Biometrics , 1–19
September 2019

Approximate Bayesian Inference for Case-Crossover Models

Alex Stringer* and Patrick Brown

Department of Statistical Sciences, University of Toronto and

Centre for Global Health Research, St. Michael's Hospital

**email:* alex.stringer@mail.utoronto.ca

and

Jamie Stafford

Department of Statistical Sciences, University of Toronto

Summary: A case-crossover analysis is used as a simple but powerful tool for estimating the effect of short-term environmental factors such as extreme temperatures or poor air quality on mortality. The environment on the day of each death is compared to the one or more "control days" in previous weeks, and higher levels of exposure on death days than control days provides evidence of an effect. Current state-of-the-art methodology and software (INLA) cannot be used to fit the most flexible case-crossover models to large datasets, because these models violate the condition imposed by INLA that observations be independent conditional on the latent variables. In this paper we develop a flexible and scalable modelling framework for case-crossover models with linear and semi-parameteric effects which retains the flexibility and computational advantages of INLA, but accomodates a broad class of models that violate this conditional independence restriction. We apply our method to quantify non-linear associations between mortality and extreme temperatures in India. An R package implementing our methods will be released publicly.

Key words: Additive models; Bayesian inference; Conditional logistic regression; Case-crossover; INLA.

1. Introduction

1.1 *Motivation and examples*

When studying the association of mortality with short-term exposures to risk factors, for example spikes in air pollution or extreme temperatures, the most readily-available data are usually daily death counts and daily exposure measurements. Inference about the relative risk of mortality at various exposures is based on the difference between subjects' exposures at time of death to their exposures at these previous times. This is especially cost effective in the common situation where the risk factors under consideration are factors that are already being recorded as part of large administrative databases available to the researcher, such as temperature or air pollution levels.

Maclure (1991) introduced the case-crossover design as a tool for inferring associations between mortality and exposure using such retrospectively sampled, within-subject measurements. They used the case-crossover to study the hypothesis that incidence of myocardial infarctions (heart attacks) are influenced by short-term exposure to risk factors such as smoking and heavy eating. Using the case-crossover design, this was achieved using only information from patients already admitted to the hospital having experienced a myocardial infarction. In a different setting, Redelmeier and Tibshirani (1997) utilized case-crossover models for quantifying association between cell phone use and motor vehicle accidents. The authors were able to quantify this association by simply contacting invidiuals who had been in an accident and asking them about their cell phone use at the time of crashing compared to previous crash-free days.

More recently, Fu et al. (2018) considered the effect of short-term exposure to both extreme and moderate temperatures on mortality in India using a large, nationally-representitive dataset. The case-crossover design here enabled the authors to associate exposure both to extremely hot or to moderately cold temperatures with an increase in mortality risk.

However, their spline-based approach to modelling nonlinear associations was very sensitive to the number and placement of knots and had unpredictable behaviour at very high and low temperatures, which were the temperature ranges of primary inferential interest. Our methodology addresses these concerns, which we will demonstrate in §5 with a novel analysis of their data.

1.2 *Proposed research*

Existing inference methodologies used for fitting case-crossover models are able to handle moderately-sized datasets and focus mainly on linear associations between mortality and exposure. However, modern datasets used for these studies are often very large, and associations between exposure and mortality are often non-linear. There is a need in practice for a flexible, scalable framework for fitting case-crossover models to modern datasets. MCMC methods are flexible in theory, but they do not scale efficiently with dataset size or model complexity and hence are often prohibitively slow. The Integrated Nested Laplace Approximations (INLA; Rue et al. (2009)) framework provides a flexible and scalable modelling framework, but works only for a limited class of models in which observations are conditionally independent given the latent variables. The case-crossover model, and other complex models such as those for survival and aggegrated spatio-temporal data, violate this restriction and hence cannot be fit using INLA. In certain cases, ad-hoc methods exist for avoiding this restriction, for example the analysis of survival data (Martino et al., 2011). In this paper, we develop an inference methodology that removes this conditional independence restriction completely while retaining the flexibility and computational advantages of INLA, and is hence able to fit a very flexible class of case-crossover models to large datasets.

1.3 *Outline of paper*

This paper is organized as follows. In §2 we describe the case-crossover model, and show that our approach allows estimation of complex non-linear associations between exposure and mortality and provides principled, model-based uncertainty estimates. In §3 we describe our inference methodology, where we avoid the conditional independence restraint by allowing for a non-diagonal Hessian matrix of the log-likelihood. In §4 we discuss computational considerations with a focus on sparse matrix algebra, and demonstrate that allowing for a non-diagonal Hessian is computationally feasible. In §5 we present three data analyses which illustrate the speed and empirical accuracy of our approach as well as its breadth. We conclude in §6 with a discussion.

2. Model

2.1 *Case-crossover models*

Define the response vector $Y = \{Y_i; i = 1 \dots n\}$ where Y_i is the random variable representing the death time of the i^{th} subject, referred to as the *case day*, and let $y = \{y_i; i = 1 \dots n\}$ be its realization in the observed data. The hazard function of the event times Y_i is modelled as $h_i(t) = \alpha_i(t)\lambda_i(t)$, where

$$\log \lambda_i(t) = x_i(t)^T\beta + \sum_{q=1}^{r} \gamma_q[u_{qi}(t)] \tag{1}$$

Here $x_i(t)$ is a vector of time-varying covariates modelled as fixed effects and $u_{qi}(t)$ are covariates modelled semi-parametrically using unknown smooth functions γ_q. The baseline mortality hazard $\alpha_i(t)$ captures all unmeasured mortality risk factors.

Case-crossover analyses involve choosing a referent frame $S_i = \{c_{i1}, \dots, c_{iJ_i}\}$ of J_i *control days* for each subject chosen such that each subject's baseline mortality hazard is similar to that on the case day; $\alpha_i(t) \approx \alpha_i(y_i)$ for each $t \in S_i$ (Janes et al., 2005). For example, in longitudinal mortality studies, control days are often chosen to be on the same day of the week as the case day. Redelmeier and Tibshirani (1997) choose control days in which accident victims had similar patterns of driving as on the case day. Under this assumption the *hazard ratios* take the form $h_i(y_i)/h_i(t) = \lambda_i(y_i)/\lambda_i(t) \equiv \exp[\Delta_i(t)]$ where $\Delta_i(t) = \log \lambda_i(y_i) - \log \lambda_i(t)$.

The vectors $\lambda = \{\lambda_i(t); t \in \{y_i\} \cup S_i, i = 1 \dots n\}$ and $\Delta = \{\Delta_i(t); t \in \{y_i\} \cup S_i, i = 1 \dots n\}$ are of inferential interest.

A partial likelihood is obtained by considering the conditional probability:

$$\begin{aligned} \pi(y_i|\lambda) = \mathbb{P}\left(y_i = Y_i \middle| Y_i \in \{y_i\} \cup S_i; \lambda\right) &= \frac{\exp\left[\lambda_i(y_i)\right]}{\exp\left[\lambda_i(y_i)\right] + \sum_{t \in S_i} \exp\left[\lambda_i(t)\right]} \\ &= \frac{1}{1 + \sum_{t \in S_i} \exp\left[-\Delta_i(t)\right]} \end{aligned} \qquad (2)$$

and hence the likelihood is $\pi(y|\Delta) = \prod_{i=1}^{n} \pi(y_i|\lambda)$. Note the likelihood and corresponding log-likelihood $\ell(\Delta; y) = \log \pi(y|\Delta)$ only depend on λ through Δ and we reflect this in our notation throughout the remainder of the paper.

2.2 *Additive predictor*

Our methodology can fit case-crossover models with a very flexible and general form for the mortality hazard due to the general form allowed for the additive predictor (1). For computational reasons, we follow Rue et al. (2009) and add a small amount of Gaussian noise to the additive predictor, defining auxillary variables $\eta_i(t)$:

$$\eta_i(t) = \log \lambda_i(t) + \epsilon_i(t) = x_i(t)^T \beta + \sum_{q=1}^{R} \gamma_q[u_{qi}(t)] + \epsilon_i(t) \qquad (3)$$

where $\epsilon_i(t) \overset{iid}{\sim} \text{Normal}(0, \tau^{-1})$. We perform inference on $\eta = \{\eta_i(t); t \in \{y_i\} \cup S_i, i = 1 \dots n\}$ rather than λ. This greatly increases the *sparsity* of the large matrices involved in calculations; see §4 for a detailed discussion. By choosing the precision τ to be a large, fixed constant, we ensure that the addition of the $\epsilon_i(t)$ terms do not appreciably change the numerical answers returned by our software.

To complete our model specification, we require prior distributions for all model parameters. A joint Gaussian prior distribution is used for the regression coefficients $\beta \sim \text{Normal}\left(0, \Sigma_\beta\right)$. To perform inference on the smooth functions γ_q, we model each with a separate Gaussian Process prior. To reduce this infinite-dimensional estimation problem to finite dimensions, define $u_q = \{u_{qi}(t) : i = 1 \dots n, t \in \{y_i\} \cup S_i\}$ and let $U_q = \{U_{q\ell}; \ell = 1 \dots M_q\}$

be an ordered vector of all *unique values* of u_q. We approximate each γ_q using a piecewise-constant function with values at the $U_{q\ell}$, defining $\Gamma_q = \{\gamma_q(U_{q\ell}); \ell = 1 \ldots M_q\}$ and $\Gamma = \{\Gamma_q; q = 1 \ldots R\}$. We regard each Γ_q as a M_q-dimensional *latent vector*. Putting a Gaussian Process prior on γ_q corresponds to putting a joint Gaussian distribution on each $\Gamma_q|\theta \sim \text{Normal}\left[0, \Sigma_q(\theta)\right]$, where θ is a low-dimensional vector of *hyperparameters* which are given prior distribution $\pi(\theta)$. The choice of model for γ_q is then reduced to the choice of the covariance matrix $\Sigma_q(\theta)$; in practice, the model is most often parametrized through the *precision matrix* $\Sigma_q^{-1}(\theta)$. A common choice that we adopt in our data analysis examples is the *second-order random walk* (RW$_2$) model (Lindgren and Rue, 2008) but others, including those with longitudinal or spatial covariance structures, are possible due to the flexibility of our approach.

Using the piecewise approximation to γ_q, we may write $\eta = X\beta + A\Gamma + \epsilon$. Here X and A are the fixed- and random-effects design matrices and $\epsilon \sim \text{Normal}(0, \tau^{-1}I)$. We re-define $\Delta_i(t) = \eta_i(y_i) - \eta_i(t)$. We obtain Δ directly from η by a simple linear transformation and hence $\pi(\Delta|\Gamma, \beta, \theta)$ is a multivariate Gaussian distribution. Our model further directly specifies each of $\pi(\Gamma|\theta)$ and $\pi(\beta)$ as Gaussian distributions. Motivated by this construction we define the vector of *latent Gaussian* quantities as $W = (\Delta, \Gamma, \beta)$ and write its conditional distribution as $W|\theta \sim \text{Normal}\left[0, Q^{-1}(\theta)\right]$ where $Q(\theta)$ is the *precision matrix*. For any fixed θ, our model specifies the *conditional posterior of the latent Gaussian variables*:

$$\begin{aligned}\pi(W|\theta, y) &\propto \pi(y|W, \theta)\pi(W|\theta) \\ &\propto \pi(y|\Delta)\pi(\Delta|\Gamma, \beta, \theta)\pi(\Gamma|\theta)\pi(\beta) \\ &\propto \exp\left[-\frac{1}{2}W^T Q(\theta) W + \ell(\Delta; y)\right]\end{aligned} \tag{4}$$

where $\pi(y|W, \theta) \equiv \pi(y|\Delta)$. Expression (4) is not useful on its own, as it is an unnormalized high-dimensional distribution and is defined for a fixed value of the unknown hyperparameter

θ. We use (4) as a basis for the marginal posterior distributions upon which we base our inferences. We describe our approach to this in detail in §3.

3. Inference Methodology

Our objects of primary inferential interest are the *marginal posterior distributions*:

$$\pi(W_j|y) = \int \pi(W_j|\theta, y)\pi(\theta|y)d\theta \tag{5}$$

which we may use to compute posterior means and modes, credible intervals, and other summaries of interest for the unknowns $W = (\Delta, \Gamma, \beta)$. We are also interested in the *joint posterior distribution*:

$$\pi(W|y) = \int \pi(W|y, \theta)\pi(\theta|y)d\theta \tag{6}$$

which we may use to compute samples, posterior summaries of non-linear functions of W, and global credible envelopes. Expression (5) depends on the marginal posterior of W for fixed θ,

$$\pi(W_j|\theta, y) = \int \pi(W|\theta, y)dW_{-j} \tag{7}$$

and the *hyperparameter posterior*

$$\pi(\theta|y) = \int \pi(\theta, W|y)dW = \frac{\int \pi(W, \theta, y)dW}{\int\int \pi(W, \theta, y)dWd\theta} \tag{8}$$

Expressions (7) and (8) both involve intractable high-dimensional integrations and require efficient, scalable approximations, which we achieve by using Laplace approximations. We will use these combined with numerical integration to approximate the low-dimensional integral (5). This will amount to approximating (5) using a *mixture of Gaussian distributions.*

3.1 *Approximations for the marginal posteriors*

In order to ensure our flexible modelling framework scales efficiently with the dimension of W, we require an approximation methodology for (7) and (8) that achieves high accuracy without sacrificing scalability. We describe our approach here.

3.1.1 *Laplace approximation for* $\pi(W_j|\theta,y)$. For fixed θ, we utilize a Laplace approximation for (4), motivated by the fact that (4) is nearly Gaussian. Define the *conditional mode*:

$$\hat{W}(\theta) = \left[\hat{\Delta}(\theta), \hat{\Gamma}(\theta), \hat{\beta}(\theta)\right] = \text{argmax}_W \log \pi(W|y,\theta) \tag{9}$$

and the *negated Hessian* of the log-posterior:

$$H_\theta(W) = -\frac{\partial^2 \log \pi(W|y,\theta)}{\partial W \partial W^T} \tag{10}$$

Details regarding the high-dimensional optimization required to find $\hat{W}(\theta)$ are given in §4 and Appendix A. We expand $\log \pi(W|y,\theta)$ to second order around $\hat{W}(\theta)$:

$$\log \pi(W|y,\theta) \approx c + \log \pi[\hat{W}(\theta)|y,\theta] - \frac{1}{2}\left[W - \hat{W}(\theta)\right]^T H_\theta[\hat{W}(\theta)]\left[W - \hat{W}(\theta)\right] \tag{11}$$

where c is a constant. This yields a Gaussian approximation to $\pi(W|y,\theta)$,

$$\tilde{\pi}(W|y,\theta) \propto \exp\left\{-\frac{1}{2}\left[W - \hat{W}(\theta)\right]^T H_\theta[\hat{W}(\theta)]\left[W - \hat{W}(\theta)\right]\right\} \tag{12}$$

from which we obtain the Laplace approximation to the desired marginal posterior,

$$\tilde{\pi}(W_j|y,\theta) = \int \tilde{\pi}(W|y,\theta) dW_{-j} \propto \exp\left\{-\frac{1}{2v(\theta)_j^2}\left[W_j - \hat{W}_j(\theta)\right]^2\right\} \tag{13}$$

where the variance is:

$$v(\theta)_j^2 = \left\{H_\theta[\hat{W}(\theta)]^{-1}\right\}_{jj} \tag{14}$$

The negated Hessian has the form $H_\theta(W) = Q(\theta) + C(W)$ where

$$C(W) = -\begin{pmatrix} \frac{\partial^2 \ell(\Delta;y)}{\partial\Delta\partial\Delta^T} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \tag{15}$$

The matrix $C(W)$ is non-diagonal because the likelihood for each observation depends on a non-linear combination of the latent variables W. The restriction imposed by the INLA methodology (Rue et al., 2009) that the observations be conditionally independent given a *linear* combination of the latent variables forces $C(W)$ to be a diagonal matrix. This excludes case-crossover and other important models from being fit using the INLA framework. By allowing a non-diagonal Hessian matrix, we remove this restriction.

3.1.2 *Laplace approximation for $\pi(\theta|y)$.* The hyperparameter posterior (8) is intractable because it involves two high-dimensional integrals, but unlike (4) there is no reason to believe it is close to Gaussian. If we consider the full vector of unknowns (W, θ), we see that (8) can be regarded as a low-dimensional *marginal posterior* of a high-dimensional latent vector. Tierney and Kadane (1986) describe an accurate and efficient approximation scheme for this setting, which yields a *Laplace approximation to the hyperparameter posterior*:

$$\tilde{\pi}(\theta|y) \propto \pi(\theta)\frac{|Q(\theta)|^{1/2}}{|H_{\theta}[\hat{W}(\theta)]|^{1/2}} \exp\left\{-\frac{1}{2}\hat{W}(\theta)^{T}Q(\theta)\hat{W}(\theta) + \ell[\hat{\Delta}(\theta); y]\right\} \tag{16}$$

Expression (16) may also be derived directly using the Gaussian approximation (12); see Appendix B. The constant of proportionality in (16) is known but difficult to compute, and due to the low dimension of θ, Tierney and Kadane (1986) recommend ignoring constants in the calculation and renormizaling via numerical integration. We take this approach here. It is important to note that in order for a Laplace approximation to be used, the parameters θ must each be supported on the whole real line, which can always be achieved via reparametrization. For example, if the model chosen for Γ involves a standard deviation σ, we may instead parametrize it using the log-precision $\theta = -2\log\sigma$.

3.1.3 *Numerical integration for $\pi(W_j|y)$ and $\pi(W|y)$.* With efficient, scalable approximations (13) and (16) available, we approximate (5) and (6) using a mixture of Gaussian distributions via numerical integration:

$$\begin{aligned}\tilde{\pi}(W_j|y) &= \sum_{k=1}^{K}\tilde{\pi}(W_j|y, \theta^k)\tilde{\pi}(\theta^k|y) \\ \tilde{\pi}(W|y) &= \sum_{k=1}^{K}\tilde{\pi}(W|y, \theta^k)\tilde{\pi}(\theta^k|y)\end{aligned} \tag{17}$$

Marginal moments, quantiles, and any other desired posterior summaries are obtained directly from (17). The $\left\{\theta^k; k = 1 \dots K\right\}$ is a regular grid of integration points. We now turn to a detailed discussion of the computational aspects of our approach.

4. Computational Considerations

Our methodology is developed with scalability and computational efficiency as fundamentally important goals. We achieve this using sparse matrix algebra and modern optimization techniques. Our major computational tasks are computing the conditional mode $\hat{W}(\theta)$ for each θ, obtaining marginal variances from $H_\theta[\hat{W}(\theta)]$, and incorporating any linear constraints imposed on the model for Γ.

4.1 *Sparse precision matrix of the latent Gaussian variables*

Conditional on the hyperparameters θ the latent vector W is jointly Gaussian and its precision matrix $Q(\theta)$ is required to compute the approximations described in §3. We obtain $Q(\theta)$ from our model specification as follows. Starting from the additive predictor η we obtain $\Delta = D\eta$ where D is a block-diagonal *differencing matrix* with blocks

$$D_i = \begin{pmatrix} -I_{j_i} & 1_{j_i} \end{pmatrix} \in \mathbb{R}^{J_i \times (J_i+1)} \tag{18}$$

where 1_{J_i} is a $J_i \times 1$ column matrix of 1s. Each block is of dimension $J_i \times (J_i + 1)$ and has rank J_i, the number of control days for subject i. The precision matrix of $W|\theta$ is:

$$Q(\theta) = \tau \begin{pmatrix} \Lambda^{-1} & -\Lambda^{-1}DA & -\Lambda^{-1}DX \\ -A^TD^T\Lambda^{-1} & \frac{1}{\tau}\Sigma_U^{-1} + A^TD^T\Lambda^{-1}DA & A^TD^T\Lambda^{-1}DX \\ -X^TD^T\Lambda^{-1} & X^TD^T\Lambda^{-1}DA & \frac{1}{\tau}\Sigma_\beta^{-1} + X^TD^T\Lambda^{-1}DX \end{pmatrix} \tag{19}$$

where $\Lambda = DD^T$.

The sparsity of $Q(\theta)$ is fundamental to the feasibility of our approach. Specifically, an n-dimensional matrix with m non-zero values is referred to as *sparse* if $m = O(n)$ (Rue et al., 2009). Sparse matrix routines allow us to store $Q(\theta)$ with $O(n)$ memory complexity rather than $O(n^2)$ and perform Cholesky decompositions and hence calculate determinants and solve linear systems in $O(n)$ time complexity instead of $O(n^3)$. The sparsity of $Q(\theta)$ is due to most of its large dimension coming from the upper-left block Λ^{-1}. The matrix Λ is block diagonal with blocks $D_iD_i^T$ and its inverse is cheap to compute because $D_iD_i^T = 1_{J_i}1_{J_i}^T + I_{J_i}$

and hence $(D_iD_i^T)^{-1} = I_{J_i} - (J_i+1)^{-1}1_{J_i}1_{J_i}^T$. In practice, the number of control days is not usually large enough for the dimension of these blocks to cause computational difficulties. The cross-products $A^TD^T\Lambda^{-1}DA$ and $X^TD^T\Lambda^{-1}DX$ are dense in general, however their dimensions are determined by the number of linear and smooth terms. Models do not usually have large numbers of linear terms, and the dimension of each Γ_q is governed by the resolution of the grid U_q, giving the user control over this aspect of computation. It follows that $Q(\theta)$ is a sparse matrix, amenable to computationally efficient sparse matrix algorithms, permitting our procedure to scale to very high dimensional problems.

4.2 *Conditional mode and marginal variances*

To determine the conditional mode $\hat{W}(\theta)$, we utilize a modern implementation of trust region optimization with conjugate-gradient updates (Braun, 2014). Trust region algorithms are ideal for cases in which the objective function is high-dimensional, nearly quadratic, and has a sparse Hessian. These considerations mean that the high-dimensional optimization required to compute $\hat{W}(\theta)$ is efficient, stable, and scales efficiently to very high dimensions.

Once $\hat{W}(\theta)$ is found, the required marginal means are obtained directly from its coordinates. The computation of marginal variances involves obtaining the diagonal elements of $H_\theta[\hat{W}(\theta)]^{-1}$. These elements are calculated by inverting the sparse Cholesky decomposition of $H_\theta[\hat{W}(\theta)]$. Because $C[\hat{W}(\theta)]$ is a non-diagonal matrix, sparsity of $Q(\theta)$ does not *immediately* imply sparsity of $H_\theta[\hat{W}(\theta)] = Q(\theta) + C[\hat{W}(\theta)]$. However, the pattern of nonzero values in $C[\hat{W}(\theta)]$ is the *same* as the pattern of nonzero values in the corresponding block of $Q(\theta)$, and hence the result of summing these matrices is still sparse. It follows that computing marginal variances is also an efficient and scalable operation, computable in $O(n)$ time.

4.3 *Models for latent Gaussian variables*

Our framework permits any *latent Gaussian* model for (Γ, β). In our data analysis examples we utilize second-order random walk models, $\Gamma_q \sim \text{RW}_2(\sigma_q^2)$ (Lindgren and Rue, 2008). RW_2 models are underparametrized and while identifiability is not an issue when using Bayesian inference, it is common practice to combine RW_2 terms with underlying global polynomials and corresponding linear constraints. For a single covariate u_1, we fit models with log-hazards of the form:

$$\begin{aligned} \log \lambda_i(t) &\approx u_{1i}(t)\beta_1 + a_{it}^T \Gamma_1 \\ \beta_1 &\sim \text{Normal}(0, \sigma_\beta^2);\ \Gamma_1 \sim \text{RW}_2(\sigma_1^2);\ \Gamma_{1r} = 0 \end{aligned} \tag{20}$$

where a_{it} is the appropriate row of the design matrix A and $r \in U_1$ is some reference value. As in §2, U_1 is the set containing the unique values of u_1.

To fit these models under such linear constraints, we follow Rue and Martino (2007) and first fit the unconstrained model and then *correct* the marginal means and variances post-hoc. Marginal variances for *linear combinations* of elements of W are also straightforward to obtain after fitting. The computational bottleneck of both of these operations is the solving of a high-dimensional linear system involving $Q(\theta)$, which can be done in $O(n)$ time by using sparse matrix algebra.

5. Examples

We illustrate the flexibility and scalability of our procedure using a simulation study and two real data analysis examples. The dimension of W in all of the examples considered is on the order of 10^3 to 10^5, and the number of control days per subject ranges from 3 to 5 and does not have to be the same for every subject in the dataset.

5.1 *Indian mortality data*

Our motivating example is the re-analysis of the data used by Fu et al. (2018) to quantify the association between mortality and exposure to extreme temperatures in India. Fu et al. (2018) found that exposure to both extremely hot and moderately cold temperatures was associated with increased mortality risk by fitting a case-crossover model using splines to capture non-linear associations. In the paper's supplementary materials, it is shown that this method is highly sensitive to the choice of the number and placement of spline knots. In particular, these choices affect the inferred associations in the regions of extreme temperature where there is less data, which unfortunately are the regions of primary inferential interest. Unlike Fu et al. (2018), for this illustrative example we focus on deaths due to stroke and analyze all climate regions simultaneously.

We applied our approach to these data, using both a global linear term with coefficient $\beta_1 \overset{iid}{\sim} \text{Normal}(0, 10^3)$ and a semi-parametric term $\Gamma \sim \text{RW}_2(\sigma^2_\Gamma)$ to model the association between temperature and mortality risk. The data were centred at 32 degrees celcius for the linear model, and we set $\Gamma_{32} = 0$ for the semi-parametric model. The fitted values are interpreted as mortality risk relative to that at 32 degrees. We bin temperature into 1-degree ranges between 10 and 38 degrees celcius and use an exponential prior for σ_Γ, or a Penalized Complexity prior (Simpson et al., 2017). Prior parameters were set such that $\mathbb{P}(\sigma_\Gamma < .2) < .75$, or there is a 75% prior probability that a one degree change in temperature changes the curvature of the underlying risk function by less than 20%. The posterior $\pi(\sigma_\Gamma|y)$ was evaluated at a grid of 45 σ_Γ values, and subsequently smoothed with a simple `loess` smoother to average out small numerical variations. The final dataset had $13,493$ deaths with 3-5 control days each and the dimension of W was $26,303$.

[Figure 1 about here.]

Figure 1 shows that our method gives broadly comparable inferences to what was reported by Fu et al. (2018). We find a higher relative risk at low temperatures, and our error bands are tighter than those reported by Fu et al. (2018). However, our method is *not sensitive* to the number and placement of the bins U_1, a major statistical advantage over the frequentist spline-based approach. A finer temperature resolution results in a more accurate approximation to the underlying continuous process, at a cost of longer computation time. This is highly desirable as it effectively separates the *statistical* properties of the procedure from the *computational* considerations. Also shown in Figure 1 is the posterior for the random walk standard deviation.

5.2 *Canadian air pollution data*

Daily measurements of air pollution and mortality counts since 1990 in Toronto, Canada were obtained from Health Canada. There were $42,274$ deaths from all causes, with five control days per subject chosen in consecutive one week intervals. Subjects' exposure to three pollutants (`PM25`, `NO2`, and `O3`) were recorded as well as the daily maximum temperature. We fit a linear model with fixed effects for each of the four covariates, with a prior precision matrix of $\Sigma_{\beta}^{-1} = 0.001I$. There are no hyperparameters or random effects, and the dimension of W is $211,369$.

[Figure 2 about here.]

Figure 2 shows our estimated densities for the marginal posterior distributions of each regression coefficient and a histogram of $4,000$ posterior samples obtained using the STAN probabilistic programming language (Carpenter et al., 2017), which implements a general Hamiltonian MCMC sampler. Our Gaussian approximations to the marginal posteriors agree closely and total computation time was 3 minutes 45 seconds, a $20\times$ speedup over the 1 hour and 10 minutes taken by STAN. Switching to smooth terms for the covariates represents a minor relative increase in computational burden for us but is infeasible to do with STAN as the MCMC methods used do not scale efficiently with dimension.

5.3 *Simulation study*

To illustrate the accuracy of our procedure for fitting case-crossover models with smooth covariate effects, we performed a simulation study. We performed two simulations, one with a covariate u from true risk function $\gamma_1(u) = 3u^2$ and one with $\gamma_1(u) = 10[.5u^2 - .5u + .05\sin(4\pi u)]$. These represent risk functions that we expect to be easy and difficult to infer, respectively. Our model in both simulations is a line with slope $\beta_1 \sim \text{Normal}(0, 20)$ plus a semi-parametric term $\Gamma_1 \sim \text{RW}_2(\sigma_1^2)$, a second order random walk with linear constraints $\Gamma_{1,l-1} = \Gamma_{1,l} = 0$ for index l corresponding to the mean observed value of u. The sample size for both was $10,000$ with 3 control days per subject. After removing simulated subjects whose values of u were the same on the case and control days (and hence do not contribute to the likelihood), the dimension of W was $26,477$.

[Figure 3 about here.]

Figure 3 shows the resulting estimated curves on the natural scale, which closely match the true curves used to generate the data. The width of the pointwise error bars reflects the difficulty of the estimation task, with the more difficult function having higher posterior uncertainty. Figure 3 also shows the approximation to the posterior distribution of the random walk smoothing standard deviation along with the integration points used. Both posteriors are nicely convex with well-defined modes.

6. Discussion

This paper has introduced a novel methodology for fitting case-crossover models with linear and non-parametric associations between mortality and exposure to risk factors. This approach is efficient and scales to high dimensions by relying on Laplace approximations and numerical integration instead of sampling algorithms to approximate marginal posterior distributions. We demonstrated the accuracy of our approach using simulation studies and its effectiveness through real data examples.

One limitation of our approach is that it is not clear how to theoretically evaluate the accuracy of our approximation, due to the various potential sources of approximation error. Laplace approximations have good relative error rates when the dimension of the vector of unknowns is fixed but in our applications it depends on the sample size. The INLA framework does correct its output for skewness using an ad-hoc skew-Normal approximation fit by moment matching. Our datasets are very large and our procedure is empirically very accurate on simulation studies and when compared to STAN, so we leave this to future work.

Our framework can be applied to many other problems. More complicated models for the non-parametric effects are immediately available, such as temporal and spatially correlated covariance structures, which fit naturally within this framework. The analysis of survival data in general involves likelihoods with the same functional form as ours and would be a feasible extension of this work. Models for aggregated spatial and spatio-temporal data involve likelihoods that depend on non-linear combinations of the latent quantities, and our use of a non-diagonal Hessian matrix opens up the possibility of extending our approach to this important problem.

Acknowledgements

The second and third authors are funded by the Natural Science and Engineering Research Council of Canada. Thanks are due to Hwashin Shin at Health Canada for providing the Canadian mortality data.

References

Braun, M. (2014). trustOptim: An R package for trust region optimization with sparse hessians. *Journal of Statistical Software* **60,** 1–16.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of statistical software* **76,**.

Fu, S. H., Gasparrini, A., Rodriguez, P. S., and Jha, P. (2018). Mortality attributable to hot and cold ambient temperatures in india: a nationally representative case-crossover study. *PLoS Med* **15,**.

Janes, H., Sheppard, L., and Lumley, T. (2005). Casecrossover analyses of air pollution exposure data: Referent selection strategies and their implications for bias. *Epidemiology* **16,** 717–726.

Lindgren, F. and Rue, H. (2008). On the second-order random walk model for irregular locations. *Scandinavian Journal of Statistics* **35,** 691–700.

Maclure, M. (1991). The case-crossover design: A method for studying transient effects on the risk of acute events. *American Journal of Epidemiology* **185,** 144–153.

Martino, S., Akerkar, R., and Rue, H. (2011). Approximate bayesian inference for survival models. *Scandinavian Journal of Statistics* **38,** 514 – 528.

Redelmeier, D. A. and Tibshirani, R. (1997). Association between cellular telephone calls and motor vehicle collisions. *The New England Journal of Medicine* **336,** 453 – 458.

Rue, H. and Martino, S. (2007). Approximate bayesian inference for hierarchical gaussian markov random field models. *Journal of Statistical Planning and Inference* **137,** 3177 – 3192.

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 (Statistical Methodology)* **71,** 319 – 392.

Simpson, D., Rue, H., Martins, T. G., Riebler, A., and Srbye, S. H. (2017). Penalising model component complexity: A principled, practical approach to constructing priors.

Statistical Science **32,**.

Tierney, L. and Kadane, J. B. (1986). Accurate approximations to posterior moments and marginal densities. *Journal of the American Statistical Association* **81,**.

Received September 2019.

Appendix A: Calculations required for optimization

In this appendix, we give details on the calculations required to implement the approximations described in §3. The goal is to find the maximum of $\log \pi(W|y,\theta)$ in W for any fixed arbitrary θ. Specifically,

$$\begin{aligned}\hat{W}(\theta) &= \text{argmax}_W \log \pi(W|y,\theta) \\ \log \pi(W|y,\theta) &= c - \frac{1}{2}W^T Q(\theta) W + \ell(\Delta; y) \\ \ell(\Delta; y) &= -\sum_{i=1}^{n} \log\left\{1 + \sum_{t \in S_i} \exp[-\Delta_i(t)]\right\}\end{aligned} \tag{A.1}$$

The gradient of the log-posterior is:

$$\begin{aligned}\frac{\partial}{\partial W}\log \pi(W|y,\theta) &= Q(\theta)W + \left[\frac{\partial \ell(\Delta_1(c_{i1});y)}{\partial \Delta_1(c_{i1})}, \ldots, \frac{\partial \ell(\Delta_n(c_{nJ_n});y)}{\partial \Delta_n(c_{nJ_n})}, 0, \ldots, 0\right]^T \\ \frac{\partial \ell(\Delta;y)}{\partial \Delta_i(t)} &= \frac{e^{-\Delta_i(t)}}{1 + \sum_{u \in S_i} e^{-\Delta_i(u)}}\end{aligned} \tag{A.2}$$

The negated Hessian is:

$$
\begin{aligned}
H_\theta(W) &= -\frac{\partial^2}{\partial W \partial W^T} \log \pi(W|y,\theta) = Q(\theta) + C(W) \\
C(W) &= \begin{pmatrix} C_1(W) & 0 & \cdots & 0 \\ 0 & \ddots & \cdots & \vdots \\ \vdots & \vdots & C_n(W) & \vdots \\ 0 & \cdots & \cdots & 0 \end{pmatrix} \qquad\qquad \text{(A.3)} \\
C_i(W) &= -\frac{\partial^2}{\partial \Delta_i \partial \Delta_i^T} \ell(\Delta; y) = \left[-\frac{\partial^2}{\partial \Delta_i(t) \partial \Delta_i(s)} \ell(\Delta; y) \right] \; ; \; t, s \in S_i \\
-\frac{\partial^2}{\partial \Delta_i(t) \partial \Delta_i(s)} \ell(\Delta; y) &= \begin{cases} \frac{e^{-\Delta_i(t)}}{1+\sum_{u \in S_i} e^{-\Delta_i(u)}} \left[1 - \frac{e^{-\Delta_i(t)}}{1+\sum_{u \in S_i} e^{-\Delta_i(u)}} \right] & t = s \\ -\frac{e^{-\Delta_i(t)} e^{-\Delta_i(s)}}{\left[1+\sum_{u \in S_i} e^{-\Delta_i(u)}\right]^2} & t \neq s \end{cases}
\end{aligned}
$$

The sparse structure of the gradient and hessian is exploited efficiently by the `trustOptim::trust.optim` function in `R`.

Appendix B: Derivation of approximation to hyperparameter posterior

In this section, we present a detailed derivation of the approximation used for the hyperparameter posterior $\pi(\theta|y)$. We show how our Gaussian approximation to $\pi(W|\theta, y)$ can be used directly to obtain an expression identical to that described by Tierney and Kadane (1986). Starting from the identity:

$$\pi(W, \theta, y) = \pi(W|\theta, y)\pi(\theta|y)\pi(y) \qquad \text{(A.4)}$$

we obtain the representation:

$$
\begin{aligned}
\pi(\theta|y) &= \frac{\pi(W, \theta, y)}{\pi(W|\theta, y)\pi(y)} \\
&= \frac{\pi(\theta)\pi(W|\theta)\pi(y|W, \theta)}{\pi(W|\theta, y)\pi(y)} \qquad \text{(A.5)} \\
&\propto \frac{\pi(\theta)\pi(W|\theta)\pi(y|W, \theta)}{\pi(W|\theta, y)}
\end{aligned}
$$

which holds for any W. Specifically, it holds for $W = \hat{W}(\theta)$. Note that $\pi(y|W,\theta) \equiv \pi(y;\Delta)$ is the likelihood. Our Gaussian approximation to $\pi(W|\theta,y)$ is:

$$\tilde{\pi}(W|\theta,y) \propto |H_\theta[\hat{W}(\theta)]|^{1/2} \exp\left\{-\frac{1}{2}\left[W - \hat{W}(\theta)\right]^T H_\theta[\hat{W}(\theta)]\left[W - \hat{W}(\theta)\right]\right\} \tag{A.6}$$

where $H_\theta[\hat{W}(\theta)] = Q(\theta) + C[\hat{W}(\theta)]$. Evaluated at its own mode, however, this approximation becomes:

$$\tilde{\pi}(\hat{W}(\theta)|\theta,y) \propto |H_\theta[\hat{W}(\theta)]|^{1/2} \tag{A.7}$$

Plugging this into (A.5) yields:

$$\begin{aligned}\tilde{\pi}(\theta|y) &\propto \frac{\pi(\theta)\pi[\hat{W}(\theta)|\theta]\pi(y|W,\theta)}{\tilde{\pi}[\hat{W}(\theta)|\theta,y]} \\ &\propto \frac{\pi(\theta)|Q(\theta)|^{1/2}\exp\left\{-\frac{1}{2}\hat{W}(\theta)Q(\theta)\hat{W}(\theta) + \ell[\hat{\Delta}(\theta);y]\right\}}{|H_\theta[\hat{W}(\theta)]|^{1/2}}\end{aligned} \tag{A.8}$$

which is expression (16). This is a low-dimensional density and is normalized using numerical integration.

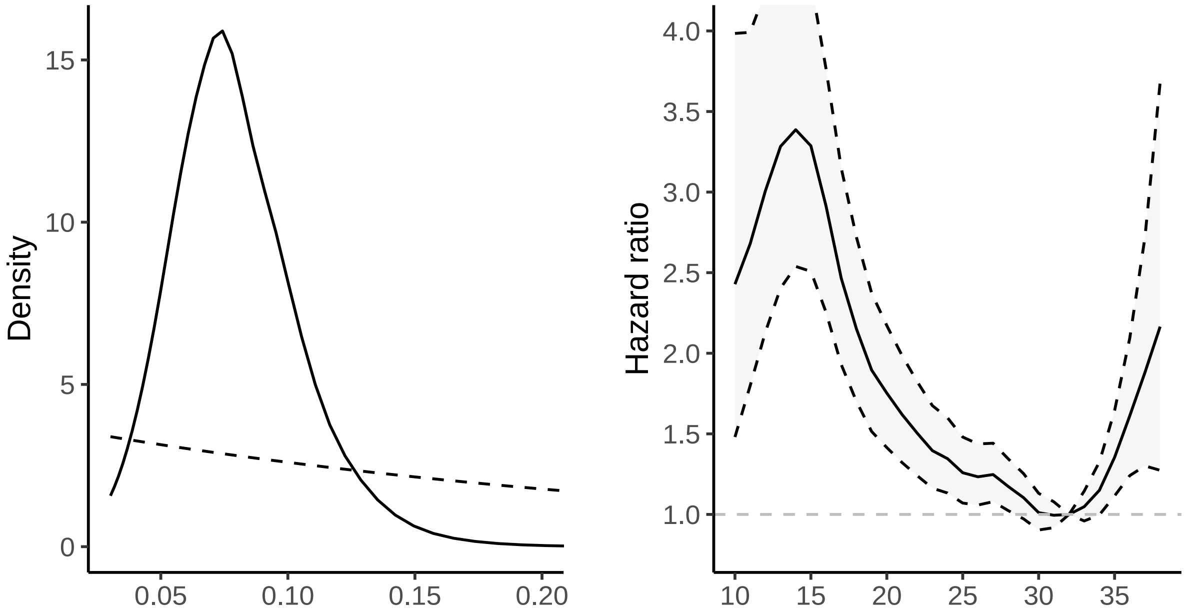
The approach of Tierney and Kadane (1986) is to apply a separate Laplace approximation to the numerator and denominator of $\pi(\theta|y)$. However, they suggest ignoring constants and renormalizing the approximation numerically, which gives an identical expression to (A.8). To see this, first note that:

$$\log\pi(W,\theta,y) = \log\pi(W|\theta,y) + \log\pi(\theta|y) + \log\pi(y) \tag{A.9}$$

A Laplace approximation to the numerator of $\pi(\theta|y)$ is:

$$\int \pi(W,\theta,y)dW \approx c \times |H_\theta[\hat{W}(\theta)]|^{-1/2}\pi[\hat{W}(\theta),\theta,y] \tag{A.10}$$

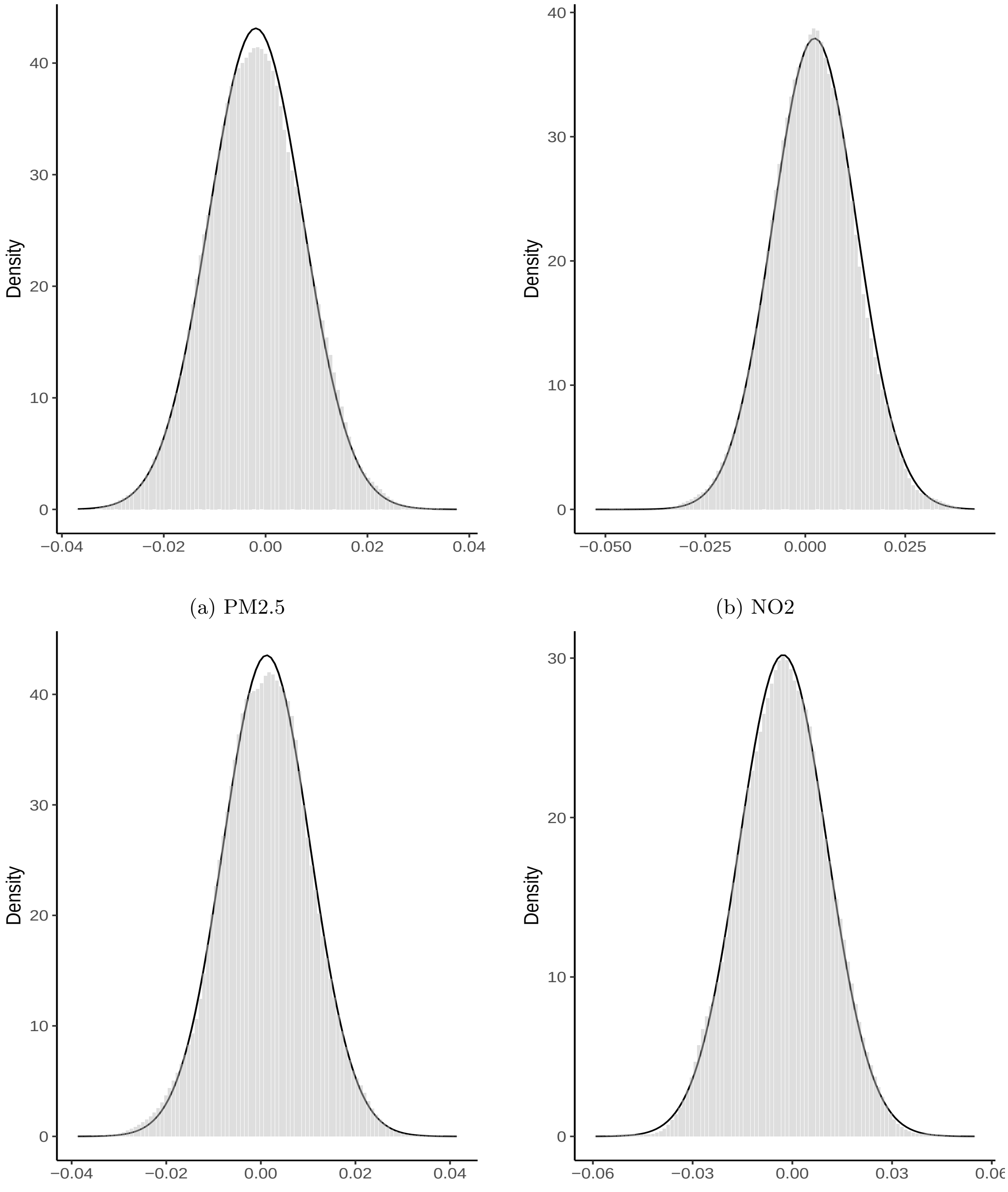
where c is a constant not depending on θ. This expression is, up to constants not depending on θ, identical to (A.8). Since a Laplace approximation to the denominator of $\pi(\theta|y)$ is constant with respect to θ, (A.10) is the result of applying the method of Tierney and Kadane (1986).



(a) σ

(b) Hazard ratio vs. temperature

Figure 1: Results from fitting the case-crossover model to the Indian temperature and mortality data. Left panel: prior (- - -) and posterior (—) distributions for the standard deviation of the random walk. Right panel: posterior mean (—) and 95% credible interval for the hazard ratio as a function of maximum temperature.



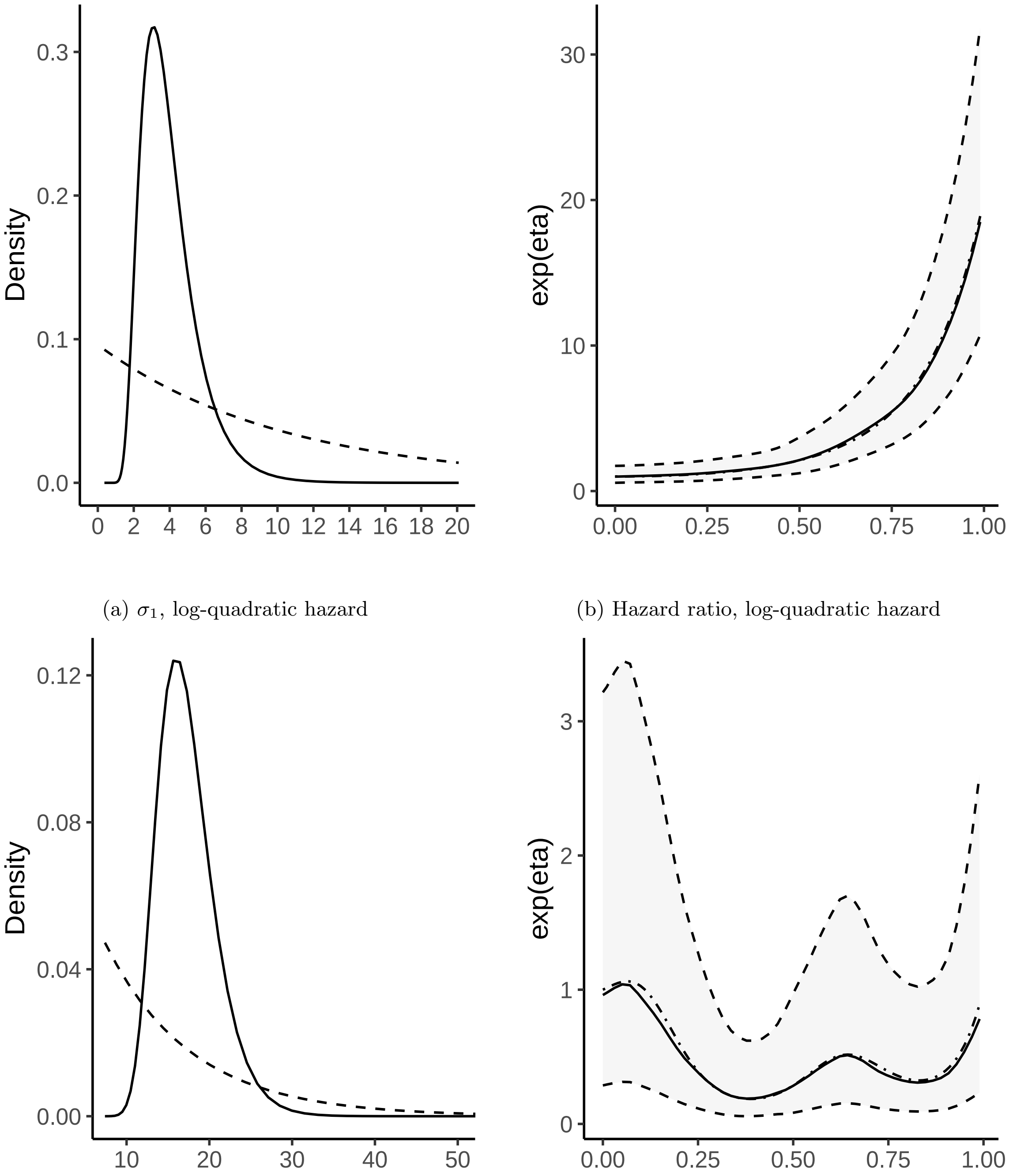
(a) PM2.5

(b) NO2

(c) O3

(d) Daily Maximum Temperature

Figure 2: Posterior distributions for regression coefficients from the analysis of the Canadian pollution data. Shown are results from STAN (shaded histograms) and the proposed methodology (solid density lines).



(a) σ_1, log-quadratic hazard

(b) Hazard ratio, log-quadratic hazard

(c) σ_1, sinusoidal hazard

(d) Hazard ratio, sinusoidal hazard

Figure 3: Results from fitting non-parametric case-crossover models to simulated data with quadratic log-hazard ratio (top panels) and sinusoidal log-hazard ratio (bottom panels). Left panels show prior (- - -) and posterior (—) for the standard deviation of the random walk term. Right panels show the true hazard ratio (- · -), posterior mean (—), and 95% credible interval.