Bayesian Analysis (2015) **10**, Number 4, pp. 991–1023

Variational Inference for Count Response Semiparametric Regression

J. Luts[*] and M. P. Wand[†]

Abstract. Fast variational approximate algorithms are developed for Bayesian semiparametric regression when the response variable is a count, i.e., a non-negative integer. We treat both the Poisson and Negative Binomial families as models for the response variable. Our approach utilizes recently developed methodology known as non-conjugate variational message passing. For concreteness, we focus on generalized additive mixed models, although our variational approximation approach extends to a wide class of semiparametric regression models such as those containing interactions and elaborate random effect structure.

Keywords: approximate Bayesian inference, generalized additive mixed models, mean field variational Bayes, penalized splines, real-time semiparametric regression.

1 Introduction

A pervasive theme impacting Statistics in the mid-2010s is the increasing prevalence of data that are big in terms of volume and/or velocity. One of many relevant articles is Michalak et al. (2012), where the need for systems that perform real-time streaming data analyses is described. The analysis of high volume data and velocity data requires approaches that put a premium on speed, possibly at the cost of accuracy. Within this context, we develop methodology for fast, and possibly online, semiparametric regression analyses in the case of count response data.

Semiparametric regression, as defined in Ruppert et al. (2009), is a fusion between parametric and nonparametric regression that integrates low-rank penalized splines and wavelets, mixed models and Bayesian inference methodology. In Luts et al. (2014), we developed semiparametric regression algorithms for high volume and velocity data using a mean field variational Bayes (MFVB) approach (e.g., Wainwright and Jordan, 2008). It was argued there that MFVB, or similar methodology, is necessary for fast batch and online semiparametric regression analyses, and that more traditional methods such as Markov chain Monte Carlo (MCMC) and sequential Monte Carlo are not feasible. However, the methodology of Luts et al. (2014) was restricted to fitting Gaussian and Bernoulli response models. Extension to various other response distributions, such as t, Skew Normal and Generalized Extreme Value is relatively straightforward using approaches described in Wand et al. (2011). However, count response distributions such as the Poisson and Negative Binomial distribution have received little attention in the MFVB literature. Recently Tan and Nott (2013) used an extension of MFVB, known as non-conjugate variational message passing, to handle Poisson mixed models for lon-

[*] University of Technology Sydney, jan@thesearchparty.com

[†] University of Technology Sydney, matt.wand@uts.edu.au

DOI: 10.1214/14-BA932

gitudinal data and their lead is followed here for more general classes of count response semiparametric regression models. The Poisson response models treated here are more general than those of Tan and Nott (2013) in that they include, for example, Poisson mixed models, additive models, varying coefficient models, additive mixed models and geoadditive models as special cases. Our MFVB methodology for Negative Binomial semiparametric regression is unprecedented.

In generalized response regression, the Poisson distribution is often bracketed with the Bernoulli distribution since both are members of the one-parameter exponential family. However, variational approximations for Poisson response models are not as forthcoming as those with Bernoulli responses. Jaakkola and Jordan (2000) derived a lower bound on the Bayesian logistic regression marginal likelihood that leads to tractable approximate variational inference. As explained in Girolami and Rogers (2006) and Consonni and Marin (2007), the Albert and Chib (1993) auxiliary variable representation of Bayesian probit regression leads to a different type of variational approximation method for binary response regression. There do not appear to be analogues of these approaches for Bayesian Poisson regression and different routes are needed. In the MCMC literature, novel strategies for handling the Poisson case include the introduction of auxiliary variables that convert the MCMC problem into one that involves sampling from Truncated Normal distributions (Damien et al., 1999) and Finite Normal Mixture density approximations of the Log-Gamma family of density functions (Frühwirth-Schnatter et al., 2009). An effective solution in the variational approximation case is afforded by an extension of MFVB, due to Knowles and Minka (2011), known as *non-conjugate variational message passing*. The Negative Binomial distribution can also be handled using non-conjugate variational message passing, via its well-known representation as a Poisson–Gamma mixture (e.g., Lawless, 1987). We adopt such an approach here and develop MFVB algorithms for both Poisson and Negative Binomial semiparametric regression models. For ease of presentation, we restrict attention to the special case of generalized additive mixed models, but extension to other semiparametric regression models is straightforward.

Section 2 contains general discussion on statistical methodological development in this era of high volume/velocity data becoming more prevalent. Section 3 describes the count response semiparametric regression models to be treated. The article's centerpiece is Section 4, which is where the variational inference algorithms for count response semiparametric regression are presented. In Section 5, we describe real-time fitting of such models. Numerical illustrations are given in Section 6, and concluding remarks are made in Section 7. Appendix A lays down required notation and distributional results. It also provides a brief synopsis of non-conjugate mean field variational Bayes. Appendix B contains derivations of the aforementioned variational algorithms.

2 Statistical Methodology and High Volume/Velocity Data

As high volume/velocity data become more prevalent in the mid-2010s and beyond it is worthwhile to reflect on how advanced statistical methodology might adapt. Even though the current article is concerned with semiparametric regression analysis when the response data are counts, it is a substantial statistical problem that is representative of ongoing methodological research. Before describing our new approach geared towards high volume/velocity situations we will, in this section, make some general remarks about statistical methodological development in the current era.

For most of its history statistical methodological development has concentrated on "batch" analysis of small to moderately-sized data sets. In many areas of application, such as demography and epidemiology, it is typical that the amount of effort required to collect the data greatly exceeds that required for its analysis. Computationally intensive methods such as bootstrapping and MCMC often provide satisfactory analyses in a fraction of the collection time. However, the applicability of computationally intensive methods to very large data sets and models arising in the 2010s is not clear-cut and involves trade-offs between factors such as the size of the problem, amount of available computing power and the time frame for actionable results. The article that introduced the bootstrap (Efron, 1979) contains a single illustration via a regression data set with 2 variables and 9 observations. As mentioned in Mittal et al. (2013), regression data sets with "hundreds of thousands of variables and even millions of observations" are now commonplace in genomics. Michalak et al. (2012) describe radio astronomy data that are so big and fast that it is "prohibitively expensive or impossible to save all of it for later analysis" and "data of interest must be identified quickly". In such circumstances, computationally intensive methods are not viable. The analyses of Michalak et al. (2012) involve computationally frugal methods such as M-estimation and exponentially weighted moving average filtering.

It is difficult to forecast the types of high volume/velocity applications that will emerge in the upcoming decades. However, trends in the first two decades of the 21st Century indicate such applications becoming increasingly prevalent. As data become cheaper, bigger and faster, we believe that some of the traditional mindsets of statistical methodologists need to change. Computationally intensive methods such as MCMC and sequential Monte Carlo will always have a place in Bayesian statistical analyses. However, "lightweight" alternatives, such as integrated nested Laplace approximation (Rue et al., 2009) are likely to have an increasingly important role.

The approach taken in this article is based on recent trends and developments in machine learning research. Bishop (2008) describes a "new framework for machine learning" that is "built on three key ideas: (i) the adoption of a Bayesian viewpoint, (ii) the use of probabilistic graphical models, and (iii) the application of fast, deterministic inference algorithms". In this article we focus on the most popular fast deterministic inference algorithm, mean field variational Bayes (MFVB). Expectation propagation (Minka, 2001) is another algorithm of this type. Some advantages of the MFVB approach to fitting and inference for high volume/velocity data are:

- It results in iterative algorithms with updates that are often closed form algebraic expressions that are easy to implement, fast to compute, parallelize and modify for real-time processing (Luts et al., 2014);

- The iterative updates required for a parameter in a large graphical model are localized around the parameter's node on the graph, which implies that MFVB methodology developed for smaller models is readily transferrable to larger more general models.

These aspects are illustrated in Sections 4 and 5. Our MFVB procedure for Poisson semiparametric regression, within Algorithm 1, requires only algebraic manipulations, some of which are identical to the more general Negative Binomial model. Algorithm 2 represents a relatively straightforward extension to real-time fitting and inference.

3 Model Descriptions

Count responses are most commonly modeled according to the Poisson and Negative Binomial distributions. The latter may be viewed as an extension of the former through the introduction of an additional parameter.

The model descriptions depend on distributional definitions and results given in Appendix A. Throughout this section we use $\overset{\text{ind.}}{\sim}$ to denote "independently distributed as".

3.1 Poisson Additive Mixed Model

We work with the following class of Bayesian Poisson additive mixed models:

$$\begin{gathered} y_i|\,\boldsymbol{\beta},\boldsymbol{u} \overset{\text{ind.}}{\sim} \text{Poisson}[\,\exp\{(\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{Z}\boldsymbol{u})_i\}], \quad 1\le i\le n,\\ \boldsymbol{u}|\,\sigma_1^2,\ldots,\sigma_r^2 \sim N(\mathbf{0},\text{blockdiag}(\sigma_1^2\,\boldsymbol{I}_{K_1},\ldots,\sigma_r^2\,\boldsymbol{I}_{K_r})), \\ \boldsymbol{\beta}\sim N(\mathbf{0},\sigma_\beta^2\boldsymbol{I}_p), \quad \text{and} \quad \sigma_\ell \overset{\text{ind.}}{\sim} \text{Half-Cauchy}(A_\ell), \quad 1\le \ell\le r. \end{gathered} \tag{1}$$

Here y_i is the ith response measurement, $\boldsymbol{\beta}$ is a $p\times 1$ vector of fixed effects, $\boldsymbol{u}$ is a vector of random effects, $\boldsymbol{X}$ and $\boldsymbol{Z}$ are corresponding design matrices, and $\sigma_1^2,\ldots,\sigma_r^2$ are variance parameters corresponding to sub-blocks of $\boldsymbol{u}$ of size $K_1,\ldots,K_r$. All distributional definitions are given in Table 2.

In model (1), the random effects component is

$$\boldsymbol{Z}\boldsymbol{u}=\sum_{\ell=1}^{r}\boldsymbol{Z}_\ell\,\boldsymbol{u}_\ell$$

where the design matrix $\boldsymbol{Z}_\ell$ has dimension $n\times K_\ell$ and its corresponding coefficient vector $\boldsymbol{u}_\ell\sim N(\mathbf{0},\sigma_r^2\boldsymbol{I})$ has length K_ℓ. In additive mixed models, the $\boldsymbol{Z}_\ell$ either contain indicators of group membership, as is common in classical longitudinal and multilevel models, or spline basis functions of a continuous predictor variable. Wand and Ormerod (2008) describe appropriate spline bases which correspond to low-rank cubic smoothing splines, based on results given in O'Sullivan (1986), and dub them O'Sullivan splines. We recommend O'Sullivan splines as a default basis, with interior knots placed at the sample quantiles of the unique predictor data, and they are used in the examples in Section 6. There remains the choice of the number of basis functions, corresponding to the K_ℓ for which $\boldsymbol{Z}_\ell$ contains splines. For most functions arising in applications, including all of those in the examples of Section 6, $K_\ell = 25$ is sufficient. However, for complicated functional effects, such as those with very many oscillations, larger spline bases may be needed.

In (1) note that the response measurements have a single subscript even though one or more of the $\boldsymbol{Z}_\ell$s may induce grouping structure. For example, if $r = 1$ and $\boldsymbol{Z}_1 = \text{blockdiag}_{1\le i\le m}(\mathbf{1}_{n_i})$, where $\mathbf{1}_{n_i}$ is the $n_i \times 1$ vector of ones, then it is common to label the response variables using double subscripts: that is, y_{ij} denotes the jth response measurement within the ith group, $1 \le j \le n_i$, $1 \le i \le m$. Whilst we use this double subscript convention in particular examples involving grouped data, we use the single subscript version here since it better handles general additive mixed models – which may or may not have grouping structure.

Result 2 of Appendix A.2 allows us to replace $\sigma_\ell \overset{\text{ind.}}{\sim} \text{Half-Cauchy}(A_\ell)$ with

$$\sigma_\ell^2 \,|\, a_\ell \overset{\text{ind.}}{\sim} \text{Inverse-Gamma}(\tfrac{1}{2}, 1/a_\ell), \quad a_\ell \overset{\text{ind.}}{\sim} \text{Inverse-Gamma}(\tfrac{1}{2}, 1/A_\ell^2), \quad 1 \le \ell \le r,$$

which is more amenable to variational inference.

Note that the $r = 1$ version of (1) is treated in Wand (2014).

3.2 Negative Binomial Additive Mixed Model

The Negative Binomial distribution is an extension of the Poisson distribution in that the former approaches a version of the latter as the shape parameter $\kappa \to \infty$ (see Table 2). The Negative Binomial shape parameter allows for a wider range of dependencies of the variance on the mean and can better handle over-dispersed count data.

The Bayesian Negative Binomial additive mixed model treated here is

$$\begin{gathered} y_i|\,\boldsymbol{\beta},\boldsymbol{u},\kappa \overset{\text{ind.}}{\sim} \text{Negative-Binomial}[\exp\{(\boldsymbol{X\beta}+\boldsymbol{Zu})_i\},\kappa], \quad 1 \le i \le n, \\ \boldsymbol{u}|\,\sigma_1^2,\ldots,\sigma_r^2 \sim N(\mathbf{0}, \text{blockdiag}(\sigma_1^2\,\boldsymbol{I}_{K_1},\ldots,\sigma_r^2\,\boldsymbol{I}_{K_r})), \quad \boldsymbol{\beta} \sim N(\mathbf{0},\sigma_\beta^2\boldsymbol{I}_p), \\ \sigma_\ell \overset{\text{ind.}}{\sim} \text{Half-Cauchy}(A_\ell), \quad 1 \le \ell \le r, \quad \text{and} \quad \kappa \sim F_{1,1}(M_\kappa). \end{gathered} \tag{2}$$

Courtesy of Result 1, given in Appendix A.2,

$$y_i|\,\boldsymbol{\beta},\boldsymbol{u},\kappa \overset{\text{ind.}}{\sim} \text{Negative-Binomial}[\exp\{(\boldsymbol{X\beta}+\boldsymbol{Zu})_i\},\kappa], \quad 1 \le i \le n,$$

can be replaced by

$$y_i|g_i \overset{\text{ind.}}{\sim} \text{Poisson}\,(g_i)\,, \quad g_i|\boldsymbol{\beta},\boldsymbol{u},\kappa \overset{\text{ind.}}{\sim} \text{Gamma}\,(\kappa, \kappa\,\exp\{-(\boldsymbol{X\beta}+\boldsymbol{Zu})_i\})\,, \quad 1 \le i \le n,$$

where $\boldsymbol{g}$ is the $n \times 1$ vector containing the g_i, $1 \le i \le n$.

The prior distribution on κ is a scaled $F_{1,1}$ distribution, defined in Table 2. The scale parameter of the $F_{1,1}$ distribution coincides with the median, so we denote it by M_κ. This prior on κ is skewed towards zero and corresponds to a prior belief of overdispersion for lower values of M_κ. As explained in Section 3 of Marley and Wand (2010) this corresponds to the coefficient of variation of the latent Gamma variables (the g_is) having a Half Cauchy prior, and is in keeping with the advice of Gelman (2006) regarding noninformative priors for strictly positive parameters. Specifically,

$$\kappa \sim F_{1,1}(M_\kappa) \quad \text{if and only if} \quad \sqrt{\text{Var}(g_i)}/E(g_i) = \kappa^{-1/2} \sim \text{Half-Cauchy}(M_\kappa^{-1/2}).$$

If we invoke Result 2 of Appendix A.2 then we can write the prior on κ as

$$\kappa^{-1}|\, a_\kappa \sim \text{Inverse-Gamma}(\tfrac{1}{2}, 1/a_\kappa), \quad a_\kappa \sim \text{Inverse-Gamma}(\tfrac{1}{2}, M_\kappa). \tag{3}$$

As with the a_ℓ, the introduction of the auxiliary variable a_κ simplifies the MFVB calculations.

The Negative Binomial response model (1) is richer than the Poisson response model (2) for semiparametric regression since it allows for the variance function to differ from the mean function. In the Poisson case, these two functions are constrained to equal each other. We elaborate on this point in the case of count response nonparametric regression via mixed model-based penalized splines. Suppose that the data are (x_i, y_i), $1 \le i \le n$, where the x_is are continuous and the y_is are non-negative integers. A Poisson nonparametric regression model is

$$y_i \,|\, x_i, \boldsymbol{\beta}, \boldsymbol{u} \overset{\text{ind.}}{\sim} \text{Poisson}[\exp\{f(x_i)\}], \quad 1 \le i \le n, \tag{4}$$

where

$$f(x) \equiv \beta_0 + \beta_1\, x + \sum_{k=1}^{K} u_k\, z_k(x), \quad u_k|\sigma^2 \overset{\text{ind.}}{\sim} N(0, \sigma^2)$$

corresponding to an $r = 1$ version of (1). Figure 10 in Section 6.3 depicts an example of (4). In (4), properties of the Poisson distribution imply that

$$E(y_i \,|\, x_i, \boldsymbol{\beta}, \boldsymbol{u}) = \text{Var}(y_i \,|\, x_i, \boldsymbol{\beta}, \boldsymbol{u}) = \exp\{f(x_i)\}, \quad 1 \le i \le n,$$

where $\boldsymbol{\beta} \equiv (\beta_0, \beta_1)$ and $\boldsymbol{u} \equiv (u_1, \ldots, u_K)$. The Negative Binomial alternative:

$$y_i \,|x_i, \boldsymbol{\beta}, \boldsymbol{u} \overset{\text{ind.}}{\sim} \text{Negative Binomial}[\exp\{f(x_i)\}], \quad 1 \le i \le n,$$

is such that

$$\begin{array}{rrcl} & E(y_i \,|\, x_i, \boldsymbol{\beta}, \boldsymbol{u}) & = & \exp\{f(x_i)\} \\ \text{and} & \text{Var}(y_i \,|\, x_i, \boldsymbol{\beta}, \boldsymbol{u}) & = & \exp\{f(x_i)\} + \exp\{2f(x_i)\}/\kappa > E(y_i \,|\, x_i, \boldsymbol{\beta}, \boldsymbol{u}) \end{array} \tag{5}$$

(e.g., Lawless, 1987) which allows better handling of the situation where count data are highly dispersed about a nonlinear signal. It is also apparent from (5) that the Negative Binomial distribution can capture overdispersion, but not underdispersion.

There are other versions of (1) where overdispersion is handled via marginalization. A simple example is the Poisson mixed model

$$y_{ij}|\,\beta_0, U_i \overset{\text{ind.}}{\sim} \text{Poisson}\{\exp(\beta_0 + U_i)\}, \quad U_i|\,\sigma^2 \overset{\text{ind.}}{\sim} N(0, \sigma^2), \tag{6}$$

where $1 \leq j \leq n_i$, $1 \leq i \leq m$. The conditional mean and variance are

$$E(y_{ij}|\beta_0, U_i) = \text{Var}(y_{ij}|\beta_0, U_i) = \exp(\beta_0 + U_i)$$

but the *marginal* mean and variance (given the fixed effect and variance parameter) are

$$\begin{array}{rrcl} & E(y_{ij}|\beta_0, \sigma^2) & = & \exp(\beta_0 + \sigma^2/2) \\ \text{and} & \text{Var}(y_{ij}|\beta_0, \sigma^2) & = & \exp(\beta_0 + \sigma^2/2)[1 + \exp(\beta_0 + \sigma^2/2)\{\exp(\sigma^2) - 1\}] \\ & & > & E(y_{ij}|\sigma^2). \end{array}$$

The essential difference between (4) and (6) is that the conditional mean and variance are of intrinsic interest in the former, whereas the marginal mean and variance may matter in the latter.

3.3 Directed Acyclic Graph Representations

Figure 1 provides a directed acyclic graph representation of models (1) and (2). Observed data are indicated by the shaded node while parameters, random effects and auxiliary variables are so-called hidden nodes. This visual representation shows that the Poisson model and Negative Binomial model have parts of their graphs in common. The locality property of MFVB (e.g., Section 2 of Wand et al., 2011) means that the variational inference algorithms for the two models share some of the updates. We take advantage of this in Section 4.

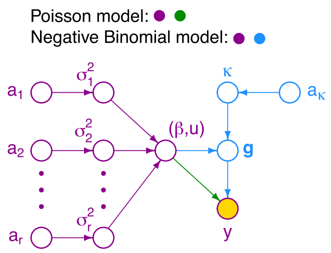

Figure 1: Directed acyclic graph corresponding to the models (1) and (2). The shaded node corresponds to the observed data. The color key at the top of the figure denotes the components of the graph corresponding to each model.

3.4 Extension to Unstructured Covariance Matrices for Random Effects

Section 2.3 of Luts et al. (2014) describes the extension to semiparametric models containing unstructured covariance matrices. Such extensions arise in the case of random intercept and slope models. A simple example of such a model having count responses is

$$y_{ij}|\beta_0,\beta_1,U_i,V_i \overset{\text{ind.}}{\sim} \text{Poisson}\{\exp(\beta_0+U_i+(\beta_1+V_i)\,x_{ij})\},\quad 1\le i\le m,\quad 1\le j\le n_i,$$

$$\text{and}\quad \left[\begin{array}{c} U_i\\ V_i\end{array}\right]\Big|\boldsymbol{\Sigma}\sim N(\mathbf{0},\boldsymbol{\Sigma}),\quad \text{where}\quad \boldsymbol{\Sigma}\equiv\left[\begin{array}{cc}\sigma_u^2 & \rho_{uv}\,\sigma_u\,\sigma_v\\ \rho_{uv}\,\sigma_u\,\sigma_v & \sigma_v^2\end{array}\right].$$

The advice given in Section 2.3 of Luts et al. (2014) concerning such extensions applies here as well.

4 Variational Inference Scheme

We are now in a position to derive a variational inference scheme for fitting the Poisson and Negative Binomial additive mixed models described in Section 3 and displayed in Figure 1. In this section we work toward a variational inference algorithm that treats both models by taking advantage of their commonalities, but also recognizing the differences. The algorithm, which we call Algorithm 1, is given in Section 4.3.

4.1 Poisson Case

We first treat the Poisson additive mixed model (1). Ordinary MFVB begins with a product restriction such as

$$p(\boldsymbol{\beta},\boldsymbol{u},\sigma_1^2,\ldots,\sigma_r^2,a_1,\ldots,a_r|\boldsymbol{y})\approx q(\boldsymbol{\beta},\boldsymbol{u},a_1,\ldots,a_r)\;q(\sigma_1^2,\ldots,\sigma_r^2). \tag{7}$$

Under (7), the optimal posterior density function of $(\boldsymbol{\beta},\boldsymbol{u})$ is

$$q^*(\boldsymbol{\beta},\boldsymbol{u})\propto\exp[E_{-(\boldsymbol{\beta},\boldsymbol{u})}\{\log p(\boldsymbol{\beta},\boldsymbol{u}|\text{rest})\}],$$

where $E_{-(\boldsymbol{\beta},\boldsymbol{u})}$ denotes expectation with respect to the density function $q(a_1,\ldots,a_r)\times q(\sigma_1^2,\ldots,\sigma_r^2)$ and 'rest' denotes all random variables in the model other than $(\boldsymbol{\beta},\boldsymbol{u})$. However, evaluation of $q^*(\boldsymbol{\beta},\boldsymbol{u})$ involves multivariate integrals that are not available in closed form. A non-conjugate variational message passing solution is one that instead works with

$$\begin{aligned}&p(\boldsymbol{\beta},\boldsymbol{u},\sigma_1^2,\ldots,\sigma_r^2,a_1,\ldots,a_r|\boldsymbol{y})\approx\\&\qquad q(\boldsymbol{\beta},\boldsymbol{u};\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})},\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})})\,q(\sigma_1^2,\ldots,\sigma_r^2)\,q(a_1,\ldots,a_r)\end{aligned} \tag{8}$$

where

$$q(\boldsymbol{\beta},\boldsymbol{u};\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})},\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})})\quad\text{is the}\quad N\left(\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})},\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}\right)\quad\text{density function.} \tag{9}$$

In Appendix B, we show that the optimal posterior densities for the variance and auxiliary parameters are:

$$\begin{aligned}
&q^*(\sigma_1^2,\ldots,\sigma_r^2)\text{ is the product of}\\
&\text{Inverse-Gamma}\left(\tfrac{K_\ell+1}{2},\mu_{q(1/a_\ell)}+\tfrac{1}{2}\left\{\|\boldsymbol{\mu}_{q(\boldsymbol{u}_\ell)}\|^2+\text{tr}\left(\boldsymbol{\Sigma}_{q(\boldsymbol{u}_\ell)}\right)\right\}\right)\\
&\text{ density functions, and}\\
&q^*(a_1,\ldots,a_r)\text{ is the product of}\\
&\text{Inverse-Gamma}\left(1,\mu_{q(1/\sigma_\ell^2)}+A_\ell^{-2}\right)\text{ density functions, } 1\le\ell\le r,
\end{aligned}\tag{10}$$

where $\mu_{q(1/\sigma_\ell^2)}\equiv\int_0^\infty(1/\sigma_\ell^2)q(\sigma_\ell^2)\,d\sigma_\ell^2$, $\mu_{q(1/a_\ell)}$ is defined analogously,

$$\boldsymbol{\mu}_{q(\boldsymbol{u}_\ell)}\equiv\text{sub-vector of }\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}\text{ corresponding to }\boldsymbol{u}_\ell$$

and

$$\boldsymbol{\Sigma}_{q(\boldsymbol{u}_\ell)}\equiv\text{sub-matrix of }\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}\text{ corresponding to }\boldsymbol{u}_\ell.$$

The interdependencies between the parameters in these optimal density functions, combined with the updates for $\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}$ and $\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}$ given by (20) in Appendix A, give rise to an iterative scheme for their solution, and is encompassed in Algorithm 1.

Algorithm 1 also uses the variational lower bound on the marginal log-likelihood. For model (1) and restriction (8), it has the explicit expression

$$\begin{aligned}
\log\underline{p}(\boldsymbol{y};q) \quad = \quad & \tfrac{P}{2}-r\log(\pi)-\tfrac{p}{2}\log(\sigma_\beta^2)+\tfrac{1}{2}\log|\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}|-\mathbf{1}^T\log(\boldsymbol{y}!)\\
&-\tfrac{1}{2\sigma_\beta^2}\{\|\boldsymbol{\mu}_{q(\boldsymbol{\beta})}\|^2+\text{tr}(\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})})\}+\textstyle\sum_{\ell=1}^r\Big[\mu_{q(1/a_\ell)}\mu_{q(1/\sigma_\ell^2)}\\
&-\log(A_\ell)-\log(\mu_{q(1/\sigma_\ell^2)}+A_\ell^{-2})+\log\left\{\Gamma\left(\tfrac{K_\ell+1}{2}\right)\right\}\\
&-\tfrac{K_\ell+1}{2}\log(\mu_{q(1/a_\ell)}+\tfrac{1}{2}\{\|\boldsymbol{\mu}_{q(\boldsymbol{u}_\ell)}\|^2+\text{tr}(\boldsymbol{\Sigma}_{q(\boldsymbol{u}_\ell)})\}\Big]\\
&+\boldsymbol{y}^T\boldsymbol{C}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}-\mathbf{1}^T\exp\left\{\boldsymbol{C}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}+\tfrac{1}{2}\text{diagonal}(\boldsymbol{C}\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}\boldsymbol{C}^T)\right\}.
\end{aligned}$$

Here and elsewhere,

$$\text{diagonal}(\boldsymbol{M})\equiv\text{vector of diagonal entries of }\boldsymbol{M}$$

for any square matrix $\boldsymbol{M}$. Also,

$$\boldsymbol{C}\equiv[\boldsymbol{X}\;\;\boldsymbol{Z}]\quad\text{and}\quad P\equiv\text{number of columns in }\boldsymbol{C}=p+\sum_{\ell=1}^r K_\ell.$$

4.2 Negative Binomial Case

We now turn our attention to the Negative Binomial response semiparametric regression model (2) and posterior density function approximations of the form

$$\begin{aligned}
&p(\boldsymbol{\beta},\boldsymbol{u},\boldsymbol{g},\kappa,\sigma_1^2,\ldots,\sigma_r^2,a_1,\ldots,a_r|\boldsymbol{y})\\
&\qquad\approx q(\boldsymbol{\beta},\boldsymbol{u};\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})},\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})})\,q(\boldsymbol{g})\,q(\kappa)\;q(\sigma_1^2,\ldots,\sigma_r^2)\,q(a_1,\ldots,a_r)
\end{aligned}$$

with $q(\boldsymbol{\beta}, \boldsymbol{u}; \boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}, \boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})})$ given by (9).

The optimal q-density functions for $\sigma_1^2, \ldots, \sigma_r^2$ and $a_1, \ldots, a_r$ are given by (10). With $\boldsymbol{c}_i$ denoting the ith row of $\boldsymbol{C}$, the optimal densities for $\boldsymbol{g}$ and κ are:

$$\begin{aligned}
&q^*(\boldsymbol{g}) \text{ is the product of}\\
&\text{Gamma}\left(\mu_{q(\kappa)} + y_i, 1 + \mu_{q(\kappa)} \exp\left(-\boldsymbol{c}_i^T \boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})} + \tfrac{1}{2}\boldsymbol{c}_i^T \boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})} \boldsymbol{c}_i\right)\right)\\
&\text{density functions over } 1 \leq i \leq n \text{ and}\\
&q^*(\kappa) \;=\; \frac{\exp[n\left\{\kappa \log(\kappa) - \log(\Gamma(\kappa))\right\} - C_1 \kappa]}{\kappa^{1/2}\, \mathcal{H}(-\frac{1}{2}, 0, 1, n, C_1)}, \quad \kappa > 0,
\end{aligned} \tag{11}$$

where

$$\mathcal{H}(p, q, r, s, t) \equiv \int_0^\infty x^p \log(1 + rx)^q \{x^x / \Gamma(x)\}^s \exp(-tx)\, dx, \tag{12}$$

for $p \in \{-\frac{1}{2}, \frac{1}{2}\}$, $q \in \{0, 1\}$ and $r, s, t > 0$. Also,

$$\begin{aligned}
C_1 \;\; \equiv \;\; & \boldsymbol{1}^T \big\{ \boldsymbol{C}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})} - \boldsymbol{\mu}_{q(\log(\boldsymbol{g}))} \big\}\\
& + \boldsymbol{\mu}_{q(\boldsymbol{g})}^T \exp\big\{ -\boldsymbol{C}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})} + \tfrac{1}{2}\,\text{diagonal}(\boldsymbol{C}\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}\boldsymbol{C}^T) \big\} + \mu_{q(1/a_\kappa)}.
\end{aligned} \tag{13}$$

Details on the derivation of (11) are given in Appendix B.

Algorithm 1 provides an iterative scheme for obtaining all q-density parameters. The marginal log-likelihood lower bound for the Negative Binomial case is

$$\begin{aligned}
\log \underline{p}(\boldsymbol{y}; q) \;=\; & \tfrac{P}{2} - r\log(\pi) - \tfrac{p}{2}\log(\sigma_\beta^2) + \tfrac{1}{2}\log|\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}| - \tfrac{1}{2\sigma_\beta^2}\left\{ \|\boldsymbol{\mu}_{q(\boldsymbol{\beta})}\|^2 + \text{tr}\left(\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}\right) \right\}\\
& + \textstyle\sum_{l=1}^r \Bigg(\mu_{q(1/a_\ell)}\mu_{q(1/\sigma_\ell^2)} - \log(A_\ell) - \log\left\{ \mu_{q(1/\sigma_\ell^2)} + A_\ell^{-2} \right\}\\
& - \tfrac{K_\ell+1}{2}\log\left[\mu_{q(1/a_\ell)} + \tfrac{1}{2}\left\{ \|\boldsymbol{\mu}_{q(\boldsymbol{u}_\ell)}\|^2 + \text{tr}(\boldsymbol{\Sigma}_{q(\boldsymbol{u}_\ell)}) \right\} \right] + \log\left\{ \Gamma\left(\tfrac{K_\ell+1}{2}\right) \right\} \Bigg)\\
& + \boldsymbol{1}^T \log\left\{ \Gamma\left(\mu_{q(\kappa)}\boldsymbol{1} + \boldsymbol{y} \right) \right\} - \mu_{q(\kappa)}\boldsymbol{1}^T\boldsymbol{\mu}_{q(\log(\boldsymbol{g}))} - \boldsymbol{1}^T\log(\boldsymbol{y}!)\\
& - (\boldsymbol{y} + \mu_{q(\kappa)}\boldsymbol{1})^T\\
& \qquad \times \log\left[\boldsymbol{1} + \mu_{q(\kappa)}\exp\left\{ -\boldsymbol{C}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})} + \tfrac{1}{2}\,\text{diagonal}(\boldsymbol{C}\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}\boldsymbol{C}^T) \right\} \right]\\
& + \mu_{q(\kappa)}\boldsymbol{\mu}_{q(\boldsymbol{g})}^T \exp\left\{ -\boldsymbol{C}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})} + \tfrac{1}{2}\,\text{diagonal}(\boldsymbol{C}\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}\boldsymbol{C}^T) \right\}\\
& - \tfrac{1}{2}\log(M_\kappa) - \log(\pi) - \{\mathcal{H}(-\tfrac{1}{2}, 1, 1/M_\kappa, n, C_1)/\mathcal{H}(-\tfrac{1}{2}, 0, 1, n, C_1)\}\\
& + \log\left\{ \mathcal{H}(-\tfrac{1}{2}, 0, 1, n, C_1) \right\}.
\end{aligned}$$

4.3 Algorithm

We now present Algorithm 1. Note that $\boldsymbol{A} \odot \boldsymbol{B}$ denotes the element-wise product of two equal-sized matrices $\boldsymbol{A}$ and $\boldsymbol{B}$. Function evaluation is also interpreted in an element-wise fashion. For example,

$$\Gamma\left(\left[\begin{array}{c}7\\3\\9\end{array}\right]\right)\equiv\left[\begin{array}{c}\Gamma(7)\\\Gamma(3)\\\Gamma(9)\end{array}\right].$$

The digamma function is given by $\text{digamma}(x)\equiv\frac{d}{dx}\log\{\Gamma(x)\}$. Most of the updates in Algorithm 1 require standard arithmetic. The exception is the function $\mathcal{H}$ defined by (12), and its logarithm is evaluated using efficient quadrature strategies as described in Appendix B of Wand et al. (2011).

An effective stopping rule for Algorithm 1 is when the absolute relative change in $\log\{\underline{p}(\boldsymbol{y};q)\}$ falls below a small tolerance threshold such as 10^{-10}. Menictas and Wand (2015) conducted extensive checks on this strategy for similar non-conjugate MFVB algorithms and found it to have excellent performance in terms of convergence to the global maximum of $\underline{p}(\boldsymbol{y};q)$.

4.4 Limitations of the Variational Inference Scheme

Mean field restrictions such as (8) impose limitations on inferential accuracy since the posterior dependencies between parameters appearing in different q-density factors are ignored, and replaced by posterior independencies for convenience. For example, in the Bayesian Poisson mixed model

$$y_{ij}|\beta_0,U_i\overset{\text{ind.}}{\sim}\text{Poisson}\{\exp(\beta_0+U_i)\},\quad U_i|\,\sigma^2\overset{\text{ind.}}{\sim}N(0,\sigma^2),$$

the overdispersion in y_{ij} compared with the Poisson regression model is the variance ratio

$$\exp(\sigma^2/2)+\exp(\beta_0+\sigma^2)\,\{\exp(\sigma^2)-1\},$$

a function of both β_0 and σ^2. Variational inference for this quantity will be adversely affected by restriction (8) since it imposes, for example, $\text{cov}_q(\beta_0,\sigma^2\,|\,\boldsymbol{y})=0$ while the actual posterior correlation is non-zero. The practical effects of ignoring such posterior dependencies can be assessed via simulation studies, with MCMC as a benchmark. Such studies are described in Section 6 and show that accuracy of the inference produced by Algorithm 1 is good to excellent for the quantities of primary interest in semiparametric regression.

5 Real-Time Count Response Semiparametric Regression

An advantage of MFVB approaches to approximate inference is their adaptability to real-time processing. As discussed in Section 1, this is important for both high volume and/or velocity data. Here we briefly present an adaptation of the Poisson component of Algorithm 1 that permits a version of real-time count response semiparametric regression.

Algorithm 1 Non-conjugate MFVB algorithm for approximate inference in either the Poisson response model (1) or the Negative Binomial response model (2).

```
Initialize: $\mu_{q(1/\sigma_\ell^2)} > 0\,(1 \le \ell \le r), \mu_{q(\kappa)} > 0, \mu_{q(a_\kappa)} > 0, \boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}$ a $P \times 1$ vector
and $\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}$ a $P \times P$ symmetric positive definite matrix.
Cycle:
    $\boldsymbol{M}_{q(1/\boldsymbol{\sigma}^2)} \leftarrow \text{blockdiag}(\sigma_\beta^{-2}\boldsymbol{I}_p, \mu_{q(1/\sigma_1^2)}\boldsymbol{I}_{K_1}, \ldots, \mu_{q(1/\sigma_r^2)}\boldsymbol{I}_{K_r})$
    If fitting the Poisson response model (1):
        $\boldsymbol{w}_{q(\boldsymbol{\beta},\boldsymbol{u})} \leftarrow \exp\{\boldsymbol{C}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})} + \frac{1}{2}\text{diagonal}(\boldsymbol{C}\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}\boldsymbol{C}^T)\}$
        $\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})} \leftarrow \left\{\boldsymbol{C}^T\text{diag}(\boldsymbol{w}_{q(\boldsymbol{\beta},\boldsymbol{u})})\,\boldsymbol{C} + \boldsymbol{M}_{q(1/\boldsymbol{\sigma}^2)}\right\}^{-1}$
        $\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})} \leftarrow \boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})} + \boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}\Big\{\boldsymbol{C}^T\left(\boldsymbol{y} - \boldsymbol{w}_{q(\boldsymbol{\beta},\boldsymbol{u})}\right)$
                        $-\boldsymbol{M}_{q(1/\boldsymbol{\sigma}^2)}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}\Big\}$
    If fitting the Negative Binomial response model (2):
        $\boldsymbol{w}_{q(\boldsymbol{\beta},\boldsymbol{u})} \leftarrow \exp\{-\boldsymbol{C}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})} + \frac{1}{2}\text{diagonal}(\boldsymbol{C}\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}\boldsymbol{C}^T)\}$
        $\boldsymbol{\mu}_{q(\boldsymbol{g})} \leftarrow (\mu_{q(\kappa)}\boldsymbol{1} + \boldsymbol{y})/(\boldsymbol{1} + \mu_{q(\kappa)}\boldsymbol{w}_{q(\boldsymbol{\beta},\boldsymbol{u})})$
        $\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})} \leftarrow \left\{\mu_{q(\kappa)}\boldsymbol{C}^T\text{diag}(\boldsymbol{\mu}_{q(\boldsymbol{g})} \odot \boldsymbol{w}_{q(\boldsymbol{\beta},\boldsymbol{u})})\,\boldsymbol{C} + \boldsymbol{M}_{q(1/\boldsymbol{\sigma}^2)}\right\}^{-1}$
        $\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})} \leftarrow \boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})} + \boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}\Big\{\mu_{q(\kappa)}\boldsymbol{C}^T\left(\boldsymbol{\mu}_{q(\boldsymbol{g})} \odot \boldsymbol{w}_{q(\boldsymbol{\beta},\boldsymbol{u})} - \boldsymbol{1}\right)$
                        $-\boldsymbol{M}_{q(1/\boldsymbol{\sigma}^2)}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}\Big\}$
        $\boldsymbol{\mu}_{q(\log(\boldsymbol{g}))} \leftarrow \text{digamma}(\boldsymbol{1}\mu_{q(\kappa)} + \boldsymbol{y}) - \log(\boldsymbol{1} + \mu_{q(\kappa)}\boldsymbol{w}_{q(\boldsymbol{\beta},\boldsymbol{u})})$
        $C_1 \leftarrow \boldsymbol{1}^T\boldsymbol{C}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})} - \boldsymbol{1}^T\boldsymbol{\mu}_{q(\log(\boldsymbol{g}))} + \boldsymbol{\mu}_{q(\boldsymbol{g})}^T\boldsymbol{w}_{q(\boldsymbol{\beta},\boldsymbol{u})} + \mu_{q(1/a_\kappa)}$
        $\mu_{q(\kappa)} \leftarrow \exp\left[\log\left\{\mathcal{H}(\frac{1}{2}, 0, 1, n, C_1)\right\} - \log\left\{\mathcal{H}(-\frac{1}{2}, 0, 1, n, C_1)\right\}\right]$
        $\mu_{q(1/a_\kappa)} \leftarrow 1/(\mu_{q(\kappa)} + M_\kappa)$
    For $\ell = 1, \ldots, r$ :
      $\mu_{q(1/a_\ell)} \leftarrow 1/(\mu_{q(1/\sigma_\ell^2)} + A_\ell^{-2})$
      $\mu_{q(1/\sigma_\ell^2)} \leftarrow \dfrac{K_\ell + 1}{2\,\mu_{q(1/a_\ell)} + \|\boldsymbol{\mu}_{q(\boldsymbol{u}_\ell)}\|^2 + \text{tr}(\boldsymbol{\Sigma}_{q(\boldsymbol{u}_\ell)})}$
until the absolute relative change in $\log\{\underline{p}(\boldsymbol{y}; q)\}$ is negligible.
```

Despite the switch to real-time processing in this section, we are not modifying our semiparametric regression models and the model parameters are assumed to remain fixed. The extension to scenarios where the model parameters drift over time is certainly worth investigating, but not within the scope of the current article.

Rather than processing $\boldsymbol{y}$ and $\boldsymbol{C}$ in batch, as done by Algorithm 1, Algorithm 2 processes each new entry of $\boldsymbol{y}$, denoted by y_{new}, and its corresponding row of $\boldsymbol{C}$, denoted by $\boldsymbol{c}_{\text{new}}$, sequentially in real time. Algorithm 2 is partially online in comparison with the real-time semiparametric regression algorithms in Luts et al. (2014). It shares the advantage of the Luts et al. (2014) algorithms of not requiring any iteration in Step 3 whenever a new observation arrives. However, it does involve full passes through the full current predictor data to compute $\boldsymbol{w}_{q(\boldsymbol{\beta},\boldsymbol{u})}$, $\boldsymbol{C}^T\boldsymbol{w}_{q(\boldsymbol{\beta},\boldsymbol{u})}$ and $\boldsymbol{C}^T\text{diag}(\boldsymbol{w}_{q(\boldsymbol{\beta},\boldsymbol{u})})\,\boldsymbol{C}$. Storage of $\boldsymbol{C}$ is also required. Alleviation of these aspects for Poisson semiparametric regression remains an open problem.

Luts et al. (2014) stress the importance of batch runs for determination of starting values for real-time semiparametric regression procedures and their Algorithm 2' formalized such a strategy and this is reflected in Algorithm 2.

Algorithm 2 Online non-conjugate variational message passing algorithm for real-time approximate inference in the Poisson response model (1).

```
1. Use Algorithm 1 to perform batch-based tuning runs, analogous to those described in Algorithm 2' of Luts et al. (2014), and determine a warm-up sample size $n_{\text{warm}}$ for which convergence is validated.
2. Set $\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}$, $\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}$ and $\mu_{q(1/\sigma_1^2)},\ldots,\mu_{q(1/\sigma_r^2)}$ to be the values for these quantities obtained in the batch-based tuning run with sample size $n_{\text{warm}}$. Then set $\boldsymbol{y}_{\text{warm}}$ and $\boldsymbol{C}_{\text{warm}}$ to be the response vector and design matrix based on the first $n_{\text{warm}}$ observations. Put $\boldsymbol{C}^T\boldsymbol{y} \leftarrow \boldsymbol{C}_{\text{warm}}^T\boldsymbol{y}_{\text{warm}}$ and $n \leftarrow n_{\text{warm}}$.
3. Cycle:
       Read in $y_{\text{new}}$ $(1\times 1)$ and $\boldsymbol{c}_{\text{new}}$ $(P\times 1)$   ;   $n \leftarrow n+1$
       $\boldsymbol{M}_{q(1/\boldsymbol{\sigma}^2)} \leftarrow \text{blockdiag}(\sigma_\beta^{-2}\boldsymbol{I}_p, \mu_{q(1/\sigma_1^2)}\boldsymbol{I}_{K_1},\ldots,\mu_{q(1/\sigma_r^2)}\boldsymbol{I}_{K_r})$
       $\boldsymbol{C}^T\boldsymbol{y} \leftarrow \boldsymbol{C}^T\boldsymbol{y} + \boldsymbol{c}_{\text{new}}y_{\text{new}}$   ;   $\boldsymbol{C} \leftarrow [\boldsymbol{C}^T\ \boldsymbol{c}_{\text{new}}]^T$
       $\boldsymbol{w}_{q(\boldsymbol{\beta},\boldsymbol{u})} \leftarrow \exp\{\boldsymbol{C}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})} + \tfrac{1}{2}\text{diagonal}(\boldsymbol{C}\,\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}\boldsymbol{C}^T)\}$
       $\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})} \leftarrow \boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})} + \boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}\Big\{\boldsymbol{C}^T\boldsymbol{y} - \boldsymbol{C}^T\boldsymbol{w}_{q(\boldsymbol{\beta},\boldsymbol{u})} - \boldsymbol{M}_{q(1/\boldsymbol{\sigma}^2)}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}\Big\}$
       $\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})} \leftarrow \Big\{\boldsymbol{C}^T\text{diag}(\boldsymbol{w}_{q(\boldsymbol{\beta},\boldsymbol{u})})\,\boldsymbol{C} + \boldsymbol{M}_{q(1/\boldsymbol{\sigma}^2)}\Big\}^{-1}$
       For $\ell = 1,\ldots,r$ :
           $\mu_{q(1/a_\ell)} \leftarrow 1/\{\mu_{q(1/\sigma_\ell^2)} + A_\ell^{-2}\}$
           $\mu_{q(1/\sigma_\ell^2)} \leftarrow \dfrac{K_\ell + 1}{2\,\mu_{q(1/a_\ell)} + \|\boldsymbol{\mu}_{q(\boldsymbol{u}_\ell)}\|^2 + \text{tr}(\boldsymbol{\Sigma}_{q(\boldsymbol{u}_\ell)})}$
   until data no longer available or analysis terminated.
```

An illustration of Algorithm 2 and assessment of its efficacy is described in Section 6.3.

Semiparametric regression for streaming data is a new area of research and the MFVB approach is one of several approaches that could be contemplated. Some comparative advantages of the MFVB approach are described in Section 4 of Luts et al. (2014). Sequential Monte Carlo is an alternative Bayesian inference approach that is amenable to real-time fitting and inference (e.g., Chopin, Jacob, and Papaspiliopoulos

(2013)). In the case of batch fitting, Fan, Leslie, and Wand (2008) explored the use of sequential Monte Carlo for generalized response semiparametric regression models but found them difficult to tune. Our experiences to date point to MFVB being the most promising approach for effective real-time semiparametric regression.

6 Numerical Results

Algorithms 1 and 2 have been tested on various synthetic and actual data-sets. We first describe the results of a simulation study that allows us to make some summaries of the accuracy of MFVB in this context. This is followed by some applications. Lastly, we describe an illustration of Algorithm 2 that takes the form of a movie on our real-time semiparametric regression web-site.

6.1 Simulation Study

We ran a simulation study involving the true mean function

$$f(x_1,x_2)\equiv\exp\{f_1(x_1)+f_2(x_2)\}$$

where

$$\begin{aligned} f_1(x) &\equiv \cos(4\pi x)\ +2\,x, \\ f_2(x) &\equiv 0.4\,\phi(x;0.38,0.08)-1.02\,x+0.018\,x^2+0.08\,\phi(x;0.75,0.03) \end{aligned}$$

and $\phi(x;\mu,\sigma)$ denotes the value of the Normal density function with mean μ and standard deviation σ evaluated at x. Next, we generated 100 data-sets, each having 500 triplets (y_i,x_{1i},x_{2i}), using the Poisson response model

$$y_i\overset{\text{ind.}}{\sim}\text{Poisson}\{f(x_{1i},x_{2i})\},\quad 1\le i\le 500, \tag{14}$$

and the Negative Binomial response model

$$y_i\overset{\text{ind.}}{\sim}\text{Negative-Binomial}\{f(x_{1i},x_{2i}),3.8\},\quad 1\le i\le 500, \tag{15}$$

where $x_{1i},x_{2i}\overset{\text{ind.}}{\sim}\text{Uniform}(0,1)$. We model $f_1(x_1)+f_2(x_2)$ using mixed model-based penalized splines (e.g., Ruppert et al., 2003):

$$\begin{gathered} \beta_0+\beta_1\,x_1+\beta_2\,x_2+\textstyle\sum_{k=1}^{K_1}u_{1k}\,z_{1k}(x_1)+\sum_{k=1}^{K_2}u_{2k}\,z_{2k}(x_2), \\ u_{1k}|\,\sigma_1^2\overset{\text{ind.}}{\sim}N(0,\sigma_1^2),\quad u_{2k}|\,\sigma_2^2\overset{\text{ind.}}{\sim}N(0,\sigma_2^2), \end{gathered} \tag{16}$$

where z_{1k} and z_{2k} represent O'Sullivan splines (Wand and Ormerod, 2008). After grouping $\boldsymbol{\beta}=[\beta_0\,\beta_1\,\beta_2]^T$, $\boldsymbol{u}=[u_{11},\ldots,u_{1K_1},u_{21},\ldots,u_{2K_2}]^T$ and creating the corresponding design matrices $\boldsymbol{X}$ and $\boldsymbol{Z}$, Algorithm 1 is used for MFVB inference. We set the number of spline basis functions to be $K_1=K_2=17$. The MFVB iterations were terminated when the relative change in $\log\underline{p}(\boldsymbol{y};q)$ was less than 10^{-10}.

The MCMC analyses involved generation of samples of size 10,000 and then the first 5,000 values being discarded as burn-in. Thinning with a factor of 5 was applied to the remaining samples, which resulted in MCMC samples of size 1,000 being retained for inference. All MCMC sampling was performed using BUGS (Spiegelhalter et al., 2003).

Accuracy Assessment

Figure 2 displays side-by-side boxplots of the accuracy scores for the parameters in the Poisson response simulation study. For a generic parameter θ, the accuracy score is defined by

$$\text{accuracy}(q^*) = 100\left(1 - \frac{1}{2}\int_{-\infty}^{\infty} |\, q^*(\theta) - p(\theta|\boldsymbol{y})|\, d\theta\right)\%.$$

Note that a kernel density estimate based on the MCMC samples is used to estimate the posterior density function $p(\theta|\boldsymbol{y})$.

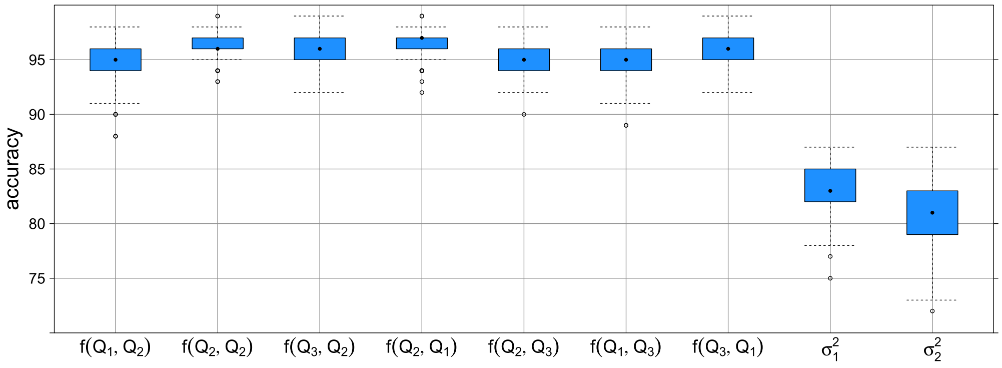

Figure 2: Side-by-side boxplots of accuracy values for MFVB against an MCMC benchmark for the Poisson response model (14).

The parameters on the horizontal axis of Figure 2 represent the estimated approximate posterior density functions for f, evaluated at the sample quartiles of the x_{1i}s and the x_{2i}s. We use Q_1, Q_2 and Q_3, generically, to denote these sample quartiles. For example, $f(Q_1, Q_2)$ denotes f evaluated at the first sample quartile of the x_{1i}s and the second sample quartile of the x_{2i}s. Also shown are the estimated approximate posterior density functions for σ_1^2 and σ_2^2. The boxplots indicate that the accuracies for $f(x_1, x_2)$ are around 95%, while accuracies between 80% and 85% are obtained for the variance parameters.

Figure 3 shows the MFVB-based approximate posterior density functions against the MCMC result for a single replicated data-set. The accuracy of MFVB is excellent for the $f(x_1, x_2)$ approximate posterior density functions.

Figure 4 displays side-by-side boxplots of the accuracies for the 100 data-sets generated according to the Negative Binomial response model (15). The parameters on the

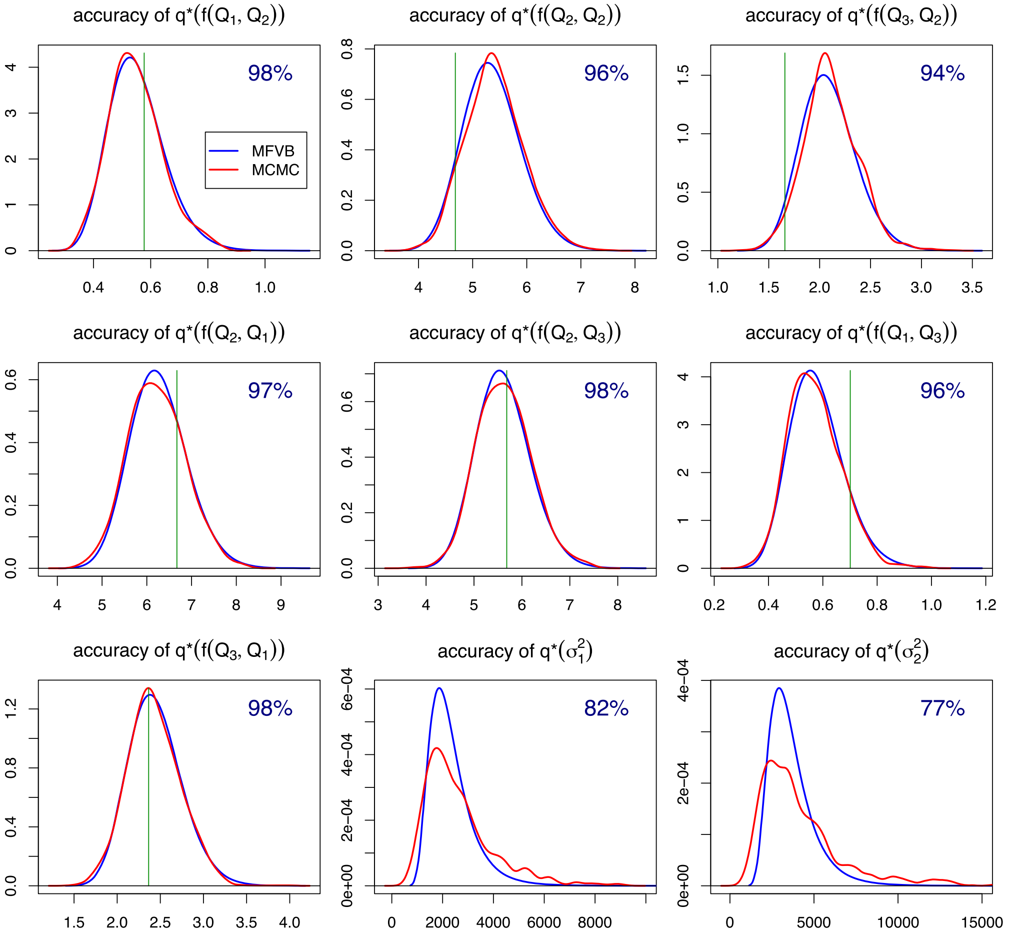

Figure 3: Approximate posterior density functions for the Poisson response model (14). Vertical lines indicate the true values.

horizontal axis in Figure 4 have similar meanings as in Figure 2, but the result for the approximate posterior density function of κ is also included. Compared to the results for the Poisson case, the accuracies for the Negative Binomial response model are lower, but still attain good performance for $f(x_1, x_2)$ with approximately values between 70 and 90%. The majority of the accuracies for the variances σ_1^2 and σ_2^2 is around 70%, while lower accuracies are obtained for κ.

Finally, Figure 5 compares the approximate posterior density functions obtained using MFVB inference and the MCMC result for a single replicated data-set. MFVB attains particularly good accuracies for the $f(x_1, x_2)$ approximate posterior density functions.

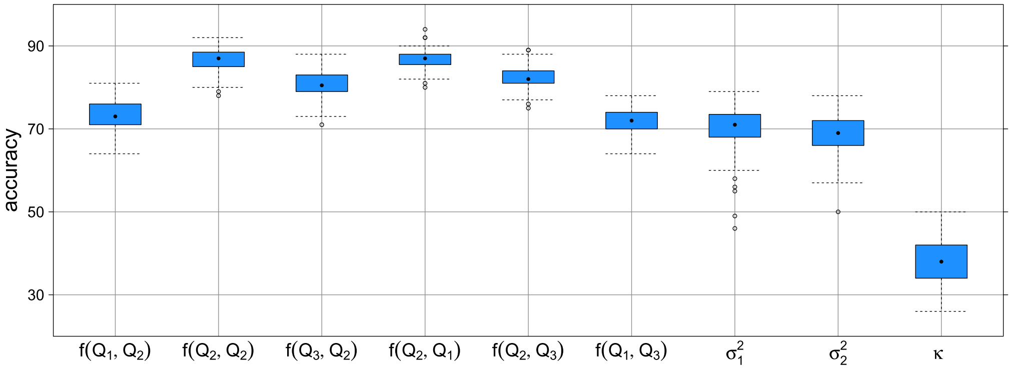

Figure 4: Side-by-side boxplots of accuracy values for MFVB against an MCMC benchmark for Negative Binomial response model (15).

Function Estimation Accuracy and Spline Basis Sensitivity

Next we investigate the accuracy of the MFVB estimates of f_1 and f_2, both in terms of closeness to the MCMC estimates, and the true functions. We also conduct a sensitivity check on the number of spline basis functions.

Figure 6 shows the functions

$$f_1(x_1) + f_2(\overline{x}_2) \text{ versus } x_1 \quad \text{and} \quad f_1(\overline{x}_1) + f_2(x_2) \text{ versus } x_2$$

where $\overline{x}_1$ is the sample mean of the x_{1i}s and $\overline{x}_2$ is the sample mean of the x_{2i}s. Also shown are MFVB and MCMC Bayes estimates of these functions and corresponding 95% pointwise credible sets for the first three replications of the Poisson additive model simulation study. These estimates were obtained for both $K_1 = K_2 = 17$, as in the simulation study, and then again with $K_1 = K_2 = 34$. All estimates and corresponding credible sets are displayed in Figure 6.

Firstly, we see that there is very close correspondence between the MFVB and MCMC estimates and credible sets, showing very high accuracy of MFVB in these cases. Also, estimation of f_1 and f_2 is shown to be very good. Lastly, there is very low sensitivity to the number of spline basis functions for penalized splines in this example, with $K_1 = K_2 = 17$ shown to be adequate. Theoretical underpinning for this last aspect is given in Li and Ruppert (2008).

Computational Cost

Table 1 summarizes the computation times for MCMC and MFVB fitting in case of the Poisson and Negative Binomial simulation study as run using an Intel Core i7 2.66 GHz processor with 4 GB of random access memory. Full details of the MCMC fitting are given in Section 6.1. The MFVB fitting was done in R (R Development Core Team,

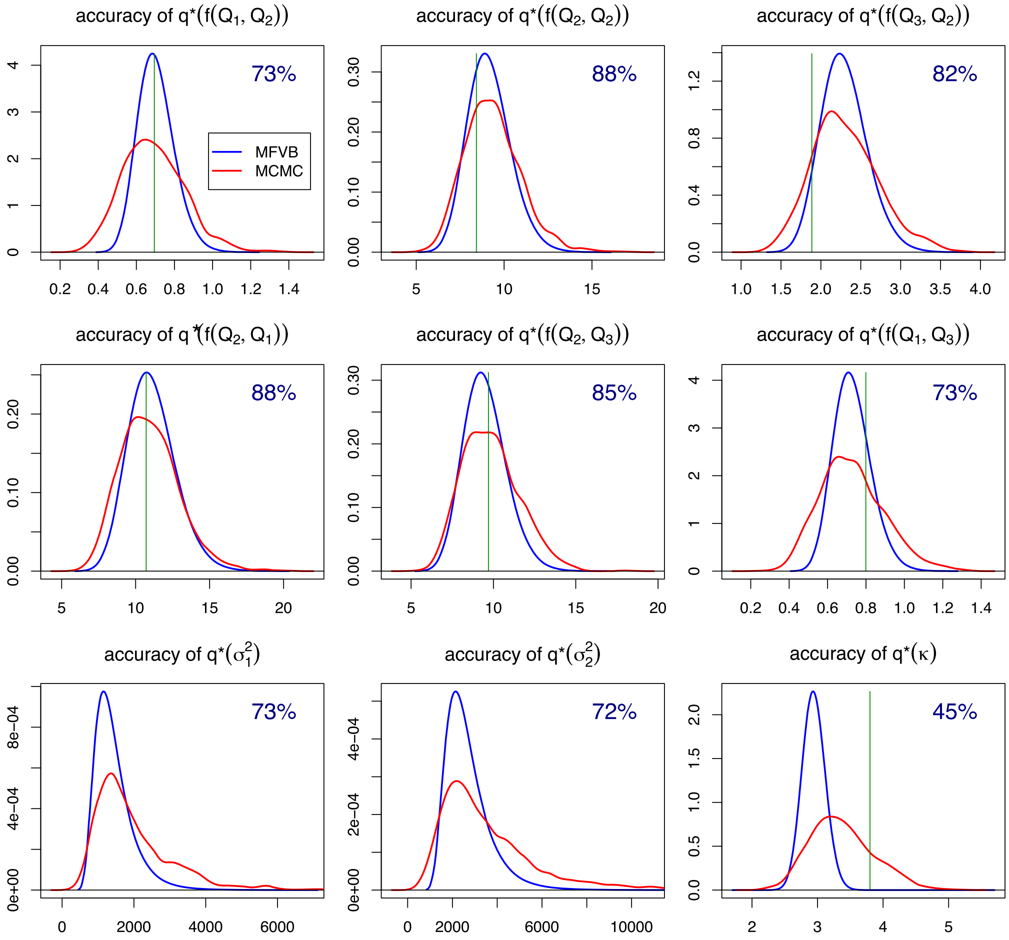

Figure 5: Approximate posterior density functions for Negative Binomial response model (15). Vertical lines indicate the true values.

	MCMC	MFVB
Poisson response model	180.23 (15.76)	0.96 (0.58)
Negative Binomial response model	224.95 (13.7)	5.55 (1.30)

Table 1: Average (standard deviation) times in seconds for MCMC and MFVB inference based on the simulation study described in Section 6.1.

2015). Timing comparisons between MCMC and MFVB are inherently difficult due to differences in stopping rules and, in this case, programming languages. Despite this difficulty, Table 1 is meaningful in that it allows appreciation and comparison of the time

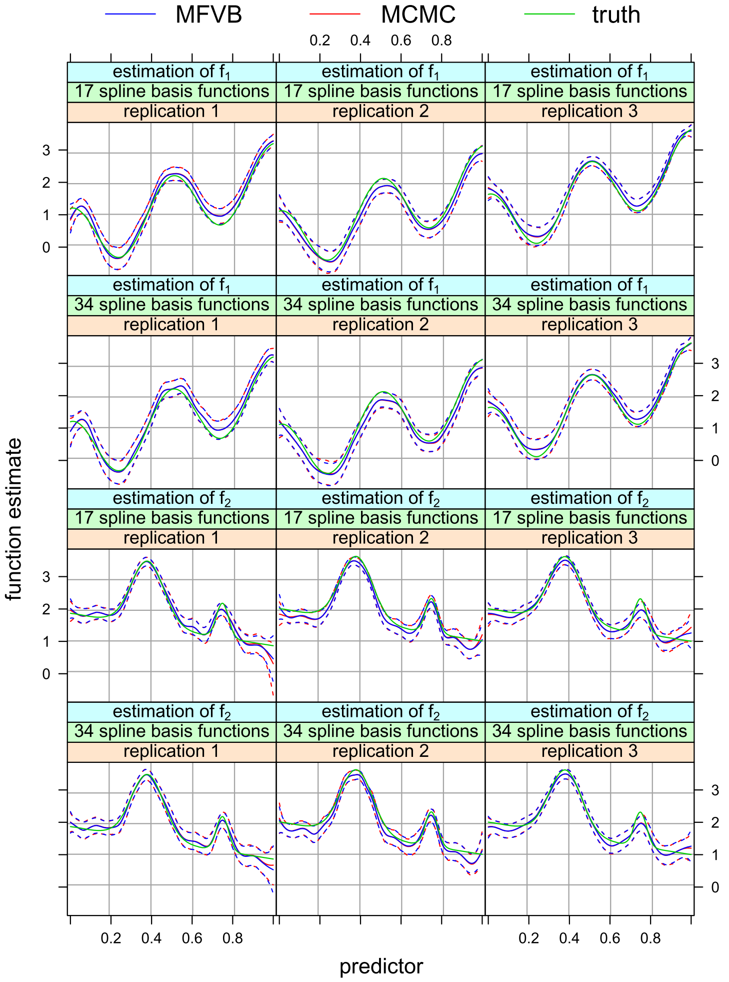

Figure 6: Function estimation accuracy and spline basis sensitivity analyses for the Poisson additive model simulation study. The upper panels show $f_1(x_1) + f_2(\overline{x}_2)$ versus x_1, its MFVB-based and MCMC-based Bayes estimates (solid curves) and corresponding 95% credible sets for two spline bases, one with 17 basis functions and the other with 34 basis functions, for data from the first three replications of the simulation study. The lower panels are analogous, but for estimation of $f_1(\overline{x}_1) + f_2(x_2)$ versus x_2.

taken for an existing popular Bayes computing approach with a new one implemented in the most common statistical programming language. The average computing time for MFVB is considerably lower than that of MCMC. Nevertheless, the speed gains of MFVB need to be traded off against accuracy losses as shown in Figures 2 and 4.

6.2 Applications

We now present some applications involving each of models (1) and (2) in turn.

North African Conflict Data

We fitted the Poisson response model (1) using Algorithm 1 to a data-set extracted from the Global Database of Events, Language and Tone (Leetaru and Schrodt, 2013). This database contains more than 200 million geo-located events, obtained from news reports, with global coverage between early 1979 and June 2012. For this example we extracted the daily number of material conflicts for each African country for the period September 2010 to June 2012. Our model is

$$\texttt{conflicts}_{ij}\,|\,\boldsymbol{\beta},\boldsymbol{u}_1,U_i \overset{\text{ind.}}{\sim} \text{Poisson}[\exp\{\beta_0+f_1(\texttt{time}_j)+U_i\}],$$

with $\texttt{conflicts}_{ij}$ the number of news reports about material conflicts for country i on date j, $\texttt{time}_j$ the time in days for date j starting from September 1, 2010 and U_i the random intercept for country i, $1 \le i \le 54$. The total number of observations for all African countries is $n = 36,126$. Note that 20 spline basis functions were used for modeling f_1.

Figure 7 shows the estimate for $\exp\{\beta_0+f_1(\texttt{time}_j)\}$ and corresponding 95% pointwise credible sets. The strong increase, starting around December 2010, in number of news reports about material conflicts coincides with the Arab Spring demonstrations and civil wars which took place in several African countries as Mauritania, Western Sahara, Morocco, Algeria, Tunisia, Libya, Egypt, Sudan, Djibouti and the related crisis in Mali. In addition, 95% credible sets for the estimates of $\exp(U_i)$ are plotted for the 15 countries with the largest random intercept estimates, i.e., showing larger numbers of material conflict-related news reports. Despite the size of the data set and model, fitting via Algorithm 1 took about 3 minutes.

Adduct Data

Illustrations of Negative Binomial semiparametric regression models have previously been given in Thurston et al. (2000) and Marley and Wand (2010) using data on adduct counts, which are carcinogen-DNA complexes, and smoking variables for 78 former smokers in the lung cancer study (Wiencke et al., 1999). Here we use Algorithm 1 to fit a version of the Bayesian penalized model that Marley and Wand (2010) fitted via MCMC.

Thurston et al. (2000) and Marley and Wand (2010) considered Negative Binomial additive models of the form:

$$\begin{aligned}\texttt{adducts}_i\,|\,\boldsymbol{\beta},\boldsymbol{u}_1,\boldsymbol{u}_2,\boldsymbol{u}_3,\boldsymbol{u}_4,\kappa \overset{\text{ind.}}{\sim}\;& \text{Negative-Binomial}(\exp\{\beta_0+f_1(\texttt{ageInit}_i)\\ &+f_2(\texttt{yearsSmoking}_i)\\ &+f_3(\texttt{yearsSinceQuit}_i)\\ &+f_4(\texttt{cigsPerDay}_i)\},\kappa),\end{aligned} \tag{17}$$

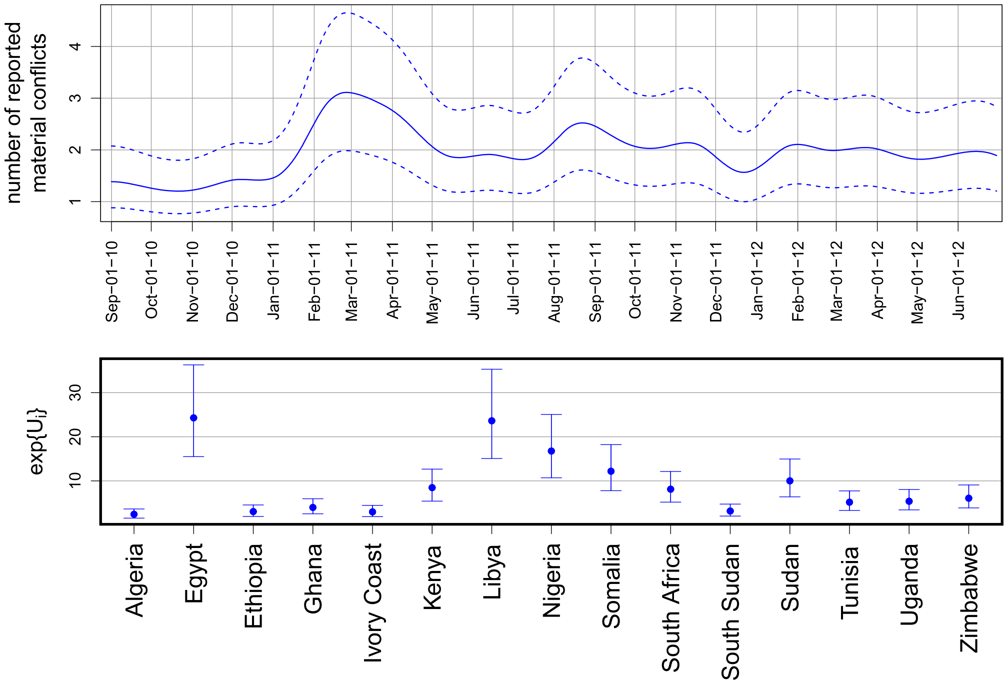

Figure 7: Poisson regression result using MFVB inference for global data on events, location and tone database. The solid curve in the top panel corresponds to the posterior means of the response, given the value of the time variable, and the dashed curves are pointwise 95% credible sets. The lower panel shows 95% credible sets for the estimates of $\exp(U_i)$ for the 15 countries with highest posterior means.

with $\texttt{ageInit}_i$ the age of smoking initiation, $\texttt{yearsSmoking}_i$ the number of years of smoking, $\texttt{yearsSinceQuit}_i$ the number of years since quitting and $\texttt{cigsPerDay}_i$ the number of cigarettes smoked per day for subject i. The f_ℓ, $1 \leq \ell \leq 4$, are modeled using mixed-model based penalized splines as in (16), with 20 basis functions each.

Figure 8 displays the fitted functions for model (17). Marley and Wand (2010) reported slow MCMC convergence for this model, so we used burn-in size of 100,000, a retained sample size of 50,000, and a thinning factor of 50. The MCMC-based fits are added as a reference to Figure 8.

Fitting of (17) via Algorithm 1 took 2 minutes whilst MCMC fitting in BUGS took 1 hour and 28 minutes. As indicated by Figure 8, the much faster MFVB estimates are quite close to the more accurate MCMC estimates.

We investigated the sensitivity of the hyperparameter M_κ for this example. The analysis summarized in Figure 8 was re-run with

$$M_\kappa \text{ taking values in } \{0.01, 0.1, 1, 10, 100\}. \tag{18}$$

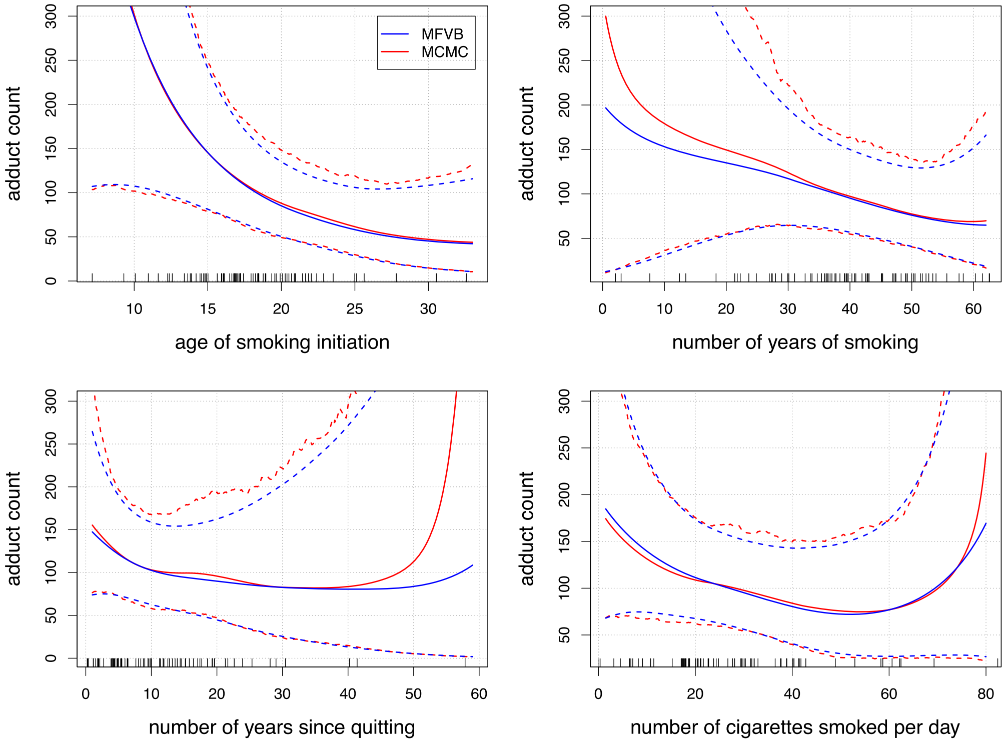

Figure 8: Negative Binomial additive model fits, using MFVB and MCMC inference, for the adduct data. Solid curves are posterior means for fitted functions while dashed curves are corresponding pointwise 95% credible sets.

We kept track of the posterior means and 95% credible sets for $f_j(Q_1)$, $f_j(Q_2)$ and $f_j(Q_3)$, $1 \le j \le 4$, obtained from Algorithm 1. Figure 9 compares the results across the hyperparameter values (18). Very low sensitivity is apparent.

6.3 Real-Time Poisson Nonparametric Regression Movie

The web-site `realtime-semiparametric-regression.net` contains a movie that illustrates Algorithm 2 in the special case of Poisson nonparametric regression with $r = 1$. The spline basis functions set-up is analogous to that given in (16).

The data are simulated according to

$$x_{\text{new}} \sim \text{Uniform}(0,1), \quad y_{\text{new}}|x_{\text{new}} \sim \text{Poisson}[\exp\{\cos(4\pi x_{\text{new}}) + 2\,x_{\text{new}}\}]$$

and the warm-up sample size is $n_{\text{warm}} = 100$. The movie is under the link titled `Poisson nonparametric regression` and shows the efficacy of Algorithm 2 for recovery of the underlying mean function in real time.

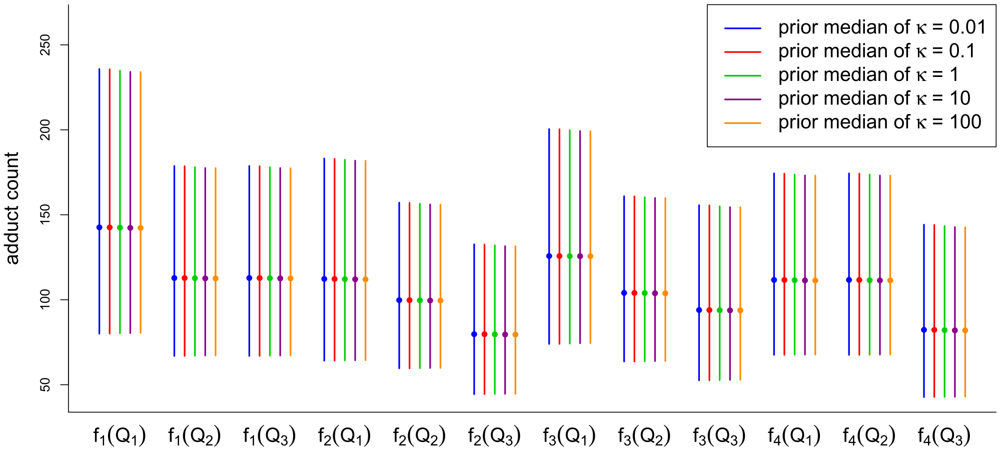

Figure 9: Sensitivity analysis of the hyperparameter M_κ for the Negative Binomial additive model analysis of the adduct data with the prior median of $p(\kappa)$ taking values in $\{0.01, 0.1, 1, 10, 100\}$. The parameters are $f_j(Q_1)$, $f_j(Q_2)$ and $f_j(Q_3)$, $1 \le j \le 4$. The line segments correspond to approximate 95% credible sets and the points correspond to posterior means.

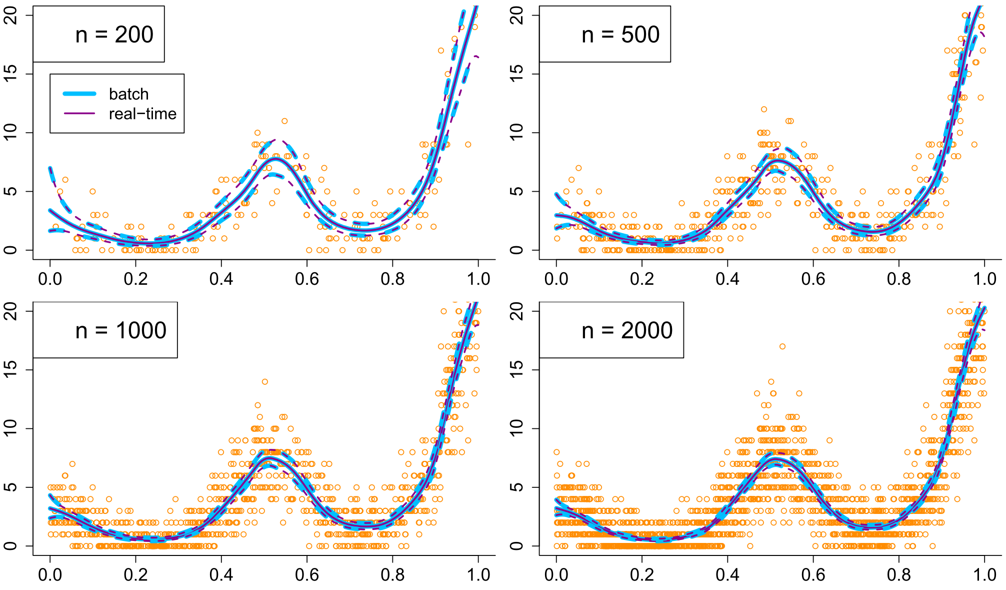

Figure 10: Verification that the fits in the real-time Poisson nonparametric regression movie are close to those obtained via batch fitting. In each panel, the fits from the real-time Algorithm 2 are compared with those by feeding the same data into the batch counterpart Algorithm 1, for four different sample sizes.

Figure 10 indicates excellent correspondence between the fits produced by the real-time Algorithm 2 and its batch counterpart, Algorithm 1. We re-ran this test over several random seeds and found this excellence correspondence to persist.

7 Conclusion

Our MFVB algorithms, based on non-conjugate variational message passing, have been demonstrated to result in good to excellent inference for parameters of interest in count response semiparametric regression. MCMC is more accurate, but does not scale as well to high volume/velocity data. Depending on the application, both approaches are likely to have a place in future analyses. The new MFVB approach is an option when MCMC becomes infeasible.

Appendix A: Background Material

The specification of the models and their fitting via variational algorithms requires several definitions and results, and are provided in this section.

A.1 Distributional Definitions

Table 2 lists all distributions used in this article. In particular, the parametrizations of the corresponding density functions and probability functions are provided.

A.2 Distributional Results

The variational inference algorithms given in Section 4 make use of the following distributional results:

Result 1. *Let x and a be random variables such that*

$$x|\,a \sim \text{Poisson}(a) \quad \textit{and} \quad a \sim \text{Gamma}(\kappa, \kappa/\mu).$$

Then $x \sim$ Negative-Binomial(μ, κ).

Result 2. *Let x and a be random variables such that*

$$x|\,a \sim \text{Inverse-Gamma}(1/2, 1/a) \quad \textit{and} \quad a \sim \text{Inverse-Gamma}(\tfrac{1}{2}, 1/A^2).$$

Then $\sqrt{x} \sim$ Half-Cauchy(A).

Result 1 is a relatively well-known distribution-theoretic result (e.g., Lawless, 1987). Result 2 is related to established results concerning the F distribution family, and this particular version is taken from Wand et al. (2011).

distribution	density/probability function in x	abbreviation
Poisson	$\lambda^x\, e^{-\lambda}/x!;\quad x=0,1,\ldots$	Poisson(λ)
Negative Binomial	$\dfrac{\kappa^\kappa\Gamma(x+\kappa)\mu^x}{\Gamma(\kappa)(\kappa+\mu)\Gamma(x+1)};\ x=0,1\ldots$ $\kappa,\mu>0$	Negative-Binomial(μ,κ)
Uniform	$1/(b-a);\quad a<x<b$	Uniform(a,b)
Multivariate Normal	$\lvert 2\pi\boldsymbol{\Sigma}\rvert^{-1/2}\exp\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T$ $\times\,\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\}$	$N(\boldsymbol{\mu},\boldsymbol{\Sigma})$
Gamma	$\dfrac{B^A\,x^{A-1}e^{-B\,x}}{\Gamma(A)};\quad x>0;\ A,B>0$	Gamma(A,B)
Inverse-Gamma	$\dfrac{B^A\,x^{-A-1}e^{-B/x}}{\Gamma(A)};\quad x>0;\ A,B>0$	Inverse-Gamma(A,B)
Half-Cauchy	$\dfrac{2}{\pi\sigma((x/\sigma)^2+1)};\quad x>0;\ \sigma>0$	Half-Cauchy(σ)
$F_{1,1}$	$\dfrac{1}{\pi\sigma\sqrt{x/\sigma}(1+x/\sigma)};\quad x>0;\ \sigma>0$	$F_{1,1}(\sigma)$

Table 2: Distributions used in this article and their corresponding density/probability functions.

A.3 Non-conjugate Variational Message Passing

Non-conjugate variational message passing (Knowles and Minka, 2011) is an extension of MFVB. It can yield tractable variational approximate inference in situations where ordinary MFVB is intractable and is a variant of earlier contributions of the same type by Barber and Bishop (1998) and Honkela et al. (2010). The Knowles and Minka (2011) approach yields particularly simple updates, based on fixed-point iteration, so we focus on their version here.

MFVB relies on approximating the joint posterior density function $p(\boldsymbol{\theta}|\boldsymbol{y})$ by a product form $q(\boldsymbol{\theta})=\prod_{i=1}^d q(\boldsymbol{\theta}_i)$, where $\boldsymbol{\theta}$ corresponds to the hidden nodes in Figure 1. The optimal q-density functions, denoted by $q^*(\boldsymbol{\theta}_i)$, are those that minimize the Kullback–Leibler divergence

$$\int q(\boldsymbol{\theta})\log\left\{\frac{q(\boldsymbol{\theta})}{p(\boldsymbol{\theta}|\boldsymbol{y})}\right\}d\boldsymbol{\theta}.$$

An equivalent optimization problem represents maximizing the lower bound on the marginal likelihood $p(\boldsymbol{y})$:

$$\underline{p}(\boldsymbol{y};q) \equiv \exp\left[\int q(\boldsymbol{\theta}) \log\left\{\frac{p(\boldsymbol{\theta},\boldsymbol{y})}{q(\boldsymbol{\theta})}\right\} d\boldsymbol{\theta}\right].$$

The optimal q-density functions can be shown to satisfy

$$q^*(\boldsymbol{\theta}_i) \propto \exp\left[E_{-\boldsymbol{\theta}_i}\left\{\log p(\boldsymbol{\theta}_i|\text{rest})\right\}\right], \quad 1 \le i \le d,$$

where $E_{-\boldsymbol{\theta}_i}$ denotes expectation with respect to the density $\prod_{j\neq i} q_j(\boldsymbol{\theta}_j)$ and 'rest' denotes all random variables in the model other than $\boldsymbol{\theta}_i$.

In the event that one of the $E_{-\boldsymbol{\theta}_i}\{\log p(\boldsymbol{\theta}_i|\text{rest})\}$ is not tractable, say the one corresponding to $q(\boldsymbol{\theta}_j)$ for some $j \in \{1,\ldots,d\}$, non-conjugate variational message passing offers a way out (Knowles and Minka, 2011). It first postulates that $q(\boldsymbol{\theta}_j)$ is an exponential family density function with natural parameter vector $\boldsymbol{\eta}_j$ and natural statistic $\boldsymbol{T}(\boldsymbol{\theta}_j)$. The optimal parameters are then obtained via updates of the form

$$\boldsymbol{\eta}_j \leftarrow \left\{\text{var}\left(\boldsymbol{T}(\boldsymbol{\theta}_j)\right)\right\}^{-1}\left[\mathsf{D}_{\boldsymbol{\eta_j}} E_{\boldsymbol{\theta}}\left\{\log p(\boldsymbol{\theta},\boldsymbol{y})\right\}\right], \tag{19}$$

where $\mathsf{D}_{\boldsymbol{x}} f$ is the derivative vector of f with respect to $\boldsymbol{x}$ and var$(\boldsymbol{v})$ denotes the covariance matrix of random vector $\boldsymbol{v}$ (Magnus and Neudecker, 1999). Wand (2014) derived fully simplified expressions for (19) in case $q(\boldsymbol{\theta}_j)$ has a Multivariate Normal density with mean $\boldsymbol{\mu}_{q(\theta_j)}$ and covariance matrix $\boldsymbol{\Sigma}_{q(\theta_j)}$. These are:

$$\begin{aligned} \boldsymbol{\Sigma}_{q(\theta_j)} &\leftarrow \left\{-2\,\text{vec}^{-1}\left(\left[\mathsf{D}_{\text{vec}(\boldsymbol{\Sigma})} E_{\boldsymbol{\theta}}\left\{\log p(\boldsymbol{\theta},\boldsymbol{y})\right\}\right]^T\right)\right\}^{-1}, \\ \boldsymbol{\mu}_{q(\theta_j)} &\leftarrow \boldsymbol{\mu}_{q(\theta_j)} + \boldsymbol{\Sigma}_{q(\theta_j)}\left[\mathsf{D}_{\boldsymbol{\mu}} E_{\boldsymbol{\theta}}\left\{\log p(\boldsymbol{\theta},\boldsymbol{y})\right\}\right]^T. \end{aligned} \tag{20}$$

Here vec$(\boldsymbol{A})$ denotes a vector formed by stacking the columns of matrix $\boldsymbol{A}$ underneath each other in order from left to right and vec$^{-1}(\boldsymbol{a})$ is a $d \times d$ matrix formed from listing the entries of the $d^2 \times 1$ vector $\boldsymbol{a}$ in a column-wise fashion in order from left to right.

Appendix B: Derivation of q^* Density Functions

B.1 Derivation of $q^*(a_\ell)$ and $q^*(\sigma_\ell^2)$ for the Poisson and Negative Binomial Response Model

Standard manipulations lead to the following full conditional distributions:

$$a_\ell|\text{rest} \overset{\text{ind.}}{\sim} \text{Inverse-Gamma}(1, \sigma_\ell^{-2} + A_\ell^{-2}) \;\; \text{and}$$

$$\sigma_\ell^2|\text{rest} \overset{\text{ind.}}{\sim} \text{Inverse-Gamma}(1/2\,(K_\ell+1), a_\ell^{-1} + 1/2\,\|\boldsymbol{u}_\ell\|^2).$$

B.2 Derivation of the $(\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}, \boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})})$ Updates for the Poisson Response Model

Adaptation of the derivations in Appendix A.3 of Wand (2014) leads to

$$\begin{aligned}E_q\left[\log p(\boldsymbol{y},\boldsymbol{\beta},\boldsymbol{u},\sigma_1^2,\ldots,\sigma_r^2,a_1,\ldots,a_r)\right] &= E_q\Bigg[\log\, p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u}) + \log\, p(\boldsymbol{\beta},\boldsymbol{u}|\sigma_1^2,\ldots,\sigma_r^2)\\ &\qquad + \sum_{\ell=1}^{r}\log\, p(\sigma_\ell^2|a_\ell) + \sum_{\ell=1}^{r}\log\, p(a_\ell)\Bigg]\\ &= S + \text{terms not involving } \boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})} \text{ or } \boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}\end{aligned}$$

where

$$\begin{aligned}S &\equiv \boldsymbol{y}^T\boldsymbol{C}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})} - \mathbf{1}^T\exp\left\{\boldsymbol{C}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})} + \tfrac{1}{2}\text{diagonal}(\boldsymbol{C}\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}\boldsymbol{C}^T)\right\}\\ &\quad - \tfrac{1}{2}\text{tr}\left(\text{blockdiag}(\sigma_\beta^{-2}\boldsymbol{I}_p, \mu_{q(1/\sigma_1^2)}\boldsymbol{I}_{K_1},\ldots,\mu_{q(1/\sigma_r^2)}\boldsymbol{I}_{K_r})\{\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}^T + \boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}\}\right)\\ &\quad - \tfrac{1}{2}P\log(2\pi) - \tfrac{1}{2}\,p\log(\sigma_\beta^2) - \tfrac{1}{2}\sum_{\ell=1}^{r}K_\ell\, E_q\{\log(\sigma_\ell^2)\} - \mathbf{1}^T\log(\boldsymbol{y}!).\end{aligned}$$

Then,

$$\begin{aligned}d_{\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}}S &= \Bigg(\left[\boldsymbol{y} - \exp\{\boldsymbol{C}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})} + \tfrac{1}{2}\text{diagonal}(\boldsymbol{C}\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}\boldsymbol{C}^T)\}\right]^T\boldsymbol{C}\\ &\quad -\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}^T\text{blockdiag}(\sigma_\beta^{-2}\boldsymbol{I}_p, \mu_{q(1/\sigma_1^2)}\boldsymbol{I}_{K_1},\ldots,\mu_{q(1/\sigma_r^2)}\boldsymbol{I}_{K_r})\Bigg)d\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}\end{aligned}$$

and by Theorem 6, Chapter 5, of Magnus and Neudecker (1999),

$$\begin{aligned}\{\mathsf{D}_{\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}}S\}^T &= \boldsymbol{C}^T\left[\boldsymbol{y} - \exp\{\boldsymbol{C}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})} + \tfrac{1}{2}\text{diagonal}(\boldsymbol{C}\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}\boldsymbol{C}^T)\}\right]\\ &\quad -\text{blockdiag}(\sigma_\beta^{-2}\boldsymbol{I}_p, \mu_{q(1/\sigma_1^2)}\boldsymbol{I}_{K_1},\ldots,\mu_{q(1/\sigma_r^2)}\boldsymbol{I}_{K_r})\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}.\end{aligned}$$

Next,

$$\begin{aligned}d_{\text{vec}(\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})})}S &= -\tfrac{1}{2}\text{vec}\Big(\boldsymbol{C}^T\text{diag}[\exp\{\boldsymbol{C}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})} + \tfrac{1}{2}\text{diagonal}(\boldsymbol{C}\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}\boldsymbol{C}^T)\}]\boldsymbol{C}\\ &\quad +\text{blockdiag}(\sigma_\beta^{-2}\boldsymbol{I}_p, \mu_{q(1/\sigma_1^2)}\boldsymbol{I}_{K_1},\ldots,\mu_{q(1/\sigma_r^2)}\boldsymbol{I}_{K_r})\Big)^T d\,\text{vec}(\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})})\end{aligned}$$

and

$$\begin{aligned}\text{vec}^{-1}\left((\mathsf{D}_{\text{vec}(\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})})}S)^T\right) &= -\tfrac{1}{2}(\boldsymbol{C}^T\text{diag}[\,\exp\{\boldsymbol{C}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}\\ &\quad + \tfrac{1}{2}\text{diagonal}(\boldsymbol{C}\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}\boldsymbol{C}^T)\}]\,\boldsymbol{C}\\ &\quad + \text{blockdiag}(\sigma_\beta^{-2}\boldsymbol{I}_p, \mu_{q(1/\sigma_1^2)}\boldsymbol{I}_{K_1},\ldots,\mu_{q(1/\sigma_r^2)}\boldsymbol{I}_{K_r})).\end{aligned}$$

The final result follows from plugging in $\{\mathsf{D}_{\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}}S\}^T$ and $\text{vec}^{-1}((\mathsf{D}_{\text{vec}(\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})})}S)^T)$ in the updating formulas (20).

B.3 Derivation of $q^*(g_i)$ and $q^*(\kappa)$ for the Negative Binomial Response Model

Standard manipulations lead to the following full conditional distribution

$$g_i|\text{rest} \overset{\text{ind.}}{\sim} \text{Gamma}(\kappa + y_i, 1 + \kappa\exp\{-\boldsymbol{c}_i^T[\boldsymbol{\beta}^T\ \boldsymbol{u}^T]^T\})$$

such that $q^*(g_i)$ is the Gamma density function specified in (11). In addition, standard distributional results for the Gamma density function lead to

$$\begin{aligned}\boldsymbol{\mu}_{q(\log(\boldsymbol{g}))} \quad = \quad & \text{digamma}(\boldsymbol{1}\mu_{q(\kappa)} + \boldsymbol{y}) \\ & \qquad - \log\big[\boldsymbol{1} + \mu_{q(\kappa)}\exp\big\{ - \boldsymbol{C}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})} + \tfrac{1}{2}\,\text{diagonal}(\boldsymbol{C}\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}\boldsymbol{C}^T)\big\}\big].\end{aligned}$$

To derive $q^*(\kappa)$, we define $\omega \equiv \kappa^{-1/2}$ and first obtain $q^*(\omega^2)$. The full conditional density function of ω^2 is

$$p(\omega^2|\text{rest}) = p(\omega^2|\boldsymbol{g},\boldsymbol{\beta},\boldsymbol{u},a_\kappa) \propto p(\boldsymbol{g}|\boldsymbol{\beta},\boldsymbol{u},\omega^2)\,p(\omega^2|a_\kappa).$$

Hence

$$\begin{aligned}\log\{p(\omega^2|\text{rest})\} \quad &= \quad \log\{p(\boldsymbol{g}|\boldsymbol{\beta},\boldsymbol{u},\omega^2)\} + \log\{p(\omega^2|\,a_\kappa)\} + \text{const} \\ &= \quad \left[\sum_{i=1}^n \log\{p(g_i|\boldsymbol{\beta},\boldsymbol{u},\omega^2)\}\right] + \log\{p(\omega^2|\,a_\kappa)\} + \text{const} \\ &= \quad \Big[\sum_{i=1}^n \log\Big\{\frac{(\omega^{-2}\exp\{-(\boldsymbol{X\beta}+\boldsymbol{Zu})_i\})^{1/\omega^2}}{\Gamma(1/\omega^2)} g_i^{1/\omega^2-1} \\ &\qquad\qquad \times\exp[-g_i/(\omega^2\exp\{(\boldsymbol{X\beta}+\boldsymbol{Zu})_i\})]\Big\}\Big] \\ &\qquad\qquad + \log\Big\{(\omega^2)^{-3/2}\exp\{-1/(\omega^2 a_\kappa)\}\Big\} \\ &= \quad -n\left\{\omega^{-2}\log(\omega^2) + \log\Gamma(1/\omega^2)\right\} - 3/2\log(\omega^2) \\ &\qquad\qquad + \frac{\sum_{i=1}^n[\log g_i - (\boldsymbol{X\beta}+\boldsymbol{Zu})_i - g_i\ \exp\{-(\boldsymbol{X\beta}+\boldsymbol{Zu})_i\}] - \frac{1}{a_\kappa}}{\omega^2} \\ &\qquad\qquad + \text{const}.\end{aligned}$$

It follows that

$$\begin{aligned}&\log q(\omega^2) = -n\left\{\omega^{-2}\log(\omega^2) + \log\Gamma(1/\omega^2)\right\} - \tfrac{3}{2}\log(\omega^2) \\ &\quad + \frac{\sum_{i=1}^n \Big(\mu_{q(\log g_i)} - (\boldsymbol{C}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})})_i + \mu_{q(g_i)}\,E_q[\exp\{-(\boldsymbol{X\beta}+\boldsymbol{Zu})_i\}]\Big) - \mu_{q(1/a_\kappa)}}{\omega^2} \\ &\quad + \text{const}.\end{aligned}$$

This leads to

$$q^*(\omega^2) \propto \left\{ \frac{(1/\omega^2)^{1/\omega^2}}{\Gamma(1/\omega^2)} \right\}^n (\omega^2)^{-3/2} \exp(-C_1/\omega^2), \quad \omega^2 > 0,$$

where C_1 is defined at (13). The change of variable $\kappa = 1/\omega^2$ leads to the expression for $q^*(\kappa)$ given at (11). It immediately follows that

$$\mu_{q(\kappa)} = \frac{\mathcal{H}(\frac{1}{2}, 0, 1, n, C_1)}{\mathcal{H}(-\frac{1}{2}, 0, 1, n, C_1)} = \exp[\log\{\mathcal{H}(\tfrac{1}{2}, 0, 1, n, C_1)\} - \log\{\mathcal{H}(-\tfrac{1}{2}, 0, 1, n, C_1)\}].$$

In practice, the second expression is preferable since it helps avoid underflow and overflow problems.

B.4 Derivation of the $(\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}, \boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})})$ Updates for the Negative Binomial Response Model

Note that

$$\begin{aligned}
&E_q[\log p(\boldsymbol{y}, \boldsymbol{g}, \boldsymbol{\beta}, \boldsymbol{u}, \kappa, \sigma_1^2, \ldots, \sigma_r^2, a_1, \ldots, a_r)] \\
&\quad = E_q\Bigg[\log\, p(\boldsymbol{y}|\boldsymbol{g}) + \log\, p(\boldsymbol{g}|\boldsymbol{\beta}, \boldsymbol{u}, \kappa) + \log\, p(\boldsymbol{\beta}, \boldsymbol{u}|\sigma_1^2, \ldots, \sigma_r^2) + \log\, p(\kappa) \\
&\qquad + \sum_{\ell=1}^r \log\, p(\sigma_\ell^2|a_\ell) + \sum_{\ell=1}^r \log\, p(a_\ell) \Bigg] \\
&\quad = S + \text{terms not involving } \boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})} \text{ or } \boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}
\end{aligned}$$

where

$$\begin{aligned}
S \equiv\ & nE_q[\kappa \log(\kappa)] - \mu_{q(\kappa)} \mathbf{1}^T \boldsymbol{C} \boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})} - nE_q[\log(\Gamma(\kappa))] + (\mu_{q(\kappa)} - 1)\mathbf{1}^T E_q[\log(\boldsymbol{g})] \\
& -\mu_{q(\kappa)} \boldsymbol{\mu}_{q(\boldsymbol{g})}^T \exp\{-\boldsymbol{C}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})} + \tfrac{1}{2}\text{diagonal}(\boldsymbol{C}\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}\boldsymbol{C}^T)\} \\
& -\tfrac{1}{2}\text{tr}\left(\text{blockdiag}(\sigma_\beta^{-2}\boldsymbol{I}_p, \mu_{q(1/\sigma_1^2)}\boldsymbol{I}_{K_1}, \ldots, \mu_{q(1/\sigma_r^2)}\boldsymbol{I}_{K_r})\{\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}^T + \boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}\}\right) \\
& -\tfrac{1}{2}P\log(2\pi) - \tfrac{1}{2}\,p\log(\sigma_\beta^2) - \tfrac{1}{2}\,\textstyle\sum_{\ell=1}^r K_\ell\, E_q\{\log(\sigma_\ell^2)\}.
\end{aligned}$$

Then,

$$\begin{aligned}
\{\mathsf{D}_{\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}} S\}^T \quad = \quad & \mu_{q(\kappa)}\boldsymbol{C}^T \left[\boldsymbol{\mu}_{q(\boldsymbol{g})} \odot \exp\{-\boldsymbol{C}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})} + \tfrac{1}{2}\text{diagonal}(\boldsymbol{C}\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}\boldsymbol{C}^T)\} - \mathbf{1} \right] \\
& -\text{blockdiag}(\sigma_\beta^{-2}\boldsymbol{I}_p, \mu_{q(1/\sigma_1^2)}\boldsymbol{I}_{K_1}, \ldots, \mu_{q(1/\sigma_r^2)}\boldsymbol{I}_{K_r})\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}
\end{aligned}$$

and

$$\begin{aligned}
d_{\mathrm{vec}(\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})})}\, S \;=\; & -\tfrac{1}{2}\mathrm{vec}\Big(\mu_{q(\kappa)}\boldsymbol{C}^T\mathrm{diag}[\boldsymbol{\mu}_{q(\boldsymbol{g})}\odot\exp\{-\boldsymbol{C}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})} \\
& +\tfrac{1}{2}\mathrm{diagonal}(\boldsymbol{C}\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}\boldsymbol{C}^T)\}]\boldsymbol{C} \\
& +\mathrm{blockdiag}(\sigma_\beta^{-2}\boldsymbol{I}_p,\mu_{q(1/\sigma_1^2)}\boldsymbol{I}_{K_1},\ldots,\mu_{q(1/\sigma_r^2)}\boldsymbol{I}_{K_r})\Big)^T d\,\mathrm{vec}(\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})})
\end{aligned}$$

such that

$$\begin{aligned}
\mathrm{vec}^{-1}\left((\mathsf{D}_{\mathrm{vec}(\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})})}\, S)^T\right) \;=\; & -\tfrac{1}{2}(\mu_{q(\kappa)}\boldsymbol{C}^T\mathrm{diag}[\boldsymbol{\mu}_{q(\boldsymbol{g})}\odot\exp\{-\boldsymbol{C}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})} \\
& +\tfrac{1}{2}\mathrm{diagonal}(\boldsymbol{C}\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}\boldsymbol{C}^T)\}]\,\boldsymbol{C} \\
& +\mathrm{blockdiag}(\sigma_\beta^{-2}\boldsymbol{I}_p,\mu_{q(1/\sigma_1^2)}\boldsymbol{I}_{K_1},\ldots,\mu_{q(1/\sigma_r^2)}\boldsymbol{I}_{K_r})).
\end{aligned}$$

The final result follows from plugging in these expressions in the updating formulas given at (20).

References

Albert, J. H. and Chib, S. (1993). "Bayesian analysis of binary and polychotomous response data." *Journal of the American Statistical Association*, 88: 669–679. MR1224394. doi: http://dx.doi.org/10.1080/01621459.1993.10476321. 992

Barber, D. and Bishop, C. M. (1998). "Ensemble learning for multi-layer networks." In: Jordan, M. I., Kearns, K. J., and Solla, S. A. (eds.), *Advances in Neural Information Processing Systems, 10*, 395–401. Cambridge, Massachusetts: MIT Press. 1015

Bishop, C. M. (2008). "A new framework for machine learning." In: J. M. Zurada et al., (eds.), *World Congress on Computational Intelligence, 2008 Plenary/Invited Lectures*, *Lecture Notes in Computer Science, 5050*, 1–24. Heidelberg, Germany: Springer-Verlag, . 993

Chopin, N., Jacob, P. E., and Papaspiliopoulos, O. (2013). "SMC^2: an efficient algorithm for sequential analysis of state space models." *Journal of the Royal Statistical Society: Series B*, 75: 397–426. doi: http://dx.doi.org/10.1111/j.1467-9868.2012.01046.x. 1003

Consonni, G. and Marin, J.-M. (2007). "Mean-field variational approximate Bayesian inference for latent variable models." *Computational Statistics and Data Analysis*, 52: 790–798. MR2418528. doi: http://dx.doi.org/10.1016/j.csda.2006.10.028. 992

Damien, P., Wakefield, J., and Walker, S. (1999). "Gibbs sampling for Bayesian nonconjugate and hierarchical models by using auxiliary variables." *Journal of the Royal Statistical Society, Series B*, 61: 331–345. MR1680334. doi: http://dx.doi.org/10.1111/1467-9868.00179. 992

Efron, B. (1979). "Bootstrap methods: another look at the jackknife." *The Annals of Statistics*, 7: 1–26. MR0515681. doi: http://dx.doi.org/10.1214/aos/1176344552. 993

Fan, Y., Leslie, D. S. and Wand, M. P. (2008). "Generalised linear mixed model analysis via sequential Monte Carlo sampling." *Electronic Journal of Statistics*, 2: 916–938. doi: http://dx.doi.org/10.1214/07-EJS158. 1004

Frühwirth-Schnatter, S., Frühwirth, R., Held, L., and Rue, H. (2009). "Improved auxiliary mixture sampling for hierarchical models of non-Gaussian data." *Statistics and Computing*, 19: 479–492. MR2565319. doi: http://dx.doi.org/10.1007/s11222-008-9109-4. 992

Gelman, A. (2006). "Prior distributions for variance parameters in hierarchical models." *Bayesian Analysis*, 1: 515–533. doi: http://dx.doi.org/10.1214/06-BA107A. 996

Girolami, M. and Rogers, S. (2006). "Variational Bayesian multinomial probit regression." *Neural Computation*, 18: 1790–1817. MR2230854. doi: http://dx.doi.org/10.1162/neco.2006.18.8.1790. 992

Honkela, A., Raiko, T., Kuusela, M., Tornio, M. and Karhunen, J. (2010). "Approximate Riemannian conjugate gradient learning for fixed-form variational Bayes." *Journal of Machine Learning Research*, 11: 3235–3268. 1015

Jaakkola, T. S. and Jordan, M. I. (2000). "Bayesian parameter estimation via variational methods." *Statistics and Computing*, 10: 25–37. doi: http://dx.doi.org/10.1023/A:1008932416310. 992

Knowles, D. A. and Minka, T. P. (2011). "Non-conjugate message passing for multinomial and binary regression." In: Shawe-Taylor, J., Zamel, R. S., Bartlett, P., Pereira, F., and Weinberger, K. Q. (eds.), *Advances in Neural Information Processing Systems 24*, 1701–1709. 992, 1015, 1016

Lawless, J. F. (1987). "Negative Binomial and mixed Poisson regression." *Canadian Journal of Statistics*, 15: 209–225. MR0926553. doi: http://dx.doi.org/10.2307/3314912. 992, 996, 1014

Leetaru, K. H. and Schrodt, P. A. (2013). "A 30-year georeferenced global event database: The Global Database of Events, Language, and Tone (GDELT)." In: *International Studies Association Conference*. San Francisco, USA. 1010

Li, Y. and Ruppert, D. (2008). "On the asymptotics of penalized splines." *Biometrika*, 95: 415–436. MR2521591. doi: http://dx.doi.org/10.1093/biomet/asn010. 1007

Luts, J., Broderick, T., and Wand, M. P. (2014). "Real-time semiparametric regression." *Journal of Computational and Graphical Statistics*, 23: 589–615. MR3224647. doi: http://dx.doi.org/10.1080/10618600.2013.810150. 991, 993, 997, 998, 1003

Magnus, J. R. and Neudecker, H. (1999). *Matrix Differential Calculus with Applications in Statistics and Econometrics, Revised Edition*. Chichester UK: Wiley. MR1698873. 1016, 1017

Marley, J. K. and Wand, M. P. (2010). "Non-standard semiparametric regression via BRugs." *Journal of Statistical Software*, 37: 1–30. 996, 1010, 1011

Menictas, M. and Wand, M. P. (2015). "Variational inference for heteroscedastic semiparametric regression." *Australian and New Zealand Journal of Statistics*, 57: in press. 1001

Michalak, S., DuBois, A., DuBois, D., Vander Wiel, S., and Hogden, J. (2012). "Developing systems for real-time streaming analysis." *Journal of Computational and Graphical Statistics*, 21: 561–580. MR2970908. doi: http://dx.doi.org/10.1080/10618600.2012.657144. 991, 993

Minka, T. P. (2001). "Expectation propagation for approximate Bayesian inference." In: *Proceedings of Conference on Uncertainty in Artificial Intelligence*, 362–369. 993

Mittal, S., Madigan, D., Burd, R. S., and Suchard, M. A. (2013). "High-dimensional, massive sample-size Cox proportional hazards regression for survival analysis." *Biostatistics*, 15: 287–294. 993

O'Sullivan, F. (1986). "A statistical perspective on ill-posed inverse problems (with discussion)." *Statistical Science*, 1: 502–527. MR0874480. 994

R Development Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/. 1007

Rue, H., Martino, S., and Chopin, N. (2009). "Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations (with discussion)." *Journal of the Royal Statistical Society, Series B*, 71: 319–392. MR2649602. doi: http://dx.doi.org/10.1111/j.1467-9868.2008.00700.x. 993

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003). *Semiparametric Regression*. New York: Cambridge University Press. MR1998720. doi: http://dx.doi.org/10.1017/CBO9780511755453. 1004

— (2009). "Semiparametric regression during 2003–2007." *Electronic Journal of Statistics*, 3: 1193–1256. MR2566186. doi: http://dx.doi.org/10.1214/09-EJS525. 991

Spiegelhalter, D. J., Thomas, A., Best, N. G., Gilks, W. R., and Lunn, D. (2003). *BUGS: Bayesian inference using Gibbs sampling*. Medical Research Council Biostatistics Unit, Cambridge, UK. http://www.mrc-bsu.cam.ac.uk/bugs 1005

Tan, L. S. L. and Nott, D. J. (2013). "Variational inference for generalized linear mixed models using partially noncentred parametrizations." *Statistical Science*, 28: 168–188. MR3112404. doi: http://dx.doi.org/10.1214/13-STS418. 991, 992

Thurston, S. W., Wand, M. P., and Weincke, J. K. (2000). "Negative binomial additive models." *Biometrics*, 56: 139–144. doi: http://dx.doi.org/10.1111/j.0006-341X.2000.00139.x. 1010

Wainwright, M. J. and Jordan, M. I. (2008). "Graphical models, exponential families, and variational inference." *Foundation and Trends in Machine Learning*, 1: 1–305. doi: http://dx.doi.org/10.1561/2200000001. 991

Wand, M. P. (2014). "Fully simplified Multivariate Normal updates in non-conjugate variational message passing." *Journal of Machine Learning Research*, 15: 1351–1369. MR3214787. 995, 1016, 1017

Wand, M. P. and Ormerod, J. T. (2008). "On O'Sullivan penalised splines and semiparametric regression." *Australian and New Zealand Journal of Statistics*, 50: 179–198. MR2431193. doi: http://dx.doi.org/10.1111/j.1467-842X.2008.00507.x. 994, 1004

Wand, M. P., Ormerod, J. T., Padoan, S. A., and Frühwirth, R. (2011). "Mean field variational Bayes for elaborate distributions." *Bayesian Analysis*, 6: 847–900. MR2869967. doi: http://dx.doi.org/10.1214/11-BA631. 991, 997, 1001, 1014

Wiencke, J., Thurston, S. W., Kelsey, K. T., Varkonyi, A., Wain, J. C., Mark, E. J., and Christiani, D. C. (1999). "Early age at smoking initiation and tobacco carcinogen DNA damage in the lung." *Journal of the National Cancer Institute*, 91: 614–619. doi: http://dx.doi.org/10.1093/jnci/91.7.614. 1010

Acknowledgments

The authors are grateful to Marianne Menictas and David Rohde for their assistance with this research and to the editor, an associate editor and a reviewer for their suggestions for improvement. This research was partially supported by Australian Research Council Discovery Project DP110100061.