Approximate inference for spatial functional data on massively parallel processors

Lars Lau Rakêt[1] and Bo Markussen[2]
[1]Department of Computer Science,
[2]Department of Mathematical Sciences
University of Copenhagen, Denmark

December 10, 2013

Abstract

With continually increasing data sizes, the relevance of the big n problem of classical likelihood approaches is greater than ever. The functional mixed-effects model is a well established class of models for analyzing functional data. Spatial functional data in a mixed-effects setting is considered, and so-called operator approximations for doing inference in the resulting models are presented. These approximations embed observations in function space, transferring likelihood calculations to the functional domain. The resulting approximated problems are naturally parallel and can be solved in linear time. An extremely efficient GPU implementation is presented, and the proposed methods are illustrated by conducting a classical statistical analysis of 2D chromatography data consisting of more than 140 million spatially correlated observation points.[1]

1 Introduction

During the last half century, functional data analysis has developed into a well-established subdiscipline of statistics (Ramsay & Silverman 2005, Ferraty & Vieu 2006, Horváth & Kokoszka 2012). The continuous sophistication of instruments gives rise to an increasing number of problems where functional aspects have to be taken into account. Statistical analysis of functional data generally involves the ill-posed problem of inferring an infinite-dimensional function from discrete data points. This requires some sort of regularization, and the type of regularization is often chosen in terms of roughness penalties that lead to sparse representations of the inferred function in terms of simple basis functions (Wahba 1990), thus reducing the computational complexity. The most typical specification, however, considers the inverse regularization process where

[1]Code for analyzing spatial functional data on graphics processing units is available as supplementary material.

a sparse basis is chosen explicitly for the given problem, which may then be further regularized through a roughness penalty (Ramsay & Silverman 2005).

This paper takes a different path for model specification; we consider functional mixed-effects models with random effects generated by Gaussian processes, and present a framework that moves the calculations needed in such analyses from the discrete domain induced by the observations to the underlying functional domain. As a consequence it is possible to efficiently compute the functions in question, even if the regularization does not lead to sparse representations. The methods are based on the one-dimensional operator approximations of Markussen (2013), and here new results and resolution strategies are presented for high-dimensional domains.

The functional viewpoint sheds new light on some of the current challenges in statistics (Jordan 2011), by both reducing the computational complexity of a large class of statistical problems dramatically, and at the same time revealing a natural link between partial differential equations and a large number of statistical models, including functional mixed-effects models, penalized likelihood, and Bayesian models.

In addition to reducing the computational complexity, the proposed resolution strategies are highly parallel, and naturally suited for implementation on massively parallel processors like graphics processing units (GPUs). While parallelization and GPUs have received some attention in the statistical community in recent years, the main focus has been on parallelizing matrix operations and sampling techniques (Suchard et al. 2010, da Silva 2010). To our knowledge, this work marks the first attempt of actively formulating solutions for classical statistical problems in a way that is particularly beneficial for implementation on massively parallel hardware.

The proposed methods are illustrated by conducting a classical statistical analysis of a dataset of 2D chromatograms with more than 140 million spatially correlated observations on a GPU.

2 Model and estimation

We consider spatial functional data on a domain $\mathscr{T} \subseteq \mathbb{R}^d$. Suppose we are given k noisy vectorized functional samples $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_k$ each consisting of n observation points. We assume that the observations are generated from the following functional mixed-effect model

$$y_i(\boldsymbol{t}) = \theta_{e(i)}(\boldsymbol{t}) + x_i(\boldsymbol{t}) + \varepsilon_i(\boldsymbol{t}) \tag{1}$$

where $e : \{1, \ldots, k\} \to \{1, \ldots, p\}$ is a factor, $\theta_{e(i)}$ is the fixed functional mean for group $e(i)$, x_i is a zero-mean Gaussian process with covariance function $\tau^2\mathcal{G}$, and ε_i is a Gaussian white noise process with variance σ^2.

A wide variety of functional mixed-effects models have previously been considered. One of the dominant approaches is to model functional effects using smoothing splines (Wahba 1990). Such constructions are considered by Wang (1998) and Guo (2002). Modeling of mixed effects in terms of penalized splines

is considered by Chen & Wang (2011), and Lee et al. (2013) propose a related method based on nested basis functions for spatial mixed-effects models. An alternative approach to functional mixed-effect models considers the problem in a nonparametric setting, where no distributional or parametric assumptions are made on the random effects. Boularan et al. (1994) considered modeling of growth curves, assuming only that population and individual effects were twice differentiable, and proposed kernel smoothing estimates for the effects. On a similar note, Núñez-Antón et al. (1999) considered a nonparametric three-level model and applied it to speech recognition data. For the use of nonparametric statistical modeling techniques for functional data we refer to the monograph by Ferraty & Vieu (2006), and for a review on functional mixed-effects models we refer to Liu & Guo (2012).

Now, let $\boldsymbol{y}$ be the concatenation of all the vectorized observations of length $N = kn$. The discrete observation $\boldsymbol{y}$ generated by function evaluation at the points $\boldsymbol{t}_1, \ldots, \boldsymbol{t}_n$ in the model (1) may be modeled by a conventional linear mixed-effects model

$$\boldsymbol{y} = \boldsymbol{\Gamma}\boldsymbol{\theta} + \boldsymbol{x} + \boldsymbol{\varepsilon}, \tag{2}$$

where $\boldsymbol{\Gamma} = \mathbb{I}_n \otimes \boldsymbol{\Gamma}_0$ is the design matrix corresponding to the factor e and $\boldsymbol{\theta} \in \mathbb{R}^{np}$ is a vector of parameters describing the group mean functions pointwise, $\boldsymbol{x}$ consists of the spatially correlated effects, $\boldsymbol{x} \sim \mathcal{N}(0, \mathbb{I}_k \otimes \tau^2\boldsymbol{\Sigma})$ with covariance matrix $\boldsymbol{\Sigma} = \big\{\mathcal{G}(\boldsymbol{t}_i, \boldsymbol{t}_j)\big\}_{i,j}$, and $\boldsymbol{\varepsilon}$ is independent, identically distributed Gaussian noise $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2\mathbb{I}_N)$. Since the design is constant across all observations, i.e. given by $\boldsymbol{\Gamma}_0$, the fixed effect $\boldsymbol{\theta}$ can be estimated pointwise. The solution strategy presented below may also be adapted to the situation with a low rank design matrix following Markussen (2013).

Functional mixed-effect models are typically modeled with fixed effects of a functional nature. For simplicity, we parametrize the fixed effect with one parameter per observation point, mimicking classical mixed-effects models. The adaption to functional fixed effects given by a limited number of basis functions can be done following the previously mentioned references. In particular, the computations needed for fixed effects parametrized in terms of smoothing splines closely follow the computations related to the spatially correlated effect $\boldsymbol{x}$, and the presented methods naturally extend to such parametrizations.

The best linear unbiased prediction for the spatially correlated effects in the model (2) is done by means of the conditional expectation (Robinson 1991)

$$\mathrm{E}\,[\,\boldsymbol{x}\,|\,\boldsymbol{y}\,] = \big(\mathbb{I}_k \otimes \tau^2\boldsymbol{\Sigma}\big)\,\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{\Gamma}\hat{\boldsymbol{\theta}}), \tag{3}$$

where $\boldsymbol{V} = \sigma^2\mathbb{I}_N + \mathbb{I}_k \otimes \tau^2\boldsymbol{\Sigma}$. The variance parameters are typically estimated by minimizing the negative log restricted likelihood (Harville 1977, Lee et al. 2006)

$$\ell_{\boldsymbol{y}}(\sigma, \tau) = \log\det\boldsymbol{V} + \log\det[\boldsymbol{\Gamma}^\top\boldsymbol{V}^{-1}\boldsymbol{\Gamma}] + (\boldsymbol{y} - \boldsymbol{\Gamma}\hat{\boldsymbol{\theta}})^\top\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{\Gamma}\hat{\boldsymbol{\theta}}). \tag{4}$$

For later use it is noted that the last term in the likelihood function can be written as

$$\begin{aligned}(\boldsymbol{y}-\boldsymbol{\Gamma}\hat{\boldsymbol{\theta}})^\top \boldsymbol{V}^{-1}(\boldsymbol{y}-\boldsymbol{\Gamma}\hat{\boldsymbol{\theta}}) &= \frac{1}{\sigma^2}(\boldsymbol{y}-\boldsymbol{\Gamma}\hat{\boldsymbol{\theta}})^\top(\boldsymbol{y}-\boldsymbol{\Gamma}\hat{\boldsymbol{\theta}}-\mathrm{E}\,[\,\boldsymbol{x}\,|\,\boldsymbol{y}\,]) \\ &= \frac{1}{\sigma^2}(\boldsymbol{y}-\boldsymbol{\Gamma}\hat{\boldsymbol{\theta}}-\mathrm{E}\,[\,\boldsymbol{x}\,|\,\boldsymbol{y}\,])^\top(\boldsymbol{y}-\boldsymbol{\Gamma}\hat{\boldsymbol{\theta}}-\mathrm{E}\,[\,\boldsymbol{x}\,|\,\boldsymbol{y}\,]) \\ &\quad + \frac{1}{\sigma^2}\mathrm{E}\,[\,\boldsymbol{x}\,|\,\boldsymbol{y}\,]^\top(\boldsymbol{y}-\boldsymbol{\Gamma}\hat{\boldsymbol{\theta}}-\mathrm{E}\,[\,\boldsymbol{x}\,|\,\boldsymbol{y}\,]).\end{aligned} \tag{5}$$

3 Operator approximations

For many common covariances $\mathcal{G}$ the underlying functional structure of the covariance matrix $\boldsymbol{\Sigma}$ can be exploited, so that one may approximate calculations involving $\boldsymbol{\Sigma}$. The functional counterpart to $\boldsymbol{\Sigma}$ is the integral operator $\mathscr{G}$ given by

$$\mathscr{G}f = \int_{\mathscr{T}} \mathcal{G}(\,\cdot\,,\boldsymbol{t})f(\boldsymbol{t})\,\mathrm{d}\boldsymbol{t}.$$

To ease notation it is assumed that $k = 1$. The general case follows easily. Furthermore, assume for simplicity that the observations are equidistantly spaced within $[0,1]^d$. For non-equidistant observations, one can introduce a normalization operator following Markussen (2013). Let $\mathscr{E}_{\cdot} : \mathbb{R}^n \to \mathcal{C}(\mathscr{T},\mathbb{R})$ be a linear embedding of the observation space into the space of piecewise linear functions on $\mathscr{T}$. For n large, one has the Riemannian sum approximation of the integral

$$\boldsymbol{\Sigma}\boldsymbol{z} \approx \{n\,\mathscr{G}\mathscr{E}_{\boldsymbol{z}}(\boldsymbol{t}_i)\}_i. \tag{6}$$

Assuming that $\mathcal{G}$ is two times continuously differentiable within the d-cubes spanned by the observation points, the approximation error can be specified explicitly by applying the trapezoidal rule on the right-hand side integrals, mimicking Proposition 1 in Markussen (2013). The error is of order $\sum_{i=1}^d O(n_i^{-1})$ where n_i denotes the number of sample points across data dimension i, i.e. $n = n_1 \cdots n_d$.

Denote by $\mathscr{L} = \mathscr{G}^{-1}$ the precision operator corresponding to $\mathscr{G}$, i.e.

$$\mathscr{L}\mathcal{G}(\,\cdot\,,\boldsymbol{t}) = \delta_{\boldsymbol{t}} \tag{7}$$

where $\delta_{\boldsymbol{t}}$ is the Dirac delta function at $\boldsymbol{t}$. In many cases $\mathscr{L}$ is a differential operator with $\mathcal{G}$ as its corresponding Green's function. For a general introduction to Green's functions we refer to the monograph by Duffy (2001). The relation between covariance functions and differential operators can be used to approximate calculations involving the covariance matrix $\boldsymbol{\Sigma}$.

First we consider the conditional expectation (3). One may rewrite the matrix product, to get

$$\mathrm{E}\,[\,\boldsymbol{x}\,|\,\boldsymbol{y}\,] = \left(\mathbb{I}_n + \frac{\sigma^2}{\tau^2}\boldsymbol{\Sigma}^{-1}\right)^{-1}(\boldsymbol{y}-\boldsymbol{\Gamma}\hat{\boldsymbol{\theta}}).$$

By using the approximation (6) and the fact that inversion is a continuous operation, one can derive (component-wise) operator approximations of the conditional expectation (3)

$$\widehat{\mathrm{E}}\left[\,\boldsymbol{x}\,|\,\boldsymbol{y}\,\right] = \left(\mathbb{I} + \frac{\sigma^2}{n\tau^2}\mathscr{L}\right)^{-1} \mathscr{E}_{\boldsymbol{y}-\boldsymbol{\Gamma}\hat{\boldsymbol{\theta}}}, \tag{8}$$

which means that the conditional expectation can be approximated by applying an integral operator with smoothing kernel corresponding to the Green's function of $\mathbb{I} + \frac{\sigma^2}{n\tau^2}\mathscr{L}$ on the continuously embedded residual $\boldsymbol{y} - \boldsymbol{\Gamma}\hat{\boldsymbol{\theta}}$. As opposed to the original conditional expectation (3) that requires inversion of a possibly dense covariance matrix, the operator approximation (8) require the inversion of an operator. This may be done explicitly, and the approximation (8) can typically be evaluated in linear time, and may in fact often be evaluated at all observation points in linear time (Markussen 2013). Furthermore, convolving high-dimensional data with possibly non-isotropic smoothing kernels can be done very efficiently on massively parallel processors (Hartung et al. 2012).

By applying the differential operator $\mathbb{I} + \frac{\sigma^2}{n\tau^2}\mathscr{L}$ on both sides of equation (8) one gets that $f = \widehat{\mathrm{E}}\left[\,\boldsymbol{x}\,|\,\boldsymbol{y}\,\right]$ is the solution to the partial differential equation

$$\mathscr{L} f = \frac{n\tau^2}{\sigma^2}(\mathscr{E}_{\boldsymbol{y}-\boldsymbol{\Gamma}\hat{\boldsymbol{\theta}}} - f). \tag{9}$$

In general, numerical solution of the differential equation (9) is the most efficient choice for obtaining the approximated conditional expectation (8). In particular, GPUs are very suited for efficiently solving differential equations based on finite difference approximations (Micikevicius 2009).

In the following, point evaluation of $\widehat{\mathrm{E}}\left[\,\boldsymbol{x}\,|\,\boldsymbol{y}\,\right]$ will be assumed to be done at all data points, giving a vector object directly comparable to $\mathrm{E}\left[\,\boldsymbol{x}\,|\,\boldsymbol{y}\,\right]$. Point evaluation is always done after applications of operators, for example differentiation.

Considering the differential equation (9), one can derive a numerically stable expression for the last part of the expanded quadratic term (5). By inserting the functional approximations of the conditional expectation in the term and using (9), one gets that

$$\widehat{\mathrm{E}}\left[\,\boldsymbol{x}\,|\,\boldsymbol{y}\,\right]^\top (\mathscr{E}_{\boldsymbol{y}-\boldsymbol{\Gamma}\hat{\boldsymbol{\theta}}} - \widehat{\mathrm{E}}\left[\,\boldsymbol{x}\,|\,\boldsymbol{y}\,\right]) = \frac{\sigma^2}{n\tau^2}\widehat{\mathrm{E}}\left[\,\boldsymbol{x}\,|\,\boldsymbol{y}\,\right]^\top \mathscr{L}\widehat{\mathrm{E}}\left[\,\boldsymbol{x}\,|\,\boldsymbol{y}\,\right].$$

Assuming that the covariance function $\mathcal{G}$ is positive definite, a square root $\mathscr{K}$ of $\mathscr{L}$ exists, such that $\mathscr{L} = \mathscr{K}^\dagger\mathscr{K}$, which means that the last term may also be written as a sum of squares

$$\frac{\sigma^2}{n\tau^2}(\mathscr{K}\widehat{\mathrm{E}}\left[\,\boldsymbol{x}\,|\,\boldsymbol{y}\,\right])^\top(\mathscr{K}\widehat{\mathrm{E}}\left[\,\boldsymbol{x}\,|\,\boldsymbol{y}\,\right]).$$

Finally, to approximate the determinant terms in the restricted likelihood function (4), one notes that

$$\frac{\mathrm{d}}{\mathrm{d}\alpha}\log\det\left[\mathbb{I}_n + \alpha\boldsymbol{\Sigma}\right] = \mathrm{tr}\big((\mathbb{I}_n + \alpha\boldsymbol{\Sigma})^{-1}\boldsymbol{\Sigma}\big),$$

which means that

$$\log\det[\mathbb{I}_n + \boldsymbol{\Sigma}] = \int_0^1 \sum_{\ell=1}^{n} \boldsymbol{e}_\ell^\top (\mathbb{I} + \alpha\boldsymbol{\Sigma})^{-1} \boldsymbol{\Sigma}\boldsymbol{e}_\ell \,\mathrm{d}\alpha,$$

where the vectors $\boldsymbol{e}_\ell$ constitute an orthonormal basis for $\mathbb{R}^n$. By approximating the matrix computations with their operator counterparts, one gets that

$$\log\det[\sigma^2\mathbb{I}_n + \tau^2\boldsymbol{\Sigma}] \approx \int_0^1 \int_{\mathscr{T}} \left(\alpha\mathbb{I} + \tfrac{\sigma^2}{n\tau^2}\mathscr{L}\right)^{-1} \delta_{\boldsymbol{t}}(\boldsymbol{t}) \,\mathrm{d}\boldsymbol{t} \,\mathrm{d}\alpha + n\log\sigma^2.$$

The integral term integrates over a family of Green's functions, and for many common covariance functions $\mathcal{G}$, the integral may be explicitly computed, resulting in constant time computation of the approximated log-determinant.

The explicit link between the covariance $\mathcal{G}$ and the differential operator $\mathscr{L}$ can be convenient in model specification. For some models, it may be natural to start out assuming that the random effect has a specific covariance function, and for others it may be straightforward to specify the differential operator.

Many well-known covariance functions $\mathcal{G}$ correspond to simple differential operators $\mathscr{L}$ with suitable boundary conditions. This will be illustrated in the following examples.

Example 3.1. Let $\mathscr{T} = [0,1]^d$ and $\mathscr{L} = \partial_{t_1}^2 \cdots \partial_{t_d}^2$. For homogeneous Dirichlet boundary conditions the corresponding Green's function is

$$\mathcal{G}(\boldsymbol{t}, \boldsymbol{t}') = (t_1 \wedge t_1' - t_1 t_1') \cdots (t_d \wedge t_d' - t_d t_d'),$$

which is the covariance of the tied-down Brownian bridge on $\mathscr{T}$. Alternatively, assuming homogeneous Dirichlet boundaries along the 0-boundaries, and corresponding Neumann boundaries along the 1-boundaries results in the Green's function

$$\mathcal{G}(\boldsymbol{t}, \boldsymbol{t}') = (t_1 \wedge t_1') \cdots (t_d \wedge t_d'),$$

which is the covariance of the Brownian sheet.

Other boundary conditions leads to e.g. the Brownian bridge on $\mathscr{T}$. Finally, assuming homogeneous Neumann boundary conditions may often be a good choice from a modeling point of view, as this corresponds to a Brownian process with a free level. Even though this will only make $\mathscr{L}$ and the corresponding covariance $\mathcal{G}$ positive semi-definite, all calculations can be done completely analogous to the cases where $\mathscr{L}$ is positive definite. ∘

Example 3.2. Let $\mathscr{T} = [0,1]^d$ and $\mathscr{L} = (-\Delta)^\ell + \varepsilon$ where Δ denotes the Laplace operator, $\epsilon > 0$, and $\ell \geq 2$. Under suitable boundary conditions and with $\varepsilon = 0$, this class of precision operators corresponds to penalizing the squares of derivatives (Wahba & Wendelberger 1980), which is commonly used for regularization.

For homogeneous Dirichlet boundary conditions one gets the covariance

$$\mathcal{G}(\boldsymbol{t}, \boldsymbol{t}') = \sum_{i_1,\ldots,i_d=1}^{\infty} \frac{2^d}{\pi^{2\ell}(i_1^2 + \cdots + i_d^2)^\ell + \varepsilon} \prod_{j=1}^{d} \sin(i_j \pi t_j) \sin(i_j \pi t_j'). \tag{10}$$

For Neumann boundaries the covariance function is similar, only with the sine functions substituted by cosines. When $\varepsilon = 0$, the covariance function is no longer positive, but the above expression is well defined, and so in practice one may choose $\varepsilon = 0$.

For some choices of d and ℓ more compact descriptions are available (Duffy 2001, chap. 5). Finally it is worth noting that these Green's functions may take the value $+\infty$ on the diagonal, corresponding to infinite variance. This happens for example when $d = 2$ and $\ell = 1$. ∘

Example 3.3. Let $\mathscr{T} = \mathbb{R}^d$ and $\mathscr{L} = (\kappa^2 - \Delta)^{\alpha/2}$ with free boundary conditons. Assume that $\alpha = \nu + d/2$, $\kappa > 0$, and $\nu > 0$. This choice of precision $\mathscr{L}$ has the Matérn covariance function (Lindgren et al. 2011) as its Green's function

$$\mathcal{G}(\boldsymbol{t}, \boldsymbol{t}') = \frac{\|\boldsymbol{t} - \boldsymbol{t}'\|^v}{2^{\nu-1}\Gamma(\nu + d/2)(4\pi)^{d/2}\kappa^v} K_v(\kappa\|\boldsymbol{t} - \boldsymbol{t}'\|).$$

∘

Example 3.4. Suppose that $\boldsymbol{x}$ from (2) is a tied-down Brownian sheet on $[0, 1]^2$, i.e.

$$\mathcal{G}((t_1, t_2), (t_1', t_2')) = (t_1 \wedge t_1' - t_1 t_1')(t_2 \wedge t_2' - t_2 t_2').$$

The Green's function $\mathcal{G}^\alpha((t_1, t_2), (t_1', t_2'))$ for the differential operator $\mathscr{L} + \alpha\mathbb{I}$ is given by

$$\sum_{i=1}^{\infty} \frac{2\sinh(\frac{\sqrt{\alpha}}{i\pi}(1 - t_2 \vee t_2'))\sinh(\frac{\sqrt{\alpha}}{i\pi}(t_2 \wedge t_2'))}{i\pi\sqrt{\alpha}\sinh(\frac{\sqrt{\alpha}}{i\pi})} \sin(i\pi t_1)\sin(i\pi t_1').$$

With this expression one can explicitly compute (8). Furthermore, one can derive the following log-determinant approximation

$$\log\det[\sigma^2\mathbb{I}_n + \tau^2\boldsymbol{\Sigma}] \approx n\log\sigma^2 + \sum_{i=1}^{\infty} \log\left(\frac{i\pi\sigma}{\sqrt{n}\tau}\sinh\left(\frac{\tau\sqrt{n}}{i\pi\sigma}\right)\right) \tag{11}$$

which can be evaluated by cutting the sum off at some sufficiently high value of i. This provides an interesting generalization to the known log-determinant approximation for the Brownian bridge under Gaussian noise (Markussen 2013) which is

$$n\log\sigma^2 + \log\left(\frac{\sigma}{\sqrt{n}\tau}\sinh\left(\frac{\tau\sqrt{n}}{\sigma}\right)\right).$$

Finally, due to the symmetry of the eigenfunctions of $\mathscr{L}$ under Dirichlet and Neumann boundary conditions, the approximation (11) is identical to the expression one would get with Neumann boundary conditions. ∘

Example 3.5. Assume that $\mathscr{L} = (-\Delta)^\ell + \varepsilon$ and $\mathscr{T} = [0, 1]^2$ with homogeneous Dirichlet or Neumann boundary conditions. Using (10) the following

log-determinant approximation is easily derived

$$\log\det[\sigma^2\mathbb{I}_n + \tau^2\boldsymbol{\Sigma}] \approx n\log\sigma^2 + \sum_{i=1}^{\infty}\sum_{j=1}^{\infty}\log\left(1 + \frac{\tau^2 n}{\sigma^2}\frac{1}{\pi^{2\ell}(i^2+j^2)^{\ell}+\varepsilon}\right). \tag{12}$$

○

Example 3.6. To compare the computation time of the conditional expectation (3) with the approximation given by the solution of the differential equation (9), the two solutions were calculated for $m \times m$ images. The matrix solution (3) was calculated by efficiently inverting the matrix $\boldsymbol{V}$ in BLAS using the Cholesky decomposition and a single thread on a 3.4 GHz Intel Core i7. The differential equation (9) was solved using the explicit diffusion scheme described in detail in Appendix A. The scheme was implemented in CUDA C and executed on an NVIDIA GeForce GTX 680MX GPU with 1536 CUDA cores. The runtime results, excluding the construction time for the matrix $\boldsymbol{V}$ for the matrix approach, can be seen in Figure 1. We note that for $m = 50$, the runtime of the matrix computation is a factor 1200 slower than the solution of the differential equation. For the given observation sizes, we note that the GPU runtimes only differ slightly, with an average runtime increase of approximately 10% from $m = 10$ to $m = 50$ despite of the factor 25 increase in observation size. This is caused by the GPU not being fully utilized for data sizes in the given range, and the runtime is dominated by memory bandwidth. The runtime increase from $m = 10$ to $m = 1000$ of the GPU implementation was found to be merely a factor of 33.

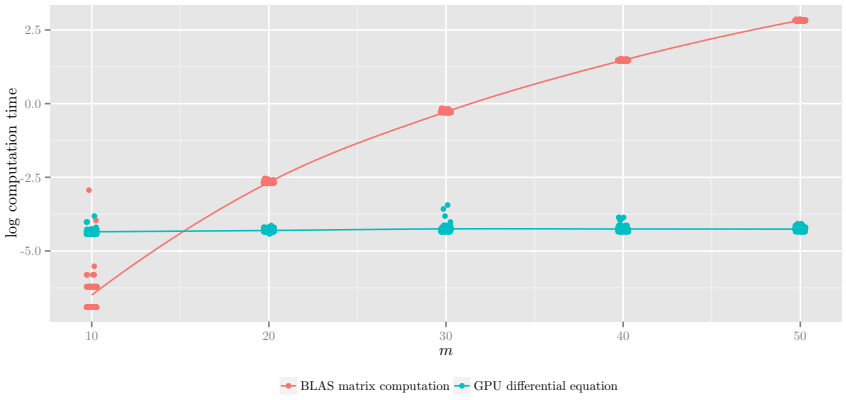

Figure 1: Runtime for the prediction of the conditional expectation using respectively the matrix formulation (3) and the differential equation (9) based on 100 replications for $m = 10, 20, 30, 40, 50$.

○

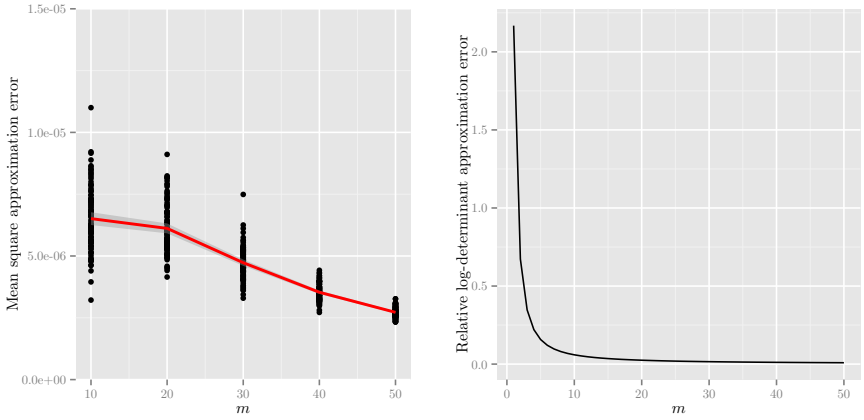

Figure 2: Mean square error of the approximated conditional expectation (8) computed by solving (9), based on 100 replications per m (left) and relative approximation error of the log-determinant (right).

Example 3.7. To assess the quality of the approximations, observations of tied-down Brownian sheets on $[0,1]^2$ with added Gaussian noise have been generated. The observation points are on an equidistant $m \times m$ grid, for varying values of m. The parameters in terms of the model (2) were $\boldsymbol{\Gamma} = \mathbf{0}$, $\sigma^2 = 0.1$, and $\tau^2 = 1$.

Figure 2 shows the mean square error of the approximated conditional expectation (8) with respect to the original conditional expectation (3), and the relative error of the log-determinant approximation (11). The approximated conditional expectation was computed by solving the differential equation (9) using the same setup as described in the previous example. The log-determinant approximation was computed using the formula (11) where n was replaced by $n+1$ in the second term to correct for the Dirichlet boundary conditions, and the sum was cut off after 10,000 terms.

Both approximations clearly improve as m increases. In particular, it is worth noting that the relative error of the log-determinant approximation seems to converge faster than $O(m^{-1})$. ∘

3.1 Related models

The model (2) is closely related to other types of models. In particular, assuming that $k = 1$ and $\boldsymbol{\Gamma} = \mathbf{0}$, one arrives at the classical functional data model (Ramsay & Silverman 2005)

$$\boldsymbol{y} = \boldsymbol{x} + \boldsymbol{\varepsilon}. \tag{13}$$

which is typically written in functional form as

$$y(\boldsymbol{t}) = x(\boldsymbol{t}) + \varepsilon(\boldsymbol{t}).$$

One can think of the model (13) as a Bayesian model with $\boldsymbol{x}$ as the prior of the observed function. In this model, the conditional expectation corresponds to the Bayes estimator of the function. Alternatively one can see the conditional expectation as the minimizer of the penalized likelihood function

$$\ell_{\boldsymbol{y}}(x) = (\boldsymbol{y} - x(\boldsymbol{t}_i)_i)^\top (\boldsymbol{y} - x(\boldsymbol{t}_i)_i) + \lambda \int_{\mathscr{T}} x(\boldsymbol{t}) \mathscr{L} x(\boldsymbol{t}) \, \mathrm{d}\boldsymbol{t}, \tag{14}$$

where the λ parameter corresponds to σ^2/τ^2 in the mixed-effects and Bayesian model, and $x(\boldsymbol{t}_i)_i$ is the column vector consisting of the function x evaluated at the points $\boldsymbol{t}_1, \ldots, \boldsymbol{t}_n$. In these cases one would typically estimate parameters by means of marginalized likelihood methods or the generalized cross validation criterion (Craven & Wahba 1978)

$$\mathrm{GCV}(\lambda) = \frac{n}{(n - \mathrm{df}(\lambda))^2} (\boldsymbol{y} - \hat{\boldsymbol{x}}_\lambda)^\top (\boldsymbol{y} - \hat{\boldsymbol{x}}_\lambda),$$

where $\hat{\boldsymbol{x}}_\lambda$ is the conditional expectation (3), with $\lambda = \sigma^2/\tau^2$, and $\mathrm{df}(\lambda)$ is the trace of the matrix $\frac{1}{2\lambda}\boldsymbol{\Sigma}(\mathbb{I} + \frac{1}{2\lambda}\boldsymbol{\Sigma})^{-1}$. Similarly to the calculations for the log-determinant, one can approximate

$$\mathrm{df}(\lambda) \approx \int_{\mathscr{T}} \mathcal{G}^*_\lambda(\boldsymbol{t}, \boldsymbol{t}) \, \mathrm{d}\boldsymbol{t},$$

where $\mathcal{G}^*_\lambda$ is the Green's function corresponding to the differential operator $\frac{2\lambda}{n}\mathscr{L} + \mathbb{I}$, and thus carry out the generalized cross validation using operator approximations. If a marginalized likelihood approach is preferred, the likelihood can be approximated using the already presented approximations.

In addition to the connection between the mentioned statistical models, the differential equation (9) naturally links mathematical models governed by this type of equation to the models described here. This in turn allows the use of the mentioned criteria to estimate parameters in such mathematical models.

3.2 Related work

It was noticed by Dolph & Woodbury (1952) that covariance functions of stochastic processes and Green's functions were related through stochastic differential equations. The solution $\boldsymbol{x}$ to the stochastic partial differential equation

$$\mathscr{L}\boldsymbol{x}(\boldsymbol{t}) = \boldsymbol{w}(\boldsymbol{t}), \tag{15}$$

where $\boldsymbol{w}$ is Gaussian white noise and $\mathscr{L}$ is positive definite, is a Gaussian random field with covariance $\mathcal{G}$—the Green's function of $\mathscr{L}$. In a somewhat similar fashion to what has been described in the present paper, Dolph & Woodbury (1952) used this representation to pose prediction problems for continuously observed curves as solutions to differential equations.

More recently, Lindgren et al. (2011) used the connection (15) with $\mathscr{L} = (\kappa^2 - \Delta)^{\alpha/2}$ as the definition of the class of MatÈrn fields, and derived a computationally efficient Markov representation of the solution. In contrast this

paper poses the prediction of the corresponding stochastic differential equation as a partial differential equation in the functional domain, and does not use any explicit representation of the data. Because of this relation to the stochastic differential equation formulation, the presented method can also be generalized to domains that are smooth manifolds, by simply changing the domain of (9), completely analogous to the manifold generalization by Lindgren et al. (2011). In addition, the presented method can handle a large class of covariance functions since the presented methods only need to identify the corresponding differential operator and solve a partial differential equation.

4 Example: Glyphosate data

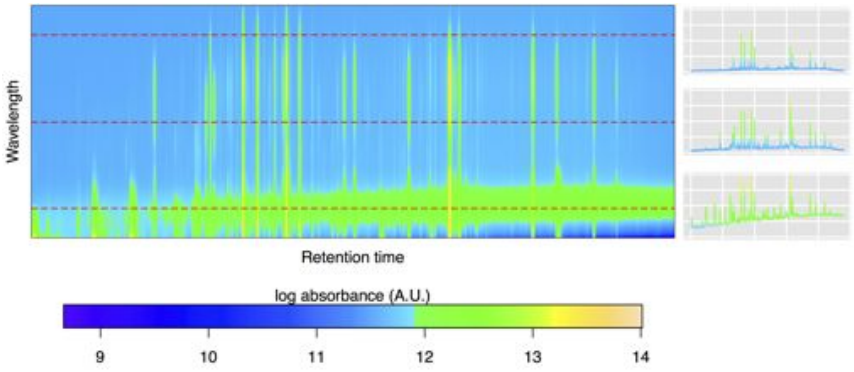

Figure 3: Example of a chromatogram along with absorbance curves for three fixed wavelengths (corresponding to the dashed red lines) on log-scale.

Consider a dataset consisting of $k = 28$ chromatograms $(\boldsymbol{y}_i)_{1 \leq i \leq 28}$, each of which consists of $n = 209 \times 24,000$ (wavelength $\times$ retention time) observations of absorbance (A.U.). The chromatograms have been generated using ultra-high-performing liquid chromatography with diode-array-detection (Petersen et al. 2011). The subjects of the analysis are rapeseed seedlings having been exposed to different levels of glyphosate, commonly known as Roundup®.

The original data have been preprocessed prior to the analysis. The chromatograms have been registered in retention time using a so-called TV-L^1 warping algorithm (Zach et al. 2007, Rakêt et al. 2011). First, the observations of each glyphosate-level group have been iteratively registered toward the group mean. Next, warping functions of all group means toward the maximum-glyphosate-level group mean are computed. Finally, these warps are applied to the intra-group registered observations, such that all samples follow a similar coordinate system. For the algorithmic details we refer to Rakêt (2013). Furthermore, the data does not have homogeneous variance; in flat regions, little or no noise is present while noise around peaks is stronger. To alleviate this

problem Gaussian noise with variance $2 \cdot 10^{-4}$ has been added to the logarithm of the registered absorbances. Figure 3 displays one of the preprocessed chromatograms, and from the scale of the log absorbance it is clear that the added noise is minuscule compared to the signal.

The logarithm of the absorbance is modeled according to (2)

$$\log(\boldsymbol{y}_i + 1) = \boldsymbol{\theta}_{e(i)} + \boldsymbol{x}_i + \boldsymbol{\varepsilon}_i \tag{16}$$

where the factor $e : \{1, \ldots, 28\} \to \{0, 1, 5, 10, 20, 30, 50\}$ with $p = 7$ levels gives the glyphosate exposure (in μM), and each $\boldsymbol{\theta}$ is $209 \times 24,000$ dimensional. The $\boldsymbol{x}_i$s are independent $209 \times 24,000$ dimensional free Brownian sheets (i.e. $n = 5,016,000$, $N = 140,448,000$ and $\mathscr{L} = \partial_s^2 \partial_t^2$ with Neumann boundary conditions) with variance parameter $\tau^2 = \sigma^2 \xi^2$, and the $\boldsymbol{\varepsilon}_i$s are independent, identically distributed Gaussian noise $\boldsymbol{\varepsilon}_i \sim \mathcal{N}(0, \sigma^2 \mathbb{I}_n)$. Brownian sheets have folds parallel to the axes, which also carry over to the associated posteriors (see e.g. Figure 10). This behavior makes the Brownian sheet a natural model for the present data, where responses at individual retention times are expected to extend along wavelengths. Furthermore, a multiplicative difference between chromatograms can be expected for this data. This gives a constant level shift after the log transformation. The Neumann boundary conditions (corresponding to a free level, see Example 3.1) are then natural for the problem, since the level shift may be captured in the prediction of the spatially correlated effects.

To approximate the restricted likelihood function (4) it is first noted that the determinant terms can be simplified

$$\det \boldsymbol{V} = \sigma^{2N} \det[\mathbb{I} + \xi^2 \boldsymbol{\Sigma}]^k, \qquad \det[\boldsymbol{\Gamma}^\top V^{-1} \boldsymbol{\Gamma}] = \sigma^{-2np} \left(\frac{k}{p}\right)^{np} \det[\mathbb{I} + \xi^2 \boldsymbol{\Sigma}]^{-p},$$

both of which are approximated using the operator approximation (11). In the given parametrization, a closed form restricted maximum likelihood estimate for σ^2 can be derived

$$\hat{\sigma}^2 = \frac{1}{N - np} \Big((\boldsymbol{y} - \boldsymbol{\Gamma}\hat{\boldsymbol{\theta}} - \widehat{\mathrm{E}}\,[\,\boldsymbol{x} \,|\, \boldsymbol{y}\,])^\top (\boldsymbol{y} - \boldsymbol{\Gamma}\hat{\boldsymbol{\theta}} - \widehat{\mathrm{E}}\,[\,\boldsymbol{x} \,|\, \boldsymbol{y}\,]) + \frac{1}{c\xi^2} (\mathscr{K} \widehat{\mathrm{E}}\,[\,\boldsymbol{x} \,|\, \boldsymbol{y}\,])^\top (\mathscr{K} \widehat{\mathrm{E}}\,[\,\boldsymbol{x} \,|\, \boldsymbol{y}\,]) \Big).$$

The conditional expectation is computed as the solution to the differential equation (9), which is solved numerically using a finite difference approximation with a stabilized explicit diffusion scheme on a GPU. We refer to Appendix A for the details.

The fixed effects $\boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \boldsymbol{\theta}_5, \boldsymbol{\theta}_{10}, \boldsymbol{\theta}_{20}, \boldsymbol{\theta}_{30}, \boldsymbol{\theta}_{50}$ are estimated pointwise, and the contrasts to baseline $\boldsymbol{\theta}_0$ can be found in Figure 4. Examples of the predicted spatially correlated effect can be found in Figure 5. We note that the range of the log absorbance values in the predicted spatially correlated effect is around one fifth of the range for the estimated fixed effect contrasts. The estimates of the variance parameters are 91.96 and $1.363 \cdot 10^{-2}$ respectively for ξ and σ.

Figure 6 displays a QQ plot of the conditional residual quantiles against normal quantiles and a scatter plot of conditional residuals against the estimated fixed effects. While the QQ plot shows non-normal tail behavior, this is caused by approximately 0.2% of the observations, and their effect on the estimate of σ is small. The residual plot shows an unnaturally large variation of the residuals corresponding to low absorbance, and for log absorbance levels of around 12.2. Nevertheless, these effects are again caused by very few observations, and the vast majority of the observations, that lie between log absorbance levels of 11.5 and 12, behave as one would expect.

Figure 7 shows the difference in log-likelihood evaluated at the maximum likelihood estimates between the original model (16) and the six models corresponding to collapsing the zero-exposure group with each of the other exposure level groups. The likelihood has been used instead of the restricted likelihood in order to invoke Wilk's likelihood ratio statistic (Pawitan 2001). Classical asymptotical behavior would prescribe twice the difference in log-likelihood to be approximately χ^2-distributed with degrees of freedom equal to n. In this example the test statistics of order $17 \cdot 10^6$ thus could be evaluated at approximately $5 \cdot 10^6$ degrees of freedom. However, since the validity of a χ^2-test with this many degrees of freedom is questionable, we have not computed p-values. However, there seems to be no doubt concerning the significant difference between the exposure groups. Apart from the $1\mu M$ exposure group, that has a somewhat irregular fixed effect (Figure 4), the log-likelihood differences behave as one would expect; differences increase with glyphosate level. The irregularity of the $1\mu M$ group is mainly caused by one observation with very strong peaks. The prediction of the corresponding spatially correlated effect can be seen in Figure 5 (top left).

5 Example: Simulated data

In this simulation example, $k = 25$ images on $\mathscr{T} = [0,1]^2$ sampled at 200×200 equidistant points have been generated from the model

$$\boldsymbol{y}_i = f_\alpha(\boldsymbol{t}_j)_j + g_{\beta_i,\gamma_i}(\boldsymbol{t}_j)_j + \boldsymbol{\varepsilon}_i, \tag{17}$$

where the functions f and g at a point $\boldsymbol{t} = (t_1, t_2)$ are given as

$$f_\alpha(t_1,t_2) = \sin(2\alpha t_1) - \sin(\alpha t_1 t_2)\cos(5t_2) + t_2,$$

$$g_{\beta,\gamma}(t_1,t_2) = g^\star_{\beta,\gamma}(t_1,t_2) - E[g^\star_{\beta,\gamma}(t_1,t_2)],$$

with

$$g^\star_{\beta,\gamma}(t_1,t_2) = \frac{1}{2}(\sin(\beta t_1 t_2)\cos(\gamma t_1)t_2^2 - \cos(\beta\gamma t_2)).$$

Here $\alpha \in \{1,\dots,10\}$ is a fixed integer, $\beta_i \sim \mathcal{N}(1,4)$, $\gamma_i \sim \mathcal{N}(1,9)$, $\boldsymbol{\varepsilon}_i \sim \mathcal{N}(0,\sigma^2\mathbb{I}_n)$ with $n = 40,000$ and variance $\sigma^2 = 0.1$, and all random variables are independent across the different samples. Images of the functions f_α and $g_{\beta,\gamma}$ with different parameters can be found in figures 8 and 9.

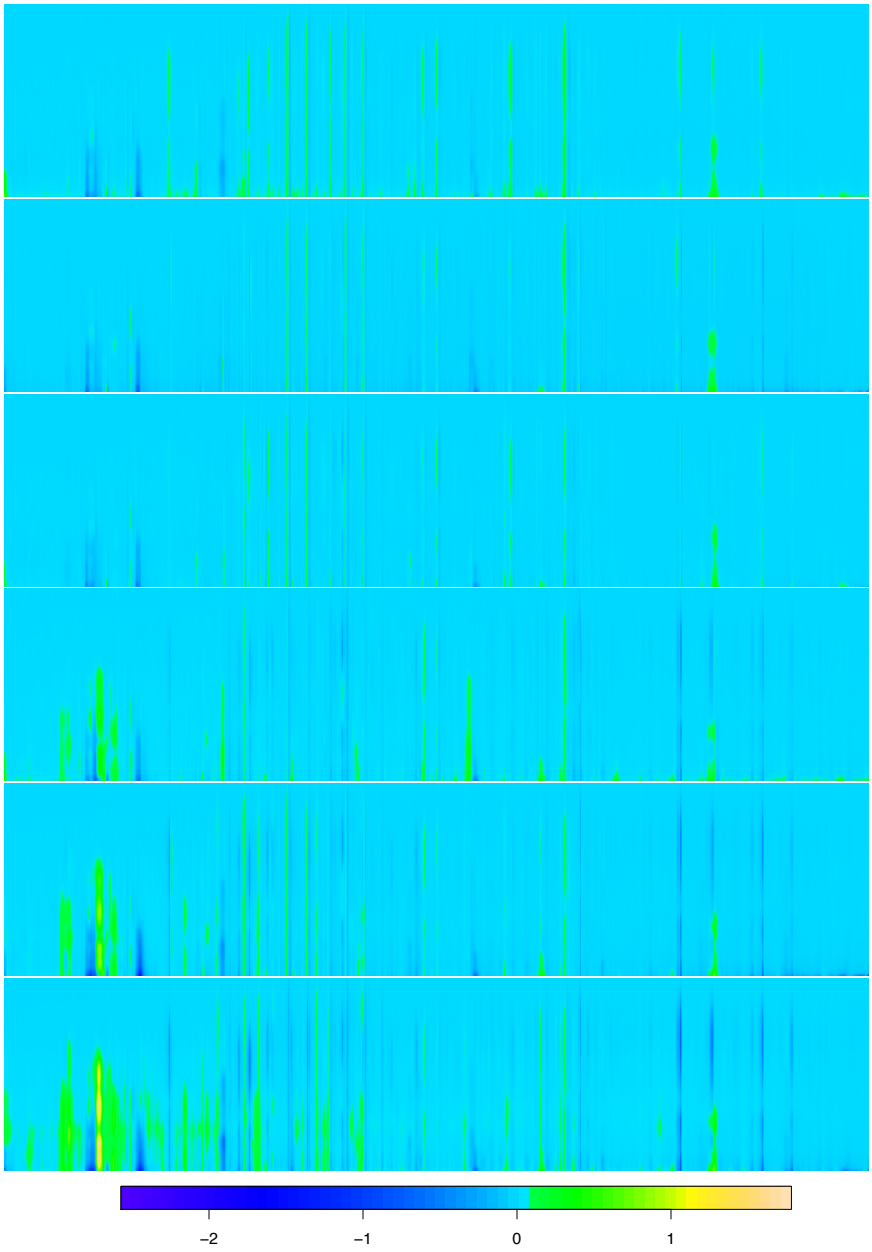

Figure 4: Differences between the estimated fixed effects $\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_5, \hat{\boldsymbol{\theta}}_{10}, \hat{\boldsymbol{\theta}}_{20}, \hat{\boldsymbol{\theta}}_{30}, \hat{\boldsymbol{\theta}}_{50}$ and baseline $\hat{\boldsymbol{\theta}}_0$ (from top to bottom).

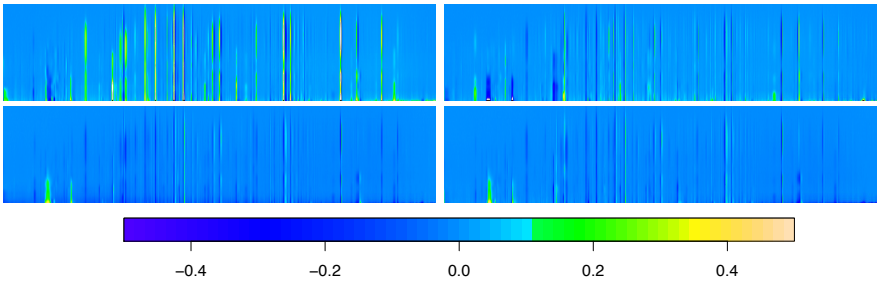

Figure 5: Predictions of the spatially correlated effects $\boldsymbol{x}_i$ for the four observations with glyphosate exposure level $1\,\mu M$.

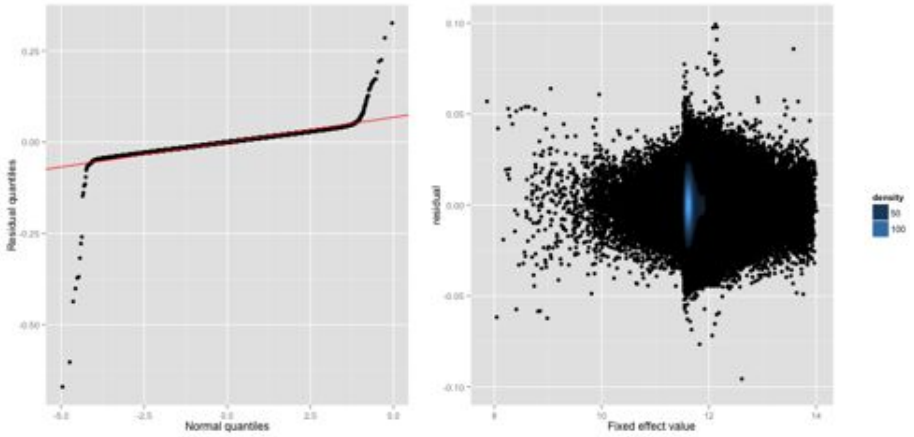

Figure 6: QQ plot and residual plot of a random sample consisting of 0.1% of the conditional glyphosate data residuals (1,404,480 data points), with the 38 most severe outliers removed from the residual plot. The line in the QQ plot shows the estimated standard deviation. For the residual plot the conditional residuals are plotted against the fitted values $\hat{\boldsymbol{\theta}}_{e(i)}$, and the point density is indicated in blue.

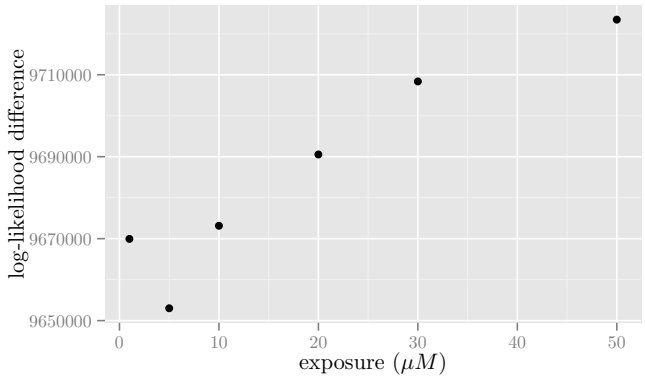

Figure 7: Log-likelihood differences between the model with the marked exposure level and zero-exposure level combined and the full model (16). The likelihood functions have been evaluated at the maximum likelihood estimates.

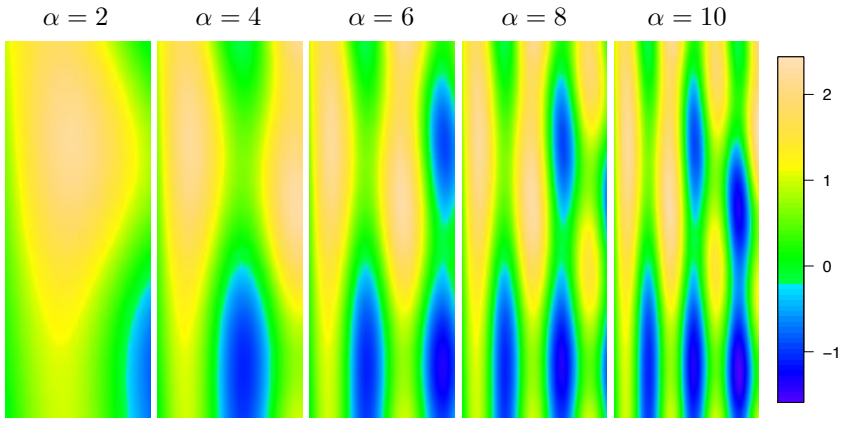

Figure 8: The function f_α for different values of α.

The spatially correlated part of the model is simulated from a parametric random effect model with two degrees of freedom, and it is investigated how the developed model performs under misspecification. This is relevant since one would expect the functional model to be misspecified in most real data applications.

The parametrization and calculations from the previous example trivially carries over to this example. Figures 10 and 11 show examples of the conditional expectation under the assumption of a free Brownian sheet effect and of an effect with biharmonic precision $\mathscr{L} = \Delta\Delta$. For the presented figures $\alpha = 6$ was used and the spatially correlated effects shown correspond to those of Figure 9. In this setting the smoother predictions from the biharmonic precision consistently

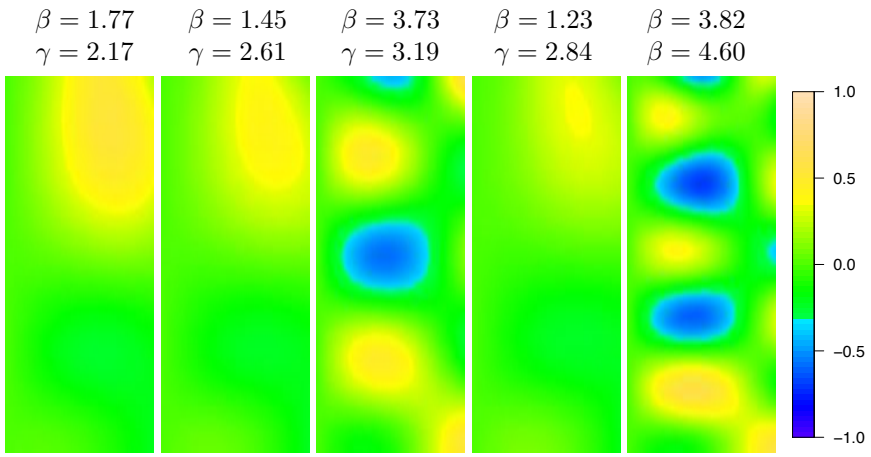

Figure 9: The function $g_{\beta,\gamma}$ with β and γ values simulated following $\beta \sim \mathcal{N}(1, 4)$, $\gamma \sim \mathcal{N}(1, 9)$.

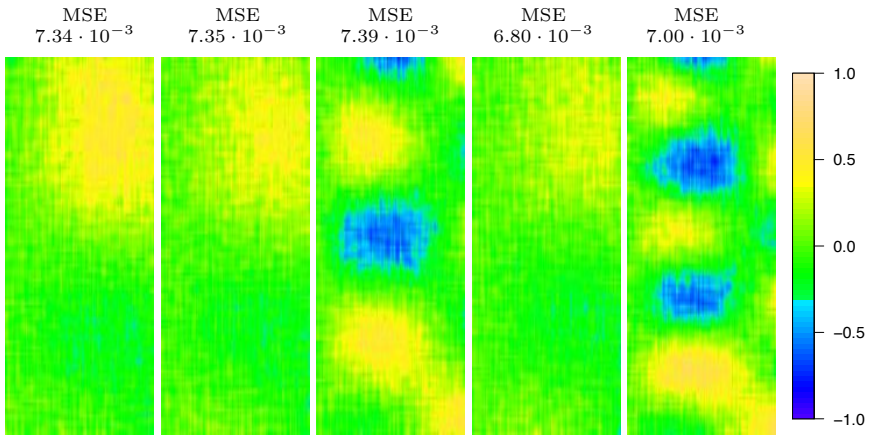

Figure 10: Predictions of the spatially correlated effects from Figure 9 in the model (17) with $\alpha = 6$ under the assumption of a free Brownian sheet effect, with $\hat{\xi} = 0.115$, along with mean squared errors (MSEs).

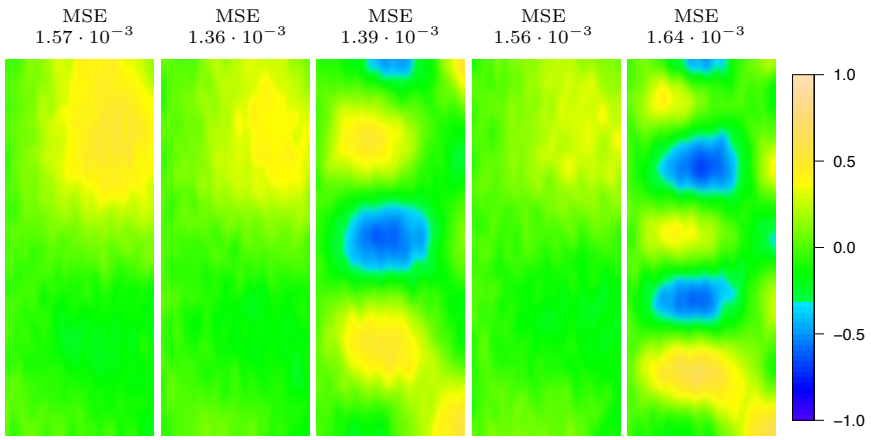

Figure 11: Predictions of the spatially correlated effects from Figure 9 in the model (17) with $\alpha = 6$ under the assumption of a precision operator $\mathscr{L} = \Delta\Delta$, with $\hat{\xi} = 0.0535$, along with mean squared errors (MSEs).

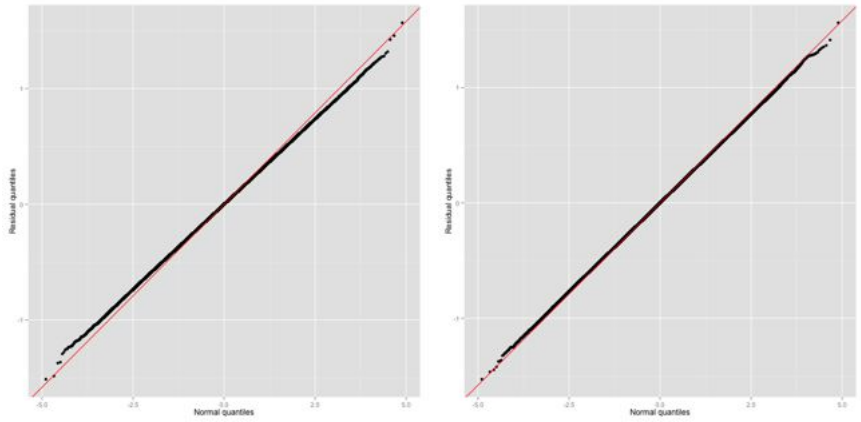

Figure 12: QQ plot of the conditional residuals from the model with a Brownian (left) and biharmonic (right) spatially correlated effect (1,000,000 data points). The lines show the true standard deviation.

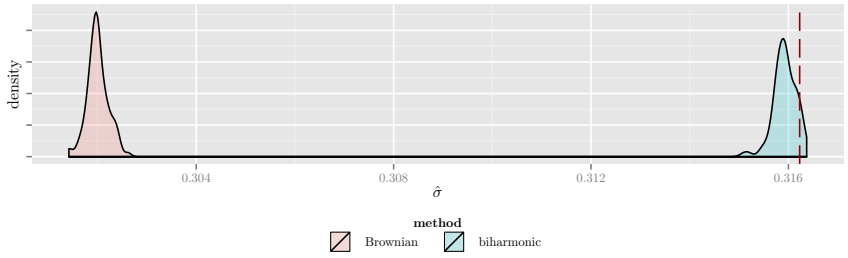

Figure 13: Histograms of parameter estimates in the model (17) under assumption of Brownian and biharmonic correlated effects. The dashed red line in the right histogram shows the true standard deviation.

lead to better predictions of the spatially correlated effects. QQ plots of the conditional residuals can be found in Figure 12. While both plots look very reasonable, it can be seen that the biharmonic model gives a better variance estimate. This is caused by the inherent roughness of the Brownian sheet prior, that will capture some of the noise in the prediction of the spatially correlated effects.

To quantify the behaviour of the variance parameter estimators 100 independent replications (10 for each value of α) of data from the model (17) have been generated. Figure 13 shows a histogram of $\hat{\sigma}^2$ under the assumption of a Brownian and biharmonic correlated effect, respectively. The previously mentioned property that the Brownian sheet effect results in underestimation of the true standard deviation (0.3162) is clearly visible. It is also seen that the biharmonic effect underestimates the standard deviation, although to a much smaller extent.

6 Discussion

This work presents a new method for conducting classical statistical analyses of functional data. By avoiding a direct representation of the data, and doing calculations in the functional domain, the computational complexities of the likelihood function and the predictions of spatially correlated effects are significantly reduced. In addition to reducing the computational complexity, the problem of predicting spatially correlated effects may be posed as a partial differential equation. Solvers for such partial differential equations are easily implemented on massively parallel processors, which drastically decrease computation times. CUDA C and R (R Core Team 2012) code for conducting the presented analyses on NVIDIA graphics hardware is available as supplementary material.

The presented methods allow for analyzing data that are orders of magnitude larger than what has previously been feasible. Using a massively parallel

implementation, it was demonstrated that statistical analysis of a dataset of 2D chromatograms, consisting of more than 140 million spatially correlated observations, can be done in a matter of minutes.

The considered model was kept simple to illustrate the computational methods, but a number of generalizations can be made. Extensions to vector valued data and more complex designs, including functional fixed effects, are straightforward, and the approximations may be useful in e.g. hierarchical functional models (Staicu et al. 2010). Furthermore, the results are also easily adapted to the case of the domain $\mathscr{T}$ being more complex than what was considered here, e.g. a smooth manifold. Further generalizations that are relevant from the perspective of achieving valid statistical models, but also require new methodological work, is to allow for variance heterogeneity (Pintore et al. 2006, Yue, Speckman & Sun 2012, Yue, Simpson, Lindgren & Rue 2012) and to incorporate data registration directly in the mixed-effects model.

References

Boularan, J., Ferré, L. & Vieu, P. (1994), 'Growth curves: a two-stage nonparametric approach', *Journal of Statistical Planning and Inference* **38**(3), 327–350.

Chen, H. & Wang, Y. (2011), 'A penalized spline approach to functional mixed effects model analysis', *Biometrics* **67**(3), 861–870.

Craven, P. & Wahba, G. (1978), 'Smoothing noisy data with spline functions', *Numerische Mathematik* **31**, 377–403.

da Silva, A. F. (2010), 'cudaBayesreg: Bayesian Computation in CUDA', *The R Journal* **2**(2), 48–55.

Dolph, C. L. & Woodbury, M. A. (1952), 'On the relation between Green's functions and covariances of certain stochastic processes and its application to unbiased linear prediction', *Transactions of the American Mathematical Society* pp. 519–550.

Duffy, D. G. (2001), *Green's Functions with Applications*, Chapman & Hall/CRC.

Ferraty, F. & Vieu, P. (2006), *Nonparametric Functional Data Analysis*, Springer.

Grewenig, S., Weickert, J. & Bruhn, A. (2010), From box filtering to fast explicit diffusion, *in* M. Goesele, S. Roth, A. Kuijper, B. Schiele & K. Schindler, eds, 'Pattern Recognition', Vol. 6376 of *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, pp. 533–542.

Guo, W. (2002), 'Functional mixed effects models', *Biometrics* **58**(1), 121–128.

Hartung, S., Shukla, H., Miller, J. P. & Pennypacker, C. (2012), GPU acceleration of image convolution using spatially-varying kernel, *in* 'Image Processing (ICIP), 2012 19th IEEE International Conference on', IEEE, pp. 1685–1688.

Harville, D. A. (1977), 'Maximum likelihood approaches to variance component estimation and to related problems', *Journal of the American Statistical Association* **72**, 320–340.

Horváth, L. & Kokoszka, P. (2012), *Inference for functional data with applications*, Vol. 200, Springer.

Jordan, M. I. (2011), 'Message from the President: What are the open problems in Bayesian statistics?', *ISBA Bulletin* **18**, 1–4.

Lee, D.-J., Durbán, M. & Eilers, P. (2013), 'Efficient two-dimensional smoothing with P-spline ANOVA mixed models and nested bases', *Computational Statistics & Data Analysis* **61**, 22 – 37.

Lee, Y., Nelder, J. A. & Pawitan, Y. (2006), *Generalized Linear Models With Random Effects: Unified Analysis Via H-Likelihood*, Chapman & Hall.

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

Liu, Z. & Guo, W. (2012), 'Functional mixed effects models', *Wiley Interdisciplinary Reviews: Computational Statistics* **4**(6), 527–534.

Markussen, B. (2013), 'Functional data analysis in an operator-based mixed-model framework', *Bernoulli* **19**, 1–17.

Micikevicius, P. (2009), 3D finite difference computation on GPUs using CUDA, *in* 'Proceedings of 2nd Workshop on General Purpose Processing on Graphics Processing Units', ACM, pp. 79–84.

Núñez-Antón, V., Rodríguez-Póo, J. M. & Vieu, P. (1999), 'Longitudinal data with nonstationary errors: a nonparametric three-stage approach', *Test* **8**(1), 201–231.

Pawitan, Y. (2001), *In All Likelihood*, Oxford University Press.

Petersen, I. L., Tomasi, G., Sørensen, H., Boll, E. S., Hansen, H. C. B. & Christensen, J. H. (2011), 'The use of environmental metabolomics to determine glyphosate level of exposure in rapeseed (Brassica napus L.) seedlings', *Environmental Pollution* **159**(10), 3071 – 3077.

Pintore, A., Speckman, P. & Holmes, C. C. (2006), 'Spatially adaptive smoothing splines', *Biometrika* **93**(1), 113–125.

R Core Team (2012), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
URL: *http://www.R-project.org/*

Rakêt, L. L. (2013), 'Duality based optical flow algorithms with applications', University of Copenhagen prize thesis in Computer Science, Copenhagen University Library.

Rakêt, L. L., Roholm, L., Nielsen, M. & Lauze, F. (2011), TV-L^1 optical flow for vector valued images, *in* Y. Boykov, F. Kahl, V. Lempitsky & F. Schmidt, eds, 'Energy Minimization Methods in Computer Vision and Pattern Recognition', Vol. 6819 of *Lecture Notes in Computer Science*, Springer, pp. 329–343.

Ramsay, J. O. & Silverman, B. W. (2005), *Functional Data Analysis*, second edn, Springer.

Robinson, G. K. (1991), 'That BLUP is a good thing: The estimation of random effects', *Statistical Science* **6**(1), pp. 15–32.

Staicu, A.-M., Crainiceanu, C. M. & Carroll, R. J. (2010), 'Fast methods for spatially correlated multilevel functional data', *Biostatistics* **11**(2), 177–194.

Suchard, M. A., Wang, Q., Chan, C., Frelinger, J., Cron, A. & West, M. (2010), 'Understanding GPU programming for statistical computation: Studies in massively parallel massive mixtures', *Journal of Computational and Graphical Statistics* **19**(2), 419–438.

Wahba, G. (1990), *Spline Models for Observational Data*, Society for Industrial and Applied Mathematics.

Wahba, G. & Wendelberger, J. (1980), 'Some new mathematical methods for variational objective analysis using splines and cross validation', *Monthly Weather Review* **108**, 1122.

Wang, Y. (1998), 'Mixed effects smoothing spline analysis of variance', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **60**(1), 159–174.

Weickert, J. & Schnörr, C. (2001), 'Variational optic flow computation with a spatio-temporal smoothness constraint', *Journal of Mathematical Imaging and Vision* **14**, 245–255.

Yue, Y. R., Simpson, D., Lindgren, F. & Rue, H. (2012), 'Bayesian adaptive smoothing spline using stochastic differential equations', *arXiv preprint arXiv:1209.2013* .

Yue, Y. R., Speckman, P. L. & Sun, D. (2012), 'Priors for Bayesian adaptive spline smoothing', *Annals of the Institute of Statistical Mathematics* **64**(3), 577–613.

Zach, C., Pock, T. & Bischof, H. (2007), A duality based approach for realtime TV-L^1 optical flow, *in* F. Hamprecht, C. Schnörr & B. Jähne, eds, 'Pattern Recognition', Vol. 4713 of *Lecture Notes in Computer Science*, Springer, pp. 214–223.

A Solving the fourth order PDEs

Consider the differential equation (9) with $\mathscr{L} = \partial_{t_1}^2 \partial_{t_2}^2$ (the case $\mathscr{L} = \Delta\Delta$ is treated similarly), i.e.

$$(\partial_{t_1}^2 \partial_{t_2}^2 + c)f = g. \tag{18}$$

This equation is solved using an explicit diffusion scheme, with an added artificial time variable. The solution to (18) is found as the steady state of the corresponding diffusion equation.

In order to get numerically stable solutions, $\partial_{t_1}^2 \partial_{t_2}^2$ is approximated using a 5×5 stencil, and furthermore the diffusion is stabilized by evaluating the center point of the stencil at the future time point (Weickert & Schnörr 2001). This scheme is stable for time steps of 0.125, but the convergence rate is greatly accelerated by using the so-called fast explicit diffusion (FED) method of Grewenig et al. (2010), which cleverly mixes stable and unstable time steps. In the following example the procedure is demonstrated for a one-dimensional example.

Example A.1. To illustrate the solution procedure, consider the one-dimensional version of the differential equation (18), i.e. the differential equation (9) with $\mathscr{L} = -\partial_t^2$. We approximate $\mathscr{L}$ by a standard five-point stencil, so assuming equidistant observations we get

$$\mathscr{L} f(t)|_{t=t_i} \approx \frac{f(t_{i-2}) - 16f(t_{i-1}) + 30f(t_i) - 16f(t_{i+1}) + f(t_{i+2})}{12(t_i - t_{i-1})^2}$$

The one-dimensional version of the differential equation (18) is given by

$$(-\partial_t^2 + c)f = g. \tag{19}$$

Instead of considering this equation directly, we introduce an artificial time variable τ and consider the diffusion equation

$$\partial_\tau f(t, \tau) = g(t) - (-\partial_t^2 + c)f(t, \tau), \tag{20}$$

where $f(t, 0)$ is initialized using the observed data values. The steady state of the differential equation (20) in τ, e.g. when $\partial_\tau f(t, \tau) = 0$ will solve the original differential equation (19). We discretize

$$\partial_\tau f(t, \tau)|_{\tau=\tau_j} \approx \frac{f(t, \tau_{j+1}) - f(t, \tau_j)}{\tau_{j+1} - \tau_j}$$

and $\mathscr{L}f(t_i,\tau)|_{t=t_i,\tau=\tau_j}$ is approximated by

$$\frac{f(t_{i-2},\tau_j) - 16f(t_{i-1},\tau_j) + 30f(t_i,\tau_{j+1}) - 16f(t_{i+1},\tau_j) + f(t_{i+2},\tau_j)}{12(t_i - t_{i-1})^2},$$

where the future time point τ_{j+1} is used in the term $f(t_i,\tau_{j+1})$ for stability. The equation (20) is now solved iteratively by considering its finite difference representation, and taking time steps of size $\tau_{j+1} - \tau_j$, where at each step we solve for $f(t_i,\tau_{j+1})$. ∘

For the glyphosate data from Section 4, the diffusion is assumed to have reached its steady state once the artificial time reaches 5,000, corresponding to 40,000 iterations using a step size of 0.125, or a mere 346 FED steps.

The presented solver has been implemented in CUDA C, in order to utilize the thousands of cores on modern GPUs. The runtime (including writing to GPU memory) for computing the solution to (18) for a single $209 \times 24,000$ chromatogram is on average 2.0 seconds on an NVIDIA GeForce GTX 680MX. The resulting average computation time for the restricted likelihood function (4) is 69 seconds on the full glyphosate dataset presented in Section 4.