Sparse Log Gaussian Processes via MCMC for Spatial Epidemiology

Jarno Vanhatalo JARNO.VANHATALO@TKK.FI
Aki Vehtari AKI.VEHTARI@TKK.FI
Laboratory of Computational Engineering
Helsinki University of Technology
P.O.Box 9203, FIN-02015 TKK, Espoo, Finland

Editor: Neil Lawrence, Anton Schwaighofer and Joaquin Quiñonero-Candela

Abstract

Log Gaussian processes are an attractive manner to construct intensity surfaces for the purposes of spatial epidemiology. The intensity surfaces are naturally smoothed by placing a Gaussian process (GP) prior over the relative log Poisson rate, and the spatial correlations between areas can be included in an explicit and natural way into the model via a correlation function. The drawback with using a Gaussian process is the computational burden of the covariance matrix calculations. To overcome the computational limitations a number of approximations for Gaussian process have been suggested in the literature. In this work a fully independent training conditional sparse approximation is used to speed up the computations. The posterior inference is conducted using Markov chain Monte Carlo simulations and the sampling of the latent values is sped up by a transformation taking into account their posterior covariance. The sparse approximation is compared to a full GP with two sets of mortality data.

Keywords: Sparse Log Gaussian Process, MCMC, FITC, pseudo-input, Poisson, HMC

1. Introduction

Spatial epidemiology concerns both describing and understanding the spatial variation in the disease risk in geographically referenced health data. One of the main classes of spatial epidemiological studies is disease mapping, where the aim is to describe the overall disease distribution on a map and, for example, highlight areas of elevated or lowered mortality or morbidity risk (e.g. Lawson, 2001; Richardson, 2003; Elliot et al., 2001). The spatially referenced health data may be point level, appointing to continuously varying co-ordinates and showing for example home residence of diseased people. More commonly, however, the data are areal level, referring to a finite sub-region of space, as for example, county or country and telling the counts of diseased people in the area (e.g. Banerjee et al., 2004).

In this work the aim is to construct a model to study the spatial variations in relative mortality risk in areally referenced health-care data. The data are aggregated from point-referenced data into lattices of various grid cell sizes. The data are geographically more accurate than areal level data where the subregions are defined by governmental districts. However, high resolution lattices usually contain empty cells that appoint to areas of no population and this kind of *areally sparse data* may lead to problems with certain models.

The mortality in areas is modeled as a Poisson process with mean intensity surface, which is the product of a standardized expected number of deaths and a relative risk. The expected number of deaths is evaluated using age, gender and scholarly degree standardization and the logarithm of

the relative risk is given a Gaussian process prior. Compared to conditional autoregressive (CAR) -models, GPs are more suitable for point referenced and areally sparse data and provide a more flexible way to describe the form of the spatial prior with a combination of different covariance functions.

The drawback with using GP is the computational burden of the required covariance matrix inversion, which limits the study either to very small areas or a coarse grid. To overcome the computational limitations a number of sparse approximations for GP have been suggested in the literature. Here the computations are sped up with a fully independent training conditional (FITC) sparse approximation (Snelson and Ghahramani, 2006; Quiñonero-Candela and Rasmussen, 2005).

In spatial epidemiology, it is very important to have good estimates of whether the spatial variation in disease risk is significant. To set a golden standard for the uncertainty estimates both the hyperparameters and the latent values of Gaussian process are marginalized out using Markov chain Monte Carlo (MCMC) methods. The sampling is conducted using the hybrid Monte Carlo (HMC) method to sample from the conditional distributions of latent values given the covariance function parameters and the covariance function parameters given the latent values. The mixing of latent value sampling is improved with a transformation taking into account their approximate conditional posterior precision. The use of the HMC method requires the gradients of the logarithm of the marginal likelihood, which in the case of the sparse approximation are evaluated without forming the full covariance matrix.

The main focus of the work is to test the usability of the sparse approximation for the disease mapping problem and give a detailed description of its implementation. To test the approach, the full and sparse Gaussian process models, with four different covariance functions, are applied to two mortality data sets and compared with 10-fold cross-validation using the log predictive density diagnostics. Maps revealing the posterior relative risk are also presented. The focus of the work is in the implementation and the results of the performance of the approximation and its effect on the spatial inference are still preliminary.

2. The Data

The data comprise of a lattice data set containing mortality and population data from the years 1970–1999. The whole country of Finland is included, spanning an area over 1100km in height and more than 600km in width. The standard population is approximately 5 million people and there are around 200 000 deceased for each five-year period. The data list every death with one month accuracy and provides snapshots of the population from census surveys conducted every five years.

The data are aggregated by Statistics Finland from point-referenced data into a lattice, formed of 250m $\times$ 250m grid cells. Background population and the number of deaths for each cause of death were provided as counts pointed to cells. At its highest accuracy the number of data points is computationally prohibitive and thus in the study the data are further aggregated in lattices of larger cells. The data consist of six covariates. 1) Age, 2) sex, 3) cause of death, 4) date of death, 5) co-ordinates of the lattice cell, within which the individual had a home and 6) scholarly degree of an individual.

3. Model

The model constructed in this work follows the general approach discussed, for example, by Best et al. (2005). The data are aggregated into areas A_i with co-ordinates $(x_{i,1}, x_{i,2})$. The mortality in

an area A_i is modeled as a Poisson process with mean $E_i\mu_i$, where E_i is the standardized expected number of deaths in the area A_i, and the μ_i is the relative risk, which is given a Gaussian process prior.

3.1 Sparse log Gaussian process model

The standardized expected number of deaths E_i is evaluated following the idea of the directly standardized rate (e.g. Ahmad et al., 2000), where the rate of death in an area is standardized according to the age distribution of the population in that area. The expected value in the area A_i is obtained by summing the products of the rate and population over the age-groups in the area

$$E_i = \sum_{r=1}^{R} \frac{Y_r}{N_r} n_{ir},$$

where Y_r and N_r are the total number of deaths and people in the whole area of study in the age-group r, and n_{ir} is the number of people in the age-group r and in the area A_i. Here, the population was first divided between genders and both genders were then partitioned into 14 age segments accounting in 28 age-gender groups. All the age-gender groups were further partitioned with respect to 3 scholarly degrees accounting into 66 groups in total, since all scholarly degrees are not present for all age-gender groups. A better approach than standardization would be to give a probabilistic model also for E_i. However, since the amount of data is large, the standardization for E_i should be a rather reliable estimate, and thus modeling of E_i is left for future improvement.

The log relative risk is given a Gaussian process prior with zero mean and different covariance functions, for example *squared exponential*

$$k_{\text{sexp}}(\mathbf{x}_i, \mathbf{x}_j) = \sigma^2_{\text{sexp}} \exp\left(-r^2/l^2\right), \tag{1}$$

where $r = |\,\mathbf{x}_i - \mathbf{x}_j\,|$, and l and σ^2_{sexp} are the length-scale and magnitude, respectively. It is a priori plausible that process variance is zero or very small and thus the prior for the covariance function parameters should be such that it enables both the length-scale and the magnitude to reach zero. To obtain these characteristics the covariance function parameters, $\theta = \left[l, \sigma^2_{\cdot}\right]$, are given a half-Student's t-prior. In case of the length-scale this is related to the choice of width of population prior in hierarchical normal model discussed by Gelman (2006). This results in the complete model

$$\begin{aligned}
\mathbf{Y} &\sim \text{Poisson}(\mathbf{E}\mu) \\
\log(\mu) = \mathbf{f} &\sim \mathcal{GP}(\mathbf{0}, \mathbf{K}_{\text{f,f}}) \\
p(\theta_l|\nu, A) &\propto \begin{cases} 0 & \text{if } \theta_l < 0, \\ \left(1 + \frac{1}{\nu}\left(\frac{\theta_l}{A}\right)^2\right)^{-(\nu+1)/2} & \text{otherwise,} \end{cases}
\end{aligned}$$

where $\mathbf{f}$ is the latent value of the Gaussian process, $\left[\mathbf{K}_{\text{f,f}}\right]_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$, and A is the scale and ν the degrees of freedom of the half-Students't distribution.

Due to the fact that the population sizes in general are large and the number of disease cases relatively small, the Poisson distribution can be considered as a good approximation for the underlying binomial distribution in the case of a large population size and a small number of disease cases. However, in certain parts of Finland the population sizes are rather small and thus Bernoulli or zero-inflated Poisson models could be considered instead. The Gaussian process should be a reasonable

choice to construct the intensity surface for the relative risk, since the surface is naturally smoothed by the process and the spatial correlations between areas can be included in an explicit and natural way into the model via the correlation function.

4. Methods

The study focus in this work is the posterior distribution of the relative risk $\mu = \exp(\mathbf{f})$, which can not be solved analytically because of the Poisson likelihood. In the case of GPs with a non Gaussian likelihood the posterior inference is often conducted by using simpler parametric approximations for the posterior of latent values, and a point estimate for the hyperparameters obtained by maximizing the approximate marginal likelihood. Here, in order to set a golden standard for the uncertainty estimates, the posterior of both the hyperparameters and the latent values of the Gaussian process, are approximated by Markov chain Monte Carlo methods.

The computational time needed in Gaussian process models could be reduced with a simple subsampling of the data, or in the case of spatial epidemiology by aggregating the data into larger cells. In these approaches, however, the approximation is given for the data and some of its information is lost. In order to maintain the high accuracy of the data we use the recently proposed fully independent training conditional (FITC) sparse approximation for the GP prior. The approximation was first introduced by Snelson and Ghahramani (2006) with the name sparse pseudo-input Gaussian process, but the name and the notation used here follow the treatment of Quiñonero-Candela and Rasmussen (2005).

4.1 Conducting the posterior inference using MCMC

The sampling from the joint posterior of hyperparameters $\theta = \left[l, \sigma_{\cdot}^2\right]$ and the latent values $\mathbf{f}$ is performed by alternate sampling from the conditional distributions, $p(\theta|\,\mathbf{f}, D)$ and $p(\mathbf{f}\,|\theta, D)$, via the hybrid Monte Carlo method (Duane et al., 1987; Neal, 1996). The HMC method uses the basic idea of the Metropolis-Hastings algorithm, where random walk behavior is reduced using the gradient information of the negative log posterior cost function,

$$E = -\log\left(p(\mathbf{y}\,|\,\mathbf{f})\right) - \log\left(p(\mathbf{f}\,|\theta)\right) - \log\left(p(\theta)\right), \tag{2}$$

with respect to the sampled parameters. Hybrid Monte Carlo becomes especially practical when sampling high dimensional distributions, since it suffers the dimensionality less than, for example, the simple Metropolis-Hastings algorithm. In the case of a GP prior, the computationally most time consuming operation is the inversion of the covariance matrix, $\mathbf{K}_{\mathrm{f,f}}$, needed in the second term of (2) and in its derivatives,

$$\begin{aligned}
\log\left(p(\mathbf{f}\,|\theta)\right) &= \frac{1}{2}\log\left|\mathbf{K}_{\mathrm{f,f}}\right| + \frac{1}{2}\,\mathbf{f}^{\mathrm{T}}\,\mathbf{K}_{\mathrm{f,f}}^{-1}\,\mathbf{f} - \frac{n}{2}\log(2\pi), \\
\frac{\partial \log\left(p(\mathbf{f}\,|\theta)\right)}{\partial \theta} &= \frac{1}{2}\mathrm{tr}\left(\mathbf{K}_{\mathrm{f,f}}^{-1}\frac{\partial\,\mathbf{K}_{\mathrm{f,f}}}{\partial\theta}\right) - \frac{1}{2}\,\mathbf{f}^{\mathrm{T}}\,\mathbf{K}_{\mathrm{f,f}}^{-1}\frac{\partial\,\mathbf{K}_{\mathrm{f,f}}}{\partial\theta}\mathbf{K}_{\mathrm{f,f}}^{-1}\,\mathbf{f}, \\
\frac{\partial \log\left(p(\mathbf{f}\,|\theta)\right)}{\partial\,\mathbf{f}} &= \mathbf{K}_{\mathrm{f,f}}^{-1}\,\mathbf{f}\,.
\end{aligned} \tag{3}$$

The inversion of the full covariance matrix needs $O(n^3)$ time , where n is the number of data points, but with the FITC approximation discussed next the required time is only $O(m^2 n)$, where $m \ll n$.

4.2 Sparse approximation for Gaussian process

The FITC approximation is based on introducing an additional set of latent values $\mathbf{u} = [u_1, ..., u_m]^{\mathrm{T}}$, called *inducing variables*, that correspond to a set of input locations $\mathbf{x}_u$, called *inducing inputs*. The inducing variables are given a zero mean Gaussian prior $\mathbf{u} \sim N(\mathbf{0}, \mathbf{K}_{\mathrm{u,u}})$ and using inducing variables the prior of the latent values is given by the approximation

$$p(\mathbf{f}) \approx q(\mathbf{f}) = \int q(\mathbf{f} \mid \mathbf{u}) p(\mathbf{u}) \mathrm{d}\mathbf{u}, \tag{4}$$

where $\mathbf{f}$ is interpreted to be conditional on $\mathbf{u}$ through the inducing conditional $q(\mathbf{f} \mid \mathbf{u})$. The exact conditional, which would leave the prior $p(\mathbf{f})$ unchanged, would be $N(\mathbf{K}_{\mathrm{f,u}} \mathbf{K}_{u,u}^{-1} \mathbf{u}, \mathbf{K}_{\mathrm{f,f}} - \mathbf{Q}_{\mathrm{f,f}})$, where $\mathbf{Q}_{\mathrm{f,f}} = \mathbf{K}_{\mathrm{f,u}} \mathbf{K}_{\mathrm{u,u}}^{-1} \mathbf{K}_{\mathrm{u,f}}$, and $\left[\mathbf{K}_{\mathrm{f,u}}\right]_{ij} = \mathbf{k}(\mathbf{x}_i, [\mathbf{x}_u]_j)$. However, in the FITC approximation the inducing conditional is approximated by

$$q_{\mathrm{FITC}}(\mathbf{f} \mid \mathbf{u}) = N(\mathbf{K}_{\mathrm{f,u}} \mathbf{K}_{u,u}^{-1} \mathbf{u}, \mathrm{diag}\left[\mathbf{K}_{\mathrm{f,f}} - \mathbf{Q}_{\mathrm{f,f}}\right]),$$

where the diagonal matrix $\mathrm{diag}\left[\mathbf{K}_{\mathrm{f,f}} - \mathbf{Q}_{\mathrm{f,f}}\right]$ will be denoted in the following by Λ. By integrating out the inducing variables from (4) an approximate prior over latent values is obtained as

$$\mathbf{f} \sim \mathcal{GP}(\mathbf{0}, \mathbf{Q}_{\mathrm{f,f}} + \Lambda). \tag{5}$$

Using m inducing inputs and the approximate prior, the inversion of the matrix $\mathbf{K}_{\mathrm{f,f}}$ is transformed into the inversion of $\mathbf{Q}_{\mathrm{f,f}} + \Lambda$, where $\mathbf{Q}_{\mathrm{f,f}}$ is of rank m. The inverse can be evaluated effectively using a *matrix inversion lemma*, or the Woodbury, Sherman and Morrison formula (e.g. Harville, 1997),

$$\left(\mathbf{Q}_{\mathrm{f,f}} + \Lambda\right)^{-1} = \Lambda^{-1} + \Lambda^{-1} \mathbf{K}_{\mathrm{f,u}} \left(\mathbf{K}_{\mathrm{u,u}} + \mathbf{K}_{\mathrm{u,f}} \Lambda^{-1} \mathbf{K}_{\mathrm{f,u}}\right)^{-1} \mathbf{K}_{\mathrm{u,f}} \Lambda^{-1}, \tag{6}$$

where the inversion of the $n \times n$ matrix Λ is easy, since it is diagonal. Here the computationally most time consuming operations are the matrix multiplications, which only need time $O(m^2 n)$.

In the case of full GP, the evaluation of gradients of the negative log posterior cost function, needed in HMC, is straightforward, since the entries of $\frac{\partial \mathbf{K}_{\mathrm{f,f}}}{\partial \theta}$ in (3) are obtained directly from the derivatives of the covariance function $k(\mathbf{x}_i, \mathbf{x}_j)$. However, in the case of FITC approximation the gradients of $\mathbf{Q}_{\mathrm{f,f}} = \mathbf{K}_{\mathrm{f,u}} \mathbf{K}_{u,u}^{-1} \mathbf{K}_{\mathrm{u,f}}$ can not be evaluated without matrix operations and thus the calculations become more awkward. Snelson and Ghahramani (2006) have used a gradient ascent method for optimizing the hyperparameters and the locations of the inducing inputs in their work. However, they have omitted the calculations from their paper and thus we included the implementation of the gradient evaluations in the Appendix A.

The inducing variables are integrated out from the approximate prior (5), but the inducing inputs do influence the solution and the choice of their locations should be considered carefully. Here, the inducing inputs were placed on a uniform grid that is sparser than the lattice grid of the data. The locations of inducing inputs should be reasonable if the distance from data inputs to the nearest inducing input is less than the length-scale. A natural approach, at least if the FITC approximation is seen as a model in its own right, would also be to integrate over the locations of the inducing inputs.

4.3 Transformation of latent values

The posterior distribution of latent values is proportional to the product of the GP prior and the Poisson likelihood. The latent values are correlated and have a wide range of variances, which in turn may lead to slow mixing in the sampling. To improve the mixing and speed up the sampling, latent values are transformed with respect to their approximate posterior covariance Σ and the sampling is conducted in the resulting $\tilde{\mathbf{f}} = \Sigma^{-1/2}\mathbf{f}$ space. In the case of a full GP, the approach follows the idea of Christensen et al. (2006) and here it is extended for the FITC sparse approximation.

By giving a normal approximation for the likelihood at its mode the approximate posterior precision can be obtained as a sum of the precisions of the prior and the likelihood, $\Sigma^{-1} = \mathbf{K}_{\mathrm{f,f}}^{-1} + \Sigma_l^{-1}$. Here the precision of the likelihood is approximated with a second derivative of the log Poisson in the mode $\Sigma_{\mathrm{l}}^{-1} \approx -\frac{\partial^2}{\partial f^2}\log(\mathrm{Poisson}(E\mu)) = \mathrm{diag}\,[E_1\mu_1, ..., E_n\mu_n]$, which is the product of the age adjusted risk and the relative risk. In the FITC approximation, $\mathbf{Q}_{\mathrm{f,f}} + \Lambda$ replaces the prior covariance $\mathbf{K}_{\mathrm{f,f}}$, and the posterior precision transforms into

$$\Sigma_{\mathrm{FITC}}^{-1} = \left(\mathbf{Q}_{\mathrm{f,f}} + \Lambda\right)^{-1} + \mathrm{diag}\,[E_1\mu_1, ..., E_n\mu_n]\,. \tag{7}$$

The transformation $\tilde{\mathbf{f}} = \Sigma_{\mathrm{FITC}}^{-1/2}\mathbf{f}$ could be done by evaluating the full matrix $\Sigma_{\mathrm{FITC}}^{-1}$ and taking a matrix square root of it, but then the advantage of the sparse approximation would be lost. To extend the transformation for FITC in a way that avoids evaluating the full covariance matrix, the scaling is done only in the direction of the m largest eigenvalues of Σ_{FITC}.

To conduct the transformation we first write the inverse of $\mathbf{Q}_{\mathrm{f,f}} + \Lambda$ as in (6), denote

$$\widehat{\Lambda}^{-1} = \Sigma_l^{-1} + \Lambda^{-1} \tag{8}$$

$$\mathbf{L} = \Lambda^{-1}\mathbf{K}_{\mathrm{f,u}}\mathrm{chol}\left[\mathbf{K}_{\mathrm{u,u}} + \mathbf{K}_{\mathrm{u,f}}\Lambda^{-1}\mathbf{K}_{\mathrm{f,u}}\right]^{-1}, \tag{9}$$

and write the posterior precision as $\Sigma_{\mathrm{FITC}}^{-1} = \widehat{\Lambda}^{-1} - \mathbf{L}\mathbf{L}^{\mathrm{T}}$. Next, $\Sigma_{\mathrm{FITC}}^{-1}$ is scaled to make the diagonal elements of $\widehat{\Lambda}$ equal, that is $\widehat{\Lambda} = \lambda\mathbf{I}$. This is done by multiplying the posterior precision by $\widehat{\Lambda}^{1/2}$ from left and right, which corresponds to transforming the latent values into $\hat{\mathbf{f}} = \widehat{\Lambda}^{-1/2}\mathbf{f}$ with approximate posterior precision $\widehat{\Sigma}_{\mathrm{FITC}}^{-1} = \mathbf{I} - \widehat{\Lambda}^{1/2}\mathbf{L}\mathbf{L}^{\mathrm{T}}\widehat{\Lambda}^{1/2}$.

In the second step of transformation we want to find the eigenvectors corresponding to the m largest eigenvalues of $\widehat{\Sigma}_{\mathrm{FITC}}^{-1}$ and scale $\hat{\mathbf{f}}$ in their direction. To do this, let $\mathbf{D}^2$ be an $m \times m$ diagonal matrix of these m largest eigenvalues and $\mathbf{U}$ an $n \times m$ matrix with corresponding eigenvectors on its columns. The matrices satisfy the following relations (for example Harville, 1997, Section 21.10)

$$\mathbf{U}\mathbf{S}\mathbf{U}^{\mathrm{T}} = \widehat{\Lambda}^{1/2}\mathbf{L}\mathbf{L}^{\mathrm{T}}\widehat{\Lambda}^{1/2}$$
$$\mathbf{D}^2 = \mathrm{diag}\,[1 - \mathbf{S}_{11}, ..., 1 - \mathbf{S}_{mm}]\,.$$

The singular value decomposition $\mathbf{U}\mathbf{S}\mathbf{U}^{\mathrm{T}}$ can be found without explicitly forming the full $n \times n$ matrix by first defining a helper matrix $\mathbf{B} = \mathbf{U}\mathbf{S}^{1/2}\mathbf{V}^{\mathrm{T}}$ and finding the eigenvalue decomposition of an $m \times m$ matrix

$$\mathbf{B}^{\mathrm{T}}\mathbf{B} = \mathbf{V}\mathbf{S}\mathbf{V}^{\mathrm{T}}, \tag{10}$$

after which the matrix of eigenvectors $\mathbf{U}$ can be obtained from $\mathbf{U} = \mathbf{B}\mathbf{V}\mathbf{S}^{-1/2}$.

After solving $\mathbf{U}$, $\widehat{\Lambda}$ and $\mathbf{D}$ the transformation into a transformed space and back to the latent value space can be summarized as follows

$$\tilde{\mathbf{f}} = \left(1 + \mathbf{U}\mathbf{D}\mathbf{U}^{\mathrm{T}} - \mathbf{U}\mathbf{U}^{\mathrm{T}}\right) \widehat{\Lambda}^{-1/2}\, \mathbf{f} \tag{11}$$

$$\mathbf{f} = \widehat{\Lambda}^{1/2} \left(1 + \mathbf{U}\mathbf{D}^{-1}\mathbf{U}^{\mathrm{T}} - \mathbf{U}\mathbf{U}^{\mathrm{T}}\right) \tilde{\mathbf{f}}. \tag{12}$$

In order to retain the reversibility of MCMC sampling the transformation should not depend on the sampled parameter, and thus relative risk $\mu = \exp(\mathbf{f})$ is approximated with its prior mean of 1 when constructing the transformation matrices in (11) and (12). This should be a reasonably good approximation since μ's posterior variance is usually moderate in spatial epidemiology. The transformation is presented in algorithmic form in the appendix B.

5. Results

To test the model, we studied the spatial variations of two different diseases in Finland, the mortality due to *cerebral vascular diseases* and *alcohol-related diseases* in the time interval 1995-1999. We used two data sets of different sizes and in the case of smaller data set we compared the FITC approximation to full GP via 10-fold cross-validation.

5.1 Case data sets and models

The cerebral vascular diseases comprised roughly 18 000 deaths and the alcohol-related diseases about 5200 deaths. The data sets were aggregated in lattice resolutions of 20km × 20km and 10km × 10km resulting in 915 and 3193 data points respectively and models with four different covariance functions were tested. In the case of smaller data set the FITC approximation was compared to the full GP and the results are shown in the Section 5.3. The 10km × 10km lattice data were studied only with FITC approximation and the resulting maps are presented in the Section 5.2.

The inducing inputs in the FITC approximation were placed on a uniform grid as shown in the Figure 1 and in the case of 10km × 10km lattice data sets the number of them was 238. In the smaller data sets the performance of the approximation was tested with 30, 36, 46, 56, 74, 100, 149 and 221 inducing inputs (see Section 5.3). The covariance functions used in the models were, in addition to the squared exponential (1), an exponential, a Mátern $\nu = 3/2$ and a Mátern $\nu = 5/2$, given respectively as

$$k_{\exp}(\mathbf{x}_i, \mathbf{x}_j) = \sigma^2_{\exp} \exp\left(-r/l\right) \tag{13}$$

$$k_{\nu=3/2}(\mathbf{x}_i, \mathbf{x}_j) = \sigma^2_{\nu=3/2} \left(1 + \sqrt{3}r/l\right) \exp\left(-\sqrt{3}r/l\right) \tag{14}$$

$$k_{\nu=5/2}(\mathbf{x}_i, \mathbf{x}_j) = \sigma^2_{\nu=5/2} \left(1 + \sqrt{5}r/l + 5r^2/(3l^2)\right) \exp\left(-\sqrt{5}r/l\right). \tag{15}$$

The covariance functions are treated more extensively, for example, by Rasmussen and Williams (2006) and Abrahamsen (1997).

5.2 Examples of maps

The final products of the disease mapping analysis are the maps representing the spatial variations in the relative risk. We choose to present the posterior knowledge about relative risk with maps of the median of the relative risk and the probability of the relative risk being over 1, $p(\mu > 1|D)$.

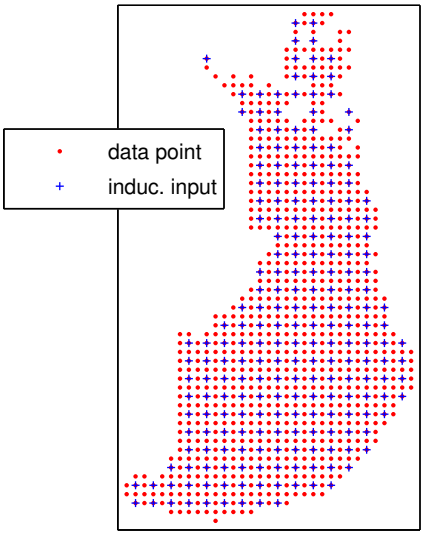


Figure 1: The 221 inducing inputs for lattice data with 20km × 20km grid cells.

The maps in the Figure 2 present the results for the cerebral vascular diseases obtained with the full GP and the FITC sparse approximation with a data aggregated into a 20km×20km lattice. In this case the models work equally well (see also Figure 4) and the maps also look similar. The resolution in these maps is the same as in the data, but it can also be increased by predicting the values of the posterior risk surfaces in a denser grid. A map created like this, however, is not more informative than a map in training resolution, but it may appear visually better due to the extra smoothing.

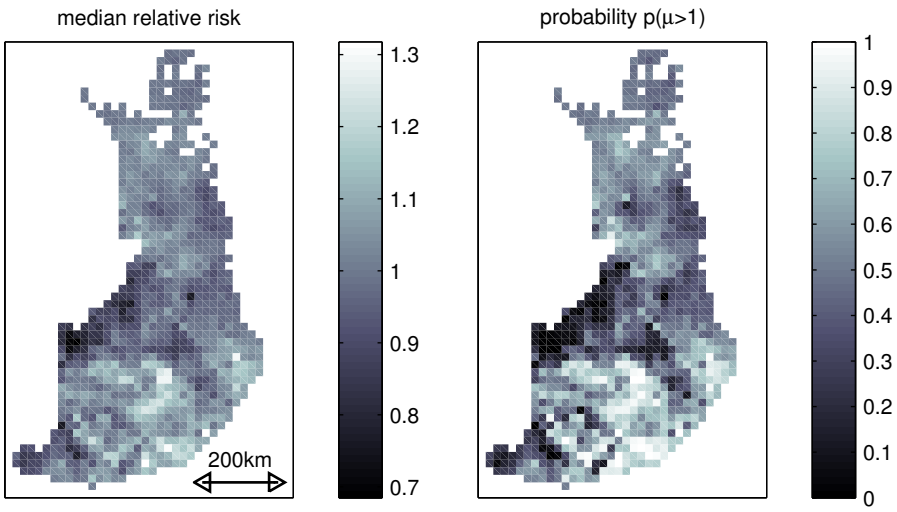
The results for alcohol related diseases are shown in 10km×10km resolution in the Figure 3. The other map is a result of the FITC approximation trained with data in a 10km×10km lattice and the other is a result of a full GP trained with data in 20km×20km grid cells and predicted into higher resolution. The overall structure of relative risk in both maps is similar, but the boundaries between different relative risk areas are sharper in the map of the FITC approximation. It also seems that in the eastern parts of Finland the map from the full GP smooths the results too much.

5.3 Model comparison

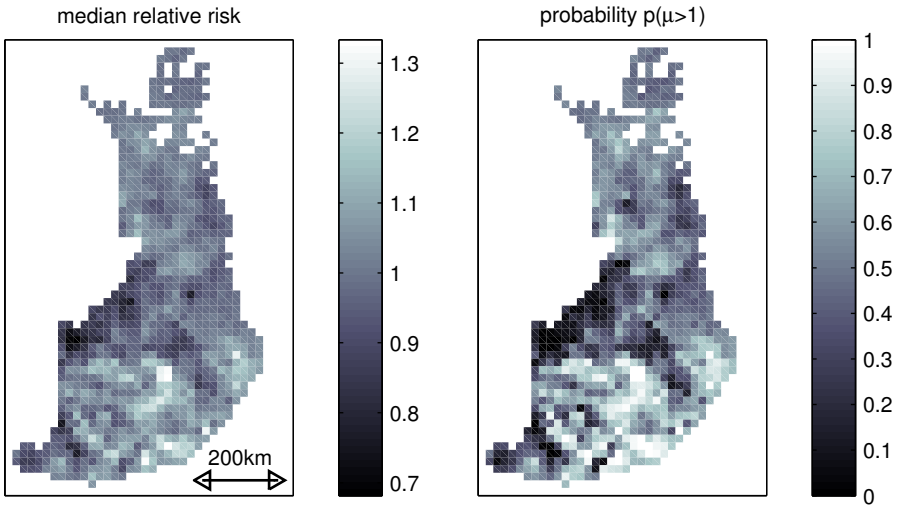
The model comparison is conducted by using 10-fold cross-validation with a bias correction (e.g. Vehtari and Lampinen, 2002). The data is divided into 10 groups so that $s(i)$ is the set of data points in group where the ith data point belongs. The comparison is done using the log predictive density diagnostics

$$\frac{1}{N}\sum_{i=1}^{N}\log p(Y_i^{\text{rep}} = Y_i \mid \mathbf{Y}^{\setminus s(i)}, \mathbf{X}^{\setminus s(i)}),$$

where Y_i^{rep} denotes the posterior predictive replicate of the number of deaths in the cell i given the data in the groups where data point Y_i does not belong, $\{\mathbf{Y}^{\setminus s(i)}, \mathbf{X}^{\setminus s(i)}\}$, and N is the number of data points. This is equivalent to conditional predictive ordinate diagnostics by Gelfand et al. (1992).

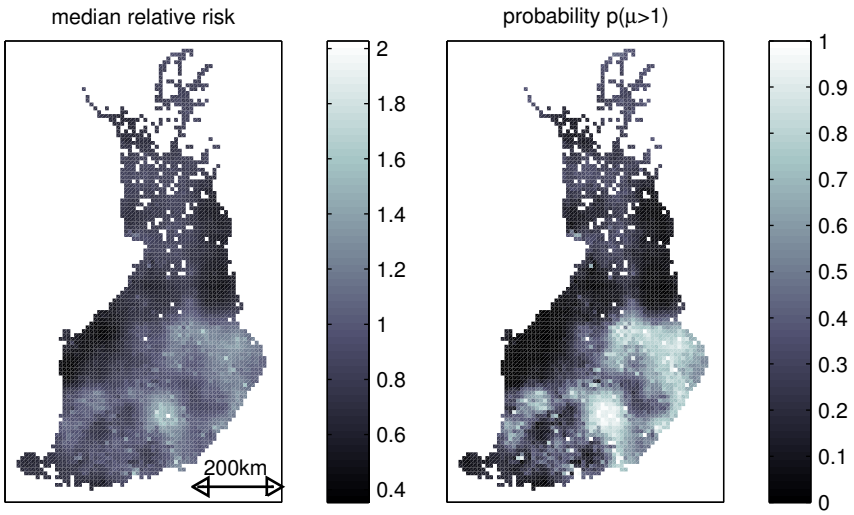


(a) FITC sparse approximation

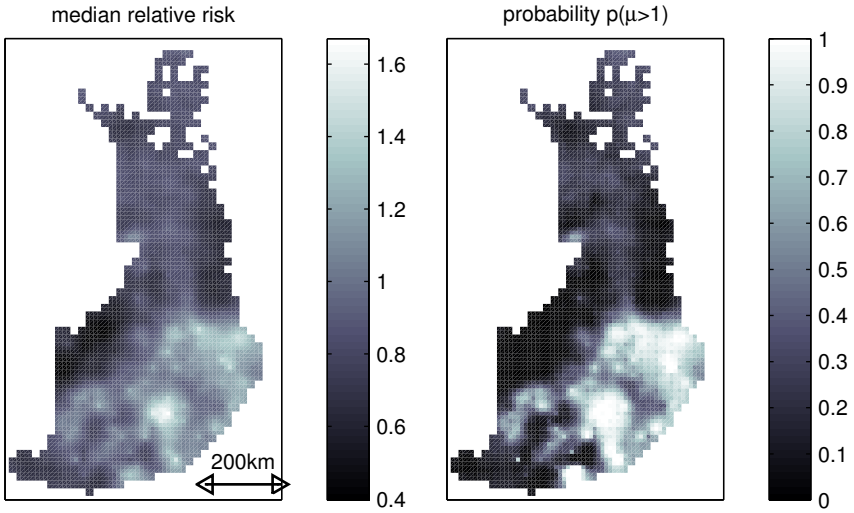


(b) Full Gaussian process

Figure 2: The relative risk of cerebral vascular diseases. The maps are results of models using the exponential covariance function trained with data aggregated in a 20km×20km lattice. In the FITC approximation there was 221 inducing inputs. The resolution in the maps is the same as in the training data. The posterior median and standard deviation of the length-scale of the covariance function were 33.0km and 9.8km in the FITC approximation and 27.0km and 7.9km in a full GP. In case of the FITC approximation and 10km×10km lattice data the median was 25.4km and the standard deviation 6.1km.



(a) FITC sparse approximation trained with 10km×10km lattice data



(b) Results of full GP trained with 20km×20km lattice data and predicted into a 10km×10km lattice

Figure 3: The relative risk of alcohol related diseases. The FITC approximation is trained with 10km×10km lattice data and the full GP is trained with 20km×20km lattice data. Both of the models are presented in a 10km×10km lattice and use the exponential covariance function. The posterior median and standard deviation of the length-scale of the covariance function were 66.2km and 24.7km in the FITC approximation and 83.5km and 43.2km in a full GP. In case of the FITC approximation trained with data in 20km×20km lattice data and 221 inducing inputs the median was 84.9km and the standard deviation 39.3km.

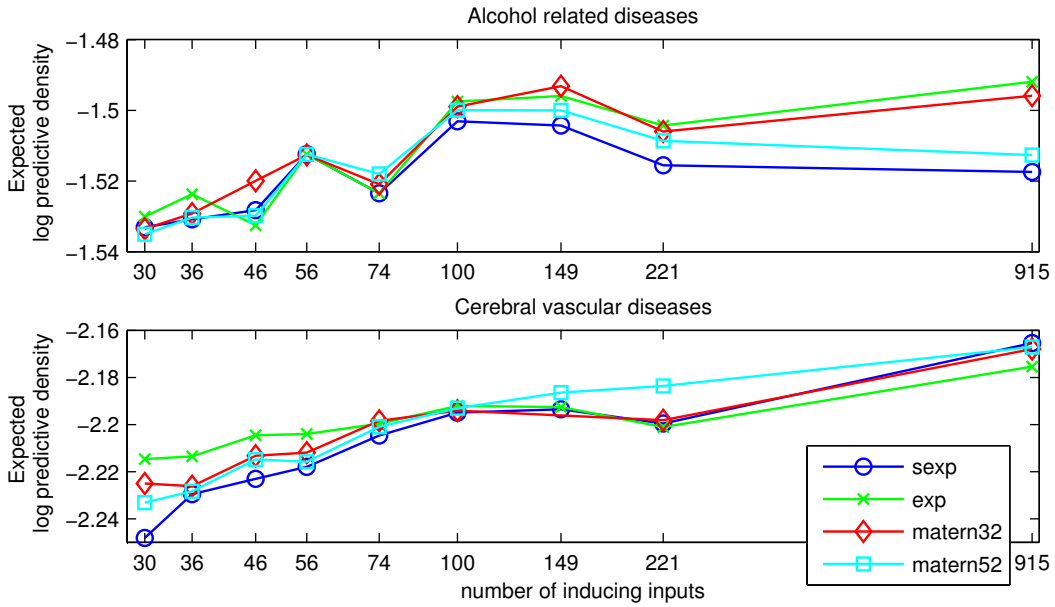


Figure 4: The log predictive diagnostics of models in 20km×20km lattice data. 915 points represents full GP models and other FITC approximations. Differences larger than 0.01 are estimated to be significant (see section 5.3 for details).

In the case of 20km×20km lattice data, we compared the models with FITC sparse approximation to the full GP models and the results are shown in the Figure 4. Pairwise model comparison was performed using Bayesian bootstrap as proposed by Vehtari and Lampinen (2002). Results of the comparison can briefly be summarized so that differences larger than approximately 0.01 are significant. Since the model with full GP was not constructed for the 10km×10km lattice data, the results of the FITC approximation in that case were compared to the full model only via the maps.

It can be concluded that in the case of cerebral vascular diseases the predictive performance of the FITC approximation decreases constantly as the number of inducing inputs is decreased. In the alcohol related diseases the predictive performance of the full GPs and the FITC approximations are practically as good with large number of inducing inputs. As the number of the inducing inputs is decreased the performance of the approximation starts decreasing also in this case. It was also noticed that as the number of inducing inputs was decreased below 100 the posterior values of the length-scale and magnitude increased in the FITC models. This can be seen as a change in the covariance function to one with a heavier tail, and thus the spatial inference of models with different number of inducing inputs might be different.

There is more overhead in the calculations with FITC than with the full GP, since the equations have more matrix multiplications. Due to the overhead, the use of FITC with a small number of data points is not reasonable, but the advantage of FITC increases as the data set increases and already with a ratio $n/m = 915/221$ the time saving was approximately 50%. With larger data sets the time saving is even more considerable and with normal office PC the use of full GP with MCMC becomes practically impossible when the number of data points gets close to 10 000. With both full and FITC GPs the efficiency of the posterior simulations was limited by the strong dependence

between latent values and hyperparameters, which caused slow mixing in the sampling of the joint posterior. Christensen et al. (2006) proposed an additional transformation to alleviate this problem, but it did not seem to be useful in our simulations.

The model comparison results are still preliminary and the approximation and model comparison techniques need still to be studied in more detail. The main focus of this work, however, was in the implementation of the FITC approximation in the disease mapping problem and the further study of the approximation and its affect on the spatial inference are left for the future.

6. Conclusions and future work

The aim of this work was to study the usability of the FITC sparse Gaussian process in disease mapping with high accuracy areally referenced healthcare data. The sparse Gaussian process was implemented for a Poisson likelihood and MCMC methods were used to conduct the posterior inference. For two test data cases the performance of the FITC approximation was similar to the full GP and significantly faster. The performance of the FITC approximation gradually decreased as the number of inducing inputs was reduced.

The sampling of the latent values was sped up with a transformation using their approximate posterior precision. The transformation worked well and enabled good mixing for the latent values. In the case of FITC the gradient evaluations needed in the hyperparameter sampling and in the latent value transformation were performed without explicitly forming the full covariance matrix.

The results obtained here were promising and thus encourage further study of the FITC approximation. As a future development we will study the practical limit of the number of regions which can be handled, sampling of the locations of inducing inputs, the performance of the models with fewer inducing inputs in more detail, various covariance functions, and accuracy of variational type approximations for marginalizing over the latent values. In the future the sparse GP model will also be tested in various other problems than disease mapping.

The work was focused on the methodology research and thus the significance of the results for the research of spatial epidemiology in Finland remains still for further study. This will be performed in collaboration with healthcare specialists.

Acknowledgments

Authors would like to thank Harri Valpola for helpful comments on latent variable transformation and Markus Siivola for data manipulation.

Appendix A. Gradients of the marginal likelihood in the case of FITC

The conditional distribution of hyperparameters is sampled with the hybrid Monte Carlo method, which needs gradients of the log marginal likelihood with respect to the covariance function parameters θ (see eq. (2)). These gradients are obtained from

$$\begin{aligned}\frac{\partial \log(p(\mathbf{f}\,|\theta))}{\partial \theta} =& \operatorname{tr}\left(\left(\mathbf{Q}_{\mathrm{f,f}}+\Lambda\right)^{-1}\frac{\partial(\mathbf{Q}_{\mathrm{f,f}}+\Lambda)}{\partial\theta}\right)\\ &-\frac{1}{2}\,\mathbf{f}^{\mathrm{T}}\left(\mathbf{Q}_{\mathrm{f,f}}+\Lambda\right)^{-1}\frac{\partial(\mathbf{Q}_{\mathrm{f,f}}+\Lambda)}{\partial\theta}\left(\mathbf{Q}_{\mathrm{f,f}}+\Lambda\right)^{-1}\mathbf{f}\,. \end{aligned} \tag{16}$$

which requires the expression of gradients of $\mathbf{Q}_{\mathrm{f,f}} = \mathbf{K}_{\mathrm{f,u}}\,\mathbf{K}_{u,u}^{-1}\,\mathbf{K}_{\mathrm{u,f}}$,

$$\frac{\partial\,\mathbf{Q}_{\mathrm{f,f}}}{\partial\theta} = \left(2\frac{\partial}{\partial\theta}\left[\mathbf{K}_{\mathrm{f,u}}\right]+\mathbf{K}_{\mathrm{f,u}}\,\mathbf{K}_{\mathrm{u,u}}^{-1}\frac{\partial}{\partial\theta}\left[\mathbf{K}_{\mathrm{u,u}}\right]\right)\left(\mathbf{K}_{\mathrm{f,u}}\,\mathbf{K}_{\mathrm{u,u}}^{-1}\right)^{\mathrm{T}}. \tag{17}$$

This is an $n\times n$ matrix and thus it is not evaluated explicitly. The gradient evaluation without explicit formation of any $n\times n$ matrix is shown below and the needed matrix algebra for the calculations are given, for example, by Harville (1997). To shorten the notation the first and second term in the right hand side of (16) are denoted by T and V respectively.

The gradient evaluation is begun with the term V. First, a vector $\mathbf{b} = \mathbf{f}^{\mathrm{T}}\left(\mathbf{Q}_{\mathrm{f,f}}+\Lambda\right)^{-1}$ is formed by using a matrix $\mathbf{L}$ from (9) and evaluating

$$\mathbf{b} = \mathbf{f}^{\mathrm{T}}\,\Lambda^{-1}+\left(\mathbf{f}^{\mathrm{T}}\,\mathbf{L}\right)\left(\mathbf{f}^{\mathrm{T}}\,\mathbf{L}\right)^{\mathrm{T}}, \tag{18}$$

where it should be noticed that Λ is diagonal. Now, by taking in the gradients of $\mathbf{Q}_{\mathrm{f,f}}$ from (17) the term V can be expressed as

$$\begin{aligned}\mathbf{b}\,\frac{\partial\,\mathbf{Q}_{\mathrm{f,f}}}{\partial\theta}\,\mathbf{b}^{\mathrm{T}}+\mathbf{b}\,\frac{\partial\,\Lambda}{\partial\theta}\,\mathbf{b}^{\mathrm{T}} =& \left(\mathbf{b}\,2\frac{\partial}{\partial\theta}\left[\mathbf{K}_{\mathrm{f,u}}\right]+\mathbf{b}\,\mathbf{K}_{\mathrm{f,u}}\,\mathbf{K}_{\mathrm{u,u}}^{-1}\frac{\partial}{\partial\theta}\left[\mathbf{K}_{\mathrm{u,u}}\right]\right)\left(\mathbf{K}_{\mathrm{f,u}}\,\mathbf{K}_{\mathrm{u,u}}^{-1}\right)^{\mathrm{T}}\mathbf{b}^{\mathrm{T}}\\ &+\mathbf{b}\,\frac{\partial}{\partial\theta}\left[\operatorname{diag}\left[\mathbf{K}_{\mathrm{f,f}}\right]\right]\mathbf{b}^{\mathrm{T}}-\mathbf{b}\,\frac{\partial}{\partial\theta}\left[\operatorname{diag}\left[\mathbf{Q}_{\mathrm{f,f}}\right]\right]\mathbf{b}^{\mathrm{T}}, \end{aligned} \tag{19}$$

where the first term can be evaluated without forming an $n\times n$ matrix if the calculations are conducted in the right order. The second term is also easy because of the diagonal matrix. In order to proceed with the third term a diagonal matrix $\mathbf{B} = \operatorname{diag}\left[b_1^2, b_2^2, ..., b_n^2\right]$ is defined so that its diagonal elements are the elements of $\mathbf{b}$ squared, after which the third term can be modified into

$$\begin{aligned}\mathbf{b}\,\frac{\partial(\operatorname{diag}\left[\mathbf{Q}_{\mathrm{f,f}}\right])}{\partial\theta}\,\mathbf{b}^{\mathrm{T}} =& \operatorname{tr}\left(\mathbf{b}\,\frac{\partial(\operatorname{diag}\left[\mathbf{Q}_{\mathrm{f,f}}\right])}{\partial\theta}\,\mathbf{b}^{\mathrm{T}}\right)\\ =& \operatorname{tr}\left(\mathbf{B}\frac{\partial(\operatorname{diag}\left[\mathbf{Q}_{\mathrm{f,f}}\right])}{\partial\theta}\right)\\ =& \operatorname{tr}\left(\mathbf{B}\frac{\partial(\mathbf{Q}_{\mathrm{f,f}})}{\partial\theta}\right), \end{aligned} \tag{20}$$

Now, by taking in the $\frac{\partial\,\mathbf{Q}_{\mathrm{f,f}}}{\partial\theta}$ and using the fact that $\operatorname{tr}(\mathbf{AC}) = \operatorname{tr}(\mathbf{CA})$, where $\mathbf{C}$ is an $m\times n$ matrix and $\mathbf{A}$ an $n\times m$ matrix, this can be modified further as follows

$$\operatorname{tr}\left(\mathbf{B}\frac{\partial(\mathbf{Q}_{\mathrm{f,f}})}{\partial\theta}\right) = 2\operatorname{tr}\left(\left(\mathbf{K}_{\mathrm{f,u}}\,\mathbf{K}_{\mathrm{u,u}}^{-1}\right)^{\mathrm{T}}\mathbf{B}\,\frac{\partial}{\partial\theta}\left[\mathbf{K}_{\mathrm{f,u}}\right]\right)-\operatorname{tr}\left(\left(\mathbf{K}_{\mathrm{f,u}}\,\mathbf{K}_{\mathrm{u,u}}^{-1}\right)^{\mathrm{T}}\mathbf{B}\,\mathbf{K}_{\mathrm{f,u}}\,\mathbf{K}_{\mathrm{u,u}}^{-1}\frac{\partial}{\partial\theta}\left[\mathbf{K}_{\mathrm{u,u}}\right]\right). \tag{21}$$

Above, the expressions inside the trace operator form an $n \times n$ matrix if the matrix multiplications are conducted and the trace is taken after that. However, this can be avoided by noticing that the trace of a matrix product between an $n \times m$ matrix $\mathbf{A}$ and an $m \times n$ matrix $\mathbf{C}$ can be written as $\mathrm{tr}(\mathbf{AC}) = \Sigma_{i=1}^{n}\Sigma_{j=1}^{m} a_{ij}c_{ji}$ which is actually a dot product of vectors $\mathbf{a} = [a_{11}, a_{12}, ..., a_{1n}, a_{21}, ..., a_{2n}, ..., a_{mn}]$ and $\mathbf{c} = [c_{11}, c_{21}, ..., c_{n1}, c_{12}, ..., c_{n2}, ..., c_{nm}]$. The evaluation of the traces in (21) can thus be handled with a dot product of two $1 \times nm$ vectors. Furthermore, by writing the term $\left(\mathbf{K}_{\mathrm{f,u}}\,\mathbf{K}_{\mathrm{u,u}}^{-1}\right)^{\mathrm{T}}\mathbf{b}^{\mathrm{T}}$ in (19) as $\left(\mathbf{b}\,\mathbf{K}_{\mathrm{f,u}}\,\mathbf{K}_{\mathrm{u,u}}^{-1}\right)^{\mathrm{T}}$, the term V is obtained from

$$
\begin{aligned}
V = &\left[2\,\mathbf{b}\,\frac{\partial}{\partial\theta}\left[\mathbf{K}_{\mathrm{f,u}}\right] + \mathbf{b}\,\mathbf{K}_{\mathrm{f,u}}\,\mathbf{K}_{\mathrm{u,u}}^{-1}\frac{\partial}{\partial\theta}\left[\mathbf{K}_{\mathrm{u,u}}\right]\right]\left(\mathbf{b}\,\mathbf{K}_{\mathrm{f,u}}\,\mathbf{K}_{\mathrm{u,u}}^{-1}\right)^{\mathrm{T}} + \mathbf{b}\,\frac{\partial(\mathrm{diag}\left[\mathbf{K}_{\mathrm{f,f}}\right])}{\partial\theta}\,\mathbf{b}^{\mathrm{T}} \\
&- 2\mathrm{tr}\left(\left(\mathbf{K}_{\mathrm{f,u}}\,\mathbf{K}_{\mathrm{u,u}}^{-1}\right)^{\mathrm{T}}\mathbf{B}\,\frac{\partial}{\partial\theta}\left[\mathbf{K}_{\mathrm{f,u}}\right]\right) + \mathrm{tr}\left(\left(\mathbf{K}_{\mathrm{f,u}}\,\mathbf{K}_{\mathrm{u,u}}^{-1}\right)^{\mathrm{T}}\mathbf{B}\,\mathbf{K}_{\mathrm{f,u}}\,\mathbf{K}_{\mathrm{u,u}}^{-1}\frac{\partial}{\partial\theta}\left[\mathbf{K}_{\mathrm{u,u}}\right]\right),
\end{aligned}
$$

which can be evaluated without forming any $n \times n$ matrices and enables the use of intermediate results in several places.

The evaluation of the term T is begun by partitioning it as following

$$
\begin{aligned}
T = &\mathrm{tr}\left(\left(\mathbf{Q}_{\mathrm{f,f}} + \Lambda\right)^{-1}\frac{\partial\,\mathbf{Q}_{\mathrm{f,f}}}{\partial\theta}\right) + \mathrm{tr}\left(\left(\mathbf{Q}_{\mathrm{f,f}} + \Lambda\right)^{-1}\frac{\partial}{\partial\theta}\mathrm{diag}\left[\mathbf{K}_{\mathrm{f,f}}\right]\right) - \\
&\mathrm{tr}\left(\left(\mathbf{Q}_{\mathrm{f,f}} + \Lambda\right)^{-1}\frac{\partial}{\partial\theta}\mathrm{diag}\left[\mathbf{Q}_{\mathrm{f,f}}\right]\right).
\end{aligned}
$$

where the first term can be evaluated using the matrix inversion lemma for $\left(\mathbf{Q}_{\mathrm{f,f}} + \Lambda\right)^{-1}$. The second term can be evaluated by first solving $\mathrm{diag}\left[\left(\mathbf{Q}_{\mathrm{f,f}} + \Lambda\right)^{-1}\right]$, which can be done efficiently using $\mathbf{L}$ from (9), and then using the fact that $\mathrm{tr}\left(\mathbf{A}\mathrm{diag}\left[\mathbf{C}\right]\right) = \mathrm{tr}\left(\mathrm{diag}\left[\mathbf{A}\right]\mathrm{diag}\left[\mathbf{C}\right]\right)$. Using the same idea as in (20) the last term can be changed into $\mathrm{tr}\left(\mathrm{diag}\left[\left(\mathbf{Q}_{\mathrm{f,f}} + \Lambda\right)^{-1}\right]\frac{\partial}{\partial\theta}\left[\mathbf{Q}_{\mathrm{f,f}}\right]\right)$. By plugging in the derivative of $\mathbf{Q}_{\mathrm{f,f}}$ from Equation (17) and using the fact that $\mathrm{tr}(\mathbf{AC}) = \mathrm{tr}(\mathbf{CA})$ as above, the expression can be modified into

$$
\begin{aligned}
T = &2\mathrm{tr}\left(\left(\mathbf{K}_{\mathrm{f,u}}\,\mathbf{K}_{\mathrm{u,u}}^{-1}\right)^{\mathrm{T}}\left(\mathbf{Q}_{\mathrm{f,f}} + \Lambda\right)^{-1}\frac{\partial}{\partial\theta}\left[\mathbf{K}_{\mathrm{f,u}}\right]\right) + \\
&\mathrm{tr}\left(\left(\mathbf{K}_{\mathrm{f,u}}\,\mathbf{K}_{\mathrm{u,u}}^{-1}\right)^{\mathrm{T}}\left(\mathbf{Q}_{\mathrm{f,f}} + \Lambda\right)^{-1}\mathbf{K}_{\mathrm{f,u}}\,\mathbf{K}_{\mathrm{u,u}}^{-1}\frac{\partial}{\partial\theta}\left[\mathbf{K}_{\mathrm{u,u}}\right]\right) + \\
&\mathrm{tr}\left(\mathrm{diag}\left[\left(\mathbf{Q}_{\mathrm{f,f}} + \Lambda\right)^{-1}\right]\frac{\partial}{\partial\theta}\left[\mathrm{diag}\left[\mathbf{K}_{\mathrm{f,f}}\right]\right]\right) - \\
&2\mathrm{tr}\left(\mathrm{diag}\left[\left(\mathbf{Q}_{\mathrm{f,f}} + \Lambda\right)^{-1}\right]\frac{\partial}{\partial\theta}\left[\mathbf{K}_{\mathrm{f,u}}\right]\left(\mathbf{K}_{\mathrm{f,u}}\,\mathbf{K}_{\mathrm{u,u}}^{-1}\right)^{\mathrm{T}}\right) + \\
&\mathrm{tr}\left(\mathrm{diag}\left[\left(\mathbf{Q}_{\mathrm{f,f}} + \Lambda\right)^{-1}\right]\mathbf{K}_{\mathrm{f,u}}\,\mathbf{K}_{\mathrm{u,u}}^{-1}\frac{\partial}{\partial\theta}\left[\mathbf{K}_{\mathrm{u,u}}\right]\left(\mathbf{K}_{\mathrm{f,u}}\,\mathbf{K}_{\mathrm{u,u}}^{-1}\right)^{\mathrm{T}}\right).
\end{aligned}
$$

Earlier it was mentioned that the evaluation of trace can be changed to a dot product of two vectors formed of the matrices. Thus by conducting the operations above in a right order, the calculation of T can be conducted without forming any $n \times n$ matrix. A pseudo code for the gradient evaluation is shown in the algorithm 1.

Algorithm 1 Calculate the gradients of minus log likelihood. Note: Here the notation $\mathbf{C}(:)$ represents a vector $[c_{11}, c_{21}, ..., c_{n1}, c_{12}, ..., c_{n2}, ..., c_{nn}]^{\mathrm{T}}$. Note: Some of the notations are from Matlab, they are ./ (elementwise division), .* (elementwise multiplication), and some of the notations that are matrices in the text are vectors here

Input: $\mathbf{K}_{\mathrm{f,u}}$, $\mathbf{K}_{\mathrm{u,u}}$, $\frac{\partial}{\partial\theta}\left[\mathbf{K}_{\mathrm{f,u}}\right]$, $\frac{\partial}{\partial\theta}\left[\mathbf{K}_{\mathrm{u,u}}\right]$ $\mathbf{f}$, $\mathbf{k} = \left[\mathbf{K}_{\mathrm{f,f}}(1,1), ..., \mathbf{K}_{\mathrm{f,f}}(n,n)\right]$

% First evaluate helper matrices
$\mathbf{b} \leftarrow \mathbf{f}^{\mathrm{T}}\left(\mathbf{Q}_{\mathrm{f,f}} + \Lambda\right)^{-1}$ (evaluate as in (18))
$\mathbf{A} \leftarrow \mathbf{K}_{\mathrm{f,u}}\,\mathbf{K}_{\mathrm{u,u}}^{-1}$
$\mathbf{F} \leftarrow \mathbf{A}^{\mathrm{T}}\mathbf{B}$
$\mathbf{G} \leftarrow \mathbf{bA}$
$\mathbf{M} \leftarrow \mathbf{A}^{\mathrm{T}}\left(\mathbf{Q}_{\mathrm{f,f}} + \Lambda\right)^{-1}$ (evaluate using matrix inversion lemma)
$\mathbf{q} \leftarrow \mathrm{diag}\left[\left(\mathbf{Q}_{\mathrm{f,f}} + \Lambda\right)^{-1}\right]$ ($1 \times n$ vector of diagonal elements)
$\mathbf{P} \leftarrow \mathbf{A}\frac{\partial}{\partial\theta}\left[\mathbf{K}_{\mathrm{u,u}}\right]$
$\mathbf{R} \leftarrow 2\mathrm{diag}\left[\left(\mathbf{Q}_{\mathrm{f,f}} + \Lambda\right)^{-1}\right]\frac{\partial}{\partial\theta}\left[\mathbf{K}_{\mathrm{f,u}}\right]$ (use vector $\mathbf{q}$)
$\mathbf{W} \leftarrow \mathrm{diag}\left[\left(\mathbf{Q}_{\mathrm{f,f}} + \Lambda\right)^{-1}\right]\mathbf{P}$

% Then evaluate the gradient
$V \leftarrow 2 * \frac{\partial}{\partial\theta}\left[\mathbf{K}_{\mathrm{f,u}}\right] + \mathbf{G} * \frac{\partial}{\partial\theta}\left[\mathbf{K}_{\mathrm{u,u}}\right]$
$V \leftarrow V * \mathbf{G}^{\mathrm{T}} + (\mathbf{b}.*\mathbf{k}) * \mathbf{b}^{\mathrm{T}}$
$V \leftarrow V + 2 * \left(\mathbf{F}^{\mathrm{T}}(:)\right)^{\mathrm{T}} * \frac{\partial}{\partial\theta}\left[\mathbf{K}_{\mathrm{f,u}}(:)\right] + \left(\mathbf{F}^{\mathrm{T}}(:)\right)^{\mathrm{T}} * \mathbf{P}(:)$
$T \leftarrow 2 * (\mathbf{M}^{\mathrm{T}}(:))^{\mathrm{T}} * \frac{\partial}{\partial\theta}\left[\mathbf{K}_{\mathrm{f,u}}(:)\right] + (\mathbf{M}^{\mathrm{T}}(:))^{\mathrm{T}} * \mathbf{P}(:) + q * \mathbf{k}^{\mathrm{T}}$
$T \leftarrow T + \mathbf{R}^{\mathrm{T}}(:))^{\mathrm{T}} * \mathbf{A}^{\mathrm{T}}(:) + \mathbf{W}^{\mathrm{T}}(:))^{\mathrm{T}} * \mathbf{A}^{\mathrm{T}}(:)$
return $T + V$

Appendix B. Algorithm for latent value transformation

Algorithm 2 Transformation and re-transformation of latent values with their approximate posterior covariance.

Input: $\mathbf{f}$, $\mathbf{E}$, $\mathbf{K}_{\mathrm{f,u}}$, $\mathbf{K}_{\mathrm{u,u}}$, $\mathbf{k} = \left[\left[\mathbf{K}_{\mathrm{f,f}}\right]_{11}, ..., \left[\mathbf{K}_{\mathrm{f,f}}\right]_{nn}\right]$

if transform from $\mathbf{f}$ to $\tilde{\mathbf{f}}$ **then**
  $\mathbf{q} \leftarrow$ diagonals of $\mathbf{Q}_{\mathrm{f,f}}$ (evaluated efficiently from chol($\mathbf{K}_{\mathrm{u,u}}$) \ $\mathbf{K}_{\mathrm{f,u}}{}^{\mathrm{T}}$)
  $\Lambda \leftarrow k - \mathbf{q}$ (vector diag[Λ] of length n)
  $\widehat{\Lambda}^{-1} \leftarrow \mathbf{E}$+1./$\Lambda$; (vector diag$\left[\widehat{\Lambda}\right]$, eq. (8)
  $\mu = \exp(\mathbf{f}) = 1$)
  $\mathbf{K} \leftarrow \Lambda^{-1}\,\mathbf{K}_{\mathrm{f,u}}$ (note that Λ is diagonal)
  $\mathbf{L} \leftarrow \mathbf{K}\left(\left(\mathrm{chol}\left[\mathbf{K}_{\mathrm{u,u}} + \mathbf{K}_{\mathrm{f,u}}\,\mathbf{K}\right]\right)^{-1}\right)^{\mathrm{T}}$ (This is faster and numerically more stable than $\mathbf{K}\mathrm{chol}\left[\left(\mathbf{K}_{\mathrm{u,u}} + \mathbf{K}_{\mathrm{f,u}}\,\mathbf{K}\right)^{-1}\right]$)
  $\mathbf{B} \leftarrow \mathbf{L}*\widehat{\Lambda}^{1/2}$
  $\mathbf{S} \leftarrow$ eigenvalues of $\mathbf{B}^{\mathrm{T}}\mathbf{B}$ (a vector of length m eq. (10))
  $\mathbf{V} \leftarrow$ eigenvectors of $\mathbf{B}^{\mathrm{T}}\mathbf{B}$ ($m \times m$ matrix eq. (10))
  $\mathbf{U} \leftarrow \mathbf{B}\mathbf{V}/\mathbf{S}^{1/2}$
  $\mathbf{D} \leftarrow (1 - \mathbf{S})^{1/2}$ (this is a vector and thus the square root can be evaluated pointwise)
  save $\mathbf{D}$, $\mathbf{U}$ and $\widehat{\Lambda}$ (for use in re-transformation)
  $\hat{\mathbf{f}} \leftarrow \widehat{\Lambda}^{-1/2}\,\mathbf{f}$
  $\tilde{\mathbf{f}} \leftarrow \hat{\mathbf{f}} + \mathbf{U}\left[\left(\mathbf{D}\mathbf{U}^{\mathrm{T}} - \mathbf{U}^{\mathrm{T}}\right)\hat{\mathbf{f}}\right]$
  return $\tilde{\mathbf{f}}$
end if

if transform from $\tilde{\mathbf{f}}$ to $\mathbf{f}$ **then**
  load $\mathbf{D}$, $\mathbf{U}$ and $\widehat{\Lambda}$
  $\mathbf{f} \leftarrow \widehat{\Lambda}^{1/2}\left[\tilde{\mathbf{f}} + \mathbf{U}\left((\mathbf{D}^{-1}\mathbf{U}^{\mathrm{T}} - \mathbf{U}^{\mathrm{T}})\tilde{\mathbf{f}}\right)\right]$
  return $\mathbf{f}$
end if

References

Petter Abrahamsen. A review of Gaussian random fields and correlation functions, second edition. Technical Report 917, Norwegian Computing Center, April 1997.

Omar B. Ahmad, Cynthia Boschi-Pinto, Alan D. Lopez, Christopher J.L. Murray, Rafael Lozano, and Mie Inoue. Age standardization of rates: A new WHO standard. *GPE Discussion Paper Series*, 31, 2000.

Sudipto Banerjee, Bradley P. Carlin, and Alan E. Gelfand. *Hierarchical Modelling and Analysis for Spatial Data*. Chapman Hall/CRC, 2004.

Nicky Best, Sylvia Richardson, and Andrew Thomson. A comparison of Bayesian spatial models for disease mapping. *Statistical Methods in Medical Research*, 14:35–59, 2005.

Ole F. Christensen, Gareth O. Roberts, and Martin Sköld. Robust Markov chain Monte Carlo methods for spatial generalised linear mixed models. *Journal of Computational and Graphical Statistics*, 15:1–17, 2006.

Simon Duane, A.D. Kennedy, Brian J. Pendleton, and Duncan Roweth. Hybrid Monte Carlo. *Physics Letters B*, 195(2):216–222, September 1987.

P. Elliot, Jon Wakefield, Nicola Best, and David Briggs, editors. *Spatial Epidemiology Methods and Applications*. Oxford University Press, 2001.

Alan E. Gelfand, D. K. Dey, and H. Chang. Model determination using predictive distributions with implementation via sampling-based methods (with discussion). In J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, editors, *Bayesian Statistics 4*, pages 147–167. Oxford University Press, 1992.

Andrew Gelman. Prior distribution for variance parameters in hierarchical models. *Bayesian Analysis*, 1(3):515–533, 2006.

David A. Harville. *Matrix Algebra From a Statistician's Perspective*. Springer-Verlag, 1997.

Andrew B. Lawson. *Statistical Methods in Spatial Epidemology*. John Wiley & Sons, Ltd, 2001.

Radford M. Neal. *Bayesian Learning for Neural Networks*. Springer, 1996.

Joaquin Quiñonero-Candela and Carl Edward Rasmussen. A unifying view of sparse approximate Gaussian process regression. *Journal Of Machine Learning Research*, 6(3):1939–1959, December 2005.

Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 2006.

Sylvia Richardson. Spatial models in epidemiological applications. In Peter J. Green, Nils Lid Hjort, and Sylvia Richardson, editors, *Highly Structured Stochastic Systems*, pages 237–259. Oxford University Press, 2003.

Edward Snelson and Zouhin Ghahramani. Sparse Gaussian process using pseudo-inputs. In Y. Weiss, B. Schölkopf, and J. Platt, editors, *Advances in Neural Information Processing Systems*, volume 18. The MIT Press, 2006.

Aki Vehtari and Jouko Lampinen. Bayesian model assessment and comparison using cross-validation predictive densities. *Neural Computation*, 14(10):2439–2468, 2002.