Biometrika (2020), **107**, 1, *pp.* 205–221
Printed in Great Britain

doi: 10.1093/biomet/asz056
Advance Access publication 23 November 2019

Bayesian sparse multiple regression for simultaneous rank reduction and variable selection

By ANTIK CHAKRABORTY, ANIRBAN BHATTACHARYA AND BANI K. MALLICK

Department of Statistics, Texas A&M University, College Station, Texas 77843, U.S.A.

antik@stat.tamu.edu anirbanb@stat.tamu.edu bmallick@stat.tamu.edu

Summary

We develop a Bayesian methodology aimed at simultaneously estimating low-rank and row-sparse matrices in a high-dimensional multiple-response linear regression model. We consider a carefully devised shrinkage prior on the matrix of regression coefficients which obviates the need to specify a prior on the rank, and shrinks the regression matrix towards low-rank and row-sparse structures. We provide theoretical support to the proposed methodology by proving minimax optimality of the posterior mean under the prediction risk in ultra-high-dimensional settings where the number of predictors can grow subexponentially relative to the sample size. A one-step post-processing scheme induced by group lasso penalties on the rows of the estimated coefficient matrix is proposed for variable selection, with default choices of tuning parameters. We additionally provide an estimate of the rank using a novel optimization function achieving dimension reduction in the covariate space. We exhibit the performance of the proposed methodology in an extensive simulation study and a real data example.

Some key words: Dimension reduction; High dimension; Posterior concentration; Scalability; Shrinkage; Variable selection.

1. Introduction

Studying the relationship between multiple response variables and a set of predictors has broad applications ranging from bioinformatics, econometrics and time series analysis to growth curve models. The least squares solution in a linear multiple response regression problem is equivalent to performing separate least squares on each of the responses (Anderson, 1984) and ignores any potential dependence among the responses. In the context of multiple response regression, a popular technique to achieve parsimony and interpretability is to consider a reduced-rank decomposition of the coefficient matrix, commonly known as reduced rank regression (Anderson, 1951; Izenman, 1975; Velu & Reinsel, 2013). Although many results exist about the asymptotic properties of reduced rank estimators (Anderson, 2002), formal statistical determination of the rank remains difficult, even with a fixed number of covariates and a large sample size, due mainly to the discrete nature of the parameter. The problem becomes substantially harder when a large number of covariates are present, and has motivated a series of recent works on penalized estimation of low-rank matrices, where either the singular values of the coefficient matrix (Yuan et al., 2007; Chen et al., 2013) or the rank itself (Bunea et al., 2011) are penalized. Theoretical evaluations of these estimators focusing on adaptation to the oracle convergence rate when the true coefficient matrix is of low rank have been conducted in Bunea et al. (2011). Bunea et al. (2012) also noted that the convergence rate can be improved when the true coefficient matrix has rows of zeros and variable selection is incorporated within the estimation procedure. Methods that simultaneously handle rank reduction and variable selection include Yuan et al. (2007), Bunea et al. (2012) and Chen & Huang (2012). To the best of our knowledge, uncertainty characterization for the parameter estimates from these procedures is not currently available.

The first fully systematic Bayesian treatment of reduced rank regression was carried out in Geweke (1996), where, conditioned on the rank, independent Gaussian priors were placed on the elements of the coefficient matrix. While formal Bayesian model selection can be performed to determine the rank (Geweke, 1996), calculation of marginal likelihoods for various candidate ranks gets computationally burdensome with increasing dimensions. The problem of choosing the rank is not unique to reduced rank regression and is ubiquitous in situations involving low-rank decompositions, with factor models being a prominent example. Lopes & West (2004) placed a prior on the number of factors and proposed a computationally intensive reversible jump algorithm (Green, 1995) for model fitting. As an alternative, Bhattacharya & Dunson (2011) proposed to increasingly shrink the factors, starting with a conservative upper bound, and adaptively collapse redundant columns inside their Markov chain Monte Carlo algorithm. Recent advancements in Bayesian matrix factorization have taken a similar approach (see, for example, Lim & Teh, 2007; Salakhutdinov & Mnih, 2008; Babacan et al., 2011; Alquier, 2013).

From a Bayesian point of view, a natural way to select variables in a single-response regression framework is to use point mass mixture priors (George & McCulloch, 1993; Scott & Berger, 2010), which allow a subset of the regression coefficients to be exactly zero. These priors were also adapted to multiple response regression by several authors (Brown et al., 1998; Lucas et al., 2006; Wang, 2010; Bhadra & Mallick, 2013). Posterior inference with such priors involves a stochastic search over an exponentially growing model space and is computationally expensive even in moderate dimensions. To alleviate the computational burden, a number of continuous shrinkage priors have been proposed in the literature which mimic the operating characteristics of the discrete mixture priors. Such priors can be expressed as Gaussian scale mixtures (Polson & Scott, 2010), leading to block updates of model parameters; see Bhattacharya et al. (2016) for a review of such priors and efficient implementations in high-dimensional settings. To perform variable selection with these continuous priors, several methods for post-processing the posterior distribution have been proposed (Bondell & Reich, 2012; Kundu et al., 2013; Hahn & Carvalho, 2015).

In this article we simultaneously address the problems of dimension reduction and variable selection in high-dimensional reduced rank models from a Bayesian perspective. We develop a novel shrinkage prior on the coefficient matrix which encourages shrinkage towards low-rank and row-sparse matrices. The shrinkage prior is induced from appropriate shrinkage priors on the components of a full-rank decomposition of the coefficient matrix, and hence bypasses the need to specify a prior on the rank. We provide theoretical understanding of the operating characteristics of the proposed prior in terms of a novel prior concentration result around rank-reduced and row-sparse matrices. The prior concentration result is utilized to prove minimax concentration rates of the posterior under the fractional posterior framework of Bhattacharya et al. (2019) in an ultra-high-dimensional setting where the number of predictor variables can grow subexponentially in the sample size.

The continuous nature of the prior enables efficient block updates of parameters inside a Gibbs sampler. In particular, we adapt an algorithm for sampling structured multivariate Gaussians from Bhattacharya et al. (2016) to efficiently sample a high-dimensional matrix in a block leading to a low per-iteration Markov chain Monte Carlo computational cost. We propose two independent post-processing schemes to achieve row sparsity and rank reduction with encouraging performance. A key feature of our post-processing schemes is to exploit the posterior summaries to offer careful default choices of tuning parameters, resulting in a procedure which is completely free of tuning parameters. The resulting row-sparse and rank-reduced coefficient estimate is called a Bayesian sparse multi-task learner. We illustrate the superiority of our estimator over its competitors through a detailed simulation study, and the methodology is applied to a yeast cell cycle dataset.

2. Bayesian sparse multitask learner

Suppose, for each observational unit $i = 1, \ldots, n$, we have a multivariate response $y_i \in \mathbb{R}^q$ on q variables of interest, along with information on p possible predictors $x_i \in \mathbb{R}^p$, a subset of which are assumed to be important in predicting the q responses. Let $X \in \mathbb{R}^{n\times p}$ denote the design matrix whose ith row is x_i^{T}, and $Y \in \mathbb{R}^{n\times q}$ the matrix of responses with the ith row as y_i^{T}. The multivariate linear regression model is

$$Y = XC + E, \qquad E = (e_1^{\mathrm{T}}, \ldots, e_n^{\mathrm{T}})^{\mathrm{T}}, \tag{1}$$

where we follow standard practice to centre the response and exclude the intercept term. The rows of the error matrix are independent, with $e_i \sim N(0, \Sigma)$. Our main motivation is the high-dimensional case where $p \geqslant \max\{n, q\}$, although the method trivially applies to $p < n$ settings as well. We shall also assume the dimension of the response q to be modest relative to the sample size.

The basic assumption in reduced rank regression is that $\mathrm{rank}(C) = r \leqslant \min(p, q)$, whence C admits a decomposition $C = B_* A_*^{\mathrm{T}}$ with $B_* \in \mathbb{R}^{p\times r}$ and $A_* \in \mathbb{R}^{q\times r}$. While it is possible to treat r as a parameter and assign it a prior distribution inside a hierarchical formulation, posterior inference on r requires calculation of intractable marginal likelihoods or resorting to complicated reversible jump Markov chain Monte Carlo algorithms. To avoid specifying a prior on r, we work within a parameter-expanded framework (Liu & Wu, 1999) to consider a potentially full-rank decomposition $C = BA^{\mathrm{T}}$ with $B \in \mathbb{R}^{p\times q}$ and $A \in \mathbb{R}^{q\times q}$, and assign shrinkage priors to A and B to shrink out the redundant columns when C is indeed low rank. This formulation embeds all reduced-rank models inside the full model; if a conservative upper bound $q^* \leqslant q$ on the rank is known, the method can be modified accordingly. The role of the priors on B and A is important to encourage appropriate shrinkage towards reduced-rank models, which is discussed below.

We consider independent standard normal priors on the entries of A. As an alternative, a uniform prior on the Stiefel manifold (Hoff, 2009) of orthogonal matrices can be used. However, our numerical results suggest significant gains in computation time using the Gaussian prior over the uniform prior with no discernible difference in statistical performance. The Gaussian prior allows an efficient block update of $\mathrm{vec}(A)$, whereas the algorithm of Hoff (2009) involves conditional Gibbs update of each column of A. Our theoretical results also suggest that the shrinkage provided by the Gaussian prior is optimal when q is modest relative to n, the regime in which we operate. We shall henceforth use Π_A to denote the prior on A, i.e., $a_{hk} \sim N(0, 1)$ independently for $h, k = 1, \ldots, q$.

Recalling that the matrix B has dimension $p \times q$, with p potentially larger than n, stronger shrinkage is warranted on the columns of B. We use independent horseshoe priors (Carvalho et al., 2010) on the columns of B, which can be represented hierarchically as

$$b_{jh} \mid \lambda_{jh}, \tau_h \sim N(0, \lambda_{jh}^2 \tau_h^2), \quad \lambda_{jh} \sim \mathrm{Ca}_+(0, 1), \quad \tau_h \sim \mathrm{Ca}_+(0, 1), \tag{2}$$

independently for $j = 1, \ldots, p$ and $h = 1, \ldots, q$, where $\mathrm{Ca}_{+}(0, 1)$ denotes the truncated standard half-Cauchy distribution with density proportional to $(1 + t^2)^{-1}\mathbb{1}_{(0,\infty)}(t)$. We shall denote the prior on the matrix B induced by the hierarchy in (2) by Π_B.

We shall primarily restrict our attention to settings where Σ is diagonal, $\Sigma = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_q^2)$, noting that extensions to nondiagonal Σ can be incorporated in a straightforward fashion. For example, for moderate q a conjugate inverse-Wishart prior can be used as a default. Furthermore, if Σ has a factor model or Gaussian Markov random field structure, they can also be incorporated using standard techniques (Rue, 2001; Bhattacharya & Dunson, 2011). The cost per iteration of the Gibbs sampler retains the same complexity as in the diagonal Σ case; see § 3 for more details. In the diagonal case, we assign independent improper priors $\pi(\sigma_h^2) \propto \sigma_h^{-2}$ $(h = 1, \ldots, q)$ on the diagonal elements, and call the resulting prior Π_Σ.

The model augmented with the above priors now takes the shape

$$Y = XBA^{\mathrm{T}} + E, \qquad e_i \sim N(0, \Sigma), \tag{3}$$

$$B \sim \Pi_B, \qquad A \sim \Pi_A, \qquad \Sigma \sim \Pi_\Sigma. \tag{4}$$

We shall refer to the induced prior on $C = BA^{\mathrm{T}}$ by Π_C, and let

$$p^{(n)}(Y \mid C, \Sigma; X) \propto |\Sigma|^{-n/2}\, \mathrm{e}^{-\mathrm{tr}\{(Y-XC)\Sigma^{-1}(Y-XC)^{\mathrm{T}}\}/2}$$

denote the likelihood for (C, Σ).

3. Posterior computation

3.1. *Gibbs sampler*

Exploiting the conditional conjugacy of the proposed prior, we develop a straightforward and efficient Gibbs sampler to update the model parameters in (3) from their full conditional distributions. We use vectorization to update parameters in blocks. Specifically, in what follows we will make multiple use of the following identity. For matrices Φ_1, Φ_2, Φ_3 with appropriate dimensions, and $\mathrm{vec}(A)$ denoting columnwise vectorization, we have

$$\mathrm{vec}(\Phi_1\Phi_2\Phi_3) = (\Phi_3^{\mathrm{T}} \otimes \Phi_1)\mathrm{vec}(\Phi_2) = (\Phi_3^{\mathrm{T}}\Phi_2^{\mathrm{T}} \otimes I_k)\mathrm{vec}(\Phi_1), \tag{5}$$

where the matrix Φ_1 has k rows and $\otimes$ denotes the Kronecker product.

Letting $\theta \mid -$ denote the full conditional distribution of a parameter θ given other parameters and the data, the Gibbs sampler cycles through the following steps, sampling parameters from their full conditional distributions.

Step 1. To sample $B \mid -$, use (5) to vectorize $Y = XBA^{\mathrm{T}} + E$ to obtain

$$y = (X \otimes A)\beta + e, \tag{6}$$

where $\beta = \mathrm{vec}(B^{\mathrm{T}}) \in \mathbb{R}^{pq\times 1}$, $y = \mathrm{vec}(Y^{\mathrm{T}}) \in \mathbb{R}^{nq\times 1}$, and $e = \mathrm{vec}(E^{\mathrm{T}}) \sim N_{nq}(0, \widetilde{\Sigma})$ with $\widetilde{\Sigma} = \mathrm{diag}(\Sigma, \ldots, \Sigma)$. Multiplying both sides of (6) by $\widetilde{\Sigma}^{-1/2}$ yields $\widetilde{y} = \widetilde{X}\beta + \widetilde{e}$, where $\widetilde{y} = \widetilde{\Sigma}^{-1/2}y$, $\widetilde{X} = \widetilde{\Sigma}^{-1/2}(X \otimes A)$ and $\widetilde{e} = \widetilde{\Sigma}^{-1/2}e \sim N_{nq}(0, \mathrm{I}_{nq})$. Thus, the full conditional distribution is $\beta \mid - \sim N_{pq}(\Omega_B^{-1}\widetilde{X}^{\mathrm{T}}\widetilde{y}, \Omega_B^{-1})$, where $\Omega_B = (\widetilde{X}^{\mathrm{T}}\widetilde{X} + \Lambda^{-1})$ with $\Lambda = \mathrm{diag}(\lambda_{11}^2\tau_1^2, \ldots, \lambda_{1q}^2\tau_q^2, \ldots, \lambda_{p1}^2\tau_1^2, \ldots, \lambda_{pq}^2\tau_q^2)$.

Naively sampling from the full conditional of β has complexity $O(p^3q^3)$, which becomes highly expensive for moderate values of p and q. Bhattacharya et al. (2016) recently developed an algorithm to sample from a class of structured multivariate normal distributions whose complexity scales linearly in the ambient dimension. We adapt the algorithm in Bhattacharya et al. (2016) as follows:

(i) Sample $u \sim N(0, \Lambda)$ and $\delta \sim N(0, \mathrm{I}_{nq})$ independently.
(ii) Set $v = \widetilde{X}u + \delta$.
(iii) Solve $(\widetilde{X}\Lambda\widetilde{X}^{\mathrm{T}} + \mathrm{I}_{nq})w = (\tilde{y} - v)$ to obtain w.
(iv) Set $\beta = u + \Lambda\widetilde{X}^{\mathrm{T}}w$.

It follows from Bhattacharya et al. (2016) that β obtained from steps (i)–(iv) above produces a sample from the desired full conditional distribution. One only requires matrix multiplications and linear system solvers to implement the above algorithm, and no matrix decomposition is required. It follows from standard results (Golub & van Loan, 1996) that the above steps have a combined complexity of $O(q^3 \max\{n^2, p\})$, a substantial improvement over $O(p^3q^3)$ when $p \gg \max\{n, q\}$.

Step 2. To sample $A \mid -$, once again vectorize $Y = XBA^{\mathrm{T}} + E$, but this time use the equality of the first and the third terms in (5) to obtain

$$y = (XB \otimes \mathrm{I}_q)a + e, \tag{7}$$

where e and y are the same as in Step 1, and $a = \mathrm{vec}(A) \in \mathbb{R}^{q^2\times 1}$. The full conditional posterior distribution is $a \mid - \sim N(\Omega_A^{-1}X_*\widetilde{y}, \Omega_A^{-1})$, where $\Omega_A = (X_*^{\mathrm{T}}X_* + \mathrm{I}_{q^2})$, $X_* = \widetilde{\Sigma}^{-1/2}(XB \otimes \mathrm{I}_{q^2})$ and $\widetilde{y} = \widetilde{\Sigma}^{-1/2}y$. To sample from the full conditional of a, we use the algorithm from § 3.1.2 of Rue (2001). Compute the Cholesky decomposition $(X_*^{\mathrm{T}}X_* + I_{q^2}) = LL^{\mathrm{T}}$. Solve the system of equations $Lv = X_*^{\mathrm{T}}\tilde{y}$, $L^{\mathrm{T}}m = v$ and $L^{\mathrm{T}}w = z$, where $z \sim N(0, \mathrm{I}_{\mathrm{q}^2})$. Finally, obtain a sample as $a = m + w$.

Step 3. To sample $\sigma_h^2 \mid -$, observe that $\sigma_h^2 \mid - \sim \mathrm{IG}(n/2, S_h/2)$ independently across h, where $S_h = \{Y_h - (XBA^{\mathrm{T}})_h\}^{\mathrm{T}}\{Y_h - (XBA^{\mathrm{T}})_h\}$, with Φ_h denoting the hth column of a matrix Φ. In the case of an unknown Σ and an $\mathrm{IW}(q, \mathrm{I}_q)$ prior on Σ, the posterior update of Σ can be easily modified due to conjugacy; we sample $\Sigma \mid -$ from $\mathrm{IW}\{n + q, (Y - XC)^{\mathrm{T}}(Y - XC) + \mathrm{I}_q\}$.

Step 4. The global and local scale parameters λ_{jh} and τ_h have independent conditional posteriors across j and h, which can be sampled via the slice sampling scheme provided in the online supplement to Polson et al. (2014). We illustrate the sampling technique for a generic local shrinkage parameter λ_{jh}; a similar scheme works for τ_h. Setting $\eta_{jh} = \lambda_{jh}^{-2}$, the slice sampler proceeds by sampling $u_{jh} \mid \eta_{jh} \sim \mathrm{Un}\{0, 1/(1 + \eta_{jh})\}$ and then sampling $\eta_{jh} \mid u_{jh} \sim \mathrm{Ex}(2\tau_h^2/b_{jh}^2)\mathrm{I}\{\eta_{jh} < (1 - u_{jh})/u_{jh}\}$, a truncated exponential distribution.

The Gibbs sampler above, when modified to accommodate nondiagonal Σ as mentioned in Step 3, retains the overall complexity. Steps 1–2 do not assume any structure for Σ. The matrix $\Sigma^{-1/2}$ can be computed in $O(q^3)$ steps using standard algorithms, which does not increase the overall complexity of Steps 1 and 2 since $q < n \ll p$ by assumption. Modifications to situations where Σ has a graphical/factor model structure are also straightforward.

Point estimates of C, such as the posterior mean or elementwise posterior median, are readily obtained from the Gibbs sampler along with a natural uncertainty quantification, which can be used for point and interval predictions. However, the continuous nature of our prior implies that such point estimates will be nonsparse and full rank with probability one, and hence not directly amenable for variable selection and rank estimation. Motivated by our concentration result in Theorem 2 that the posterior mean $X\bar{C}$ increasingly concentrates around XC_0, we propose two simple post-processing schemes for variable selection and rank estimation below. The procedures are completely automated and do not involve any input of tuning parameters from the user's end.

3.2. *Post-processing for variable selection*

We define a row-sparse estimate $\hat{C}_{\mathrm{R}}$ for C as the solution to the optimization problem

$$\hat{C}_{\mathrm{R}} = \underset{\Gamma\in\mathbb{R}^{p\times q}}{\arg\min}\left\{\|X\bar{C} - X\Gamma\|_{\mathrm{F}}^2 + \sum_{j=1}^{p}\mu_j\|\Gamma^{(j)}\|_2\right\}, \tag{8}$$

where $\Phi^{(j)}$ represents the jth row of a matrix Φ, and the μ_j are predictor-specific regularization parameters. The objective function aims to find a row-sparse solution close to the posterior mean in terms of the prediction loss, with the sparsity driven by the group lasso penalty (Yuan & Lin, 2006). For a derivation of the objective function in (8) from a utility function perspective as in Hahn & Carvalho (2015), refer to the Supplementary Material.

To solve (8), we set the subgradient of (8) with respect to $\Gamma^{(j)}$ to zero and replace $\|\Gamma^{(j)}\|$ by a data-dependent quantity to obtain the soft thresholding estimate

$$\hat{C}_{\mathrm{R}}^{(j)} = \frac{1}{X_j^{\mathrm{T}}X_j}\left(1 - \frac{\mu_j}{2\|X_j^{\mathrm{T}}R_j\|}\right)_+ X_j^{\mathrm{T}}R_j, \tag{9}$$

where, for $x\in\mathbb{R}$, $x_+ = \max(x,0)$, and R_j is the residual matrix obtained after regressing $X\bar{C}$ on X leaving out the jth predictor, $R_j = X\bar{C} - \sum_{k\neq j}X_k\hat{C}_{\mathrm{R}}^{(k)}$. See the Supplementary Material for the derivation of (9). For practical implementation, we use $\bar{C}$ as our initial estimate and make a single pass through each variable to update the initial estimate according to (9). With this initial choice, $R_j = X_j\bar{C}^{(j)}$ and $\|X_j^{\mathrm{T}}R_j\| = \|X_j\|^2\|\bar{C}_j\|$.

While the p tuning parameters μ_j can be chosen by cross-validation, the computational cost of searching a grid in p dimensions explodes with p. Exploiting the presence of an optimal initial estimate in the form of $\bar{C}$, we recommend default choices for the hyperparameters as $\hat{\mu}_j = 1/\|\bar{C}_j\|^{-2}$, which in spirit is similar to the adaptive lasso (Zou, 2006). When predictor j is not important, the minimax ℓ_2-risk for estimating $C_0^{(j)}$ is $(\log q)/n$, so that $\|\bar{C}^{(j)}\| \asymp (\log q)/n$. Since $\|X_j\|^2 \asymp n$ by assumption, see § 6, $\hat{\mu}_j/\|X_j^{\mathrm{T}}R_j\| \asymp n^{1/2}/(\log q)^{3/2} \gg 1$, implying a strong penalty for all irrelevant predictors.

Following Hahn & Carvalho (2015), posterior uncertainty in variable selection can be gauged if necessary by replacing $\bar{C}$ with the individual posterior samples for C in (8).

3.3. *Post-processing for rank estimation*

To estimate the rank, we threshold the singular values of $X\hat{C}_{\mathrm{R}}$, with $\hat{C}_{\mathrm{R}}$ obtained from (9). In situations where row sparsity is not warranted, $\bar{C}$ can be used instead of $\hat{C}_{\mathrm{R}}$. For $s_1,\ldots,s_q$ the singular values of $X\hat{C}_{\mathrm{R}}$, and a threshold $\omega > 0$, define the thresholded singular values as $\nu_h = s_h\,\mathrm{I}(s_h > \omega)$ for $h = 1,\ldots,q$. We estimate the rank as the number of nonzero thresholded singular values; that is, $\hat{r} = \sum_{h=1}^{q}\mathrm{I}(\nu_h > 0) = \sum_{h=1}^{q}\mathrm{I}(s_h > \omega)$. We use the largest singular value of $Y - X\hat{C}_{\mathrm{R}}$ as the default choice of the threshold parameter ω, a natural candidate for the maximum noise level in the model.

4. Simulation results

We performed a thorough simulation study to assess the performance of the proposed method across different settings. For all our simulation settings the sample size n was fixed at 100. We considered three different (p, q) combinations, $(p, q) = (500, 10), (200, 30), (1000, 12)$. The data were generated from the model $Y = XC_0 + E$. Each row of the matrix E was generated from a multivariate normal distribution with the diagonal covariance matrix having diagonal entries uniformly chosen between 0.5 and 1.75. The columns of the design matrix X were independently generated from $N(0, \Sigma_X)$. We considered two cases, $\Sigma_X = \mathrm{I}_p$, and $\Sigma_X = (\sigma_{ij}^X)$, $\sigma_{jj}^X = 1$, $\sigma_{ij}^X = 0.5$ for $i \neq j$. The true coefficient matrix $C_0 = B_* A_*^{\mathrm{T}}$, with $B_* \in \mathbb{R}^{p \times r_0}$ and $A_* \in \mathbb{R}^{r \times r_0}$, with the true rank $r_0 \in \{3, 5, 7\}$. The entries of A_* were independently generated from a standard normal distribution. We generated the entries in the first $s = 10$ rows of B_* independently from $N(0, 1)$, and the remaining $(p - s)$ rows were set equal to zero.

As a competitor, we considered the sparse partial least squares approach due to Chun & Keleş (2010). Partial least squares minimizes the least square criterion between the response Y and design matrix X in a projected lower-dimensional space, where the projection direction is chosen to preserve the correlation between Y and X as well as the variation in X. Chun & Keleş (2010) suggested adding lasso-type penalties while optimizing for the projection vectors for sparse high-dimensional problems. We call their estimate $\widetilde{C}$. Since the estimator from $\widetilde{C}$ returns a coefficient matrix which is both row sparse and rank reduced, we create a rank reduced matrix $\hat{C}_{\mathrm{RR}}$ from $\hat{C}_{\mathrm{R}}$ for a fair comparison. Recalling that $\hat{C}_{\mathrm{R}}$ has zero rows, let $\hat{S}_{\mathrm{R}}$ denote the submatrix corresponding to the nonzero rows of $\hat{C}_{\mathrm{R}}$. Truncate the singular value decomposition of $\hat{S}_{\mathrm{R}}$ to the first $\hat{r}$ terms, where $\hat{r}$ is as obtained in § 3.3. Insert back the zero rows corresponding to $\hat{C}_{\mathrm{R}}$ in the resulting matrix to obtain $\hat{C}_{\mathrm{RR}}$. Clearly, $\hat{C}_{\mathrm{RR}} \in \mathbb{R}^{p \times q}$ so created is row sparse and has rank at most $\hat{r}$.

For an estimator $\hat{C}$ of C, we consider the mean square error, $\textsc{mse} = \|\hat{C} - C_0\|_{\mathrm{F}}^2/(pq)$, and the mean square prediction error, $\textsc{mspe} = \|X\hat{C} - XC_0\|_{\mathrm{F}}^2/(nq)$, to measure its performance. The squared estimation and prediction errors of $\widetilde{C}$ and $\hat{C}_{\mathrm{RR}}$ for different settings are reported in Table 1 along with the estimates of rank. In our simulations we used the default ten-fold cross-validation in the `cv.spls` function from the `R` package `spls` (R Development Core Team 2020). The sparse partial least squares estimator of the rank is the one for which the minimum cross-validation error is achieved. We observed highly accurate estimates of the rank for the proposed method, whereas the rank of $\widetilde{C}$ overestimated the true rank in all the settings considered. The proposed method also achieved superior performance in terms of the two squared errors, improving upon $\widetilde{C}$ by as much as five times in some cases. Additionally, we observed that the performance of $\widetilde{C}$ deteriorated relative to $\hat{C}_{\mathrm{RR}}$ with an increasing number of covariates.

In terms of variable selection, both methods had specificity and sensitivity close to one in all the simulation settings listed in Table 1. Since $\widetilde{C}$ consistently overestimated the rank, we further investigated the effect of the rank on variable selection. We focused on the simulation case $(p, q, r_0) = (1000, 12, 3)$, and fitted both methods with different choices of the postulated rank between 3 and 9. For the proposed method, we set q^* in § 2 to be the postulated rank; that is, we considered $B \in \mathbb{R}^{p \times q^*}$ and $A \in \mathbb{R}^{q \times q^*}$ for $q^* \in \{3, \ldots, 9\}$. For the sparse partial least squares estimator, we simply input q^* as the number of hidden components inside the function `spls`. Figure 1 plots the sensitivity and specificity of $\hat{C}_{\mathrm{RR}}$ and $\widetilde{C}$ as a function of the postulated rank. While the specificity is robust for either method, the sensitivity of $\widetilde{C}$ turned out to be highly dependent on the rank. Figure 1(a) reveals that at the true rank, $\widetilde{C}$ only identifies 40% of the significant variables, and only achieves a similar sensitivity to $\hat{C}_{\mathrm{RR}}$ when the postulated rank is substantially overfitted. The proposed estimate $\hat{C}_{\mathrm{RR}}$, on the other hand, exhibits a decoupling effect wherein the overfitting of the rank does not impact the variable selection performance.

Table 1. *Estimation and predictive performance of the proposed estimator $\hat{C}_{\text{RR}}$ versus $\widetilde{C}$ across different simulation settings. We report the average estimated rank, $\hat{r}$, mean square error,* MSE *$(\times 10^{-4})$ and mean square predictive error,* MSPE, *across* 50 *replications. For each setting the true number of signals was* 10 *and sample size was* 100. *For each combination of (p, q, r_0) the columns of the design matrix were generated from $N(0, \Sigma_X)$. Two different choices of Σ_X were considered: $\Sigma_X = \mathrm{I}_p$ (independent) and $\Sigma_X = (\sigma_{ij}^X)$, $\sigma_{jj}^X = 1$, $\sigma_{ij}^X = 0.5$ for $i \neq j$ (correlated)*

						(p,q)							
		(200, 30)				(500, 10)				(1000, 12)			
		Independent		Correlated		Independent		Correlated		Independent		Correlated	
Rank	Measures	$\hat{C}_{\text{RR}}$	$\widetilde{C}$	$\hat{C}_{\text{RR}}$	$\widetilde{C}$	$\hat{C}_{\text{RR}}$	$\widetilde{C}$	$\hat{C}_{\text{RR}}$	$\widetilde{C}$	$\hat{C}_{\text{RR}}$	$\widetilde{C}$	$\hat{C}_{\text{RR}}$	$\widetilde{C}$
	$\hat{r}$	3.0	7.9	3.0	9.4	3.0	9.7	3.0	8.8	3.2	9.4	3.4	8.9
3	MSE	3.0	14.0	5.0	15.0	3.0	7.0	5.0	30.0	3.0	50.0	3.0	38.0
	MSPE	0.07	0.25	0.06	0.17	0.22	0.15	0.34	0.21	0.35	4.19	0.30	1.51
	$\hat{r}$	5.0	9.7	4.9	12.2	4.9	9.9	4.8	9.8	5.1	9.9	5.1	9.9
5	MSE	5.0	69.0	9.0	61.0	3.0	10.0	6.0	24.0	2.0	108.0	4.0	129.0
	MSPE	0.11	3.8	0.09	4.6	0.17	0.41	0.20	0.38	0.32	9.54	0.32	4.63
	$\hat{r}$	6.9	10.3	6.9	15.8	6.8	10.0	6.7	9.7	6.8	10.2	6.6	11.5
7	MSE	6.0	116.0	10.0	112.0	3.0	20.0	5.0	49.0	2.0	195.0	4.0	261.0
	MSPE	0.12	10.81	0.11	9.01	0.16	0.72	0.16	0.92	0.32	16.70	0.31	7.44

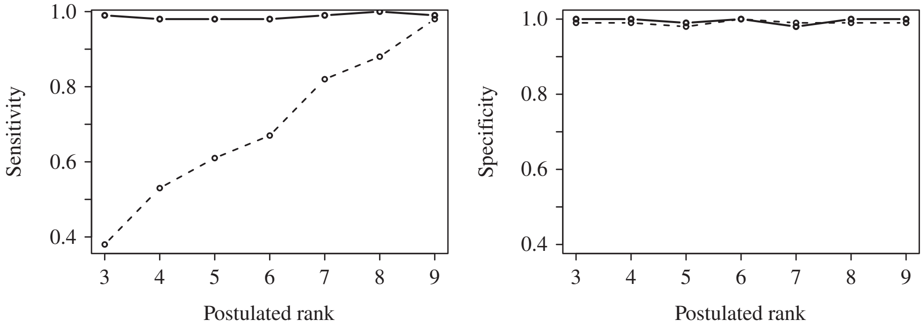

Fig. 1. Average sensitivity and specificity across 50 replicates plotted for different choices of the postulated rank, with $(p, q, r_0) = (1000, 12, 3)$: $\hat{C}_{\text{RR}}$ (solid), $\widetilde{C}$ (dashed).

We conclude this section with a simulation experiment carried out in a correlated response setting. Keeping the true rank r_0 fixed at 3, the data were generated as before, except that the individual rows e_i of the matrix E were generated from $N(0, \Sigma)$, with $\Sigma_{ii} = 1$, $\Sigma_{ij} = 0.5$, $1 \leqslant i \neq j \leqslant q$. To accommodate the nondiagonal error covariance, we placed an $\text{IW}(q, \mathrm{I}_q)$ prior on Σ. An associate editor pointed out a recent article (Ruffieux et al., 2017) which used spike-slab priors on the coefficients in a multiple response regression setting. They implemented a variational algorithm for posterior inclusion probabilities of each covariate, which is available from the `R` package `locus`. To select a model using the posterior inclusion probabilities we used the median probability model (Barbieri & Berger, 2004); predictors with a posterior inclusion probability less than 0.5 were deemed irrelevant. We implemented their procedure with the prior average number of predictors to be included in the model conservatively set to 25, a fairly well-chosen value in this context. We observed a fair degree of sensitivity to this parameter in estimating the sparsity of the model; when set to the true value of 10 it resulted in comparatively poor performance,

Table 2. *Variable selection performance of the proposed method in a nondiagonal error structure setting with independent and correlated predictors $e_i \sim \Sigma$, $\sigma_{ii} = 1$, $\sigma_{ij} = 0.5$. The sensitivity and specificity of our approach is compared with Ruffieux et al. (2017)*

			Our approach		Ruffieux et al. (2017)	
	(p, q)	Measure	Independent	Correlated	Independent	Correlated
$r_0 = 3$	$(200, 30)$	Sensitivity	1.0	1.0	0.96	0.87
		Specificity	0.90	0.84	0.77	0.67
	$(500, 10)$	Sensitivity	1.0	0.99	0.9	0.8
		Specificity	0.99	0.99	0.80	0.64
	$(1000, 12)$	Sensitivity	0.99	0.99	0.92	0.63
		Specificity	0.99	0.99	0.80	0.64

whereas a value of 100 resulted in much better performance. Table 2 reports the sensitivity and specificity of this procedure and ours, averaged over 50 replicates. While the two methods performed almost identically in the relatively low-dimensional setting $(p, q) = (200, 30)$, $\hat{C}_{\mathrm{RR}}$ consistently outperformed Ruffieux et al. (2017) when the dimension was higher.

5. Yeast cell cycle data

Identifying transcription factors which are responsible for cell cycle regulation is an important scientific problem (Chun & Keleş, 2010). The yeast cell cycle data from Spellman et al. (1998) contains information from three different experiments on mRNA levels of 800 genes on an α-factor-based experiment. The response variable is the amount of transcription, mRNA, which was measured every 7 minutes in a period of 119 minutes, a total of 18 measurements, Y, covering two cell cycle periods. The ChIP-chip data from Lee et al. (2002) on chromatin immunoprecipitation contains the binding information of the 800 genes for 106 transcription factors, X. We analyse this data, which is publicly available from the R package `spls`. The yeast cell cycle data was also analysed in Chen & Huang (2012) via sparse reduced rank regression. We call their estimator $\hat{C}_*$. Scientifically, 21 transcription factors of the 106 were verified by Wang et al. (2007) to be responsible for cell cycle regulation.

The proposed estimator $\hat{C}_{\mathrm{RR}}$ identified 33 transcription factors; the corresponding numbers for $\widetilde{C}$ and $\hat{C}_*$ were 48 and 69, respectively. Of the 21 verified transcription factors, the proposed method selected 14, whereas Chun & Keleş (2010) and Chen & Huang (2012) selected 14 and 16, respectively. Ten additional transcription factors that regulate cell cycle were identified by Lee et al. (2002), out of which three transcription factors were selected by our proposed method. Figure 2 plots the posterior mean, $\hat{C}_{\mathrm{RR}}$, and 95% symmetric pointwise credible intervals for two common effects, ACE2 and SW14, which are identified by all the methods. Similar periodic patterns of the estimated effects are observed for the other two methods, perhaps unsurprisingly due to the two cell cycles during which the mRNA measurements were taken. Similar plots for the remaining 19 effects identified by our method are provided in the Supplementary Material.

The proposed automatic rank detection technique estimated a rank of 1, which is significantly different from Chen & Huang (2012), who estimated it to be 4, and Chun & Keleş (2010), who estimated it to be 8. The singular values of $Y - X\hat{C}_{\mathrm{R}}$ showed a significant drop in magnitude after the first four values, which agrees with the findings in Chen & Huang (2012). The ten-fold cross-validation error with a postulated rank of 4 for $\hat{C}_{\mathrm{RR}}$ was 0.009, and that of $\widetilde{C}$ was 0.19.

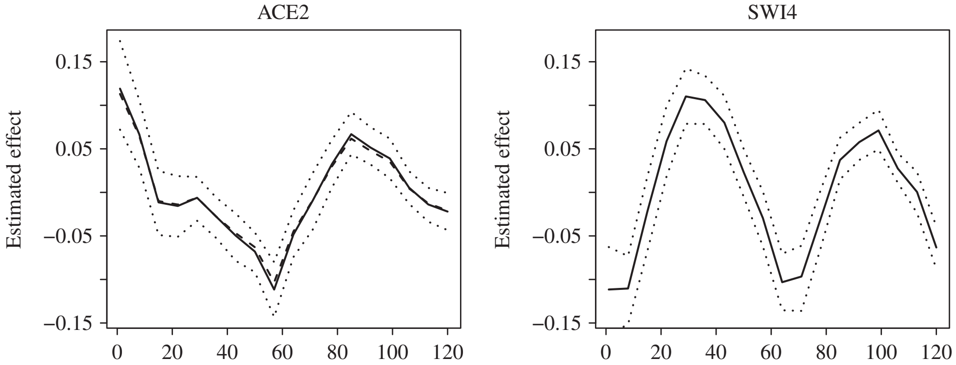

Fig. 2. Estimated effects of ACE2 and SWI4, two of 33 transcription factors with nonzero effects on cell cycle regulation. Both have been scientifically verified by Wang et al. (2007). The dotted lines correspond to 95% posterior symmetric credible intervals, the bold lines represent the posterior mean and the dashed lines plot values of the estimate $\hat{C}_{\text{RR}}$.

We repeated the entire analysis with a nondiagonal Σ, which was assigned an inverse-Wishart prior. No changes in the identification of transcription factors or rank estimation were detected.

6. Concentration results

In this section we establish a minimax posterior concentration result under the prediction loss when the number of covariates is allowed to grow subexponentially in n. To the best of our knowledge, this is the first such result in Bayesian reduced rank regression models. We are also not aware of a similar result involving the horseshoe or another polynomial-tailed shrinkage prior in ultra-high-dimensional settings beyond the generalized linear model framework. Armagan et al. (2013) applied the general theory of posterior consistency (Ghosal et al., 2000) to linear models with a growing number of covariates and established consistency for the horseshoe prior with a sample-size-dependent hyperparameter choice when $p = o(n)$. Results (van der Pas et al., 2014; Ghosh & Chakrabarti, 2017) that quantify rates of convergence focus exclusively on the normal means problem, with their proofs crucially exploiting an exact conjugate representation of the posterior mean.

A key ingredient of our theory is a novel non-asymptotic prior concentration bound for the horseshoe prior around sparse vectors. The prior concentration or local Bayes complexity (Ghosal et al., 2000; Bhattacharya et al., 2019) is a key component in the general theory of posterior concentration. Let $\ell_0[s;p] = \{\theta_0 \in \mathbb{R}^p : \#(1 \leqslant j < p : \theta_{0j} \neq 0) \leqslant s\}$ denote the space of p-dimensional vectors with at most s nonzero entries.

Lemma 1. *Let Π_{HS} denote the horseshoe prior on $\mathbb{R}^p$ given by the hierarchy $\theta_j \mid \lambda_j, \tau \sim N(0, \lambda_j^2\tau^2)$, $\lambda_j \sim \text{Ca}_+(0,1)$, $\tau \sim \text{Ca}_+(0,1)$. Fix $\theta_0 \in \ell_0[s;p]$ and let $S = \{j : \theta_{0j} \neq 0\}$. Assume $s = o(p)$ and $\log p \leqslant n^\gamma$ for some $\gamma \in (0,1)$, and $\max|\theta_{0j}| \leqslant M$ for some $M > 0$ for $j \in S$. Define $\delta = \{(s \log p)/n\}^{1/2}$. Then*

$$\Pi_{\text{HS}}\big(\theta : \|\theta - \theta_0\|_2 < \delta\big) \geqslant \mathrm{e}^{-Ks\log p}$$

for some positive constant K.

A proof of the result is provided in the Supplementary Material. We believe Lemma 1 will be of independent interest in various other models involving the horseshoe prior, for example high-dimensional regression and factor models. The only other instance of a similar prior concentration result for a continuous shrinkage prior in $p \gg n$ settings that we are aware of is for the Dirichlet–Laplace prior (Pati et al., 2014).

We now study concentration properties of the posterior distribution in model (3) in $p \gg n$ settings. To aid the theoretical analysis we adopt the fractional posterior framework of Bhattacharya et al. (2019), where a fractional power of the likelihood function is combined with a prior using the usual Bayes formula to arrive at a fractional posterior distribution. Specifically, fix $\alpha \in (0, 1)$, recall the prior Π_C on C defined after equation (4) and set Π_Σ as the inverse-Wishart prior for Σ. The α-fractional posterior for (C, Σ) under model (3) is then given by

$$\Pi_{n,\alpha}(C, \Sigma \mid Y) \propto \{p^{(n)}(Y \mid C, \Sigma; X)\}^{\alpha}\, \Pi_C(C)\, \Pi_\Sigma(\Sigma). \tag{10}$$

Assuming the data is generated with a true coefficient matrix C_0 and a true covariance matrix Σ_0, we now study the frequentist concentration properties of $\Pi_{n,\alpha}(\cdot \mid Y)$ around (C_0, Σ_0). The adoption of the fractional framework is primarily for technical convenience; refer to the Supplementary Material for a detailed discussion. We additionally discuss the closeness of the fractional posterior to the usual posterior in the next subsection.

We first list our assumptions on the truth.

Assumption 1 (Growth of number of covariates). The sample size n and the number of covariates p satisfy $\log p/n^{\gamma} \leqslant 1$ for some $\gamma \in (0, 1)$.

Assumption 2. The number of response variables q is fixed.

Assumption 3 (True coefficient matrix). The true coefficient matrix C_0 admits the decomposition $C_0 = B_0 A_0^{\mathrm{T}}$, where $B_0 \in \mathbb{R}^{p\times r_0}$ and $A_0 \in \mathbb{R}^{q\times r_0}$ for some $r_0 = \kappa q$, $\kappa \in \{1/q, 2/q, \ldots, 1\}$. We additionally assume that A_0 is semi-orthogonal, i.e., $A_0^{\mathrm{T}} A_0 = \mathrm{I}_{r_0}$, and all but s rows of B_0 are identically zero for some $s = o(p)$. Finally, $\max\limits_{j,h} |C_{0jh}| < T$ for some $T > 0$.

Assumption 4 (Response covariance). The covariance matrix Σ_0 satisfies, for some a_1 and a_2, $0 < a_1 < s_{\min}(\Sigma_0) < s_{\max}(\Sigma_0) < a_2 < \infty$, where $s_{\min}(P)$ and $s_{\max}(P)$ are respectively the minimum and maximum singular values of a matrix P.

Assumption 5 (Design matrix). For X_j the jth column of X, $\max_{1\leqslant j\leqslant p} \|X_j\| \asymp n$.

Assumption 1 allows the number of covariates p to grow at a subexponential rate of $\mathrm{e}^{n^{\gamma}}$ for some $\gamma \in (0, 1)$. Assumption 2 can be relaxed to let q grow slowly with n. Assumption 3 posits that the true coefficient matrix C_0 admits a reduced-rank decomposition with the matrix B_0 row-sparse. The orthogonality assumption on true A_0 is made to ensure that B_0 and C_0 have the same row-sparsity (Chen & Huang, 2012). The positive definiteness of Σ_0 is ensured by Assumption 4. Finally, Assumption 5 is a standard minimal assumption on the design matrix and is satisfied with large probability if the elements of the design matrix are independently drawn from a fixed probability distribution, such as $N(0, 1)$ or any sub-Gaussian distribution. It also encompasses situations when the columns of X are standardized.

Let $p_0^{(n)}(Y \mid X) \equiv p^{(n)}(Y \mid C_0, \Sigma_0; X)$ denote the true density. For two densities q_1, q_2 with respect to a dominating measure μ, recall the squared Hellinger distance $h^2(q_1, q_2) = \{(1/2)\int (q_1^{1/2} - q_2^{1/2})^2\, \mathrm{d}\mu\}$. As a loss function to measure closeness between (C, Σ) and (C_0, Σ_0), we consider the squared Hellinger distance h^2 between the corresponding densities $p(\cdot \mid C, \Sigma; X)$ and $p_0(\cdot \mid X)$. It is common to use h^2 to measure the closeness of the fitted density to the truth in high-dimensional settings; see, e.g., Jiang (2007). In the following theorem we provide a non-asymptotic bound to the squared Hellinger loss under the fractional posterior $\Pi_{n,\alpha}$.

Theorem 1. *Suppose $\alpha \in (0,1)$ and let $\Pi_{n,\alpha}$ be defined as in (10). Suppose Assumptions 1–5 are satisfied. Let the joint prior on (C,Σ) be defined by the product prior Π_C and Π_Σ, where Π_Σ is the inverse-Wishart prior with parameters (q, I_q). Define $\widetilde{\epsilon}_n = \max\{K_1 \log \rho / s^2_{\min}(\Sigma_0), 4/s^2_{\min}(\Sigma_0)\}\epsilon_n$, where $\rho = s_{\max}(\Sigma_0)/s_{\min}(\Sigma_0)$, K_1 is an absolute positive constant and $\epsilon_n = \{(qr_0 + r_0 s \log p)/n\}^{1/2}$. Then, for for any $D \geqslant 1$ and $t > 0$,*

$$\Pi_{n,\alpha}\left[(C,\Sigma) : h^2\left\{p^{(n)}(Y \mid C, \Sigma; X), p_0^{(n)}(Y \mid X)\right\} \geqslant \frac{(D+3t)}{2(1-\alpha)} n\widetilde{\epsilon}_n^{\,2} \mid Y\right] \leqslant \mathrm{e}^{-tn\widetilde{\epsilon}_n^{\,2}},$$

with $P^{(n)}_{(C_0,\Sigma_0)}$-probability at least $1 - K_2/\{(D-1+t)n\widetilde{\epsilon}_n^{\,2}\}$ for sufficiently large n and some positive constant K_2.

The proof of Theorem 1, provided in the Appendix, hinges upon establishing sufficient prior concentration around C_0 and Σ_0 for our choices of Π_C and Π_Σ, which in turn drives the concentration of the fractional posterior. Specifically, building upon Lemma 1 we prove in the Supplementary Material that for our choice of Π_C we have sufficient prior concentration around row and rank sparse matrices.

Bunea et al. (2012) obtained $n\epsilon_n^2 = (qr_0 + r_0 s \log p)$ as the minimax risk under the loss $\|XC - XC_0\|_{\mathrm{F}}^2$ for model (1) with $\Sigma = \mathrm{I}_q$ and when C_0 satisfies Assumption 3. Theorem 1 can then be viewed as a more general result with unknown covariance. Indeed, if $\Sigma = \mathrm{I}_q$, we recover the minimax rate ϵ_n as the rate of contraction of the fractional posterior as stated in the following theorem. Furthermore, we show that the fractional posterior mean as a point estimator is rate optimal in the minimax sense. For a given $\alpha \in (0,1)$ and $\Sigma = \mathrm{I}_q$, the fractional posterior simplifies to $\Pi_{n,\alpha}(C \mid Y) \propto \{p(Y \mid C, \mathrm{I}_q; X)\}^\alpha \Pi_C$.

Theorem 2. *Fix $\alpha \in (0,1)$. Suppose Assumptions 1–5 are satisfied and assume that Σ is known and, without loss of generality, equals I_q. Let ϵ_n be defined as in Theorem 1. Then, for any $D \geqslant 2$ and $t > 0$,*

$$\Pi_{n,\alpha}\left\{C \in \mathbb{R}^{p\times q} : \frac{1}{nq}\|XC - XC_0\|_{\mathrm{F}}^2 \geqslant \frac{2(D+3t)}{\alpha(1-\alpha)}\epsilon_n^2 \mid Y\right\} \leqslant \mathrm{e}^{-tn\epsilon_n^2}$$

holds with $P^{(n)}_{C_0}$-probability at least $1 - 2/\{(D-1+t)n\epsilon_n^2\}$ for sufficiently large n. Moreover, if $\bar{C} = \int C\Pi_{n,\alpha}(\mathrm{d}C)$, then, with $P^{(n)}_{C_0}$-probability at least $1 - K_1/\{n\epsilon_n^2\}$,

$$\|X\bar{C} - XC_0\|_{\mathrm{F}}^2 \leqslant K_2\,(qr_0 + r_0 s \log p)$$

for some positive constants K_1 and K_2 independent of α.

The proof of Theorem 2 is provided in the Appendix. The optimal constant multiple of ϵ_n^2 is attained at $\alpha = 1/2$. This is consistent with the optimality of the half-power in Leung & Barron (2006) in the context of a pseudolikelihood approach for model aggregation of least squares estimates, which shares a Bayesian interpretation as a fractional posterior.

7. Fractional and standard posteriors

From a computational point of view, for model (1), raising the likelihood to a fractional power only results in a change in the (co)variance term, and hence our Gibbs sampler discussed subsequently can be easily adapted to sample from the fractional posterior. We conducted numerous simulations with values of α close to 1 and obtained virtually indistinguishable point estimates compared to the full posterior; details are provided in the Supplementary Material. In this subsection we study the closeness between the fractional posterior $\Pi_{n,\alpha}(\cdot \mid Y)$ and the standard posterior $\Pi_n(\cdot \mid Y)$ for model (1) with prior $\Pi_C \otimes \Pi_\Sigma$ in terms of the total variation metric. Proofs of the results in this section are collected in the Supplementary Material.

Recall that for two densities g_1 and g_2 with respect to some measure μ, the total variation distance between them is given by $\|g_1 - g_2\|_{\text{TV}} = \int |g_1 - g_2| \, \mathrm{d}\mu = \sup_{B \in \mathcal{B}} |G_1(B) - G_2(B)|$, where G_1 and G_2 denote the corresponding probability measures.

Theorem 3. *Consider model* (1) *with* $C \sim \Pi_C$ *and* $\Sigma \sim \Pi_\Sigma$. *Then*

$$\lim_{\alpha \to 1_-} \left\| \Pi_{n,\alpha}(C, \Sigma \mid Y) - \Pi_n(C, \Sigma \mid Y) \right\|_{\text{TV}} = 0$$

for every $Y \sim P_{C_0}$.

Bhattacharya et al. (2019) proved a weak convergence result under a more general set-up, whereas Theorem 3 provides a substantial improvement to show strong convergence for the Gaussian likelihood function considered here. The total variation distance is commonly used in Bayesian asymptotics to justify posterior merging of opinion, i.e., the total variation distance between two posterior distributions arising from two different priors vanish as the sample size increases. Theorem 3 has a similar flavour, with the exception that the merging of opinion takes place under small perturbations of the likelihood function.

We conclude this section by showing that the regular posterior $\Pi_n(C, \Sigma \mid Y)$ is consistent, leveraging on the contraction of the fractional posteriors $\Pi_{n,\alpha}(C, \Sigma \mid Y)$ for any $\alpha < 1$ in combination with Theorem 3 above. For ease of exposition we assume $\Sigma = \text{diag}(\sigma_1^2, \ldots, \sigma_q^2)$ and $\Pi_\Sigma(\cdot)$ is a product prior with components $\text{IG}(a, b)$ for some $a, b > 0$. Similar arguments can be made for the inverse-Wishart prior when Σ is nondiagonal.

Theorem 4. *Assume* $\Sigma = \textit{diag}(\sigma_1^2, \ldots, \sigma_q^2)$ *in model* (1) *with priors* Π_C *and a product* $\text{IG}(a, b)$ *prior on* Σ *with* $a, b > 0$. *For any* $\epsilon > 0$ *and sufficiently large* M,

$$\lim_{n \to \infty} \Pi_n\left[(C, \Sigma) : \frac{1}{n} h^2 \left\{ p^{(n)}(\cdot \mid C, \Sigma; X), p_0^{(n)}(\cdot \mid X) \right\} \geqslant M\epsilon \mid Y \right] \to 0$$

almost surely under $P_{(C_0, \Sigma_0)}$.

Theorem 4 establishes consistency of the regular posterior under the average Hellinger metric $n^{-1} h^2 \left\{ p^{(n)}(\cdot \mid C, \Sigma; X), p_0^{(n)}(\cdot \mid X) \right\}$. This is proved using a novel representation of the regular posterior as a fractional posterior under a different prior, a trick which we believe will be useful to arrive at similar consistency results in various other high-dimensional Gaussian models. For any $\alpha \in (0,1)$,

$$
\begin{aligned}
\Pi_n(C,\Sigma \mid Y) &\propto |\Sigma|^{-n/2}\,\mathrm{e}^{-\mathrm{tr}\{(Y-XC)\Sigma^{-1}(Y-XC)^{\mathrm{T}}\}/2}\,\Pi_C(\mathrm{d}C)\Pi_\Sigma(\mathrm{d}\Sigma)\\
&\propto |\Sigma|^{-n\alpha/2}\,\mathrm{e}^{-\alpha\mathrm{tr}\{(Y-XC)(\alpha\Sigma)^{-1}(Y-XC)^{\mathrm{T}}\}/2}\,\Pi_C(\mathrm{d}C)\,|\Sigma|^{-n(1-\alpha)/2}\,\Pi_\Sigma(\mathrm{d}\Sigma)\\
&\propto |\Sigma_*|^{-n\alpha/2}\,\mathrm{e}^{-\alpha\mathrm{tr}\{(Y-XC)\Sigma_*^{-1}(Y-XC)^{\mathrm{T}}\}/2}\,\Pi_C(\mathrm{d}C)\Pi_{\Sigma_*}(\mathrm{d}\Sigma_*)\\
&\propto \Pi_{n,\alpha}(C,\Sigma_* \mid Y),
\end{aligned}
$$

where $\Sigma_* = \alpha\Sigma$ and, from a simple change of variable, $\Pi_{\Sigma_*}(\cdot)$ is again a product of inverse-Gamma densities with each component a, $\textsc{ig}\{n(1-\alpha)/2 + a, \alpha b\}$. Since the first and last expressions in the above display are both probability densities, we conclude that $\Pi_n(C,\Sigma \mid Y) = \Pi_{n,\alpha}(C,\Sigma_* \mid Y)$. This means that the regular posterior distribution of (C,Σ) can be viewed as the α-fractional posterior distribution of (C,Σ_*), with the prior distribution of Σ_* dependent on both n and α. Following an argument similar to Theorem 1, we only need to show the prior concentration of (C,Σ_*) around the truth to obtain posterior consistency of $\Pi_{n,\alpha}(C,\Sigma_* \mid Y)$, and hence equivalently of $\Pi_n(C,\Sigma \mid Y)$. Some care is needed to show the prior concentration of Σ_* with an n-dependent prior, which can be managed by setting $\alpha = 1 - 1/(\log n)^t$ for some appropriate $t > 1$.

Acknowledgement

The authors thank the editor, an associate editor and two reviewers for their constructive comments, which substantially helped improve the paper. Bhattacharya acknowledges support from the National Science Foundation. Mallick was supported by the National Cancer Institute.

Supplementary material

Supplementary material available at *Biometrika* online includes further simulation results, discussion of the fractional framework and proof of Lemma 1. Code for implementation of our method is available at `https://github.com/antik015`.

Appendix

Lemmas numbered S1, S2, etc. refer to technical lemmas included in the Supplementary Material.

For two densities p_θ and p_{θ_0} with respect to a common dominating measure μ and indexed by parameters θ and θ_0, respectively, the Rényi divergence of order $\alpha \in (0,1)$ is $D_\alpha(\theta,\theta_0) = (\alpha-1)^{-1}\log\int p_\theta^\alpha p_{\theta_0}^{1-\alpha}\,\mathrm{d}\mu$. The α-affinity between p_θ and p_{θ_0} is denoted by $A_\alpha(p_\theta,p_{\theta_0}) = \int p_\theta^\alpha p_{\theta_0}^{1-\alpha}\,\mathrm{d}\mu = \mathrm{e}^{-(1-\alpha)D_\alpha(p_\theta,p_{\theta_0})}$. See Bhattacharya et al. (2019) for a review of Rényi divergences.

Proof of Theorem 1

Fix $\alpha \in (0,1)$. Define $U_n = \left[(C,\Sigma) : \dfrac{1}{n}D_\alpha\{(C,\Sigma),(C_0,\Sigma_0)\} > \dfrac{D+3t}{1-\alpha}\tilde{\epsilon}_n^2\right]$. Let $\eta = (C,\Sigma)$ and $\eta_0 = (C_0,\Sigma_0)$. For convenience we abbreviate $p^{(n)}(Y \mid C,\Sigma;X)$ by $p_\eta^{(n)}$ and write $p_{\eta_0}^{(n)}$ for $p_0^{(n)}(Y \mid X)$. Finally, let Π_η denote the joint prior $\Pi_C \times \Pi_\Sigma$. Then, the α-fractional posterior probability assigned to the set U_n can be written as

$$
\Pi_{n,\alpha}(U_n \mid Y) = \frac{\int_{U_n}\mathrm{e}^{-\alpha r_n(\eta,\eta_0)}\,\mathrm{d}\Pi_\eta}{\int \mathrm{e}^{-\alpha r_n(\eta,\eta_0)}\,\mathrm{d}\Pi_\eta} := \frac{N_n}{D_n},
$$

where $r_n(\eta, \eta_0) = \log p^{(n)}_{\eta_0}/p^{(n)}_{\eta}$. We prove in Lemma S6 of the Supplementary Material that, with $P^{(n)}_{(C_0,\Sigma_0)}$-probability at least $1 - K_2/\{(D-1+t)^2 n\widetilde{\epsilon}_n^2\}$, $D_n \geqslant \mathrm{e}^{-\alpha(D+t)n\widetilde{\epsilon}_n^2}$ for some positive constant K_2. For the numerator, proceeding similarly to the proof of Theorem 3.2 in Bhattacharya et al. (2019), we arrive at $P^{(n)}_{(C_0,\Sigma_0)}\{N_n \leqslant \mathrm{e}^{-(D+2t)n\widetilde{\epsilon}_n^2}\} \geqslant 1 - 1/\{(D-1+t)^2 n\widetilde{\epsilon}_n^2\}$. Combining the upper bound for N_n and lower bound for D_n, we then have

$$\Pi_{n,\alpha}\left[(C,\Sigma) : \frac{1}{n}D_\alpha\{(C,\Sigma),(C_0,\Sigma_0)\} \geqslant \frac{(D+3t)}{1-\alpha}\widetilde{\epsilon}_n{}^2 \mid Y\right] \leqslant \mathrm{e}^{-tn\widetilde{\epsilon}_n{}^2},$$

with $P^{(n)}_{\eta_0}$-probability at least $1 - K_2/\{(D-1+t)^2 n\widetilde{\epsilon}_n^2\}$. The result then follows from the equivalence of Rényi divergences given in equations (R2) and (R3) in Bhattacharya et al. (2019). □

Proof of Theorem 2

For $C \in \mathbb{R}^{p\times q}$, we write $p^{(n)}_C$ to denote the density of $Y \mid X$, which is proportional to $\mathrm{e}^{-\operatorname{tr}\{(Y-XC)(Y-XC)^{\mathrm{T}}\}/2}$. For any $C^* \in \mathbb{R}^{p\times q}$ we define a ϵ-neighbourhood as

$$B_n(C^*,\epsilon) = \left\{C \in \mathbb{R}^{p\times q} : \int p^{(n)}_{C^*}\log(p^{(n)}_{C^*}/p^{(n)}_{C})\,\mathrm{d}Y \leqslant n\epsilon^2, \int p^{(n)}_{C^*}\log^2(p^{(n)}_{C^*}/p^{(n)}_{C})\,\mathrm{d}Y \leqslant n\epsilon^2\right\}.$$

Observe that $B_n(C_0,\epsilon) \supset A_n(C_0,\epsilon) = \left\{C \in \mathbb{R}^{p\times q} : n^{-1}\|XC - XC_0\|_{\mathrm{F}}^2 \leqslant \epsilon^2\right\}$ for all $\epsilon > 0$, and the Rényi divergence $D_\alpha(p^{(n)}_C, p^{(n)}_{C_0}) = (\alpha/2)\|XC - XC_0\|_{\mathrm{F}}^2$. By a similar argument to step 1 of the proof of Lemma S6 of the Supplementary Material, we have that $\Pi_C\{A_n(C_0,\epsilon_n)\} \geqslant \mathrm{e}^{-Kn\epsilon_n^2}$ for positive K. Hence, the first part follows from Bhattacharya et al. (2019, Theorem 3.2).

For the second part, first observe that from Bhattacharya et al. (2019, Corollary 3.3) we get $\int (nq)^{-1}\|XC - XC_0\|_{\mathrm{F}}^2\Pi_{n,\alpha}(\mathrm{d}C \mid Y) \leqslant K_2\{\alpha(1-\alpha)\}^{-1}\epsilon_n$ with $P^{(n)}_{C_0}$-probability at least $1 - K_1/\{n\epsilon_n^2\}$, where K_1 and K_2 are positive constants independent of α; see the Supplementary Material for a precise statement of the corollary. Using the convexity of the Frobenius norm and applying Jensen's inequality, we get $(\alpha/2)\|X\bar{C} - XC_0\|_{\mathrm{F}}^2 = (\alpha/2)\|X\int C\,\Pi_{n,\alpha}(\mathrm{d}C) - X\int C_0\,\Pi_{n,\alpha}\|_{\mathrm{F}}^2 \leqslant \alpha/2\int\|XC - XC_0\|_{\mathrm{F}}^2\Pi_{n,\alpha}(\mathrm{d}C) \leqslant K_2 n(1-\alpha)^{-1}\epsilon_n^2$. □

References

Alquier, P. (2013). Bayesian methods for low-rank matrix estimation: Short survey and theoretical study. In *Proc. 24th Int. Conf. Algorithmic Learning Theory*, S. Jain, R. Munos, F. Stephan & Th. Zeugmann, eds, pp. 309–23. New York: Springer.

Anderson, T. (1984). *Multivariate Statistical Analysis*. New York: Wiley.

Anderson, T. (2002). Specification and misspecification in reduced rank regression. *Sankhyā* **64**, 193–205.

Anderson, T. W. (1951). Estimating linear restrictions on regression coefficients for multivariate normal distributions. *Ann. Math. Statist.* **22**, 327–51.

Armagan, A., Dunson, D. B., Lee, J., Bajwa, W. U. & Strawn, N. (2013). Posterior consistency in linear models under shrinkage priors. *Biometrika* **100**, 1011–18.

Babacan, S. D., Molina, R. & Katsaggelos, A. K. (2011). Variational Bayesian super resolution. *IEEE Trans. Image Proces.* **20**, 984–99.

Barbieri, M. M. & Berger, J. O. (2004). Optimal predictive model selection. *Ann. Statist.* **32**, 870–97.

Bhadra, A. & Mallick, B. K. (2013). Joint high-dimensional Bayesian variable and covariance selection with an application to eQTL analysis. *Biometrics* **69**, 447–57.

Bhattacharya, A., Chakraborty, A. & Mallick, B. K. (2016). Fast sampling with Gaussian scale mixture priors in high-dimensional regression. *Biometrika* **103**, 985–91.

Bhattacharya, A. & Dunson, D. (2011). Sparse Bayesian infinite factor models. *Biometrika* **98**, 291–306.

Bhattacharya, A., Pati, D. & Yang, Y. (2019). Bayesian fractional posteriors. *Ann. Statist.* **47**, 39–66.

Bondell, H. D. & Reich, B. J. (2012). Consistent high-dimensional Bayesian variable selection via penalized credible regions. *J. Am. Statist. Assoc.* **107**, 1610–24.

Brown, P. J., Vannucci, M. & Fearn, T. (1998). Multivariate Bayesian variable selection and prediction. *J. R. Statist. Soc.* B **60**, 627–41.

Bunea, F., She, Y. & Wegkamp, M. H. (2011). Optimal selection of reduced rank estimators of high-dimensional matrices. *Ann. Statist.* **39**, 1282–309.

Bunea, F., She, Y. & Wegkamp, M. H. (2012). Joint variable and rank selection for parsimonious estimation of high-dimensional matrices. *Ann. Statist.* **40**, 2359–88.

Carvalho, C., Polson, N. & Scott, J. (2010). The horseshoe estimator for sparse signals. *Biometrika* **97**, 465–80.

Chen, K., Dong, H. & Chan, K.-S. (2013). Reduced rank regression via adaptive nuclear norm penalization. *Biometrika* **100**, 901–20.

Chen, L. & Huang, J. Z. (2012). Sparse reduced-rank regression for simultaneous dimension reduction and variable selection. *J. Am. Statist. Assoc.* **107**, 1533–45.

Chun, H. & Keleş, S. (2010). Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *J. R. Statist. Soc.* B **72**, 3–25.

George, E. I. & McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *J. Am. Statist. Assoc.* **88**, 881–9.

Geweke, J. (1996). Bayesian reduced rank regression in econometrics. *J. Economet.* **75**, 121–46.

Ghosal, S., Ghosh, J. K. & van der Vaart, A. W. (2000). Convergence rates of posterior distributions. *Ann. Statist.* **28**, 500–31.

Ghosh, P. & Chakrabarti, A. (2017). Asymptotic optimality of one-group shrinkage priors in sparse high-dimensional problems. *Bayesian Anal.* **12**, 1133–61.

Golub, G. H. & van Loan, C. F. (1996). *Matrix Computations*, 3rd ed. Baltimore, MA: Johns Hopkins University Press.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82**, 711–32.

Hahn, P. R. & Carvalho, C. M. (2015). Decoupling shrinkage and selection in Bayesian linear models: A posterior summary perspective. *J. Am. Statist. Assoc.* **110**, 435–48.

Hoff, P. (2009). Simulation of the matrix Bingham–von Mises-Fisher distribution, with applications to multivariate and relational data. *J. Comp. Graph. Statist.* **18**, 438–56.

Izenman, A. J. (1975). Reduced-rank regression for the multivariate linear model. *J. Mult. Anal.* **5**, 248–64.

Jiang, W. (2007). Bayesian variable selection for high-dimensional generalized linear models: Convergence rates of the fitted densities. *Ann. Statist.* **35**, 1487–511.

Kundu, S., Baladandayuthapani, V. & Mallick, B. K. (2013). Bayes regularized graphical model estimation in high dimensions. *arXiv:* 1308.3915.

Lee, T. I., Rinaldi, N. J., Robert, F., Odom, D. T., Bar-Joseph, Z., Gerber, G. K., Hannett, N. M., Harbison, C. T., Thompson, C. M., Simon, I. et al. (2002). Transcriptional regulatory networks in *saccharomyces cerevisiae*. *Science* **298**, 799–804.

Leung, G. & Barron, A. R. (2006). Information theory and mixing least-squares regressions. *IEEE Trans. Info. Theory* **52**, 3396–410.

Lim, Y. J. & Teh, Y. W. (2007). Variational Bayesian approach to movie rating prediction. In *Proc. KDD Cup and Workshop*.

Liu, J. S. & Wu, Y. N. (1999). Parameter expansion for data augmentation. *J. Am. Statist. Assoc.* **94**, 1264–74.

Lopes, H. F. & West, M. (2004). Bayesian model assessment in factor analysis. *Statist. Sinica* **14**, 41–67.

Lucas, J., Carvalho, C., Wang, Q., Bild, A., Nevins, J. & West, M. (2006). Sparse statistical modelling in gene expression genomics. In *Bayesian Inference for Gene Expression and Proteomics* K.-A. Do, P. Müller & M. Vannucci, eds, pp. 155–76. Cambridge: Cambridge University Press.

Pati, D., Bhattacharya, A., Pillai, N. S. & Dunson, D. (2014). Posterior contraction in sparse Bayesian factor models for massive covariance matrices. *Ann. Statist.* **42**, 1102–30.

Polson, N. G. & Scott, J. G. (2010). Shrink globally, act locally: Sparse Bayesian regularization and prediction. *Bayesian Statist.* **9**, 501–38.

Polson, N. G., Scott, J. G. & Windle, J. (2014). The Bayesian bridge. *J. R. Statist. Soc.* B **76**, 713–33.

R Development Core Team (2020). *R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing*, Vienna, Austria. ISBN 3-900051-07-0. http://www.R-project.org.

Rue, H. (2001). Fast sampling of Gaussian Markov random fields. *J. R. Statist. Soc.* **B 63**, 325–38.

Ruffieux, H., Davison, A. C., Hager, J. & Irincheeva, I. (2017). Efficient inference for genetic association studies with multiple outcomes. *Biostatistics* **18**, 618–36.

Salakhutdinov, R. & Mnih, A. (2008). Bayesian probabilistic matrix factorization using Markov chain Monte Carlo. In *Proc. 25th Int. Conf. Machine learning*. New York: Association for Computing Machinery.

Scott, J. G. & Berger, J. O. (2010). Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *Ann. Statist.* **38**, 2587–619.

Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D. & Futcher, B. (1998). Comprehensive identification of cell cycle–regulated genes of the yeast *saccharomyces cerevisiae* by microarray hybridization. *Molec. Biol. Cell* **9**, 3273–97.

van der Pas, S., Kleijn, B. & van der Vaart, A. (2014). The horseshoe estimator: Posterior concentration around nearly black vectors. *Electron. J. Statist.* **8**, 2585–618.

Velu, R. & Reinsel, G. C. (2013). *Multivariate Reduced-Rank Regression: Theory and Applications*, Lecture Notes in Statistics vol. 136. New York: Springer.

WANG, H. (2010). Sparse seemingly unrelated regression modelling: Applications in finance and econometrics. *Comp. Statist. Data Anal.* **54**, 2866–77.

WANG, L., CHEN, G. & LI, H. (2007). Group SCAD regression analysis for microarray time course gene expression data. *Bioinformatics* **23**, 1486–94.

YUAN, M., EKICI, A., LU, Z. & MONTEIRO, R. (2007). Dimension reduction and coefficient estimation in multivariate linear regression. *J. R. Statist. Soc.* B **69**, 329–46.

YUAN, M. & LIN, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Statist. Soc.* B **68**, 49–67.

ZOU, H. (2006). The adaptive lasso and its oracle properties. *J. Am. Statist. Assoc.* **101**, 1418–29.

[*Received on* 23 *September* 2017. *Editorial decision on* 15 *May* 2019]