Scalable Bayesian spatial analysis with Gaussian Markov random fields

Per Sidén
At the Faculty of Arts and Sciences at Linköping University, research and doctoral studies are carried out within broad problem areas. Research is organized in interdisciplinary research environments and doctoral studies mainly in graduate schools. Jointly, they publish the series Linköping Studies in Arts and Sciences. This thesis comes from Statistics and Machine Learning at the Department of Computer and Information Science.
to Filippa
POPULÄRVETENSKAPLIG SAMMANFATTNING

Observationer som har någon typ av geografisk eller rumslig koppling brukar inom statistiken benämnas som spatiala data. Sådana data uppstår inom många olika tillämpningar och ställer ofta särskilda krav på de statistiska analyserna. Till exempel: regionala temperaturmätningar, medicinska bilder av hjärnan och huspriserna inom en stad visar alla prov på det som kallas spatial autokorrelation – att närliggande mätningar är mer lika varandra än dem som ligger längre ifrån. Om man inte tar hänsyn till sådan korrelation riskerar analyserna att leda till felaktiga slutsatser. Samtidigt ökar ständig storleken på de spatiala datamaterialen vilket leder till en beräkningsmässig utmaning – en enda medicinsk bild kan innehålla miljontals pixlar eller datapunkter, medan många klassiska metoder för spatiala data är begränsade till att kunna hantera några tusental.

Den här avhandlingen undersöker hur en speciell typ av matematisk statistisk modell, så kallade Gaussiska Markov-fält (eng: Gaussian Markov random fields, GMRF) kan användas för mer skalbara analyser av spatiala data. GMRF-modellen liknar de klassiska spatiala metoderna, men har så kallade gleshetsegenskaper, som gör den mer beräkningseffektiv för stora datamängder.

Vi utvecklar nya algoritmer som gör det möjligt att använda GMRF-modellen för att analysera hjärnaktivitet i medicinska bilder framtagna med tekniken fMRI. Detta är tredimensionella bilder av hjärnan med miljontals datapunkter, utifrån vilka den underliggande hjärnaktiviteten kan stimuleras, samtidigt som antagandet att hjärnaktiviteten är ungefär densamma i närliggande områden uppfylls med hjälp av GMRF-modellen. Vi visar att våra metoder är både snabbare och mer korrekta än tidigare forskning. Vi gör också en kopppling mellan GMRF-modellen och så kallade djupa neurala faltningsnätverk som på senare år haft stora framgångar inom artificial intelligens och maskininlärning. Denna kopppling öppnar upp för en ny, djup GMRF-modell som uppvissar enastående prediktionsförmåga i satellitbilder där vissa observationer saknas på grund av täckande moln. Slutligen visar vi också hur GMRF-modellen kan användas av självflygande drönare för räddningsinsatser i katastrofområden för att prediktera var de skadade personerna befinner sig och fatta beslut i realtid.
ABSTRACT

Accurate statistical analysis of spatial data is important in many applications. Failing to properly account for spatial autocorrelation may often lead to false conclusions. At the same time, the ever-increasing sizes of spatial datasets pose a great computational challenge, as many standard methods for spatial analysis are limited to a few thousand data points.

In this thesis, we explore how Gaussian Markov random fields (GMRFs) can be used for scalable analysis of spatial data. GMRFs are closely connected to the commonly used Gaussian processes, but have sparsity properties that make them computationally cheap both in time and memory. The Bayesian framework enables a GMRF to be used as a spatial prior, comprising the assumption of smooth variation over space, and gives a principled way to estimate the parameters and propagate uncertainty.

We develop new algorithms that enable applying GMRF priors in 3D to the brain activity inherent in functional magnetic resonance imaging (fMRI) data, with millions of observations. We show that our methods are both faster and more accurate than previous work. A method for approximating selected elements of the inverse precision matrix (i.e. the covariance matrix) is also proposed, which is important for evaluating the posterior uncertainty. In addition, we establish a link between GMRFs and deep convolutional neural networks, which have been successfully used in countless machine learning tasks for images, resulting in a deep GMRF model. Finally, we show how GMRFs can be used in real-time robotic search and rescue operations, for modeling the spatial distribution of injured persons.
Acknowledgments

The writing of this thesis has been a long journey for me, academically and personally, and I want to express my gratitude to a number of persons which have given me the inspiration, feedback and support needed to follow out this process.

First of all, I want to thank Mattias Villani, who has been an incredible supervisor. Mattias has taught me so much. He has shaped my way of reasoning about statistics, of doing research, of writing and of teaching. After a meeting with Mattias, you feel a joyful hunger to go back to your desk and continue learning and uncovering new knowledge. He is probably the best teacher I ever had. I also want to thank my co-supervisor Anders Eklund, who has always been there to support me, and who has taught me a lot about medical imaging and what you can accomplish with the help of (a small army of) fast CPUs and GPUs. David Bolin introduced me to the wonderful world of Gaussian Markov random fields, and as a co-author he has been very helpful and inspiring, and he has given me many good ideas. Finn Lindgren kindly received me as a visiting PhD student at the University of Edinburgh during the autumn of 2017. He taught me the value of solid mathematical understanding of stochastic processes and optimization algorithms, and also a few steps of Balboa. During the latter part of my PhD project, I have had the pleasure of working with the brilliant Fredrik Lindsten, who has given me a better perspective on the field of machine learning. The collaboration with Fredrik has been very exciting and has led to new insights. I want to thank Olov Andersson and Johan Dahlin with whom I had a nice collaboration resulting in Paper V of the thesis, and as more senior PhD students they were good role models and taught me a lot.

Without doubt, Bertil Wegmann has been my best friend at the office during these years. We have shared many great moments in the fika room and at the golf restaurant, discussing Bayesian statistics and everything else. Coming to work has been so much more fun with Bertil around, and he has given me many wisdoms about research and teaching. A special thanks to Matias Quiroz who, during our time as colleagues at the Riksbank, strongly encouraged me to pursue a PhD, and in particular to do it in Bayesian statistics with Mattias Villani. This has turned out to be one of the best advices I ever got. As an academical big brother, Matias has continued to give me great advice. It has been fun seeing him at conferences, and I only wish Sydney was closer. I have had the pleasure to share the path as a PhD student with Josef Wilzen, and through our many discussions he has helped me cope with the struggles of Bayesian neuroimaging, PhD life and parenting. Josef is a really nice person and friend. I also want to thank the other PhD students at STIMA: Mans, Sarah, Hector, Caroline, Amanda, Amir and Hao Chi for good times. Thanks to Joaquin Martinez-Minaya for making my time in Edinburgh more enjoyable. Thanks to all the other co-workers at STIMA, and to the administrators who have made my life easier, in particular Anne, Inger and Karin.

Finally, I want to thank my family and friends for all their support. My parents, Karin and Göran, have always been there for me and helped guiding me through the jungle of career and study paths. My sisters, Lisa and Anna-Karin, have given me care and intellectual stimulation of the non-statistical kind. My son, Joel, has been a cheerful little guy who has taken my mind
off my concerns by ignoring my serious face and requiring me to play. The last person that I wish to thank, but the most important person for this thesis and in my life is Filippa. She somewhat reluctantly followed me to Linköping to pursue this dream, and has made our life here really pleasant, and she is now my wife and the mother of my child (very soon children). Thank you, Filippa, for always showing interest in my work, and for putting up with my long working hours and my occasionally distracted mind. You mean so much to me.

Linköping, June 2020
# Contents

| Section | Page |
|---------|------|
| Abstract | v |
| Acknowledgments | viii |
| Contents | ix |
| List of Figures | x |
| 1 Introduction | 1 |
| 1.1 Background | 1 |
| 1.2 Motivation | 3 |
| 1.3 Thesis outline | 5 |
| 1.4 Summary of papers | 5 |
| 2 Bayesian analysis of spatial data | 9 |
| 2.1 Bayesian statistics | 9 |
| 2.2 Spatial statistics | 13 |
| 2.3 Gaussian Markov random fields | 19 |
| 3 Inference methods | 27 |
| 3.1 Conditional exact inference for the latent variables | 28 |
| 3.2 Empirical Bayes | 31 |
| 3.3 Markov chain Monte Carlo | 32 |
| 3.4 Variational inference | 34 |
| 3.5 Integrated nested Laplace approximations | 37 |
| 4 Functional magnetic resonance imaging | 39 |
| 5 Concluding remarks | 45 |
| 5.1 Summary of contributions | 45 |
| 5.2 Outlook on future research | 46 |
| Bibliography | 49 |
| Paper I | 57 |
| Paper II | 75 |
| Paper III | 89 |
| Paper IV | 129 |
| Paper V | 147 |
List of Figures

1.1 Wheat yield data from the agricultural field trials in Mercer and Hall (1911). Each square represents the produced wheat (in pounds) in a $3.30 \times 2.59$ meters rectangle in the field. Clearly, there is spatial correlation in the data. ............................................. 2

2.1 Bayesian inference for the simple temperature example. The plots show how the posterior distribution of $\theta$ depends on the prior and likelihood in the case of 1 observation (left) and 10 observations (right). ................................................................. 11

2.2 Three types of spatial data. Top left: Continuously referenced data, displayed with size and color proportional to the measurements. Precipitation in mm from weather stations in and nearby Colorado during the spring months (March, April, May) in 1997 (Nychka et al., 2017). Top right: Grid data. Three slices (oriented along the three different axes) for a single 3D volume of fmri data from the face repetition experiment (Penny et al., 2005) used in Paper I and Paper III. Bottom: Point pattern data. Cancer cases (lung and larynx) recorded in the Chorley and South Ribble Health Authority of Lancashire (England) between 1974 and 1983 (Baddeley et al., 2015; Diggle, 1990). ................................................................. 14

2.3 Matérn correlation functions for various combinations of hyperparameters and 1D realizations of the corresponding GPs. ................................................................. 17

2.4 A small part of the graph $\mathcal{G}$ of a GMRF defined on a regular grid with first order neighborhoods. The nodes $i$ and $j$ are conditionally independent given the other nodes, since they are not neighbors. ................................................................. 20

4.1 Illustration of fmri data. A dataset consists of a series of 3D images of the living human brain, collected over time. Each image consists of a large number (~100000) of voxels (3D pixels). The intensity of a specific voxel over time creates a signal known as the BOLD signal. The variability of the BOLD signal partially depends on nearby neural activity. Therefore, analysing the BOLD signal a can give information regarding which parts of the brain that are active during an experiment. ............... 40

4.2 A HRF models how the BOLD signal responds to rapid neural activation spike. Here the commonly used canonical HRF is shown. ................................................................. 41

4.3 Illustration of the measured BOLD signal, the experimental paradigm (events), and the predicted BOLD signal. ................................................................. 41

4.4 Posterior probability maps (PPMs) taken from the results in Paper III. Posterior probabilities of brain activity larger than 0.9 are shown for a single slice of the brain. The left image shows the resulting PPM when no spatial prior is included in the model, and the right image shows the PPM when using a spatial 3D second order Matérn GMRF prior. ................................................................. 43
1 Introduction

1.1 Background

Spatial statistics deals with describing the statistical patterns that exist in data measured across space. Data which are references by spatial location are common and arise naturally in a wide range of areas. Many applications are on a geographic scale, for example describing the distribution of plants and animals, the spread of diseases or the variability of house prices in a city. One important aspect with spatial data is that nearby measurements tend to be more similar than those further apart, which can be described as spatial autocorrelation. Plants of the same species are more often found close to each other, and houses tend to sell for prices similar to other houses in the same area rather than to houses in other areas. Properly accounting for these sort of dependencies is key to draw correct conclusions and make credible predictions when analyzing spatial data.

The statistical analysis of spatial data dates back at least to the agricultural field trials conducted at the Rothamsted Experimental Station, England in the early 1900’s. Figure 1.1 shows data from an experiment in a wheat field reported in Mercer and Hall (1911). The field is represented by 20 × 25 squares, where each square shows the produced wheat (in pounds) in a 3.30 × 2.59 meters rectangle. The simple assumption of independent measurements is clearly false. The wheat yield from a single square is in general much more similar to the neighboring squares than to more distant squares. The underlying reason for this smooth spatial variation is presumably external factors: how the field was watered, the soil fertility, etc. However, as long as these factors are unknown or not measured, the observed dependencies must be accounted for in another way in order to make correct inferences from the data. Ronald Fisher was employed at Rothamsted between 1919 and 1933, and made some early contributions to the analysis of spatial data. For example he advocated dividing the measurement into blocks, for instance arrays of size 5 × 5 squares, and assuming independence between these larger blocks, an assumption that is less clearly violated by the data (Diggle, 2010).
1. INTRODUCTION

![Wheat yield data from the agricultural field trials in Mercer and Hall (1911). Each square represents the produced wheat (in pounds) in a 3.30 \times 2.59 meters rectangle in the field. Clearly, there is spatial correlation in the data.](image)

Figure 1.1: Wheat yield data from the agricultural field trials in Mercer and Hall (1911). Each square represents the produced wheat (in pounds) in a 3.30 \times 2.59 meters rectangle in the field. Clearly, there is spatial correlation in the data.

A more model based direction of spatial statistics emerged from applications in mining engineering and forestry in the middle of the 1900’s. Danie Krige worked in the South African mining industry, on methods for making spatial predictions for the yield of valuable minerals. In Krige (1951) he made the first steps for the general spatial statistical framework that would later become known as kriging. Bertil Matérn was a Swedish forest statistician whose PhD thesis (Matérn, 1960) became very influential and widely cited in the spatial statistics community. A fundamental idea in both works is that of a spatial autocorrelation function that depends only on the distance between points. Put differently, the dependence between the measurements in two different points is decided only by how far they are apart, and not by the location or the direction between the points, which is a property known as stationarity. One of the main legacies of Matérn is the Matérn correlation function, which remains a standard choice to this day.

Bayesian statistics is a discipline in statistics which treats the unknown model parameters as random. Based on ideas of Thomas Bayes and Pierre-Simon Laplace in the 1700’s and 1800’s, much of Bayesian methodology was developed in the beginning of the 1900’s, but became widely popular from the 1990’s and forward with the development of computational methods such as Markov chain Monte Carlo. Given a prior belief about the parameters, observing the data allows for updating the belief using Bayes’ rule, and obtaining a posterior belief, represented by a posterior distribution for the parameters. With an increasing number of observations, the uncertainty about the parameters gets smaller and smaller, and the posterior distribution becomes more and more concentrated. Bayesian statistics gives a principled framework for specifying models, estimating parameters, and reasoning about uncertainty.

In spatial statistics, the Bayesian view enables the use of spatial priors. For the wheat yield example in Figure 1.1, one can specify a spatial prior that comprises the assumption that the measured values vary “smoothly” over space, avoiding the false assumption of independent measurements. The information in such a prior can then, using Bayesian methods, be propagated into a posterior distribution that more correctly reflects ones beliefs after observing the data. Furthermore, the “smoothness” can be represented in the model, by one or several parameters, which
1.2 Motivation

The usefulness of spatial priors becomes even more clear when considering that they can be applied not only to the measured data directly, but also to some other latent, unobserved phenomenon. In the analysis of functional magnetic resonance imaging (fMRI) data, which is central to this thesis, one can put spatial priors on the brain activity, to impose that the activity should appear in spatially contiguous patterns. The fMRI technique does not measure the brain activity directly, but instead a blood oxygen level dependent (BOLD) signal, which partially depends on the activity in different locations in the brain. Even so, the Bayesian modelling allows making posterior predictions about the brain activity, incorporating both the information from the BOLD signal and from the spatial prior.

The wheat yield data in Figure 1.1 is an early example of a certain type of spatial data, which are of central interest in this thesis – data that are located on a discrete grid or regular lattice. Nowadays, such data are very common in the form of digital images. Since this type of data is a special case of spatial data with observations in arbitrary locations, standard methods from spatial statistics can be used to analyze this type of data as well. However, the regular structure also allows for different methods to be used, for example by defining the model directly on the graph induced by nodes in the lattice, instead of defining a spatial correlation function. The regular structure can also be leveraged for computational advantages. A model that enables both is the Gaussian Markov random field (GMRF). GMRF models where popularized by Besag (1974), and have been widely used in image processing, and in spatial statistics, not necessarily restricted to data on the grid. In this thesis, the GMRF is a main building block, enabling scalable and flexible Bayesian analysis of spatial data.

The amount of digital images produced every day is enormous. They come from cameras, MR scanners, satellites, microscopes, and many other sources. At the same time, the resolution of images is getting finer and finer. In combination, this means a rapid increase in the total volume of image data, and poses a computational challenge for the algorithms used to analyze such data, both in handling the large number of images, and in the processing and storage in computer memory of each individual image. This thesis is focused more towards the second of these tasks, the efficient analysis of large-scale images.

Modelling the spatial dependence between measurements is especially expensive. One way to understand this is to consider the 500 squares in the wheat field in Figure 1.1. These measurements give rise to roughly $500^2/2 = 125,000$ pairwise relations. That means that just storing a representation of all these dependencies, for example a correlation matrix, would have been infeasible in the days when the data were collected. Using a modern computer, this is of course manageable, but if the number of pixels in each dimension is increased by a factor 100, resulting in a grid of size $2000 \times 2500$ which is normal in modern applications, the corresponding number of pairwise relations exceed 10 trillion ($10^{13}$), or around 100 terabyte of computer storage. This rapid growth in storage can be described as quadratic in the number of measurements, and the cost of computations is even greater, cubic, for standard naive methods. By exploiting structural properties of the problem, such as the sparsity in the graph of a GMRF, these costs can be considerably reduced, yet further improving the scalability of these methods is greatly desirable.

As a practical example, fMRI data consists of three-dimensional images of the brain of living human subjects. Each subject is scanned over time, collecting one image roughly every second, resulting in a time series normally consisting of a couple of hundred images. A typical image consists of a hundred thousand voxels (three-dimensional pixels). In total, fMRI data from a
1. INTRODUCTION

A single subject can be thought of as a single four-dimensional image with millions of data points. The direct use of classical methods from spatial statistics for images of this size is prohibitive, and many articles divide the brain into parcels or two-dimensional slices and run separate analyses instead of using a single spatial model (e.g. Groves et al. (2009), Lee et al. (2014), Penny et al. (2005)). So there is a great need for methods that are more computationally efficient for larger models.

At the same time, in order to draw correct conclusions about the probability of activation in different parts of the brain, a rigorous spatial statistical model is necessary. Experience has shown that replacing spatial models with simpler methods, as smoothing images in a preprocessing step, which is much used in practice, often leads to spurious results (Eklund et al., 2016). For a simple illustration, see Sidén and Villani (2018). We address the issue of enabling whole-brain spatial priors for single subject task-fMRI analysis in Paper I and Paper III.

Many applications that use GMRF models for computational efficiency require the calculation of posterior variances and covariances. This is non-trivial since GMRFs parameterize the precision (inverse covariance) matrix, and inverting this matrix to obtain the covariance matrix is computationally infeasible for large models. For example, in fMRI analysis, the posterior covariances are needed for computing the probability of brain activity in different parts of the brain. We address this problem in Paper II.

The ever increasing size of image data is a computational challenge, but also enables the use of more flexible and adaptive models, as more data decreases the risk of over-fitting with high-capacity models. In fact, the abundance of image data is one of the reasons behind the success story of deep learning in the last decade, in particular using deep convolutional neural networks (CNNs). CNNs are extremely flexible machine learning models, typically having thousands or millions of parameters. Nevertheless, using a large amount of images, normally thousands or millions, these parameters can be trained for the CNN to learn complicated tasks, for example to classify different objects in a natural image. There exist countless examples of imaging applications in which the use of CNNs have led to substantial improvements in performance.

In spatial statistics, however, the adoption of CNNs has been moderate. There could be several reasons for this: many spatial datasets are not on a fixed grid like images, CNNs are black-box models which do not correspond to an interpretable correlation function, and reliable predictive uncertainties are non-trivial to obtain. Still, the main reason in many applications is probably that the number of images that can be used as training data is simply too small. A map of mineral yields over an area or house prices in a city usually only come as one, or as a few temporally correlated observations. For fMRI, one can increase the number of images by considering data from multiple subjects, but only if one assume some spatial properties to be constant across subjects, which may not be the case. In Paper IV, we attempt to bridge the gap between spatial statistics and deep learning by exploring a formal connection between GMRFs and CNNs.

The methods developed in this thesis are mainly demonstrated using spatial data on the grid, but rigorous and scalable methods are needed for general spatial datasets as well. In Paper V, we study robotic search and rescue operations during disasters using unmanned aerial vehicles (UAVs), where the data consists of the location of detected persons on the ground and whether they are injured or not. In statistical language, this can be described as marked point pattern data.

In summary, the motivation for this thesis is to enable Bayesian analysis with spatial priors for medical images and other large-scale spatial data. Many applications need hierarchical, structural and flexible Bayesian spatial models in order to appropriately describe the data, to correctly propagate uncertainty, and to draw correct conclusions. We address this problem by developing fast algorithms for Bayesian inference, and demonstrating their performance in several applications.
1.3 Thesis outline

The thesis is divided into two parts, where the first part is a basic introduction to the research field, and the second part is a collection of research articles. This chapter ends with a summary of these articles. In Chapter 2, we review Bayesian and spatial statistical modelling, with special focus on GMRFs. Chapter 3 covers methods for Bayesian inference. Chapter 4 introduces the statistical analysis of fMRI data, in particular with regard to spatial priors. Finally, Chapter 5 gives a summary of the contributions of the thesis as well as an outlook of future research.

1.4 Summary of papers

**Paper I**

Sidén, P., Eklund, A., Bolin, D., & Villani, M. (2017). Fast Bayesian whole-brain fMRI analysis with spatial 3D priors. *NeuroImage, 146*, 211–225.

**Source code**: [https://github.com/psiden/BFAST3D](https://github.com/psiden/BFAST3D)

**Summary**: This article introduces new methods for Bayesian spatial analysis of brain activity patterns in task-fMRI data, leading to substantial improvements in computational speed and accuracy. We consider the spatial model that is default for Bayesian single subject analysis in the popular neuroimaging software SPM. The inference method in SPM processes the data slice-by-slice and uses an approximate variational Bayes (VB) estimation algorithm that enforces posterior independence between activity coefficients in different voxels. We introduce several new, sparsity-exploiting techniques, which enable Bayesian inference with Markov chain Monte Carlo (MCMC) or with a less restrictive spatial VB method, both slice-wise and for the whole brain in a single 3D model. Using MCMC, we are for the first time able to evaluate VB against the exact MCMC posterior, and demonstrate that SPM’s VB can lead to spurious activation, while our spatial VB shows negligible difference compared to the MCMC posterior.

**Background and contributions**: Mattias Villani had the original idea to evaluate SPM’s VB method by comparing to full MCMC. In 2015, I took David Bolin’s PhD course about GMRFs. This course, and further discussions with David, made me aware of computationally efficient methods for GMRFs, such as sparse reordered Cholesky decomposition and preconditioned conjugate gradient (PCG), which made Mattias’ idea feasible in both 2D and 3D. The spatial VB method was my idea. I did the mathematical derivations and wrote the code, loosely based on SPM’s code and some code from David. I ran the experiments with help with fMRI data expertise from Anders Eklund. I wrote the manuscript and the other authors edited.

**Paper II**

Sidén, P., Lindgren, F., Bolin, D., & Villani, M. (2018). Efficient covariance approximations for large sparse precision matrices. *Journal of Computational and Graphical Statistics, 27*(4), 898–909.

**Source code**: [https://github.com/psiden/CovApprox](https://github.com/psiden/CovApprox)

**Summary**: The use of sparse precision matrices has become popular because they allow for efficient algorithms for joint inference in high-dimensional models. Many applications require the computation of certain elements of the covariance matrix, such as the marginal variances, which may be non-trivial to obtain when the dimension is large. This paper introduces a fast Rao-Blackwellized Monte Carlo (RBMC) sampling-based method for efficiently approximating selected elements of the covariance matrix. The variance and confidence bounds of the approximations can be precisely estimated without additional computational costs. Furthermore, a
1. INTRODUCTION

A method that iterates over subdomains is introduced, and it is shown to additionally reduce the approximation errors to practically negligible levels in an application on fMRI data. Both methods have low memory requirements, which is typically the bottleneck for competing direct methods.

**Background and contributions:** I got the original idea of using RBMC estimates for marginal variances during the work with Paper I, as a response to the practical problem of computing posterior probability maps, which requires these variances. After discussing the idea with David Bolin, he connected me with Finn Lindgren, who had the same idea for general elements of the covariance matrix, and also the idea to improve the estimates by iterating over subdomains. I did the derivations, wrote the code and ran the experiments. I wrote the manuscript and the other authors edited.

**Paper III**

Sidén, P., Lindgren, F., Bolin, D., Eklund, A., & Villani, M. (2019). Spatial 3D Matérn priors for fast whole-brain fMRI analysis. *arXiv preprint arXiv:1906.10591*. Preprint submitted to journal.

**Source code:** [https://github.com/psiden/BFAST3D](https://github.com/psiden/BFAST3D)

**Summary:** This paper extends the model for fMRI data in Paper I to spatial priors with Matérn covariance function, using the stochastic partial differential equation (SPDE) link between Gaussian processes and GMRFs. Unlike the prior in Paper I, the Matérn prior is proper and can have interpretable attributes, such as spatial correlation range. It is also more flexible, while maintaining the sparsity required for fast inference in the whole-brain model. We enable empirical Bayes (EB) inference, by developing an accelerated stochastic gradient descent optimization algorithm for the spatial hyperparameters, and make a fully Bayesian treatment of the brain activity coefficients, which are the main parameters of interest. We use both experimental and simulated task-fMRI to show that the Matérn prior is preferable over the prior in Paper I, by comparisons using prior simulation, cross validation and visual inspection of the resulting activation maps. Additionally, to illustrate the potential of the SPDE formulation, we derive an anisotropic version of the Matérn 3D prior.

**Background and contributions:** The original idea to extend the spatial prior in Paper I to Matérn came from David Bolin. Using the Matérn prior makes it more difficult to infer the spatial hyperparameters, and I worked on this in collaboration with Finn Lindgren during my visit in Edinburgh 2017, also exploring some other ideas that did not make it into the paper. I did the derivations, wrote the code and ran the experiments. I wrote the manuscript and the other authors edited.

**Paper IV**

Sidén, P., & Lindsten, F. (2020). Deep Gaussian Markov random fields, In *Proceedings of the 37th International Conference on Machine Learning (ICML)*.

**Source code:** [https://bitbucket.org/psiden/deepgmrf](https://bitbucket.org/psiden/deepgmrf)

**Summary:** This article shows a formal connection between GMRFs and CNNs. Common GMRFs are special cases of a generative model where the inverse mapping from data to latent variables is given by a 1-layer linear CNN. This connection allows us to generalize GMRFs to multi-layer CNN architectures, effectively increasing the order of the corresponding GMRF in a way which has favorable computational scaling. We describe how well-established tools, such as autodiff and variational inference, can be used for simple and efficient inference and learning of the deep GMRF. We demonstrate the flexibility of the proposed model and show that it outperforms the state-of-the-art on a dataset of satellite temperatures, in terms of prediction accuracy.
and predictive uncertainty.

**Background and contributions:** I had the original idea to use CNNs for problems in spatial statistics, and found the formal connection to GMRFs. This idea was developed together with Fredrik Lindsten, especially the inference part. I did the derivations, wrote the code and ran the experiments. I wrote the manuscript, apart from the abstract and introduction, and Fredrik edited.

**Paper V**

Andersson, O., Sidén, P., Dahlin, J., Doherty, P., & Villani, M. (2019). Real-time robotic search using structural spatial point processes, In *Proceedings of the 35th Conference on Uncertainty in Artificial Intelligence (UAI)*.

**Summary:** In this paper we show how probabilistic spatial models can be used for automated planning of aerial robots in search and rescue operations over large areas, such as during natural disasters. In traditional approaches, robots typically search an area exhaustively, thereby ignoring that the density of victims varies based on predictable factors, such as the terrain type and population density. Our proposed model is a spatial point process with three interacting spatial fields for: i) the point pattern of persons in the area, ii) the probability of detecting persons and iii) the probability of injury. This structure allows inclusion of informative priors from e.g. geographic or cell phone traffic data, while falling back to latent Gaussian processes when priors are missing or inaccurate. To solve the planning problem in real-time, we propose a combination of fast approximate inference using Integrated Nested Laplace Approximation (INLA), and a novel Monte Carlo tree search (MCTS) method tailored to the problem, minimizing the expected time to victim discovery. Experiments using data simulated from real world maps show that the framework outperforms competing approaches, finding many more injured in the crucial first hours.

**Background and contributions:** The idea to use spatial models for automated planning of aerial robots in disaster sites originated from Mattias Villani in 2017. Johan Dahlin developed an initial version, in which all persons were assumed to be detected and injured, and which used simpler search strategies than MCTS. In 2018-2019, Olov Andersson and I continued to work on the model, in close collaboration with Mattias. Me and Mattias came up with the structure of the spatial model, with the interacting population, detection and injury models, and I wrote the main part of Section 2 and Section 3.1 in the manuscript. Olov and I developed the code together, based on Johan’s code, where I did the main part related to the spatial model and INLA inference, and Olov did among other the part related to planning, MCTS and data collection. The rest of the paper was mainly written by Olov and Mattias, and all the authors edited.

---

1Equal contribution.
Bayesian analysis of spatial data

This chapter gives a brief introduction to Bayesian statistics and spatial statistics. In particular, we review GMRF models, which can be favorably used as spatial priors for large-scale datasets due to their computational advantages.

2.1 Bayesian statistics

Probabilistic models

The general purpose of almost any statistical application is to analyse some dataset $y$, where $y = (y_1, \ldots, y_N)^\top$ denotes a vector with $N$ observations. We will consider probabilistic models for the data, that is, we assume that the data are generated through some random process which can be described by a probability distribution $p(y|\theta)$, where $\theta$ is a vector of parameters. The data $y$ have been observed, and are known with some fixed value, while the parameters $\theta$ are unknown, and a common goal is to infer $\theta$, that is, to specify plausible values for $\theta$.

As a simple example, assume that $y$ is a single measurement of the temperature (in degrees Celsius) outside your house at an (somewhat unrealistically) unknown time point in the last year. A model for the data is

$$y|\theta \sim N(\theta, 5^2),$$

where $\theta$ has the interpretation as the yearly mean temperature and it is assumed that the standard deviation is 5 degrees. Assume that the measurement was $y = 6$. If our goal is to estimate $\theta$, without any further information, the best guess must obviously be that $\theta = 6$, since the Normal distribution is symmetric. However, we might also be interested in answering questions like

- How certain is this estimate?
2. BAYESIAN ANALYSIS OF SPATIAL DATA

- How can we include prior knowledge about the problem to improve the estimate?
- What is a good prediction of the temperature tomorrow?

The Bayesian view of statistics gives a natural and principled way to answer these types of questions.

Bayes’ theorem

The fundamental idea of Bayesian statistics is that information about the parameters \( \theta \) is obtained by inverting the probability model \( p(y|\theta) \) using Bayes’ theorem

\[
p(\theta|y) = \frac{p(y|\theta) p(\theta)}{p(y)}.
\]

Here, \( p(\theta|y) \) is known as the posterior distribution, which describes the knowledge about the parameters \( \theta \) given the data \( y \). The likelihood function \( p(y|\theta) \) is the model of the data given the parameters. The prior distribution \( p(\theta) \) describes the information about \( \theta \) before observing the data. Bayes’ theorem describes the essence of Bayesian statistics in a very compact form. One could also describe Bayesian analysis as a three-step procedure:

1. Specify the likelihood \( p(y|\theta) \). This is the model of how the data \( y \) is generated given the unknown parameters \( \theta \).
2. Specify the prior \( p(\theta) \). Depending on the interpretation of \( \theta \), the prior knowledge may be easier or harder to specify in terms of a distribution. If there is no information, we can use a so called non-informative prior, for example a completely flat distribution, where all values for \( \theta \) are equally likely.
3. Compute the posterior \( p(\theta|y) \) using Bayes’ theorem. This is straightforward in theory, but can be difficult in practice, due to large datasets or complicated models. A key reason for this is that the fourth component of Bayes’ theorem, \( p(y) \), known as the marginal likelihood, is usually unknown or hard to compute. Fortunately, several algorithms exist that allow for computation or approximation of the posterior without using the marginal likelihood. However, in practice, step 1 and 2 will have a large impact on which inference algorithm that is more efficient, and carefully choosing the likelihood and prior can simplify the third step. Chapter 3 looks into how to do efficient Bayesian inference for the models relevant to this thesis.

Finally, when we have computed the posterior distribution, we can use it to make various assessments regarding \( \theta \). We can for example compute a point estimate of \( \theta \) as the posterior mean \( E(\theta|y) \), or as \( \text{argmax}_{\theta} p(\theta|y) \), known as the maximum-a-posteriori (MAP) estimate. We can compute a credible (uncertainty) interval based on the quantiles of \( p(\theta|y) \), or the probability that \( \theta \) exceeds some threshold \( \delta \)

\[
\mathbb{P}(\theta > \delta) = \int_{\delta}^{\infty} p(\theta|y) d\theta.
\]

Returning to our temperature example, step 1 of the Bayesian procedure was performed by specifying the likelihood in Equation (2.1). For step 2, if I for example believe that, before seeing the measurement, the yearly mean temperature \( \theta \) is “somewhere around 10, but I’m not really sure”, I could set the prior to

\[
\theta \sim \mathcal{N}(10, 4^2).
\]

By specifying both the likelihood and prior as Gaussian distributions, we have made step 3 very simple, as it can be shown that in this case, the posterior is Gaussian as well. This is a univariate special case of the GMRF posterior, derived later in Equation (2.20). The likelihood, prior and posterior, for the single measurement \( y = 6 \), are depicted in the left-hand side of Figure 2.1. We see that the observation updates our prior belief, that \( \theta \) is around 10, towards lower values in the posterior. In fact, the posterior can be seen as a weighted average between the prior and the likelihood.
2.1. Bayesian statistics

Now, consider a different scenario, where we have \( N \) different temperature measurements at unknown time points in the last year. A typical assumption would then be that of independent observations, with each observation distributed as in Equation (2.1), in which case the likelihood becomes

\[
p(y|\theta) = \prod_{i=1}^{N} p(y_i|\theta) \propto \exp\left( -\frac{1}{2} \frac{(\bar{y} - \theta)^2}{\sigma^2/N} \right),
\]

where \( \bar{y} \) denotes the mean of \( y \), and the proportionality sign \( \propto \) indicates that the likelihood is proportional to the right-hand expression (which requires a short derivation) up to some constant with respect to \( \theta \). It is common to use the \( \propto \) sign in Bayesian derivations, as the inversion of the probability model implies that we start considering the likelihood as a function of \( \theta \) instead of a function of \( y \), in which case it no longer integrates to 1. The right-hand side expression has the same analytic form as a Gaussian distribution, and it can be shown that the posterior is Gaussian also in this scenario. The situation can be seen in the right-hand side of Figure 2.1, for the case with \( N = 10 \) observations and \( \bar{y} = 6 \). It can be seen that the posterior is now closer to the likelihood, which reflects that the increased amount of information from the data weighs heavier than the prior information. It is known that the posterior gets closer and closer to the likelihood with more and more data.

**Comparison to classical statistics**

One can compare to classical (or frequentist) statistics, in which the model is normally defined through the likelihood function only, and standard procedure is to estimate the parameters \( \theta \) by maximizing the likelihood with respect to \( \theta \). The Bayesian method gives the same point estimate when using a non-informative prior and the MAP estimate. However, classical uncertainty measures build on a different philosophical view. As the value of \( \theta \) is seen as fixed, classical statistics instead uses the idea of a sampling distribution of the estimate, that is, if we could generate new datasets from the model, what variation would that imply for the estimate? This is quite complicated compared to the Bayesian view, where the data is fixed and conditioned on, while the parameter uncertainty is described directly by the posterior distribution. The fact that we represent our knowledge about \( \theta \) using probability distributions does not imply that \( \theta \) is random or cannot have a true fixed value; the distribution merely reflects the uncertainty in our belief about the value of \( \theta \).
2. **Bayesian analysis of spatial data**

**Conjugate priors**

Our temperature example above also illustrates an important Bayesian tool, known as a conjugate prior. By choosing the prior for $\theta$ to be normally distributed, we saw that the posterior became normally distributed as well. This means that we get a closed form analytical expression for the posterior, which is computationally much simpler to handle. Conjugate priors exist for several commonly used likelihood functions, but far from all, and can be used to simplify Bayesian analysis in cases where the prior belief can be reasonably well represented by the conjugate prior. In this thesis, I make use of the fact that the GMRF is a conjugate prior for the mean parameter in Gaussian likelihoods, and also that the gamma distribution is a conjugate prior to the precision parameter.

**Prediction**

The Bayesian approach to making prediction for some new unseen observations $y^*$ is to compute the posterior predictive distribution

$$p(y^*|y) = \int p(y^*|\theta) p(\theta|y) \, d\theta,$$

where $p(y^*|\theta)$ is the distribution of $y^*$ given the model, here assumed to be independent of $y$ given $\theta$, and $p(\theta|y)$ is the posterior computed based on the previously seen data $y$. This is elegant, as it properly accounts for the uncertainty about $\theta$, by integrating out $\theta$ with respect to the posterior. This can be compared to making predictions based on $p(y^*|\hat{\theta})$, using some estimate $\hat{\theta}$, where the parameter uncertainty is neglected. Given the full posterior predictive $p(y^*|y)$, just as with the posterior distribution, we can use it to make various assessments about $y^*$. We can for example compute a point prediction as $E(y^*|y)$ and a measure of the predictive uncertainty with $\text{Var}(y^*|y)$. Moreover, the posterior predictive distribution is central in Bayesian decision theory, which can be used to automate decision-making, by considering a loss function $L(y^*)$ and choosing the action that minimizes the expected loss with respect to the posterior belief about $y^*$.

**Dependent data**

We return to the temperature measurement example one final time, to prepare for the spatial models presented in next section. We have so far somewhat unrealistically assumed that the observation time points are unknown. If we instead assume that the time points are known, and say that we have 365 observations, one for each day of the year, we have to drop the assumption of independent measurements, as the temperatures in two consecutive days must be highly dependent. If we stay with the Gaussian model, the likelihood function can now be described by

$$y|\theta \sim \mathcal{N}(\mu, \Sigma),$$

where $\theta = \{\mu, \Sigma\}$ and $\Sigma$ is no longer a diagonal matrix. Let $y_t$ denote the measured temperature at day $t$. Here are two common ways to model $\Sigma$:

1. Model the covariance between day $t$ and $t'$, that is $\Sigma_{t,t'}$, as a function that depends on the time distance $|t - t'|$. Normally this function decays with distance and will be close to zero for distant time points. The covariance function commonly depends on parameters that can be added to $\theta$.

2. Use an autoregressive model, for example a so called AR(1) process

$$y_t - \mu = \phi(y_{t-1} - \mu) + \epsilon_t$$

$$\epsilon_t \sim \mathcal{N}(0, \sigma^2),$$

where $\mu$, $\phi$ and $\sigma^2$ are parameters to be included in $\theta$. For $0 < \phi < 1$, if $y_{t-1}$ deviates from the steady state mean $\mu$, $y_t$ is likely to do so as well, which introduces a positive
correlation. This formulation implies a precision (inverse covariance) matrix $\Sigma^{-1}$ that is sparse, which has computational advantages, and is also equivalent to a sequential formulation of the likelihood

$$p(y|\theta) = p(y_1|\theta) \prod_{t=2}^{365} p(y_t|y_{t-1}, \theta),$$

where each $y_t$ only depends on $y_{t-1}$.

Moreover, we can extend the example to a spatial problem by also considering temperature measurements at your neighbor’s house or in arbitrary locations at the planet. We can then model the dependence between spatially proximate locations based on a covariance function that depends on the geographic distance, similar to the first option for $\Sigma$ above, however, in this thesis we mainly use GMRF models which can be seen as a spatial extensions of the autoregressive model in option 2.

Regardless of the model, once we have specified the likelihood, the Bayesian continuation is clear: specify priors for the parameters and compute the posterior. If we have global measurements, one could imagine having a parameter representing the global mean temperature, and use the posterior to draw conclusions about the magnitude of global warming. We could also compute the posterior predictive for making predictions for unknown locations, or future time points.

We have given a brief introduction to Bayesian modelling. For a more detailed presentation, see e.g. Gelman et al. (2013) or McElreath (2016).

### 2.2 Spatial statistics

**Spatial data**

Spatial data are commonly divided into three types, as illustrated in Figure 2.2, namely:

1. Continuously referenced data. These are data for which each measurement is associated with a fixed point in a real continuous space. The value of every measurement may be real or discrete. This situation is common in classical applications, as for example in the precipitation data in Figure 2.2, where the measurements were made in weather stations with positions without any particular structure.

2. Grid data. The positions of all measurements are on a fixed discrete grid. These data often come from images, but we saw an example of an exception in the wheat yield data in Figure 1.1. The fMRI data in Figure 2.2 is an example of data on a three-dimensional grid, which is less common than two-dimensional data, and also an example of an incomplete grid, where not every node in the lattice is observed.

3. Point pattern data. These data are also continuously referenced, but there are no measurements in the given points. Instead, the main information is the locations of the points themselves, which are seen as stochastic, and reflects the presence of some phenomenon, for example cancer cases as in Figure 2.2.

This thesis is primarily focused on methods for grid data. However, since many spatial methods are designed for continuously referenced data, we begin by reviewing such methods in this section, and continue with GMRF models for grid data in the next section. Moreover, the notion of continuously referenced spatial stochastic processes is helpful also when considering grid data. In fact, grid data can be seen as a special case of continuously referenced data, with the positions of measurements fixed to the grid. However, as a side-note, a more accurate interpretation of
2. BAYESIAN ANALYSIS OF SPATIAL DATA

grid data would in many cases be as some sort of average over a continuously referenced process in the rectangles or cuboids (pixels or voxels) in the grid, but for simplicity, we will view the measurements as referenced to single points.

Moreover, in Paper V, we consider also non-grid data in the form of marked point patterns, which can be seen as combination of the other two data types. The location of points can be seen as point pattern data, and the mark or label of each point follows a binary valued process, which can be seen as continuously referenced data. However, for analysing this data, we first transform it into grid data by dividing the domain into equally sized rectangles and counting the number of points of each mark within each rectangle, which allows us to use methods for grid data.

For a more comprehensive introduction to spatial statistics, see Gelfand et al. (2010).
Spatial stochastic processes

A spatial stochastic process \( X(s) : s \in D \subseteq \mathbb{R}^d \) is defined for arbitrary spatial locations \( s \) in some domain \( D \) embedded in a \( d \)-dimensional real space. The process is stochastic in the sense that it can be seen as a collection of random variables with a well-defined joint distribution. In particular, for fixed locations \( s_1, \ldots, s_N \), we have that \( (X(s_1), \ldots, X(s_N))^T \) is a random vector, with some joint distribution function

\[
F(x_1, \ldots, x_N; s_1, \ldots, s_N) = P(X(s) \leq x_1, \ldots, X(s_N) \leq x_N).
\]

An important special case is the Gaussian process (GP), which is discussed more below, where the joint distribution is multivariate Gaussian.

It is common to assume a stationary spatial stochastic process, which means that the distributional properties of the process do not vary over space. Strong stationarity can be defined by the condition

\[
F(x_1, \ldots, x_N; s_1 + h, \ldots, s_N + h) = F(x_1, \ldots, x_N; s_1, \ldots, s_N),
\]

for a vector \( h \). Weak stationarity is based only on the first two moments of the process, the mean function \( m(s) \) and a covariance function \( C(s, s') \), defined as

\[
m(s) = E(X(s))
\]

\[
C(s, s') = Cov(X(s), X(s')).
\]

Weak stationarity is defined by the two conditions

\[
E(X(s)) = E(X(s + h)) = m
\]

\[
Cov(X(s), X(s')) = Cov(X(s + h), X(s' + h)) = C(s - s'),
\]

that is, the mean is constant over space, and the covariance\(^1\) between two points \( s \) and \( s' \) depends only on their relative position \( s - s' \). Importantly, for GPs it can be shown that the two definitions of stationarity are equivalent. A third type of stationarity, known as intrinsic stationarity, is used to describe certain processes that do not have a well-defined covariance function, for example by having infinite variance, but for which some increments, such as \( X(s) - X(s') \), are weakly stationary.

Gaussian processes

GPs have been very widely used in spatial statistics and also in machine learning; see Rasmussen and Williams (2006) for an introduction for an introduction in a machine learning context. GPs are incredibly powerful in their simplicity. Even though they describe high-dimensional distribution, they can be completely specified using only a mean function and a covariance function, and they stay within the Gaussian family under marginalization and conditioning. We may specify a GP as

\[
X(s) \sim GP(m(s), C(s, s')).
\]

The joint distribution of \( X \) evaluated at any finite set of fixed locations \( s_1, \ldots, s_N \) is multivariate Gaussian, which can be written as

\[
\begin{bmatrix}
X(s_1) \\
\vdots \\
X(s_N)
\end{bmatrix}
\sim
N
\begin{pmatrix}
m(s_1) \\
\vdots \\
m(s_N)
\end{pmatrix}
\begin{pmatrix}
C(s_1, s_1) & \cdots & C(s_1, s_N) \\
\vdots & \ddots & \vdots \\
C(s_N, s_1) & \cdots & C(s_N, s_N)
\end{pmatrix}.
\]

\[^1\text{We allow different types of inputs to the covariance function } C(\cdot), \text{ and e.g. when the input is a single vector the process is implicitly stationary.}\]
2. Bayesian Analysis of Spatial Data

This means that the marginalization and conditioning properties mentioned follows directly from the same properties for the multivariate Gaussian distribution. For example it is clear that \( X(s_1) \sim \mathcal{N}(m(s_1), C(s_1, s_1)) \) and

\[
X(s_1)|X(s_2) = x_2 \sim \mathcal{N}\left(m(s_1) + C(s_1, s_2)C(s_2, s_2)^{-1}(x_2 - m(s_2)), C(s_1, s_1) - C(s_1, s_2)C(s_2, s_2)^{-1}C(s_2, s_1)\right),
\]

(2.3)

for any \( s_1 \) and \( s_2 \), with obvious extensions replacing \( s_1 \) and \( s_2 \) with groups of locations.

GPs are commonly used as spatial priors, in which case stationarity is often a reasonable assumption. As pointed out earlier, for GPs, stationarity is achieved by having a constant mean \( m(s) = m \) and covariance function \( C(s, s') = C(s - s') = C(r) \). When specifying a spatial GP prior, it is thus sufficient to decide upon a constant mean \( m \) and a covariance function \( C(r) \). One can avoid specifying \( m \) by setting it as a parameter to be learned from the data, but in practice it is common to assume \( m = 0 \). The covariance function can be chosen in many ways, but is required to be positive definite in order to give rise to a valid (positive semi-definite) covariance matrix in Equation (2.2) for every choice of points \( s_1, \ldots, s_N \). For practical applications, there exists a number of well-known families of covariance functions that can be used, which fulfill this requirement. Such covariance functions are controlled with some set of learnable parameters, often referred to as hyperparameters, which allows them to adapt to different datasets.

The Matérn Covariance Function

One of the most common choices is the Matérn covariance function, defined as

\[
C(r) = \frac{\sigma^2}{2^{-\nu} \Gamma(\nu)} (\kappa |r|)^\nu K_\nu(\kappa |r|),
\]

(2.4)

where \( K_\nu(\cdot) \) is a modified Bessel function of the second kind, and \( \Gamma(\cdot) \) is the Gamma function. The Matérn covariance function has three hyperparameters:

- \( \sigma^2 \) controls the marginal variance of the process, \( \text{Var}(X(s)) = \sigma^2 \) for all \( s \). For \( \sigma^2 = 1 \), the covariance function coincides with the correlation function.
- \( \kappa \) controls the spatial correlation range, which can be defined as \( \rho = \sqrt{6\nu}/\kappa \). For inputs \( r \) with \( |r| = \rho \), the correlation function will have values close to 0.13.
- \( \nu \) controls the smoothness, where higher values leads to processes that are more smooth.

One property that makes this understandable is that a random realization of the process is \( n \) times differentiable if and only if \( n < \nu \). Two interesting special cases exists for \( \nu = 1/2 \), which corresponds to the exponential covariance function \( (C(r) \propto \exp(-|r|)) \), and for \( \nu \to \infty \), which corresponds to the squared exponential covariance function \( (C(r) \propto \exp(-|r|^2)) \).

The impact of \( \kappa \) and \( \nu \) on the behaviour of the process is illustrated in Figure (2.3). In the left-hand side we see that varying \( \kappa \), for a fixed \( \nu \), changes how quickly the tail of the correlation function goes to zero with increasing distance. A larger \( \kappa \) leads to more fluctuating realizations. The right-hand side shows the effect of varying \( \nu \) when the spatial correlation range \( \rho \) is fixed. The most important differences between the correlation functions is here for distances close to zero, and in the realizations we see how the smoothness appears to increase with \( \nu \). In total, GPs with Matérn covariance function are extremely flexible and can successfully be used to model many different kinds of data, by varying the hyperparameters, which has made the Matérn covariance a popular choice. The hyperparameters also have useful interpretations. In the practical suggestions for prediction of spatial data by Stein (1999), the author famously summarizes his suggestions for the choice of covariance function by “Use the Matérn model”.

16
2.2. Spatial statistics

Figure 2.3: Matérn correlation functions for various combinations of hyperparameters and 1D realizations of the corresponding GPs.

Anisotropy

The Matérn covariance function is also an example of an isotropic covariance function. This refers to that the covariance only depends on the distance \( r = |\mathbf{r}| \) between points, and we could just as well define \( C(|\mathbf{r}|) \) instead. This is practical, but an unreasonable assumption is numerous applications. Isotropic covariance functions can be made anisotropic by the introduction of a \( d \times d \) positive definite matrix \( \mathbf{H} \), and replacing the argument \(|\mathbf{r}|\) in the isotropic covariance function with \( (\mathbf{r}^T \mathbf{H}^{-1} \mathbf{r})^{1/2} \). This extension makes the isocovariance curves (the set of points for which the covariance with a fixed point is constant) become ellipsoids rather than spheres. The elements of \( \mathbf{H} \) can be added to the other hyperparameters and can be learned jointly. In Paper III, we use this sort of construction with the constraint that \( \mathbf{H} \) is a diagonal matrix, which corresponds to scaling the range of the covariance function along the dimension axes.

Spectral representation

It is in some situations useful to consider the spectral representation of a stochastic process. The spectral density can be defined through the inverse Fourier transform of the covariance function

\[
S(\omega) = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} C(\mathbf{r}) e^{-i \omega^T \mathbf{r}} d\mathbf{r}.
\]
2. **Bayesian analysis of spatial data**

for frequency $\omega \in \mathbb{R}^d$, with the corresponding forward transform

$$C(r) = \int_{\mathbb{R}^d} S(\omega)e^{i\omega^\top r}d\omega.$$  

Since the distribution of a constant-mean GP is completely defined by the covariance function, and functions with the same Fourier transform are almost everywhere identical, the spectral density can be used to analyse the distributions of certain GPs. As an example, a Matérn GP has spectral density

$$S(\omega) = \frac{1}{(2\pi)^d \tau^2 (\omega^2 + \kappa^2)^{d/2}},$$

where $\tau$ is a constant with respect to $\omega$, and is defined by Equation (2.17). In this case $S(\omega)$ is in some sense simpler than $C(r)$, since it does not have any Bessel functions. Two special cases are that the spectral density is Cauchy for $\nu = 1/2$ and $d = 1$ and Gaussian as $\nu \to \infty$.

The properties of the spectral density gives information about the stochastic process and more density mass at higher frequencies corresponds to rougher processes. The spectral density can also be used to simplify mathematical derivations and to construct fast algorithms based on the fast Fourier transform.

**Hierarchical models**

In many applications, it is normal to not model the data using a GP directly, but to instead use the GP as a building block in a larger model. In classical geostatistical models the measurement process $Y(s)$ is decomposed as

$$Y(s) = \mu(s) + X(s) + \epsilon(s), \quad (2.5)$$

where

- $\mu(s)$ is a deterministic mean function. This may include a constant mean $\beta_1$ and spatial covariates $F_2(s), \ldots, F_K(s)$ such as linear spatial trends, or for example altitude in the precipitation example in Figure 2.2. One can then write

  $$\mu(s) = \sum_{i=1}^K F_i(s) \beta_i,$$

  where $F_1(s) = 1$ and $\beta = (\beta_1 \ldots \beta_K)^\top$ are parameters to be estimated.

- $X(s)$ is a zero-mean spatial stochastic process, typically a stationary GP.

- $\epsilon(s)$ is zero-mean spatial white noise, that is

  $$\text{Cov}(\epsilon(s), \epsilon(s')) = \begin{cases} \sigma^2, & s = s' \\ 0, & s \neq s' \end{cases},$$

  independent of $X(s)$.

In the case where all distributions are Gaussian, and a Gaussian prior is assumed also for $\beta$, this model can be written in a hierarchical fashion as

$$\beta \sim \mathcal{N}(m_\beta, \Sigma_\beta)$$  

$$X(s) \sim \mathcal{GP}(0, C(s, s'))$$  

$$Y(s)|X(s), \beta \sim \mathcal{N}(\mu(s) + X(s), \sigma^2),$$

where the last row assumes conditional independence at every point $s$. We may add an additional layer of hierarchy by assuming a prior also for the hyperparameters $\theta$, in which we may
2.3. Gaussian Markov random fields

Given observations \( y = (y_1, \ldots, y_N) \) at \( s_1, \ldots, s_N \), one can show that the likelihood function \( p(y|\theta) \) can be computed on closed form, by integrating out \( X(s) \) and \( \beta \), and that it will be Gaussian. In the machine learning literature, \( p(y|\theta) \) this is sometimes referred to as the marginal likelihood, since \( X(s) \) is marginalized, but should not be confused with \( p(y) \). This makes Bayesian inference simple and straightforward, but is associated with a computational cost of \( O(N^3) \) time and \( O(N^2) \) storage for standard methods, which may be prohibitive when the number of observations \( N \) is large. In Section 2.3, we will see how this problem can be somewhat remedied by replacing the GP in Equation (2.7) with a GMRF.

For fixed hyperparameters \( \theta \), making predictions with this model is also straightforward. This is due to the fact that the posterior predictive process \( Y^*|y \) can be described as a (non-stationary) GP itself. Thus, for a number of \( N^* \) test points \( s^*_1, \ldots, s^*_{N^*} \), similar to Equation (2.3), the conditional distribution \( p(y^*|y) \) is normal, which gives closed form expressions for the predictive mean and variance. In the geostatistical literature this is known as kriging. Again, for large datasets these computations may however be costly. More details are given in Section 2.3, in the analog GMRF case, see Equation (2.22).

Hierarchical models are extremely useful, and can be flexibly adapted to different kinds of data, by replacing the Gaussian measurement equation, Equation (2.8), with some other distribution. In this case \( p(y|\theta) \) is no longer available on closed form and more approximate inference methods must be considered. However, modelling the spatial dependence through a latent GP is still very efficient compared to the alternatives. Bayesian inference is further discussed in Chapter 3.

Log Gaussian Cox processes

We conclude this section by also mentioning a model for spatial point pattern data, the log Gaussian Cox process (LGCP) introduced in (Møller et al., 1998), which we use in Paper V. The LGCP can also be seen as a hierarchical model, where the point pattern in a spatial domain \( D \subset \mathbb{R}^d \) is modelled as a Poisson process where the intensity of points is specified by the exponential function of a latent GP. We let \( N_p(D) \) denote the number of points in an arbitrary subset \( D \subset \mathbb{D} \). Now the LGCP can be defined by

\[
N_p(\tilde{D})|X \sim \text{Poisson} \left( \int_{\tilde{D}} \exp X(s)ds \right)
\]

\[
X(s) \sim \mathcal{GP} (m(s), C(s, s'))
\]

This construction conveniently enables the use GPs to model point patterns, and allows the intensity of points to vary slowly across space, where clusters of points correspond to large values of \( X(s) \) and the absence of points corresponds to small values of \( X(s) \). We can extend this model to allow for fixed effects, by including a deterministic mean function \( \mu(s) \) to \( X(s) \), similar to in Equation (2.5).

2.3 Gaussian Markov random fields

Gaussian Markov random fields (GMRFs) are fields defined for discretely referenced data. They are the main building blocks of the spatial models considered in this thesis. In particular, we use GMRFs for defining spatial models on the grid. The main reason for using GMRFs rather than GPs, introduced in the previous section, is computational. GPs are also known as Gaussian random fields (GRFs), so what essentially makes GMRFs different is a Markov property. This Markov property has the effect for GMRFs to have sparse precision matrices, which are computationally much cheaper to handle than the dense covariance matrices associated with standard GPs. This section gives a brief background on GMRFs, and a more comprehensive presentation can be found in Rue and Held (2005).
Introduction

We can define a GMRF as a Gaussian random vector \( \mathbf{x} = (x_1, \ldots, x_N)^\top \) whose distribution is defined through the mean \( \mu \) and precision matrix \( Q \) as

\[
\mathbf{x} \sim \mathcal{N}(\mu, Q^{-1}).
\]

We here assume that \( Q \) is invertible and positive definite so that \( Q^{-1} \) is a valid covariance matrix. We will extend this definition also to semi-positive definite \( Q \) later. We can also define a graph \( \mathcal{G} = (\mathcal{V}, \mathcal{E}) \) with vertices \( \mathcal{V} \), corresponding to the elements in \( \mathbf{x} \), and edges \( \mathcal{E} \), corresponding to the conditional independencies between these elements, which can be written as

\[
\{i,j\} \notin \mathcal{E} \iff x_i \perp x_j | \mathbf{x}_{-ij}, \quad \text{for all } i \neq j,
\]

where \( -ij \) denotes all elements but \( i \) and \( j \). In other words, an element \( x_i \) is conditionally independent to all non-neighboring elements of \( \mathbf{x} \), with the neighborhoods defined by \( \mathcal{E} \). For example, when \( \mathbf{x} \) is defined on a regular grid, as an image with height \( H \) pixels and width \( W \) pixels, we may represent it using a matrix

\[
\mathbf{X} = \begin{bmatrix}
X_{1,1} & \cdots & X_{1,W} \\
\vdots & \ddots & \vdots \\
X_{H,1} & \cdots & X_{H,W}
\end{bmatrix},
\]

with \( \mathbf{x} = \text{vec}(\mathbf{X}) \) being a vectorized version of \( \mathbf{X} \). In this situation, a reasonable model could be to define all adjacent pixels to be neighbors in the graph, that is, define the neighborhood of \( X_{ij} \) as the four elements

\[
n(X_{ij}) = \{X_{i-1,j}, X_{i+1,j}, X_{i,j-1}, X_{i,j+1}\}.
\]

Such a graph is illustrated in Figure 2.4. Equation (2.9) then implies that all pairs of non-adjacent pixels of the image are conditionally independent given the other pixels in the image, but also that a specific pixel given its neighbors is conditionally independent to all the other pixels. In other words, the values of the neighbors is the only information needed to make predictions about a pixel, and additional information from other pixels does not improve the prediction. This is a Markov property, and the neighborhood of a pixel is sometimes referred to as its Markov blanket.

Moreover, the structure of the graph also determines the zero-pattern of the precision matrix \( Q \) of the GMRF, by the property

\[
\{i,j\} \notin \mathcal{E} \iff Q_{ij} = 0, \quad \text{for all } i \neq j.
\]
2.3. Gaussian Markov random fields

This has a great computational impact as it means that a sparsely connected graph $G$ results in a sparse precision matrix $Q$, and working with a sparse $Q$ automatically results in less storage and fewer computes required for every operation.

Conditional distributions

GMRFs are simple to work with under conditioning. If we partition $x = (x_A, x_B)^T$ and correspondingly $\mu = (\mu_A, \mu_B)^T$ and

$$Q = \begin{bmatrix} Q_{AA} & Q_{AB} \\ Q_{BA} & Q_{BB} \end{bmatrix},$$

the conditional distribution of $x_A|x_B$ is then also a GMRF with mean $\mu_{A|B}$ and precision matrix $Q_{A|B}$ given by

$$\mu_{A|B} = \mu_A - Q_{AA}^{-1}Q_{AB}(x_B - \mu_B)$$

$$Q_{A|B} = Q_{AA}. \quad (2.10)$$

Note that $Q_{A|B}$ is given directly as just a subblock of $Q$. The expression for the conditional mean $\mu_{A|B}$ involves an inverse, but if the set $A$ is small, this inverse will be quite cheap to compute. More importantly, since $Q_{AB}(x_B - \mu_B)$ is a vector, the expression $Q_{AA}^{-1}Q_{AB}(x_B - \mu_B)$ can be computed by solving the sparse matrix equation $Q_{AA}u = Q_{AB}(x_B - \mu_B)$ for $u$, which is typically much cheaper than computing $Q_{AA}^{-1}$ explicitly. Equation (2.10) can be compared to the corresponding equation when the normal distribution is parameterized in terms of the covariance matrix $\Sigma = Q_{AA}^{-1}$ instead with

$$\Sigma = \begin{bmatrix} \Sigma_{AA} & \Sigma_{AB} \\ \Sigma_{BA} & \Sigma_{BB} \end{bmatrix}.$$

In this case the mean and covariance matrix of $x_A|x_B$ are given by

$$\mu_{A|B} = \mu_A + \Sigma_{AB}\Sigma_{BB}^{-1}(x_B - \mu_B)$$

$$\Sigma_{A|B} = \Sigma_{AA} - \Sigma_{AB}\Sigma_{BB}^{-1}\Sigma_{BA},$$

which is in general much more computationally demanding, due to the inverse $\Sigma_{BB}^{-1}$. However, when considering marginalization, that is, computing the distribution of $x_A$, the covariance matrix parameterization is much simpler as $x_A \sim N(\mu_A, \Sigma_{AA})$, which only requires subblocking, whereas a GMRF requires the full inversion of $Q$.

Intrinsic GMRFs on the grid

We now concentrate on a special type of GMRFs known as intrinsic GMRFs (IGMRFs), which are particularly well suited as spatial priors on the grid. However, they bring some complexity by having a precision matrix $Q$ that is not positive definite. For simplicity, let us assume that the spatial domain $D$ of interest is an integer valued cuboid, $D \subseteq \mathbb{Z}^d$.

As a first example, consider the first order IGMRF in one dimension, which can be defined by a distribution for the increments

$$x_{i+1} - x_i \sim \mathcal{N}(0, (\tau^2)^{-1}), \quad i = 1, \ldots, H - 1. \quad (2.11)$$

This defines a random walk process for the $H$-dimensional vector $x$, where the difference between two subsequent elements is distributed as a zero-mean Gaussian with variance $\tau^{-2}$. We can collect all $H - 1$ distribution statements and write this in matrix formulation as

$$Gx \sim \mathcal{N}(0, \tau^{-2}I), \quad G = \begin{bmatrix} -1 & 1 \\ -1 & 1 \\ \vdots & \vdots \\ -1 & 1 \end{bmatrix}.$$
where $G$ is a $(H - 1) \times H$-matrix and empty elements corresponds to 0. The distribution for $x$ can informally be derived by

$$
p(x; \tau) \propto p_{\mathcal{N}(0, \tau^{-2})}(Gx) \propto \exp \left( -\frac{1}{2} (Gx)^\top \tau^2 Gx \right) = \exp \left( -\frac{1}{2} x^\top \tau^2 G^\top G x \right). \tag{2.12}
$$

This form looks like a normal distribution with covariance matrix $(\tau^2 G^\top G)^{-1}$, but by inspecting

$$
G^\top G = \begin{bmatrix}
1 & -1 & & & \\
-1 & 2 & -1 & & \\
& -1 & 2 & \ddots & \\
& & \ddots & \ddots & -1 \\
& & & -1 & 2 \\
& & & & -1 & 1
\end{bmatrix},
$$

one see that it is not invertible, since $G^\top G \mathbf{1} = 0$, so the normalizing constant in the normal distribution $\det(G^\top G) = 0$. However, we may work with $x$ anyway by seeing it as an improper GMRF with improper distribution defined by Equation (2.12). By viewing this as the limiting case of a proper GMRF, we can make a more formal derivation, and we may still use its Markov properties, see Chapter 3 of Rue and Held (2005) for details. For example, this means that

$$
x_i | x_{-i} \sim \mathcal{N} \left( \frac{1}{2} (x_{i-1} + x_{i+1}) \frac{1}{\tau^2}, \frac{1}{\tau^2} \right),
$$

which shows both why the first order IGMRF is a reasonable spatial prior and its Markov properties; the predictive distribution of a single missing value is centered around the mean of its two nearest neighbors. Even though the first order IGMRF does not correspond to any covariance matrix, one can see $x$ as intrinsically stationary, as $x_{i+1} - x_i$ has a stationary distribution.

We may extend the first order IGMRF to grids of arbitrary dimensionality. We do this by extending the definition of $G$ so that it contains one row for every pair of adjacent nodes in the $d$-dimensional grid. Let $(i,j)$ be the $k$th element of $\mathcal{E}$, then set $G_{ij} = -1$, $G_{ij} = 1$ and the rest of row $k$ to zero. This definition implies a simple structure for $R = G^\top G$, namely

$$
R_{ij} = \begin{cases}
|n_i|, & \text{for } i = j \\
-1, & \text{for } i \sim j \\
0, & \text{otherwise}
\end{cases}, \tag{2.13}
$$

where $|n_i|$ is the number of neighbors of node $i$, for example $|n_i| = 4$ for non-boundary nodes in 2D, $|n_i| = 6$ in 3D etc.

We may also consider IGMRFs of higher order. In 1D, these can be constructed by sequentially applying the forward difference operator $F$, defined by $F x_i = x_{i+1} - x_i$. A $k$th order IGMRF can then be defined as

$$
F^k x_i \sim \mathcal{N} \left( 0, \tau^{-2} \right),
$$

for which Equation (2.11) is the special case for $k = 1$. Similar constructions are possible for $d$-dimensional grids. An important special case is the second order IGMRF, defined by

$$
D^2_{i,j} x_i x_j \sim \mathcal{N} \left( 0, \tau^{-2} \right), \tag{2.14}
$$

where $D^2_{i,j}$ is the discrete Laplace operator in $d$ dimensions, which for example in 2D amounts to

$$
-4X_{i,j} + X_{i-1,j} + X_{i+1,j} + X_{i,j-1} + X_{i,j+1} \sim \mathcal{N} \left( 0, \tau^{-2} \right). \tag{2.15}
$$

Equivalently, one can define the second order IGMRF using the previously defined matrix $R$ by $Rx \sim \mathcal{N} \left( 0, \tau^{-2} \right)$ or by saying that $x$ has precision matrix $\tau^2 R^\top R$. The second order IGMRF gives rise to more smooth realizations than the first order IGMRF and is therefore the preferred
2.3. Gaussian Markov random fields

choice in many application, even though it corresponds to a less sparse precision matrix and thus more computation.

The IGMRF models go under many names. In Paper I we refer to the first order IGMRF as the unweighted graph Laplacian (UGL) prior, since this is the name used in the fMRI software SPM. In Paper III we instead refer to it as a first order intrinsic conditional autoregression ICAR(1), and correspondingly use ICAR(2) for the second order model.

The stochastic partial differential equation approach

The stochastic partial differential equation (SPDE) approach presented in Lindgren et al. (2011) offers a valuable link between GMRFs and GPs. It enables the view of a certain type of GMRF as a discretization of a GP with Matérn covariance function, allowing for using the flexibility and interpretability of this covariance function also for GMRF models. Likewise, it enables the computational advantages of GMRFs to be used for continuously referenced spatial data.

Consider the SPDE

\[
(k^2 - \Delta)^{\nu/2} \tau X(s) = W(s), \quad s \in \mathbb{R}^d, \tag{2.16}
\]

where \(\Delta\) is the continuous Laplace operator defined by

\[
\Delta X(s) = \frac{\partial^2}{\partial s_1^2} X(s) + \ldots + \frac{\partial^2}{\partial s_d^2} X(s),
\]

for \(d\) spatial dimensions \(s_1, \ldots, s_d\), and \(W(s)\) is spatial white noise with mean zero and unit variance. It was shown by Whittle (Whittle, 1954, 1963) that a GP \(X(s)\) with Matérn covariance function is a solution to the SPDE. There is a one-to-one correspondence between \(\kappa\), \(\tau\) and \(\alpha\) in the SPDE and the hyperparameters \(\kappa\), \(\alpha^2\) and \(\nu\) of the Matérn covariance function in Equation (2.4). The range parameter \(\kappa\) is the same, the smoothness parameter \(\alpha = \nu + d/2\), and the marginal variance

\[
\alpha^2 = \frac{\Gamma(\nu)}{\Gamma(\nu + d/2) (4\pi)^{d/2} \tau^2 \kappa^{2\nu}}. \tag{2.17}
\]

The solutions to the SPDE in Equation (2.16) only have Matérn covariance for \(\kappa > 0\) and \(\nu > 0\), but for \(\kappa = 0\) or \(\nu = 0\) the SPDE still has solutions which are well-defined random measures.

In order to obtain the link to GMRFs we consider discretized solutions to the SPDE. Lindgren et al. (2011) suggest using a set of compact basis functions based on a triangulation of the spatial domain. They can then use what is known in the numerical analysis literature as the finite difference method, which represents the solution only on this grid. This is essentially done by replacing the Laplace operator \(\Delta\) in the SPDE with the discrete Laplace operator \(D^2_{i,\ldots,j}\) used in Equation (2.14). For example, for \(\alpha = 2\), \(d = 2\) and a discretization on the integer grid it is easily seen that the SPDE can be written as

\[
(k^2 - D^2_{ij}) \tau X_{ij} \sim \mathcal{N}(0, 1)
\]

\[
\iff (k^2 + 4)X_{ij} - X_{i-1,j} - X_{i+1,j} - X_{ij-1} - X_{ij+1} \sim \mathcal{N}
\]

\[
(0, \tau^{-2}) \tag{2.18}
\]

for all \(i\) and \(j\) on the integer grid. This bears a striking resemblance with Equation (2.15), with the only difference being \(k^2\) and a minus sign that could be flipped without changing the distribution, considering the symmetry of the normal distribution. Similar to the derivations of the IGMRFs, it is straightforward to show that Equation (2.18) has a GMRF solution

\[
x \sim \mathcal{N}
\]

\[
(0, \left(\tau^2 (k^2 I + R)^{-1} (k^2 I + R)\right)^{-1})
\].

23
2. BAYESIAN ANALYSIS OF SPATIAL DATA

with \( \mathbf{R} \) defined as in Equation (2.13). This is a convenient result, as it gives a discrete GMRF representation of a continuous Matérn GP, and the computational benefits that come with the sparse precision matrix.

The use of the discrete Laplacian results in some approximation error of the GMRF compared to the GP, but this error can be reduced by using a finer grid. The result is easily generalized to arbitrary dimension \( d \), as long as \( \nu = \alpha \cdot \frac{d}{2} \geq 0 \). For simple solutions we are restricted to integer values of \( \alpha \), but also fractional values can be used (Bolin et al., 2020). For even values of \( \alpha \), the SPDE can be solved with the same strategy as in Equation (2.18), while for odd values of \( \alpha \) we have to interpret the operator \((\kappa^2 \Delta)^{\alpha/2}\) in a pseudodifferential sense, in terms of its Fourier transform, see Lindgren et al. (2011) for details. One can show that for \( \alpha = 1 \) the solution to the SPDE is a GMRF with precision matrix \( \tau^2 (\kappa^2 \mathbf{I} + \mathbf{R}) \), and for higher values of \( \alpha \) the precision matrix \( Q_\alpha \) can be computed in a recursive manner using the following rule:

\[
\begin{align*}
\alpha &= 1 : & Q_\alpha &= \tau^2 (\kappa^2 \mathbf{I} + \mathbf{R}) \\
\alpha &= 2 : & Q_\alpha &= \tau^2 (\kappa^2 \mathbf{I} + \mathbf{R})^T (\kappa^2 \mathbf{I} + \mathbf{R}) \\
\alpha &= 3, 4, \ldots : & Q_\alpha &= (\kappa^2 \mathbf{I} + \mathbf{R})^T Q_{\alpha-2} (\kappa^2 \mathbf{I} + \mathbf{R}).
\end{align*}
\]

Given the link between Matérn GPs and GMRFs, we can reinterpret the intrinsic GMRFs as Matérn GMRFs in the limiting case when \( \kappa \rightarrow 0 \). This corresponds to infinite spatial correlation range \( \rho \), which is a concerning factor for the use of IGMRFs as spatial priors.

In practice, when using the SPDE approach one also have to account for boundary effects. The SPDE only exactly has solutions with Matérn covariance when defined on the whole \( \mathbb{R}^d \), which means that the discretization should also be done with an infinite discrete grid, which is of course impossible to store in a computer. If instead using a finite grid, for example a \( d \)-dimensional cuboid, the marginal distributions of the elements on the edge of the cuboid will be biased, for example having larger variances than the corresponding Matérn GP. However, the further away from the boundary one goes, the smaller this bias will be. A practical remedy for this problem is therefore to extend the grid outside the region of interest, normally where the data are, to make the bias negligible for the important inferences. This will come with a greater computational cost, however.

Hierarchical models

When the data are located on a grid, we may readily replace the GP in the hierarchical models presented in Section 2.2 with a spatial GMRF prior. This leads to substantially reduced computational costs, for example \( \mathcal{O}(N^{3/2}) \) for 2D problems and \( \mathcal{O}(N^2) \) for 3D problems (Rue & Held, 2005) compared to \( \mathcal{O}(N^3) \) for standard GPs, and also to greatly reduced memory requirements.

For completeness, we derive the posterior of the latent field for the simple model where we have a GMRF prior and measurements with Gaussian error

\[
x \sim \mathcal{N}(\mu, \mathbf{Q}^{-1}) \quad y_i | x_i \overset{i.i.d.}{\sim} \mathcal{N}(x_i, \sigma^2), \quad i \in \mathcal{M},
\]

where \( x \) is a vector of length \( N \) and \( \mathcal{M} \subseteq \{1, \ldots, N\} \). We further define \( \mathbf{m} \) as the measurement mask, the vector of length \( N \) with \( m_i = 1_{(i \in \mathcal{M})} \), \( \mathbf{L} \) as the diagonal matrix with diagonal \( \mathbf{m} \), and \( \mathbf{y} \) is the \( N \)-dimensional vector of observations, with \( y_i \) in non-observed locations \( i \notin \mathcal{M} \) set to 0.
2.3. Gaussian Markov random fields

for convenience. The posterior is obtained through Bayes’ theorem as

\[
p(x | y) \propto p(y | x) p(x) \\
\propto \exp \left\{ -\frac{1}{2} \left( \mathbf{y} - \mathbf{x} \right)^T \mathbf{I}_m \left( \mathbf{y} - \mathbf{x} \right) \right\} \exp \left\{ -\frac{1}{2} \left( \mathbf{x} - \mu \right)^T \mathbf{Q} \left( \mathbf{x} - \mu \right) \right\}
\]

(2.20)

which can be identified as the form of a multivariate normal, so that we have

\[
x | y \sim \mathcal{N}(\tilde{\mu}, \tilde{\mathbf{Q}}^{-1}), \quad \text{with} \\
\tilde{\mathbf{Q}} = \mathbf{Q} + \frac{1}{\sigma^2} \mathbf{I}_m, \quad \tilde{\mu} = \mathbf{Q}^{-1} \left( \mathbf{Q} \mu + \frac{1}{\sigma^2} \mathbf{y} \right).
\]

(2.21)

This shows that a GMRF is a conjugate prior for a Gaussian likelihood, where the posterior precision matrix \( \tilde{\mathbf{Q}} \) has the same sparsity structure as the prior precision matrix \( \mathbf{Q} \). Thus the posterior is also spatially correlated, but with different spatial characteristics than the prior. In particular, it will be non-stationary, even if the prior is stationary.

Computing the posterior mean \( \tilde{\mu} \) requires solving an equation system with \( \tilde{\mathbf{Q}} \), which is normally feasible to do for quite large \( N \), at least for the applications considered in this thesis, and this is discussed further in Chapter 3. It is also possible to draw samples from the posterior. However, many applications also require the computation of selected elements of the posterior covariance matrix \( \tilde{\mathbf{Q}}^{-1} \), for example the marginal variances \( \text{Var}(x_i | y) \). Computing the full inverse \( \mathbf{Q}^{-1} \) is prohibitive in general, but selected element can be computed or approximated and this problem is addressed extensively in Paper II. To make predictions for new observations \( y_* \), one can use that the predictive distribution will also be Gaussian, with marginal distributions given by the moments

\[
\begin{align*}
E(y_* | y) &= E_{x|y}[E(y_* | x, y)] = E_{x|y}[x_i] = \tilde{\mu}_i, \\
\text{Var}(y_* | y) &= \text{Var}_{x|y}[E(y_* | x, y)] + E_{x|y}[\text{Var}(y_* | x, y)] = \text{Var}(x_i | y) + \sigma^2.
\end{align*}
\]

(2.22)

This ends our overview of GMRFs, and how they can be used as spatial priors in Bayesian hierarchical models. In the next chapter we will discuss Bayesian inference, in particular the problem of estimating the spatial hyperparameters jointly with the latent spatial field, and we will provide different algorithms for solving this problem.
This chapter deals with methods for inferring the unknown parameters in Bayesian hierarchical models, introduced in the previous chapter. Such models can largely be divided into two types of variables, in addition to the observed data $y$:

- The latent variables $x$. This can be represented as a vector, with a length that typically grows with the size of the data. The data $y$ are modeled as spatially conditionally independent given $x$.
- The parameters $\theta$. This is typically a smaller set of parameters, whose size does not depend on the amount of data, and commonly include “hyperparameters” in the spatial prior, for example $\tau^2$ and $\kappa^2$ for the Matérn GMRF prior, or the measurement noise variance $\sigma^2$.

With this notation, a hierarchical model is specified by the measurement model $p(y|x, \theta)$, the spatial prior $p(x|\theta)$, and a prior for the parameters $p(\theta)$. Normally, real applications have research questions that can be answered by (functions of) the posterior for the latent variables $p(x|y)$, the posterior for the parameters $p(\theta|y)$, and the posterior predictive $p(y^*|y)$, so our goal will be to compute these. Unfortunately, these quantities are rarely available on a simple form. However, in some situations the conditional posterior $p(x|y, \theta)$ will have closed form, for example when the likelihood is Gaussian as in Equation (2.21). This is a key property that will be used to compute the other quantities as well.

This chapter begins with a section which describes algorithms for computing this conditional posterior, and how to draw samples from it. The following sections presents different inference methods for the larger problem: empirical Bayes, Markov chain Monte Carlo, variational inference, and integrated nested Laplace approximations.
3. INFEREN CE METHODS

3.1 Conditional exact inference for the latent variables

When conditioning on the parameters \( \theta \), and the measurement model is Gaussian, a GMRF prior distribution \( p(x|\theta) \) is conjugate, so that the conditional posterior \( p(x|y, \theta) \) is a GMRF as well. An example of such a model was given in Equation (2.19) and the conditional posterior in Equation (2.21), and we will stick to this example throughout this section. Just as in the example, we will drop \( \theta \) in the notation, as this is assumed to be fixed.

Even though the posterior \( x|y \sim \mathcal{N}(\mu, Q^{-1}) \) is theoretically known in closed form, using it in calculations can be computationally challenging when \( N \) is large. Two particular problems are how to compute the posterior mean \( \hat{\mu} = Q^{-1}(Q\mu + \sigma^{-2}y) \) and how to draw a new random sample \( x_s \sim \mathcal{N}(\mu, Q^{-1}) \) from the posterior. We present two solutions to these problems: the sparse Cholesky factorization and the conjugate gradient method.

Sparse Cholesky factorization

The Cholesky factor \( L \) of a symmetric positive definite square matrix \( A \) is a lower triangular matrix such that \( A = LL^T \). It can be shown that such a factorization always exists and is unique. When the factorization is applied to a sparse matrix, such as the posterior precision matrix \( \hat{Q} \), the Cholesky factor will be sparse as well, which is important for computational efficiency. However, \( L \) is normally not as sparse as \( \hat{Q} \), as discussed in more detail below, which may be a limitation for large \( N \).

An equation system involving \( L \), for example \( Lv = b \) for some vector \( b \), is comparatively cheap to solve for \( v \). Since \( L \) is lower triangular, the first equation corresponds to \( L_{11}v_1 = b_1 \), so \( v_1 \) can be obtained immediately. This value can be substituted into the next equation, to obtain \( v_2 \) and so on, resulting in an algorithm known as forward substitution. The complexity of this algorithm is proportional to the number of non-zero elements in \( L \), which is \( O(N^2) \) in general, but for sparse \( L \) normally closer to \( O(N) \). Equation systems with \( LL^T \) are solved equally fast in a reversed order.

Given the Cholesky factorization \( \hat{Q} = LL^T \), the posterior mean \( \hat{\mu} = \hat{Q}^{-1}(Q\mu + \sigma^{-2}y) \) can thus be cheaply computed with the following steps

1. Solve \( Lv = (Q\mu + \sigma^{-2}y) \) for \( v \)
2. Solve \( L^T \hat{\mu} = v \) for \( \hat{\mu} \)

as this implies that

\[
\hat{\mu} = L^{-T}v = L^{-T}L^{-1}(Q\mu + \sigma^{-2}y) = (LL^T)^{-1}(Q\mu + \sigma^{-2}y) = Q^{-1}(Q\mu + \sigma^{-2}y),
\]

where \( L^{-T} = (L^T)^{-1} \). Moreover, we can produce a random sample \( x_s \) from the posterior through the following steps

1. Sample \( z_s \sim \mathcal{N}(0, I) \)
2. Solve \( L^T v_s = z_s \) for \( v_s \)
3. Compute \( x_s = \hat{\mu} + v_s \).

It can be readily verified that \( x_s \) has the correct distribution. This last procedure can also be effectively parallelized in the occasion that we want multiple samples.

The computational bottleneck for this approach is in the actual computation of the Cholesky factor \( L \), and the complexity depends on its resulting sparsity, which will now be discussed. This can be analyzed through the graph \( G = (V, E) \) introduced in Section 2.3.
A cycle in a graph is a path that starts in one of the nodes and passes by some other nodes through connecting edges until it reaches the starting node again. A chordal graph is a graph where all possible cycles of length 4 or longer have a chord, an edge connecting two non-subsequent nodes in the cycle. For example, a 1D grid with edges connecting neighboring grid points corresponds to a chordal graph since it contains no cycles. On the other hand, a rectangular 2D grid does not correspond to a chordal graph, since the cycle that follows the boundary of the rectangle has no chord.

A non-chordal graph can be made chordal by adding edges so that the condition on the cycles is fulfilled. Given an ordering of the nodes, this can be done in a sequential manner. One can start with the first four nodes and if necessary add edges so that their subgraph becomes chordal. The following nodes are included one by one, and for each node additional edges are added to fulfill the condition. The resulting graph is known as the chordal extension of the original graph $G$.

As mentioned in Section 2.3 the sparsity of $\tilde{Q}$ is directly related to the edges $E$. One can show that the sparsity of $L$ is in the same way related to the edges $\bar{E}$ in the chordal extension by $\{(i, j) \notin \bar{E} \implies \bar{L}_{ij} = 0, \text{ for all } i > j\}$.

This means that non-zero elements in $L$ that were zero in $\tilde{Q}$, which can be referred to as fill-in, correspond to the added edges in the chordal extension. The number of edges $|\bar{E}|$ in the chordal extension depends greatly on the ordering of the nodes (which is equivalent to the ordering of the elements in the GMRF $x$) and reordering algorithms exist that aim for minimizing the amount of fill-in, for example the approximate minimum degree reordering (Amestoy et al., 1996). Using such algorithms, the computational complexity of the sparse Cholesky factorization can be reduced to $O(N^{3/2})$ for 2D problems and $O(N^2)$ for 3D problems (Rue & Held, 2005), which is a great reduction compared to standard GPs. Nevertheless, for large $N$ this may be prohibitive, as illustrated in Paper I, and also the memory requirements can be overwhelming. A cheaper alternative is presented next. For a more in-depth introduction to chordal graph and sparse Cholesky factorization we refer to Vandenberghe and Andersen (2015) and Rue and Held (2005).

**Conjugate gradient method**

The linear conjugate gradient (CG) method is an iterative method for solving equation systems. Instead of solving a matrix equation $Qx = b$ through the direct inversion of $Q$ or its Cholesky factors, the problem can be recast as an optimization problem

$$\arg \min_x \frac{1}{2} x^T \tilde{Q} x - b^T x.$$  

Since the first derivative of this expression

$$\nabla \left( \frac{1}{2} x^T \tilde{Q} x - b^T x \right) = \tilde{Q} x - b$$

is zero for $x = Q^{-1} b$, the solution to the original matrix equation is also an optimum to the optimized function. Furthermore, since the second derivative

$$\nabla^2 \left( \frac{1}{2} x^T \tilde{Q} x - b^T x \right) = \tilde{Q}$$

is a positive definite symmetric matrix, it is clear that this optimum is a unique global minimum. Thus, the solution to the optimization problem and the solution to the matrix equation are equal. The procedure of solving a matrix equation by using CG on the corresponding optimization problem can be referred to as a CG solve.

The CG algorithm solves the optimization problem by starting in some arbitrary point $x = x_0$ and in each iteration taking a step towards the optimum. Rather than taking each step in the
direction of the negative gradient, as in the common steepest descent method, each step taken by the CG method is conjugate with respect to the directions of all steps previously taken, that is, the step directions are all orthogonal with respect to the inner product $\langle x_1, x_2 \rangle_Q = x_1^T Q x_2$. This results in an algorithm that avoids the oscillating behavior often seen with steepest descent, and that has faster convergence rate. The algorithm only uses matrix-vector-multiplications with $Q$, which means that method can be implemented in matrix free fashion, that is, without actually storing the matrix $Q$, if the multiplication transform can be carried in some other way, e.g. using convolution. For a more detailed and pedagogical introduction to CG, see Shewchuk (1994).

Even though CG is an iterative method that in each iteration only partially reduces the error of $x$ relative to the exact solution to the matrix equation, one can argue that it is "as exact" as some direct method. The first argument for this is that one can show that CG converges to $x$ of the solution $\hat{x}$ even though CG is an iterative method that in each iteration only partially reduces the error convolution. For a more detailed and pedagogical introduction to CG, see Shewchuk (1994).

The preconditioned conjugate gradient (PCG) method is an extended version of the CG method which means that method can be implemented in matrix free fashion, that is, without actually storing the matrix $Q$, which means that method can be implemented in matrix free fashion, that is, without actually storing the matrix $Q$, if the multiplication transform can be carried in some other way, e.g. using convolution. For a more detailed and pedagogical introduction to CG, see Shewchuk (1994).

The computational complexity of the CG method can be described as $O(N\sqrt{\gamma})$, where $\gamma$ is the condition number of $Q$. Exactly how $\gamma$ depends on $N$ is hard to answer in general, but for a second-order elliptic boundary value problem, for example the Matérn GMRF with $\alpha = 2$, $\gamma$ is often $O(N^{3/2})$ in which case CG is $O(N^{3/2})$ in 2D and $O(N^{4/3})$ in 3D. That means the same complexity as the sparse Cholesky factorization in 2D and much faster in 3D. In addition, the memory requirements for CG are much lower, $O(N)$. The preconditioned conjugate gradient (PCG) method is an extended version of the CG method using the technique in Papandreou and Yuille (2010) by computing $\hat{x}$.

The second line is a matrix equation with the matrix $M^{-1}Q M^{-\top}$, which typically has a much smaller condition number and the equation can therefore be solved much faster. From its solution $\hat{x}$, the solution to the original matrix equation $\hat{x} = M^{-\top}b$ can then be easily obtained. How to choose a good preconditioner $M$ is a large research field in its own, but one alternative is the incomplete Cholesky factor (Manteuffel, 1980).

Finally, once we have the CG (or PCG) method in our toolbox, we might readily use it to compute the posterior mean $\hat{\mu} = Q^{-1}(Q\mu + \sigma^{-2} y)$. We can also use it to sample from the posterior using the technique in Papandreou and Yuille (2010) by computing

$$x_c = Q^{-1} \left( S^\top (z_1 + S\mu) + \frac{1}{\sigma^2} (y + \sigma^2 z_2) \right),$$

where $z_1$ and $z_2$ are standard Gaussian random vectors, and $S$ is such that $S^\top S$ equals the prior precision matrix $Q$. Such a matrix can often be easily found, e.g. $S = \tau (\mathbf{1}^\top + R)$ for the Matérn GMRF with $\alpha = 2$. It can be easily verified that $x_c$ is Gaussian, and that

$$E(x_c) = Q^{-1} \left( S^\top (0 + S\mu) + \frac{1}{\sigma^2} (y + \sigma^2 \mathbf{1}_m) \right) = Q^{-1} \left( Q\mu + \frac{1}{\sigma^2} y \right) = \hat{\mu}$$

$$\text{Cov}(x_c) = Q^{-1} \left( S^\top S + \frac{1}{\sigma^2} \mathbf{1}_m \right) Q^{-\top} = Q^{-1},$$

so this is a valid sampling procedure.

3. Inference Methods
3.2 Empirical Bayes

Empirical Bayes (EB) is, despite its name, not an altogether Bayesian method, since it aims to optimize the parameters $\theta$, rather than computing the full posterior distribution. However, it is fully Bayesian in its treatment of the latent variables $x$. EB can be seen as a two-step procedure:

1. Find $\hat{\theta} = \arg \max_{\theta} p(\theta|y)$.
2. Compute $p(x|y, \hat{\theta})$.

The second step is normally simpler, and was covered in Section 3.1 for the Gaussian case. We will here focus on step 1.

In our setting, the primary elements of $\theta$ are the spatial hyperparameters in the covariance function or GMRF prior, e.g. $\kappa$ and $\tau$ in the Matérn GMRF. Historically, in classical geostatistics the method for deciding upon these parameters was to look at the semivariogram, a transform of the covariance function. The idea was to search for parameters $\theta$ for the model’s covariance function such that the analytical semivariogram made a good fit to the empirical semivariogram, estimated from data at different ranges. A more principled way to estimate $\theta$ is to use maximum (marginal) likelihood, that is, to directly optimize the likelihood of the data $p(y|\theta)$, where $x$ has been marginalised. EB is related to maximum likelihood, since

$$\log p(\theta|y) = \log p(y|\theta) + \log p(\theta) + \text{const},$$

(3.2)

which means that maximizing the left hand side, the log posterior, is equivalent to maximizing the right hand side, the log likelihood with a penalty given by the log prior. When the prior is non-informative, $p(\theta) \propto \text{const}$, the two methods coincide. However, using an informative prior $p(\theta)$ is often a reasonable thing to do. For example, it is known that the range and marginal variance parameters of the Matérn covariance function cannot be consistently estimated under in-fill asymptotics (Stein, 1999; H. Zhang, 2004), meaning that regardless of the amount of data in a bounded domain, the optimum of the likelihood will not converge to a unique point. Instead, the optimal value of the likelihood is attained at a ridge, a set of values for the two parameters for which the likelihood is constant. Given prior knowledge about e.g. the spatial range, this problem can be remedied by using an informative prior $p(\theta)$, for example the penalised complexity prior introduced by Fuglstad et al. (2019).

The optimization in step 1, requires computing $p(\theta|y)$ for various values of $\theta$, or equivalently computing $p(y|\theta)$ and $p(\theta)$ by Equation (3.2). Since the prior $p(\theta)$ is normally simple to compute, the challenging part is the likelihood $p(y|\theta)$. Given the hierarchical model specification (Equation (2.19)), one way to compute this likelihood is through

$$p(y|\theta) = \left. \frac{p(y|x, \theta) p(x|\theta)}{p(x|y, \theta)} \right|_{x=x^{*}},$$

(3.3)

which is a rewriting of Bayes’ theorem, and holds for any value of $x^{*}$. When the posterior $p(x|y, \theta)$ is has closed form, as in the Gaussian case, this equation can be readily used. However, for large $N$, the determinants $\det(Q)$, appearing in the prior $p(x|\theta)$, and $\det(Q)$, appearing in the posterior $p(x|y, \theta)$, can be computationally costly. When we have the Cholesky factorization $Q = LL^{T}$, the corresponding determinant can be cheaply computed using

$$\det(Q) = \left( \prod_{i=1}^{N} L_{ii} \right)^{2},$$

but when the Cholesky factorization is computationally prohibitive, so is also the determinant.

One strategy to overcome this problem is by performing the optimization using steepest descent (in our case ascent), following the gradient $\nabla_{\theta} p(\theta|y)$, whose computation does not require these
3. **Inference Methods**

determinants, but only their derivatives. For example, using matrix derivative rules (Petersen & Pedersen, 2012) one can show that

\[
\frac{\partial}{\partial B_j} \log \det (Q) = \text{tr} \left( Q^{-1} \frac{\partial Q}{\partial B_j} \right),
\]

where \(\text{tr}(\cdot)\) denotes the trace, for which there exists an unbiased approximation

\[
\text{tr} \left( Q^{-1} \frac{\partial Q}{\partial B_j} \right) \approx \frac{1}{N_s} \sum_{s=1}^{N_s} v_s^T Q^{-1} \frac{\partial Q}{\partial B_j} v_s,
\]

which is known as the Hutchinson estimator (Hutchinson, 1990). Here, \(v_1, \ldots, v_{N_s}\) are \(N_s\) vectors of length \(N\) with independent random elements 1 or \(-1\) with equal probability, and we may efficiently compute each \(v_s^T Q^{-1}\) using CG. Since the trace is approximated, the resulting algorithm will be a stochastic gradient descent (ascent) method, and the number of samples \(N_s\) should be chosen to get a good balance between the gradient variance and computation time. Nevertheless, the method can be shown to converge to a local optimum, so it is exact whenever the posterior \(p(\theta)\) is convex. This type of algorithm is used in Paper III, but is enhanced to include also second order derivative information, and we refer there for further details.

Even though this optimization method is exact and faster than using Cholesky factorization, it can be time-consuming for large \(N\) as it needs multiple CG solves in every iteration. Also, it requires deriving the expressions of several matrix derivatives by hand, which limits how simple it is to try out different models. Under certain circumstances, it is possible to achieve faster, but approximate, inference by maximizing a variational lower bound on the posterior \(p(\theta|y)\), instead of the posterior itself. This is presented in more detail in Section 3.4, but the idea is that if the lower bound is reasonably tight, the parameters \(\theta\) maximizing the lower bound will provide a good approximation to the parameters that maximize the posterior. Paper IV uses such an approach, which avoids CG solves in every iteration, thereby attaining linear time complexity. Also, tedious manual matrix derivatives are sidestepped by using automatic differentiation, which enables gradients to be computed algorithmically by the computer with the chain rule.

The main weakness of EB is that it disregards the posterior uncertainty in the parameters \(\theta\). However, in applications where this uncertainty is of minor practical interest, and where the posterior of the latent variables \(p(x|y)\) is well-approximated by \(p(x|y, \hat{\theta})\), this weakness may be rationally ignored. Alternative approaches that properly account for the posterior uncertainty in \(\theta\), for example MCMC and INLA presented below, generally require the costly Cholesky factorization that the EB method avoids. Therefore, EB can often be regarded as fast and sufficiently accurate in practice.

### 3.3 Markov chain Monte Carlo

Markov chain Monte Carlo (MCMC) is a gold standard method for Bayesian inference. The basic Monte Carlo idea is to generate random samples from the joint posterior distribution \(p(x, \theta|y)\), and to use these samples to compute quantities of interest. For example, given \(N_s\) iid. posterior samples \((x^{(1)}, \theta^{(1)}), \ldots, (x^{(N_s)}, \theta^{(N_s)})\) we may approximate the posterior mean of \(x\) as

\[
E(x|y) \approx \frac{1}{N_s} \sum_{i=1}^{N_s} x^{(i)},
\]

and similarly for \(\theta\). More generally, we may approximate the expectation of any function \(g(x)\) as

\[
E(g(x)|y) \approx \frac{1}{N_s} \sum_{i=1}^{N_s} g \left( x^{(i)} \right),
\]

which makes the computation of any posterior statistics such as standard deviations, quantiles etc. very convenient. However, simulating iid. samples from the joint posterior \(p(x, \theta|y)\) is
3.3. Markov chain Monte Carlo

normally not feasible unless the model has a pretty simple structure. The ingenious idea behind MCMC is to instead construct a Markov chain, i.e. a random process where each sample \((x(i), \theta(i))\) is generated out of the previous sample \((x(i-1), \theta(i-1))\) in such a way that the marginal stationary distribution of the process, \(p(x(i), \theta(i))\), is identical to the posterior distribution \(p(x, \theta|y)\). In fact, several algorithms to construct such Markov chains are available, perhaps most famously the Metropolis-Hastings algorithm. These can be readily used for sampling from the posterior of most statistical models, in particular all the models considered in this thesis, at least in theory.

A side effect of the Markov chain construction is that the produced random samples from the different iterations will no longer be independent. Nevertheless, the approximation in Equation (3.4) is still a simulation consistent estimator, but the variance of this estimator is directly related to autocorrelation in the sequence of random draws. Therefore, an important aspect of any MCMC algorithm is how to make this correlation small.

One MCMC algorithm that is especially simple to implement is known as Gibbs sampling. The requirement for using this method is easily sampled conditional posterior distributions. In Section 3.1 how the conditional posterior \(p(x|y, \theta)\) could be efficiently sampled from using CG in the Gaussian case. If the elements of \(\theta = \{\theta_1, \ldots, \theta_p\}\) have easily sampled full conditionals \(p(\theta_i|y, x, \theta_{-i})\) for all \(i\), then Gibbs sampling is possible. The Gibbs sampling algorithm is presented in Algorithm 1.

**Algorithm 1** Gibbs sampling

**Require:** starting value \(\theta(0)\)

1: for \(j = 1\) to \(N_s\) do
2: Draw sample \(x(i)\) from \(p(x|y, \theta^{(j-1)})\)
3: for \(i = 1\) to \(p\) do
4: Draw sample \(\theta_i(i)\) from \(p(\theta_i(i)|y, x(i), \theta_i^{(j-1)}, \theta_{i+1}^{(j-1)}, \ldots, \theta_p^{(j-1)})\)
5: end for
6: end for
7: Return posterior samples \((x(1), \theta(1)), \ldots, (x(N_s), \theta(N_s))\)

The intuition why this algorithm works is as follows: if \(\theta^{(j-1)}\) is truly a random sample from the posterior \(p(\theta|y)\), then \(x(i)\) sampled in step 2 is a sample from the posterior \(p(x|y, \theta)\) as \(p(x|y, \theta)p(\theta|y) = p(x, \theta|y)\). The corresponding reversed argument can be given for all \(\theta_i\) in step 3-5. Any starting value \(\theta(0)\) selected where the posterior \(p(\theta|y)\) has positive support could in theory have been generated as a random sample from \(p(\theta|y)\), although some values are very unlikely. However, if the algorithm is run long enough for any correlation with this starting value to die off, new samples will be correctly sampled from the posterior distribution and practically independent from the starting value. The dependent samples from the first iterations can be discarded from the final output as what is usually referred to as burn-in. A formal proof can be given by noting that Gibbs sampling is a special case of the Metropolis-Hastings algorithm, which can be shown to converge under certain assumptions, but this will not be done here.

One situation which enables easily sampled parameters is for the Gaussian model in Equation (2.19) when the first order IGMRF is used as spatial prior

\[ x \sim N(0, (\tau^2 R)^{-1}). \]

In this case, a Gamma prior for \(\tau^2\) is conjugate with the full conditional \(p(\tau^2|y, x, \sigma^2)\), which then also follows a Gamma distribution. A similar conjugacy exists for the \(\sigma^2\) parameter. This
3. Inference Methods

Type of model is used in Paper I, and the parameters are inferred using Gibbs sampling. When more complicated spatial priors are used, for example the Matérn GMRF, conjugate priors are no longer available for all parameters and Gibbs sampling cannot easily be used. The common strategy to use a so called Metropolis-within-Gibbs step is in our setting not computationally feasible as it requires the computation of the determinant of the prior precision matrix $\det(Q)$ for various values of $\theta$.

In a more primitive version of the Gibbs sampling algorithm, the sampling of $x$ in step 2 would not be done with CG but by iterating over the elements $x_1, \ldots, x_N$ instead, for each element $x_i$ sampling from its full conditional posterior, in the same way that $\theta$ is sampled in step 3-5. Even though such an algorithm would be correct in theory, due to the large spatial correlation between the elements of $x$, the autocorrelation of the simulated Markov chain would be impractically high, leading to slow convergence. Using CG to sample from the joint conditional posterior of $x$ is much more efficient.

The standard deviation of the estimator in Equation (3.4) can be shown to decrease with the number of samples $N_s$ as $O(1/\sqrt{N_s})$, with a hidden constant that depends on the autocorrelation in the posterior samples. This relationship with $N_s$ means that in order to gain an additional digit of accuracy (decrease the standard deviation by a factor 10) we need to produce a hundred times as many samples. This is quite a tough request, and means that we often need to run MCMC algorithms for a large number of iterations to produce satisfactory estimates, resulting in long computation times. Also, in our problems $x$ is often a high-dimensional vector, which leads to large memory requirements for storing the samples $x(1), \ldots, x(N_s)$. Nevertheless, MCMC is invaluable as an asymptotically exact method that can be used to verify the correctness of more approximate methods.

This was a very brief introduction to MCMC; for a deeper review, see e.g. Brooks et al. (2011), Gelman et al. (2013).

3.4 Variational inference

In variational inference (VI), also known as variational Bayes (VB), an intractable posterior distribution $p(x, \theta | y)$ is approximated by a simpler distribution $q(x, \theta)$. Given some assumptions about the structure of the variational approximate posterior $q(x, \theta)$, the goal is to find the $q(x, \theta)$ that minimizes the Kullback-Leibler (KL) divergence

$$D_{KL}(q||p) = - \int q(x, \theta) \log \left( \frac{p(x, \theta | y)}{q(x, \theta)} \right) dx d\theta.$$

The KL divergence $D_{KL}$ can be interpreted as a distance between the two distributions, with $D_{KL} \geq 0 \forall q$ and $D_{KL} = 0$ if and only if $q = p$. The variational approximation can be formulated as the solution to the optimization problem

$$\arg \min_{q \in \mathcal{Q}} D_{KL}(q||p),$$

where $\mathcal{Q}$ is the variational family, i.e. the set of distributions that we have restricted $q$ to belong to.
3.4. Variational inference

Consider that the log marginal likelihood of the data can be rewritten as

$$
\log p(y) = \int q(x, \theta) \log p(y|x, \theta) dx \theta = \ldots \text{[use Bayes' theorem]} \ldots
$$

$$
= \int q(x, \theta) \log \left[ \frac{p(y|x, \theta) p(x, \theta)}{q(x, \theta)} \right] dx \theta
$$

$$
= \frac{\mathcal{L}(q)}{D_{KL}(q||p)} = \int q(x, \theta) \log \left[ \frac{p(x, \theta|y)}{q(x, \theta)} \right] dx \theta - \int q(x, \theta) \log \left[ \frac{p(x, \theta)}{q(x, \theta)} \right] dx \theta
$$

(3.5)

where $\mathcal{L}(q)$ is a lower bound on the marginal likelihood, since $D_{KL} \geq 0$, and is often referred to as the ELBO (evidence lower bound). Thus, minimizing the KL divergence is equivalent to maximizing a lower bound on the marginal likelihood. Since the expression for $\mathcal{L}(q)$ does not contain the intractable posterior $p(x, \theta|y)$, this value can often be computed and optimized.

Two main types of assumptions are normally made for the structure of $q$: mean field VI and parametric VI, which are discussed in turn below.

Mean field VI

The posterior is assumed to factorize as

$$
q(x, \theta) = q(x) \prod_{i=1}^{p} q_{\theta_i}(\theta_i),
$$

so that the parameters are independent in the posterior. Even though this assumption does not say anything about the form of the factor distributions, for certain models the optimal distributions turn out to be tractable despite of $p(x, \theta|y)$ being intractable. This assumption can be modified by assuming independence between different blocks of parameters, for example by factorizing also the distribution of $x$. However, such an assumption is more restrictive and ignores the spatial dependence present between the elements in $x$.

There exists a simple coordinate ascent optimization algorithm that can be used for finding the best posterior approximation $q$. The optimal distribution for one of the factors in Equation (3.4), e.g. $q_1(x)$, given the other factors $q_\theta(\theta_1), \ldots, q_\theta(\theta_p)$, can be computed by

$$
\log q_1(x) = \mathcal{E}_x [\log (p(x|y, \theta))] + \text{const},
$$

where the constant is with respect to $x$. Inserting the conditional posterior expression from Equation (2.20) gives

$$
\log q_1(x) = -\frac{1}{2} x^T E_{\theta} \left[ \sigma^2 + 1 \right] x + x^T E_{\theta} \sigma y + \text{const},
$$

showing that $q_1(x)$ is a GMRF given $q_\theta(\theta)$, and the expectations with respect to $\theta$ can be easily computed for certain models and parameterizations. The corresponding approach can be used to compute the optimal posterior approximation $q_\theta(\theta)$ given $q_1(x)$ and $q_{\theta_{-1}}(\theta_{-1})$. For models where the full conditionals are on closed form, each $q_\theta(\theta)$ is generally on closed form as well. However, involved expectations with respect to $q_1(x)$ may raise computational issues, but unbiased Monte Carlo approximations can be obtained by sampling from the GMRF $q_1(x)$ using CG. The resulting coordinate ascent algorithm is presented below.
3. Inference Methods

Algorithm 2 Coordinate ascent VI

Require: Initial posterior approximation \( q_\theta(x) \)

1. while \( q(x, \theta) \) not converged do
2. Update \( q(x) \propto \exp \left( E_{\phi} \left[ \log (p(x|y, \theta)) \right] \right) \)
3. for \( i = 1 \) to \( P \) do
4. Update \( q_{\theta_i}(\theta_i) \propto \exp \left( E_{q_{\theta_{-i}}} \left[ \log (p(\theta_i|y, x, \theta_{-i})) \right] \right) \)
5. end for
6. end while
7. Return posterior approximation \( q(x, \theta) = q_x(x) \prod_{i=1}^{P} q_{\theta_i}(\theta_i) \)

The algorithm is similar to Gibbs sampling, but instead of sampling from the different parameters in each iteration, their optimal posterior approximation for each parameter is computed instead. This method is used in Paper I.

Parametric VI

In parametric VI, the posterior approximation is assumed to have a certain parametric form \( q_\phi(x, \theta) \), parameterized with variational parameters \( \phi \). To find the best approximation \( q_\phi(x, \theta) \), the ELBO \( L(q) \) in Equation (3.5) is optimized with respect to \( \phi \). The distributional form for \( q_\phi(x, \theta) \) should be chosen so that a good approximation to \( p(x|\theta, y) \) is contained within the parametric family.

In Paper IV, we use parametric VI only for the latent variables \( x \), approximating the conditional posterior \( p(x|\theta, y) \) by \( q_\phi(x) \), and the parameters \( \theta \) are optimized similar to EB. Similar to Equation (3.5) we can write

\[
\log p(y|\theta) \leq L(\theta, \phi, y) = \int q_\phi(x) \log \left[ \frac{p(y|x, \theta) p(x|\theta)}{q_\phi(x)} \right] dx \tag{3.6}
\]

where we have denoted the ELBO as \( L(\theta, \phi, y) \) to emphasize that it is a function of both \( \theta \) and \( \phi \) which we wish to optimize with respect to. Depending on the form of \( q_\phi(x) \) we might in some cases be able to compute the expectation in Equation (3.6) on closed form. In other cases, we might be able to use the reparameterization trick (Kingma & Welling, 2013), which means that we approximate the expectation using random samples from \( q_\phi(x) \), but in such a way that we can still take gradients with respect to \( \phi \) for the optimization. We use a combination of these approaches in Paper IV. Another approach worth mentioning is known as stochastic VI (Hoffman et al., 2013), which approximates the gradient of the ELBO directly.

The optimization returns the optimal values \( \hat{\theta} \) and \( \hat{\phi} \), and the posterior approximation \( q_\phi(x) \) can be used for various posterior assessments. Note that \( \hat{\theta} \) is now optimizing the ELBO, however, if the ELBO is reasonably tight (and we use a non-informative prior \( p(\theta) \)), \( \hat{\theta} \) is also close to the mode of the marginal posterior \( p(\theta|y) \) as in EB. In Paper IV we use this idea, and make posterior predictions using \( p(x|y, \theta) \) which are empirically found to outperform predictions using \( q_\phi(x) \). This approach is a bit unusual; \( q_\phi(x) \) is normally introduced as an approximation because \( p(x|y, \theta) \) is intractable, which is also the case here, but the difference is that we can here sample from \( p(x|y, \theta) \) using Equation (3.1).
3.5 Integrated nested Laplace approximations

We have so far mainly considered the case when the measurement model $p(y|x, \theta)$ is Gaussian, but this is not true for many models. In that case, we will often not have conjugacy, ruling out using Gibbs sampling or mean-field VI for inference. Even though other MCMC or VI approaches could be viable, another popular method for Bayesian inference in latent Gaussian models is the integrated nested Laplace approximations (INLA) (Rue et al., 2009). We use this method in Paper V for Bayesian inference for point process models, where INLA has been shown to be a very accurate approximation and much faster than MCMC (Teng et al., 2017).

Bayesian inference with INLA can be described as a three-step procedure

1. Compute the Laplace approximation $\tilde{p}(\theta | y)$ for the posterior for the parameters $\theta$
2. Compute additional simplified Laplace approximations $\tilde{p}(x_i | y, \theta)$ for the marginal posteriors of the latent variables $x$ conditioned on $\theta$
3. Numerically integrate out $\theta$ to obtain the approximation of the marginal posteriors of the latent variables $\tilde{p}(x_i | y) = \int \tilde{p}(x_i | \theta(x^{(k)}), y) \tilde{p}(\theta(x^{(k)}) | y) \Delta \theta^{(k)}$

The Laplace approximation in the first step is

$$p(\theta | y) \propto \frac{p(y | x, \theta)p(x | \theta)p(\theta)}{p(x | y, \theta)} = \frac{p(y | x, \theta)p(x | \theta)p(\theta)}{p_G(x | y, \theta)} |_{x=x^*},$$

where $p_G(x | y, \theta)$ is a Gaussian approximation of $p(x | y, \theta)$, with the same mode $x^*$ and curvature (observed information). This is essentially the same equation as the one used for EB in Equation (3.3), but in that setting the denominator $p(x | y, \theta)$ was in fact Gaussian already. The simplified Laplace approximations $\tilde{p}(x_i | y, \theta)$ for the second step are more involved, and we refer to Rue et al. (2009) for details, but largely corresponds to marginally correcting the Gaussian approximation $p_G(x | y, \theta)$ for location and skewness. The numerical integration in step 3 is carried out using a central composite design integration scheme, and works well as long as the dimension of $\theta$ is reasonably small.

The resulting approximation error has been empirically shown to be extremely small for many latent Gaussian models. If the posterior uncertainty of $\theta$ is of minor importance and believed to have little effect on the posteriors $p(x_i | y)$, an EB version of INLA can be used, essentially only saving the mode for $\theta$ in step 1 and computing posteriors in step 2, which gives additional speedups. The main reason why we do not use INLA for analyzing fMRI data in Paper I and Paper III that the readily available C++ implementation, accessible through the R package R-INLA, requires the Cholesky factorization which is not computationally feasible for our problems, and that R-INLA does not support 3D spatial models.
Functional magnetic resonance imaging (fMRI) is a non-invasive technique for studying brain activity, and is the primary application used in this thesis to motivate the need for computationally efficient methods for spatial statistical modelling of large datasets. It therefore deserves its own chapter, and we will in what follows give a short review of what an fMRI dataset is, how it is collected, what it is used for, and how its analysis can benefit from spatial priors. For a more comprehensive statistical introduction we refer to Lindquist et al. (2008).

fMRI data

One conception of fMRI data is illustrated in Figure 4.1, which is as a time series of three-dimensional images of the living human brain. The data are collected using an MR scanner, which is essentially a very large magnet which creates a magnetic field of strength typically between 1.5 and 7 Tesla. The temporal resolution with which the images are acquired is normally around 1 second, and each image is collected slice by slice. Within each slice, the spatial inhomogeneity of the magnetic field is measured, by affecting the field with radio frequency pulses, and observing how it returns to the original state. By sequentially varying the frequencies at which the measurements are collected, an image then can be retrieved using the inverse Fourier transform. Finally, the slice-wise images are stacked to make up the full 3D image. In fMRI, the image intensities measure the $T_2^*$ contrast, which strongly depends on the blood oxygen level and the resulting signal is referred to as the BOLD (blood oxygen level dependent) signal (Ogawa et al., 1990). When a neuron in the brain activates it causes a local influx of oxygenated blood, so by analysing the BOLD signal in different voxels one can calculate estimates of the brain activity as well. The spatial resolution of the voxels is typically a few millimeters in each dimension.

After the data acquisition, the fMRI volumes (3D images) are normally preprocessed in various ways in order to remove different artifacts. One preprocessing step is known as slice timing correction, which accounts for the fact that the individual slices in each volume were collected
4. FUNCTIONAL MAGNETIC RESONANCE IMAGING

Figure 4.1: Illustration of fMRI data. A dataset consists of a series of 3D images of the living human brain, collected over time. Each image consists of a large number (~100000) of voxels (3D pixels). The intensity of a specific voxel over time creates a signal known as the BOLD signal. The variability of the BOLD signal partially depends on nearby neural activity. Therefore, analysing the BOLD signal can give information regarding which parts of the brain that are active during an experiment.

At slightly different time points. By interpolating between different slices and time points all the data in a single volume are altered to appear as simultaneously collected. Another important step is head motion correction, where the volumes are realigned to a common template so that the location of the different voxels stays the same over time. Other common steps include masking out the brain from the surroundings, registering the brain also to a common brain atlas if the data from multiple subjects are to be compared in a group analysis, and spatially smoothing the data which will be discussed more later. A single volume of preprocessed fMRI data was shown in Figure 2.2, displaying three perpendicular slices.

In spite of the preprocessing, remaining artifacts from e.g. distortions of the magnetic field from head motion will inevitably still be present in the data, and other noise sources such as respiratory/cardiac effects etc. are also difficult to remove. Overall, fMRI data can be described as quite noisy, which is one motivation for the use of statistical methods. The raw MR signal is in fact complex valued, and a statistical noise distribution for the real and imaginary part are commonly assumed to be Gaussian, which implies a Rician distribution for the magnitude, which is normally the signal studied (Gudbjartsson & Patz, 1995). However, for the signal-to-noise ratios regularly seen in fMRI data, the Rice distribution is well approximated by the Gaussian, which is why Gaussian models are commonplace for fMRI analysis.

The effect on the BOLD signal from the brain activity is not instantaneous. When the neurons fire, oxygenated blood is transported to the site but the increase in the BOLD signal starts about 2 seconds after the activation, sometimes preceded by a small dip. The rise continues until 5-8 seconds after the activation, and then starts to decrease again as the oxygen is consumed by the metabolism. This normally results in an undershoot after roughly 15 seconds, and the BOLD signal finally reaches its original level after about 30 seconds. This whole phenomenon is referred to as the hemodynamics, and the BOLD signal’s dependence on the brain activity is often modeled using a linear time-invariant (LTI) system, with a response function known as the hemodynamic response function (HRF). One common HRF, known as the canonical HRF, is parameterized using a linear combination of two gamma functions. The canonical HRF is shown in Figure 4.2. However, many other models exist and in reality the hemodynamics are likely to vary between brain regions and subjects. This indirect measurement of the neural activity makes fMRI analysis more sensitive to incorrect model assumptions compared to other
popular brain imaging modalities as e.g. electroencephalography (EEG), but can on the other hand provide much higher spatial resolution.

**Hemodynamic response function (HRF)**

![Hemodynamic response function](image)

Figure 4.2: A HRF models how the BOLD signal responds to rapid neural activation spike. Here the commonly used canonical HRF is shown.

![Illustration of the measured BOLD signal, the experimental paradigm (events), and the predicted BOLD signal.](image)

Figure 4.3: Illustration of the measured BOLD signal, the experimental paradigm (events), and the predicted BOLD signal.

**Task-fMRI**

A common setup for fMRI experiments is to let the observed subject in the scanner perform some task, which is referred to as task-fMRI. The task can basically be any task that does not cause the subject to move the head too much, but a simple example is to show images of faces to the subject on a little screen inside the scanner. In this case the purpose is to find out which parts of the subject’s brain is active in processing these images. An illustration is given in Figure 4.3. The blue curve is the measured BOLD signal in a single voxel. The red curve corresponds to
4. Functional Magnetic Resonance Imaging

the experimental paradigm, or the events taking in place. In this case an image is shown to
the subject after 20 seconds, for 20 seconds, then removed for 20 seconds, then shown again for
20 seconds and so on. Given the assumption about the LTI system and the HRF, the predicted
BOLD signal from these events can be computed by simply convolving the red curve with the
HRF. The resulting predicted BOLD signal is shown in yellow. Given the strong similarities in
the patterns of the predicted BOLD and the measured BOLD signal in this simple illustration,
it would be safe to conclude that neurons in this voxel is activated during the processing of the
face images. Another voxel might show a completely different BOLD signal pattern than the
predicted BOLD in which case the conclusion would be the opposite.

Statistically, if we let \( y_t \) denote the measured BOLD in a single voxel and \( x_t \) the predicted BOLD
at time point \( t \), we can formulate the problem using a regression model (Friston et al., 1994)

\[
y_t = x_t \beta + e_t, \quad t = 1, \ldots, T,
\]

where \( e_t \) is zero mean Gaussian noise, usually assumed to be temporally autocorrelated, for
example using an AR process. Given this model, we can compute the posterior distribution of
the regression coefficient \( \beta \), and in particular the posterior probability \( P(\beta > \gamma) \) that \( \beta \) exceeds
some predefined activity threshold \( \gamma \). If this probability is high, say more than 0.9, we can
decide to view the voxel as active. The same analysis can be done for every measured voxel,
at every location resulting in a different posterior probability. We can now plot an image of
the brain where the intensity of each voxel represents the corresponding posterior probability,
which is known as a posterior probability map (PPM). This can be seen as the final output of the
experiment; a map showing which parts of the brain that are active during a specific task and
which are not. Examples of PPMs are shown in Figure 4.4.

In real studies, the regression model will normally have additional covariates, and it can then
be written on matrix form

\[
y = X\beta + e, \quad (4.1)
\]

where \( y \) and \( e \) are \( T \)-dimensional vectors, \( X \) is a \( T \times K \) so called design matrix, and \( \beta \) is a \( K \)-
dimensional regression coefficient vector. The extra covariates can be divided into three groups:

- Additional tasks. For example, we may in addition to showing the subject images of
  faces also show images of houses, and the first and second column of \( X \) will then be
  the corresponding predicted BOLD signal to each task. This allows us to also study the
  contrast \( c = \beta_1 - \beta_2 \), which lets us examine which voxels that get more activated from
  faces than from houses, isolating the effect from face processing itself and not merely
  visual processing.
- HRF temporal derivatives. It is common to also convolve the first and second temporal
derivatives of the HRF with each event signal and add as covariates. This can make the
model more robust to cases when the HRF is misspecified, for example if the peak has a
longer delay.
- Nuisance regressors. These covariates try to capture different sources of explainable
noise. Most commonly, head motion is represented by 6 covariates corresponding to the
time series of parameters describing the rigid body transformation of the head (3 angles
and 3 shifts) which were estimated in the preprocessing. It is also common to include an
intercept. Other regressors may include for example linear trends or respiratory/cardiac
signals if these were measured.

Spatial modelling

As mentioned, we can fit the regression model presented above independently for every voxel
to compute a PPM. However, by doing so we implicitly assume that the brain activity is spatially
non-correlated, which is an assumption that goes against a fact that has been observed in many
neuroscientific studies, that activated brain regions are typically spatially contiguous and locally homogeneous.

Historically, common practice for “solving” this issue has been to spatially smooth the data as a preprocessing step. The smoothing is performed by convolving every fMRI volume with a filter shaped as a Gaussian kernel, with some selected width. The matched filter theorem (Rosenfield & Kak, 1982) states that by smoothing the data with a kernel with frequency structure that matches that of the signal, the signal to noise ratio can be improved. The spatial dependence is then indirectly accounted for by using random field theory in the multiple comparison correction of the entire map of test statistics, see Lindquist et al. (2008) for details. Although this approach is fast, it has several drawbacks, and is not the same as having a spatial generative model for the data. Crucially, the smoothing will affect not only the signal, but also the noise, worsening the common assumption of independent noise between voxels, and may lead to confusing noise with signal, see Sidén and Villani (2018) for an illustration. Also, the filter width will typically be selected by the user rather than estimated from the data.

A more principled approach to incorporate this knowledge into the model, is to introduce a spatial prior over the regression coefficients \( \beta \). We begin by reformulating the regression models as a single, multi-voxel regression model

\[
Y \in [T \times N] = X \in [T \times K] W \in [K \times N] + E \in [T \times N]
\]

where the size of each matrix is indicated underneath it. \( Y \) contains the measured BOLD signal in the \( N \) voxels in its \( N \) columns. \( X \) is the same design matrix as in Equation (4.1). The regression coefficients are now stored in the matrix \( W \), where each column contains the coefficients related to a single voxel, and \( E \) is the noise. Without further assumptions, estimating the parameters of this model is equivalent to fitting the model for each voxel independently. However, we may now assign a spatial GMRF prior for the regression coefficients of covariate \( k \) as

\[
W_{k} \sim N(0, Q_{k}^{-1}),
\]

where \( Q_{k} \) may contain additional parameters to be estimated, for example \( \tau_{k}^{2} \) and \( \kappa_{k}^{2} \) for a Matérn GMRF. We may further assign different priors for the different covariates, for example non-spatial priors for the nuisance regressors, and independence are assumed over covariates so
that we have

$$p(W) = \prod_{k=1}^{K} p(W_k^T).$$

In an alternative formulation, we can write this prior as a single GMRF for $\beta = \text{vec}(W^T)$, with a precision matrix with block-diagonal structure. Since the measurement model is Gaussian and spatially independent, although temporally correlated, one can show that the full conditional posterior of $\beta$ is a GMRF as well. This means we can follow the inference methods described in Chapter 3, treating $\beta$ as the latent variables, and the remaining unknown variables as the parameters. For further model choices and details, see Paper I and III. The right-hand side of Figure 4.4 shows the resulting PPMs from using a second order Matérn GMRF prior, where the estimated brain activity is more concentrated to spatially contiguous regions.
This chapter summarizes the contributions of the thesis and presents some possible directions for future research.

5.1 Summary of contributions

This thesis clearly demonstrates the benefit of using GMRF based statistical models for large-scale spatial datasets. We show how such models can be successfully applied in various applications, with relatively fast computation times. We develop several new inference algorithms, that lead to improved accuracy and computational efficiency. We design novel GMRF based models, and verify their appropriateness by comparing to the existing literature.

We enable whole-brain spatial priors to be used for single subject task-fMRI analysis in Paper I and Paper III. As a starting point, we use the default Bayesian spatial model in the popular neuroimaging software SPM (Penny et al., 2007; Penny et al., 2005). This is a highly structured and hierarchical spatio-temporal model with GMRF priors, but uses an approximate variational inference algorithm. The method is in practice only applicable to single 2D slices or small parcels of the brain due to computational constraints. By using the techniques described in Chapter 3, we are able to do less approximate VI, as well as exact inference with MCMC. The new algorithms are shown to be orders of magnitude faster for some large-scale datasets. Paper III extends upon the same model, by allowing the use of GMRF priors of the flexible and interpretable Matérn family, which are linked to the classical Matérn correlation function using the SPDE approach. These priors have previously been successfully used in many applications with two-dimensional spatial data, but we verify their suitability for three-dimensional fMRI data through model comparison.

In Paper II, we address the problem of computing selected elements of the inverse of a large sparse precision matrix, that is, the covariance matrix. This is for example important for eval-
5. CONCLUDING REMARKS

Evaluating the posterior uncertainty in large-scale GMRF models. We develop several new computationally efficient methods for approximating the selected elements of the covariance matrix. We analyse the theoretical properties of the approximations, and empirically demonstrate their performance on different datasets, showing improvements over existing methods. We also give practical guidelines for how to find a good balance between approximation error and computing time.

In Paper IV, we introduce a formal connection between GMRFs and the CNN models commonly used in machine learning and deep learning. We show that common GMRF models can be written as shallow linear CNNs, and thereby trained using techniques and toolboxes commonly used for CNNs. This connection also opens up for a generalization to deep GMRF models which are more flexible and possibly non-linear models, but that have few parameters and can be trained using a single image. This is in contrast to common CNN applications which require training on a large set of images to avoid overfitting. Our method gives state-of-the-art predictive performance on a dataset of satellite temperatures. Hopefully, this paper can be a first step towards bridging the gap between spatial statistics and deep learning.

Paper V demonstrates how an advanced model with GMRF priors can be used in a different type of application, where the spatial data is not on a grid, but instead a marked point pattern. The data consists of the location and injury status of detected persons on the ground collected by UAVs during a robotic search and rescue operation. By dividing the domain into rectangles of equal size, and counting the number of persons within each rectangle, we transform the data to the grid, making grid adapted GMRF priors feasible. Moreover, we design a structural spatial model for the population spread, the detection probability and the injury probability. We also make use of geographical covariates. Using a combination of fast inference with INLA and search planning using Monte Carlo tree search, we are able to effectively solve the problem in real-time.

5.2 Outlook on future research

In this section, we present some possibilities for future research in spatial statistics, particularly for fMRI analysis. The full set of unexplored extensions and applications where our presented methods could be useful is of course enormous, but we give a few suggestions. Further ideas can be found in the latter parts of the articles in the second part of the thesis.

Scalable spatial fMRI analysis

The field of brain imaging, in particular using fMRI, has great potential for revealing the secrets behind the inner workings of the human brain. The constant development of the technique opens up for new types of analyses to be made and new findings are constantly reported. As 7 Tesla MR scanners become more common, the supply of extremely high-resolution fMRI data increases. Although the methods presented in this thesis are relatively fast, for such datasets, where each brain volume can contain millions of voxels, the computation time could be impractically slow for everyday use. To increase the applicability of our methods for such datasets and for the ones that will come in the future, further improvements or approximations are needed.

A common strategy in computation-heavy methods for fMRI analysis is to use parcellation, that is, to divide the brain into parcels, and to run the analysis separately in each parcel, possibly in parallel. If the computational complexity is super-linear in the number of voxels, which is typically the case for spatial methods, this may substantially reduce the computation time. However, there is an obvious drawback with this strategy, namely that spatial dependencies are not modeled over the parcel boundaries. Nevertheless, assuming independence or conditional independence between brain regions could be an effective approach to reduce computing times, if done in a reasonable way. Here are a few ideas:
5.2. Outlook on future research

- A common recommendation for methods that use parcellation is to use parcels such that the voxels within each parcel behaves similarly. This seems reasonable, but a question is of course how to make sure that this assumption holds for a given parcellation and an arbitrary dataset. A possible solution is to compute the parcellation from the dataset as well, or even better, to infer the brain activity and parcellation jointly. See Chaari et al. (2012) for one attempt.

- A related approach could be to base the independence assumption on anatomical brain boundaries. Abramian et al. (2020) use information from a $T_1$-weighted brain image to decide the direction of spatial smoothing for a spatial GMRF prior. This leads to an anatomically informed spatial prior, but one step further would be to truncate elements in the precision matrix to zero based on the anatomical information. This would give more sparsity and thereby faster computations.

- To address the worry that the results depends on the chosen parcellation, one could let the model include a few different parcellations, which could also be described as a mixture of parcellations, or even integrate out the parcellation using MCMC.

- Another possibility is to not restrict the spatial prior used, but to instead put the independence assumption in a variational posterior approximation. This type of independence assumption was used between every voxel in Penny et al. (2005), and shown to lead to spurious results in Paper I, but using the independence assumption on the parcel level rather than the voxel level could have less severe effects.

Another common strategy for speeding up computation in spatial models is to use spatial basis functions, where the number of basis functions is typically much smaller than the number of data points, with variants in the GP literature such as inducing points methods (Quiñonero-Candela & Rasmussen, 2005) and fixed rank kriging (Cressie & Johannesson, 2008). However, such methods often suffer from over-smoothing, unless the scale of spatial variation is much larger than the distance between the data points. The SPDE approach (Lindgren et al., 2011) suggests placing local basis functions more sparsely where there are less data points to decrease the computational load. For fMRI data, this is difficult since the voxel positions are equally spaced over the domain, but one could consider an adaptive method that makes an initial rough estimate of regions which are unlikely to contain any brain activity, and which ratios the basis functions based on that information for further analysis. A method that implicitly leaves out uninteresting parts of the brain is cortical surface fMRI, which is analyzed with a spatial prior on the two-dimensional surface by Mejia et al. (2019). Additional recent scalable methods in the GP literature that could be considered for fMRI data include multiresolution approximations (Katzfuss, 2017), Vecchia approximations (Katzfuss & Guinness, 2017), and matrix-free GPU methods (Wang et al., 2019).

**Improved spatial modelling of fMRI data**

In addition to computational efficiency, spatial fMRI analysis could also benefit from more flexible spatial models, as demonstrated by the improvements in using the more flexible Matérn model in Paper III. Further steps into this direction would be to use e.g. non-stationary models (Dunlop et al., 2018; Fuglstad et al., 2015), spectral mixture kernels (Wilson & Adams, 2013), or the deep GMRFs in Paper IV.

A separate branch in the field of spatial fMRI analysis focuses on spatially modelling the probability of activation rather than the activation effects as in our work (Bezener et al., 2018; Smith & Fahrmeir, 2007; L. Zhang et al., 2014), which provides a different interesting perspective on spatial modeling. Such models can also be formulated by using latent GMRFs (Quírós et al., 2010), and for that case the computational advances presented in this thesis could be useful.

To make our methods applicable to a wider set of fMRI studies, they should be adapted also to group studies, and not only single-subject analyses. This is a computational challenge, and
one should investigate which approximations that can be made without sacrificing accuracy in a 
group analysis setting. One should also explore better methods for multiple comparison correc-
tion, for example by developing more efficient ways to compute joint PPMs (Bolin & Lindgren, 
2015; Mejia et al., 2019) in large-scale settings.

Deep GMRFs
The deep GMRFs presented in Paper IV have large room for further exploration. A better un-
derstanding of the theoretical of properties of these processes, e.g. stationarity, could facilitate 
learning and help motivate their use as spatial priors. Two possible starting points can be found 
in Bolin and Lindgren (2011) and Heaps (2020). They could also be further empirically com-
pared to other spatial methods in different applications, e.g. with non-Gaussian observations. 
To release the full potential of non-linear deep GMRFs, better inference algorithms or vari-a-
tional approximations are necessary. Additional ideas from the field of deep generative models, 
e.g. flow-based models (Dinh et al., 2017; Kingma & Dhariwal, 2018; Rezende & Mohamed, 
2015) and generative adversarial networks (Goodfellow et al., 2014), could help improve the 
capacity and computational efficiency of these models.
Abramian, D., Sidén, P., Knutsson, H., Villani, M., & Eklund, A. (2020). Anatomically informed Bayesian spatial priors for fMRI analysis, In 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI).

Amestoy, P. R., Davis, T. A., & Duff, I. S. (1996). An approximate minimum degree ordering algorithm. SIAM Journal on Matrix Analysis and Applications, 17(4), 886–905.

Andersson, O., Sidén, P., Dahlin, J., Doherty, P., & Villani, M. (2019). Real-time robotic search using structural spatial point processes, In Proceedings of the 35th Conference on Uncertainty in Artificial Intelligence (UAI).

Baddeley, A., Rubak, E., & Turner, R. (2015). Spatial point patterns: Methodology and applications with R. CRC Press.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. Journal of the Royal Statistical Society. Series B (Methodological), 36(2), 192–236.

Bezener, M., Hughes, J., Jones, G., Et al. (2018). Bayesian spatiotemporal modeling using hierarchical spatial priors, with applications to functional magnetic resonance imaging (with discussion). Bayesian Analysis, 13(4), 1261–1313.

Bolin, D., Kirchner, K., & Kovács, M. (2020). Numerical solution of fractional elliptic stochastic PDEs with spatial white noise. IMA Journal of Numerical Analysis, 40(2), 1051–1073.

Bolin, D., & Lindgren, F. (2011). Spatial models generated by nested stochastic partial differential equations, with an application to global ozone mapping. The Annals of Applied Statistics, 5(1), 523–550.
Bolin, D., & Lindgren, F. (2015). Excursion and contour uncertainty regions for latent Gaussian models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77(1), 85–106.

Brooks, S., Gelman, A., Jones, G., & Meng, X.-L. (2011). *Handbook of Markov chain Monte Carlo*. CRC Press.

Chaari, L., Forbes, F., Vincent, T., & Ciuciu, P. (2012). Hemodynamic-informed parcellation of fMRI data in a joint detection estimation framework, In *International conference on medical image computing and computer-assisted intervention*. Springer.

Cressie, N., & Johannesson, G. (2008). Fixed rank kriging for very large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1), 209–226.

Diggle, P. J. (1990). A point process modelling approach to raised incidence of a rare phenomenon in the vicinity of a prespecified point. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 153(3), 349–362.

Diggle, P. J. (2010). Historical Introduction (A. E. Gelfand, P. Diggle, M. Fuentes, & P. Guttorp, Eds.). In A. E. Gelfand, P. Diggle, M. Fuentes, & P. Guttorp (Eds.), *Handbook of spatial statistics*. CRC Press.

Dinh, L., Sohl-Dickstein, J., & Bengio, S. (2017). Density estimation using real NVP. *arXiv preprint arXiv:1605.08803*, arXiv arXiv:1605.08803v3.

Dunlop, M. M., Girolami, M. A., & Stuart, A. M. (2018). How Deep Are Deep Gaussian Processes? *Journal of Machine Learning Research*, 19, 1–46.

Eklund, A., Nichols, T. E., & Knutsson, H. (2016). Cluster failure: why fMRI inferences for spatial extent have inflated false positive rates. *Proceedings of the National Academy of Sciences*, 113(28), 7900–7905. https://doi.org/10.1073/pnas.1602413113

Friston, K. J., Holmes, A. P., Worsley, K. J., Poline, J.-P., Frith, C. D., & Frackowiak, R. S. (1994). Statistical parametric maps in functional imaging: A general linear approach. *Human brain mapping*, 2(4), 189–210.

Fuglstad, G.-A., Lindgren, F., Simpson, D., & Rue, H. (2015). Exploring a new class of non-stationary spatial Gaussian random fields with varying local anisotropy. *Statistica Sinica*, 25(1), 115–133.

Fuglstad, G.-A., Simpson, D., Lindgren, F., & Rue, H. (2019). Constructing priors that penalize the complexity of Gaussian random fields. *Journal of the American Statistical Association*, 114(525), 445–452.

Gelfand, A. E., Diggle, P., Guttorp, P., & Fuentes, M. (2010). *Handbook of spatial statistics*. CRC Press.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis*. CRC Press.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets, In *Advances in Neural Information Processing Systems*.
Groves, A. R., Chappell, M. A., & Woolrich, M. W. (2009). Combined spatial and non-spatial prior for inference on MRI time-series. *Neuroimage, 45*(3), 795–809.

Gudbjartsson, H., & Patz, S. (1995). The rician distribution of noisy mri data. *Magnetic resonance in medicine, 34*(6), 910–914.

Heaps, S. E. (2020). Enforcing stationarity through the prior in vector autoregressions. *arXiv preprint arXiv:2004.09455*.

Hoffman, M. D., Blei, D. M., Wang, C., & Paisley, J. (2013). Stochastic variational inference. *The Journal of Machine Learning Research, 14*(1), 1303–1347.

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

Katzfuss, M. (2017). A multi-resolution approximation for massive spatial datasets. *Journal of the American Statistical Association, 112*(517), 201–214.

Katzfuss, M., & Guinness, J. (2017). A general framework for Vecchia approximations of Gaussian processes. *arXiv preprint arXiv:1708.06302*.

Kingma, D. P., & Dhariwal, P. (2018). Glow: Generative flow with invertible 1x1 convolutions, In *Advances in Neural Information Processing Systems*.

Kingma, D. P., & Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Krige, D. G. (1951). A statistical approach to some basic mine valuation problems on the Witwatersrand. *Journal of the Southern African Institute of Mining and Metallurgy, 52*(6), 119–139.

Lee, K.-J., Jones, G. L., Caffo, B. S., & Bassett, S. S. (2014). Spatial Bayesian variable selection models on functional magnetic resonance imaging time-series data. *Bayesian Analysis, 9*(3), 699–732.

Lindgren, F., Rue, H., & Lindström, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: The SPDE approach. *Journal of the Royal Statistical Society Series B, 73*(4), 423–498.

Lindquist, M. A. Et al. (2008). The statistical analysis of fMRI data. *Statistical science, 23*(4), 439–464.

Manteuffel, T. A. (1980). An incomplete factorization technique for positive definite linear systems. *Mathematics of computation, 34*(150), 473–497.

Matérn, B. (1960). *Spatial variation* (Doctoral dissertation).

McElreath, R. (2016). *Statistical rethinking: A Bayesian course with examples in R and Stan* (Vol. 122). CRC Press.

Mejia, A. F., Yue, Y., Bolin, D., Lindgren, F., & Lindquist, M. A. (2019). A Bayesian general linear modeling approach to cortical surface fMRI data analysis. *Journal of the American Statistical Association, 1–20*.

Mercer, W., & Hall, A. (1911). The experimental error of field trials. *The Journal of Agricultural Science, 4*(2), 107–132.
Møller, J., Syversveen, A. R., & Waagepetersen, R. P. (1998). Log Gaussian Cox processes. *Scandinavian Journal of Statistics, 25*(3), 451–482.

Nychka, D., Furrer, R., Paige, J., & Sain, S. (2017). Fields: Tools for spatial data [R package version 10.0]. https://doi.org/10.5065/D6W957CT

Ogawa, S., Lee, T.-M., Kay, A. R., & Tank, D. W. (1990). Brain magnetic resonance imaging with contrast dependent on blood oxygenation. *Proceedings of the National Academy of Sciences, 87*(24), 9868–9872.

Papandreou, G., & Yuille, A. L. (2010). Gaussian sampling by local perturbations, In *Advances in Neural Information Processing Systems*.

Penny, W., Flandin, G., & Trujillo-Barreto, N. (2007). Bayesian comparison of spatially regularised general linear models. *Human brain mapping, 28*(4), 275–293.

Penny, W., Trujillo-Barreto, N. J., & Friston, K. J. (2005). Bayesian fMRI time series analysis with spatial priors. *NeuroImage, 24*(2), 350–362.

Petersen, K. B., & Pedersen, M. S. (2012). *The matrix cookbook*. Version 20121115. https://doi.org/10.1111/j.1365-294X.2006.03161.x

Quiñonero-Candela, J., & Rasmussen, C. E. (2005). A unifying view of sparse approximate Gaussian process regression. *Journal of Machine Learning Research, 6*(Dec), 1939–1959.

Quirós, A., Diez, R. M., & Gamerman, D. (2010). Bayesian spatiotemporal model of fMRI data. *NeuroImage, 49*(1), 442–456.

Rasmussen, C. E., & Williams, K. I. (2006). *Gaussian processes for machine learning*. MIT Press. https://doi.org/10.1142/S0129065704001899

Rezende, D., & Mohamed, S. (2015). Variational inference with normalizing flows, In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*.

Rosenfield, A., & Kak, A. C. (1982). *Digital picture processing*. Academic Press.

Rue, H., & Held, L. (2005). *Gaussian Markov random fields: theory and applications*. CRC Press.

Rue, H., Martino, S., & Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the royal statistical society: Series b (statistical methodology), 71*(2), 319–392.

Shewchuk, J. R. (1994). An introduction to the conjugate gradient method without the agonizing pain. Technical Report CS-94-125, Carnegie Mellon University, arXiv 1102.0183.

Sidén, P., Eklund, A., Bolin, D., & Villani, M. (2017). Fast Bayesian whole-brain fMRI analysis with spatial 3D priors. *NeuroImage, 146*, 211–225.

Sidén, P., Lindgren, F., Bolin, D., Eklund, A., & Villani, M. (2019). Spatial 3D Matérn priors for fast whole-brain fMRI analysis. *arXiv preprint arXiv:1906.10591*.

Sidén, P., Lindgren, F., Bolin, D., & Villani, M. (2018). Efficient covariance approximations for large sparse precision matrices. *Journal of Computational and Graphical Statistics, 27*(4), 898–909.
Sidén, P., & Lindsten, F. (2020). Deep Gaussian Markov random fields, In Proceedings of the 37th International Conference on Machine Learning (ICML).

Sidén, P., & Villani, M. (2018). Discussion on the paper Bayesian spatiotemporal modeling using hierarchical spatial priors, with applications to functional magnetic resonance imaging by Bezener, Hughes and Jones. *Bayesian Analysis, 13*(4), 1261–1313.

Smith, M., & Fahrmeir, L. (2007). Spatial Bayesian variable selection with application to functional magnetic resonance imaging. *Journal of the American Statistical Association, 102*(478), 417–431.

Stein, M. L. (1999). *Interpolation of spatial data. Some theory for kriging.* Springer-Verlag.

Teng, M., Nathoo, F., & Johnson, T. D. (2017). Bayesian computation for log-Gaussian Cox processes: A comparative analysis of methods. *Journal of Statistical Computation and Simulation, 87*(11), 2227–2252.

Vandenberghe, L., & Andersen, M. S. (2015). Chordal graphs and semidefinite optimization. *Foundations and Trends® in Optimization, 1*(4), 241–433.

Wang, K., Pleiss, G., Gardner, J., Tyree, S., Weinberger, K. Q., & Wilson, A. G. (2019). Exact Gaussian processes on a million data points, In Advances in neural information processing systems.

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

Whittle, P. (1963). Stochastic processes in several dimensions. *Bulletin of the International Statistical Institute, 40*(2), 974–994.

Wilson, A., & Adams, R. (2013). Gaussian process kernels for pattern discovery and extrapolation, In Proceedings of the 30th International Conference on Machine Learning (ICML).

Zhang, H. (2004). Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics. *Journal of the American Statistical Association, 99*(465), 250–261.

Zhang, L., Guindani, M., Versace, F., & Vannucci, M. (2014). A spatiotemporal nonparametric bayesian variable selection model of fMRI data for clustering correlated time courses. *NeuroImage, 95*, 162–175.
Papers

The papers associated with this thesis have been removed for copyright reasons. For more details about these see:

http://urn.kb.se/resolve?urn=urn:nbn:se:liu:diva-165872
Scalable Bayesian spatial analysis with Gaussian Markov random fields