Journal of Machine Learning Research 24 (2023) 1-39 Submitted 1/23; Revised 10/23; Published 10/23

Fast Expectation Propagation for Heteroscedastic, Lasso-Penalized, and Quantile Regression

Jackson Zhou jzho4727@uni.sydney.edu.au
School of Mathematics and Statistics
University of Sydney
NSW 2006, Australia

John T. Ormerod john.ormerod@sydney.edu.au
School of Mathematics and Statistics
University of Sydney
NSW 2006, Australia

Clara Grazian clara.grazian@sydney.edu.au
School of Mathematics and Statistics
University of Sydney
and ARC Training Centre in Data Analytics for Resources & Environments
NSW 2006, Australia

Editor: Chris Oates

Abstract

Expectation propagation (EP) is an approximate Bayesian inference (ABI) method which has seen widespread use across machine learning and statistics, owing to its accuracy and speed. However, it is often difficult to apply EP to models with complex likelihoods, where the EP updates do not have a tractable form and need to be calculated using methods such as multivariate numerical quadrature. These methods increase run time and reduce the appeal of EP as a fast approximate method. In this paper, we demonstrate that EP can still be made fast for certain models in this category. We focus on various types of linear regression, for which fast Bayesian inference is becoming increasingly important in the transition to big data. Fast EP updates are achieved through analytic integral reductions in certain moment computations. EP is compared to other ABI methods across simulations and benchmark datasets, and is shown to offer a good balance between accuracy and speed.

Keywords: Expectation propagation, approximate Bayesian inference, Laplace approximation, variational Bayes, linear regression

1. Introduction

The demand for scalable statistical inference is on the rise as datasets grow larger and models become increasingly complex, especially in big data fields such as bioinformatics, signal processing, and computer vision. In the Bayesian sphere, this has been seen through growing interest in approximate Bayesian inference (ABI) methods as a faster alternative to traditional Markov chain Monte Carlo (MCMC), which is exact in the limit but typically does not scale well with the size of the data (Robert et al., 2018). Such methods include the Laplace approximation, a whole family of variational approaches (Blei et al., 2017), expectation propagation (EP) (Minka, 2001), and integrated nested Laplace approximations (INLA) (Rue et al., 2009).

We focus on the EP methodology. Originally introduced in Minka (2001) as a generalized message passing algorithm on factor graphs, EP extended and unified assumed density filtering (Maybeck, 1982), and loopy belief propagation (Frey and MacKay, 1997). Since then, theoretical properties have been developed regarding asymptotic theory and convergence in special cases (Ribeiro and Opper, 2011; Dehaene and Barthelmé, 2017), and the method has been successfully used in a wide range of applications. These include Gaussian process modeling (Bui et al., 2016; Hernandez-Lobato and Hernandez-Lobato, 2016), signal processing (Wang et al., 2020), and microarray data classification (Hernández-Lobato et al., 2010). More generally, EP has seen use in likelihood-free inference (Barthelmé and Chopin, 2014), partitioned inference for big data (Vehtari et al., 2020), feature selection (Hernández-Lobato et al., 2013, 2015), and frequentist inference (Hall et al., 2020).

The success of EP as a statistical inference paradigm can be attributed to its advantages over other methods in accuracy and speed. EP is generally more accurate than competing ABI methods for the clutter problem and mixture weight estimation (Minka, 2001), generally outperforms existing methods in various applications relating to the sparse linear model estimation (Seeger et al., 2007), and produces negligible approximation error in certain binary regression models (Chopin and Ridgway, 2017), for instance. Furthermore, parallel EP schemes can be derived which take advantage of the localized nature of EP updates; this can lead to large run time savings (Hasenclever et al., 2017; Vehtari et al., 2020).

However, for models with complex likelihoods, EP is slowed down by tractability issues, somewhat blunting one of its main advantages. The standard EP algorithm calculates the moments of so-called tilted distributions, intermediate distributions formed by combining the joint likelihood with the current EP approximation. When the likelihood is simple, the tilted distributions are often one-dimensional (Rasmussen and Williams, 2005) or are multi-dimensional but have a simplifying structure (Chopin and Ridgway, 2017) where it is straightforward to compute their moments using univariate quadrature, after potentially performing dimension reduction (Chen and Wand, 2020; Vehtari et al., 2020). When the likelihood is complex, these techniques usually break down, necessitating alternate moment computation approaches such as multivariate numerical quadrature (Seeger and Jordan, 2004) Monte Carlo (Li et al., 2018), or the Laplace approximation (Smola et al., 2003); these come at the cost of increased run time (for a comparison of standard and Laplace-based implementations, see Wang et al. 2020 for example), reducing the appeal of EP as a fast approximate method.

We demonstrate that EP can still be made fast for certain models with complex likelihoods where the standard moment computation techniques (dimension reduction and univariate quadrature) are insufficient for fast EP updates. The Bayesian versions of heteroscedastic, scale-augmented lasso-penalized, and scale-augmented quantile linear regression is considered. Scale-augmented simply refers to the addition of a nuisance scale parameter into the model. An emphasis is placed on linear regression as it is the most ubiquitous modeling technique in modern statistics; fast inference in such models is therefore an important goal when dealing with the massive datasets of the current day. This can be seen, for example, in the development of sub-sampling procedures when the number of observations is large (Ma and Sun, 2015; Wang et al., 2019; Wang and Ma, 2021), along with fast inference schemes for specific linear regression variants (Clarkson et al., 2016; Fujiwara et al., 2016; Fasiolo et al., 2021). Fast EP updates are accomplished through analytic integral reductions in the moment computations of the tilted distributions; this is the main contribution of this paper.

The outline of this paper is as follows. In Section 2, EP is introduced in its classical Gaussian implementation, along with the dimension reduction technique. Section 3 uses analytic integral reductions within this framework to derive fast EP updates for the aforementioned linear regression models. In Section 4, a set of experiments is conducted to evaluate the performance of the proposed EP implementation on smaller datasets; this is compared to that of other popular ABI methods, including a more conventional EP implementation. In Section 5, a similar set of experiments is conducted using a big data example. In Section 6, the results are discussed, and future directions of study are given. Finally, derivations, additional implementation details, and supplementary figures are provided in the appendices.

2. Expectation propagation

This section provides a digestible summary of the basic EP methodology, including the dimension reduction technique. While no original content is developed here, the notation introduced will be used for the rest of the paper.

2.1 Standard EP

At a high level, EP introduces normalized approximations to the factors (commonly referred to as sites) of a particular factorization of the joint likelihood. Each site approximation is associated with its own site parameters, and the product of the site approximations forms the EP approximation of the posterior distribution. Over the course of the EP algorithm, the site parameters are iteratively refined until convergence is reached and their change between iterations is negligible. This process may be illustrated more concretely as follows. Let the data be coded as $\mathcal{D}$, and $\boldsymbol{\theta} \in \mathbb{R}^d$ be the parameter of interest. Assume that the joint likelihood admits the factorization

$$p(\boldsymbol{\theta}, \mathcal{D}) = \prod_{k=1}^{K} f_k(\boldsymbol{\theta}),$$

where K is the total number of sites and the sites $f_k(\boldsymbol{\theta})$ are the components of either the likelihood (with their dependence on the data suppressed) or the prior(s). The corresponding EP approximation has the form

$$q(\boldsymbol{\theta}) = \prod_{k=1}^{K} q_k(\boldsymbol{\theta}).$$

Often it is convenient to set the site approximations q_k to be multivariate Gaussian densities, with the resulting algorithm being called Gaussian EP; we will follow this convention throughout this paper. Other densities are possible, provided they belong to the exponential family of distributions since this family is closed under multiplication; see Seeger (2005) for more details. Our strategy in dealing with parameters with a constrained domain will be to transform these to an unconstrained domain, as in Kucukelbir et al. (2015) for example. Writing in terms of natural parameters,

$$q_k(\boldsymbol{\theta}) = (2\pi)^{-d/2}|\mathbf{Q}|^{1/2}\exp\left(-\tfrac{1}{2}\boldsymbol{\theta}^\mathsf{T}\mathbf{Q}_k\boldsymbol{\theta} + \boldsymbol{\theta}^\mathsf{T}\mathbf{r}_k\right),$$

so that each site approximation q_k has associated site parameters $\mathbf{Q}_k$ (a precision matrix) and $\mathbf{r}_k$ (a linear shift vector). Here, the EP approximation q is a multivariate Gaussian with precision $\mathbf{Q}_\bullet = \sum_k \mathbf{Q}_k$ and linear shift $\mathbf{r}_\bullet = \sum_k \mathbf{r}_k$. The EP algorithm aims to iteratively refine the $\mathbf{Q}_k$ and $\mathbf{r}_k$ until convergence. For an update to q_k, the k-th *cavity distribution* is first defined as $q_{-k}(\boldsymbol{\theta}) = \prod_{j\neq k} q_j(\boldsymbol{\theta})$; let $\boldsymbol{\Sigma}_{-k}$, $\boldsymbol{\mu}_{-k}$, $\mathbf{Q}_{-k}$, and $\mathbf{r}_{-k}$ be its covariance, mean, precision, and linear shift respectively. The k-th *tilted distribution* (sometimes called the *hybrid distribution*) is then defined as

$$h_k(\boldsymbol{\theta}) \propto f_k(\boldsymbol{\theta})q_{-k}(\boldsymbol{\theta}).$$

Let $\boldsymbol{\Sigma}_{h_k}$, $\boldsymbol{\mu}_{h_k}$, $\mathbf{Q}_{h_k}$, and $\mathbf{r}_{h_k}$ be the covariance, mean, precision, and linear shift respectively of h_k, and let $\widetilde{h}_k(\boldsymbol{\theta}) = f_k(\boldsymbol{\theta})q_{-k}(\boldsymbol{\theta})$ be its kernel. The tilted distributions are combinations of the joint likelihood and the current EP approximation, and are the mechanism by which information is propagated from the former to the latter. The site parameters $\mathbf{Q}_k$ and $\mathbf{r}_k$ are then updated such that the Kullback-Leibler divergence from the tilted distribution to the EP approximation q is minimized. When using Gaussian site approximations, this is equivalent to calculating the first two moments of the tilted distribution ($\boldsymbol{\Sigma}_{h_k}$ and $\boldsymbol{\mu}_{h_k}$) to get the moment parameters of its Gaussian approximation, transforming into natural parameters ($\mathbf{Q}_{h_k}$ and $\mathbf{r}_{h_k}$), and subtracting off the corresponding natural parameters of the cavity distribution. The EP update may be written compactly as

$$\mathbf{Q}_k \leftarrow \mathbf{Q}_{h_k} - \mathbf{Q}_{-k} \quad \text{and} \quad \mathbf{r}_k \leftarrow \mathbf{r}_{h_k} - \mathbf{r}_{-k}.$$

Naïvely, the update requires solutions to d-dimensional integrals. In particular, define

$$I_{h_k,0} = \int \widetilde{h}_k(\boldsymbol{\theta})\,d\boldsymbol{\theta}, \quad \mathbf{I}_{h_k,1} = \int \boldsymbol{\theta}\widetilde{h}_k(\boldsymbol{\theta})\,d\boldsymbol{\theta}, \quad \text{and} \quad \mathbf{I}_{h_k,2} = \int \boldsymbol{\theta}\boldsymbol{\theta}^\mathsf{T}\widetilde{h}_k(\boldsymbol{\theta})\,d\boldsymbol{\theta}$$

to be the d-dimensional integrals corresponding to the 0th, 1st, and 2nd unnormalized raw moments respectively of the tilted distribution. We have that

$$\boldsymbol{\Sigma}_{h_k} = \frac{\mathbf{I}_{h_k,2}}{I_{h_k,0}} - \left(\frac{\mathbf{I}_{h_k,1}}{I_{h_k,0}}\right)\left(\frac{\mathbf{I}_{h_k,1}}{I_{h_k,0}}\right)^\mathsf{T} \quad \text{and} \quad \boldsymbol{\mu}_{h_k} = \frac{\mathbf{I}_{h_k,1}}{I_{h_k,0}},$$

from which we can recover $\mathbf{Q}_{h_k}$ and $\mathbf{r}_{h_k}$.

The EP updates are repeated across the sites (these can be performed in parallel) until convergence, where the change in the site parameters is sufficiently small and the algorithm is terminated. If a site has a Gaussian form, then its approximation is set to be itself during initialization and is not refined. Additional details, especially those on algorithmic considerations, can be found in Vehtari et al. (2020).

2.2 Dimension reduction

Papers such as Chen and Wand (2020) and Vehtari et al. (2020) describe a general dimension reduction approach for the computation of the tilted distribution moments, which we describe here. Suppose that we would like to update q_k. It is sometimes the case that f_k can be rewritten as

$$f_k(\boldsymbol{\theta}) = f_k^*(\boldsymbol{\vartheta}_k(\boldsymbol{\theta})), \quad \text{with} \quad \boldsymbol{\vartheta}_k(\boldsymbol{\theta}) = (\vartheta_{k,1}(\boldsymbol{\theta}), \ldots, \vartheta_{k,d_k}(\boldsymbol{\theta})) \quad \text{and} \quad \vartheta_{k,j}(\boldsymbol{\theta}) = \mathbf{a}_{k,j}^\mathsf{T}\boldsymbol{\theta},$$

where $f_k^* : \mathbb{R}^{d_k} \to \mathbb{R}$, $\vartheta_{k,j}(\boldsymbol{\theta}) : \mathbb{R}^d \to \mathbb{R}$, and $\mathbf{a}_{k,j} \in \mathbb{R}^d$ being determined from the model and data. The vector $\boldsymbol{\vartheta}$ represents the low dimensional version of $\boldsymbol{\theta}$, and we say that the k-th site has d_k linear components, with coefficient vectors $\mathbf{a}_{k,j}$. When $d_k < d$, the dimension of the problem has been reduced.

For instance, consider the Bayesian logistic regression model. If we let $\boldsymbol{\theta}$ be the p-vector of regression coefficients, then the model may be specified as

$$y_i|\boldsymbol{\theta} \sim \text{Bernoulli}((1 + \exp(-\mathbf{x}_i^\mathsf{T}\boldsymbol{\theta}))^{-1})$$

for $i = 1, \ldots, n$. If we assume a multivariate Gaussian prior on $\boldsymbol{\theta}$, that is, $\boldsymbol{\theta} \sim \mathcal{N}_d(\boldsymbol{\mu}_{\boldsymbol{\theta}}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}})$, then the joint likelihood can be factored as the product of n likelihood sites and one Gaussian prior site:

$$p(\boldsymbol{\theta}, \mathcal{D}) = \phi_p(\boldsymbol{\theta}; \boldsymbol{\mu}_{\boldsymbol{\theta}}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}}) \prod_{i=1}^{n} [1 + \exp(-2(y_i - 0.5)\mathbf{x}_k^\mathsf{T}\boldsymbol{\theta})]^{-1}.$$

If we additionally let the indices $k = 1, \ldots, n$ correspond to the likelihood sites and $k = n+1$ correspond to the prior site, then for $k = 1, \ldots, n$, we have $d_k = 1$,

$$f_k^*(\vartheta) = [1 + \exp(-\vartheta)]^{-1}, \quad \text{and} \quad \mathbf{a}_{k,1} = 2(y_k - 0.5)\mathbf{x}_k,$$

which is an example of univariate dimension reduction.

This alternate form allows for moment-based integrals involving the tilted distribution to be reduced in dimension, improving the speed and tractability of the EP algorithm; the details are given here, with the derivations deferred to Appendix A. Let $\mathbf{A}_k \in \mathbb{R}^{d \times d_k}$ be the matrix where the j-th column of $\mathbf{A}_k$ is $\mathbf{a}_{k,j}$ for $j = 1, \ldots, d_k$. The k-th *low-dimensional cavity distribution* q_{-k}^* is first defined as the multivariate Gaussian distribution with covariance $\boldsymbol{\Sigma}_{-k}^* = \mathbf{A}_k^\mathsf{T}\boldsymbol{\Sigma}_{-k}\mathbf{A}_k$ and mean $\boldsymbol{\mu}_{-k}^* = \mathbf{A}_k^\mathsf{T}\boldsymbol{\mu}_{-k}$; let $\mathbf{Q}_{-k}^*$ and $\mathbf{r}_{-k}^*$ be its precision and linear shift respectively. The k-th *low-dimensional tilted distribution* is then defined as

$$h_k^*(\boldsymbol{\vartheta}) \propto f_k^*(\boldsymbol{\vartheta})\phi_{d_k}(\boldsymbol{\vartheta}; \boldsymbol{\mu}_{-k}^*, \boldsymbol{\Sigma}_{-k}^*).$$

Let $\boldsymbol{\Sigma}_{h_k}^*$, $\boldsymbol{\mu}_{h_k}^*$, $\mathbf{Q}_{h_k}^*$, and $\mathbf{r}_{h_k}^*$ be the covariance, mean, precision, and linear shift respectively of h_k^*, and let $\widetilde{h}_k^*(\boldsymbol{\vartheta}) = f_k^*(\boldsymbol{\vartheta})\phi_{d_k}(\boldsymbol{\vartheta}; \boldsymbol{\mu}_{-k}^*, \boldsymbol{\Sigma}_{-k}^*)$ be its kernel. The EP update can be shown to be written as

$$\mathbf{Q}_k \leftarrow \mathbf{A}_k \left(\mathbf{Q}_{h_k}^* - \mathbf{Q}_{-k}^*\right) \mathbf{A}_k^\mathsf{T} \quad \text{and} \quad \mathbf{r}_k \leftarrow \mathbf{A}_k \left(\mathbf{r}_{h_k}^* - \mathbf{r}_{-k}^*\right). \tag{1}$$

We see that the dimensionality of the moment-based integrals involving the tilted distribution has been reduced from d to d_k. In particular, define

$$I_{h_k,0}^* = \int \widetilde{h}_k^*(\boldsymbol{\vartheta})\, d\boldsymbol{\vartheta}, \quad \mathbf{I}_{h_k,1}^* = \int \boldsymbol{\vartheta}\widetilde{h}_k^*(\boldsymbol{\vartheta})\, d\boldsymbol{\vartheta}, \quad \text{and} \quad \mathbf{I}_{h_k,2}^* = \int \boldsymbol{\vartheta}\boldsymbol{\vartheta}^\mathsf{T}\widetilde{h}_k^*(\boldsymbol{\vartheta})\, d\boldsymbol{\vartheta} \tag{2}$$

to be the d_k-dimensional integrals corresponding to the 0th, 1st, and 2nd unnormalized raw moments respectively of the low-dimensional tilted distribution. We have that

$$\boldsymbol{\Sigma}^*_{h_k} = \frac{\mathbf{I}^*_{h_k,2}}{I^*_{h_k,0}} - \left(\frac{\mathbf{I}^*_{h_k,1}}{I^*_{h_k,0}}\right)\left(\frac{\mathbf{I}^*_{h_k,1}}{I^*_{h_k,0}}\right)^{\mathsf{T}} \quad \text{and} \quad \boldsymbol{\mu}^*_{h_k} = \frac{\mathbf{I}^*_{h_k,1}}{I^*_{h_k,0}}, \tag{3}$$

from which we can recover $\mathbf{Q}^*_{h_k}$ and $\mathbf{r}^*_{h_k}$. The updates given in (1) rely on not only $\boldsymbol{\Sigma}^*_{-k}$ being invertible, but also positive definite, so that $\mathbf{Q}^*_{-k}$ exists, and that $\boldsymbol{\mu}^*_{h_k}$ and $\boldsymbol{\Sigma}^*_{h_k}$ are able to be evaluated. (The latter implies that $\mathbf{Q}^*_{h_k}$ exists, as $\boldsymbol{\Sigma}^*_{h_k}$ is a covariance matrix.) In general, it is not true that $\boldsymbol{\Sigma}^*_{-k} = \mathbf{A}_k^{\mathsf{T}}\boldsymbol{\Sigma}_{-k}\mathbf{A}_k$ is positive definite, given that $\boldsymbol{\Sigma}_{-k}$ is positive definite. However, it is well-known that this is the case when $\mathbf{A}_k$ has full rank, which is true for the models covered in the following sections.

Algorithm 1 summarizes the Gaussian EP algorithm when the dimension reduction technique is used, assuming all sites can be reduced in dimension. The rank d_k updates in (1) are exploited to reduce memory usage, where we only store a $d_k \times d_k$ matrix and a d_k-vector at the k-th site rather than the full $d \times d$ matrix and d-vector; the $\mathbf{Q}^*_k$ and $\mathbf{r}^*_k$ can be thought of as low-dimensional versions of the $\mathbf{Q}_k$ and $\mathbf{r}_k$ respectively. In the same spirit, the $\mathbf{A}_k$ (used to recover the full parameters) are computed from $\mathcal{D}$ on demand.

A combination of power EP updates (Minka, 2004) and damped updates (e.g., Minka 2001 and Seeger et al. 2007) were also incorporated into Algorithm 1 to improve convergence stability. Power EP with $\eta \in (0, 1]$ (a power parameter) redefines the k-th cavity distribution as $q_{-k}(\boldsymbol{\theta}) = q_k(\boldsymbol{\theta})^{1-\eta}\prod_{j\neq k} q_j(\boldsymbol{\theta})$ and the corresponding tilted distribution as $h_k(\boldsymbol{\theta}) \propto f_k(\boldsymbol{\theta})^{\eta} q_{-k}(\boldsymbol{\theta})$. This has the effect of balancing out the variance of the two components of the tilted distribution and can help with convergence in high-dimensional cases (Seeger et al., 2007). The updates given in (1) now correspond to fractional sites, and need to be scaled up by dividing by η. The derivations for the power EP downdate (calculation of the cavity distribution) using the dimension reduction technique can be found in Appendix B. Damped updates with $\alpha \in (0, 1]$ as the damping factor further modify the power EP updates by multiplying the updated site parameters by α, adding on $(1 - \alpha)$ times the previous site parameters, and assigning those instead. Damping reduces the likelihood of improper site parameters (e.g., non-positive-definite covariance matrices for Gaussian EP), and has been shown to decrease approximation error for parallel implementations of EP and solve oscillation issues in the site approximations (Minka and Lafferty, 2012; Vehtari et al., 2020).

For each update, the main operations to take into account are matrix multiplications of the form $\mathbf{A}_k\mathbf{M}\mathbf{A}_k^{\mathsf{T}}$ and $\mathbf{A}_k^{\mathsf{T}}\mathbf{M}\mathbf{A}_k$ with complexity $\mathcal{O}(d^2 d_k)$, inversions of $(d_k \times d_k)$-dimensional matrices with complexity $\mathcal{O}(d_k^3)$, and an $\mathcal{O}(d_k^2)$ number of d_k-dimensional integrals with combined complexity $\mathcal{O}(d_k^2 G^{d_k})$, assuming d_k-dimensional numerical quadrature is used with G quadrature/grid points in each dimension. The overall time complexity is therefore $\mathcal{O}(M(\sum_k d^2 d_k + d_k^2 G^{d_k}))$, where M is the total number of passes through the data. On the other hand, the data $\mathcal{D}$ with complexity $\mathcal{O}(Kd)$, the global parameters $\mathbf{Q}_\bullet$ and $\mathbf{r}_\bullet$ with complexity $\mathcal{O}(d^2)$, and the site parameters $\mathbf{Q}^*_k$ and $\mathbf{r}^*_k$ with combined complexity $\mathcal{O}(\sum_k d_k^2)$ need to be stored in memory. The overall space complexity is therefore $\mathcal{O}\left(Kd + d^2 + \sum_k d_k^2\right)$. When $d_k = 1$ (a common case), the time complexity becomes $\mathcal{O}(MK(d^2 + G))$ and the space complexity becomes $\mathcal{O}(Kd + d^2)$.

Algorithm 1 Gaussian power EP with damping and dimension reduction, assuming all sites can be reduced in dimension.

```
Require: $\mathcal{D}$, $K$, $\eta$, $\alpha$, $[d_k,\, f_k^*]_{k=1}^K$
$\mathbf{Q}_k^* \leftarrow \mathbf{I}_{d_k}$   and   $\mathbf{r}_k^* \leftarrow \mathbf{0}_{d_k}$   for   $k = 1, \ldots, K$        ▷ Initialization
$\mathbf{Q}_\bullet \leftarrow \sum_k \mathbf{A}_k \mathbf{Q}_k^* \mathbf{A}_k^\top$   and   $\mathbf{r}_\bullet \leftarrow \sum_k \mathbf{A}_k \mathbf{r}_k^*$
while $[\mathbf{Q}_k^*]_{k=1}^K$ and $[\mathbf{r}_k^*]_{k=1}^K$ have not converged do
    $\mathbf{\Sigma}_\bullet \leftarrow \mathbf{Q}_\bullet^{-1}$   and   $\boldsymbol{\mu}_\bullet \leftarrow \mathbf{\Sigma}_\bullet \mathbf{r}_\bullet$
    parallel for $k = 1, \ldots, K$ do
    Phase 1 – EP downdate
        $\mathbf{\Sigma}_{\bullet,k}^* \leftarrow \mathbf{A}_k^\top \mathbf{\Sigma}_\bullet \mathbf{A}_k$   and   $\boldsymbol{\mu}_{\bullet,k}^* \leftarrow \mathbf{A}_k^\top \boldsymbol{\mu}_\bullet$
        $\mathbf{Q}_{\bullet,k}^* \leftarrow {\mathbf{\Sigma}_{\bullet,k}^*}^{-1}$   and   $\mathbf{r}_{\bullet,k}^* \leftarrow \mathbf{Q}_{\bullet,k}^* \boldsymbol{\mu}_{\bullet,k}^*$
        $\mathbf{Q}_{-k}^* \leftarrow \mathbf{Q}_{\bullet,k}^* - \eta \mathbf{Q}_k^*$   and   $\mathbf{r}_{-k}^* \leftarrow \mathbf{r}_{\bullet,k}^* - \eta \mathbf{r}_k^*$
        $\mathbf{\Sigma}_{-k}^* \leftarrow {\mathbf{Q}_{-k}^*}^{-1}$   and   $\boldsymbol{\mu}_{-k}^* \leftarrow \mathbf{\Sigma}_{-k}^* \mathbf{r}_{-k}^*$
    Phase 2 – Tilted distribution inference
        $\widetilde{h}_k^*(\boldsymbol{\vartheta}) \leftarrow [f_k^*(\boldsymbol{\vartheta})]^\eta \phi_{d_k}(\boldsymbol{\vartheta}; \boldsymbol{\mu}_{-k}^*, \mathbf{\Sigma}_{-k}^*)$
        Compute $I_{h_k,0}^*$, $\mathbf{I}_{h_k,1}^*$, and $\mathbf{I}_{h_k,2}^*$ using (2).
        Compute $\mathbf{\Sigma}_{h_k}^*$ and $\boldsymbol{\mu}_{h_k}^*$ using (3).
    Phase 3 – EP update
        $\mathbf{Q}_{h_k}^* \leftarrow {\mathbf{\Sigma}_{h_k}^*}^{-1}$   and   $\mathbf{r}_{h_k}^* \leftarrow \mathbf{Q}_{h_k}^* \boldsymbol{\mu}_{h_k}^*$
        $\widetilde{\mathbf{Q}}_k^* \leftarrow (1-\alpha)\mathbf{Q}_k^* + \frac{\alpha}{\eta}\left(\mathbf{Q}_{h_k}^* - \mathbf{Q}_{-k}^*\right)$   and   $\widetilde{\mathbf{r}}_k^* \leftarrow (1-\alpha)\mathbf{r}_k^* + \frac{\alpha}{\eta}\left(\mathbf{r}_{h_k}^* - \mathbf{r}_{-k}^*\right)$
        $\mathbf{Q}_\bullet \leftarrow \mathbf{Q}_\bullet + \mathbf{A}_k(\widetilde{\mathbf{Q}}_k^* - \mathbf{Q}_k^*)\mathbf{A}_k^\top$   and   $\mathbf{r}_\bullet \leftarrow \mathbf{r}_\bullet + \mathbf{A}_k(\widetilde{\mathbf{r}}_k^* - \mathbf{r}_k^*)$
        $\mathbf{Q}_k^* \leftarrow \widetilde{\mathbf{Q}}_k^*$   and   $\mathbf{r}_k^* \leftarrow \widetilde{\mathbf{r}}_k^*$
    end parallel for
end while
return $\mathbf{\Sigma}_\bullet$ and $\boldsymbol{\mu}_\bullet$
```

3. Analytic integral reductions

For many models, $1 < d_k < d$ such that dimension reduction can be performed but it is insufficient to achieve fast EP updates. Either multivariate numerical quadrature or Monte Carlo is required for the computation of $I^*_{h_k,0}$, $\mathbf{I}^*_{h_k,1}$, and $\mathbf{I}^*_{h_k,2}$, if the accuracy of the EP approximation is to be preserved; this comes at the cost of increased run time, as indicated by the time complexities in Section 2.2. In the following sections, we show that it is possible to perform model-specific analytic reductions inside these integrals to allow for a fast EP implementation. In particular, we consider linear regression models where $d_k = 2$ and analytic reductions result in only the evaluation of univariate integrals. This reduces the overall time complexity from $\mathcal{O}(MK(d^2+G^2))$ to $\mathcal{O}(MK(d^2+G))$, using the notation from the previous section. For convenience, throughout this section define the functions

$$\begin{aligned}&\widetilde{a}_k(\vartheta_2) = Q^*_{-k,11}, \quad \widetilde{b}_k(\vartheta_2) = 2\left[Q^*_{-k,12}(\vartheta_2 - \mu^*_{-k,2}) - Q^*_{-k,11}\mu^*_{-k,1}\right], \quad \text{and}\\ &\widetilde{c}_k(\vartheta_2) = Q^*_{-k,11}(\mu^*_{-k,1})^2 + 2Q^*_{-k,12}\mu^*_{-k,1}(\mu^*_{-k,2} - \vartheta_2) + Q^*_{-k,22}(\vartheta_2 - \mu^*_{-k,2})^2.\end{aligned}$$

3.1 Heteroscedastic linear regression

Assuming constant variance in linear regression models is occasionally questionable; see for example Leedan and Meer (2000) and Rosopa et al. (2013). Consider the standard linear regression model

$$y_i|\boldsymbol{\beta}_1, \sigma^2 \overset{\text{ind.}}{\sim} N(\mathbf{x}_{i,1}^\mathsf{T}\boldsymbol{\beta}_1, \sigma^2) \quad \text{for} \quad i = 1, \ldots, n,$$

where y_i is the response, $\boldsymbol{\beta}_1$ is a p_1-vector of coefficients, $\mathbf{x}_{i,1}$ is a p_1-vector of predictors, and σ^2 is the residual variance. Replacing σ^2 with

$$\sigma_i^2 = \exp(2\,\mathbf{x}_{i,2}^\mathsf{T}\boldsymbol{\beta}_2) \quad \text{for} \quad i = 1, \ldots, n$$

is a natural approach to incorporating heteroscedasticity where $\mathbf{x}_{i,2}$ is a second p_2-vector of predictors and $\boldsymbol{\beta}_2$ is a second p_2-vector of coefficients. The factor of two in the exponent in the expression for σ_i^2 leads to a model where the log standard deviation (SD) is a linear function of predictors, and so assumes that each predictor has a multiplicative effect on the residual SD. When $\mathbf{x}_{i,2} = 1$, the fitted $\sigma_i^2 \equiv \sigma^2$ is a constant, the model becomes homoscedastic and the Gaussian EP approximation for σ^2 takes on a log-normal form. For such a model, $d_k = 2$, but it can be shown that analytic reduction is possible at each site.

We let $\boldsymbol{\theta} = (\boldsymbol{\beta}_1^\mathsf{T}, \boldsymbol{\beta}_2^\mathsf{T})$ with $d = p_1 + p_2$, and assume a multivariate Gaussian prior on $\boldsymbol{\theta}$ such that $\boldsymbol{\theta} \sim \mathcal{N}_d(\boldsymbol{\mu}_{\boldsymbol{\theta}}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}})$. The joint likelihood can be factored as the product of n likelihood sites and one Gaussian prior site:

$$p(\boldsymbol{\theta}, \mathcal{D}) = \phi_d(\boldsymbol{\theta}; \boldsymbol{\mu}_{\boldsymbol{\theta}}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}}) \prod_{i=1}^{n} \exp\left[-\mathbf{x}_{i,2}^\mathsf{T}\boldsymbol{\beta}_2 - \frac{(y_i - \mathbf{x}_{i,1}^\mathsf{T}\boldsymbol{\beta}_1)^2}{2\exp(2\,\mathbf{x}_{i,2}^\mathsf{T}\boldsymbol{\beta}_2)}\right].$$

Suppose that the likelihood sites are associated with the indices $k = 1, \ldots, n$, with the Gaussian prior site corresponding to $k = n + 1$. For $k = 1, \ldots, n$, we see that $d_k = 2$, with

$$f_k^*(\boldsymbol{\vartheta}) = \exp\left[-\vartheta_2 - \frac{(y_k - \vartheta_1)^2}{2\exp(2\vartheta_2)}\right], \quad \mathbf{a}_{k,1} = (\mathbf{x}_{k,1}^\mathsf{T}, \mathbf{0}_{p_2}^\mathsf{T}), \quad \text{and} \quad \mathbf{a}_{k,2} = (\mathbf{0}_{p_1}^\mathsf{T}, \mathbf{x}_{k,2}^\mathsf{T}).$$

It is clear that ϑ_1 can be integrated out analytically in $I^*_{h_k,0}$, $\mathbf{I}^*_{h_k,1}$, and $\mathbf{I}^*_{h_k,2}$. In particular, we are only concerned with the evaluation of Gaussian integrals

$$\mathcal{G}_{k,r}(\vartheta_2) = \int \vartheta_1^r \exp\left[-\tfrac{1}{2}\left(a_k(\vartheta_2)\vartheta_1^2 + b_k(\vartheta_2)\vartheta_1 + c_k(\vartheta_2)\right)\right] d\vartheta_1,$$

with $r = 0, 1, 2$ and coefficients

$$a_k(\vartheta_2) = \widetilde{a}_k(\vartheta_2) + \frac{1}{\exp(2\vartheta_2)}, \quad b_k(\vartheta_2) = \widetilde{b}_k(\vartheta_2) - \frac{2y_k}{\exp(2\vartheta_2)}, \quad \text{and}$$
$$c_k(\vartheta_2) = \widetilde{c}_k(\vartheta_2) + 2\vartheta_2 + \frac{y_k^2}{\exp(2\vartheta_2)},$$

from which we can recover the original integrals via

$$I^*_{h_k,0} = C_k \int \mathcal{G}_{k,0}(\vartheta_2)\, d\vartheta_2, \quad \mathbf{I}^*_{h_k,1} = C_k \int \begin{bmatrix} \mathcal{G}_{k,1}(\vartheta_2) \\ \vartheta_2\mathcal{G}_{k,0}(\vartheta_2) \end{bmatrix} d\vartheta_2, \quad \text{and}$$
$$\mathbf{I}^*_{h_k,2} = C_k \int \begin{bmatrix} \mathcal{G}_{k,2}(\vartheta_2) & \vartheta_2\mathcal{G}_{k,1}(\vartheta_2) \\ \vartheta_2\mathcal{G}_{k,1}(\vartheta_2) & \vartheta_2^2\mathcal{G}_{k,0}(\vartheta_2) \end{bmatrix} d\vartheta_2,$$

where the constant $C_k = |2\pi\mathbf{\Sigma}^*_{-k}|^{-1/2}$ cancels when evaluating $\mathbf{\Sigma}^*_{h_k}$ and $\boldsymbol{\mu}^*_{h_k}$. These final integrals can be evaluated numerically using univariate quadrature. For numerical stability, it is recommended that the log scale is used and that a minimum lower bound of integration is set, as the coefficients blow up when $\vartheta_2 \ll 0$, leading to unstable calculations. Expressions for $\mathcal{G}_{k,r}$ can be found in Appendix C.

3.2 Lasso-penalized linear regression

A lasso penalty can be interpreted as introducing independent Laplace priors on the regression parameters, and generally leads to a fast EP implementation with $d_k = 1$ if the scale parameter and penalty coefficient are considered as hyperparameters; see Seeger et al. (2007) for details. However, a more complete Bayesian treatment might be to view these quantities as parameters themselves; this gives rise to a model with $d_k = 2$, but fast EP updates can still be derived using analytic reduction. Consider the simpler case where we only treat the scale as an additional nuisance parameter.

Let $\boldsymbol{\theta} = (\boldsymbol{\beta}^\mathsf{T}, \kappa)$ with $d = p + 1$, where $\boldsymbol{\beta}$ is a p-vector of regression coefficients and κ is the logarithm of the scale parameter. The model may be specified as

$$y_i|\boldsymbol{\beta}, \kappa \overset{\text{ind.}}{\sim} \mathcal{N}(\mathbf{x}_i^\mathsf{T}\boldsymbol{\beta}, \exp(2\kappa)), \quad \beta_j|\kappa \overset{\text{ind.}}{\sim} \text{Laplace}(0, \exp(\kappa)/\lambda), \quad \text{and} \quad \kappa \overset{\text{ind.}}{\sim} \mathcal{N}(\mu_\kappa, \sigma_\kappa^2),$$

where $i = 1, \ldots, n$, $j = 1, \ldots, p$, and λ is the standard lasso penalty parameter (assumed fixed). The joint likelihood can be factored as the product of n likelihood sites, p Laplace prior sites, and one Gaussian prior site:

$$p(\boldsymbol{\theta}, \mathcal{D}) = \phi(\kappa; \mu_\kappa, \sigma_\kappa^2) \prod_{i=1}^n \exp\left[-\kappa - \frac{(y_i - \mathbf{x}_i^\mathsf{T}\boldsymbol{\beta})^2}{2\exp(2\kappa)}\right] \prod_{j=1}^p \exp\left[-\kappa - \frac{\lambda|\beta_j|}{\exp(\kappa)}\right].$$

Suppose that the likelihood sites are associated with the indices $k = 1, \ldots, n$, the Laplace prior sites are associated with $k = n+1, \ldots, n+p$, and the Gaussian prior site corresponds to $k = n+p+1$. For $k = 1, \ldots, n$, we have that $d_k = 2$,

$$f_k^*(\boldsymbol{\vartheta}) = \exp\left[-\vartheta_2 - \frac{(y_k - \vartheta_1)^2}{2\exp(2\vartheta_2)}\right], \quad \mathbf{a}_{k,1} = (\mathbf{x}_k^\mathsf{T}, 0), \quad \text{and} \quad \mathbf{a}_{k,2} = (\mathbf{0}_p^\mathsf{T}, 1).$$

Similar to the heteroscedastic linear regression model, ϑ_1 is able to be integrated out analytically in $I_{h_k,0}^*$, $\mathbf{I}_{h_k,1}^*$, and $\mathbf{I}_{h_k,2}^*$, and the same formulas from Section 3.1 apply. For $k = n+1, \ldots, n+p$, we have $d_k = 2$,

$$f_k^*(\boldsymbol{\vartheta}) = \exp\left[-\vartheta_2 - \frac{\lambda|\vartheta_1|}{\exp(\vartheta_2)}\right], \quad \mathbf{a}_{k,1} = (\mathbf{e}_{k-n}^\mathsf{T}, 0), \quad \text{and} \quad \mathbf{a}_{k,2} = (\mathbf{0}_p^\mathsf{T}, 1).$$

Again, ϑ_1 can be integrated out analytically in $I_{h_k,0}^*$, $\mathbf{I}_{h_k,1}^*$, and $\mathbf{I}_{h_k,2}^*$. In particular, we only require solutions to the truncated Gaussian integrals

$$\begin{aligned}
\mathcal{T}_{k,r}^-(\vartheta_2) &= \int_{-\infty}^0 \vartheta_1^r \exp\left[-\tfrac{1}{2}\left(a_k^-(\vartheta_2)\vartheta_1^2 + b_k^-(\vartheta_2)\vartheta_1 + c_k^-(\vartheta_2)\right)\right] d\vartheta_1 \quad \text{and} \\
\mathcal{T}_{k,r}^+(\vartheta_2) &= \int_0^{\infty} \vartheta_1^r \exp\left[-\tfrac{1}{2}\left(a_k^+(\vartheta_2)\vartheta_1^2 + b_k^+(\vartheta_2)\vartheta_1 + c_k^+(\vartheta_2)\right)\right] d\vartheta_1,
\end{aligned}$$

with $r = 0, 1, 2$ and coefficients

$$a_k^\pm(\vartheta_2) = \widetilde{a}_k(\vartheta_2), \quad b_k^\pm(\vartheta_2) = \widetilde{b}_k(\vartheta_2) \pm \frac{2\lambda}{\exp(\vartheta_2)}, \quad \text{and} \quad c_k^\pm(\vartheta_2) = \widetilde{c}_k(\vartheta_2) + 2\vartheta_2,$$

from which we can recover the original integrals via

$$\begin{aligned}
I_{h_k,0}^* &= C_k \int \mathcal{T}_{k,0}^-(\vartheta_2) + \mathcal{T}_{k,0}^+(\vartheta_2)\, d\vartheta_2, \\
\mathbf{I}_{h_k,1}^* &= C_k \int \begin{bmatrix} \mathcal{T}_{k,1}^-(\vartheta_2) + \mathcal{T}_{k,1}^+(\vartheta_2) \\ \vartheta_2\left\{\mathcal{T}_{k,0}^-(\vartheta_2) + \mathcal{T}_{k,0}^+(\vartheta_2)\right\} \end{bmatrix} d\vartheta_2, \quad \text{and} \\
\mathbf{I}_{h_k,2}^* &= C_k \int \begin{bmatrix} \mathcal{T}_{k,2}^-(\vartheta_2) + \mathcal{T}_{k,2}^+(\vartheta_2) & \vartheta_2\left\{\mathcal{T}_{k,1}^-(\vartheta_2) + \mathcal{T}_{k,1}^+(\vartheta_2)\right\} \\ \vartheta_2\left\{\mathcal{T}_{k,1}^-(\vartheta_2) + \mathcal{T}_{k,1}^+(\vartheta_2)\right\} & \vartheta_2^2\left\{\mathcal{T}_{k,0}^-(\vartheta_2) + \mathcal{T}_{k,0}^+(\vartheta_2)\right\} \end{bmatrix} d\vartheta_2,
\end{aligned}$$

where again $C_k = |2\pi\boldsymbol{\Sigma}_{-k}^*|^{-1/2}$ cancels in the evaluation of $\boldsymbol{\Sigma}_{h_k}^*$ and $\boldsymbol{\mu}_{h_k}^*$. These final integrals can be evaluated numerically using univariate quadrature, and the same stability considerations from Section 3.1 are recommended. Expressions for $\mathcal{T}_{k,r}^\pm$ can be found in Appendix C.

It is straightforward to extend the work in this section to augmented versions of the elastic net and potentially other penalties. If λ is to be treated as a parameter, care should be taken in choosing its prior distribution so as to ensure the posterior distribution is proper.

3.3 Quantile linear regression

EP for quantile linear regression shares many of the same implementation details as EP for lasso-penalized linear regression. This is because quantile linear regression is equivalent to modeling the data as coming from an asymmetric Laplace (AL) distribution. For this section, we use the parameterization given by Yu and Moyeed (2001). Consider a Bayesian τ-quantile regression model. For such a model, $d_k = 2$, but again it can be shown that analytic reduction is possible at each site.

Let $\boldsymbol{\theta} = (\boldsymbol{\beta}^\mathsf{T}, \kappa)$ with $d = p + 1$, where $\boldsymbol{\beta}$ is a p-vector of quantile regression coefficients and κ is the logarithm of the scale parameter. The model may be specified as

$$y_i | \boldsymbol{\beta}, \kappa \overset{\text{ind.}}{\sim} \text{AL}(\rho = \tau, \mu = \mathbf{x}_i^\mathsf{T}\boldsymbol{\beta}, \sigma = \exp(\kappa)) \quad \text{and} \quad \boldsymbol{\theta} \overset{\text{ind.}}{\sim} \mathcal{N}_d(\boldsymbol{\mu}_{\boldsymbol{\theta}}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}}),$$

where $i = 1, \dots, n$. The joint likelihood can be factored as the product of n likelihood sites and one Gaussian prior site:

$$p(\boldsymbol{\theta}, \mathcal{D}) = \phi_d(\boldsymbol{\theta}; \boldsymbol{\mu}_{\boldsymbol{\theta}}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}}) \prod_{i=1}^{n} \exp\left[-\kappa - \frac{\rho_\tau(y_i - \mathbf{x}_i^\mathsf{T}\boldsymbol{\beta})}{\exp(\kappa)}\right],$$

where $\rho_\tau(x) = (|x| + (2\tau - 1)x)/2$ is the quantile loss function. The quantity $\tau \in (0, 1)$ is a fixed constant determining the quantile to be estimated. Suppose that the likelihood sites are associated with the indices $k = 1, \dots, n$, with the Gaussian prior site corresponding to $k = n + 1$. For $k = 1, \dots, n$, we see that $d_k = 2$,

$$f_k^*(\boldsymbol{\vartheta}) = \exp\left[-\vartheta_2 - \frac{\rho_\tau(y_k - \vartheta_1)}{\exp(\vartheta_2)}\right], \quad \mathbf{a}_{k,1} = (\mathbf{x}_k^\mathsf{T}, 0), \quad \text{and} \quad \mathbf{a}_{k,2} = (\mathbf{0}_p^\mathsf{T}, 1).$$

Similar to lasso-penalized linear regression, ϑ_1 is able to be integrated out analytically when calculating $I^*_{h_k,0}$, $\mathbf{I}^*_{h_k,1}$, and $\mathbf{I}^*_{h_k,2}$. The same formulas in Section 3.2 apply, using the modified truncated Gaussian integrals

$$\mathcal{S}^-_{k,r}(\vartheta_2) = \int_{-\infty}^{y_k} \vartheta_1^r \exp\left[-\tfrac{1}{2}\left(a_k^-(\vartheta_2)\vartheta_1^2 + b_k^-(\vartheta_2)\vartheta_1 + c_k^-(\vartheta_2)\right)\right] d\vartheta_1 \quad \text{and}$$
$$\mathcal{S}^+_{k,r}(\vartheta_2) = \int_{y_k}^{\infty} \vartheta_1^r \exp\left[-\tfrac{1}{2}\left(a_k^+(\vartheta_2)\vartheta_1^2 + b_k^+(\vartheta_2)\vartheta_1 + c_k^+(\vartheta_2)\right)\right] d\vartheta_1,$$

with $r = 0, 1, 2$ and new coefficients

$$a_k^\pm(\vartheta_2) = \widetilde{a}_k(\vartheta_2), \quad b_k^\pm(\vartheta_2) = \widetilde{b}_k(\vartheta_2) + \frac{1 \pm 1 - 2\tau}{\exp(\vartheta_2)}, \quad \text{and}$$
$$c_k^\pm(\vartheta_2) = \widetilde{c}_k(\vartheta_2) + 2\vartheta_2 + \frac{(2\tau - 1 \mp 1)y_k}{\exp(\vartheta_2)}.$$

Expressions for $\mathcal{S}^\pm_{k,r}$ can be found in Appendix C.

4. Experiments with smaller datasets

The speed and accuracy of the fast EP approximations proposed in Section 3 were evaluated and compared to that of other methods in the context of smaller datasets. For each combination of model and method, evaluation was performed across three simulation settings and three benchmark datasets. For each simulation setting, methods were evaluated against five simulated datasets and the average result was taken. Thirty repetitions were conducted for each combination of model, method, and either simulation setting or benchmark dataset. All experiments were executed across 10 computational cores running at 2.5 GHz each, with a combined 32 gigabytes of random access memory. The code for these experiments can be found at https://github.com/jackson-zhou-sydney/EP-multicomp.

The gold standard which all other methods were evaluated against was Markov Chain Monte Carlo (MCMC) using the No-U-Turn Sampler from the R package `rstan` (Stan Development Team, 2023). For each of 10 chains, we set 1000 warm-up iterations and 10,000 sampling iterations, for a total of 10,000 warm-up and 100,000 sampling iterations. Convergence was verified by checking that $\widehat{R} < 1.1$, as per the recommendation in Gelman et al. (1995); this was always the case.

Four common methods were implemented across all three models in the experiments. These were a long run of MCMC, a short run of MCMC, the proposed EP implementation, and an alternate EP implementation where bivariate numerical quadrature was used to evaluate the moments of the low-dimensional tilted distributions. These are coded as MCMC, MCMC-S, EP-1D, and EP-2D in the text, and as ML, MS, E1, and E2 in the figures/tables respectively. The long run of MCMC used the same settings as the gold standard (but a different seed), and was used to provide a rough upper bound on performance. Note that it is sometimes possible for other methods to outperform the gold standard, especially for sampling-based evaluation metrics which do not use the full sample. For the short run of MCMC, the number of sampling iterations per chain was minimized within the set $\{100, 200, 400, 1000, 2000, 4000, 6000, 8000, 10000\}$, such that $\widehat{R} < 1.1$. For each chain, the number of warm-up iterations was always one-tenth the number of sampling iterations. Both versions of EP were implemented in C++ via the R package `Rcpp` (Eddelbuettel and François, 2011) to facilitate fairer comparisons with other methods that also were implemented in C++, and parallel computing was handled using OpenMP (Chandra et al., 2001). We set $\eta = 0.5$ based on Seeger et al. (2007), $\alpha = 0.5$ as a balanced damping value that is unlikely to result in either improper site parameters (α too high) or slow convergence (α too low), and choose 400 as the number of quadrature points in each dimension. Refinements to the site approximations were grouped into different passes through the data, so as to ensure that all site approximations are updated an equal number of times. In each pass, all site approximations are refined exactly once, with the order of these refinements randomized. For each refinement, the norms of the differences in natural parameters between the old and updated site approximations were calculated, and maximums were taken across sites for each combination of pass and natural parameter. Convergence was determined to be satisfied after a given pass (with the algorithm terminated) if each of these "maximum absolute deltas" was less than 0.05 times the corresponding values of the first pass. The minimum and maximum number of passes was set to 6 and 200 respectively. It was always the case that EP converged.

Alternate ABI methods were also implemented for each model. For heteroscedastic linear regression, the derivatives of the joint likelihood are able to be evaluated and the Laplace (coded as Laplace in the text and as LA in the figures/tables) and Pathfinder (Zhang et al., 2022) approximations were considered; both have multivariate Gaussian forms. The Laplace approximation was implemented using the L-BFGS algorithm with a maximum iteration count of 20,000. Convergence was determined to be satisfied if the difference between the current and new function values was less than 10^{-6} times the current function value, or if the norm of the current gradient was less than 10^{-5} times the maximum of one and the norm of the current solution. It was always the case that the L-BFGS optimization converged. The Pathfinder approximation was run with three different values for the number of samples returned. These were 100 (the default number of samples returned), 1000, and 10,000, and were coded as Pathfinder-A, Pathfinder-B, and Pathfinder-C in the text, and as PA, PB, and PC in the figures/tables respectively. The remaining settings for Pathfinder, including those related to optimization, were left as their suggested default values in the paper, where it is noted that Pathfinder results are not sensitive to its settings. Note that the optimizations need not converge for Pathfinder to give sensible results, and so convergence was not checked. The derivatives of the joint likelihood are not always defined for lasso-penalized and quantile linear regression, and so a mean-field variational Bayes (coded as MFVB in the text and as MF in the figures/tables) approximation was implemented instead; the derivations can be found in Appendix D and Appendix E respectively. Convergence was determined to be satisfied after a given iteration of the while loop (with the algorithm terminated) if the norms of the differences between the current and new values of all parameters were less than 0.05 times the corresponding values of the first iteration. The minimum and maximum number of iterations was set to 6 and 200 respectively, to match that of EP. It was always the case that MFVB converged. The Laplace and MFVB approximations were implemented in C++ via the R package `Rcpp` (Eddelbuettel and François, 2011) to facilitate fair comparisons, and the Pathfinder R code was obtained from https://github.com/LuZhangstat/Pathfinder. We expect Pathfinder run times to be comparable to a full C++ implementation since the computationally intensive optimization stage is already performed in C++ via the R package `optimx` (Nash and Varadhan, 2011).

To evaluate multivariate performance, we considered three different metrics. The first was the maximum mean discrepancy (MMD) of Gretton et al. (2012). In particular, we used $M^* = -\log(\text{MMD}_u^2 + 10^{-5})$, where MMD_u^2 is the unbiased estimate of the squared MMD, given by

$$\text{MMD}_u^2 = \frac{1}{m(m-1)} \sum_{i \neq j}^{m} \left[k(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) + k(\mathbf{y}^{(i)}, \mathbf{y}^{(j)}) - k(\mathbf{x}^{(i)}, \mathbf{y}^{(j)}) - k(\mathbf{x}^{(j)}, \mathbf{y}^{(i)}) \right].$$

The quantity m is the total number of samples from each of the two distributions to be evaluated; these samples are denoted as $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(m)}$ and $\mathbf{y}^{(1)}, \dots, \mathbf{y}^{(m)}$, and are required to be independent from each other. The function k is the kernel function, and was chosen to be the radial basis function. We evaluated M^* with $m = 500$ samples from each of the gold standard and method to be evaluated. If the number of samples from the latter was less than 500, then that was used instead. When MMD_u^2 was calculated to be negative, it was set to zero. Higher values of M^* indicate better multivariate performance. The second multivariate metric was the computed log pointwise predictive density (lppd) described in Gelman et al. (1995); it is given by

$$\text{computed lppd} = \sum_{i=1}^{n} \log \left(\frac{1}{S} \sum_{s=1}^{S} p(y_i | \boldsymbol{\theta}^{(s)}) \right),$$

where n is the total number of data points to be evaluated (these are denoted as $y_1, \ldots, y_n$), S is the total number of samples from the distribution to be evaluated (these are denoted as $\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(S)}$). For convenience, we will refer to the computed lppd as just the lppd. The lppd metric was evaluated out-of-sample using a train-test split of 87.5-12.5, where $S = 500$ samples from the method to be evaluated were taken (across many repetitions this is comparable to cross-validation). If the number of samples was less than 500, then that was used instead. Higher values of the lppd indicate better multivariate performance. The third multivariate metric was the Frobenius norm of the difference in the estimated covariance matrices of the gold standard and method to be evaluated. In particular, we used $F^* = -\log(\|\boldsymbol{\Sigma}_{\text{Gold}} - \boldsymbol{\Sigma}_{\text{Method}}\|_{\text{F}} + 10^{-5})$, where higher values of F^* indicate better multivariate performance.

To evaluate marginal performance, we used the L^1 accuracy of Faes et al. (2011). The L^1 accuracy for the j-th marginal component is given by

$$L^1 \text{ accuracy} = 100 \left(1 - \frac{1}{2}\int_{-\infty}^{\infty} |p_{\text{Gold}}(\theta_j) - p_{\text{Method}}(\theta_j)| \, d\theta_j\right)\%. \tag{4}$$

In (4), kernel density estimates for the MCMC approaches were made using a solve-the-equation version of the Sheather-Jones bandwidth selection method (Sheather and Jones, 1991) coupled with a Gaussian kernel (as implemented by the R function `density`) for its general reliability and favorable asymptotics (Jones et al., 1996). Performance may slightly deteriorate for the short MCMC runs (due to smaller sample sizes) as well as for the lasso-penalized and quantile linear regression models (as the posterior distribution deviates from the Gaussian reference density). In these cases, L^1 accuracies should be supplemented with the previous multivariate performance metrics. After the density estimation, the enclosing integral was computed numerically, with the domain of integration centered at the j-th marginal mean of the gold standard, and having a radius of five times the j-th marginal SD of the gold standard. The domain of integration was uniformly split into 1024 intervals for use with the composite trapezoidal rule. The L^1 accuracy ranges from 0 to 100, with higher values indicating better marginal performance. These accuracies were averaged over appropriate blocks of the parameter vector $\boldsymbol{\theta}$ (for example, $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ for heteroscedastic regression).

Across the simulations, the numeric entries of the design matrices were sampled from independent standard Gaussian distributions. For all datasets, the response vector and numeric columns of the design matrix were centered at zero and scaled to have unit variance (potentially after sampling) to standardize the effect of each predictor.

4.1 Heteroscedastic linear regression

Simulations were based on the settings

$$(n, p_1, p_2) \in \{(200, 40, 10), (200, 20, 20), (200, 10, 40)\}.$$

In each case, we set for $l \in \{1, 2\}$

$$\boldsymbol{\beta}_l = (2/p_l, -2/p_l, \ldots, -2/p_l) \in \mathbb{R}^{p_l}.$$

	Setting 1	Setting 2	Setting 3	Setting 1	Setting 2	Setting 3
		$\boldsymbol{\beta}_1$ L^1 accuracy			$\boldsymbol{\beta}_2$ L^1 accuracy	
ML	99.4 ± 0.0	99.4 ± 0.0	99.4 ± 0.0	99.4 ± 0.0	99.4 ± 0.0	99.4 ± 0.0
MS	**99.3 ± 0.0**	**99.3 ± 0.0**	**99.3 ± 0.0**	**99.3 ± 0.0**	**99.3 ± 0.0**	**99.3 ± 0.0**
E1	99.1 ± 0.0	99.2 ± 0.0	98.8 ± 0.0	96.9 ± 0.0	98.8 ± 0.0	99.0 ± 0.0
E2	99.1 ± 0.0	99.2 ± 0.0	98.8 ± 0.0	96.9 ± 0.0	98.8 ± 0.0	99.0 ± 0.0
LA	90.3 ± 0.0	95.1 ± 0.0	97.9 ± 0.0	72.4 ± 0.0	90.1 ± 0.0	96.0 ± 0.0
PA	71.7 ± 3.1	84.5 ± 1.3	77.2 ± 3.4	69.2 ± 2.6	84.0 ± 1.8	77.7 ± 3.1
PB	75.9 ± 3.3	89.6 ± 0.7	86.6 ± 1.4	73.5 ± 2.6	89.4 ± 0.9	86.6 ± 1.2
PC	81.8 ± 1.7	93.2 ± 0.7	91.0 ± 1.1	79.6 ± 1.5	92.8 ± 0.4	91.2 ± 0.7
		M^*			lppd	
ML	9.12 ± 0.66	8.47 ± 0.60	9.25 ± 0.67	-31.01 ± 1.55	-32.71 ± 1.47	-33.25 ± 1.26
MS	**9.30 ± 0.71**	8.41 ± 0.58	**9.32 ± 0.69**	-30.99 ± 1.55	-32.72 ± 1.46	-33.24 ± 1.24
E1	7.59 ± 0.50	**8.68 ± 0.76**	9.19 ± 0.59	-31.30 ± 1.43	-32.36 ± 1.78	**-33.03 ± 1.58**
E2	7.59 ± 0.50	**8.68 ± 0.76**	9.19 ± 0.59	-31.30 ± 1.43	-32.36 ± 1.78	**-33.03 ± 1.58**
LA	2.49 ± 0.02	3.54 ± 0.03	4.06 ± 0.03	**-30.57 ± 1.85**	**-32.19 ± 2.01**	-33.09 ± 1.76
PA	1.92 ± 0.19	2.92 ± 0.19	2.25 ± 0.27	-31.57 ± 2.02	-32.97 ± 1.81	-34.09 ± 1.68
PB	2.11 ± 0.24	3.62 ± 0.14	3.09 ± 0.22	-31.03 ± 1.89	-32.77 ± 1.68	-33.68 ± 1.56
PC	2.53 ± 0.22	4.48 ± 0.20	3.94 ± 0.21	-30.86 ± 1.57	-32.73 ± 1.62	-33.60 ± 1.52
		F^*			Run time (seconds)	
ML	6.94 ± 0.02	7.19 ± 0.01	6.95 ± 0.01	64.98 ± 4.05	59.68 ± 3.82	64.57 ± 4.15
MS	**6.88 ± 0.01**	**7.13 ± 0.02**	**6.88 ± 0.01**	49.63 ± 6.51	45.36 ± 6.04	49.17 ± 6.39
E1	6.47 ± 0.01	7.00 ± 0.01	6.80 ± 0.02	0.02 ± 0.01	0.02 ± 0.00	0.02 ± 0.00
E2	6.47 ± 0.01	7.00 ± 0.01	6.80 ± 0.02	5.95 ± 0.03	5.61 ± 0.03	5.72 ± 0.03
LA	4.96 ± 0.00	5.94 ± 0.00	6.29 ± 0.01	**0.01 ± 0.00**	**0.00 ± 0.00**	**0.01 ± 0.00**
PA	3.29 ± 0.08	3.55 ± 0.07	3.16 ± 0.11	0.84 ± 0.10	0.79 ± 0.09	0.88 ± 0.12
PB	3.31 ± 0.07	3.89 ± 0.08	3.42 ± 0.08	0.92 ± 0.11	0.86 ± 0.10	0.98 ± 0.13
PC	3.40 ± 0.06	4.21 ± 0.09	3.71 ± 0.09	2.08 ± 0.22	1.73 ± 0.21	1.97 ± 0.24

Table 1: Results (mean ± SD) across heteroscedastic linear regression simulations. Each cell is based on thirty repetitions. Settings are organized by order of appearance in the main text. The best non-MCMC performances are highlighted in bold.

	Food	Salary	Sniffer	Food	Salary	Sniffer
		$\boldsymbol{\beta}_1$ L^1 accuracy			$\boldsymbol{\beta}_2$ L^1 accuracy	
ML	99.3 ± 0.1	99.4 ± 0.1	99.3 ± 0.1	99.4 ± 0.1	99.3 ± 0.1	99.3 ± 0.0
MS	**99.3 ± 0.1**	99.3 ± 0.1	97.1 ± 0.6	**99.3 ± 0.1**	**99.3 ± 0.1**	97.7 ± 0.3
E1	98.7 ± 0.1	**99.5 ± 0.0**	**99.0 ± 0.1**	98.4 ± 0.1	99.0 ± 0.1	**99.0 ± 0.1**
E2	98.7 ± 0.1	**99.5 ± 0.0**	**99.0 ± 0.1**	98.4 ± 0.1	99.0 ± 0.1	**99.0 ± 0.1**
LA	98.6 ± 0.1	99.3 ± 0.1	98.1 ± 0.1	93.8 ± 0.1	96.3 ± 0.1	94.3 ± 0.1
PA	95.4 ± 2.0	94.4 ± 1.3	89.1 ± 6.1	95.8 ± 1.6	93.7 ± 2.6	88.7 ± 5.2
PB	98.2 ± 0.5	98.2 ± 0.4	94.4 ± 2.0	98.5 ± 0.5	97.8 ± 0.8	93.7 ± 1.9
PC	99.1 ± 0.1	99.0 ± 0.2	97.1 ± 1.4	99.1 ± 0.3	98.8 ± 0.3	96.8 ± 0.9
		M^*			lppd	
ML	8.77 ± 2.40	8.52 ± 2.15	8.24 ± 2.61	-5.87 ± 0.73	-84.43 ± 6.93	-7.41 ± 1.57
MS	8.14 ± 2.10	8.58 ± 2.17	6.51 ± 1.39	-5.86 ± 0.73	**-84.44 ± 6.92**	-7.42 ± 1.56
E1	7.77 ± 2.02	**9.72 ± 2.14**	**8.28 ± 2.16**	-5.88 ± 0.72	-84.77 ± 7.85	-7.33 ± 1.01
E2	7.77 ± 2.02	9.71 ± 2.14	**8.28 ± 2.16**	-5.88 ± 0.72	-84.77 ± 7.85	-7.33 ± 1.01
LA	6.65 ± 1.49	8.04 ± 1.97	5.86 ± 0.52	**-5.84 ± 0.74**	-84.73 ± 7.96	-7.04 ± 1.06
PA	8.32 ± 3.10	6.12 ± 2.36	3.68 ± 1.19	-5.88 ± 0.73	-87.54 ± 11.03	**-7.03 ± 1.22**
PB	**9.29 ± 2.45**	8.95 ± 2.38	5.20 ± 1.03	-5.85 ± 0.72	-87.56 ± 11.12	-7.05 ± 1.20
PC	9.15 ± 1.99	9.24 ± 2.18	6.87 ± 2.06	-5.87 ± 0.72	-87.59 ± 11.23	-7.05 ± 1.22
		F^*			Run time (seconds)	
ML	7.82 ± 0.28	9.31 ± 0.20	6.25 ± 0.51	41.47 ± 1.01	43.73 ± 0.61	44.32 ± 0.85
MS	**7.85 ± 0.29**	9.27 ± 0.18	4.58 ± 0.57	32.63 ± 0.74	34.43 ± 0.73	2.74 ± 0.12
E1	6.65 ± 0.11	9.39 ± 0.16	**5.48 ± 0.31**	0.02 ± 0.04	0.05 ± 0.03	0.02 ± 0.05
E2	6.64 ± 0.11	**9.48 ± 0.18**	5.46 ± 0.31	1.11 ± 0.07	18.45 ± 0.98	5.52 ± 0.08
LA	6.40 ± 0.08	9.23 ± 0.19	4.62 ± 0.14	**0.00 ± 0.00**	**0.00 ± 0.00**	**0.00 ± 0.00**
PA	5.18 ± 0.33	6.30 ± 0.18	2.90 ± 0.34	0.69 ± 0.03	3.46 ± 1.85	0.64 ± 0.02
PB	6.26 ± 0.27	7.41 ± 0.19	3.62 ± 0.23	0.66 ± 0.03	3.66 ± 1.90	0.67 ± 0.02
PC	7.29 ± 0.31	8.07 ± 0.21	3.94 ± 0.49	1.15 ± 0.03	4.49 ± 1.96	1.03 ± 0.03

Table 2: Results (mean ± SD) across heteroscedastic linear regression benchmarks. Each cell is based on thirty repetitions. The best non-MCMC performances are highlighted in bold.

The benchmark datasets used were `Food` ($n = 40$, $p_1 = 2$, $p_2 = 2$), `Salary` ($n = 725$, $p_1 = 6$, $p_2 = 2$), and `Sniffer` ($n = 125$, $p_1 = 5$, $p_2 = 5$). `Food` (Hill et al., 2018) records the food expenditure (the response) and income of various households. `Salary` (DeMaris, 2007) contains salary data (the response) on a large number of university faculty members, along with predictors such as prior experience and years worked. Both `Food` and `Salary` are used as illustrative examples in the documentation for Stata's `hetregress` function. `Sniffer` (Weisberg, 2013) records the amount of hydrocarbon emitted from a tank when gasoline is poured in (the response), along with the temperature and pressure of the tank and gasoline as predictors. For all simulations and benchmarks, we set $\mu_{\boldsymbol{\theta}} = \mathbf{0}$ and $\mathbf{\Sigma}_{\boldsymbol{\theta}} = \text{diag}((\mathbf{1}_{p_1}, 0.01\mathbf{1}_{p_2}))$, where diag creates a diagonal matrix with its argument as the diagonal and $\mathbf{1}_n$ is the n-vector containing only ones.

The results for the simulated and benchmark datasets are shown in Table 1 and Table 2 respectively. Supplementary L^1 accuracy plots can be found in Appendix F. Across the simulated datasets, the accuracy of EP-1D was generally higher compared to that of the Laplace and various Pathfinder approximations, equal to that of EP-2D, and slightly lower compared to that of the short MCMC runs. Across the benchmark datasets, the differences in accuracy were mostly the same, with the exception that the performance of the Pathfinder approximation has slightly improved, now occasionally beating EP in some metrics. Overall, EP-1D tended to be slower than the Laplace approximation, but was much faster than all other methods, including EP-2D.

One point of interest that is not clear from the tables is that the Laplace approximation tended to consistently underestimate the intercept component of $\boldsymbol{\beta}_2$ compared to the other methods. We believe that this is due multivariate skewness in the posterior distribution towards higher values of this intercept. This underestimation may explain in part the poor performance of the Laplace approximation for this model.

4.2 Lasso-penalized linear regression

Simulations were based on the settings

$$(n, p) \in \{(200, 40), (40, 40), (10, 40)\}.$$

In each case, we set

$$\boldsymbol{\beta} = (0, \dots, 0, 2/p, -2/p, \dots, -2/p) \in \mathbb{R}^p, \quad \kappa = -1, \quad \text{and} \quad \lambda = 0.5,$$

where the first $p/2$ entries of $\boldsymbol{\beta}$ are zero. The benchmark datasets used were `Diabetes` ($n = 442$, $p = 11$), `Prostate` ($n = 97$, $p = 9$), and `Eye` ($n = 120$, $p = 201$). `Diabetes` (Efron et al., 2004) stores data on diabetes progression (the response) for a large number of patients, along with clinical predictors such as age, body mass index, and average blood pressure. `Prostate` (Hodge et al., 1989) records the logged amount of a prostate-specific antigen (the response) on a moderate number of patients, and also includes clinical predictors such as age and logged prostate weight. `Eye` (Scheetz et al., 2006) contains TRIM32 gene expressions (the response) for a moderate number of rats, with the predictors being the results of 200 gene probes. For all simulations and benchmarks, we set $\mu_\kappa = 0$ and $\sigma_\kappa^2 = 0.01$.

The results for the simulated and benchmark datasets are shown in Table 3 and Table 4 respectively. Supplementary L^1 accuracy plots can be found in Appendix F. Across all

	Setting 1	Setting 2	Setting 3	Setting 1	Setting 2	Setting 3
		$\boldsymbol{\beta}$ L^1 accuracy			κ L^1 accuracy	
ML	99.3 ± 0.0	99.3 ± 0.0	98.9 ± 0.0	99.4 ± 0.1	99.3 ± 0.1	99.3 ± 0.1
MS	96.9 ± 0.1	96.6 ± 0.2	**95.5 ± 0.1**	**96.8 ± 0.4**	**96.8 ± 0.4**	**96.6 ± 0.6**
E1	**99.4 ± 0.0**	**97.5 ± 0.0**	91.3 ± 0.0	94.9 ± 0.1	90.0 ± 0.1	87.4 ± 0.1
E2	**99.4 ± 0.0**	**97.5 ± 0.0**	91.3 ± 0.0	94.9 ± 0.1	90.0 ± 0.1	87.4 ± 0.1
MF	99.2 ± 0.0	95.1 ± 0.0	89.9 ± 0.0	95.7 ± 0.1	86.4 ± 0.1	85.6 ± 0.1
		M^*			lppd	
ML	8.25 ± 0.83	6.97 ± 0.76	7.43 ± 0.43	-34.22 ± 1.75	-10.61 ± 0.27	-3.11 ± 0.07
MS	8.42 ± 0.67	7.26 ± 0.86	**7.38 ± 0.36**	**-34.21 ± 1.75**	-10.63 ± 0.29	-3.11 ± 0.06
E1	**8.77 ± 0.71**	**7.43 ± 0.91**	5.07 ± 0.16	**-34.21 ± 1.55**	-10.72 ± 0.23	-3.08 ± 0.05
E2	**8.77 ± 0.71**	**7.43 ± 0.91**	5.07 ± 0.16	**-34.21 ± 1.55**	-10.72 ± 0.23	-3.08 ± 0.05
MF	8.93 ± 0.72	5.35 ± 0.16	3.53 ± 0.07	-34.24 ± 1.58	**-10.20 ± 0.27**	**-2.93 ± 0.05**
		F^*			Run time (seconds)	
ML	6.96 ± 0.02	3.67 ± 0.07	0.01 ± 0.02	59.75 ± 2.79	58.91 ± 2.72	58.43 ± 2.67
MS	4.98 ± 0.02	1.71 ± 0.09	-1.94 ± 0.02	3.10 ± 0.69	3.02 ± 0.69	3.02 ± 0.72
E1	**7.12 ± 0.01**	**3.38 ± 0.06**	**-1.91 ± 0.01**	0.15 ± 0.06	0.06 ± 0.02	0.03 ± 0.01
E2	**7.12 ± 0.01**	**3.38 ± 0.06**	**-1.91 ± 0.01**	6.59 ± 0.46	2.70 ± 0.16	1.33 ± 0.09
MF	6.91 ± 0.02	0.90 ± 0.01	-2.49 ± 0.00	**0.01 ± 0.01**	**0.00 ± 0.00**	**0.01 ± 0.01**

Table 3: Results (mean ± SD) across lasso-penalized linear regression simulations. Each cell is based on thirty repetitions. Settings are organized by order of appearance in the main text. The best non-MCMC performances are highlighted in bold.

	Diabetes	Prostate	Eye	Diabetes	Prostate	Eye
		$\boldsymbol{\beta}$ L^1 accuracy			κ L^1 accuracy	
ML	99.3 ± 0.0	99.3 ± 0.0	99.1 ± 0.0	99.3 ± 0.1	99.3 ± 0.1	99.3 ± 0.2
MS	96.9 ± 0.5	96.7 ± 0.4	**97.1 ± 0.1**	97.6 ± 0.8	97.0 ± 0.9	**96.9 ± 1.2**
E1	**99.0 ± 0.1**	**99.3 ± 0.0**	93.0 ± 0.0	98.5 ± 0.2	97.3 ± 0.2	61.8 ± 0.3
E2	**99.0 ± 0.1**	**99.3 ± 0.0**	93.0 ± 0.0	98.5 ± 0.2	97.3 ± 0.2	61.8 ± 0.3
MF	98.4 ± 0.1	98.8 ± 0.1	92.6 ± 0.0	**98.9 ± 0.2**	**97.6 ± 0.2**	11.5 ± 0.1
		M^*			lppd	
ML	8.41 ± 2.51	8.80 ± 2.20	8.27 ± 1.21	-58.78 ± 4.34	-12.48 ± 1.71	-30.38 ± 1.13
MS	6.79 ± 2.42	7.85 ± 1.90	**7.88 ± 0.87**	**-58.77 ± 4.32**	-12.47 ± 1.74	-30.17 ± 0.92
E1	8.88 ± 2.56	**9.17 ± 2.14**	6.54 ± 0.61	-58.78 ± 4.33	-12.47 ± 1.46	-31.70 ± 0.57
E2	**8.90 ± 2.57**	**9.17 ± 2.15**	6.54 ± 0.61	-58.78 ± 4.33	-12.47 ± 1.46	-31.70 ± 0.57
MF	8.42 ± 2.63	9.10 ± 2.15	5.99 ± 0.41	-58.78 ± 4.34	**-12.46 ± 1.47**	**-28.51 ± 0.63**
		F^*			Run time (seconds)	
ML	7.09 ± 0.57	7.39 ± 0.16	1.69 ± 0.01	49.09 ± 2.66	46.31 ± 2.31	369.17 ± 8.55
MS	4.78 ± 0.69	5.46 ± 0.17	0.09 ± 0.03	3.40 ± 0.73	2.39 ± 0.33	10.12 ± 0.35
E1	**7.19 ± 0.63**	**7.51 ± 0.09**	**0.98 ± 0.01**	0.10 ± 0.03	0.05 ± 0.01	0.67 ± 0.05
E2	7.16 ± 0.66	7.50 ± 0.09	**0.98 ± 0.01**	12.59 ± 0.85	3.13 ± 0.16	27.97 ± 1.51
MF	4.31 ± 0.07	6.31 ± 0.06	0.91 ± 0.01	**0.01 ± 0.03**	**0.00 ± 0.00**	**0.07 ± 0.00**

Table 4: Results (mean ± SD) across lasso-penalized linear regression benchmarks. Each cell is based on thirty repetitions. The best non-MCMC performances are highlighted in bold.

	Setting 1	Setting 2	Setting 3	Setting 1	Setting 2	Setting 3
		$\boldsymbol{\beta}$ L^1 accuracy			κ L^1 accuracy	
ML	99.3 ± 0.0	99.3 ± 0.0	99.3 ± 0.0	99.3 ± 0.1	99.4 ± 0.0	99.3 ± 0.1
MS	96.6 ± 0.1	96.6 ± 0.1	96.6 ± 0.1	96.5 ± 0.5	**96.8 ± 0.5**	**96.8 ± 0.6**
E1	**99.0 ± 0.0**	**99.0 ± 0.0**	**99.0 ± 0.0**	**96.7 ± 0.1**	96.6 ± 0.1	**96.8 ± 0.1**
E2	**99.0 ± 0.0**	**99.0 ± 0.0**	**99.0 ± 0.0**	**96.7 ± 0.1**	96.6 ± 0.1	**96.8 ± 0.1**
MF	84.1 ± 0.0	82.4 ± 0.0	83.1 ± 0.0	81.6 ± 0.1	81.5 ± 0.1	81.4 ± 0.1
		M^*			lppd	
ML	7.35 ± 0.41	7.06 ± 0.39	7.14 ± 0.47	-29.37 ± 0.60	-39.84 ± 1.01	-36.07 ± 0.58
MS	6.66 ± 0.25	6.73 ± 0.29	6.73 ± 0.25	-29.36 ± 0.60	-39.83 ± 0.98	-36.06 ± 0.59
E1	**7.98 ± 0.51**	**7.80 ± 0.51**	**7.96 ± 0.67**	-29.36 ± 0.52	-39.67 ± 0.80	-35.98 ± 0.79
E2	7.97 ± 0.51	**7.80 ± 0.51**	**7.96 ± 0.67**	-29.36 ± 0.52	-39.67 ± 0.80	-35.98 ± 0.79
MF	3.20 ± 0.03	3.02 ± 0.04	3.09 ± 0.04	**-29.00 ± 0.63**	**-39.56 ± 0.90**	**-35.82 ± 0.89**
		F^*			Run time (seconds)	
ML	6.90 ± 0.02	6.16 ± 0.02	6.43 ± 0.02	61.76 ± 2.80	61.19 ± 2.93	60.91 ± 2.86
MS	4.91 ± 0.02	4.18 ± 0.02	4.44 ± 0.02	2.95 ± 0.21	2.92 ± 0.22	2.92 ± 0.20
E1	**6.58 ± 0.02**	**5.95 ± 0.02**	**6.16 ± 0.02**	0.26 ± 0.09	0.16 ± 0.05	0.20 ± 0.05
E2	**6.58 ± 0.02**	**5.95 ± 0.02**	**6.16 ± 0.02**	11.46 ± 1.58	7.20 ± 0.90	9.26 ± 1.20
MF	4.26 ± 0.00	3.41 ± 0.00	3.72 ± 0.00	**0.02 ± 0.01**	**0.02 ± 0.00**	**0.02 ± 0.00**

Table 5: Results (mean ± SD) across quantile linear regression simulations. Each cell is based on thirty repetitions. Settings are organized by order of appearance in the main text. The best non-MCMC performances are highlighted in bold.

datasets, the accuracy of EP-1D was somewhat higher compared to that of the MFVB approximation, equal to that of EP-2D, and similar to that of the short MCMC runs. In general, EP performed similarly compared to MFVB when $n > p$, but performed much better than MFVB when $n < p$. For the $\boldsymbol{\beta}$ L^1 accuracy and F^* metrics, EP always had an advantage over MFVB. The MFVB approximation was seen to have better predictive performance (as measured by the lppd value) compared to other methods; this can potentially be explained by underestimation of the posterior variance (a known issue with mean-field methods), concentrating samples in regions of high density. Overall, EP-1D tended to be slower than the MFVB approximation, but was much faster than all other methods, including EP-2D.

4.3 Quantile linear regression

All three simulations were based on $n = 200$ and $p = 40$, with

$$\boldsymbol{\beta} = (2/p, -2/p, \dots, -2/p) \in \mathbb{R}^p \quad \text{and} \quad \tau = 0.5.$$

The responses were sampled as

$$y_k \sim \mathcal{N}(\mathbf{x}_k^\mathsf{T}\boldsymbol{\beta}, 0.2^2), \quad y_k \sim \text{Poisson}(\exp(\mathbf{x}_k^\mathsf{T}\boldsymbol{\beta})), \quad \text{and} \quad y_k \sim \text{Binomial}(10, \Phi(\mathbf{x}_k^\mathsf{T}\boldsymbol{\beta}))$$

for the first, second, and third sets of simulations respectively. The benchmark datasets used were `IgG` ($n = 298$, $p = 2$), `Engel` ($n = 235$, $p = 2$), and `Stack` ($n = 21$, $p = 4$). `IgG` (Isaacs et al., 1983) measures the serum concentration of immunoglobulin G (the response) for a large number of children, along with their age as a predictor. `Engel` (Koenker and Bassett, 1982) is very similar to `Food`, and contains observations on food expenditure (the response)

	IgG	Engel	Stack	IgG	Engel	Stack
		$\boldsymbol{\beta}$ L^1 accuracy			κ L^1 accuracy	
ML	99.3 ± 0.1	99.3 ± 0.1	99.3 ± 0.1	99.3 ± 0.1	99.3 ± 0.1	99.3 ± 0.2
MS	96.7 ± 0.8	96.2 ± 0.9	96.3 ± 0.5	96.8 ± 1.0	96.1 ± 1.3	96.8 ± 1.0
E1	**99.1 ± 0.1**	**97.8 ± 0.1**	**97.6 ± 0.1**	**99.2 ± 0.1**	**99.2 ± 0.2**	**99.1 ± 0.2**
E2	99.0 ± 0.1	**97.8 ± 0.1**	**97.6 ± 0.1**	**99.2 ± 0.1**	**99.2 ± 0.2**	**99.1 ± 0.2**
MF	65.8 ± 0.1	71.9 ± 0.1	90.5 ± 0.1	87.1 ± 0.2	89.4 ± 0.2	97.4 ± 0.2
		M^*			lppd	
ML	9.08 ± 2.40	8.45 ± 2.44	8.23 ± 2.45	-51.08 ± 2.69	-19.42 ± 2.59	-3.00 ± 0.15
MS	9.03 ± 2.32	7.86 ± 2.63	6.99 ± 1.62	-51.10 ± 2.69	**-19.42 ± 2.56**	-3.00 ± 0.16
E1	**9.08 ± 2.34**	9.75 ± 2.34	**8.39 ± 2.22**	**-51.08 ± 2.66**	**-19.42 ± 2.35**	-3.03 ± 0.12
E2	9.07 ± 2.32	**9.76 ± 2.34**	**8.39 ± 2.22**	**-51.08 ± 2.66**	**-19.42 ± 2.35**	-3.03 ± 0.12
MF	2.38 ± 0.10	3.17 ± 0.17	4.69 ± 0.37	-51.12 ± 2.66	**-19.42 ± 2.38**	**-2.97 ± 0.12**
		F^*			Run time (seconds)	
ML	9.32 ± 0.32	9.56 ± 0.36	5.69 ± 0.34	43.15 ± 2.92	42.01 ± 2.51	42.90 ± 3.01
MS	7.53 ± 0.45	7.74 ± 0.55	3.80 ± 0.28	2.79 ± 0.55	2.33 ± 0.11	2.34 ± 0.13
E1	**9.46 ± 0.22**	**9.22 ± 0.22**	**4.64 ± 0.14**	0.32 ± 0.11	0.32 ± 0.09	0.03 ± 0.01
E2	9.39 ± 0.22	9.17 ± 0.21	**4.64 ± 0.14**	13.05 ± 2.18	13.38 ± 1.32	0.79 ± 0.09
MF	5.30 ± 0.01	5.97 ± 0.01	2.41 ± 0.02	**0.01 ± 0.00**	**0.00 ± 0.00**	**0.00 ± 0.00**

Table 6: Results (mean ± SD) across quantile linear regression benchmarks. Each cell is based on thirty repetitions. The best non-MCMC performances are highlighted in bold.

and income for various Belgian working class households. Stack (Brownlee, 1965) records stack loss data (the response) at assorted time points during the operation of a factory, along with the predictors of airflow, water temperature, and acid concentration. For all simulations and benchmarks, we set $\mu_{\boldsymbol{\theta}} = \mathbf{0}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\theta}} = \text{diag}((\mathbf{1}_p, 0.01))$.

The results for the simulated and benchmark datasets are shown in Table 5 and Table 6 respectively. Supplementary L^1 accuracy plots can be found in Appendix F. Across all datasets, the accuracy of EP-1D was generally higher compared to the MFVB approximation and short MCMC runs, and equal to that of EP-2D. Similar to lasso-penalized case, the MFVB approximation was occasionally seen to have better predictive performance compared to the other methods, which was possibly caused by underestimation of the posterior variance. As with the previous section, EP-1D tended to be slower than the MFVB approximation, but was much faster than all other methods, including EP-2D.

5. Experiments with big data

As an approximate Bayesian inference technique, the main appeal of EP lies in fast yet relatively accurate statistical inference for massive datasets, where savings in time are magnified. To evaluate the performance of EP under this important regime for the chosen linear regression models, the experiments in Section 4 were repeated for a big data example. In this paper, we use the term big data to refer to datasets where exact methods of inference such as MCMC require an inconvenient amount of computation time (say, more than one day). As with Section 4, the code for these experiments can be found at https://github.com/jackson-zhou-sydney/EP-multicomp. The settings and evaluation metrics are carried over from Section 4, with a few changes. Firstly, for the gold standard and long run of MCMC, we now set 2000 warm-up iterations and 20,000 sampling iterations per chain in order to ensure convergence, for a total of 20,000 warm-up and 200,000 sampling iterations. Secondly, the set of iterations considered for the short run of MCMC now includes $\{12000, 14000, 16000, 18000, 20000\}$. Finally, the total number of repetitions was reduced to eight and the same training and test sets were used across repetitions in the computation of the lppd (with all models fitted to this training set) to limit computation time. For all three models, the benchmark dataset used was `Energy` (Candanedo 2017; $n = 19735$), which records appliance energy usage (the response) across households, given temperature and humidity predictors. The fixed training and test sets contained $n = 17,268$ and $n = 2467$ observations respectively. For each model, the number of variables was adjusted such that the run time for the gold standard was in the order of days. For heteroscedastic linear regression, the pairwise interactions of all internal variables was considered for both the mean and SD components ($p_1 = 172$, $p_2 = 172$). For lasso-penalized linear regression, the pairwise interactions of all variables was considered ($p = 301$). Finally, for quantile linear regression, the pairwise interactions of all internal variables was considered ($p = 172$). The results are shown in Table 7. As with Section 4, supplementary L^1 accuracy plots can be found in Appendix F.

For heteroscedastic linear regression, EP-1D generally surpassed all other methods (apart from EP-2D) across all performance metrics. In particular, the Pathfinder approximations broke down for the energy dataset, where in some cases only a small number of unique samples was returned. This was most likely caused by the the posterior distribution being markedly non-Gaussian, as is noted in Zhang et al. (2022). It is interesting that the accuracy did not increase with more samples; we expect that this is caused by the erratic behavior of the Pathfinder algorithm when approximating such distributions. EP-1D was also much faster than all other methods, including the Laplace approximation (which was not able to be parallelized).

For lasso-penalized linear regression, EP-1D performed slightly worse than the short MCMC runs, equal to that of EP-2D, and similar to that of the MFVB approximation. In particular, EP performed better than MFVB for the L^1 accuracy of the $\boldsymbol{\beta}$ block and F^*, but worse for the L^1 accuracy of the κ block, M^*, and the lppd. This agrees with the results of the experiments using the smaller datasets, where EP performed similarly to MFVB in the low-dimensional cases, but better than MFVB in the high-dimensional cases. We expect to see better results for EP if we increase the value of p, but due to the computational demands of MCMC for fitting such a dataset, this was not implemented. In terms of time, EP-1D was faster than all other methods apart from MFVB.

Finally, for quantile linear regression, EP-1D generally surpassed all other methods (apart from EP-2D) across all performance metrics, while generally taking much less time. Note that the MFVB L^1 accuracies presented here roughly agree with the results shown in Wand et al. (2011).

To further emphasize the speed of our proposed EP implementation compared to one using bivariate quadrature, an additional small experiment was conducted comparing the run times of both EP implementations under a more conservative scenario. In particular, the same datasets and settings from the main big data experiment were used, apart from the following EP settings: the minimum number of passes, which was increased from 6 to 30; the convergence threshold, which was decreased from 0.05 to 0.01; and the number of quadrature points in each dimension, which was increased from 400 to 800. In practice, a

	Hetero.	Lasso	Quantile	Hetero.	Lasso	Quantile
	$\boldsymbol{\beta}_1$ L^1 acc.	$\boldsymbol{\beta}$ L^1 acc.	$\boldsymbol{\beta}$ L^1 acc.	$\boldsymbol{\beta}_2$ L^1 acc.	κ L^1 acc.	κ L^1 acc.
ML	98.7 ± 0.1	97.9 ± 0.1	96.6 ± 0.2	98.9 ± 0.0	98.9 ± 0.4	97.9 ± 0.9
MS	97.1 ± 0.1	**96.9 ± 0.3**	96.3 ± 0.2	97.6 ± 0.1	98.0 ± 0.8	**97.5 ± 0.9**
E1	**99.1 ± 0.0**	93.4 ± 0.1	**97.5 ± 0.1**	**99.2 ± 0.0**	77.7 ± 0.7	82.5 ± 0.8
E2	**99.1 ± 0.0**	93.4 ± 0.1	**97.5 ± 0.1**	**99.2 ± 0.0**	77.7 ± 0.7	82.4 ± 0.9
LA	81.9 ± 0.1	—	—	97.4 ± 0.0	—	—
PA	25.5 ± 26.2	—	—	26.2 ± 25.5	—	—
PB	0.7 ± 0.6	—	—	3.1 ± 0.7	—	—
PC	9.8 ± 18.0	—	—	6.8 ± 12.0	—	—
MF	—	91.1 ± 0.1	78.6 ± 0.1	—	**98.9 ± 0.4**	81.9 ± 1.0
		M^*			lppd	
ML	2.86 ± 0.09	2.44 ± 0.09	2.28 ± 0.12	-2830.3 ± 5.1	-3208.4 ± 1.0	-1971.0 ± 0.7
MS	2.83 ± 0.12	2.43 ± 0.08	2.24 ± 0.05	-2831.7 ± 5.6	-3209.9 ± 2.0	-1970.7 ± 0.8
E1	**3.48 ± 0.14**	2.46 ± 0.17	**2.98 ± 0.11**	**-2829.1 ± 3.6**	-3209.6 ± 0.5	-1970.5 ± 0.1
E2	**3.48 ± 0.14**	2.46 ± 0.17	**2.98 ± 0.11**	**-2829.1 ± 3.6**	-3209.6 ± 0.5	-1970.5 ± 0.1
LA	1.82 ± 0.05	—	—	-2843.2 ± 3.7	—	—
PA	-0.29 ± 0.02	—	—	-2932.2 ± 4.6	—	—
PB	0.01 ± 0.02	—	—	-2932.8 ± 5.0	—	—
PC	0.09 ± 0.10	—	—	-2933.0 ± 5.8	—	—
MF	—	**2.56 ± 0.11**	2.19 ± 0.11	—	**-3209.4 ± 0.5**	**-1970.2 ± 0.1**
		F^*			Run time (seconds)	
ML	0.22 ± 0.02	-2.39 ± 0.02	-0.22 ± 0.01	$(3.3 \pm 0.2) \times 10^5$	$(2.5 \pm 0.5) \times 10^5$	$(2.5 \pm 0.5) \times 10^5$
MS	-0.63 ± 0.04	-2.82 ± 0.07	-0.28 ± 0.07	$(3.4 \pm 0.3) \times 10^4$	$(6.5 \pm 1.1) \times 10^4$	$(1.9 \pm 0.5) \times 10^5$
E1	**0.58 ± 0.03**	**-2.71 ± 0.03**	**0.14 ± 0.02**	**11.1 ± 0.3**	7.2 ± 0.6	**35.2 ± 5.1**
E2	**0.58 ± 0.03**	**-2.71 ± 0.03**	**0.14 ± 0.02**	630.3 ± 15.6	390.4 ± 23.6	1234.5 ± 52.1
LA	**0.58 ± 0.03**	—	—	99.9 ± 73.0	—	—
PA	-1.68 ± 0.00	—	—	1438.6 ± 121.5	—	—
PB	-1.68 ± 0.00	—	—	1478.8 ± 107.5	—	—
PC	-2.18 ± 0.92	—	—	1450.5 ± 102.0	—	—
MF	—	-3.16 ± 0.02	-0.65 ± 0.01	—	**1.1 ± 0.1**	750.0 ± 22.1

Table 7: Results (mean ± SD) for the energy dataset. Each cell is based on eight repetitions. The best non-MCMC performances are highlighted in bold.

	Hetero.	Lasso	Quantile
E1	28.7 ± 0.3	33.3 ± 0.5	78.1 ± 0.8
E2	6994.1 ± 327.2	6923.2 ± 65.6	6861.8 ± 100.8

Table 8: EP times in seconds (mean ± SD) under conservative settings for the energy dataset. Each cell is based on eight repetitions.

researcher might use these more conservative parameters if they did not want to compromise on accuracy for an important regression task. The timing results are shown in Table 8. We see that in this case, EP-2D takes hours (a potentially impractical amount of time) to run, while EP-1D only takes around a minute.

Recall that for the models we considered where $d_k = 2$, using analytic integral reductions reduces the overall time complexity of Gaussian EP with the dimension reduction technique from $\mathcal{O}(MK(d^2+G^2))$ to $\mathcal{O}(MK(d^2+G))$, where M is the total number of passes through the data, K is the total number of sites, d is the dimension of the parameter, and G is the number of quadrature points in each dimension. Therefore, the reduction in run time for EP when using analytic integral reductions is magnified for datasets with more observations (sites), or for ill-conditioned datasets/site approximation initializations, such that more passes through the data are required before EP converges. In general, we expect that this approach is much more scalable compared to EP using bivariate quadrature.

6. Closing discussion

In this paper, we showed that fast Gaussian EP updates are possible for models with seemingly complex likelihoods, by combining the multivariate version of the standard dimension reduction technique with analytic integral reductions. Experiments were conducted to compare the performance of such an EP implementation to that of standard methods, including a version of EP which used bivariate quadrature in the evaluation of the tilted distribution moments. The models of interest were the Bayesian variants of heteroscedastic, lasso-penalized, and quantile linear regression. For all three models, EP generally performed as well or better than competing ABI methods, for a moderate increase in run time. One notable exception was that MFVB seemed to have better predictive performance than EP, as measured by the lppd metric, in the lasso-penalized and quantile linear regression models. However, this was most likely caused by underestimation of the posterior variance in MFVB, as indicated by the poor performance of MFVB relative to EP in the other inference-based accuracy metrics. Thus, EP should still be preferred over MFVB for these models if accurate parameter inference is more important than good predictions for the proposed application. Even in terms of predictive power, we expect MFVB to perform worse than EP for extreme data points/outliers, as a result of its underestimation of the posterior variance. On the other hand, EP tended to perform similarly well (sometimes worse and sometimes better) compared to the short MCMC runs, while taking less time. Additionally, the proposed EP implementation was always faster than its counterpart using bivariate quadrature, which was often much slower than the alternate approaches. Overall, EP with analytic integral reductions offered a good balance of accuracy and speed, and is much more appealing than using bivariate quadrature instead.

In practice, the majority of the run time for EP was taken up by univariate numerical quadrature; this is costly and the process needs to be repeated an $\mathcal{O}(d_k^2)$ amount of times for an update to the k-th site. Significant computational savings may be achieved by employing cheaper quadrature schemes; we found in our numerical study that in some cases, EP stably converged even when using the trapezoidal rule with only 50 intervals. We also noticed that the scale of the data was a potential issue for EP in practice. For example, the coefficients a_k, b_k, and c_k from Section 3.1 can get very large when y_k is large (as was the case for the `Salary` dataset), leading to numerical instability in the evaluation of the required Gaussian integrals. This was easily solved, however, by limiting the domains of integration sensibly and/or normalizing the data.

As EP approximates each site individually and is modular in nature, the work presented in this paper offers high potential for generalizability. In particular, it is simple to mix and match the derivations in Section 3 to allow for multi-component Gaussian EP to be applied to more complicated models. For example, heteroscedastic linear regression may be combined with a lasso penalty. Here, it may be more appropriate to work with a message passing framework to modularize the calculations for the updates (Kim and Wand, 2018).

Other potential avenues of future work include: investigating models where more substantial analytic integral reductions can be made; exploiting sparsity in the site parameters from structured design matrices; and combining with approximate tilted distribution inference schemes to further speed up the algorithm. A more comprehensive numerical study may also be conducted, comparing the proposed implementation of EP with faster (but less accurate) EP variants such as that found in Heess et al. (2013).

Acknowledgments

The authors would like to thank the editor and reviewers for their comments leading to an improvement of the paper. The authors acknowledge the Sydney Informatics Hub and the University of Sydney's high performance computing cluster Artemis for providing the high performance computing resources that have contributed to the research results reported within this paper. Finally, the following sources of funding are gratefully acknowledged: the Australian Government Research Training Program Scholarship to Jackson Zhou and the Australian Research Council Discovery Project Grant (DP210100521) to John T. Ormerod.

Appendix A. Dimension reduction derivations

We prove the dimension reduction result for Gaussian EP. To start, write down the moment generating function of the tilted distribution kernel $\widetilde{h}_k$ as

$$\begin{aligned}
M_{\widetilde{h}_k}(\mathbf{t}) &= \int \exp(\mathbf{t}^\top \boldsymbol{\theta}) f_k^*(\boldsymbol{\vartheta}_k(\boldsymbol{\theta}))\phi_d(\boldsymbol{\theta}; \boldsymbol{\mu}_{-k}, \boldsymbol{\Sigma}_{-k})\, d\boldsymbol{\theta} \\
&= \exp\left(\mathbf{t}^\top \boldsymbol{\mu}_{-k} + \tfrac{1}{2}\mathbf{t}^\top \boldsymbol{\Sigma}_{-k}\mathbf{t}\right) \int f_k^*(\boldsymbol{\vartheta})\phi_{d_k}\left(\boldsymbol{\vartheta}; \mathbf{A}_k^\top\left(\boldsymbol{\mu}_{-k} + \boldsymbol{\Sigma}_{-k}\mathbf{t}\right), \mathbf{A}_k^\top \boldsymbol{\Sigma}_{-k}\mathbf{A}_k\right) d\boldsymbol{\vartheta},
\end{aligned}$$

where we have used the Gaussian linear subspace property. If we define

$$\begin{aligned} g_k(\mathbf{t}) &= \exp\left(\mathbf{t}^\top \boldsymbol{\mu}_{-k} + \tfrac{1}{2}\mathbf{t}^\top \boldsymbol{\Sigma}_{-k}\mathbf{t}\right), \\ n_k(\mathbf{t}) &= \int f_k^*(\boldsymbol{\vartheta})\phi_{d_k}\left(\boldsymbol{\vartheta}; \mathbf{A}_k^\top\left(\boldsymbol{\mu}_{-k} + \boldsymbol{\Sigma}_{-k}\mathbf{t}\right), \mathbf{A}_k^\top \boldsymbol{\Sigma}_{-k}\mathbf{A}_k\right) d\boldsymbol{\vartheta}, \end{aligned}$$

and let $I_{h_k,0} \in \mathbb{R}$, $\mathbf{I}_{h_k,1} \in \mathbb{R}^d$, and $\mathbf{I}_{h_k,2} \in \mathbb{R}^{d\times d}$ be the 0th, 1st, and 2nd unnormalized raw moments of h_k, then using the product rule for the gradient and Hessian, we have that

$$\begin{aligned} I_{h_k,0} &= g_k(\mathbf{0}_d)n_k(\mathbf{0}_d), \\ \mathbf{I}_{h_k,1} &= \left[g_k(\mathbf{0}_d)\nabla n_k(\mathbf{0}_d) + \nabla g_k(\mathbf{0}_d) n_k(\mathbf{0}_d)\right], \quad \text{and} \\ \mathbf{I}_{h_k,2} &= \left[\nabla^2 g_k(\mathbf{0}_d)n_k(\mathbf{0}_d) + \nabla g_k(\mathbf{0}_d)\nabla n_k(\mathbf{0}_d)^\top + \nabla n_k(\mathbf{0}_d)\nabla g_k(\mathbf{0}_d)^\top + g_k(\mathbf{0}_d)\nabla^2 n_k(\mathbf{0}_d)\right]. \end{aligned}$$

Routine calculations lead to $g_k(\mathbf{0}_d) = 1$, $\nabla g_k(\mathbf{0}_d) = \boldsymbol{\mu}_{-k}$, and $\nabla^2 g_k(\mathbf{0}_d) = \boldsymbol{\mu}_{-k}\boldsymbol{\mu}_{-k}^\top + \boldsymbol{\Sigma}_{-k}$, while differentiation under the integral sign yields

$$\begin{aligned} n_k(\mathbf{0}_d) &= I_{h_k,0}^*, \\ \nabla n_k(\mathbf{0}_d) &= \mathbf{U}_k\left(\mathbf{I}_{h_k,1}^* - \boldsymbol{\mu}_{-k}^* I_{h_k,0}^*\right), \quad \text{and} \\ \nabla^2 n_k(\mathbf{0}_d) &= \mathbf{U}_k\left(\mathbf{I}_{h_k,2}^* + \boldsymbol{\mu}_{-k}^*{\boldsymbol{\mu}_{-k}^*}^\top I_{h_k,0}^* - \mathbf{I}_{h_k,1}^*{\boldsymbol{\mu}_{-k}^*}^\top - \boldsymbol{\mu}_{-k}^*{\mathbf{I}_{h_k,1}^*}^\top - I_{h_k,0}^*\boldsymbol{\Sigma}_{-k}^*\right)\mathbf{U}_k^\top, \end{aligned}$$

with $\mathbf{U}_k = \boldsymbol{\Sigma}_{-k}\mathbf{A}_k\mathbf{Q}_{-k}^* \in \mathbb{R}^{d\times d_k}$. Let $\boldsymbol{\mu}_{h_k}$ and $\boldsymbol{\Sigma}_{h_k}$ be the mean and covariance respectively of h_k. Combining everything, we see after some algebra that

$$\begin{aligned} \boldsymbol{\Sigma}_{h_k} &= \frac{\mathbf{I}_{h_k,2}}{I_{h_k,0}} - \left(\frac{\mathbf{I}_{h_k,1}}{I_{h_k,0}}\right)\left(\frac{\mathbf{I}_{h_k,1}}{I_{h_k,0}}\right)^\top = \boldsymbol{\Sigma}_{-k} + \mathbf{U}_k\left(\boldsymbol{\Sigma}_{h_k}^* - \boldsymbol{\Sigma}_{-k}^*\right)\mathbf{U}_k^\top \quad \text{and} \\ \boldsymbol{\mu}_{h_k} &= \frac{\mathbf{I}_{h_k,1}}{I_{h_k,0}} = \boldsymbol{\mu}_{-k} + \mathbf{U}_k\left(\boldsymbol{\mu}_{h_k}^* - \boldsymbol{\mu}_{-k}^*\right). \end{aligned}$$

For standard EP, the update to $\mathbf{Q}_k$ then has the form

$$\begin{aligned} \mathbf{Q}_k \leftarrow \mathbf{Q}_{h_k} - \mathbf{Q}_{-k} &= -\mathbf{A}_k\mathbf{Q}_{-k}^*\left(\left(\boldsymbol{\Sigma}_{h_k}^* - \boldsymbol{\Sigma}_{-k}^*\right)^{-1} + \mathbf{Q}_{-k}^*\right)^{-1}\mathbf{Q}_{-k}^*\mathbf{A}_k^\top \\ &= \mathbf{A}_k\left(\mathbf{Q}_{h_k}^* - \mathbf{Q}_{-k}^*\right)\mathbf{A}_k^\top, \end{aligned}$$

as required, where we have applied the Woodbury matrix identity twice. Similarly, after some algebra, the update to $\mathbf{r}_k$ can be written as

$$\begin{aligned} \mathbf{r}_k \leftarrow \mathbf{r}_{h_k} - \mathbf{r}_{-k} &= \left[\mathbf{Q}_{-k} + \mathbf{A}_k\left(\mathbf{Q}_{h_k}^* - \mathbf{Q}_{-k}^*\right)\mathbf{A}_k^\top\right]\left[\boldsymbol{\mu}_{-k} + \mathbf{U}_k\left(\boldsymbol{\mu}_{h_k}^* - \boldsymbol{\mu}_{-k}^*\right)\right] - \mathbf{r}_{-k} \\ &= \mathbf{A}_k\left(\mathbf{r}_{h_k}^* - \mathbf{r}_{-k}^*\right), \end{aligned}$$

as required. Refer to the main body for the modified updates corresponding to damped power EP.

Appendix B. EP downdate derivations

We justify the EP downdate equations from Algorithm 1. For power EP with η as the power, using the definitions from the main text we have that

$$\begin{aligned}
\boldsymbol{\Sigma}^*_{-k} &= \mathbf{A}_k^\top \left(\mathbf{Q}_\bullet - \eta \mathbf{A}_k \mathbf{Q}_k^* \mathbf{A}_k^\top\right)^{-1} \mathbf{A}_k \\
&= (\mathbf{A}_k^\top \mathbf{Q}_\bullet^{-1} \mathbf{A}_k) + \eta (\mathbf{A}_k^\top \mathbf{Q}_\bullet^{-1} \mathbf{A}_k) \left[\mathbf{Q}_k^{*-1} - \eta (\mathbf{A}_k^\top \mathbf{Q}_\bullet^{-1} \mathbf{A}_k)\right]^{-1} (\mathbf{A}_k^\top \mathbf{Q}_\bullet^{-1} \mathbf{A}_k) \\
&= \left[(\mathbf{A}_k^\top \mathbf{Q}_\bullet^{-1} \mathbf{A}_k)^{-1} - \eta \mathbf{Q}_k^*\right]^{-1},
\end{aligned}$$

as required, where we have applied the Woodbury matrix identity twice. We also see that

$$\begin{aligned}
\boldsymbol{\mu}^*_{-k} &= \mathbf{A}_k^\top \left(\mathbf{Q}_\bullet - \eta \mathbf{A}_k \mathbf{Q}_k^* \mathbf{A}_k^\top\right)^{-1} (\mathbf{r}_\bullet - \eta \mathbf{A}_k \mathbf{r}_k^*) \\
&= \mathbf{A}_k^\top \left(\mathbf{Q}_\bullet - \eta \mathbf{A}_k \mathbf{Q}_k^* \mathbf{A}_k^\top\right)^{-1} \mathbf{r}_\bullet - \left[(\mathbf{A}_k^\top \mathbf{Q}_\bullet^{-1} \mathbf{A}_k)^{-1} - \eta \mathbf{Q}_k^*\right]^{-1} \eta \mathbf{r}_k^* \\
&= \left[\mathbf{I}_{d_k} + \eta (\mathbf{A}_k^\top \mathbf{Q}_\bullet^{-1} \mathbf{A}_k) \left\{\mathbf{Q}_k^{*-1} - \eta (\mathbf{A}_k^\top \mathbf{Q}_\bullet^{-1} \mathbf{A}_k)\right\}\right] \mathbf{A}_k^\top (\mathbf{Q}_\bullet^{-1} \mathbf{r}_\bullet) \\
&\quad - \left[(\mathbf{A}_k^\top \mathbf{Q}_\bullet^{-1} \mathbf{A}_k)^{-1} - \eta \mathbf{Q}_k^*\right]^{-1} \eta \mathbf{r}_k^* \\
&= \left[(\mathbf{A}_k^\top \mathbf{Q}_\bullet^{-1} \mathbf{A}_k)^{-1} - \eta \mathbf{Q}_k^*\right]^{-1} \left[(\mathbf{A}_k^\top \mathbf{Q}_\bullet^{-1} \mathbf{A}_k)^{-1} \mathbf{A}_k^\top (\mathbf{Q}_\bullet^{-1} \mathbf{r}_\bullet) - \eta \mathbf{r}_k^*\right],
\end{aligned}$$

as required, where we again have used the Woodbury matrix identity, along with the result from the previous set of equations. It may be possible to simplify these identities further (and only work in the low dimensional space) if $\mathbf{Q}_\bullet = \sum_k \mathbf{A}_k^\top \mathbf{Q}_k^* \mathbf{A}_k$ and $\mathbf{r}_\bullet = \sum_k \mathbf{A}_k^\top \mathbf{r}_k^*$, but this is generally infeasible as the prior sites need to be accounted for.

Appendix C. Gaussian integral results

We provide expressions corresponding to the Gaussian integrals $\mathcal{G}_{k,r}$, $\mathcal{T}^\pm_{k,r}$ and $\mathcal{S}^\pm_{k,r}$ mentioned in the main section. For conciseness, the dependence on ϑ_2 for the integral and coefficient functions is suppressed throughout this section, in addition to the dependence on ± and k for the coefficient functions. It can be shown that

$$\mathcal{G}_{k,0} = \sqrt{\frac{2\pi}{a}} \exp\left[-\frac{1}{2}\left(c - \frac{b^2}{4a}\right)\right], \quad \mathcal{G}_{k,1} = \mathcal{G}_{k,0}\left(-\frac{b}{2a}\right), \quad \text{and} \quad \mathcal{G}_{k,2} = \mathcal{G}_{k,0}\left(\frac{1}{a} + \frac{b^2}{4a^2}\right).$$

For the truncated Gaussian integrals, start by defining $\widetilde{y}_k = \sqrt{a} y_k + b/(2\sqrt{a})$. We have

$$\begin{aligned}
\mathcal{S}^-_{k,0} &= \mathcal{G}_{k,0} \Phi(\widetilde{y}_k), \quad \mathcal{S}^+_{k,0} = \mathcal{G}_{k,0} - \mathcal{S}^-_{k,0}, \\
\mathcal{S}^-_{k,1} &= \frac{\mathcal{G}_{k,0}}{\sqrt{a}} \left[-\frac{b}{2\sqrt{a}} \Phi(\widetilde{y}_k) - \phi(\widetilde{y}_k)\right], \quad \mathcal{S}^+_{k,1} = \mathcal{G}_{k,1} - \mathcal{S}^-_{k,1}, \\
\mathcal{S}^-_{k,2} &= \frac{\mathcal{G}_{k,0}}{a} \left[\left(\frac{b^2}{4a} + 1\right) \Phi(\widetilde{y}_k) - \left(\widetilde{y}_k - \frac{b}{\sqrt{a}}\right) \phi(\widetilde{y}_k)\right], \quad \text{and} \quad \mathcal{S}^+_{k,2} = \mathcal{G}_{k,2} - \mathcal{S}^-_{k,2}.
\end{aligned}$$

Note that $\mathcal{T}^\pm_{k,r}$ can be recovered by setting $y_k = 0$ in the previous set of equations.

Appendix D. MFVB for lasso-penalized linear regression

We implement a mean-field variational Bayes algorithm for the lasso-penalized linear regression model from Section 3.2. Introduce auxiliary variables a_j for $j = 1, \dots, p$; the new parameter vector is $\boldsymbol{\theta} = (\boldsymbol{\beta}^\mathsf{T}, \mathbf{a}^\mathsf{T}, \kappa)$ and the model can be rewritten as

$$y_i|\boldsymbol{\beta}, \kappa \overset{\text{ind.}}{\sim} \mathcal{N}(\mathbf{x}_i^\mathsf{T}\boldsymbol{\beta}, \exp(2\kappa)), \quad \beta_j|a_j, \kappa \overset{\text{ind.}}{\sim} \mathcal{N}\left(0, \frac{\exp(2\kappa)}{a_j\lambda^2}\right),$$
$$a_j \overset{\text{ind.}}{\sim} \text{Inverse-Gamma}\left(1, \tfrac{1}{2}\right), \quad \text{and} \quad \kappa \overset{\text{ind.}}{\sim} \mathcal{N}(\mu_\kappa, \sigma_\kappa^2).$$

This alternate representation of the (asymmetric) Laplace distribution was originally introduced in the context of mean-field approximations by Wand et al. (2011). Here, the log joint likelihood is

$$f(\boldsymbol{\theta}) = -\frac{1}{2\exp(2\kappa)}\Big(\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda^2\boldsymbol{\beta}^\mathsf{T}\text{diag}(\mathbf{a})\boldsymbol{\beta}\Big) - \frac{3}{2}\sum_{j=1}^p \log(a_j) - \frac{1}{2}\sum_{j=1}^p \frac{1}{a_j}$$
$$- (n+p)\kappa - \frac{(\kappa - \mu_\kappa)^2}{2\sigma_\kappa^2} + \text{constants in } \boldsymbol{\theta}.$$

Impose the product restriction $q(\boldsymbol{\theta}) = q(\boldsymbol{\beta})q(\mathbf{a})q(\kappa)$. For the regression parameters, the optimal form of the density is given by

$$\begin{aligned} q(\boldsymbol{\beta}) &\propto \exp\left[\mathbb{E}_{-q(\boldsymbol{\beta})}\left\{f(\boldsymbol{\theta})\right\}\right] \\ &\propto \exp\Big[\mathbb{E}_{-q(\boldsymbol{\beta})}\Big\{-\tfrac{1}{2}\boldsymbol{\beta}^\mathsf{T}\Big(\exp(-2\kappa)\Big(\mathbf{X}^\mathsf{T}\mathbf{X} + \lambda^2\text{diag}(\mathbf{a})\Big)\Big)\boldsymbol{\beta} + \mathbf{y}^\mathsf{T}\mathbf{X}\boldsymbol{\beta}\Big\}\Big] \\ &= \exp\Big[-\tfrac{1}{2}\boldsymbol{\beta}^\mathsf{T}\Big\{\mathbb{E}_q\left(\exp(-2\kappa)\right)\Big(\mathbf{X}^\mathsf{T}\mathbf{X} + \lambda^2\text{diag}(\mathbb{E}_q(\mathbf{a}))\Big)\Big\}\boldsymbol{\beta} + \mathbf{y}^\mathsf{T}\mathbf{X}\boldsymbol{\beta}\Big], \end{aligned}$$

which after completing the square is Gaussian with mean and covariance parameters

$$\widetilde{\boldsymbol{\mu}} = \widetilde{\boldsymbol{\Sigma}}\mathbf{X}^\mathsf{T}\mathbf{y} \quad \text{and} \quad \widetilde{\boldsymbol{\Sigma}} = \left[\mathbb{E}_q\left(\exp(-2\kappa)\right)\right]^{-1}\Big(\mathbf{X}^\mathsf{T}\mathbf{X} + \lambda^2\text{diag}(\mathbb{E}_q(\mathbf{a}))\Big)^{-1}.$$

For the auxiliary variables, the optimal density is given by

$$\begin{aligned} q(\mathbf{a}) &\propto \exp\left[\mathbb{E}_{-q(\mathbf{a})}\left\{f(\boldsymbol{\theta})\right\}\right] \\ &\propto \exp\left[\mathbb{E}_{-q(\mathbf{a})}\left\{-\frac{\lambda^2\,\boldsymbol{\beta}^\mathsf{T}\text{diag}(\mathbf{a})\boldsymbol{\beta}}{2\exp(2\kappa)} - \frac{3}{2}\sum_{j=1}^p \log(a_j) - \frac{1}{2}\sum_{j=1}^p \frac{1}{a_j}\right\}\right]. \end{aligned}$$

We see that the $q(a_j)$'s are independent, with density

$$q(a_j) \propto \exp\left[-\frac{\lambda^2}{2}\mathbb{E}_q\left(\frac{\beta_j^2}{\exp(2\kappa)}\right)a_j - \frac{3}{2}\log(a_j) - \frac{1}{2a_j}\right],$$

which is inverse Gaussian with mean $\widetilde{d}_j$ and shape $\widetilde{\lambda}_j$ given by

$$\widetilde{d}_j = \frac{1}{\lambda}\left[\mathbb{E}_q\left(\frac{\beta_j^2}{\exp(2\kappa)}\right)\right]^{-1/2} \quad \text{and} \quad \widetilde{\lambda}_j = 1.$$

Finally, the optimal density for the scale parameter is

$$
\begin{aligned}
q(\kappa) &\propto \exp\left[\mathbb{E}_{-q(\kappa)}\left\{f(\boldsymbol{\theta})\right\}\right] \\
&\propto \exp\left[\mathbb{E}_{-q(\kappa)}\left\{-\frac{\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|^2+\lambda^2\boldsymbol{\beta}^\mathsf{T}\text{diag}(\mathbf{a})\boldsymbol{\beta}}{2\exp(2\kappa)}-(n+p)\kappa-\frac{(\kappa-\mu_\kappa)^2}{2\sigma_\kappa^2}\right\}\right] \\
&= \exp\left[-\frac{\mathbb{E}_q\left\{\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|^2+\lambda^2\boldsymbol{\beta}^\mathsf{T}\text{diag}(\mathbf{a})\boldsymbol{\beta}\right\}}{2\exp(2\kappa)}-(n+p)\kappa-\frac{(\kappa-\mu_\kappa)^2}{2\sigma_\kappa^2}\right],
\end{aligned}
$$

which is not a standard distribution for κ. If we define the integral ratio

$$
\mathcal{F}(p,q,r,s)=\frac{\displaystyle\int_{-\infty}^{\infty}\exp\left[-p\exp(-2x)-(q-2)x-\tfrac{(x-r)^2}{2s}\right]dx}{\displaystyle\int_{-\infty}^{\infty}\exp\left[-p\exp(-2x)-qx-\tfrac{(x-r)^2}{2s}\right]dx},
$$

then the MFVB procedure for lasso-penalized linear regression can be written as in Algorithm 2. We evaluate the integrals appearing in $\mathcal{F}$ numerically.

Algorithm 2 MFVB for lasso-penalized linear regression.

```
Require: $\widetilde{\boldsymbol{\mu}}$, $\widetilde{\boldsymbol{\Sigma}}$, $\widetilde{\mathbf{d}}$, and $\mathbb{E}_q\left[\exp(-2\kappa)\right]$
while change in parameters do is non-negligible
    $\mathbf{Q} \leftarrow \mathbf{X}^\mathsf{T}\mathbf{X}+\lambda^2\,\text{diag}(\widetilde{\mathbf{d}})$
    $\widetilde{\boldsymbol{\Sigma}} \leftarrow \left[\mathbb{E}_q\left(\exp(-2\kappa)\right)\right]^{-1}\mathbf{Q}^{-1}$
    $\widetilde{\boldsymbol{\mu}} \leftarrow \widetilde{\boldsymbol{\Sigma}}\mathbf{X}^\mathsf{T}\mathbf{y}$
    $\mathbb{E}_q\left[\exp(-2\kappa)\right] \leftarrow \mathcal{F}\left(\frac{1}{2}\|\mathbf{y}-\mathbf{X}\widetilde{\boldsymbol{\mu}}\|^2+\frac{\lambda^2}{2}\widetilde{\boldsymbol{\mu}}^\mathsf{T}\,\text{diag}(\widetilde{\mathbf{d}})\widetilde{\boldsymbol{\mu}}+\frac{1}{2}\text{tr}(\mathbf{Q}\widetilde{\boldsymbol{\Sigma}}), n+p,\, \mu_\kappa,\, \sigma_\kappa^2\right)$
    $\widetilde{\mathbf{d}} \leftarrow \frac{1}{\lambda}\left[\mathbb{E}_q\left(\exp(-2\kappa)\right)\right]^{-1/2}\left(\widetilde{\boldsymbol{\mu}}^{\odot 2}+\text{dg}(\widetilde{\boldsymbol{\Sigma}})\right)^{\odot -1/2}$
end while
```

Appendix E. MFVB for quantile linear regression

We implement a mean-field variational Bayes algorithm for the quantile linear regression (asymmetric Laplace) model from Section 3.3. For convenience, consider a slightly simplified version of the model presented, where the dependence between $\boldsymbol{\beta}$ and κ is removed from the prior distribution. This can be written as

$$
y_i|\boldsymbol{\beta},\kappa \overset{\text{ind.}}{\sim} \text{AL}(\rho=\tau,\mu=\mathbf{x}_i^\mathsf{T}\boldsymbol{\beta},\sigma=\exp(\kappa)),\quad \boldsymbol{\beta}\overset{\text{ind.}}{\sim}\mathcal{N}_p(\boldsymbol{\mu}_{\boldsymbol{\beta}},\boldsymbol{\Sigma}_{\boldsymbol{\beta}}),\quad\text{and}\quad \kappa\overset{\text{ind.}}{\sim}\mathcal{N}(\mu_\kappa,\sigma_\kappa^2).
$$

Using the same representation by Wand et al. (2011) mentioned in Appendix D, introduce auxiliary variables a_i for $i=1,\ldots,n$. The new parameter vector is $\boldsymbol{\theta}=(\boldsymbol{\beta}^\mathsf{T},\mathbf{a}^\mathsf{T},\kappa)$ and the model can be rewritten as

$$
\begin{gathered}
y_i|\boldsymbol{\beta},a_i,\kappa \overset{\text{ind.}}{\sim}\mathcal{N}\left(\mathbf{x}_i^\mathsf{T}\boldsymbol{\beta}+\frac{(\frac{1}{2}-\tau)\exp(\kappa)}{\tau(1-\tau)a_i},\frac{\exp(2\kappa)}{\tau(1-\tau)a_i}\right),\\
a_i\overset{\text{ind.}}{\sim}\text{Inverse-Gamma}\left(1,\,\tfrac{1}{2}\right),\quad \boldsymbol{\beta}\overset{\text{ind.}}{\sim}\mathcal{N}_p(\boldsymbol{\mu}_{\boldsymbol{\beta}},\boldsymbol{\Sigma}_{\boldsymbol{\beta}}),\quad\text{and}\quad \kappa\overset{\text{ind.}}{\sim}\mathcal{N}(\mu_\kappa,\sigma_\kappa^2).
\end{gathered}
$$

The log joint likelihood may be expressed as

$$\begin{aligned} f(\boldsymbol{\theta}) \quad &= -\frac{\tau(1-\tau)}{2}\left[(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^{\mathsf{T}}\frac{\text{diag}(\mathbf{a})}{\exp(2\kappa)}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) - \frac{(1-2\tau)\mathbf{1}_n^{\mathsf{T}}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})}{\tau(1-\tau)\exp(\kappa)}\right. \\ &\qquad \left. + \left(\frac{1/2-\tau}{\tau(1-\tau)}\right)^2 \sum_{i=1}^{n}\frac{1}{a_i}\right] - \frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\mu}_{\boldsymbol{\beta}})^{\mathsf{T}}\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}(\boldsymbol{\beta}-\boldsymbol{\mu}_{\boldsymbol{\beta}}) \\ &\qquad -\frac{3}{2}\sum_{i=1}^{n}\log(a_i) - \frac{1}{2}\sum_{i=1}^{n}\frac{1}{a_i} - n\kappa - \frac{(\kappa-\mu_\kappa)^2}{2\sigma_\kappa^2} + \text{constants in } \boldsymbol{\theta}. \end{aligned}$$

Similar to Appendix D, impose the product restriction $q(\boldsymbol{\theta}) = q(\boldsymbol{\beta})q(\mathbf{a})q(\kappa)$. For the regression parameters, the optimal form of the density is given by

$$\begin{aligned} q(\boldsymbol{\beta}) &\propto \exp\left[\mathbb{E}_{-q(\boldsymbol{\beta})}\{f(\boldsymbol{\theta})\}\right] \\ &\propto \exp\left[\mathbb{E}_{-q(\boldsymbol{\beta})}\left\{-\frac{1}{2}\left(\boldsymbol{\beta}^{\mathsf{T}}\left(\tau(1-\tau)\mathbf{X}^{\mathsf{T}}\frac{\text{diag}(\mathbf{a})}{\exp(2\kappa)}\mathbf{X}+\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}\right)\boldsymbol{\beta}\right.\right.\right. \\ &\quad \left.\left.\left. -2\left(\tau(1-\tau)\mathbf{y}^{\mathsf{T}}\frac{\text{diag}(\mathbf{a})}{\exp(2\kappa)}\mathbf{X} - \frac{(\frac{1}{2}-\tau)\mathbf{1}_n^{\mathsf{T}}\mathbf{X}}{\exp(\kappa)} + \boldsymbol{\mu}_{\boldsymbol{\beta}}^{\mathsf{T}}\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}\right)\boldsymbol{\beta}\right)\right\}\right] \\ &= \exp\left[-\frac{1}{2}\left\{\boldsymbol{\beta}^{\mathsf{T}}\left(\tau(1-\tau)\mathbf{X}^{\mathsf{T}}\mathbb{E}_q\left(\frac{\text{diag}(\mathbf{a})}{\exp(2\kappa)}\right)\mathbf{X}+\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}\right)\boldsymbol{\beta}\right.\right. \\ &\quad \left.\left. -2\left(\tau(1-\tau)\mathbf{y}^{\mathsf{T}}\mathbb{E}_q\left(\frac{\text{diag}(\mathbf{a})}{\exp(2\kappa)}\right)\mathbf{X} - \left(\frac{1}{2}-\tau\right)\mathbb{E}_q\left(\frac{1}{\exp(\kappa)}\right)\mathbf{1}_n^{\mathsf{T}}\mathbf{X} + \boldsymbol{\mu}_{\boldsymbol{\beta}}^{\mathsf{T}}\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}\right)\boldsymbol{\beta}\right\}\right], \end{aligned}$$

which is Gaussian with parameters

$$\begin{aligned} \widetilde{\boldsymbol{\mu}} &= \widetilde{\boldsymbol{\Sigma}}\left[\tau(1-\tau)\mathbf{X}^{\mathsf{T}}\mathbb{E}_q\left(\frac{\text{diag}(\mathbf{a})}{\exp(2\kappa)}\right)\mathbf{y} - \left(\frac{1}{2}-\tau\right)\mathbb{E}_q\left(\frac{1}{\exp(\kappa)}\right)\mathbf{X}^{\mathsf{T}}\mathbf{1}_n + \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}\boldsymbol{\mu}_{\boldsymbol{\beta}}\right], \text{ and} \\ \widetilde{\boldsymbol{\Sigma}} &= \left[\tau(1-\tau)\mathbf{X}^{\mathsf{T}}\mathbb{E}_q\left(\frac{\text{diag}(\mathbf{a})}{\exp(2\kappa)}\right)\mathbf{X}+\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}\right]^{-1}. \end{aligned}$$

For the auxiliary variables, the optimal density is given by

$$\begin{aligned} q(\mathbf{a}) &\propto \exp\left[\mathbb{E}_{-q(\mathbf{a})}\{f(\boldsymbol{\theta})\}\right] \\ &\propto \exp\left[\mathbb{E}_{-q(\mathbf{a})}\left\{-\frac{\tau(1-\tau)}{2}\left((\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^{\mathsf{T}}\frac{\text{diag}(\mathbf{a})}{\exp(2\kappa)}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) + \left(\frac{1/2-\tau}{\tau(1-\tau)}\right)^2\sum_{i=1}^{n}\frac{1}{a_i}\right)\right.\right. \\ &\quad \left.\left. -\frac{3}{2}\sum_{i=1}^{n}\log(a_i) - \frac{1}{2}\sum_{i=1}^{n}\frac{1}{a_i}\right\}\right]. \end{aligned}$$

We see that the $q(a_i)$'s are independent, with density

$$q(a_i) \propto \exp\left[-\frac{\tau(1-\tau)}{2}\mathbb{E}_q\left(\frac{(y_k-\mathbf{x}_k^{\mathsf{T}}\boldsymbol{\beta})^2}{\exp(2\kappa)}\right)a_i - \frac{3}{2}\log(a_i) - \left(\frac{(1/2-\tau)^2}{2\tau(1-\tau)}+\frac{1}{2}\right)\frac{1}{a_i}\right],$$

which is inverse Gaussian with mean $\widetilde{d}_k$ and shape $\widetilde{\lambda}_k$ given by

$$\widetilde{d}_k = \frac{1}{2\tau(1-\tau)}\left[\mathbb{E}_q\left(\frac{(y_k-\mathbf{x}_k^{\mathsf{T}}\boldsymbol{\beta})^2}{\exp(2\kappa)}\right)\right]^{-1/2} \quad \text{and} \quad \widetilde{\lambda}_k = \frac{1}{4\tau(1-\tau)}.$$

Finally, the optimal density for the scale parameter is

$$
\begin{aligned}
q(\kappa) &\propto \exp\left[\mathbb{E}_{-q(\kappa)}\left\{f(\boldsymbol{\theta})\right\}\right] \\
&\propto \exp\Bigg[\mathbb{E}_{-q(\kappa)}\Bigg\{-\frac{\tau(1-\tau)}{2}\Bigg((\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^\top\frac{\text{diag}(\mathbf{a})}{\exp(2\kappa)}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) \\
&\qquad -\frac{(1-2\tau)\mathbf{1}_n^\top(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})}{\tau(1-\tau)\exp(\kappa)}\Bigg) - n\kappa - \frac{(\kappa-\mu_\kappa)^2}{2\sigma_\kappa^2}\Bigg\}\Bigg] \\
&= \exp\Bigg[-\frac{\tau(1-\tau)}{2}\Bigg\{\frac{\mathbb{E}_q\left((\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^\top\text{diag}(\mathbf{a})(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\right)}{\exp(2\kappa)} \\
&\qquad -\frac{(1-2\tau)\mathbb{E}_q\left(\mathbf{1}_n^\top(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\right)}{\tau(1-\tau)\exp(\kappa)}\Bigg\} - n\kappa - \frac{(\kappa-\mu_\kappa)^2}{2\sigma_\kappa^2}\Bigg],
\end{aligned}
$$

which is not a standard distribution. If we define the integral ratio

$$
\mathcal{H}(p,q,r_1,r_2,s,t) = \frac{\displaystyle\int_{-\infty}^{\infty}\exp\left[-p\exp(-2x)+q\exp(-x)-r_1x+\frac{(x-s)^2}{2t}\right]dx}{\displaystyle\int_{-\infty}^{\infty}\exp\left[-p\exp(-2x)+q\exp(-x)-r_2x+\frac{(x-s)^2}{2t}\right]dx},
$$

then the MFVB procedure for quantile linear regression can be written as in Algorithm 3. We evaluate the integrals appearing in $\mathcal{H}$ numerically.

Algorithm 3 MFVB for quantile linear regression.

```
Require: $\widetilde{\boldsymbol{\mu}}$, $\widetilde{\boldsymbol{\Sigma}}$, $\widetilde{\mathbf{d}}$, $\mathbb{E}_q[\exp(-\kappa)]$, $\mathbb{E}_q[\exp(-2\kappa)]$
while change in parameters is non-negligible do
    $\widetilde{\boldsymbol{\Sigma}} \leftarrow \left[\tau(1-\tau)\mathbf{X}^\top\text{diag}(\widetilde{\mathbf{d}})\mathbb{E}_q(\exp(-2\kappa))\mathbf{X} + \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}\right]^{-1}$
    $\widetilde{\boldsymbol{\mu}} \leftarrow \widetilde{\boldsymbol{\Sigma}}\Big[\tau(1-\tau)\mathbf{X}^\top\text{diag}(\widetilde{\mathbf{d}})\mathbb{E}_q(\exp(-2\kappa))\mathbf{y}$
            $-(\tfrac{1}{2}-\tau)\mathbb{E}_q(\exp(-\kappa))\mathbf{X}^\top\mathbf{1}_n + \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}\boldsymbol{\mu}_{\boldsymbol{\beta}}\Big]$
    $\mathbf{y}^* \leftarrow (\mathbf{y}-\mathbf{X}\widetilde{\boldsymbol{\mu}})^{\odot 2} + \text{dg}(\mathbf{X}\widetilde{\boldsymbol{\Sigma}}\mathbf{X}^\top)$
    $\mathbb{E}_q[\exp(-\kappa)] \leftarrow \mathcal{H}\left(\tfrac{\tau(1-\tau)}{2}\widetilde{\mathbf{d}}^\top\mathbf{y}^*,\ \tfrac{1-2\tau}{2}\mathbf{1}_n^\top(y-\mathbf{X}\widetilde{\boldsymbol{\mu}}),\ n+1,\ n,\ \mu_\kappa,\ \sigma_\kappa^2\right)$
    $\mathbb{E}_q[\exp(-2\kappa)] \leftarrow \mathcal{H}\left(\tfrac{\tau(1-\tau)}{2}\widetilde{\mathbf{d}}^\top\mathbf{y}^*,\ \tfrac{1-2\tau}{2}\mathbf{1}_n^\top(y-\mathbf{X}\widetilde{\boldsymbol{\mu}}),\ n+2,\ n,\ \mu_\kappa,\ \sigma_\kappa^2\right)$
    $\widetilde{\mathbf{d}} \leftarrow \tfrac{1}{2\tau(1-\tau)}\left[\mathbf{y}^*\mathbb{E}_q(\exp(-2\kappa))\right]^{\odot -1/2}$
end while
```

Appendix F. Supplementary figures

This section contains figures which complement the tables in Section 4 and Section 5. For all three models, we plot the mean L^1 accuracies across the benchmarks (including the big data example), and also plot the relationship between L^1 accuracies and run time.

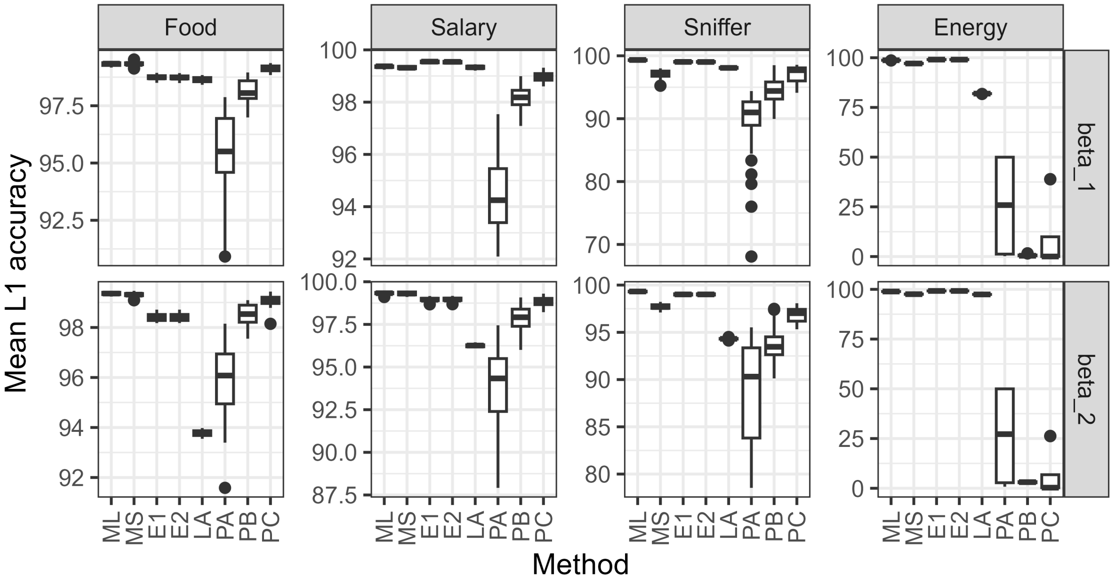

Figure 1: Mean L^1 accuracies across marginals for heteroscedastic linear regression benchmark datasets. Each point is based on a single repetition.

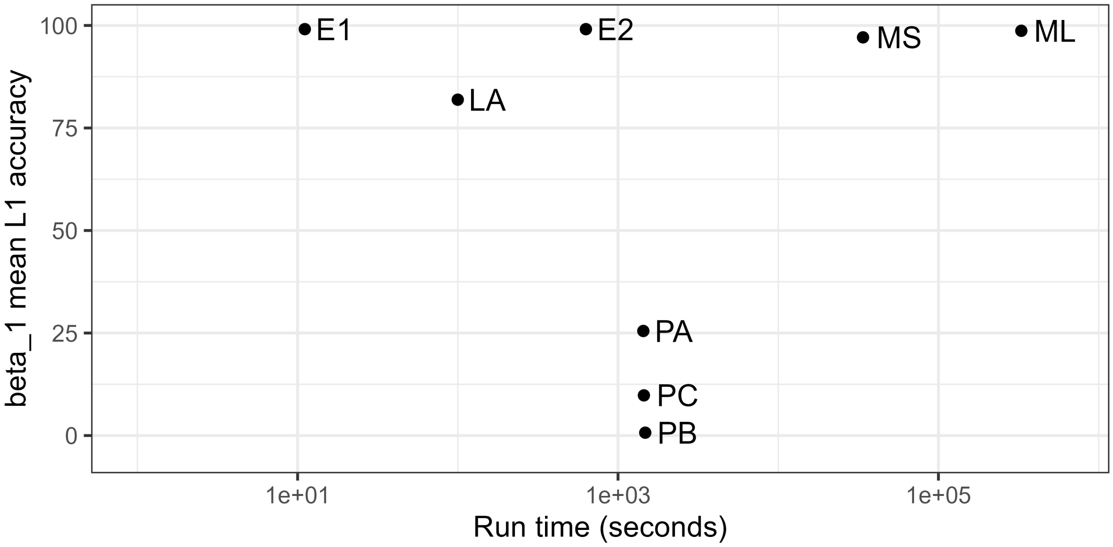

Figure 2: $\boldsymbol{\beta}_1$ L^1 accuracies vs. run times for heteroscedastic linear regression using the energy dataset. Each point represents an average over eight repetitions.

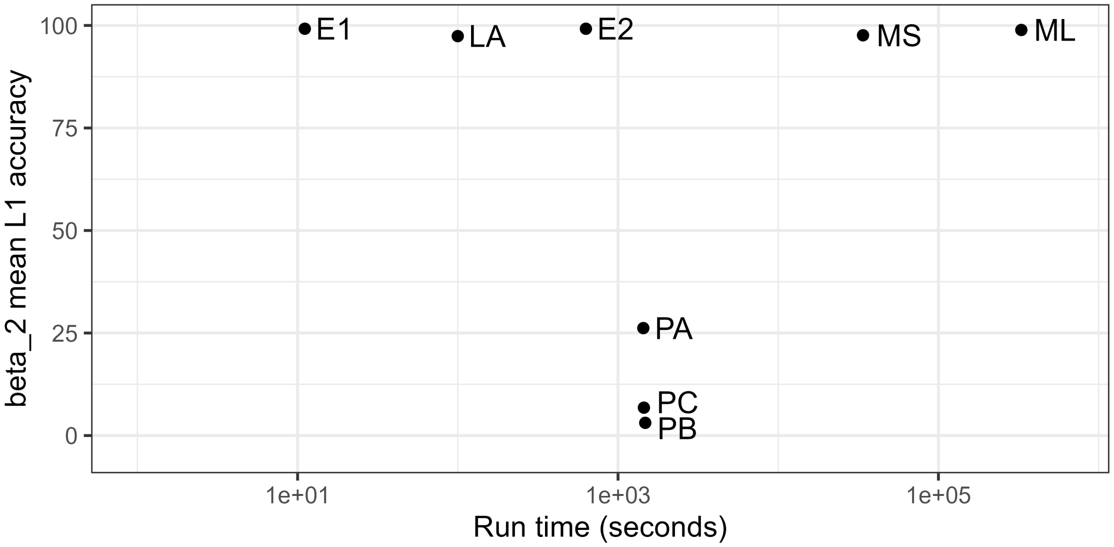

Figure 3: $\boldsymbol{\beta}_2$ L^1 accuracies vs. run times for heteroscedastic linear regression using the energy dataset. Each point represents an average over eight repetitions.

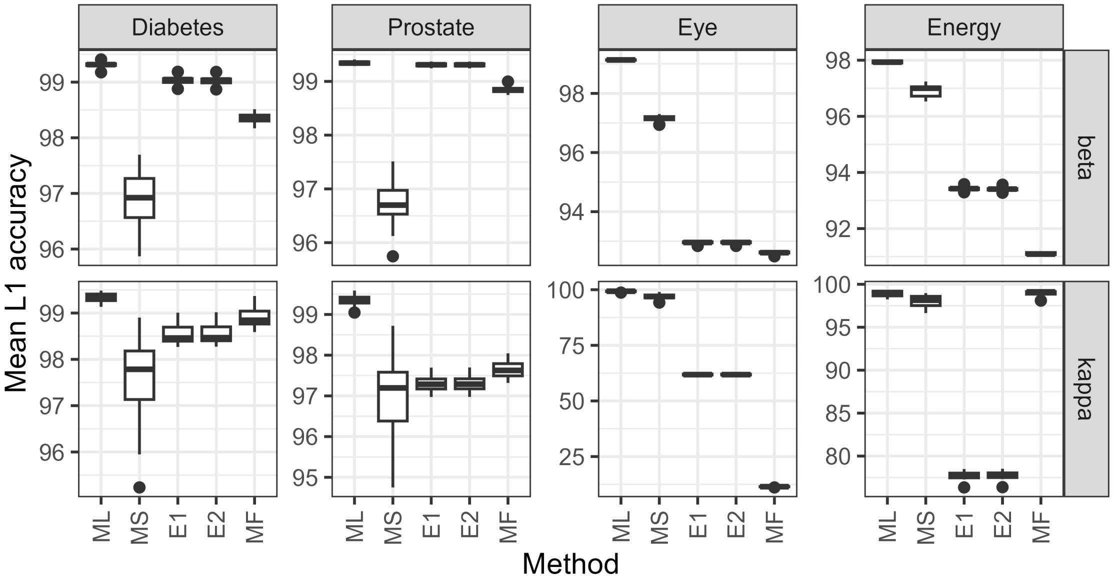

Figure 4: Mean L^1 accuracies across marginals for lasso-penalized linear regression benchmark datasets. Each point is based on a single repetition.

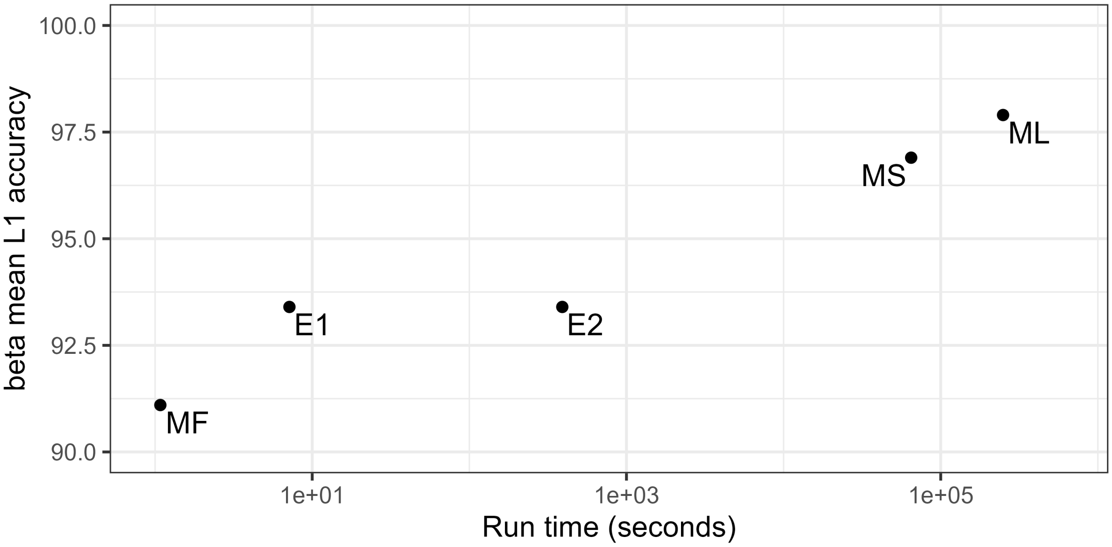

Figure 5: $\boldsymbol{\beta}$ L^1 accuracies vs. run times for lasso-penalized linear regression using the energy dataset. Each point represents an average over eight repetitions.

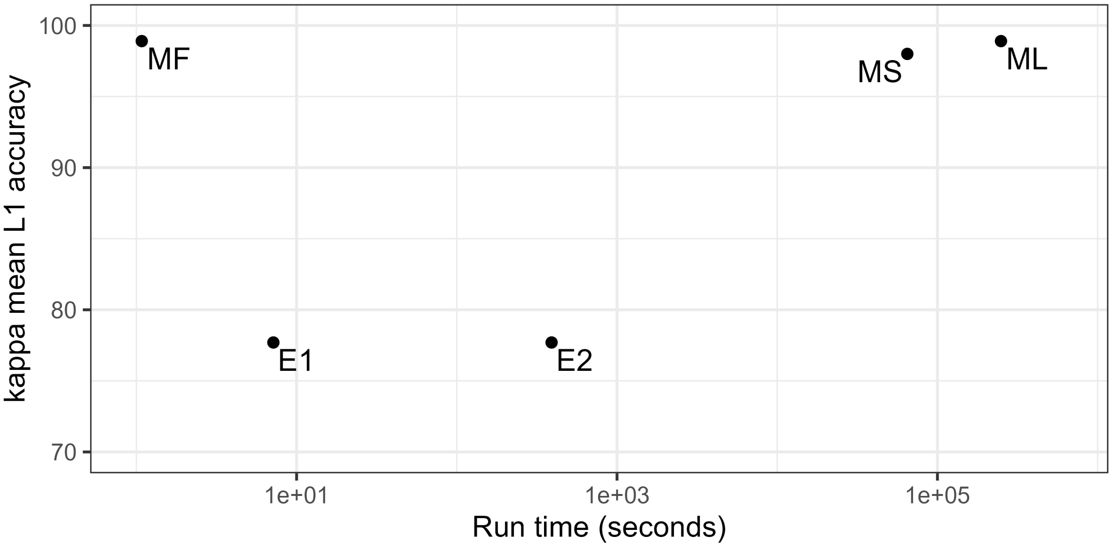

Figure 6: κ L^1 accuracies vs. run times for lasso-penalized linear regression using the energy dataset. Each point represents an average over eight repetitions.

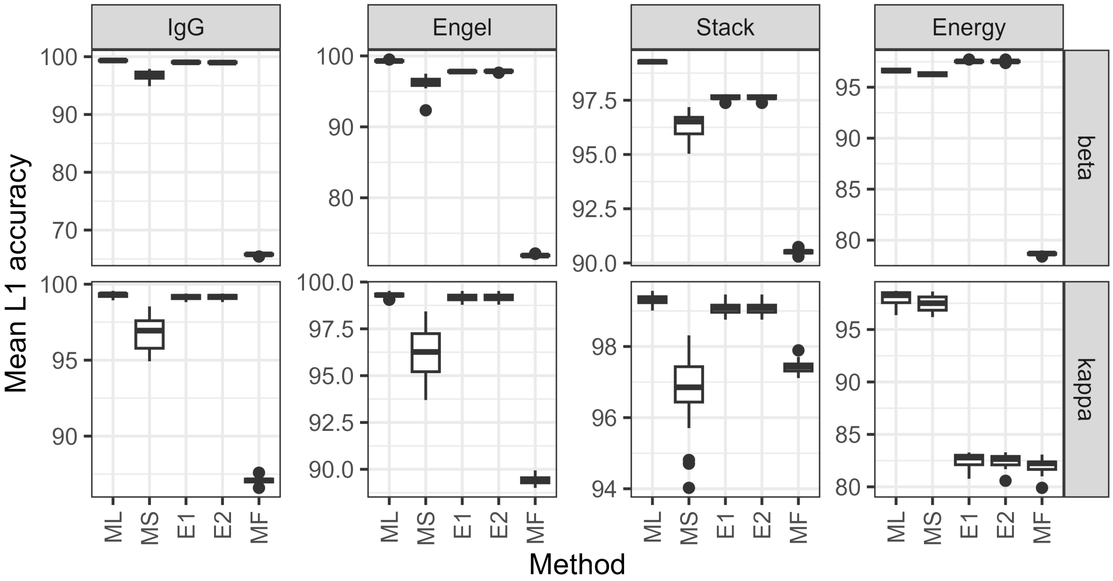

Figure 7: Mean L^1 accuracies across marginals for quantile linear regression benchmark datasets. Each point represents a single repetition.

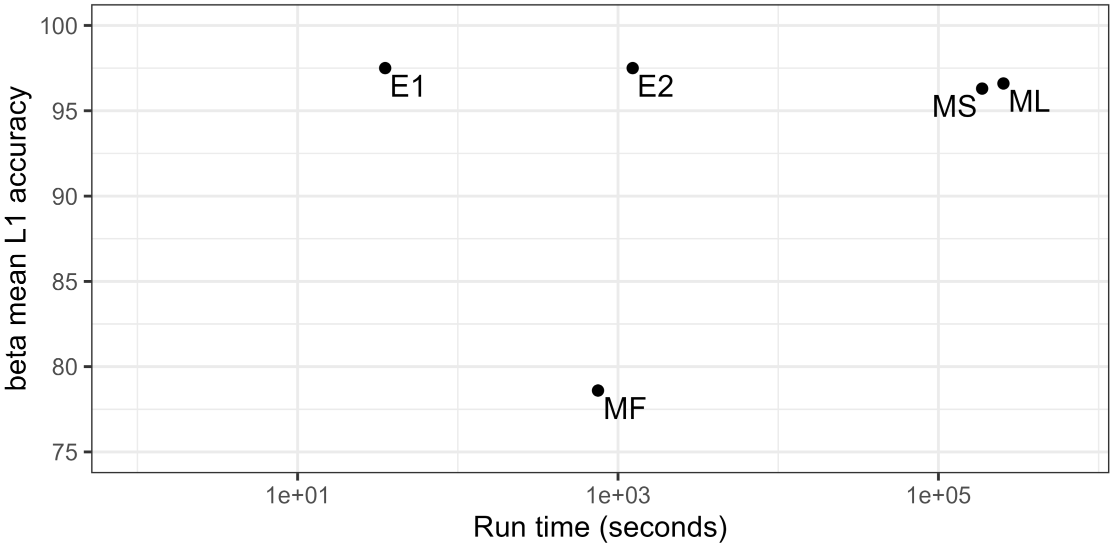

Figure 8: $\boldsymbol{\beta}$ L^1 accuracies vs. run times for quantile linear regression using the energy dataset. Each point represents an average over eight repetitions.

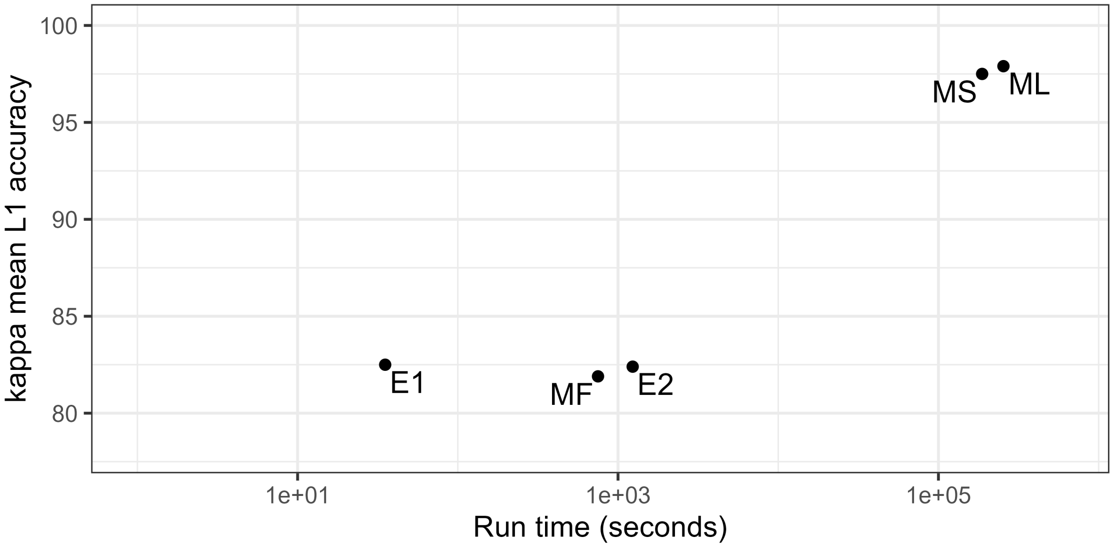

Figure 9: κ L^1 accuracies vs. run times for quantile linear regression using the energy dataset. Each point represents an average over eight repetitions.

References

Simon Barthelmé and Nicolas Chopin. Expectation propagation for likelihood-free inference. *Journal of the American Statistical Association*, 109(505):315–333, 2014.

David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. Variational inference: a review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

K. A. Brownlee. *Statistical Theory and Methodology in Science and Engineering.* John Wiley & Sons, 1 1965. ISBN 978-0471113553.

Thang Bui, Daniel Hernandez-Lobato, Jose Hernandez-Lobato, Yingzhen Li, and Richard Turner. Deep Gaussian processes for regression using approximate expectation propagation. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1472–1481, New York, New York, USA, 20–22 Jun 2016. PMLR.

Luis Candanedo. Appliances energy prediction. UCI Machine Learning Repository, 2017. DOI: https://doi.org/10.24432/C5VC8G.

Rohit Chandra, Leo Dagum, David Kohr, Ramesh Menon, Dror Maydan, and Jeff McDonald. *Parallel programming in OpenMP.* Morgan Kaufmann, 2001.

Wilson Y. Chen and Matt P. Wand. Factor graph fragmentization of expectation propagation. *Journal of the Korean Statistical Society*, 49(3):722–756, January 2020.

Nicolas Chopin and James Ridgway. Leave Pima Indians alone: binary regression as a benchmark for Bayesian computation. *Statistical Science*, 32(1):64–87, 2017.

Kenneth L Clarkson, Petros Drineas, Malik Magdon-Ismail, Michael W Mahoney, Xiangrui Meng, and David P Woodruff. The fast Cauchy transform and faster robust linear regression. *SIAM Journal on Computing*, 45(3):763–810, 2016.

Guillaume Dehaene and Simon Barthelmé. Expectation propagation in the large data limit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(1):199–217, August 2017.

Alfred DeMaris. *Regression With Social Data: Modeling Continuous and Limited Response Variables*. Wiley-Interscience, 8 2007.

Dirk Eddelbuettel and Romain François. Rcpp: Seamless R and C++ integration. *Journal of Statistical Software*, 40(8):1–18, 2011.

Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *The Annals of Statistics*, 32(2), April 2004.

C. Faes, J. T. Ormerod, and M. P. Wand. Variational Bayesian inference for parametric and nonparametric regression with missing data. *Journal of the American Statistical Association*, 106(495):959–971, 2011.

Matteo Fasiolo, Simon N Wood, Margaux Zaffran, Raphaël Nedellec, and Yannig Goude. Fast calibrated additive quantile regression. *Journal of the American Statistical Association*, 116(535):1402–1412, 2021.

Brendan J Frey and David MacKay. A revolution: belief propagation in graphs with cycles. In M. Jordan, M. Kearns, and S. Solla, editors, *Advances in Neural Information Processing Systems*, volume 10. MIT Press, 1997.

Yasuhiro Fujiwara, Yasutoshi Ida, Hiroaki Shiokawa, and Sotetsu Iwamura. Fast lasso algorithm via selective coordinate descent. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, 2016.

Andrew Gelman, John B. Carlin, Hal S. Stern, and Donald B. Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, June 1995.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25): 723–773, 2012.

P. Hall, I.M. Johnstone, J.T. Ormerod, M.P. Wand, and J.C.F. Yu. Fast and accurate binary response mixed model analysis via expectation propagation. *Journal of the American Statistical Association*, 115(532):1902–1916, 2020.

Leonard Hasenclever, Stefan Webb, Thibaut Lienart, Sebastian Vollmer, Balaji Lakshminarayanan, Charles Blundell, and Yee Whye Teh. Distributed Bayesian learning with stochastic natural gradient expectation propagation and the posterior server. *Journal of Machine Learning Research*, 18(1):3744–3780, 2017.

Nicolas Heess, Daniel Tarlow, and John Winn. Learning to pass expectation propagation messages. *Advances in Neural Information Processing Systems*, 26, 2013.

Daniel Hernandez-Lobato and Jose Miguel Hernandez-Lobato. Scalable Gaussian process classification via expectation propagation. In Arthur Gretton and Christian C. Robert, editors, *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51 of *Proceedings of Machine Learning Research*, pages 168–176, Cadiz, Spain, 09–11 May 2016. PMLR.

Daniel Hernández-Lobato, José Miguel Hernández-Lobato, and Pierre Dupont. Generalized spike-and-slab priors for Bayesian group feature selection using expectation propagation. *Journal of Machine Learning Research*, 14(7), 2013.

José Miguel Hernández-Lobato, Daniel Hernández-Lobato, and Alberto Suárez. Expectation propagation in linear regression models with spike-and-slab priors. *Machine Learning*, 99 (3):437–487, 2015.

Daniel Hernández-Lobato, José Miguel Hernández-Lobato, and Alberto Suárez. Expectation propagation for microarray data classification. *Pattern Recognition Letters*, 31(12):1618–1626, 2010. ISSN 0167-8655. Pattern Recognition of Non-Speech Audio.

R. Carter Hill, William E. Griffiths, and Guay C. Lim. *Principles of Econometrics.* Wiley, 2 2018. ISBN 978-1118452271.

Kathryn K. Hodge, John E. McNeal, Martha K. Terris, and Thomas A. Stamey. Random systematic versus directed ultrasound guided transrectal core biopsies of the prostate. *Journal of Urology*, 142(1):71–74, 1989. ISSN 0022-5347.

D. Isaacs, D. G. Altman, C. E. Tidmarsh, H. B. Valman, and A. D. Webster. Serum immunoglobulin concentrations in preschool children measured by laser nephelometry: reference ranges for IgG, IgA, IgM. *Journal of Clinical Pathology*, 36(10):1193–1196, October 1983.

M Chris Jones, James S Marron, and Simon J Sheather. A brief survey of bandwidth selection for density estimation. *Journal of the American Statistical Association*, 91(433): 401–407, 1996.

Andy S.I. Kim and Matt P. Wand. On expectation propagation for generalised, linear and mixed models. *Australian & New Zealand Journal of Statistics*, 60(1):75–102, March 2018.

Roger Koenker and Gilbert Bassett. Robust tests for heteroscedasticity based on regression quantiles. *Econometrica*, 50(1):43, January 1982.

Alp Kucukelbir, Rajesh Ranganath, Andrew Gelman, and David Blei. Automatic variational inference in Stan. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

Yoram Leedan and Peter Meer. Heteroscedastic regression in computer vision: problems with bilinear constraint. *International Journal of Computer Vision*, 37:127–150, 2000.

Ximing Li, Changchun Li, Jinjin Chi, Jihong Ouyang, and Wenting Wang. Black-box expectation propagation for Bayesian models. In *Proceedings of the 2018 SIAM International Conference on Data Mining*, pages 603–611. Society for Industrial and Applied Mathematics, May 2018.

Ping Ma and Xiaoxiao Sun. Leveraging for big data regression. *Wiley Interdisciplinary Reviews: Computational Statistics*, 7(1):70–76, 2015.

Peter S. Maybeck. *Stochastic Models, Estimation and Control.* Academic Press, 8 1982. ISBN 978-0124807037.

Thomas Minka. Power EP. Technical report, Microsoft Research, Cambridge, 2004.

Thomas P Minka and John Lafferty. Expectation-propogation for the generative aspect model. *arXiv preprint arXiv:1301.0588*, 2012.

Thomas Peter Minka. *A family of algorithms for approximate Bayesian inference.* PhD thesis, Massachusetts Institute of Technology, 2001.

John C. Nash and Ravi Varadhan. Unifying optimization algorithms to aid software system users: optimx for R. *Journal of Statistical Software*, 43(9):1–14, 2011. doi: 10.18637/jss.v043.i09.

Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning.* The MIT Press, 11 2005. ISBN 978-0262182539.

Fabiano Ribeiro and Manfred Opper. Expectation propagation with factorizing distributions: A Gaussian approximation and performance results for simple models. *Neural Computation*, 23(4):1047–1069, 2011.

Christian P. Robert, Víctor Elvira, Nick Tawn, and Changye Wu. Accelerating MCMC algorithms. *WIREs Computational Statistics*, 10(5), June 2018.

Patrick J Rosopa, Meline M Schaffer, and Amber N Schroeder. Managing heteroscedasticity in general linear models. *Psychological Methods*, 18(3):335, 2013.

Håvard Rue, Sara Martino, and Nicolas Chopin. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(2):319–392, April 2009.

Todd E. Scheetz, Kwang-Youn A. Kim, Ruth E. Swiderski, Alisdair R. Philp, Terry A. Braun, Kevin L. Knudtson, Anne M. Dorrance, Gerald F. DiBona, Jian Huang, Thomas L. Casavant, Val C. Sheffield, and Edwin M. Stone. Regulation of gene expression in the mammalian eye and its relevance to eye disease. *Proceedings of the National Academy of Sciences*, 103(39):14429–14434, September 2006.

Matthias Seeger. Expectation propagation for exponential families. Technical report, Laboratory for Probabilistic Machine Learning, 2005.

Matthias Seeger and Michael Jordan. Sparse Gaussian process classification with multiple classes. Technical report, Laboratory for Probabilistic Machine Learning, 2004.

Matthias Seeger, Florian Steinke, and Koji Tsuda. Bayesian inference and optimal design in the sparse linear model. In Marina Meila and Xiaotong Shen, editors, *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics*, volume 2 of *Proceedings of Machine Learning Research*, pages 444–451, San Juan, Puerto Rico, 21–24 Mar 2007. PMLR.

Simon J Sheather and Michael C Jones. A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society: Series B (Methodological)*, 53(3):683–690, 1991.

Alex Smola, S.V.N. Vishwanathan, and Eleazar Eskin. Laplace propagation. In S. Thrun, L. Saul, and B. Schölkopf, editors, *Advances in Neural Information Processing Systems*, volume 16. MIT Press, 2003.

Stan Development Team. RStan: the R interface to Stan, 2023. R package version 2.21.8.

Aki Vehtari, Andrew Gelman, Tuomas Sivula, Pasi Jylänki, Dustin Tran, Swupnil Sahai, Paul Blomstedt, John P. Cunningham, David Schiminovich, and Christian P. Robert. Expectation propagation as a way of life: a framework for Bayesian inference on partitioned data. *Journal of Machine Learning Research*, 21(17):1–53, 2020.

Matthew P. Wand, John T. Ormerod, Simone A. Padoan, and Rudolf Frühwirth. Mean field variational Bayes for elaborate distributions. *Bayesian Analysis*, 6(4), December 2011.

Hai Ying Wang, Min Yang, and John Stufken. Information–based optimal subdata selection for big data linear regression. *Journal of the American Statistical Association*, 114(525): 393–405, 2019.

Haiying Wang and Yanyuan Ma. Optimal subsampling for quantile regression in big data. *Biometrika*, 108(1):99–112, 2021.

Hanqing Wang, Alva Kosasih, Chao-Kai Wen, Shi Jin, and Wibowo Hardjawana. Expectation propagation detector for extra-large scale massive MIMO. *IEEE Transactions on Wireless Communications*, 19(3):2036–2051, 2020.

Sanford Weisberg. *Applied Linear Regression.* Wiley, 12 2013. ISBN 978-1118386088.

Keming Yu and Rana A. Moyeed. Bayesian quantile regression. *Statistics & Probability Letters*, 54(4):437–447, 2001. ISSN 0167-7152.

Lu Zhang, Bob Carpenter, Andrew Gelman, and Aki Vehtari. Pathfinder: parallel quasi-Newton variational inference. *Journal of Machine Learning Research*, 23(1):13802–13850, 2022.