Computational Statistics
https://doi.org/10.1007/s00180-021-01135-x

ORIGINAL PAPER

A fast regression via SVD and marginalization

Philip Greengard[1] · Andrew Gelman[1] · Aki Vehtari[2]

Received: 23 February 2021 / Accepted: 10 July 2021

Abstract

We describe a numerical scheme for evaluating the posterior moments of Bayesian linear regression models with partial pooling of the coefficients. The principal analytical tool of the evaluation is a change of basis from coefficient space to the space of singular vectors of the matrix of predictors. After this change of basis and an analytical integration, we reduce the problem of finding moments of a density over $k + 2$ dimensions, to finding moments of a 2-dimensional density, where k is the number of coefficients. Moments can then be computed using, for example, MCMC, the trapezoid rule, or adaptive Gaussian quadrature. An evaluation of the SVD of the matrix of predictors is the dominant computational cost and is performed once during the precomputation stage. We demonstrate numerical results of the algorithm.

Keywords Bayesian Regression · Singular Value Decomposition · Marginalization · Fast Algorithms

1 Introduction

Linear regression is a ubiquitous tool for statistical modeling in a range of applications including social sciences, epidemiology, biochemistry, and environmental sciences (Gelman et al. 2013; Gelman and Hill 2007; Greenland 2000; Merlo et al. 2005; Bardini et al. 2017).

A common bottleneck for applied statistical modeling workflow is the computational cost of model evaluation. Since posterior distributions in statistical models are often high dimensional and computationally intractable, various techniques have been used to approximate posterior moments. Standard approaches often involve a variety of techniques including Markov chain Monte Carlo (MCMC) or using a suitable approximation of the posterior.

✉ Philip Greengard
pg2118@columbia.edu

1 Columbia University, New York, USA

2 Aalto University, Espoo, Finland

SPI Journal: 180 Article No.: 1135 TYPESET DISK LE CP Disp.:2021/7/29 Pages: 20 Layout: Small-Ex

In this paper, we describe an approach for reducing the computational costs for a particular class of regression models — those that contain parameters $\theta \in \mathbb{R}^k$ such that θ has a normal prior and normal likelihood. These models represent only a subset of regression models that appear in applications. We focus our attention in this paper on normal-normal models because they have well known analytical properties and are more computationally tractable than the vast majority of multilevel models. A broader class of models, including logistic regression, contain distributions that are less amenable to the techniques of this paper and will require other analytical and computational tools. Mathematically, marginalization of normal-normal parameters is well-known and has been applied to the posterior by, for example, Lindley and Smith (1972). Our contribution is to provide a stable, accurate, and fast algorithm for marginalization.

The primary numerical tool used in the algorithm is the singular value decomposition (SVD) of the data matrix. As a mathematical and statistical tool, SVD has been known since at least 1936 (see Eckart and Young (1936)). Use of the SVD as a practical and efficient numerical algorithm only started gaining popularity much later, with the first widely used scheme introduced in Golub and Kahan (1965). Due in large part to advances in computing power, use of the SVD as a tool in applied mathematics, statistics, and data science has been gaining significant popularity in recent years, however efficient evaluation of SVDs and related matrix decompositions is still an active area of research (see Hastie et al. 2015; Halko et al. 2011; Shamir et al. 2016).

Similar schemes to ours are used in the software packages lme4 (Bates et al. 2015) and INLA (Rue et al. 2017). There are several differences between the problems they address and their computational techniques, and those that we shall discuss here. While lme4 finds maximum likelihood and restricted maximum likelihood estimates, our goal is to find posterior moments. The software package INLA uses Laplace approximation on the posterior for a general choice of likelihood functions, whereas our algorithm is focused on fast and accurate solutions for only a particular class of densities: those with normal-normal parameters.

The approach presented in this paper analytically marginalizes the normal-normal parameters of a model using a change of variables. After marginalization, posterior moments can be computed using standard techniques on the lower-dimensional density. In particular, for a model that contains $k+2$ total variables, k of which are normal-normal, our scheme converts the problem of evaluating expectations of a density in $k+2$ dimensions to finding expectations of a 2-dimensional density. After marginalization, we evaluate the 2-dimensional posterior density in $O(k)$ operations.

We illustrate our scheme on the problem of evaluating the marginal expectations of the unnormalized density

$$q(\sigma_1, \sigma_2, \beta) = \sigma_1^{-(k+1)} \sigma_2^{-n} \exp\Bigg(-\gamma(\log(\sigma_1))^2 - \frac{\sigma_2^2}{2} - \frac{\|X\beta - y\|^2}{2\sigma_2^2} - \frac{\|\beta\|^2}{2\sigma_1^2} \Bigg), \tag{1}$$

SPI Journal: 180 Article No.: 1135 TYPESET DISK LE CP Disp.:2021/7/29 Pages: 20 Layout: Small-Ex

where $\gamma > 0$ is a constant, $\sigma_1, \sigma_2 > 0$, and $\beta \in \mathbb{R}^k$. We assume that X is a fixed $n \times k$ matrix, $y \in \mathbb{R}^n$ is fixed, and the normalizing constant of (1) is unknown. For fixed $n, k \in \mathbb{N}$, the algorithm is nearly identical when X is an $n \times k$ matrix to when X is a $k \times n$ matrix. In the case where $k \gg n$, Kwon et al. (2011) also use SVD for marginalization. There are three main distinctions between their method and ours. (i) Our method applies to $n \times k$ matrices X for $k < n$ and $k > n$. (ii) We use the SVD to analytically compute conditional second moments with respect to β, not only first moments. (iii) While they use MCMC for computing posterior moments, we use a high-order quadrature scheme.

Using the standard notation of Bayesian models, density q is the unnormalized posterior of the model

$$\begin{aligned} \sigma_1 &\sim \text{lognormal}(0, \sqrt{\gamma}) \\ \sigma_2 &\sim \text{normal}^+(0, 1) \\ \beta &\sim \text{normal}(0, \sigma_1) \\ y &\sim \text{normal}(X\beta, \sigma_2). \end{aligned} \tag{2}$$

In Appendix A, we include Stan code that can be used to sample from density (1) using MCMC. We also include Stan code that samples from the marginalized 2-dimensional posterior obtained via the algorithm of this paper.

Statistical model (2) is a standard model of Bayesian statistics and appears when seeking to model an outcome, y, as a linear combination of related predictors, the columns of X. In Gelman and Hill (2007), these models are described in detail and are used in the estimation of the distribution of radon levels in houses in Minnesota. See (Dias et al. 2013; Rover et al. 2020) for further examples.

Density (1) is also closely related to posterior densities that appear in genome-wide association studies (GWAS; see Zhu and Stephens 2017; Meuwissen, et al. 2001; Azevedo et al. 2015) which can be used to identify genomic regions containing genes linked with a specific trait, such as height. Using the notation of (1), each row of matrix X corresponds to a person, each column of X represents a genomic location, entries of X indicate genotypes, and y corresponds to the trait. Due to technical advances in genome sequencing over the last ten years, it is now feasible to collect large amounts of sequencing data. GWAS models can contain data on up to millions of people and often between hundreds and thousands of genome locations (see Linner et al. 2019). As a result, efficient computational tools are required for model evaluation.

The number of operations required by the scheme of this paper scales like $O(nk^2)$ with a small constant. The key analytical tool is a change of variables of β such that the terms,

$$-\frac{1}{2\sigma_2^2}\|X\beta - y\|^2 - \frac{1}{2\sigma_1^2}\|\beta\|^2, \tag{3}$$

in (1) are converted to a diagonal quadratic form in $\mathbb{R}^k$. After that change of variables, expectations over q are analytically converted from integrals over $\mathbb{R}^{k+2}$ to integrals over $\mathbb{R}^2$. The remaining 2-dimensional integrals can be computed to high accuracy using classical numerical techniques including, for example, adaptive Gaussian quadrature or even the 2-dimensional trapezoid rule.

SPI Journal: 180 Article No.: 1135 TYPESET DISK LE CP Disp.:2021/7/29 Pages: 20 Layout: Small-Ex

The tools used in this paper to evaluate the expectations of (1) can also be used in the evaluation of expectations of multilevel and multigroup posterior distributions including, for example, the two-group posterior of the form,

$$q(\sigma_1, \sigma_2, \sigma_3, \beta) = \exp\Bigg(-\frac{1}{2\sigma_1^2}\|X\beta - y\|^2 - \frac{1}{2\sigma_2^2}\sum_{i=1}^{k_1}\beta_i^2 - \frac{1}{2\sigma_3^2}\sum_{i=k_1+1}^{k_1+k_2}\beta_i^2\Bigg), \tag{4}$$

where X is a $n \times k$ matrix, $y \in \mathbb{R}^n$, k_1 and k_2 are non-negative integers satisfying $k_1 + k_2 = k$, and $\sigma_1, \sigma_2, \sigma_3 > 0$.

The structure of this paper is as follows. In the following section we describe the analytic integration that transforms (1) from a $k + 2$-dimensional problem to a 2-dimensional problem. Section 3 includes formulas that will allow for the evaluation of posterior moments using the 2-dimensional density. In Sects. 4 and 5 we provide formulas for evaluating covariances of (1). In Sect. 6, we discuss the numerical results of the implementation of the algorithm. Conclusions and generalizations of the algorithm of this paper are presented in Sect. 7. Appendix A provides Stan code that can be used to sample from (1), and Appendix B includes proofs of the formulas of this paper.

2 Analytic integration of β

In this section, we describe how we analytically marginalize the normal-normal parameter β of density (1). We include proofs of all formulas in Appendix B.

We start by in marginalizing β using a change of variables that converts the quadratic forms in (1) into diagonal quadratic forms. The resulting integral in the new variable, z, is Gaussian, and the coefficients of z_i and z_i^2 are available analytically. The change of variables is given by the right orthogonal matrix of the singular value decomposition (SVD) of X. That is, we set

$$z = V^t\beta \tag{5}$$

where the SVD of X is

$$X = UDV^t. \tag{6}$$

We define λ_i to be the i^{th} element of the diagonal of D. The elements of diagonal need not be sorted. After this change of variables, we obtain the following identity for the last two terms of (1). A proof can be found in Lemma 5 in Appendix B.

Formula 2.1

$$-\frac{1}{2\sigma_2^2}\|X\beta - y\|^2 - \frac{1}{2\sigma_1^2}\|\beta\|^2 = a_0 + \sum_{i=1}^{k} a_{2,i}\left(z_i - \frac{a_{1,i}}{2a_{2,i}}\right)^2 + \frac{a_{1,i}^2}{4a_{2,i}} \tag{7}$$

SPI Journal: 180 Article No.: 1135 TYPESET DISK LE CP Disp.:2021/7/29 Pages: 20 Layout: Small-Ex

where

$$a_{2,i} = \frac{\lambda_i^2}{2\sigma_2^2} + \frac{1}{2\sigma_1^2}, \tag{8}$$

$$a_{1,i} = \frac{w_i}{\sigma_2^2}, \tag{9}$$

and

$$a_0 = -\frac{y^t y}{2\sigma_2^2} \tag{10}$$

where

$$w = V^t X^t y. \tag{11}$$

After performing the change of variables $z = V^t \beta$ and using (7), we now have an expression for density (1) in a form that allows us to use the well-known properties of a Gaussian with diagonal covariance. The following identity uses these properties and provides a formula for analytically reducing expectations of (1) from integrals over $k+2$ dimensions to integrals over 2 dimensions. After the formula is applied, we have a new density, $\tilde{q}$, over only 2 dimensions. See Theorem 1 in Appendix B for a proof.

Formula 2.2 *For all* $\sigma_1, \sigma_2 > 0$ *we have*

$$\int_{\mathbb{R}^k} q(\sigma_1, \sigma_2, \beta) d\beta = \tilde{q}(\sigma_1, \sigma_2) \tag{12}$$

where $\tilde{q}(\sigma_1, \sigma_2)$ *is defined by the formula*

$$\tilde{q}(\sigma_1, \sigma_2) = \sigma_1^{-(k+1)} \sigma_2^{-n} \exp\Bigg(-\gamma \log^2(\sigma_1) - \frac{\sigma_2^2}{2} + a_0 + \sum_{i=1}^{k} \frac{a_{1,i}^2}{4a_{2,i}} \Bigg) \prod_{i=1}^{k} \frac{1}{\sqrt{2a_{2,i}}} \tag{13}$$

where $a_{2,i}$ *is defined in (8),* $a_{1,i}$ *is defined in (9),* a_0 *is defined in (10), and* γ *is a constant.*

In (58) we provide a formula for $\tilde{q}$ in the case where both scale parameters have half-normal priors.

Remark 1 Certain Bayesian models might contain correlated priors on β that will result in posteriors such as (28) of Sect. 4. For such models, we perform the change of variables that uses the fact that two diagonal forms over β can be simultaneously diagonalized.

We include in Fig. 1 a plot of the density of q as a function of σ_1 and β_1 for fixed σ_2 and randomly chosen X and y. Figure 2 shows a plot of q as a function of σ_2 and β for fixed σ_1. Figure 3 provides an illustration of $\tilde{q}$, obtained after the change of variables and marginalization described in this section.

SPI Journal: 180 Article No.: 1135 TYPESET DISK LE CP Disp.:2021/7/29 Pages: 20 Layout: Small-Ex

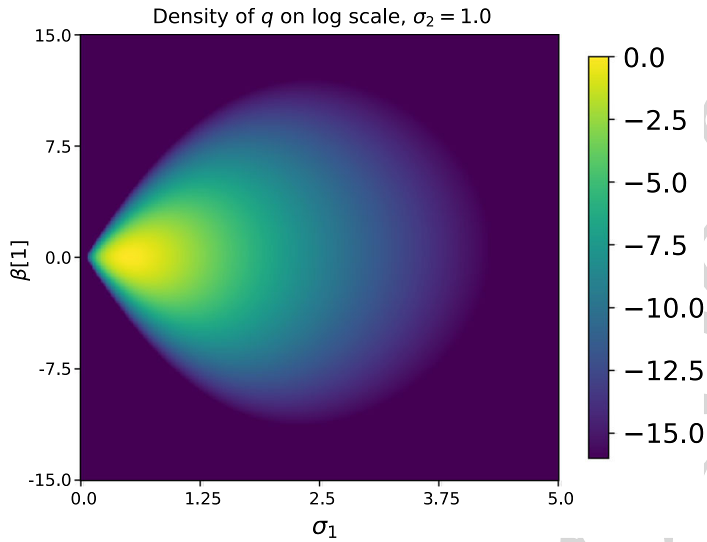

Fig. 1 Density of q (see (1)) with respect to σ_1 and β_1, where $\gamma = 8$, $n = 100$, $k = 10$, and data were randomly generated

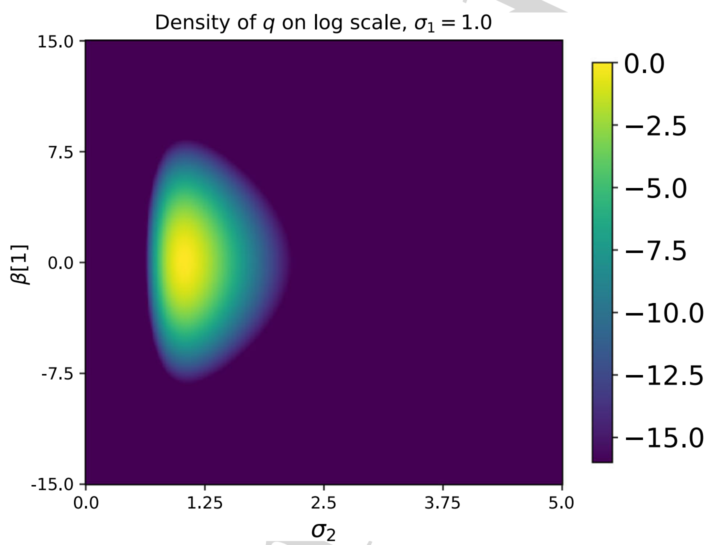

Fig. 2 Density of q (see (1)) with respect to σ_2 and β_1, where $\gamma = 8$, $n = 100$, $k = 10$, and data were randomly generated

3 Evaluation of posterior means

Now that we have reduced the $k+2$-dimensional density q to the 2-dimensional density $\tilde{q}$, it remains to recover the posterior moments of q using $\tilde{q}$. We first observe that moments of σ_1 and σ_2 with respect to q are equivalent to moments of σ_1 and σ_2 over $\tilde{q}$. That is,

$$\mathbb{E}_q(\sigma_1) = \mathbb{E}_{\tilde{q}}(\sigma_1) \tag{14}$$

and

$$\mathbb{E}_q(\sigma_2) = \mathbb{E}_{\tilde{q}}(\sigma_2). \tag{15}$$

SPI Journal: 180 Article No.: 1135 TYPESET DISK LE CP Disp.:2021/7/29 Pages: 20 Layout: Small-Ex

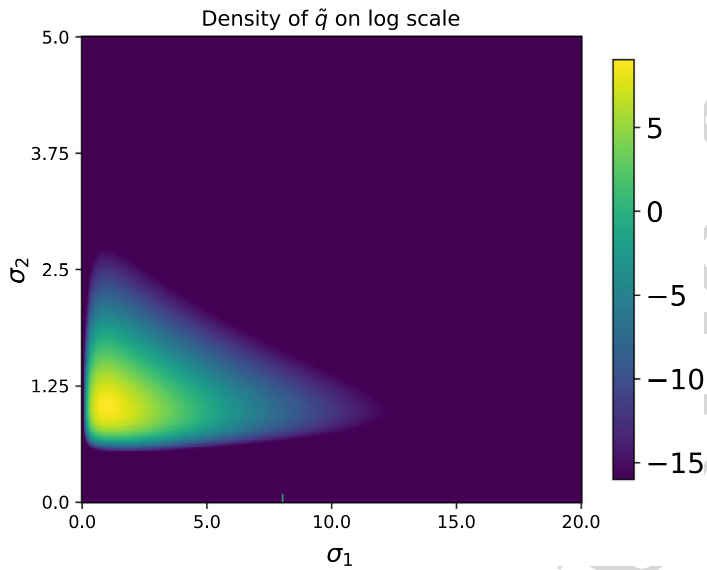

Fig. 3 Log density of $\tilde{q}$ (see (13)) using the same q as Fig. 1, where $n = 100$, $k = 10$, and data were randomly generated

As for moments of β, we use (13) and standard properties of Gaussians to obtain the following formula.

Formula 3.1 *For all* $\sigma_1, \sigma_2 > 0$,

$$\int_{\mathbb{R}^k} z_i q(\sigma_1, \sigma_2, \beta) d\beta = \frac{a_{1,i}}{2a_{2,i}} \tilde{q}(\sigma_1, \sigma_2) \tag{16}$$

where q *is defined in (1),* $\tilde{q}$ *is defined in (13),* $a_{2,i}$ *is defined in (8), and* $a_{1,i}$ *is defined in (9).*

As an immediate consequence of (16), we are able to evaluate the posterior expectation of z as an expectation of a 2-dimensional density:

$$\mathbb{E}_q(z_i) = \mathbb{E}_{\tilde{q}}\left(\frac{a_{1,i}}{2a_{2,i}}\right). \tag{17}$$

We then transform those expectations back to expectations over the desired basis, β using the matrix V computed in (6). Specifically, using linearity of expectation and (17), we know

$$\begin{aligned}
\mathbb{E}_q((\beta_1, \ldots, \beta_k)^t) &= \mathbb{E}_q(VV^t(\beta_1, \ldots, \beta_k)^t) \\
&= V\mathbb{E}_q(V^t(\beta_1, \ldots, \beta_k)^t) \\
&= V\mathbb{E}_q((z_1, \ldots, z_k)^t) \\
&= V\mathbb{E}_{\tilde{q}}\left(\left(\frac{a_{1,1}}{2a_{2,1}}, \ldots, \frac{a_{1,k}}{2a_{2,k}}\right)^t\right). \qquad (18)
\end{aligned}$$

SPI Journal: 180 Article No.: 1135 TYPESET DISK LE CP Disp.:2021/7/29 Pages: 20 Layout: Small-Ex

4 Covariance of β

In addition to facilitating the rapid evaluation of posterior means, the change of variables described in Sect. 2 is also useful for the evaluation of higher moments.

Equation (7) shows that after the change of variables from β to z, the resulting density is a Gaussian in z with a diagonal covariance matrix. Additionally, for each z_i, using Eq. (7) and standard properties of Gaussians, we have the following identity.

Formula 4.1 *For all $\sigma_1, \sigma_2 > 0$, we have*

$$\int_{\mathbb{R}^k} (z_i - \mu_{z_i})^2 q(\sigma_1, \sigma_2, \beta) d\beta = (2a_{2,i})^{-1} \tilde{q}(\sigma_1, \sigma_2) \tag{19}$$

where μ_{z_i} is the expectation of z_i, $\tilde{q}$ is defined in (13), and $a_{2,i}$ is defined in (8).

The second moments of the posterior of β are obtained as a linear transformation of the posterior variances of z. In particular, denoting the expectation of z by μ_z, we have

$$\begin{aligned} \mathbb{E}(\beta\beta^t) &= VV^t \mathbb{E}(\beta\beta^t) VV^t \\ &= V\mathbb{E}(zz^t)V^t \\ &= V(E((z-\mu_z)(z-\mu_z)^t) + \mu_z\mu_z^t)V^t \end{aligned} \tag{20}$$

We observe that due to the independence of all z_i,

$$\mathbb{E}((z-\mu_z)(z-\mu_z)^t) \tag{21}$$

is diagonal and we can therefore evaluate the $k \times k$ posterior covariance matrix of β by evaluating $\operatorname{var}(z_i)$ and μ_{z_i} for $i = 1, \ldots, k$ and then applying two orthogonal matrices. Specifically, combining Formula 4.1, (17), and (20), we obtain

$$\begin{aligned} \operatorname{cov}(\beta) = {} & V\mathbb{E}_{\tilde{q}}\left(\left((2a_{2,1})^{-1}, \ldots, (2a_{2,k})^{-1}\right)^t\right)V^t \\ & + VE_{\tilde{q}}\left(\frac{a_{1,i}}{2a_{2,i}}\right)E_{\tilde{q}}\left(\frac{a_{1,i}}{2a_{2,i}}\right)^t V^t - \mu_\beta\mu_\beta^t. \end{aligned} \tag{22}$$

5 Variance of σ_1 and σ_2

Higher moments of σ_1 and σ_2 with respect to q can be evaluated directly as higher moments of σ_1 and σ_2 with respect to $\tilde{q}$. That is, for all $j \in \{2, 3, \ldots, \}$, we have

$$\mathbb{E}_q((\sigma_1 - \mu_{\sigma_1})^j) = \mathbb{E}_{\tilde{q}}((\sigma_1 - \mu_{\sigma_1})^j) \tag{23}$$

and

$$\mathbb{E}_q((\sigma_2 - \mu_{\sigma_2})^j) = \mathbb{E}_{\tilde{q}}((\sigma_2 - \mu_{\sigma_2})^j). \tag{24}$$

In particular, for $j = 2$, we obtain

$$\mathrm{var}_q(\sigma_1) = \mathrm{var}_{\tilde{q}}(\sigma_1) \tag{25}$$

and

$$\mathrm{var}_q(\sigma_2) = \mathrm{var}_{\tilde{q}}(\sigma_2). \tag{26}$$

Algorithm 1: *Evaluation of posterior expectations of normal-normal models*

Compute SVD of matrix X
Compute w (see (11))
Compute $V^t \mathbb{1}$ (see (9))
Construct evaluator for density $\tilde{q}$ of (13)
Evaluate first and second moments with respect to $\tilde{q}$: $\mathbb{E}_{\tilde{q}}(\sigma_1)$, $\mathbb{E}_{\tilde{q}}(\sigma_2)$, $\mathbb{E}_{\tilde{q}}(\frac{a_{1,i}}{2a_{2,i}})$
Compute $\mathbb{E}(\beta)$ via formula (18)

6 Numerical experiments

Algorithm 1 was implemented in Fortran. We used the GFortran compiler on a 2.6 GHz 6-Core Intel Core i7 MacBook Pro. All examples were run in double precision arithmetic. The matrix X and vector y were randomly generated as follows. Each entry of X was generated with an independent Gaussian with mean 0 and variance 1. The vector y was created by first randomly generating a vector $\beta \in \mathbb{R}^k$, each entry of which is an independent Gaussian with mean 0 and variance 1. The vector y was set to the value of $X\beta + \epsilon$ where $\epsilon \in \mathbb{R}^n$ contains standard normal iid entries. We generated y this way in order to ensure that the $\mathbb{E}(\beta_i)$ were not all small in magnitude. We set γ of (1) to 8 for all subsequent experiments and note that in practice the value of γ would be set according to some problem-specific knowledge.

In Table 1 and Fig. 5, we compare the performance of Algorithm 1 to two alternative schemes for computing posterior expectations — one in which we analytically marginalize via Eq. (12) and then integrate the 2-dimensional density via MCMC using Stan. In the other, we use Stan's MCMC sampling on the full $k + 2$ dimensional posterior. When using MCMC with Stan, we took 10,000 posterior draws. In Table 1 and Fig. 5 we denote Algorithm 1 by "SVD-Trap". The algorithm that uses Stan on the marginal 2-dimensional density is labeled "SVD-MCMC", and "MCMC" corresponds to the algorithm that uses only MCMC sampling in Stan. We observe that both the time for evaluation and the accuracy is drastically improved when using Algorithm 1 over full MCMC and MCMC with marginalization. In particular, for large n, the algorithm of this paper is faster by a factor of thousands compared to full MCMC via Stan.

In the appendix, we include Stan code to sample from the marginal density $\tilde{q}$ of (13).

SPI Journal: 180 Article No.: 1135 TYPESET DISK LE CP Disp.:2021/7/29 Pages: 20 Layout: Small-Ex

Table 1 Accuracy of evaluation of expectations of q (see (1)) using three different algorithms: (i) SVD-Trap: Algorithm 1 of this paper, (ii) SVD-MCMC: marginalization with MCMC integration of $\tilde{q}$ using Stan, and (iii) MCMC: full MCMC integration of q using Stan

n	k	SVD-Trap max error	SVD-MCMC max error	MCMC max error
100	100	0.9×10^{-14}	0.4×10^{-4}	0.1×10^{-1}
200	100	0.9×10^{-14}	0.3×10^{-2}	0.8×10^{-2}
500	100	0.9×10^{-13}	0.2×10^{-2}	0.8×10^{-2}
1000	100	0.2×10^{-13}	0.6×10^{-3}	0.7×10^{-2}
5000	100	0.4×10^{-13}	0.2×10^{-3}	0.3×10^{-2}
10,000	100	0.2×10^{-13}	0.4×10^{-3}	0.2×10^{-2}

Table 2 Scaling of computation times for evaluation of expectations of q (see (1)) using Algorithm 1

n	k	max error	precompute time (s)	integrate time (s)	total (s)
50	5	0.22×10^{-13}	0.01	0.01	0.02
100	10	0.26×10^{-13}	0.02	0.01	0.03
500	20	0.30×10^{-13}	0.04	0.01	0.05
1000	50	0.34×10^{-13}	0.09	0.03	0.12
5000	100	0.37×10^{-13}	0.29	0.05	0.34
10000	500	0.26×10^{-13}	14	0.3	14.2
10,000	1000	0.39×10^{-13}	54	0.6	54.5

Remark 2 In the numerical integration stage of algorithm 1, we use the trapezoid rule with 200 nodes in each direction. See Sect. C for a brief description of the 2-dimensional trapezoid rule. Because the integrand is smooth and vanishes near the boundary, convergence of the integral is super-algebraic when using the trapezoid rule (see Stoer and Bulirsch 1992). A rectangular grid with 200 points in each direction is satisfactory for obtaining approximately double precision accuracy. In problems with large numbers of non-normal-normal parameters, MCMC algorithms such as Hamiltonian Monte Carlo or other methods can be used.

In Tables 1 and 2, n and k represent the size of the $n \times k$ random matrix X.

The column labeled "max error" provides the maximum absolute error of the expectations of σ_1, σ_2, and β_i for $i \in \{1, 2, \ldots, n\}$. The true solution was evaluated using trapezoid rule with 500 nodes in each direction in extended precision.

In Table 2, "Precompute time (s)" denotes the time in seconds of all computations until numerical integration. These times are dominated by the cost of SVD (36). The total time of the numerical integration in addition to the matrix-vector product (18) is given in "integrate time (s)." The final column of Table 2, "total time (s)", provides the total time of precomputation and integration.

Notably, Table 2 demonstrates that the dominant cost of the algorithm of this paper is the SVD in the precomputation stage. Additionally, even for large regression problems with 10,000 observations and 1000 predictors, evaluation time is under a minute.

SPI Journal: 180 Article No.: 1135 TYPESET DISK LE CP Disp.:2021/7/29 Pages: 20 Layout: Small-Ex

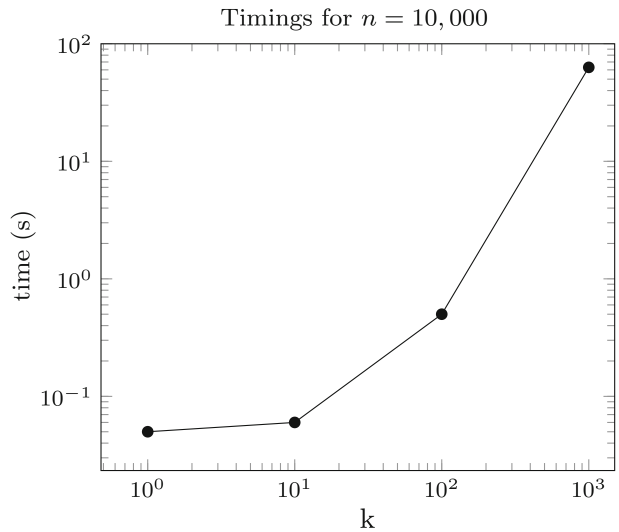

Fig. 4 Scaling of computation times for evaluation of posterior expectations of q (see (1)) using Algorithm 1 as a function of k with $n = 10{,}000$

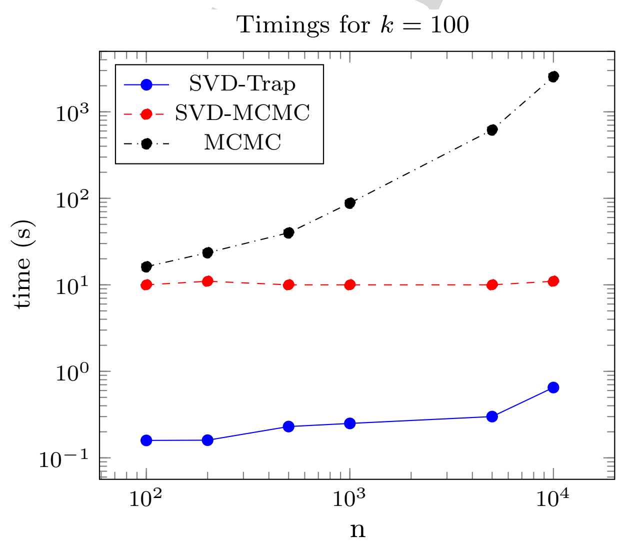

Fig. 5 Scaling of computation times for evaluation of posterior expectations of q (see (1)) as a function of sample size n with $k = 100$. The three algorithms comapred are (i) SVD-Trap: Algorithm 1 of this paper, (ii) SVD-MCMC: marginalization with MCMC integration of $\tilde{q}$ using Stan, and (iii) MCMC: full MCMC integration of q using Stan

7 Generalizations and conclusions

In this paper, we present a numerical scheme for the evaluation of the expectations of a particular class of distributions that appear in Bayesian statistics; posterior dsitributions of linear regression problems with normal-normal parameters.

The tools used in the numerical scheme of this paper generalize to several related classes of distributions that appear frequently in Bayesian statistics. We list several examples of posterior densities whose expectations can be evaluated using this method.

1. The choice of priors for σ_1, and σ_2 in this document were log normal and half-normal. This choice did not substantially impact the algorithm and can be generalized. Adaptive Gaussian quadrature (see, e.g. Trefethen 2020) can be used for the numerical integration step of the algorithm for a more general choice of prior on σ_1 and σ_2.

SPI Journal: 180 Article No.: 1135 TYPESET DISK LE CP Disp.:2021/7/29 Pages: 20 Layout: Small-Ex

2. Regression problems with multiple groups such as the two-group model with posterior

$$\exp\left(-\frac{1}{2\sigma_1^2}\|X\beta - y\|^2 - \frac{1}{2\sigma_2^2}\sum_{i=1}^{k_1}\beta_i^2 - \frac{1}{2\sigma_3^2}\sum_{i=k_1+1}^{k_1+k_2}\beta_i^2\right) \tag{27}$$

where X is a $n \times k$ matrix, $y \in \mathbb{R}^n$, and k_1 and k_2 are non-negative integers satisfying $k_1 + k_2 = k$.

3. Regression problems with correlated priors on β:

$$\exp\left(-\frac{1}{2\sigma_2^2}\|X_1\beta - y\|^2 - \frac{1}{2\sigma_1^2}\|X_2\beta\|^2\right) \tag{28}$$

For regression problems with large numbers of non-normal-normal parameters, marginal expectations can be computed using, for example, MCMC in Stan. For such problems, the algorithm of this paper would convert an MCMC evaluation from $k+m$ dimensions to m dimensions, where k is the number of normal-normal parameters.

Acknowledgements The authors are grateful to Ben Bales, Bob Carpenter, and Mitzi Morris for many useful discussions.

A Code

The following Stan code allows for sampling from the distribution corresponding to the probability density function proportional to (1).

```
data {
  int n;
  int k;
  vector[n] y;
  matrix[n,k] X;
}
parameters {
  real<lower=0> sigma1;
  real<lower=0> sigma2;
  vector<offset=0, multiplier=sigma1>[k] beta;
}
model {
  y ~ normal(X*beta, sigma2);
  beta ~ normal(0, sigma1);
  sigma1 ~ lognormal(0, 0.25);
  sigma2 ~ normal(0, 1);
}
```

The following Stan program samples from the marginal density $\tilde{q}$ (see (13)). The data input `yty` corresponds to $y^t y$ of (10), `lam` is the vector of singular values of X,

SPI Journal: 180 Article No.: 1135 TYPESET DISK LE CP Disp.:2021/7/29 Pages: 20 Layout: Small-Ex

and `w` is the vector w in Eq. (11). We include R code for computing `yty`, `lam`, and `w` after the following Stan code.

```
functions {
  real q_tilde_lpdf(real sig1, real sig2, vector w,
                    vector lam, real yty, int k,
                    int n) {
    vector[min(n,k)] a2 = lam^2/(sig2^2) + 1/(sig1^2);
    real sol = sum(w^2 ./a2)/2/sig2^4 - sum(log(a2))/2
      -yty/(2*sig2^2);
    sol += -min(n,k)*log(sig1) - n*log(sig2);
    return sol;
  }
}
data {
  int n;
  int k;
  vector[min(n,k)] w;
  vector[min(n,k)] lam;
  real yty;
  matrix[min(n,k),k] V;
}
parameters {
  real<lower=0> sigma1;
  real<lower=0> sigma2;
}
model {
  sigma1 ~ q_tilde(sigma2, w, lam, yty, k, n);
  sigma1 ~ lognormal(0, 0.25);
  sigma2 ~ normal(0, 1);
}
generated quantities {
 vector[k] beta;
 {
   vector[min(n,k)] zvar = 1 ./(2*(lam^2 ./(2*sigma2^2)
     + 1/(2*sigma1^2)));
   vector[min(n,k)] zmu = w./sigma2^2 .* zvar;
   vector[min(n,k)] z =
      to_vector(normal_rng(zmu, sqrt(zvar)));
   beta = V * z;
 }
}
```

The following is a sample of code from R that can be used for the precomputation stage of Algorithm 1.

```
udv <- svd(X)
```


SPI Journal: 180 Article No.: 1135 TYPESET DISK LE CP Disp.:2021/7/29 Pages: 20 Layout: Small-Ex

```
V <- udv$v
lam <- as.vector(udv$d)
w <- t(V) %*% t(X) %*% y
w <- as.vector(w)
yty <- t(y) %*% y
yty <- yty[1]
```

B Proofs

In this appendix, we include proofs of the formulas provided in this paper. For increased readability, this appendix is self-contained.

B.1 Mathematical preliminaries and notation

In this section, we introduce notation and elementary mathematical identities that will be used throughout the remainder of this section.

We define $C \in \mathbb{R}$ by the Eq.

$$C = \int_{\sigma_1 \in \mathbb{R}^+} \int_{\sigma_2 \in \mathbb{R}^+} \int_{\beta \in \mathbb{R}^k} q(\sigma_1, \sigma_2, \beta) d\beta d\sigma_2 d\sigma_1, \tag{29}$$

and define $\mathbb{E}(\sigma_1)$, $\mathbb{E}(\sigma_2)$, and $\mathbb{E}(\beta_i)$ by the formulas

$$\mathbb{E}(\sigma_1) = \frac{1}{C} \int_{\sigma_1 \in \mathbb{R}^+} \int_{\sigma_2 \in \mathbb{R}^+} \int_{\beta \in \mathbb{R}^k} \sigma_1 q(\sigma_1, \sigma_2, \beta) d\beta d\sigma_2 d\sigma_1, \tag{30}$$

$$\mathbb{E}(\sigma_2) = \frac{1}{C} \int_{\sigma_1 \in \mathbb{R}^+} \int_{\sigma_2 \in \mathbb{R}^+} \int_{\beta \in \mathbb{R}^k} \sigma_2 q(\sigma_1, \sigma_2, \beta) d\beta d\sigma_2 d\sigma_1, \tag{31}$$

and

$$\mathbb{E}(\beta_i) = \frac{1}{C} \int_{\sigma_1 \in \mathbb{R}^+} \int_{\sigma_2 \in \mathbb{R}^+} \int_{\beta \in \mathbb{R}^k} \beta_i q(\sigma_1, \sigma_2, \beta) d\beta d\sigma_2 d\sigma_1 \tag{32}$$

for $i \in \{1, 2, \ldots, k\}$.

We provide algorithms for the evaluation of (29), (30), (31), and (32).

We will be denoting by $\mathbb{1}$ the vector of ones

$$\mathbb{1} = (1, 1, \ldots, 1)^t. \tag{33}$$

We denote the i^{th} component of a vector v by v_i.

The following two well-known identities give the normalizing constant and expectation of a Gaussian distribution.

Lemma 1 *For all $\sigma > 0$ we have*

$$\sqrt{2\pi}\sigma = \int_{\mathbb{R}} e^{\frac{-(\beta-\mu)^2}{2\sigma^2}} d\beta \tag{34}$$

SPI Journal: 180 Article No.: 1135 TYPESET DISK LE CP Disp.:2021/7/29 Pages: 20 Layout: Small-Ex

Lemma 2 *For all μ in $\mathbb{R}$ and $\sigma > 0$, we have*

$$\mu\sqrt{2\pi}\sigma = \int_{\mathbb{R}} \beta e^{\frac{-(\beta-\mu)^2}{2\sigma^2}} d\beta \tag{35}$$

B.2 Analytic integration of β

We denote the singular value decomposition of X by

$$X = UDV^t \tag{36}$$

where U is an orthogonal $n \times k$ matrix, V is an orthogonal $k \times k$ matrix, and D is a $k \times k$ diagonal matrix. We define $z \in \mathbb{R}^k$ by the formula

$$z = V^t \beta. \tag{37}$$

The following lemma, which will be used in the proof of Lemma 5, gives an expression for the second to last term of the exponent in (1) after a change of variables.

Lemma 3 *For all $\beta \in \mathbb{R}^k$, and $y \in \mathbb{R}^n$,*

$$-\frac{1}{2\sigma_2^2}\|X\beta - y\|^2 = -\frac{y^t y}{2\sigma_2^2} + \sum_{i=1}^{k} -\frac{\lambda_i^2}{2\sigma_2^2} z_i^2 + \frac{w_i}{\sigma_2^2} z_i \tag{38}$$

where

$$w = V^t X^t y, \tag{39}$$

z is defined in (37), and λ_i is the i^{th} entry on the diagonal of D (see (36)).

Proof Clearly,

$$\|X\beta - y\|^2 = \beta^t X^t X\beta - 2y^t X\beta + y^t y. \tag{40}$$

Substituting (36) and (37) into (40), we obtain

$$\begin{aligned}\|X\beta - y\|^2 &= \beta^t (UDV^t)^t (UDV^t)\beta - 2y^t XVV^t\beta + y^t y \\ &= (\beta^t V)D^2(V^t\beta) - 2y^t (V^t X^t)^t z + y^t y.\end{aligned} \tag{41}$$

where z is defined in (37). Substituting (39) and (37) into (41), we have

$$\|X\beta - y\|^2 = z^t D^2 z - 2w^t z + y^t y \tag{42}$$

Equation (38) follows immediately from (42). □

The following lemma provides an equation for the last term of the exponent in (1). The identity will be used in Lemma 5.

Lemma 4 *For all* $\sigma_1 > 0$,

$$-\frac{\|\beta\|^2}{2\sigma_1^2} = \sum_{i=1}^{k} -\frac{z_i^2}{2\sigma_1^2} \tag{43}$$

where $\beta \in \mathbb{R}^k$, z *is defined in (37), and* V *is defined in (36).*

Proof Clearly,

$$\frac{\|\beta\|^2}{2\sigma_1^2} = \frac{1}{2\sigma_1^2}(Vz)^t(Vz) = \frac{z^t z}{2\sigma_1^2} \tag{44}$$

where V is defined in (36). Equation (43) follows immediately from (44). □

The following formula combines Lemmas 3 and 4 to convert the final two terms of (1) into a Gaussian in k dimensions.

Lemma 5

$$-\frac{\|X\beta - y\|^2}{2\sigma_2^2} - \frac{\|\beta\|^2}{2\sigma_1^2} = a_0 + \sum_{i=1}^{k} a_{2,i}(z_i - \frac{a_{1,i}}{2a_{2,i}})^2 + \frac{a_{1,i}^2}{4a_{2,i}} \tag{45}$$

where

$$a_{2,i} = \frac{\lambda_i^2}{2\sigma_2^2} + \frac{1}{2\sigma_1^2}, \tag{46}$$

$$a_{1,i} = \frac{w_i}{\sigma_2^2} \tag{47}$$

and

$$a_0 = -\frac{y^t y}{2\sigma_2^2} \tag{48}$$

where z *is defined in (37),* w *is defined in (39) and* V *is defined in (36).*

Proof By combining Lemmas 3 and 4, we have

$$-\frac{1}{2\sigma_2^2}\|X\beta - y\|^2 - \frac{1}{2\sigma_1^2}\|\beta\|^2 = a_0 + \sum_{i=1}^{k}\left(a_{1,i}z_i - a_{2,i}z_i^2\right). \tag{49}$$

We obtain Eq. (45) by completing the square in Eq. (49). □

The following theorem is the principal analytical apparatus of this note. It provides a formula for the k-dimensional integrals that appear in (29), (30), and (31).

Theorem 1 *For all* $\sigma_1, \sigma_2 > 0$

$$\int_{\mathbb{R}^k} q(\sigma_1, \sigma_2, \beta)d\beta = \tilde{q}(\sigma_1, \sigma_2) \tag{50}$$

where $\tilde{q}(\sigma_1, \sigma_2)$ is defined by the formula

$$\tilde{q}(\sigma_1, \sigma_2) = \sigma_1^{-(k+1)} \sigma_2^{-n} \exp\Bigg(-\log^2(\sigma_1) - \frac{\sigma_2^2}{2} + a_0 + \sum_{i=1}^{k} \frac{a_{1,i}^2}{4a_{2,i}} \Bigg) \sqrt{2\pi}^k \prod_{i=1}^{k} \frac{1}{\sqrt{2a_{2,i}}} \tag{51}$$

where $a_{2,i}$ is defined in (46), $a_{1,i}$ is defined in (47) and a_0 is defined in (48).

Proof Using (1), clearly

$$\int_{\mathbb{R}^k} q(\sigma_1, \sigma_2, \beta) d\beta = \sigma_1^{-(k+1)} \int_{\mathbb{R}^k} \exp\Bigg(-\log^2(\sigma_1) - \frac{\sigma_2^2}{2} - \frac{1}{2\sigma_2^2} \|X\beta - y\|^2 - \frac{1}{2\sigma_1^2} \|\beta\|^2 \Bigg) d\beta \tag{52}$$

Performing the change of variables (37) and substituting (45) into (52), we have

$$\int_{\mathbb{R}^k} q(\sigma_1, \sigma_2, \beta) d\beta = \sigma_1^{-(k+1)} \exp\Bigg(-\log^2(\sigma_1) - \frac{\sigma_2^2}{2} + a_0 + \sum_{i=1}^{k} \frac{a_{1,i}^2}{4a_{2,i}} \Bigg) \int_{\mathbb{R}^k} \exp\Bigg(\sum_{i=1}^{k} a_{2,i} (z_i - \frac{a_{1,i}}{2a_{2,i}})^2 \Bigg) dz \tag{53}$$

Since the integrand on the right side of (53) is a Gaussian in z_i, Eq. (50) follows from applying Lemma 1 to (53). □

Remark 3 When adjusting the priors on the scale parameter to both become half-normal, we have the model

$$\sigma_1 \sim \text{normal}^+(0, 1) \tag{54}$$
$$\sigma_2 \sim \text{normal}^+(0, 1) \tag{55}$$
$$\beta \sim \text{normal}(0, \sigma_1) \tag{56}$$
$$y \sim \text{normal}(X\beta, \sigma_2). \tag{57}$$

For the corresponding posterior, we note that $\tilde{q}$ becomes

$$\int_{\mathbb{R}^k} q(\sigma_1, \sigma_2, \beta) d\beta = \exp\Bigg(-\frac{\sigma_1^2 + \sigma_2^2}{2} + a_0 + \sum_{i=1}^{k} \frac{a_{1,i}^2}{4a_{2,i}} \Bigg) \int_{\mathbb{R}^k} \exp\Bigg(\sum_{i=1}^{k} a_{2,i} (z_i - \frac{a_{1,i}}{2a_{2,i}})^2 \Bigg) dz \tag{58}$$

SPI Journal: 180 Article No.: 1135 TYPESET DISK LE CP Disp.:2021/7/29 Pages: 20 Layout: Small-Ex

The following theorem provides a formula for the expectation of z (see (37)). We use this formula, in combination with an orthogonal transformation, to obtain the expectation of β.

Theorem 2 *For all $\sigma_1 > 0$ and $\sigma_2 \in \mathbb{R}$,*

$$\int_{\mathbb{R}^k} (V^t x)_i q(\sigma_1, \sigma_2, \beta) d\beta = \frac{a_{1,i}}{2a_{2,i}} \tilde{q}(t) \tag{59}$$

where q is defined in (1), $\tilde{q}$ is defined in (51), $a_{2,i}$ is defined in (46), $a_{1,i}$ is defined in (47), a_0 is defined in (48).

Proof Combining (53) and (37), we have

$$\begin{aligned}
&\int_{\mathbb{R}^k} (V^t \beta)_i q(\sigma_1, \sigma_2, \beta) d\beta \\
&= \exp\left(-\log^2(\sigma_1) - \frac{\sigma_2^2}{2} + a_0 + \sum_{i=1}^{k} \frac{a_{1,i}^2}{4a_{2,i}} \right) \\
&\times \int_{\mathbb{R}^k} z_i \exp\left(\sum_{i=1}^{k} a_{2,i} (z_i - \frac{a_{1,i}}{2a_{2,i}})^2 \right) dz.
\end{aligned} \tag{60}$$

Applying Lemma 2 to (60), we obtain (59). □

C Trapezoid rule

The trapezoid rule (see, e.g. Stoer and Bulirsch 1992) is a quadrature scheme that is used to approximate the integral

$$\int_a^b f(x) dx \tag{61}$$

with the sum

$$\sum_{k=1}^{n} \frac{f(x_{k-1}) + f(x_k)}{2} \Delta_x \tag{62}$$

where $\Delta_x = (b-a)/(n-1)$ and

$$x_k = a + k\frac{b-a}{n} \tag{63}$$

SPI Journal: 180 Article No.: 1135 TYPESET DISK LE CP Disp.:2021/7/29 Pages: 20 Layout: Small-Ex

for $k = 0, ..., n$. In the 2-dimensional analogue of the trapezoid rule we approximate the integral

$$\int_c^d \int_a^b f(x, y) dx dy \tag{64}$$

with the sum

$$\sum_{k=1}^{n} \frac{g(y_{k-1}) + g(y_k)}{2} \Delta_y \tag{65}$$

where

$$g(y) = \sum_{k=1}^{m} \frac{f(x_{k-1}, y) + f(x_k, y)}{2} \Delta_x \tag{66}$$

and

$$\Delta_y = (d - c)/(n - 1), \tag{67}$$

$$y_k = c + k \frac{d - c}{n}, \tag{68}$$

$$\Delta_x = (b - a)/(m - 1), \tag{69}$$

$$x_k = a + k \frac{b - a}{m}. \tag{70}$$

References

Bardini R, Politano G, Benso A, Di Carlo S (2017) Multi-level and hybrid modelling approaches for systems biology. Comput Struct Biotechnol J 15:396–402

Bates D, Martin M, Ben B, Steve W (2015) Fitting linear mixed-effects models using lme4. J Stat Softw

Carpenter B, Gelman A, Hoffman MD, Lee D, Goodrich B, Betancourt M, Brubaker M, Guo J, Li P, Riddell A (2017) Stan: a probabilistic programming language. J Stat Softw 76(1):1–32

Dias S, Sutton AJ, Welton NJ, Ades AE (2013) Evidence synthesis for decision making 3: heterogeneity-subgroups, meta-regression, bias, and bias-adjustment. Med Decis Mak 33(5):618–640

Eckart C, Young G (1936) The approximation of one matrix with another of lower rank. Psychometrika 1:3

Ferreira AC et al (2015) Ridge, lasso and Bayesian additive-dominance genomic models. BMC Genet 16:105

Gelman A, Hill J (2007) Data analysis using regression and multilevel/hierarchical models. Cambridge University Press, Cambridge, UK

Gelman A, Carlin JB, Stern SH, Dunson BD, Vehtari A, Rubin BD (2013) Bayesian Data Analysis, 3rd edn. CRC, New York, U.S

Golub G, Kahan W (1965) Calculating the singular values and psuedo-inverse of a matrix. J SIAM Numer Anal 2:3

Greenland S (2000) Principles of multilevel modelling. Int J Epidemiol 29(1):158–167

Halko N, Martinsson PG, Tropp JA (2011) Finding structure with randomness: probabilisitc algorithms for constructing approximate matrix decompositions. SIAM Rev 53:2

Hastie T, Rahul M, Lee JD, Zadeh R (2015) Matrix completion and low-rank SVD via fast alternating least squares. J Mach Learn Res 16(1):3367–3402

SPI | Journal: 180 Article No.: 1135 TYPESET DISK LE CP Disp.:2021/7/29 Pages: 20 Layout: Small-Ex

Kwon S, Yan X, Cui J, Yao J, Yang K, Tsiand D, Li X, Rotter J, Guo X (2011) Application of Bayesian regression with singular value decomposition method in association studies for sequence data. BMC Proc 5:9

Lindley DV, Smith AFM (1972) Bayes estimates for the linear model. J Royal Stat Soc B 34:1–41

Linner K et al (2019) Genome-wide association analyses of risk tolerance and risky behaviors in over 1 million individuals identify hundreds of loci and shared genetic influences. Nat Genet 51:2

Merlo J, Chaix B, Yang M, Lynch J, Rastam L (2005) A brief conceptual tutorial of multilevel analysis in social epidemiology: linking the statistical concept of clusetering to the idea of contextual phenomenon. J Epidemiol Community Health 59:3367–3402

Meuwissen TH, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. Genet 157:4

Rover C., Ralf B., Sofia D., Christopher H.S., Heinz S., Sibylle S., Sebastian W., Tim F. (2020) "On weakly informative prior distributions for the heterogeneity parameter in Bayesian random-effects meta-analysis." arXiv:2007.08352v3

Rue H, Riebler A, Sorbye SH, Illian JB, Simpson DP, Lindgren FK (2017) Bayesian computing with INLA: a review. Annual Rev Stat Appl 4:395–421

Shamir O. (2016) Fast Stochastic Algorithms for SVD and PCA: Convergence Properties and Convexity, Proceedings of the 33rd ICML, New York, NY,

Stoer J, Bulirsch R (1992) Introduction to numerical analysis, 2nd edn. Springer-Verlag, Berlin

Trefethen LN (2020) Approximation Theory and Approximation Practice. SIAM, Extended

Xiang Z, Matthew S (2017) Bayesian large-scale multiple regression with summary statistics from genome-wide association studies. Annals Appl Stat 11:3