where the existence of the solution to the stochastic Poisson equation on a bounded domain in $\mathbb{R}^d$ was demonstrated.

Proposition 3.1. *Assume that M is an independently scattered L_2-valued random measure with $\mathsf{E}(|M(\mathrm{d}\mathbf{x})|^2) = C\,\mathrm{d}\mathbf{x}$. Then for $\kappa > 0$, $\alpha > 0$, there exists a random functional $X : H_n \times \Omega \to \mathbb{R}$ such that for a certain set Ω_0, $P(\Omega_0) = 1$ and for all $\omega \in \Omega_0$ and all $\varphi \in H_n$*

$$X(\varphi, \omega) = \int G^\alpha \varphi(\mathbf{x}) M(\mathrm{d}\mathbf{x}, \omega), \tag{12}$$

where $G^\alpha \varphi(\mathbf{x}) = \int G_\alpha(\mathbf{s}, \mathbf{x})\varphi(\mathbf{s})\,\mathrm{d}\mathbf{s}$ and G_α is given by (5). This is the unique H_n-solution to (9) if $n > d/2$, and moreover we have $X \in H_m$ almost surely for $m < \alpha - d/2$.

Proof. From the standard theory of fractional differential equations, one has that G^α maps H_n isomorphically onto $H_{n+\alpha}$ (see e.g. Samko et al., 1992, p.547). Let X be any H_n-solution to (9) and let $\psi = G^\alpha \varphi$. Applying (11) to ψ and using that $\mathcal{T} G^\alpha \varphi = \varphi$ one gets that

$$X(\varphi) = X(\mathcal{T} G^\alpha \varphi) = X(\mathcal{T}\psi) = \int \psi(\mathbf{y}) M(\mathrm{d}\mathbf{y}) = \int G^\alpha \varphi(\mathbf{y}) M(\mathrm{d}\mathbf{y}).$$

Thus this solution also satisfies (12) and the solution is unique if it exists.

To prove existence, let X be defined by (12) and take $\varphi \in L_2(\mathbb{R}^d)$. Then

$$\begin{aligned}
\mathsf{E}(|X(\varphi)|^2) &= \mathsf{E}\left[\left(\int G^\alpha \varphi(\mathbf{y}) M(\mathrm{d}\mathbf{y})\right)^2\right] \\
&= \int G^\alpha \varphi(\mathbf{x}) G^\alpha \varphi(\mathbf{y}) \mathsf{E}\left[M(\mathrm{d}\mathbf{x}) M(\mathrm{d}\mathbf{y})\right] \\
&= C \int \left(G^\alpha \varphi(\mathbf{x})\right)^2 \mathrm{d}\mathbf{x} = C\|G^\alpha \varphi\|_0^2 \le C_2 \|\varphi\|_{-\alpha}^2,
\end{aligned}$$

where the last inequality follows from that G^α maps $H_{n-\alpha} \to H_n$ boundedly for $\alpha > 0$. Thus, it follows that X is a random linear functional that is continuous in probability on $H_{-\alpha}$. The embedding maps $H_{n_1} \to H_{n_2}$ are of Hilbert-Schmidt type if $n_1 > n_2 + d/2$ (see e.g. Example 1a in Walsh, 1986), and using this with $n_2 = -\alpha$ together with Theorem 4.1 in Walsh (1986) one gets that there exists a version of X which is almost surely in H_{-n} if $n > d/2 > d/2 - \alpha$. From now on, X is such a version and we note that $X \in H_m$ almost surely for $m < \alpha - d/2$.

What is left to show now is that X with probability one satisfies (11) for each $\varphi \in H_n$ for $n > d/2$. To that end, first note that if $\varphi \in H_n$, then by the definitions of $\mathcal{T}$ and G^α one has

$$\begin{aligned} G^\alpha \mathcal{T}\varphi(\mathbf{s}) &= \int G_\alpha(\mathbf{y}, \mathbf{s}) \mathcal{T}\varphi(\mathbf{y}) \, \mathrm{d}\mathbf{y} \\ &= \mathcal{F}^{-1}\left((\kappa^2 + \mathbf{k}^\top \mathbf{k})^{-\frac{\alpha}{2}} (\kappa^2 + \mathbf{k}^\top \mathbf{k})^{\frac{\alpha}{2}} \hat{\varphi}(\mathbf{k}) \right)(\mathbf{s}) \\ &= \varphi(\mathbf{s}). \end{aligned}$$

Let $n > d/2$ and fix $\varphi \in H_n$. If $M(\varphi, \omega)$ denotes the functional $\int \varphi(\mathbf{s}) M(\, \mathrm{d}\mathbf{s}, \omega)$, one has by the definition of X and by the equation above that

$$\int |X(\mathcal{T}\varphi, \omega) - M(\varphi, \omega)|^2 \, \mathrm{d}\mathsf{P}(\omega) = 0.$$

Hence, there is a set $\Omega_\varphi \subset \Omega$ with $\mathsf{P}(\Omega_\varphi) = 1$ such that for each $\omega \in \Omega_\varphi$ one has $X(\mathcal{T}\varphi, \omega) = M(\varphi, \omega)$. Now, H_n is separable, so we can chose a countable base $B = \{b_i\}_{i=1}^\infty$ in H_n and define $\bar{\Omega}_0 = \cap_{i=1}^\infty \Omega_{b_i}$. Then equality holds for each $f \in B$ and for each $\omega \in \bar{\Omega}_0$ and $\mathsf{P}(\bar{\Omega}_0) = 1$ by the countability of B.

The map $\varphi \to \mathcal{T}\varphi \to X(\mathcal{T}\varphi)$ of $H_n \to H_{n-\alpha} \to \mathbb{R}$ is continuous since X is continuous on H_n for $n > d/2 - \alpha$ and $\mathcal{T}$ is a continuous map from H_n to $H_{n-\alpha}$. Thus, both $X(\mathcal{T}\cdot, \omega)$ and $M(\cdot, \omega)$ are continuous functionals on H_n for ω in some full probability set $\tilde{\Omega}_0$ and equality therefore holds in (11) for each $\varphi \in H_n$ for $\omega \in \Omega_0 = \bar{\Omega}_0 \cap \tilde{\Omega}_0$ since B is linearly dense in H_n. $\square$

Remark 1. By similar arguments one can show that the solution X defined in Proposition 3.1 also is a solution to the SPDE (9) in the sense that with probability one $X \in E'$ and (11) holds for every $\varphi \in E$. This is, however, a weaker statement since $E = \cap_n H_n$ and $E' = \cup_n H_n$.

Remark 2. The solution X defined in Proposition 3.1 is in general a random linear functional. However, it can be identified with a random function if $\alpha > d/2$ since

$X \in H_m$ almost surely for $m < \alpha - d/2$. Using the relation between α and the parameter ν in the Matérn covariance function, $\alpha = \nu + d/2$, we see that $X \in H_m$ almost surely for $m < \nu$. Thus, ν acts as a smoothness parameter for the solution since the sample paths almost surely will be differentiable if $\nu > 1$, two times differentiable if $\nu > 2$ etc.

Remark 3. The previous remark can be strengthened using the Sobolev embedding theorem which shows that H_n can be embedded in the Hölder space $C_k^r(\mathbb{R}^d)$ where $n - (r + k) = d/2$ and $r \in (0, 1)$ (see e.g. Adams, 1975). The space $C_k^r(\mathbb{R}^d)$ consists of functions such that all partial derivatives up to order k are continuous and such that the kth partial derivatives are Hölder continuous with exponent r. Thus, if $\nu > d/2$, we almost surely have $X \in C_k^r(\mathbb{R}^d)$ (after possibly redefining it on a set of measure zero) where k is the integer part of $\nu - d/2$ and $r = \nu - d/2 - k$.

We now go back to the special case of Laplace noise and since the main interest here is ordinary random fields with Matérn covariance functions, we from now on assume that $\alpha > d/2$ in (8). One sometimes uses $m(A) = l(A)\tau$, where l is the Lebesgue measure and τ some constant, as a control measure for Λ. By the definition of the differential operator $\mathcal{T}$, it is then easy to see that the spectrum for the solution X is

$$R_X(\mathbf{k}) = \frac{\tau(\sigma^2 + \mu^2)}{(2\pi)^d} \frac{1}{(\kappa^2 + \mathbf{k}^\top \mathbf{k})^\alpha}.$$

Thus, the covariance function for X is a Matérn covariance of the form (1) with $\phi^2 = \tau(\sigma^2 + \mu^2)$. Since X is Laplace noise convolved with a Green function, which also has the form of a Matérn covariance function, the model is equivalent to the Laplace moving average models in Åberg et al. (2009) and Åberg and Podgórski (2011). Thus, using Theorem 1 in Åberg and Podgórski (2011), the marginal distribution for $X(\mathbf{s})$ is given by the characteristic function

$$\phi_X(u) = \exp\left(\tau \int i\gamma G_\alpha(\mathbf{s}, \mathbf{t})u - \log\left(1 - i\mu u G_\alpha(\mathbf{s}, \mathbf{t}) + \frac{\sigma^2 u^2}{2} G_\alpha^2(\mathbf{s}, \mathbf{t})\right) \mathrm{d}\mathbf{t}\right). \tag{13}$$

A few examples of the marginal distributions for symmetric and asymmetric cases are shown in Figure 1.

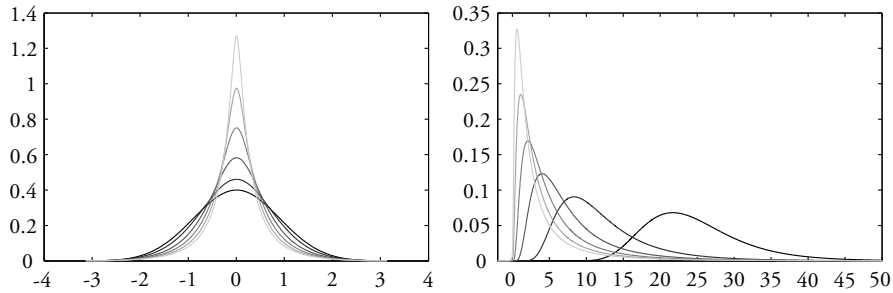

Figure 1: Marginal distributions of the solution $X(\mathbf{s})$ to (8) in the case of symmetric (left panel) and asymmetric (right panel) Laplace noise.

4 Hilbert space approximations

To obtain a computationally efficient representation of a Matérn field, the Hilbert space approximation technique by Lindgren et al. (2011) can be used. The starting point is to consider the stochastic weak formulation (10) of the SPDE. A finite element approximation of the solution X is then obtained by representing it as a finite basis expansion $\tilde{X} = \sum_{i=1}^{n} w_i \varphi_i(\mathbf{s})$, where the stochastic weights are calculated by requiring (10) to hold for only a specific set of test functions $\{\psi_i, i = 1, \ldots, n\}$ and $\{\varphi_i\}$ is a set of predetermined basis functions. To simplify the presentation, we first look at the case $\alpha/2 \in \mathbb{N}$ and then turn to the case of a general $\alpha > d/2$.

4.1 The case $\alpha/2 \in \mathbb{N}$

To construct the approximation for $\alpha = 2, 4, \ldots$, we first look at the fundamental case $\alpha = 2$. Lindgren et al. (2011) then use $\psi_i = \varphi_i$, and one then has

$$(\kappa^2 - \Delta)\tilde{X}(\varphi_i) = \sum_{j=1}^{n} w_j \left\langle \varphi_i, (\kappa^2 - \Delta)\varphi_j \right\rangle ,$$

where $\langle f, g \rangle = \int f(s)g(s)\,\mathrm{d}\mathbf{s}$. By introducing the vector $\mathbf{w} = (w_1, \ldots, w_n)^\top$ and a matrix $\mathbf{K}$ with elements $\mathbf{K}_{ij} = \left\langle \varphi_i, (\kappa^2 - \Delta)\varphi_j \right\rangle$, the left hand side of (10) can

be written as $\mathbf{Kw}$. Under mild conditions on the basis functions, one has

$$\begin{aligned}\left\langle \varphi_i,\, (\kappa^2 - \Delta)\varphi_j \right\rangle &= \kappa^2 \left\langle \varphi_i,\, \varphi_j \right\rangle - \left\langle \varphi_i,\, \Delta\varphi_j \right\rangle \\ &= \kappa^2 \left\langle \varphi_i,\, \varphi_j \right\rangle + \left\langle \nabla\varphi_i,\, \nabla\varphi_j \right\rangle .\end{aligned}$$

Hence, the matrix $\mathbf{K}$ can be written as the sum $\mathbf{K} = \kappa^2\mathbf{C} + \mathbf{G}$ where $\mathbf{C}$ and $\mathbf{G}$ are matrices with elements $\mathbf{C}_{ij} = \left\langle \varphi_i,\, \varphi_j \right\rangle$ and $\mathbf{G}_{ij} = \left\langle \nabla\varphi_i,\, \nabla\varphi_j \right\rangle$ respectively.

4.1.1 Gaussian noise

In the Gaussian case, when $\dot{M}$ is Gaussian white noise, the right hand side of (10) under the finite element approximation can be shown to be Gaussian with mean zero and covariance $\mathbf{C}$. Thus, one has

$$\mathbf{w} \sim \mathsf{N}\left(0, \mathbf{K}^{-1}\mathbf{C}\mathbf{K}^{-1}\right). \tag{14}$$

For higher order $\alpha/2 \in \mathbb{N}$, the weak solution is obtained recursively. If, for example, $\alpha = 4$ the solution to $(\kappa^2 - \Delta)^2 X_0 = \mathcal{W}$ is obtained by solving $(\kappa^2 - \Delta)X_0 = \tilde{X}$, where $\tilde{X}$ is the solution for the case $\alpha = 2$. This results in replacing the matrix $\mathbf{K}$ with a matrix $\mathbf{K}_\alpha$ defined recursively as $\mathbf{K}_\alpha = \mathbf{K}\mathbf{C}^{-1}\mathbf{K}_{\alpha-2}$, where $\mathbf{K}_2 = \mathbf{K}$. For more details about these representations in the Gaussian case, see Lindgren et al. (2011).

So far, we have not specified how the basis functions $\{\varphi_i\}$ should be chosen, but this choice will determine the quality of the approximation as well as some computational properties. If, for example, Daubechies wavelets are used as basis functions, the precision matrix (inverse covariance matrix) $\mathbf{Q}$ for the weights is a sparse matrix (Bolin and Lindgren, 2009), which facilitates the use of efficient sparse matrix techniques when using this model. Lindgren et al. (2011) used piecewise linear basis functions induced by triangulating the domain, and in this case $\mathbf{C}$ is a sparse matrix, but its inverse is dense. To obtain a sparse precision matrix in this case (which is needed for efficient GMRF computations), one can approximate $\mathbf{C}$ with a diagonal matrix $\tilde{\mathbf{C}}$ with elements $\tilde{\mathbf{C}}_{ii} = \int \varphi_i(\mathbf{s})\,\mathrm{d}\mathbf{s}$. To simplify the notation later, we denote the ith element on the diagonal by $\mathbf{a}_i$ as it is the area where $\varphi_i > \varphi_j$ for $j \neq i$. For more details on this approximation and the choice of basis functions, see Bolin and Lindgren (2009).

4.1.2 Laplace noise

For the Laplace case, one has $\dot{M} = \dot{\Lambda}$ in the weak formulation (10). Under the finite element approximation, the left-hand side can, as in the Gaussian case, be written as $\mathbf{K}_\alpha \mathbf{w}$. Using Theorem 1 in Åberg and Podgórski (2011), the distribution of the right-hand side in the case of Laplace noise is given by the characteristic function

$$\phi_\Lambda(\mathbf{u}) = \exp\left(\tau \int_\Omega i\gamma\varphi(\mathbf{s})^\top \mathbf{u} - \log\left(1 - i\mu\varphi(\mathbf{s})^\top \mathbf{u} + \frac{\sigma^2}{2}(\varphi(\mathbf{s})^\top \mathbf{u})^2\right) \mathrm{d}\mathbf{s}\right),$$

where $\varphi(\mathbf{s}) = \left(\varphi_1(\mathbf{s}), \ldots, \varphi_n(\mathbf{s})\right)^\top$. This representation is not very convenient for approximation and simulation of the model. Instead we will use a representation based on the series expansion (6) of Λ. However, we for a moment turn to the more general setup of type-G processes to hint at how this technique could be applied also for this broader class of random fields.

Recall that a Lévy process is type G if its increments can be represented as a Gaussian variane mixture $V^{1/2}Z$ where Z is a standard Gaussian variable and V is a non-negative infinitely divisible random variable. Clearly, the Laplace fields are of type G as their increments are of the form $\Gamma^{1/2}Z$ where Γ is a gamma variable. Rosiński (1991) showed that every Lévy process of type G can be represented as a series expansion similar to the expansion (6) for the Laplace fields. This expansion also holds in $\mathbb{R}^d$, and for a compact domain $D \in \mathbb{R}^d$ it can be written as

$$M(\mathbf{s}) = \sum_{k=1}^{\infty} G_k g(\gamma_k)^{\frac{1}{2}} \mathbf{1}(\mathbf{s} \geq \mathbf{s}_k),$$

where the function g is the generalized inverse of the tail Lévy measure for V and the other variables are the same as in the Laplace case (6). Since V is infinitely divisible, there exists a non-decreasing Lévy process $V(\mathbf{s})$ with increments distributed the same as V. This process has the series representation

$$V(\mathbf{s}) = \sum_{k=1}^{\infty} g(\gamma_k)^{\frac{1}{2}} \mathbf{1}(\mathbf{s} \geq \mathbf{s}_k). \tag{15}$$

Now, consider the integral of some basis function φ_i with respect to M, which can be represented as

$$\int_D \varphi_i(\mathbf{s}) M(\mathrm{d}\mathbf{s}) \stackrel{d}{=} \sum_{k=1}^{\infty} \varphi_i(\mathbf{s}_k) G_k \sqrt{g(\gamma_k)}. \tag{16}$$

Thus, the distribution of $(\int_D \varphi_1(\mathbf{s})M(d\mathbf{s}), \ldots, \int_D \varphi_n(\mathbf{s})M(d\mathbf{s}))$ can be approximated in distribution by taking partial sums of the series in (16). Another way of calculating the distribution is to evaluate the integrals by conditioning on the variance process $V(\mathbf{s})$ (Wiktorsson, 2002); given that $\int_D \varphi_i^2(\mathbf{s})V(\,\mathrm{d}\mathbf{s}) < \infty$, the integral conditionally on V is simply a Gaussian variable

$$\int_D \varphi_i(\mathbf{s})M(\,\mathrm{d}\mathbf{s})|V \sim \mathsf{N}\left(0, \int_D \varphi_i^2(\mathbf{s})V(\,\mathrm{d}\mathbf{s})\right).$$

Going back to the case of Laplace noise. If M is a Laplace field corresponding to the Laplace measure Λ, the variance process is a gamma process, $\Gamma(\mathbf{s})$, so by the argument above one has that the right hand side of (10) under the finite element approximation and conditionally on the gamma process is $\mathsf{N}(\tilde{\mathbf{m}}, \tilde{\mathbf{\Sigma}})$, where the elements of $\tilde{\mathbf{m}}$ and $\tilde{\mathbf{\Sigma}}$ are given by

$$\tilde{\mathbf{\Sigma}}_{ij} = \mathsf{C}\left(\int_D \varphi_i(\mathbf{s})\Lambda(\,\mathrm{d}\mathbf{s})\,, \int_D \varphi_j(\mathbf{s})\Lambda(\,\mathrm{d}\mathbf{s})\middle| \Gamma\right) = \int_D \varphi_i(\mathbf{s})\varphi_j(\mathbf{s})\Gamma(\,\mathrm{d}\mathbf{s}),$$
$$\tilde{\mathbf{m}}_i = \mathsf{E}\left(\int_D \varphi_i(\mathbf{s})\Lambda(\,\mathrm{d}\mathbf{s})\middle| \Gamma\right) = \gamma\int_D \varphi_i(\mathbf{s})\,\mathrm{d}\mathbf{s} + \int_D \varphi_i(\mathbf{s})\Gamma(\,\mathrm{d}\mathbf{s}).$$

Given this, the weights $\mathbf{w}$ can be calculated conditionally on the gamma process, $\Gamma(\mathbf{s})$, as

$$\mathbf{w}|\Gamma \sim \mathsf{N}\left(\mathbf{K}_\alpha^{-1}\tilde{\mathbf{m}}, \mathbf{K}_\alpha^{-1}\tilde{\mathbf{\Sigma}}\mathbf{K}_\alpha^{-1}\right), \tag{17}$$

where $\mathbf{K}_\alpha$ is defined recursively as in the Gaussian case.

It would seem as one has not gained much by using the conditional representation since the conditional mean and covariances, $\tilde{\mathbf{m}}_i$ and $\tilde{\mathbf{\Sigma}}_{ij}$, do not have any simple distributions. One way of approximating them is to approximate the integrals with respect to the Gamma process using the right hand side of (15) with a finite number of terms. However, by using compactly supported linear basis functions, one can simplify things further. Thus, now assume that the basis functions are piecewise linear functions induced by some triangulation of the domain. One can then perform the same Markov approximation as in the Gaussian case. This results in an approximation of the right-hand side of (10) conditionally on the gamma process distributed as $\mathsf{N}(\mathbf{m}, \mathbf{\Sigma})$ with $\mathbf{m} = \gamma\tau\mathbf{a} + \mu\mathbf{\Gamma}$, and $\mathbf{\Sigma} = \mathrm{diag}(\mathbf{\Gamma})$. Here, the gamma variables $\mathbf{\Gamma}_i \sim \Gamma(\tau\mathbf{a}_i, 1)$ are independent and $\mathbf{a}_i = \int \varphi_i(\mathbf{s})\,\mathrm{d}\mathbf{s}$, and these can be calculated without numerically estimating the integrals with respect to the gamma process.

Bolin and Lindgren (2009) studies how this approximation affects the resulting covariance function of the process in the Gaussian case, and it is shown that the error is small if the approximation is used for piecewise linear basis functions. Although additional studies are needed in the non-Gaussian case, the results are likely similar so that the simplification has no large impact on the approximation. Figures 2-4 show that the approximation is accurate in one and two dimensions as explained in Section 5.

4.2 The solution for general $\alpha > d/2$

If one could approximate the solution to (8) for $\alpha = 1$, the recursive scheme discussed above could be used to represent the solutions for all positive odd α. In the Gaussian case, Lindgren et al. (2011) use a least-squares method where the test functions are chosen as $\psi_i = (\kappa^2 - \Delta)^{\frac{1}{2}}\varphi_i$. The left-hand side of (10) can then be expressed as $\mathbf{Kw}$ and the right-hand side is a mean zero Gaussian variable with covariance matrix $\mathbf{K}$. This follows from Lemma 2 in Lindgren et al. (2011), which shows that the covariance between element i and element j on the right-hand side can be written as

$$\mathbf{\Sigma}_{ij} = \left\langle (\kappa^2 - \Delta)^{\frac{1}{2}}\varphi_i,\, (\kappa^2 - \Delta)^{\frac{1}{2}}\varphi_j \right\rangle = \left\langle (\kappa^2 - \Delta)\varphi_i,\, \varphi_j \right\rangle = \mathbf{K}_{ij}.$$

The stochastic weights therefore form a GMRF $\mathbf{w} \sim \mathsf{N}(\mathbf{0}, \mathbf{K}^{-1})$. This argument is unfortunately not applicable in the non-Gaussian case as the covariance between the elements given the gamma process $\Gamma(\mathbf{s})$ is

$$\begin{aligned}
\mathbf{\Sigma}_{ij} &= \mathsf{C}\left(\int_D (\kappa^2 - \Delta)^{\frac{1}{2}}\varphi_i(\mathbf{s})\Lambda(\,\mathrm{d}\mathbf{s}), \int_D (\kappa^2 - \Delta)^{\frac{1}{2}}\varphi_j(\mathbf{s})\Lambda(\,\mathrm{d}\mathbf{s}) \,\middle|\, \Gamma \right) \\
&= \int_D \left((\kappa^2 - \Delta)^{\frac{1}{2}}\varphi_i(\mathbf{s}) \right) \left((\kappa^2 - \Delta)^{\frac{1}{2}}\varphi_j(\mathbf{s}) \right) \Gamma(\,\mathrm{d}\mathbf{s}) \\
&\neq \int_D \left((\kappa^2 - \Delta)\varphi_i(\mathbf{s}) \right) \varphi_j(\mathbf{s})\Gamma(\,\mathrm{d}\mathbf{s}).
\end{aligned}$$

We have not been able to find an easy way of evaluating $\mathbf{\Sigma}_{ij}$ in the non-Gaussian case, and it seems as this least-squares procedure is not extendable to the non-Gaussian case. However, if one instead uses $\psi_i = \varphi_i$, the right-hand side of (10) conditionally on the variance process is $\mathsf{N}(\mathbf{m}, \mathbf{\Sigma})$, as in the case $\alpha = 2$. With this as a starting point, one can use a finite element matrix transfer technique (FE-MTT) to obtain a discretized approximation of the solution. Simpson (2008)

studied such methods for sampling generalized Matérn fields on locally planar Riemannian manifolds, and argued that one could sample the stochastic weights for a general α using the matrix transfer equation $(\mathbf{C}^{-1}\mathbf{K})^{\alpha/2}\mathbf{w} \sim \mathsf{N}(0, \mathbf{C}^{-1})$. To simplify the notations in later sections, denote $\mathbf{K}_\alpha = (\mathbf{C}^{-1}\mathbf{K})^{\alpha/2}$ and note that we now have changed the definition of $\mathbf{K}_\alpha$ from the one that was used for even α. The weights $\mathbf{w}$ are then mean zero Gaussian with a precision matrix $\mathbf{Q}_\alpha = \mathbf{K}_\alpha\mathbf{C}^{-1}\mathbf{K}_\alpha$. In the case $\alpha = 2$, this discretization coincides with the approximation described above, but it can be used for any $\alpha > d/2$.

Now in the non-Gaussian case, the results from the case $\alpha = 2$ can be used directly to get a right-hand side that is Gaussian with mean $\mathbf{m}$ and covariance $\mathbf{\Sigma}$ conditionally on the variance process. As in the Gaussian case, this should be multiplied with $\mathbf{C}^{-1}$ to get consistency in the FE-MTT procedure. Hence, in the case of Laplace noise the weights are given by

$$\mathbf{w}|\Gamma \sim \mathsf{N}\left(\mathbf{K}_\alpha^{-1}\mathbf{C}^{-1}\mathbf{m}, \mathbf{K}_\alpha^{-1}\mathbf{C}^{-1}\mathbf{\Sigma}\mathbf{C}^{-1}\mathbf{K}_\alpha^{-1}\right). \tag{18}$$

Again, for the case $\alpha = 2$, this coincides with the procedure described in the section above, and because of this we will from now on use this FE-MTT procedure for all $\alpha > d/2$. Consistency of the FE-MTT procedure follows from similar arguments as in Simpson (2008). These arguments do not provide a rate of convergence as the number of basis functions are increased, and as for the Gaussian case, the rate of convergence and the numerical properties of the approximation are strongly dependent on α.

5 Sampling from the model

Using the finite element representation obtained in the previous section it is easy to generate samples from the SPDE (8). Assume that we want sample the model at locations $\mathbf{s} = (\mathbf{s}_1, \ldots, \mathbf{s}_n)$, and let $\mathbf{\Phi}$ be a matrix with elements $\mathbf{\Phi}_{ij} = \varphi_j(\mathbf{s}_i)$. Samples can now be generated using the following three-step algorithm.

Algorithm 5.1. *Sampling the Laplace driven SPDE* (8).

1. *Generate two independent random vectors* $\mathbf{\Gamma}$ *and* $\mathbf{Z}$*, where* $\mathbf{\Gamma}_i \sim \Gamma(\tau\mathbf{a}_i, 1)$ *and* $\mathbf{Z}_i \sim \mathsf{N}(0, 1)$.
2. *Let* $\mathbf{\Lambda} = \gamma\tau\mathbf{a} + \mu\mathbf{\Gamma} + \mathrm{diag}(\sqrt{\mathbf{\Gamma}})\mathbf{Z}$ *and calculate* $\mathbf{w} = \mathbf{C}^{-1}\mathbf{\Lambda}$.
3. $\mathbf{X} = \mathbf{\Phi}\mathbf{K}_\alpha^{-1}\mathbf{w}$ *is now a sample of the random field at the locations* $\mathbf{s}$.

The last step could potentionally be computationally expensive for large simulations. However, if α is even, one can take advantage of the sparsity of $\mathbf{K}_\alpha$ and solve the equation system $\mathbf{v} = \mathbf{K}_\alpha^{-1}\mathbf{w}$ efficiently without calculating the inverse by using Cholesky factorization and back substitution as suggested by Rue and Held (2005). For other $\alpha > d/2$, $\mathbf{K}_\alpha$ is not sparse and the Cholesky method will not improve the computational efficiency. However, as Simpson (2008) shows, one can instead use Krylov subspace methods in the calculations to obtain efficient sampling schemes. The basic problem for general α is to solve the matrix equation $\mathbf{v} = (\mathbf{C}^{-1}\mathbf{K})^{-\frac{\alpha}{2}}\mathbf{w}$, and there are a number of methods with different computational properties that can be used. In this work we use the method by Hale et al. (2008), which is based on combining contour integrals evaluated by the periodic trapezoid rule with conformal maps involving Jacobi elliptic functions.

In Figure 2, a simulation of a process on $\mathbb{R}$ with parameters $\mu = \gamma = \sigma = 1$, $\tau = 2$, $\kappa = 15$, and $\alpha = 2$ is shown. Since $\mathbf{K}_\alpha$ is sparse in this case, the Cholesky method is used for the simulation. In the upper left panel, a histogram of the samples from 1000 simulations is shown together with the theoretical density, calculated using numerical Fourier inversion of the characteristic function (13). In the upper right panel, the empirical covariance function of the samples is shown together with the theoretical Matérn covariance function. Two more examples of densities and covariance functions for different parameter settings are shown in Figure 3. In the upper panels, we have $\alpha = 1$, which results in an exponential covariance function. The other parameters are $\mu = \gamma = 0$, $\sigma = 1$, and $\tau = \kappa = 10$, which results in a symmetric distribution. In the lower panels, we have $\alpha = 3.5$ which results in a smoother field. The other parameters are $\mu = \sigma = 0.1$, $\gamma = 0$, $\tau = 10$, and $\kappa = 20$, which results in an asymmetric distribution. In both cases in Figure 3, the Krylov subspace method is used for the simulations.

In Figure 4 and Figure 5, two simulations of fields on $\mathbb{R}^2$ are shown together with the corresponding covariance functions, densities, and empirically estimated versions based on 1000 simulations each. As seen in the figures for all five examples, there is a close agreement between the histograms and the true densities, and between the true covariance functions and the empirically estimated covariance functions for all these parameter settings, indicating that the approximation procedure works as intended. A more detailed analysis of the simulation procedure is outside the scope of this article, but it should be noted that the SPDE approximation using piecewise linear basis functions does not provide convergence of higher-order derivatives, and the simulation procedure is therefore not appropriate for applications where such properties are important.

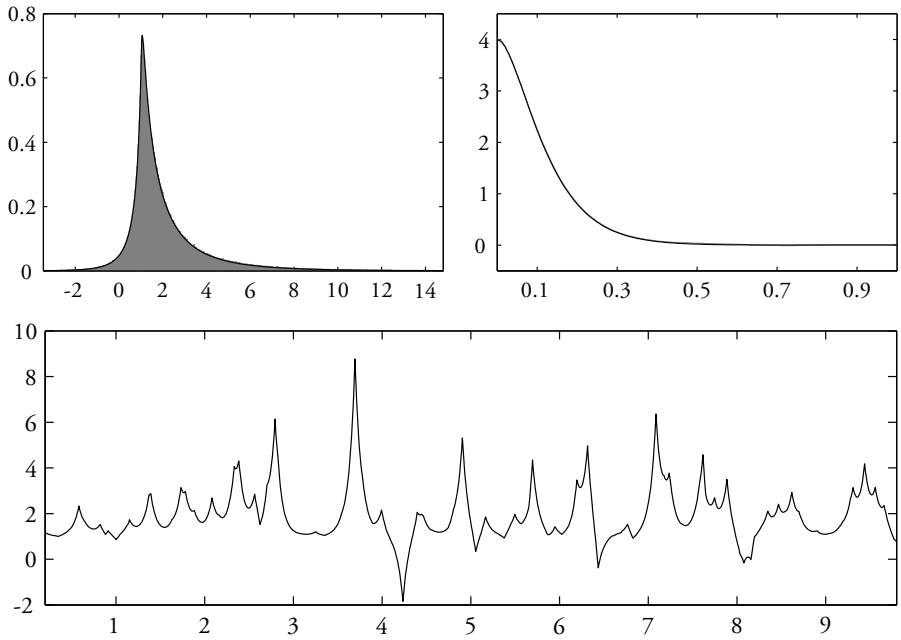

Figure 2: The lower panel shows a simulation of the Laplace driven SPDE (8) on $\mathbb{R}$ with parameters $\mu = \gamma = \sigma = 1$, $\tau = 2$, $\kappa = 15$, and $\alpha = 2$. The upper left panel shows a histogram of the samples from 1000 simulations together with the true density. The upper right panel shows the empirical covariance function for the samples (grey curve) together with the true Matérn covariance function (black curve). It is difficult to see the grey curve since the two curves are very similar.

6 Parameter estimation

Parameter estimation for Laplace moving average models is not easy since there is no closed form expression for the parameter likelihood. Recently, Podgórski and Wegener (2011) derived a method of moments-based estimation procedure for these types of models. In their method, the convolution kernel is first estimated from the spectral density of the data, and given the estimated kernel, the parameters in the Laplace distribution are estimated by fitting the theoretical moments of the Laplace distribution to the sample moments. The method is quite simple although some special care has to be taken to handle the cases when the method of

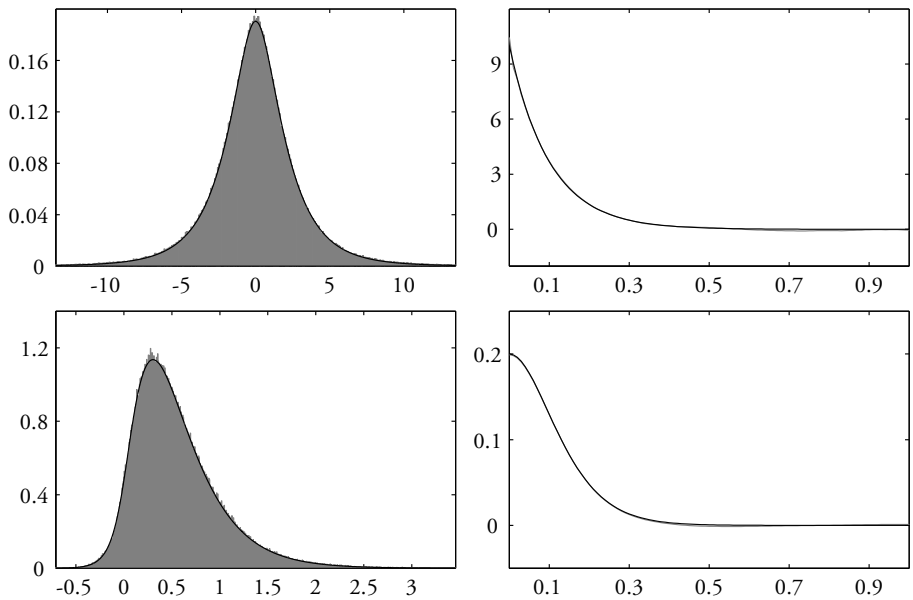

Figure 3: Simulation results as in Figure 2 with different parameters. The top row shows a symmetric case with parameters $\mu = \gamma = 0$, $\sigma = 1$, $\kappa = \tau = 10$, and $\alpha = 1$. The bottom row shows an asymmetric case with parameters $\mu = \sigma = 0.1$, $\gamma = 0$, $\kappa = 20$, $\tau = 10$ and $\alpha = 3.5$.

moments equation system does not have a solution, which can happen for certain values of the sample skewness and excess kurtosis.

Using the SPDE formulation, parameter estimation can instead be performed in a likelihood framework. One of the advantages with this is that maximum likelihood parameter estimates always are in the allowed parameter space. Another advantage is that the estimates will account for all relevant information in the data, which might not be the case for method of moment estimates.

To be able to estimate the parameters in a maximum likelihood framework, the problem is interpreted as a missing data problem which facilitates use of the Expectation Maximization (EM) algorithm (Dempster et al., 1977). The proposed EM algorithm is based on the same ideas as the ones in Lange et al. (1989) and Protassov (2004) which looked at EM estimation in the case of iid observations of certain Gaussian mixtures. Our main contribution is the extension of

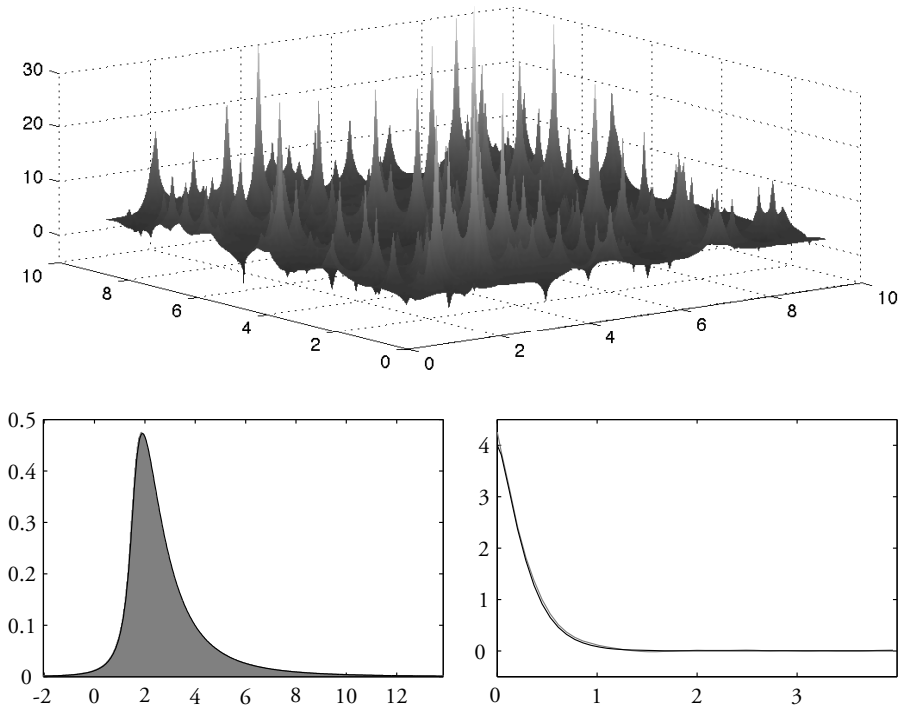

Figure 4: A simulation of an asymmetric model (8) in $\mathbb{R}^2$ where the parameters are $\kappa = 5$, $\sigma = \mu = \gamma = 1$, $\tau = 2$, and $\alpha = 2$. The covariance functions and densities for these fields can be seen in the second row. The empirically estimated versions are based on 1000 simulations.

these ideas to the random field setting.

Assume we have measurements $\mathbf{X}$ of the process $X(\mathbf{s})$ taken at some locations and that the Hilbert space approximation procedure is used with a basis obtained by triangulating the measurement locations. In this case, the matrix $\mathbf{\Phi}$ is diagonal and conditioning on the measurements and the parameters is equivalent to conditioning on $\mathbf{\Lambda}$ and the parameters as there is a one-to-one correspondence between the two through $\mathbf{\Lambda} = \mathbf{C}\mathbf{K}_\alpha\mathbf{\Phi}^{-1}\mathbf{X}$, see Algorithm 5.1. To obtain simpler updating expressions, we first make a change of variables by introducing the parameter $\bar{\gamma} = \gamma\tau$ and estimate this parameter instead of γ. As for Gaussian Matérn models, the shape parameter ν is difficult to estimate accurately and it is therefore assumed to be known throughout this section and no attempt is made at estimating it.

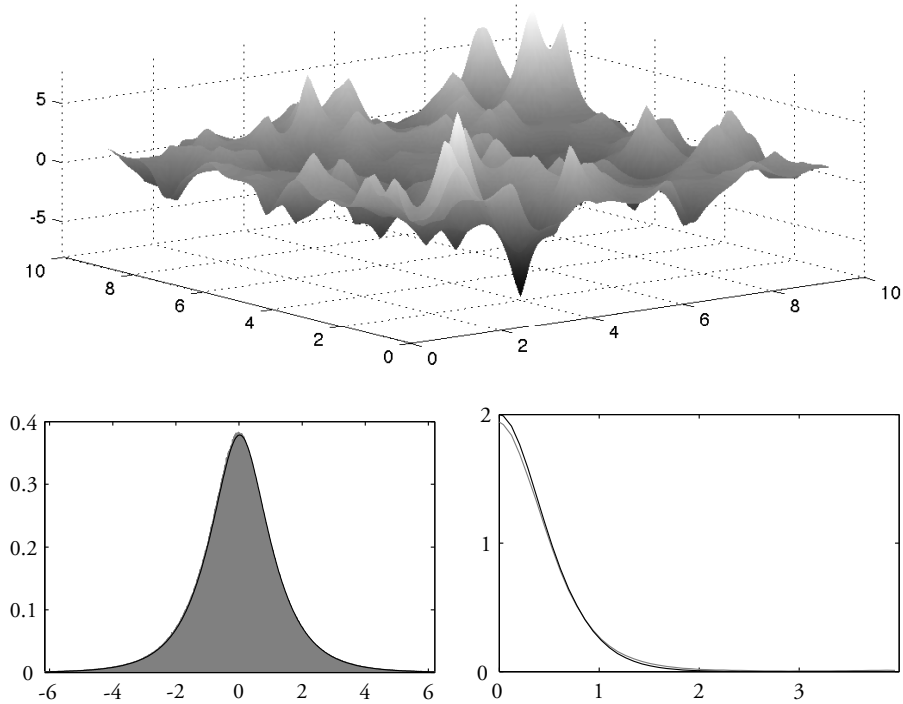

Figure 5: A simulation of a symmetric model (8) in $\mathbb{R}^2$ with parameters $\kappa = 5$, $\sigma = 1$, $\mu = \gamma = 0$, $\tau = 2$, and $\alpha = 4$. The covariance function and density are shown in the second row. The empirically estimated versions are based on 1000 simulations.

Augmenting the data with the unknown (missing) gamma variables, the augmented likelihood is $L(\boldsymbol{\theta}|\mathbf{X}, \boldsymbol{\Gamma}) = \pi(\mathbf{X}|\boldsymbol{\Gamma}, \boldsymbol{\theta})\pi(\boldsymbol{\Gamma}|\boldsymbol{\theta})$, and the loss-function that is needed for the EM-procedure is

$$\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(j)}) = \mathsf{E}\left(\log L(\boldsymbol{\theta}|\mathbf{X}, \boldsymbol{\Gamma})\middle|\mathbf{X}, \boldsymbol{\theta}^{(j)}\right),$$

where $\boldsymbol{\theta}^{(j)}$ is an estimate of $\boldsymbol{\theta} = \left(\kappa, \sigma, \mu, \bar{\gamma}, \tau\right)$ at iteration j, and the expectation is taken according to the distribution of $\boldsymbol{\Gamma}$ given $\mathbf{X}$. We have $\mathbf{X}|\boldsymbol{\Gamma}, \boldsymbol{\theta} \sim \mathsf{N}(\mathbf{m}, \sigma^2\boldsymbol{\Sigma})$, where $\mathbf{m} = \boldsymbol{\Phi}\mathbf{K}_\alpha^{-1}\mathbf{C}(\bar{\gamma}\mathbf{a} + \mu\boldsymbol{\Gamma})$, $\boldsymbol{\Sigma} = \boldsymbol{\Phi}\mathbf{K}_\alpha^{-1}\mathbf{C}\mathbf{D}_\Gamma\mathbf{C}\mathbf{K}_\alpha^{-1}\boldsymbol{\Phi}$, and $\mathbf{D}_\Gamma$ is the diagonal matrix with the vector $\boldsymbol{\Gamma}$ on the main diagonal. The second part of the augmented likelihood can be written as $\pi(\boldsymbol{\Gamma}|\boldsymbol{\theta}) = \prod \pi(\boldsymbol{\Gamma}_i|\boldsymbol{\theta})$ since the compon-

ents in $\mathbf{\Gamma}$ are independent gamma variables $\mathbf{\Gamma}_i$ with shape parameters $\tau\mathbf{a}_i$ and scale one, where $\mathbf{a}_i$ are known constants depending on the basis used. The log-likelihood is

$$\log L(\boldsymbol{\theta}|\mathbf{X},\mathbf{\Gamma}) = -\, n\log(\sigma) + \log(|\mathbf{K}_\alpha|) - \frac{1}{2\sigma^2}(\mathbf{X}-\mathbf{m})^\top \mathbf{\Sigma}^{-1}(\mathbf{X}-\mathbf{m}) + \sum_{i=1}^{n}\left(\tau\mathbf{a}_i \log \mathbf{\Gamma}_i - \log \Gamma(\tau\mathbf{a}_i)\right) + C,$$

where the constant C does not depend on the unknown parameters. Thus, using the relation between $\mathbf{X}$ and $\mathbf{\Lambda}$, the loss-function is

$$\begin{aligned}\mathcal{Q}(\boldsymbol{\theta},\boldsymbol{\theta}^{(j)}) = &-\, n\log(\sigma) + \log(|\mathbf{K}_\alpha|) - \frac{1}{2\sigma^2}\Big((\mathbf{\Lambda}-\bar{\gamma}\mathbf{a})^\top \mathbf{D}_{\mathsf{E}(\mathbf{\Gamma}^{-1}|\star)}(\mathbf{\Lambda}-\bar{\gamma}\mathbf{a}) \\ &+ \mu^2\mathbf{1}^\top \mathsf{E}(\mathbf{\Gamma}|\star) + 2\bar{\gamma}\mu\mathbf{a}^\top\mathbf{1} - 2\mu\mathbf{\Lambda}^\top\mathbf{1}\Big) \\ &+ \sum_{i=1}^{n}\left(\tau\mathbf{a}_i\mathsf{E}(\log\mathbf{\Gamma}_i|\star) - \log\Gamma(\tau\mathbf{a}_i)\right) + C,\end{aligned}$$

where $\mathsf{E}(\cdot|\star)$ denotes $\mathsf{E}(\cdot|\theta^{(j)},\mathbf{X})$. The expectations needed to evaluate the loss-function are $\mathsf{E}(\mathbf{\Gamma}|\star)$, $\mathsf{E}(\mathbf{\Gamma}^{-1}|\star)$, and $\mathsf{E}(\log\mathbf{\Gamma}_i|\star)$. To calculate these, first note that (see Gradshteyn and Ryzhik, 2000, formula 3.472.9)

$$\mathcal{I}(a,b,c) = \int_0^\infty x^{a-1}e^{-\frac{b}{x}-cx}\,\mathrm{d}x = 2\left(\frac{b}{c}\right)^{\frac{a}{2}} K_a\left(2\sqrt{bc}\right). \tag{19}$$

Using this expression, the expectation $\mathsf{E}(\mathbf{\Gamma}_i|\star)$ can be written as

$$\begin{aligned}\mathsf{E}(\mathbf{\Gamma}_i|\star) &= \int \mathbf{\Gamma}_i\pi(\mathbf{\Gamma}_i|\mathbf{X},\boldsymbol{\theta})\,\mathrm{d}\mathbf{\Gamma}_i = \frac{\int \mathbf{\Gamma}_i\pi(\mathbf{X}|\mathbf{\Gamma}_i,\boldsymbol{\theta})\pi(\mathbf{\Gamma}_i|\boldsymbol{\theta})\,\mathrm{d}\mathbf{\Gamma}_i}{\pi(\mathbf{X}|\boldsymbol{\theta})} \\ &= \frac{\int \mathbf{\Gamma}_i\pi(\mathbf{X}|\mathbf{\Gamma}_i,\boldsymbol{\theta})\pi(\mathbf{\Gamma}_i|\boldsymbol{\theta})\,\mathrm{d}\mathbf{\Gamma}_i}{\int \pi(\mathbf{X}|\mathbf{\Gamma}_i,\boldsymbol{\theta})\pi(\mathbf{\Gamma}_i|\boldsymbol{\theta})\,\mathrm{d}\mathbf{\Gamma}_i} = \frac{\mathcal{I}\left(\tau\mathbf{a}_i+\frac{1}{2},\frac{(\mathbf{\Lambda}_i-\bar{\gamma}\mathbf{a}_i)^2}{2\sigma^2},1+\frac{\mu^2}{2\sigma^2}\right)}{\mathcal{I}\left(\tau\mathbf{a}_i-\frac{1}{2},\frac{(\mathbf{\Lambda}_i-\gamma\tau\mathbf{a}_i)^2}{2\sigma^2},1+\frac{\mu^2}{2\sigma^2}\right)} \\ &= \frac{|\mathbf{\Lambda}_i-\bar{\gamma}\mathbf{a}_i|}{\sqrt{2\sigma^2+\mu^2}}\frac{K_{\tau\mathbf{a}_i+\frac{1}{2}}\left(\sigma^{-2}|\mathbf{\Lambda}_i-\bar{\gamma}\mathbf{a}_i|\sqrt{2\sigma^2+\mu^2}\right)}{K_{\tau\mathbf{a}_i-\frac{1}{2}}\left(\sigma^{-2}|\mathbf{\Lambda}_i-\bar{\gamma}\mathbf{a}_i|\sqrt{2\sigma^2+\mu^2}\right)}.\end{aligned}$$

If the argument in the Bessel functions is very small or very large one might get numerical problems when evaluating this expression depending on how it is

implemented. In the case of small arguments, one can use the following approximation to improve the numerical stability

$$K_a(x) \approx \frac{\Gamma(|a|)}{2}\left(\frac{2}{x}\right)^{|a|}, \text{ if } a \neq 0 \text{ and } x \ll \sqrt{|a|+1}.$$

The expectation $\mathsf{E}(\mathbf{\Gamma}_i|\star)$ then simplifies to

$$\mathsf{E}(\mathbf{\Gamma}_i|\star) \approx \begin{cases} \left(\tau\mathbf{a}_i - \frac{1}{2}\right)\frac{2\sigma^2}{2\sigma^2+\mu^2}, & \tau > \frac{1}{2\mathbf{a}_i}, \\ \frac{\Gamma\left(\tau\mathbf{a}_i+\frac{1}{2}\right)}{\Gamma\left(\frac{1}{2}-\tau\mathbf{a}_i\right)}\frac{\left(2\sigma^2\right)^{2\tau\mathbf{a}_i}}{\left(2\sigma^2+\mu^2\right)^{\frac{2\tau\mathbf{a}_i+1}{2}}}|\mathbf{\Lambda}_i - \bar{\gamma}\mathbf{a}_i|^{1-2\tau\mathbf{a}_i}, & \tau < \frac{1}{2\mathbf{a}_i}. \end{cases}$$

In the case of large arguments, one can instead use the approximation

$$\frac{K_a(x)}{K_{a-1}(x)} \approx 1 + \left(a - \frac{1}{2}\right)\frac{1}{x},$$

which gives the following approximation for $\mathsf{E}(\mathbf{\Gamma}_i|\star)$

$$\mathsf{E}(\mathbf{\Gamma}_i|\star) \approx \frac{|\mathbf{\Lambda}_i - \bar{\gamma}\mathbf{a}_i|}{\sqrt{2\sigma^2+\mu^2}} + \frac{\tau\mathbf{a}_i\sigma^2}{2\sigma^2+\mu^2}.$$

The expectation $\mathsf{E}(\mathbf{\Gamma}_i^{-1}|\star)$ is calculated similarly using (19) and can be written as

$$\mathsf{E}(\mathbf{\Gamma}_i^{-1}|\star) = \frac{\sqrt{2\sigma^2+\mu^2}}{|\mathbf{\Lambda}_i - \bar{\gamma}\mathbf{a}_i|}\frac{K_{\tau\mathbf{a}_i-\frac{3}{2}}\left(\sigma^{-2}|\mathbf{\Lambda}_i - \bar{\gamma}\mathbf{a}_i|\sqrt{2\sigma^2+\mu^2}\right)}{K_{\tau\mathbf{a}_i-\frac{1}{2}}\left(\sigma^{-2}|\mathbf{\Lambda}_i - \bar{\gamma}\mathbf{a}_i|\sqrt{2\sigma^2+\mu^2}\right)}. \tag{20}$$

Evaluating modified Bessel functions numerically is computationally expensive and should therefore be avoided as much as possible when implementing the estimation procedure. To that end, one can express $K_{\tau\mathbf{a}_i-\frac{3}{2}}(\cdot)$ using the following recurrence relationship for modified Bessel functions

$$K_a(x) = K_{a+2}(x) - \frac{2(a+1)}{x}K_{a+1}(x),$$

giving the following expression for $\mathsf{E}(\mathbf{\Gamma}_i^{-1}|\star)$ in terms of $\mathsf{E}(\mathbf{\Gamma}_i|\star)$

$$\mathsf{E}(\mathbf{\Gamma}_i^{-1}|\star) = \frac{(\mu^2+2\sigma^2)\mathsf{E}(\mathbf{\Gamma}_i|\star) - \sigma^2(2\tau\mathbf{a}_i - 1)}{(\mathbf{\Lambda}_i - \bar{\gamma}\mathbf{a}_i)^2}.$$

Using this expression instead of (20), one only have to evaluate two modified Bessel functions instead of three.

Finally, the expectation $\mathsf{E}(\log(\boldsymbol{\Gamma}_i)|\star)$ is similarly written as

$$\mathsf{E}(\log(\boldsymbol{\Gamma}_i)|\star) = \frac{\int \log(\boldsymbol{\Gamma}_i)\pi(\mathbf{X}|\boldsymbol{\Gamma}_i, \boldsymbol{\theta})\pi(\boldsymbol{\Gamma}_i|\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\Gamma}_i}{\pi(\mathbf{X}|\boldsymbol{\theta})}.$$

The denominator is the same as in the previous expectations, while calculating the nominator requires evaluating an integral on the form

$$\mathcal{I}_{log}(a, b, c) = \int_0^\infty \log(x)x^{a-1}\exp\left(-\frac{b}{x} - cx\right)\,\mathrm{d}x. \tag{21}$$

To calculate this integral, we differentiate (19) with respect to a and obtain

$$\begin{aligned}\mathcal{I}_{log}(a, b, c) &= \frac{\partial}{\partial a}\int_0^\infty x^{a-1}e^{-\frac{b}{x}-cx}\,\mathrm{d}x = \frac{\partial}{\partial a}\left(2\left(\frac{b}{c}\right)^{\frac{a}{2}}K_a\left(2\sqrt{bc}\right)\right)\\ &= 2\left(\frac{b}{c}\right)^{\frac{a}{2}}\left(\log\left(\frac{b}{c}\right)K_a\left(2\sqrt{bc}\right) + \frac{\partial}{\partial a}K_a\left(2\sqrt{bc}\right)\right).\end{aligned}$$

The derivative of $K_a(2\sqrt{bc})$ with respect to a can be expressed using infinite sums of gamma- and polygamma functions; however, in this case it is easier to numerically approximate the derivative using for example forward differences:

$$\frac{\partial}{\partial a}K_a\left(2\sqrt{bc}\right) \approx \frac{K_{a+\varepsilon}\left(2\sqrt{bc}\right) - K_a\left(2\sqrt{bc}\right)}{\varepsilon}.$$

Using this expression, we approximate $\mathsf{E}(\log(\Gamma_i)|\star)$ as

$$\begin{aligned}\mathsf{E}(\log(\boldsymbol{\Gamma}_i)|\star) =& \frac{\mathcal{I}_{log}\left(\tau\mathbf{a}_i - \frac{1}{2}, \frac{(\boldsymbol{\Lambda}_i - \bar{\gamma}\mathbf{a}_i)^2}{2\sigma^2}, 1 + \frac{\mu^2}{2\sigma^2}\right)}{\mathcal{I}\left(\tau\mathbf{a}_i - \frac{1}{2}, \frac{(\boldsymbol{\Lambda}_i - \bar{\gamma}\mathbf{a}_i)^2}{2\sigma^2}, 1 + \frac{\mu^2}{2\sigma^2}\right)}\\ \approx& \log\left(\frac{|\boldsymbol{\Lambda}_i - \bar{\gamma}\mathbf{a}_i|}{\sqrt{\mu^2 + 2\sigma^2}}\right) - \frac{1}{\varepsilon}\\ &+ \frac{1}{\varepsilon}\frac{K_{\tau\mathbf{a}_i - \frac{1}{2} + \varepsilon}\left(\sigma^{-2}|\boldsymbol{\Lambda}_i - \bar{\gamma}\mathbf{a}_i|\sqrt{2\sigma^2 + \mu^2}\right)}{K_{\tau\mathbf{a}_i - \frac{1}{2}}\left(\sigma^{-2}|\boldsymbol{\Lambda}_i - \bar{\gamma}\mathbf{a}_i|\sqrt{2\sigma^2 + \mu^2}\right)}.\end{aligned}$$

To obtain the updating equations for the parameters, the loss-function should be maximized with respect to each parameter, for example by differentiating it with respect to the parameters and setting the derivatives equal to zero. Since the system of equations obtained from this procedure is not analytically solvable, one would have to iterate numerically in each step to obtain the parameter updates if the EM algorithm is used without modifications. A better alternative is to use an Expectation Conditional Maximization (ECM) algorithm (Meng and Rubin, 1993) where the M-step is divided into two conditional maximization steps. In the first step, the parameters of the Laplace noise is updated conditionally on the current value of κ, and in the second step κ is updated conditionally on the other parameters. Differentiating the loss-function with respect to μ, $\bar{\gamma}$, and σ and setting the derivatives equal to zero yields the following updating rules

$$
\begin{aligned}
\mu^{(j+1)} &= \frac{(\boldsymbol{\Lambda}^\top \mathbf{1})(\mathbf{a}^\top \mathbf{D}_{\mathsf{E}(\boldsymbol{\Gamma}^{-1}|\star)}\mathbf{a}) - (\mathbf{a}^\top \mathbf{1})(\boldsymbol{\Lambda}^\top \mathbf{D}_{\mathsf{E}(\boldsymbol{\Gamma}^{-1}|\star)}\mathbf{a})}{(\mathbf{1}^\top \mathsf{E}(\boldsymbol{\Gamma}|\star))(\mathbf{a}^\top \mathbf{D}_{\mathsf{E}(\boldsymbol{\Gamma}^{-1}|\star)}\mathbf{a}) - (\mathbf{1}^\top \mathbf{a})^2}, \\
\bar{\gamma}^{(j+1)} &= \frac{(\mathbf{1}^\top \mathsf{E}(\boldsymbol{\Gamma}|\star))(\boldsymbol{\Lambda}^\top \mathbf{D}_{\mathsf{E}(\boldsymbol{\Gamma}^{-1}|\star)}\mathbf{a}) - (\boldsymbol{\Lambda}^\top \mathbf{1})(\mathbf{a}^\top \mathbf{1})}{(\mathbf{1}^\top \mathsf{E}(\boldsymbol{\Gamma}|\star))(\mathbf{a}^\top \mathbf{D}_{\mathsf{E}(\boldsymbol{\Gamma}^{-1}|\star)}\mathbf{a}) - (\mathbf{1}^\top \mathbf{a})^2}, \\
\sigma^{(j+1)} &= \frac{1}{\sqrt{n}} \left(\boldsymbol{\Lambda}^\top \mathbf{D}_{\mathsf{E}(\boldsymbol{\Gamma}^{-1}|\star)}\boldsymbol{\Lambda} + 2\frac{(\boldsymbol{\Lambda}^\top \mathbf{D}_{\mathsf{E}(\boldsymbol{\Gamma}^{-1}|\star)}\mathbf{a})(\boldsymbol{\Lambda}^\top \mathbf{1})(\mathbf{1}^\top \mathbf{a})}{(\mathbf{1}^\top \mathsf{E}(\boldsymbol{\Gamma}|\star))(\mathbf{a}^\top \mathbf{D}_{\mathsf{E}(\boldsymbol{\Gamma}^{-1}|\star)}\mathbf{a}) - (\mathbf{1}^\top \mathbf{a})^2} \right. \\
&\quad \left. - \frac{(\boldsymbol{\Lambda}^\top \mathbf{D}_{\mathsf{E}(\boldsymbol{\Gamma}^{-1}|\star)}\mathbf{a})^2(\mathbf{1}^\top \mathsf{E}(\boldsymbol{\Gamma}|\star)) + (\mathbf{a}^\top \mathbf{D}_{\mathsf{E}(\boldsymbol{\Gamma}^{-1}|\star)}\mathbf{a})(\boldsymbol{\Lambda}^\top \mathbf{1})^2}{(\mathbf{1}^\top \mathsf{E}(\boldsymbol{\Gamma}|\star))(\mathbf{a}^\top \mathbf{D}_{\mathsf{E}(\boldsymbol{\Gamma}^{-1}|\star)}\mathbf{a}) - (\mathbf{1}^\top \mathbf{a})^2} \right)^{\frac{1}{2}}.
\end{aligned}
$$

In general, there is no closed form expression for the conditional updating equation for τ, so the following equation is maximized numerically to obtain $\tau^{(j+1)}$

$$
\mathcal{Q}_\tau = \sum_{i=1}^{n} \left(\tau \mathbf{a}_i \mathsf{E}(\log \boldsymbol{\Gamma}_i|\star) - \log \Gamma(\tau \mathbf{a}_i) \right).
$$

In the special case when all $\mathbf{a}_i$ are equal to some value a, which for example is the case if a triangulation induced by a regular lattice is used in the Hilbert space approximation, the solution can be written as

$$
\tau^{(j+1)} = \frac{1}{a}\psi^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} \mathsf{E}(\log \boldsymbol{\Gamma}_i|\star) \right),
$$

where $\psi^{-1}(\cdot)$ is the inverse of the digamma function. Finally κ is updated conditionally on the other parameters. There is no closed form expression for the updating equation for κ either, so the following expression is maximized numerically with respect to κ,

$$\begin{aligned}\mathcal{Q}_\kappa = &\log(|\mathbf{K}_\alpha|) - \frac{1}{2(\sigma^{(i+1)})^2}\Big(\mathbf{\Lambda}^\top \mathbf{D}_{\mathsf{E}(\Gamma^{-1}|\star)}\mathbf{\Lambda} \\ &- 2\bar{\gamma}^{(j+1)}\mathbf{\Lambda}^\top \mathbf{D}_{\mathsf{E}(\Gamma^{-1}|\star)}\mathbf{a} - 2\mu^{(j+1)}\mathbf{\Lambda}^\top \mathbf{1}\Big).\end{aligned}$$

By the construction of $\mathbf{K}_\alpha$, its log-determinant can be written as

$$\log(|\mathbf{K}_\alpha|) = \frac{\alpha}{2}\log|\mathbf{C}^{-1}\mathbf{G} + \kappa^2\mathbf{I}| = \frac{\alpha}{2}\sum_{i=1}^{n}\log(\lambda_i + \kappa^2),$$

where λ_i denotes the ith eigenvalue of $\mathbf{C}^{-1}\mathbf{G}$. If the size of $\mathbf{K}_\alpha$ is small, these eigenvalues can be pre-calculated as they do not depend on the parameters. For larger problems is it most efficient to calculate the log-determinant in each iteration using a sparse Cholesky factorization of $\mathbf{K} = \mathbf{G} + \kappa^2\mathbf{C}$.

As shown by Meng and Rubin (1993), the ECM algorithm has the same convergence properties as the ordinary EM algorithm. The likelihood is increasing for each iteration and the convergence is linear. Hence, we do not lose any rate of convergence by using the ECM algorithm instead of the EM algorithm.

7 A simulation study

In this section, a simulation study is performed to test the accuracy of the parameter estimation algorithm presented above. The algorithm is tested for twelve different parameter settings corresponding to marginal distributions shown in Figure 6 for processes in one dimension with $\alpha = 2$. For Matérn covariance functions, one sometimes defines the approximate range as $r = \sqrt{8\nu}\kappa^{-1}$, which is the value where the correlation is approximately 0.1. For the first six test cases, we have $\kappa = 1$ which corresponds to an approximate range of 3.5, and for the last six cases we have $\kappa = 0.1$ which corresponds to an approximate range of 35. For each value of κ, three symmetric distributions and three asymmetric distributions are used. In Figure 6, the distributions for the short range are shown in the two upper panels, and the distributions for the long range are shown in the two bottom panels.

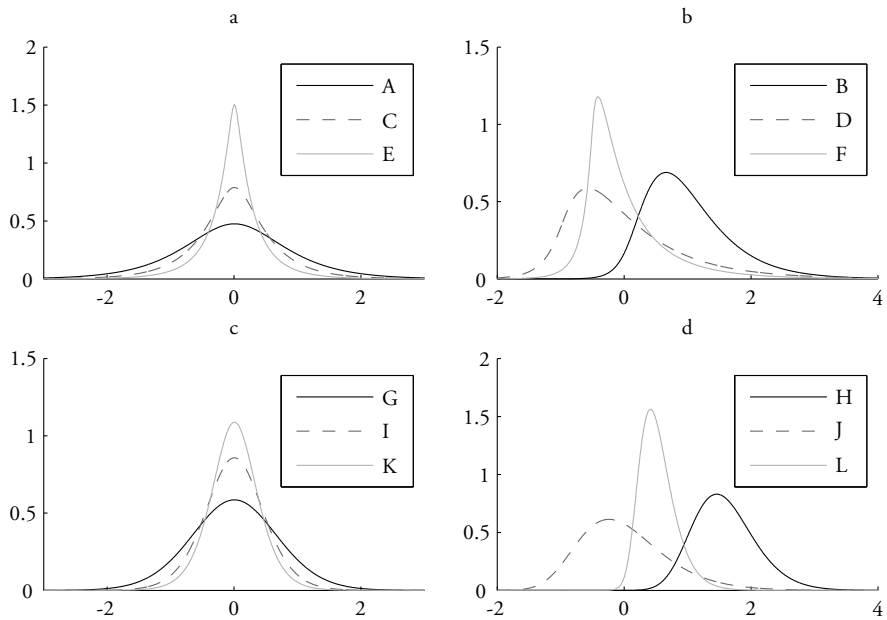

Figure 6: Marginal distributions for the twelve test cases in the simulation study. In Panel a and Panel b, the approximate covariance range is 3.5 and in Panel c and d, the range is 35.

For each set of parameters, 500 data sets are simulated using Algorithm 5.1, where each data set contains 1000 equally spaced observations on $[1, 1000]$. The basis used in the Hilbert space approximations consists of 1000 piecewise linear basis functions centered at $1, 2, \ldots, 1000$. For each data set, the starting value for κ is set to $\sqrt{8\nu}\hat{r}^{-1}$, where $\hat{r}$ is the approximate range for the empirical covariance function for the data set. To obtain good starting values for the other parameters, an initial run of the EM estimator is made with κ fixed to the starting value and where the starting values for μ and γ are drawn independently from a $\mathsf{N}(0, 1)$ distribution, and the starting values for σ and τ^{-1} are drawn from a $\chi^2(1)$-distribution. After 100 steps, this initial run is ended, and the estimates are used as starting values for the full EM-estimator.

In Table 1, the 10% 50%, and 90% percentiles of 500 Monte Carlo samples are shown for each parameter setting, together with the true values of the parameters. One can note that all estimates are more or less unbiased and have fairly

small variances, indicating that the estimation procedure works as intended. The only case where the estimator seems to have a bias is in case L, where most of the estimated values of τ are above the true value. The cause of this bias is probably that the estimation procedure is not very stable for small values of τ because some of the expectations $\mathsf{E}(\boldsymbol{\Gamma}_i^{-1}|\star)$ can be infinite in this case. More precisely, for $\tau < 3/(2\min(\mathbf{a}_1, \ldots, \mathbf{a}_n))$, the likelihood is unbounded for any $\bar{\gamma} = \boldsymbol{\Lambda}_i/\mathbf{a}_i$ and the ML procedure thus has to be modified. To improve the stability of the algorithm, the expectations $\mathsf{E}(\boldsymbol{\Gamma}_i^{-1}|\star)$ are truncated to 1000 in the first iteration, and for each iteration this bound is made larger so that it has little to no effect after a few hundred iterations of the algorithm. This greatly improves the stability for $\tau < 3/(2\min(\mathbf{a}_1, \ldots, \mathbf{a}_n))$, but it is left for future research to justify this modified maximum likelihood procedure theoretically, to derive large sample properties of the estimator, and to investigate other improvements for the case of small values of τ.

It should finally be noted that the parameters are estimated assuming the same finite element approximation as is used for simulating the data. Estimating the model parameters using a different numbers of basis functions in the approximation can possibly give biased estimates, as the parameters are estimated to maximize the likelihood for the approximate model instead of the exact SPDE. The size of this bias depends on the specific parameters of the model, and especially on the true covariance range in relation to the spacing of the basis functions, as discussed in Bolin and Lindgren (2011) in the case of Gaussian models. It is, however, outside the scope of this work to investigate this issue further here.

8 Discussion and extensions

We have showed how the SPDE approach by Lindgren et al. (2011) can be extended to the case of Laplace noise and how this can be used to obtain an efficient estimation procedure as well as an accurate estimation technique for the Laplace moving average models. This is indented as a demonstration that the methods in Lindgren et al. (2011) are applicable to more general situations than the ordinary Gaussian models. There are also a number of extensions that can be made to this work which are discussed below.

First of all, the Hilbert space approximation technique in Section 4 was derived using theory for Lévy processes of type G, and although we only used this for the case of Laplace noise, the methods work equally well for this larger class of

models. All that is changed are the distributions of the integrals conditionally on the variance process. These techniques are also applicable to the case when more general SPDEs are used, one could for example use the nested SPDEs by Bolin and Lindgren (2011) to achieve more general covariance structures without any additional work needed, or one could include drift terms in the operator on the left-hand side to mimic the effects of asymmetric kernels in the Laplace moving average models. The methods are in fact not restricted to $\mathbb{R}^d$ or stationary SPDEs, but can be extended to non-stationary SPDEs on general Riemann manifolds.

Secondly, the estimation procedure in Section 6 assumed that one basis function was used for each observation of the process. The reason being that this gives us a one-to-one correspondence between the observations and the Laplace variables $\mathbf{\Lambda}$ which simplified the estimation procedure. For practical applications this is not ideal as one would like to be able to choose the basis independently of the measurement locations, and it would also be useful if one could assume that the measurements are taken under measurement noise. If the estimation procedure could be extended to handle these cases, the practical usefulness of these models would greatly improve.

As mentioned in Section 7, the estimation procedure is sensitive to the value of τ. Too large values will result in a model which is very similar to a standard Gaussian model, and it might be difficult to accurately estimate the parameters in this case without a very large data set. This is not a big problem as if the data is Gaussian, one should not use these models but a standard Gaussian model. The estimation procedure is also unstable for small values of τ, and modifications to further improve the stability in this case are currently being investigated.

Acknowledgements

The author is grateful to Krzysztof Podgórski for many helpful comments and discussions regarding the theoretical aspects of this work, to Daniel Simpson for providing some of his code for the Krylov subspace methods used in Section 5, and to Jonas Wallin for numerous discussions regarding the parameter estimation problem and for suggesting the truncation of the expectations mentioned at the end of Section 7.

	κ	τ	σ	μ	γ
A	1 (0.95 1.00 1.06)	2 (1.63 2.03 3.02)	1 (0.78 0.98 1.13)	0 (-0.07 -0.01 0.06)	0 (-0.06 0.00 0.07)
B	1 (0.96 1.00 1.05)	2 (1.68 1.99 2.41)	$\frac{1}{2}$ (0.42 0.50 0.57)	$\frac{1}{2}$ (0.43 0.50 0.57)	0 (-0.06 0.00 0.07)
C	1 (0.96 1.00 1.05)	1 (0.85 0.99 1.21)	1 (0.87 1.00 1.11)	0 (-0.07 0.00 0.05)	0 (-0.04 0.00 0.05)
D	1 (0.96 1.00 1.04)	1 (0.90 1.00 1.14)	1 (0.90 1.00 1.10)	1 (0.87 1.00 1.13)	-1 (-1.11 -1.00 -0.89)
E	1 (0.97 1.00 1.02)	$\frac{1}{2}$ (0.45 0.49 0.54)	1 (0.93 1.00 1.08)	0 (-0.06 0.00 0.06)	0 (-0.01 0.00 0.01)
F	1 (0.98 1.00 1.01)	$\frac{1}{2}$ (0.46 0.50 0.54)	1 (0.91 1.00 1.08)	1 (0.89 1.00 1.11)	-1 (-1.09 -1.01 -0.92)
G	$\frac{1}{10}$ (0.09 0.10 0.11)	1 (0.86 1.00 1.24)	1 (0.87 1.00 1.11)	0 (-0.06 0.00 0.06)	0 (-0.04 0.00 0.04)
H	$\frac{1}{10}$ (0.09 0.10 0.11)	1 (0.89 0.99 1.13)	$\frac{1}{2}$ (0.45 0.50 0.54)	$\frac{1}{2}$ (0.44 0.50 0.56)	0 (-0.05 0.01 0.13)
I	$\frac{1}{10}$ (0.10 0.10 0.10)	$\frac{1}{2}$ (0.45 0.49 0.54)	1 (0.91 1.01 1.09)	0 (-0.06 0.00 0.07)	0 (-0.01 0.00 0.01)
J	$\frac{1}{10}$ (0.10 0.10 0.10)	$\frac{1}{2}$ (0.46 0.50 0.54)	1 (0.91 0.99 1.08)	1 (0.90 1.01 1.13)	-1 (-1.09 -1.01 -0.93)
K	$\frac{1}{10}$ (0.10 0.10 0.10)	$\frac{1}{3}$ (0.31 0.33 0.36)	1 (0.91 0.99 1.06)	0 (-0.07 0.00 0.07)	0 (-0.00 0.00 0.00)
L	$\frac{1}{10}$ (0.09 0.10 0.12)	$\frac{1}{3}$ (0.33 0.36 0.43)	$\frac{1}{2}$ (0.42 0.47 0.51)	$\frac{1}{2}$ (0.39 0.50 0.52)	0 (-0.05 0.00 0.13)

Table 1: Parameter settings for the twelve cases shown in Figure 6 the estimation procedure is tested for. In the parentheses, the 10% 50%, and 90% percentiles of 500 Monte Carlo samples are shown. Note that most estimates seem to be unbiased, perhaps with the exception of the estimates of τ in case L.

References

Åberg, S. and Podgórski, K. (2011). A class of non-Gaussian second order random fields. *Extremes*, 14:187–222.

Åberg, S., Podgórski, K., and Rychlik, I. (2009). Fatigue damage assessment for a spectral model of non-Gaussian random loads. *Probab. Eng. Mech.*, 24:608–617.

Adams, R. A. (1975). *Sobolev Spaces*. Academic Press.

Bogsjö, K., Podgórski, K., and Rychlik, I. (2012). Models for road surface roughness. *Vehicle System Dynamics*, 50(5):725–747.

Bolin, D. and Lindgren, F. (2009). Wavelet Markov approximations as efficient alternatives to tapering and convolution fields (submitted). *Preprints in Math. Sci. Lund University*, 2009:13.

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

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

Gradshteyn, I. S. and Ryzhik, I. M. (2000). *Table of integrals, series and products*. Elsevier Inc., 6 edition.

Hale, N., Higham, N. J., and Trefethen, L. N. (2008). Computing A^α, $\log(A)$ and related matrix functions by contour integrals. *SIAM J. Numer. Anal.*, 46(5):2505–2523.

Higdon, D. (2001). Space and space-time modeling using process convolutions. Technical report.

Lange, K., Little, R., and Taylor, J. (1989). Robust statistical modeling using the t distribution. *J. Amer. Statist. Assoc.*, 84(408):881–896.

Lindgren, F., Rue, H., and Lindström, J. (2011). 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.

Matérn, B. (1960). Spatial variation. *Meddelanden från statens skogsforskningsinstitut*, 49(5).

Meng, X.-L. and Rubin, D. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika*, 80(2):267–78.

Podgórski, K. and Wegener, J. (2011). Estimation for stochastic models driven by Laplace motion. *Commun. Statist.- Theory Methods*, 40:3281–3302.

Protassov, R. (2004). EM-based maximum likelihood parameter estimation for multivariate generalized hyperbolic distributions with fixed λ. *Statist. and Comput.*, 14(1):67–77.

Røislien, J. and Omre, H. (2006). T-distributed random fields: A parametric model for heavy-tailed well-log data. *Math. Geol.*, 38(7):821–849.

Rosiński, J. (1991). On a class of infinitely divisible processes represented as mixtures of Gaussian processes. In *Stable Processes and Related Topics*, volume 25 of *Progress in Probability*, pages 27–41. Birkhauser, Boston.

Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields; Theory and Applications*, volume 104 of *Monographs on Statistics and Applied Probability*. Chapman & Hall/CRC.

Samko, S. G., Kilbas, A. A., and Maricev, O. I. (1992). *Fractional integrals and derivatives: theory and applications*. Gordon and Breach Science Publishers, Yveron.

Simpson, D., Lindgren, F., and Rue, H. (2010). In order to make spatial statistics computationally feasible, we need to forget about the covariance function. *Preprint, statistics, Trondheim, Norway*, 16/2010.

Simpson, D. P. (2008). Krylov subspace methods for approximating functions of symmetric positive definite matrices with applications to applied statistics and anomalous diffusion. *PhD thesis, Queensland University of Technology*.

Walsh, J. (1986). An introduction to stochastic partial differential equations. In *École d'Été de Probabilités de Saint Flour XIV - 1984*, volume 1180 of *Lecture Notes in Mathematics*, chapter 3, pages 265–439. Springer Berlin / Heidelberg.

Whittle, P. (1963). Stochastic processes in several dimensions. *Bull. Internat. Statist. Inst.*, 40:974–994.

Wiktorsson, M. (2002). Simulation of stochastic integrals with respect to Lévy processes of type G. *Stoch. Proc. Appl.*, 101:113–125.

E

Excursion and contour uncertainty regions for latent Gaussian models

DAVID BOLIN[1] AND FINN LINDGREN[2]

[1] *Centre for Mathematical Sciences, Lund University, Lund, Sweden*
[2] *Department of Mathematical Sciences, Norwegian University of Science and Technology, Trondheim, Norway*

Abstract: An interesting statistical problem is to find regions where some studied process exceeds a certain level. Estimating these regions so that the probability for exceeding the level in the entire set is equal to some predefined value is a difficult problem that occurs in several areas of applications ranging from brain imaging to astrophysics. In this work, a method for solving this problem, as well as the related problem of finding uncertainty regions for contour curves, for latent Gaussian models is proposed. The method is based on using a parametric family for the excursion sets in combination with INLA and an importance sampling algorithm for estimating joint probabilities. The accuracy of the method is investigated using simulated data and two environmental applications are presented. In the first, areas where the air pollution in the Piemonte region in northern Italy exceeds the daily limit value, set by the European Union for human health protection, are estimated. In the second, regions in the African Sahel that experienced an increase in vegetation after the drought period in the early 1980s are estimated.

Key words: Latent Gaussian models; excursion sets; contour curves; multiple testing; INLA

1 Introduction

In many statistical applications, one is interested in finding areas where the studied process exceeds a certain level or is significantly different from some reference level. A typical example is in studies of air pollution, where one is interested in testing if, and where, the pollution level exceeds some given limit value set by some regulatory agency (Cameletti et al., 2012), and similar examples can be found in a wide range of scientific fields including brain imaging (Marchini and Presanis,

2003) and astrophysics (Beaky et al., 1992). In spatio-temporal applications one might be interested in finding regions that have experienced significant changes over the studied time period. This is a common problem in climate science and the studied quantity can for example be temperature (Furrer et al., 2007), precipitation (Sain et al., 2011), or vegetation (Eklundh and Olsson, 2003, Bolin et al., 2009).

The quintessential problem is that one has observations $\mathbf{y}$ of some latent stochastic field $x(\mathbf{s})$ and wants to find a region D such that, with a certain given probability, $x(\mathbf{s}) > u$ for all $\mathbf{s} \in D$ for a given level u. The easiest, and most common, way of specifying D is to choose it as the set of locations

$$D_m = \{\mathbf{s} : \mathsf{P}(x(\mathbf{s}) > u) \geq 1 - \alpha\}, \tag{1}$$

where the probability is taken under the posterior distribution for $x|\mathbf{y}$. Thus, D is specified as the set of locations where the marginal probability for exceeding the level exceeds some given value $1 - \alpha$. The set can be calculated using multiple hypothesis testing and the parameter α then acts as the type 1 error parameter, and should thus control how certain one is that the level is exceeded in the entire set. The problem with this definition of D is that of multiple hypothesis testing; the confidence level α does not give us any information about the family-wise error rate. That is, the probability $\mathsf{P}(x(\mathbf{s}) > u, \mathbf{s} \in D_m)$ is in general not equal to $1 - \alpha$. If one instead wants to choose D so that this simultaneous probability is $1 - \alpha$, one has to modify the procedure for constructing the set.

The more general problem of multiple hypothesis testing is an active research area and there exists a number of proposed solutions for problems in various contexts. Most of these solutions are based on first calculating the marginal probabilities $\mathsf{P}(x(\mathbf{s}) > u)$, then calculating a single threshold value t, and finally defining the exceedance region as $D = \{\mathbf{s} : \mathsf{P}(x(\mathbf{s}) > u) > t\}$. The methods differ in how the threshold t is calculated, and can basically be divided into three main categories; type 1 error control thresholding, false discovery rate thresholding, and posterior probability thresholding (Marchini and Presanis, 2003). The most popular method is likely the method by Adler (1981) using the Euler characteristic of the latent field to control the family-wise error rate when defining the threshold t. Though this method is simple to use, one has to be careful to check whether the required assumptions are satisfied. Typically the method is accurate for large values of u, and the latent field is assumed to be stationary.

In this work we will focus on the problem where the latent spatial field $x(\mathbf{s})$ is Gaussian and measured at a set of irregular locations. This means that the

posterior distribution $\pi(x|\mathbf{y})$ is non-stationary and typically non-Gaussian unless the measurements are Gaussian and the model parameters are known a priori. One of our motivating examples is the problem of finding regions with significant changes in vegetation in the African Sahel studied by Eklundh and Olsson (2003) and Bolin et al. (2009), and in this example the threshold u is zero, so methods based on asymptotic arguments when u goes to infinity are unlikely to perform well.

The structure of the article is as follows. In Section 2, the problem is formulated and definitions for excursion sets and uncertainty regions for contour curves are given. In Section 3, a method for estimating these sets is proposed. Estimating the sets is the most difficult problem as one easily runs into computational difficulties arising from having to evaluate high-dimensional integrals. In Section 4, the methods are tested on a few simulated examples to test the method's accuracy. Two applications to real data are covered in Section 5, the first considers air pollution data from the North-Italian region Piemonte, and the second considers estimation of spatially dependent vegetation trends in the African Sahel. Finally, a few remarks and comments are given in Section 6.

2 Problem formulation

There are a number of different ways one could formulate excursion sets, and not all of them are useful from a practical point of view. Hence, in this section we will formalise the problem and discuss how the results should be interpreted. More precisely we look at two connected problems. The first one is to find areas where a stochastic process exceeds a given level with some probability and the second one is to quantify the uncertainty in contour curves of stochastic fields.

Throughout this section, let Ω be a bounded domain of $\mathbb{R}^n$, or have a well-defined area $|\Omega| < \infty$. First some notation for excursion sets of a function and contour sets is needed.

Definition 2.1 (Excursion sets for functions)**.** Given a function $f(\mathbf{s})$, $\mathbf{s} \in \Omega$, the positive excursion set A_u^+ for a level u is given by

$$A_u^+(f) = \{\mathbf{s} \in \Omega; f(\mathbf{s}) > u\}.$$

Similarly

$$A_u^-(f) = \{\mathbf{s} \in \Omega; f(\mathbf{s}) < u\}.$$

is the negative excursion set.

In a similar fashion one could now define the set of contour points for the level u as the set of points $\mathbf{s}$ for which $f(\mathbf{s}) = u$; however, a contour curve consists not only of these points but also discontinuous crossings of the level u. In order to incorporate both continuous and discontinuous crossings, a contour point is defined as a point $\mathbf{s}$ such that in every neighborhood B of $\mathbf{s}$

$$\exists\, \mathbf{s}_1, \mathbf{s}_2 \in B : f(\mathbf{s}_1) \geq u, f(\mathbf{s}_2) \leq u.$$

The set of all such points is the complement of the union of the interior sets of the positive and negative excursion sets.

Definition 2.2 (Contour sets for functions)**.** Given a function $f(\mathbf{s})$, $\mathbf{s} \in \Omega$, the contour set A_u^o for a level u is given by

$$A_u^c(f) = \left(A_u^+(f)^o \cup A_u^-(f)^o\right)^c.$$

where A^o is the interior, relative to Ω, of the set A and A^c is the complement.

Remark 4. Taking the interiors of the sets $A_u^+(f)$ and $A_u^-(f)$ is important. Consider for example the following function on $\Omega = [0, 1]$

$$f(s) = \begin{cases} -1 & 0 \leq s < 0.5 \\ 1 & 0.5 \leq s \leq 1. \end{cases}$$

In this case $A_0^+(f) \cup A_0^-(f) = \Omega$, so without taking the interiors of the sets $A_0^c(f)$ would be empty when we want to include the discontinuous crossing at 0.5 in the contour set. It is also important to take the interiors with respect to Ω and not $\mathbb{R}$, since the endpoints 0 and 1 always would be included in the countour set otherwise. This may seem as only a theoretical nicety, but the problem with discontinuous functions occurs frequently in environmental applications when discontinuous covariates are used for the mean value function of the field. This makes it essential to not treat contour sets as regions where the function lies close to a level, but rather as regions where level *crossings* occur.

The statistical problem is now to find a region D such that the function $x(\mathbf{s})$ exceeds the level u with a certain probability $1 - \alpha$ for all $\mathbf{s} \in D$. There might be many such regions, so if one is interested in a single answer one might look for the largest of these.

Definition 2.3 (Excursion sets)**.** Let $x(\mathbf{s})$, $\mathbf{s} \in \Omega$ be a random field (or process). The positive level u excursion set with probability $1-\alpha$ is given by

$$E^+_{u,\alpha}(x) = \arg\max_D \{|D| : \mathsf{P}(D \subseteq A^+_u(x)) \geq 1-\alpha\}.$$

Similarly

$$E^-_{u,\alpha}(x) = \arg\max_D \{|D| : \mathsf{P}(D \subseteq A^-_u(x)) \geq 1-\alpha\}.$$

is the negative level u excursion set with probability $1-\alpha$.

Remark 5. The set $E^+_{u,\alpha}(x)$ can also be formulated as the largest set D for which $\mathsf{P}(\inf_{\mathbf{s}\in D} x(\mathbf{s}) \leq u) \leq \alpha$, which can be useful when calculating the set in practice. Also note that for deterministic functions f one has $E^+_{u,\alpha}(f) = A^+_u(f)$ and $E^-_{u,\alpha}(f) = A^-_u(f)$ for any $\alpha \in [0,1]$.

It is important to realize how the excursion set $E^+_{u,\alpha}(x)$ should be interpreted: It is the largest set so that the level u is exceeded *at all locations* in the set with probability $1-\alpha$. It will thus be a smaller set than D_m defined in (1), which is the set of points where the marginal probability for exceeding the level is at least $1-\alpha$. Another possible definition of an excursion set would be a set that contains *all excursions* with probability $1-\alpha$. This is a larger set than D_m, given by $E^-_{u,\alpha}(x)^c$. Which set one is interested in depends on the application, but it can be a good idea to show both to get a better unterstanding of the uncertainties in the problem.

In certain applications, one might be interested in joint positive and negative excursions from some level, for example when doing simultaneous regressions and one is interested in finding regions where the slopes are significantly different from zero (see Section 5.2 for a possible scenario of this kind).

Definition 2.4 (Level avoiding sets)**.** Let $x(\mathbf{s})$, $\mathbf{s} \in \Omega$ be a random field. The pair of level u avoiding sets with probability $1-\alpha$ is given by

$$\begin{aligned}&(M^+_{u,\alpha}(x), M^-_{u,\alpha}(x))\\ &= \arg\max_{(D^+,D^-)} \{|D^- \cup D^+| : \mathsf{P}(D^- \subseteq A^-_u(x),\ D^+ \subseteq A^+_u(x)) \geq 1-\alpha\}.\end{aligned}$$

Denote the union of these two sets the level avoiding set $M_{u,\alpha}$:

$$M_{u,\alpha}(x) = M^+_{u,\alpha}(x) \cup M^-_{u,\alpha}(x).$$

Remark 6. The sets $M^+_{u,\alpha}(x)$ and $M^-_{u,\alpha}(x)$ must be non-overlapping for the probability to be non-zero. The set $M_{u,\alpha}$ can be calculated as an excursion set itself. To see this, define a new random process $y(\mathbf{s})$ by

$$y(\mathbf{s}) = \begin{cases} u - x(\mathbf{s}), & \mathbf{s} \in D^-, \\ x(\mathbf{s}) - u, & \mathbf{s} \notin D^-. \end{cases}$$

The probability calculation in Definition 2.4 can now be reformulated as an ordinary excursion probability in y:

$$\mathsf{P}(D^- \subseteq A^-_u(x),\, D^+ \subseteq A^+_u(x)) = \mathsf{P}(D^- \cup D^+ \subseteq A^+_0(y)).$$

Also in this case, the set can be found using a reformulation using the infimum over the region as $\mathsf{P}(\inf_{\mathbf{s}\in D^-\cup D^+} y(\mathbf{s}) \leq 0) \leq \alpha$.

Simlarly to the contour sets for deterministic functions were defined, the pair of level avoiding sets can now be used to define uncertainty regions for contour curves.

Definition 2.5 (Uncertainty region for contour sets)**.** Let $x(\mathbf{s})$, $\mathbf{s} \in \Omega$ be a random field, and let $(M^+_{u,\alpha}(x), M^-_{u,\alpha}(x))$ be the pair of level avoiding sets from Definition 2.4. The set

$$M^c_{u,\alpha}(x) = \left(M^+_{u,\alpha}(x)^o \cup M^-_{u,\alpha}(x)^o\right)^c$$

is then an uncertainty region for the contour set of level u.

The interpretation of this uncertainty region is important. The set $M^c_{u,\alpha}$ is the smallest set such that *all* level u crossings of x are in the set with probability $1-\alpha$. One should note that this definition of the uncertainty region for level curves is different from some other definitions in the literature. For example, Lindgren and Rychlik (1995) define uncertainty regions as a union of intervals where each interval contains a single level crossing with probability $1-\alpha$.

It is somewhat unsatisfactory that the sets defined here are made unique by finding the largest set satisfying a certain restriction. The set $E^+_{u,\alpha}(x)$ is for example defined as the largest set D satisfying $\mathsf{P}(D \subseteq A^+_u(x)) \geq 1-\alpha$, but there are also many other smaller sets satisfying the requirement, and these are not seen if only $E^+_{u,\alpha}(x)$ is reported. Also, if one wants to know where the field likely exceeds the level u, the set $E^+_{u,\alpha}(x)$ might not be sufficient since it does not provide any information about the locations not contained in the set.

It would instead be good to have something similar to p-values, i.e. the marginal probabilities of exceeding the level, but which can be interpreted simultaneously. To that end we introduce the excursion function, level avoidance function, and contour function as visual tools for answering such questions.

Definition 2.6 (Excursion functions)**.** The positive and negative u excursion functions are given by

$$F^{+}(u, \mathbf{s}) = \sup\{1 - \alpha; \mathbf{s} \in E^{+}_{u,\alpha}\},$$
$$F^{-}(u, \mathbf{s}) = \sup\{1 - \alpha; \mathbf{s} \in E^{-}_{u,\alpha}\}.$$

Similarly, the level avoidance and contour functions are given by

$$F(u, \mathbf{s}) = \sup\{1 - \alpha; \mathbf{s} \in M_{u,\alpha}\},$$
$$F^{c}(u, \mathbf{s}) = 1 - F(u, \mathbf{s}).$$

These functions will take values between zero and one, and for a fixed level u and a fixed location $\mathbf{s}$, this value is equal to $1 - \alpha$ for the smallest α such that the location is a member of the excursion set. Thus, the set $E^{\bullet}_{u,\alpha}$ can be retrieved as the $1 - \alpha$ excursion set of the function $F^{\bullet}(u, s)$. The interpretation of the excursion function is therefore that if, for a given location $\mathbf{s}$, the function takes a value close to one, this indicates that this location is a member of the excursion sets for alsmost all values of α, whereas if the value of the function is close to zero, the location is only a member of excursion sets with large values of α and it is therefore more unlikely that the process exceeds the level at that location.

3 Computations

So far, no assumptions have been made regarding the distribution of $x(\mathbf{s})$, but to be able to calculate the excursion sets in practice we will now restrict ourselves to the class of latent Gaussian models, which is a popular model class with many practical applications (see e.g. Rue et al., 2009). Thus, the following problem setup is assumed. Let $x(\mathbf{s})$ be a random field that can be written on the form

$$x(\mathbf{s}) = \sum_{i=1}^{k} \beta_i f_i(\mathbf{s}) + z(\mathbf{s})$$

where $f_i(\mathbf{s})$ are fixed effects and $z(\mathbf{s})$ is a zero mean random field with covariance parameters $\boldsymbol{\theta}_1$. Further assume that both $z(\mathbf{s})$ and the parameter vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)^\top$ are a priori Gaussian.

Let $\mathbf{y} = (y_1, \ldots, y_m)^\top$ be a vector of measurements of the latent field with some distribution $\pi(\mathbf{y}|\mathbf{x}_{obs}, \boldsymbol{\theta}_2)$, where $\mathbf{x}_{obs}$ is a vector containing the latent field evaluated at the measurement locations and $\boldsymbol{\theta}_2$ is a vector of parameters for the measurement distribution. Finally let $\mathbf{s}_1, \ldots, \mathbf{s}_n$ be the set of locations where predictions of the latent field should be calculated and let $\mathbf{x} = (x(\mathbf{s}_1), \ldots, x(\mathbf{s}_n))^\top$. The posterior distribution for $\mathbf{x}$ can then be written as

$$\pi(\mathbf{x}|\mathbf{y}) = \int \pi(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta})\pi(\boldsymbol{\theta}|\mathbf{y})\,\mathrm{d}\boldsymbol{\theta}, \tag{2}$$

where $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^\top, \boldsymbol{\theta}_2^\top)^\top$, and this is the distribution that should be used in the probability calculations when estimating the excursion sets.

There are now, in principle, two main problems that have to be solved in order to find the excursion sets, level avoidance sets, or contour uncertainty sets:

integration For excursion sets, calculate $\mathsf{P}(D \subseteq A_u^+(\mathbf{x}))$ or $\mathsf{P}(D \subseteq A_u^-(\mathbf{x}))$ for a given set D, or in the case of level avoidance sets or uncertainty regions for contour curves, calculate $\mathsf{P}(D^- \subseteq A_u^-(\mathbf{x}), D^+ \subseteq A_u^+(\mathbf{x}))$ for the pair of sets (D^+, D^-).

optimization Use shape optimization to find largest region D satisfying the required probability constraint.

Hence, given a method to solve each of the two problems, one could simply run the shape optimization algorithm and in each iteration calculate the required probability using the integration method. In theory there are no problems doing this, but in practice the integration method will be computationally demanding and it may not be feasible to use this strategy for applications involving large data sets. Therefore, we instead propose a slightly different strategy that will minimize the number of calls to the integration method by solving the problem sequentially. We first outline the strategy in the simplest possible situation, which will be used as a basis for all other more complicated strategies.

The method is based on using an increasing parametric family for the excursion sets in combination with a sequential integration routine for calculating the probabilities. The advantage with using a sequential integration routine is that if

the required probability has been calculated for some set D_1, then the calculation for a larger set $D_2 \supset D_1$ can be based on the result for D_1, resulting in large computational savings.

Algorithm 3.1 (Calculating excursion sets using a one-parameter family)**.** *Assume that the model parameters $\boldsymbol{\theta}$ are known and that the posterior distribution $\pi(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta})$ is Gaussian. Further assume that $D(\rho)$ is a parametric family for the possible excursion sets, such that $D(\rho_1) \subseteq D(\rho_2)$ if $\rho_1 < \rho_2$. The following strategy is then used to calculate $E^+_{u,\alpha}$.*

1. *Choose a suitable (sequential) integration method for the problem.*

2. *Reorder the nodes to the order they will be added to the excursion set when the parameter ρ is increased.*

3. *sequentially add nodes to the set D according to the ordering given above and in each step update the probability $\mathsf{P}(D \subseteq A^+_u(\mathbf{x}))$. Stop as soon as this probability falls below $1 - \alpha$.*

4. *$E^+_{u,\alpha}$ is given by the last set D for which $\mathsf{P}(D \subseteq A^+_u(\mathbf{x})) \geq 1 - \alpha$.*

The computational savings of this sequential strategy are large. For example, assume that we want to find the positive level u excursion set $E^+_{u,\alpha}(x)$, and have m candidates $D(\rho_1), \ldots, D(\rho_m)$ to choose from. Using the naïve optimization method, we would then have to check whether $\mathsf{P}(D(\rho_i) \subseteq A^+_u(x)) > 1 - \alpha$ for each of these sets, and among the sets that satisfy the condition select the largest. Thus doing the probability calculation m times. However, by reordering the nodes and adding them sequentially we only have to run the integration routine once.

Before extending this method to more general situations, we go into more detail on how to do the steps in Algorithm 3.1 in practice. In Section 3.1, a few sequential integration methods are presented. In Section 3.2, some different parametric families for the excursion sets and level avoidance sets are introduced and Algorithm 3.1 is extended using two-parameter families. The problem of how to optimally reorder the nodes is also discussed in this section. Finally in Section 3.3, three different methods are proposed for calculating excursion sets under the full posterior distribution (2).

3.1 Gaussian probability calculations

For a Gaussian vector $\mathbf{x}$, the probabilities $\mathsf{P}(D \subseteq A_u^+(\mathbf{x}))$, $\mathsf{P}(D \subseteq A_u^-(\mathbf{x}))$, and $\mathsf{P}(D^+ \subseteq A_u^+(\mathbf{x}), D^- \subseteq A_u^-(\mathbf{x}))$ can all be written on the form

$$I(\mathbf{a}, \mathbf{b}, \mathbf{\Sigma}) = \frac{1}{(2\pi)^{d/2}|\mathbf{\Sigma}|^{1/2}} \int_{\mathbf{a} \leq \mathbf{x} \leq \mathbf{b}} \exp(-\frac{1}{2}\mathbf{x}^\top \mathbf{\Sigma}^{-1}\mathbf{x})\, \mathrm{d}\mathbf{x}, \tag{3}$$

where $\mathbf{a}$ and $\mathbf{b}$ are vectors depending on the mean value of $\mathbf{x}$, the domain D, and on u. There have been considerable research efforts devoted to approximating integrals of this form in recent years, and we will in this section briefly describe a few techniques that can be used.

3.1.1 Quasi Monte Carlo methods

The simplest way of approximating (3) is to use Monte-Carlo (MC) integration. However, estimating the probability with any reasonable accuracy using standard MC integration is often too computationally expensive. Fortunately there are a number of variance reduction techniques that can be used to increase the efficiency.

A key step in many numerical integration techniques is to transform the integral to make it more suitable for integration. Notably, Genz (1992) derived such a transformation for the Gaussian integral (3), though similar transformations have been suggested by other authors as well (see e.g. Geweke, 1991). Genz (1992) begins by calculating the Cholesky factor $\mathbf{L}$ of $\mathbf{\Sigma}$ and then transforms the integral as

$$\begin{aligned} I(\mathbf{a}, \mathbf{b}, \mathbf{\Sigma}) &= \frac{1}{(2\pi)^{d/2}} \int_{\mathbf{a} \leq \mathbf{L}\mathbf{x} \leq \mathbf{b}} \exp(-\frac{1}{2}\mathbf{y}^\top \mathbf{y})\, \mathrm{d}\mathbf{y} \\ &= \int_{\tilde{a}_1}^{\tilde{b}_1} \phi(y_1) \int_{\tilde{a}_2(y_1)}^{\tilde{b}_2(y_1)} \phi(y_2) \cdots \int_{\tilde{a}_d(\mathbf{y}_{1:d-1})}^{\tilde{b}_d(\mathbf{y}_{1:d-1})} \phi(y_d)\, \mathrm{d}\mathbf{y} \end{aligned}$$

where $\tilde{a}_i = L_{ii}^{-1}(a_i - \sum_{j=1}^{i-1} L_{ij}y_j)$, $\tilde{b}_i = L_{ii}^{-1}(b_i - \sum_{j=1}^{i-1} L_{ij}y_j)$, $\phi(x)$ is the standard Gaussian probability density function, and $\mathbf{y}_{1:d-1} = \{y_1, \ldots, y_{d-1}\}$. After this, two more transformations are made to transform the integral to the unit hyper cube $[0, 1]^d$. Firstly, let $y_i = \Phi^{-1}(z_i)$, where $\Phi(x)$ is the standard Gaussian

cumulative distribution function, and secondly let $z_i = d_i + (e_i - d_i)w_i$ where

$$\begin{aligned} d_i(\mathbf{w}_{1:i-1}) =& \Phi(\tilde{a}_i(\Phi^{-1}(z_1(w_1)), \ldots, \Phi^{-1}(z_{i-1}(w_{i-1}))) \\ e_i(\mathbf{w}_{1:i-1}) =& \Phi(\tilde{b}_i(\Phi^{-1}(z_1(w_1)), \ldots, \Phi^{-1}(z_{i-1}(w_{i-1}))). \end{aligned}$$

Then the integral $I(\mathbf{a}, \mathbf{b}, \mathbf{\Sigma})$ can be rewritten as

$$(e_1 - d_1) \int_0^1 (e_2(w_1) - d_2(w_1)) \cdots \int_0^1 (e_{d-1}(\mathbf{w}_{1:d-1}) - d_{d-1}(\mathbf{w}_{1:d-1}) \int_0^1 \mathrm{d}\mathbf{w}.$$

Besides having transformed the integral to the unit hyper cube, the transformation has also achieved a separation of the variables so that the full problem can be calculated sequentially.

This integral can then efficiently be evaluated using a quasi MC (QMC) method where the uniform random numbers in the ordinary MC integrator are replaced by some deterministic sequence of points chosen to reduce the probabilistic error bound of the crude MC integrator. There are a number of ways such deterministic sequences can be chosen in, and it is outside the scope of this article to cover these, see Genz and Bretz (2009) for details.

A final variance reduction technique for the general integration problem can be achieved by reordering the variables before calculating the integral. Schervish (1984) originally proposed sorting the variables so that the first variable has the shortest integration interval and inte innermost integral has the widest interval. Gibson et al. (1994) improved this reordering by sorting the variables so that the innermost integral has the largest expected value. This reordering can reduce the error by an order of magnitude, as shown by Genz and Bretz (2002). However, the technique will not be applicable in our situation since the reordering will be determined by a parametric family for the excursion sets.

3.1.2 Methods for Markov random fields

A common assumption in spatial statics and image analysis is that the latent field can be modeled, or approximated, using a Gaussian Markov random field (GMRF). One of the motivating reasons for using GMRFs is that it reduces the computational cost for parameter estimation and spatial prediction, and because of this one would also like to be able to use the Markov property in the calculation of (3).

The main difference between latent GMRF models and standard Gaussian models is that the distribution is specified using the (sparse) precision matrix instead of the covariance matrix:

$$I(\mathbf{a}, \mathbf{b}, \mathbf{Q}) = \frac{|\mathbf{Q}|^{1/2}}{(2\pi)^{d/2}} \int_{\mathbf{a} \leq \mathbf{x} \leq \mathbf{b}} \exp(-\frac{1}{2}\mathbf{x}^\top \mathbf{Q}\mathbf{x})\, \mathrm{d}\mathbf{x}, \tag{4}$$

Using the QMC methods from the previous section directly is difficult without first inverting the precision matrix and then ignoring the sparsity of $\mathbf{Q}$ in the calculations. To take advantage of the sparsity of $\mathbf{Q}$ one can use the fact that any GMRF can be viewed as a non-homogeneous auto-regressive process defined backwards in the indices of $\mathbf{x}$ (see Rue and Held, 2005, Theorem 2.7), that is, if $\mathbf{x}$ is a GMRF with mean $\boldsymbol{\mu}$ and precision matrix $\mathbf{Q}$, then

$$x_i | x_{i+1}, \ldots, x_n \sim \mathsf{N}\left(\mu_i - \frac{1}{L_{ii}} \sum_{j=i+1}^{n} L_{ji}(x_j - \mu_j), L_{ii}^{-2}\right), \tag{5}$$

where L_{ij} are the elements of the Cholesky factor of $\mathbf{Q}$. The integral can thus be written as

$$I(\mathbf{a}, \mathbf{b}, \mathbf{Q}) = \int_{a_1}^{b_1} \pi(x_1|\mathbf{x}_{2:d}) \int_{a_2}^{b_2} \pi(x_2|\mathbf{x}_{3:d}) \cdots \int_{a_{d-1}}^{b_{d-1}} \pi(x_{d-1}|x_d) \int_{a_d}^{b_d} \pi(x_d)\, \mathrm{d}\mathbf{x}$$

where, because of the Markov structure, $x_i|\mathbf{x}_{i+1:d}$ only depends on the elements in $x_{\mathcal{N}_i \cap \{i+1:d\}}$, and $\mathcal{N}_i$ is the neighborhood of i in the graph of the GMRF.

If $\mathbf{Q}$ is a band-matrix, the integral can be efficiently calculated as a sequence of iterated one-dimensional integrals as discussed in Genz and Kahaner (1986). However, the band width of $\mathbf{L}$ will often be too large for this method to be efficient, and a better alternative is then of use a particle filter algorithm based on the GHK simulator (Geweke, 1991, Hajivassiliou, 1991, Keane, 1993). Denote the integral of the last $d - i$ components by I_i,

$$I_i = \int_{a_i}^{b_i} \pi(x_i|\mathbf{x}_{i+1:d}) \cdots \int_{a_{d-1}}^{b_{d-1}} \pi(x_{d-1}|x_d) \int_{a_d}^{b_d} \pi(x_d)\, \mathrm{d}\mathbf{x},$$

and note that the integral is the normalizing constant to the truncated density $f_i(\mathbf{x}_{i:d}) = 1(\mathbf{a}_{i:d} < \mathbf{x}_{i:d} < \mathbf{b}_{i:d})\pi(\mathbf{x}_{i:d})$. The integrals $I_d, \ldots, I_1$ are now estimated sequentially using importance sampling. In the first step, calculate

$I_d = \Phi(L_{ii}(b_d - \mu_d)) - \Phi(L_{ii}(a_d - \mu_d))$, simulate N samples $\{x_d^j\}_{j=1}^N$ from the truncated normal distribution $h_d(x_d) = 1(a_d < x_d < b_d)\pi(x_d)$, and set $w_d^j = I_d$. Next, simulate x_{d-1}^j from $h_{d-1}(x_{d-1}|x_d^j) = 1(a_{d-1} < x_{d-1} < b_{d-1})\pi(x_{d-1}|x_d^j)$ and set $\mathbf{x}_{d-1:d}^j = \{x_{d-1}^j, x_d^j\}$. The integral I_{d-1} is estimated as $I_{d-1} \approx \sum_{j=1}^N w_{d-1}^j$ where w_{d-1}^j are the importance weights $w_{d-1}^j = f_{d-1}(\mathbf{x}_{d-1:d})/h_{d-1}(\mathbf{x}_{d-1:d})$. Proceed like this, simulating from the truncated conditional distributions, and in each step update the importance weights recursively through

$$w_i^j = \left[\Phi\left(L_{ii}(b_i - \mu_i) + \sum_{j=i+1}^n L_{ji}(X_j - \mu_j) \right) - \Phi\left(L_{ii}(a_i - \mu_i) + \sum_{j=i+1}^n L_{ji}(X_j - \mu_j) \right) \right] w_{i+1}^j.$$

A common problem with sequential importance sampling is weight degeneration, i.e. that most of the importance weights will become very small after a few steps and the integral approximation will be determined by only a few particles with large weights. To reduce the variance of the estimator when the target probability is small, a resampling step can be performed after having calculated the weights w_i^j. The sample $\{\mathbf{x}_{i:d}^j\}$ is then updated by selecting N particles from the set, where $x_{i:d}^j$ is selected with probability $w_i^j / \sum_{k=1}^N w_i^k$. To avoid resampling too often, one can do the resampling only if some criterion is met, for example if the effective sample size is below some given threshold (see for example Doucet et al., 2001, for an introduction to particle filter techniques).

3.2 Parametric families

In theory one can use any shape optimization technique to find the largest region D. However, since evaluating the probability $\mathsf{P}(x(\mathbf{s}) > u, \mathbf{s} \in D)$ for a given set D is computationally expensive, one would like to do as few iterations as possible in this step. As discussed previously, we will solve this by assuming a parametric form of the sets D. The optimization can then be reduced to a standard optimization of only a few variables instead of doing a full shape optimization procedure. The parametric families will be based on the marginal quantiles of $x(\mathbf{s})$,

$$\mathsf{P}(x(\mathbf{s}) \leq q_\rho(\mathbf{s})) = \rho,$$

which are easy to calculate using only the marginal posterior distributions. The simplest one-parameter family based on the marginal quantiles is given in the following definition.

Definition 3.2 (One-parameter family)**.** Let $q_\rho(\mathbf{s})$ be the marginal quantiles for $x(\mathbf{s})$, then a one-parameter family for the positive and negative u excursion sets is given by

$$D_1^+(\rho) = \{\mathbf{s}; \mathsf{P}(x(\mathbf{s}) > u) \geq 1 - \rho\} = \{\mathbf{s}; \mathsf{P}(x(\mathbf{s}) \leq u) \leq \rho\} = A_u^+(q_\rho),$$
$$D_1^-(\rho) = \{\mathbf{s}; \mathsf{P}(x(\mathbf{s}) < u) \geq 1 - \rho\} = \{\mathbf{s}; \mathsf{P}(x(\mathbf{s}) \geq u) \leq \rho\} = A_u^-(q_{1-\rho}).$$

Using this one-parameter family in Algorithm 3.1 is equivalent to finding a threshold value for the marginal excursion probabilities to get the correct simultaneous significance level. It is thus similar to the thresholding algorithms discussed in Marchini and Presanis (2003) but with the important difference that the correct joint, often non-stationary, posterior density is used when finding the threshold.

The simple one-parameter family can be extended in a number of ways, for example by considering other levels in the excursion sets.

Definition 3.3 (Two-parameter family)**.** Let $q_\rho(\mathbf{s})$ be the marginal quantiles for $x(\mathbf{s})$, then a two-parameter family for the positive and negative u excursion sets is given by

$$D_1^+(v, \rho) = \{\mathbf{s}; \mathsf{P}(x(\mathbf{s}) > v) \geq 1 - \rho\} = \{\mathbf{s}; \mathsf{P}(x(\mathbf{s}) \leq v) \leq \rho\} = A_v^+(q_\rho),$$
$$D_1^-(v, \rho) = \{\mathbf{s}; \mathsf{P}(x(\mathbf{s}) < v) \geq 1 - \rho\} = \{\mathbf{s}; \mathsf{P}(x(\mathbf{s}) \geq v) \leq \rho\} = A_v^-(q_{1-\rho}).$$

The sets $D_1^+(v, \rho)$ and $D_1^-(v, \rho)$ are increasing in ρ for a fixed v.

One drawback with this parametric family is that it does not take the spatial dependency of the data into account directly. Therefore certain sets which might seem reasonable to test are not included in the family. Consider the example in Section 4.1, Figure 1, Panel (a), where the marginal excursion probabilities are shown in grey for an example in one dimension where the model is a Gaussian process with exponential covariance function. The estimated posterior mean in this example is shown as the black curve in Panel (b) in the figure, and in this situation a reasonable candidate for the 0-excursion set might be a contiguous set centered at 1, $[1 - \lambda_1, 1 + \lambda_2]$ for some positive λ_1, λ_2. However, looking at the

marginal probabilities we see that sets on this form will not be included in the parametric family. One way of including such sets is to first smooth the marginal excursion probabilities $p_i = \mathsf{P}(x(\mathbf{s}_i) > u)$ using some parametric smoother and then consider sets on the form $\{\mathbf{s}; p_i^\tau \geq 1 - \rho\}$ where p_i^τ are the smoothed excursion probabilities.

Definition 3.4 (Two-parameter smoothing family)**.** Let p_i^τ be the smoothed marginal u excursion probabilities, using a circular averaging filter with radius τ. A two-parameter family for the positive and negative u excursion sets is then given by

$$D_2^+(\tau, \rho) = \{\mathbf{s}; p_i^\tau \geq 1 - \rho\},$$
$$D_2^-(\tau, \rho) = \{\mathbf{s}; p_i^\tau \geq 1 - \rho\}.$$

The parameter τ determines how close p_i^τ is to the original excursion probabilities. For $\tau = 0$, no smoothing is done and for a general τ, p_i^τ is equal to the average of the marginal excursion probabilities in the disk with radius τ centered at $\mathbf{s}_i$. As τ increases p_i^τ becomes smoother and approaches a constant function equal to the average excursion probability. One could also use other types of parametric smoothers instead of the simple averaging filter.

Using the two-parameter families requires a modification to Algorithm 3.5, resulting in a slightly more computationally demanding method.

Algorithm 3.5 (Calculating excursion sets using a two-parameter family)**.** *Assume that the model parameters $\boldsymbol{\theta}$ are known and that the posterior distribution $\pi(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta})$ is Gaussian. Further assume that $D(\nu, \rho)$ is a parametric family for the possible excursion sets, such that $D(\nu, \rho_1) \subseteq D(\nu, \rho_2)$ if $\rho_1 < \rho_2$ for a fixed ν. The following strategy is then used to calculate $E_{u,\alpha}^+$.*

1. *Choose a suitable (sequential) integration method for the problem.*
2. *Select a suitable one-dimensional optimization strategy.*
3. *Do optimization of the size of $D(\nu, \bullet)$ over ν:*
   - *For the current value of ν, reorder the nodes to the order they will be added to the excursion set when the parameter ρ is increased.*
   - *sequentially add nodes to the set D according to the ordering given above and in each step update the probability $\mathsf{P}(D \subseteq A_u^+(\mathbf{x}))$. Stop as soon as this probability falls below $1 - \alpha$.*

- *return the last set D for which* $\mathsf{P}(D \subseteq A_u^+(\mathbf{x})) \geq 1 - \alpha$.

4. $E_{u,\alpha}^+$ *is given by the largest set D found in the optimization over ν.*

The optimization can in this case be done using a Golden section search or a similar fast optimization procedure for one-dimensional problems. Algorithm 3.5 can also be used to estimate uncertainty regions for contour curves by using the following two-parameter family for the pair of level avoiding sets.

Definition 3.6 (Parametric family for level avoiding sets)**.** Let $D_1^+(\rho_1)$ and $D_1^-(\rho_2)$ be given by Definition 3.2. A two-parameter family for the pair of level avoiding sets is obtained as $(D_1^+(\rho_1), D_1^-(\rho_2))$. A one-parameter family is obtained by requiring that $\rho_1 = \rho_2 = \rho$.

The one-parameter family in Definition 3.6 can be used in Algorithm 3.1 to estimate level avoiding sets and uncertainty regions for contour curves without having to use the more computationally expensive Algorithm 3.5.

3.2.1 Domain bounds and reorderings

In the case of a GMRF posterior, it is desirable to make the Cholesky factor of the precision matrix as sparse as possible, because it reduces the number of floating point calculations that have to be done and reduces the error of the estimator. Reordering the nodes according to a parametric family does not guarantee good sparsity of the Cholesky factor, but the reordering can be improved by finding upper and lower bounds for the region.

The simplest upper bound for the region is to use

$$U_1 = \{\mathbf{s} : \mathsf{P}(x(\mathbf{s}) > u) \geq 1 - \alpha\},$$

which is calculated using only the marginal probabilities, and which is the largest region D if $x(\mathbf{s})$ is a perfectly correlated field. The domain D cannot contain any locations $\mathbf{s}$ which are not in U_1 because all points not in U_1 have marginal probabilities lower than $1 - \alpha$ of exceeding the level u.

A simple lower bound for the region is obtained using Boole's inequality as

$$L_1 = \{\mathbf{s} : \mathsf{P}(x(\mathbf{s}) > u) \geq 1 - \alpha/n\}$$

where n is the number of points in the discretization of the domain. In terms of multiple hypothesis testing, this lower bound is obtained from the classical

Bonferroni correction method and an improved lower bound can be obtained using the Holm-Bonferroni method (Holm, 1979) as

$$L_2 = \{\mathbf{s} : p_{(k)} > 1 - \alpha/k\}$$

where $p_{(k)}$ is the kth largest probability in the set $\{\mathsf{P}(x(\mathbf{s}_i) > u), i = 1, \ldots, n\}$. If the stochastic variables $x(\mathbf{s}_i)$ are independent, L_2 is the largest domain D. If the variables are not independent or perfectly correlated, one has $L_2 \subset D \subset U_1$.

The nodes can now be categorized into three classes, the first class contains the nodes included in the upper bound U_1, the second class contains the nodes in the set $L_2 \setminus U_1$ and the third class contains all other nodes. Since one knows that all nodes in U_1 will be included in D, these can be reordered to maximize the sparsity of the Cholesky factor, for example using an approximate minimum degree permutation. The nodes in the second class are then added in the order determined by the parametric family. Finally, since the nodes in the third class will not be included in the domain, these can be reordered to maximize the sparsity or integrated out of the posterior distribution. Making the bounds more precise will improve the sparsity of the problem and therefore reduce the Monte-Carlo error and the computational complexity.

3.3 Probability calculations for the latent Gaussian setting

In practice, we cannot use the computations from the previous sections directly unless we are in a purely Gaussian setting with known parameters. In the latent Gaussian setting with posterior (2), the method has to be modified. Since this is a latent Gaussian setting, Integrated Nested Laplace Approximations (INLA) (Rue et al., 2009) are used to estimate the posterior distributions $\pi(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta})$ and $\pi(\boldsymbol{\theta}|\mathbf{y})$. First we propose three methods for calculating the excursion probabilities, assuming that $\pi(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta})$ is Gaussian:

EB: (Empirical Bayes) Ignore the parameter uncertainty and calculate the probability conditionally on a parameter estimate. That is, estimate the excursion sets under the conditional posterior $\pi(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta}_0)$ where $\boldsymbol{\theta}_0$ for example is the maximum a posteriori estimate or the maximum likelihood estimate of $\boldsymbol{\theta}$.

QC: (Quantile correction) Do a correction to the Gaussian probability calculations based only on the marginal posteriors in the following way. For each

i use the marginal posterior to calculate $\mathsf{P}(x_i > a_i|\mathbf{y})$ and $\mathsf{P}(x_i < b_i|\mathbf{y})$ and calculate $\tilde{a}_i$ and $\tilde{b}_i$ so that $\mathsf{P}(z_i > \tilde{a}_i|\mathbf{y}, \boldsymbol{\theta}_0) = \mathsf{P}(x_i > a_i|\mathbf{y})$ and $\mathsf{P}(z_i < \tilde{b}_i|\mathbf{y}, \boldsymbol{\theta}_0) = \mathsf{P}(x_i < b_i|\mathbf{y})$. An estimate of the probability is then given by $I(\tilde{\mathbf{a}}, \tilde{\mathbf{b}}, \mathbf{Q}(\boldsymbol{\theta}_0))$, where $\mathbf{Q}(\boldsymbol{\theta}_0)$ is the posterior precision matrix for $\mathbf{x}$ given the estimated parameters.

NI: (Numerical integration) Use the same integration strategy as is used in INLA when estimating the marginal posterior distributions. The idea is to numerically approximate the excursion probability by approximating the integral in (2) as

$$\mathsf{P}(\mathbf{a} < \mathbf{x} < \mathbf{b}|\mathbf{y}) = \mathsf{E}(\mathsf{P}(\mathbf{a} < \mathbf{x} < \mathbf{b}|\mathbf{y}, \boldsymbol{\theta})) \approx \sum_{i=1}^{k} w_i \mathsf{P}(\mathbf{a} < \mathbf{x} < \mathbf{b}|\mathbf{y}, \boldsymbol{\theta}_i)$$

where the configuration of the points $\boldsymbol{\theta}_i$ is taken from the integration in INLA and the weights w_i are chosen proportional to $\pi(\boldsymbol{\theta}_i|\mathbf{y})$.

The Empirical Bayes estimator is the simplest, and may be sufficient in many situations. The quantile correction method is based on correcting the limits of the integral so that the probability would be correct if the x_i's were independent. This method is as easy to implement as the empirical Bayes method and should perform better in most scenarios. Finally, the numerical integration strategy is k times more computationally demanding as the probability has to be calculated for each parameter configuration $\boldsymbol{\theta}_i$, but should also be the most exact method. If the number of parameters is small one can often obtain accurate results with only a few parameter configurations, but the accuracy of the estimator will depend on how these configurations are chosen.

A second modification is required if the conditional posterior $\pi(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta}_0)$ is not Gaussian. The simplest solution to this problem is to do a Gaussian approximation $\tilde{\pi}_G(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta}_0)$, for example using Laplace approximations or simplified Laplace approximations as suggested by Rue et al. (2009). If a Gaussian approximation is not sufficient, the sequential integration method has to be modified, and how to do this will depend on the posterior distribution. For example, Genz and Bretz (2009) outline how the quasi Monte Carlo methods can be extended to t-distributions, and the GHK-based particle filter method can be extended to other types of distributions as well.

4 Tests on simulated data

In this section, three examples using simulated data are presented to illustrate the methods and test their accuracy. In the first example, we look at a problem in one dimension with known model parameters, where a latent Gaussian process with an exponential covariance is observed under Gaussian measurement noise. In the second example, we compare the different parametric families for contour uncertainty sets for a model in two dimensions with known parameters, where a latent Gaussian Matérn field is observed under Gaussian measurement noise. In the third example, the same spatial model setup is used, but this time the model parameters are estimated from data and the three methods for handling the full posterior distribution are compared.

4.1 Example 1: 1d Gaussian data with known parameters

We begin with a simple one-dimensional example to illustrate the different sets we have previously defined. Let $x(s)$, $s \in [0, 2]$ be a Gaussian process with an exponential covariance function with scaling parameter $\lambda = 1$ and mean

$$\mu(s) = \begin{cases} s - 0.5 & \text{if } s < 1 \\ 1.5 - s & \text{if } s \geq 1. \end{cases}$$

We generate a trajectory from the model and observe it at 500 locations $s_1, \ldots, s_n$ drawn at random in the interval under Gaussian measurement noise, giving us observations $y_i = x(s_i) + \epsilon_i$ where $\epsilon_i \sim \mathsf{N}(0, \sigma^2)$. We do spatial prediction (kriging) to 1000 equally spaced locations in the interval given the parameters $\boldsymbol{\theta}$ and the measurements $\mathbf{y}$, and then estimate the positive 0-excursion function $F^+(0, s)$ using the parametric family $D_1^+(0, \rho)$. In Figure 1, Panel (a), $F^+(0, s)$ is shown in red together with the marginal excursion probabilities $\mathsf{P}(x(s) > 0)$ in grey.

By the definition of $F^+(0, s)$, the positive 0-excursion set $E_{0,\alpha}^+(x)$, is obtained by calculating the $1 - \alpha$ excursion set of the function $F^+(0, s)$, and this set is shown for $\alpha = 0.05$ in red in Figure 1, Panel (b). The grey set shows the upper bound U_1, which is the set where $\mathsf{P}(x(s) > 0) \geq 1 - \alpha$, and the dark red set shows the Holm-Bonferroni lower bound L_2. The black curve shows the kriging estimate of the process given the data. Note that the grey and red sets are obtained as excursion sets of the grey and red functions in Panel (a), and also note that $L_2 \subset E_{0,\alpha}^+(x) \subset U_1$.

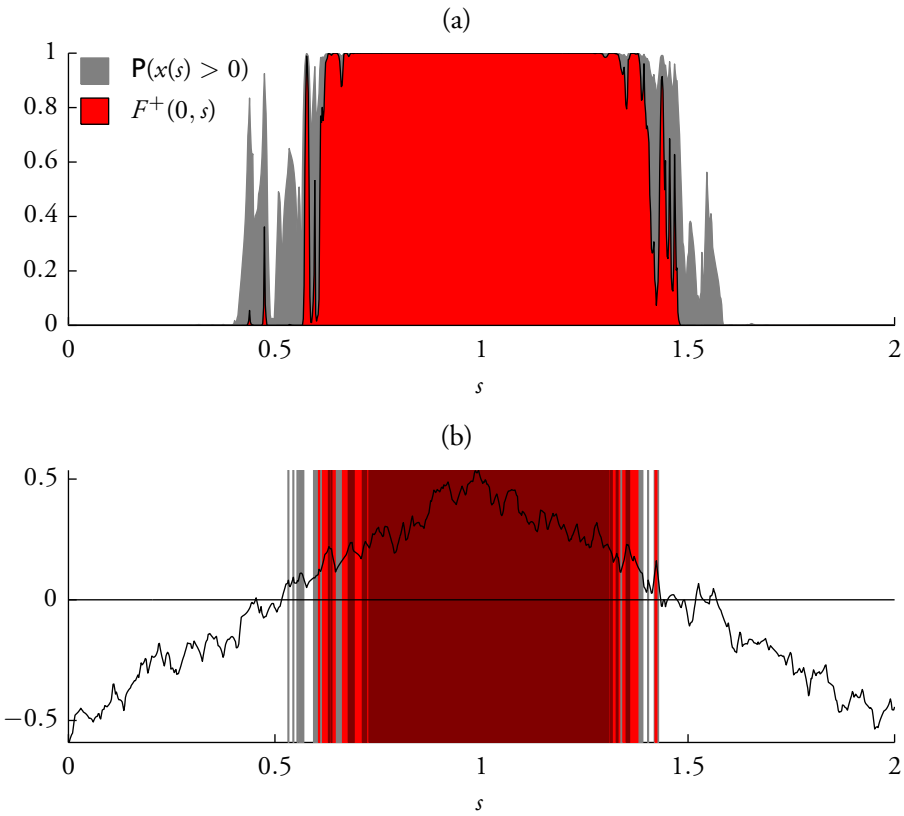

Figure 1: Results from Example 1. Panel (a) shows the excursion function $F^+(0,s)$ (red) and the marginal excursion probabilities $p(s) = \mathsf{P}(x(s) > 0)$ (grey). Panel (b) shows $E^+_{0,0.05}(x)$ in red, obtained as $A^+_{0.95}(F^+)$. The grey area shows $A^+_{0.95}(p)$, which is the upper bound U_1, and the dark red set is the lower bound L_2. The black curve is the kriging estimate of $x(s)$.

We now want to verify that the estimated sets $E^+_{0,\alpha}(x)$ have the correct excursion probability, that is, that $\mathsf{P}(x(s) > 0, s \in E^+_{0,\alpha}(x)) = 1 - \alpha$. To that end, draw N samples, $x_1(s), \ldots, x_N(s)$ from $\pi(x|\mathbf{y}, \boldsymbol{\theta})$, count the number of samples for which $\inf\{x(s), s \in E^+_{0,\alpha}(x)\} \geq 0$, and denote this number by N_s. Further let $\hat{p}(\alpha)$ denote the proportion of samples, N_s/N, that satisfies the requirement. If $E^+_{0,\alpha}(x)$ is correctly estimated, $\hat{p}(\alpha)$ should be close to $1 - \alpha$. In Figure 2, Panel (a), the difference $1 - \alpha - \hat{p}(\alpha)$ is shown as a function of $1 - \alpha$. The difference is

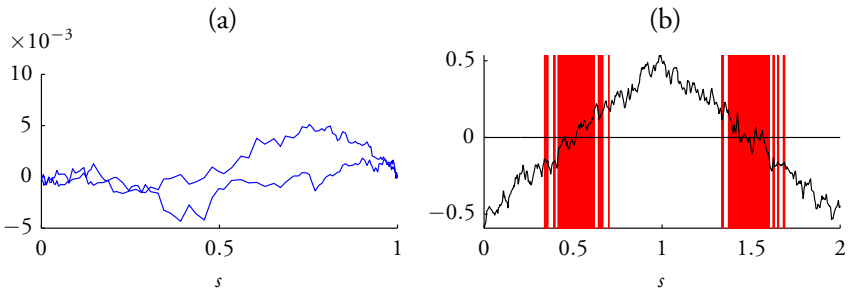

Figure 2: Results from Example 1. Panel (a) shows $1-\alpha-\mathsf{P}(x(s) > 0, s \in E^{+}_{0,\alpha}(x))$ as a function of $1-\alpha$, estimated twice using Monte-Carlo simulation of x. Panel (b) shows the estimated contour uncertainty set $M^{c}_{0,0.05}(x)$.

calculated twice, using two different estimates $\hat{p}(\alpha)$, each based on $N = 50000$ samples. As can be seen in the figure, the difference is very small for all values of α, and the difference that can be seen is mostly due to the Monte-Carlo error in the estimation of $\hat{p}(\alpha)$. Thus the sets $E^{+}_{0,\alpha}$ indeed have the correct excursion probabilities.

Finally in Figure 2, Panel (b), the 0-contour uncertainty region $M^{c}_{0,0.05}(x)$ is shown in red and the kriging estimate of $x(s)$ is again shown in black. The set was estimated using the two-parameter family for level avoidance sets from Definition 3.6 and Algorithm 3.5. The complement of this set is the union of the level avoiding sets $(M^{-}_{0,0.05}(x), M^{+}_{0,0.05}(x))$, which is the largest pair of sets (D^{+}, D^{-}) satisfying $\mathsf{P}(D^{-} \subseteq A^{-}_{u}(x),\ D^{+} \subseteq A^{+}_{u}(x)) \geq 0.95$.

4.2 Example 2: 2d Gaussian data with known parameters

In this example, we change to a spatial model to test the parametric families for contour sets. Let $x(\mathbf{s})$, $\mathbf{s} \in [0,10] \times [0,10]$, be a Gaussian field with a constant mean $\mu = 0$ and a Matérn covariance function

$$C(\|\mathbf{h}\|) = \frac{2^{1-\nu}\phi^2}{(4\pi)^{\frac{d}{2}}\Gamma(\nu+\frac{d}{2})\kappa^{2\nu}}(\kappa\|\mathbf{h}\|)^{\nu}K_{\nu}(\kappa\|\mathbf{h}\|), \tag{6}$$

where ν is a shape parameter, κ^2 a scale parameter, ϕ^2 a variance parameter, K_ν is a modified Bessel function of the second kind of order $\nu > 0$, and $\|\cdot\|$ denotes the Euclidean spatial distance. We use the SPDE representation by Lindgren et al.

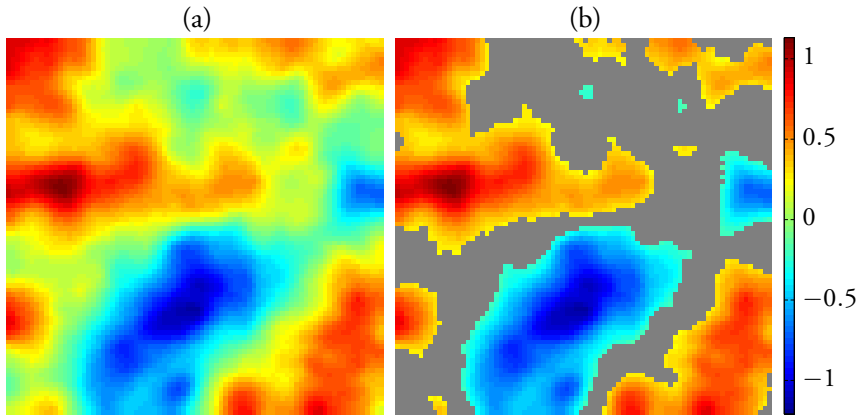

Figure 3: Results from Example 2. Panel (a) shows a kriging estimate and Panel (b) shows the same estimate where the corresponding estimated contour uncertanty set $M^c_{0,0.05}(x)$ is superimposed in grey.

(2011) of the field using a triangulation based on an 80×80 regular lattice in the region. The representation is a piecewise linear approximation $x(\mathbf{s}) \approx \sum_i x_i \varphi_i(\mathbf{s})$ of the field using 6400 piecewise linear functions $\varphi_i(\mathbf{s})$, each centered at one of the nodes in the lattice. The advantage with this representation is that it allows us to do all calculations using the weights $\mathbf{x}$ of the basis expansion, which form a Gaussian Markov random field.

We set $\nu = \phi = 1$, and $\kappa^2 = 0.5$, and generate a sample of the field and measure it at 1000 locations in the square, chosen at random, under Gaussian measurement noise, giving us observations $y_i = x(\mathbf{s}_i) + \epsilon_i$ where $\epsilon_i \sim \mathsf{N}(0, \sigma^2)$ and $\sigma = 0.1$. The posterior estimate (kriging) of $\mathbf{x}|\mathbf{y}$ can be seen in Figure 3, Panel (a), and the uncertainty region $M^c_{0,0.05}(x)$ for the 0-contour can be seen in Panel (b). In this case the $M^c_{0,0.05}(x)$ was estimated using the one-parameter family in Definition 3.6, and it is now of interest to test how much is gained by using the two-parameter family from the same definition instead.

To that end, we generate 50 data sets using the same setup, and for each data set estimate $M^c_{0,0.05}(x)$, first using the one-parameter family $(D_1^+(\rho), D_1^-(\rho))$, and then using the more general two-parameter family $(D_1^+(\rho_1), D_1^-(\rho_2))$. Since the one-parameter family is a special case of the two-parameter family where $\rho_1 = \rho_2 = \rho$, the contour sets estimated with the two-parameter family should always be smaller than the one-parameter sets. However, using the two-parameter

family, the estimated sets are on average only 0.2% smaller than if the one-parameter family is used, so in this case it is arguably not worth the extra computational effort to use the two-parameter family, although for other levels u, or other latent models, the difference might be larger.

4.3 Example 3: 2d Gaussian data with unknown parameters

In this example we compare the three methods,described in Section 3.3, for handling the full posterior distribution (2) in the calculations. The same Gaussian Matérn model is used as in Example 2, with the difference that we now also estimate the parameters from the data.

We set $\nu = \phi = 1$, and $\kappa^2 = 2$ in the covariance function (6), and generate a sample of the field and measure it at 1000 locations in the square, chosen at random, under Gaussian measurement noise, giving us observations $y_i = x(\mathbf{s}_i)+\epsilon_i$ where $\epsilon_i \sim \mathsf{N}(0, \sigma^2)$ and $\sigma = 0.5$. Given the measurements, we estimate the parameters and the marginal posterior distributions using INLA. The posterior estimate (kriging) of $\mathbf{x}|\mathbf{y}$ can be seen in the lower right panel of Figure 4, and in the lower left panel, the marginal probabilities $\mathsf{P}(x(\mathbf{s}) > 0|\mathbf{y})$ are shown.

We now estimate the excursion function $F^+(0, \mathbf{s})$ using the three different methods described in Section 3.3 and the one-parameter family from Definition 3.2 for the excursion sets. These can be seen in the upper panels of Figure 4. Visually it is in this case difficult to see any differences between the three estimates of the excursion function. To compare the accuracy of the estimates we will do a simular comparison to the one performed in Example 1, where Monte-Carlo simulation was used to estimate $\hat{p}(\alpha)$, the proportion of samples satisfying $\inf\{x(\mathbf{s}), \mathbf{s} \in E^+_{0,\alpha}(x)\} \geq 0$, which should be close to $1 - \alpha$ if $E^+_{0,\alpha}(x)$ is correct.

There are three possible sources of errors in this comparison. The first one is the Monte-Carlo error from the estimation of $\hat{p}(\alpha)$, which has nothing to do with the accuracy of the method. The second error is the Monte-Carlo error in the probability estimation when estimating the excursion distribution functions. This error is, however many orders of magnitude smaller in this case. The final error is the approximation error induced by using any of the three methods EB, QC, or NI for handling the full posterior distribution.

To investigate this approximation error, the difference $1 - \alpha - \hat{p}(\alpha)$ is estimated for the three estimates of $F^+(0, \mathbf{s})$. First we base the estimate on 20000 samples from the posterior $\pi(\mathbf{x}|\mathbf{y})$, obtained using the MCMC sampler described in Appendix A. In Figure 5, Panel (a), the results can be seen for the EB method

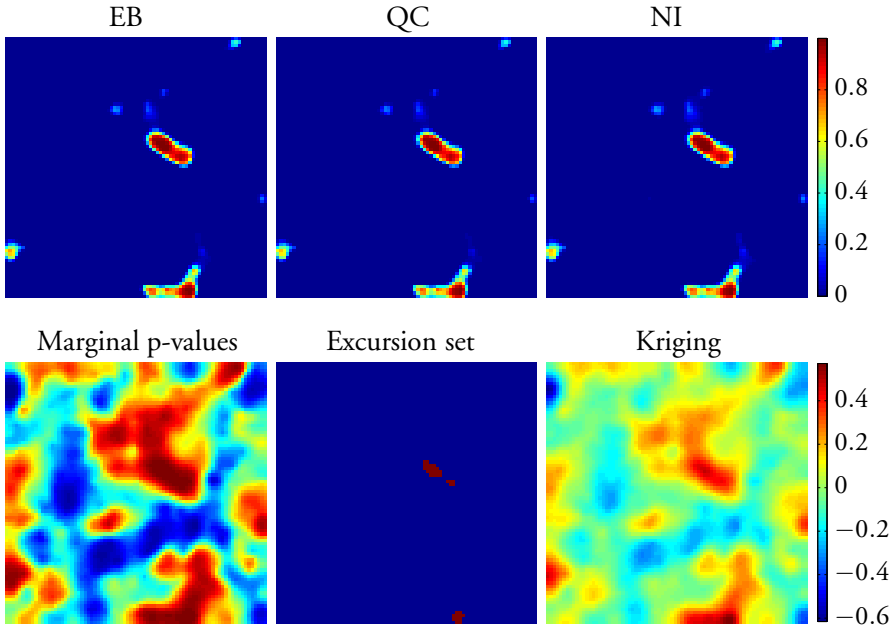

Figure 4: Results from Example 3. In the top row, three estimates of the excursion function can be seen using the EB method (left), the QC method (middle), and the NI method with 15 parameter configurations (right). In the bottom row, the marginal p-values for exceeding the limit can be seen in the left panel, using the same color scale as for the top row. The middle panel shows the set $E^{+}_{0,0.05}(x)$ given by excursion function estimated by the NI method. Finally the right panel shows the kriging estimate of the latent field.

(blue), the QC method (green), the NI method with $k = 45$ parameter settings (red), and the NI method with $k = 15$ parameter settings (cyan). The comparison was done twice, with two different samples of size 20000 when calculating $\hat{p}(\alpha)$, and the curves of the same color show these two and give an indication of the size of the Monte-Carlo error in the comparison. As seen in the figure, the NI method performs best, as expected.

The error using the NI method comes from the fact that only finitely many points are used in the integration when approximating the posterior distribution for the parameters. That is, the full posterior $\pi(\boldsymbol{\theta}|\mathbf{y})$ is approximated by a discrete distribution with point masses at the parameter configurations $\boldsymbol{\theta}_i$ used in the

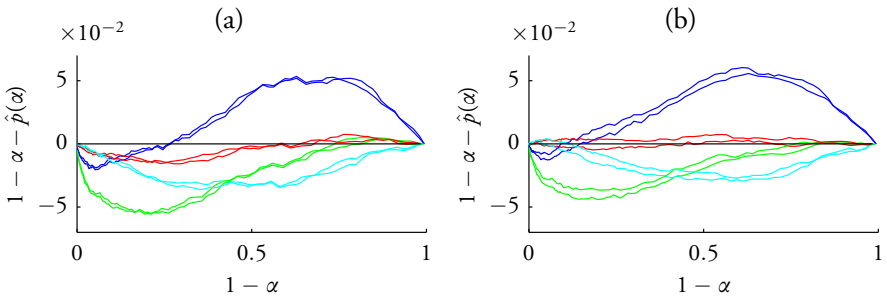

Figure 5: Results from Example 3 showing the difference $1-\alpha-\hat{p}(\alpha)$ as a function of $1-\alpha$ for the different approximation methods, EB (blue), QC (green), NI using 45 parameter configurations (red), and NI using 15 parameter configurations (cyan). $\hat{p}(\alpha)$ is an estimate of $\mathsf{P}(x(\mathbf{s})>0, \mathbf{s}\in E^{+}_{0,\alpha}(x))$ based on MCMC simulation of $x(\mathbf{s})$, which should be close to $1-\alpha$ if $E^{+}_{0,\alpha}(x)$ is correctly estimated. In Panel (a), $\hat{p}(\alpha)$ is estimated using the full posterior distribution, and in Panel (b), $\hat{p}(\alpha)$ is estimated using the discrete posterior distribution defined by the 45 parameter configurations from the NI method. The comparison was done twice, with the two different estimates of $\hat{p}(\alpha)$, each based on 20000 samples of $x(\mathbf{s})$, and the curves of the same color shows these two.

integration, $\pi(\boldsymbol{\theta}_i) = w_i$. To verify that this indeed is the source of the error in, we construct a second Monte-Carlo sampler where we instead of sampling $\boldsymbol{\theta}$ from the full posterior $\pi(\boldsymbol{\theta}|\mathbf{y})$ sample it from the discrete distribution defined by the 45 parameter configurations used in the first NI method. Panel (b) in Figure 5 shows the same comparison as Panel (a) but where $\boldsymbol{\theta}$ is sampled from the discrete distribution. As expected, the error for the NI method with 45 parameter configurations is now smaller.

The Monte-Carlo error from estimating $\hat{p}(\alpha)$ is quite large in Figure 5, so to get a better understanding of the other errors, a larger study was also performed where the procedure in Figure 5 was repeated 50 times for 50 different simulated data sets, and for each data set $N = 60000$ draws from the posterior was used when estimating $\hat{p}(\alpha)$. The average errors of these 50 runs can be seen in Figure 6. In Panel (a) the results using samples from the full posterior is shown, and in Panel (b) the results using the discrete distribution for $\boldsymbol{\theta}$ is shown. Note that the red curve is very close to zero in Panel (b), indicating that the error in the NI method mostly depends on choosing the integration points for $\boldsymbol{\theta}$ so that

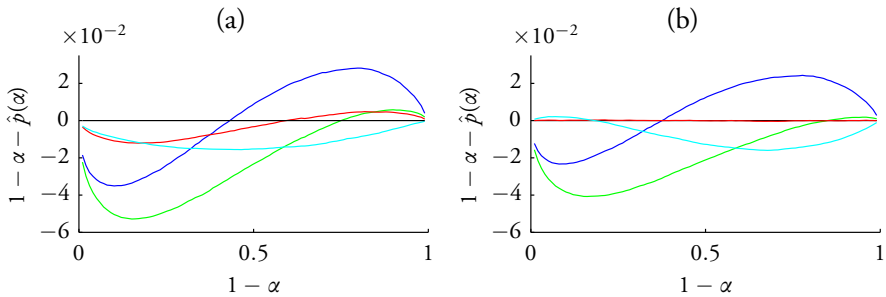

Figure 6: Results from Example 3 showing the difference $1-\alpha-\hat{p}(\alpha)$ as a function of $1-\alpha$ for the different approximation methods, EB (blue), QC (green), NI using 45 parameter configurations (red), and NI using 15 parameter configurations (cyan). The same setup as in Figure 5 was repeated 50 times for 50 different datasets and 60000 samples were used when estimating $\hat{p}(\alpha)$. This figure shows the average error of these 50 runs.

they capture the true posterior distribution well. Also note that the QC method performs well for large values of $1-\alpha$, and since one most often is interested in finding the excursion sets for small values of α, this method is then a good way of finding such sets with less computational effort than using the NI method.

5 Applications

In this section, we will use the techniques described above in two different applications. In the first, we study air pollution data from Piemonte region in northern Italy and estimate regions where the daily limit for PM_{10} (particulate matter with an aerodynamic diameter of less than 10 μm) is exceeded. In the second application, we study vegetation index data from the African Sahel and estimate areas that experienced a significant increase in vegetation after the drought period in the early 1980's. In both of these applications the data sets are large, and the Markov structure of the latent Gaussian models has to be used in the calculations.

5.1 Air pollution data

High levels of air pollution can be harmful for the ecosystems and the human health. The effects on human health ranges from minor effects to the cardio-

respiratory system to premature mortality (Cohen et al., 2009, Cameletti et al., 2012). Because of this, environmental agencies have to assess the air quality in order to take proper actions for improving the situation in polluted areas, and an important tool in this process is the ability to produce continuous maps of air pollution.

A region where the daily limit values fixed by the European Union for human health protection (see EU Council Directive 1999/30/EC) are periodically exceeded is the Piemonte region in northern Italy. Recently, Cameletti et al. (2012) proposed a statistical model to capture the complex spatio-temporal dynamics of PM_{10} concentration in the region and used it to produce daily maps of PM_{10}. They also produced daily maps of exceedance probabilities of the value $50\mu g/m^3$, which is the value fixed by the European directive 2008/50/EC for the daily mean concentration that cannot be exceeded more than 35 days in a year. These probability maps only considered the marginal excursion probabilities, and no attempts of producing maps of simultaneous exceedance probabilities were made. In the following, we will therefore consider the same model and data but also estimate the excursion functions for the $50\mu g/m^3$ limit value.

Cameletti et al. (2012) considers daily PM_{10} data measured at 24 monitoring stations by the Piemonte monitoring network during 182 days in the period October 2005 - March 2006. Denoting the measurements made at location $\mathbf{s}_i$ at time t by $y(\mathbf{s}_i, t)$, the following measurement equation is assumed,

$$y(\mathbf{s}_i, t) = x(\mathbf{s}_i, t) + \epsilon(\mathbf{s}_i, t), \tag{7}$$

where $\epsilon(\mathbf{s}_i, t) \sim \mathsf{N}(0, \sigma_\epsilon^2)$ is Gaussian measurement noise, both spatially and temporally uncorrelated, and $x(\mathbf{s}_i, t)$ is the latent field of true unobserved air pollutions. The latent field is assumed to be on the form

$$x(\mathbf{s}_i, t) = \sum_{k=1}^{p} z_k(\mathbf{s}_i, t)\beta_k + \xi(\mathbf{s}_i, t), \tag{8}$$

where the $p = 9$ covariates z_k are used and ξ is a spatio-temporal Gaussian random field. Based on the work of Cameletti et al. (2011) the following covariates were used: *1*) Daily mean wind speed; *2*) daily maximum mixing height; *3*) daily precipitation; *4*) daily mean temperature; *5*) daily emissions; *6*) altitude; *7*) longitude; *8*) latitude; and *9*) intercept. These covariates are provided with hourly temporal resolution on a 4 km $\times$ 4 km regular grid by the environmental agency

of Piemonte region (Arpa Piemonte). The spatio-temporal process ξ is assumed to follow first order autoregressive dynamics in time with spatially dependent innovations:

$$\xi(\mathbf{s}_i, t) = a\xi(\mathbf{s}_i, t-1) + \omega(\mathbf{s}_i, t), \tag{9}$$

where $|a| < 1$ and $\omega(\mathbf{s}_i, t)$ is a zero-mean temporally independent Gaussian process characterized by the spatio-temporal covariance function

$$\mathrm{Cov}(\omega(\mathbf{s}_i, t_1), \omega(\mathbf{s}_j, t_2)) = \begin{cases} 0 & \text{if } t_1 \neq t_2 \\ C(\|\mathbf{s}_i - \mathbf{s}_j\|) & \text{otherwise,} \end{cases} \tag{10}$$

where $C(\cdot)$ is a Matérn covariance function given by (6). The model parameters and the posterior distribution for the latent field are then estimated using INLA in combination with the SPDE representation of Lindgren et al. (2011), see Cameletti et al. (2012) for details.

The map of marginal excursion probabilities for the level $50\mu g/m^3$ for January 30, 2006, based on the estimated posterior distribution for x, can be seen in the left panel of Figure 7. To avoid inappropriate linear extrapolation of the effect of elevation beyond the range of the elevation of the observations, the results are only shown for areas below 1km. Based on these results, we now estimate the positive excursion function for the level $50\mu g/m^3$, $F^+(50, \mathbf{s})$, using the NI method from Section 3.3 and the parametric family of excursion sets from Definition 3.2. A total of 25 parameter configurations are used in the integration. The result can be seen in the right panel of Figure 7. As seen in the figure, there are three regions where the level is clearly exceeded, and a fourth that possibly contains too high pollution levels. As expected, these areas coincide with the locations of the main metropolitan areas in the region; Turin, Novara, Vercelli, and Alessandria. In this case, it would have been desirable to make the predictions on a finer spatial scale, but since the covariates were given on a 4 km $\times$ 4 km grid, this spatial resolution had to be used in the prediction.

To get a better understanding of the results, it is also of interest to find the regions where the pollution level is simultaneously below the limit value with some given probability. The marginal probabilities for being below be the level $50\mu g/m^3$, based on the estimated posterior distribution for x, can be seen in the left panel of Figure 8. The results are again only shown for areas below 1km altitude. Using the same method as for the positive excursion function, we now

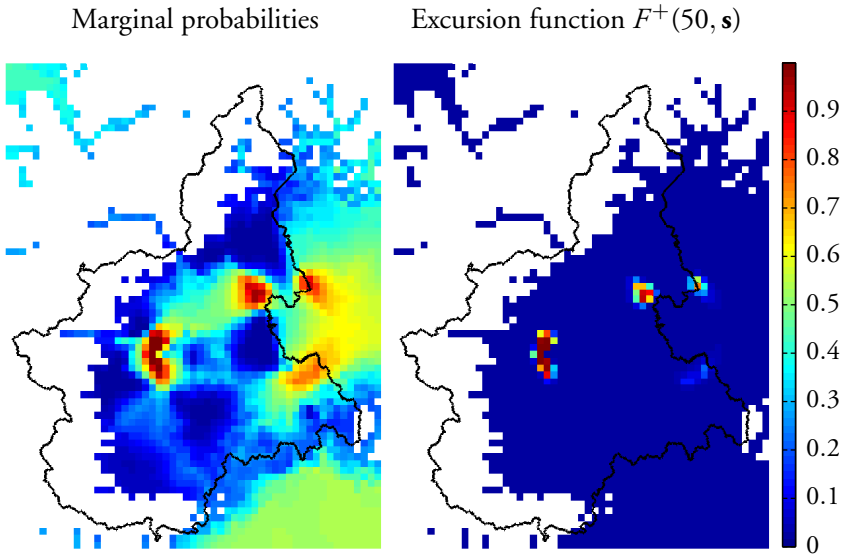

Figure 7: Results from the PM_{10} application for January 30, 2006. A map of the marginal exceedance probabilities for $50\mu g/m^3$ (left), and the joint excursion distribution function for the level (right).

estimate the negative excursion function for the level $50\mu g/m^3$, $F^-(50, \mathbf{s})$. The result can be seen in the right panel of Figure 8.

Note that the union of $E^+_{50,0.1}(x)$ and $E^-_{50,0.1}(x)$ covers only a small part of the region, indicating that the uncertainty in the problem is large. See the red and blue sets in the left panel of Figure 9. Also, by taking the complement of the set $E^-_{50,0.1}(x)$, we get the region that contains all exceedances of the level with certainty 0.9, indicated in grey in the left panel of Figure 9. This set is large, indicating that there are many regions where the level possibly is exceeded. Hence, it is important to note that the positive excursion set $E^+_{50,0.1}(x)$ is small because the uncertainty is large in the problem, and not because the other regions certainly have concentrations below the level.

To verify that the uncertainty is large, we finally calculate the contour function for the level $50\mu g/m^3$, $F^c(50, \mathbf{s})$, using the NI method and the one-parameter family from Definition 3.6 for the pair of level avoiding sets. The result can be seen in the right panel of Figure 9, and it can be seen that the uncertainty regions for the countour curve indeed cover a large part of the region.

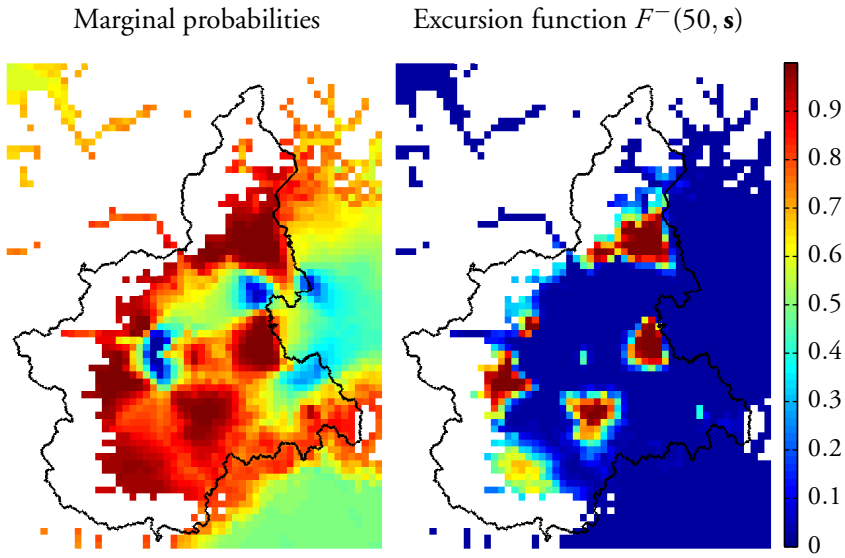

Figure 8: Results from the PM_{10} application for January 30, 2006. A map of the marginal probabilities for the field being below the level $50\mu g/m^3$ (right), and the joint negative excursion distribution function for the level (left).

5.2 Spatially dependent temporal trends in vegetation data

Trends in vegetation cover are related to changes in climatic drivers, feedback mechanisms between the atmosphere and land surface, and human interaction. A region with rapid recent changes is the African Sahel. This zone has received much attention regarding desertification and climatic variations (Olsson, 1993, Nicholson, 2000, Lamb, 1982). Recently, Eklundh and Olsson (2003) observed a strong increase in seasonal vegetation index over parts of the Sahel using Advanced Very High Resolution Radiometer (AVHRR) data from the NOAA/NASA Pathfinder AVHRR Land (PAL) database (Agbu and James, 1994, James and Kalluri, 1994), for the period 1982-1999. The study was based on ordinary least squares linear regression on individual time series extracted for each pixel in the satellite images. The results of Eklundh and Olsson (2003) were later improved by Bolin et al. (2009) where a spatial model for the vegetation was used in the analysis to capture the spatial dependencies in the trend estimation.

To find regions where changes in the vegetation have occurred over the course

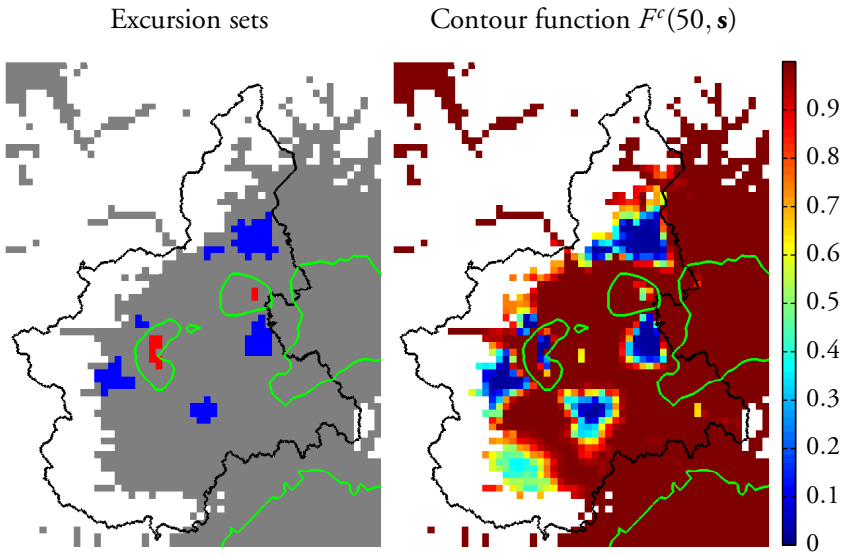

Figure 9: Results from the PM_{10} application for January 30, 2006. In the left panel, the set $E^{+}_{50,0.1}(x)$ is shown in red, $E^{-}_{50,0.1}(x)$ in blue, and it's complement $E^{+}_{50,0.1}(x)^c$ in grey. The contour curve for the level $50\mu g/m^3$ is shown in green. The right panel shows the contour function for the level $50\mu g/m^3$, $F^c(50, s)$.

of the studied time period, both Eklundh and Olsson (2003) and Bolin et al. (2009) used significance testing for the individual pixels in the field. Thus, pixels that individually had significant changes in vegetation were found, but no attempts were made to find simultaneous excursion regions. Here, we will use a similar model to that of Bolin et al. (2009) but also estimate joint excursion regions for the vegetation trends.

Assume that the vegetation measurements year t are generated as,

$$\mathbf{Y}_t|\mathbf{X}_t, \mathbf{\Sigma}_{\varepsilon_t} \in \mathsf{N}(\mathbf{A}_t\mathbf{X}_t, \mathbf{\Sigma}_{\varepsilon_t}),$$

where $\mathbf{X}_t$ is the latent vegetation field with prior distribution $\pi(\mathbf{X}_t)$, $\mathbf{\Sigma}_{\varepsilon_t}$ is a measurement noise covariance matrix, and $\mathbf{A}_t$ is an observation matrix determining which pixels in the field that are observed. To estimate time varying trends in the observations, $\mathbf{X}$ is restricted to follow a field of spatially varying linear trends:

$$\mathbf{X}_t = \mathbf{K}_1 + t\mathbf{K}_2 \tag{11}$$

The prior distribution for $\mathbf{K} = [\mathbf{K}_1^\top, \mathbf{K}_2^\top]^\top$ is obtained by evaluating the joint distribution for $\mathbf{X} = [\mathbf{X}_1^\top, \ldots, \mathbf{X}_T^\top]^\top$ conditionally on the restriction (11). We choose a second-order polynomial IGMRF (Rue and Held, 2005, Section 3.4.2) prior for $\mathbf{X}$ and then calculate the corresponding prior distribution for $\mathbf{K}$, (see Bolin et al., 2009, for details).

To complete the model, the structure of $\mathbf{\Sigma}_\varepsilon$ needs to be determined. Many of the factors that the measurement noise should model are local phenomena, such as aerosol and cloud cover. Since it seems unreasonable that the scale of these disturbances would be the same over the entire region, Bolin et al. (2009) assumed that the measurement noise was uncorrelated with a different noise variance at each pixel in the field. This results in a large number of parameters for the measurements noise, one for each pixel in the field, so here we instead use a different slightly simplified noise model. We divide the region into five different land cover categories using the Africa Land Cover Characteristics Data Base Version 2.0 (`http://edc2.usgs.gov/glcc/glcc.php`): *1*) Bare desert; *2*) Semi desert; *3*) Savanna; *4*) Crops, grass, and shrubs; and *5*) Forests and wetlands. The measurement noise variance at pixel $\mathbf{s}_i$ is then modeled as

$$\log \sigma^2(\mathbf{s}_i) = \sum_{k=1}^{5} \theta_k b_k(\mathbf{s}_i), \tag{12}$$

where $b_k(\mathbf{s})$ is the spatial basis function with values equal to the proportion of vetetation type k at each pixel $\mathbf{s}$. The parameters of the model are thus the scale parameter κ and the five measurement noise parameters $\theta_1, \ldots, \theta_5$. A gamma prior is assumed for κ and gaussian priors are used for θ_k.

We choose to study the western part of the Sahel region, and this area is divided into 35463 pixels of size 8 km $\times$ 8 km, so the field $\mathbf{K}$ has 70926 elements, and there are 547832 measurements from 17 years of data starting in 1982 and ending in 1999.

The model parameters and the marginal posterior distributions are estimated using the INLA framework and the excursion sets are estimated using the QC method from Section 3.3. The results can be seen in Figure 10 and Figure 11. The top panel in Figure 10 shows the posterior estimates of the intercepts, $\mathbf{K}_1$, and the slopes, $\mathbf{K}_2$, is shown in the bottom panel. As expected, intercepts are larger in the savanna regions to the south, and smaller in the semi desert regions to the north. The top panel of Figure 11 shows the estimated standard deviation of the measurement noise using model (12). It is worth noting that these results

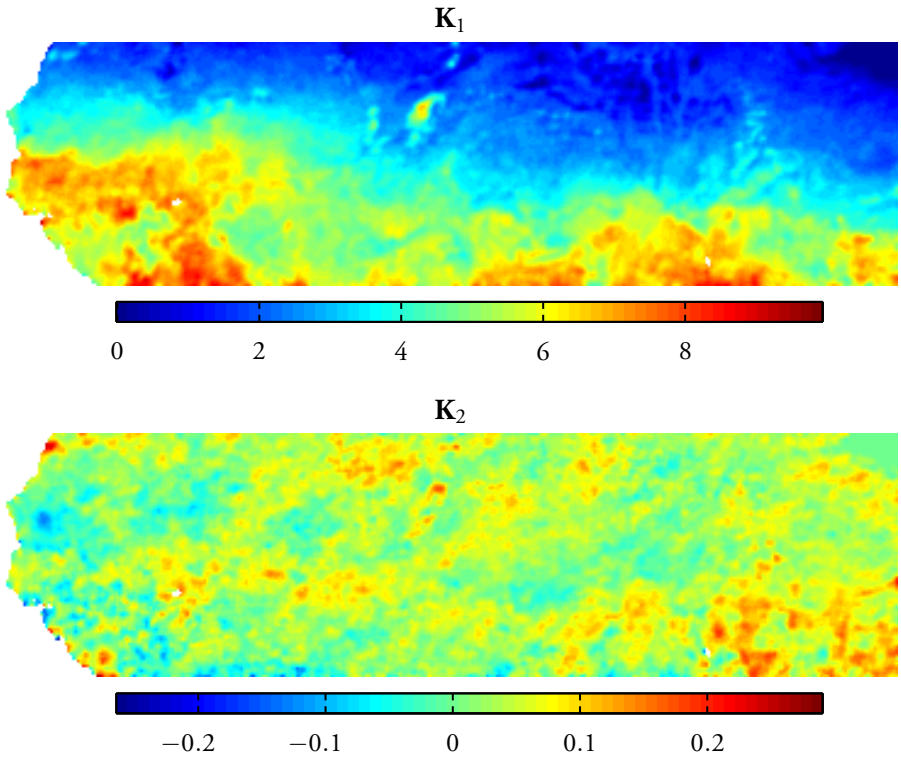

Figure 10: Results from the Sahel vegetation data. The top panel shows the posterior estimates of the regression intercepts, $\mathbf{K}_1$, and the botom panel shows the estimated slopes, $\mathbf{K}_2$.

look reasonable, with larger measurement errors in the coastal region and where there are forests and wetlands, and smaller measurement errors in desert and semi desert regions. Finally the bottom panel of Figure 11 shows the estimated excursion set $E^{+}_{0,0.05}(\mathbf{K}_2)$ in red and the point-wise positive significant trends in green. The interpretation of the result is that one with high certainty can conclude that the areas indicated in red have experienced an increase in vegetation over the studied time period. Hence, conclusions drawn by Eklundh and Olsson (2003) seem valid, also when taking the spatial dependency of the vegetation into account and when estimating the excursion sets controlling the family-wise error.

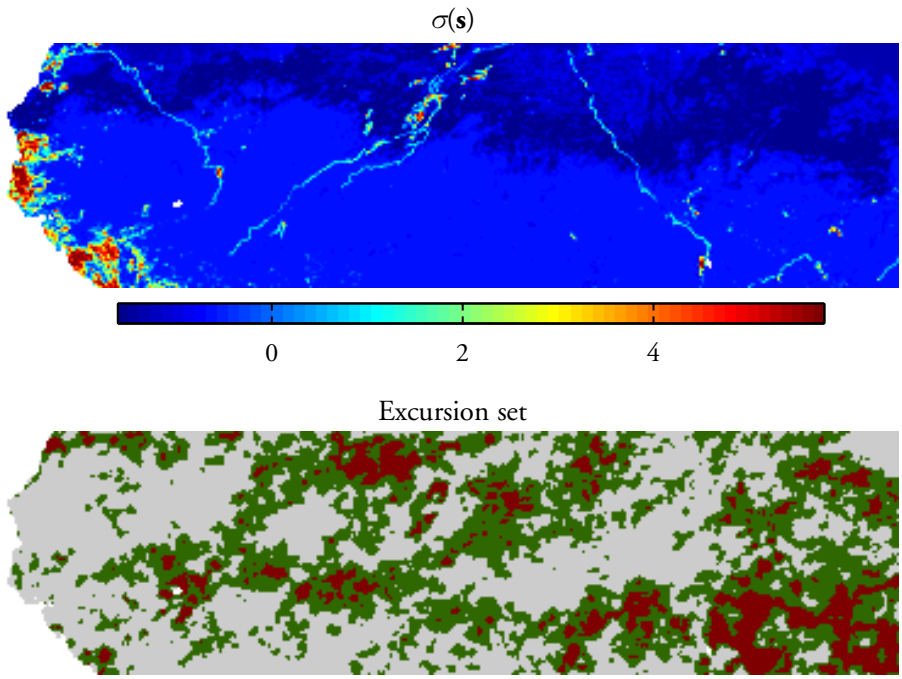

Figure 11: Results from the Sahel vegetation data. The top panel shows the estimated standard deviation of the measurement noise and the bottom panel shows the estimated excursion set $E^{+}_{0,0.05}(\mathbf{K}_2)$ in red and the point-wise positive significant trends in green.

6 Discussion

Estimating excursion sets and uncertainty regions for contour curves for stochastic fields are difficult problems. In this work, we have presented a method for calculating such sets for latent Gaussian models. The main idea is to use a parametric family for the excursion sets in combination with a sequential integration method to reduce the computational effort required when estimating the sets in practise. Tests on simulated data showed that the method is accurate, and two applications were presented to show that the method is applicable even to large environmental problems.

There are a number of extensions that could be made to this work. First of all, using the one-parameter family for the excursion sets gives a method that falls

into the broad category of p-value thresholding methods for estimating simultaneous excursion sets. It would therefore be interesting to do a comparison with other similar methods with respect to the accuracy and computational complexity. Another interesting comparison would be to compare the uncertainty regions for contour curves produced by these methods to those of Lindgren and Rychlik (1995). One could potentially also combine these methods with the work by Polfeldt (1999) to make statements on the quality of contour maps.

We also presented other parametric families that can be used to obtain more complicated methods for estimating the excursion sets, with the possibility of finding more precise estimates under the cost of higher computational complexity. Initial comparisons showed that there is not much gain in using these more complicated methods, but so far these comparisons have only been made using fairly simple latent models, and the gain is likely higher when the latent models are more complex. Hence, more studies are required to verify if this is the case and to investigate in what situations it is appropriate to use the simple one-parameter families.

As the method is valid for the same class of models as the INLA framework is used for, it is our intention to integrate these two in order to minimize the coding effort required of the user for using these methods in practice. Therefore, our main focus at the moment is to implement these methods in the R-INLA package (`r-inla.org`).

Acknowledgements

The authors are grateful to Johan Lindström and Daniel Simpson for valuable discussions on the subject of excursions and contour curve uncertainty sets, and to Peter Guttorp for highlighting the need for a thourough treatment of the subject.

A Notes on the MCMC algorithm used in Example 3

To generate samples from the posterior distribution $\pi(\mathbf{x}|\mathbf{y})$ in Example 3, an MCMC algorithm is used. The algorithm is a random-walk Metropolis Hastings algorithm (Metropolis et al., 1953, Hastings, 1970) with proposal kernel

$$q(\{\mathbf{x}_{\text{old}}, \boldsymbol{\theta}_{\text{old}}\}, \{\mathbf{x}, \boldsymbol{\theta}\}) = \pi(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta}) q_{\theta}(\boldsymbol{\theta}_{\text{old}}, \theta).$$

A new proposal of the parameters $\boldsymbol{\theta} = (\log(\sigma), \log(\kappa), \log(\phi))$ is proposed based on the old value $\boldsymbol{\theta}_{\text{old}}$ using $\boldsymbol{\theta} \sim \mathsf{N}(\boldsymbol{\theta}_{\text{old}}, \boldsymbol{\Sigma}_\theta)$. Here $\boldsymbol{\Sigma}_\theta$ is a scaled version of the Hessian matrix evaluated at the maximum posterior estimate of $\boldsymbol{\theta}$. The scaling is selected as suggested by Gelman et al. (1996). A new value for $\mathbf{x}$ is then proposed using the marginal posterior distribution $\pi(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta})$ given by $\mathbf{x}|\mathbf{y}, \boldsymbol{\theta} \sim \mathsf{N}(\frac{1}{\sigma^2}\hat{\mathbf{Q}}^{-1}\mathbf{A}^\top\mathbf{y}, \hat{\mathbf{Q}})$. Here $\hat{\mathbf{Q}} = \mathbf{Q} + \frac{1}{\sigma^2}\mathbf{A}^\top\mathbf{A}$, where $\mathbf{Q}$ is the precision matrix for $\mathbf{x}$ and $\mathbf{A}$ is an observation matrix determined by the measurement locations. The acceptance probability simplifies to

$$\alpha_{\text{MCMC}} = \min\left(1, \frac{\pi(\boldsymbol{\theta}_{\text{old}}|\mathbf{y})}{\pi(\boldsymbol{\theta}|\mathbf{y})}\right),$$

where the posterior $\pi(\boldsymbol{\theta}|\mathbf{y})$ is given by

$$\pi(\boldsymbol{\theta}|\mathbf{y}) \propto \frac{|\mathbf{Q}|^{\frac{1}{2}}\pi(\boldsymbol{\theta})}{|\hat{\mathbf{Q}}|^{\frac{1}{2}}|\sigma\mathbf{I}|} \exp\left(\frac{1}{2\sigma^2}\mathbf{y}^\top\left(\frac{\mathbf{A}\hat{\mathbf{Q}}^{-1}\mathbf{A}^\top}{\sigma^2} - \mathbf{I}\right)\mathbf{y}\right). \tag{13}$$

Since the proposal for $\mathbf{x}$ does not affect the acceptance probability, a new proposal for $\mathbf{x}$ is only generated if $\boldsymbol{\theta}$ is accepted. With only three parameters in the model, we achieve good mixing this way, but being an MCMC-procedure it is still highly computationally demanding since the calculation of the acceptance probability requires a few Cholesky factorizations and back substitutions based on the posterior precision matrix for $\mathbf{x}$.

References

Adler, R. J. (1981). *The Geometry of Random Fields*. Wiley, New York.

Agbu, P. and James, M. (1994). *The NOAA/NASA Pathfinder AVHRR Land Data Set User's Manual*. Goddard Distributed Active Archive Center, NASA, Goddard Space Flight Center, Greenbelt.

Beaky, M. M., Scherrer, R. J., and Villumsen, J. V. (1992). Topology of large-scale structure in seeded hot dark matter models. *Astrophys. J.*, 387:443–448.

Bolin, D., Lindström, J., Eklundh, L., and Lindgren, F. (2009). Fast estimation of spatially dependent temporal vegetation trends using Gaussian Markov random fields. *Comput. Statist. and Data Anal.*, 53:2885–2896.

Cameletti, M., Ignaccolo, R., and Bande, S. (2011). Comparing spatio-temporal models for particulate matter in piemonte. *Environmetrics*, 22(8):985–996.

Cameletti, M., Lindgren, F., Simpson, D., and Rue, H. (2012). Spatio-temporal modeling of particulate matter concentration through the SPDE approach (submitted).

Cohen, M. A., Adar, S. D., Allen, R. W., Avol, E., Curl, C. L., Gould, T., Hardie, D., Ho, A., Kinney, P., Larson, T. V., Sampson, P., Sheppard, L., Stukovsky, K. D., Swan, S. S., Liu, L.-J. S., and Kaufman, J. D. (2009). Approach to estimating participant pollutant exposures in the Multi-Ethnic Study of Atherosclerosis and air pollution (MESA air). *Environ. Sci. Technol.*, 43(13):4687–4693.

Doucet, A., De Freitas, N., and Gordon, N. (2001). *Sequential Monte Carlo Methods in Practice*. Springer.

Eklundh, L. and Olsson, L. (2003). Vegetation index trends for the African Sahel 1982–1999. *J. Geophys. Res.*, 30:1430–1433.

Furrer, R., Knutti, R., Sain, S. R., Nychka, D., and A., M. G. (2007). Spatial patterns of probabilistic temperature change projections from a multivariate bayesian analysis. *Geophys. Res. Lett.*, 34.

Gelman, A., Roberts, G. O., and Gilks, W. R. (1996). Efficient Metropolis jumping rules. *Bayesian Statist.*, 5:599–607.

Genz, A. (1992). Numerical computation of multivariate normal probabilities. *J. Comput. Graph. Statist.*, 1(2):pp. 141–149.

Genz, A. and Bretz, F. (2002). Comparison of methods for the computation of multivariate t probabilities. *J. Comput. Graph. Statist.*, 11(4):pp. 950–971.

Genz, A. and Bretz, F. (2009). *Computation of Multivariate Normal and t Probabilities*, volume 195 of *Lecture Notes in Statistics*. Springer.

Genz, A. and Kahaner, D. (1986). The numerical evaluation of certain multivariate normal integrals. *J. Comput. Appl. Math.*, 16:255–258.

Geweke, J. (1991). Efficient simulation from the multivariate normal and student-t distributions subject to linear constraints and the evaluation of constraint probabilities.

Gibson, G. J., Glasbey, C., and D.A., E. (1994). Monte Carlo evaluation of multivariate normal integrals and sensitivity to variate ordering. In Dimov, I., Sendov, B., and Vassilevski, P., editors, *Advances in Numerical Methods and Aplications*, pages 120–126. World Scientific Publishing, River Edge.

Hajivassiliou, V. (1991). Simulation estimation methods for limited dependent variable models. Cowles Foundation Discussion Papers 1007, Cowles Foundation for Research in Economics, Yale University.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57:97–109.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scand. J. Statist.*, 6(2):pp. 65–70.

James, M. and Kalluri, S. (1994). The Pathfinder AVHRR Land data set: An improved coarse resolution data set for terrestrial monitoring. *Internat. J. Remote Sensing*, 15:3347–3363.

Keane, M. (1993). 20 simulation estimation for panel data models with limited dependent variables. In *Econometrics*, volume 11 of *Handbook of Statistics*, pages 545 – 571. Elsevier.

Lamb, P. (1982). Persistence of Subsaharan drought. *Nature*, 299:46–47.

Lindgren, F., Rue, H., and Lindström, J. (2011). 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.

Lindgren, G. and Rychlik, I. (1995). How reliable are contour curves? confidence sets for level contours. *Bernoulli*, 4(1):301–319.

Marchini, J. and Presanis, A. (2003). Comparing methods of analyzing fMRI statistical parametric maps. *NeuroImage*, 22:1203–1213.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equations of state calculations by fast computing machines. *J. Chem. Phys.*, 21:1087–1092.

Nicholson, S. (2000). Land surface process and Sahel climate. *Rev. of Geophys.*, 38:117–140.

Olsson, L. (1993). On the causes of famine – drought, desertification and market failure in the Sudan. *Ambio*, 22:395–403.

Polfeldt, T. (1999). On the quality of contour maps. *Environmetrics*, 10:785–790.

Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields; Theory and Applications*, volume 104 of *Monographs on Statistics and Applied Probability*. Chapman & Hall/CRC.

Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations (with discussion). *J. Roy. Statist. Soc. Ser. B Stat. Methodol.*, 71(2):319–392.

Sain, S. R., Furrer, R., and Cressie, N. (2011). A spatial analysis of multivariate output from regional climate models. *Ann. Appl. Statist.*, 5:150–175.

Schervish, M. J. (1984). Algorithm as 195: Multivariate normal probabilities with error bound. *J. Roy. Statist. Soc. Ser. C Appl. Statist.*, 33(1):pp. 81–94.