Spline smoothing in Bayesian disease mapping

Ying C MacNab\textsuperscript{1} and Paul Gustafson\textsuperscript{2}

\textsuperscript{1} Division of Epidemiology and Biostatistics, Department of Health Care and Epidemiology, University of British Columbia, Centre for Healthcare Innovation and Improvement, British Columbia Research Institute for Children’s and Women’s Health, Room E417, 4480 Oak Street, Vancouver, BC, Canada V6H 3V4.

\textsuperscript{2} Department of Statistics, University of British Columbia

Abstract: In this paper, a class of Bayesian hierarchical disease mapping models with spline smoothing are motivated and developed for sequential disease mapping and for surveillance of disease risk trends and clustering. The methodological development aims to provide reliable information about the patterns (both over space and time) of disease risk and to quantify uncertainty. Bayesian disease mapping models with B-splines, smoothing splines and P-splines are developed respectively and a comparison of the three smoothing methods in the context of risks ensemble prediction is presented. The methods are illustrated through a Bayesian analysis of iatrogenic injuries to hospital in-patients in British Columbia, Canada.

Keywords: Bayesian Disease mapping; Semiparametric models; Adaptive regression B-spline; Smoothing spline; P-spline.

1 The Bayesian disease mapping model framework

Recent development of Bayesian disease mapping methodology has been motivated by the need to address statistical issues of assessing the geographic variation of crude rates or relative risks that were frequently displayed in thematic maps. These crude rates and relative risks are often subject to large chance variation when rare diseases are investigated in small areas. Literature on Bayesian disease mapping presents mixed effects Poisson models that are characterized as ‘spatial smoothing’. The methods assume ‘spatially varying’ or randomly varying relative risks and the associated joint prior probability for ‘pooling data’ and ‘borrowing strength’. When spatio-temporal disease rates are available for sequential risk mapping of several time periods, the ‘smoothing’ issue may be explored by considering spatial smoothing, temporal smoothing and spatio-temporal interaction. In this paper, these considerations are motivated and explored through development of a class of semi-parametric mixed effects models.
with spline smoothing. The methodological development aims to provide reliable information about the patterns (both over space and time) of disease risks and to quantify uncertainty.

Let $y_{it}$ denote disease occurrence and $n_{it}$ the ‘at risk’ population for the $i$th local region at time $t$; $i = 1, ..., N$ and $t = t_1, ..., t_T$. To facilitate spatial and temporal smoothing concurrently for spatio-temporal disease mapping, we discuss three-stage hierarchical semiparametric models and assume that conditioning on a vector $b$ of random effects the data $\{y_{it}\}$ arise independently from Poisson family with intensity $\mu_{it}$:

$$
\log(\mu_{it}) = \log(n_{it}) + a_0 + S_0(t) + b_{0i} + RS_i(t).
$$

Here, $\log(n_{it})$ is an offset; $a_0 + S_0(t)$ and $b_{0i} + RS_i(t)$ are smoothing functions: with $t$ being centered, $m = \exp(a_0)$ represents mid-period ‘global’ rate, $b_{0i}$ the mid-period random area effect, $a_0 + S_0(t)$ the ‘global’ rate trend, and \{$b_{0i} + RS_i(t)$\} the regional risk trend relative to the ‘global’ rates\[1\].

2 Temporal smoothing with splines

In this study we present three spline methods for fitting of the smoothing functions $S_0$ and $RS_i$: an adaptive regression B-spline, smoothing/natural spline, and P-spline. From a Bayesian perspective, a simple and computationally efficient adaptive regression B-spline method is proposed so that the smoothing functions are fit by mixed-effects regression in terms of the B-splines; the resulting splines are piecewise polynomials with a minimal number of interior knots.

Specifically, we assume a regression cubic B-spline for the arbitrary smoothing function $S_0(t)$ and a family of regression cubic B-splines for $RS_i(t)\[1\]: $\sum_{k=1}^{K} a_k B_k(t)$ and $\sum_{k=1}^{K} b_{ik} B_k(t)$, where $\{a_k, k = 1, ..., K\}$ are fixed effects; $\{B_k\}$, $k = 1, 2, ..., K$, is a set of basis functions (without the intercept) for a $K$-dimensional space $Sp(t_1, ..., t_L, 3)$ of B-splines of degree 3, $(t_1, ..., t_L)$ denoting $L$ pre-specified inner knots, $B_k(t)$ denoting the $k$th B-spline basis function evaluated at time $t$, $K = L + 3$; $b_{ik} = (b_{ik1}, ..., b_{iKN})^\top$ are random spline coefficients. We illustrate a simple and computationally efficient method for knot selection as well as model selection using recently developed Bayesian hierarchical model selection criterion, the Deviance Information Criterion (DIC). In particular, we begin with B-spline of one interior knot at mid-period and pay specific attention to the assessment of temporal smoothing and goodness-of-fit for B-splines of increasing number of inner knots; the knots are positioned uniformly over the range of time $t$. Knot and model selections are based on examining the corresponding deviance residuals, deviance, and deviance information criterion (DIC). We note that the $NK$-vector of regression parameters $b = (\beta_{10}, ..., \beta_{N0}, ..., \beta_{1K}, ..., \beta_{NK})^\top$ represents a random effects ensemble that governs the spatio-temporal risks.
smoothing via specifications of the second-stage prior. In this study, we explore several prior options which accommodate spatially varying as well as randomly varying smoothing functions for the modeling of temporal risk trends[2]. We also explore and examine several hyperprior choices.

In the regression spline approach, the use of a small number of uniformly distributed knots over the data domain (i.e., the design points) may have limited flexibility in the resulting smoothing, as the small and discrete number of knots may facilitates only limited control over the fit and smoothness. The smoothing spline [3] and P-spline[4], on the other hand, are penalty splines with a relatively large number of knots. With recent work on mixed-effects model representation of the smoothing splines [5] and P-splines [6], additive fittings of smoothing splines and P-splines can be formulated and implemented within the generalized linear mixed model (GLMM) framework; the corresponding Bayesian justification and formulation for fully Bayesian implementation can also be developed.

Both the smoothing spline and P-spline approaches can be readily casted into the GLMM context and the ensuing Bayesian framework for disease mapping. For example, the smoothing spline GLMM formulation of model (1) can be written as

$$\log(\mu_{it} | b) = \log(n_{it}) + u_0 + u_1 t + \sum_{k=2}^{T-1} u_k C_k(t), \quad \nu_{i0} + \nu_{i1} t + \sum_{k=2}^{T-1} \nu_{ik} C_k(t), \quad (2)$$

where \((u_0, u_1)\) are fixed effects, \(u = (u_2, ..., u_{T-1})^\top\) are random effects, \(u \sim N(0, \sigma_u^2 I_{T-1})\), and \(\sigma_u^2\) is the smoothing parameter; \((\nu_{i0}, \nu_{i1})\) and \(\nu_i = (\nu_{i2}, ..., \nu_{iT-1})^\top\) are random effects, \(\nu_i \sim N(0, \sigma_i^2 I_{T-1})\), and \(\sigma_i^2\) is the smoothing parameter, \(i = 1, ..., N\); \(C = (C_k(t))\) is the corresponding \(T\) by \(T - 2\) design matrix for the random effects, \(C = L(L^\top L)^{-1}, L = QU^{-1}, U\) is the Choleski decomposition of \(R, R\) and \(Q\) are the two band matrices defined in [3].

Similarly, the P-spline GLMM formulation of model (1) can be written as

$$\log(\mu_{it} | b) = \log(n_{it}) + \alpha_0 + \alpha_1 t + \sum_{k=2}^{T} \alpha_k A_k(t) + \beta_{i0} + \beta_{i1} t + \sum_{k=2}^{T} \beta_{ik} A_k(t), \quad (3)$$

where \((\alpha_0, \alpha_1)\) are fixed effects, \(\alpha = (\alpha_2, ..., \alpha_T)^\top\) are random effects, \(\alpha \sim N(0, \sigma_\alpha^2 I_T)\), \(\sigma_\alpha^2\) is the smoothing parameter; \((\beta_{i0}, \beta_{i1})\) and \(\beta_i = (\beta_{i2}, ..., \beta_{iT})^\top\) are random effects, \(\beta_i \sim N(0, \sigma_i^2 I_T)\), \(\sigma_i^2\) is the smoothing parameter, \(i = 1, ..., N\); \(A = (A_k(t))\) is the corresponding \(T\) by \(T - 1\) design matrix for the random effects, \(A = BD^\top (DD^\top)^{-1}, B\) is the corresponding B-spline bases matrix, \(D\) is the difference matrix defined in [4].

For both the smoothing spline and P-spline, we explore several second-stage and third-stage prior choices.
3 Results and conclusion

The Bayesian disease mapping models with spline smoothing were applied to the analysis of hospital admission/separation data on non-fatal iatrogenic injuries among male patients 1-19 years of age in 84 local health areas in the Province of British Columbia (BC), Canada. For the B-spline models, the DIC results favored the non-spatial prior and a randomly varying B-splines ensemble for risk estimation; the results also showed that the 4-knot model offered best balance between goodness-of-fit and smoothness. The analysis suggested that the 1- and 2-knot models provided sufficient flexibility to expose gradually changing temporal trends. A sensitivity analysis on several hyperprior alternatives indicated considerable Bayesian robustness with respect to the rate and risk inference. The temporal risk smoothing unveiled important temporal risk trends and spatial risk clustering and informed on the magnitude of the iatrogenic injury problem at the LHA level.

For comparable prior and hyperprior specifications, the smoothing spline and P-spline methods led to similar posterior prediction and inference of the relative risks ensemble. Both methods, however, were shown to be sensitive to hyperprior specification. Specifically, hyperprior specifications of the random effects variances were shown to influence the degree of smoothing in the resulting rate/risk splines, an indication that the data provided limited information about the smoothing parameters. In particular, for ‘lesser informed’ hyperprior, the penalty splines were shown to impose considerably lesser degree of smoothing on the rates/risks and have noticeably higher posterior uncertainty (i.e. larger posterior standard deviations and wider credible intervals).

Acknowledgments: We gratefully acknowledge the support of the Canadian Institute for Health Research (CIHR MOP-42498) and the Natural Sciences and Engineering Research Council of Canada (NSERC RG-238660).

References


