

## BAYESIAN GENERALIZED TWO-WAY ANOVA MODELING FOR FUNCTIONAL DATA USING INLA

Yu (Ryan) Yue, David Bolin, Håvard Rue and Xiao-Feng Wang

*Baruch College, Chalmers University of Technology,  
King Abdullah University of Science and Technology  
and Cleveland Clinic Lerner Research Institute*

*Abstract:* Functional analysis of variance (ANOVA) modeling has been proved particularly useful to investigate the dynamic changes of functional data according to certain categorical factors and their interactions. The current existing methods often encounter difficulties when the functional data are high-dimensional, non-Gaussian, and/or exhibit certain shape characteristics that vary with spatial locations. In this paper, we investigate the functional two-way ANOVA models from a novel Bayesian perspective. A class of highly flexible Gaussian Markov random fields (GMRF) are taken as priors on the functions in the model, which allows us to model various types of functional effects, such as (discrete or continuous) temporal effects and (point-level or areal) spatial effects. The resulting posterior distributions are obtained by an efficient computational tool based on integrated nested Laplace approximations (INLA) Rue, Martino and Chopin (2009). We then employ the excursion method introduced by Bolin and Lindgren (2015) to build simultaneous credible intervals of functional effects and test their significance from a Bayesian point of view. A simulation study and multiple data examples are presented to demonstrate the merits of our method.

*Key words and phrases:* Contour avoiding functions, functional data, Gaussian Markov random fields, generalized ANOVA, integrated nested Laplace approximations, simultaneous credible intervals.

### 1. Introduction

Functional data analysis plays an important role in modern statistics. Being associated with continuous time/space monitoring processes, functional data are usually a sample of curves or surfaces and are often treated as realizations of underlying random functions. The basic philosophy of functional data is to consider observed data functions as single entities, rather than only as a sequence of individual observations. Functional data are becoming increasingly common in a variety of scientific and medical areas. A considerable effort has been made in

statistical fields to adapt standard statistical methods to functional data. The general ideas of functional data analysis were laid out by Ramsay and Dalzell (1991). Müller (2005) developed some concepts for estimating functional principal component scores in the sparse situation, and proposed an extension of the generalized functional linear models and functional logistic discrimination for the classification of functional data. Functional modeling methods have also become more and more popular in applications. For instance, Wang et al. (2009) applied a functional random-effects model to human brain electroencephalogram (EEG) data. Comprehensive surveys of parametric and nonparametric techniques for analyzing functional data can be found in Ramsay and Silverman (2005) and Ferraty and Vieu (2006). Bayesian approaches to functional data analysis have also drawn increasing attention. Morris and Carroll (2006) proposed a Bayesian wavelet-based approach for functional mixed models. Crainiceanu and Goldsmith (2010) presented functional linear regression and functional principle component analysis using WinBUGS. Goldsmith, Wand and Crainiceanu (2011) studied functional regression via variational Bayes.

This paper focuses on the analysis of variance (ANOVA) models for functional data, which have proven particularly useful when the functional data are expected to differ with certain categorical factors (see e.g., Faraway (1997); Brumback and Rice (1998)). We start from a simple one-way functional ANOVA model, where there is only a single factor of interest. If  $y_{ik}(\mathbf{x})$  is the  $k$ th functional observation in the  $i$ th group, then the model can be expressed as

$$y_{ik}(\mathbf{x}) = \mu(\mathbf{x}) + \alpha_i(\mathbf{x}) + \varepsilon_{ik}(\mathbf{x}), \mathbf{x} \in \mathcal{X} \subset \mathbb{R}^d, \quad (1.1)$$

where  $\mu(\mathbf{x})$  is the grand mean,  $\alpha_i(\mathbf{x})$  is the  $i$ th group effect function, and  $\varepsilon_{ik}(\mathbf{x})$  is the residual function with mean zero, for  $i = 1, \dots, m_A$  and  $k = 1, \dots, n$ . To be able to identify them, we require the group effect functions to satisfy certain constraint, e.g.,  $\alpha_{m_A}(\mathbf{x}) = 0$  for all  $\mathbf{x}$ . There are two main questions to address in functional ANOVA: the first is “How do we estimate the functions in the model?” and the second is “How do we test the functional effects of various factors, e.g., test  $H_0 : \alpha_1(\mathbf{x}) = \dots = \alpha_{m_A}(\mathbf{x}) = 0$  in model (1.1)?” To estimate the functions, Brumback and Rice (1998) used the correspondence between penalized regression and mixed-effects models to extend the smoothing spline model (see e.g., Wahba (1990)) for individual curves. Ramsay and Silverman (2005) suggested a penalized least squares approach with regularized basis expansions. The basis functions commonly used are Fourier series or B-spline functions; they allow a continuous control over smoothness while still permitting as much high frequency

detail in the model as the functional data require. Regarding hypothesis tests in functional ANOVA, Faraway (1997) and Zhang and Chen (2007) proposed  $L^2$ -norm-based tests while Shen and Faraway (2004) discussed an  $F$ -type test. Cuevas, Febrero and Fraiman (2004) derived the random expression of the test statistics for a  $L^2$ -norm-based test, and used a parametric bootstrap method to approximate the null distribution. Zhang and Liang (2014) proposed globalizing the point-wise  $F$ -test suggested by Ramsay and Silverman (2005).

Bayesian functional ANOVA has been given less attention. Kaufman and Sain (2010) introduced a Gaussian process ANOVA modeling approach under a Bayesian framework, with Gaussian process distributions assigned to each batch of functional effects as priors, and Markov chain Monte Carlo (MCMC) algorithms used for posterior sampling. They emphasized graphical summaries based on the posterior distribution of model-based analogues of traditional ANOVA decompositions of variance. In the spirit of Kaufman and Sain, Kang and Cressie (2013) proposed a spatial ANOVA model, with spatial priors assigned for the functional effects, to analyze regional climate-change projections. These approaches are only able to deal with Gaussian functional data, and the MCMC algorithms are slow and not easily implemented.

In this work, we present a general and efficient Bayesian modeling approach to conducting ANOVA for a variety of (Gaussian and non-Gaussian) functional data. A class of highly flexible Gaussian Markov random fields (GMRF) (see, e.g., Rue and Held (2005)) are assigned as priors for each batch of functional effects in the model. As shown in Section 6, the GMRF priors allow for modeling various underlying functions, such as one-dimensional curves in the example of diffusion tensor imaging (DTI), two-dimensional graphical regions in the example of Apgar scores, and even locations in a spherical coordinate system in the example of global solar radiation. To the best of our knowledge, no functional ANOVA model has been applied to analyze the functions of the spherical coordinates.

In our approach, the resulting posterior distributions are obtained by an efficient computational tool based on integrated nested Laplace approximations (INLA) (Rue, Martino and Chopin (2009)). INLA provides accurate and fast approximations to the posterior marginals, and thus avoids time-consuming MCMC sampling. INLA also allows for a broad range of distributions as the functional data. For instance, we use the beta distribution in the DTI example and the binomial distribution in the Apgar score example. Most distributions from the exponential family can be specified. Our method is flexible in the choice of likelihoods as well as the distributions of functional effects.

To see whether or not the functional effects are statistically significant, we traditionally conduct a hypothesis test with, e.g.,  $H_0 : \alpha_1(\mathbf{x}) = \cdots = \alpha_{m_A}(\mathbf{x}) = 0$  for all  $\mathbf{x} \in \mathcal{X}$ , and correct the family-wise error rate. It is difficult to transfer the same concept to a Bayesian hierarchical modeling framework. We therefore resort to simultaneous credible intervals and contour avoiding sets derived from the posteriors for finding the significant functional effects. When the functions of interest are curves, as in the DTI example, one can use simultaneous intervals, and conclude that the functional effect is significant if any part of its interval does not contain 0. When the functions are of higher dimensions, as in the other examples, it is preferable to find the contour avoiding set, in which the function is nonzero with a high probability for all  $\mathbf{x} \in \mathcal{X}$ . Both approaches require integrating the joint posterior distribution of  $[\alpha(\mathbf{x}_1), \dots, \alpha(\mathbf{x}_n)]$  given  $n$  observations of  $\mathbf{x}$ . This is difficult, as the integration is usually of a very high dimension due to the large sample size of a typical functional dataset, as demonstrated in our examples. Fortunately, every posterior approximation given by INLA can be viewed as a mixture of normals, and therefore the *excursion* method introduced by Bolin et al. (2015) can be used to efficiently estimate the simultaneous credible intervals and the contour avoiding sets.

The proposed functional ANOVA modeling approach can be implemented by general users. It is easy to use INLA under a Linux, Windows or Macintosh environment via R interface (R Core Team, 2017). A R package, INLA, can be downloaded from [www.r-inla.org](http://www.r-inla.org) for free. The excursion method for calculating the simultaneous credible intervals and the contour avoiding sets is implemented using the *excursions* package, available at [cran.r-project.org/web/packages/excursions](http://cran.r-project.org/web/packages/excursions).

In the rest of the paper, we address a generalized two-way functional ANOVA model, probably the most commonly-used one in practice. The method is readily extended to the cases with more factors. The generalized modeling framework is introduced in Section 2. The prior distributions taken on the functional effects are specified in Section 3. Posterior inference and the construction of simultaneous credible intervals and contour avoiding sets is addressed in Section 4. Simulation studies are presented in Section 5, and three case studies are discussed in Section 6, followed by closing remarks in Section 7. The analysis software programs and the data sets used in this paper are available in the Supplementary Material, and we hope they will help general users adapt the programs to their own functional ANOVA models.

## 2. Generalized Two-way Functional ANOVA Models

The one-way ANOVA model as (1.1) can be easily extended to accommodate additional categorical variables. Consider a two-way functional ANOVA model here. Suppose we have a two-by-two factorial design: each level of the factor A (with  $m_A$  levels) is crossed with each level of the factor B (with  $m_B$  levels). Let  $\{y_{ijk}(\mathbf{x}), \mathbf{x} \in \mathcal{X} \subset \mathbb{R}^d\}$  be a functional response at  $i$ th level of A and  $j$ th level of B from  $k$ th subject, for  $i = 1, \dots, m_A, j = 1, \dots, m_B$  and  $k = 1, \dots, n_{ij}$ . We assume that  $y_{ijk}(\mathbf{x})$  has a distribution from the exponential family with the density

$$\pi(y_{ijk}(\mathbf{x})) = \exp\left(\frac{y_{ijk}(\mathbf{x})\theta_{ijk}(\mathbf{x}) - b(\theta_{ijk}(\mathbf{x}))}{a(\psi_{ijk}(\mathbf{x}))} + c(y_{ijk}(\mathbf{x}), \psi_{ijk}(\mathbf{x}))\right),$$

where  $\theta_{ijk}(\mathbf{x})$  and  $\psi_{ijk}(\mathbf{x})$  are canonical function and dispersion function, respectively. The mean and variance are given by  $\mu_{ijk}(\mathbf{x}) = E\{y_{ijk}(\mathbf{x})\} = b'(\theta_{ijk}(\mathbf{x}))$  and  $\text{Var}(y_{ijk}(\mathbf{x})) = b''(\theta_{ijk}(\mathbf{x}))a(\psi_{ijk}(\mathbf{x}))$ , where  $b'(\theta)$  and  $b''(\theta)$  are the first two derivatives of  $b(\theta)$  with respect to  $\theta$ . We further denote the grand mean function by  $\mu$ , the main effect of A at  $i$ th level by  $\alpha_i$ , the main effect of B at  $j$ th level by  $\beta_j$ , and their interaction effect by  $\gamma_{ij}$ . Then, the mean response function  $\mu_{ijk}(\mathbf{x})$  is partitioned into

$$g(\mu_{ijk}(\mathbf{x})) = \mu(\mathbf{x}) + \alpha_i(\mathbf{x}) + \beta_j(\mathbf{x}) + \gamma_{ij}(\mathbf{x}), \tag{2.1}$$

where  $g$  is the *link function* used to associate the mean response with the two-way ANOVA model. For identifiability, we choose to restrict the batches of functions  $\alpha_i, \beta_j$  and  $\gamma_{ij}$  as

$$\alpha_{m_A}(\mathbf{x}) = 0, \beta_{m_B}(\mathbf{x}) = 0, \gamma_{m_A j}(\mathbf{x}) = 0, \gamma_{i m_B}(\mathbf{x}) = 0, \tag{2.2}$$

for  $\mathbf{x} \in \mathcal{X}$ . These constraints make  $\mu$  the (transformed) mean response function at  $m_A$ th level of A and  $m_B$ th level of B, and the main effect and interaction functions represent the deviations of other levels from this ‘baseline’ level.

We explicitly write out two generalized functional ANOVA models and take them as example. Suppose the functions are observed at locations  $\mathbf{x}_1, \dots, \mathbf{x}_n$ . Let  $\mathbf{y}_{ijk} = (y_{ijk}(\mathbf{x}_1), \dots, y_{ijk}(\mathbf{x}_n))'$ ,  $\boldsymbol{\mu} = (\mu(\mathbf{x}_1), \dots, \mu(\mathbf{x}_n))'$ ,  $\boldsymbol{\alpha}_i = (\alpha_i(\mathbf{x}_1), \dots, \alpha_i(\mathbf{x}_n))'$ ,  $\boldsymbol{\beta}_j = (\beta_j(\mathbf{x}_1), \dots, \beta_j(\mathbf{x}_n))'$  and  $\boldsymbol{\gamma}_{ij} = (\gamma_{ij}(\mathbf{x}_1), \dots, \gamma_{ij}(\mathbf{x}_n))'$ . When each observation is assumed normal, the two-way functional ANOVA model is

$$\mathbf{y}_{ijk} | \boldsymbol{\mu}_{ij}, \boldsymbol{\theta}_\varepsilon \sim N(\boldsymbol{\mu}_{ij}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}_\varepsilon}), \boldsymbol{\mu}_{ij} = \boldsymbol{\mu} + \boldsymbol{\alpha}_i + \boldsymbol{\beta}_j + \boldsymbol{\gamma}_{ij}, \tag{2.3}$$

where  $\mathbf{y}_{ijk}$  follows a multivariate normal distribution with mean vector  $\boldsymbol{\mu}_{ij}$  and a covariance matrix  $\boldsymbol{\Sigma}_{\boldsymbol{\theta}_\varepsilon}$  that depends on a set of parameters  $\boldsymbol{\theta}_\varepsilon$ . The link function

in (2.1) is the identity link in this Gaussian case. When each observation is binary and assumed to follow a Bernoulli distribution, the functional ANOVA model can be written as

$$\begin{aligned} y_{ijk}(\mathbf{x}_\ell) | \mu_{ij}(\mathbf{x}_\ell) &\sim \text{Bin}(1, \mu_{ij}(\mathbf{x}_\ell)), \\ \text{logit}(\mu_{ij}(\mathbf{x}_\ell)) &= \mu(\mathbf{x}_\ell) + \alpha_i(\mathbf{x}_\ell) + \beta_j(\mathbf{x}_\ell) + \gamma_{ij}(\mathbf{x}_\ell), \end{aligned} \quad (2.4)$$

for  $\ell = 1, \dots, n$ , where  $\text{Bin}(n, \mu)$  denotes the binomial distribution with  $n$  trials and mean  $\mu$ , and  $\text{logit}(\mu) = \log(\mu/(1-\mu))$  is the link function used in this case. Both models are subject to the linear constraints in (2.2).

There are challenges to fitting the generalized functional ANOVA models, such as models (2.3) and (2.4), under a Bayesian framework. One needs to model functional effects, so that the priors taken on them are able to impose sufficient smoothness to remove noise while being flexible enough to work for a variety of  $\mathcal{X}$  domains. There is the issue of computation. The functional data are often large and there are many functions to estimate in the functional ANOVA models. For example, in model (2.3) the sample size is  $nm_Am_B$  if it is a balanced design, and we need to estimate  $(m_A + m_B + m_Am_B)$  functions. Those numbers can grow quickly as we consider more factors. The computational issue is much worse when the functional response is non-Gaussian, as in model (2.4). As well, one needs to make simultaneous inference on the functional effects. It is often desirable to simultaneously test the functional effects to see if they are significant or not. Such tests are difficult to implement from both frequentist and Bayesian perspectives. We look to address the challenges using a novel Bayesian modeling approach.

### 3. Priors on Functional Effects

To make Bayesian inference on model (2.3), we need to specify prior distributions for the functional effects  $\boldsymbol{\mu}$ ,  $\boldsymbol{\alpha}_i$ ,  $\boldsymbol{\beta}_j$  and  $\boldsymbol{\gamma}_{ij}$ , and their hyperparameters. We assume each batch of functions to be independent of the other batches and assign each batch its own set of hyperparameters. The priors we choose for the functions are GMRFs, a class of flexible Gaussian processes that have been extensively used in a variety of applications (e.g., Rue and Held (2005); Yue and Wang (2014)).

A GMRF is a Gaussian random vector  $\mathbf{f} = (f_1, \dots, f_n)'$  with the Markov property: for some  $i \neq j$ ,  $f_i$  and  $f_j$  are independent conditional on other variables  $\mathbf{f}_{-ij}$ . It is defined over a set of discrete indexed locations connected by a graph labelled by  $\mathcal{G}$ . We say  $\mathbf{f}$  is a GMRF with respect to  $\mathcal{G}$  if its density function has

the form

$$\pi(\mathbf{f}) \propto |\mathbf{Q}_\theta|_+^{1/2} \exp\left(-\frac{1}{2}\mathbf{f}'\mathbf{Q}_\theta\mathbf{f}\right), \tag{3.1}$$

where  $\mathbf{Q}_\theta$  is the precision matrix that depends on a set of parameters  $\theta$ , and  $|\mathbf{Q}_\theta|_+$  is the product of non-zero eigenvalues of  $\mathbf{Q}_\theta$ . The Markov property in  $\mathbf{f}$  is conveniently encoded in the matrix  $\mathbf{Q}_\theta$ , with  $\mathbf{Q}_\theta[i, j] = 0$  if and only if  $f_i$  and  $f_j$  are conditionally independent. In most cases only  $\mathcal{O}(n)$  of the  $n^2$  entries of  $\mathbf{Q}_\theta$  are not zeroes, making  $\mathbf{Q}_\theta$  a highly sparse matrix. This allows for a fast Cholesky decomposition of  $\mathbf{Q}_\theta$  as  $\mathbf{L}\mathbf{L}'$ , where  $\mathbf{L}$  is a lower triangular matrix inheriting sparseness from  $\mathbf{Q}_\theta$ . Only non-zero terms in  $\mathbf{L}$  are computed and the nodes can be reordered to decrease the number of the non-zero terms. The typical cost of this decomposition is  $\mathcal{O}(n)$  for one dimension,  $\mathcal{O}(n^{3/2})$  for two dimensions and  $\mathcal{O}(n^2)$  for three dimensions. Using the Cholesky triangle, it is easy to produce random samples from a GMRF, and compute log-density of (3.1) and marginal variances. See Rue and Held (2005) for more technical details. In the following sections, we briefly describe a few GMRF models that are used to analyze the datasets in this paper.

### 3.1. Random walk models

To estimate one-dimensional effect functions, we use so-called the second-order random walk (RW2) model. It is constructed by the independent Gaussian increments

$$f_i - 2f_{i+1} + f_{i+2} \sim N(0, \tau^{-1}),$$

where  $\tau$  is the precision (inverse of variance) parameter. The resulting density of  $\mathbf{f}$  can be shown to have form (3.1) with  $\mathbf{Q}_\theta = \tau\mathbf{R}$ , where  $\mathbf{R}$  is the sparse *structure matrix* reflecting the Markov property of the model. Here  $\mathbf{R}$  has a rank of  $n - 2$ , so the RW2 model has an improper Gaussian density. It can also be viewed as a discretized version of cubic smoothing splines (e.g., Yue, Speckman and Sun (2012)).

The RW2 model can be extended to estimating two-dimensional functions by defining  $\mathcal{G}$  on a regular lattice. The full conditional distributions for the nodes in the interior of the lattice are Gaussian with mean and precision given by

$$\begin{aligned} E(f_{ij}|\mathbf{f}_{-ij}) &= \frac{1}{20} \begin{pmatrix} \circ & \circ \\ \circ & \circ & \bullet & \circ & \circ & \circ & \circ & \circ & \circ \\ \circ & \bullet & \bullet & \bullet & \bullet & \circ & \circ & \circ & \circ \\ \circ & \circ & \bullet & \bullet & \bullet & \bullet & \circ & \circ & \circ \\ \circ & \circ \\ \circ & \circ \\ \circ & \circ \\ \circ & \circ \end{pmatrix}, \\ \text{Prec}(f_{ij}|\mathbf{f}_{-ij}) &= 20\tau, \end{aligned} \tag{3.2}$$

where the locations denoted by ‘•’ represent the neighbors that  $f_{ij}$  depends on,

and the number in front of each grid denotes the weight given to the corresponding ‘•’ locations. We need to apply necessary boundary corrections to (3.2) to fix rank deficiency issue (see Rue and Held (2005) Sec. 3.4.2). It can be proved that this two-dimensional random walk (RW2D) model has a density of form (3.1) and its  $\mathbf{Q}_\theta$  has a rank of  $n - 3$ . Yue and Speckman (2010) showed that the RW2D model is closely related to the thin-plate spline model (e.g., Wahba (1990)).

An important type of spatial data is *areal* data, which are assumed to be the realizations from a function of geographic regions (e.g., the states of the United States) with adjacency information. For the areal data we often define region  $i$  and  $j$  are neighbors if they share a common border, and consider their Gaussian increments  $f_i - f_j \sim N(0, (w_{ij}\tau)^{-1})$ , where  $w_{ij}$  are positive and symmetric weights. We can let  $w_{ij} = 1$  if region  $i$  depends on its neighbors equally. Otherwise, let  $w_{ij}$  be, for instance, the inverse Euclidean distance between region centroids. The resulting full conditional distribution of  $f_i$  is

$$f_i | \mathbf{f}_{-i}, \tau \sim N\left(\frac{\sum_{j \sim i} w_{ij} f_j}{w_{i+}}, \frac{1}{\tau w_{i+}}\right), \quad (3.3)$$

where  $i \sim j$  denotes the set of all unordered pairs of neighbors, and  $w_{i+} = \sum_{j: j \sim i} w_{ij}$  is the summation over neighbors of region  $i$ . We can prove that the density of  $\mathbf{f}$  has form (3.1) with

$$\mathbf{Q}_\theta[i, j] = \tau \begin{cases} w_{i+} & \text{if } i = j, \\ -w_{ij} & \text{if } i \sim j, \\ 0 & \text{otherwise.} \end{cases}$$

Since the sum of each row is zero,  $\mathbf{Q}_\theta$  has rank of  $n - 1$ . We call (3.3) as the *Besag* model because it is closely related to the intrinsic autoregressive model introduced by Besag and Kooperberg (1995).

### 3.2. SPDE models

The random walk models are discrete indexed in nature and thus work best for the data on regularly-spaced locations. For the observations measured at irregularly-spaced locations, a good candidate is the class of Matérn Gaussian fields, which have been extensively used in spatial statistics due to its flexible covariance function between locations  $\mathbf{u}, \mathbf{v} \in \mathbb{R}^2$ ,

$$\text{Cov}(\mathbf{u}, \mathbf{v}) = \frac{\sigma^2}{2^{\nu-1}\Gamma(\nu)} (\kappa\|\mathbf{u} - \mathbf{v}\|)^\nu K_\nu(\kappa\|\mathbf{u} - \mathbf{v}\|),$$

where  $K_\nu(\cdot)$  is the modified Bessel function of the second kind with order  $\nu > 0$ ,

$\Gamma(\cdot)$  is the Gamma function,  $\nu$  is the smoothness,  $\kappa > 0$  is the spatial scale, and  $\sigma^2 > 0$  is the variance.

While a Matérn field is not computationally feasible for large data sets because its covariance matrix is completely dense and therefore difficult to invert, Lindgren, Rue and Lindström (2011) derived an explicit GMRF representation for Matérn Gaussian fields by solving the stochastic partial differential equation (SPDE) given by

$$(\kappa^2 - \Delta)^{\alpha/2} \{\tau f(\mathbf{x})\} = \mathcal{W}(\mathbf{x}), \quad (3.4)$$

where  $\Delta = \sum_{i=1}^d \partial^2 / \partial x_i^2$  is the  $d$ -dimensional Laplacian operator,  $\kappa > 0$  is the spatial scale,  $\alpha > 0$  controls the smoothness of the realizations,  $\tau > 0$  controls the variance, and  $\mathcal{W}$  is the spatial Gaussian white noise. The link to the Matérn smoothness  $\nu$  and variance  $\sigma^2$  is  $\nu = \alpha - d/2$  and  $\sigma^2 = \Gamma(\nu) \{ \Gamma(\alpha) (4\pi)^{d/2} \kappa^{2\nu} \tau^2 \}^{-1}$ . A measure of the spatial range can be empirically derived as  $\sqrt{8\nu}/\kappa$ . Spectral theory shows that an integer  $\alpha$  must be chosen to obtain a Markov field. We thus let  $\alpha = 2$ , resulting in  $\nu = 1$  for the smoothness of a two-dimensional field. Lindgren, Rue and Lindström solve (3.4) using the *finite element method*: they first approximate  $f(\mathbf{x})$  as a piecewise linear basis expansion defined on a triangular mesh, and then turn (3.4) into a system of linear equations to derive a GMRF solution with density of form (3.1) based on an equality in distribution criterion. Given a triangulation the resulting SPDE model is the best piecewise linear approximation to the continuous solution to (3.4) (Lindgren, Rue and Lindström (2011)).

The SPDE models are ideal for functional ANOVA when the data can be taken as the realizations of a spatial function. They provide fast computation and accurate estimation for a variety of spatial functions, whose locations can be either regularly spaced or irregularly spaced, and domains are in one-dimensional, two-dimensional, or even three-dimensional spherical spaces. They are also equipped with the flexibility for being extended for the functions that are non-stationary (Yue et al. (2014); Bolin and Lindgren (2011)), multivariate (Hu et al. (2013)) and spatial-temporal (Cameletti et al. (2013)).

### 3.3. Hyperpriors

For the random walk models there is only one hyperparameter  $\tau$ , which is the precision parameter of the Gaussian increments used to build the models. The prior on  $\tau$  is often selected in an ad-hoc and subjective manner as a diffuse gamma prior, and the same hyperprior is often taken for different random walk

models assigned for different effect functions in one functional ANOVA model. This is a rather important and delicate issue because the value of  $\tau$  controls the smoothness of the effect function and a fixed hyperprior on  $\tau$  does not have the same interpretation for different effect functions. Sørbye and Rue (2014) proposed to scale the models in a way that the  $\tau$ 's in different models have a similar interpretation, making it more reasonable to take the same prior on  $\tau$  across those models. More specifically, they consider marginal variances of  $\mathbf{f}$  given by the diagonal entries of the generalized inverse matrix  $\mathbf{Q}^- = \tau^{-1}\mathbf{R}^-$ , denoted by  $\sigma^2(f_i)/\tau$  for  $i = 1, \dots, n$ . To characterize the level of the marginal variances, the generalized variance of  $\mathbf{f}$  is taken to be the geometric mean

$$\sigma_{GV}^2(\mathbf{f}) = \exp\left(\frac{1}{n} \sum_{i=1}^n \log(\sigma^2(f_i))\right).$$

The random walk models are scaled such that  $\sigma_{GV}^2(\mathbf{f}) = 1$ , which implies that the precision parameters of different models have similar interpretation. The hyperpriors for scaled models are invariant to the shape and size of the graph for a specific model and to different scalings of covariates (Sørbye and Rue (2014)).

For SPDE models, there are hyperparameters  $\kappa$  (spatial scale) and  $\tau$  (precision). Since both parameters influence the marginal variances of the resulting field, it is more reasonable to model the parameters using the standard deviation  $\sigma$  and range  $\rho$ . Specifically, we let  $\log \delta = \log \delta_0 + \theta_1$  and  $\log \rho = \log \rho_0 + \theta_2$ , where  $\delta_0$  and  $\rho_0$  are baseline standard deviation and range values. Based on the explicit mapping between  $(\tau, \kappa)$  and  $(\sigma, \rho)$ , we can have the parameterization  $\log \tau = \log \tau_0 - \theta_1 + \nu \theta_2$  and  $\log \kappa = \log \kappa_0 - \theta_2$ , where

$$\begin{aligned} \log \kappa_0 &= \frac{\log(8\nu)}{2} - \log \rho_0 \\ \log \tau_0 &= \frac{1}{2} \log\left(\frac{\Gamma(\nu)}{\Gamma(\alpha)(4\pi)^{d/2}}\right) - \log \sigma_0 - \nu \log \kappa_0, \end{aligned}$$

and  $\theta_1$  and  $\theta_2$  are assigned a joint normal prior distribution. The hyperparameters are chosen heuristically to suit the specific situation. See Lindgren and Rue (2015) for more details.

#### 4. Posterior Inference

Consider model (2.3) for example. The joint posterior distribution of this generalized two-way functional ANOVA model is given by

$$\pi(\boldsymbol{\mu}, \boldsymbol{\alpha}_i, \boldsymbol{\beta}_j, \boldsymbol{\gamma}_{ij}, \boldsymbol{\theta} | \mathbf{y}) \propto \prod_{ijk} \pi(\mathbf{y}_{ijk} | \boldsymbol{\mu}, \boldsymbol{\alpha}_i, \boldsymbol{\beta}_j, \boldsymbol{\gamma}_{ij}, \boldsymbol{\theta}_\varepsilon)$$

$$\begin{aligned} & \pi(\boldsymbol{\mu}|\boldsymbol{\theta}_\mu) \prod_i \pi(\boldsymbol{\alpha}_i|\boldsymbol{\theta}_\alpha) \prod_j \pi(\boldsymbol{\beta}_j|\boldsymbol{\theta}_\beta) \prod_{ij} \pi(\boldsymbol{\gamma}_{ij}|\boldsymbol{\theta}_\gamma) \\ & \pi(\boldsymbol{\theta}_\varepsilon)\pi(\boldsymbol{\theta}_\mu)\pi(\boldsymbol{\theta}_\alpha)\pi(\boldsymbol{\theta}_\beta)\pi(\boldsymbol{\theta}_\gamma), \end{aligned} \tag{4.1}$$

subject to the linear constraints as (2.2). Here  $\pi(\mathbf{y}_{ijk}|\cdot)$  is the likelihood function from the exponential family,  $\pi(\boldsymbol{\mu}|\cdot)$ ,  $\pi(\boldsymbol{\alpha}_i|\cdot)$ ,  $\pi(\boldsymbol{\beta}_j|\cdot)$  and  $\pi(\boldsymbol{\gamma}_{ij}|\cdot)$  are the GMRF priors taken on the effect functions, and the remaining  $\pi(\cdot)$  functions denote hyperpriors. The standard MCMC approach has been proposed by Kaufman and Sain (2010) to fit a similar model but using Gaussian process priors on the effect functions. With model complexity and large data size, the MCMC simulation procedure is slow and often has mixing problems. We use an approximate Bayesian inference tool based on the integrated nested Laplace approximations (INLA) introduced by Rue, Martino and Chopin (2009). The main benefit of using INLA is computational: where MCMC algorithms need hours or days to run, the approximations provide more precise estimates in seconds or minutes. It can be used to fit a large class of *latent Gaussian models*, of which model (4.1) is a special case. The asymptotic properties of INLA are given in Section 4 of Rue, Martino and Chopin (2009). We briefly describe how we use INLA to approximate the posterior distributions of the functional ANOVA model in (4.1).

#### 4.1. INLA methodology

Let  $\mathbf{y} = (\mathbf{y}'_{111}, \dots, \mathbf{y}'_{11n_{11}}, \dots, \mathbf{y}'_{m_A m_B n_{m_A m_B}})'$  denote an  $n_y$ -dimensional vector of response values,  $\mathbf{f} = (\boldsymbol{\mu}', \boldsymbol{\alpha}'_1, \dots, \boldsymbol{\alpha}'_{m_A}, \boldsymbol{\beta}'_1, \dots, \boldsymbol{\beta}'_{m_B}, \boldsymbol{\gamma}'_{11}, \dots, \boldsymbol{\gamma}'_{m_A m_B})'$  an  $n_f$ -dimensional vector of functional effect parameters, and  $\boldsymbol{\theta} = (\boldsymbol{\theta}'_\varepsilon, \boldsymbol{\theta}'_\mu, \boldsymbol{\theta}'_\alpha, \boldsymbol{\theta}'_\beta, \boldsymbol{\theta}'_\gamma)'$  an  $n_\theta$ -dimensional vector of hyperparameters. Since the GMRF priors are taken on all the effect functions,  $\mathbf{f}$  is another GMRF of form (3.1) with precision matrix  $\mathbf{Q}_\theta$ . This  $\mathbf{Q}_\theta$  is a large but sparse block diagonal matrix, in which the diagonal elements are the precision matrices of all effect functions. Then, the joint posterior distribution in (4.1) can be written as

$$\begin{aligned} \pi(\mathbf{f}, \boldsymbol{\theta}|\mathbf{y}) & \propto \pi(\boldsymbol{\theta}) \pi(\mathbf{f}|\boldsymbol{\theta}) \prod_{i=1}^{n_y} \pi(y_i|f_i, \boldsymbol{\theta}) \\ & \propto \pi(\boldsymbol{\theta})|\mathbf{Q}_\theta|_+^{n_f/2} \exp\left(-\frac{1}{2}\mathbf{f}'\mathbf{Q}_\theta\mathbf{f} + \sum_{i=1}^{n_y} \log \pi(y_i|f_i, \boldsymbol{\theta})\right), \end{aligned}$$

and the posterior marginals of interest can be derived as

$$\pi(f_i|\mathbf{y}) = \int \pi(f_i|\boldsymbol{\theta}, \mathbf{y}) \pi(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}, i = 1, \dots, n_f,$$

$$\pi(\theta_j|\mathbf{y}) = \int \pi(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}_{-j}, j = 1, \dots, n_\theta. \quad (4.2)$$

There are three stages in the implementation of the INLA method: propose an approximation  $\tilde{\pi}(\boldsymbol{\theta}|\mathbf{y})$  to the joint posterior of the hyperparameters  $\pi(\boldsymbol{\theta}|\mathbf{y})$ ; propose an approximation  $\tilde{\pi}(f_i | \boldsymbol{\theta}, \mathbf{y})$  to the marginals of the conditional distribution of  $f_i$  given the data and the hyperparameters  $\pi(f_i|\boldsymbol{\theta}, \mathbf{y})$ ; explore  $\tilde{\pi}(\boldsymbol{\theta}|\mathbf{y})$  on a grid and use it to numerically integrate out  $\boldsymbol{\theta}$  in (4.2).

To approximate  $\pi(\boldsymbol{\theta}|\mathbf{y})$ , we use the Laplace approximation

$$\tilde{\pi}(\boldsymbol{\theta}|\mathbf{y}) \propto \frac{\pi(\mathbf{f}, \boldsymbol{\theta}, \mathbf{y})}{\tilde{\pi}_G(\mathbf{f}|\boldsymbol{\theta}, \mathbf{y})} \Big|_{\mathbf{f}=\mathbf{f}^*(\boldsymbol{\theta})}, \quad (4.3)$$

where  $\tilde{\pi}_G(\mathbf{f}|\boldsymbol{\theta}, \mathbf{y})$  is the Gaussian approximation to  $\pi(\mathbf{f}|\boldsymbol{\theta}, \mathbf{y})$ , obtained by matching the modal configuration and the curvature at the mode  $\mathbf{f}^*(\boldsymbol{\theta})$ . Expression (4.3) yields the exact posterior of  $\boldsymbol{\theta}$  if the likelihood is Gaussian. Since  $\tilde{\pi}(\boldsymbol{\theta}|\mathbf{y})$  is only used to integrate out uncertainty in  $\boldsymbol{\theta}$  when approximating  $\pi(f_i|\mathbf{y})$ , there is no need of a detailed exploration of this density, as long as we are able to select good evaluation points for the numerical integration. Rue, Martino and Chopin (2009) propose two different exploration schemes depending on the number of hyperparameters. For simplicity we assume that  $\boldsymbol{\theta} = (\theta_1, \dots, \theta_{n_\theta}) \in \mathbb{R}^{n_\theta}$  and reparameterize  $\boldsymbol{\theta}$ -space to make the density more regular, as follows. First, find the mode  $\boldsymbol{\theta}^*$  of  $\tilde{\pi}(\boldsymbol{\theta}|\mathbf{y})$  and compute the negative Hessian matrix  $\mathbf{H}$  at the modal configuration. Then standardize  $\boldsymbol{\theta}$  to obtain a new variable  $\mathbf{z} = (\mathbf{V}\boldsymbol{\Lambda}^{1/2})^{-1}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)$ , where  $\mathbf{H}^{-1} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}'$  is the eigen-decomposition. When the dimension of  $\boldsymbol{\theta}$  is small ( $n_\theta \leq 5$ ), the  $\mathbf{z}$ -parametrization is used to build a grid covering the bulk of the probability mass of  $\tilde{\pi}(\boldsymbol{\theta}|\mathbf{y})$ . When  $n_\theta$  is a moderate number, a so-called CCD integration is used, where some ‘points’ in the  $n_\theta$ -dimensional space are found to approximate the unknown function with a second order surface. This requires much less computation and is still able to capture variability in the  $\boldsymbol{\theta}$  space if our purpose is to make inference on  $\pi(f_i|\mathbf{y})$  only.

To approximate  $\pi(\theta_j|\mathbf{y})$  we use the density points that have already been evaluated in the grid exploration of  $\tilde{\pi}(\boldsymbol{\theta}|\mathbf{y})$  to construct an interpolation, based on which the marginals are computed with numerical integration. The so-called asymmetric Gaussian interpolation is used to correct the Gaussian approximation for possible skewness in  $\pi(\theta_j|\mathbf{y})$ . The multi-dimensional numerical integration is unstable when fitting models with a high number of hyperparameters, resulting in approximated posterior marginals densities with undesirable spikes. To relieve this issue, Martins et al. (2013) proposed a new integration-free algorithm, where  $\tilde{\pi}(\theta_j|\mathbf{y})$  is directly approximated by a mixture of Gaussian distributions. It gives

sensible results with almost no extra computation time.

The simplest way to approximate  $\pi(f_i|\boldsymbol{\theta}, \mathbf{y})$  is by using the Gaussian approximation  $\tilde{\pi}_G(f_i|\boldsymbol{\theta}, \mathbf{y})$ , which can be easily obtained from the  $\tilde{\pi}_G(\mathbf{f}|\boldsymbol{\theta}, \mathbf{y})$  already computed in evaluating (4.3). This approximation, however, could yield errors due to the lack of skewness. It can be largely improved by doing a Laplace approximation again

$$\pi_{LA}(f_i|\boldsymbol{\theta}, \mathbf{y}) \propto \frac{\pi(\mathbf{f}, \boldsymbol{\theta}, \mathbf{y})}{\tilde{\pi}_{GG}(\mathbf{f}_{-i}|f_i, \boldsymbol{\theta}, \mathbf{y})} \Big|_{\mathbf{f}_{-i}=\mathbf{f}_{-i}^*(f_i, \boldsymbol{\theta})}, \tag{4.4}$$

where  $\tilde{\pi}_{GG}$  is the Gaussian approximation to the full conditional of  $\mathbf{f}_{-i}$  and  $\mathbf{f}_{-i}^*(f_i, \boldsymbol{\theta})$  is the modal configuration. Unfortunately,  $\tilde{\pi}_{GG}$  must be recomputed for each  $f_i$ , making this approximation too expensive. Rue, Martino and Chopin (2009) proposed two modifications to (4.4) for making it computationally feasible. A third option is called *simplified Laplace* approximation, achieved by doing a third-order Taylor expansion on the numerator and denominator of (4.4), and it thus corrects the Gaussian approximation for location and skewness at a much lower cost (see Rue, Martino and Chopin (2009) for details).

## 4.2. Identification of significant functional effects

### 4.2.1. Contour avoiding set

Let  $h(\mathbf{x})$ ,  $\mathbf{x} \in \mathcal{X} \subset \mathbb{R}^d$ , denote a certain effect function in a functional ANOVA model. It is of interest to test the hypothesis  $H_0 : h(\mathbf{x}) \neq 0$  for all  $\mathbf{x} \in \mathcal{X}$ , in order to find out whether or not the functional effect is statistically significant. From a Bayesian point of view, we need to find a set of  $\mathbf{x}$  such that  $h(\mathbf{x}) \neq 0$  with a posterior probability  $1 - \alpha$  for all  $\mathbf{x}$  in that set. To define this set, we need the definitions regarding excursion and contour uncertainty as given by Bolin and Lindgren (2015). The *positive excursion set* of  $h(\mathbf{x})$  is defined by  $A^+(h) = \{\mathbf{x} \in \mathcal{X}; h(\mathbf{x}) > 0\}$ . The *negative excursion set* is similarly defined by  $A^-(h) = \{\mathbf{x} \in \mathcal{X}; h(\mathbf{x}) < 0\}$ . We then define the pair of *joint contour excursion sets* with probability  $1 - \alpha$  to be

$$\begin{aligned} & (M_\alpha^+(h), M_\alpha^-(h)) \\ & = \arg \max_{(D^+, D^-)} \{ |D^- \cap D^+| : P(D^- \subseteq A^-(h), D^+ \subseteq A^+(h)) \geq 1 - \alpha \}, \end{aligned}$$

where the sets  $(D^+, D^-)$  are open. The union of these two sets  $E_\alpha(h) = M_\alpha^+(h) \cup M_\alpha^-(h)$  is termed the level-zero *contour avoiding set*, which is the largest set such that, for *all*  $\mathbf{x}$  in that set, the effect function  $h(\mathbf{x})$  is different from zero with probability  $1 - \alpha$ . We are not able to see other smaller sets satisfying

the requirement if only  $E_\alpha(h)$  is reported; the set  $E_\alpha(h)$  does not provide any information about the locations not contained in the set. We hence introduce a level-zero *contour avoiding function* defined as  $F(\mathbf{x}) = \sup\{1 - \alpha; \mathbf{x} \in E_\alpha\}$ . It is a graphical tool that is similar to  $p$ -values, the marginal probabilities of exceeding the level, but can be interpreted simultaneously.

We briefly describe the procedures of computing the contour avoiding set, and refer readers to Bolin and Lindgren (2015) for details. There are two problems in the computations. The first is integration: we must calculate probability  $P(D^- \subseteq A^-(h), D^+ \subseteq A^+(h))$  for the pair of sets  $(D^+, D^-)$ . The second is shape optimization: find the largest region satisfying the required probability constraint. In practice the integration is computationally intensive when the dimension of  $\mathbf{h} = (h(\mathbf{x}_1), \dots, h(\mathbf{x}_n))'$  is large (which is usually the case in functional ANOVA). Therefore, we use a method based on a parametric family for the contour sets in combination with a sequential integration for calculating the probabilities. Assume that  $\mathbf{h}$  is Gaussian and  $D(\rho)$  is a parametric family for the possible excursion sets such that  $D(\rho_1) \supset D(\rho_2)$  if  $\rho_1 < \rho_2$ . The following strategy is then used to calculate  $E_\alpha(h)$ . First, find the reordering given by the order that the nodes are added to  $D(\rho)$  when  $\rho$  is increased. Second, sequentially add nodes to  $D$  according to this ordering and update probability  $P(D^- \subseteq A^-(h), D^+ \subseteq A^+(h))$  in each step. Third, stop as soon as the probability falls below  $1 - \alpha$ , and then  $E_\alpha(h)$  is given by the last set  $D$  for which the probability is greater than equal  $1 - \alpha$ . We obtain  $F(\mathbf{x})$  by letting  $\alpha = 1$  and saving the probabilities in each step. The parametric families are based on the marginal quantiles of  $h(\mathbf{x})$ , which can be easily calculated using the marginal posterior distributions. The sequential integration is greatly facilitated by the fact that  $\mathbf{h}$  is a sparse GMRF and can be viewed as a non-homogeneous autoregressive process defined backwards in the indices of  $\mathbf{h}$  (Rue and Held (2005)).

This algorithm is in the Gaussian setting, and can be extended to the latent Gaussian setting with posterior  $\pi(\mathbf{h}|\mathbf{y}) = \int \pi(\mathbf{h}|\mathbf{y}, \boldsymbol{\theta})\pi(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}$ , where the full conditional  $\pi(\mathbf{h}|\mathbf{y}, \boldsymbol{\theta})$  can be non-Gaussian if the likelihood is non-Gaussian. The basic idea to estimate the non-Gaussian posterior distributions is to use Gaussian approximations of either the posterior  $\pi(\mathbf{h}|\mathbf{y})$  or the conditional posterior  $\pi(\mathbf{h}|\mathbf{y}, \boldsymbol{\theta})$ , which can be provided by the INLA method introduced in the previous section. Given estimates of the posterior distributions, we can then compute the excursion probabilities by one of the five methods proposed by Bolin and Lindgren (2015).

### 4.2.2. Simultaneous credible interval

When  $h(x)$  is in one dimension, it is useful to construct the credible interval to visualize the uncertainty of the estimated function. Given an  $x$  a *point-wise* credible interval of  $h(x)$  is given by  $[q_{\rho/2}(x), q_{1-\rho/2}(x)]$ , where  $q_{\rho}(x)$  denotes the  $\rho$ th-quantile of the marginal posterior distribution of  $f(x)$ . The quantiles can be easily computed by INLA because the marginal posterior distributions are approximated as one-dimensional Gaussian mixtures. To identify the significance of this functional effect we need a *simultaneous* credible interval, inside which  $f(x)$  is with probability  $1 - \rho$  for *all*  $x$ . It can be constructed by finding a value of  $\eta$  such that  $P(q_{\eta}(\mathbf{x}) < h(\mathbf{x}) < q_{1-\eta}(\mathbf{x}), \mathbf{x} \in \mathcal{X} | \mathbf{y}) = 1 - \rho$ . Finding such an  $\eta$  requires one to integrate the joint posterior distribution of  $\mathbf{h}$  given  $n$  observations of  $x$ , which is often computationally intensive. Fortunately, one can find this simultaneous credible interval using the method of estimating credible intervals for Gaussian mixtures in Bolin et al. (2015). It is based on the sequential integration method for Gaussian integrals that is described above. We therefore skip the description of the method here.

## 5. Simulation

We conducted a simulation study to evaluate the proposed method. We considered a one-way generalized ANOVA framework, where the mean response function was given by

$$\mu_{ik}(x) = g^{-1}(\mu(x) + \alpha_i(x)), x \in \mathcal{X}, \quad (5.1)$$

for  $i = 1, \dots, m_A$  and  $k = 1, \dots, n$ . For simplicity we let the main effect have only two levels,  $m_A = 2$ . Therefore  $\alpha_2(x) = -\alpha_1(x)$  for any  $x$ , and model (5.1) is a generalized additive model (see, e.g., Wood (2006)). Both normal and binary data were simulated. For normal data, we let the error standard deviation  $\sigma = 0.5$ . Regarding binary data, we used the logistic link function to compute ‘success’ probability as  $P(y_{ik} = 1|x) = 1/\{1 + \exp(-\mu_{ik}(x))\}$ , based on which the data were simulated from independent Bernoulli distributions. There were 500 data sets generated for each type of response variables.

We considered two scenarios with regard to the effect functions. In the first scenario, the effect functions are smooth curves given by

$$\mu(x_{\ell}) = \sin(x_{\ell}), \alpha_1(x_{\ell}) = \frac{\sin(2x_{\ell})}{2}$$

for  $\ell = 1, \dots, 100$ , where  $x_{\ell}$ s are equally spaced within  $[0, 6]$ . At each level of the main effect we simulated  $n = 100$  such curves. The RW2 priors were taken on  $\mu$

and  $\alpha_1$  functions. In the second scenario, we considered two-dimensional effect functions given by

$$\begin{aligned}\mu(x_1, x_2) &= \frac{0.75}{\pi\sigma_{x_1}\sigma_{x_2}} \exp\left(-\frac{1}{\sigma_{x_1}^2}(x_1 - 0.2)^2 - \frac{1}{\sigma_{x_2}^2}(x_2 - 0.3)^2\right) \\ &\quad + \frac{0.45}{\pi\sigma_{x_1}\sigma_{x_2}} \exp\left(-\frac{1}{\sigma_{x_1}^2}(x_1 - 0.7)^2 - \frac{1}{\sigma_{x_2}^2}(x_2 - 0.8)^2\right) \\ \alpha_1(x_1, x_2) &= \frac{0.75}{\pi\sigma_{x_1}\sigma_{x_2}} \exp\left(-\frac{1}{\sigma_{x_1}^2}(x_1 - 0.5)^2 - \frac{1}{\sigma_{x_2}^2}(x_2 - 0.5)^2\right)\end{aligned}$$

where  $\sigma_{x_1} = 0.3$  and  $\sigma_{x_2} = 0.4$ . Both  $x_1$  and  $x_2$  had 40 equally-spaced realizations within  $[0, 1]$ , therefore the effect functions were built on  $40 \times 40$  grid. At each level of the main effect we simulated  $n = 10$  such surfaces. The RW2D priors were taken on  $\mu$  and  $\alpha_1$  functions.

Model (5.1) can be formulated as a generalized additive model (GAM). A GAM, from a frequentist point of view, is usually fitted by penalized likelihood maximization, where the model likelihood is modified by the addition of a penalty for each smooth function to prevent it from being overfitted. A smoothing parameter was used for each penalty to balance the fidelity to data against the smoothness of fit. In `mgcv` R package (Wood (2006)), the smooth functions in a GAM were represented using penalized regression splines (Wood (2003)) with the basis functions designed to be optimal for these splines. The smoothing parameter was estimated by generalized cross validation (GCV) (e.g., Wahba (1990)). For comparison purposes, we also used `mgcv` to fit model (5.1) for both scenarios, and compared it to our method using mean squared error (MSE) of each estimated function for each simulated data set with

$$\text{MSE} = \frac{\sum_{\ell=1}^m \left\{ f(x_\ell) - \hat{f}(x_\ell) \right\}^2}{m},$$

where  $f$  and  $\hat{f}$  denote the true function and its estimator (posterior mean), respectively.

The figures of the distributions of  $\log(\text{MSEs})$  provided by INLA and `mgcv` methods for normal and binary simulated data are given in the supplementary materials of this paper. The INLA method apparently outperforms `mgcv` by yielding overall smaller MSEs for both effect functions in every simulated situation. The MSEs of `mgcv` have more variability than those of INLA. Also, the outperformance of INLA is more significant when it is more difficult to estimate the effect function. For example, in Scenario II it is harder to estimate the grand mean function than the main effect function, and therefore INLA outperforms

`mgcv` more in estimating the former than the latter. We believe the reason that `mgcv` was outperformed by INLA in the simulation study is that the penalized regression splines tend to provide too smooth fits because they use low-rank spline basis functions that are not able to capture the varying features that may exist in the true functions.

## 6. Examples

In this section we present three data examples to demonstrate how to do functional ANOVA using the proposed method. The functional data in the examples differ with respect to their response distributions (normal and non-normal), the dimensions of function domain (1D, 2D and 3D), and the number of potential factors (one-way and two-way). It shows how versatile and flexible our method is. The R code with illustration of those examples is provided in the supplementary document of this paper.

### 6.1. Diffusion tensor imaging

Our first example is a study of functional quantitative imaging metrics of multiple sclerosis (MS) patients over multiple clinical visits (Goldsmith et al. (2012)). The MS is a common chronic demyelinating disease in white matter tracts, which is often investigated by diffusion tensor imaging (DTI) methods. The DTI is a popular quantitative magnetic resonance imaging technique that measures the three-dimensional diffusion of water molecules within tissue through the application of multiple diffusion gradients. Because water diffuses anisotropically in the white matter and isotropically elsewhere, the DTI is used to generate images of the white matter specifically. It allows us to quantitatively measure tissue integrity and thus provide information about damage to the parts of nervous system.

The data in this study consist of 100 subjects, aged between 21 and 70 years at first visit. The number of visits per subject ranged from 2 to 8; a total of 340 visits were recorded. Since we are interested in the visit effect of MS patients, we restricted our attention to the 18 subjects who had completed four visits within approximately one year. For each visit the DTI scans were obtained and used to create tract profiles. The derived outcome is known as fractional anisotropy (FA), which is computed from the diffusion tensor's eigenvalues in the image processing. It is a measure describing the degree of anisotropy of a diffusion process in a brain and has been routinely used as an index of white matter integrity. The region of our interest was the right cortico-spinal tract (RCST),

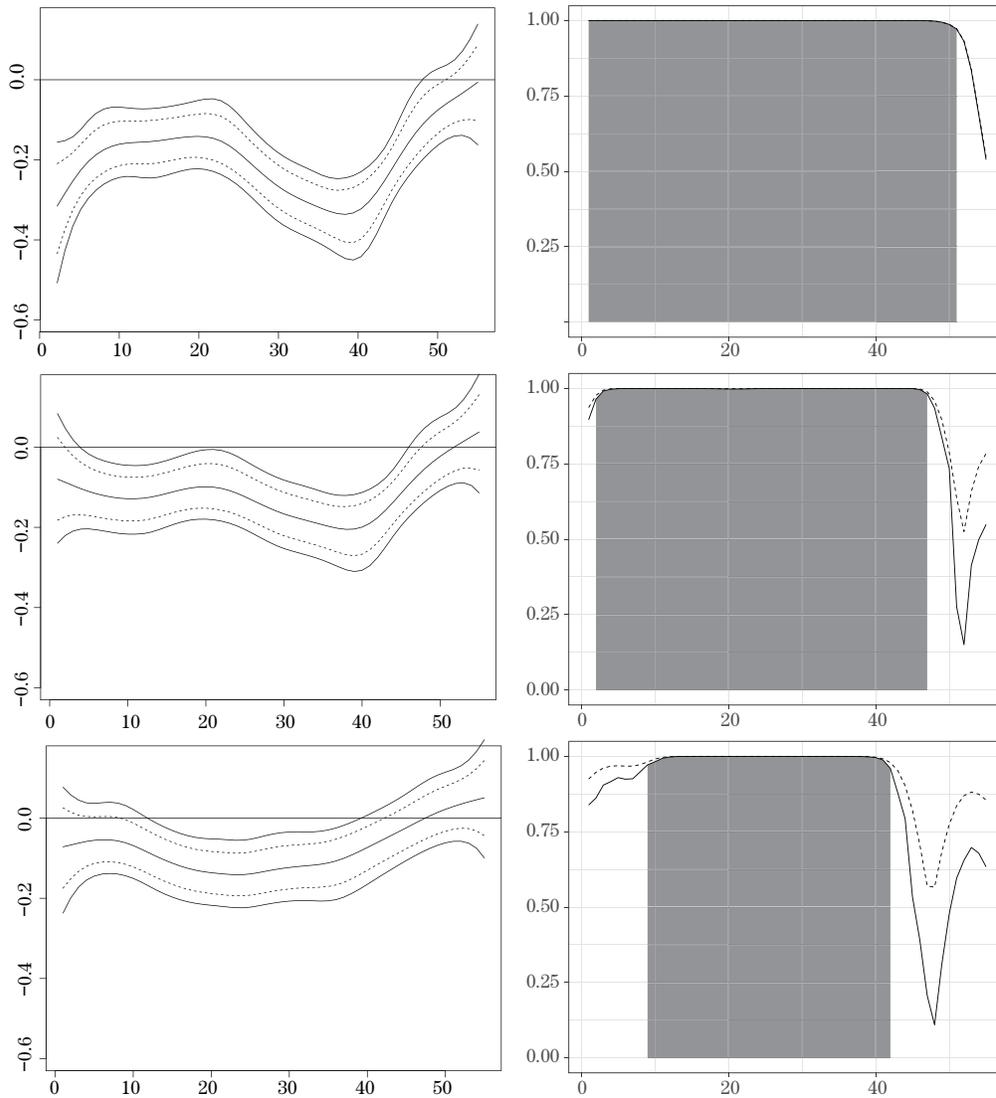


Figure 1. Results of DTI example: The posterior means (solid), point-wise credible bands (dashed) and simultaneous credible bands (solid) of main effect functions (left), and the level-zero contour avoiding functions (solid), marginal probabilities (dashed) and contour avoiding set (grey) of main effect functions (right), for Visit #1 (top), Visit #2 (middle) and Visit #3 (bottom).

an anatomical structure that connects the motor cortex to the brain stem. The damage to the structure can be linked with a decline in cognitive performance in MS patients (Goldsmith et al. (2012)). The data set is available in **refund**

package in R (<http://cran.r-project.org/web/packages/refund/index.html>).

Let  $y_{ij}(x)$  denote the FA measure at trace location  $x$  for  $j$ th subject in  $i$ th visit. Because it is bounded between 0 and 1,  $y_{ij}(x)$  is assumed to follow a beta distribution, which is capable of flexibly modeling this type of data (Wang and Li (2014)). Being interested in the visit effect on FA profiles in the RCST region, we considered the one-way functional ANOVA model

$$y_{ij}(x)|\mu, \alpha_i, \lambda \sim \text{Beta}(p_{ij}(x), \lambda),$$

$$\text{logit}(p_{ij}(x)) = \mu(x) + \alpha_i(x),$$

for  $x \in \mathcal{X} \subset \mathbb{R}$ , where  $\text{Beta}(\mu, \lambda)$  denotes the beta distribution with mean  $\mu$  and variance  $\mu(1 - \mu)/(1 + \lambda)$ . The function  $\mu$  is the grand mean tract profile and  $\alpha_i$  is the main effect function of  $i$ th visit. They are associated with  $p_{ij}$  via logit link. For identifiability we let  $\alpha_4(x) = 0$  for any  $x$ . The independent RW2 priors were taken on  $\mu$  and  $\alpha_i$  for all  $i$ . A diffuse gamma prior was taken on  $\lambda$ .

In Figure 1 we present the results. In the left panel we show the posterior mean, point-wise credible band as well as simultaneous credible band for the main effect function of each visit. We can see that the effect function becomes flatter as the patients pay more visits. This makes sense because the last visit is modeled as baseline. In terms of its dynamic change, the main effect is more significant at the locations in the middle than those at two ends. It indicates that for MS patients the middle part of RCST may be damaged more quickly than the two end parts. In the right panel we show the joint and marginal probabilities that each main effect function is non-zero. Each gray region shows the set of locations such that the main effect function is not zero with at least 0.95 probability for all locations in that region. As one can see, the region keeps shrinking as the patients pay more visits, and the three visits have the common region between 10 and 40. One concludes that the functional effect of ‘visit’ is statistically significant.

## 6.2. Apgar score

There is a study on mortality and live births in Brazil ([http://tabnet.datasus.gov.br/cgi/sinasc/dados/nov\\_indice.htm#docs](http://tabnet.datasus.gov.br/cgi/sinasc/dados/nov_indice.htm#docs)). It is intended to produce indicators and analysis that not only support the planning and evaluation of actions, programs and health services, but also strengthen the use of epidemiological evidence for decision-making. The information of newborns was recorded and the data were published for the period from 1996 to 2012 by the Ministry of Health in Brazil. We chose the dataset of Paraná state in 2012 for demonstration purposes.

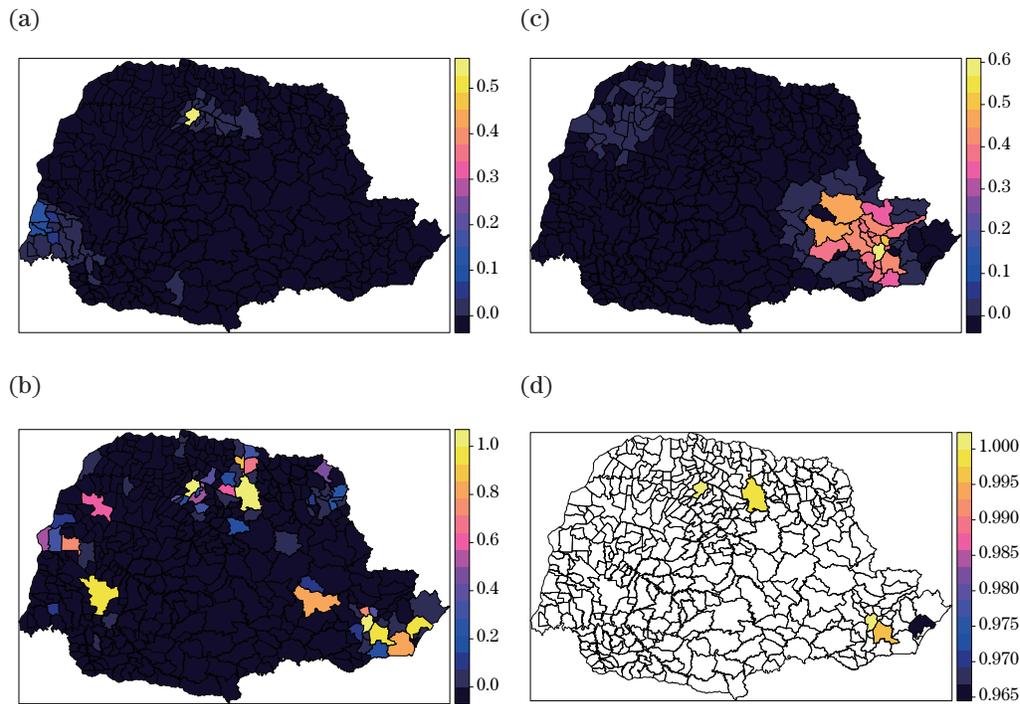


Figure 2. Apgar scores of newborns: Level-zero contour avoiding functions of (a) main effect of consultation; (b) main effect of mother education level; (c) effect of interaction; and 95% level-zero avoiding set of (d) main effect of mother education level.

The response variable of interest is the Apgar score, invented by Virginia Apgar in 1952, which is often used as a method to quickly summarize the health of newborn children. It is determined by evaluating the newborn on five simple criteria (appearance, pulse, grimace, activity, respiration) on a scale from 0 to 2, then summing up the five values thus obtained. Scores 7 and above are generally normal, 4 to 6 fairly low, and 3 and below are regarded as critically low. To identify health status of each newborn, we used a binary response variable  $y = 1$  if the score is less than 7 and  $y = 0$  otherwise. In addition, we associated each score with mother's municipality of residence to make it a functional response with a spatial domain.

The explanatory variables that have potential impact on the Apgar score are weight of the newborn, gestation duration, mother education level, and number of consultations during gestation. The latter three are categorical variables with more than two categories. As for factors, we used the education level and the number of consultations. We made the education level a binary variable with

1 = higher education and 0 otherwise, and the number of consultations with 1 = full consultations ( $\# = 4$ ) and 0 otherwise, for easy interpretation. Then the functional binary response was decomposed according to the main effects of these two factors and their interaction. The weight and gestation duration were taken as linear effects.

As a result, we used the two-way functional ANOVA model

$$\begin{aligned}
 y_{ijk}(\mathbf{x})|p_{ijk} &\sim \text{Bin}(1, p_{ijk}(\mathbf{x})), \mathbf{x} \in \mathcal{X} \subset \mathbb{R}^2, \\
 \text{logit}(p_{ijk}(\mathbf{x})) &= \phi_0 + z_{1ijk}\phi_1 + z'_{2ijk}\phi_2 + \mu_{ij}(\mathbf{x}), \\
 \mu_{ij}(\mathbf{x}) &= \mu(\mathbf{x}) + \alpha_i(\mathbf{x}) + \beta_j(\mathbf{x}) + \gamma_{ij}(\mathbf{x}), \tag{6.1}
 \end{aligned}$$

for  $i = 1, 2, j = 1, 2, k = 1, \dots, n_i$ , where  $\text{Bin}(n, p)$  denotes the binomial distribution with  $n$  trials and success probability  $p$  for each trial,  $\text{logit}(p) = \log(p/(1-p))$ ,  $\phi_0$  is the intercept, and  $\phi_1$  and  $\phi_2$  are linear effects of weight and gestation duration, respectively. Regarding the ANOVA structure,  $\alpha_i$  is the main effect function of the  $i$ th consultation level,  $\beta_j$  is the main effect function of  $j$ th education level, and  $\gamma_{ij}$  is their interaction. Because there are only two levels per factor, we let  $i = 1$  represent the complete consultations,  $i = -1$  the incomplete consultations,  $j = 1$  the higher education, and  $j = -1$  otherwise. Then, we have  $\alpha_i = i\alpha$ ,  $\beta_j = j\beta$  and  $\gamma_{ij} = ij\gamma$ . This reparameterization simplifies the model as well as satisfies the constraints for identifiability. With respect to the priors we assumed (independent) diffuse normal distributions on the linear effects, and the Besag model on the effect functions.

The figure of the posterior means of the grand mean function, main effect function of consultation, main effect function of education level, and their interaction function is given in the supplementary materials. Consultation and interaction effects are spatially smoother but with smaller scale than the education level effect. Figure 2 presents the level-zero avoiding functions for the two main effects and interaction, and the 95% level-zero avoiding set of education level effect. There are five cities identified as having significant effects of education level on the Apgar score; the mother with higher education tends to give birth to a normal baby in those cities. Neither the consultation effect nor the interaction was detected to be significant in any city.

### 6.3. Global solar radiation

Global irradiance data were used to investigate the possible exploitation of solar energy in different areas of the planet. The data were collected at a global spatial lattice, where the longitude ranges from  $-180$  to  $179$  and the

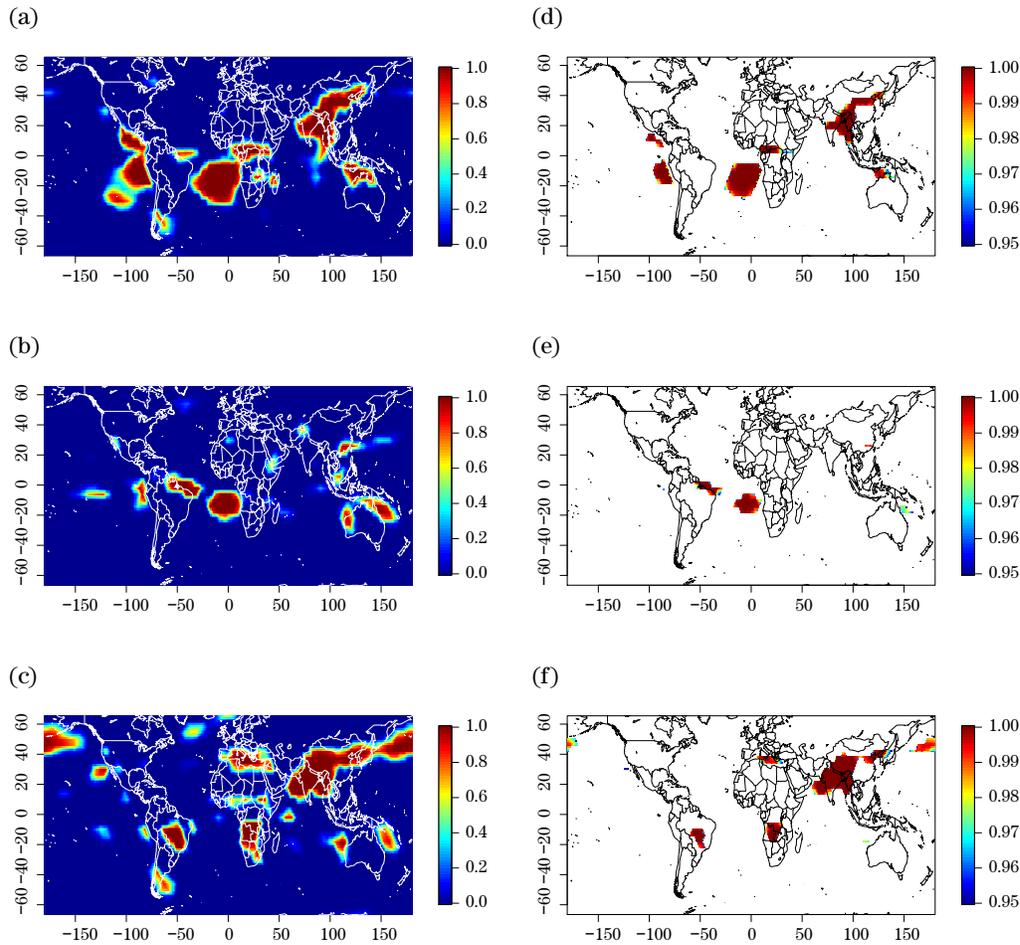


Figure 3. Global solar radiation: level-zero contour avoiding functions (a-c) and the functions thresholded at 95% level (d-f). The seasons are winter (a and d), spring (b and e), and summer (c and f).

latitude from  $-90$  to  $89$  (negative values are south and west; positive values are north and east), yielding totally 64,800 observations of regional averages for each month. The detailed data description can be found at NASA website (<https://eosweb.larc.nasa.gov/sse/>). We focused on monthly averaged insolation incident over a 22-year period (July 1983 - June 2005). Insolation is a measure of solar radiation energy received on a given surface area in a given time. It is often expressed as average irradiance in kilowatt-hours per square meter per day  $\{kWh/(m^2\text{day})\}$ . We considered direct insolation, measured by a surface element perpendicular to the sunbeams at a given location on earth.

One is interested in finding areas that are suitable for installing solar thermal collectors. Because they employ only direct sunlight, the collectors should be positioned in the area, where the direct insolation is constantly high through the whole year and has no strong seasonal pattern. The direct insolation is equal to the solar constant minus the atmospheric losses due to absorption and scattering. The solar constant varies with the earth-sun distance and solar cycles, while the losses depend on the cloud cover and other impurities. We considered the model

$$y_{ij}(\mathbf{x})|\mu_i, \phi_i, \sigma^2 \sim N\left(\mu_i(\mathbf{x}) + \phi_i(z), \sigma^2\right),$$

$$\mu_i(\mathbf{x}) = \mu(\mathbf{x}) + \alpha_i(\mathbf{x}), \mathbf{x} \subset \mathbb{R}^3, z \subset \mathbb{R}, \quad (6.2)$$

for  $i = 1, 2, 3, 4$  and  $j = 1, 2, 3$ . Here  $y_{ij}$  denotes the direct insolation in  $j$ th month of  $i$ th season,  $\phi_i(z)$  denotes the effect of latitude  $z$  for each season,  $\mu(\mathbf{x})$  and  $\alpha_i(\mathbf{x})$  denote the grand mean function and the main effect function of the  $i$ th season, respectively, at the three-dimensional location  $\mathbf{x}$  on the surface of the earth. We expect  $\phi_i$  to capture the seasonal pattern in solar constant while  $\alpha_i$  in atmospheric loss. The seasons were indexed as 1 = winter, 2 = spring, 3 = summer, and 4 = fall. For identifiability we let  $\alpha_4(\mathbf{x}) = 0$  for any  $\mathbf{x}$ , making  $\mu(\mathbf{x})$  the mean function of the fall season, and  $\alpha_i(\mathbf{x})$  ( $i = 1, 2, 3$ ) the difference between  $i$ th season and the fall season. We constructed SPDE priors in spherical space, and took them on the  $\mu$  and  $\alpha_i$  functions. The RW2 priors were assigned for  $\phi_i$  functions.

The figure of the posterior means of baseline mean function and the main effect function for each season is given in the supplementary materials of this paper. Not surprisingly, the seasonal pattern of direct insolation is more significant in winter and summer than in spring and fall, because the radiation should be the same during spring and fall. It seems that Southwest China, India, Central Africa and the west coast of South America tend to receive more direct insolation in winter, while North Africa, South Arica and Brazil do so in summer. Figure 3 presents the contour avoiding function and the largest region where the main effect function is not zero jointly with probability 0.95. This region, which includes Southwest China, North India, Midwest Brazil and a few areas in Africa, is identified with significant seasonal pattern. These plots show how atmospheric losses, which are closely related to the cloud cover, change over the seasons. After comparing the cloud maps produced by the MODIS sensors on NASA's Terra and Aqua satellites (<http://modis.gsfc.nasa.gov/>), we find that over the areas in the identified region the cloud cover shows a clear seasonal pattern that is similar to the one found in our direct insolation data. Based on the results shown above,

south Pacific ocean, Sahara desert, southwestern US and western Australia appear to have constantly high direct insolation and can be used for installing solar thermal collectors.

## 7. Closing Remarks

Functional ANOVA is a powerful tool to investigate how the underlying dynamic patterns in functional data are influenced by categorical factors and their interactions. A variety of approaches for functional ANOVA have been proposed. Bayesian analysis with MCMC sampler has been successfully applied in a few case studies, but their application has never become part of standard analyses.

We illustrate how generalized functional ANOVA models can be solved using the INLA framework that gives fast and accurate approximations. This approach allows us to analyze various types of functional data (both Gaussian and non-Gaussian) with (discrete or continuous) temporal effects or (point-level or areal) spatial effects easily. Both point-wise and simultaneous credible intervals for functional effects can be obtained using the excursion method with a fairly fast and accurate computation. We are able to quantitatively analyze the contribution of each factor over various regions of the functional domain by finding contour avoiding functions at those regions. The provided R code makes such models accessible to a broader audience. In addition, the INLA library provides a much larger number of likelihood functions and latent models than we illustrate here, and thus adds to its flexibility in modeling even more complex functional data.

One-way and two-way ANOVA models are shown here, but the proposed methodology is readily extended where three or more factors are of interest.

The main disadvantage of applying the INLA approach to functional ANOVA, as in other types of latent Gaussian models, is that its computational cost increases exponentially with the number of hyperparameters. It takes INLA a relatively long time to fit a model with a moderate number (e.g., 15) of hyperparameters, due to the expensive cost of evaluating the joint density of those hyperparameters. It is our future work to make INLA feasible for the higher-order functional ANOVA models with many hyperparameters.

## Supplementary Materials

In the supplementary document posted online, we demonstrate how to use R INLA package to perform functional ANOVA step-by-step in the data examples

given in the paper. We also present an extra example that was originally studied in Kaufman and Sain (2010), and show the additional figures that are referred to, but not shown, in the paper.

## Acknowledgment

We are grateful to an associate editor and two reviewers for their valuable suggestions, which substantially improved the paper. For any questions regarding the paper and/or the R code, please contact the corresponding author: Xiao-Feng Wang (wangx6@ccf.org).

## References

- Besag, J. and Kooperberg, C. (1995). On conditional and intrinsic autoregressions. *Biometrika* **82**, 733–746.
- Bolin, D., Guttorp, P., Januzzi, A., Jones, D., Novak, M., Podschwit, H., Richardson, L., Särkkä, A., Sowder, C. and Zimmerman, A. (2015). Statistical prediction of global sea level from global temperature. *Statistica Sinica* **25**, 351–367.
- Bolin, D. and Lindgren, F. (2011). Spatial models generated by nested stochastic partial differential equations, with an application to global ozone mapping. *The Annals of Applied Statistics* **5**, 523–550.
- Bolin, D. and Lindgren, F. (2015). Excursion and contour uncertainty regions for latent Gaussian models. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **77**, 85–106.
- Brumback, B. and Rice, J. (1998). Smoothing spline models for the analysis of nested and crossed samples of curves. *Journal of the American Statistical Association* **93**, 961–976.
- Cameletti, M., Lindgren, F., Simpson, D. and Rue, H. (2013). Spatio-temporal modeling of particulate matter concentration through the spde approach. *AStA Advances in Statistical Analysis* **97**, 109–131.
- Crainiceanu, C. M. and Goldsmith, A. J. (2010). Bayesian functional data analysis using WinBUGS. *Journal of Statistical Software* **32**, 1–33.
- Cuevas, A., Febrero, M. and Fraiman, R. (2004). An ANOVA test for functional data. *Computational Statistics and Data Analysis* **47**, 111–122.
- Faraway, J. J. (1997). Regression analysis for a functional response. *Technometrics* **39**, 254–261.
- Ferraty, F. and Vieu, P. (2006). *Nonparametric Functional Data Analysis: Theory and Practice* Springer.
- Goldsmith, J., Crainiceanu, C. M., Caffo, B. and Reich, D. (2012). Longitudinal penalized functional regression for cognitive outcomes on neuronal tract measurements. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **61**, 453–469.
- Goldsmith, J., Wand, M. P. and Crainiceanu, C. (2011). Functional regression via variational bayes. *Electronic Journal of Statistics* **5**, 572.
- Hu, X., Simpson, D., Lindgren, F. and Rue, H. (2013). Multivariate Gaussian random fields using systems of stochastic partial differential equations. *Arxiv Preprint Arxiv: 1307.1379*.

- Kang, E. L. and Cressie, N. (2013). Bayesian hierarchical ANOVA of regional climate-change projections from NARCCAP phase II. *International Journal of Applied Earth Observation and Geoinformation* **22**, 3–15.
- Kaufman, C. G. and Sain, S. R. (2010). Bayesian functional ANOVA modeling using Gaussian process prior distributions. *Bayesian Analysis* **5**, 123–150.
- Lindgren, F. and Rue, H. (2015). Bayesian spatial modelling with R-INLA. *Journal of Statistical Software* **63**, 1–25.
- Lindgren, F., Rue, H. and Lindström, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach (with discussion). *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **73**, 423–498.
- Martins, T. G., Simpson, D., Lindgren, F. and Rue, H. (2013). Bayesian computing with INLA: New features. *Computational Statistics and Data Analysis* **67**, 68–83.
- Morris, J. S. and Carroll, R. J. (2006). Wavelet-based functional mixed models. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **68**, 179–199.
- Müller, H.-G. (2005). Functional modelling and classification of longitudinal data. *Scandinavian Journal of Statistics* **32**, 223–240.
- Ramsay, J. O. and Dalzell, C. J. (1991). Some tools for functional data analysis. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **53**, 539–572.
- Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis*. 2nd Edition Springer.
- Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields: Theory and Applications*, Volume 104 of *Monographs on Statistics and Applied Probability*. London: Chapman & Hall.
- Rue, H., Martino, S. and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations (with discussion). *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **71**, 319–392.
- Shen, Q. and Faraway, J. (2004). An F test for linear models with functional responses. *Statistica Sinica* **14**, 1239–1258.
- Sørbye, S. H. and Rue, H. (2014). Scaling intrinsic Gaussian Markov random field priors in spatial modelling. *Spatial Statistics* **8**, 39–51.
- Wahba, G. (1990). *Spline Models for Observational Data*. Philadelphia: SIAM [Society for Industrial and Applied Mathematics].
- Wang, X.-F. and Li, Y. (2014). Bayesian inferences for beta semiparametric-mixed models to analyze longitudinal neuroimaging data. *Biometrical Journal* **56**, 662–677.
- Wang, X.-F., Yang, Q., Fan, Z., Sun, C.-K. and Yue, G. H. (2009). Assessing time-dependent association between scalp EEG and muscle activation: A functional random-effects model approach. *Journal of Neuroscience Methods* **177**, 232–240.
- Wood, S. N. (2003). Thin plate regression splines. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **65**, 95–114.
- Wood, S. N. (2006). *Generalized Additive Models: An Introduction with R*. Chapman and Hall/CRC Press.
- Yue, Y., Simpson, D., Lindgren, F. and Rue, H. (2014). Bayesian adaptive smoothing spline using stochastic differential equations. *Bayesian Analysis* **9**, 397–424.
- Yue, Y., Speckman, P. and Sun, D. (2012). Priors for Bayesian adaptive spline smoothing. *Annals of the Institute of Statistical Mathematics* **64**, 577–613.

- Yue, Y. and Speckman, P. L. (2010). Nonstationary spatial Gaussian Markov random fields. *Journal of Computational and Graphical Statistics* **19**, 96–116.
- Yue, Y. and Wang, X.-F. (2014). Spatial Gaussian Markov random fields: Modelling, applications and efficient computations. *Journal of Biometrics and Biostatistics* **5**, e128. doi: 10.4172/2155-6180.1000e128.
- Zhang, J.-T. and Chen, J. (2007). Statistical inferences for functional data. *The Annals of Statistics* **35**, 1052–1079.
- Zhang, J.-T. and Liang, X. (2014). One-way ANOVA for functional data via globalizing the pointwise F-test. *Scandinavian Journal of Statistics* **41**, 51–71.

One Bernard Baruch Way, Box B11-220, New York, NY 10010, USA.

E-mail: Yu.Yue@baruch.cuny.edu

Mathematical Statistics, Chalmers University of Technology, SE-412 96 Göteborg, Sweden.

E-mail: david.bolin@chalmers.se

MailBox 4318, King Abdullah University of Science & Technol (KAUST), P.O.Box 4700, Postal Code 23955-6900 Thuwal - Jeddah, Saudi Arabia.

E-mail: Haavard.Rue@kaust.edu.sa

Department of Quantitative Health Sciences, Cleveland Clinic, 9500 Euclid / JJN3-01, Cleveland, OH 44195, USA.

E-mail: wangx6@ccf.org

(Received January 2016; accepted September 2017)