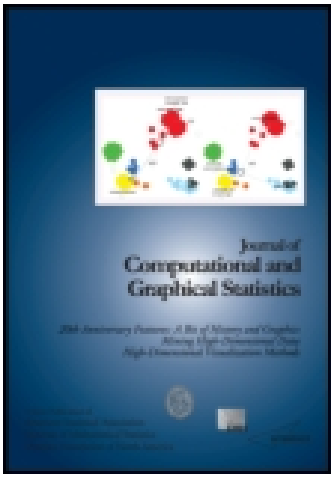

Journal of Computational and Graphical Statistics

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/ucgs20

Real-Time Semiparametric Regression

J. Luts, T. Broderick & M. P. Wand

Accepted author version posted online: 27 Jun 2013.Published online: 23 Jun 2014.

To cite this article: J. Luts, T. Broderick & M. P. Wand (2014) Real-Time Semiparametric Regression, Journal of Computational and Graphical Statistics, 23:3, 589-615, DOI: 10.1080/10618600.2013.810150

To link to this article: http://dx.doi.org/10.1080/10618600.2013.810150

Supplementary materials for this article are available online.
Please go to www.tandfonline.com/r/JCGS

Real-Time Semiparametric Regression

J. LUTS, T. BRODERICK, and M. P. WAND

We develop algorithms for performing semiparametric regression analysis in real time, with data processed as it is collected and made immediately available via modern telecommunications technologies. Our definition of semiparametric regression is quite broad and includes, as special cases, generalized linear mixed models, generalized additive models, geostatistical models, wavelet nonparametric regression models and their various combinations. Fast updating of regression fits is achieved by couching semiparametric regression into a Bayesian hierarchical model or, equivalently, graphical model framework and employing online mean field variational ideas. An Internet site attached to this article, *realtime-semiparametric-regression.net*, illustrates the methodology for continually arriving stock market, real estate, and airline data. Flexible real-time analyses based on increasingly ubiquitous streaming data sources stand to benefit. This article has online supplementary material.

Key Words: Approximate Bayesian inference; Generalized additive models; Mean field variational Bayes; Mixed models; Online variational Bayes; Penalized splines; Wavelets.

1. INTRODUCTION

Ongoing technological advancements mean that data are being collected and made available for inference with rapidly increasing volume and speed. There are numerous examples of this explosion of data, but two that have established connections with semiparametric regression, our focus in this article, are Internet auction analysis (e.g., Jank and Shmueli 2007) and real-time spatial epidemiology (e.g., Kaimi and Diggle 2011).

Semiparametric regression refers to a large class of regression models that provide for nonlinear predictor effects using spline and wavelet basis functions, as well as dependencies arising in grouped data such as within-subject correlation. An arsenal of both frequentist and Bayesian fitting and inference procedures now exist. Recent overviews are contained in Ruppert, Wand, and Carroll (2009) and Wand and Ormerod (2011).

Virtually all semiparametric regression methodology proposed to date assume that the data are processed in *batch*; that is, all at the same time. Summaries such as function estimates, confidence intervals, and posterior density functions are then outputted. Downsides

J. Luts is Postdoctoral Research Fellow (E-mail: *jan.luts@uts.edu.au*) and M. P. Wand is Distinguished Professor, School of Mathematical Sciences, University of Technology Sydney, Broadway 2007, Australia (E-mail: *matt.wand@uts.edu.au*). T. Broderick is Doctoral Candidate, Department of Statistics, University of California, Berkeley, CA 94720 (E-mail: *tab@stat.berkeley.edu*).

Journal of Computational and Graphical Statistics, Volume 23, Number 3, Pages 589–615
DOI: 10.1080/10618600.2013.810150

to batch processing include the requirement that statistical analysis wait until an entire dataset has been assembled and, sometimes, the necessity of storing the entire dataset in memory. In the *online* case, the procedure updates as each new data point (or subset of data points) is obtained. Online updates use only the new data and summary statistics from previous iterations rather than the full set of available data. A particular advantage of online processing is that summaries, such as those just mentioned, are updated throughout the data collection process and, therefore, are available immediately upon demand. Online processing also has the advantage of not requiring storage of potentially very large datasets.

While a number of batch procedures exist for performing semiparametric regression, we focus on a particular methodology here due to the ease of adapting it to the online framework as well as its wide range of applicability. Consider single predictor nonparametric regression, a special case of semiparametric regression with a long history and large literature. Fully automatic nonparametric regression batch procedures include: (a) local linear kernel smoother with cross-validation bandwidth selection, (b) local linear kernel smoother with direct plug-in bandwidth selection, (c) frequentist low-rank smoothing spline with restricted maximum likelihood smoothing parameter selection, (d) Bayesian low-rank smoothing spline with Markov chain Monte Carlo approximate inference, and (e) Bayesian low-rank smoothing spline with mean field variational Bayesian (MFVB) approximate inference. Details of (a) are in Härdle (1990), details of (b) are in Wand and Jones (1995), while (c) and (d) are described in Ruppert, Wand, and Carroll (2003). Section 2.7 of Wand and Ormerod (2011) explains (e). Approaches (a)–(d) are more established, but none has a viable online modification. However, (e) is relatively easy to modify for this purpose.

Another advantage of the Bayesian low-rank smoothing spline approach to nonparametric regression is its extendibility. As explained in Wand (2009), couching semiparametric regression in a graphical models framework permits arbitrarily sophisticated models to be handled elegantly, efficiently, and cohesively. This approach can handle generalized additive models, geostatistical models, wavelet nonparametric regression models, and their various combinations, as well complications such as outliers and missingness. Inference in these models is often accomplished by applying Markov chain Monte Carlo procedures using the directed acyclic graph of variable dependencies. While versatile and accurate, such inference procedures can be unacceptably slow. MFVB approaches, as demonstrated in Faes, Ormerod, and Wand (2011) and Wand and Ormerod (2011), are a much faster alternative. Some accuracy and versatility must be sacrificed in return for the increased speed of MFVB. Nonetheless, for the models treated in this article MFVB accuracy ranges from good to excellent.

Iterative algorithms that make a single pass through the data—with one iteration per data point or per some small, fixed number of data points—have recently been developed for variational Bayesian inference. In the machine learning literature, Hoffman, Blei, and Bach (2010) introduced such an MFVB algorithm for latent Dirichlet allocation and applied their algorithm to topic modeling. This procedure was extended to the hierarchical Dirichlet process by Wang, Paisley, and Blei (2011). Tchumtchoua, Dunson, and Morris (2012) further developed online MFVB approximate inference for high-dimensional correlated data. The methodology in these articles is referred to as *online mean field variational Bayes* or often with the shorter name *online variational Bayes*. While they are indeed single-pass and require storing at most a small, fixed number of data points in memory,

they do, however, require knowledge of the number of data points from the start of the algorithm. Our focus in this work, by contrast, is not on transforming MFVB algorithms that require multiple data passes into single-pass algorithms. Rather, we are, in some sense, pursuing a more classical definition of an "online algorithm" in that each iteration of our procedure uses past data only in the form of sufficient statistics and future data not at all. The proposed methodology is especially relevant for fitting models possessing both fixed and random effects, also known as mixed models, for datasets that do not fit in memory or data that are continuously being generated.

Online MFVB has not been entertained previously for nonparametric and semiparametric regression, but there is an old and large literature involving other online approaches. For nonparametric regression and the related density estimation problem Wolverton and Wagner (1969), Yamato (1971), Devroye and Wagner (1980), and Krzyzak and Pawlak (1984) are examples of early articles on online analysis using kernel estimators. However, they are chiefly concerned with theoretical properties of the estimators and are devoid of practical automatic smoothing parameter selection strategies.

Outside of semiparametric regression, there are also large literatures on online analysis. A few recent examples are: Ng, McLachlan, and Lee (2006) on prediction of hospital resource utilization, Fricker and Chang (2008) on biosurveillance, and Kaimi and Diggle (2011) on monitoring of variation in risk of infections. A very recent article by Michalak et al. (2012) describes the development of systems for real-time streaming analysis.

Semiparametric regression is a highly visual branch of statistics, with graphics being a crucial means of conveying and diagnosing regression fits. The norm for such graphical display are ink drawings on pieces of paper or figures in PDF file. Real-time semiparametric regression represents a paradigm shift in graphical display, where regression summaries are best thought of as dynamic graphics on web pages or iDevice apps. We have organized an Internet site that illustrates real-time semiparametric regression graphical display.

Section 2 introduces the notion of real-time semiparametric regression with online MFVB via increasingly more sophisticated Gaussian response models. Both classical and sparse shrinkage are treated. The more challenging binary response case is dealt with in Section 3. In Section 4, we justify our approach to real-time semiparametric regression in relation to various other online learning methods such as stochastic gradient descent. Some discussion about inferential accuracy is given in Section 5. Dynamic web pages that illustrate the new methodology on live data are the focus of Section 6.

2. GAUSSIAN RESPONSE MODELS

The conversion of a batch MFVB semiparametric regression procedure to one that does online processing is particularly straightforward in the Gaussian response case. We start by explaining such conversion for the multiple linear regression model, since it has minimal notational overhead.

2.1 MULTIPLE LINEAR REGRESSION

Let $\boldsymbol{X}$ be a $n \times p$ design matrix and consider the Bayesian regression model

$$\boldsymbol{y} \,|\, \boldsymbol{\beta}, \sigma^2 \sim N(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}), \quad \boldsymbol{\beta} \sim N\left(\boldsymbol{0}, \sigma_\beta^2 \boldsymbol{I}\right), \quad \sigma \sim \text{Half-Cauchy}(A), \tag{1}$$

Figure 1. Directed acyclic graph for the model conveyed by (1) and (2). The shading corresponds to the observed data.

where the Half-Cauchy(A) prior is such that the prior density function of σ satisfies $p(\sigma) \propto \{1 + (\sigma/A)^2\}^{-1}$, $\sigma > 0$. An equivalent, but more tractable model, is that where $\sigma \sim$ Half-Cauchy(A) is replaced by the auxiliary variable representation

$$\sigma^2 | a \sim \text{Inverse-Gamma}\left(\frac{1}{2}, 1/a\right), \quad a \sim \text{Inverse-Gamma}\left(\frac{1}{2}, 1/A^2\right), \tag{2}$$

where the random variable $v \sim$ Inverse-Gamma(A, B) if and only if its density function is

$$p(v) = B^A \Gamma(A)^{-1} v^{-A-1} \exp(-B/v), \quad v > 0.$$

A pertinent result for this distribution is $E(1/v) = A/B$. Figure 1 displays the directed acyclic graph corresponding to the model conveyed by (1) and (2).

Mean field variational Bayesian (MFVB) is a general prescription for approximation of posterior density functions in a graphical model. General references on MFVB include Bishop (2006) and Wainwright and Jordan (2008). Mean field approximation of the joint posterior density function $p(\boldsymbol{\beta}, a, \sigma^2 | \boldsymbol{y})$ is founded upon this function being restricted to have a product form such as

$$q(\boldsymbol{\beta}, a)\, q(\sigma^2) \tag{3}$$

for density functions $q(\boldsymbol{\beta}, a)$ and $q(\sigma^2)$. We then choose these so-called q-density functions to minimize the Kullback-Leibler distance between $p(\boldsymbol{\beta}, a, \sigma^2 | \boldsymbol{y})$ and $q(\boldsymbol{\beta}, a)\, q(\sigma^2)$:

$$\int q(\boldsymbol{\beta}, a)\, q(\sigma^2) \log\left\{\frac{q(\boldsymbol{\beta}, a)\, q(\sigma^2)}{p(\boldsymbol{\beta}, a, \sigma^2 | \boldsymbol{y})}\right\} d\boldsymbol{\beta}\, da\, d\sigma^2.$$

Standard manipulations show that an equivalent optimization problem is that of maximizing

$$\underline{p}(\boldsymbol{y}; q) \equiv \exp \int q(\boldsymbol{\beta}, a)\, q(\sigma^2) \log\left\{\frac{p(\boldsymbol{\beta}, a, \sigma^2, \boldsymbol{y})}{q(\boldsymbol{\beta}, a)\, q(\sigma^2)}\right\} d\boldsymbol{\beta}\, da\, d\sigma^2$$

and that $\underline{p}(\boldsymbol{y}; q)$ is a lower bound on the marginal likelihood $p(\boldsymbol{y})$ for all q-densities. The solutions can be shown to satisfy

$$\begin{aligned} q^*(\boldsymbol{\beta}, a) &\propto \exp\left[E_{q(\sigma^2)}\{\log\{p(\boldsymbol{\beta}, a | \boldsymbol{y}, \sigma^2)\}\right], \\ \text{and} \qquad & \\ q^*(\sigma^2) &\propto \exp\left[E_{q(\boldsymbol{\beta}, a)}\{\log\{p(\sigma^2 | \boldsymbol{y}, \boldsymbol{\beta}, a)\}\right] \end{aligned} \tag{4}$$

(see, e.g., Section 2.2 of Ormerod and Wand 2010). Application of standard distribution theory to (4) shows that

$$\begin{aligned} &q^*(\boldsymbol{\beta}, a) \text{ is the product of the } N\left(\boldsymbol{\mu}_{q(\boldsymbol{\beta})}, \boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}\right) \text{ density function} \\ &\text{and the Inverse-Gamma}\left(1, B_{q(a)}\right) \text{ density function;} \\ &q^*(\sigma^2) \text{ is the Inverse-Gamma}\left(\frac{1}{2}(n+1), B_{q(\sigma^2)}\right) \text{ density function} \end{aligned} \tag{5}$$

for parameters $\boldsymbol{\mu}_{q(\boldsymbol{\beta})}$ and $\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}$, the mean vector and covariance matrix of $q^*(\boldsymbol{\beta})$, $B_{q(a)}$, the rate parameter of $q^*(a)$ and $B_{q(\sigma^2)}$, the rate parameter of $q^*(\sigma^2)$. The MFVB solution is also such that $q^*(\boldsymbol{\beta}, a) = q^*(\boldsymbol{\beta})\, q^*(a)$ even though (3) does not assume this.

The symbols $\boldsymbol{\mu}_{q(\boldsymbol{\beta})}$ and $\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}$ in (5) are instances of the following general notation that we use throughout this article. If v is a random variable having density function $q(v)$ then

$$\mu_{q(v)} \equiv E_q(v) \quad \text{and} \quad \sigma^2_{q(v)} \equiv \text{var}_q(v).$$

If $\boldsymbol{v}$ is a random vector having density function $q(\boldsymbol{v})$, then

$$\boldsymbol{\mu}_{q(\boldsymbol{v})} \equiv E_q(\boldsymbol{v}) \quad \text{and} \quad \boldsymbol{\Sigma}_{q(\boldsymbol{v})} \equiv \text{cov}_q(\boldsymbol{v}).$$

The optimal parameters in the q^*-density functions are interrelated. For example,

$$\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})} = \left\{\mu_{q(1/\sigma^2)}\, \boldsymbol{X}^T \boldsymbol{X} + \sigma_\beta^{-2}\, \boldsymbol{I}\right\}^{-1}.$$

Hence, they must be obtained via an iterative coordinate ascent algorithm, in which equalities between the parameters are replaced by updates. This leads to Algorithm 1 for batch MFVB fitting of (1) and (2). Each update is guaranteed to increase the value of $\underline{p}(\boldsymbol{y}; q)$ (e.g., Luenberger and Ye 2008).

The lower bound on the marginal log-likelihood, used to monitor convergence in Algorithm 1, has explicit expression

$$\begin{aligned}\log\, \underline{p}(\boldsymbol{y}; q) = {} & \frac{1}{2}\, p - \frac{1}{2}\, n \log(2\pi) - 2\log(\pi) + \log\Gamma\left(\frac{1}{2}(n+1)\right) \\ & - \frac{1}{2}\, p\, \log\left(\sigma^2_{\boldsymbol{\beta}}\right) - \log(A) - \frac{1}{2\sigma^2_{\boldsymbol{\beta}}}\left\{\left\|\boldsymbol{\mu}_{q(\boldsymbol{\beta})}\right\|^2 + \text{tr}\left(\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}\right)\right\} \\ & + \frac{1}{2}\log\left|\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}\right| - \frac{1}{2}(n+1)\log\left[(n+1)/\left\{2\mu_{q(1/\sigma^2)}\right\}\right] \\ & - \log\left(\mu_{q(1/\sigma^2)} + A^{-2}\right) + \mu_{q(1/\sigma^2)}\mu_{q(1/a)}.\end{aligned}$$

Algorithm 1 Batch mean field variational Bayes algorithm for approximate inference in the Gaussian response linear regression models (1) and (2)

Initialize: $\mu_{q(1/\sigma^2)} > 0$.

Read in $\boldsymbol{y}$ $(n \times 1)$ and $\boldsymbol{X}$ $(n \times p)$.

Cycle:

$$\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})} \leftarrow \left\{\mu_{q(1/\sigma^2)}\, \boldsymbol{X}^T \boldsymbol{X} + \sigma_\beta^{-2}\, \boldsymbol{I}\right\}^{-1}$$

$$\boldsymbol{\mu}_{q(\boldsymbol{\beta})} \leftarrow \mu_{q(1/\sigma^2)}\, \boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}\, \boldsymbol{X}^T \boldsymbol{y}; \quad \mu_{q(1/a)} \leftarrow 1/\left\{\mu_{q(1/\sigma^2)} + A^{-2}\right\}$$

$$\mu_{q(1/\sigma^2)} \leftarrow \frac{n+1}{2\,\mu_{q(1/a)} + \boldsymbol{y}^T\boldsymbol{y} - 2\boldsymbol{\mu}^T_{q(\boldsymbol{\beta})}\boldsymbol{X}^T\boldsymbol{y} + \text{tr}\left[(\boldsymbol{X}^T\boldsymbol{X})\left\{\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})} + \boldsymbol{\mu}_{q(\boldsymbol{\beta})}\boldsymbol{\mu}^T_{q(\boldsymbol{\beta})}\right\}\right]}$$

until the increase in $\underline{p}(\boldsymbol{y}; q)$ is negligible.

Produce summaries based on $q^*(\boldsymbol{\beta}) \sim N\left(\boldsymbol{\mu}_{q(\boldsymbol{\beta})}, \boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}\right)$ and

$$q^*(\sigma^2) \sim \text{Inverse-Gamma}\left(\frac{1}{2}(n+1), (n+1)/\left\{2\mu_{q(1/\sigma^2)}\right\}\right).$$

In Algorithm 1, dependence on the data is only through the quantities $\boldsymbol{y}^T\boldsymbol{y}$, $\boldsymbol{X}^T\boldsymbol{y}$, and $\boldsymbol{X}^T\boldsymbol{X}$, and each of these have simple updates when a new response y_{new} and its corresponding $p \times 1$ vector of predictors $\boldsymbol{x}_{\text{new}}$ arrives. For example, the new $\boldsymbol{X}^T\boldsymbol{X}$ matrix is

$$\boldsymbol{X}_{\text{new}}^T\boldsymbol{X}_{\text{new}} = \boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{x}_{\text{new}}\boldsymbol{x}_{\text{new}}^T.$$

Based on these observations Algorithm 2, the *online* modification of Algorithm 1, ensues.

Algorithm 2 Online mean field variational Bayes algorithm for approximate inference in the Gaussian response linear regression models (1) and (2)

Initialize: $\mu_{q(1/\sigma^2)} > 0$, $\boldsymbol{y}^T\boldsymbol{y} \leftarrow 0$, $\boldsymbol{X}^T\boldsymbol{y} \leftarrow \boldsymbol{0}_{p\times 1}$, $\boldsymbol{X}^T\boldsymbol{X} \leftarrow \boldsymbol{0}_{p\times p}$, $n \leftarrow 0$.

Cycle:

read in y_{new} (1×1) and $\boldsymbol{x}_{\text{new}}$ $(p \times 1)$; $n \leftarrow n + 1$

$\boldsymbol{y}^T\boldsymbol{y} \leftarrow \boldsymbol{y}^T\boldsymbol{y} + y_{\text{new}}^2$; $\boldsymbol{X}^T\boldsymbol{y} \leftarrow \boldsymbol{X}^T\boldsymbol{y} + \boldsymbol{x}_{\text{new}}\, y_{\text{new}}$; $\boldsymbol{X}^T\boldsymbol{X} \leftarrow \boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{x}_{\text{new}}\,\boldsymbol{x}_{\text{new}}^T$

$\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})} \leftarrow \left\{\mu_{q(1/\sigma^2)}\,\boldsymbol{X}^T\boldsymbol{X} + \sigma_\beta^{-2}\,\boldsymbol{I}\right\}^{-1}$

$\boldsymbol{\mu}_{q(\boldsymbol{\beta})} \leftarrow \mu_{q(1/\sigma^2)}\,\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}\,\boldsymbol{X}^T\boldsymbol{y}$; $\mu_{q(1/a)} \leftarrow 1/\{\mu_{q(1/\sigma^2)} + A^{-2}\}$

$$\mu_{q(1/\sigma^2)} \leftarrow \frac{n+1}{2\,\mu_{q(1/a)} + \boldsymbol{y}^T\boldsymbol{y} - 2\boldsymbol{\mu}_{q(\boldsymbol{\beta})}^T\boldsymbol{X}^T\boldsymbol{y} + \text{tr}\left[(\boldsymbol{X}^T\boldsymbol{X})\left\{\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})} + \boldsymbol{\mu}_{q(\boldsymbol{\beta})}\boldsymbol{\mu}_{q(\boldsymbol{\beta})}^T\right\}\right]}$$

produce summaries based on $q^*(\boldsymbol{\beta}) \sim N\left(\boldsymbol{\mu}_{q(\boldsymbol{\beta})}, \boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}\right)$ and

$$q^*(\sigma^2) \sim \text{Inverse-Gamma}\left(\frac{1}{2}(n+1), (n+1)/\left\{2\mu_{q(1/\sigma^2)}\right\}\right)$$

until data no longer available or analysis terminated.

Algorithm 2 differs from Algorithm 1 in that the data are processed on arrival and the approximate posterior densities of the model parameters are continually updated. In the case of streaming data, there is the option of dynamic graphical displays of the approximate posterior density functions of the regression coefficients and error variance and corresponding approximate Bayes estimates and credible sets. Dynamic regression diagnostic plots could also be entertained.

Figure 2 provides rudimentary illustration of online regression inference when data from the Vietnam World Bank Living Standards Survey (source: Cameron and Trivedi 2005) are fed into Algorithm 2. These data are in the `VietNamI` data-frame of the R package `Ecdat` (Croissant 2011). The response variable is the logarithm of total medical expenses. Description of the predictor variables is given in the `VietNamI` documentation of Crossaint (2011). Each variable was transformed to lie inside the unit interval before being processed. The scaling is determined using an initialization batch just as for the initial parameter tuning described in Section 2.1.1. The posterior density functions were then back-transformed to correspond to the original units. The hyperparameters were set at $\sigma_\beta^2 = 10^{10}$ and $A = 10^5$ to impose noninformativity.

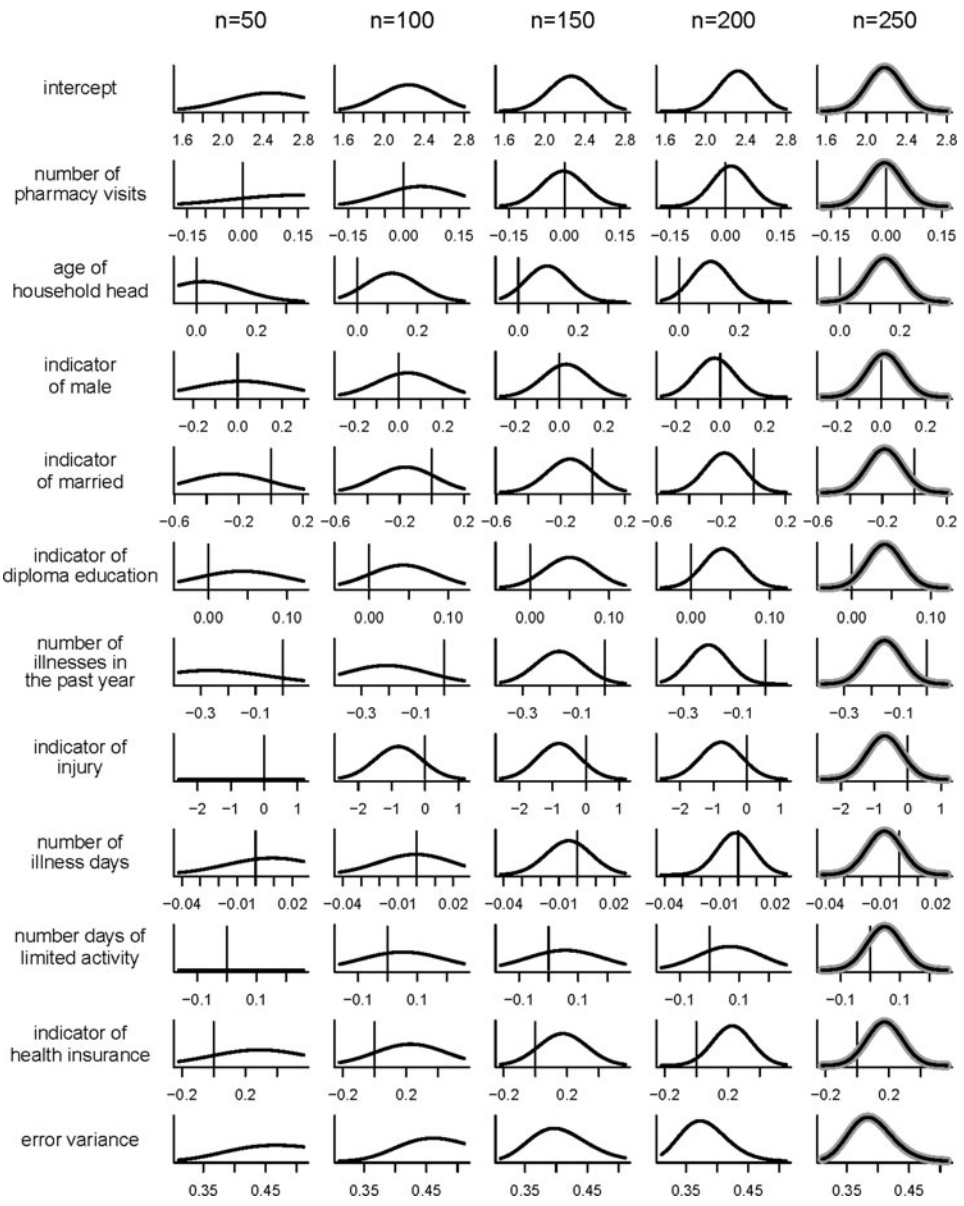

Figure 2. Successive approximate posterior density functions of regression coefficients and the logarithm of error variance for the Vietnam medical expenses data described in the text. The predictors corresponding to each regression coefficient are listed in the left-hand columns. The posterior density functions are based on online MFVB as detailed in Algorithm 2. The axis limits are the same across each row and a vertical line is positioned at zero. For $n = 250$ the batch MFVB approximate fits are shown as thick gray curves.

Note, for example, the approximate posterior density functions for β_2, the regression coefficient attached to `age of household head`. For $n \leq 100$, the posterior density function is relatively flat and β_2 is not statistically significant. As n increases, the posterior density functions become narrower and, by $n = 250$, the lower

limit of the 95% credible set is positive—indicating statistical significance of this predictor.

The right-most column of Figure 2 shows the batch MFVB posterior density functions for $n = 250$. In this case, the batch and online MFVB results are seen to be virtually identical. However, as demonstrated later, such agreement is not guaranteed in general.

2.1.1 Batch-Based Tuning and Convergence Diagnosis. Ideally, the online Algorithm 2 will mimic the results of the batch Algorithm 1 as the sample size n increases. However, we know of no guarantees that this will happen and it is possible that the online parameters will diverge from their batch counterparts. For the more elaborate models studied later in this article, such divergence is very common. Therefore, convergence diagnosis at the start of the online iterations is essential. The principal idea is to start by running a small subset of initial data points in the batch algorithm to obtain starting values for both data sufficient statistics and, more importantly, estimated parameters of the model. A second, small validation subset of data is used to compare the batch and online algorithm results. If convergence of the online iterations to their batch counterparts is not verified by this comparison then larger initial batch runs are required to tune the online algorithm.

The idea of collecting streaming data into a small subset before processing it to improve performance of a single-pass algorithm is reminiscent of the "mini-batches" of Hoffman, Blei, and Bach (2010). However, in our approach, the batching of data happens only with a small subset at the very beginning of the algorithm rather than throughout. Also, we develop an alternative tuning method for this subset batch size below; notably, our tuning method requires batching only some initial subset of the data rather than the full dataset.

We will now provide details via the Figure 2 example. Figure 3 shows the posterior means and 95% credible sets for each β_j, $0 \le j \le 11$, and $\log(\sigma^2)$ and sample sizes $n = 100, 110, \ldots, 200$ when the Vietnam medical expenses data are fitted via both batch and online MFVB. The batch MFVB summary statistics (shown as gray lines in Figure 3) correspond to simply inputting the first $n_{\text{warm}} = 100$ observations into Algorithm 1 and then repeating this process for 10 additional equally spaced sample sizes that are $n_{\text{valid}} = 100$ greater than n_{warm}. The largest sample size is then $n_{\text{warm}} + n_{\text{valid}} = 200$. The online results (shown as gray lines in Figure 3) were obtained via online MFVB updating steps of Algorithm 2 but with $\mu_{q(1/\sigma^2)}$, $\boldsymbol{y}^T\boldsymbol{y}$, $\boldsymbol{X}^T\boldsymbol{y}$, $\boldsymbol{X}^T\boldsymbol{X}$ and n initialized at *the values obtained when the first* $n_{\text{warm}} = 100$ observations are inputted into Algorithm 1. This implies that all the results are identical at $n = n_{\text{warm}} = 100$, but there are some small discrepancies for $n > 100$. In this example, the discrepancies are negligible, and hard to discern from Figure 3 – indicating convergence of the online MFVB algorithm. Figure 5 in Section 3 shows an example where convergence is not achieved with $n_{\text{warm}} = 100$ and a larger warm-up is required.

Algorithm 2′ is a modification of Algorithm 2 that incorporates batch-based tuning and convergence diagnostics. While such modification is not necessary for the example depicted in Figures 2 and 3, it is crucial for more sophisticated semiparametric models such as those described later in this article.

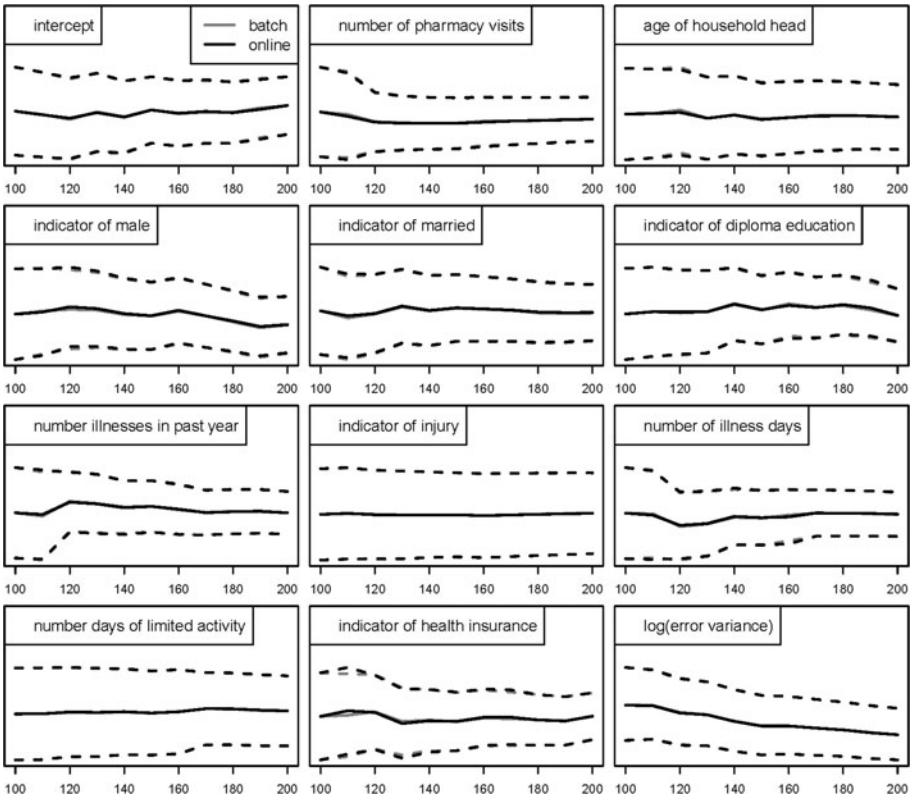

Figure 3. Convergence diagnostics for the example given in Figure 2. The solid lines track the posterior means, while the dashed lines show corresponding 95% credible sets. The horizontal axes show the sample sizes between a warm-up batch sample of size $n_{\text{warm}} = 100$ and validation sample sizes up to $n_{\text{valid}} = 100$ greater than n_{warm}.

Algorithm 2′ Modification of Algorithm 2 to include batch-based tuning and convergence diagnosis

1. Set n_{warm} to be the warm-up sample size and n_{valid} to be size of the validation period. Read in the first $n_{\text{warm}} + n_{\text{valid}}$ response and predictor values.
2. Create $\boldsymbol{y}_{\text{warm}}$ and $\boldsymbol{X}_{\text{warm}}$ consisting of the first n_{warm} response and predictor values.
3. Feed $\boldsymbol{y}_{\text{warm}}$ and $\boldsymbol{X}_{\text{warm}}$ into the batch MFVB Algorithm 1 to obtain a starting value for $\mu_{q(1/\sigma^2)}$.
4. Set $\boldsymbol{y}^T\boldsymbol{y} \leftarrow \boldsymbol{y}_{\text{warm}}^T\boldsymbol{y}_{\text{warm}}$, $\boldsymbol{X}^T\boldsymbol{y} \leftarrow \boldsymbol{X}_{\text{warm}}^T\boldsymbol{y}_{\text{warm}}$, $\boldsymbol{X}^T\boldsymbol{X} \leftarrow \boldsymbol{X}_{\text{warm}}^T\boldsymbol{X}_{\text{warm}}$ and $n \leftarrow n_{\text{warm}}$.
5. Run the online MFVB Algorithm 2 until $n = n_{\text{warm}} + n_{\text{valid}}$.
6. Use convergence diagnostic graphics to assess whether the online parameters are converging to the batch parameters.
   (a) If not converging then return to Step 1 and increase n_{warm}.
   (b) If converging then continue running the online MFVB Algorithm 2 until data no longer available or analysis terminated.

One could contemplate automating Step 6 of Algorithm 2′, to save the user from having to conduct diagnostic checks. However, we have not yet explored automatic convergence diagnosis and, instead, flag this as a problem worthy of future research.

2.1.2 Model Assumptions. The online MFVB Algorithm 2′ is founded upon the same assumptions as its batch counterpart Algorithm 1. Both algorithms fit the Bayesian linear regression model (1), but the latter has the option to do the fitting in real time for sequentially arriving data.

Throughout this article, we are not allowing for the model parameters to change as new data arrive. Colloquially, we assume "fixed targets" rather than "moving targets." Extensions to semiparametric regression scenarios where the model parameters drift over time, and real-time algorithms that adapt to such drifts, are certainly worthy of future investigation—but beyond this article's scope.

2.2 LINEAR MIXED MODELS

A very useful structure for semiparametric regression is the class of Bayesian linear mixed models of the form

$$\begin{aligned} \boldsymbol{y} \,|\, \boldsymbol{\beta}, \boldsymbol{u}, \sigma_\varepsilon^2 &\sim N\big(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u}, \sigma_\varepsilon^2 \boldsymbol{I}\big) \\ \boldsymbol{u} \,|\, \sigma_{u1}^2, \ldots, \sigma_{ur}^2 &\sim N\big(\boldsymbol{0}, \text{blockdiag}\big(\sigma_{u1}^2 \boldsymbol{I}_{K_1}, \ldots, \sigma_{ur}^2 \boldsymbol{I}_{K_r}\big)\big), \end{aligned} \tag{6}$$

where $\boldsymbol{y}$ is an $n \times 1$ vector of response variables, $\boldsymbol{\beta}$ is a $p \times 1$ vector of fixed effects, $\boldsymbol{u}$ is a vector of random effects, $\boldsymbol{X}$ and $\boldsymbol{Z}$ corresponding design matrices, σ_ε^2 is the error variance and $\sigma_{u1}^2, \ldots, \sigma_{ur}^2$ are variance parameters corresponding to sub-blocks of $\boldsymbol{u}$ of size $K_1, \ldots, K_r$. Here the priors are taken to be

$$\boldsymbol{\beta} \sim N\big(\boldsymbol{0}, \sigma_\beta^2 \boldsymbol{I}\big), \quad \sigma_{u\ell} \sim \text{Half-Cauchy}(A_{u\ell}), \quad 1 \le \ell \le r, \quad \sigma_\varepsilon \sim \text{Half-Cauchy}(A_\varepsilon) \tag{7}$$

with the hyperparameters satisfying $\sigma_\beta^2, A_\varepsilon, A_{u\ell} > 0$ for $1 \le \ell \le r$. As in Section 2, tractability considerations motivate the introduction of auxiliary variables

$$a_{u\ell} \sim \text{Inverse-Gamma}\left(\frac{1}{2}, 1/A_{u\ell}^2\right) \quad \text{and} \quad a_\varepsilon \sim \text{Inverse-Gamma}\left(\frac{1}{2}, 1/A_\varepsilon^2\right) \tag{8}$$

and use of the analog of (2) to induce Half-Cauchy priors on the standard deviation parameters.

As spelled out in Section 2 of Zhao et al. (2006), models (6) and (7) encompass a rich class of models including (with example number from Zhao et al. 2006 added):

1. Simple random effects models (Examples 1 and 2)
2. Cross-random effects models (Example 3)
3. Nested random effects models (Example 4)
4. Generalized additive models (Example 6)

5. Semiparametric mixed models (Example 7)
6. Bivariate smoothing and geoadditive models extensions (Example 8).

Examples 2 and 6 of Zhao et al. (2006) actually involve 2×2 and 3×3 unstructured covariance matrix parameters which are not covered by (7). However, as discussed in Section 2.3, the unstructured covariance matrix extension is quite straightforward.

We seek a mean field approximation to the joint posterior density function:

$$p\left(\boldsymbol{\beta}, \boldsymbol{u}, a_{u1}, \ldots, a_{ur}, a_\varepsilon, \sigma_{u1}^2, \ldots, \sigma_{ur}^2, \sigma_\varepsilon^2 \middle| \boldsymbol{y}\right) \approx q\left(\boldsymbol{\beta}, \boldsymbol{u}, a_{u1}, \ldots, a_{ur}, a_\varepsilon, \sigma_{u1}^2, \ldots, \sigma_{ur}^2, \sigma_\varepsilon^2\right).$$

The product form

$$\begin{aligned} &q\left(\boldsymbol{\beta}, \boldsymbol{u}, a_{u1}, \ldots, a_{ur}, a_\varepsilon, \sigma_{u1}^2, \ldots, \sigma_{ur}^2, \sigma_\varepsilon^2\right) \\ &\quad = q(\boldsymbol{\beta}, \boldsymbol{u}, a_{u1}, \ldots, a_{ur}, a_\varepsilon)\, q\left(\sigma_{u1}^2, \ldots, \sigma_{ur}^2, \sigma_\varepsilon^2\right) \end{aligned} \tag{9}$$

has the advantage of being minimally restrictive while also yielding closed form MFVB updates. The analog of (4) leads to

$q^*(\boldsymbol{\beta}, \boldsymbol{u}, a_{u1}, \ldots, a_{ur}, a_\varepsilon)$ is the product of the $N\left(\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}, \boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}\right)$ density function, Inverse-Gamma$\left(1, B_{q(a_{u\ell})}\right)$ density functions, $1 \le \ell \le r$, and the Inverse-Gamma$\left(1, B_{q(a_\varepsilon)}\right)$ density function;

$q^*\left(\sigma_{u1}^2, \ldots, \sigma_{ur}^2, \sigma_\varepsilon^2\right)$ is the product of Inverse-Gamma $\left(\frac{1}{2}(K_\ell + 1), B_{q(\sigma_{u\ell}^2)}\right)$ density functions for $1 \le \ell \le r$ and the Inverse-Gamma $\left(\frac{1}{2}(n+1), B_{q(\sigma_\varepsilon^2)}\right)$ density function.

The subscripted B's are rate parameters. Batch MFVB fitting of (6), but with slightly different prior distributions, is given by Algorithm 3 of Ormerod and Wand (2010), where the notation

$$\boldsymbol{C} = [\boldsymbol{X}\ \boldsymbol{Z}]$$

is used. Let P be the number of columns in $\boldsymbol{C}$. Then each pass of the corresponding online MFVB algorithm involves arrival and processing of a new scalar response measurement, y_{new}, and a $P \times 1$ vector $\boldsymbol{c}_{\text{new}}$, corresponding to the new row of $\boldsymbol{C}$. This results in Algorithm 3 for real-time fitting of (6).

The $\boldsymbol{c}_{\text{new}}$ vector will have different forms depending on the type of linear mixed model. To better understand the nature of these forms, consider the following two special cases of (6):

$$\begin{aligned} y_{ij} | \beta_0, U_i, \beta_1, \sigma_\varepsilon^2 &\overset{\text{ind.}}{\sim} N\left(\beta_0 + U_i + \beta_1 x_{ij}, \sigma_\varepsilon^2\right), \quad 1 \le i \le m, \quad 1 \le j \le n_i, \\ U_i | \sigma_u^2 &\overset{\text{ind.}}{\sim} N\left(0, \sigma_u^2\right), \quad \beta_0, \beta_1 \overset{\text{ind.}}{\sim} N\left(0, \sigma_{\boldsymbol{\beta}}^2\right), \\ \sigma_u &\sim \text{Half-Cauchy}(A_u), \quad \sigma_\varepsilon \sim \text{Half-Cauchy}(A_\varepsilon) \end{aligned} \tag{10}$$

Algorithm 3 Online mean field variational Bayes algorithm for approximate inference in the Gaussian response linear mixed model (6)

1. Perform batch-based tuning runs analogous to those described in Algorithm 2′ and determine a warm-up sample size n_{warm} for which convergence is validated.
2. Set $\boldsymbol{y}_{\text{warm}}$ and $\boldsymbol{C}_{\text{warm}}$ to be the response vector and design matrix based on the first n_{warm} observations. Then set $\boldsymbol{y}^T\boldsymbol{y} \leftarrow \boldsymbol{y}_{\text{warm}}^T\boldsymbol{y}_{\text{warm}}$, $\boldsymbol{C}^T\boldsymbol{y} \leftarrow \boldsymbol{C}_{\text{warm}}^T\boldsymbol{y}_{\text{warm}}$, $\boldsymbol{C}^T\boldsymbol{C} \leftarrow \boldsymbol{C}_{\text{warm}}^T\boldsymbol{C}_{\text{warm}}$, $n \leftarrow n_{\text{warm}}$. Also, set $\mu_{q(1/\sigma_\varepsilon^2)}$ and $\mu_{q(1/\sigma_{u1}^2)}, \ldots, \mu_{q(1/\sigma_{ur}^2)}$ to be the values for these quantities obtained in the batch-based tuning run with sample size n_{warm}.
3. Cycle:

   read in y_{new} (1×1) and $\boldsymbol{c}_{\text{new}}$ $(P \times 1)$; $n \leftarrow n+1$

   $$\boldsymbol{y}^T\boldsymbol{y} \leftarrow \boldsymbol{y}^T\boldsymbol{y} + y_{\text{new}}^2;\ \boldsymbol{C}^T\boldsymbol{y} \leftarrow \boldsymbol{C}^T\boldsymbol{y} + \boldsymbol{c}_{\text{new}}\, y_{\text{new}};\ \boldsymbol{C}^T\boldsymbol{C} \leftarrow \boldsymbol{C}^T\boldsymbol{C} + \boldsymbol{c}_{\text{new}}\,\boldsymbol{c}_{\text{new}}^T$$

   $$\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})} \leftarrow \Big[\mu_{q(1/\sigma_\varepsilon^2)}\,\boldsymbol{C}^T\boldsymbol{C} + \text{blockdiag}\Big\{\sigma_\beta^{-2}\,\boldsymbol{I}_p, \mu_{q(1/\sigma_{u1}^2)}\boldsymbol{I}_{K_1}, \ldots, \mu_{q(1/\sigma_{ur}^2)}\boldsymbol{I}_{K_r}\Big\}\Big]^{-1}$$

   $$\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})} \leftarrow \mu_{q(1/\sigma_\varepsilon^2)}\,\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}\,\boldsymbol{C}^T\boldsymbol{y};\quad \mu_{q(1/a_\varepsilon)} \leftarrow 1/\Big\{\mu_{q(1/\sigma_\varepsilon^2)} + A_\varepsilon^{-2}\Big\}$$

   $$\mu_{q(1/\sigma_\varepsilon^2)} \leftarrow \frac{n+1}{2\,\mu_{q(1/a_\varepsilon)} + \boldsymbol{y}^T\boldsymbol{y} - 2\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}^T\boldsymbol{C}^T\boldsymbol{y} + \text{tr}\Big[(\boldsymbol{C}^T\boldsymbol{C})\Big\{\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})} + \boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}^T\Big\}\Big]}$$

   For $\ell = 1, \ldots, r$:

   $$\mu_{q(1/a_{u\ell})} \leftarrow 1/\Big\{\mu_{q(1/\sigma_{u\ell}^2)} + A_{u\ell}^{-2}\Big\}$$

   $$\mu_{q(1/\sigma_{u\ell}^2)} \leftarrow \frac{K_\ell + 1}{2\,\mu_{q(1/a_{u\ell})} + \big\|\boldsymbol{\mu}_{q(\boldsymbol{u}_\ell)}\big\|^2 + \text{tr}\big(\boldsymbol{\Sigma}_{q(\boldsymbol{u}_\ell)}\big)}$$

   until data are no longer available or analysis terminated.

and

$$y_i|\beta_0, \beta_s, \beta_t, \boldsymbol{u}_s, \boldsymbol{u}_t, \sigma_\varepsilon^2 \overset{\text{ind.}}{\sim} N\left(\beta_0 + \beta_s\, s_i + \beta_t\, t_i + \sum_{k=1}^{K_s} u_{s,k}\, z_k^s(s_i) + \sum_{k=1}^{K_t} u_{t,k}\, z_k^t(t_i), \sigma_\varepsilon^2\right),$$

$$1 \le i \le n, \quad \boldsymbol{u}_s = [u_{s,1}, \ldots, u_{s,K_s}]^T, \quad \boldsymbol{u}_t = [u_{t,1}, \ldots, u_{t,K_t}]^T,$$

$$\boldsymbol{u}_s|\,\sigma_{u,s}^2 \sim N\big(0, \sigma_{u,s}^2\,\boldsymbol{I}\big), \quad \boldsymbol{u}_t|\,\sigma_{u,t}^2 \sim N\big(0, \sigma_{u,t}^2\,\boldsymbol{I}\big),$$

$$\beta_0, \beta_s, \beta_t \overset{\text{ind.}}{\sim} N\big(0, \sigma_{\boldsymbol{\beta}}^2\big), \quad \sigma_{u,s} \sim \text{Half-Cauchy}(A_{u,s}), \quad \sigma_{u,t} \sim \text{Half-Cauchy}(A_{u,t}),$$

$$\sigma_\varepsilon \sim \text{Half-Cauchy}(A_\varepsilon). \tag{11}$$

Here and throughout $\overset{\text{ind.}}{\sim}$ denotes "distributed independently."

Model (10) is the random intercept extension of simple linear regression for longitudinal data with (x_{ij}, y_{ij}) denoting the jth predictor/response pair for the ith group, with m denoting the number of groups. There is no intrinsic reason to insist that the observations arrive in

order with respect to the i, j subscripting. Hence, $\boldsymbol{c}_{\text{new}}$ will have the form:

$$\boldsymbol{c}_{\text{new}} = \begin{bmatrix} 1 \\ x_{\text{new}} \\ \boldsymbol{e}_{\text{new}} \end{bmatrix}$$

where x_{new} is the new predictor measurement that partners y_{new} and $\boldsymbol{e}_{\text{new}}$ is a $m \times 1$ vector with an entry of 1 in position i_{new}, corresponding to the group that $(x_{\text{new}}, y_{\text{new}})$ is from, and zeros elsewhere.

Model (11) is a mixed model-based penalized spline version of the additive model

$$y_i = \beta_0 + f_s(s_i) + f_t(t_i) + \varepsilon_i, \quad 1 \le i \le n,$$

where the s_i and t_i are continuous predictor measurements and f_s and f_t are smooth functions. The functions $z_k^s(\cdot)$, $1 \le k \le K_s$, are spline basis functions. A simple example is the truncated line basis

$$z_k^s(s) = \left(s - \kappa_k^s\right)_+, \tag{12}$$

where $\kappa_1^s, \ldots, \kappa_{K_s}^s$ are a set of knots within the domain of the s_i values. More sophisticated, and numerically stable, options for $z_k(s)$ are described in, for example, Wood (2006), Welham et al. (2007), and Wand and Ormerod (2008). We use the last of these, known as O'Sullivan splines, in our examples. The $z_k^t(\cdot)$, $1 \le k \le K_t$, are defined similarly. A key feature of the $z_k^s(\cdot)$ and $z_k^t(\cdot)$ is that the multiple-of-diagonal covariance matrices are appropriate under mixed model representations of penalized splines. This subtlety is explained in Section 4 of Wand and Ormerod (2008).

Online fitting of (11) involves reading in vectors of the form

$$\boldsymbol{c}_{\text{new}} = [1, s_{\text{new}}, t_{\text{new}}, z_1^s(s_{\text{new}}), \ldots, z_{K_s}^s(s_{\text{new}}), z_1^t(t_{\text{new}}), \ldots, z_{K_t}^t(t_{\text{new}})]^T,$$

where s_{new} and t_{new} are the new predictor measurements that partner y_{new}. There is, however, the issue of having to set the spline basis functions in advance. For instance, if the truncated line basis (12) is used then the knots have to be set at or near the start of the algorithm. For many applications, this is not a major problem. For example, if the s_{new} values correspond to age, in years, of human adults then the range of possible s_i values is easy to specify and a reasonable spline basis can be set in advance. In a similar vein, for longitudinal data, Algorithm 3 assumes that the number of groups is set in advance. If the groups correspond to the counties of a geographical entity then this should not pose a problem. If the data are from a medical study then Algorithm 3 assumes that the number of patients and their identity numbers are fixed in advance. If this is not a reasonable assumption then some adjustment is required.

Finally, we mention the possibility of speeding up the most expensive update:

$$\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})} \leftarrow \left[\mu_{q(1/\sigma_\varepsilon^2)}\boldsymbol{C}^T\boldsymbol{C} + \text{blockdiag}\left\{\sigma_\beta^{-2}\boldsymbol{I}_p, \mu_{q(1/\sigma_{u1}^2)}\boldsymbol{I}_{K_1}, \ldots, \mu_{q(1/\sigma_{ur}^2)}\boldsymbol{I}_{K_r}\right\}\right]^{-1}. \tag{13}$$

For Model (10), the matrix requiring inversion has dimension $(2+m) \times (2+m)$. If the number of groups is high then naïve implementation could lead to a bottleneck at (13). In the batch case, it is well known (e.g., Smith and Wand 2008) that $\boldsymbol{C}^T\boldsymbol{C}$ contains diagonal forms that allow $O(m)$ computation of the right-hand side of (13). Such efficiencies are

available in the online case, but require careful rearrangement of the entries of $\boldsymbol{C}^T\boldsymbol{C}$ during the updates.

2.3 Extension to Unstructured Covariance Matrices for Random Effects

A *random intercepts and slopes* extension of (10) is one with the first two hierarchical levels set to

$$y_{ij}|\beta_0,\beta_1,U_i,V_i,\sigma_\varepsilon^2 \overset{\text{ind.}}{\sim} N\big(\beta_0+U_i+(\beta_1+V_i)\,x_{ij},\sigma_\varepsilon^2\big),\quad 1\le i\le m,\quad 1\le j\le n_i,$$

$$\text{and}\quad \begin{bmatrix}U_i\\V_i\end{bmatrix}\Big|\boldsymbol{\Sigma}\sim N(\boldsymbol{0},\boldsymbol{\Sigma}),\quad \text{where}\quad \boldsymbol{\Sigma}\equiv\begin{bmatrix}\sigma_u^2 & \rho_{uv}\,\sigma_u\,\sigma_v\\ \rho_{uv}\,\sigma_u\,\sigma_v & \sigma_v^2\end{bmatrix}$$

is an unstructured 2×2 covariance matrix. The conjugate prior for $\boldsymbol{\Sigma}$ is the Inverse Wishart distribution. However, the specification

$$\boldsymbol{\Sigma}|\,a_{uv1},a_{uv2}\sim\text{Inverse-Wishart}\left(\nu+1,2\nu\begin{bmatrix}1/a_{uv1}&0\\0&1/a_{uv2}\end{bmatrix}\right),$$

$$a_{uv1},a_{uv2}\overset{\text{ind.}}{\sim}\text{Inverse-Gamma}\left(\frac{1}{2},1/A_{uv}\right),\quad \nu,A_{uv}>0$$

provides a covariance matrix extension of $\sigma_u\sim$ Half-Cauchy(A_u). The choice $\nu=2$ is particularly attractive since it imposes a Uniform$(-1,1)$ distribution on ρ_{uv} and Half-t_2 distributions on σ_u and σ_v. This is laid out in Huang and Wand (2013), including the definition of the Inverse-Wishart$(a,\boldsymbol{B})$ distribution.

Extensions to more sophisticated models, possibly having larger unstructured covariance matrices, can be done in a similar fashion.

2.4 Extension to Sparse Shrinkage Penalties

Model (6) involves the following Gaussian penalization on subvectors of $\boldsymbol{u}$:

$$\boldsymbol{u}_\ell|\,\sigma_{u\ell}^2\sim N\big(\boldsymbol{0},\sigma_{u\ell}^2\boldsymbol{I}\big),\quad 1\le\ell\le r. \tag{14}$$

However, many models of current-day interest, such as *wide data ("$p\gg n$")* and *wavelet* regression, require an assumption that the regression coefficients are sparse. Under such sparseness assumptions, the Gaussian priors (14) are not appropriate since they induce a relatively gentle amount of penalization that lacks the ability to annihilate regression coefficients during fitting and inference.

For simplicity of exposition we will confine discussion of the sparse shrinkage extension to the $r=1$ version of (6). Hence, we retain

$$y\,|\,\boldsymbol{\beta},\boldsymbol{u},\sigma_\varepsilon^2\sim N\big(\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{Z}\,\boldsymbol{u},\sigma_\varepsilon^2\boldsymbol{I}\big)$$

without any subdivision of $\boldsymbol{u}$. Let K be the dimension of $\boldsymbol{u}$ and consider general mutually independent prior penalizations of the form

$$u_k\overset{\text{ind.}}{\sim}p(u;\sigma_u,\boldsymbol{\theta}),$$

where $p(\cdot;\sigma_u,\boldsymbol{\theta})$ is a density function with scale parameter σ_u and shape parameter $\boldsymbol{\theta}$. Options for $p(u;1,\boldsymbol{\theta})$ include:

$$
\begin{aligned}
p(u;1,w) &= w\left\{\frac{1}{2}\exp(-|u|)\right\} + (1-w)\,\delta_0(u) && \text{(Laplace-Zero)},\\
p(u;1) &= (2\pi^3)^{-1/2}\exp(u^2/2)E_1(u^2/2) && \text{(Horseshoe)},\\
p(u;1,\lambda) &= \frac{\lambda\, 2^{\lambda}\Gamma\left(\lambda+\frac{1}{2}\right)}{\pi^{1/2}}\exp(u^2/4)\,D_{-2\lambda-1}(|u|) && \text{(Normal-Exponential-Gamma)}\\
&\text{and}\\
p(u;1,\lambda) &= \frac{1}{2(1+|u|/\lambda)^{\lambda+1}} && \text{(Generalized Double Pareto).}
\end{aligned}
\tag{15}
$$

Here δ_0 denotes the Dirac delta function with mass at zero. Also, E_1 denotes the exponential integral function of order 1 and D_ν denotes the parabolic cylinder function of order ν according to the definitions of Gradshteyn and Ryzhik (1994). References for the development of these sparse shrinkage priors are Johnstone and Silverman (2005) (Laplace-Zero), Carvalho, Polson, and Scott (2010) (Horseshoe), Griffin and Brown (2011) (Normal-Exponential-Gamma), and Armagan, Dunson, and Lee (2013) (Generalized Double Pareto).

Batch MFVB algorithms for the priors (15) recently have been derived by Wand and Ormerod (2011) (Laplace-Zero prior) and Neville, Ormerod, and Wand (2013) (Horseshoe, Normal-Exponential-Gamma and Generalized Double Pareto priors).

Algorithm 4 is the online adaptation of Algorithm 4 of Wand and Ormerod (2011) for the Laplace-Zero prior model:

$$
\begin{aligned}
&\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{v},\sigma_\varepsilon^2 \sim N\big(\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{Z}(\boldsymbol{\gamma}\odot\boldsymbol{v}),\sigma_\varepsilon^2\boldsymbol{I}\big), \quad \boldsymbol{v}|\,\sigma_u^2,\boldsymbol{b}\sim N\big(\boldsymbol{0},\sigma_u^2\,\mathrm{diag}(\boldsymbol{b})^{-1}\big),\\
&\sigma_u^2|\,a_u \sim \text{Inverse-Gamma}\left(\frac{1}{2},1/a_u\right), \quad \sigma_\varepsilon^2|\,a_\varepsilon \sim \text{Inverse-Gamma}\left(\frac{1}{2},1/a_\varepsilon\right),\\
&\boldsymbol{\beta}\sim N\big(\boldsymbol{0},\sigma_\beta^2\boldsymbol{I}\big),\ a_u\sim\text{Inverse-Gamma}\left(\frac{1}{2},1/A_u^2\right),\ a_\varepsilon\sim\text{Inverse-Gamma}\left(\frac{1}{2},1/A_\varepsilon^2\right),\\
&b_k \overset{\text{ind.}}{\sim} \text{Inverse-Gamma}\left(1,\frac{1}{2}\right), \quad \gamma_k|\,\rho \overset{\text{ind.}}{\sim} \text{Bernoulli}(\rho), \quad \rho\sim\text{Beta}(A_\rho,B_\rho).
\end{aligned}
\tag{16}
$$

Note that $\boldsymbol{A}\odot\boldsymbol{B}$ denotes the element-wise product of matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ having the same dimensions. Model (16) is a reproduction of (30) in Wand and Ormerod (2011) and the additional notation is explained there. Note, in particular, that the Laplace-Zero prior is handled via the introduction of auxiliary variables $\boldsymbol{b}$, $\boldsymbol{\gamma}$, and $\boldsymbol{v}$. Section 3.6 of Wand and Ormerod (2011) provides the necessary details. Similar online MFVB algorithms for the continuous sparse signal shrinkage priors listed in (15) follow from the batch MFVB algorithms of Neville, Ormerod, and Wand (2013).

As with the spline-based semiparametric regression models described in Section 2.2, the wavelet-based models described here benefit from the *low-rank* property laid out in Section 3.1 of Wand and Ormerod (2011). This property entails that the basis functions are fixed once and for all during the warm-up period. This permits fast updating of wavelet nonparametric fits as new data arrive. A cost of this approach is that the domain of predictors

Algorithm 4 Mean field variational Bayes algorithm for the determination of the optimal parameters in $q^*(\boldsymbol{\beta}, v)$, $q^*(\boldsymbol{\gamma})$, $q^*(\sigma_u^2)$, and $q^*(\sigma_\varepsilon^2)$ for the Bayesian sparse signal regression model (16)

1. Perform batch-based tuning runs analogous to those described in Algorithm 2′ and determine a warm-up sample size n_{warm} for which convergence is validated. The batch MFVB algorithm is Algorithm 4 of Wand and Ormerod (2011).

2. Set $\boldsymbol{y}_{\text{warm}}$ and $\boldsymbol{C}_{\text{warm}} = [\mathbf{1}\ \boldsymbol{Z}_{\text{warm}}]$ to be the response vector and design matrix based on the first n_{warm} observations. Then set $\boldsymbol{y}^T\boldsymbol{y} \leftarrow \boldsymbol{y}_{\text{warm}}^T\boldsymbol{y}_{\text{warm}}$, $\boldsymbol{Z}^T\mathbf{1} \leftarrow \boldsymbol{Z}_{\text{warm}}^T\mathbf{1}$, $\boldsymbol{Z}^T\boldsymbol{y} \leftarrow \boldsymbol{Z}_{\text{warm}}^T\boldsymbol{y}$, $\boldsymbol{Z}^T\boldsymbol{Z} \leftarrow \boldsymbol{Z}_{\text{warm}}^T\boldsymbol{Z}_{\text{warm}}$, $\boldsymbol{C}^T\boldsymbol{y} \leftarrow \boldsymbol{C}_{\text{warm}}^T\boldsymbol{y}_{\text{warm}}$, $\boldsymbol{C}^T\boldsymbol{C} \leftarrow \boldsymbol{C}_{\text{warm}}^T\boldsymbol{C}_{\text{warm}}$, $n \leftarrow n_{\text{warm}}$. Set K to be the number of columns in $\boldsymbol{Z}_{\text{warm}}$. Also, set $\mu_{q(1/\sigma_\varepsilon^2)}$, $\mu_{q(1/\sigma_u^2)}$, $\mu_{q(1/a_\varepsilon)}$, $\mu_{q(1/a_u)}$, $\boldsymbol{\mu}_{q(\boldsymbol{b})}$, $\boldsymbol{\mu}_{q(\boldsymbol{w}_\gamma)}$ and $\boldsymbol{\Omega}_{q(\boldsymbol{w}_\gamma)}$ to be the values for these quantities obtained in the batch-based tuning run with sample size n_{warm}.

3. Cycle:

read in y_{new} (1×1) and $\boldsymbol{z}_{\text{new}}$ $(K \times 1)$; $n \leftarrow n+1$; $\boldsymbol{c}_{\text{new}} \leftarrow \begin{bmatrix} 1 \\ \boldsymbol{z}_{\text{new}} \end{bmatrix}$

$$\boldsymbol{y}^T\boldsymbol{y} \leftarrow \boldsymbol{y}^T\boldsymbol{y} + y_{\text{new}}^2;\ \boldsymbol{Z}^T\mathbf{1} \leftarrow \boldsymbol{Z}^T\mathbf{1} + \boldsymbol{z}_{\text{new}};\ \boldsymbol{Z}^T\boldsymbol{y} \leftarrow \boldsymbol{Z}^T\boldsymbol{y} + \boldsymbol{z}_{\text{new}}\, y_{\text{new}}$$

$$\boldsymbol{Z}^T\boldsymbol{Z} \leftarrow \boldsymbol{Z}^T\boldsymbol{Z} + \boldsymbol{z}_{\text{new}}\boldsymbol{z}_{\text{new}}^T;\ \boldsymbol{C}^T\boldsymbol{y} \leftarrow \boldsymbol{C}^T\boldsymbol{y} + \boldsymbol{c}_{\text{new}}\, y_{\text{new}};\ \boldsymbol{C}^T\boldsymbol{C} \leftarrow \boldsymbol{C}^T\boldsymbol{C} + \boldsymbol{c}_{\text{new}}\boldsymbol{c}_{\text{new}}^T$$

$$\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},v)} \leftarrow \left(\mu_{q(1/\sigma_\varepsilon^2)}(\boldsymbol{C}^T\boldsymbol{C}) \odot \boldsymbol{\Omega}_{q(\boldsymbol{w}_\gamma)} + \begin{bmatrix} \sigma_\beta^{-2} & \mathbf{0} \\ \mathbf{0} & \mu_{q(1/\sigma_u^2)}\text{diag}\left(\boldsymbol{\mu}_{q(\boldsymbol{b})}\right) \end{bmatrix}\right)^{-1}$$

$$\boldsymbol{\mu}_{q(\boldsymbol{\beta},v)} \leftarrow \mu_{q(1/\sigma_\varepsilon^2)}\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},v)}\text{diag}\{\boldsymbol{\mu}_{q(\boldsymbol{w}_\gamma)}\}\boldsymbol{C}^T\boldsymbol{y}$$

$$\mu_{q(\boldsymbol{b})} \leftarrow [\mu_{q(1/\sigma_u^2)}\{\text{diagonal}(\Sigma_{q(v)}) + \boldsymbol{\mu}_{q(v)}^2\}]^{-1/2}$$

$$\begin{aligned}\boldsymbol{\eta}_{q(\boldsymbol{\gamma})} \leftarrow\ & -\frac{1}{2}\mu_{q(1/\sigma_\varepsilon^2)}[\text{diagonal}(\boldsymbol{Z}^T\boldsymbol{Z}) \odot \{\boldsymbol{\sigma}_{q(v)}^2 + \boldsymbol{\mu}_{q(v)}^2\} - 2(\boldsymbol{Z}^T\boldsymbol{y}) \odot \boldsymbol{\mu}_{q(v)} \\ & + 2(\boldsymbol{Z}^T\mathbf{1}) \odot \{[\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},v)})]_{i=1,2\le j\le K+1} + \mu_{q(\beta)}\boldsymbol{\mu}_{q(v)}\} \\ & + 2\,\text{diagonal}\left\{\boldsymbol{Z}^T\boldsymbol{Z}\,\text{diag}\left(\boldsymbol{\mu}_{q(\boldsymbol{\gamma})}\right)\boldsymbol{\Sigma}_{q(v)}\right\} \\ & - 2\,\text{diagonal}(\boldsymbol{Z}^T\boldsymbol{Z}) \odot \boldsymbol{\mu}_{q(\boldsymbol{\gamma})} \odot \text{diagonal}(\boldsymbol{\Sigma}_{q(v)}) \\ & + 2\boldsymbol{\mu}_{q(v)} \odot \left\{\boldsymbol{Z}^T\boldsymbol{Z}(\boldsymbol{\mu}_{q(\boldsymbol{\gamma})} \odot \boldsymbol{\mu}_{q(v)}) - \text{diagonal}(\boldsymbol{Z}^T\boldsymbol{Z}) \odot \boldsymbol{\mu}_{q(\boldsymbol{\gamma})} \odot \boldsymbol{\mu}_{q(v)}\right\}] \\ & + \psi(A_\rho + \mu_{q(\gamma_\bullet)}) - \psi(B_\rho + K - \mu_{q(\gamma_\bullet)})\end{aligned}$$

$$\mu_{q(\boldsymbol{\gamma})} \leftarrow \frac{\exp(\eta_{q(\boldsymbol{\gamma})})}{1+\exp(\eta_{q(\boldsymbol{\gamma})})};\ \boldsymbol{\mu}_{q(\boldsymbol{w}_\gamma)} \leftarrow \begin{bmatrix} 1 \\ \boldsymbol{\mu}_{q(\boldsymbol{\gamma})} \end{bmatrix};\ \mu_{q(\gamma_\bullet)} \leftarrow \textstyle\sum_{k=1}^K \mu_{q(\gamma_k)}$$

$$\boldsymbol{\Omega}_{q(\boldsymbol{w}_\gamma)} \leftarrow \text{diag}\{\boldsymbol{\mu}_{q(\boldsymbol{w}_\gamma)} \odot (\mathbf{1} - \boldsymbol{\mu}_{q(\boldsymbol{w}_\gamma)})\} + \boldsymbol{\mu}_{q(\boldsymbol{w}_\gamma)}\boldsymbol{\mu}_{q(\boldsymbol{w}_\gamma)}^T$$

$$\mu_{q(1/a_\varepsilon)} \leftarrow 1/\{\mu_{q(1/\sigma_\varepsilon^2)} + A_\varepsilon^{-2}\};\ \mu_{q(1/a_u)} \leftarrow 1/\{\mu_{q(1/\sigma_u^2)} + A_u^{-2}\}$$

$$\begin{aligned}B_{q(\sigma_\varepsilon^2)} \leftarrow\ & \mu_{q(1/a_\varepsilon)} + \frac{1}{2}\boldsymbol{y}^T\boldsymbol{y} - (\boldsymbol{\mu}_{q(\boldsymbol{w}_\gamma)} \odot \mu_{q(\boldsymbol{\beta},v)})^T\boldsymbol{C}^T\boldsymbol{y} \\ & + \frac{1}{2}\text{tr}(\boldsymbol{C}^T\boldsymbol{C}\ [\boldsymbol{\Omega}_{q(\boldsymbol{w}_\gamma)} \odot \{\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},v)} + \boldsymbol{\mu}_{q(\boldsymbol{\beta},v)}\boldsymbol{\mu}_{q(\boldsymbol{\beta},v)}^T\}])\end{aligned}$$

$$B_{q(\sigma_u^2)} \leftarrow \mu_{q(1/a_u)} + \frac{1}{2}\boldsymbol{\mu}_{q(\boldsymbol{b})}^T\{\text{diagonal}\left(\boldsymbol{\Sigma}_{q(v)}\right) + \boldsymbol{\mu}_{q(v)}^2\}$$

$$\mu_{q(1/\sigma_u^2)} \leftarrow \frac{1}{2}(K+1)/B_{q(\sigma_u^2)};\ \mu_{q(1/\sigma_\varepsilon^2)} \leftarrow \frac{1}{2}(n+1)/B_{q(\sigma_\varepsilon^2)}$$

until data no longer available or analysis terminated.

needs to be specified based on the warm-up data. As explained in Section 2.2, this will often be reasonable. Of course, there is always the possibility of new predictor values landing outside domain of the basis functions, in which case some modification may be necessary.

Figure 4 illustrates online wavelet nonparametric regression for data generated via

$$x_{\text{new}} \sim \text{Uniform}(0,1), \quad y_{\text{new}} | \, x_{\text{new}} \sim N(f_{\text{WO}}(x_{\text{new}}), 1),$$

where f_{WO} is defined by (20) of Wand and Ormerod (2011). The warm-up sample size is $n_{\text{warm}} = 300$. The desired improvement in the estimate of f_{WO} as n increases is clearly apparent. Convergence to the batch MFVB estimate was found to be excellent in this case.

3. BINARY RESPONSE MODELS

The binary response model we consider here takes the same form as (6) and (7), but with σ_ε removed and

$$\boldsymbol{y} | \, \boldsymbol{\beta}, \boldsymbol{u} \sim \text{Bernoulli}\{\text{logit}^{-1}(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u})\}. \tag{17}$$

Note that (17) is a convenient shorthand for the entries of $\boldsymbol{y}$, conditional on $(\boldsymbol{\beta}, \boldsymbol{u})$, being independent and with ith entry $\text{Bernoulli}[\text{logit}^{-1}\{(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u})_i\}]$.

Batch MFVB algorithms for approximate inference in (17), (6), and (7) start with the product restriction

$$p\left(\boldsymbol{\beta}, \boldsymbol{u}, a_{u1}, \ldots, a_{ur}, \sigma^2_{u1}, \ldots, \sigma^2_{ur} \big| \boldsymbol{y}\right) \approx q(\boldsymbol{\beta}, \boldsymbol{u}, a_{u1}, \ldots, a_{ur})\, q\left(\sigma^2_{u1}, \ldots, \sigma^2_{ur}\right).$$

The resultant updates for the $\sigma^2_{u\ell}$ and $a_{u\ell}$ are the same as in the Gaussian response case. The optimal q-density for $(\boldsymbol{\beta}, \boldsymbol{u})$ satisfies

$$q^*(\boldsymbol{\beta}, \boldsymbol{u}) \propto \exp\left\{\boldsymbol{y}^T(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u}) - \boldsymbol{1}^T \log(1 + e^{\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u}}) - \frac{1}{2\sigma^2_\beta}\|\boldsymbol{\beta}\|^2 \right.$$
$$\left. -\frac{1}{2}\sum_{\ell=1}^{L} \mu_{q(1/\sigma^2_{u\ell})} \|\boldsymbol{u}_\ell\|^2 \right\}. \tag{18}$$

However, this is a nonstandard form and poses tractability problems with regard to approximate inference for $(\boldsymbol{\beta}, \boldsymbol{u})$. A reasonable remedy is to replace (18) by a member of the following family of multivariate Normal approximations:

$$\underline{q}^*(\boldsymbol{\beta}, \boldsymbol{u}) \sim N(\boldsymbol{\mu}_{\underline{q}(\boldsymbol{\beta},\boldsymbol{u};\boldsymbol{\xi})}, \boldsymbol{\Sigma}_{\underline{q}(\boldsymbol{\beta},\boldsymbol{u};\boldsymbol{\xi})}),$$

where

$$\boldsymbol{\Sigma}_{\underline{q}(\boldsymbol{\beta},\boldsymbol{u})} \equiv \left[2\boldsymbol{C}^T \text{diag}\{\lambda(\boldsymbol{\xi})\}\boldsymbol{C} + \text{blockdiag}\left\{\sigma_\beta^{-2}\boldsymbol{I}_p, \mu_{q(1/\sigma^2_{u1})}\boldsymbol{I}_{K_1}, \ldots, \mu_{q(1/\sigma^2_{ur})}\boldsymbol{I}_{K_r}\right\}\right]^{-1},$$

$\boldsymbol{\xi}$ is an $n \times 1$ vector of positive *variational* parameters, $\lambda(x) \equiv \tanh(x/2)/(4x)$, and

$$\boldsymbol{\mu}_{\underline{q}(\boldsymbol{\beta},\boldsymbol{u})} \equiv \boldsymbol{\Sigma}_{\underline{q}(\boldsymbol{\beta},\boldsymbol{u})}\boldsymbol{C}^T\left(\boldsymbol{y} - \frac{1}{2}\boldsymbol{1}\right)$$

with $\boldsymbol{C} = [\boldsymbol{X}\ \boldsymbol{Z}]$ as before. This family of approximations is due to Jaakkola and Jordan (2000) and its genesis is given there. Section 3.1 of Ormerod and Wand (2010) explains

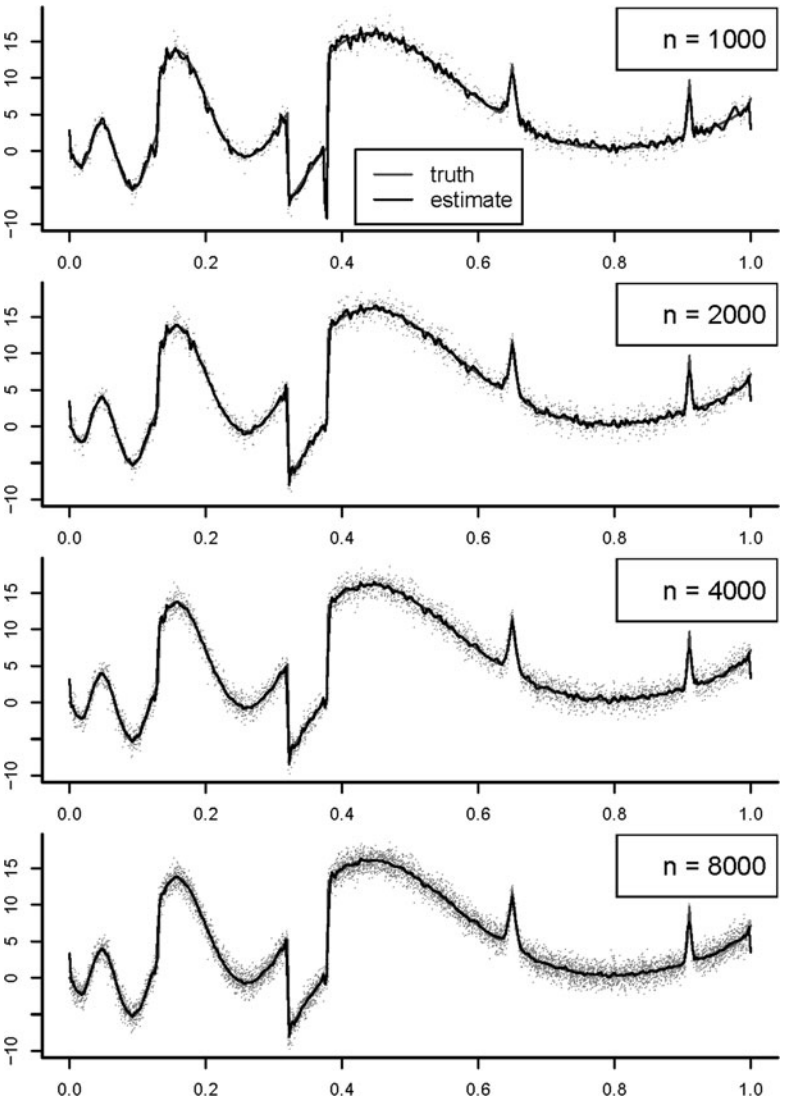

Figure 4. Examples of online MFVB wavelet fits based on Algorithm 4. The true regression curve is the function f_{WO} defined in Wand and Ormerod (2011).

this approximation strategy using notation similar to that used here. Jaakkola and Jordan (2000) also present an expectation-maximization argument that results in

$$\boldsymbol{\xi} \leftarrow \sqrt{\text{diagonal}\left[\boldsymbol{C}\left\{\boldsymbol{\Sigma}_{\underline{q}(\boldsymbol{\beta},\boldsymbol{u};\boldsymbol{\xi}))} + \boldsymbol{\mu}_{\underline{q}(\boldsymbol{\beta},\boldsymbol{u};\boldsymbol{\xi})}\boldsymbol{\mu}^T_{\underline{q}(\boldsymbol{\beta},\boldsymbol{u};\boldsymbol{\xi})}\right\}\boldsymbol{C}^T\right]}$$

being the optimal update for the $\boldsymbol{\xi}$ vector. Algorithm 5 is the online MFVB algorithm that arises from appropriately modifying the batch MFVB algorithm for (17) with the Jaakkola and Jordan (2000) strategy.

An alternative route to an online MFVB algorithm for binary response linear mixed models involves the probit link and the Albert and Chib (1993) auxiliary variable strategy.

Batch MFVB algorithms for models of this general type have been developed by Girolami and Rogers (2006) and Consonni and Marin (2007). Modification of these algorithms for the probit link version of (17) should lead to an algorithm that performs online approximate inference similar to that performed by Algorithm 5.

Algorithm 5 Online mean field variational Bayes algorithm for approximate inference in the binary response logistic mixed model (17)

1. Perform batch-based tuning runs analogous to those described in Algorithm 2′ and determine a warm-up sample size n_{warm} for which convergence is validated.
2. Set $\boldsymbol{y}_{\text{warm}}$ and $\boldsymbol{C}_{\text{warm}}$ to be the response vector and design matrix, and $\boldsymbol{\xi}_{\text{warm}}$ to be the vector of variational parameters, based on the first n_{warm} observations. Then set $\boldsymbol{C}^T\left(\boldsymbol{y} - \frac{1}{2}\mathbf{1}\right) \leftarrow \boldsymbol{C}_{\text{warm}}^T(\boldsymbol{y}_{\text{warm}} - \frac{1}{2}\mathbf{1})$, $\boldsymbol{C}^T \text{diag}\{\lambda(\boldsymbol{\xi})\}\boldsymbol{C} \leftarrow \boldsymbol{C}_{\text{warm}}^T \text{diag}\{\lambda(\boldsymbol{\xi}_{\text{warm}})\}\boldsymbol{C}_{\text{warm}}$, $n \leftarrow n_{\text{warm}}$. Also, set $\boldsymbol{\mu}_{\underline{q}(\boldsymbol{\beta},\boldsymbol{u};\boldsymbol{\xi})}$, $\boldsymbol{\Sigma}_{\underline{q}(\boldsymbol{\beta},\boldsymbol{u};\boldsymbol{\xi})}$, $\mu_{q(1/\sigma_{u1}^2)}, \ldots, \mu_{q(1/\sigma_{ur}^2)}$ to be the values for these quantities obtained in the batch-based tuning run with sample size n_{warm}.
3. Cycle:

   read in y_{new} (1×1) and $\boldsymbol{c}_{\text{new}}$ $(P \times 1)$; $n \leftarrow n + 1$

   $$\xi \leftarrow \sqrt{\boldsymbol{c}_{\text{new}}^T\left\{\boldsymbol{\Sigma}_{\underline{q}(\boldsymbol{\beta},\boldsymbol{u};\boldsymbol{\xi})} + \boldsymbol{\mu}_{\underline{q}(\boldsymbol{\beta},\boldsymbol{u};\boldsymbol{\xi})}\boldsymbol{\mu}_{\underline{q}(\boldsymbol{\beta},\boldsymbol{u};\boldsymbol{\xi})}^T\right\}\boldsymbol{c}_{\text{new}}}$$

   $$\boldsymbol{C}^T\left(\boldsymbol{y} - \frac{1}{2}\mathbf{1}\right) \leftarrow \boldsymbol{C}^T\left(\boldsymbol{y} - \frac{1}{2}\mathbf{1}\right) + \boldsymbol{c}_{\text{new}}\left(y_{\text{new}} - \frac{1}{2}\right)$$

   $$\boldsymbol{C}^T \text{diag}\{\lambda(\boldsymbol{\xi})\}\boldsymbol{C} \leftarrow \boldsymbol{C}^T \text{diag}\{\lambda(\boldsymbol{\xi})\}\boldsymbol{C} + \lambda(\xi)\,\boldsymbol{c}_{\text{new}}\,\boldsymbol{c}_{\text{new}}^T$$

   $$\boldsymbol{\Sigma}_{\underline{q}(\boldsymbol{\beta},\boldsymbol{u})} \leftarrow \left[2\boldsymbol{C}^T \text{diag}\{\lambda(\boldsymbol{\xi})\}\,\boldsymbol{C} + \text{blockdiag}\{\sigma_\beta^{-2}\,\boldsymbol{I}_p, \mu_{q(1/\sigma_{u1}^2)}\boldsymbol{I}_{K_1}, \ldots, \mu_{q(1/\sigma_{ur}^2)}\boldsymbol{I}_{K_r}\}\right]^{-1}$$

   $$\boldsymbol{\mu}_{\underline{q}(\boldsymbol{\beta},\boldsymbol{u})} \leftarrow \boldsymbol{\Sigma}_{\underline{q}(\boldsymbol{\beta},\boldsymbol{u})}\,\boldsymbol{C}^T\left(\boldsymbol{y} - \frac{1}{2}\mathbf{1}\right)$$

   For $\ell = 1, \ldots, r$:

   $$\mu_{q(1/a_{u\ell})} \leftarrow 1 \Big/ \left\{\mu_{q(1/\sigma_{u\ell}^2)} + A_{u\ell}^{-2}\right\}$$

   $$\mu_{q(1/\sigma_{u\ell}^2)} \leftarrow \frac{K_\ell + 1}{2\,\mu_{q(1/a_{u\ell})} + \left\|\boldsymbol{\mu}_{q(\boldsymbol{u}_\ell)}\right\|^2 + \text{tr}\left(\boldsymbol{\Sigma}_{q(\boldsymbol{u}_\ell)}\right)}$$

   until data no longer available or analysis terminated.

Figure 5 performs batch-based convergence diagnostics for a binary response nonparametric regression example. This is a special case of (17) with $r = 1$ and $\boldsymbol{Z}$ containing spline basis functions. New predictor/response pairs $(x_{\text{new}}, y_{\text{new}})$ were generated according to

$$x_{\text{new}} \sim \text{Uniform}(0, 1), \quad y_{\text{new}}|x_{\text{new}} \sim \text{Bernoulli}(\text{logit}^{-1}(\cos(4\pi\, x_{\text{new}}) + 2\,x_{\text{new}} - 1)). \tag{19}$$

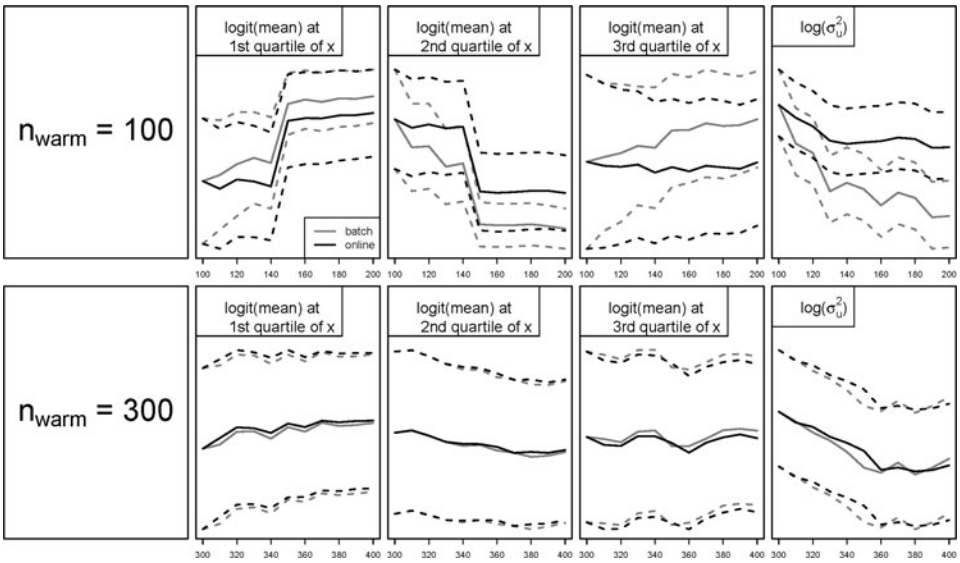

Figure 5. Convergence diagnostics for a binary response nonparametric regression example with data generated according to (19). The solid lines track the posterior means, while the dashed lines show corresponding 95% credible sets. First row: the horizontal axes show the sample sizes between a warm-up batch sample of size $n_{\text{warm}} = 100$ and validation sample sizes up to $n_{\text{valid}} = 100$ greater than n_{warm}. Second row: as for the first row, but with $n_{\text{warm}} = 300$.

The analogs of Steps 1–5 of Algorithm 2′ were applied with an initial trial involving $n_{\text{warm}} = 100$ and $n_{\text{valid}} = 100$. The Bayes estimates and 95% credible sets of the logit-transformed mean function at each of the quartiles of the x-values, as well as $\log(\sigma_u^2)$, are shown in the upper row of Figure 5. However, they have noticeable disagreement, which indicates nonconvergence of the online MFVB results to their batch counterparts and that n_{warm} should be increased. Setting $n_{\text{warm}} = 300$ leads to the more concordant results shown in the lower row of Figure 5, indicating adequacy of this warm-up size. We have found this behavior typical for binary response online MFVB and this simple example demonstrates the importance of batch-based tuning and convergence diagnostics.

4. JUSTIFICATION FOR USING MEAN FIELD VARIATIONAL BAYES

Our use of online mean field variational Bayes is founded upon it being the only approach of which we are aware that (a) is readily extendible to a wide range of semiparametric regression models and (b), in the case of streaming data, has the ability to perform fast approximate inference for all model parameters.

Various other approaches, such as stochastic gradient descent, Markov chain Monte Carlo, and expectation-maximization, can be ruled out since they fall short on at least one of these criteria. We now provide brief reasoning for their elimination from contention for real-time semiparametric regression.

Stochastic gradient descent (e.g., Zhang 2004) allows for regularized regression models to be fitted in an online fashion. Recently Langford, Li, and Zhang (2009) devised stochastic gradient methodology for sparse signal regression. However, in both Zhang (2004) and Langford, Li, and Zhang (2009), the regularization parameters need to be input. This is in contrast to Algorithms 3 and 4 in which the regularization parameters are embedded in the underlying Bayesian model in the form of variance parameters. This allows online estimation of the optimal amount of regularization. It appears that current stochastic gradient descent technology does not support online estimation of regularization parameters.

Markov chain Monte Carlo (MCMC) has analogs with MFVB but is much more computationally expensive. The full conditional distributions depend on the same matrix algebraic forms, such as $\boldsymbol{y}^T\boldsymbol{y}$, $\boldsymbol{C}^T\boldsymbol{y}$ and $\boldsymbol{C}^T\boldsymbol{C}$, that appear in the batch MFVB algorithms for our semiparametric regression models. As shown in Algorithms 3–5, these forms are simple to update whenever a new vector of observations arrives. But MCMC then requires multiple sampling from the resulting full conditional distributions. This is much more expensive than MFVB's arithmetic updates. For streaming data, this heavy computational burden will tend to rule out MCMC.

Expectation-maximization (EM) analogs of Algorithm 3, but for frequentist linear mixed models, are given in Sections 14.2a and 14.2b of McCulloch, Searle, and Neuhaus (2008). They are similar in nature to batch MFVB algorithms such as Algorithm 3 of Ormerod and Wand (2010) and, therefore, can be readily adapted for online processing. Estimates of the precision are not included and further computing, possibly involving the Louis (1982) methodology, is required for online inference. Moreover, the handling of sparse shrinkage penalties and binary response variables requires considerably more complicated EM algorithms, and require approximation, such as Laplace's method, to be computationally feasible. In summary, an EM approach may lead to viable real-time semiparametric regression algorithms, but they would be much more complicated than Algorithms 2–5.

Finally, we mention Newton-Raphson optimization of the likelihood within a frequentist framework (e.g., Section 14.2c of McCulloch, Searle, and Neuhaus 2008). For streaming data, there is the problem of how to keep track of convergence of the Newton-Raphson schemes as data continually arise. The modification for sparse signal penalties looks particularly challenging. The binary response case also involves intractable forms which necessitate approximations such as those based on Laplace's method.

5. INFERENTIAL ACCURACY

Algorithms 2–5 perform real-time approximate Bayesian inference for the model parameters. We now discuss the quality of the approximations induced by the mean field assumptions.

Inferential accuracy of MFVB is a relatively new and modestly studied area of statistical research. There have been a few theoretical contributions, such as Wang and Titterington (2005), and simulation studies, such as those presented in Faes, Ormerod, and Wand (2011) and Menictas and Wand (2013) for Gaussian response models. Menictas and Wand (2013) also provided some heuristic arguments, based on likelihood theory, for why mean field approximations such as (3) and (9) can be highly accurate for Gaussian response models of Section 2. The essential reason is parameter orthogonality between the coefficient

parameters and variance parameters. We also conducted a small simulation study for the single predictor binary response model corresponding to Figure 5 and found MFVB to have very good accuracy in this case.

A broad summary is that MFVB exhibits good to excellent inferential accuracy for all models considered in the present article. However, we acknowledge that the simulation studies on which we base this summary are relatively limited.

6. LIVE INTERNET DEMONSTRATIONS

We have launched the website: *realtime-semiparametric-regression.net* for displaying live real-time semiparametric regression analyses. Links on this website point to several examples, and we anticipate that the set of examples will grow during the next few years. At the time of this writing, the examples involve simulated data and three types of real-time data: stock prices from the U.S. National Association of Securities Dealers Automated Quotations (NASDAQ) and the London Stock Exchange in the United Kingdom, features of property rentals in Sydney, Australia, and data on delays in U.S. domestic flights.

6.1 SIMULATED DATA

Our lead-off examples involve synthetic data. First consider the Gaussian additive model

$$y_i|\boldsymbol{\beta}, \boldsymbol{u}_4, \boldsymbol{u}_5, \boldsymbol{u}_6, \sigma_\varepsilon^2 \sim N\big(\beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + f_4(x_{4i}) + f_5(x_{5i}) + f_6(x_{6i}), \sigma_\varepsilon^2\big), \tag{20}$$

where, for $j = 4, 5, 6$, $\boldsymbol{u}_j$ is vector of spline coefficients for f_j. We generated 30,000 observations from (20) with $x_{1i}, x_{2i}, x_{3i} \overset{\text{ind.}}{\sim} \text{Bernoulli}(\frac{1}{2})$ and $x_{4i}, x_{5i}, x_{6i} \overset{\text{ind.}}{\sim} N(0, 1)$. Truth was set according to $\beta_1 = 0.2$, $\beta_2 = -0.3$, $\beta_3 = 0.6$, $f_4(x) = 2\Phi(6x - 3)$, $f_5(x) = \sin(3\pi x^3)$, $f_6(x) = \cos(4\pi x)$ and $\sigma_\varepsilon^2 = 1$. The link `Gaussian additive model` on the above-mentioned website points to a movie showing summaries of the regression fits when the data are sequentially fed into Algorithm 3.

The `Logistic additive model` link points to a similar movie, but with data generated from the logistic additive model

$$y_i \,|\, \boldsymbol{\beta}, \boldsymbol{u}_2, \boldsymbol{u}_3 \sim \text{Bernoulli}(\text{logit}^{-1}(\beta_1 x_{1i} + f_2(x_{2i}) + f_3(x_{3i})))$$

with $x_{1i} \overset{\text{ind.}}{\sim} \text{Bernoulli}(\frac{1}{2})$, $x_{2i}, x_{3i} \overset{\text{ind.}}{\sim} N(0, 1)$ and truth set at $\beta_1 = 0.2$, $f_2(x) = \cos(4\pi x) + 2x$ and $f_3(x) = \sin(2\pi x^2)$.

Finally, the `Wavelet regression` link corresponds to the simulation setting used to produce Figure 4, with description given in Section 2.4.

6.2 STOCK PRICE DATA

In this set of examples, the predictor and response variable pairs correspond to pairs of stock prices. An example nonparametric regression model is

$$\begin{aligned}&(\texttt{Microsoft stock price})_i \,|\, \boldsymbol{\beta}, \boldsymbol{u}, \sigma_\varepsilon^2 \\ &\overset{\text{ind.}}{\sim} N\left(\beta_0 + f((\texttt{Intel stock price})_i), \sigma_\varepsilon^2\right),\end{aligned} \tag{21}$$

where $f(x) = \beta_1 x + \sum_{k=1}^{K} u_k z_k(x)$ is a penalized spline function as described in Section 2.2 with the same distributional structures imposed on the model parameters. In addition, $(\texttt{Microsoft stock price})_i$ and $(\texttt{Intel stock price})_i$ denote the ith stock price for the U.S. companies Microsoft Corporation and Intel Corporation, respectively, for the current trading day. The website displays fitting of (21) in real time during the NASDAQ opening hours (9:30 a.m. to 4:00 p.m. North American Eastern Standard Time). The R package `quantmod` (Ryan 2012) is used to obtain the NASDAQ data from the Yahoo! Finance website (`finance.yahoo.com`).

A similar set of examples is based on London Stock Exchange data during its opening hours (8:00 a.m. to 4:20 p.m. Greenwich Mean Time). Note that Yahoo! Finance delays London Stock Exchange data by 20 minutes.

Depending on the example and the live dataset, the appropriateness of the nonparametric regression model (21) may be questionable and more sophisticated models could be entertained. Hence, these examples should only be viewed as simple illustrations of the concept of real-time semiparametric regression.

6.3 SYDNEY PROPERTY RENTAL DATA

This example involves real-time semiparametric regression analysis of data from the property rental market in Sydney, Australia. Each day, hundreds of properties come on the Sydney market and these fresh data are usually advertised on rental agency websites and real estate websites as *realestate.com.au*. This offers the possibility to perform real-time analysis and produce live and up-to-date summaries of the rental market status. An attractive approach to model such data is the special case of semiparametric regression known as geoadditive models (Kammann and Wand 2003). Explicitly, we work with the model

$$\begin{aligned}
&\log((\texttt{weekly rent})_{ij}) \,|\, \boldsymbol{\beta}, U_i, \boldsymbol{u}_2, \boldsymbol{u}_3, \boldsymbol{u}_4, \boldsymbol{u}_5, \sigma_\varepsilon^2 \\
&\quad \overset{\text{ind.}}{\sim} N\big(\beta_0 + \beta_1 \,\texttt{house}_{ij} + f_2(\texttt{(number of bedrooms)}_{ij}) \\
&\qquad + f_3(\texttt{(number of bathrooms)}_{ij}) + f_4(\texttt{(number of car spaces)}_{ij}) \\
&\qquad + f_5(\texttt{longitude}_{ij}, \texttt{latitude}_{ij}) + U_i, \sigma_\varepsilon^2\big), \qquad U_1, \ldots, U_{992} \,|\, \sigma_U^2 \overset{\text{ind.}}{\sim} N\big(0, \sigma_U^2\big).
\end{aligned} \tag{22}$$

Here, $(\texttt{weekly rent})_{ij}$ is the weekly rental amount in Australian dollars of the jth property for the ith real estate agency (hereafter called the (i, j)th property), and $\texttt{house}_{ij}$ is an indicator of the (i, j)th property being a house, townhouse or villa (rather than an apartment). The variable $\texttt{(number of bedrooms)}_{ij}$ is the number of bedrooms in the (i, j)th property. Variables concerning the numbers of bathrooms and car spaces are defined similarly. The geographical location of the (i, j)th property is conveyed by the variables $\texttt{longitude}_{ij}$ and $\texttt{latitude}_{ij}$. The U_i, $1 \le i \le 992$, are random intercepts for each of the 992 agencies. The fixed effect regression coefficients β_0, β_1 and the linear contribution to $f_2, \ldots, f_5$ are stored in $\boldsymbol{\beta}$. Similarly, the spline basis coefficients for $f_2, \ldots, f_5$ are stored in $\boldsymbol{u}_2, \ldots, \boldsymbol{u}_5$. The estimate of f_5 is based on bivariate thin plate splines as explained in Chapter 13 of Ruppert, Wand, and Carroll (2003).

The website for this example displays fitting of (22) in real time based on data collected since May 9, 2012. Several regression summaries are presented. First, a geographical map

is listed with processed properties as small black dots and recently (i.e., during the last hour) added ones as yellow circles. The total number of processed properties is included at the bottom right. Next, a color-coded geographical map displays the weekly rent for a two-bedroom apartment with one bathroom and one car space for various geographical locations. The approximate posterior density function for β_1 shows the impact of the property being a house or not. Regression fits and 95% credible sets for the number of bedrooms, bathrooms, and car spaces for apartments are presented. Finally, a list of rental agencies with the least and most expensive properties, after correcting for all other covariates, is provided. All these regression summaries are computed in real time and the figures are updated every hour.

6.4 U.S. DOMESTIC FLIGHT DATA

Air traffic delays represent a critical problem for both airlines and passengers. In this section, we will demonstrate the proposed methodology for real-time analysis of U.S. domestic flights. We use the website *www.flightstats.com* to obtain real-time data on flight delay, flight distance, operating airline, and flight path. Data on temperature, wind speed, and aviation flight category are obtained through the *aviationweather.gov* website. This example is inspired by a recent competition, titled GE Flight Quest, run by the kaggle platform (*www.kaggle.com*).

The real-time data consist of flight delay, flight distance, operating airline, and flight path. In addition, data on temperature, wind speed, and aviation flight category are available. The aviation flight categories are based on the North American conventions known as METAR and are based on the ceiling (height above ground of the base of the lowest layer of cloud) and visibility. Table 1 provides the aviation flight categories definitions.

Our demonstration uses the semiparametric regression model:

$$\begin{aligned}
&\log(\texttt{delay}_{ijk} + 120) | \boldsymbol{\beta}, U_i, V_j, \boldsymbol{u}_7, \boldsymbol{u}_8, \boldsymbol{u}_9, \boldsymbol{u}_{10}, \boldsymbol{u}_{11}, \sigma_\varepsilon^2 \\
&\overset{\text{ind.}}{\sim} N\big(\beta_0 + \beta_1 \texttt{MVFRdep}_{ijk} + \beta_2 \texttt{IFRdep}_{ijk} + \beta_3 \texttt{LIFRdep}_{ijk} + \beta_4 \texttt{MVFRarr}_{ijk} \\
&\quad + \beta_5 \texttt{IFRarr}_{ijk} + \beta_6 \texttt{LIFRarr}_{ijk} + f_7((\texttt{flight distance})_j) \\
&\quad + f_8((\texttt{departure temperature})_{ijk}) + f_9((\texttt{arrival temperature})_{ijk}) \\
&\quad + f_{10}((\texttt{departure wind speed})_{ijk}) + f_{11}((\texttt{arrival wind speed})_{ijk}) \\
&\quad + U_i + V_j, \sigma_\varepsilon^2\big), \quad U_1, \ldots, U_{171} | \sigma_U^2 \overset{\text{ind.}}{\sim} N\big(0, \sigma_U^2\big), \quad V_1, \ldots, V_{2,000} | \sigma_V^2 \overset{\text{ind.}}{\sim} N\big(0, \sigma_V^2\big).
\end{aligned} \tag{23}$$

Table 1. Definitions of North American aviation flight categories

Category	Ceiling	and/or visibility
Visual flight rules	Above 3000 feet	Above 5 miles
Marginal visual flight rules	1000–3000 feet	3–5 miles
Instrument flight rules	500–1000 feet	1–3 miles
Low instrument flight rules	Below 500 feet	Below 1 mile

Here $\texttt{delay}_{ijk}$ is the difference between the actual and scheduled runway arrival time in minutes for the kth flight of airline i on flight path j and

$$\texttt{MVFRdep}_{ijk} = \begin{cases} 1 & \text{if marginal visual flight rules apply at the scheduled runway} \\ & \text{departure time of the } k\text{th flight of airline } i \text{ on flight path } j \\ 0 & \text{otherwise.} \end{cases}$$

The variable $\texttt{MVFRarr}_{ijk}$ is defined analogously, but for the scheduled runway arrival time. The other aviation flight category variables are defined similarly, with `IFR` denoting "instrument flight rules" and `LIFR` denoting "low instrument flight rules." The variable $\texttt{(flight distance)}_j$ denotes the distance of flight path j in kilometers. Variables $\texttt{(departure temperature)}_{ijk}$ and $\texttt{(arrival temperature)}_{ijk}$ are the temperature in degrees Celsius at the scheduled runway departure and arrival time of the kth flight of airline i on flight path j, respectively. Variables $\texttt{(departure wind speed)}_{ijk}$ and $\texttt{(arrival wind speed)}_{ijk}$ are the wind speed in knots at the scheduled runway departure and arrival time of the kth flight of airline i on flight path j, respectively. The $U_i, 1 \leq i \leq 171$, are random intercepts for each of the 171 airlines, while $V_j, 1 \leq j \leq 2000$, are random effects for each of the 2000 flight paths. The fixed effect regression coefficients $\beta_0, \ldots, \beta_6$ and the linear contribution to $f_7, \ldots, f_{11}$ are stored in $\boldsymbol{\beta}$. Similarly, the spline basis coefficients for $f_7, \ldots, f_{11}$ are stored in $\boldsymbol{u}_7, \ldots, \boldsymbol{u}_{11}$.

The link `U.S. domestic flight data` on our live demonstrations website displays fitting of (23) in real time based on data collected since January 25, 2013. A map shows the flight paths that have most recently been processed and the number of processed flights is given at the bottom of the map. Various regression summaries are provided. Of particular interest are tables of airlines and flight paths with the lowest and highest delays. All these regression summaries are computed in real time and the figures are updated every few minutes.

ACKNOWLEDGMENTS

This research was partially supported by Australian Research Council Discovery Project DP110100061. T. Broderick's research was supported by a U.S. National Science Foundation Graduate Research Fellowship. We are grateful to Jeff Morris and Paul Murrell for discussions related to this research.

[Received September 2012. Revised May 2013.]

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. [606]

Armagan, A., Dunson, D. B., and Lee, J. (2013), "Generalized Double Pareto Shrinkage," *Statistica Sinica*, 23, 119–143. [603]

Bishop, C. M. (2006), *Pattern Recognition and Machine Learning*, New York: Springer. [592]

Cameron, A. C., and Trivedi, P. K. (2005), *Microeconometrics: Methods and Applications*, New York: Cambridge University Press. [594]

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010), "The Horseshoe Estimator for Sparse Signals," *Biometrika*, 97, 465–480. [603]

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. [607]

Croissant, Y. (2011), Ecdat 0.1. "Datasets for Econometrics," R package, Available at *cran.r-project.org* [594]

Devroye, L., and Wagner, T. J. (1980), "On the L_1 Convergence of Kernel Estimators of Regression Functions With Application to Discrimination," *Zeitschrift für Wahrscheinlichkeitstheorie und Vervandte Gebiete*, 51, 15–25. [591]

Faes, C., Ormerod, J. T., and Wand, M. P. (2011), "Variational Bayesian Inference for Parametric and Nonparametric Regression With Missing Data," *Journal of the American Statistical Association*, 106, 959–971. [590,609]

Fricker, R. D., and Chang, J. T. (2008), "A Spatio-Temporal Methodology for Real-time Biosurveillance," *Quality Engineering*, 20, 465–477. [591]

Girolami, M., and Rogers, S. (2006), "Variational Bayesian Multinomial Probit Regression," *Neural Computation*, 18, 1790–1817. [607]

Gradshteyn, I. S., and Ryzhik, I. M. (1994), *Tables of Integrals, Series, and Products* (5th ed.), San Diego, California: Academic Press. [603]

Griffin, J. E., and Brown, P. J. (2011), "Bayesian Hyper Lassos With Non-convex Penalization," *Australian and New Zealand Journal of Statistics*, 53, 423–442. [603]

Härdle, W. (1990), *Applied Nonparametric Regression*, Cambridge: Cambridge University Press. [590]

Hoffman, M., Blei, D., and Bach, F. (2010), "Online Learning for Latent Dirichlet Allocation," in *Advances in Neural Information Processing Systems 23*, eds. J. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, Vancouver, Canada: Neural Information Processing Systems Foundation, pp. 856–864. [590,596]

Huang, A., and Wand, M. P. (2013), "Simple Marginally Noninformative Prior Distributions for Covariance Matrices," *Bayesian Analysis*, 2, Number 2, 439–452. [602]

Jaakkola, T. S., and Jordan, M. I. (2000), "Bayesian Parameter Estimation via Variational Methods," *Statistics and Computing*, 10, 25–37. [605,606]

Jank, W., and Shmueli, G. (2007), "Modelling Concurrency of Events in Online Auctions via Spatiotemporal Semiparametric Models," *Applied Statistics*, 56, 1–27. [589]

Johnstone, I. M., and Silverman, B. W. (2005), "Empirical Bayes Selection of Wavelet Thresholds," *The Annals of Statistics*, 33, 1700–1752. [603]

Kaimi, I., and Diggle, P. J. (2011), "A Hierarchical Model for Real-time Monitoring of Variation in Risk of Non-Specific Gastro-intestinal Infections," *Epidemiology and Infection*, 139, 1854–1862. [589,591]

Kammann, E. E., and Wand, M. P. (2003), "Geoadditive Models," *Journal of the Royal Statistical Society,* Series C, 52, 1–18. [611]

Krzyzak, A., and Pawlak, M. (1984), "Almost Everywhere Convergence of a Recursive Regression Function Estimate and Classification," *IEEE Transactions on Information Theory*, IT-30, 91–93. [591]

Langford, J., Li, L., and Zhang, T. (2009), "Sparse Online Learning Via Truncated Gradient," *Journal of Machine Learning Research*, 10, 777–801. [609]

Louis, T. A. (1982), "Finding the Observed Information Matrix When Using the EM Algorithm," *Journal of the Royal Statistical Society,* Series B, 44, 226–233. [609]

Luenberger, D. G., and Ye, Y. (2008), *Linear and Nonlinear Programming* (3rd ed.), New York: Springer. [593]

McCulloch, C. E., Searle, S. R., and Neuhaus, J. M. (2008), *Generalized, Linear, and Mixed Models* (2nd ed.), New York: Wiley. [609]

Menictas, M., and Wand, M. P. (2013), "Variational Inference for Marginal Longitudinal Semiparametric Regression," *Stat*, 2, 61–71. [609]

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. [591]

Neville, S. E., Ormerod, J. T., and Wand, M. P. (2013), "Mean Field Variational Bayes for Continuous Sparse Signal Shrinkage: Pitfalls and Remedies," available at *matt-wand.utsacademics.info/papers.html*. [603]

Ng, S.-K., McLachlan, G. J., and Lee, A. H. (2006), "An Incremental EM-based Learning Approach for Online Prediction of Hospital Resource Utilization," *Artificial Intelligence in Medicine*, 36, 257–267. [591]

Ormerod, J. T., and Wand, M. P. (2010), "Explaining Variational Approximations," *The American Statistician*, 64, 140–153. [592,599,605,609]

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003), *Semiparametric Regression*, New York: Cambridge University Press. [590,611]

——— (2009), "Semiparametric Regression During 2003–2007," *Electronic Journal of Statistics*, 3, 1193–1256. [589]

Ryan, J. A. (2012), "quantmod 0.3. Quantitative Financial Modelling Framework," R package, Available at *cran.r-project.org*. [611]

Smith, A. D. A. C., and Wand, M. P. (2008), "Streamlined Variance Calculations for Semiparametric Mixed Models," *Statistics in Medicine*, 27, 435–448. [601]

Tchumtchoua, S., Dunson, D. B., and Morris, J. S. (2012), "Online Variational Bayes Inference for High-dimensional Correlated Data," unpublished manuscript. Available at *www.stat.duke.edu/~dunson/submitted.html*. [590]

Wainwright, M. J., and Jordan, M. I. (2008), "Graphical Models, Exponential Families, and Variational Inference," *Foundation and Trends in Machine Learning*, 1, 1–305. [592]

Wand, M. P. (2009), "Semiparametric Regression and Graphical Models," *Australian and New Zealand Journal of Statistics*, 51, 9–41. [590]

Wand, M. P., and Jones, M. C. (1995), *Kernel Smoothing*, London: Chapman and Hall. [590]

Wand, M. P., and Ormerod, J. T. (2008), "On Semiparametric Regression With O'Sullivan Penalized Splines," *Australian and New Zealand Journal of Statistics*, 50, 179–198. [601]

——— (2011), "Penalized Wavelets: Embedding Wavelets Into Semiparametric Regression," *Electronic Journal of Statistics*, 5, 1654–1717. [589,590,603,605]

Wang, C., Paisley, J., and Blei, D. M. (2011), "Online Variational Inference for the Hierarchical Dirichlet Process," in *International Conference on Artificial Intelligence and Statistics, 2011*, Fort Lauderdale, FL, USA. [590]

Wang, B., and Titterington, D. M. (2005), "Inadequacy of Interval Estimates Corresponding to Variational Bayesian Approximations," in *Proceedings of the 10th International Workshop on Artificial Intelligence*, eds. R. G. Cowell and Z. Ghahramani, Barbados: Society for Artificial Intelligence and Statistics, pp. 373–380. [609]

Welham, S. J., Cullis, B. R., Kenward, M. G., and Thompson, R. (2007), "A Comparison of Mixed Model Splines for Curve Fitting," *Australian and New Zealand Journal of Statistics*, 49, 1–23. [601]

Wolverton, C. T., and Wagner, T. J. (1969), "Asymptotically Optimal Discriminant Functions for Pattern Recognition," *IEEE Transactions on Information Theory*, IT-15, 258–265. [591]

Wood, S. N. (2006), *Generalized Additive Models: An Introduction With R*. Boca Raton, Florida: Chapman & Hall/CRC. [601]

Yamato, H. (1971), "Sequential Estimation of a Continuous Probability Density Function and Model," *Bulletin of Mathematical Statistics*, 14, 1–12. [591]

Zhang, T. (2004), "Solving Large Scale Linear Prediction Problems Using Stochastic Gradient Descent Algorithms," in *Proceedings of the Twenty-First International Conference on Machine Learning*, ed. C. E. Brodley, Banff, Canada: The International Machine Learning Society, pp. 919–926. [609]

Zhao, Y., Staudenmayer, J., Coull, B. A., and Wand, M. P. (2006), "General Design Bayesian Generalized Linear Mixed Models," *Statistical Science*, 21, 35–51. [598,599]