Bayesian Nonparametric Regression and Density Estimation Using Integrated Nested Laplace Approximations

Xiao-Feng Wang*

Department of Quantitative Health Sciences/Biostatistics Section, Cleveland Clinic Lerner Research Institute, Cleveland, OH 44195, USA

Abstract

Integrated nested Laplace approximations (INLA) are a recently proposed approximate Bayesian approach to fit structured additive regression models with latent Gaussian field. INLA method, as an alternative to Markov chain Monte Carlo techniques, provides accurate approximations to estimate posterior marginals and avoid time-consuming sampling. We show here that two classical nonparametric smoothing problems, nonparametric regression and density estimation, can be achieved using INLA. Simulated examples and R functions are demonstrated to illustrate the use of the methods. Some discussions on potential applications of INLA are made in the paper.

Keywords: Nonparametric regression; Density estimation; Approximate Bayesian inference; Integrated nested Laplace approximations; Markov chain Monte Carlo

INLA and Approximate Bayesian Inference

Structured additive regression models have been extensively used in many fields, such as medicine, public health, and economics. In these models, the response variable y_i is assumed to belong to an exponential family and its mean μ_i is linked to a structured additive predictor η_i through a link function $g(\cdot)$, where

$$g(\mu_i) = \eta_i = \alpha_0 + \sum_{k=1}^{n_\alpha} \alpha_k x_{ki} + \sum_{l=1}^{n_f} f_l(u_{li}) + \varepsilon_i. \qquad (1)$$

Here α_0 is the intercept, α_k's are the linear effects of covariates x_k's, $f_l(\cdot)$'s are unknown functions of the covariates μ_l's, and ε is unstructured error term. In Bayesian statistics, the common solution to obtain estimates for the models is the use of Markov chain Monte Carlo (MCMC) techniques. Although it is always possible to implement them in theory, MCMC methods could come with a wide range of issues in practice: parameter samples could be highly correlated, computational time is very long, and estimates may content large Monte Carlo errors, etc.

Approximate Bayesian inference using integrated nested Laplace approximations (INLA) is a recently proposed method for solving the structured additive regression models where the latent field of the models is Gaussian [1]. The methodology is particularly useful, when the latent Gaussian model is a Gaussian Markov random field (GMRF) with precision matrix G controlled by a few hyperparameters. Denote that β is the vector of all unknown regression parameters in the structured additive predictor (1), assigned Gaussian priors; ξ is the vector of hyperparameters for which non-Gaussian priors are assigned in the model. In a Bayesian framework, one is to estimate the marginal posterior density

$$\pi(\beta_j, y) = \int \pi(\beta_j, \xi | y) d\xi = \int \pi(\beta_j | \xi, y)\pi(\xi | y) d\xi. \qquad (2)$$

given data y for each component β_j of the Gaussian vector β. The key of INLA is to construct a nested approximation of (2). First, the marginal posterior $\pi(\xi | y)$ is approximated using

$$\tilde{\pi}(\xi | y) \propto \left. \frac{\pi(\beta, \xi, y)}{\tilde{\pi}(\beta | \xi, y)} \right|_{\beta = \beta^*(\xi)},$$

where $\tilde{\pi}(\beta | \xi, y)$ is the Gaussian approximation to the full conditional distribution of β evaluated in the mode $\beta^*(\xi)$ for a given ξ. Then, one computes $\tilde{\pi}(\beta_j | \xi, y)$, an approximation of the posterior $\pi(\beta_j | \xi, y)$. Rue et al. [1] suggested three approximation approaches: a Gaussian approximation, a full Laplace approximation, and a simplified Laplace approximation. Lastly, INLA approach combines the previous two steps with a numerical integration to reach the goal. The approximate posterior of $\pi(\beta_j | y)$ is obtained by

$$\tilde{\pi}(\beta_j | y) = \sum_q \tilde{\pi}(\beta_j | \xi_q, y)\tilde{\pi}(\xi_q | y)\Delta_q,$$

where the sum is over values of ξ with area weights Δ_q.

INLA method provides accurate approximations to the posterior marginals of the latent variables which are extremely fast to compute. No samples of the posterior marginal distributions need to be drawn using INLA, so it is a computationally convenient alternative to MCMC. All computations required by the INLA methodology have been efficiently implemented by a R package INLA (available on www.r-inla.org). It integrates GMRFLib, a C-based library for fast and exact simulation of GMRF.

Rue et al. [1] demonstrated that INLA can be applied to solve a variety of popular statistical models, which includes generalized linear mixed models, stochastic volatility models, spatial disease mapping, and Log-Gaussian Cox processes. In this paper, we show that two classical nonparametric smoothing problems, nonparametric regression and probability density estimation, can be achieved by using INLA. Simulated examples and R functions are demonstrated to illustrate the use of the methods. The future research on the applications of INLA to other complex statistical models is discussed in the paper.

*Corresponding author: Xiao-Feng Wang, Department of Quantitative Health Sciences, Cleveland Clinic Lerner Research Institute, 9500 Euclid Avenue/JJN3, Cleveland, OH 44195, USA; E-mail: wangx6@ccf.org

Nonparametric Regression Using INLA

Nonparametric regression is a classical problem in statistics, which does not require to specify a parametric functional form for the relationship between the response y and the predictor x. We assume a model of the form

$$y_i = m(x_i) + \varepsilon_i, \qquad i = 1, \ldots, n,$$

where $m(\cdot)$ is an unknown smooth function and the errors ε_i's are assumed to be independent and identically distributed as $N(0, \sigma^2)$.

There are several popular techniques to estimate nonparametric regression models, such as, local polynomial regression methods [2], and smoothing spline approaches [3]. Here we examine the use of spline techniques. Smoothing spline is motivated by considering the penalized residual sum of squares as a criterion for the goodness of fit of $m(x)$ to the data. It minimizes

$$S(\lambda) = \sum_{i=1}^{n} (y_i - m(x_i))^2 + \lambda \int m''(x)^2 dx,$$

where λ is a smoothing parameter to control the trade-off between fit and the penalty $\int m''(x)^2 dx$. The solution to this minimization problem is a natural cubic spline with knots at the distinct observed values of X_i.

Wahba [4] showed that the smoothing spline is equivalent to Bayesian estimation with a partially improper prior. $m(x)$ has the prior distribution which is the same as the distribution of the stochastic process

$$Z(x) = \theta_0 + \theta_1 x + b^{1/2} V(x),$$

where $\theta_0, \theta_1 \sim N(0, \varsigma)$, $b = \sigma^2 / \lambda$ is fixed, and $V(x)$ is the one-fold integrated Wiener process,

$$V(x) = \int_0^x (x-t) dW(t).$$

Thus, estimating $m(x)$ becomes to seek the solution to the stochastic differential equation

$$\frac{d^2 m(x)}{dx^2} = \lambda^{-1/2} \sigma \frac{dW(x)}{dx}, \tag{3}$$

as a prior over m. Note that such a differential equation of order two is the continuous-time version of a second-order random walk. However, the solution of (3) does not have any Markov properties. The precision matrix is dense, hence it is computationally intensive. Lindgren and Rue [5] suggested a Galerkin approximation to $m(x)$, as the solution of (3). To be specific, let $x_1 < x_2 < \ldots < x_n$ be the set of (unequal-spaced) fixed points, a finite element representation of $m(x)$ is constructed as

$$\tilde{m}(x) = \sum_{i=1}^{n} \psi_i(x) w_i,$$

for the piecewise linear basis functions ψ_i's and random weights w_i's.

In order to estimate the smooth function $m(x)$, one needs to determine the joint distribution of the weights $w = (w_1, \ldots, w_n)^T$. Using the Galerkin method, w is derived as a GMRF with mean zero and precision matrix G. Let $d_i = x_{i+1} - x_i$ for $i = 1, \ldots, n-1$ and $d_{-1} = d_0 = d_n = d_{n+1} = \infty$.

The $n \times n$ symmetric matrix G is defined as

$$G = \begin{bmatrix}
g_{11} & g_{12} & g_{13} & & & & \\
g_{21} & g_{22} & g_{23} & g_{24} & & \mathbf{0} & \\
g_{31} & g_{32} & g_{33} & g_{34} & g_{35} & & \\
\ddots & \ddots & \ddots & \ddots & \ddots & & \\
& & g_{n-2,n-4} & g_{n-2,n-3} & g_{n-2,n-2} & g_{n-2,n-1} & g_{n-2,n} \\
\mathbf{0} & & & g_{n-1,n-3} & g_{n-1,n-2} & g_{n-1,n-1} & g_{n-1,n} \\
& & & & g_{n,n-2} & g_{n,n-1} & g_{n,n}
\end{bmatrix},$$

where the non-zero elements of row i are given by

$$g_{i,i-2} = \frac{2}{d_{i-2} d_{i-1} (d_{i-2} + d_{i-1})},$$

$$g_{i,i-1} = \frac{-2}{d_{i-1}^2} \left(\frac{1}{d_{i-2}} + \frac{1}{d_i} \right),$$

$$g_{i,i} = \frac{2}{d_{i-1}^2 (d_{i-2} + d_{i-1})} + \frac{2}{d_{i-1} d_i} \left(\frac{1}{d_{i-1}} + \frac{1}{d_i} \right) + \frac{2}{d_i^2 (d_i + d_{i+1})},$$

with $g_{i,i+1} \equiv g_{i+1,i}$ and $g_{i,i+2} \equiv g_{i+2,i}$ due to symmetry. G is a sparse matrix with rank $n-2$, making the model computationally effective.

If we assign $\tilde{m}$ as a smoothness prior over m, the cubic smoothing spline $\hat{m}(x)$ at x coincides with the posterior expectation of $m(x)$ give the data, i.e. $\hat{m}(x) \approx E(m(x)|y)$. Therefore, the nonparametric regression problem becomes to fit a latent Gaussian model. It can be accomplished using INLA since w is a GMRF. The implementation of the method needs some extra programming in R to define a user-specified GMRF. A R function called npr.INLA is listed in Appendix. The following simple code is to call the fit using INLA, where "x" and "y" are inputs of numeric vector of predictors and responses, respectively.

R> library (INLA)

R> library (splines)

R> fit1<-npr.INLA (x,y)

The outputs of npr.INLA include a vector of sorted x values at which the estimate is computed, a vector of smoothed estimates for the regression at the corresponding x, and two vectors of the corresponding 2.5% and 97.5% credible bands for the regression.

We now show two simulated examples to compare the INLA approach and the conventional cubic smoothing spline regression. Generalized cross-validation method is applied to select the smoothing parameters for smoothing splines. The Bayesian method takes the advantage that the smoothing parameter is automatically determined by the model fitting without any user input. In the first example, $m(x) = 3\sin(2.5x) + 2\exp(-5x^2)$, $x \sim U(0, 1.5)$, and $\varepsilon \sim N(0, 0.5^2)$ with sample size $n=50$. In the second example, $m(x) = 3x^{2/5} - \cos(\pi x)$, $x \sim U(-3.0, 0)$, and $\varepsilon \sim N(0, 0.8^2)$ with sample size $n=100$. Figure 1 shows the estimation results. The estimates using INLA are denoted by the solid lines, and the corresponding 95% credible bands are denoted by the dashed lines. The estimates using the cubic smoothing spline regression are denoted by the dash-dotted lines. The true functions are denoted by the dotted lines. Unsurprisingly, the estimates using INLA are almost identical to those using cubic smoothing spline. The 95% credible bands from the INLA fits completely cover the true functions in these two examples. Approximate Bayesian inference through INLA
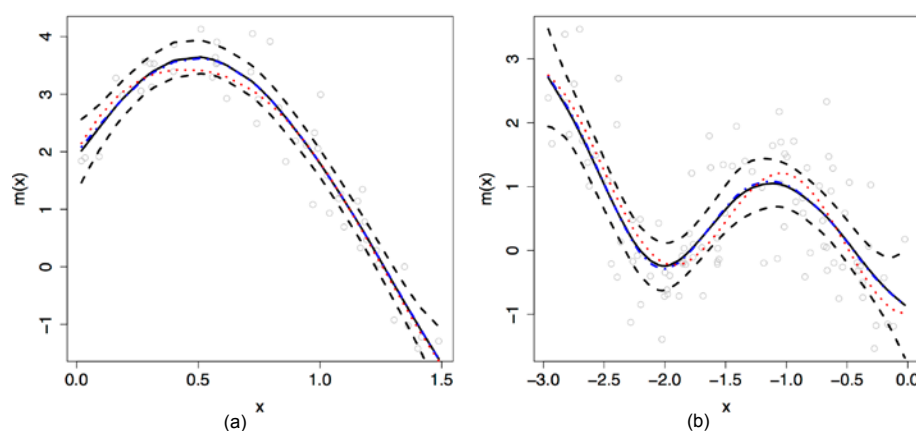
Figure 1: Simulated examples for nonparametric regression: (a) $m(x) = 3\sin(2.5x) + 2\exp(-5x)$, $n = 50, \sigma = 0.5$; (b) $m(x) = x^{2/5} - \cos(\pi x)$, $n = 100, \sigma = 0.8$. The estimates (solid lines) with credible bands (dashed lines) using INLA are compared to the estimates (dash-dotted lines) using cubic smoothing spline regression. The true functions are denoted by the dotted lines.
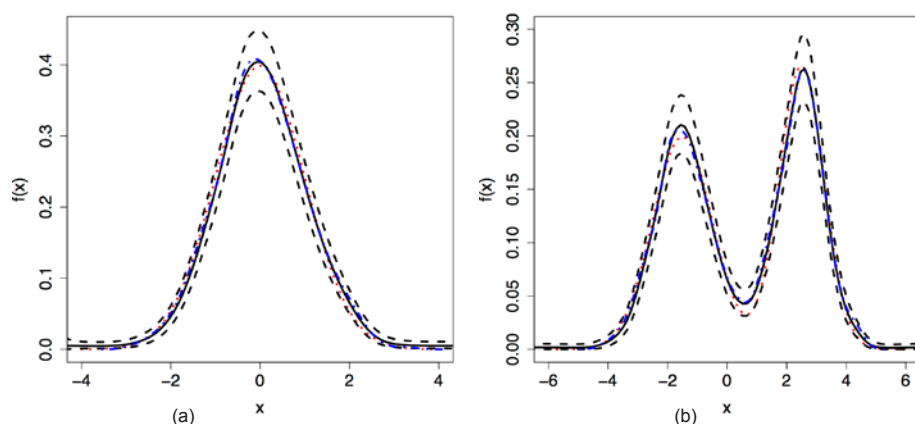

Figure 2: Simulated examples for density estimation: (a) $X \sim N(0,1), n = 500$; (b) $X \sim 0.5N(-1.5,1) + 0.5N(2.5,0.75^2), n = 1000$. The estimates (solid lines) using INLA are compared to the estimates (dash-dotted lines) using kernel density estimation with Sheather and Jones [8]'s plug-in bandwidth. The true functions are denoted by the dotted lines.

allows fast Bayesian computation and makes it possible to perform analysis in an automatic way.

Density Estimation Using INLA

Nonparametric density estimation can be also implemented using INLA. Brown et al. [6] proposed a "root-unroot" density estimation procedure, which turned density estimation into a nonparametric regression problem. The regression problem was created by binning the original observations into suitable size of bins and applying a mean-matching variance stabilizing root transform to the binned data counts. Then, a wavelet block thresholding regression was used to obtain the density estimate. Here we adopt Brown et al. [6]'s root-unroot procedure but use a second-order random walk model with INLA for the regression step. The second-order random walk model is particularly suitable for an equi-spaced nonparametric time series regression problem [7]. In addition, there are two advantages to use the Bayesian nonparametric approach. First, we avoid the smoothing

parameter selection, where the smoothness of curve is automatically determined by the Bayesian model fitting. Second, it is straightforward to construct credible bounds of a regression curve from INLA output. As a result, constructing credible bands for the probability density function becomes a natural by-product in the density estimation. Let $\{X_1, ..., X_n\}$ be a random sample from a distribution with the density function f_x. The estimation algorithm is summarized as follows.

1. Poissonization. Divide $\{X_1, ..., X_n\}$ in T equal length intervals. Let $C_1, ..., C_T$ be the count of observations in each of the intervals.

2. Root Transformation. Apply the mean-matching variance stabilizing root transform, $Y_j = \sqrt{C_j + 1/4}, j = 1, ..., T$.

3. Bayesian Smoothing with INLA. Consider the time series $Y = (Y_1, ..., Y_T)$ to be the sum $Y_j = m_j + \varepsilon_j, j = 1, ..., T$ of a smooth trend function m and a noise component   . Fit a second-order random walk model with INLA for the equi-spaced time series to obtain an posterior

mean estimate $\hat{m}$ of m, and $\alpha/2$ and $1-\alpha/2$ quantiles, $\hat{m}_{\alpha/2}$ and $\hat{m}_{1-\alpha/2}$.

4. Unroot Transformation and Normalization. The density function f_x is estimated by

$$\hat{f}_X(x) = \gamma[\hat{m}(x)]^2,$$

and the $100(1-\alpha)\%$ credible bands of $f(x)$ is

$$(\gamma[\hat{m}_{\alpha/2}(x)]^2, \gamma[\hat{m}_{1-\alpha/2}(x)]^2)$$

where $\gamma = (\int \hat{f}_X dx)^{-1}$ is a normalization constant.

The R function, density.INLA in Appendix implements the above root-unroot algorithm. The following code is to call the density fit using INLA, where "x" is a numeric vector of data values, "m" is the number of equally spaced points at which the density is to be estimated, "from" and "to" are the left and right-most points of the grid at which the density is to be estimated. If "from" and "to" are missing, "from" equals to the minima of the data values minus "cut" times the range of the data values and "to" equals to the maxima of the data values plus "cut" times the range of the data values.

```
R> library (INLA)

R> fit1 <-density.INLA (x, m=101, from=min (x), to=max (x), cut=0.1)
```

Figure 2 shows the two simulated examples to compare INLA method and conventional kernel density estimation. In the first example data were generated from the standard normal distribution, $X \sim N(0,1)$, with sample size $n=500$; In the second example data were generated from a normal mixture model, $X \sim 0.5N(-1.5,1) + 0.5N(2.5, 0.75^2)$, with sample size $n=1000$. In the figure, the estimates using INLA are denoted by the solid lines, and the 95% credible bands are denoted by the dashed lines. The kernel density estimates with Sheather and Jones [8]'s plug-in bandwidth are denoted by the dash-dotted lines. The true functions are denoted by the dotted lines. We note that the INLA estimates are very close to the kernel density estimates. The INLA approach allows us to compute the credible bands of the density function without additional computational efforts.

Discussion

MCMC techniques were commonly choices to fit structured additive regression models, however they may often suffer the issues of convergence and computational time. INLA approach provides a novel approximate Bayesian inference method to fit a large class of structured additive models with latent Gaussian field. It, as an alternative to MCMC, provides accurate approximations to estimate posterior marginals and avoid time-consuming sampling. INLA method has been implemented in C and a R-interfaced package is available under Linux, Windows and Macintosh. The package provides a user-friendly interface and makes latent Gaussian models applicable in a general way.

We have shown that two classical nonparametric models can be fit through the approximate Bayesian inferential procedure. Nonparametric regression is treated within a general second-order random walk model by assigning appropriate GMRF priors. Density estimation is transformed to an equally-spaced nonparametric Bayesian time series regression problem by adapting Brown et al. [6]'s root-unroot algorithm. The approximate Bayesian inference of the nonparametric models enjoys the advantage of fast computation, automatic selection of smoothing parameter and construction of credible bands.

Recently, Lindgren et al. [9] have addressed the explicit link between Gaussian fields and Gaussian Markov random fields through the stochastic partial differential equation approach. INLA approach has been applied to a variety of complex models, such as spatial-temporal disease mapping [10], additive mixed quantile regression models [11], and beta semiparametric mixed models [12]. INLA, as new Bayesian computation tools, have a great potential to be used in many scientific fields. There are many open problems of applying INLA to advanced statistical models. For instance, it would be of interest to use INLA for joint modeling of longitudinal and time-to-event data, functional data analysis, measurement error models, and sparse ultra-high-dimensional modeling. Such problems need further investigation and should be detailed in the future.

Appendix: R Functions

We present here the key compute code to call INLA for nonparametric regression and density estimation. Simulation examples and code files can be obtained from the journal's webpage or http://filer.case.edu/xxw17/Software/npINLA/.

```
## Nonparametric regression using INLA

npr.INLA <- function(x, y, diagonal = 1e-03, constr = T, ...){

    if (any(is.na(x))) stop("'x' contains missing values!")

    if (any(is.na(y))) stop("'y' contains missing values!")

    if (length(x) != length(y)) stop("'x' and 'y' have different lengths!")

    y <- y[order(x)]

    x <- sort(x)

    B <- bs(x, degree = 1, intercept = TRUE)

    idx <- 1:length(x)

    Q <- Gmatrix(x)

    inla.fit <- inla(y ~ B.1 + B.2 + f(idx, model = "generic0", Cmatrix = Q,

        diagonal = diagonal, constr = constr, ...) - 1,

        data = as.data.frame(list(y = y, idx = idx, B = B)),

        control.predictor = list(compute = T))

        return(structure(list(x = x, y = inla.fit$summary.linear.predictor[,1],

        y.lower = inla.fit$summary.linear.predictor[,3],

        y.upper = inla.fit$summary.linear.predictor[,5])))

}


## Density estimation using INLA

density.INLA <- function(x, m = 101, from, to, cut = 0.1,

        diagonal = 1e-03, constr = T, ...){

    if (any(is.na(x))) stop("'x' contains missing values!")
```

```
if (missing(from)) from <- min(x) - cut * diff(range(x))

if (missing(to)) to <- max(x) + cut * diff(range(x))

bins=seq(from, to, length.out = m)

x.bins <- hist(x, breaks = bins, plot=FALSE)

x.bins.root <- sqrt(x.bins$counts+1/4)

idx <- 1:length(x.bins.root)

Q <- Gmatrix(idx)

 inla.fit <- inla(x.bins.root ~ f(idx, model = "generic0", Cmatrix
= Q,

    diagonal = diagonal, constr = constr),

    data = as.data.frame(list(x.bins.root = x.bins.root, idx = idx)),

    control.predictor = list(compute = T))

inla.est <- inla.fit$summary.linear.predictor[,1]

inla.lower <- inla.fit$summary.linear.predictor[,3]

inla.upper <- inla.fit$summary.linear.predictor[,5]

SimpsonInt <- function (x, f, subdivisions = 256){

    ap <- approx(x, f, n = 2 * subdivisions + 1)

    integral <- diff(ap$x)[1] * (ap$y[2 * (1:subdivisions) - 1]

      + 4 * ap$y[2 * (1:subdivisions)] + ap$y[2 * (1:subdivisions)
+ 1])/3

    return(sum(integral))

    }

normalized <- SimpsonInt(x.bins$mids, inla.est^2)

f <- inla.est^2/normalized

f.lower <- inla.lower^2/normalized

f.upper <- inla.upper^2/normalized

  return(structure(list(x = x.bins$mids, y = f, y.lower= f.lower,
y.upper=f.upper)))

}


## Function to compute G matrix

Gmatrix <- function(x, sparse = TRUE){

if (any(is.na(x))) stop("'x' contains missing values!")

x <- sort(x)

n <- length(x)

d <- diff(x)

d <- c(Inf, Inf, d, Inf, Inf)

k <- 3:(n + 2)

g <- (2/((d[k - 1]^2) * (d[k - 2] + d[k - 1]))

    + 2/(d[k - 1]*d[k]) * (1/d[k - 1] + 1/d[k])

    + 2/((d[k]^2) * (d[k] + d[k + 1]))))

k <- 4:(n + 2)

g1 <- -2/(d[k - 1]^2) * (1/d[k - 2] + 1/d[k])

k <- 5:(n + 2)

g2 <- 2/(d[k - 2] * d[k - 1] * (d[k - 2]+d[k - 1]))

G <- diag(g)

G[row(G) == col(G) + 1] <- g1

G[col(G) == row(G) + 1] <- g1

G[row(G) == col(G) + 2] <- g2

G[col(G) == row(G) + 2] <- g2

if (sparse == TRUE) G <- as(G, "dgTMatrix")

return(G)

}
```

References

1. Rue H, Martino S, Chopin N (2009) Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. J R Stat Soc Series B Stat Methodol 71: 319-392.

2. Fan J, Gijbels I (1996) Local Polynomial Regression. Chapman and Hall, London. UK.

3. Gu C (2002) Smoothing Spline ANOVA Models. Springer Verlag, New York, USA.

4. Wahba G (1978) Improper priors, spline smoothing and the problem of guarding against model errors in regression. Journal of the Royal Statistical Society. Series B (Methodological) 40: 364-372.

5. Lindgren F, Rue H (2008) On the second-order random walk model for irregular locations. Scand Stat Theory Appl 35: 691-700.

6. Brown L, Cai T, Zhang R, Zhao L, Zhou H (2010) The root-unroot algorithm for density estimation as implemented via wavelet block thresholding. Probability theory and related fields 146: 401-433.

7. Fahrmeir L, Knorr-Held L (2000) Dynamic and semiparametric models. In Smoothing and Regression: Approaches, Computation, and Application (MG Schimek edn), Wiley, New York, USA.

8. Sheather SJ, Jones MC (1991) A reliable data-based bandwidth selection method for kernel density estimation. Journal of the Royal Statistical Society. Series B (Methodological) 53: 683-690.

9. Lindgren F, Rue H, Lindström J (2011) An explicit link between Gaussian fields and Gaussian markov random fields: the stochastic partial differential equation approach. J R Stat Soc Series B Stat Methodol 73: 423-498.

10. Schrödle B, Held L (2011) Spatio-temporal disease mapping using INLA. Environmetrics 22: 725-734.

11. Yue YR, Rue H (2011) Bayesian inference for additive mixed quantile regression models. Comput Stat Data Anal 55: 84-96.

12. Wang XF, Li Y (2013) Bayesian inferences for beta semiparametric mixed models to analyze longitudinal neuroimaging data. In Review.