On Bayesian P-splines in disease mapping

A. Adin[1], M.D. Ugarte[2], T. Goicoa[3]

[1]aritz.adin@unavarra.es, Department of Statistics and O.R and Institute for Advanced Materials (INAMAT), Public University of Navarre, Spain.
[2]lola@unavarra.es, Department of Statistics and O.R and Institute for Advanced Materials (INAMAT), Public University of Navarre, Spain.
[3]tomas.goicoa@unavarra.es, Department of Statistics and O.R and Institute for Advanced Materials (INAMAT), Public University of Navarre, Spain.

Abstract

In recent years, models incorporating splines have been considered for smoothing mortality risks in disease mapping. Although these models are very flexible, they can be computationally expensive when analyzing spatio-temporal data. In this work, alternative models are proposed to avoid the dimension of the three-dimensional B-spline basis. Pancreatic cancer mortality data in continental Spain during the period 1988-2012 will be used for illustration purposes.

Keywords: INLA; pancreatic cancer; space-time disease mapping.

1. Introduction

In recent years, models incorporating splines have been considered for smoothing mortality risks in spatio-temporal disease mapping as an alternative to conditional autoregressive (CAR) models. A sensible approach consists in using CAR distributions for spatial random effects and B-splines for temporal smoothing [1]. Very recently, Lee and Durbán [2] consider two-dimensional P-splines with B-spline bases for spatial count data. These authors also propose three-dimensional P-splines to smooth ozone levels in space and time [3], whereas [4, 5] use three-dimensional P-splines to smooth risks in space and time. These P-splines models have been embedded within a generalized linear mixed model (GLMM) framework and model fitting and inference has been carried out using the well-known penalized quasi-likelihood (PQL) technique [6]. From a fully Bayes approach, P-splines have been also used to smooth risks in spatio-temporal disease mapping [7]. Although these models are quite flexible, they rely on Markov chain Monte Carlo (McMC) methods for model fitting and inference, with the inconvenience of large computing time. Very recently, integrated nested Laplace approximations (INLA) [8] have been proposed for Bayesian inference to reduce computational burden.

In this work, spatially structured (or unstructured) one-dimensional temporal P-splines as well as temporally structured (or unstructured) two-dimensional spatial P-splines are proposed to smooth risks and to avoid three-dimensional P-splines of large dimension. The spatial and temporal correlation will be specified giving appropriate prior distributions to the basis coefficients. The models will be fitted using INLA.

2. P-splines models for spatio-temporal count data

Let us assume that the region under study is divided into n contiguous small areas labeled as $i = 1, \ldots, n$ and data are available for several time periods $t = 1, \ldots, T$. Then, conditional to the relative risks r_{it}, the number of counts in each area and time period, O_{it}, is assumed to be Poisson distributed with mean $\mu_{it} = E_{it} r_{it}$, where E_{it} represent the number of expected cases for area i and time t. Namely

$$O_{it}|r_{it} \sim Poisson(\mu_{it} = E_{it} r_{it}) \quad \text{and} \quad \log r_{it} = \log E_{it} + \log r_{it}.$$

Depending on the specification of $\log r_{it}$, different models are defined.

2.1. Spatially structured temporal P-splines

An spatially structured temporal P-spline model is defined as

$$\log r_{it} = \eta + \xi_i + f(x_t) + f_i(x_t) \quad \text{for} \quad i = 1, \ldots, n; \quad t = 1, \ldots, T,$$

where η quantifies the logarithm of the global risk, ξ_i is a spatially structured random effect, $f(x_t)$ is a temporal smooth function common to all areas and $f_i(x_t)$ is a spatially structured temporal smooth function specific for each area. That is, temporal trends from neighbouring regions tend to be similar [7]. As suggested by [9], the Leroux et al. CAR prior [10] is considered for the spatial effects ξ_i

$$\boldsymbol{\xi} \sim N\left(\mathbf{0}, \sigma_s^2(\lambda_s \mathbf{Q}_s + (1-\lambda_s)\mathbf{I}_s)^{-1}\right),$$

where λ_s is a spatial smoothing parameter taking values between 0 and 1, $\mathbf{I}_s$ is an $n \times n$ identity matrix, and $\mathbf{Q}_s$ is the spatial neighborhood matrix. The common temporal trend is specified as $f(\mathrm{x}_t) = \mathbf{B}_t \boldsymbol{\theta}_t$, where a random walk prior of first (RW1) or second (RW2) order is considered for $\boldsymbol{\theta}_t$, and $\mathbf{B}_t$ is the temporal B-spline basis of dimension $T \times k$ (with k depending on the number of knots and the degree of the B-spline basis). Finally, the spatially structured temporal trend is defined as $f_i(\mathrm{x}_t) = \mathbf{B}_{st} \boldsymbol{\theta}_{st}$, where $\mathbf{B}_{st} = \mathbf{I}_n \otimes \mathbf{B}_t$ is a block-diagonal matrix of dimension $nT \times nk$. Spatial correlation is included through a CAR prior distribution on the coefficients $\boldsymbol{\theta}_{st} = (\theta_{11}, \ldots, \theta_{1k}, \ldots, \theta_{n1}, \ldots, \theta_{nk})'$ such that

$$\boldsymbol{\theta}_{\cdot j} = (\theta_{1j}, \ldots, \theta_{nj}) \sim N(\mathbf{0}, \sigma_{st}^2 \mathbf{Q}_s^-) \quad \text{for} \quad j = 1, \ldots, k.$$

If spatial correlation on the coefficients is ignored, then the area-specific temporal trend will vary randomly. Depending on the prior for neighbor coefficients in space and in time, different models arise resembling the four types of interaction models defined in Knorr-Held [11].

2.2. Temporally structured spatial P-splines

Similarly, a temporally structured spatial P-spline model is defined as

$$\log r_{it} = \eta + f(x_{1i}, x_{2i}) + \gamma_t + f_t(x_{1i}, x_{2i}) \quad \text{for} \quad i = 1, \ldots, n; \quad t = 1, \ldots, T,$$

where here $f(x_{1i}, x_{2i})$ is a smooth surface constant along the time periods, γ_t is a temporally structured random effect and $f_t(x_{1i}, x_{2i})$ is a temporally structured spatial smooth function specific for each time point. Here, random walks priors, RW1 or RW2, are considered for the temporal effects γ_t, defined as

$$\boldsymbol{\gamma} \sim N(\mathbf{0}, \sigma_t^2 \mathbf{Q}_t^-),$$

where $\mathbf{Q}_t$ is the structure matrix of a RW1/RW2 and the symbol $^-$ denotes the Moore-Penrose generalized inverse of a matrix. The constant spatial smooth surface is specified as $f(\mathbf{x}_1, \mathbf{x}_2) = \mathbf{B}_s\boldsymbol{\theta}_s$, where $\mathbf{B}_s = \mathbf{B}_2 \Box \mathbf{B}_1$ is the two-dimensional B-spline basis of dimension $n \times K$ (with $K = k_1 k_2$ depending on the number of knots and the degree of the marginal B-spline basis) obtained from the row-wise tensor product of marginal B-spline bases for longitude $\mathbf{x}_1 = (x_{11}, \dots, x_{1n})^{'}$ and latitude $\mathbf{x}_2 = (x_{21}, \dots, x_{2n})^{'}$. Finally, the temporally correlated spatial smooth function is defined as $f_t(\mathbf{x}_1, \mathbf{x}_2) = \mathbf{B}_{st}\boldsymbol{\theta}_{st}$, where $\mathbf{B}_{st} = \mathbf{I}_t \otimes \mathbf{B}_s$ is a block-diagonal matrix of dimension $nT \times KT$. Temporal correlation is included through a RW1/RW2 prior on the coefficients $\boldsymbol{\theta}_{st} = (\theta_{11}, \dots, \theta_{K1}, \dots, \theta_{1T}, \dots, \theta_{KT})^{'}$ such that

$$\boldsymbol{\theta}_{j\cdot} = (\theta_{j1}, \dots, \theta_{jT}) \sim N(\mathbf{0}, \sigma_{st}^2 \mathbf{Q}_t^-) \quad \text{for} \quad j = 1, \dots, K.$$

3. Illustration

Pancreatic cancer mortality data in continental Spain during the period 1988-2012 will be considered to illustrate these models. The INLA approach will be used for model fitting and inference. It provides accurate approximations to the posterior marginals of the quantity of interest in relatively short computational time (see for example [12]). The models can be implemented in `R` using the package R-INLA.

4. Acknowledgments

This research has been supported by the Spanish Ministry of Science and Innovation (project MTM 2011-22664, jointly sponsored with FEDER grants, and project MTM2014-51992-R), and by the Health Department of the Navarre Government (project 113, Res.2186/2014). The authors would like to thank the Spanish National Epidemiology Center (area of Environmental Epidemiology and Cancer) for providing the data.

5. Bibliography

[1] MacNab, Y.C., and Dean, C. (2001). Autoregressive spatial smoothing and temporal spline smoothing for mapping rates. *Biometrics*, **57**(3), 949 – 956.

[2] Lee, D.-J. and Durbán, M. (2009). Smooth-CAR mixed models for spatial count data. *Computational Statistics and Data Analysis*, **53**, 2968-2979.

[3] Lee, D.J., and Durbán, M. (2011). P-spline ANOVA-type interaction models for spatio-temporal smoothing. *Statistical Modelling*, **11**(1), 49 – 69.

[4] Ugarte, M.D., Goicoa, T., and Militino, A.F. (2010). Spatio-temporal modeling of mortality risks using penalized splines. *Environmetrics*, **21**(3-4), 270 – 289.

[5] Ugarte, M.D., Goicoa, T., Etxeberria J., and Militino, A.F. (2012). A P-spline ANOVA type model in space-time disease mapping. *Stochastic Environmental Research and Risk Assessment*, **26**, 835 – 845.

[6] Breslow, N.E., and Clayton, D.G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, **88**, 9 – 25.

[7] MacNab, Y.C. (2007). Spline smoothing in Bayesian disease mapping. *Environmetrics*, **18**, 728 – 744.

[8] Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society, Series B*, **71**(2), 319 – 392.

[9] Ugarte, M.D., Adin, A., Goicoa, T., and Militino, A.F. (2014). On fitting spatio-temporal disease mapping models using approximate Bayesian inference. *Statistical Methods in Medical Research*, **23**(6), 507 – 530.

[10] Leroux, B.G., Lei, X., and Breslow, N. (1999). Estimation of disease rates in small areas: A new mixed model for spatial dependence. In Halloran, M. and Berry, D. editors, *Statistical Models in Epidemiology, the Environment, and Clinical Trials*, pages 179 – 192. Springer-Verlag: New York.

[11] Knorr-Held, L. (2000). Bayesian modelling of inseparable space-time variation in disease risk. *Statistics in Medicine*, **19**(17-18), 2555 – 2567.

[12] Schrödle, B., Held, L., Riebler, A. and Danuser, J. (2011). Using integrated nested Laplace approximations for evaluation of the veterinary surveillance data from Switzerland: a case study. *Journal of the Royal Statistical Society, Series C*, **60**(2), 261 – 279.