*Biometrika* (2020), **xx**, x, *pp*. 1–14
*Printed in Great Britain*

# Approximating posteriors with high-dimensional nuisance parameters via integrated rotated Gaussian approximation

BY W. VAN DEN BOOM

*Yale-NUS College, National University of Singapore, 16 College Avenue West #01-220, Singapore 138527, Singapore*
willem@yale-nus.edu.sg

G. REEVES AND D. B. DUNSON

*Department of Statistical Science, Duke University, Box 90251, Durham, North Carolina 27708, U.S.A.*
galen.reeves@duke.edu dunson@duke.edu

SUMMARY

Posterior computation for high-dimensional data with many parameters can be challenging. This article focuses on a new method for approximating posterior distributions of a low- to moderate-dimensional parameter in the presence of a high-dimensional or otherwise computationally challenging nuisance parameter. The focus is on regression models and the key idea is to separate the likelihood into two components through a rotation. One component involves only the nuisance parameters, which can then be integrated out using a novel type of Gaussian approximation. We provide theory on approximation accuracy that holds for a broad class of forms of the nuisance component and priors. Applying our method to simulated and real data sets shows that it can outperform state-of-the-art posterior approximation approaches.

*Some key words*: Bayesian statistics; Dimensionality reduction; Marginal inclusion probability; Nuisance parameter; Posterior approximation; Support recovery; Variable selection

## 1. INTRODUCTION

Consider the regression model

$$y \sim N\big(X\beta + \eta,\ \sigma^2 I_n\big), \tag{1}$$

where $y$ is an $n$-dimensional vector of observations, $X$ is an $n \times p$ design matrix, $\beta$ is a $p$-dimensional parameter of interest, $\eta$ is an $n$-dimensional nuisance parameter, and $\sigma^2$ is the error variance. The nuisance parameter can for instance capture the effect of a large set of covariates not included in $X$, or of non-Gaussian errors. Our goal is Bayesian inference on the model in (1) when $p$ is of moderate size such that $p \ll n - p$ with the focus on the posterior

$$\pi(\beta \mid y) = \int \pi(\beta, \eta \mid y)\, d\eta = \frac{1}{\pi(y)} \int \pi(y \mid \beta, \eta)\, \pi(\beta, \eta)\, d\eta. \tag{2}$$

The integrals in (2) and $\pi(y)$ are intractable to approximate accurately for certain priors $\pi(\beta, \eta)$, with direct approximations such as Laplace's method producing inaccurate results and Monte Carlo sampling being daunting computationally. Our key idea is to transform the hard problem with nuisance parameter $\eta$ in a principled way to a $p$-dimensional one which can be written as

a linear model including only $\beta$. Then, a low-dimensional inference technique can be applied to this $p$-dimensional model. The transformation uses a novel type of Gaussian approximation using a data rotation to integrate out $\eta$ from (1).

Section 3 discusses special cases of the model in (1). Applications include epidemiology studies in which $y$ is a health outcome, $X$ consists of exposures and key clinical or demographic factors of interest, and $\eta$ is the effect of high-dimensional biomarkers. The goal is inference on the effect of the exposures and the clinical or demographic covariates, but adjusting for the high-dimensional biomarkers. For example, $\eta$ may result from genetic factors, such as single-nucleotide polymorphisms (SNPs), and we want to control for these in identifying an environmental main effect. It is often impossible to isolate the impact of individual genetic factors so we consider these effects as nuisance parameters. Another use of (1) is computation of posterior inclusion probabilities in high-dimensional Bayesian variable selection as detailed in § 3·2.

Data with a complex component $\eta$ that is not of primary interest and only a moderate number $p$ of parameters of interest, are more and more common. Unfortunately, the complexity of $\eta$ can make accurate approximation of $\pi(\beta \mid y)$ in (2) challenging even when $p = 1$. One naive approach is to ignore the nuisance parameter $\eta$ by setting it to zero. The result can be problematic as omitting $\eta$ changes the interpretation of the parameter of interest $\beta$, which therefore might take on a different value. For example, $\eta$ might capture the effect of covariates with it being important to adjust for them to avoid misleading conclusions on $\beta$.

Many posterior approximation methods exist, including Monte Carlo (George & McCulloch, 1993, 1997; O'Hara & Sillanpää, 2009), variational Bayes (Carbonetto & Stephens, 2012; Ormerod et al., 2017), integrated nested Laplace approximations (Rue et al., 2009), and expectation propagation (Hernández-Lobato et al., 2015). However, these methods can be computationally expensive, do not apply to our setting, or lack theoretical results regarding approximation accuracy. A notable exception to the latter is the fast posterior approximation algorithm of Huggins et al. (2017) which comes with bounds on the approximation error under conditions on the prior such as log-concavity, Gaussianity, and smoothness. The class of priors that we allow on $\beta$ and $\eta$ is much larger. Our method and its analysis for instance apply to dimensionality reduction and shrinkage priors such as spike-and-slab, horseshoe, and Laplace distributions.

The main computational bottleneck of our method is calculation of the mean and variance of a nuisance term, for which one can choose any suitable algorithm. As a result, the computational cost of our method is comparable to that of the fast algorithm chosen for this step.

## 2. INTEGRATED ROTATED GAUSSIAN APPROXIMATION

### 2·1. *Notation and assumptions*

Denote the multivariate Gaussian distribution with mean $\mu$ and covariance $\Sigma$ by $N(\mu,\ \Sigma)$, and its density function evaluated at $a$ by $N(a \mid \mu,\ \Sigma)$. Denote the distribution of $a$ conditional on $b$ by $\Pi(a \mid b)$ and its density, with respect to some dominating measure, evaluated at $a$ by $\pi(a \mid b)$. We assume that $\beta$ and $\eta$ are a priori independent so that $\Pi(\beta, \eta) = \Pi(\beta)\Pi(\eta)$. We treat $X$ and $\sigma^2$ as known constants unless otherwise noted. Assume that $p \leq n$. We assume that $X$ is full rank to simplify the exposition, but our method also applies to rank deficient $X$.

### 2·2. *Description of the method*

We integrate out $\eta$ from (1) by splitting the model into two parts, one of which does not involve $\beta$. A data rotation provides such a model split. Specifically, consider as rotation matrix the $n \times n$ orthogonal matrix $Q$ from the QR decomposition of $X$. Define the $n \times p$ matrix $M$ and the $n \times (n - p)$ matrix $S$ by $(M,\ S) = Q$. Then, the columns of $M$ form an orthonormal

basis for the column space of $X$ since $X$ is full rank by assumption (Golub & Van Loan, 1996, § 5.2). Since $Q$ is orthogonal, the columns of $S$ form an orthonormal basis for the orthogonal complement of the column space of $X$. Therefore, $S^{\mathrm{T}}X = 0_{(n-p)\times p}$, an $(n-p)\times p$ matrix of zeros, which can also be derived from the fact that $Q^{\mathrm{T}}X$ is upper triangular.

By the rotational invariance of the Gaussian distribution and $Q^{\mathrm{T}}Q = I_n$, $Q^{\mathrm{T}}y \sim N\big(Q^{\mathrm{T}}X\beta + Q^{\mathrm{T}}\eta,\ \sigma^2 I_n\big)$ is distributionally equivalent to (1). This rotated model splits as

$$M^{\mathrm{T}}y \sim N\big(M^{\mathrm{T}}X\beta + M^{\mathrm{T}}\eta,\ \sigma^2 I_p\big), \tag{3a}$$

$$S^{\mathrm{T}}y \sim N\big(S^{\mathrm{T}}\eta,\ \sigma^2 I_{n-p}\big); \tag{3b}$$

using $S^{\mathrm{T}}X = 0_{(n-p)\times p}$. This transformation motivates a two-stage approach in which one first computes $\Pi(\eta \mid S^{\mathrm{T}}y)$ from submodel (3b) and then uses this distribution as an updated prior for the projected nuisance term $M^{\mathrm{T}}\eta$ in submodel (3a). Following this approach, the posterior of $\beta$ can be expressed as $\Pi(\beta \mid y) \propto \Pi(\beta) \int N(M^{\mathrm{T}}y \mid M^{\mathrm{T}}X\beta + M^{\mathrm{T}}\eta,\ \sigma^2 I_p)\, d\Pi(M^{\mathrm{T}}\eta \mid S^{\mathrm{T}}y)$.

In practice, $\Pi(M^{\mathrm{T}}\eta \mid S^{\mathrm{T}}y)$ may be intractable to compute exactly because of the complexity of $\Pi(\eta)$. To alleviate this challenge, we consider an approximation $\hat{\Pi}(M^{\mathrm{T}}\eta \mid S^{\mathrm{T}}y)$, which then leads to an approximation for the posterior of $\beta$:

$$\hat{\Pi}(\beta \mid y) \propto \Pi(\beta) \int N(M^{\mathrm{T}}y \mid M^{\mathrm{T}}X\beta + M^{\mathrm{T}}\eta,\ \sigma^2 I_p)\, d\hat{\Pi}(M^{\mathrm{T}}\eta \mid S^{\mathrm{T}}y). \tag{4}$$

All distributions, densities, and probabilities resulting from this approximation carry a hat to distinguish them from their exact counterparts.

A Gaussian approximation is analytically convenient:

$$\hat{\Pi}(M^{\mathrm{T}}\eta \mid S^{\mathrm{T}}y) = N(\hat{\mu},\ \hat{\Sigma}), \tag{5}$$

where $\hat{\mu}$ and $\hat{\Sigma}$ are estimates of the mean and covariance of $\Pi(M^{\mathrm{T}}\eta \mid S^{\mathrm{T}}y)$, respectively. In this case, (4) simplifies as

$$\hat{\Pi}(\beta \mid y) \propto \Pi(\beta)\, N(M^{\mathrm{T}}y \mid M^{\mathrm{T}}X\beta + \hat{\mu},\ \sigma^2 I_p + \hat{\Sigma}). \tag{6}$$

Only $\beta$ is unknown such that the computational problems with (1) resulting from the complexity of $\Pi(\eta)$ have been resolved in (6). Furthermore, (6) is equivalent to a Gaussian linear model with observations $M^{\mathrm{T}}y - \hat{\mu}$, design matrix $M^{\mathrm{T}}X$, and parameter $\beta$. We have reduced a model with a potentially challenging nuisance parameter to a low-dimensional one with the nuisance integrated out while controlling for the effect of the nuisance parameter in a principled manner. Algorithm 1 summarizes our method when the Gaussian approximation from (5) is used.

*Algorithm* 1. Integrated rotated Gaussian approximation.
Input: Data $(y,\ X)$

1. Compute the QR decomposition of $X$ to obtain the rotation matrix $Q = (M,\ S)$.
2. Compute the estimates $\hat{\mu}$ and $\hat{\Sigma}$ for the mean and covariance of $\Pi(M^{\mathrm{T}}\eta \mid S^{\mathrm{T}}y)$ based on submodel (3b) using an algorithm of choice.
3. Approximate the posterior $\Pi(\beta \mid y)$ according to (6).

Output: The approximate posterior $\hat{\Pi}(\beta \mid y)$

### 2·3. *Relation to other methods*

Algorithm 1 has resemblances with other approximation methods. Integrated nested Laplace approximations (Rue et al., 2009) also approximate a nested part of a Bayesian model by a Gaussian distribution but with important differences. A Laplace approximation is applied without a

data rotation and is done at two, rather than one, nested levels of the model. Moreover, a Laplace approximation matches the mode and curvature of the approximating Gaussian while (5) matches the moments. Laplace's method (Tierney & Kadane, 1986) requires a continuous target distribution and integrated nested Laplace approximations assume a conditionally Gaussian prior on some parameters. Our Gaussian approximation needs no such conditions on priors but assumes a Gaussian error distribution. For instance, § 3 considers examples of priors on $\eta$ that are not continuous or are non-Gaussian.

The approximation in (5) aims to match the first two moments of the exact $\Pi(M^{\mathrm{T}}\eta \mid S^{\mathrm{T}}y)$. Such matching is the principle behind expectation consistent inference (Opper & Winther, 2005). Our method matches moments for the nuisance parameter but not for the parameter of interest $\beta$. This differs from applications of the expectation consistent framework in which moment matching is pervasive such as in expectation propagation (Hernández-Lobato et al., 2015). Implementations of expectation propagation are usually not able to capture dependence among dimensions of the posterior while our method allows for dependence in the $p$-dimensional $\beta$.

Effectively, our method integrates out the nuisance parameter $\eta$ approximately. Integrating out nuisance parameters from the likelihood is not new (Berger et al., 1999), including doing so approximately (Severini, 2011). Previous approximations, however, do not apply a data rotation and consider cases where the distribution on the nuisance parameter is regular enough so that a Laplace approximation can be applied. Our method does not need such regularity conditions.

The rotation $Q$ is similar to the projection in the Frisch-Waugh-Lovell theorem (Stachurski, 2016, Theorem 11.2.1) for least-squares estimation of a parameter subset. Our method applies beyond least squares. Also, our estimation of the nuisance parameter through the rotation is merely an intermediate step for inference on $\beta$. Our method reduces to the algorithm from van den Boom et al. (2015) when considering the example in § 3·2 with $p = 1$.

### 2·4. *Estimating $\sigma^2$ and hyperparameters*

So far, we have treated $\sigma^2$ as fixed and known. In practice, $\sigma^2$ usually needs to be estimated, as well as any unknown parameters in the prior on $\eta$. This estimation fits naturally into Step 2 of Algorithm 1 as the methods that can be used there frequently come with such estimation procedures: See for instance § S5·3 and § S6 of the Supplementary Material. The resulting estimates can then be plugged into Step 3. By doing so, only the $(n-p)$-dimensional submodel (3b) informs the estimates of these parameters and not the $p$-dimensional submodel (3a). We expect (3b) to contain the vast majority of information on the unknown parameters if $(n-p) \gg p$, which is often the case in scenarios of interest.

## 3. EXAMPLES OF NUISANCE PARAMETERS $\eta$

### 3·1. *Adjusting for high-dimensional covariates*

Section 3 provides examples of the general setting of model (1) that demonstrate the utility of the integrated rotated Gaussian approximation in Algorithm 1. As a first example, consider $\eta = Z\alpha$ with $Z$ a known $n \times q$ feature matrix and $\alpha$ an unknown $q$-dimensional parameter with $q \gg n$. Then, the model in (1) becomes $y \sim N\big(X\beta + Z\alpha,\, \sigma^2 I_n\big)$, so that we are adjusting for high-dimensional covariates $Z$ in performing inference on the coefficients $\beta$ on the predictors $X$ of interest. One way to deal with the fact that the number of covariates $q$ exceeds the number of observations $n$ is by inducing sparsity in $\alpha$ via its prior $\Pi(\alpha)$. We consider the spike-and-slab prior, $\alpha_j \sim \lambda\, N(0,\, \psi) + (1-\lambda)\, \delta(0)$ independently for $j = 1, \ldots, q$, where $\lambda = \mathrm{pr}(\alpha_j \neq 0)$ is the prior inclusion probability, $\psi$ the slab variance, and $\delta(0)$ a point mass at zero. By specifying $\Pi(\alpha)$, we have also defined $\Pi(\eta) = \Pi(Z\alpha)$. Since each $\Pi(\alpha_j)$ is a mixture of a point mass and a

Gaussian, $\Pi(\alpha)$ and thus $\Pi(\eta)$ are mixtures of $2^q$ Gaussians. As a result, computation of $\pi(\beta \mid y)$ in (2) involves summing over these $2^q$ components. This is infeasible for large $q$.

Algorithm 1 provides an approximation $\hat{\Pi}(\beta \mid y)$ while avoiding the exponential computational cost. Step 2 in Algorithm 1 requires choice of an estimation algorithm. Substituting $\eta = Z\alpha$ into (3b) yields $S^{\mathrm{T}}y \sim N(S^{\mathrm{T}}Z\alpha,\, \sigma^2 I_{n-p})$, which is a linear model with $(n-p)$ observations and design matrix $S^{\mathrm{T}}Z$. As such, methods for linear regression with spike-and-slab priors can produce an approximation to $\Pi(\alpha \mid S^{\mathrm{T}}y)$ and thus the estimates $\hat{\mu}$ and $\hat{\Sigma}$ in (5). We choose vector approximate message passing (Rangan et al., 2017), detailed in § S5 of the Supplementary Material, to approximate $\Pi(\alpha \mid S^{\mathrm{T}}y)$ because of its computational scalability and accuracy. The computational scalability limits the size of $q$. For instance, § 5·3 considers a subset of $q = 10,000$ SNPs as using all SNPs was computationally infeasible. As a more scalable alternative, we consider the debiased lasso (Javanmard & Montanari, 2013) in § 5·2 as it can also approximate $\Pi(\alpha \mid S^{\mathrm{T}}y)$ as detailed in § S6 of the Supplementary Material. A $q$ in the millions is feasible with embarrassingly parallel split-and-merge strategies (Song & Liang, 2014). The $q$-dimensional distribution $\Pi(\alpha \mid S^{\mathrm{T}}y)$ is possibly highly non-Gaussian, being a mixture of Gaussians. At the same time, the $p$-dimensional distribution $\Pi(M^{\mathrm{T}}Z\alpha \mid S^{\mathrm{T}}y) = \Pi(M^{\mathrm{T}}\eta \mid S^{\mathrm{T}}y)$ can be nearly Gaussian such that the approximation in (5) is accurate as discussed in § 4·2.

### 3·2. *Bayesian variable selection*

For a second application of (1), consider the linear model $y \sim N(A\theta,\, \sigma^2 I_n)$ where $A$ is a known $n \times r$ design matrix and $\theta$ an unknown $r$-dimensional parameter. Variable selection is the problem of determining which entries of $\theta$ are non-zero. Modeling the data in a Bayesian fashion provides a natural framework to evaluate statistical evidence via the posterior $\Pi(\theta \mid y)$. A standard variable selection prior $\Pi(\theta)$ is the spike-and-slab prior defined by $\theta_j \sim \lambda\, N(0,\, \psi) + (1-\lambda)\, \delta(0)$ independently for $j = 1, \ldots, p$. As in § 3·1, the cost of computing the exact posterior with a spike-and-slab prior grows exponentially in $r$. Therefore, computation of $\Pi(\theta \mid y)$ is infeasible for $r$ beyond moderate size. A variety of approximation methods exist for larger $r$ including Monte Carlo (George & McCulloch, 1993, 1997; O'Hara & Sillanpää, 2009), variational Bayes (Carbonetto & Stephens, 2012; Ormerod et al., 2017), and expectation propagation (Hernández-Lobato et al., 2015).

Monte Carlo methods do not scale well with the number of predictors $r$. For $r$ even moderately large, the $2^r$ possible non-zero subsets of $\theta$ is so huge that there is no hope of visiting more than a vanishingly small proportion of models. The result is high Monte Carlo error in estimating posterior probabilities, with almost all models assigned zero probability as they are never visited. As an alternative to Monte Carlo sampling, fast approximation approaches for Bayesian variable selection include variational Bayes (Carbonetto & Stephens, 2012; Ormerod et al., 2017) and expectation propagation (Hernández-Lobato et al., 2015). Their accuracy, however, does not come with theory guarantees. Our method, which applies to variable selection as detailed in the next paragraph, allows for theoretical analysis as § 4 shows.

In variable selection, often the main question asked is whether $\theta_j \neq 0$ $(j = 1, \ldots, r)$ as measured by the posterior inclusion probability $\mathrm{pr}(\theta_j \neq 0 \mid y)$. Algorithm 1 can estimate $\mathrm{pr}(\theta_j \neq 0 \mid y)$: Let $p < r$ elements from $\theta$ constitute $\beta$ and let the other $q = r - p$ elements in $\theta$ constitute $\alpha$. Then, $A\theta = X\beta + Z\alpha$ where $X$ and $Z$ consist of the respective columns in $A$, and $\Pi(\alpha, \beta) = \Pi(\alpha)\, \Pi(\beta)$ since $\Pi(\theta) = \prod_{j=1}^{r} \Pi(\theta_j)$. This set-up is the same as in § 3·1 and Algorithm 1 approximates $\Pi(\beta \mid y)$ as in § 3·1. Assuming $\theta_j$ is contained in $\beta$, an approximation of $\Pi(\theta_j \mid y)$ can be obtained as a marginal distribution of $\hat{\Pi}(\beta \mid y)$. Repeating Algorithm 1 with different splits of $\theta$ into $\beta$ and $\alpha$ provides estimates of all $\mathrm{pr}(\theta_j \neq 0 \mid y)$ $(j = 1, \ldots, r)$. Computations for these different splits can run in parallel.

The approximation accuracy is not very sensitive to how $\theta$ is split into $\alpha$ and $\beta$, and to $p$ per § S9 of the Supplementary Material. We therefore use simple sequential splitting, where the first $p$ elements of $\theta$ constitute $\beta$ in the first split, and recommend choosing $p$ based on computational complexity. Assume that the number of CPU cores is less than the number of variables $r$. Then, computation time to obtain all $\hat{\text{pr}}(\theta_j \neq 0 \mid y)$ is a trade-off between the length $p$ of $\beta$, which affects the cost of each execution of Algorithm 1, and the number $r/p$ of executions of Algorithm 1. The order of $r$ is limited by the order of $q$, which is again limited by the algorithm chosen for Step 2 of Algorithm 1 as discussed in § 3·1. The complexity in terms of $p$ and $r$ of computing all $\hat{\text{pr}}(\theta_j \neq 0 \mid y)$ is $O(r^2 \log^2 r)$ if $p = O(\log r)$ and vector approximate message passing is used as detailed in the next paragraph.

Step 1 of Algorithm 1 is the QR decomposition of an $n \times p$ matrix which has complexity $O(np^2)$ (Golub & Van Loan, 1996, § 5.2). Step 2 involves vector approximate message passing on $n - p$ observations and $q$ parameters, which has a complexity of $O\{(n - p + K)\, q \min(n - p, q)\}$ where $K$ is the number of message passing iterations as detailed in § S5·2 of the Supplementary Material. Additionally for Step 2, computation of $S^{\mathrm{T}}y$ and $S^{\mathrm{T}}Z$, which are the observations and design matrix in (3b), and computing $\hat{\mu}$ and $\hat{\Sigma}$ in (5) from the message passing output is $O(n^2 q)$. Computing Step 3 with the spike-and-slab prior $\Pi(\beta)$ is $O(2^p\, p^3)$, ignoring dependence on $n$. The complexity of obtaining all $\hat{\text{pr}}(\theta_j \neq 0 \mid y)$ by applying Algorithm 1 $r/p$ times is thus $O\{(r/p)(q + 2^p\, p^3)\} = O\{(r/p)(r - p + 2^p\, p^3)\}$, ignoring dependence on $n$ and $K$. For $p = O(\log r)$, this complexity reduces to $O(r^2 \log^2 r)$.

### 3·3. *Non-parametric adjustment for covariates*

As a last example, let $\eta_i = (g \circ f)(z_i)$ $(i = 1, \ldots, n)$ where $g : \mathbb{R} \to \mathbb{R}$ is a known, differentiable, non-linear function, $f : \mathbb{R}^q \to \mathbb{R}$ is an unknown function, $g \circ f : \mathbb{R}^q \to \mathbb{R}$ is $g$ composed with $f$, and $z_i$ is a $q$-dimensional feature vector. Then, $\eta_i$ provides a non-parametric adjustment for the covariate $z_i$ in performing inferences on the effect of $x_i$ on $y_i$. Take $f$'s prior as a Gaussian process that induces a prior $\Pi(\eta)$. Algorithm 1 applies if a Gaussian approximation $\hat{\Pi}(M^{\mathrm{T}}\eta \mid S^{\mathrm{T}}y)$ is available: Submodel (3b) reduces to $S^{\mathrm{T}}y \sim N\{S^{\mathrm{T}}G(F),\, \sigma^2 I_{n-p}\}$ where $F = \{f(z_1), \ldots, f(z_n)\}^{\mathrm{T}}$ and $G(F) = \{g(F_1), \ldots, g(F_n)\}^{\mathrm{T}}$, which is a non-linear Gaussian model as studied in Steinberg & Bonilla (2014). Linearizing $G$ using a first-order Taylor series yields a Gauss-Newton algorithm for a Laplace approximation of $\Pi(F \mid S^{\mathrm{T}}y)$ as detailed in § S7 of the Supplementary Material. Based on that approximation, compute $\hat{\mu}$ and $\hat{\Sigma}$ in (5), for instance by sampling $F$ from a Laplace approximation $\hat{\Pi}(F \mid S^{\mathrm{T}}y)$ and computing the sample mean and covariance of $M^{\mathrm{T}}G(F)$ since $M^{\mathrm{T}}\eta = M^{\mathrm{T}}G(F)$.

## 4. ANALYSIS OF INTEGRATED ROTATED GAUSSIAN APPROXIMATION

### 4·1. *Approximation accuracy*

This section provides theoretical guarantees on the accuracy of our posterior approximation framework. We begin with a general upper bound in terms of the accuracy of the approximation for the projected nuisance parameter. For this, denote the distribution of the $p$-dimensional $a + b$ where $b \sim N(0,\, \sigma^2 I_p)$ by $\Pi(a) * N_{\sigma^2}$. Define the Kullback-Leibler divergence from $\Pi(b)$ to $\Pi(a)$ as $D\{\Pi(a) \parallel \Pi(b)\} = \int \log\{\pi(a)/\pi(b)\} d\Pi(a)$.

At a high level, it is clear that the accuracy of the approximation $\hat{\Pi}(\beta \mid y)$ defined in (4) depends on the accuracy of the approximation $\hat{\Pi}(M^{\mathrm{T}}\eta \mid S^{\mathrm{T}}y)$. The following result quantifies the nature of this dependence in the setting where the data are generated from the prior predictive dis-

tribution. This result applies generally for any approximation $\hat{\Pi}(M^{\mathrm{T}}\eta \mid S^{\mathrm{T}}y)$ and thus includes the Gaussian approximation (5) used in Algorithm 1 as a special case.

THEOREM 1. *Let $y$ be distributed according to the model in* (1) *with $\beta \sim \Pi(\beta)$ and $\eta \sim \Pi(\eta)$ distributed according to their priors. Conditional on any realization $S^{\mathrm{T}}y$, the posterior approximation $\hat{\Pi}(\beta \mid y)$ described in* (4) *satisfies*

$$E\Big[D\Big\{\Pi(\beta \mid y) \parallel \hat{\Pi}(\beta \mid y)\Big\} \;\Big|\; S^{\mathrm{T}}y\Big] \leq D\Big\{\Pi(M^{\mathrm{T}}\eta \mid S^{\mathrm{T}}y) * N_{\sigma^2} \parallel \hat{\Pi}(M^{\mathrm{T}}\eta \mid S^{\mathrm{T}}y) * N_{\sigma^2}\Big\},$$

*where the expectation on the left is with respect to the conditional distribution of $y$ given $S^{\mathrm{T}}y$.*

A particularly useful property of Theorem 1 is that the upper bound does not depend in any way on the prior $\Pi(\beta)$. This differs from some of the related work on posterior approximation, such as Huggins et al. (2017), which requires additional smoothness constraints, and thus excludes certain priors such as the spike-and-slab prior in § 3·1. Another useful property of Theorem 1 is that it does not require any assumptions about the extent to which the exact posterior $\Pi(M^{\mathrm{T}}\eta \mid S^{\mathrm{T}}y)$ is concentrated about the ground truth. As a consequence, this result is relevant for non-asymptotic settings where there may be high uncertainty about $\eta$.

### 4·2. *Accuracy of the Gaussian approximation*

Next, we provide theoretical justification for a Gaussian approximation to $\Pi(M^{\mathrm{T}}\eta \mid S^{\mathrm{T}}y)$ by showing that such an approximation can be accurate even when the prior on $\eta$ is highly non-Gaussian. Without loss of generality, we focus on the set-up of § 3·1 where the nuisance term has the form $\eta = Z\alpha$ with a known $n \times q$ feature matrix $Z$ and unknown parameter vector $\alpha$. In this setting, the projected nuisance parameter $M^{\mathrm{T}}\eta$ can be expressed as $M^{\mathrm{T}}Z\alpha$ where $M^{\mathrm{T}}Z$ is a $p \times q$ matrix with $p \ll q$. There are no constraints on the dimension $n$ other than $n \geq p$.

As motivation for a Gaussian approximation to the projected nuisance term, consider the special case where the conditional distribution $\Pi(\alpha \mid S^{\mathrm{T}}y)$ is a product measure with uniformly bounded second moments. Under regularity assumptions on the columns of $M^{\mathrm{T}}Z$, the multivariate central limit theorem combined with the assumption $p \ll q$ implies that the distribution of the projection $M^{\mathrm{T}}Z\alpha$ is close to the Gaussian distribution with the same mean and covariance. By contrast, the unprojected $n$-dimensional nuisance term $\eta = Z\alpha$ can be very far from Gaussian, particularly if $n$ is of a similar order to $q$.

More realistically, one may envision settings where the entries of $\Pi(\alpha \mid S^{\mathrm{T}}y)$ are not independent but are weakly correlated on average. In this case, the usual central limit theorem does not hold because one can construct counterexamples in which the normalized sum of dependent but uncorrelated variables is far from Gaussian. Nevertheless, a classic result due to Diaconis & Freedman (1984) suggests that these counterexamples are atypical. Specifically, if one considers a weighted linear combination of the entries in $\alpha$, then approximate Gaussianity holds for most choices of the weights, where most is quantified with respect to the uniform measure on the sphere. The implications of this phenomenon have been studied extensively in the context of statistical inference (Hall & Li, 1993; Leeb, 2013), and Meckes (2012) and Reeves (2017) provide approximation bounds for the setting of multidimensional linear projections.

In the context of our approximation framework, these results imply that a Gaussian approximation is accurate for most, but not necessarily all, instances of the $p \times q$ feature matrix $M^{\mathrm{T}}Z$. To make this statement mathematically precise, we consider the expected behavior when the rows of $Z$ are drawn independently from the $q$-dimensional Gaussian distribution $N(0,\, \Lambda)$ where $\Lambda$ is positive definite. As in the rest of the paper, we assume that $X$ is fixed and arbitrary. Under these

assumptions, the rows of the projected matrices $M^{\mathrm{T}}Z$ and $S^{\mathrm{T}}Z$ are independent with the same distribution as in $Z$.

Our results depend on certain properties of the conditional distribution $\Pi(\alpha \mid S^{\mathrm{T}}y, S^{\mathrm{T}}Z)$. Let $\xi$ and $\Psi$ denote the mean and covariance of $\Pi(\alpha \mid S^{\mathrm{T}}y, S^{\mathrm{T}}Z)$, respectively. Define

$$m_1 = E\left\{ \left| \frac{\|\Lambda^{\frac{1}{2}}(\alpha - \xi)\|^2}{\mathrm{tr}(\Lambda\Psi)} - 1 \right| \mid S^{\mathrm{T}}y,\, S^{\mathrm{T}}Z \right\}, \qquad m_2 = \frac{\mathrm{tr}\{(\Lambda\Psi)^2\}}{\mathrm{tr}(\Lambda\Psi)^2}.$$

The term $m_1$ provides a measure of the concentration of $\|\Lambda^{1/2}(\alpha - \xi)\|^2$ about its mean and satisfies $0 \leq m_1 \leq 2$. The term $m_2$ provides a measure of the average correlation between the entries of $\Lambda^{1/2}\alpha$ and satisfies $1/q \leq m_2 \leq 1$ with equality on the left when $\Lambda\Psi$ is proportional to the identity matrix and equality on the right when $\Lambda\Psi$ has rank one.

Given estimates $\hat{\xi}$ and $\hat{\Psi}$ that are functions of $S^{\mathrm{T}}y$ and $S^{\mathrm{T}}Z$, we consider the Gaussian approximation

$$\hat{\Pi}(M^{\mathrm{T}}\eta \mid S^{\mathrm{T}}y, S^{\mathrm{T}}Z) = N\{M^{\mathrm{T}}Z\hat{\xi},\, \mathrm{tr}(\Lambda\hat{\Psi})I_p\}. \tag{7}$$

The covariance is chosen independently of $M^{\mathrm{T}}Z$ and depends only on a scalar summary of the estimated covariance. The following result bounds the accuracy of this approximation in terms of the terms $m_1$ and $m_2$ and the accuracy of the estimated mean and covariance.

THEOREM 2. *Conditional on any $S^{\mathrm{T}}y$ and $S^{\mathrm{T}}Z$, the Gaussian approximation in* (7) *satisfies*

$$E_{M^{\mathrm{T}}Z}\left[ D\left\{ \Pi(M^{\mathrm{T}}\eta \mid S^{\mathrm{T}}y, S^{\mathrm{T}}Z) * N_{\sigma^2} \parallel \hat{\Pi}(M^{\mathrm{T}}\eta \mid S^{\mathrm{T}}y, S^{\mathrm{T}}Z) * N_{\sigma^2} \right\} \right] \leq \delta_1 + \delta_2,$$

*where the expectation is with respect to $M^{\mathrm{T}}Z$ and*

$$\delta_1 = 3p\left[ m_1 \log\left\{ 1 + \frac{\mathrm{tr}(\Lambda\Psi)}{\sigma^2} \right\} + m_2^{\frac{1}{4}} + m_2^{\frac{1}{2}} \left\{ 1 + \frac{3\,\mathrm{tr}(\Lambda\Psi)}{\sigma^2} \right\}^{\frac{p}{4}} \right],$$

$$\delta_2 = \frac{p\,\|\Lambda^{\frac{1}{2}}(\xi - \hat{\xi})\|^2}{2\sigma^2} + \frac{p}{2\sigma^2}\left\{ \mathrm{tr}(\Lambda\Psi)^{\frac{1}{2}} - \mathrm{tr}(\Lambda\hat{\Psi})^{\frac{1}{2}} \right\}^2.$$

This result is meaningful when $p \ll q$ and the noise variance is non-negligible compared to the covariance of the nuisance term such the ratio $\mathrm{tr}(\Lambda\Psi)/\sigma^2$ is bounded from above. Then, $\delta_1$ converges to zero as $m_1$ and $m_2$ become small. The term $\delta_2$ quantifies the effect of mismatch between the first and second moments of $\Pi(\alpha \mid S^{\mathrm{T}}y, S^{\mathrm{T}}Z)$ and their approximations. The dependence on the second moments appears only in the terms $\mathrm{tr}(\Lambda\Psi)$ and $\mathrm{tr}(\Lambda\hat{\Psi})$. Thus, this bound can be small even if the approximation $\hat{\Psi}$ is very different from the true covariance $\Psi$.

To illustrate the significance of our results, consider two scaling regimes. First, if $n \ll q$ then the same arguments used in the proof of Theorem 2 can be used to show that the distribution of the $n$-dimensional nuisance term $\eta$ is also approximately Gaussian. Then, our approximation framework is well motivated, but does not differ fundamentally from existing approaches that apply a Laplace approximation directly on the unrotated data. The second, and more interesting, regime occurs when $n \approx q$ or $n \gg q$. Then, the $n$-dimensional nuisance term is non-Gaussian in general, because there exists a near isometry between $\eta$ and $\alpha$. Our approximation framework can provide significant gains by taking this non-Gaussianity into account when estimating the mean and covariance. Moreover, combining Theorems 1 and 2 provides an upper bound on the error of the approximation to the posterior of $\beta$ described in Algorithm 1. In particular, if the approximations of the mean and covariance are accurate enough, then this approximation error converges to zero as the terms $m_1$ and $m_2$ become small.

### 4·3. *Variable selection consistency*

Finally, we provide guarantees for variable selection consistency of (6), which only considers $\beta$ in contrast to § 3·2. Let the set $\gamma \subset \{1, \ldots, p\}$ contain all indices $j$ such that $\beta_j \neq 0$. Define $\gamma^0$ analogously for a non-random $\beta^0$. Variable selection consistency as in Fernández et al. (2001) and Liang et al. (2008) means that, for $y \sim N(X\beta^0 + \eta^0,\ \sigma^2 I_n)$, the posterior probability of the true model $\gamma^0$ converges to one, $\mathrm{pr}(\gamma = \gamma^0 \mid y) \to 1$ as $n \to \infty$ where $p$ does not change with $n$. It is desirable for a posterior approximation to inherit this property. Monte Carlo approximations do, but only if they are run for an infinite amount of time. Our approximation bypasses the need for such sampling, instead requiring mean and variance estimation for (5), while inheriting the consistency property if $\hat{\Pi}(M^{\mathrm{T}}\eta \mid S^{\mathrm{T}}y)$ concentrates appropriately. Relatedly, Ormerod et al. (2017) established such consistency for their variational Bayes algorithm. More recently, K. Ray and B. Szabó (arXiv:1904.07150) showed optimal convergence rates of variable selection using variational Bayes with different priors than we consider here.

Let the $|\gamma|$-dimensional vector $\beta_\gamma$ consist of the elements in $\beta$ with indices in $\gamma$, and the $n \times |\gamma|$ matrix $X_\gamma$ consist of the columns of $X$ with indices in $\gamma$. Then, specifying $\Pi(\gamma)$ and $\Pi(\beta_\gamma \mid \gamma)$ defines $\Pi(\beta)$. We consider $g$-priors (Zellner, 1986):

$$\beta_\gamma \mid \gamma \sim N\Big\{0,\ \sigma^2 g_n \big(X_\gamma^{\mathrm{T}} X_\gamma\big)^{-1}\Big\}, \qquad g_n \in (0,\ \infty). \tag{8}$$

Liang et al. (2008) showed variable selection consistency for priors of this form. Our approximation inherits this property under the additional assumption (9) on $\hat{\Pi}(M^{\mathrm{T}}\eta \mid S^{\mathrm{T}}y)$ and $g_n$. This is an assumption on $g_n$ and $\sigma^2$ jointly since $\hat{\Pi}(M^{\mathrm{T}}\eta \mid S^{\mathrm{T}}y)$ depends on $\sigma^2$. Otherwise, the sensitivity on $\sigma^2$ is limited since the property considers the asymptotic regime $n \to \infty$, when the signal-to-noise ratio goes to infinity regardless of $\sigma^2$.

THEOREM 3. *Let $\Pi(\beta_\gamma)$ be the g-prior on $\beta_\gamma$ from* (8)*. Assume that $g_n$ in* (8)*, $\Pi(\gamma)$, and $X$ satisfy $\mathrm{pr}(\gamma = \gamma^0) > 0$, $\lim_{n\to\infty} \|\{I_n - X_\gamma(X_\gamma^{\mathrm{T}}X_\gamma)^{-1}X_\gamma^{\mathrm{T}}\}X\beta^0\|/n > 0$ for any $\gamma$ not containing $\gamma^0$, $g_n \to \infty$, and $\log(g_n)/n \to 0$, which are standard assumptions used in Fernández et al. (2001) and Liang et al. (2008) as detailed in § S4 of the Supplementary Material. Let $y$ be distributed according to the data-generating model in* (1) *with $\beta$ and $\eta$ fixed to $\beta^0$ and $\eta^0$, respectively. Assume that $\hat{\Pi}(M^{\mathrm{T}}\eta \mid S^{\mathrm{T}}y)$ concentrates appropriately in that*

$$\frac{\|M^{\mathrm{T}}\eta - M^{\mathrm{T}}\eta^0\|^2}{\log g_n} \to 0, \tag{9}$$

*in probability with respect to $M^{\mathrm{T}}\eta \sim \hat{\Pi}(M^{\mathrm{T}}\eta \mid S^{\mathrm{T}}y)$ and $y$. Let $\hat{\Pi}(\beta \mid y)$ be as in* (4)*. Then, $\hat{\mathrm{pr}}(\gamma = \gamma^0 \mid y) \to 1$ in probability with respect to $y$ as $n \to \infty$.*

## 5. SIMULATION STUDIES AND APPLICATIONS

### 5·1. *Non-parametric adjustment for covariates*

Consider the set-up from § 3·3 with $g(a) = a^2$ and $q = 1$. We assign $f : \mathbb{R} \to \mathbb{R}$ a zero-mean Gaussian process prior with a squared exponential covariance function such that $\mathrm{cov}\{f(z_i),\ f(z_j)\} = \exp\{-(z_i - z_j)^2/10\}$ $(i, j = 1, \ldots, n)$, and $\beta \sim N(0, 16I_p)$. Set $n = 100$, $p = 3$, and $\sigma^2 = 1$. We draw the rows of $X$ independently from $N(0_{p\times 1}, \Phi)$ where $\Phi$ is a Toeplitz matrix defined so that its first row equals $(0.9^0, \ldots, 0.9^p)$. Then, the columns of $X$ are correlated. The features $z_i$ $(i = 1, \ldots, n)$ equal the $i$th element of the first column of $X$. Generate $y$ according to (1) with $f$ equal to a draw from its prior distribution and $\beta = (4, -4, 4)^{\mathrm{T}}$.

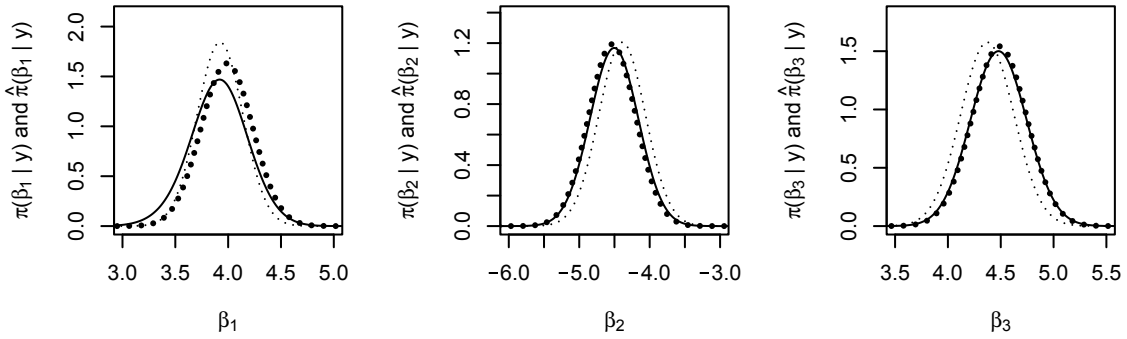


Fig. 1. Marginal posterior density estimates from the simulation in § 5·1 with the solid line representing the Gibbs estimate $\pi(\beta_j \mid y)$, the thick dotted line the estimate $\hat{\pi}(\beta_j \mid y)$ from Algorithm 1, and the thin dotted line the estimate resulting from ignoring the nuisance parameter.

We approximate the posterior $\Pi(\beta \mid y)$ using a random walk Metropolis-Hastings algorithm on $f$ with 10,000 burnin and 90,000 recorded iterations. We marginalize out $\beta$ since $\Pi(\beta \mid f, y)$ is analytically available, allowing approximation of $\pi(\beta \mid y)$ with samples from $\Pi(f \mid y)$. Algorithm 1 also provides $\hat{\Pi}(\beta \mid y)$ per § 3·3. Lastly, ignoring the non-parametric nuisance parameter by setting $\eta_i = (g \circ f)(z_i) = 0$ yields a simpler approximation. The Metropolis-Hastings algorithm took 6 minutes while our method finished in 2 seconds. The resulting posterior density estimates for $\beta_j$ $(j = 1, \ldots, p)$ are in Fig. 1. Taking the Metropolis-Hastings estimate as the gold standard, our method yields an approximation that matches the location and spread of the posterior better than the result from ignoring the non-parametric nuisance term $\eta$.

### 5·2. *Bayesian variable selection*

We consider the diabetes data from Efron et al. (2004) as it is a popular example of variable selection with collinear predictors (Park & Casella, 2008; Polson et al., 2013). The outcome $y$ measures disease progression one year after baseline for $n = 442$ patients with diabetes. The $r = 64$ predictors come from 10 covariates with their squares and interactions. The outcome and predictors are standardized to have zero mean and unit norm. Consider the variable selection set-up from § 3·2 with prior inclusion probability $\lambda = 1/2$ and $\psi = 1$. Usually, one would not use scalable approximations for such a moderate-dimensional problem as a Gibbs sampler can provide accurate estimates. The latter is why we include it here as these accurate estimates enable assessment of the approximation accuracy of scalable methods.

We estimate the posterior inclusion probabilities $\mathrm{pr}(\theta_j \neq 0 \mid y)$ $(j = 1, \ldots, r)$ using 1) a Gibbs sampler with 10,000 burnin and 90,000 recorded iterations, Algorithm 1 as described in § 3·2 using 2) vector approximate message passing and 3) the debiased lasso in Step 2 with $p = 4$ as suggested by $p = O(\log r)$ and parallelization across 8 CPU cores, 4) expectation propagation as in Hernández-Lobato et al. (2015), and 5) variational Bayes as in Carbonetto & Stephens (2012). To implement expectation propagation and variational Bayes, we used the R code from `https://jmhl.org/publications/` dated January 2010 and the R package 'varbvs' version 2.5-7, respectively. Results from the variational Bayes algorithm by Ormerod et al. (2017) are omitted as the method from Carbonetto & Stephens (2012) outperforms it in the scenarios that we consider. Since the error variance is unknown, we assign it the prior $1/\sigma^2 \sim \mathrm{Ga}(1, 1)$, a gamma distribution with unit shape and rate parameter. The Gibbs sampler incorporates this prior. Algorithm 1 estimates $\sigma^2$ as described in § 2·4, and § S5·3 and § S6 of the Supplementary Material. Expectation propagation estimates $\sigma^2$ by maximizing approximate ev-

Table 1. *Summary statistics of the absolute difference between the Gibbs sampler estimates and the approximations of the posterior log odds of inclusion for the application in* § 5·2 *with computation times. IRGA and VAMP stand for integrated rotated Gaussian approximation and vector approximate message passing, respectively.*

| Method | Min | Q1 | Median | Q3 | Max | Mean | Computation time (seconds) |
|---|---|---|---|---|---|---|---|
| IRGA with VAMP | 0.003 | 0.036 | 0.076 | 0.133 | 10.7 | 0.599 | 4.1 |
| IRGA with the debiased lasso | 0.003 | 0.100 | 0.142 | 0.199 | 7.85 | 0.470 | 3.8 |
| Expectation propagation | 0.003 | 0.061 | 0.109 | 0.168 | 11.9 | 0.666 | 0.8 |
| Variational Bayes | 0.002 | 0.093 | 0.124 | 0.166 | 11.6 | 0.667 | 1.0 |

idence (Hernández-Lobato et al., 2015). The R package 'varbvs' (Carbonetto & Stephens, 2012) uses approximate maximum likelihood for $\sigma^2$ within the variational Bayes method.

As discussed in § 3·2, determining whether posterior inclusion probabilities from a Gibbs sampler are accurate is non-trivial. Overlapping batch means (Flegal & Jones, 2010, § 3) estimates their average Monte Carlo standard error as 0.0015 in this application.

Table 1 focuses on the errors in the posterior inclusion probability estimates. An approximation error of $0.01$ is worse when the inclusion probability is $0.01$ versus $0.5$. We therefore transform the probabilities to log odds. Our method with vector approximate message passing outperforms expectation propagation and variational Bayes as its error is lowest in Table 1, though at a higher computational cost. Our method is slowest but still considerably faster than the Gibbs sampler which took 11 minutes to run. Since the debiased lasso yielded the worst approximation, we do not consider it in the remainder of this article.

### 5·3. *Controlling for single-nucleotide polymorphisms*

The Geuvadis dataset from Lappalainen et al. (2013), available at `https://www.ebi.ac.uk/Tools/geuvadis-das`, contains gene expression data from lymphoblastoid cell lines of $n = 462$ individuals from the 1000 Genomes Project along with roughly 38 million SNPs. We focus on the gene E2F2, ensemble ID ENSG00000007968, as it plays a key role in the cell cycle (Attwooll et al., 2004). Our focus is on assessing whether expression differs between populations, even after adjusting for genetic variation between individuals. Specifically, we compare people from British descent with the four other populations given in Table 2. If such differences occur, they can be presumed to be due to environmental factors that differ between these populations and that relate to E2F2 expression. We therefore consider the set-up from § 3·1 with $y$ the E2F2 gene expressions, $X$ demographic factors, and $Z$ containing SNPs we would like to control for.

The demographics in $X$ are gender and the 4 populations with British as the reference group. The matrix $X$ thus has $p = 5$ columns. The covariates $Z$ consist of $q = 10,000$ SNPs selected using sure independence screening (Fan & Lv, 2008) as vector approximate message passing on all 38 million SNPs was infeasible. We standardize $y$ and the columns of $X$ and $Z$ to have zero mean and unit variance. To complete the set-up from § 3·1, set $\lambda = n/(10\,q)$ and $\psi = 1/n$ for the spike-and-slab prior on $\alpha$ while $\Pi(\beta)$ is a spike-and-slab with prior inclusion probability $1/2$ and slab variance 1 such that, a priori, the SNPs do not capture more variation in the outcome than the demographic factors. This may provide a reasonable default for SNP data, but in other settings, hyperparameter values should be reconsidered. Vector approximate message passing estimates $\sigma^2$ using the prior $1/\sigma^2 \sim \text{Ga}(1, 1)$ and employs damping to achieve convergence in this application, as described in § S5·3 and § S5·4 of the Supplementary Material, respectively.

Table 2. *Posterior inclusion probabilities for the demographic factors from the application in* § 5·3. *IRGA stands for integrated rotated Gaussian approximation.*

| | | Population | | | |
|---|---|---|---|---|---|
| Method | Gender | Utahn of European ancestry | Finnish | Tuscan | Yoruba |
| IRGA | 0.83 | 0.96 | 0.96 | 0.92 | 0.00 |
| Ignoring the SNPs | 0.73 | 0.07 | 0.04 | 0.20 | 0.49 |

Table 2 contains the resulting posterior inclusion probabilities for the demographic factors, also when not controlling for the SNPs. The results vary hugely by whether SNPs are controlled for, with more evidence of a difference in the expression of gene E2F2 by population when controlling for SNPs using Algorithm 1. Section S10 of the Supplementary Material contains additional comparisons with other high-dimensional inference methods.

Section S8 of the Supplementary Material contains additional simulation studies. They further show that integrated rotated Gaussian approximation outperforms variational Bayes and either beats or is on par with expectation propagation in terms of approximation accuracy. This improved accuracy comes with increased computational cost for our method in certain scenarios.

## 6. DISCUSSION

Although our focus was Bayesian inference, our method marginalizes out nuisance parameters from the likelihood for $\beta$ as an intermediate step. This approximate likelihood from (6) can be useful in frequentist inference. It is well known that priors used in Bayesian inference correspond to penalties in frequentist inference. One can think of the log prior for the nuisance parameter $\eta$ as a penalty on $\eta$. An $L_2$ penalty might not be ideal due to the complex or high-dimensional nature of $\eta$. Instead, one might want to use sparsity-inducing penalties, such as $L_1$ or the non-convex smoothly clipped absolute deviation, which come with attractive theoretical properties (Pötscher & Leeb, 2009) but can be computationally challenging. Our method obtains the marginal likelihood for $\beta$ with such penalties on $\eta$, resolving the main computational bottleneck for frequentist inference on $\beta$ in the model of interest in (1).

## ACKNOWLEDGMENT

This work was partially supported by the National Institute of Environmental Health Sciences of the U.S. National Institutes of Health, the Singapore Ministry of Education Academic Research Fund, and the Laboratory for Analytic Sciences. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the Laboratory for Analytic Sciences and/or any agency or entity of the United States Government.

## SUPPLEMENTARY MATERIAL

Supplementary material available at *Biometrika* online includes the proofs for § 4, a corollary to Theorem 1, details of vector approximate message passing and the Laplace approximation for § 3·3, and additional simulation studies. The R code for the numerical results is available at `https://github.com/willemvandenboom/IRGA`.

## REFERENCES


ATTWOOLL, C., DENCHI, E. L. & HELIN, K. (2004). The E2F family: Specific functions and overlapping interests. *The EMBO Journal* **23**, 4709–4716.

BERGER, J. O., LISEO, B. & WOLPERT, R. L. (1999). Integrated likelihood methods for eliminating nuisance parameters. *Statistical Science* **14**, 1–28.

CARBONETTO, P. & STEPHENS, M. (2012). Scalable variational inference for Bayesian variable selection in regression, and its accuracy in genetic association studies. *Bayesian Anal.* **7**, 73–108.

DIACONIS, P. & FREEDMAN, D. (1984). Asymptotics of graphical projection pursuit. *Ann. Stat.* **12**, 793–815.

EFRON, B., HASTIE, T., JOHNSTONE, I. & TIBSHIRANI, R. (2004). Least angle regression. *Ann. Stat.* **32**, 407–499.

FAN, J. & LV, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *J. R. Statist. Soc.* B **70**, 849–911.

FERNÁNDEZ, C., LEY, E. & STEEL, M. F. (2001). Benchmark priors for Bayesian model averaging. *J. Econom.* **100**, 381–427.

FLEGAL, J. M. & JONES, G. L. (2010). Batch means and spectral variance estimators in Markov chain Monte Carlo. *Ann. Stat.* **38**, 1034–1070.

GEORGE, E. I. & MCCULLOCH, R. E. (1993). Variable selection via Gibbs sampling. *J. Am. Statist. Assoc.* **88**, 881–889.

GEORGE, E. I. & MCCULLOCH, R. E. (1997). Approaches for Bayesian variable selection. *Stat. Sin.* **7**, 339–374.

GOLUB, G. H. & VAN LOAN, C. F. (1996). *Matrix Computations*. Baltimore: Johns Hopkins University Press, 3rd ed.

HALL, P. & LI, K.-C. (1993). On almost linearity of low dimensional projections from high dimensional data. *Ann. Stat.* **21**, 867–889.

HERNÁNDEZ-LOBATO, J. M., HERNÁNDEZ-LOBATO, D. & SUÁREZ, A. (2015). Expectation propagation in linear regression models with spike-and-slab priors. *Mach. Learn.* **99**, 437–487.

HUGGINS, J., ADAMS, R. P. & BRODERICK, T. (2017). PASS-GLM: Polynomial approximate sufficient statistics for scalable Bayesian GLM inference. In *Advances in Neural Information Processing Systems 30*. pp. 3611–3621.

JAVANMARD, A. & MONTANARI, A. (2013). Confidence intervals and hypothesis testing for high-dimensional statistical models. In *Advances in Neural Information Processing Systems 26*. pp. 1187–1195.

LAPPALAINEN, T., , SAMMETH, M., FRIEDLÄNDER, M. R., 'T HOEN, P. A. C., MONLONG, J., RIVAS, M. A., GONZÀLEZ-PORTA, M., KURBATOVA, N., GRIEBEL, T., FERREIRA, P. G., BARANN, M., WIELAND, T., GREGER, L., VAN ITERSON, M., ALMLÖF, J., RIBECA, P., PULYAKHINA, I., ESSER, D., GIGER, T., TIKHONOV, A., SULTAN, M., BERTIER, G., MACARTHUR, D. G., LEK, M., LIZANO, E., BUERMANS, H. P. J., PADIOLEAU, I., SCHWARZMAYR, T., KARLBERG, O., ONGEN, H., KILPINEN, H., BELTRAN, S., GUT, M., KAHLEM, K., AMSTISLAVSKIY, V., STEGLE, O., PIRINEN, M., MONTGOMERY, S. B., DONNELLY, P., MCCARTHY, M. I., FLICEK, P., STROM, T. M., LEHRACH, H., SCHREIBER, S., SUDBRAK, R., CARRACEDO, Á., ANTONARAKIS, S. E., HÄSLER, R., SYVÄNEN, A.-C., VAN OMMEN, G.-J., BRAZMA, A., MEITINGER, T., ROSENSTIEL, P., GUIGÓ, R., GUT, I. G., ESTIVILL, X. & DERMITZAKIS, E. T. (2013). Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**, 506–511.

LEEB, H. (2013). On the conditional distributions of low-dimensional projections from high-dimensional data. *Ann. Stat.* **41**, 464–483.

LIANG, F., PAULO, R., MOLINA, G., CLYDE, M. A. & BERGER, J. O. (2008). Mixtures of $g$ priors for Bayesian variable selection. *J. Am. Statist. Assoc.* **103**, 410–423.

MECKES, E. (2012). Projections of probability distributions: A measure-theoretic Dvoretzky theorem. In *Lecture Notes in Mathematics*. Berlin: Springer, pp. 317–326.

O'HARA, R. B. & SILLANPÄÄ, M. J. (2009). A review of Bayesian variable selection methods: What, how and which. *Bayesian Anal.* **4**, 85–117.

OPPER, M. & WINTHER, O. (2005). Expectation consistent approximate inference. *J. Mach. Learn. Res.* **6**, 2177–2204.

ORMEROD, J. T., YOU, C. & MÜLLER, S. (2017). A variational Bayes approach to variable selection. *Electron. J. of Stat.* **11**, 3549–3594.

PARK, T. & CASELLA, G. (2008). The Bayesian lasso. *J. Am. Statist. Assoc.* **103**, 681–686.

POLSON, N. G., SCOTT, J. G. & WINDLE, J. (2013). The Bayesian bridge. *J. R. Statist. Soc.* B **76**, 713–733.

PÖTSCHER, B. M. & LEEB, H. (2009). On the distribution of penalized maximum likelihood estimators: The LASSO, SCAD, and thresholding. *J. of Multivar. Anal.* **100**, 2065–2082.

RANGAN, S., SCHNITER, P. & FLETCHER, A. K. (2017). Vector approximate message passing. In *IEEE International Symposium on Information Theory*. pp. 1588–1592.

REEVES, G. (2017). Conditional central limit theorems for Gaussian projections. In *IEEE International Symposium on Information Theory*. pp. 3045–3049.

RUE, H., MARTINO, S. & CHOPIN, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested laplace approximations. *J. R. Statist. Soc.* B **71**, 319–392.

SEVERINI, T. A. (2011). Frequency properties of inferences based on an integrated likelihood function. *Stat. Sin.* **21**, 433–447.

SONG, Q. & LIANG, F. (2014). A split-and-merge Bayesian variable selection approach for ultrahigh dimensional regression. *J. R. Statist. Soc.* B **77**, 947–972.

STACHURSKI, J. (2016). *A Primer in Econometric Theory*. Cambridge: MIT Press.

STEINBERG, D. M. & BONILLA, E. V. (2014). Extended and unscented Gaussian processes. In *Advances in Neural Information Processing Systems 27*. pp. 1251–1259.

TIERNEY, L. & KADANE, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *J. Am. Statist. Assoc.* **81**, 82–86.

VAN DEN BOOM, W., DUNSON, D. & REEVES, G. (2015). Quantifying uncertainty in variable selection with arbitrary matrices. In *2015 IEEE 6th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*. pp. 385–388.

ZELLNER, A. (1986). On assessing prior distributions and Bayesian regression analysis with $g$-prior distributions. In *Bayesian inference and decision techniques: Essays in Honor of Bruno de Finetti*, P. K. Goel & A. Zellner, eds. Amsterdam: North-Holland/Elsevier, pp. 233–243.

*Biometrika* (2020), **xx**, x, *pp*. 1–21
*Printed in Great Britain*

# Supplementary material for Approximating posteriors with high-dimensional nuisance parameters via integrated rotated Gaussian approximation

BY W. VAN DEN BOOM

*Yale-NUS College, National University of Singapore, 16 College Avenue West #01-220, Singapore 138527, Singapore*
willem@yale-nus.edu.sg

G. REEVES AND D. B. DUNSON

*Department of Statistical Science, Duke University, Box 90251, Durham, North Carolina 27708, U.S.A.*
galen.reeves@duke.edu dunson@duke.edu

## S1. PROOF OF THEOREM 1

LEMMA S1. *Let $P(a,b)$ and $Q(a,b)$ be probability measures defined on the same space that have the same $a$-marginal, that is, $P(a) = Q(a)$. Then,*

$$E_{P(b)}[D\{P(a \mid b) \parallel Q(a \mid b)\}] \leq E_{P(a)}[D\{P(b \mid a) \parallel Q(b \mid a)\}].$$

*Proof.* Using the chain rule for Kullback-Leibler divergence (Cover & Thomas, 2006, Theorem 2.5.3) two different ways leads to

$$\begin{aligned} D\{P(a,b) \parallel Q(a,b)\} &= E_{P(b)}[D\{P(a \mid b) \parallel Q(a \mid b)\}] + D\{P(b) \parallel Q(b)\} \\ &= E_{P(a)}[D\{P(b \mid a) \parallel Q(b \mid a)\}] + D\{P(a) \parallel Q(a)\}. \end{aligned}$$

Hence, the desired result follows from the fact that $D\{P(b) \parallel Q(b)\}$ is non-negative, and the assumption $P(a) = Q(a)$ which implies that $D\{P(a) \parallel Q(a)\} = 0$. □

*Proof of Theorem* 1. The distributions $\Pi(\beta, M^{\mathrm{T}}y \mid S^{\mathrm{T}}y) = \Pi(M^{\mathrm{T}}y \mid S^{\mathrm{T}}y, \beta)\,\Pi(\beta \mid S^{\mathrm{T}}y)$ and $\hat{\Pi}(\beta, M^{\mathrm{T}}y \mid S^{\mathrm{T}}y) = \hat{\Pi}(M^{\mathrm{T}}y \mid S^{\mathrm{T}}y, \beta)\,\Pi(\beta \mid S^{\mathrm{T}}y)$ have the same $\beta$-marginal $\Pi(\beta \mid S^{\mathrm{T}}y)$. Hence, we can apply Lemma S1 with $P(a,b) = \Pi(\beta, M^{\mathrm{T}}y \mid S^{\mathrm{T}}y)$ and $Q(a,b) = \hat{\Pi}(\beta, M^{\mathrm{T}}y \mid S^{\mathrm{T}}y)$:

$$\begin{aligned} &E\Big[D\Big\{\Pi(\beta \mid y) \parallel \hat{\Pi}(\beta \mid y)\Big\} \;\Big|\; S^{\mathrm{T}}y\Big] \\ &\qquad = E_{\Pi(M^{\mathrm{T}}y \mid S^{\mathrm{T}}y)}\Big[D\Big\{\Pi(\beta \mid M^{\mathrm{T}}y, S^{\mathrm{T}}y) \parallel \hat{\Pi}(\beta \mid M^{\mathrm{T}}y, S^{\mathrm{T}}y)\Big\}\Big] \\ &\qquad \leq E_{\Pi(\beta \mid S^{\mathrm{T}}y)}\Big[D\Big\{\Pi(M^{\mathrm{T}}y \mid \beta, S^{\mathrm{T}}y) \parallel \hat{\Pi}(M^{\mathrm{T}}y \mid \beta, S^{\mathrm{T}}y)\Big\}\Big], \end{aligned}$$

Let $\Pi(a) * \Pi(b)$ denote the distribution of $a + b$. Then, (3a) provides

$$\begin{aligned} \Pi(M^{\mathrm{T}}y \mid \beta, S^{\mathrm{T}}y) &= \Pi(M^{\mathrm{T}}\eta \mid S^{\mathrm{T}}y) * N(M^{\mathrm{T}}X\beta,\, \sigma^2 I_p), \\ \hat{\Pi}(M^{\mathrm{T}}y \mid \beta, S^{\mathrm{T}}y) &= \hat{\Pi}(M^{\mathrm{T}}\eta \mid S^{\mathrm{T}}y) * N(M^{\mathrm{T}}X\beta,\, \sigma^2 I_p). \end{aligned}$$

Combining the last two displays yields

$$E\Big[D\Big\{\Pi(\beta \mid y) \parallel \hat{\Pi}(\beta \mid y)\Big\} \;\Big|\; S^{\mathrm{T}}y\Big] \leq$$
$$E_{\Pi(\beta \mid S^{\mathrm{T}}y)}\Big[D\Big\{\Pi(M^{\mathrm{T}}\eta \mid S^{\mathrm{T}}y) * N(M^{\mathrm{T}}X\beta,\, \sigma^2 I_p) \parallel \hat{\Pi}(M^{\mathrm{T}}\eta \mid S^{\mathrm{T}}y) * N(M^{\mathrm{T}}X\beta,\, \sigma^2 I_p)\Big\}\Big].$$

Since the Kullback-Leibler divergence is invariant to one-to-one transformations (Kullback & Leibler, 1951, Corollary 4.1), the Kullback-Leibler divergence is constant with respect to $M^{\mathrm{T}}X\beta$. The required result follows from setting $M^{\mathrm{T}}X\beta$ equal to zero and dropping the expectation in the right-hand side of the last display. □

## S2. COROLLARY TO THEOREM 1

Theorem 1 considered how close our approximation $\hat{\Pi}(\beta \mid y)$ is to the posterior $\Pi(\beta \mid y)$. Alternatively, one may be interested in a scenario where the nuisance parameter $\eta$ equals $\eta^0$, and one would like to do inference without interference from the nuisance term using $\Pi(\beta \mid y, \eta^0)$, even though $\eta^0$ is unknown.

Define the squared quadratic Wasserstein distance between the distributions $\Pi(a)$ and $\Pi(b)$ as $W_2^2\{\Pi(a),\, \Pi(b)\} = \inf E(\|a - b\|^2)$ where $\|\cdot\|$ denotes the Euclidean norm and the infimum is over all joint distributions on $(a, b)$ such that $a \sim \Pi(a)$ and $b \sim \Pi(b)$.

LEMMA S2. *Let $P$ and $Q$ be distributions on $\mathbb{R}^p$. For any $\sigma^2 > 0$,*

$$D\big\{P * N(0,\, \sigma^2 I_p) \parallel Q * N(0,\, \sigma^2 I_p)\big\} \leq \frac{1}{2\sigma^2}\, W_2^2(P,\, Q).$$

*Proof.* Let $\Pi(a, b)$ be any coupling on $\mathbb{R}^p \times \mathbb{R}^p$ satisfying the marginal constraints $\Pi(a) = P(a)$ and $\Pi(b) = Q(b)$. By the convexity of Kullback-Leibler divergence (Cover & Thomas, 2006, Theorem 2.7.2), Jensen's inequality provides

$$\begin{aligned} D\big\{P * N(0,\, \sigma^2 I_p) \parallel Q * N(0,\, \sigma^2 I_p)\big\} &\leq E_{\Pi(a,b)}\big[D\big\{N(a,\, \sigma^2 I_p)] \parallel N(b,\, \sigma^2 I_d)\big\}\big] \\ &= \frac{1}{2\sigma^2}\, E_{\Pi(a,b)}\big(\|a - b\|^2\big), \end{aligned}$$

where the equality follows from inserting the Gaussian densities into the definition of the Kullback-Leibler divergence. Recalling the definition of the quadratic Wasserstein distance and choosing the infimum over all couplings $\Pi(a, b)$ of $P$ and $Q$ gives the stated result. □

COROLLARY S1. *Let $\hat{\Pi}(\beta \mid y)$ be as in* (4)*. Let $y$ be distributed according to the data-generating model in* (1) *with $\beta \sim \Pi(\beta)$ distributed according to its prior and $\eta$ fixed to $\eta^0$. Then,*

$$E\Big[D\Big\{\Pi(\beta \mid y, \eta^0) \parallel \hat{\Pi}(\beta \mid y)\Big\} \;\Big|\; S^{\mathrm{T}}y\Big] \leq \frac{1}{2\sigma^2}\, E_{\hat{\Pi}(M^{\mathrm{T}}\eta \mid S^{\mathrm{T}}y)}\Big(\big\|M^{\mathrm{T}}\eta^0 - M^{\mathrm{T}}\eta\big\|^2 \;\Big|\; S^{\mathrm{T}}y\Big).$$

*In particular, under the Gaussian approximation $\hat{\Pi}(M^{\mathrm{T}}\eta \mid S^{\mathrm{T}}y)$ from* (5)*,*

$$E\Big[D\Big\{\Pi(\beta \mid y, \eta^0) \parallel \hat{\Pi}(\beta \mid y)\Big\} \;\Big|\; S^{\mathrm{T}}y\Big] \leq \frac{1}{2\sigma^2}\, \Big\{\big\|M^{\mathrm{T}}\eta^0 - \hat{\mu}\big\|^2 + \mathrm{tr}(\hat{\Sigma})\Big\}.$$

*Proof.* Evaluating Theorem 1 with Lemma S2, $\Pi(\eta) = \delta(\eta^0)$, a point mass at $\eta^0$, and recalling the definition of the quadratic Wasserstein distance provides the first inequality. For the second equality, (5) provides $M^{\mathrm{T}}\eta^0 - M^{\mathrm{T}}\eta \mid S^{\mathrm{T}}y \sim \mathcal{N}(M^{\mathrm{T}}\eta^0 - \hat{\mu},\, \hat{\Sigma})$. Evaluating the right-hand side of the first inequality with this distribution provides the second inequality. □

Corollary S1 links two different quantities of interest. The left-hand side is the difference between our approximation $\hat{\Pi}(\beta \mid y)$ and the exact posterior $\Pi(\beta \mid y, \eta^0)$. The right-hand side involves the average squared deviation of the distribution $\hat{\Pi}(M^{\mathrm{T}}\eta \mid S^{\mathrm{T}}y)$ from $M^{\mathrm{T}}\eta^0$. This deviation can be small while the average squared deviation of $\Pi(\eta \mid S^{\mathrm{T}}y)$ from $\eta^0$ is large: The $n$-dimensional $\eta$ can have a potentially high-dimensional distribution while the $p$-dimensional term $M^{\mathrm{T}}\eta$ is a projection onto the low-dimensional column space of $M$. In Corollary S1, $y$ is distributed according to (1) with $\beta \sim \Pi(\beta)$ while $\eta$ is fixed to $\eta^0$. That $\beta$ and $\eta$ are treated differently is a result of their different treatment in Algorithm 1.

Consider asymptotic analysis where, for a sequence of instances of (1), $n \to \infty$ and interest is in the properties of $\hat{\Pi}(\beta \mid y)$ as $n \to \infty$. If $\hat{\Pi}(M^{\mathrm{T}}\eta \mid S^{\mathrm{T}}y)$ contracts around the value $M^{\mathrm{T}}\eta^0$ as $n \to \infty$, Corollary S1 shows that the posterior approximation from our method converges to the posterior $\Pi(\beta \mid y, \eta^0)$ based on the likelihood from (1) with $\eta$ equal to $\eta^0$. This convergence is in terms of Kullback-Leibler divergence which bounds dissimilarity measures commonly used in asymptotic analyses of Bayesian posteriors. For instance, Bernstein-von Mises theorems often use total variation distance (Bontemps, 2011) which Pinsker's inequality bounds by the square root of the Kullback-Leibler divergence. The finite-sample analysis of Corollary S1 therefore gives rise to asymptotic properties of the approximate posterior $\hat{\Pi}(\beta \mid y)$ if $E_{\hat{\Pi}(M^{\mathrm{T}}\eta \mid S^{\mathrm{T}}y)}(\left\|M^{\mathrm{T}}\eta^0 - M^{\mathrm{T}}\eta\right\|^2 \mid S^{\mathrm{T}}y) \to 0$. Such asymptotic results for $\hat{\Pi}(\beta \mid y)$ differ from usual Bayesian asymptotics due to the set-up of Corollary S1: The data-generating process involves $\beta \sim \Pi(\beta)$ rather than fixing $\beta$ to a value. By contrast, $\eta$ is fixed to $\eta^0$ in the data-generating process of Corollary S1 rather than distributed according to its prior.

## S3. Proof of Theorem 2

To simplify notation, define $a = q^{1/2}\Lambda^{1/2}(\alpha - \xi)$, $b_a = q^{1/2}\Lambda^{1/2}(\hat{\xi} - \xi)$, and $H = q^{-1/2}M^{\mathrm{T}}Z\Lambda^{-1/2}$ such that the entries of $H$ are independent with distribution $N(0,\ 1/q)$ and $Ha = M^{\mathrm{T}}Z(\alpha - \xi)$. Also,

$$\begin{aligned}
\Delta &= D(\Pi(Ha \mid S^{\mathrm{T}}y, S^{\mathrm{T}}Z) * N_{\sigma^2} \parallel N[Hb_a,\ \{\mathrm{tr}(\Lambda\hat{\Psi}) + \sigma^2\}I_p]),\\
\Delta_1 &= D(\Pi(Ha \mid S^{\mathrm{T}}y, S^{\mathrm{T}}Z) * N_{\sigma^2} \parallel N[0,\ \{\mathrm{tr}(\Lambda\Psi) + \sigma^2\}I_p]),\\
\Delta_2 &= D(N[0,\ \{\mathrm{tr}(\Lambda\Psi) + \sigma^2\}I_p] \parallel N[Hb_a,\ \{\mathrm{tr}(\Lambda\hat{\Psi}) + \sigma^2\}I_p]).
\end{aligned}$$

Here, $N[Hb_a,\ \{\mathrm{tr}(\Lambda\hat{\Psi}) + \sigma^2\}I_p]$ is a shifted version of the Gaussian approximation in (7). We will show that $\Delta$ equals the divergence in Theorem 2. $\Delta_1$ is the Kullback-Leibler divergence from the target distribution to the Gaussian approximation evaluated with the true and approximated mean and covariance. $\Delta_2$ depends on the mismatch in the estimates $\hat{\xi}$, captured by $b_a$, and $\hat{\Psi}$.

Lemma S3. *Conditional on any $S^{\mathrm{T}}y$ and $S^{\mathrm{T}}Z$, $E_H(\Delta_1) \leq \delta_1$ with $\delta_1$ as in Theorem 2.*

*Proof.* Since $E(Ha \mid S^{\mathrm{T}}y, S^{\mathrm{T}}Z) = 0$, $\mathrm{cov}(Ha \mid S^{\mathrm{T}}y, S^{\mathrm{T}}Z) = E(Haa^{\mathrm{T}}H^{\mathrm{T}})$, where we drop the condition on $S^{\mathrm{T}}y$ and $S^{\mathrm{T}}Z$ for notation convenience. By the law of total expectation, $E(Haa^{\mathrm{T}}H^{\mathrm{T}}) = E\{E(Haa^{\mathrm{T}}H^{\mathrm{T}} \mid H)\} = E\{H\mathrm{cov}(a)H^{\mathrm{T}}\}$. Inserting the definition of $a$ and recalling $\mathrm{cov}(\alpha) = \Psi$ yields $E\{H\mathrm{cov}(a)H^{\mathrm{T}}\} = E(Hq\Lambda^{1/2}\Psi\Lambda^{1/2}H^{\mathrm{T}})$. Since $E(H_{ij}H_{kl})$ equals $1/q$ if $(i,j) = (k,l)$ and 0 otherwise, $E(Hq\Lambda^{1/2}\Psi\Lambda^{1/2}H^{\mathrm{T}}) = \mathrm{tr}(q\Lambda^{1/2}\Psi\Lambda^{1/2})I_p/q$. The cyclic property of the trace now provides $\mathrm{cov}(Ha \mid S^{\mathrm{T}}y, S^{\mathrm{T}}Z) = \mathrm{tr}(q\Lambda^{1/2}\Psi\Lambda^{1/2})I_p/q = \mathrm{tr}(\Lambda\Psi)I_p$. Thus, the mean and covariance of both distributions in $\Delta_1$ are matched. Therefore, Theorem 2

from Reeves (2017) evaluated with $\epsilon = 1$ and $C = 3$ yields

$$E_H(\Delta_1) \leq 3p \log\left\{1 + \frac{\frac{1}{q}E(\|a\|^2)}{\sigma^2}\right\} \frac{\frac{1}{q}E\{|\|a\|^2 - E(\|a\|^2)|\}}{\frac{1}{q}E(\|a\|^2)} + 3p^{\frac{3}{4}}\left\{\frac{\frac{1}{q}E(|a^{\mathrm{T}}a'|)}{\frac{1}{q}E(\|a\|^2)}\right\}^{\frac{1}{2}} + 3p^{\frac{1}{4}}\left\{1 + \frac{\frac{3}{q}E(\|a\|^2)}{\sigma^2}\right\}^{\frac{p}{4}} \frac{\frac{1}{q}E(|a^{\mathrm{T}}a'|^2)^{\frac{1}{2}}}{\frac{1}{q}E(\|a\|^2)}, \quad \text{(S1)}$$

where $a'$ is an independent copy of $a$. The remainder of this proof is simplifying this bound.

Since $E(a) = 0$ and $\operatorname{cov}(a) = q\Lambda^{1/2}\Psi\Lambda^{1/2}$,

$$\begin{aligned} q^{-2}E(|a^{\mathrm{T}}a'|^2) &= q^{-2}E(a^{\mathrm{T}}a'a'^{\mathrm{T}}a) = q^{-2}\operatorname{tr}\{E(aa^{\mathrm{T}}a'a'^{\mathrm{T}})\} \\ &= q^{-2}\operatorname{tr}\{\operatorname{cov}(a)^2\} = \operatorname{tr}(\Lambda^{\frac{1}{2}}\Psi\Lambda\Psi\Lambda^{\frac{1}{2}}) = \operatorname{tr}\{(\Lambda\Psi)^2\}, \end{aligned}$$

and

$$\frac{1}{q}E(\|a\|^2) = \frac{1}{q}\operatorname{tr}\{\operatorname{cov}(a)\} = \frac{1}{q}\operatorname{tr}(q\Lambda^{\frac{1}{2}}\Psi\Lambda^{\frac{1}{2}}) = \operatorname{tr}(\Lambda\Psi).$$

Therefore,

$$\frac{\frac{1}{q}E\{|\|a\|^2 - E(\|a\|^2)|\}}{\frac{1}{q}E(\|a\|^2)} = E\left\{\left|\frac{\|a\|^2}{E(\|a\|^2)} - 1\right|\right\} = E\left\{\left|\frac{\|q^{\frac{1}{2}}\Lambda^{\frac{1}{2}}(\alpha - \xi)\|^2}{q\operatorname{tr}(\Lambda\Psi)} - 1\right|\right\} = m_1,$$

and, by Jensen's inequality,

$$\left\{\frac{\frac{1}{q}E(|a^{\mathrm{T}}a'|)}{\frac{1}{q}E(\|a\|^2)}\right\}^2 \leq \left\{\frac{\frac{1}{q}E(|a^{\mathrm{T}}a'|^2)^{\frac{1}{2}}}{\frac{1}{q}E(\|a\|^2)}\right\}^2 = \frac{q^{-2}E(|a^{\mathrm{T}}a'|^2)}{\operatorname{tr}(\Lambda\Psi)^2} = m_2.$$

Inserting the last three displays and $p^{1/4} \leq p^{3/4} \leq p$ into (S1) provides the required result. □

LEMMA S4. *Conditional on any $S^{\mathrm{T}}y$ and $S^{\mathrm{T}}Z$, $E_H(\Delta_2) \leq \delta_2$ with $\delta_2$ as in Theorem 2.*

*Proof.* Combining (7), Lemma S2, and the evaluation of the quadratic Wasserstein distance between two Gaussians from Dowson & Landau (1982) yields

$$\begin{aligned} \Delta_2 &\leq \frac{1}{2\sigma^2}\Big[\|Hb_a\|^2 + \operatorname{tr}\Big\{\operatorname{tr}(\Lambda\Psi)I_p + \operatorname{tr}(\Lambda\hat{\Psi})I_p - 2\operatorname{tr}(\Lambda\hat{\Psi})^{\frac{1}{2}}\operatorname{tr}(\Lambda\Psi)^{\frac{1}{2}}I_p\Big\}\Big] \\ &= \frac{1}{2\sigma^2}\Big[\|Hb_a\|^2 + p\{(\Lambda\Psi)^{\frac{1}{2}} - (\Lambda\hat{\Psi})^{\frac{1}{2}}\}^2\Big]. \end{aligned} \quad \text{(S2)}$$

Recalling $b_a = q^{1/2}\Lambda^{1/2}(\hat{\xi} - \xi)$,

$$\begin{aligned} E_H\big(\|Hb_a\|^2\big) &= E_H\Big\{\|q^{\frac{1}{2}}H\Lambda^{\frac{1}{2}}(\xi - \hat{\xi})\|^2\Big\} \\ &= q\,\{\Lambda^{\frac{1}{2}}(\xi - \hat{\xi})\}^{\mathrm{T}}E_H(H^{\mathrm{T}}H)\,\Lambda^{\frac{1}{2}}(\xi - \hat{\xi}) = p\,\|\Lambda^{\frac{1}{2}}(\xi - \hat{\xi})\|^2, \end{aligned}$$

where the last equality follows from $E(H^{\mathrm{T}}H) = pI_q/q$. Taking the expectation of (S2) with respect to $H$ and inserting the last display yields the required result. □

*Proof of Theorem 2.* Let $\pi_0$ denote the density function of $\Pi(Ha \mid S^{\mathrm{T}}y, S^{\mathrm{T}}Z) * N_{\sigma^2}$ and let $E_0(\cdot)$ denote the expectation with respect to this distribution. Let $\upsilon \sim \pi_0$. By the definition of

the Kullback-Leibler divergence,

$$\begin{aligned}\Delta &= E_0\left\{\log\left(\frac{\pi_0(\upsilon)}{N[\upsilon \mid Hb_a,\, \{\mathrm{tr}(\Lambda\hat{\Psi})+\sigma^2\}I_p]}\right)\right\} \\ &= \underbrace{E_0\left\{\log\left(\frac{\pi_0(\upsilon)}{N[\upsilon \mid 0,\, \{\mathrm{tr}(\Lambda\Psi)+\sigma^2\}I_p]}\right)\right\}}_{\Delta_1} + E_0\left\{\log\left(\frac{N[\upsilon \mid 0,\, \{\mathrm{tr}(\Lambda\Psi)+\sigma^2\}I_p]}{N[\upsilon \mid Hb_a,\, \{\mathrm{tr}(\Lambda\hat{\Psi})+\sigma^2\}I_p]}\right)\right\}.\end{aligned}$$

Taking the expectation with respect to $H$ yields

$$E_H(\Delta) = E_H(\Delta_1) + E_H\left[E_0\left\{\log\left(\frac{N[\upsilon \mid 0,\, \{\mathrm{tr}(\Lambda\Psi)+\sigma^2\}I_p]}{N[\upsilon \mid Hb_a,\, \{\mathrm{tr}(\Lambda\hat{\Psi})+\sigma^2\}I_p]}\right)\right\}\right]. \tag{S3}$$

Denote the expectation with respect to $\upsilon \sim N[0,\, \{\mathrm{tr}(\Lambda\Psi)+\sigma^2\}I_p]$ by $E_2(\cdot)$. The mean and covariance of $E_H\{\Pi(Ha \mid S^{\mathrm{T}}y, S^{\mathrm{T}}Z) * N_{\sigma^2}\}$ and $N[0,\, \{\mathrm{tr}(\Lambda\Psi)+\sigma^2\}I_p]$ are the same as confirmed in the proof of Lemma S3, and the expectation of the logarithm of the Gaussian density only depends on the mean and covariance of $\upsilon$. Therefore,

$$E_H\big[E_0\big\{\log\big(N[\upsilon \mid 0,\, \{\mathrm{tr}(\Lambda\Psi)+\sigma^2\}I_p]\big)\big\}\big] = E_2\big\{\log\big(N[\upsilon \mid 0,\, \{\mathrm{tr}(\Lambda\Psi)+\sigma^2\}I_p]\big)\big\}. \tag{S4}$$

Also, expanding the square inside the Gaussian density and noting $E_0(\upsilon) = 0$ yields

$$\begin{aligned}&E_H\left[E_0\left\{\log\Big(N[\upsilon \mid Hb_a,\, \{\mathrm{tr}(\Lambda\hat{\Psi})+\sigma^2\}I_p]\Big)\right\}\right] \\ &\qquad = E_H\left[E_0\left\{\log\Big(N[\upsilon \mid 0,\, \{\mathrm{tr}(\Lambda\hat{\Psi})+\sigma^2\}I_p]\Big) + \frac{\|Hb_a\|^2}{2\{\mathrm{tr}(\Lambda\hat{\Psi})+\sigma^2\}}\right\}\right].\end{aligned}$$

Again using that the logarithm of a Gaussian density only depends on the mean and covariance of $\nu$ provides

$$\begin{aligned}&E_H\left[E_0\left\{\log\Big(N[\upsilon \mid Hb_a,\, \{\mathrm{tr}(\Lambda\hat{\Psi})+\sigma^2\}I_p]\Big)\right\}\right] \\ &\qquad = E_H\left[E_2\left\{\log\Big(N[\upsilon \mid 0,\, \{\mathrm{tr}(\Lambda\hat{\Psi})+\sigma^2\}I_p]\Big) + \frac{\|Hb_a\|^2}{2\{\mathrm{tr}(\Lambda\hat{\Psi})+\sigma^2\}}\right\}\right] \\ &\qquad = E_H\left[E_2\left\{\log\Big(N[\upsilon \mid Hb_a,\, \{\mathrm{tr}(\Lambda\hat{\Psi})+\sigma^2\}I_p]\Big)\right\}\right],\end{aligned}$$

where the last equality follows from completing the square and $E_2(\upsilon) = 0$. Inserting the last display and (S4) into (S3), and recalling the definition of the Kullback-Leibler divergence shows

$$E_H(\Delta) = E_H(\Delta_1) + E_H(\Delta_2).$$

Both distributions in the Kullback-Leibler divergence $\Delta$ are equal to their respective distributions in the divergence in Theorem 2 shifted by $Hq^{1/2}\Lambda^{1/2}\xi = M^{\mathrm{T}}Z\xi$. Since the Kullback-Leibler divergence is invariant to one-to-one transformations (Kullback & Leibler, 1951, Corollary 4.1), $\Delta$ equals the divergence in Theorem 2. Also, $H$ is a deterministic function of $M^{\mathrm{T}}Z$ such that taking the expectation with respect to one or the other yields the same result. Therefore, $E_H(\Delta)$ equals the left-hand side of Theorem 2. The required result is thus $E_H(\Delta) \leq \delta_1 + \delta_2$ which inserting Lemmas S3 and S4 into the last display provides. □

## S4. PROOF OF THEOREM 3

Let $P_{\gamma'} = X_\gamma(X_\gamma^{\mathrm{T}}X_\gamma)^{-1}X_\gamma^{\mathrm{T}}$ denote the orthogonal projection onto the column space of $X_\gamma$. The assumptions in Theorem 3 in addition to (9) are

$$\mathrm{pr}(\gamma = \gamma^0) > 0, \tag{S5a}$$

$$\lim_{n\to\infty} \frac{\|(I_n - P_\gamma)X\beta^0\|^2}{n} > 0 \text{ for any } \gamma \text{ not containing } \gamma^0, \tag{S5b}$$

$$g_n \to \infty, \tag{S5c}$$

$$\frac{\log g_n}{n} \to 0. \tag{S5d}$$

Assumption (S5a) is a basic prerequisite as otherwise $\mathrm{pr}(\gamma = \gamma^0 \mid y) = 0$. Assumption (S5b) is analogous to Equation A.4 from Fernández et al. (2001). Previous literature (Fernández et al., 2001; Liang et al., 2008) required $g_n$ to grow appropriately with $n$, estimates $g_n$ via empirical Bayes, or places an appropriate prior on $g_n$ to obtain consistency. We focus on the first case by assuming (S5c) and (S5d). Condition (S5b) ensures that any model that does not contain the true one has posterior probability converging to zero. The fact that supersets of the true model are also discarded follows from the $g$-prior, which favors smaller subsets.

LEMMA S5. $P_{\gamma^0} - P_\gamma = MM^{\mathrm{T}}(P_{\gamma^0} - P_\gamma)$.

*Proof.* Recall from § 2·2 that $S^{\mathrm{T}}X = 0_{(n-p)\times p}$ and $Q$ is orthogonal so that $QQ^{\mathrm{T}} = I_n$. Therefore,

$$MM^{\mathrm{T}}X = MM^{\mathrm{T}}X + S\underbrace{(S^{\mathrm{T}}X)}_{0_{(n-p)\times p}} = (MM^{\mathrm{T}} + SS^{\mathrm{T}})X = (QQ^{\mathrm{T}})X = I_nX = X,$$

where the third equality follows from $Q = (M,\ S)$. Considering $MM^{\mathrm{T}}X = X$ columnwise and recalling $P_\gamma = X_\gamma(X_\gamma^{\mathrm{T}}X_\gamma)^{-1}X_\gamma^{\mathrm{T}}$ yields $MM^{\mathrm{T}}P_\gamma = P_\gamma$, for any $\gamma$ including $\gamma^0$. □

*Proof of Theorem* 3. Conditional on $\gamma$ and $\eta$, the set-up is a normal-normal model as both the prior $\Pi(\beta_\gamma \mid \gamma)$ from (8) and the likelihood from (1) are Gaussian. The corresponding marginal likelihood follows as

$$\begin{aligned}\pi(y \mid \gamma, \eta) &= \int \pi(y \mid \beta_\gamma, \gamma, \eta)\, \pi(\beta_\gamma \mid \gamma)\, d\beta_\gamma \\ &= \big(2\pi\sigma^2\big)^{-\frac{p}{2}}(g_n+1)^{-\frac{|\gamma|}{2}} \exp\left\{-\frac{1}{2\sigma^2}\left(\|z\|^2 - \frac{g_n}{g_n+1}z^{\mathrm{T}}P_\gamma z\right)\right\},\end{aligned}$$

where $z = y - \eta$ and $|\gamma|$ denotes the number of elements in $\gamma$. The logarithm of the Bayes factor of the true model $\gamma^0$ over $\gamma$ conditional on $\eta$ is thus

$$\log \mathrm{BF}_{\gamma^0:\gamma} = \log\left\{\frac{\pi\big(y \mid \gamma^0, \eta\big)}{\pi(y \mid \gamma, \eta)}\right\} = \frac{|\gamma| - |\gamma^0|}{2}\log(g_n+1) + \frac{g_n}{2\sigma^2(g_n+1)}\, h_\gamma(z), \tag{S6}$$

where $h_\gamma(z) = z^{\mathrm{T}}(P_{\gamma^0} - P_\gamma)z$. By assumption (S5a), the required result follows if $\log \mathrm{BF}_{\gamma^0:\gamma} \to \infty$ in probability, except for $\gamma = \gamma^0$ when $\log \mathrm{BF}_{\gamma^0:\gamma^0} = 0$.

Since $z = y - \eta$, $z \sim N(\nu,\ \sigma^2 I_n)$ where $\nu = X\beta^0 + \eta^0 - \eta$. Then, by Theorems 5.2a and 5.2c from Rencher & Schaalje (2008) and the fact that the trace of a projection matrix equals the

dimensionality of its target space,

$$\begin{aligned} E\{h_\gamma(z) \mid \eta\} &= \sigma^2 \operatorname{tr}(P_{\gamma^0} - P_\gamma) + \nu^{\mathrm{T}}(P_{\gamma^0} - P_\gamma)\nu \\ &= \sigma^2(|\gamma^0| - |\gamma|) + \nu^{\mathrm{T}}(P_{\gamma^0} - P_\gamma)\nu, \end{aligned} \tag{S7a}$$

$$\begin{aligned} \operatorname{var}\{h_\gamma(z) \mid \eta\} &= 2\sigma^4 \operatorname{tr}\big\{(P_{\gamma^0} - P_\gamma)^2\big\} + 4\sigma^2\nu^{\mathrm{T}}(P_{\gamma^0} - P_\gamma)^2\nu \\ &\leq 2\sigma^4\big\{\operatorname{tr}(P_{\gamma^0}) + \operatorname{tr}(P_\gamma)\big\} + 4\sigma^2\|(P_{\gamma^0} - P_\gamma)\nu\|^2 \\ &= 2\sigma^4(|\gamma^0| + |\gamma|) + 4\sigma^2\|(P_{\gamma^0} - P_\gamma)\nu\|^2. \end{aligned} \tag{S7b}$$

We analyze the asymptotic behavior of $h_\gamma(z)$ by bounding this expectation and variance.

The first term of each right-hand side in (S7) is independent of $n$. Let us bound the second terms. Inserting $\nu = X\beta^0 + \zeta$ where $\zeta = \eta^0 - \eta$ and expanding the square yields

$$\nu^{\mathrm{T}}(P_{\gamma^0} - P_\gamma)\nu = (X\beta^0)^{\mathrm{T}}(P_{\gamma^0} - P_\gamma)X\beta^0 + 2\zeta^{\mathrm{T}}(P_{\gamma^0} - P_\gamma)X\beta^0 + \zeta^{\mathrm{T}}(P_{\gamma^0} - P_\gamma)\zeta.$$

Inserting $P_{\gamma^0}X\beta^0 = X\beta^0$ and Lemma S5 provides

$$\begin{aligned} \nu^{\mathrm{T}}(P_{\gamma^0} - P_\gamma)\nu &= (X\beta^0)^{\mathrm{T}}(I_n - P_\gamma)X\beta^0 + 2\zeta^{\mathrm{T}}MM^{\mathrm{T}}(P_{\gamma^0} - P_\gamma)X\beta^0 + \zeta^{\mathrm{T}}(P_{\gamma^0} - P_\gamma)\zeta \\ &= \|(I_n - P_\gamma)X\beta^0\|^2 + 2\zeta^{\mathrm{T}}MM^{\mathrm{T}}(I_n - P_\gamma)X\beta^0 + \zeta^{\mathrm{T}}MM^{\mathrm{T}}(P_{\gamma^0} - P_\gamma)\zeta. \end{aligned}$$

Applying the Cauchy-Schwarz inequality and $|\zeta^{\mathrm{T}}MM^{\mathrm{T}}(P_{\gamma^0} - P_\gamma)\zeta| \leq \zeta^{\mathrm{T}}MM^{\mathrm{T}}\zeta = \|M^{\mathrm{T}}\zeta\|^2$,

$$\begin{aligned} \nu^{\mathrm{T}}(P_{\gamma^0} - P_\gamma)\nu &\geq \|(I_n - P_\gamma)X\beta^0\|^2 - 2\|M^{\mathrm{T}}\zeta\|\,\|M^{\mathrm{T}}(I_n - P_\gamma)X\beta^0\| - \|M^{\mathrm{T}}\zeta\|^2, \\ \nu^{\mathrm{T}}(P_{\gamma^0} - P_\gamma)\nu &\leq \|(I_n - P_\gamma)X\beta^0\|^2 + 2\|M^{\mathrm{T}}\zeta\|\,\|M^{\mathrm{T}}(I_n - P_\gamma)X\beta^0\| + \|M^{\mathrm{T}}\zeta\|^2. \end{aligned}$$

Since the columns of $M$ form an orthonormal basis for the column space of $X$, $\|M^{\mathrm{T}}(I_n - P_\gamma)X\beta^0\| = \|(I_n - P_\gamma)X\beta^0\|$ such that

$$\begin{aligned} \nu^{\mathrm{T}}(P_{\gamma^0} - P_\gamma)\nu &\geq \|(I_n - P_\gamma)X\beta^0\|^2 - 2\|M^{\mathrm{T}}\zeta\|\,\|(I_n - P_\gamma)X\beta^0\| - \|M^{\mathrm{T}}\zeta\|^2 \\ &= \big\{\|(I_n - P_\gamma)X\beta^0\| - 2\|M^{\mathrm{T}}\zeta\|\big\}\,\|(I_n - P_\gamma)X\beta^0\| - \|M^{\mathrm{T}}\zeta\|^2, \end{aligned} \tag{S8a}$$

$$\begin{aligned} \nu^{\mathrm{T}}(P_{\gamma^0} - P_\gamma)\nu &\leq \|(I_n - P_\gamma)X\beta^0\|^2 + 2\|M^{\mathrm{T}}\zeta\|\,\|(I_n - P_\gamma)X\beta^0\| + \|M^{\mathrm{T}}\zeta\|^2 \\ &= \big\{\|(I_n - P_\gamma)X\beta^0\| + 2\|M^{\mathrm{T}}\zeta\|\big\}\,\|(I_n - P_\gamma)X\beta^0\| + \|M^{\mathrm{T}}\zeta\|^2. \end{aligned} \tag{S8b}$$

For the second term of the right-hand side in (S7b), consider $\nu = X\beta^0 + \zeta$ and

$$\|(P_{\gamma^0} - P_\gamma)\nu\| = \|(I_n - P_\gamma)X\beta^0 + (P_{\gamma^0} - P_\gamma)\zeta\|.$$

By the triangle inequality,

$$\begin{aligned} \|(P_{\gamma^0} - P_\gamma)\nu\| &\leq \|(P_{\gamma^0} - P_\gamma)X\beta^0\| + \|(P_{\gamma^0} - P_\gamma)\zeta\| \\ &= \|(I_n - P_\gamma)X\beta^0\| + \|(P_{\gamma^0} - P_\gamma)MM^{\mathrm{T}}\zeta\|, \end{aligned}$$

where the equality follows from $P_{\gamma^0}X\beta^0 = X\beta^0$ and Lemma S5. Also, $\|(P_{\gamma^0} - P_\gamma)MM^{\mathrm{T}}\zeta\| \leq \|MM^{\mathrm{T}}\zeta\| = \|M^{\mathrm{T}}\zeta\|$ since $M^{\mathrm{T}}M = I_p$. Therefore,

$$\|(P_{\gamma^0} - P_\gamma)\nu\| \leq \|(I_n - P_\gamma)X\beta^0\| + \|M^{\mathrm{T}}\zeta\|.$$

Inserting into (S7b) provides

$$\begin{aligned} \operatorname{var}\{h_\gamma(z) \mid \eta\}^{\frac{1}{2}} &\leq \big\{2\sigma^4(|\gamma^0| + |\gamma|) + 4\sigma^2\|(P_{\gamma^0} - P_\gamma)\nu\|^2\big\}^{\frac{1}{2}} \\ &\leq 2^{\frac{1}{2}}\,\sigma^2(|\gamma^0| + |\gamma|)^{\frac{1}{2}} + 2\sigma\|(P_{\gamma^0} - P_\gamma)\nu\| \\ &\leq 2^{\frac{1}{2}}\,\sigma^2(|\gamma^0| + |\gamma|)^{\frac{1}{2}} + 2\sigma\big\{\|(I_n - P_\gamma)X\beta^0\| + \|M^{\mathrm{T}}\zeta\|\big\}. \end{aligned} \tag{S9}$$

Since $\zeta = \eta^0 - \eta$, assumptions (9) and (S5d) imply

$$\frac{\|M^{\mathrm{T}}\zeta\|^2}{\log g_n} \to 0, \qquad \frac{\|M^{\mathrm{T}}\zeta\|^2}{n} \to 0; \tag{S10}$$

in probability. Let us consider $\gamma \neq \gamma^0$ that contain $\gamma^0$, that is $\gamma^0 \subsetneq \gamma$, and $\gamma$ that do not contain $\gamma^0$, that is $\gamma^0 \not\subset \gamma$, separately.

First, consider the case where $\gamma$ does not contain $\gamma^0$. Assumption (S5b), (S7a), (S8a), and (S10) imply $E\{h_\gamma(z) \mid \eta\}/\|(I_n - P_\gamma)X\beta^0\| \to \infty$. On the other hand, $\lim_{n\to\infty} \mathrm{var}\{h_\gamma(z) \mid \eta\}^{1/2}/\|(I_n - P_\gamma)X\beta^0\| \leq 2\sigma$ by (S5b), (S9), and (S10). Therefore, $\lim_{n\to\infty} h_\gamma(z \mid \eta)/n > 0$ with probability tending to one by Chebyshev's inequality and (S5b). Under assumption (S5d), it then follows from (S6) that $\log \mathrm{BF}_{\gamma^0:\gamma} \to \infty$ in probability.

Next, consider the case where $\gamma$ contains $\gamma_0$. In this setting, $P_\gamma X\beta^0 = X\beta^0$ and thus $(I_n - P_\gamma)X\beta^0 = 0_{n\times 1}$. Therefore, (S7a) with (S8b), and (S9) reduce to

$$\begin{aligned} E\{h_\gamma(z) \mid \eta\} &\leq \sigma^2(|\gamma^0| - |\gamma|) + \|M^{\mathrm{T}}\zeta\|^2, \\ \mathrm{var}\{h_\gamma(z) \mid \eta\}^{\frac{1}{2}} &\leq 2^{\frac{1}{2}}\,\sigma^2(|\gamma^0| + |\gamma|)^{\frac{1}{2}} + 2\sigma\,\|M^{\mathrm{T}}\zeta\|. \end{aligned}$$

Chebyshev's inequality and (S10) provide thus $\lim_{n\to\infty} h_\gamma(z \mid \eta)/\log g_n = 0$ with probability tending to one. We conclude from (S6) that $\mathrm{BF}_{\gamma^0:\gamma} \to \infty$ in probability because of assumption (S5c) and $|\gamma| > |\gamma_0|$.

We have shown $\mathrm{BF}_{\gamma^0:\gamma} \to \infty$ whenever $\gamma \neq \gamma^0$. The required result follows from this result as noted earlier in this proof. □

## S5. VECTOR APPROXIMATE MESSAGE PASSING

### S5·1. *Derivation*

To give a motivation for the steps of vector approximate message passing in Algorithm S1 on page 12, we derive the algorithm as an approximation to sum-product message passing (Bishop, 2006, § 8.4.4) similar to what is done in Rangan et al. (2016, § III-B). Consider the linear model $y \sim N(X\beta, \sigma^2 I_n)$ where $y$ is an $n$-dimensional vector of observations, $X$ an $n \times p$ design matrix, $\beta$ a $p$-dimensional vector of parameters, and $\sigma^2$ the error variance. We assume that the entries of $\beta$ are a priori independent such that $\pi(\beta) = \prod_{j=1}^p \pi(\beta_j)$. The goal is to approximate the posterior

$$\begin{aligned} \pi(\beta \mid y) \propto \pi(\beta)\,\pi(y \mid \beta) &= \pi(\beta)\,N(y \mid X\beta,\ \sigma^2 I_n) \\ &= \pi(\beta)\,\delta(\beta - \tilde{\beta})\,N(y \mid X\tilde{\beta},\ \sigma^2 I_n), \end{aligned} \tag{S11}$$

where $\delta$ is the Dirac delta function and $\tilde{\beta}$ is thus a copy of $\beta$. This copying of $\beta$ gives rise to an extra variable node in the corresponding factor graph in Fig. S1.

Let $\mu_{\pi\to\beta}$ and $\mu_{\delta\to\beta}$ denote the messages to the variable node $\beta$, $\mu_{\delta\to\tilde{\beta}}$ and $\mu_{N\to\tilde{\beta}}$ the messages to the variable node $\tilde{\beta}$, and $\mu_{\beta\to\delta}$ and $\mu_{\tilde{\beta}\to\delta}$ the messages to the factor node $\delta(\beta - \tilde{\beta})$. By the general expression for a message from a factor to a variable node (Bishop, 2006, Equation 8.69),

$$\begin{aligned} \mu_{\delta\to\tilde{\beta}}(\tilde{\beta}) &= \int \delta(\beta - \tilde{\beta})\mu_{\beta\to\delta}(\beta)d\beta = \mu_{\beta\to\delta}(\tilde{\beta}), \\ \mu_{\delta\to\beta}(\beta) &= \int \delta(\beta - \tilde{\beta})\mu_{\tilde{\beta}\to\delta}(\tilde{\beta})d\tilde{\beta} = \mu_{\tilde{\beta}\to\delta}(\beta). \end{aligned} \tag{S12}$$

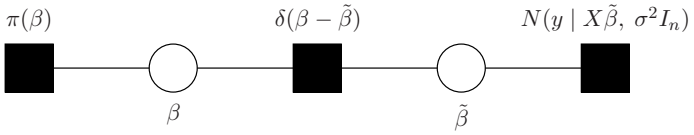


Fig. S1. The factor graph representation of (S11). The squares and circles are factor and variable nodes, respectively. This figure is an edited version of Fig. 1 from Rangan et al. (2016).

The beliefs at the variable nodes are the products of the incoming messages,

$$b(\beta) \propto \mu_{\pi\to\beta}(\beta)\,\mu_{\delta\to\beta}(\beta) = \pi(\beta)\,\mu_{\delta\to\beta}(\beta),$$
$$b(\tilde{\beta}) \propto \mu_{\delta\to\tilde{\beta}}(\tilde{\beta})\,\mu_{N\to\tilde{\beta}}(\tilde{\beta}) = \mu_{\beta\to\delta}(\tilde{\beta})\,N(y \mid X\tilde{\beta},\ \sigma^2 I_n);$$

where the last equality uses (S12). Combining these beliefs with the general expression for a message from a variable to a factor node (Bishop, 2006, Equation 8.66) and Fig. S1 yields

$$\mu_{\beta\to\delta}(\beta) = \mu_{\pi\to\beta}(\beta) \propto \frac{b(\beta)}{\mu_{\delta\to\beta}(\beta)},$$
$$\mu_{\tilde{\beta}\to\delta}(\tilde{\beta}) = \mu_{N\to\tilde{\beta}}(\tilde{\beta}) \propto \frac{b(\tilde{\beta})}{\mu_{\delta\to\tilde{\beta}}(\tilde{\beta})} = \frac{b(\tilde{\beta})}{\mu_{\beta\to\delta}(\tilde{\beta})};$$

where the last equality follows from (S12).

The last two displays provide a message-passing algorithm. Initialize $\mu_{\delta\to\beta}(\beta)$. Then, iterate the updates

$$b(\beta) \propto \pi(\beta)\,\mu_{\delta\to\beta}(\beta), \tag{S13a}$$
$$\mu_{\beta\to\delta}(\beta) \propto \frac{b(\beta)}{\mu_{\delta\to\beta}(\beta)}, \tag{S13b}$$
$$b(\tilde{\beta}) \propto \mu_{\beta\to\delta}(\tilde{\beta})\,N(y \mid X\tilde{\beta},\ \sigma^2 I_n), \tag{S13c}$$
$$\mu_{\delta\to\beta}(\tilde{\beta}) = \mu_{\tilde{\beta}\to\delta}(\tilde{\beta}) \propto \frac{b(\tilde{\beta})}{\mu_{\beta\to\delta}(\tilde{\beta})}, \tag{S13d}$$

where the last equality is from (S12). Since the graph in Fig. S1 is a tree, the beliefs $b(p)$ converge to the exact posterior $\pi(\beta \mid y)$ after one iteration. This exact algorithm can however be expensive to compute for certain $\pi(\beta)$ if $p$ is large. Vector approximate message passing approximates (S13) to reduce computational cost:

Initialize $\mu_{\delta\to\beta}(\beta) = N(\beta \mid r_0,\ \ t_0^2 I_p)$. At the $k$th iteration, approximate $b(\beta)$ by $N(\beta \mid \hat{\beta}_k,\ s_k^2 I_p)$ where $\hat{\beta}_k = E_{b(\beta)}(\beta)$ and $s_k^2 = \mathrm{Tr}\{\mathrm{cov}_{b(\beta)}(\beta)\}/p$. Applying (S13a) provides Step 3a of Algorithm S1.

Since $\mu_{\delta\to\beta}(\beta) \approx N(\beta \mid r_k,\ t_k^2 I_p)$ and $b(\beta) \approx N(\beta \mid \hat{\beta}_k,\ s_k^2 I_p)$, the resulting approximation to $\mu_{\beta\to\delta}(\beta)$ is Gaussian too by (S13b). Denote this Gaussian approximation by $N(\tilde{\beta} \mid \tilde{r}_k,\ \tilde{t}_k^2 I_p)$. Step 3b states the update equations for $\tilde{r}_k$ and $\tilde{t}_k^2$ derived from (S13b).

With $\mu_{\beta\to\delta}(\tilde{\beta}) \approx N(\tilde{\beta} \mid \tilde{r}_k,\ \tilde{t}_k^2 I_p)$, $b(\tilde{\beta})$ from (S13c) is Gaussian too. We further approximate $b(\tilde{\beta})$ by requiring its covariance to be proportional to the identity matrix. Let $b(\tilde{\beta}) \approx$

$N(\tilde{\beta} \mid \hat{\tilde{\beta}}_k,\ \tilde{s}_k^2 I_p)$. The updates follow from (S13c) as

$$\hat{\tilde{\beta}}_k = \big(\tilde{t}_k^2 X^{\mathrm{T}} X + \sigma^2 I_p\big)^{-1}\big(\tilde{t}_k^2 X^{\mathrm{T}} y + \sigma^2 \tilde{r}_k\big), \tag{S14a}$$

$$\tilde{s}_k^2 = \frac{\sigma^2 \tilde{t}_k^2}{p} \operatorname{Tr}\Big\{\big(\tilde{t}_k^2 X^{\mathrm{T}} X + \sigma^2 I_p\big)^{-1}\Big\}. \tag{S14b}$$

These involve an inversion of a $p \times p$ matrix which is expensive to compute if $p$ is large. We can however rewrite these expressions to make their computation faster.

Let $X = UDV^{\mathrm{T}}$ denote a singular-value decomposition with an $n \times \min(n,p)$ matrix $U$, a $\min(n,p) \times \min(n,p)$ diagonal matrix $D$, and $V$ an $p \times \min(n,p)$ matrix such that $U^{\mathrm{T}}U = V^{\mathrm{T}}V = I_{\min(n,p)}$. Substituting $X = UDV^{\mathrm{T}}$ yields

$$\begin{aligned}
\big(\tilde{t}_k^2 X^{\mathrm{T}} X + \sigma^2 I_p\big)^{-1} = \big(\tilde{t}_k^2 V D^2 V^{\mathrm{T}} + \sigma^2 I_p\big)^{-1} &= \frac{1}{\sigma^2}\left(\frac{\tilde{t}_k^2}{\sigma^2} V D^2 V^{\mathrm{T}} + I_p\right)^{-1} \\
&= \frac{1}{\sigma^2}\left\{I_p - V\left(\frac{\sigma^2}{\tilde{t}_k^2} D^{-2} + V^{\mathrm{T}} V\right)^{-1} V^{\mathrm{T}}\right\} \\
&= \frac{1}{\sigma^2}\left[I_p - V\left\{\frac{\sigma^2}{\tilde{t}_k^2} D^{-2} + I_{\min(n,p)}\right\}^{-1} V^{\mathrm{T}}\right],
\end{aligned} \tag{S15}$$

where $D^2 = DD$, $D^{-2} = D^{-1}D^{-1}$ and the third equality follows from the Woodbury matrix identity. Substituting $X^{\mathrm{T}} = VDU^{\mathrm{T}}$ and (S15) provide

$$\begin{aligned}
\big(\tilde{t}_k^2 X^{\mathrm{T}} X + \sigma^2 I_p\big)^{-1} \tilde{t}_k^2 X^{\mathrm{T}} y &= \frac{\tilde{t}_k^2}{\sigma^2}\left[I_p - V\left\{\frac{\sigma^2}{\tilde{t}_k^2} D^{-2} + I_{\min(n,p)}\right\}^{-1} V^{\mathrm{T}}\right] V D U^{\mathrm{T}} y \\
&= \frac{\tilde{t}_k^2}{\sigma^2}\left[V - V\left\{\frac{\sigma^2}{\tilde{t}_k^2} D^{-2} + I_{\min(n,p)}\right\}^{-1} V^{\mathrm{T}} V\right] D U^{\mathrm{T}} y \\
&= \frac{\tilde{t}_k^2}{\sigma^2} V\left[I_{\min(n,p)} - \left\{\frac{\sigma^2}{\tilde{t}_k^2} D^{-2} + I_{\min(n,p)}\right\}^{-1}\right] D U^{\mathrm{T}} y,
\end{aligned}$$

where the last equality uses $V^{\mathrm{T}}V = I_{\min(n,p)}$. Since the expression inside the square brackets consists only of diagonal matrices, we can write it as a single fraction to obtain

$$\begin{aligned}
\big(\tilde{t}_k^2 X^{\mathrm{T}} X + \sigma^2 I_p\big)^{-1} \tilde{t}_k^2 X^{\mathrm{T}} y &= \frac{\tilde{t}_k^2}{\sigma^2} V\left[\frac{\sigma^2}{\tilde{t}_k^2} D^{-2}\left\{\frac{\sigma^2}{\tilde{t}_k^2} D^{-2} + I_{\min(n,p)}\right\}^{-1}\right] D U^{\mathrm{T}} y \\
&= \frac{\tilde{t}_k^2}{\sigma^2} V\left\{I_{\min(n,p)} + \frac{\tilde{t}_k^2}{\sigma^2} D^2\right\}^{-1} D U^{\mathrm{T}} y \\
&= V\left\{\frac{\sigma^2}{\tilde{t}_k^2} I_{\min(n,p)} + D^2\right\}^{-1} D U^{\mathrm{T}} y,
\end{aligned}$$

where the second equality follows from multiplying both the numerator and the denominator of the diagonal-matrices faction by $(\tilde{t}_k^2/\sigma^2)D^2$. Combining the last display with (S14a) and (S15)

provides

$$\begin{aligned}
\hat{\tilde{\beta}}_k &= V\left\{\frac{\sigma^2}{\tilde{t}_k^2} I_{\min(n,p)} + D^2\right\}^{-1} DU^{\mathrm{T}} y + \left[I_p - V\left\{\frac{\sigma^2}{\tilde{t}_k^2} D^{-2} + I_{\min(n,p)}\right\}^{-1} V^{\mathrm{T}}\right]\tilde{r}_k \\
&= \tilde{r}_k + V\left\{\frac{\sigma^2}{\tilde{t}_k^2} I_{\min(n,p)} + D^2\right\}^{-1} DU^{\mathrm{T}} y - V\left\{\frac{\sigma^2}{\tilde{t}_k^2} D^{-2} + I_{\min(n,p)}\right\}^{-1} V^{\mathrm{T}}\tilde{r}_k \\
&= \tilde{r}_k + V\left\{\frac{\sigma^2}{\tilde{t}_k^2} I_{\min(n,p)} + D^2\right\}^{-1} DU^{\mathrm{T}} y - V\left\{\frac{\sigma^2}{\tilde{t}_k^2} I_{\min(n,p)} + D^2\right\}^{-1} D^2 V^{\mathrm{T}}\tilde{r}_k \qquad \text{(S16)} \\
&= \tilde{r}_k + V\left\{\frac{\sigma^2}{\tilde{t}_k^2} I_{\min(n,p)} + D^2\right\}^{-1} \left(DU^{\mathrm{T}} y - D^2 V^{\mathrm{T}}\tilde{r}_k\right) \\
&= \tilde{r}_k + V\left(\frac{\sigma^2}{\tilde{t}_k^2} D^{-1} + D\right)^{-1} (U^{\mathrm{T}} y - DV^{\mathrm{T}}\tilde{r}_k),
\end{aligned}$$

where we used that $D$ is a diagonal matrix. This update for $\hat{\tilde{\beta}}_k$ only involves matrix multiplications and inversions of diagonal matrices.

For $\tilde{s}_k^2$, substitute (S15) into (S14b) such that

$$\begin{aligned}
\tilde{s}_k^2 &= \frac{\tilde{t}_k^2}{p}\mathrm{Tr}\left[I_p - V\left\{\frac{\sigma^2}{\tilde{t}_k^2} D^{-2} + I_{\min(n,p)}\right\}^{-1} V^{\mathrm{T}}\right] \\
&= \tilde{t}_k^2\left(1 - \frac{1}{p}\mathrm{Tr}\left[\left\{\frac{\sigma^2}{\tilde{t}_k^2} D^{-2} + I_{\min(n,p)}\right\}^{-1} V^{\mathrm{T}} V\right]\right),
\end{aligned}$$

where the last equality uses that the trace is invariant under cyclic permutations. Since $V^{\mathrm{T}}V = I_{\min(n,p)}$, the last expression reduces to

$$\begin{aligned}
\tilde{s}_k^2 &= \tilde{t}_k^2\left(1 - \frac{1}{p}\mathrm{Tr}\left[\left\{\frac{\sigma^2}{\tilde{t}_k^2} D^{-2} + I_{\min(n,p)}\right\}^{-1}\right]\right) \\
&= \tilde{t}_k^2\left[1 - \frac{1}{p}\,\mathrm{Tr}\left\{D\left(\frac{\sigma^2}{\tilde{t}_k^2} D^{-1} + D\right)^{-1}\right\}\right],
\end{aligned}$$

where the last equality uses that the argument of the trace is diagonal. This display with (S16) constitutes Step 3c.

Recall $\mu_{\beta\to\delta}(\tilde{\beta}) \approx N(\tilde{\beta} \mid \tilde{r}_k,\ \tilde{t}_k^2 I_p)$ and $b(\tilde{\beta}) \approx N(\tilde{\beta} \mid \hat{\tilde{\beta}}_k,\ \tilde{s}_k^2 I_p)$. We would like to update $\mu_{\delta\to\beta}(\beta) \approx N(\beta \mid r_{k+1},\ t_{k+1}^2 I_p)$ where we have incremented the iteration counter. Step 3d follows now from (S13d) in the same way as Step 3b followed from (S13b).

*Algorithm S*1. Vector approximate message passing.
Input: Data $(y,\ X)$

1. Compute the singular-value decomposition $X = UDV^{\mathrm{T}}$.
2. Initialize $r_0$ and $t_0^2$.
3. For $k = 0, \ldots, K$ do:
   a. Set $\hat{\beta}_{k,j} = E(\beta_j \mid r_{k,j},\ t_k^2)$ and $s_k^2 = \sum_{j=1}^p \mathrm{var}(\beta_j \mid r_{k,j},\ t_k^2)/p$ where the density of $\beta_j$ is proportional to $\pi(\beta_j)\, N(\beta_j \mid r_{k,j},\ t_k^2)$ for $j = 1, \ldots, p$.
   b. Set $1/\tilde{t}_k^2 = 1/s_k^2 - 1/t_k^2$ and $\tilde{r}_k = (t_k^2 \hat{\beta}_k - s_k^2 r_k)/(t_k^2 - s_k^2)$.
   c. Set $\hat{\tilde{\beta}}_k = \tilde{r}_k + V(\sigma^2 D^{-1}/\tilde{t}_k^2 + D)^{-1}(U^{\mathrm{T}} y - DV^{\mathrm{T}} \tilde{r}_k)$ and $\tilde{s}_k^2 = \tilde{t}_k^2\, [1 - \mathrm{Tr}\{D\,(\sigma^2 D^{-1}/\tilde{t}_k^2 + D)^{-1}\}/p]$.
   d. Set $1/t_{k+1}^2 = 1/\tilde{s}_k^2 - 1/\tilde{t}_k^2$ and $r_{k+1} = (\tilde{t}_k^2 \hat{\tilde{\beta}}_k - \tilde{s}_k^2 \tilde{r}_k)/(\tilde{t}_k^2 - \tilde{s}_k^2)$.

Output: Approximate posterior $N(\hat{\beta}_K,\ s_K^2 I_p)$

## S5·2. *Computational complexity*

The computational complexity of the singular-value decomposition is $O\{n\, p\, \min(n, p)\}$ (Rangan et al., 2016, § I-E). The steps inside each iteration are $O(p)$ except for Step 3c which is $O\{p\, \min(n, p)\}$ if $U^{\mathrm{T}} y$ is precomputed. The computational complexity of Algorithm S1 is thus $O\{(n + K)\, p\, \min(n, p)\}$.

In practice, we do not always run Algorithm S1 for all $K$ iterations. We stop it once the innovation $\|\hat{\beta}_k - \hat{\beta}_{k-1}\|^2$ becomes small enough, indicating convergence.

## S5·3. *Estimating $\sigma^2$*

So far, we have treated $\sigma^2$ as fixed and known. As § 2·4 notes, applications like those in § 5·2 and § 5·3 often require estimation of $\sigma^2$ and methods available for Step 2 of Algorithm 1 often provide such estimation. For instance, Vila & Schniter (2011) detail how $\sigma^2$ can be estimated when using approximate message passing. We add a step to Algorithm S1 to estimate $\sigma^2$ when required: Consider the prior $1/\sigma^2 \sim \mathrm{Ga}(a_0, b_0)$ for some shape parameter $a_0$ and rate parameter $b_0$. Then, the full conditional posterior for $1/\sigma^2$ of Algorithm S1 at iteration $k$ is

$$\frac{1}{\sigma^2} \mid \hat{\beta}_k \sim \mathrm{Ga}\left(a_0 + \frac{n}{2},\, b_0 + \frac{\|y - X\hat{\beta}_k\|^2}{2}\right).$$

At each iteration, we update $\sigma^2$ such that $1/\sigma^2$ matches the mean of this full conditional:

$$\sigma_k^2 = \frac{b_0 + \|y - X\hat{\beta}_k\|^2/2}{a_0 + n/2} \quad (k = 1, \ldots, K),$$

between Steps 3a and 3b of Algorithm S1.

## S5·4. *Dampened updates*

If vector approximate message passing fails to converge, which can happen for certain matrices $X$ which have a challenging collinearity structure, damping of updates can induce convergence,

like it does in approximate message passing (Rangan et al., 2014). In this article, we only dampen the updates for the SNP application in § 5·3 to ensure convergence.

Let $\rho \in (0, 1]$ denote the damping constant with $\rho = 1$ representing no damping. Then, the dampened version of Algorithm S1 follows by replacing Steps 3a and 3c by

$$\hat{\beta}_{k,j} = (1-\rho)\,\hat{\beta}_{k-1,j} + \rho\,E(\beta_j \mid r_{k,j},\, t_k^2),$$
$$s_k^2 = (1-\rho)\,s_{k-1}^2 + \rho \sum_{j=1}^{p} \mathrm{var}(\beta_j \mid r_{k,j},\, t_k^2)/p;$$

and

$$\hat{\tilde{\beta}}_k = (1-\rho)\,\hat{\tilde{\beta}}_{k-1} + \rho\,\{\tilde{r}_k + V(\sigma^2 D^{-1}/\tilde{t}_k^2 + D)^{-1}(U^{\mathrm{T}}y - DV^{\mathrm{T}}\tilde{r}_k)\},$$
$$\tilde{s}_k^2 = (1-\rho)\,\tilde{s}_{k-1}^2 + \rho\,\tilde{t}_k^2\,[1 - \mathrm{Tr}\{D\,(\sigma^2 D^{-1}/\tilde{t}_k^2 + D)^{-1}\}/p];$$

respectively, for $k > 0$.

## S6. Debiased lasso

Consider the linear model $y \sim N(X\beta, \sigma^2 I_n)$ as in § S5 with the spike-and-slab prior $\beta_j \sim \lambda\,N(0,\,\psi) + (1-\lambda)\,\delta(0)$ independently for $j = 1, \ldots, p$. Denote the lasso estimator of $\beta$ by $\hat{\beta}^{\text{lasso}}(y)$: Here, we use the smallest lasso regularization parameter that results in at most $\lfloor \lambda p \rfloor$ nonzero coefficients where $\lambda p$ is the number of expected nonzero elements in $\beta$ under its spike-and-slab prior. The lasso algorithm from Efron et al. (2004) allows for efficient computation of $\hat{\beta}^{\text{lasso}}(y)$ under this constraint on the regularization parameter.

As the number of predictors is less than the sample size in § 5·2, we assume $p \leq n$. Then, we can set the matrix $M$ in Javanmard & Montanari (2013) equal to $\hat{\Sigma}^{-1}$ where $\hat{\Sigma} = X^{\mathrm{T}}X/n$. The debiased lasso estimator follows as (Javanmard & Montanari, 2013, Equation 5)

$$\hat{\beta}^{\text{unbiased}}(y) = \hat{\beta}^{\text{lasso}}(y) + \frac{1}{n} M X^{\mathrm{T}} \Big\{ y - \hat{\beta}^{\text{lasso}}(y) \Big\}.$$

Theorem 2.1 from Javanmard & Montanari (2013) implies

$$\beta \mid y \sim \mathcal{N}\left\{ \hat{\beta}^{\text{unbiased}}(y),\, \frac{\sigma^2}{n}\, M\hat{\Sigma}M^{\mathrm{T}} \right\},$$

as posterior approximation based on the debiased lasso.

The error variance $\sigma^2$ is unknown in the application from § 5·2. We therefore estimate it by

$$\frac{b_0 + \|y - X\hat{\beta}^{\text{unbiased}}(y)\|^2/2}{a_0 + n/2},$$

analogously to § S5·3.

## S7. Laplace approximation for § 3·3

We follow Steinberg & Bonilla (2014, § 2.3). Recall $F = \{f(z_1), \ldots, f(z_n)\}^{\mathrm{T}}$. Since $f$ has a Gaussian process prior, $F \sim N(\mu,\, \Sigma)$ for some $n$-dimensional mean $\mu$ and an $n \times n$ covariance matrix $\Sigma$. The first-order Taylor series of $G$ around an $n$-dimensional vector $m$ is $G(F) \approx G(m) + J_m(F - m)$ where $J_m$ is the Jacobian matrix of $G(F)$ evaluated at $m$. The corresponding approximate likelihood from the non-linear Gaussian model $S^{\mathrm{T}}y \sim N\{S^{\mathrm{T}}G(F),\, \sigma^2 I_{n-p}\}$

follows as $\hat{\pi}_m(S^{\mathrm{T}}y \mid F) = N\{S^{\mathrm{T}}y \mid S^{\mathrm{T}}G(m) + S^{\mathrm{T}}J_m(F-m),\, \sigma^2 I_{n-p}\}$, which yields the approximate posterior

$$\hat{\pi}_m(F \mid S^{\mathrm{T}}y) = N\left(\Sigma_m^*\left[\frac{1}{\sigma^2}J_m^{\mathrm{T}}SS^{\mathrm{T}}\{y - G(m) + J_m m\} + \Sigma^{-1}\mu\right],\, \Sigma_m^*\right),$$

where $\Sigma_m^* = (J_m^{\mathrm{T}}SS^{\mathrm{T}}J_m/\sigma^2 + \Sigma^{-1})^{-1}$. The posterior mean suggests the iterative update

$$m_{t+1} = (1-\rho_t)\, m_t + \rho_t\, \Sigma_{m_t}^*\left[\frac{1}{\sigma^2}J_{m_t}^{\mathrm{T}}SS^{\mathrm{T}}\{y - G(m_t) + J_{m_t}m_t\} + \Sigma^{-1}\mu\right],$$

where $t$ is the iteration number and $\rho_t$ the learning rate. This update produces a dampened Gauss-Newton algorithm. Since the mean of a Gaussian equals its mode, $m_\infty$ targets the mode of the exact posterior as $t \to \infty$. Therefore, $\hat{\pi}_{m_\infty}(F \mid S^{\mathrm{T}}y)$ provides a Laplace approximation $\hat{\Pi}(F \mid S^{\mathrm{T}}y)$.

## S8. ADDITIONAL SIMULATION STUDIES

### S8·1. *Variable selection on a correlated subset*

Consider the set-up from § 3·1 with the same spike-and-slab prior on the elements of $\beta$ as on the elements of $\alpha$, $n = 100$, $p = 2$, $\psi = 1$, $\lambda = p/(p+q)$, and $\sigma^2 = 1/2$. Generate the elements in $X$ and $Z$ independently from $N(0,1)$, then reassign the second column of $X$, denoted by

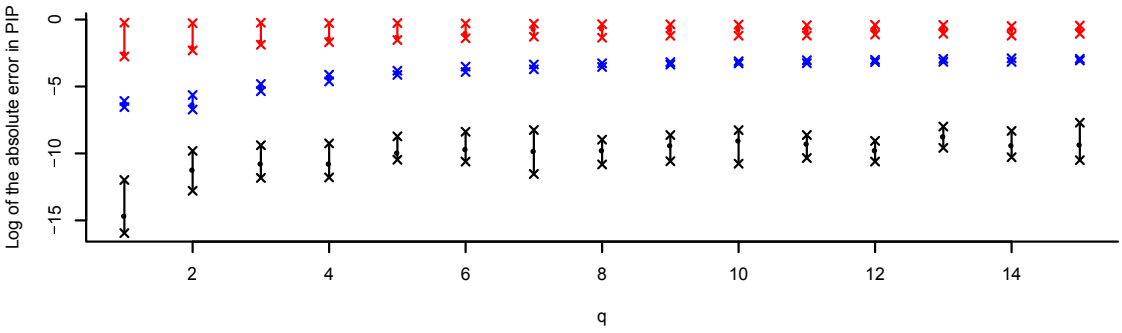


Fig. S2. Median (dot) and interquartile ranges (x) of the absolute differences between posterior inclusion probability (PIP) and their approximation from the simulation in § S8·1. Integrated rotated Gaussian approximation is in black, expectation propagation in blue, and variational Bayes in red.

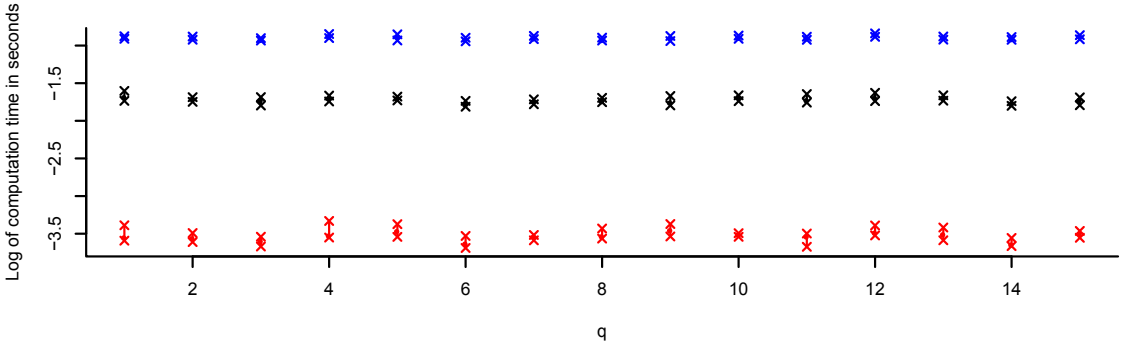


Fig. S3. Median (dot) and interquartile ranges (x) of the computation times for the results in Fig. S2. Integrated rotated Gaussian approximation is in black, expectation propagation in blue, and variational Bayes in red.

$X_{*2}$, to equal $0.01X_{*2} + 0.99X_{*1}$ to induce correlation, and lastly standardize the columns of $X$ and $Z$ to have zero mean and unit standard deviation. Generate $y$ according to (1) with $\alpha = (0, \ldots, 0)^{\mathrm{T}}$ and $\beta = (1, 2)^{\mathrm{T}}$. Then, compute the posterior inclusion probabilities for $\beta$ using Algorithm 1 with vector approximate message passing in Step 2 as described in § 3·1, and using expectation propagation and variational Bayes as in § 5·2 but with $\sigma^2$ known. Do this for $q = 1, 2, \ldots, 15$ with exact computation of the posterior inclusion probabilities as reference. For large $q$, exact computation takes too long. Therefore, use a Gibbs sampler with 10,000 burnin and 90,000 recorded iterations to compute reference posterior inclusion probabilities for $q = 15, 30, \ldots, 480, 960$. Repeat the above 20 times for each $q$.

Figs. S2 through S5 contain the results and computation times. Integrated rotated Gaussian approximation has the lowest approximation error, although the difference with expectation propagation is less pronounced in Fig. S4 as approximation error from the method and Monte Carlo error from the Gibbs sampler are mixed. Comparing $q = 15$ in Fig. S2 and Fig. S4 shows that the Monte Carlo error is of noticeable size compared to the approximation error of our method and expectation propagation. Integrated rotated Gaussian approximation deals with the fact that the columns of $X$ are correlated since $\beta$ is treated separately in Algorithm 1. Expectation propagation and variational Bayes do not make such a distinction between the elements of $\alpha$ and $\beta$. Variational Bayes consistently has the highest approximation error and the shortest computation time. Expectation propagation has a higher computation time than integrated rotated Gaussian approximation which is a result of how quickly vector approximate message passing converges in this set-up.

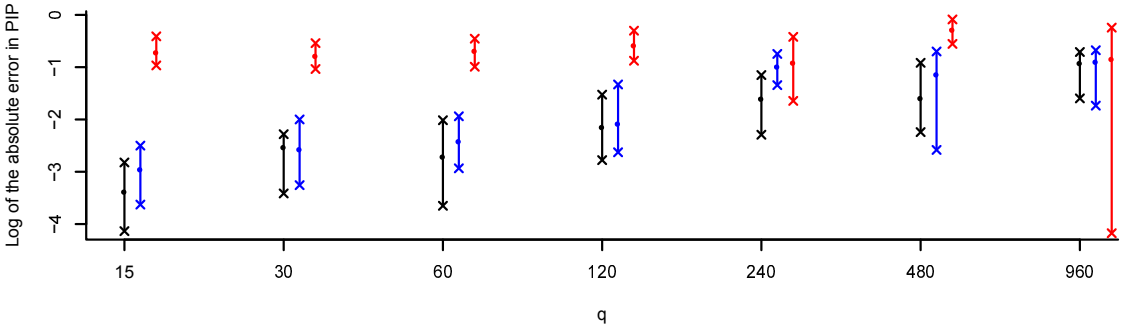


Fig. S4. Median (dot) and interquartile ranges (x) of the absolute differences between the posterior inclusion probability (PIP) approximation and their Gibbs sampler estimate from the simulation in § S8·1. Integrated rotated Gaussian approximation is in black, expectation propagation in blue, and variational Bayes in red.

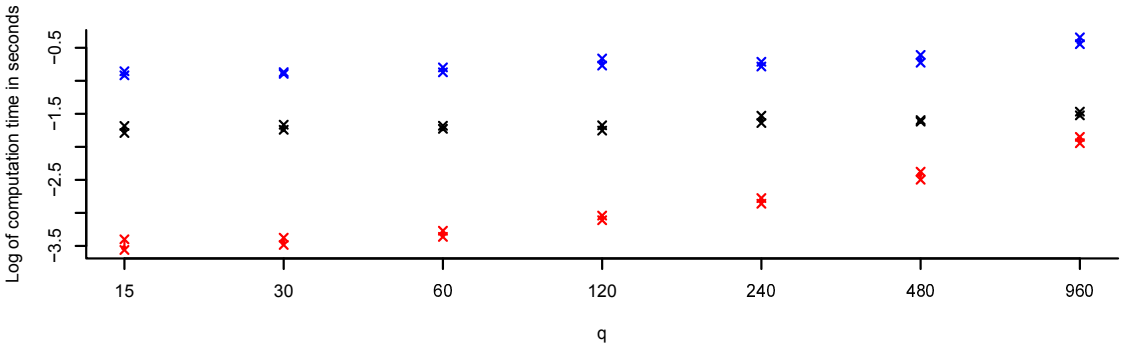


Fig. S5. Median (dot) and interquartile ranges (x) of the computation times for the results in Fig. S4. Integrated rotated Gaussian approximation is in black, expectation propagation in blue, and variational Bayes in red.

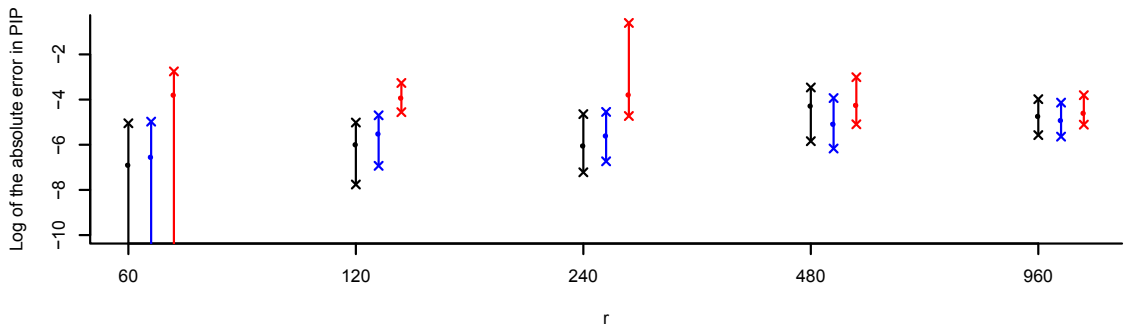


Fig. S6. Median (dot) and interquartile ranges (x) of the absolute differences between the posterior inclusion probability (PIP) approximation and their Gibbs sampler estimate from the simulation in § S8·2. Integrated rotated Gaussian approximation is in black, expectation propagation in blue, and variational Bayes in red.

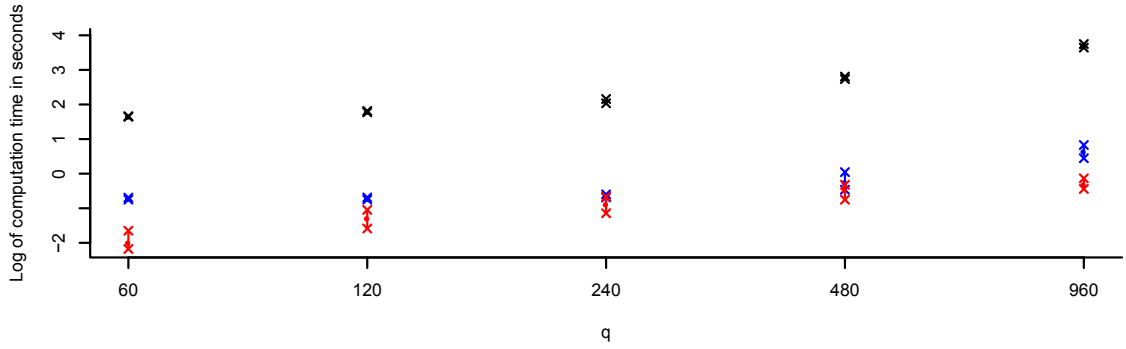


Fig. S7. Median (dot) and interquartile ranges (x) of the computation times for the results in Fig. S6. Integrated rotated Gaussian approximation is in black, expectation propagation in blue, and variational Bayes in red.

## S8·2. *Variable selection with a random design matrix*

Consider the set-up from § 3·2 with $n = 100$, $\lambda = 40/r$, $\psi = 1$, and $\sigma^2 = 1/2$. Generate $\theta$ by randomly selecting 40 elements in $\theta$ to be non-zero and drawing them from $N(0, 1)$. The elements of $A$ are drawn independently from $N(0, 1)$ after which the columns of $A$ are standardized to have zero mean and unit standard deviation. Sample $y$ according to $y \sim N(A\theta,\ \sigma^2 I_n)$. We repeat the random generation of $\theta$, $A$, and $y$ 20 times for each $r = 60, 120, 240, 480, 960$. Estimate the posterior inclusion probabilities using the same methods as in § 5·2 but with $\sigma^2 = 1/2$ known and without considering the debiased lasso. Algorithm 1 is used with $p = \lfloor \log(r) \rfloor$ as in § 5·2.

The results and computation times are in Figs. S6 and S7, respectively. There is no clear separation between the methods in terms of their approximation errors in Fig. S6. This might be a result of the smoothening effect of the Monte Carlo error which adds to the reported error as in Fig. S4. Our method seems to yield slightly more accurate posterior inclusion probabilities for $r = 60, 120, 240$, when Monte Carlo error is also lower because $r$ is smaller, albeit at a higher computational cost. The higher computational cost results from our method having to repeat Algorithm 1 $\lceil r/p \rceil$ times to obtain all $r$ posterior inclusion probabilities, while we use only $8 \ll \lceil r/p \rceil$ CPU cores for parallelization here. Expectation propagation and variational Bayes target all posterior inclusion probabilities at once.

## S8·3. *Variable selection with gene expressions*

In § S8·2, $A$ was simulated. Let us instead set $A$ equal to 3,571 expression levels from the leukemia data from Golub (1999) available in the supplementary data of Friedman et al. (2010).

Table S1. *Summary statistics of the absolute difference between the Gibbs sampler estimates and the approximations of the posterior log odds of inclusion for the simulation study in § S8·3 with median computation times. IRGA stands for integrated rotated Gaussian approximation.*

| Method | Min | Q1 | Median | Q3 | Max | Mean | Median computation time (seconds) |
|---|---|---|---|---|---|---|---|
| IRGA | 0.000 | 0.231 | 0.456 | 0.713 | 50.1 | 0.540 | 54 |
| Expectation propagation | 0.000 | 0.192 | 0.404 | 0.699 | 46.9 | 0.529 | 21 |
| Variational Bayes | 0.003 | 1.50 | 1.83 | 2.25 | 56.7 | 2.44 | 1.5 |

Then, $n = 72$ and $r = 3{,}571$. More importantly, the predictors are now highly dependent in a complex, non-linear, and non-Gaussian way: For instance, the maximum correlation between columns of $A$ equals $0.988$. The rest of the simulation, which we repeat 10 times, is the same as in § S8·2. The results are in Table S1.

Expectation propagation and our method achieve similar performance, with similar error sizes. On the other hand, expectation propagation is over twice as fast. Variational Bayes takes an order of magnitude less computation time than expectation propagation but yields worse approximations. As in § S8·2, the longer computation time of our method stems from having to repeat Algorithm 1 to obtain all posterior inclusion probabilities.

## S9. Options for splitting $\theta$ into $\alpha$ and $\beta$

### S9·1. *Motivation*

Using Algorithm 1 for Bayesian variable selection as detailed in § 3·2 requires repeatedly splitting the $r$-dimensional coefficient vector $\theta$ into the $q$-dimensional $\alpha$ and $p$-dimensional $\beta$. Different splits yield different correlation structures between the columns of $X$ and $Z$. Such correlation might affect the quality of the approximation $\hat{\Pi}(\beta \mid y)$. Therefore, one might want to choose splits that minimize the correlation between the columns of $X$ and $Z$. Also, one can average the obtained posterior inclusion probabilities over multiple splits to reduce dependence on any one splitting of $\theta$.

This section considers multiple options for splitting $\theta$. The resulting approximation accuracy of these options is empirically compared. For ease of exposition, we assume that $r$ is divisible by $p$. Then, $r/p$ splits are required to compute $\hat{\mathrm{pr}}(\theta_j \neq 0 \mid y)$ $(j = 1, \ldots, r)$ by repeating Algorithm 1. The methods are readily modified for when $r$ is not divisible by $p$ by using $\lceil r/p \rceil$ splits where in the $\lceil r/p \rceil p - r$ last splits $\beta$ is $(p-1)$-dimensional instead of $p$-dimensional.

### S9·2. *Methods of splitting $\theta$*

One method for splitting $\theta$ is sequential. The first split is $\alpha = (\theta_{p+1}, \ldots, \theta_r)^{\mathrm{T}}$ and $\beta = (\theta_1, \ldots, \theta_p)^{\mathrm{T}}$. The $k$th split is $\alpha = (\theta_1, \ldots, \theta_{(k-1)p}, \theta_{kp+1}, \ldots, \theta_r)^{\mathrm{T}}$ and $\beta = (\theta_{(k-1)p+1}, \ldots, \theta_{kp})^{\mathrm{T}}$ for $k = 2, \ldots, r/p - 1$. The final split is $\alpha = (\theta_1, \ldots, \theta_{r-p})^{\mathrm{T}}$ and $\beta = (\theta_{r-p+1}, \ldots, \theta_r)^{\mathrm{T}}$.

Splitting can also be done randomly. Sequential splitting depends on the ordering of the columns in the design matrix $A$. Instead, we can randomly permute the elements of $\theta$ and the respective columns of $A$ before splitting sequentially. This breaks the dependence on the ordering but introduces dependence on the permutation. To reduce dependence on a single permutation, one can use multiple random permutations, and then take the average of the multiple $\hat{\mathrm{pr}}(\theta_j \neq 0 \mid y)$ obtained. Here, we use 10 random permutations.

Sequential and random splitting do not minimize the correlation between the columns of $X$ and $Z$. We present two options that aim to minimize this correlation. The first option, which we call Belsley splitting, is based on Belsley et al. (1980, § 3.2) and only applies if $r \leq n$. Let $A = UDV^{\mathrm{T}}$ denote a singular-value decomposition with an $n \times r$ matrix $U$, an $r \times r$ diagonal matrix $D$ of singular values in decreasing order, and an $r \times r$ matrix $V$. Then, we compute $\phi_{kj} = V_{kj}^2 / D_{jj}^2$ and $\pi_{jk} = \phi_{kj} / \sum_{j=1}^{p} \phi_{kj}$ for $j, k = 1, \ldots, r$ as in Belsley et al. (1980, Equation 3.11). A larger $\pi_{jk}$ indicates collinearity between the $j$th and the $k$th column of $A$ if $j \neq k$ and each row of the matrix $\pi$ corresponds with a different potential near linear dependency between the columns of $A$ in decreasing order of severity as discussed in Belsley et al. (1980, § 3.2). To group collinear columns of $A$ together, the $r/p$ splits of $\theta$ are as follows. In the $j$th split, $\beta$ consists of the $p$ elements $\theta_k$ for which $\pi_{jk}$ is the largest, and $\alpha$ consists of the other elements in $\theta$.

The second option, which we call spectral splitting, is based on spectral clustering (von Luxburg, 2007) and also applies if $r > n$. Consider a weighted graph with the columns of $A$ as nodes and the absolute value of the correlation between two columns as the edge weight. Then, the $r \times r$ similarity matrix $W$ has $W_{jk}$ equal to the absolute value of the correlation between the $j$th and the $k$th column of $A$. Here, $W_{jk} = 0$ and $W_{jk} = 1$ correspond with least and most similarity, respectively, between the $j$th and the $k$th column of $A$. The $r/p$ splits of $\theta$ follow from spectral clustering using $W$. The Laplacian matrix of the graph is $L = D - W$ where $D$ is a diagonal matrix with $D_{kk} = \sum_{j=1}^{r} W_{jk}$ $(k = 1, \ldots, r)$. Let the $r \times (r/p)$ matrix $U$ consist of the first $r/p$ eigenvalues of $L$. Then, the rows of $U$ constitute $r$ points in $\mathbb{R}^{r/p}$. We cluster the points into $r/p$ clusters using $k$-means clustering. These clusters are not necessarily of equal size and $\beta$ cannot contain too many elements from $\theta$ as then evaluation of (6) is computationally expensive. Therefore, we reassign points in clusters of size greater than $r/p$ to other nearby clusters such that no cluster contains more than $r/p$ points. Each cluster corresponds with columns of $A$ and thus elements of $\theta$. Each split of $\theta$ follows by having $\beta$ consist of the elements of $\theta$ from one cluster while the other elements in $\theta$ constitute $\alpha$.

## S9·3. *Empirical comparison*

This section investigates how the various options for splitting $\theta$ affect the approximation accuracy. Firstly, consider the set-up from § 5·2. We run Algorithm 1 with vector approximate message passing, and with sequential, random, Belsley and spectral splitting. Table S2 contains the results. It shows that the approximation accuracy does not vary notably with the various methods for splitting $\theta$.

Secondly, repeat this comparison but now with the set-up from § S8·2 with $r = 60$. The results are in Table S3. Again, the approximation accuracy does not vary notably with the various methods for splitting $\theta$.

Table S2. *Summary statistics of the absolute difference between the Gibbs sampler estimates and the approximations of the posterior log odds of inclusion for the diabetes data. The approximations come from Algorithm* 1 *with different methods of splitting* $\theta$.

| Method for splitting $\theta$ | Min | Q1 | Median | Q3 | Max | Mean |
|---|---|---|---|---|---|---|
| Sequential | 0.003 | 0.036 | 0.076 | 0.133 | 10.7 | 0.599 |
| Random | 0.005 | 0.027 | 0.061 | 0.113 | 8.66 | 0.475 |
| Belsley | 0.005 | 0.034 | 0.060 | 0.140 | 12.5 | 0.588 |
| Spectral | 0.000 | 0.032 | 0.071 | 0.139 | 9.93 | 0.601 |

Table S3. *Summary statistics of the absolute difference between the Gibbs sampler estimates and the approximations of the posterior inclusion probability for data simulated as in* § *S*8·2 *with* $r = 60$*. The approximations come from Algorithm* 1 *with different methods of splitting* $\theta$*.*

| Method for splitting $\theta$ | Min | Q1 | Median | Q3 | Max | Mean |
|---|---|---|---|---|---|---|
| Sequential | 0.000 | 0.000 | 0.001 | 0.007 | 0.368 | 0.007 |
| Random | 0.000 | 0.000 | 0.000 | 0.006 | 0.123 | 0.005 |
| Belsley | 0.000 | 0.000 | 0.001 | 0.007 | 0.134 | 0.006 |
| Spectral | 0.000 | 0.000 | 0.001 | 0.006 | 0.130 | 0.006 |

Table S4. *Summary statistics of the absolute difference between the Gibbs sampler estimates and the approximations of the posterior log odds of inclusion for data simulated as in* § *S*8·3*. The approximations come from Algorithm* 1 *with different methods of splitting* $\theta$*.*

| Method for splitting $\theta$ | Min | Q1 | Median | Q3 | Max | Mean |
|---|---|---|---|---|---|---|
| Sequential | 0.000 | 0.253 | 0.500 | 0.834 | 52.9 | 0.769 |
| Random | 0.000 | 0.280 | 0.550 | 0.954 | 50.7 | 0.838 |
| Spectral | 0.000 | 0.247 | 0.494 | 0.828 | 51.8 | 0.760 |

Table S5. *Summary statistics of the absolute difference between the Gibbs sampler estimates and the approximations of the posterior log odds of inclusion for data simulated as in* § *S*8·3*. The approximations come from Algorithm* 1 *with sequential splitting using different split sizes* $p$*.*

| $p$ | Min | Q1 | Median | Q3 | Max | Mean |
|---|---|---|---|---|---|---|
| 1 | 0.000 | 0.244 | 0.472 | 0.730 | 55.2 | 0.608 |
| 2 | 0.000 | 0.242 | 0.475 | 0.735 | 55.5 | 0.632 |
| 4 | 0.000 | 0.239 | 0.466 | 0.733 | 55.1 | 0.632 |
| 8 | 0.000 | 0.229 | 0.450 | 0.711 | 54.5 | 0.588 |
| 16 | 0.000 | 0.212 | 0.422 | 0.673 | 16.6 | 0.522 |

Lastly, we run the comparison on the gene expression data from § S8·3. Here, we cannot use Belsley splitting since $r > n$ in these data. Once more, the results in Table S4 do not show notable variation in approximation accuracy across the various methods for splitting $\theta$.

### S9·4. *Choice of split size* $p$

So far, we have used $p = \lfloor \log(r) \rfloor$ as suggested by $p = O(\log r)$ from § 3·2 as a trade-off between approximation accuracy and speed. Here, we investigate how the approximation accuracy can vary with $p$. We run Algorithm 1 with sequential splitting and $p = 1, 2, 4, 8, 16$ on the gene expression data from § S8·3. Table S5 contains the results. Approximation accuracy is better at $p = 16$. However, $p$ larger than $O(\log r)$ can increase computational cost exponentially since the cost of computing (6) is exponential in $p$ for Bayesian variable selection.

## S10. Additional comparisons for § 5·3

This section provides additional results for the SNP application from § 5·3. In addition to integrated rotated Gaussian approximation and ignoring the SNPs, we consider the following methods for inference on $\beta$. 1) Expectation propagation from Hernández-Lobato et al. (2015) is readily extended to allow for different prior inclusion probabilities and slab variances per coefficient. As such, we can use it in the current set-up where the spike-and-slab prior on $\beta$ is

Table S6. *Posterior inclusion probabilities for the demographic factors from the application in § 5·3. EP and IRGA stand for integrated rotated Gaussian approximation and expectation propagation, respectively.*

| | | Population | | | |
|---|---|---|---|---|---|
| Method | Gender | Utahn of European ancestry | Finnish | Tuscan | Yoruba |
| IRGA | 0.83 | 0.96 | 0.96 | 0.92 | 0.00 |
| EP | 0.18 | 0.05 | 0.05 | 0.05 | 1.00 |
| Gibbs sampler | 0.19 | 0.05 | 0.05 | 0.06 | 1.00 |

Table S7. *Posterior mean or estimates of $\beta$ corresponding with the demographic factors from the application in § 5·3. EP and IRGA stand for integrated rotated Gaussian approximation and expectation propagation, respectively.*

| | | Population | | | | |
|---|---|---|---|---|---|---|
| Method | Gender | Utahn of European ancestry | Finnish | Tuscan | Yoruba | Computation time |
| IRGA | -0.012 | -0.001 | 0.001 | 0.004 | 0.189 | 15 seconds |
| Ignoring the SNPs | -0.083 | 0.003 | 0.000 | 0.015 | -0.050 | 82 millisecs. |
| EP | -0.014 | -0.001 | 0.002 | 0.002 | 0.336 | 21 minutes |
| Gibbs sampler | -0.015 | -0.001 | 0.002 | 0.002 | 0.323 | 5.2 days |
| Mixed effects model | -0.124 | 0.010 | 0.012 | 0.031 | -0.049 | 1.6 seconds |

different from the spike-and-slab prior on $\alpha$. 2) We run a Gibbs sampler with 10,000 burnin and 90,000 recorded iterations. 3) The model in (1) can be used as a mixed effects model with fixed effects $\beta$ and random effects $\eta$. Here, we fit a mixed effects model (Bates et al., 2015) to exemplify this interpretation of (1) and to provide a comparison with a frequentist method. The clusters for the random effects are 17 groups of individuals that are genetically distinct according to the 2,000 SNPs with the highest sure independence screening score (Fan & Lv, 2008). The mixed effects model provides estimates of $\beta$ but no posterior inclusion probabilities since it is not a Bayesian method.

The resulting posterior inclusion probabilities and estimates for $\beta$ with computation times are in Tables S6 and S7, respectively. The results from the Gibbs sampler and expectation propagation are consistent and different from our method. Either our method's approximations are inaccurate, those from the Gibbs sampler and expectation propagation are, or they all are. The Gibbs sampler and expectation propagation being consistent suggests that our method failed to give accurate approximations for this posterior. Though, it is also possible that both the Gibbs sampler and the expectation propagation struggle with the high-dimensional posterior in such a way that results in similar but inaccurate approximations. Our method requires substantially less computation time than expectation propagation and the Gibbs sampler.

The mixed effects model yields estimates that are not aligned with the posterior means from any of the Bayesian options. This is unsurprising since the mixed effects model is adjusting for the SNPs in a fundamentally different manner.

## REFERENCES

BATES, D., MÄCHLER, M., BOLKER, B. & WALKER, S. (2015). Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* **67**, 1–48.

BELSLEY, D. A., KUH, E. & WELSCH, R. E. (1980). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. Wiley-Interscience.

BISHOP, C. M. (2006). *Pattern Recognition and Machine Learning*. New York: Springer.
BONTEMPS, D. (2011). Bernstein–von Mises theorems for Gaussian regression with increasing number of regressors. *Ann. Stat.* **39**, 2557–2584.
COVER, T. M. & THOMAS, J. A. (2006). *Elements of Information Theory*. Wiley Series in Telecommunications and Signal Processing. Wiley-Interscience, 2nd ed.
DOWSON, D. & LANDAU, B. (1982). The Fréchet distance between multivariate normal distributions. *J. Multivar. Anal.* **12**, 450–455.
EFRON, B., HASTIE, T., JOHNSTONE, I. & TIBSHIRANI, R. (2004). Least angle regression. *Ann. Stat.* **32**, 407–499.
FAN, J. & LV, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *J. R. Statist. Soc.* B **70**, 849–911.
FERNÁNDEZ, C., LEY, E. & STEEL, M. F. (2001). Benchmark priors for Bayesian model averaging. *J. Econom.* **100**, 381–427.
FRIEDMAN, J., HASTIE, T. & TIBSHIRANI, R. (2010). Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33**.
GOLUB, T. R. (1999). Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science* **286**, 531–537.
HERNÁNDEZ-LOBATO, J. M., HERNÁNDEZ-LOBATO, D. & SUÁREZ, A. (2015). Expectation propagation in linear regression models with spike-and-slab priors. *Mach. Learn.* **99**, 437–487.
JAVANMARD, A. & MONTANARI, A. (2013). Confidence intervals and hypothesis testing for high-dimensional statistical models. In *Advances in Neural Information Processing Systems 26*. pp. 1187–1195.
KULLBACK, S. & LEIBLER, R. A. (1951). On information and sufficiency. *Ann. Math. Stat.* **22**, 79–86.
LIANG, F., PAULO, R., MOLINA, G., CLYDE, M. A. & BERGER, J. O. (2008). Mixtures of $g$ priors for Bayesian variable selection. *J. Am. Statist. Assoc.* **103**, 410–423.
RANGAN, S., SCHNITER, P. & FLETCHER, A. (2014). On the convergence of approximate message passing with arbitrary matrices. In *IEEE International Symposium on Information Theory*. pp. 236–240.
RANGAN, S., SCHNITER, P. & FLETCHER, A. K. (2016). Vector approximate message passing. arXiv:1610.03082v2.
REEVES, G. (2017). Conditional central limit theorems for Gaussian projections. In *IEEE International Symposium on Information Theory*. pp. 3045–3049.
RENCHER, A. C. & SCHAALJE, G. B. (2008). *Linear Models in Statistics*. Hoboken: Wiley-Interscience, 2nd ed.
STEINBERG, D. M. & BONILLA, E. V. (2014). Extended and unscented Gaussian processes. In *Advances in Neural Information Processing Systems 27*. pp. 1251–1259.
VILA, J. & SCHNITER, P. (2011). Expectation-maximization Bernoulli-Gaussian approximate message passing. In *45th Asilomar Conference on Signals, Systems and Computers*. pp. 799–803.
VON LUXBURG, U. (2007). A tutorial on spectral clustering. *Stat. Comput.* **17**, 395–416.