BAYESIAN SHRINKAGE: COMPUTATION, METHODS AND THEORY

A Dissertation

by

ANTIK CHAKRABORTY

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Bani K. Mallick
Co-Chair of Committee,	Anirban Bhattacharya
Committee Members,	Raymond J. Carroll
	Natarajan Sivakumar
Head of Department,	Valen E. Johnson

August 2018

Major Subject: Statistics

ABSTRACT

Sparsity is a standard structural assumption that is made while modeling high-dimensional statistical parameters. This assumption essentially entails a lower dimensional embedding of the high-dimensional parameter thus enabling sound statistical inference. Apart from this obvious statistical motivation, in many modern applications of statistics such as Genomics, Neuroscience etc. parameters of interest are indeed of this nature.

For over almost two decades, spike and slab type priors have been the Bayesian gold standard for modeling of sparsity. However, due to their computational bottlenecks shrinkage priors have emerged as a powerful alternative. This family of priors can almost exclusively be represented as a scale mixture of Gaussian distribution and posterior Markov chain Monte Carlo (MCMC) updates of related parameters are then relatively easy to design. Although shrinkage priors were tipped as having computational scalability in high-dimensions, when the number of parameters is in thousands or more, they do come with their own computational challenges. Standard MCMC algorithms implementing shrinkage priors generally scale cubic in the dimension of the parameter making real life application of these priors severely limited. The first chapter of this document addresses this computational issue and proposes an alternative exact posterior sampling algorithm complexity of which that linearly in the ambient dimension.

The algorithm developed in the first chapter is specifically designed for regression problems. However, simple modifications of it allows tackling other high-dimensional problems where these priors have found little application. In the second chapter, we develop a Bayesian method based on shrinkage priors for high-dimensional multiple response response regression. We show how proper shrinkage may be used for modeling high-dimensional low-rank matrices. Unlike spike and slab type priors, shrinkage priors are unable to produce exact zeros in the posterior. In this chapter we also devise two independent post MCMC processing schemes based on the idea of soft-thresholding with default choices of tuning parameters. This post processing steps provide exact estimates of the row and rank sparsity in the parameter matrix.

Theoretical study of the posterior convergence rates using shrinakge priors are relatively underdeveloped. While we do not attempt to provide a unifying foundation to study these properties, in chapter three we choose a specific member of the shrinkage family known as the horseshoe prior and study its convergence rates in several high-dimensional models. These results are completely new in the literature and also establish the horseshoe priors optimiality in the minimax sense in high-dimensional problems.

DEDICATION

To my family, Srabana and my deceased yet still the dearest friend Kanji,

ACKNOWLEDGMENTS

I would like to begin by thanking Dr. Bani K. Mallick for graciously welcoming me as one of his students. His enormous knowledge of the subject has never failed to amaze me. Over time, our discussions have grown to be more and more informal and have always provided a way out whenever I faced a roadblock.

About two weeks into the PhD program Dr.Mallick introduced me to Dr.Anirban Bhattacharya. To simply put, my doctoral journey wouldn't have been possible without Dr.Bhattacharya. He is by far the best teacher I have had in my life and I am sure all of his current and future students will feel the same way. In particular, his patience is worth a special mention; looking back I can recall many occasions where I have repeated the same mistake over and over again and yet he had always trusted me to come up with the goods no matter how much time it took.

Dr. Carroll is an inspiration. His work ethic and insatiable quest for statistical innovation is highly contiguous. Every single member of the Statistics department is in some way inspired by him. I was fortunate that during my stay at the department Dr. Carroll taught one course on Measurement error models. Dr. Carroll's email praising my work after my preliminary examination is one possession I will forever cherish in my life. I am also very grateful to him for introducing me to the work of Dr. Abhra Sarkar. Dr. Sarkar's support has helped me pass through many difficult times.

Dr. Sivakumar has dedicated his entire life towards mathematics education and I would highly recommend graduate students to sit in his Real Analysis course. Coming out of the course I have at least started to appreciate the beauty of mathematical rigor if nothing else.

I would also like to thank our head of the department Dr. Valen Johnson whose course on Bayesian statistics is still one of my favorite courses. I believe our department is very fortunate to have him as its leader. Thanks also to Dr. Samiran Sinha for helping me out during my application process. Dr. Debdeep Pati, whom I have come to know only from last year is a great addition to our department. I hope to collaborate with him in my future work. Special thanks goes to Dr.

Jeffery Hart and Dr. Darren Cline for helping me out whenever necessary.

The statistics department at Texas A&M is unthinkable without Dr. Michael Longnecker. His relentless love for the students is a testament to his great personality and I am thankful to him for all the support over the years. Like any other institution our department's support system was its incredible staff. Specifically, I would mention Sandra Wood, Athena Robinson, Andrea Dawson for all the help.

In my five years of PhD, I have developed many friendships which are very dear to me. Soutrik Mandal, with whom I started my PhD and shared an apartment for all this time is a wonderful human being. I wish him all the success in his upcoming years as a postdoc. My fellow graduate students Riddhi Pratim Ghosh, Shahina Rahman, Satwik Acharya, Subhodeep Chakraborty have made my days in College Station very memorable. Along this journey I have developed strong bonds with Siddhartha Dey, Sabyasachi Chakraborty, Dhruv Kathuria. I wish we stay connected with each other in the years to come.

Finally, I would like to thank my family for being there with me during this time, for constantly motivating me and believing in me. It is hard to put into words the contribution of Srabana, my fiancée, to this endeavor from the other corner of the globe. My decisions to come to Texas A&M and to spend the rest of my life with her are the two most wisest decisions I have ever taken and I am very proud of them. For my dearest childhood friend Kanji, I would like to say that I have missed you all this time. If there is afterlife, I would definitely see you there and continue our timeless tea sessions.

CONTRIBUTORS AND FUNDING SOURCES

Funding Sources

I am grateful to the National Science Foundation (NSF), International Society for Bayesian Analysis (ISBA) for their help in financing my graduate studies and sponsoring my conference travels.

In my last year as a graduate student I was supported by the Texas A&M Multidisciplinary research grant. I would like to thank the university for that.

NOMENCLATURE

MCMC	Markov chain Monte Carlo
BSML	Bayesian sparse multiple shrinkage
SPLS	Sparse partial least squares
LASSO	Least absolute shrinkage and selection operator
HS	Horseshoe prior
SCAD	Smoothly clipped absolute deviation
MCP	Minimax concave penalty
$\Re^p$	p dimensional Euclidean space
$l_0[s;p]$	Subspace of $\Re^p$ containing vectors with at most s non-zero entries
$s_{\min}(A)$	Smallest singular value of the matrix A
$s_{\max}(A)$	Largest singular value of the matrix A
mle	Maximum likelihood estimator
$indp.$	Independently
$\|x\|_2$	Euclidean norm of a vector x
$a \lesssim b$	$a \leq Cb$ for some positive constant C
$\|A\|_F$	Frobenius norm of the matrix A
$\|A\|_2$	Operator norm of the matrix A
$h^2(p,q)$	Squared Hellinger distance between densities p and q
$D_\alpha(p,q)$	Rényi divergence between densities p and q for $\alpha \in (0,1)$
$\mathbf{1}_A$	Indicator function on the set A

TABLE OF CONTENTS

Page

ABSTRACT ii

DEDICATION iv

ACKNOWLEDGMENTS v

CONTRIBUTORS AND FUNDING SOURCES vii

NOMENCLATURE viii

TABLE OF CONTENTS ix

LIST OF FIGURES xi

LIST OF TABLES xii

1. BAYESIAN SHRINKAGE 1

    1.1 Introduction 1
    1.2 Research questions and motivation 6
    1.3 Outilne 7

2. FAST SAMPLING OF GAUSSIAN SCALE MIXTURE PRIORS IN HIGH-DIMENSIONAL REGRESSION 9

    2.1 Introduction 9
    2.2 The algorithm 11
        2.2.1 Algorithm Derivation 14
    2.3 Application to Bayesian shrinkage priors 15
    2.4 Frequentist operating characteristics in high dimensions 17
    2.5 Discussion 20

3. BAYESIAN SPARSE MULTIPLE REGRESSION FOR SIMULTANEOUS RANK REDUCTION AND VARIABLE SELECTION 22

    3.1 Introduction 22
    3.2 Bayesian sparse multitask learner 24
        3.2.1 Model and Prior Specification 24
    3.3 Posterior Computation 26
        3.3.1 Post processing for variable selection 29

3.3.2 Post processing for rank estimation 30
3.4 Simulation Results 30
3.5 Yeast Cell Cycle Data 34
3.5.1 Concentration results 35
3.5.2 Fractional and usual posterior for reduced rank models with prior Π_C 43
3.6 Definitions required for proofs of Theorem 3.5.7, 3.5.8, 3.5.9 45
3.7 Derivation of equation (3.8) 46
3.8 Derivation of equation (3.9) 46
3.9 Further results on Yeast cell cycle data 47

4. RELATED PROOFS FROM CHAPTER 3 49

4.1 Prior concentration results 49
4.2 Denominator in the proof of Theorem 3.5.7 55

5. CONVERGENCE RATES OF FRACTIONAL HORSESHOE POSTERIORS IN HIGH-DIMENSIONS 59

5.1 Introduction 59
5.2 Notation 61
5.3 The horseshoe prior 62
5.4 Fractional posteriors 67
5.5 High-dimensional sparse linear regression 69
5.6 High-dimensional sparse factor models 71

6. SUMMARY AND CONTRIBUTIONS 75

REFERENCES 76

APPENDIX A. FRACTIONAL VERSUS USUAL POSTERIOR 86

LIST OF FIGURES

FIGURE Page

1.1 Marginal distribution for the Dirichlet-Laplace prior with hyperparameter $1/2$ and the horseshoe prior are plotted around a neighborhood of zero. The Laplace and Cauchy densities are also plotted for reference. 5

1.2 Same densities as in figure 1.1. Here the focus is on the tail behavior of the priors. .. 5

2.1 Linear regression with horseshoe prior [1] on the regression coefficients. Logarithm of time (in seconds) to complete 6000 iterations of a Gibbs sampler is reported. Algorithm 1 and proposed Algorithm 2 were applied to sample from the conditional posterior of β in (2.5). Sample size n was fixed at 100 and the dimension p was varied from 500 to 5000, with 500 step size. Additionally the case $p = 200$ was included. 17

2.2 Boxplots of ℓ_1, ℓ_2 and prediction error across 100 simulation replicates. HS_{me} and HS_{m} respectively denote posterior point wise median and mean for the horeshoe prior. True β_0 is 5-sparse with non-zero entries from setting (a). Top row: $\Sigma = \text{I}_p$ (independent). Bottom row: $\Sigma_{jj} = 1, \Sigma_{jj'} = 0.5, j \neq j'$ (compound symmetry). 18

2.3 Same setting as in Fig 2.2. True β_0 is 5-sparse with non-zero entries from setting (b). 19

3.1 Average sensitivity and specificity across 50 replicates is plotted for different choices of the postulated rank. Here $(p, q, r_0) = (1000, 12, 3)$. Values for BSML (SPLS) are in bold (dashed). 35

3.2 Estimated effects of ACE2 and SWI4, two of 33 transcription factors with non-zero effects on cell cycle regulation. Both have been scientifically verified by [2]. Dotted lines correspond to 95% posterior symmetric credible intervals, bold lines represent the posterior mean and the dashed lines plot values of the BSML estimate $\widehat{C}_{RR}$. 36

3.3 Estimated effects of the 19 of 21 scientifically verified transcription factors selected by the proposed method. Effects of other two, viz. ACE2 and SWI4 are included in the main manuscript. Red lines correspond to 95% posterior symmetric credible intervals, black lines represent the posterior mean and the blue dashed line plots values of the BSML estimate $\widehat{C}_{RR}$..... 48

LIST OF TABLES

TABLE Page

2.1 Same setting as in Fig. 2.1. Absolute time (in seconds) to run 6000 iterations of the Gibbs sampler reported for the two algorithms for chosen values of p. 17

2.2 Frequentist coverage probabilities and lengths of 95% intervals for the LASSO and the horseshoe. The confidence intervals for the LASSO were constructed using the method in [3]. The intervals for the horseshoe are the usual symmetric posterior credible intervals. For both methods, the average coverage probabilities and lengths of the intervals are reported after averaging across all signal variables (row 1 and 2) and noise variables (row 3 and 4). Numbers in the subscript denote the standard errors corresponding to the average coverage probabilities. 21

3.1 Estimation and predictive performance of the proposed method (BSML) versus SPLS 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 were 10 and sample size was 100. For each combination of (p, q, r_0) the columns of the design matrix were generated from $\mathrm{N}(0, \Sigma_X)$. Two different choices of Σ_X was considered. $\Sigma_X = \mathrm{I}_p$ (independent) and $\Sigma_X = (\sigma^X_{ij})$,$\sigma^X_{jj} = 1$,$\sigma^X_{ij} = 0.5$ for $i \neq j$ (correlated). The method achieving superior performance for each setting is highlighted in bold. 33

3.2 Variable selection performance of the proposed method in a non-diagonal error structure setting with independent and correlated predictors; $e_i \sim \Sigma$, $\sigma_{ii} = 1, \sigma_{ij} = 0.5$. Sensitivity and specificity of BSML is compared with [4]. 34

A.1 Empirical results comparing $\hat{r}$, MSPE = $(nq)^{-1}\|XC - XC_0\|^2_F$ and MSE=$(pq)^{-1}\|C - C_0\|^2_F$ for different choices of the fractional power α. $\alpha = 1$ corresponds to the usual posterior. The data used in this table was generated in a similar manner as described in section 3.4 of this document. 89

1. BAYESIAN SHRINKAGE

1.1 Introduction

Recent technological advances in different branches of science has enabled scientists to gather, collect and process extremely complex high-dimensional data. Classical statistical techniques are known to be inadequate in analyzing these data sets. However, in the last twenty years statisticians have made a conscious effort to develop statistical techniques specially designed to handle complex high-dimensional data. Popularly known as the the 'small n, large p' regime, these techniques have also ushered in new theoretical works basing them on sound foundations.

When considering a high-dimensional statistical model, one often assumes certain structure in the parameters for developing inferential tools. For example, in a linear regression problem $y = X\beta + \epsilon$, when the design matrix X has more columns than rows, traditional least squares estimates do not exist. A common assumption made in this case is that the parameter vector β is *sparse*. This means only a small proportion of entries of β are non-zero and most of its entries are zero or very close to zero. If X has n rows and p columns, then assuming a sparse β essentially shrinks the parameter space from $\Re^p$ to a subspace close to the origin. Restricting the parameter space to such a subspace then lays the groundwork for developing statistical methods. Frequentist penalized likelihood methods essentially optimize the classical least squares objective function subject to the constraint that the parameter lives in a possibly much lower dimensional subspace of $\Re^p$. One popular example of this idea is the LASSO [5] where the least squares objective function is minimized subject to a ℓ_1 penalty on the parameter β, i.e. the LASSO estimator of β is defined as,

$$\hat{\beta}_{\text{LASSO}} = \arg\min_{\beta} \left(\|y - X\beta\|_2^2 + \lambda \sum_{j=1}^{p} \beta_j \right). \tag{1.1}$$

In (1.1) λ is a tuning parameter which optimized over a set of plausible values and controls the sparsity in the estimate $\hat{\beta}_{\text{LASSO}}$. Of course, setting $\lambda = 0$ recovers the least squares estimator and $\lambda = \infty$ corresponds to the sparsest estimate $\hat{\beta}_{\text{LASSO}} = 0$. Several other penalized estimators

for β in the regression model have been proposed in the literature. See [6, 7] for the SCAD and MCP estimators. The book [8] is an excellent monograph for a compact overview of the works on penalized estimation in high-dimensional models.

From a Bayesian perspective, a natural way to model sparse parameters is to consider a prior which is a mixture of two densities, one of which is a continuous density for the non-zero entries and the other is a point mass at zero. The mixture proportion controls the degree of sparsity in the parameter. A typical point mass mixture prior for the j^{th} entry of a parameter $\beta \in \Re^p$ is displayed below:

$$\pi(\beta_j) = \omega\, g(\beta_j) + (1-\omega)\delta_0(\beta_j). \tag{1.2}$$

Here $g(\cdot)$ is a continuous density over the real line and is popularly known as the slab part of the prior π. $\delta_0(\cdot)$ is point mass at 0, known as the spike part and ω controls the proportion of non-zeros in the parameter β. Originally developed by [9] and further improved by the seminal work of [10], the spike slab priors have been remarkably successful in statistical models involving sparsity. For example, [11] used them in the context of high-dimensional factor models, [12] addressed the high-dimensional multiple response regression problem. [13] analyzed prior (1.2) in the context of Gaussian sequence model. The authors considered an empirical Bayes estimate of ω and showed the posterior median is an optimal estimator in minimax sense when the continuous density g is sufficiently heavy-tailed. When consider a fully Bayesian analysis [14] showed how the rather seemingly innocuous Uniform prior on ω may lead to suboptimal inference. They suggested using a Beta-Binomial prior on ω wherein the parameter are chosen so as to assign exponentially decaying mass on increasing model size for optimal inference. In a more recent work [15] showed that the Beta-Binomial prior on ω in conjunction with a Laplace density for $g(\cdot)$ lead to posteriors which contract at the optimal rate in a frequentist sense when the truth belongs certain sparsity classes. While the anlysis in [15] was carried out for the Gaussian sequence model, the authors extended their results in [16] with the same choice of prior in a linear regression framework. [16] also that the posterior distribution using prior (1.2) in linear regression leads to posteriors which are low-dimensional Gaussian distribution in an asymptotic sense.

Although spike and slab priors are known to enjoy theoretical optimality for quite some time now, their use in real world applications have been limited due to the computational challenges involved. When $\beta \in \Re^p$, posterior MCMC sampling involves model search over an exponentially growing model size; for with p variables there are 2^p possible models to search over. Moreover, the need to properly integrate estimates from different model sizes have garnered a separate line of research known as Bayesian model averaging. See [17] for more details and the intricacies involved in the problem. As a result, their impact on the broader scientific community has been rather limited, whereas frequentist methods enjoy an overwhelming popularity in day to day practice of high-dimensional statistics.

[18], in his seminal work introduced the idea of shrinkage. Consider a sample $y \sim \mathrm{N}(\theta, \mathrm{I})$, $\theta \in \Re^p$. The maximum likelihood estimator of θ in this case is y. However, the following estimator known as Stein's estimator,

$$\hat{\theta}_{\text{Stein}} = y\left(1 - \frac{p-2}{\|\theta\|_2^2}\right) \tag{1.3}$$

Stein showed that $\hat{\theta}_{\text{mle}}$ dominates $\hat{\theta}_{\text{Stein}}$ under the squared error loss: $\left\|\hat{\theta} - \theta\right\|_2^2$. [19] established an empirical Bayes interpretation of Stein's estimator. Full Bayes estimators such as the posterior mean automatically provides some shrinkage. For example, in a Gaussian model with Gaussian prior on the mean, the posterior mean is a convex combination of the sample mean and prior mean. Anchoring on the idea that Bayes estimators provide automatic shrinkage, a new class of priors started to emerge in the last two decades. Popularly known as shrinkage priors, these priors aim to aggressively shrink the noise components of a sparse parameter towards zero while retaining the signal components as much as possible. This is achieved by having an individual scale parameter for each component called the local scale parameter and a global scale parameter is introduced which shrinks the entire parameter towards the origin [20]. A further advantage of these priors is due to their conditionally Gaussian representation. Suppose $\theta \in \Re^p$, a shrinkage prior is defined

hierarchically as follows,

$$\begin{aligned} \theta_j \mid \lambda_j, \tau &\overset{indp.}{\sim} \mathrm{N}(0, \lambda_j^2 \tau^2),\ j = 1, \ldots, p \\ \lambda_j &\overset{indp.}{\sim} f, \\ \tau &\sim h. \end{aligned} \tag{1.4}$$

In the above display λ_j correspond to the local scale parameter for the j^{th} component of θ and τ is the global shrinkage parameter. f and h are densities on the positive real line. Different choices of the mixing densities f and h lead to different marginal densities of θ. The choice of f and h are obviously very crucial. Recipes for choosing f and h are provided in [20] where the authors suggest an exponentially light-tailed distribution for h allowing for sparse estimates and f should be sufficiently heavy-tailed to capture the signal/non-zero components.

Several authors have suggested different choices of f and h. [21, 1, 22, 23] are members of a long list of shrinkage priors. In this thesis we will mainly be focusing on methodological and theoretical aspects of the horseshoe prior from [1] which is obtained when f and h are both set to the standard Half-Cauchy distribution on the positive real line. The density of the standard Half-Cauchy distribution is $f(x) \propto (1 + x^2)^{-1} \mathrm{I}_{(0,\infty)}$. Marginal distributions for the Dirichlet-Laplace prior [23] and the horseshoe prior [1] are provided in figure 1.1 and figure 1.2. Both these priors exhibit a sharp spike in around the origin which mimics the point mass of spike and slab priors and helps shrink the noise variables aggressively. The tail of the horseshoe prior decays very slowly similar to the Cauchy distribution, while Dirichlet-Laplace tail is slightly lighter. These operating characteristics effectively provides a continuous analogue of spike slab priors.

The continuous nature of shrinkage priors comes as a boon in posterior computation. Since, due to their conjugate Gaussian formulation as shown in (1.4), designing MCMC chains for posterior sampling are relatively straightforward - closed form full conditional distributions allows for Gibbs sampling and the parameter can be updated in a block thus resulting in better mixing of the MCMC

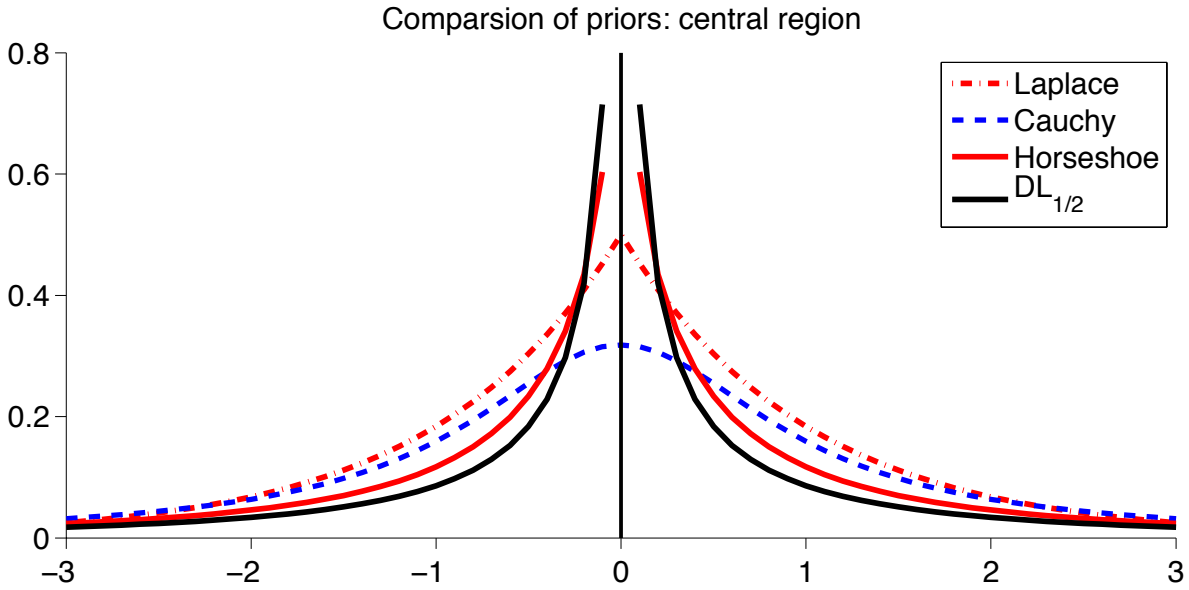


Figure 1.1: Marginal distribution for the Dirichlet-Laplace prior with hyperparameter $1/2$ and the horseshoe prior are plotted around a neighborhood of zero. The Laplace and Cauchy densities are also plotted for reference.

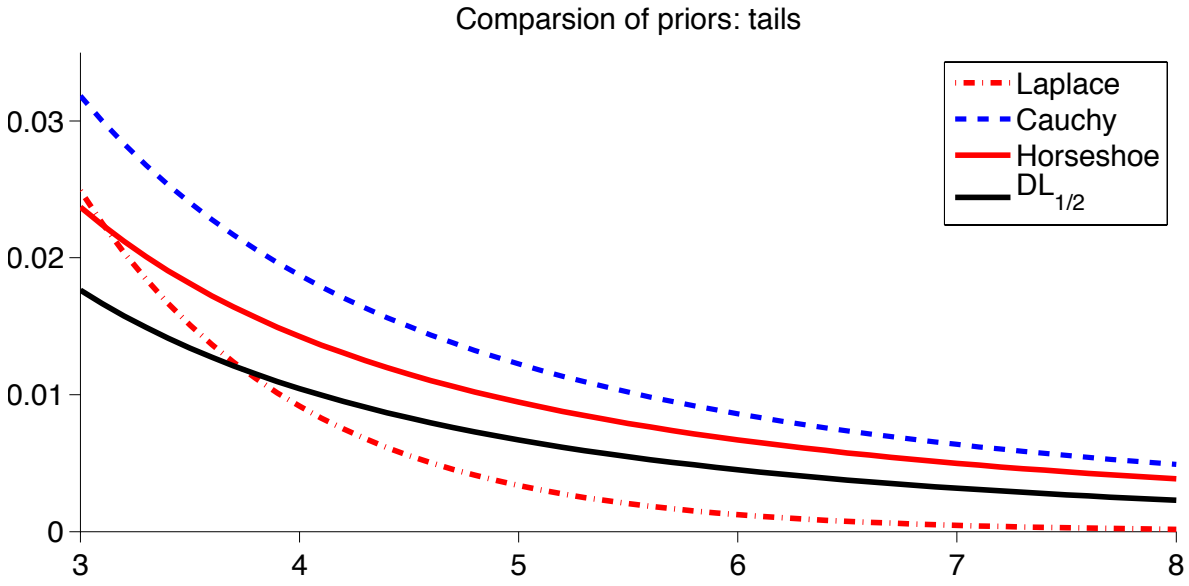


Figure 1.2: Same densities as in figure 1.1. Here the focus is on the tail behavior of the priors.

chain. For example, consider the canonical linear regression model,

$$y = X\beta + \epsilon,\ \epsilon \sim \mathrm{N}(0, 1). \tag{1.5}$$

When β is believed to be sparse, one possible choice for a prior distribution on β is the horseshoe prior. The full conditional posterior distribution of β then is $\mathrm{N}(A^{-1}X^{\mathrm{T}}y, A^{-1})$, where $A = X^{\mathrm{T}}X + \Lambda^{-1}$, where Λ is a diagonal matrix with $\lambda_j^2\tau^2$ as its j^{th} entry.

Posterior summaries from these MCMC chains such as posterior mean or median can then be used as point estimates for β. The fully Bayesian treatment of the problem also allows for quantifying the uncertainty by means of credible intervals which is not the case for Bayesian spike slab priors and frequentist penalized methods. Seemingly, shrinkage priors provide a unified foundation for high-dimensional analysis with its appealing computational tractability and the ability to simultaneously provide point estimates and uncertainty intervals.

1.2 Research questions and motivation

In the previous discussion we hope to have provided a convincing argument for the use of shrinkage priors. However, one major computational bottleneck has restricted their usage in data sets of real scientific interest. For example, in genome wide association studies it is common to record activities of thousands of genes related to a particular response. In such scenarios, implementing a linear regression model with a shrinkage prior would require sampling from a very high-dimensional Gaussian distribution at each and every MCMC iteration.Furthermore, this high-dimensional Gaussian distribution involves a matrix inversion. Standard MCMC algorithms for this problem have complexity scaling cubic in the number of parameters which can be very time consuming even when there is only a few hundred parameters. The first question that we will be asking ourselves is,

- can we develop a faster algorithm?

The idea of shrinkage is not only limited to regression problems. When multiple responses are observed for a set of subjects or individuals along with possibly thousands of predictors, proper

shrinkage is warranted. However, priors for multiple response problems based on shrinkage are relatively underdeveloped due to their computational challenges. Provided our answer to the previous question is a yes, in the next chapter we ask,

- can we develop a computationally tractable, sound methodology for handling multiple response at the same using the idea of shrinkage?

If again the answer is a yes, we ask,

- what theoretical guarantees can we provide for these methods in a high-dimensional setting?

1.3 Outilne

Here we provide a brief snapshot of the rest of our work motivated from the research questions discussed above.

In the first chapter we consider the model (1.5) with a shrinkage prior on β. Suppose n is the sample size and p is the number of predictors. Existing algorithms [24] which rely on computing Cholesky factors have $O(p^3)$ complexity. As a solution to this problem we develop an exact sampling algorithm for Gaussian distributions which relies on data augmentation. Computational complexity of the proposed algorithm is $O(n^2p)$ providing huge gains in time when $p \gg n$. To provide a perspective, for $(n, p) = (100, 5000)$ and 6000 MCMC iterations our algorithm offers a speed-up factor of 250 times over the algorithm in [24]. A MATLAB script implementing the algorithm with the horseshoe prior [1] is publicly available here. An `R` package is also available at CRAN.

The next chapter focuses on the multiple response regression problem. Often several variables of interest, possibly correlated, are measured on different individuals/subjects along with a set of predictors and it is known that statistical procedures ignoring the correlation may lead to incorrect inferential decisions. A simple yet powerful model for studying multiple responses and a set of predictors is the multi-response linear regression model: $Y = XC + E$, $Y \in \Re^{n\times q}$, $C \in \Re^{p\times q}$. In the presence of many predictors it is then natural to ask in this setting which variables are important and whether there is a latent dependence among the important ones. Here we simultaneously

address the problems of rank 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 prior is placed on a full-rank decomposition of C bypassing need to specify a prior on the unknown rank. Since shrinkage priors are unable to select variables, 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. When compared with existing frequentist methods our proposed methodology showed a decoupling effect for rank estimation and variable selection, whereas for the frequentist methods overfitting of the rank is necessary for optimal variable selection. The methodology is available for MATLAB implementation here.

Finally, we investigate the theoretical properties of the posterior obtained from a horseshoe prior in several high-dimensional models. Such studies of the horseshoe prior quantifying rates of convergence [25, 26] focus *exclusively* on the normal means problem with their proofs crucially exploiting an exact conjugate representation of the posterior mean. However, other than [27] who studied the case $p < n$, there are no posterior consistency results for the horseshoe prior in regression or general statistical models of practical ineterest where such representation is not possible. Furthermore, [27] did not quantify the rate of convergence. To aid theoretical analysis, we adopt the fractional posterior framework developed in [28] where only a prior mass condition is sufficient to guarantee consistency of the fractional posterior. In this work, our key contribution is in providing a novel non-asymptotic prior concentration result for the prior in around minimax neighborhoods of low-rank and row sparse matrices which drives the concentration of the posterior. The result is established allowing p to grow sub-exponentially in n , a first of its kind in this setting. To exhibit our result's full generality we extend our results to high-dimensional single response regression models and factor models.

2. FAST SAMPLING OF GAUSSIAN SCALE MIXTURE PRIORS IN HIGH-DIMENSIONAL REGRESSION

2.1 Introduction

Continuous shrinkage priors have received significant attention in recent years as a mechanism to induce approximate sparsity in high-dimensional parameters. Such priors can be almost exclusively expressed as global-local scale mixtures of Gaussians [20]; examples include the relevance vector machine [29], normal/Jeffrey's prior [30], the Bayesian lasso [31, 32], the horseshoe [1], normal/inverse-Gaussian priors [22], generalized double Pareto priors [33] and Dirichlet–Laplace priors [34]. These *global-local* priors [20] aim to aggressively shrink out the noise coefficients while retaining the signals, thereby providing an approximation to the operating characteristic of the more traditional discrete mixture priors [9, 35, 14], which allow a subset of the parameters to be exactly zero.

[1] A major attraction of the global-local priors has been computational efficiency and simplicity. Posterior inference poses a stiff challenge for discrete mixture priors in exploring very large model spaces in moderate to high-dimensional settings. On the contrary, the scale-mixture representation of global-local priors allows parameters to be updated in blocks via a fairly automatic Gibbs sampler in a wide variety of statistical problems; examples include regression [32, 33], wavelet denoising [20], factor models and covariance estimation [36, 37], dependent time series [38] name a few. Moreover, recent results suggest that a subclass of global-local priors can achieve the same level of statistical accuracy as the discrete mixture priors in high-dimensional estimation problems. For the normal means problem, [34] and [25] established optimal posterior concentration of the Dirichlet–Laplace and horseshoe priors respectively. In the context of large covariance estimation via factor models, [37] showed that both appropriately chosen discrete mixture priors and global-local priors lead to a minimax rate of posterior concentration. While a general theory for

[1] *Reprinted with permission from "Fast sampling with Gaussian scale-mixture priors in high-dimensional regression" by Anirban Bhattacharya, Antik Chakraborty and Bani K. Mallick, 2016, Biometrika, 103(4), 985 - 991, Copyright [2016] by Biometrika Trust.

global-local priors is yet lacking, more results are likely to appear in the near future.

In this article, we focus on computational aspects of global-local priors in the high-dimensional linear regression setting

$$y = X\beta + \epsilon, \quad \epsilon \sim \mathrm{N}(0, \sigma^2 \mathrm{I_n}), \tag{2.1}$$

where $X \in \Re^{n \times p}$ is a $n \times p$ matrix of covariates with the number of variables p potentially much larger than the sample size n. In such $p \gg n$ settings, one expects the vector of regression coefficients $\beta \in \Re^p$ to be sparse. A global-local prior on β assumes the following hierarchical structure:

$$\beta_j \mid \lambda_j, \tau, \sigma \sim \mathrm{N}(0, \lambda_j^2 \tau^2 \sigma^2), \quad (j = 1, \ldots, p), \tag{2.2}$$

$$\lambda_j \sim f, \quad (j = 1, \ldots, p) \tag{2.3}$$

$$\tau \sim g, \quad \sigma \sim h, \tag{2.4}$$

where f, g and h are densities supported on $(0, \infty)$. In the above display, the λ_js are usually referred to as local scale parameters while τ is a global scale parameter. Different choices of f and g lead to different classes of priors. For instance, a half-Cauchy distribution for f and g leads to the horseshoe prior of [1]. General guidelines for choices of the densities f and g can be found in [20] and [37].

Exploiting the scale-mixture representation (2.2), it is straightforward to write down a Gibbs sampler for the above class of priors. In particular, the conditional posterior of β given y and $\lambda = (\lambda_1, \ldots, \lambda_p)^{\mathrm{T}}, \tau$ and σ is Gaussian:

$$\beta \mid y, \lambda, \tau, \sigma \sim \mathrm{N}(A^{-1} X^{\mathrm{T}} y, \sigma^2 A^{-1}), \quad A = (X^{\mathrm{T}} X + \Lambda_*^{-1}), \quad \Lambda_* = \tau^2 \mathrm{diag}(\lambda_1^2, ..., \lambda_p^2). \tag{2.5}$$

Further, the p local scale parameters λ_j have conditionally independent posteriors and hence $\lambda = (\lambda_1, \ldots, \lambda_p)^{\mathrm{T}}$ can be updated in a block. Even in cases where λ_js and τ do not admit conditionally

conjugate posteriors, a slice sampling algorithm [39] can be adapted to update these parameters efficiently.

A standard algorithm to sample from Gaussian distributions as in (2.5) can be found in [24], which avoids inverting A and instead performs a Cholesky decomposition of A and a series of linear system solutions to generate samples. While this is efficient for moderate values of p, obtaining a Cholesky decomposition of A at each MCMC step becomes highly expensive for large p. One cannot resort to precomputing the Cholesky factors since the matrix Λ^* in (2.5) changes from one iteration to the other. The resulting computational bottleneck obscures the computational advantages of global-local priors when p is large. In this article, we present an exact sampling algorithm for Gaussian distributions as in (2.5) which relies on data augmentation and block matrix manipulations. We show that the computational complexity of the proposed algorithm scales linearly in the dimension p. Computational gains with increasing dimensionality is illustrated through a simulation example on linear regression with the horseshoe prior [1].

2.2 The algorithm

We first state our algorithm in a general setting. Suppose the goal is to sample from $\mathrm{N}_p(\mu, \Sigma)$, with

$$\Sigma = (\Phi^{\mathrm{T}}\Phi + D^{-1})^{-1}, \quad \mu = \Sigma\Phi^{\mathrm{T}}\alpha, \tag{2.6}$$

where $\Phi \in \Re^{n\times p}$, $D \in \Re^{p\times p}$ is symmetric positive definite and $\alpha \in \Re^{n\times 1}$. It is straightforward to observe that (2.5) is a special case of (2.6) with $\Phi = X/\sigma$, $D = \sigma^2\Lambda_*$ and $\alpha = y/\sigma$. Further, the need to sample from a distribution as in (2.6) arises in almost all the applications of continuous shrinkage priors mentioned in the introduction and the proposed approach can be seamlessly integrated into such settings. In the sequel, we do not require D to be diagonal, however we implicitly assume that D^{-1} is easy to calculate and it is straightforward to sample from $\mathrm{N}(0, D)$. This is the case, for example, if D corresponds to the covariance matrix of an AR(q) process or a Gaussian Markov random field.

Letting $Q = \Sigma^{-1} = (\Phi^{\mathrm{T}}\Phi + D^{-1})$ denote the precision (or inverse covariance) matrix and $b = \Phi^{\mathrm{T}}\alpha$, we can write $\mu = Q^{-1}b$. [24] proposed an efficient algorithm to sample from a Gaussian distribution with precision matrix Q and mean $Q^{-1}b$ that avoids explicitly calculating the inverse of Q which is computationally expensive and numerically unstable. Instead, the algorithm in Section 3.1.2. of [24] performs a Cholesky decomposition of Q and uses the Cholesky factor to solve a series of linear systems to arrive at a sample from the desired Gaussian distribution. The algorithm adapted to the present setting can be expressed as follows:

Algorithm 1. **[24]**

(i) Compute the Cholesky decomposition $(\Phi^{\mathrm{T}}\Phi + D^{-1}) = LL^{\mathrm{T}}$.
(ii) Solve $Lv = \Phi^{\mathrm{T}}\alpha$.
(iii) Solve $L^{\mathrm{T}}m = v$.
(iv) Solve $L^{\mathrm{T}}w = z$, where $z \sim \mathrm{N}(0, \mathrm{I_p})$.
(vi) Set $\theta = m + w$. Then, $\theta \sim \mathrm{N}(\mu, \Sigma)$.

The algorithm in [24] was originally developed to efficiently sample from Gaussian Markov random fields where the precision matrix Q has a banded structure and a Cholesky factor can be computed efficiently. Even though the precision matrix $(\Phi^{\mathrm{T}}\Phi + D^{-1})$ does not have any special structure in the present setting, Algorithm 1 remains practically useful as long as p is around 500; however, for larger values of p, there is an inevitable breakdown point. In fact, the complexity of the algorithm scales as $O(p^3)$ since the Cholesky decomposition in step (i) requires $O(p^3)$ floating point operations and the subsequent linear system solutions can be performed in $O(p^2)$ floating point operations [40].

The $O(p^3)$ complexity is prohibiting in high dimensional settings where p is in the order of tens of thousands or even higher. We present an alternative exact mechanism to sample from a normal distribution with parameters as in (2.6) below:

Proposition 2.2.1. *Suppose θ is obtained by following steps (i) - (iv) of Algorithm 2. Then, $\theta \sim \mathrm{N}(\mu, \Sigma)$, where μ and Σ are as in* (2.6).

Algorithm 2. **Proposed algorithm**

(i) Sample $u \sim \mathrm{N}(0, D)$ and $\delta \sim \mathrm{N}(0, \mathrm{I_n})$.
(ii) Set $v = \Phi u + \delta$.
(iii) Solve $(\Phi D \Phi^{\mathrm{T}} + \mathrm{I_n})w = (\alpha - v)$.
(iv) Set $\theta = u + \Phi^{\mathrm{T}} D w$.

A proof of Proposition 2.2.1 is provided in Section 2.2.1. While the sampling mechanism is valid for all n and p, the computational gains are most prominent when $p \gg n$ and the assumptions regarding D made at the beginning of the Section are satisfied. Indeed, the primary motivation behind the algorithm derivation is to scale down the problem from p to n dimensions using the Sherman–Woodbury–Morrison matrix identity [41] and then using Gauss–Jordan type factorizations and linear system solvers to scale back to p dimensions. When D is diagonal as case of continuous shrinkage priors (2.2), the complexity of the proposed algorithm can be accurately calculated.

Proposition 2.2.2. *Assume D is diagonal and $p \geq n$. Then, steps (i) - (iv) in Algorithm 2 can be carried out via $O(n^2 p)$ floating point operations.*

Proof. We use a couple of facts from [40]: (i) if $B_1 \in \Re^{m_1 \times m_2}$ and $B_2 \in \Re^{m_2 \times m_3}$, then the product $B_1 B_2$ can be computed in $O(m_1 m_2 m_3)$ floating point operations; (ii) if $B \in \Re^{m \times m}$, then the linear system $Bx = y$ can be solved in $O(m^3)$ operations. Since D is diagonal, sampling u in step (i) can be carried out in $O(p)$ floating point operations and the matrix multiplication Φu in step (ii) takes additional $O(np)$ operations. Once again using that D is diagonal, the matrix $\widetilde{\Phi} = \Phi D$ can be computed in $O(np)$ floating point operations and hence the complexity of calculating $\Phi D \Phi^{\mathrm{T}} = \widetilde{\Phi} \Phi^{\mathrm{T}}$ is $O(n^2 p)$. The $n \times n$ linear system in step (iii) can be computed in $O(n^3)$ operations. Finally, since $\widetilde{\Phi}$ is already calculated in step (iii), the matrix multiplication $\Phi^{\mathrm{T}} D w = (\widetilde{\Phi})^{\mathrm{T}} w$ takes $O(n^2 p)$ operations. Since $p \geq n$, $n^2 p \geq n^3$, and the overall complexity is $O(n^2 p)$. □

In comparison to the $O(p^3)$ complexity of Algorithm 1, the proposed Algorithm 2 therefore offers substantial gains in $p \gg n$ settings with a reduction in complexity from cubic to linear in p.

Our numerical results in Section 2.4 indeed show that when $p \asymp n$, both algorithms have similar computation time while Algorithm 2 is significantly faster when p is large.

When D is not diagonal, the complexity of the proposed algorithm needs to be calculated on a case by case basis depending on the structure of D. Without any assumptions on D, calculating $\tilde{\Phi}$ above would take $O(np^2)$ operations which is the dominating term in the complexity calculations. Even in this case, the complexity reduces from cubic to quadratic in p which can be substantial when p is large.

2.2.1 Algorithm Derivation

We now prove Proposition 2.2.1 by deriving Algorithm 1 in a constructive fashion. We begin with an identity for Σ derived from the Sherman–Woodbury–Morrison formula [41].

$$\Sigma = (\Phi^{\mathrm{T}}\Phi + D^{-1})^{-1} = D - D\Phi^{\mathrm{T}}(\Phi D\Phi^{\mathrm{T}} + \mathrm{I_n})^{-1}\Phi D. \tag{2.7}$$

Using (2.7), we first obtain a more amenable expression for μ as

$$\mu = D\Phi^{\mathrm{T}}(\Phi D\Phi^{\mathrm{T}} + \mathrm{I_n})^{-1}\alpha. \tag{2.8}$$

To see this, recall from (2.6) that $\mu = \Sigma\Phi^{\mathrm{T}}\alpha$. Using (2.7), $\Sigma\Phi^{\mathrm{T}} = D\Phi^{\mathrm{T}} - D\Phi^{\mathrm{T}}(\Phi D\Phi^{\mathrm{T}} + \mathrm{I_n})^{-1}\Phi D\Phi^{\mathrm{T}} = D\Phi^{\mathrm{T}} - D\Phi^{\mathrm{T}}(\Phi D\Phi^{\mathrm{T}} + \mathrm{I_n})^{-1}(\Phi D\Phi^{\mathrm{T}} + \mathrm{I_n} - \mathrm{I_n}) = D\Phi^{\mathrm{T}}(\Phi D\Phi^{\mathrm{T}} + \mathrm{I_n})^{-1}$.

We now provide a data augmentation scheme to sample from $\mathrm{N}(0, \Sigma)$. Clearly, if $\eta \sim \mathrm{N}(0, \Sigma)$, then $\theta = \mu + \eta$ gives us a sample from $\mathrm{N}(\mu, \Sigma)$. We first record a matrix identity below that can be easily verified:

$$\begin{pmatrix} P & S \\ S^{\mathrm{T}} & R \end{pmatrix} = \begin{pmatrix} \mathrm{I_n} & 0 \\ S^{\mathrm{T}}P^{-1} & \mathrm{I_p} \end{pmatrix} \begin{pmatrix} P & 0 \\ 0 & R - S^{\mathrm{T}}P^{-1}S \end{pmatrix} \begin{pmatrix} \mathrm{I_n} & P^{-1}S \\ 0 & \mathrm{I_p} \end{pmatrix}, \tag{2.9}$$

where $P \in \Re^{n\times n}$ and $R \in \Re^{p\times p}$ are invertible and $S \in \Re^{n\times p}$. Letting

$$\Omega = \begin{pmatrix} P & S \\ S^{\mathrm{T}} & R \end{pmatrix}, \quad L = \begin{pmatrix} \mathrm{I_n} & 0 \\ S^{\mathrm{T}}P^{-1} & \mathrm{I_p} \end{pmatrix}, \quad \Gamma = \begin{pmatrix} P & 0 \\ 0 & R - S^{\mathrm{T}}P^{-1}S \end{pmatrix},$$

Ω, L, Γ are $(n+p)\times(n+p)$ matrices and (2.9) can be conveniently expressed as $\Omega = L\Gamma L^{\mathrm{T}}$. Further, L is invertible, with $L^{-1} = \begin{pmatrix} \mathrm{I_n} & 0 \\ -S^{\mathrm{T}}P^{-1} & \mathrm{I_p} \end{pmatrix}$ easily derived since L is lower triangular. Therefore, $\Gamma = L^{-1}\Omega(L^{-1})^{\mathrm{T}}$.

Now, set $P = (\Phi D\Phi^{\mathrm{T}} + \mathrm{I_n}), S = \Phi D$ and $R = D$. Let $u \sim \mathrm{N}(0, D), \delta \sim \mathrm{N}(0, \mathrm{I_n})$ and $v = \Phi u + \delta$ as in steps (i) - (ii) of Algorithm 1 and set $\zeta = (v^{\mathrm{T}}, u^{\mathrm{T}})^{\mathrm{T}} \in \Re^{n+p}$. By construction, $\mathrm{cov}(v, u) = \Phi D = S$ and $\mathrm{var}(v) = (\Phi D\Phi^{\mathrm{T}} + \mathrm{I_n}) = P$, which implies the covariance matrix of ζ is Ω. Moreover, since u and δ are *independent*, ζ has a joint Gaussian distribution. Since both u and v are zero mean, we conclude that $\zeta \sim \mathrm{N}(0, \Omega)$.

Next, using the identity $\Gamma = L^{-1}\Omega(L^{-1})^{\mathrm{T}}$, we have $\zeta_* = L^{-1}\zeta \sim \mathrm{N}(0, \Gamma)$. The crucial observation is that with the present choices of P, S and R, the lower $p \times p$ diagonal block of Γ, $R - S^{\mathrm{T}}P^{-1}S = \Sigma$, using the identity (2.7). Therefore, if we collect the last p entries of ζ_* in a vector η, then $\eta \sim \mathrm{N}(0, \Sigma)$. Exploiting the structure of L^{-1}, we have $\eta = u - S^{\mathrm{T}}P^{-1}v$. Note further from (2.8) that $\mu = S^{\mathrm{T}}P^{-1}\alpha$. Finally,

$$\theta = \mu + \eta = S^{\mathrm{T}}P^{-1}\alpha + u - S^{\mathrm{T}}P^{-1}v = u + S^{\mathrm{T}}P^{-1}(\alpha - v).$$

The proof is completed by noting that $P^{-1}(\alpha - v)$ is identical to w in step (iii) of Algorithm 1 and $S^{\mathrm{T}} = D\Phi^{\mathrm{T}}$.

2.3 Application to Bayesian shrinkage priors

Returning to our original motivation, we illustrate the scalability of the proposed algorithm through an application to linear regression with a horseshoe prior [1] on the regression coefficients β in (5.15). The horseshoe prior is obtained by placing independent half-Cauchy priors on the λ_js

and τ in the hierarchical specification (2.2) - (2.4), with the half-Cauchy density given by $(1 + t^2)^{-1}\mathbf{1}_{(0,\infty)}(t)$. To complete the prior specification, we consider an improper prior $\pi(\sigma^2) \propto \sigma^{-2}$ on σ. Posterior computation proceeds in a straightforward manner via a Gibbs sampler which cycles through sampling from (i) $\beta \mid y, \lambda, \tau, \sigma$, (ii) $\lambda \mid \beta, \tau, \sigma$, (iii) $\tau \mid \beta, \lambda, \sigma$ and $\sigma \mid y, \beta, \lambda, \sigma$, where recall that $\lambda = (\lambda_1, \ldots, \lambda_p)^{\mathrm{T}}$. We use a slice sampling algorithm from the online supplement to [39] to update the λ_js and τ, while σ^2 has an inverse-gamma conditional posterior. The conditional posterior of β is Gaussian as given in (2.5) and is the main computational bottleneck when p is large. As noted in the beginning of Section 2.2, letting $\Phi = X/\sigma$, $D = \sigma^2\Lambda_*$ and $\alpha = y/\sigma$, we are in the setting of (2.6) and the algorithms in Section 2.2 can be applied to sample from the conditional posterior of β.

We borrow a simulation setting from §3.3.1 of [42] to generate the data. Given n and p, the p columns of the design matrix X were generated independently from a standard n dimensional Gaussian distribution. All but the first 5 entries of β were set to zero, while the non-zero entries were sampled independently as $(-1)^r(a + |z|)$, with $r \sim$ Bernoulli(0.4), $z \sim \mathrm{N}(0, 1)$ and $a = 5 \log n/\sqrt{n}$. Finally, y was generated from a $\mathrm{N}(X\beta, \sigma^2\mathrm{I}_n)$ distribution with $\sigma = 2$. We fixed $n = 100$ and varied p from 500 to 5000 with a step size of 500. We additionally included the case where $(n, p) = (100, 200)$. For each value of p, 10 datasets were generated as above.

The Gibbs sampler mentioned in the previous paragraph was run for 6000 iterations where we alternatively used the standard Algorithm 1 and the proposed Algorithm 2 to sample from the conditional posterior of β in step (i). Steps (ii) - (iv) of the Gibbs sampler were identically implemented for both methods. All computations were implemented in MATLAB on a INTEL(E5-2690) 2.9 GHz machine with 64 GB DDR3 memory. Figure 2.1 plots the time in seconds averaged over the 10 datasets to run 6000 iterations of the Gibbs sampler for the two methods in a *logarithmic scale* against p. The actual times in seconds for selected values of p are presented in Table 2.1. From Fig. 2.1 and Table 2.1, it is evident that both algorithms have similar computation time when $p \asymp n$, while the proposed algorithm produces massive gains for larger values of p. Table 2.1 also shows that the absolute run time for 6000 iterations of the Gibbs sampler in our case takes only

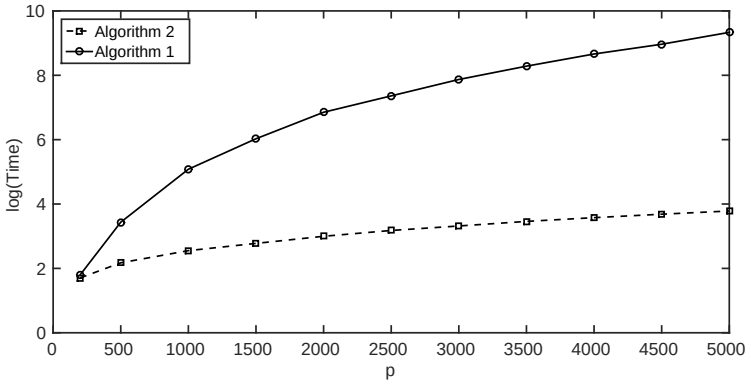


Figure 2.1: Linear regression with horseshoe prior [1] on the regression coefficients. Logarithm of time (in seconds) to complete 6000 iterations of a Gibbs sampler is reported. Algorithm 1 and proposed Algorithm 2 were applied to sample from the conditional posterior of β in (2.5). Sample size n was fixed at 100 and the dimension p was varied from 500 to 5000, with 500 step size. Additionally the case $p = 200$ was included. Reprinted with permission from [65].

40 seconds. We have tried p as large as $20,000$ in our case which takes less than 5 minutes to run 6000 iterations.

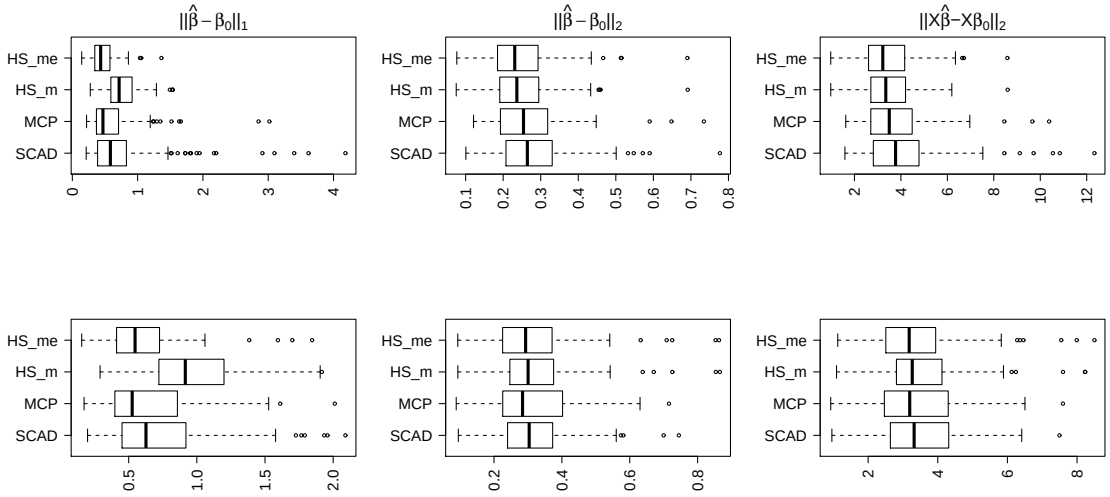
Table 2.1: Same setting as in Fig. 2.1. Absolute time (in seconds) to run 6000 iterations of the Gibbs sampler reported for the two algorithms for chosen values of p. Reprinted with permission from [65].

p	Time (in seconds)	
	Algorithm 1	Algorithm 2
200	5.50	6.05
500	8.79	31.03
1000	12.83	160.92
2000	20.04	944.78
3000	27.60	2616.80
4000	35.76	5775.70
5000	43.99	11314.28

2.4 Frequentist operating characteristics in high dimensions

The proposed algorithm provides an opportunity to compare the frequentist operating characteristics of shrinkage priors in high-dimensional regression problems, hitherto unexplored poten-

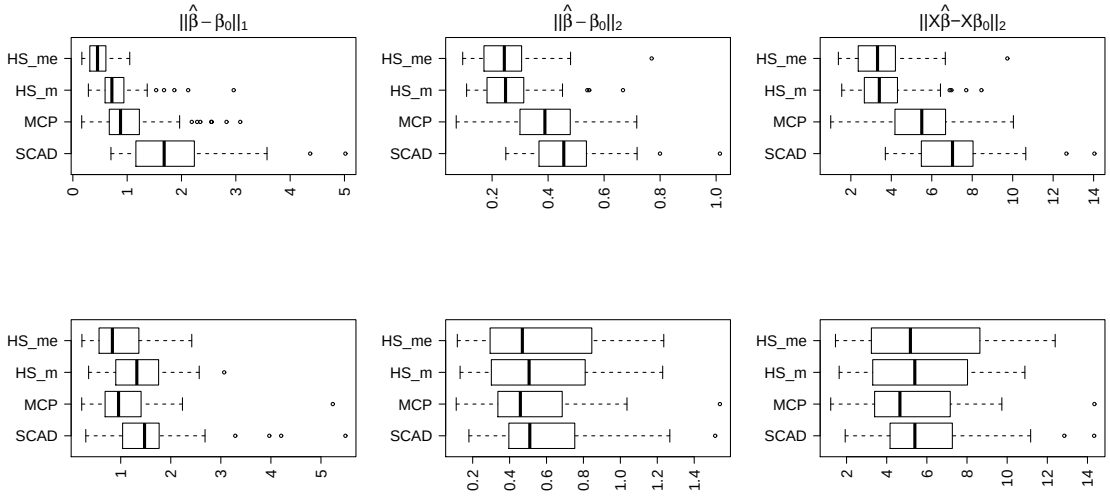
Figure 2.2: Boxplots of ℓ_1, ℓ_2 and prediction error across 100 simulation replicates. HS_{me} and HS_{m} respectively denote posterior point wise median and mean for the horeshoe prior. True β_0 is 5-sparse with non-zero entries from setting (a). Top row: $\Sigma = \text{I}_p$ (independent). Bottom row: $\Sigma_{jj} = 1, \Sigma_{jj'} = 0.5, j \neq j'$ (compound symmetry). Reprinted with permission from [65].



tially due to the computational bottleneck. We compare various aspects of the horseshoe prior [1] to a host of frequentist procedures and obtain highly promising results. While space constraints prevent us from a more detailed study, we expect similar results for the Dirichlet-Laplace [23], normal-gamma [22] and generalized double-Pareto [33] priors, which we hope to report elsewhere.

We first report comparisons with SCAD [6] and MCP [7] penalties in terms of estimation and prediction accuracy. We considered model (5.15) with $n = 200$, $p = 5000$ and $\sigma = 1.5$. Letting x_i denote the ith row of X, the x_is were independently generated from $\text{N}_p(0, \Sigma)$, with (i) $\Sigma = \text{I}_p$ (independent) and (ii) $\Sigma_{jj} = 1, \Sigma_{jj'} = 0.5, j \neq j' = 1, \ldots, p$ (compound symmetry). The true β_0 had 5 non-zero entries in all cases, with the non-zero entries having magnitude (a) $\{1.5, 1.75, 2, 2.25, 2.5\}$ and (b) $\{0.75, 1, 1.25, 1.5, 1.75\}$, multiplied by a random sign. For each of the four cases, we considered 100 simulation replicates. The frequentist penalization approaches were implemented using the `R` package `ncvreg` via 10-fold cross-validation. For the horseshoe prior, we considered both the posterior mean and the point wise posterior median as a

Figure 2.3: Same setting as in Fig 2.2. True β_0 is 5-sparse with non-zero entries from setting (b) Reprinted with permission from [65].



point estimate. Figures 1 and 2 report boxplots for ℓ_1 ($||\hat{\beta} - \beta_0||_1$), ℓ_2 ($||\hat{\beta} - \beta_0||_2$) and prediction ($||X\hat{\beta} - X\beta_0||_2$) errors across the 100 replicates for the two signal strengths. The horseshoe prior can be seen to be highly competitive across all simulation settings, in particular when the signal strength is weaker. An interesting observation is the somewhat superior performance of the point wise median compared to the mean even under an ℓ_2 loss; a similar fact has been observed about point mass mixture priors [15] in high dimensions. We repeated the entire simulation with $p = 2500$ and obtained similar conclusions. Overall, out of the 24 = 2 (choices of p) $\times$ 2 (covariate designs) $\times$ 2 (signal strength) $\times$ 3 (error criterion) settings, the horseshoe prior had the best average performance over the simulation replicates in 22 cases.

While there is now a huge literature on penalized point estimation, the question of uncertainty characterization in $p > n$ settings has started receiving attention only very recently [3]. Although Bayesian procedures provide an automatic characterization of uncertainty through the posterior distribution on parameters, the resulting credible intervals are not guaranteed to possess the correct frequentist coverage in nonparametric/high-dimensional problems [43]. This compelled us to

empirically investigate the frequentist coverage of shrinkage priors in $p > n$ settings; it is trivial to obtain element-wise credible intervals for the β_js from the posterior samples. We compared the horseshoe prior with the recently proposed approach of [3], which can be used to obtain asymptotically optimal element wise confidence intervals for the β_js. We considered a similar simulation scenario as before. We let $p \in \{500, 1000\}$, and considered a Toeplitz structure ($\Sigma_{jj'} = 0.9^{|j-j'|}$) for the covariate design [3] in addition to the independent and compound symmetry cases stated already. The first two rows of Table 2.2 report the average coverage probabilities (over 100 simulation replicates) and lengths of confidence intervals for the horseshoe and [3], averaged over the 5 signal variables. The last two rows report the same averaged over the $(p-5)$ noise variables. The standard deviations corresponding to the averages are provided in the subscripts.

It can be readily seen from Table 2.2 that the horseshoe had a superior performance. Several observations are worth mentioning. First, an attractive adaptive property of shrinkage priors emerge, where the length of the intervals automatically adapt between the signal and noise variables, maintaining the nominal coverage. The procedure of [3] on the other hand seems to obtain approximately equal sized intervals for the signals and noise variables. We should mention here that the default choice of the LASSO tuning parameter $\lambda \asymp \sqrt{\log p/n}$ suggested in [3] seemed to provide substantially poorer coverage for the signal variables at the cost of improved coverage for the noise. For each setting, we separately tuned the parameter (assuming the knowledge of the truth) to arrive at the coverage probabilities reported. The horseshoe (and other shrinkage priors) on the other hand are free of tuning parameters. The same procedure that was used for estimation automatically provides valid frequentist uncertainty characterization.

2.5 Discussion

The numerical results in the previous section warrant additional numerical and theoretical investigations into properties of shrinkage priors in high dimensions. The proposed algorithm can be used for essentially all the shrinkage priors used in the literature and should prove to be useful in an exhaustive comparison of existing priors. Further, as stated previously, the scope of the proposed algorithm extends well beyond the linear regression setting. For example, extensions to logistic and

Table 2.2: Frequentist coverage probabilities and lengths of 95% intervals for the LASSO and the horseshoe. The confidence intervals for the LASSO were constructed using the method in [3]. The intervals for the horseshoe are the usual symmetric posterior credible intervals. For both methods, the average coverage probabilities and lengths of the intervals are reported after averaging across all signal variables (row 1 and 2) and noise variables (row 3 and 4). Numbers in the subscript denote the standard errors corresponding to the average coverage probabilities. Reprinted with permission from [65].

Dimension	**500**						**1000**					
Design	**Independent**		**Comp Symm**		**Toeplitz**		**Independent**		**Comp Symm**		**Toeplitz**	
	LASSO	**HS**	**LASSO**	**HS**	**LASSO**	**HS**	**LASSO**	**HS**	**LASSO**	**HS**	**LASSO**	**HS**
Signal Coverage	$0.75_{0.12}$	$0.93_{0.01}$	$0.73_{0.04}$	$0.95_{0.00}$	$0.80_{0.07}$	$0.94_{0.04}$	$0.78_{0.12}$	$0.94_{0.02}$	$0.77_{0.02}$	$0.94_{0.01}$	$0.76_{0.03}$	$0.95_{0.01}$
Signal Length	0.46	0.42	0.71	0.85	0.79	0.86	0.41	0.39	0.76	0.82	0.96	1.05
Noise Coverage	$0.99_{0.008}$	$1_{0.000}$	$0.98_{0.01}$	$1_{0.000}$	$0.98_{0.01}$	$1_{0.000}$	$0.99_{0.01}$	$0.99_{0.000}$	$0.99_{0.01}$	$1_{0.00}$	$0.99_{0.01}$	$1_{0.00}$
Noise Length	0.43	0.02	0.69	0.04	0.78	0.05	0.42	0.006	0.76	0.007	0.98	0.003

probit regression are immediate using standard data augmentation tricks [44, 45]. The same is true for other generalized linear models and survival models. Multivariate regression problems where one has a matrix of regression coefficients can be handled by updating the vectorized version of the coefficient matrix in a block; note this is an example where even if the number of variables is not bigger than the sample size, the number of regression coefficients may be large if the dimension of the response if moderate. Factor models offer another attractive area of application of the proposed algorithm. Shrinkage priors have been used as a prior of factor loadings in [36]. While [36] update the $p > n$ rows of the factor loadings independently, exploiting the assumption of independence in the idiyosyncratic components, their algorithm does not extend to *approximate factor models*, where the idiyosyncratic errors are dependent. The proposed algorithm can be adapted to such situations by updating the vectorized loadings in a block. Finally, we envision applications in high dimensional additive models where each of a large number of functions is expanded in a basis, and the large collection of basis functions are updated in a block.

3. BAYESIAN SPARSE MULTIPLE REGRESSION FOR SIMULTANEOUS RANK REDUCTION AND VARIABLE SELECTION

3.1 Introduction

Studying the relationship between multiple response variables and a set of predictors has broad applications ranging from bioinformatics, econometrics, 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 [46] 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 [47, 48, 49]. Although many results exist about the asymptotic properties of reduced rank estimators [50], formal statistical determination of the rank remains difficult even with fixed number of covariates and 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 work on penalized estimation of low rank matrices, where either the singular values of the coefficient matrix [51, 52], or the rank itself [53] is penalized. Theoretical evaluations of these estimators focusing on adaptation to the oracle convergence rate when the true coefficient matrix is of low rank has been conducted [53]. It has also been noted [54] that the convergence rate can be improved when the true coefficient matrix has zero rows and variable selection is incorporated within the estimation procedure. Methods that simultaneously handle rank reduction and variable selection include [51, 54, 55]. To best of our knowledge, uncertainty characterization for the parameter estimates from these procedures is currently not available.

The first fully systematic Bayesian treatment of reduced rank regression was carried out in [56], 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

[56], 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. [11] placed a prior on the number of factors and proposed a computationally intensive reversible jump algorithm [57] for model fitting. As an alternative, [36] proposed to increasingly shrink the factors starting with a conservative upper bound and adaptively collapsing redundant columns inside their MCMC algorithm. Recent advancements in Bayesian matrix factorization have taken a similar approach; see for example, [58, 59, 60, 61].

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 [10, 14] which allow a subset of the regression coefficients to be exactly zero. These priors were also adapted to multiple response regression by several authors [12, 62, 63, 64]. 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 [20], leading to block updates of model parameters; see [65] 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 [66, 67, 68].

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 into the operating characteristics of the proposed prior in terms of a novel prior concentration result around rank-reduced and low-sparse matrices. The prior concentration result is utilized to prove minimax concentration rates of

the posterior under the fractional posterior framework of [28] in a ultrahigh-dimensional setting where the number of predictor variables can grow sub-exponentially 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 [65] to efficiently sample a high-dimensional matrix in a block leading to a low per-iteration MCMC 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 (BSML). We illustrate the superiority of BSML over its competitors through a detailed simulation study and the methodology is applied to a Yeast cell cycle data set. Code for implementation is available at `www.stat.tamu.edu/~antik`.

3.2 Bayesian sparse multitask learner

3.2.1 Model and Prior Specification

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

$$Y = XC + E, \quad E = (e_1^{\mathrm{T}}, \ldots, e_n^{\mathrm{T}})^{\mathrm{T}}, \tag{3.1}$$

where we follow standard practice to center the response and exclude the intercept term. The rows of the error matrix are independent, with $e_i \sim \mathrm{N}(0, \Sigma)$. Our main motivation is the high-dimensional case where $p \geq \max\{n, q\}$, 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 rank$(C) = r \leq \min(p, q)$, whence C admits a decomposition $C = B_* A_*^{\mathrm{T}}$ with $B_* \in \Re^{p \times r}$ and $A_* \in \Re^{q \times r}$. 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 [69] to consider a potentially full-rank decomposition $C = BA^{\mathrm{T}}$ with $B \in \Re^{p \times q}$ and $A \in \Re^{q \times q}$, 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 $q^* \leq q$ 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 [70] of orthogonal matrices can be used. However, our numerical results suggested 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 vec(A), whereas the algorithm of [70] 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 we operate in. We shall henceforth denote Π_A to denote the prior on A, i.e., $a_{hk} \sim \mathrm{N}(0, 1)$ 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 [1] on the columns of B, which can be represented hierarchically as

$$b_{jh} \mid \lambda_{jh}, \tau_h \sim \mathrm{N}(0, \lambda_{jh}^2 \tau_h^2), \quad \lambda_{jh} \sim \mathrm{Ca}_+(0, 1), \quad \tau_h \sim \mathrm{Ca}_+(0, 1), \tag{3.2}$$

independently for $j = 1, \ldots, p$ and $h = 1, \ldots, q$, where $\text{Ca}_+(0, 1)$ denotes the truncated standard half-Cauchy distribution with density proportional to $(1+t^2)^{-1}\mathbf{1}_{(0,\infty)}(t)$. We shall denote the prior on the matrix B induced by the hierarchy in (3.2) by Π_B.

We shall primarily restrict attention to settings where Σ is diagonal, $\Sigma = \text{diag}(\sigma_1^2, \ldots, \sigma_q^2)$, noting that extensions to non-diagonal 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 [36, 24]. The cost-per-iteration of the Gibbs sampler retains the same complexity as in the diagonal Σ case; see §3.3 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

$$Y = XBA^{\mathrm{T}} + E, \quad e_i \sim \mathrm{N}(0, \Sigma), \tag{3.3}$$

$$B \sim \Pi_B, \quad A \sim \Pi_A, \quad \Sigma \sim \Pi_\Sigma. \tag{3.4}$$

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

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

denote the likelihood for (C, Σ).

3.3 Posterior Computation

Exploiting the conditional conjugacy of the proposed prior, we develop a straightforward and efficient Gibbs sampler to update the model parameters in (3.3) from their full conditional distributions. We use vectorization to update parameters in blocks. Specifically, in what follows, we will make multiple usage of the following identity. For matrices Φ_1, Φ_2, Φ_3 with appropriate

dimensions, and vec(A) denoting column-wise vectorization, we have,

$$\text{vec}(\Phi_1\Phi_2\Phi_3) = (\Phi_3^{\mathrm{T}} \otimes \Phi_1)\text{vec}(\Phi_2) = (\Phi_3^{\mathrm{T}}\Phi_2^{\mathrm{T}} \otimes I_k)\text{vec}(\Phi_1), \tag{3.5}$$

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 (3.5) to vectorize $Y = XBA^{\mathrm{T}} + E$ to obtain,

$$y = (X \otimes A)\beta + e, \tag{3.6}$$

where $\beta = \text{vec}(B^{\mathrm{T}}) \in \Re^{pq\times 1}$, $y = \text{vec}(Y^{\mathrm{T}}) \in \Re^{nq\times 1}$, and $e = \text{vec}(E^{\mathrm{T}}) \sim \mathrm{N}_{nq}(0, \widetilde{\Sigma})$ with $\widetilde{\Sigma} = \text{diag}(\Sigma, \ldots, \Sigma)$. Multiplying both sides of (3.6) 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 $\widetilde{e} = \widetilde{\Sigma}^{-1/2}e \sim \mathrm{N}_{nq}(0, \mathrm{I}_{nq})$. Thus, the full conditional distribution $\beta \mid - \sim \mathrm{N}_{pq}(\Omega_B^{-1}\widetilde{X}^{\mathrm{T}}\widetilde{y}, \Omega_B^{-1})$, where $\Omega_B = (\widetilde{X}^{\mathrm{T}}\widetilde{X} + \Lambda^{-1})$ with $\Lambda = \text{diag}(\lambda_{11}^2\tau_1^2, \ldots, \lambda_{1q}^2\tau_q^2, \ldots, \lambda_{p1}^2\tau_1^2, \ldots, \lambda_{pq}^2\tau_q^2)$.

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. [65] 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 [65] as follows:

(i) Sample $u \sim \mathrm{N}(0, \Lambda)$ and $\delta \sim \mathrm{N}(0, \mathrm{I}_{nq})$ 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 [65] that β obtained from steps (i) - (iv) above produce 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 [40] 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 (3.5) to obtain,

$$y = (XB \otimes \mathrm{I}_q)a + e, \tag{3.7}$$

where e and y are the same as in step 1, and $a = \mathrm{vec}(A) \in \Re^{q^2 \times 1}$. The full conditional posterior distribution $a \mid - \sim \mathrm{N}(\Omega_A^{-1} X_* \widetilde{y}, \Omega_A^{-1})$, 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 [24]. 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 $z \sim \mathrm{N}(0, \mathrm{I}_{\mathrm{q}^2})$. Finally obtain a sample as $a = m + w$.

Step 3. To sample $\sigma_h^2 \mid -$, observe that $\sigma_h^2 \mid - \sim$ inverse-Gamma$(n/2, S_h/2)$ 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 inverse-Wishart(q, I_q) prior on Σ, the posterior update of Σ can be easily modified due to conjugacy; we sample $\Sigma \mid -$ from inverse-Wishart$\{n + q, (Y - XC)^{\mathrm{T}}(Y - XC) + \mathrm{I}_q\}$.

Step 4. The global and local scale parameters λ_{jh}'s and τ_h's have independent conditional posteriors across j and h, which can be sampled via a slice sampling scheme provided in the online supplement to [39]. 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 $u_{jh} \mid \eta_{jh} \sim \mathrm{Unif}(0, 1/(1 + \eta_{jh}))$ and then sampling $\eta_{jh} \mid u_{jh} \sim \mathrm{Exp}(2\tau_h^2/b_{jh}^2)\mathrm{I}\{\eta_{jh} < (1 - u_{jh})/u_{jh}\}$, a truncated exponential distribution.

The Gibbs sampler above when modified to accommodate non-diagonal 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 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 element-wise 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 non-sparse and full rank with probability one, and hence not directly amenable for variable selection and rank estimation. Motivated by our concentration result in Theorem 3.5.7 that the posterior mean $X\overline{C}$ 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.

3.3.1 Post processing for variable selection

We first focus on variable selection. We define a row-sparse estimate $\widehat{C}_R$ for C as the solution to the optimization problem

$$\widehat{C}_R = \underset{\Gamma \in \Re^{p \times q}}{\arg\min} \left\{ \|X\overline{C} - X\Gamma\|_F^2 + \sum_{j=1}^{p} \mu_j \|\Gamma^{(j)}\|_2 \right\}, \tag{3.8}$$

where $\Phi^{(j)}$ represents the j^{th} row of a matrix Φ, and the μ_js 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 [71]. For a derivation of the objective function in (3.8) from a utility function perspective as in [68], refer to section 3.7 and 3.8.

To solve (3.8), we set the sub-gradient of (3.8) with respect to $\Gamma^{(j)}$ to zero and replace $\|\Gamma^{(j)}\|$ by a data dependent quantity to obtain the soft thresholding estimate,

$$\widehat{C}_R^{(j)} = \frac{1}{X_j^{\mathrm{T}} X_j} \left(1 - \frac{\mu_j}{2\|X_j^{\mathrm{T}} R_j\|} \right)_+ X_j^{\mathrm{T}} R_j, \tag{3.9}$$

where for $x \in \Re$, $x_+ = \max(x, 0)$, and R_j is the residual matrix obtained after regressing $X\overline{C}$ on X leaving out the j^{th} predictor, $R_j = X\overline{C} - \sum_{k \neq j} X_k \widehat{C}_R^{(k)}$. For practical implementation, we use $\overline{C}$ as our initial estimate and make a single pass through each variable to update the initial estimate

according to (3.9). With this initial choice, $R_j = X_j \overline{C}^{(j)}$ and $\|X_j^{\mathrm{T}} R_j\| = \|X_j\|^2 \|\overline{C}_j\|$.

While the p tuning parameters μ_j can be chosen by cross-validation, the computational cost explodes with p to search over a grid in p dimensions. Exploiting the presence of an optimal initial estimate in the form of $\overline{C}$, we recommend default choices for the hyperparameters as $\widehat{\mu}_j = 1/\|\overline{C}_j\|^{-2}$ which in spirit is similar to the adaptive lasso [72]. When predictor j is not important, the minimax ℓ_2-risk for estimating $C_0^{(j)}$ is $(\log q)/n$, so that $\|\overline{C}^{(j)}\| \asymp (\log q)/n$. Since $\|X_j\|^2 \asymp n$ by assumption, $\widehat{\mu}_j / \|X_j^{\mathrm{T}} R_j\| \asymp n^{1/2}/(\log q)^{3/2} \gg 1$, implying a strong penalty for all irrelevant predictors.

Following [68], posterior uncertainty in variable selection can be gauged if necessary by replacing $\overline{C}$ with the individual posterior samples for C in (3.8).

3.3.2 Post processing for rank estimation

To estimate the rank, we threshold the singular values of $X\widehat{C}_R$, with $\widehat{C}_R$ obtained from (3.9). In situations where row sparsity is not warranted, $\overline{C}$ can be used instead of $\widehat{C}_R$. For $s_1, \ldots, s_q$ the singular values of $X\widehat{C}_R$, 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, $\widehat{r} = \sum_{h=1}^{q} \mathrm{I}(\nu_h > 0) = \sum_{h=1}^{q} \mathrm{I}(s_h > \omega)$. We use the largest singular value of $Y - X\widehat{C}_R$ as the default choice of the threshold parameter ω, a natural candidate for the maximum noise level in the model.

3.4 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 3 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 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 $\mathrm{N}(0, \Sigma_X)$. 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 $B_* \in \Re^{p \times r_0}$ and $A_* \in \Re^{r \times r_0}$, 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 $\mathrm{N}(0, 1)$, and the remaining $(p - s)$ rows were set equal to zero.

As a competitor, we considered the sparse partial least squares (SPLS) approach due to [73]. 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. [73] suggested adding lasso type penalties while optimizing for the projection vectors for sparse high dimensional problems. Since SPLS returns a coefficient matrix which is both row sparse and rank reduced, we create a rank reduced matrix $\widehat{C}_{RR}$ from $\widehat{C}_R$ for a fair comparison. Recalling that $\widehat{C}_R$ has zero rows, let $\widehat{S}_R$ denote the sub-matrix corresponding to the non-zero rows of $\widehat{C}_R$. Truncate the singular value decomposition of $\widehat{S}_R$ to the first $\widehat{r}$ terms as obtained in §3.3.2. Insert back the zero rows corresponding to $\widehat{C}_R$ in the resulting matrix to obtain $\widehat{C}_{RR}$. Clearly, $\widehat{C}_{RR} \in \Re^{p \times q}$ so created is row sparse and has rank at most $\widehat{r}$; we shall refer to $\widehat{C}_{RR}$ as the Bayesian sparse multi-task learner (BSML). Matlab implementation of the proposed method can be found online at `www.stat.tamu.edu/~antik`.

For an estimator $\widehat{C}$ of C, we consider the mean square error, $\mathrm{MSE} = \|\widehat{C} - C_0\|_F^2/(pq)$, and the mean square prediction error, $\mathrm{MSPE} = \|X\widehat{C} - XC_0\|_F^2/(nq)$ to measure its performance. The squared estimation and prediction errors of SPLS and $\widehat{C}_{RR}$ for different settings are reported in table A.1 along with the estimates of rank. In our simulations we used the default 10 fold cross validation in the `cv.spls` function from the `R` package `spls`. The SPLS 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 SPLS overestimated the rank in all the settings considered. The proposed method also achieved superior performance in terms of the two squared errors, improving upon SPLS by as much as 5 times in some cases. Additionally, we observed that the performance of SPLS deteriorated relative to BSML with increasing number of

covariates.

In terms of variable selection, both methods had specificity and sensitivity both close to one in all the simulation settings listed in table A.1. Since SPLS 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 fit both methods with different choices of the postulated rank between 3 and 9. For the proposed method, we set q^* in §3.2.1 to be the postulated rank, that is, considered $B \in \Re^{p\times q^*}$ and $A \in \Re^{q\times q^*}$ for $q^* \in \{3, \ldots, 9\}$. For SPLS, we simply input q^* as the number of hidden components inside the function `spls`. Figure 3.1 plots the sensitivity and specificity of BSML and SPLS as a function of the postulated rank. While the specificity is robust for either method, the sensitivity of SPLS turned out to be highly dependent on the rank. The left panel of figure 3.1 reveals that at the true rank, SPLS only identifies 40% of the significant variables, and only achieves a similar sensitivity as BSML when the postulated rank is substantially overfitted. BSML, on the other hand, exhibits a decoupling effect wherein the overfitting of the rank does not impact the variable selection performance.

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 similarly as before except that the individual rows e_i of the matrix E was generated from $\mathrm{N}(0, \Sigma)$, with $\Sigma_{ii} = 1, \Sigma_{ij} = 0.5, 1 \leq i \neq j \leq q$. To accommodate the non-diagonal error covariance, we placed a inverse-Wishart(q, I_q) prior on Σ. An associate editor pointed out the recent article [4] which used spike-slab priors on the coefficients in a multiple response regression setting. They implemented a variational algorithm to 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 [74]; 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, which when set to the true value 10, resulted in comparatively poor performance. Table 3.2 reports 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 $(p, q) = (200, 30)$, BSML consistently outperformed [4] when the dimension was higher.

Table 3.1: Estimation and predictive performance of the proposed method (BSML) versus SPLS 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 were 10 and sample size was 100. For each combination of (p, q, r_0) the columns of the design matrix were generated from $\mathrm{N}(0, \Sigma_X)$. Two different choices of Σ_X was 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). The method achieving superior performance for each setting is highlighted in bold.

		(p,q)											
		(200,30)				**(500,10)**				**(1000,12)**			
		Independent		**Correlated**		**Independent**		**Correlated**		**Independent**		**Correlated**	
Rank	**Measures**	BSML	SPLS	BSML	SPLS	BSML	SPLS	BSML	SPLS	BSML	SPLS	BSML	SPLS
	$\hat{r}$	**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**	14	**5**	15	**3**	7	**5**	30	**3**	50	**3**	38
	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**	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**	69	**9**	61	**3**	10	**6**	24	**2**	108	**4**	129
	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	**6.7**	9.7	**6.8**	10.2	**6.6**	11.5
7	MSE	**6**	116	**10**	112	**3**	20	**5**	49	**2**	195	**4**	261
	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

Table 3.2: Variable selection performance of the proposed method in a non-diagonal error structure setting with independent and correlated predictors; $e_i \sim \Sigma$, $\sigma_{ii} = 1, \sigma_{ij} = 0.5$. Sensitivity and specificity of BSML is compared with [4].

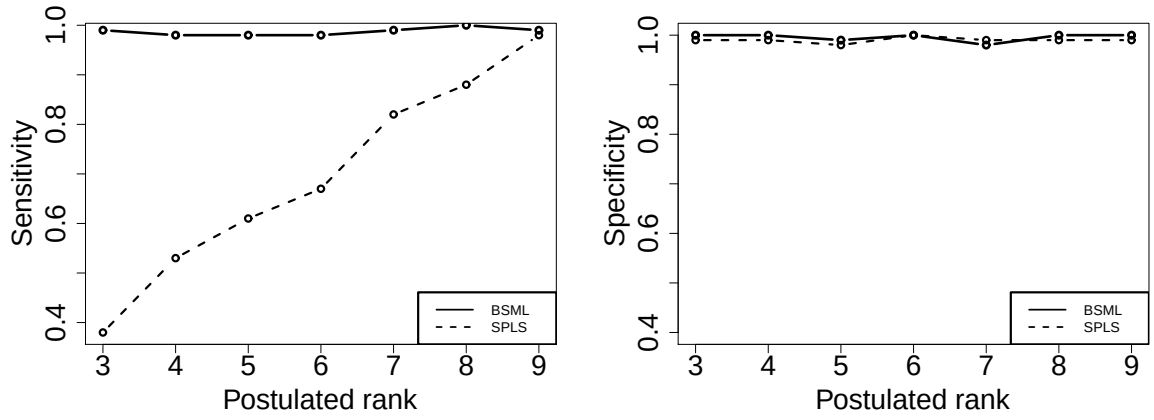
			BSML		[4]	
	$(\mathbf{p}, \mathbf{q})$	**Measure**	**Independent**	**Correlated**	**Independent**	**Correlated**
$\mathbf{r_0 = 3}$	**(200,30)**	Sensitivity	1	1	0.96	0.87
		Specificity	0.90	0.84	0.77	0.67
	(500,10)	Sensitivity	1	0.99	0.9	0.8
		Specificity	0.99	0.99	0.80	0.64
	(1000,12)	Sensitivity	0.99	0.99	0.92	0.63
		Specificity	0.99	0.99	0.80	0.64

3.5 Yeast Cell Cycle Data

Identifying transcription factors which are responsible for cell cycle regulation is an important scientific problem [73]. The yeast cell cycle data from [75] 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 [76] on chromatin immunoprecipitation contains the binding information of the 800 genes for 106 transcription factors (X). We analyze this data available publicly from the `R` package `spls` which has the above information completed for 542 genes. The yeast cell cycle data was also analyzed in [55] via sparse reduced rank regression (SRRR). Scientifically 21 transcription factors of the 106 were verified by [2] to be responsible for cell cycle regulation.

The proposed BSML procedure identified 33 transcription factors. Corresponding numbers for SPLS and SRRR were 48 and 69 respectively. Of the 21 verified transcription factors, the proposed method selected 14, whereas SPLS and SRRR selected 14 and 16 respectively. 10 additional tran-

Figure 3.1: Average sensitivity and specificity across 50 replicates is plotted for different choices of the postulated rank. Here $(p, q, r_0) = (1000, 12, 3)$. Values for BSML (SPLS) are in bold (dashed).



scription factors that regulate cell cycle were identified by [76], out of which 3 transcription factors were selected by our proposed method. Figure 3.2 plots the posterior mean, BSML estimate $\widehat{C}_{RR}$, and 95 % symmetric pointwise credible intervals for two common effects ACE2 and SW14 which are identified by all the methods. Similar periodic pattern of the estimated effects are observed as well for all the other two methods in contention, 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 placed inside the supplemental document.

The proposed automatic rank detection technique estimated a rank of 1 which is significantly different from SRRR (4) and SPLS (8). The singular values of $Y - X\widehat{C}_R$ showed a significant drop in magnitude after the first four values which agrees with the findings in [55]. The 10-fold cross validation error with a postulated rank of 4 for BSML was 0.009 and that of SPLS was 0.19.

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

3.5.1 Concentration results

In this section, we establish a minimax posterior concentration result under the prediction loss when the number of covariates are allowed to grow sub-exponentially in n. To best of our knowl-

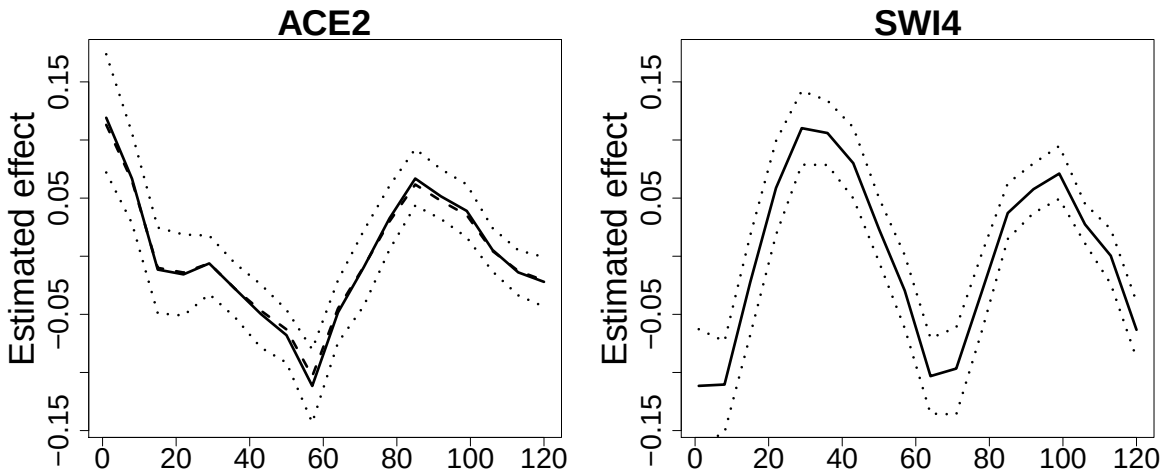


Figure 3.2: Estimated effects of ACE2 and SWI4, two of 33 transcription factors with non-zero effects on cell cycle regulation. Both have been scientifically verified by [2]. Dotted lines correspond to 95% posterior symmetric credible intervals, bold lines represent the posterior mean and the dashed lines plot values of the BSML estimate $\widehat{C}_{RR}$.

edge, this is the first such result in Bayesian reduced rank regression models. We are also not aware of a similar result in any model involving the horseshoe or another polynomial tailed shrinkage prior in ultrahigh-dimensional settings. [27] applied the general theory of posterior consistency [77] to linear models with growing number of covariates and established consistency for the horseshoe prior with a sample size dependent hyperparameter choice when $p = o(n)$. Results [25, 26] 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 [77, 28] is a key component in the general theory of posterior concentration. Let $\ell_0[s;p] = \{\theta_0 \in \Re^p : \#(1 \leq j < p : \theta_{0j} \neq 0) \leq s\}$ denote the space of p-dimensional vectors with at most s non-zero entries.

Lemma 3.5.1. *Let* Π_{HS} *denote the horseshoe prior on* $\Re^p$ *given by the hierarchy* $\theta_j \mid \lambda_j, \tau \sim \mathrm{N}(0, \lambda_j^2\tau^2)$, $\lambda_j \sim \mathrm{Ca}_+(0,1)$, $\tau \sim \mathrm{Ca}_+(0,1)$. *Fix* $\theta_0 \in \ell_0[s;p]$ *and let* $S = \{j : \theta_{0j} \neq 0\}$. *Assume* $s = o(p)$ *and* $\log p \leq n^\gamma$ *for some* $\gamma \in (0,1)$ *and* $\max \mid \theta_{0j} \mid \leq M$ *for some* $M > 0$ *for* $j \in S$.

Define $\delta = \{(s \log p)/n\}^{1/2}$. *Then we have, for some positive constant* K,

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

Proof. Using the conditional formulation of prior $\Pi_{\mathbf{HS}}$ defined in equation (3.2) we have,

$$\begin{aligned} \Pi_{\mathbf{HS}}(\|\beta - \beta_0\|_2 < \delta) &= \int_{\tau} pr(\|\beta - \beta_0\|_2 < \delta \mid \tau) f(\tau) d\tau \\ &\geq \int_{I_{\tau_*}} pr(\|\beta - \beta_0\|_2 < \delta \mid \tau) f(\tau) d\tau, \end{aligned} \tag{3.10}$$

where $I_{\tau_*} = [\tau_*/2, \tau_*]$ with $\tau_* = (s/p)^{3/2} \{(\log p)/n\}^{1/2}$. Let $S = \{1 \leq j \leq p : \theta_{0j} \neq 0\}$. We first provide a lower bound of the conditional probability $pr(\|\beta - \beta_0\|_2 < \delta \mid \tau \in I_{\tau_*})$. For $\tau \in I_{\tau^*}$, we have,

$$\begin{aligned} pr(\|\beta - \beta_0\|_2 < \delta \mid \tau) &\geq pr\left(\|\beta_S - \beta_{0S}\|_2 < \delta/2 \mid \tau\right) pr\left(\|\beta_{S^c}\|_2 < \delta/2 \mid \tau\right) \\ &\geq \prod_{j \in S} pr\left(|\beta_j - \beta_{0j}| < \frac{\delta}{2\sqrt{s}} \mid \tau\right) \prod_{j \in S^c} pr\left(|\beta_j| < \frac{\delta}{2\sqrt{p}} \mid \tau\right). \end{aligned} \tag{3.11}$$

For a fixed $\tau \in I_{\tau_*}$, we will provide lower bounds for each of the terms in the right hand side of (5.7); $pr\{|\beta_j - \beta_{0j}| < \delta/(2s^{1/2})\}$ for any $j \in S$, and $pr\{|\beta_j| < \delta/(2p^{1/2})\}$ for any $j \in S^c$.

We first consider $pr\{|\ \beta_j\ | < \delta/(2p^{1/2}) \mid \tau\}$ with $\tau \in I_{\tau_*}$. Since given τ and λ, $\beta_j \sim \mathrm{N}(0, \lambda_j^2 \tau^2)$, we use the Chernoff bound for a Gaussian random variable to obtain,

$$pr\big\{|\beta_j| > \delta/(2p^{1/2}) \mid \lambda_j, \tau\big\} \leq 2e^{-\delta^2/(8p\lambda_j^2\tau^2)} \leq 2e^{-\delta^2/(8p\lambda_j^2\tau_*^2)} = 2e^{-p^2/(8s^2\lambda_j^2)},$$

since $n\delta^2 = s\log p$. Thus,

$$\begin{aligned} pr\big\{|\beta_j| < \delta/(2p^{1/2}) \mid \tau\big\} &= \int_{\lambda_j} pr\big\{|\beta_j| < \delta/(2p^{1/2}) \mid \lambda_j, \tau\big\}\, f(\lambda_j)\, d\lambda_j \\ &\geq \int_{\lambda_j} \left\{1 - 2\exp\left(-\frac{p^2}{8s^2\lambda_j^2}\right)\right\} f(\lambda_j)\, d\lambda_j \\ &= 1 - \frac{4}{\pi}\int_{\lambda_j} \exp\left(-\frac{p^2}{8s^2\lambda_j^2}\right)(1+\lambda_j^2)^{-1} d\lambda_j = 1 - \frac{4}{\pi}\mathcal{I}, \end{aligned}$$

where $\mathcal{I} = \int_{\lambda_j} \exp\big\{-p^2/(8s^2\lambda_j^2)\big\}\,(1+\lambda_j^2)^{-1} d\lambda_j$. We then bound the integrand from above as follows,

$$\begin{aligned} \mathcal{I} = \int_{\lambda_j} \exp\left(-\frac{p^2}{8s^2\lambda_j^2}\right)(1+\lambda_j^2)^{-1} d\lambda_j &\leq \int_{\lambda_j} \exp\left(-\frac{p^2}{8s^2\lambda_j^2}\right)\lambda_j^{-2} d\lambda_j \\ &= \frac{1}{2}\int_0^\infty z^{-1/2}\exp\left(-\frac{p^2 z}{8s^2}\right) dz, \\ &= \frac{\Gamma(1/2)}{\{2p^2/(8s^2)\}^{1/2}} = \frac{s\sqrt{2\pi}}{p}, \end{aligned}$$

where we made the substitution $z = 1/\lambda^2$ at the third step. Thus, for $\tau \in I_{\tau_*}$, $pr(|\ \beta_j\ | < \delta/2p^{1/2} \mid \tau) \geq 1 - Rs/p$, where $R = (32/\pi)^{1/2}$.

Next, for $pr(|\ \beta_j - \beta_{0j}\ | < \delta_0 | \tau)$ with $\tau \in I_{\tau_*}$, we have, letting $\delta_0 = s^{-1/2}(\delta/2) = 2^{-1}\{(\log p)/n\}^{1/2}$,

$$\begin{aligned} pr(|\ \beta_j - \beta_{0j}\ | < \delta_0 \mid \tau) &= (2/\pi^3)^{1/2}\int_{\lambda_j}\int_{|\beta_j-\beta_0|<\delta_0} \exp\{-\beta_j^2/(2\lambda_j^2\tau^2)\}\frac{1}{\lambda_j\tau(1+\lambda_j^2)}\, d\beta_j d\lambda_j \\ &\geq (2/\pi^3)^{1/2}\int_{|\beta_j-\beta_0|<\delta_0}\int_{1/\tau}^{2/\tau} \exp\{-\beta_j^2/(2\lambda_j^2\tau^2)\}\frac{1}{\lambda_j\tau(1+\lambda_j^2)}\, d\lambda_j d\beta_j \\ &\geq (2/\pi^3)^{1/2}\int_{|\beta_j-\beta_0|<\delta_0} \exp(-\beta_j^2/2)\left(\int_{1/\tau}^{2/\tau}\frac{1}{1+\lambda_j^2}\, d\lambda_j\right) d\beta_j, \end{aligned}$$

since for $\lambda_j \in [1/\tau, 2/\tau]$, $1/(\lambda_j \tau) \geq 1/2$ and $\exp\{-\beta_j^2/(2\lambda_j^2\tau^2)\} \geq \exp(-\beta_j^2/2)$. Continuing,

$$\begin{aligned} pr(\mid \beta_j - \beta_{0j} \mid < \delta_0 \mid \tau) &\geq (2/\pi^3)^{1/2} \ \frac{\tau}{4+\tau^2} \int_{|\beta_j - \beta_0| < \delta_0} \exp(-\beta_j^2/2) d\beta_j \\ &\geq (2/\pi^3)^{1/2} \ \frac{\tau}{4+\tau^2} \ \exp\{-(M+1)^2/2\} \ \delta_0 \\ &\geq K \, \tau \, \delta_0 \\ &\geq K \left(\frac{s}{p}\right)^{3/2} \frac{\log p}{n} \geq K_* p^{-5/2}, \end{aligned}$$

where in the third step, we used $4 + \tau^2 < 5$ and in the final step we used $n < p$. Substituting these bounds in (5.7), we have for $\tau \in I_{\tau_*}$

$$pr(\|\beta - \beta_0\|_2 < \delta \mid \tau) \geq (1 - Rs/p)^{p-s} K_* e^{-(5s/2) \log p} \geq e^{-Ks \log p}, \tag{3.12}$$

where K is a positive constant. The proof is completed upon observing that $pr(\tau \in I_{\tau_*}) \geq \tau_*/(2\pi)$, so that we get,

$$\Pi_{\mathbf{HS}}(\|\beta - \beta_0\|_2 < \delta) \geq e^{-Ks \log p}, \tag{3.13}$$

for some positive constant K. □

We believe Lemma 3.5.1 will be of independent interest in various other models involving the horseshoe prior. The only other instance of a similar prior concentration result in $p \gg n$ settings that we are aware of is for the Dirichlet–Laplace prior [37].

We now study concentration properties of the posterior distribution in model (3.3) in $p \gg n$ settings. To aid the theoretical analysis, we adopt the fractional posterior framework of [28], 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 $\alpha \in (0, 1)$ and recall the prior Π_C on C defined after equation (3.4) and set Π_Σ as the inverse-Wishart prior for Σ. The α-fractional

posterior for (C, Σ) under model (3.3) is then given by

$$\Pi_{n,\alpha}(C, \Sigma \mid Y) \propto \{p(Y \mid C, \Sigma; X)\}^{\alpha} \, \Pi_C(C) \, \Pi_\Sigma(\Sigma). \tag{3.14}$$

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 Appendix A document 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 3.5.2 (Growth of number of covariates)**.** $\log p/n^{\gamma} \leq 1$ *for some* $\gamma \in (0, 1)$.

Assumption 3.5.3. *The number of response variables* q *is fixed.*

Assumption 3.5.4 (True coefficient matrix)**.** *The true coefficient matrix* C_0 *admits the decomposition* $C_0 = B_0 A_0^{\mathrm{T}}$ *where* $B_0 \in \Re^{p \times r_0}$ *and* $A_0 \in \Re^{q \times r_0}$ *for some* $r_0 = \kappa q$, $\kappa \in (0, 1]$. *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,* $\max_{j,h} \mid C_{0jh} \mid < T$ *for some* $T > 0$.

Assumption 3.5.5 (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 the minimum and maximum singular values of a matrix* P *respectively.*

Assumption 3.5.6 (Design matrix)**.** *For* X_j *the* j*th column of* X, $\max_{1 \leq j \leq p} \|X_j\| \asymp n$.

Assumption 1 allows the number of covariates p to grow at a sub-exponential rate of $e^{n^{\gamma}}$ 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 [55]. The positive definiteness of Σ_0 is ensured by assumption 4. Finally, assumption

4 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 $\mathrm{N}(0,1)$ or any sub-Gaussian distribution. It also encompasses situations when the columns of X are standardized.

Let $p_0(Y \mid X) \equiv p(Y \mid C_0, \Sigma_0; X)$ denote the true density. For two densities q_1, q_2 with respect to a dominating measure μ, recall the squared Hellinger distance $h^2(q_1, q_2) = \{(1/2)\int (q_1^{1/2} - q_2^{1/2})^2 d\mu\}$. 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., [78]. In the following theorem, we provide a non-asymptotic bound to the squared Hellinger loss under the fractional posterior $\Pi_{n,\alpha}$.

Theorem 3.5.7. *Suppose $\alpha \in (0,1)$ and let $\Pi_{n,\alpha}$ be defined as in* (5.11)*. 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 $\rho = s_{\max}(\Sigma_0)/s_{\min}(\Sigma_0)$ and K_1 is an absolute positive constant related to Lemma 4.0.2 in chapter 4 and $\epsilon_n = \{(qr_0 + r_0 s \log p)/n\}^{1/2}$. Then for for any $D \geq 1$ and $t > 0$,*

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

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.

Proof. Fix $\alpha \in (0,1)$. Define $U_n = \left[(C,\Sigma) : \frac{1}{n} D_\alpha\{(C,\Sigma),(C_0,\Sigma_0)\} > \frac{D+3t}{1-\alpha}\widetilde{\epsilon}_n^{\,2}\right]$. Let $\eta = (C,\Sigma)$ and $\eta_0 = (C_0,\Sigma_0)$. Also let $p_\eta^{(n)}$ denote the density of $Y \mid X$ with parameter value η under model (3.3). 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,

$$\Pi_{n,\alpha}(U_n \mid Y) = \frac{\int_{U_n} e^{-\alpha r_n(\eta,\eta_0)} d\Pi_\eta}{\int e^{-\alpha r_n(\eta,\eta_0)} d\Pi_\eta} := \frac{N_n}{D_n}, \tag{3.15}$$

where $r_n(\eta, \eta_0) = \log p_{\eta_0}^{(n)}/p_{\eta}^{(n)}$. We prove in Lemma 4.1.1 of chapter 4, with $P_{(C_0,\Sigma_0)}^{(n)}$-probability at least $1 - K_2/\{(D-1+t)^2 n\widetilde{\epsilon}_n^2\}$, the denominator $D_n \geq e^{-\alpha(D+t)n\widetilde{\epsilon}_n^2}$ for some positive constant K_2. For the numerator proceeding similarly as in the proof of theorem 3.2 in [28] we arrive at, $P_{(C_0,\Sigma_0)}^{(n)}\left\{N_n \leq e^{-(D+2t)n\widetilde{\epsilon}_n^2}\right\} \geq 1 - 1/\{(D-1+t)^2 n\widetilde{\epsilon}_n^2\}$. Combining the upper bound for N_n and lower bound for D_n we then have,

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

with $P_{\eta_0}^{(n)}$-probability at least $1 - K_2/\{(D-1+t)^2 n\widetilde{\epsilon}_n^2\}$. Finally, we combine the following two inequalities [28] for two densities p and q to arrive at the conclusion, $D_{1/2}(p,q) \geq 2h^2(p,q)$, and $\frac{\alpha}{\beta}\frac{1-\beta}{1-\alpha} D_\beta(p,q) \leq D_\alpha \leq D_\beta,\ 0 < \alpha \leq \beta < 1$. □

The proof of theorem 3.5.7 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 3.5.1 we prove in Lemma 4.0.6 of chapter 4 that for our choice of Π_C we have sufficient prior concentration around row and rank sparse matrices.

[54] obtained $n\epsilon_n^2 = (qr_0 + r_0 s \log p)$ as the minimax risk under the loss $||\ XC - XC_0\ ||_F^2$ for model (5.15) with $\Sigma = \mathrm{I}_q$ and when C_0 satisfies assumption 3. Theorem 3.5.7 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 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 3.5.8. *Fix $\alpha \in (0,1)$. Suppose Assumptions 1-5 are satisfied and assume that Σ is known to be Σ_0. Without loss of generality assume $\Sigma_0 = \mathrm{I}_q$. Let ϵ_n be defined as in theorem 3.5.7. Then for any $D \geq 2$ and $t > 0$,*

$$\Pi_{n,\alpha}\left\{C \in \Re^{p\times q} : \frac{1}{nq}\|XC - XC_0\|_F^2 \geq \frac{2(D+3t)}{\alpha(1-\alpha)}\epsilon_n^2 \mid Y\right\} \leq e^{-tn\epsilon_n^2}$$

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

$$|| X\overline{C} - XC_0 ||_F^2 \leq K_2 n\epsilon_n^2,$$

for some positive constants K_1 and K_2 independent of α.

Proof. For $C \in \Re^{p\times q}$, we write $p_C^{(n)}$ to denote the density of $Y|X$ which is proportional to $e^{-\text{tr}\{(Y-XC)(Y-XC)^{\text{T}}\}/2}$. For any $C^* \in \Re^{p\times q}$ we define a ϵ-neighborhood as,

$$B_n(C^*, \epsilon) = \left\{ C \in \Re^{p\times q} : \int p_{C^*}^{(n)} \log(p_{C^*}^{(n)}/p_C^{(n)}) dY \leq n\epsilon^2, \int p_{C^*}^{(n)} \log^2(p_{C^*}^{(n)}/p_C^{(n)}) dY \leq n\epsilon^2 \right\}. \tag{3.16}$$

Observe that $B_n(C_0, \epsilon) \supset A_n(C_0, \epsilon) = \left\{C \in \Re^{p\times q} : \frac{1}{n}\|XC - XC_0\|_F^2 \leq \epsilon^2\right\}$ for all $\epsilon > 0$ and the Rényi divergence $D_\alpha(p_C^{(n)}, p_{C_0}^{(n)}) = \frac{\alpha}{2}\|XC - XC_0\|_F^2$. By a similar argument as in step 1 of the proof of Lemma 4.1.1 of chapter 4, we have $\Pi_C\{A_n(C_0, \epsilon_n)\} \geq e^{-Kn\epsilon_n^2}$ for positive K. Hence the first part follows from [28, theorem 3.2].

For the second part first observe that from [28, corollary 3.3] we get $\int (nq)^{-1} || XC - XC_0 ||_F^2$ $\Pi_{n,\alpha}(dC \mid Y) \leq K_2\{\alpha(1-\alpha)\}^{-1}\epsilon_n$ with $P_{C_0}^{(n)}$-probability at least $1 - K_1/\{n\epsilon_n^2\}$, where K_1 and K_2 are positive constants independent of α. using the convexity of the Frobenius norm and applying Jensen's inequality, we get $\alpha/2 \,||\, X\overline{C} - XC_0 \,||_F^2 = \alpha/2 \,||\, X\int C\,\Pi_{n,\alpha}(dC) - X\int C_0\,\Pi_{n,\alpha} \,||_F^2 \leq$ $\alpha/2 \int || XC - XC_0 ||_F^2\, \Pi_{n,\alpha}(dC) \leq Kn(1-\alpha)^{-1}\epsilon_n^2$ for some positive K. □

For theorem 3.5.8 the optimal bound is obtained for $\alpha = 1/2$ which is consistent with [79] where the authors consider a pseudo-likelihood approach for weighted model aggregation of several least squares estimates.

3.5.2 Fractional and usual posterior for reduced rank models with prior Π_C

While [28] proved that $\Pi_{n,\alpha}(\cdot \mid Y)$ converges to $\Pi_n(\cdot \mid Y)$ weakly as $\alpha \to 1-$ for general statistical models and priors, in this section we investigate what more can be said in reduced rank models with Π_C as the prior and given some rate information is available for the fractional pos-

terior. As a measure of discrepancy between the fractional posterior $\Pi_{n,\alpha}(\cdot \mid Y)$ and the usual posterior $\Pi(\cdot \mid Y)$ we use the Kullback-Liebler divergence defined as $D(p \mid\mid q) = \int p \log(p/q)\, d\mu$ for two densities p and q with some common dominating measure μ. Since for two k- dimensional Gaussian distributions $\mathrm{N}(\theta_1, \mathrm{I}_k)$ and $\mathrm{N}(\theta_2, \mathrm{I}_k)$, the Kullback-Liebler divergence is $\mid\mid \theta_1 - \theta_2 \mid\mid_2^2 \asymp k$, we consider a Cesáro average of the form $p^{-1}\, D\{\Pi_{n,\alpha}(\cdot \mid Y), \Pi(\cdot \mid Y)\}$. For a discussion on information theoretic motivation for using Cesáro averages of Kullback-Liebler distances see [80].

Theorem 3.5.9. *Consider model* (5.15) *with* $\Sigma = \mathrm{I}_q$ *and the prior on* C *is* Π_C *as defined in section 2.1. Recall* ϵ_n *from theorem 3.5.8. Then,*

$$\lim_{\alpha \to 1-} p^{-1} D\{\Pi_{n,\alpha}(\cdot \mid Y) \mid\mid \Pi(\cdot \mid Y)\} \leq K_1 \frac{(qr_0 + r_0 s \log p)}{p},$$

with $P_{C_0}^{(n)}$*-probability at least* $1 - K_2/\{n\epsilon_n^2\}$*, where* K_1 *and* K_2 *are positive constants independent of* α*.*

Proof. Let us write $m_\alpha(Y) = \int l_n(C)^\alpha \Pi(dC)$ and $m(Y) = \int l_n(C)\Pi(dC)$ where $l_n(\cdot)$ is the likelihood function for model (1) from the main document. From Theorem 3.4 of [28] there exists A_n such that $A_n = \{Y : \lim_{\alpha \to 1-} m_\alpha(Y) = m(Y)\}$ and $\mathbb{P}_{C_0}(A_n) = 1$. Let us now turn our attention to $D\{\Pi_{n,\alpha}(\cdot \mid Y), \Pi(\cdot \mid Y)\}$.

$$\begin{aligned}
&D\{\Pi_{n,\alpha}(\cdot \mid Y), \Pi(\cdot \mid Y)\} = \int \log\{\Pi_{n,\alpha}(\cdot \mid Y)/\Pi(\cdot \mid Y)\}\, \Pi_{n,\alpha}(\cdot \mid Y) dC \\
&= \log m(Y)/m_\alpha(Y) + \frac{(1-\alpha)}{2} \int \|Y - XC\|_F^2 \Pi_{n,\alpha}(\cdot \mid Y) dC \\
&\leq \log m(Y)/m_\alpha(Y) + (1-\alpha) \int \|Y - XC_0\|_F^2 \Pi_{n,\alpha}(\cdot \mid Y) dC \\
&\qquad + (1-\alpha) \int \|XC - XC_0\|_F^2 \Pi_{n,\alpha}(\cdot \mid Y) dC.
\end{aligned}$$

Using theorem 3.5.8 and from corollary 3.3 of [28] we have the following following result,

$$\alpha/2 \int \|XC - XC_0\|_F^2 \, \Pi_{n,\alpha}(\cdot \mid Y) \, dC \leq K_1 n\epsilon_n^2 (1-\alpha)^{-1}$$

with P_{C_0} probability at least $1 - K_2/(n\epsilon_n^2)^{-1}$, where K_1, K_2 are positive constants independent of α. Let $B_n \in \sigma^{Y^{(n)}}$ be the set where the above result holds. Then for $Y \in A_n \cap B_n$ we have,

$$D\{\Pi_{n,\alpha}(\cdot \mid Y), \Pi(\cdot \mid Y)\} \leq \log m(Y)/m_\alpha(Y) + (1-\alpha)\|Y - XC_0\|_F^2 + (2/\alpha)K_1 n\epsilon_n^2$$

Letting $\alpha \to 1-$ in the above display, we get for $y \in A_n \cap B_n$

$$\lim_{\alpha \to 1_-} p^{-1} D\{\Pi_{n,\alpha}(\cdot \mid Y), \Pi(\cdot \mid Y)\} \leq 2K_1(n\epsilon_n^2)/p.$$

The result then follows from a union probability bound on $A_n \cap B_n$. □

Theorem 3.5.9 suggests in the limit $\alpha \to 1-$ the fractional posterior and the usual posterior for Π_C and model (5.15) are only ϵ_n apart in the average Kullback-Leibler sense. Thus posterior/fractional posterior summaries such as mean, median are expected to be close with high probability under $P_{C_0}^{(n)}$. Further empirical evidence of virtually indistinguishable results from $\Pi_{n,\alpha}(\cdot \mid Y)$ and $\Pi(\cdot \mid Y)$ is provided in appendix A justifying our choice of the theoretical environment. From a computational point of view, for model (5.15), 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.

3.6 Definitions required for proofs of Theorem 3.5.7, 3.5.8, 3.5.9

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 defined as $D_\alpha(\theta, \theta_0) = (\alpha-1)^{-1} \log \int p_\theta^\alpha p_{\theta_0}^{1-\alpha} d\mu$. The α-affinity between p_θ and p_{θ_0} is denoted by $A_\alpha(p_\theta, p_{\theta_0}) = \int p_\theta^\alpha p_{\theta_0}^{1-\alpha} d\mu = e^{-(1-\alpha)D_\alpha(p_\theta, p_{\theta_0})}$. See [28] for a review of Rényi divergences.

3.7 Derivation of equation (3.8)

Set $\Sigma = \mathrm{I}_q$. Suppose $Y^* \in \Re^{n\times q}$ be n future observations with design points X so that given C, Y^* can be decomposed into $Y^* = XC + E^*$ where E^* where the individual rows of E^* follow $\mathrm{N}(0, \Sigma)$. We define the utility function in terms of loss of predicting these n new future observations. To encourage sparsity in rows of a coefficient matrix Γ that balances the prediction we add a group lasso penalty [71] to this utility function. We define the utility function as,

$$\mathcal{L}(Y^*, \Gamma) = \|Y^* - X\Gamma\|_F^2 + \sum_{j=1}^{p} \mu_j \|\Gamma^{(j)}\|_2 \tag{3.17}$$

where the p tuning parameters $\{\mu_j\}_{j=1}^p$ control the penalty for selecting each predictor variable and $\Phi^{(j)}$ represents the j^{th} row of any matrix Φ. Intuitively we want μ_j to be small if the j^{th} predictor is important and vice versa. The expected risk, $\mathbb{E}\{\mathcal{L}(Y^*, \Gamma)\}$, after integrating over the space of all such future observations given C and Σ, is

$$\mathcal{L}(\Gamma, C, \Sigma) = q\ \mathrm{tr}(\Sigma) + \|XC - X\Gamma\|_F^2 + \sum_{j=1}^{p} \mu_j \|\Gamma^{(j)}\|_2. \tag{3.18}$$

Finally we take expectation of this quantity with respect to $\pi(C \mid Y, X)$ and drop the constant terms to obtain (3.8).

3.8 Derivation of equation (3.9)

We let Φ_j and $\Phi^{(j)}$ denote the j^{th} column and row of a generic matrix Φ. Using the subgradient of (10) with respect to $\Gamma^{(j)}$ [81], we have

$$2X_j^{\mathrm{T}}(X\Gamma - X\overline{C}) + \mu_j \alpha_j = 0, \quad j = 1, \ldots, p, \tag{3.19}$$

where $\alpha_j = \Gamma^{(j)}/\|\Gamma^{(j)}\|$ if $\|\Gamma^{(j)}\| \neq 0$ and $\|\alpha_j\| < 1$ when $\|\Gamma^{(j)}\| = 0$. For $\Gamma^{(j)} = 0$ we can rewrite (3.19) as, $2X_j^{\mathrm{T}}(\sum_{k\neq j} X_k\Gamma^{(k)} - X\overline{C}) + \mu_j\alpha_j = 0$ which imply that $\alpha_j = -2X_j^{\mathrm{T}}R_j/\mu_j$, where R_j is the residual matrix obtained after regressing $X\overline{C}$ on X leaving out the j^{th} predictor,

$R_j = X\overline{C} - \sum_{k \neq j} X_k \Gamma^{(k)}$. We can use this to set $\Gamma^{(j)}$ to zero: if $\alpha_j < 1$ set $\Gamma^{(j)} = 0$. Otherwise we have $2X_j^{\mathrm{T}}(X_j\Gamma^{(j)} - R_j) + \mu_j\Gamma^j/\|\Gamma^{(j)}\| = 0$. Solving for $\Gamma^{(j)}$ in the above equation we then get,

$$\Gamma^{(j)} = \left(X_j^{\mathrm{T}}X_j + \frac{\mu_j}{2\|\Gamma^{(j)}\|}\right)^{-1} X_j^{\mathrm{T}}R_j. \tag{3.20}$$

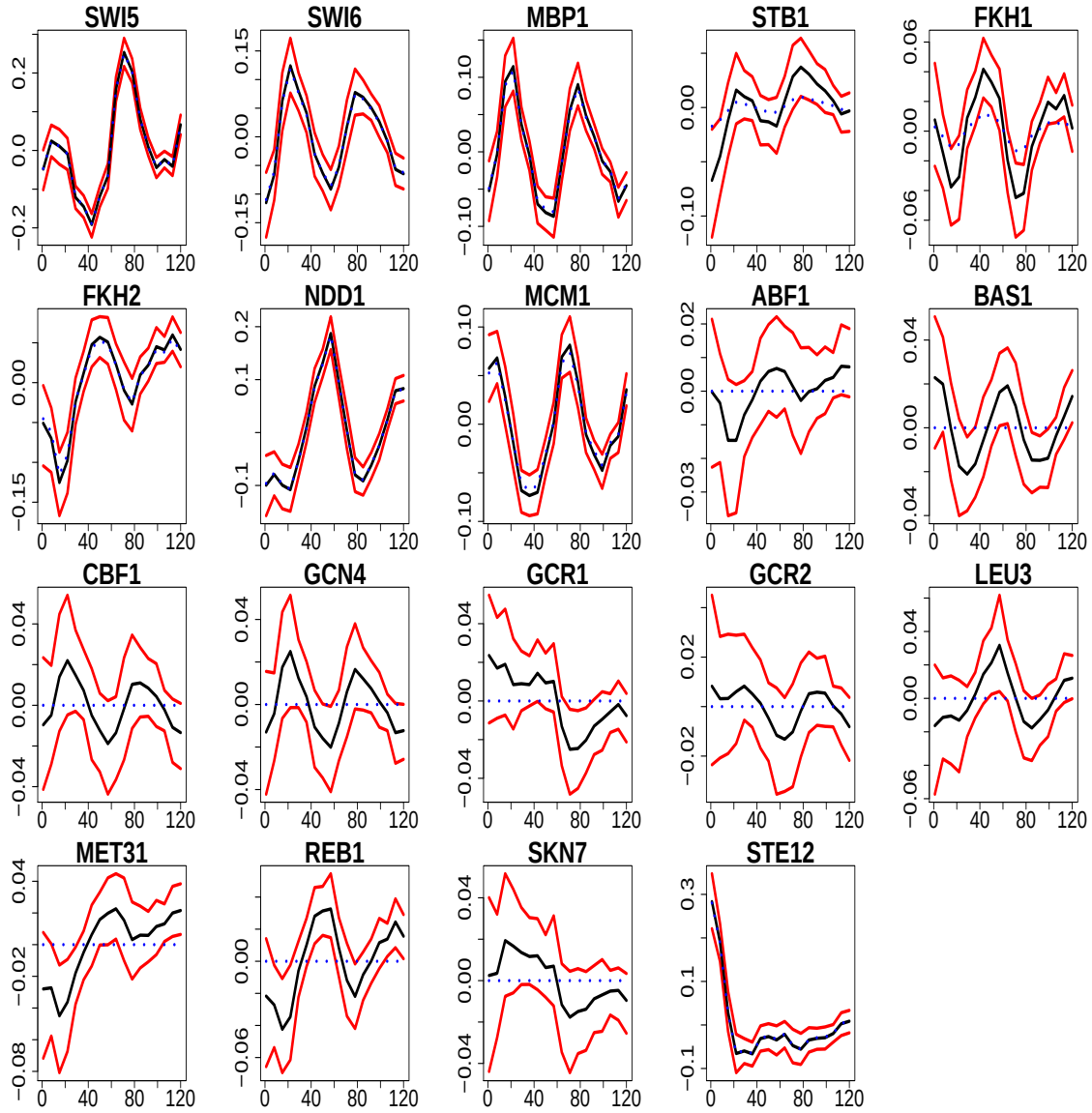
This solution is dependent on the unknown quantity $\|\Gamma^{(j)}\|$. However, taking norm on both sides in (3.20) we get a value of $\|\Gamma^{(j)}\|$ which does not involve any unknown quantities: $\|\Gamma^{(j)}\| = (\|X_j^{\mathrm{T}}R_j\| - \mu_j/2)/X_j^{\mathrm{T}}X_j$. Substituting this in (3.20) we get, $\Gamma^{(j)} = (X_j^{\mathrm{T}}X_j)^{-1}\left(1 - \mu_j/2\|X_j^{\mathrm{T}}R_j\|\right)X_j^{\mathrm{T}}R_j$.

Finally, combining the case when $\Gamma^{(j)} = 0$, we have (3.9).

3.9 Further results on Yeast cell cycle data

The yeast cell cycle data consists of mRNA measurements Y, measured every 7 minutes in a period of 119 minutes. The covariates X are binding information on 106 transcription factors. When applied to this data, the proposed method identified 33 transcription factors out of 106 that driving the variation in mRNA measurements. 14 of the identified transcription factors are among the 21 scientifically verified [76]. In section 3.5 we provided estimated effects of two of the 21 scientifically verified transcription factors. Here we plot the estimated effects of the remaining transcriptions factors that were scientifically verified.

Figure 3.3: Estimated effects of the 19 of 21 scientifically verified transcription factors selected by the proposed method. Effects of other two, viz. ACE2 and SWI4 are included in the main manuscript. Red lines correspond to 95% posterior symmetric credible intervals, black lines represent the posterior mean and the blue dashed line plots values of the BSML estimate $\widehat{C}_{RR}$.



4.1 Prior concentration results

We establish a number of results in the following sequence of Propositions and Lemmas to prove Theorem 3.5.7 and 3.5.8. The main goal here would be to establish prior concentration results around true model parameters. Recall the definitions of $\widetilde{\epsilon}_n$ and ϵ_n from Theorem 3.5.7 and 3.5.8 in the previous chapter respectively. For Theorem 3.5.7 we need a lower bound on the prior probability assigned to the set $B_n^*(\eta_0, \widetilde{\epsilon}_n) = \{\eta = (C, \Sigma) : \int p_{\eta_0}^{(n)} \log(p_{\eta_0}^{(n)}/p_{\eta}^{(n)}) dY \leq n\widetilde{\epsilon}_n^2\}$ by the product prior $\Pi_\eta = \Pi_C \otimes \Pi_\Sigma$. Similarly, for Theorem 3.5.8, we need the prior probability of the set $B_n(C_0, \epsilon_n) = \{C \in \Re^{p\times q} : \int p_{C_0}^{(n)} \log(p_{C_0}^{(n)}/p_{C}^{(n)}) dY \leq n\epsilon_n^2, \int p_{C_0}^{(n)} \log^2(p_{C_0}^{(n)}/p_{C}^{(n)}) dY \leq n\epsilon_n^2\}$. We start by characterizing $B_n^*(\eta_0, \epsilon_n)$ and $B_n(C_0, \epsilon_n)$ in terms of $\|\Sigma - \Sigma_0\|_F^2$ and $\|XC - XC_0\|_F^2$. But first, we record few inequalities which we will use frequently in our subsequent analysis. For any two matrices A & B,

$$s_{\min}(A)\|B\|_F \leq \|AB\|_F \leq \|A\|_2\|B\|_F \tag{4.1}$$

$$s_{\min}(A)\|B\|_2 \leq \|AB\|_2 \leq \|A\|_2\|B\|_2 \tag{4.2}$$

For a proof of these inequalities, see the supplementary document of [37].

Proposition 4.1.1. *Consider model 3.1 in chapter 3, $Y = XC + E$, $e_i \sim \mathrm{N}(0, \Sigma)$. Then,*

$$\int p_{\eta_0}^{(n)} \log(p_{\eta_0}^{(n)}/p_{\eta}^{(n)}) dY = \frac{n}{2}\log\frac{|\Sigma|}{|\Sigma_0|} + \frac{n}{2}\mathrm{tr}(\Sigma^{-1}\Sigma_0 - \mathrm{I}_q) + \frac{1}{2}\|(XC - XC_0)\Sigma^{-1}(XC - XC_0)^{\mathrm{T}}\|_F^2 \tag{4.3}$$

Moreover, when $\Sigma_0 = \Sigma = \mathrm{I}_q$, we have $\int p_{C_0}^{n} \log(p_{C_0}^{(n)}/p_{C}^{(n)}) dY = 2^{-1}\|XC - XC_0\|_F^2$ and if $\|C - C_0\|_F^2 < 1$, then $\int p_{C_0}^{(n)} \log^2(p_{C_0}^{(n)}/p_{C}^{(n)}) \leq 2^{-1}\|XC - XC_0\|_F^2$.

Proof. The expression for $\int p_{\eta_0}^{(n)} \log(p_{\eta_0}^{(n)}/p_{\eta}^{(n)}) dY$ follows directly from the formula of Kullback-Liebler divergence between two Normal distributions. Setting $\Sigma = \Sigma_0$ on the right hand side of

4.3 yields the second assertion. Using formulas for variance and covariance of quadratic forms of Normal random vectors the third assertion is proved by noting that $\|C-C_0\|_F^4 \leq \|C-C_0\|_F^2$ when $\|C-C_0\|_F^2 < 1$.

□

Lemma 4.1.2. *Let Σ, Σ_0 be $q \times q$ positive definite matrices and $\delta \in (0,1)$. If $\|\Sigma - \Sigma_0\|_F \leq \delta$ and $\delta / s_{\min}(\Sigma_0) < 1/2$, then*

$$\mathrm{tr}(\Sigma_0 \Sigma^{-1} - \mathrm{I}_q) - \log \mid \Sigma_0 \Sigma^{-1} \mid \leq \frac{(K \log \rho)\delta^2}{s_{\min}^2(\Sigma_0)},$$

where K is some absolute positive constant and $\rho = 2s_{\max}(\Sigma_0)/s_{\min}(\Sigma_0)$.

Furthermore,

$$\|(XC - XC_0)\Sigma^{-1}(XC - XC_0)^{\mathrm{T}}\|_F^2 \leq \{4/s_{\min}^2(\Sigma_0)\}\|XC - XC_0\|_F^2$$

.

Proof. For the first claim see Lemma 1.3 in the supplementary document of [37]. To prove the second claim, by (4.1) we have, $\|(XC - XC_0)\Sigma^{-1}(XC - XC_0)^{\mathrm{T}}\|_F^2 \leq \|XC - XC_0\|_F^2 \|\Sigma^{-1}\|_2^2$ where $\|P\|_2$ is the largest singular value of the matrix P. Lemma 1.3 from [37] also provides a lower bound of $s_{\min}(\Sigma)$ as $s_{\min}(\Sigma_0)/2$. Since $\|\Sigma^{-1}\|_2 = 1/s_{\min}(\Sigma_0)$, the result follows immediately. □

According to Lemma 3.5.1, it is equivalent to consider prior concentration of the Frobenius balls $\|XC - XC_0\|_F$ and $\|\Sigma - \Sigma_0\|_F$ for sufficient prior concentration around Kullback-Leibler neighborhoods. In the following sequence of lemmas we prove Π_C and $\Pi_\Sigma \equiv \text{inv-Wishart}(q, \mathrm{I}_q)$ satisfies such concentration.

Lemma 4.1.3. *Suppose the $q \times q$ matrix $\Sigma \sim \Pi_\Sigma$ where Π_Σ is the inverse-Wishart distribution with parameters (q, I_q). Let Σ_0 be any fixed symmetric positive definite matrix. Let δ be such that*

$2q^{-1/2}\delta/s_{\min}(\Sigma_0) \in (0,1)$. *Then,*

$$\Pi_\Sigma(\Sigma : \|\Sigma - \Sigma_0\|_F < \delta) \geq e^{-Tn\delta^2},$$

where T is a positive constant.

Proof. By inequality (4.1), we have the following,

$$\begin{aligned}
\Pi_\Sigma(\Sigma : \|\Sigma - \Sigma_0\|_F < \delta) &\geq \Pi_\Sigma(\Sigma : \|\Sigma_0\Sigma^{-1} - \mathrm{I}_q\|_F < \delta/\|\Sigma\|_2) \\
&\geq \Pi_\Sigma\{\Sigma : \|\Sigma_0\Sigma^{-1} - \mathrm{I}_q\|_F < \delta/s_{\min}(\Sigma)\} \\
&\geq \Pi_\Sigma\{\Sigma : \|\Sigma_0\Sigma^{-1} - \mathrm{I}_q\|_F < 2\delta/s_{\min}(\Sigma_0)\} \\
&\geq \Pi_\Sigma\{\Sigma : \|\Sigma_0^{1/2}\Sigma^{-1}\Sigma_0^{1/2} - \mathrm{I}_q\|_F < 2\delta/s_{\min}(\Sigma_0)\},
\end{aligned}$$

where we have used the the lower bound $s_{\min}(\Sigma) > s_{\min}(\Sigma_0)/2$ from the previous lemma and the similarity of the two matrices $\Sigma_0\Sigma^{-1}$ and $\Sigma_0^{1/2}\Sigma^{-1}\Sigma_0^{1/2}$. Let ϕ_j be the j^{th} eigenvalue of $H = \Sigma_0^{1/2}\Sigma^{-1}\Sigma_0^{1/2}$ where $j = 1, \ldots, q$. Then $\|H - \mathrm{I}_q\|_F^2 = \sum_{j=1}^q (\phi_j - 1)^2 \leq 4\delta^2/s_{\min}^2(\Sigma_0)$. Letting $\delta_* = 2q^{-1/2}\delta/s_{\min}(\Sigma_0)$, we then have,

$$\begin{aligned}
&\Pi_\Sigma\left\{\phi_j : \sum_{j=1}^q (\phi_j - 1)^2 < 4\delta^2/s_{\min}^2(\Sigma_0),\ j = 1, \ldots, q\right\} \\
&\geq \Pi_\Sigma\left\{\phi_j : (\phi_j - 1)^2 < \delta_*^2,\ j = 1, \ldots, q\right\} \\
&= \Pi_\Sigma\left\{\phi_j : \frac{1-\delta_*}{1+\delta_*} < \phi_j < 1,\ j = 1, \ldots, q\right\} \\
&= \Pi_\Sigma\left\{\phi_j : \frac{1-\delta_*}{1+\delta_*} < \phi_j < \frac{1-\delta_*}{1+\delta_*}(1+t),\ j = 1, \ldots, q\right\}, \qquad (4.4)
\end{aligned}$$

where $t = 2\delta_*(1-\delta_*)^{-1}$ which by assumption lies in $(0,1)$. Noting that $H \sim \mathrm{Wishart}(q, \Sigma_0)$ and invoking Lemma 1 of [82] we have the following lower bound on the probability assigned by Π_Σ

to the event in (4.4),

$$\Pi_\Sigma\left\{\phi_j : \frac{1-\delta_*}{1+\delta_*} < \phi_j < \frac{1-\delta_*}{1+\delta_*}(1+t),\ j=1,\ldots,q\right\}$$
$$\geq b_1\left(\frac{1-\delta_*}{1+\delta_*}\right)^{b_2}\left(\frac{2\delta_*}{1-\delta_*}\right)^{b_3} e^{-b_4\left(\frac{1-\delta_*}{1+\delta_*}\right)}. \tag{4.5}$$

Hence, for δ_* small enough, (4.5) can be lower bounded by $e^{-Tn\delta^2}$ for some positive constant T and sufficiently large n. □

Recall the prior Π_B from chapter 3. If a matrix $B \in \Re^{p\times q} \sim \Pi_B$ then each column of B is a draw from $\Pi_{\mathbf{HS}}$. In the following Lemma we generalize Lemma 1 to provide a lower bound on the probability the prior Π_B assigns to Frobenius neighborhoods of $B_0 \in \Re^{p\times r_0}$. By assumption 3 the h^{th} column of B_0, $b_h \in \ell_0[s;p]$. In order to make the Frobenius neighborhood well defined, we append $(q-r_0)$ zero columns to the right of B_0 and set $B_{0*} = (B_0 \mid O^{p\times(q-r_0)})$.

Lemma 4.1.4. *Let the entries of $B_{0*} \in \Re^{p\times q}$ satisfy $\max \mid B_{0*} \mid \leq M$ for some positive constant M. Suppose B is a draw from Π_B. Define $\delta_B = \{(r_0 s\log p)/n\}^{1/2}$. Then for some positive constant K we have,*

$$\Pi_B(\|B - B_{0*}\|_F < \delta_B) \geq e^{-Kr_0 s\log p}.$$

Proof. Observe that, since by assumption 3, $r_0 = \kappa q$ for some $\kappa \in (0,1]$,

$$\begin{aligned}
\Pi_B(\|B - B_{0*}\|_F < \delta_B) &\geq \prod_{h=1}^{q}\Pi_{\mathbf{HS}}\left(\|b_h - b_{0h}\|_2 < q^{-1/2}\delta_B\right)\\
&= \prod_{h=1}^{q}\Pi_{\mathbf{HS}}\left[\|b_h - b_{0h}\|_2 < \{(\kappa s\log p)/n\}^{1/2}\right]\\
&= \prod_{h=1}^{r_0}\Pi_{\mathbf{HS}}\left[\|b_h - b_{0h}\|_2 < \{(\kappa s\log p)/n\}^{1/2}\right]\prod_{h=r_0+1}^{q}\Pi_{\mathbf{HS}}\left[\|b_h - b_{0h}\|_2 < \{(\kappa s\log p)/n\}^{1/2}\right]\\
&= \prod_{h=1}^{r_0}\Pi_{\mathbf{HS}}\left[\|b_h - b_{0h}\|_2 < \{(\kappa s\log p)/n\}^{1/2}\right]\prod_{h=r_0+1}^{q}\Pi_{\mathbf{HS}}\left[\|b_h\|_2 < \{(\kappa s\log p)/n\}^{1/2}\right].
\end{aligned}$$

From Lemma 3.5.1 we have, $\Pi_{\mathbf{HS}}[\|b_h - b_{0h}\|_2 < \{(\kappa s \log p)/n\}^{1/2}] \geq e^{-K_1 s \log p}$ for some positive K_1. Arguments along the same line of first part of Lemma 3.5.1 can also be applied to obtain that, $\Pi_{\mathbf{HS}}\left(\|b_h\|_2 < \{(\kappa s \log p)/n\}^{1/2}\right) \geq (1 - Rs/p)^p \geq e^{-K_2 \log p}$ for some positive K_2 and R as defined in Lemma 3.5.1. Combining these two lower bounds in the above display we have,

$$\Pi_B(\|B - B_{0*}\|_F < \delta_B) \geq e^{-K_1 r_0 s \log p} e^{-K_2 (q - r_0) s \log p}.$$

Since $r_0 = \kappa q$, the result follows immediately with $K = K_1 + (1/\kappa - 1)K_2$. □

Similar to Lemma 4.1.4, the following result provides a lower bound on the probability assigned to Frobenius neighborhoods of A_0 by the prior Π_A. Again we append $(q - r_0)$ columns at the right of A_0 and set $A_{0*} = (A_0 \mid O^{q \times (q - r_0)})$.

Lemma 4.1.5. *Suppose the matrix $A \sim \Pi_A$. Let $\delta_A = (qr_0/n)^{1/2}$. Then for some positive constant K we have,*

$$\Pi_A(\|A - A_{0*}\|_F < \delta_A) \geq e^{-Kqr_0}.$$

Proof. First we use vectorization to obtain $\Pi_A(\|A - A_0\|_F < \delta_A) = \Pi_A(\|a - a_0\|_2 < \delta_A)$, where $a, a_0 \in \Re^{q^2}$. Using Anderson's lemma [83] for multivariate Gaussian distributions, we then have,

$$\begin{aligned} \Pi_A(\|a - a_0\|_2 < \delta_A) &\geq e^{-\|a_0\|^2/2} pr(\|a\|_2 < \delta_A/2) \\ &= e^{-r_0/2} pr(\|a\|_2 < \delta_A/2). \end{aligned}$$

The quantity $pr(\|a\| < \delta_A/2)$ can be bounded from below as,

$$pr(\|a\|_2 < \delta_A/2) \geq \{pr(\mid a_j \mid < \delta_A/q)\}^{q^2} \geq T e^{-\delta_A^2} (\delta_A/q)^{q^2} \geq e^{-Kq^2},$$

where K is a positive constant. Since $r_0 = \kappa q$, it follows that $\Pi_A(\|A - A_{0*}\|_F < \delta_A) \geq e^{-Kqr_0}$. □

Our final result will concern the prior mass assigned to Frobenius neighborhoods of $C \in \Re^{p \times q}$.

As in chapter 3 we write Π_C for prior on $C = BA^{\mathrm{T}}$ induced from Π_B and Π_A.

Lemma 4.1.6. *Suppose C_0 satisfies assumption 3. Let $C \sim \Pi_C$ with Π_C as defined above. Define $\delta_C = \{(qr_0 + r_0 s \log p)/n\}^{1/2}$. Then for some positive constant K,*

$$\Pi_C(\|C - C_0\|_F < \delta_C) \geq e^{-K(qr_0 + r_0 s \log p)}.$$

Proof. Recall the definition of B_{0*} and A_{0*} from Lemma 4.1.4 and 4.1.5 respectively. Using the triangle inequality followed by (4.1), we have,

$$\begin{aligned}
\|C - C_0\|_F = \|BA^{\mathrm{T}} - B_{0*}A_{0*}^{\mathrm{T}}\|_F &= \|BA^{\mathrm{T}} - B_{0*}A^{\mathrm{T}} + B_{0*}A^{\mathrm{T}} - B_{0*}A_{*0}^{\mathrm{T}}\|_F \\
&= \|(B - B_{0*})A^{\mathrm{T}} + B_{0*}(A - A_{0*})^{\mathrm{T}}\|_F \\
&\leq \|(B - B_{0*})A^{\mathrm{T}}\|_F + \|B_0(A - A_{0*})^{\mathrm{T}}\|_F \\
&\leq s_{\max}(A)\|(B - B_{0*})\|_F + s_{\max}(B_0)\|A - A_{0*}\|_F.
\end{aligned}$$

From standard random matrix theory [84] it is well known that for a random matrix of dimension $m_1 \times m_2$ with independent Gaussian entries, the largest singular admits a high probability upper bound; for every $t \geq 0$, $s_{\max(A)} \leq \sqrt{m_1} + \sqrt{m_2} + t$ with probability at least $1 - 2\exp(-t^2/2)$. Also since the elements of B_0 are bounded, so is $s_{\max}(B_0)$, say by ξ. For a sufficiently large positive number L and for $A \in E = \{A : s_{\max}(A) \leq 2\sqrt{q} + L\}$ we then have,

$$\|C - C_0\|_F \leq (2\sqrt{q} + L)\,\|B - B_{0*}\|_F + \xi\,\|A - A_{0*}\|_F.$$

Thus we have, $\Pi_C(\|C - C_0\|_F < \delta_C) \geq \Pi_C\{(2\sqrt{q} + L)\|B - B_0\|_F + \xi\|A - A_0\|_F < \delta_C\}$. Since $\delta_C = \{(qr_0 + r_0 s \log p)/n\}^{1/2} \geq 2^{-1/2}\left[\{(r_0 s(\log p)/n)\}^{1/2} + (qr_0)^{1/2}\right] =$ $2^{-1/2}(\delta_B + \delta_A)$, the probability $\Pi_C(\|C - C_0\|_F < \delta_C) \geq \Pi_B(\|B - B_{0*}\|_F < K_1\delta_B)\,\Pi_A(\|A - A_{0*}\|_F < K_2\delta_A)$, where K_1 and K_2 are positive constants.

From Lemma 4.1.4 it follows that, $\Pi_B(\|B - B_0\|_F < K_1\delta_A) \geq e^{-Kr_0 s \log p}$ and from Lemma 4.1.5 we have, $\Pi_A(\|A - A_0\|_F < K_2\delta_A) \geq e^{-Tqr_0}$. Hence $\Pi_C(E \cap \{C : \|C - C_0\|_F < \delta_C\}) \geq$

$e^{-K(qr_0+r_0 s\log p)}$. Since for two sets E_1 and E_2, $pr(E_1 \cup E_2) \geq pr(E_1) + pr(E_2) - 1$, the Lemma is proved. □

4.2 Denominator in the proof of Theorem 3.5.7

Recall D_n from the proof of Theorem 3.5.7 in chapter **??**. The following lemma establishes a high probability lower bound for D_n under the true data generating distribution $P_{\eta_0}^{(n)}$.

Lemma 4.2.1. *Let* $D_n = \int e^{-\alpha r_n(\eta,\eta_0)} d\Pi_\eta$. *Let* $B_n^*(\eta_0, \widetilde{\epsilon_n}) = \{\eta = (C, \Sigma) : \int p_{\eta_0}^{(n)} (\log p_{\eta_0}^{(n)} / p_\eta^{(n)}) dY \leq n\widetilde{\epsilon_n}^{\,2}\}$. *Then, with* $P_{\eta_0}^{(n)}$*-probability at least* $1 - K/(D-1+t)n\widetilde{\epsilon}_n^2$, *we have,*

$$D_n \geq e^{-T\alpha(D+t)n\widetilde{\epsilon}_n^2},$$

for any $D > 1$ *and* $t > 0$ *for some positive constants* K.

Proof. The proof is divided into three parts. First we show that $\Pi_\eta(B_n^*\{\eta_0, \widetilde{\epsilon}_n)\} \geq e^{-Tn\widetilde{\epsilon}_n^2}$ for some positive T. Then after noting $D_n \geq \Pi_n\{B_n^*(\eta_0, \widetilde{\epsilon}_n)\}\, D_n^*$, where $D_n^* = \int \int_{B_n(\eta_0,\widetilde{\epsilon}_n)} e^{-\alpha r_n(\eta,\eta_0)} d\Pi_\eta^B$, we bound the expectation and variance of Z where Z is such that $\log D_n^* \geq Z$ and Π_η^B is the restriction of Π_η to $B_n^*(\eta_0, \widetilde{\epsilon}_n)$. Finally, we provide a high probability lower bound of D_n^*.

Step 1. From proposition 4.1.1 if $\|\Sigma - \Sigma_0\|_F < \epsilon_n$ then $(n/2)\{\mathrm{tr}(\Sigma^{-1}\Sigma_0 - \mathrm{I}_q) - \log \mid \Sigma_0\Sigma^{-1} \mid \} \leq n\widetilde{\epsilon}_n^2/2$. Also if $\|XC - XC_0\|_F^2 \leq n\epsilon_n^2$ then $\|(XC - XC_0)\Sigma^{-1}(XC - XC_0)^{\mathrm{T}}\|_F^2 \leq n\widetilde{\epsilon}_n^2/2$. Hence $B_n^*(\eta_0, \widetilde{\epsilon}_n) \supset A_n^*(\eta_0, \widetilde{\epsilon}_n)\{\eta = (C, \Sigma) : \|XC - XC_0\|_F^2 \leq n\widetilde{\epsilon}_n^2,\ \|\Sigma - \Sigma_0\|_F < \widetilde{\epsilon}_n\}$ and $\Pi_\eta\{A_n^*(\eta_0, \widetilde{\epsilon}_n)\} = \Pi_C\{C : \|XC - XC_0\|_F^2 \leq n\widetilde{\epsilon}_n^2\}\, \Pi_\Sigma\{\Sigma : \|\Sigma - \Sigma_0\|_F < \widetilde{\epsilon}_n\}$.

Using Lemma 4.1.2 we get that $\Pi_\Sigma\{\Sigma : \|\Sigma - \Sigma_0\|_F < \widetilde{\epsilon}_n\} \geq e^{-Tn\widetilde{\epsilon}_n^2}$. Next,

$$\|X(C - C_0)\|_F^2 \leq \|X\|_F^2 \|C - C_0\|_F^2 \leq q \max_{1\leq j\leq p} \|X_j\|^2 \|C - C_0\|_F^2 = nq\|C - C_0\|_F^2,$$

where the first inequality follows from the Cauchy-Schwartz inequality and the last equality holds due to assumption 5 for a sufficiently large n. Due to lemma 4.1.6 we have $\Pi_C\{C : \|C - C_0\|_F^2 \leq \epsilon_n^2\} \geq e^{-Kn\epsilon_n^2}$ for some positive constant K. Since q is fixed, for large n, nq is of the order n and ϵ_n and

$\widetilde{\epsilon}_n$ varies only by constants, therefore by Lemma 4.1.6 $\Pi_C(\|C - C_0\|_F^2 \leq \widetilde{\epsilon_n}^2/q) \geq e^{-Kn\widetilde{\epsilon}_n^2}$. Thus we get $\Pi_C\{C : \|XC - XC_0\|_F^2 \leq n\widetilde{\epsilon}_n^2\} \geq e^{-Kn\widetilde{\epsilon}_n^2}$. Finally collecting the lower bounds for the individual probabilities we have $\Pi_\eta(B_n^*\{\eta_0, \widetilde{\epsilon}_n\}) \geq e^{-Tn\widetilde{\epsilon}_n^2}$ for some positive T as desired.

Step 2. It is obvious that,

$$\begin{aligned} D_n \geq & \Pi_\eta\{B_n(\eta_0, \widetilde{\epsilon_n})\} \int_{B_n(\eta_0, \widetilde{\epsilon_n})} e^{-\alpha r_n(\eta, \eta_0)} \Pi_\eta\{B_n(\eta_0, \widetilde{\epsilon}_n)\}^{-1} d\Pi_\eta \\ = & \Pi_\eta\{B_n(\eta_0, \widetilde{\epsilon}_n)\} D_n^*, \end{aligned}$$

where $D_n^* = \int_{B_n(\eta_0, \widetilde{\epsilon_n})} e^{-\alpha r_n(\eta, \eta_0)} d\Pi_\eta^B$. Let B be a shorthand for $B_n^*(\eta_0, \widetilde{\epsilon}_n)$. By Jensen's inequality applied to the concave logarithm function we then have $\log D_n^* \geq \alpha \int_B \log p_\eta^{(n)}/p_{\eta_0}^{(n)} d\Pi_\eta^B = Z$ (say). Then,

$$E_{\eta_0}^{(n)}(Z) = -\alpha \int_B \mathrm{KL}(p_{\eta_0}^{(n)}, p_\eta^{(n)}) d\Pi_\eta^B \geq -Tn\alpha\widetilde{\epsilon}_n^2,$$

for some positive T, where the last inequality follows from the definition of B.

Next we compute the variance of Z under $P_{\eta_0}^{(n)}$.

$$\begin{aligned} \mathrm{var}_{\eta_0}(Z) &= \alpha^2 E_{\eta_0}^{(n)}\{Z - E_{\eta_0}^{(n)}(Z)\}^2 \\ &= \alpha^2 \int \left[\int_B \left\{ \log(p_{\eta_0}^{(n)}/p_\eta^{(n)}) - E_{\eta_0}^{(n)}(\log p_{\eta_0}^{(n)}/p_\eta^{(n)}) \right\} d\Pi_\eta \right]^2 p_{\eta_0}^{(n)} dY \\ &\leq \alpha^2 \int \int_B \left\{ \log(p_{\eta_0}^{(n)}/p_\eta^{(n)}) - E_{\eta_0}^{(n)}(\log p_{\eta_0}^{(n)}/p_\eta^{(n)}) \right\}^2 p_{\eta_0}^{(n)} d\Pi_\eta dY \\ &= \alpha^2 \int_B \left[\int \left\{ \log(p_{\eta_0}^{(n)}/p_\eta^{(n)}) - E_{\eta_0}^{(n)}(\log p_{\eta_0}^{(n)}/p_\eta^{(n)}) \right\}^2 p_{\eta_0}^{(n)} dY \right] d\Pi_\eta \\ &= \alpha^2 \int_B \{\mathrm{var}_{\eta_0}(Z^*)\} d\Pi_\eta, \end{aligned}$$

where $Z^* = \log p_{\eta_0}^{(n)}/p_\eta^{(n)} = \sum_{i=1}^n \log p_{\eta_0}(Y_i)/p_\eta(Y_i) = \sum_{i=1}^n Z_i^*$. Hence due to independence

$\text{var}_{\eta_0}(Z^*) = n\text{var}_{\eta_0}(Z_1^*)$. Now,

$$\text{var}_{\eta_0}(Z_1^*) = \text{var}_{\eta_0}\left\{\frac{1}{2}(Y_1 - C^{\mathrm{T}}x_1)\Sigma^{-1}(Y_1 - C^{\mathrm{T}}x_1) - \frac{1}{2}(Y_1 - C_0^{\mathrm{T}}x_1)\Sigma_0^{-1}(Y_1 - C_0^{\mathrm{T}}x_1)\right\}.$$

Let $(Y_1 - C_0^{\mathrm{T}}x_1) = u_0$ and $(Y_1 - C^{\mathrm{T}}x_1) = u_1$ and $(C_0^{\mathrm{T}}x_1 - C^{\mathrm{T}}x_1) = u$. Then,

$$\begin{aligned}
\text{var}_{\eta_0}(Z_1^*) &= \frac{1}{4}\text{var}_{\eta_0}\left(u_1^{\mathrm{T}}\Sigma^{-1}u_1 - u_0^{\mathrm{T}}\Sigma^{-1}u_0 + u_0^{\mathrm{T}}\Sigma^{-1}u_0 - u_0^{\mathrm{T}}\Sigma_0^{-1}u_0\right) \\
&= \frac{1}{4}\text{var}_{\eta_0}\left\{u^{\mathrm{T}}\Sigma^{-1}u + 2u^{\mathrm{T}}\Sigma^{-1}u_0 + u_0^{\mathrm{T}}(\Sigma^{-1} - \Sigma_0^{-1})u_0\right\} \\
&= \frac{1}{4}\text{var}_{\eta_0}\left\{2u^{\mathrm{T}}\Sigma^{-1}u_0 + u_0^{\mathrm{T}}(\Sigma^{-1} - \Sigma_0^{-1})u_0\right\} \\
&= \text{var}_{\eta_0}(u^{\mathrm{T}}\Sigma^{-1}u_0) + \frac{1}{4}\text{var}_{\eta_0}\left\{u_0^{\mathrm{T}}(\Sigma^{-1} - \Sigma_0^{-1})u_0\right\} + \frac{1}{2}\text{cov}_{\eta_0}\left\{u^{\mathrm{T}}\Sigma^{-1}u_0, u_0^{\mathrm{T}}(\Sigma^{-1} - \Sigma_0^{-1})u_0\right\} \\
&= u^{\mathrm{T}}\Sigma^{-1}\Sigma_0\Sigma^{-1}u + \frac{1}{2}\text{tr}\{(\Sigma^{-1}\Sigma_0 - \mathrm{I}_q)^2\} \\
&\leq \|C_0^{\mathrm{T}}x_1 - C^{\mathrm{T}}x_1\|_2^2\,\|\Sigma_0\|_2\|\Sigma^{-1}\|_2^2 + \frac{1}{2}\|\Sigma^{-1/2}\Sigma_0\Sigma^{-1/2} - \mathrm{I}_q\|_F^2 \\
&\leq \|C_0^{\mathrm{T}}x_1 - C^{\mathrm{T}}x_1\|_2^2\,\|\Sigma_0\|_2\|\Sigma^{-1}\|_2^2 + \frac{1}{2}\|\Sigma_0 - \Sigma\|_F^2\|\Sigma^{-1}\|_2^2 \\
&\leq \frac{4\|\Sigma_0\|_2}{s_{\min}^2(\Sigma_0)}\|C_0^{\mathrm{T}}x_1 - C^{\mathrm{T}}x_1\|_2^2 + \frac{1}{2}\|\Sigma^{-1}\|_2^2\|\Sigma - \Sigma_0\|_F^2 \\
&\leq \frac{4\|\Sigma_0\|_2}{s_{\min}^2(\Sigma_0)}\|C_0^{\mathrm{T}}x_1 - C^{\mathrm{T}}x_1\|_2^2 + \frac{2}{s_{\min}^2(\Sigma_0)}\|\Sigma - \Sigma_0\|_F^2,
\end{aligned}$$

where we have used the lower bound on $s_{\min}(\Sigma)$ from lemma 4.1.2. Therefore, $\text{var}_{\eta_0}(Z^*) \leq \alpha^2\frac{4\|\Sigma_0\|_2}{s_{\min}^2(\Sigma_0)}\|XC - XC_0\|_F^2 + \alpha^2\frac{2n}{s_{\min}^2(\Sigma_0)}\|\Sigma - \Sigma_0\|_F^2$. Since $B \supset A_n^*(\eta_0, \widetilde{\epsilon}_n)$ from step 1, and $\Pi\{A_n^*(\eta_0, \widetilde{\epsilon}_n)\}$, we finally get $\text{var}_{\eta_0}(Z) \leq K\alpha^2 n\widetilde{\epsilon}_n^2$ for some positive constant K.

Step 3. For any $D > 1$ and $t > 0$, by Chebyshev's inequality

$$\begin{aligned}
P_{\eta_0}^{(n)}\{Z \leq -T\alpha(D+t)n\widetilde{\epsilon}_n^2\} &= P_{\eta_0}^{(n)}\{Z \leq -T\alpha(D-1+t+1)n\widetilde{\epsilon}_n^2\} \\
&= P_{\eta_0}^{(n)}\{Z - (-T\alpha n\widetilde{\epsilon}_n^2) \leq -T\alpha(D-1+t)n\widetilde{\epsilon}_n^2\} \\
&\leq \frac{\text{var}_{\eta_0}(Z)}{\{T\alpha(D-1+t)n\widetilde{\epsilon}_n^2\}^2} \\
&\leq \frac{K}{T^2(D-1+t)^2 n\widetilde{\epsilon}_n^2},
\end{aligned}$$

where we have used the fact that $\mathrm{var}_{\eta_0}(Z) \leq Kn\widetilde{\epsilon}_n^2$. Thus we get with $P_{\eta_0}^{(n)}$-probability at least $1 - K/(D-1+t)^2 n\widetilde{\epsilon}_n^2$,

$$\log D_n^* \geq -T\alpha(D+t)n\widetilde{\epsilon}_n^2 \Leftrightarrow D_n^* \geq e^{-T\alpha(D+t)n\widetilde{\epsilon}_n^2},$$

for some positive constant K. Since $D_n \geq \Pi_\eta(B) D_n^*$ and $\Pi_\eta(B) \geq e^{-Tn\widetilde{\epsilon}_n^2} \geq e^{-T\alpha(D+t)n\widetilde{\epsilon}_n^2}$ $(D > 1)$, we finally obtain,

$$D_n \geq e^{-T\alpha(D+t)n\widetilde{\epsilon}_n^2},$$

with $P_{\eta_0}^{(n)}$-probability at least $1 - K/(D-1+t)^2 n\widetilde{\epsilon}_n^2$ for some positive constant K. □

5. CONVERGENCE RATES OF FRACTIONAL HORSESHOE POSTERIORS IN HIGH-DIMENSIONS

5.1 Introduction

Nowadays data arising from various scientific fields like Genetics, Physics, Bioinformatics, Psychology nowadays are inherently high-dimensional in nature. For example, in Genetics it is of interest to study gene behaviors for thousands of genes and typically measurements are only available for a few hundred patients. In such scenarios it is frequently assumed that the high-dimensional parameter possesses a low-dimensional structure. One popular way of providing the parameter with a low dimensional structure is by assuming it is 'sprase', wherein most of the elements of the parameters are exactly equal to or very close to zero. In the literature, the non-zero component of the parametr is known as the signal and zero or small part is known as noise. The presence of high-dimensional parameters have motivated a series of works based on thresholding or regularization in common statistical problems of regression [5, 6, 7, 72, 85], covariance estimation [86, 87, 88] among others. While there is now a plethora of frequentist methods that rapidly produce point estimates in such problems, the uncertainty associated with such estimates, a decidedly non-trivial problem, has only gained limited attraction until recently [3, 89, 90].

On the other hand, by interpreting regularization as priors over parameter spaces, Bayesian methods have the innate ability to produce a posterior distribution over the space of parameters instead of just providing a point estimate. There is now a rich body of literature on various possible choices of such priors, commonly referred to as shrinkage priors - see [22, 33, 1, 23] and references therein. A unifying theme of these priors lies in the fact that these priors can mostly be expressed as global-local scale mixture of Gaussian distributions [20], wherein a single global scale parameter controls the overall shrinkage in the estimation procedure and local scale parameters allow for capturing signals with large magnitude. The Gaussian representation also allows for straightforward Markov chain Monte Carlo sampling from the posterior distribution [65]. For

a class of shrinkage priors [91, 92] establishted rapid mixing and convergence at a geometric rate of these sampling procedures. More recently, [65] developed an exact sampling scheme for sampling from a high-dimensional Gaussian distribution frequently encountered in the computational process involving shrinkage priors. Further computationally tractable algorithms are discussed and explored in [93]. The choice of the mixing distributions for the scale distribution essentially determines the priors ability to recover sparse parameters. Indeed, [83] proved sub-optimal risk properties of several routinely used mixing distributions such as the Inverse Gamma distribution. [20] prescribed a heavy-tailed prior on the local scale parameters, while the prior on the global scale parameter should have sufficient mass near zero.

In a seminal work [94] obtained the minimax lower bound for estimating a sparse vector of a given dimension and for $\mathcal{L}^q$ risks. Several frequentist estimators have been proven to be optimal in the sense of attaining the minimax risk for models ranging from Gaussian regression, generalized linear models to graphical models [8]. Analyzing the posterior obtained from shrinkage priors requires obtaining the convergence rate of the posterior to a point mass at the true parameter. We say a prior is optimal for sparse estimation if this convergence rate coincides with the minimax rate for the sparse class.

[77] provided sufficient conditions for posterior convergence rates in a general statistical model where the number of parameters are allowed to grow with the sample size or the parameter space itself can be infinite. The conditions include a prior concentration result and careful construction of sieves with a control on their entropy. Furthermore, the prior probability assigned to the complement of the sieves is required to be exponentially small. Intuitively, the prior mass condition is used to control the marginal distribution of the parameter under the specified prior and the sieve condition guarantees existence of exponentially consistent tests which are then used to control the posterior probability. In the canonical sparse normal means problem [23] proved the posterior contracts at the minimax rate for the Dirichlet-Laplace prior using the aforementioned tools. Their results can be extended to more sophisticated statistical models analogously, see [37] for a treatment involving high-dimensional factor models. However, for polynomial tailed priors such

as the horseshoe [1] the exponentially decaying prior mass assigned to the complement of sieves is difficult to verify. This technical difficulty has restricted its theoretical treatment to the normal means model where the conditionally independent Gaussian scale mixture representation is crucially exploited to achieve a tractable expression of coordinatewise posterior means. In particular, [25] showed that the horseshoe posterior contracts at the minimax rate for a suitably chosen global scale parameter, while [95] established similar results with an empirical Bayes and full Bayes approach by specifying a prior on the global scale parameter. In their work [25], the authors interpret the global scale parameter as the proportion of signal variables upto logarithmic factors. See also [26] for a study of the horseshoe prior from a testing perspective.

As noted earlier, the conditional independence structure is usually not valid for many practical statistical models. For instance, in a linear regression problem, the design matrix induces correlation in the posterior distribution of the coefficient vector. Thus results from [25, 95] cannot be readily adopted to more complex models although empirical studies have shown the superior performance of the horseshoe across different settings [65].

In this article, we adopt the fractional posterior framework [28]. Here a fractional power of the likelihood function is combined with the prior using Bayes theorem to obtain a fractional posterior. [28] show only a prior mass condition is sufficient to prove convergence rate of the fractional posterior. Simulation results in Appendix A and Theorem 3.5.9 show in Gaussian models, as far as estimation and prediction is concerned, fractional and usual posterior are almost indistinguishable. Here we focus on analyzing fractional posterior obtained from horseshoe related priors in regression and factor models.

5.2 Notation

Let $l_0[s;p] = \{\theta \in \Re^p : \#(1 \leq j \leq p : \theta_j \neq 0) \leq s\}$ be the subset of $\Re^p$ with at most s non-zero entries. For a vector $\theta \in \Re^p$, let $S_\theta = \{1 \leq j \leq p_n : \theta_j \neq 0\}$, called the support of θ. We write $|| \cdot ||_2$ to denote the l_2 norm and $|| \cdot ||_1$ denote the l_1 norm. The Kullback-Liebler divergence between two distributions P and Q is written as $D(P \mid\mid Q)$. We will use I_k to denote the k-dimensional identity matrix.

5.3 The horseshoe prior

The horseshoe prior was originally introduced in [1] as a sparse prior over $\theta \in \Re^p$ in the model

$$y_j = \theta_j + \epsilon_j,\ \epsilon_j \overset{iid}{\sim} \mathrm{N}(0,1), \quad j = 1, \ldots, p, \tag{5.1}$$

when the true mean $\theta_0 \in l_0[s;p]$. A draw $\eta \in \Re^p$ from the horseshoe prior can be hierarchically represented as,

$$\eta_j \mid \lambda_j, \tau \overset{indp.}{\sim} \mathrm{N}(0, \lambda_j^2\tau^2) \tag{5.2}$$

$$\lambda_j \overset{indp.}{\sim} \mathrm{C}^+(0,1) \quad j = 1, \ldots, p \tag{5.3}$$

$$\tau \sim g, \tag{5.4}$$

where $\mathrm{C}^+(0,1)$ is the standard Half-Cauchy distribution on the positive real line $\Re^+$ with density function $f(x) \propto (1+x^2)^{-1}\mathbf{1}_{(0,\infty)}$ and g is the density of a distribution on $\Re^+$ with respect to the Lebesgue measure. Let the joint prior thus induced on η be $\Pi_{\mathbf{HS}}$. The global scale parameter τ controls the overall shrinkage while the local scale parameters λ_j allow to capture the coordinates with large magnitude. Choices of g are discussed below. First, we give a brief overview of the properties of the hierarchy (5.2)-(5.4).

Fix $\tau = 1$. Due to independence the conditional posterior mean of θ_j in model (5.1) with *apriori* $\theta \sim \Pi_{\mathbf{HS}}$, is given by,

$$E(\theta_j \mid y_j, \lambda_j) = \left(1 - \frac{1}{1+\lambda_j^2}\right) y_j.$$

Evidently, the amount of shrinkage for each coordinate j is then controlled by the variable $\omega_j = (1+\lambda_j^2)^{-1}$. When $\lambda_j \sim \mathrm{C}^+(0,1)$, the induced prior on ω_j is a Beta(1/2,1/2) distribution [1] which has its peak near 0 and 1 and a plateau in between. Thus the prior $\Pi_{\mathbf{HS}}$ has the ability to shrink noise coordinates heavily towards 0 with ω_j close to 1 yet at the same time retain signal coordinates

with small ω_j values. Furthermore, because there are no hyperparameters involved, the horseshoe prior has enjoyed wide popularity among the class of global-local shrinkage priors.

When estimating θ_0, previous theoretical works [25, 95] have interpreted τ to be the proportion of non-zero components in θ_0 upto logarithmic factors. Indeed, small values of τ will result in sparser draws of η, see figure 1 of [25]. In this article we set g to be the Half-Cauchy density [1]. However, simple modifications with an $\mathrm{Unif}(0,1)$ prior on τ lead to similar conclusions.

For $\theta_0 \in l_0[s;p]$ with $s = o(p)$, [94] provided the minimax squared error loss as,

$$\inf_{\hat{\theta}} \sup_{\theta \in l_0[s;p]} E_{\theta_0} \mid\mid \hat{\theta} - \theta_0 \mid\mid_2^2 \asymp s \log(p/s), \tag{5.5}$$

where the infimum is taken over all possible estimators $\hat{\theta}$. In what follows, we will be working with the approximate minimax rate of $s \log p$. Next, consider n independent copies $(Y_i)_{i=1}^n$ of the p-dimensional vector $Y_i = (y_{ij})_{j=1}^p$. We work in a high-dimensional regime where the dimension p is allowed to grow with the sample size n. Further, the sparsity s of θ_0 may also change with n. In terms of notation, we will use p_n and s_n to make their dependence explicit on the sample size. Bayesian asymptotic results rely heavily on the prior mass assigned to neighborhoods of the true parameter [96, 77, 28]. Given a sparse vector $\theta_0 \in l_0[s_n; p_n]$, our first result is a non-asymptotic concentration result of $\Pi_{\mathbf{HS}}$ on shrinking l_2 neighborhoods of θ_0.

Theorem 5.3.1. *Suppose n, p_n and s_n is such that $s_n/p_n \leq 1/2$ and $\log p_n = o(n)$ and $s_n \log p_n = o(n)$. Let $\theta_0 \in l_0[s_n; p_n]$ be such that for all $j \in S_{\theta_0}$ we have $\frac{1}{2}\sqrt{\frac{\log p_n}{n}} \leq \theta_{0j} \leq p_n^m$, for some positive constant m. Then for some positive constant C,*

$$\Pi_{\boldsymbol{HS}}(\mid\mid \theta - \theta_0 \mid\mid_2 \leq \delta_n) \geq e^{-Cn\delta_n^2} = e^{-Cs_n \log p_n}.$$

Proof. Using the conditional formulation of prior (5.2)-(5.3) we have,

$$\begin{aligned}\Pi_{\mathbf{HS}}(\|\theta-\theta_0\|_2<\delta_n) &= \int_\tau pr(\|\theta-\theta_0\|_2<\delta_n \mid \tau)g(\tau)d\tau \\ &\geq \int_{I_{\tau_*}} pr(\|\theta-\theta_0\|_2<\delta_n \mid \tau)g(\tau)d\tau,\end{aligned} \tag{5.6}$$

where $I_{\tau_*}=[\tau_*/2,\tau_*]$ with $\tau_*=(s_n/p_n)^{3/2}(\log p_n/n)^{1/2}$. Let $S=\{1\leq j\leq p_n : \theta_{0j}\neq 0\}$. We first provide a lower bound of the conditional probability $pr(\|\theta-\theta_0\|_2<\delta_n \mid \tau\in I_{\tau_*})$ by dividing it into two terms. For $\tau\in I_{\tau^*}$ we have,

$$\begin{aligned}pr(\|\theta-\theta_0\|<\delta_n \mid \tau) &\geq pr\left(\|\theta_S-\theta_{0S}\|_2<\frac{\delta_n}{2}\mid\tau\right)pr\left(\|\theta_{S^c}\|_2<\frac{\delta_n}{2}\mid\tau\right) \\ &\geq \prod_{j\in S} pr\left(|\theta_j-\theta_{0j}|<\frac{\delta_n}{2\sqrt{s_n}}\mid\tau\right)\prod_{j\in S^c} pr\left(|\theta_j|<\frac{\delta_n}{2\sqrt{p_n}}\mid\tau\right)\end{aligned} \tag{5.7}$$

Now we will provide lower bounds for each of the terms $pr\{\mid \theta_j-\theta_{0j}\mid < \delta_n/(2s_n^{1/2})\}$ for any $j\in S$ and $pr\{\mid\theta_j\mid<\delta_n/(2p_n^{1/2})\}$ for any $j\in S^c$ and for $\tau\in I_{\tau_*}$.

We first consider $pr\{\mid\theta_j\mid<\delta_n/(2p_n^{1/2})\mid\tau\}$ with $\tau\in I_{\tau_*}$. Since given τ and λ, $\theta_j\sim \mathrm{N}(0,\lambda_j^2\tau^2)$, we use Chernoff type bounds for a Gaussian random variable to obtain,

$$pr\left(\mid\theta_j\mid>\delta_n/2p_n^{1/2}\mid\lambda_j,\tau\right)\leq 2e^{-\frac{\delta_n^2}{8\lambda_j^2\tau^2}}\leq 2e^{-\frac{\delta^2}{8\lambda_j^2\tau_*^2}}=2e^{-\frac{p_n^2}{8s_n^2\lambda_j^2}}.$$

Hence,

$$\begin{aligned}pr(\mid\theta_j\mid<\delta_n/2p_n^{1/2}\mid\tau) &= \int_{\lambda_j} pr(\mid\theta_j\mid<\delta_n/2p_n^{1/2}\mid\lambda_j,\tau)f(\lambda_j)d\lambda_j \\ &\geq \int_{\lambda_j}\left\{1-2\exp\left(-\frac{p_n^2}{8s_n^2\lambda_j^2}\right)\right\}f(\lambda_j)d\lambda_j \\ &= 1-\frac{4}{\pi}\int_{\lambda_j}\exp\left(-\frac{p_n^2}{8s_n^2\lambda_j^2}\right)(1+\lambda_j^2)^{-1}d\lambda_j = 1-\frac{4}{\pi}\mathrm{I},\end{aligned}$$

where $\mathrm{I} = \int_{\lambda_j} \exp\left(-\frac{p_n^2}{8s_n^2\lambda_j^2}\right)(1+\lambda_j^2)^{-1}d\lambda_j$. We then bound the integrand from above as follows,

$$\begin{aligned} I = \int_{\lambda_j} \exp\left(-\frac{p_n^2}{8s_n^2\lambda_j^2}\right)(1+\lambda_j^2)^{-1}d\lambda_j &\leq \int_{\lambda_j} \exp\left(-\frac{p_n^2}{8s_n^2\lambda_j^2}\right)\lambda_j^{-2}d\lambda_j \\ &= \frac{1}{2}\int_0^\infty z^{-1/2}\exp\left(-\frac{p_n^2 z}{8s_n^2}\right)dz, \quad z = 1/\lambda^2, \\ &= \frac{\Gamma(1/2)}{2(p_n^2/8s_n^2)^{1/2}} = \frac{s_n\sqrt{2\pi}}{p_n}. \end{aligned}$$

Thus for $\tau \in I_{\tau_*}$, $pr(\mid \theta_j \mid < \delta_n/2p_n^{1/2} \mid \tau) \geq 1 - R\frac{s_n}{p_n}$, where $R = (32/\pi)^{1/2}$.

For $pr(\mid \theta_j - \theta_{0j} \mid < \delta_0 | \tau)$ with $\tau \in I_{\tau_*}$, we have, letting $\delta_0 = \delta_n/2\sqrt{s_n}$

$$\begin{aligned} pr(\mid \theta_j - \theta_{0j} \mid < \delta_0 \mid \tau) &= (2/\pi^3)^{1/2}\int_{\lambda_j}\int_{|\theta_j-\theta_0|<\delta_0} \exp\{-\theta_j^2/(2\lambda_j^2\tau^2)\}\frac{1}{\lambda_j\tau(1+\lambda_j^2)}d\lambda_j d\theta_j \\ &\geq (2/\pi^3)^{1/2}\int_{|\theta_j-\theta_0|<\delta_0}\int_{\theta_{0j}/\tau}^{2\theta_{0j}/\tau} \exp\{-\theta_j^2/(2\lambda_j^2\tau^2)\}\frac{1}{\lambda_j\tau(1+\lambda_j^2)}d\lambda_j d\theta_j \\ &\geq (2/\pi^3)^{1/2}\int_{|\theta_j-\theta_0|<\delta_0} \exp\left(-\frac{\theta_j^2}{2\theta_{0j}^2}\right)\frac{1}{2\theta_{0j}}\int_{\theta_{0j}/\tau}^{2\theta_{0j}/\tau}\frac{1}{1+\lambda_j^2}d\lambda_j d\theta_j, \end{aligned}$$

since for $\lambda_j \in [\theta_{0j}/\tau, 2\theta_{0j}/\tau]$, $1/\lambda_j\tau \geq 1/(2\theta_{0j})$ and $\exp\{-\theta_j^2/(2\lambda_j^2\tau^2)\} \geq \exp(-\theta_j^2/2\theta_{0j}^2)$. Moreover, in the interval $[\theta_{0j}/\tau, 2\theta_{0j}/\tau]$, $(1+\lambda_j^2)^{-1} \geq (1+4\theta_{0j}^2/\tau^2)$. Thus,

$$\begin{aligned} pr(\mid \theta_j - \theta_{0j} \mid < \delta_0 \mid \tau) &\geq (2/\pi^3)^{1/2}\frac{1}{2\theta_{0j}}\frac{\theta_{0j}}{\tau}\frac{1}{1+4\theta_{0j}^2/\tau^2}\int_{|\theta_j-\theta_0|<\delta_0}\exp\left(-\frac{\theta_j^2}{2\theta_{0j}^2}\right)d\theta_j \\ &= (2/\pi^3)^{1/2}\frac{\tau}{4\theta_{0j}^2+\tau^2}\int_{|\theta_j-\theta_0|<\delta_0}\exp\left(-\frac{\theta_j^2}{2\theta_{0j}^2}\right)d\theta_j. \end{aligned}$$

By the assumption that $\theta_{0j} > \delta_0$, it follows that $\theta_j/\theta_{0j} < 2$ when $\mid \theta_j - \theta_{0j} \mid < \delta_0$. Furthermore, since for every $\tau \in I_{\tau_*}$, $\tau^2 < \tau_*^2 < \delta_0^2$ because $s_n/p_n \leq 1/2$, we get $4\theta_{0j}^2 + \tau^2 < 5\theta_{0j}^2$. Putting

together these bounds in the above display and using $\theta_{0j}^2 \leq p_n^m$ we get,

$$pr(\mid \theta_j - \theta_{0j} \mid < \delta_0 \mid \tau) \geq (2/\pi^3)^{1/2} e^{-2} \frac{1}{5\theta_{0j}^2} \tau\delta_0 = K(\theta_{0j})^{-2} \left(\frac{s_n}{p_n}\right)^{3/2} \frac{\log p_n}{n} \geq K p_n^{-(5/2+m)}, \tag{5.8}$$

assuming $s_n \log p_n \geq 1$ where $K = \frac{1}{5}(2/\pi^3)^{1/2} e^{-2}$. Substituting these bounds in (5.7), we have for $\tau \in I_{\tau_*}$

$$pr(\|\theta - \theta_0\| < \delta \mid \tau) \geq \left(1 - R\frac{s_n}{p_n}\right)^{p_n - s_n} K e^{-\frac{5}{2} s_n \log p_n} \geq e^{-Cs \log p}, \tag{5.9}$$

for some positive constant C. Then the unconditional prior mass $\Pi_{\mathbf{HS}}(\mid\mid \theta - \theta_0 \mid\mid_2 < \delta_n)$ admits the following bound,

$$\begin{aligned} \Pi_{\mathbf{HS}}(\mid\mid \theta - \theta_0 \mid\mid_2 < \delta_n) \geq e^{-Cs_n \log p_n} \int_{I_{\tau_*}} g(\tau) d\tau &\geq e^{-Cs_n \log p_n} \frac{\tau_*}{2(1+\tau_*^2)} \\ &\geq e^{-Cs_n \log p_n} \frac{\tau_*}{4} \geq e^{-Cs_n \log p_n} \end{aligned}$$

for some positive constant C.

□

A version of theorem 5.3.1 appeared in [97] for bounded uniformly θ_0. One key feature of our proof is in choosing the global shrinkage parameter τ carefully so that sufficiently large number of elements of a draw from $\Pi_{\mathbf{HS}}$ are shrunken towards zero. This happens if τ is restricted within a constant interval of $(s_n/p_n)^{3/2}\{(\log p_n)/n\}^{1/2}$. The lower threshold $2^{-1}\sqrt{(\log p_n)/n}$ represents a certain minimum signal strength below which the prior is unable to distinguish it from a noise variable. Such minimum thresholds exist in the literature [25] who call it the *detection boundary*. Theorem 5.3.1 will be the key building block in our subsequent studies of several high-dimensional models.

5.4 Fractional posteriors

In this section we introduce fractional posteriors and related concepts developed recently in [28]. Suppose we observe n independent but not necessarily identically distributed data points $X^{(n)} = (X_1, \ldots, X_n)$ from a distribution $\mathbb{P}_\theta$ where $\theta \in \Theta$ is the parameter of interest. We assume p_θ to be the density of $\mathbb{P}_\theta$ with respect to some dominating measure μ. Let $(\mathbb{P}_\theta^{(n)}, p_\theta^{(n)})$ be the corresponding product measure and density of the n random variables with respect to $\mu^{(n)}$. For $\alpha \in (0, 1)$, the α-fractional likelihood is defined as,

$$L_{n,\alpha}(\theta) = \left\{ p_\theta^{(n)} \right\}^\alpha, \tag{5.10}$$

the usual likelihood raised to the power α. Suppose Π_n is a prior distribution on Θ. Then the α-fractional posterior $\Pi_{n,\alpha}$ is defined by substituting the α-fractional likelihood in equation (5.10) in the usual Bayes theorem. For $B \in \mathcal{B}$, the σ-field of Θ,

$$\Pi_{n,\alpha}(B \mid X^{(n)}) = \frac{\int_B L_{n,\alpha}(\theta)\Pi_n(d\theta)}{\int_\Theta L_{n,\alpha}(\theta)\Pi_n(d\theta)}. \tag{5.11}$$

We assume θ_0 to be true value of the parameter and also assume that the model is correctly specified, i.e. $\theta_0 \in \Theta$. To measure the recovery of θ_0 we will use the Rényi divergence $D_\alpha(\theta, \theta_0) = D_\alpha(p_\theta^{(n)}, p_{\theta_0}^{(n)})$ of order $\alpha \in (0, 1)$ which for two densities q_1 and q_2 with respect to some common dominating measure ν is defined as,

$$D_\alpha(q_1, q_2) = \frac{1}{\alpha - 1} \log \int q_1^\alpha q_2^{(1-\alpha)} d\nu. \tag{5.12}$$

[28] show that the rate of contraction of $\Pi_{n,\alpha}$ at θ_0 is determined by the concentration of Π_n around certain Kullback-Leibler type neighborhoods of θ. We state here the concentration result of $\Pi_{n,\alpha}$ from [28] for the sake of completeness. Define for every positive ϵ the set,

$$B_n(\theta_0, \epsilon) = \left\{ \theta \in \Theta : \int p_{\theta_0}^{(n)} \log \frac{p_{\theta_0}^{(n)}}{p_\theta^{(n)}} d\mu^{(n)} \leq n\epsilon^2 \,, \int p_{\theta_0}^{(n)} \log^2 \frac{p_{\theta_0}^{(n)}}{p_\theta^{(n)}} d\mu^{(n)} \leq n\epsilon^2 \right\}. \tag{5.13}$$

The following theorem provides a *nonasymptotic* upper bound to the probability assigned by $\Pi_{n,\alpha}$ to D_α neighborhoods of θ_0.

Theorem 5.4.1 ([28]). *Fix $\alpha \in (0,1)$. Suppose ϵ_n is such that $n\epsilon_n^2 \geq 2$. Let Π_n satisfies*

$$\Pi_n(B_n(\theta_0, \epsilon_n)) \geq e^{-n\epsilon_n^2}.$$

Then for any $D \geq 2$ and $t > 0$,

$$\Pi_{n,\alpha}\left(\theta \in \Theta : \frac{1}{n} D_\alpha(\theta, \theta_0) \geq \frac{D+3t}{1-\alpha}\epsilon_n^2 \big| X^{(n)}\right) \leq e^{-tn\epsilon_n^2}$$

holds with $\mathbb{P}_{\theta_0}^{(n)}$-probability at least $1 - 2/(D-1+t)^2 n\epsilon_n^2$.

A simple application of theorem 5.3.1 and 5.4.1 gives the following fractional posterior concentration result for model (5.1) with $\theta_0 \in l_0[s_n; p_n]$ and *apriori* $\theta \sim \Pi_{\mathbf{HS}}$. With $\alpha \in (0,1)$ and for n independent copies $Y^{(n)} = (Y_1, \ldots, Y_n)$ where each Y_i is generated according to model (5.1), the α-fractional posterior is given by,

$$\Pi_{n,\alpha}(\theta \mid Y^{(n)}) \propto \exp\left(-\frac{\alpha}{2}\sum_{i=1}^{n} || Y_i - \theta ||_2^2\right)\Pi_{\mathbf{HS}}(d\theta) \tag{5.14}$$

Theorem 5.4.2. *Consider model (5.1). Let the true parameter $\theta_0 \in l_0[s_n; p_n]$ be such that for $j \in S_{\theta_0}$, $\frac{1}{2}\sqrt{\frac{\log p_n}{n}} \leq \theta_{0j} \leq p_n^m$ and s_n, p_n and n satisfy the conditions of theorem 5.3.1. Fix $\alpha \in (0,1)$. Then the α-fractional posterior $\Pi_{n,\alpha}$ defined in (5.14) satisfies,*

$$\Pi_{n,\alpha}\left(\theta \in \Theta : \frac{\alpha}{2} || \theta - \theta_0 ||_2^2 \geq \frac{D+3t}{1-\alpha} s_n \log p_n\right) \leq e^{-ts_n \log p_n}$$

for any $D \geq 2$ and $t > 0$ with $\mathbb{P}_{\theta_0}^{(n)}$-probability at least $1 - 2/\{(D-1+t)^2 s_n \log p_n\}$.

Proof. Let $\epsilon_n = \{(s_n \log p_n)/n\}^{1/2}$. Since $D\left(\mathrm{N}(\theta_0, \mathrm{I}_{p_n}) \,||\, \mathrm{N}(\theta, \mathrm{I}_{p_n})\right) = 2^{-1} \,||\, \theta - \theta_0 \,||_2^2$, the set $B_n(\theta_0, \epsilon_n) \supset \{\theta : || \theta - \theta_0 ||_2^2 \leq \epsilon_n^2\}$. From theorem 5.3.1 we then get $\Pi_{\mathbf{HS}}\{B_n(\theta_0, \epsilon_n)\} \geq$

$\Pi_{\mathbf{HS}}(\theta :|| \theta - \theta_0 ||_2^2 \leq \epsilon_n^2) \geq e^{-Cn\epsilon_n^2}$. Therefore the theorem follows from theorem 5.4.1 noting that $D_\alpha(\theta, \theta_0) = \frac{n\alpha}{2} || \theta - \theta_0 ||_2^2$. □

Given $\Pi_{n,\alpha}$ satisfies theorem 5.4.1 with rate ϵ_n, [28, Corollary 3.3] prove that with $\mathbb{P}_{\theta_0}^{(n)}$-probability at least $1 - C_1/(n\epsilon_n^2)$, $\int n^{-1} D_\alpha(\theta, \theta_0) \Pi_{n,\alpha}(d\theta) \leq C_2(1-\alpha)^{-1}\epsilon_n$ for positive constant C_1, C_2 independent of α. As a result, we get from theorem 5.4.2 that $\int || \theta - \theta_0 ||_2^2 \Pi_{n,\alpha}(d\theta) \leq C_2\{\alpha(1-\alpha)\}^{-1} s_n \log p_n$ with high $\mathbb{P}_{\theta_0}^{(n)}$-probability. Then using the convexity of the l_2-norm followed by Jensen's inequality it immediately follows that the α-fractional posterior mean $\bar{\theta}_\alpha = \int \theta \Pi_{n,\alpha}(d\theta)$ is a rate optimal estimator in the minimax sense: $|| \bar{\theta}_\alpha - \theta_0 ||_2^2 \leq C_2\{\alpha(1-\alpha)\}^{-1} s_n \log p_n$. [25] proved minimax optimality under l_2 loss of the usual posterior mean obtained from a horseshoe prior in the normal means model.

5.5 High-dimensional sparse linear regression

While the normal means model in equation (5.1) certainly provides insight into the operating characteristics of the prior $\Pi_{\mathbf{HS}}$, practical statistical models seldom admit such simple parameterization. We now consider the Gaussian linear regression model defined as,

$$y_n = X_n \beta_n + \epsilon_n,\ \epsilon_n \sim \mathrm{N}_n(0, \sigma_n^2 \mathrm{I}_n), \tag{5.15}$$

where $y_n \in \Re^n$ and X_n is a $n \times p_n$ dimensional deterministic matrix. We focus on the high-dimensional case where $n \leq p_n$ and allow p_n to grow with n exponentially fast; $\log p_n \leq n^\gamma$, for some $\gamma \in (0, 1)$. Our object of interest will be the contraction rate of fractional posteriors $\Pi_{n,\alpha}$ when the true parameter $\beta_{0n} \in l_0[s_n; p_n]$ and β_n is endowed with the prior $\Pi_{\mathbf{HS}}$. Fix $\alpha \in (0, 1)$. Since the additive error in model (5.15) is assumed to Gaussian, for any Borel set $B \in \Re^p$, $\Pi_{n,\alpha}(B \mid y_n)$ is,

$$\Pi_{n,\alpha}(B \mid y_n) = \frac{\int_B e^{-\frac{\alpha}{2}||y - X_n\beta_n||_2^2} \Pi_{\mathbf{HS}}(d\beta_n)}{\int e^{-\frac{\alpha}{2}||y - X_n\beta_n||_2^2} \Pi_{\mathbf{HS}}(d\beta_n)}. \tag{5.16}$$

In the high-dimensional setting $n \leq p_n$, the structure of the design matrix plays a central role in evaluating the performance of a statistical procedure [8]. However, for the relatively easier problem

of prediction optimality is achieved by imposing the following minimal assumption on X_n.

Assumption 5.5.1. *The entries of X_n satisfy $\|X\| = \max_{1\le j\le p_n} \|X_j\|_2^2 \asymp n$. This holds if the entries are standardized, a common practice. Also if the entries are generated from a distribution which is independent of n and p_n, for example from a $\mathrm{N}(0,1)$ distribution.*

We now generalize Theorem 5.4.1 to model (5.15).

Theorem 5.5.2. *(Prediction recovery) Fix $\alpha \in (0,1)$. Assume that $\beta_{0n} \in l_0[s_n; p_n]$ satisfies the conditions in theorem 5.4.2 and X_n satisfies assumption 5.5.1. Then for any $D \ge 2$ and $t > 0$,*

$$\Pi_{n,\alpha}\left(\beta \in \Re^p : \frac{\alpha}{2} \mid\mid X_n\beta - X_n\beta_{0n} \mid\mid_2^2 \ge \frac{D+3t}{1-\alpha} s_n \log p_n\right) \le e^{-ts_n \log p_n},$$

with $\mathbb{P}^{(n)}_{\beta_{0n}}$-probability at least $1 - 2/\{(D-1+t)^2 s_n \log p_n\}$.

The proof of Theorem 5.5.2 follows along the same lines as that of Theorem 5.4.2. Assumption 5.5.1 is made to ensure sufficient prior concentration around β_0 propagates to prior concentration around $X_n\beta_{0n}$. The precise arguments are collected below.

Proof. For model (5.15) and for every $\epsilon > 0$ we have $B_n \supset \{\beta : \|X\beta - X\beta_0\|_2^2 < n\epsilon^2\}$. Now from Theorem 5.3.1 we have $\Pi_{\mathbf{HS}}\{\beta : \|\beta - \beta_0\|_2^2 < (s_n \log p_n)/n\} \ge e^{-Cs_n \log p_n}$. Since $\|X\beta - X\beta_0\|_2^2 \le \|X\| \, \|\beta - \beta_0\|_2^2$ and $\|X\| \asymp n$, the result immediately follows from 5.4.1. □

Connections with previous work: The most comparable result to Theorem 5.5.2 for the prior $\Pi_{\mathbf{HS}}$ is provided in [98] with the rate $\epsilon_n = s_n \log p_n$. The authors prove the equivalent of Theorem 5.5.2 for the usual posterior with prior $\Pi_{\mathbf{HS}}$ where they allow a deterministic choice of the global shrinkage parameter τ. Thus, Theorem 5.5.2 establishes that the fractional posterior also achieves the optimal convergence rate. In [99] convergence properties of the fractional posterior is studied with a joint prior on the model size and the coefficients. [16] extensively studied the usual posterior in model (5.15) with a similar prior as considered in [99].

5.6 High-dimensional sparse factor models

Factor models are a widely popular tool to model the dependence structure of multi-dimensional observations through a linear combination of unobserved underlying factors [100, 101]. When the number of factors is sufficiently small compared to the dimension then it can also be interpreted as a dimension reduction tool [88]. Suppose we observe $y_1, \ldots, y_n \sim \mathrm{N}(0, \Omega_n)$ where $y_i \in \Re^{p_n}$, $i = 1, \ldots, n$ independently and our parameter of interest is Ω_n. A k factor model proposes the following decomposition,

$$y_i = \Lambda_n \eta_i + \epsilon_i, \tag{5.17}$$

where $k \ll p_n$, $\eta_i \in \Re^{k \times 1}$ are the unobserved factors, $\Lambda_n \in \Re^{p_n \times k}$ is the factor loading matrix and $\epsilon_i \sim \mathrm{N}(0, \Sigma_n)$, $\Sigma_n = \sigma_n^2 \mathrm{I}_{p_n}$. When $\eta_i \sim \mathrm{N}(0, \mathrm{I}_k)$, then $\Omega_n = \Lambda_n \Lambda_n^{\mathrm{T}} + \Sigma_n$. Thus model (5.17) reduces the number of parameters from $\{p_n(p_n - 1)/2\}$ to $p_n(k + 1)$ - a reduction of quadratic to linear in p_n. While for small to moderate p_n model (5.17) may provide sufficient dimension reduction, in many modern applications such as Genomics where p_n may be in thousands, further sparsity structure is warranted for devising efficient statistical procedures. Especially, in the Bayesian framework [74] introduced *sparse factor models* where the loading matrix Λ is modeled to be sparse via proper point mass mixture priors. See also [11, 102]. Computational challenges involving point mass priors have motivated the an alternative class of priors known as shrinkage priors. [36] used independent shrinkage priors on the columns of the loading matrix Λ starting with a conservative estimate of the number of unknown factors. [37] elicited a prior on Λ by having independent Dirichlet-Laplace priors [23] on the columns of Λ. The authors show minimax (upto log factors) convergence rates of the posterior distribution in Frobenius and operator norm loss when the dimension p_n grows exponentially with n; $\log p_n \leq n^{\gamma}$, $\gamma \in (0, 1)$.

Suppose $\Omega_{0n} \in \mathbb{C}_n$ is the true data generating parameter, i.e. we observe,

$$y^{(n)} = (y_1, \ldots, y_n), \text{ where } y_i \overset{iid}{\sim} \mathrm{N}(0, \Omega_{0n}),$$

where $\mathbb{C}_n$ is the cone of covariance matrices of order $p_n \times p_n$. In order to estimate Ω_{0n}, we model the data as $y_i \sim \mathrm{N}(0, \Omega_n)$ and use (5.17) to decompose Ω_n as $\Omega_n = \Lambda_n \Lambda_n^{\mathrm{T}} + \Sigma_n$. Before we begin our prior prior specification, we list the assumption we make on Ω_{0n}.

Assumption 5.6.1. *Ω_{0n} is of the form, $\Omega_{0n} = \Lambda_{0n}\Lambda_{0n}^{\mathrm{T}} + \Sigma_{0n}$, where $\Lambda_{0n} \in \Re^{p_n \times k_{0n}}$, $k_{0n} \leq p_n$ and $\Sigma_{0n} = \sigma_{0n}^2 \mathrm{I}_{p_n}$.*

Assumption 5.6.2. *Each column of $\Lambda_{0n} \in \ell\,[s_n; p_n]$ with $s_n = o(p_n)$ and the entries of the columns are uniformly bounded by some positive T.*

Assumption 5.6.3. *There are positive constants a and b such that $a < s_{\min}(\Omega_{0n}) < s_{\max}(\Omega_{0n}) < b$ for every n.*

Suppose k_* is a conservative estimate of the number of factors k_{0n}. Although k_{0n} is not known, assuming $k_* > k_{0n}$, we will set our prior in order to shrink out the redundant $(k_* - k_{0n})$ columns. To that end, letting $\Lambda_j,\ j = 1, \ldots, k_*$ to be the j^{th} column of Λ_n, we specify a the horseshoe prior $\Pi_{\mathbf{HS}}$ with individual global shrinkage parameter τ_j for each column,

$$\lambda_{hj} \mid \psi_{hj}, \tau_j \sim \mathrm{N}(0, \psi_{jh}^2 \tau_j^2),\ \ \psi_{hj} \sim \mathrm{C}^+(0,1),\ \ \tau_j \sim \mathrm{C}^+(0,1). \tag{5.18}$$

We call the joint prior on Λ_n as Π_{Λ_n}. As for the residual variance σ_n^2, we use a $\mathrm{Gamma}(a, b)$ prior which will be referred as Π_{σ_n}. The joint prior thus elicited on Ω_n through Λ_n and σ_n will be denoted by $\Pi_{\Omega_n} = \Pi_{\Lambda_n} \otimes \Pi_{\sigma_n}$. Under this setup, for a fixed $\alpha \in (0,1)$, the α-fractional posterior for $B \in \mathbb{C}_n$ is,

$$\Pi_{n,\alpha}(B \mid y^{(n)}) = \frac{\int_B \exp\left\{-\frac{\alpha}{2}\sum_{i=1}^n y_i \Omega_n^{-1} y_i\right\} \Pi_{\Omega_n}(d\Omega_n)}{\int_{\mathbb{C}_n} \exp\left\{-\frac{\alpha}{2}\sum_{i=1}^n y_i \Omega_n^{-1} y_i\right\} \Pi_{\Omega_n}(d\Omega_n)} \tag{5.19}$$

The set B_n from Theorem 5.4.1 on which we study the prior concentration in the current context is characterized in the following lemma.

Lemma 5.6.4. *For every positive ϵ, we have $B_n(\Omega_{0n}, \epsilon) \supset \widetilde{A}_n\{(\Lambda_{0n}, \sigma_{0n}), \epsilon\}$, where*

$$\widetilde{A}_n\{(\Lambda_{0n}, \sigma_{0n}), \epsilon\} = \{(\Lambda_{0n}, \sigma_{0n}) : \|\Lambda_{0n} - \Lambda_n\|_F^2 < C\epsilon^2, (\sigma_{0n} - \sigma_n)^2 < C\epsilon^2/p_n^2\},$$

for some positive constant C.

Proof. First we note that for every $\epsilon > 0$, $B_n(\Omega_{0n}, \epsilon) \supset A_n(\Omega_{0n}, \epsilon)$, where

$$A_n(\Omega_{0n}, \epsilon) = \left\{ \Omega_n : \int p^{(n)}_{\Omega_{0n}} \log \frac{p^{(n)}_{\Omega_{0n}}}{p^{(n)}_{\Omega_n}} \leq n\epsilon^2, \int p^{(n)}_{\Omega_{0n}} \left(\log \frac{p^{(n)}_{\Omega_{0n}}}{p^{(n)}_{\Omega_n}} - \int p^{(n)}_{\Omega_{0n}} \log \frac{p^{(n)}_{\Omega_{0n}}}{p^{(n)}_{\Omega_n}} \right)^2 \leq n\epsilon^2 \right\}. \tag{5.20}$$

The Kullback-Liebler divergence between $\mathrm{N}(0, \Omega_{0n})$ and $\mathrm{N}(0, \Omega_n)$ is $2^{-1}\mathrm{tr}(\Omega_n^{-1}\Omega_{0n} - \mathrm{I}_{p_n}) + \log|\Omega\Omega_{0n}^{-1}|$. Hence we have,

$$\int p^{(n)}_{\Omega_{0n}} \log \frac{p^{(n)}_{\Omega_{0n}}}{p^{(n)}_{\Omega_n}} = (n/2)\,\mathrm{tr}(\Omega_n^{-1}\Omega_{0n} - \mathrm{I}_{p_n}) + \log|\Omega\Omega_{0n}^{-1}|.$$

By similar calculations we also have,

$$\int p_{\Omega^{(n)}_{0n}} \left(\log \frac{p^{(n)}_{\Omega_{0n}}}{p^{(n)}_{\Omega_n}} - \int p_{\Omega^{(n)}_{0n}} \log \frac{p^{(n)}_{\Omega_{0n}}}{p^{(n)}_{\Omega_n}} \right)^2 = (n/2)\,\mathrm{tr}\left(\Omega_n^{-1}\Omega_{0n} - \mathrm{I}_{p_n}\right)^2 = \left\|\Omega_n^{-1}\Omega_{0n} - \mathrm{I}_{p_n}\right\|_F^2.$$

From Lemma 1.3 of [37] we have that for $\delta \in (0,1)$ and $\delta/s_{\min}(\Omega_{0n}) < 1/2$, $\|\Omega_{0n} - \Omega_n\|_F^2 < \delta$ implies $\mathrm{tr}(\Omega_n^{-1}\Omega_{0n} - \mathrm{I}_{p_n}) + \log|\Omega\Omega_{0n}^{-1}| \leq C\delta^2$, for some positive constant C. Also, from the same result we get for $\|\Omega_{0n} - \Omega_n\| < \delta$ $\|\Omega_n^{-1}\Omega_{0n} - \mathrm{I}_{p_n}\|_F^2 \leq K\delta^2$, for some positive K. Therefore, the set $A_n(\Omega_{0n}, \epsilon)$ after adjusting for constants contains the set $A_n^*(\Omega_{0n}, \epsilon) = \left\{\Omega_n : \|\Omega_{0n} - \Omega_n\|_F^2 \leq C\epsilon^2\right\}$, for some positive C.

Our next goal is to write $A_n^1(\Omega_{0n}, \epsilon)$ in terms of Λ_{0n} and σ_0. If $\|\Lambda_{0n} - \Lambda_n\| < \delta$ for sufficiently small δ then,

$$\begin{aligned}
\|\Omega_{0n} - \Omega_n\|_F^2 &= \|\Lambda_{0n}\Lambda_{0n}^{\mathrm{T}} - \Lambda_n\Lambda_n^{\mathrm{T}} + (\sigma_{0n} - \sigma)\mathrm{I}_{p_n}\|_F^2 \\
&\leq \|\Lambda_{0n}\Lambda_{0n}^{\mathrm{T}} - \Lambda_n\Lambda_n^{\mathrm{T}}\|_F^2 + (\sigma_{0n} - \sigma)^2 p_n \\
&\leq \|\Lambda_{0n} - \Lambda_n\|_F \|\Lambda_{0n} - \Lambda\|_2 + (\sigma_{0n} - \sigma)^2 \mathrm{I}_{p_n} \\
&\leq \|\Lambda_{0n} - \Lambda\|_F^2\, 2\, \|\Lambda_{0n}\|_2^2 + (\sigma_{0n} - \sigma)^2 \mathrm{I}_{p_n},
\end{aligned}$$

where to obtain the first inequality we have used $(a-b)^2 \leq 2(a^2+b^2)$ and for the last two steps we used $\|AB\|_F \leq \|A\|_2 \|B\|_F$ and $\|A\|_2 \leq \|A\|_F$. Hence we have the following set containment,

$$A_n^*(\Omega_{0n}, \epsilon) \supset \widetilde{A}_n\{(\Lambda 0n, \sigma_{0n}), \epsilon\} = \{(\Lambda_{0n}, \sigma_{0n}) : \|\Lambda_{0n} - \Lambda_n\|_F^2 < C\epsilon^2, (\sigma_{0n} - \sigma_n)^2 < C\epsilon^2/p_n^2\}. \tag{5.21}$$

□

Suppose our initial guess of the number of factors k_* is such that $k_* = Mk_{0n}$ for some constant $M > 1$. Then we have the following prior concentration result for $\widetilde{A}_n\{(\Lambda 0n, \sigma_{0n}), \epsilon\}$ for the prior Π_{Ω_n}.

Theorem 5.6.5. *Consider the prior Π_{Ω_n}. Then,*

$$\Pi_{\Omega_n}\{B_n(\Omega_{0n}, \epsilon)\} \geq e^{-Ck_{0n}s_n \log p_n},$$

for some positive C.

Proof. From Lemma 5.6.4 we get $\Pi_{\Omega_n}\{B_n(\Omega_{0n}, \epsilon)\} \geq \Pi_{\Omega_n}[\widetilde{A}_n\{(\Lambda 0n, \sigma_{0n}), \epsilon\}]$. The rest of the proof is very similar to Lemma 4.0.4 and is omitted here. □

Based on 5.6.5 we get the following convergence rate for the fractional posterior distribution in Hellinger distance.

Theorem 5.6.6. *For prior Π_{Ω_n} we have for any $\alpha \in (0,1)$ and $t > 0$*

$$\Pi_{n,\alpha}\left\{\Omega_n : h^2(p_{\Omega_{n0}}, p_{\Omega_n}) \geq \frac{D+3t}{(1-\alpha)}\right\} \leq e^{-tk_{0n}s_n \log p_n},$$

with $P_{\Omega_{0n}}^{(n)}$-probability at least $1 - 2/\{(D-1+t)^2 k_{0n}s_n \log p_n\}$ for any $D \geq 2$.

Proof. The result holds for any Rényi divergence of order $\alpha \in (0,1)$ by Theorem 5.4.1. The claim then follows from the equivalence of Rényi divergences [28]. □

6. SUMMARY AND CONTRIBUTIONS

In this thesis we have attempted to lay a solid foundation for applications of shrinkage priors in various high-dimensional problems. Our contributions cover the computational, methodological and theoretical aspects of this class of priors.

In particular, in chapter 2 and 3 we focus on the computational and methodological part. We develop an exact sampling algorithm for MCMC updates in chapter 2 which scales linearly in the dimension. Then building on this fast algorithm, we extend shrinkage priors to model low-rank, row-sparse matrices in 3. Finally, in chapter 5 we establish the minimax optimality of fractional horseshoe posteriors in high-dimensional regression and factor models. Our prior concentration result can be easily adopted to cover other sparse situations such as dictionary learning, approximate sparse factor models etc.

REFERENCES

[1] C. Carvalho, N. Polson, and J. Scott, "The horseshoe estimator for sparse signals," *Biometrika*, vol. 97, no. 2, pp. 465–480, 2010.

[2] L. Wang, G. Chen, and H. Li, "Group scad regression analysis for microarray time course gene expression data," *Bioinformatics*, vol. 23, no. 12, pp. 1486–1494, 2007.

[3] S. Van de Geer, P. Bühlmann, Y. Ritov, R. Dezeure, *et al.*, "On asymptotically optimal confidence regions and tests for high-dimensional models," *The Annals of Statistics*, vol. 42, no. 3, pp. 1166–1202, 2014.

[4] H. Ruffieux, A. C. Davison, J. Hager, and I. Irincheeva, "Efficient inference for genetic association studies with multiple outcomes," *Biostatistics*, p. To appear, 2017.

[5] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

[6] J. Fan and R. Li, "Variable selection via nonconcave penalized likelihood and its oracle properties," *Journal of the American statistical Association*, vol. 96, no. 456, pp. 1348–1360, 2001.

[7] C.-H. Zhang *et al.*, "Nearly unbiased variable selection under minimax concave penalty," *The Annals of statistics*, vol. 38, no. 2, pp. 894–942, 2010.

[8] P. Bühlmann and S. Van De Geer, *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011.

[9] T. J. Mitchell and J. J. Beauchamp, "Bayesian variable selection in linear regression," *Journal of the American Statistical Association*, vol. 83, no. 404, pp. 1023–1032, 1988.

[10] E. I. George and R. E. McCulloch, "Variable selection via Gibbs sampling," *Journal of the American Statistical Association*, vol. 88, no. 423, pp. 881–889, 1993.

[11] H. F. Lopes and M. West, "Bayesian model assessment in factor analysis," *Statistica Sinica*, pp. 41–67, 2004.

[12] P. J. Brown, M. Vannucci, and T. Fearn, "Multivariate Bayesian variable selection and prediction," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 60, no. 3, pp. 627–641, 1998.

[13] I. M. Johnstone, B. W. Silverman, *et al.*, "Needles and straw in haystacks: Empirical bayes estimates of possibly sparse sequences," *The Annals of Statistics*, vol. 32, no. 4, pp. 1594–1649, 2004.

[14] J. G. Scott, J. O. Berger, *et al.*, "Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem," *The Annals of Statistics*, vol. 38, no. 5, pp. 2587–2619, 2010.

[15] I. Castillo and A. van der Vaart, "Needles and straw in a haystack: Posterior concentration for possibly sparse sequences," *The Annals of Statistics*, vol. 40, no. 4, pp. 2069–2101, 2012.

[16] I. Castillo, J. Schmidt-Hieber, A. Van der Vaart, *et al.*, "Bayesian linear regression with sparse priors," *The Annals of Statistics*, vol. 43, no. 5, pp. 1986–2018, 2015.

[17] J. A. Hoeting, D. Madigan, A. E. Raftery, and C. T. Volinsky, "Bayesian model averaging: a tutorial," *Statistical science*, pp. 382–401, 1999.

[18] C. Stein, "Inadmissibility of the usual estimator for the mean of a multivariate normal distribution," tech. rep., STANFORD UNIVERSITY STANFORD United States, 1956.

[19] B. Efron and C. Morris, "Stein's estimation rule and its competitorsâĂŤan empirical bayes approach," *Journal of the American Statistical Association*, vol. 68, no. 341, pp. 117–130, 1973.

[20] N. G. Polson and J. G. Scott, "Shrink globally, act locally: sparse Bayesian regularization and prediction," *Bayesian Statistics*, vol. 9, pp. 501–538, 2010.

[21] L. Kuo and B. Mallick, "Variable selection for regression models," *Sankhyā: The Indian Journal of Statistics, Series B*, pp. 65–81, 1998.

[22] J. Griffin and P. Brown, "Inference with normal-gamma prior distributions in regression problems," *Bayesian Analysis*, vol. 5, no. 1, pp. 171–188, 2010.

[23] A. Bhattacharya, D. Pati, N. S. Pillai, and D. B. Dunson, "Dirichlet–laplace priors for optimal shrinkage," *Journal of the American Statistical Association*, vol. 110, no. 512, pp. 1479–1490, 2015.

[24] H. Rue, "Fast sampling of Gaussian markov random fields," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 63, no. 2, pp. 325–338, 2001.

[25] S. van der Pas, B. Kleijn, and A. van der Vaart, "The horseshoe estimator: Posterior concentration around nearly black vectors," *Electronic Journal of Statistics*, vol. 8, no. 2, pp. 2585–2618, 2014.

[26] P. Ghosh and A. Chakrabarti, "Asymptotic optimality of one-group shrinkage priors in sparse high-dimensional problems," *Bayesian Analysis*, 2017. To appear.

[27] A. Armagan, D. B. Dunson, J. Lee, W. U. Bajwa, and N. Strawn, "Posterior consistency in linear models under shrinkage priors," *Biometrika*, vol. 100, no. 4, pp. 1011–1018, 2013.

[28] A. Bhattacharya, D. Pati, and Y. Yang, "Bayesian fractional posteriors," *Annals of Statistics*, p. To appear, 2018.

[29] M. E. Tipping, "Sparse bayesian learning and the relevance vector machine," *Journal of machine learning research*, vol. 1, no. Jun, pp. 211–244, 2001.

[30] K. Bae and B. K. Mallick, "Gene selection using a two-level hierarchical bayesian model," *Bioinformatics*, vol. 20, no. 18, pp. 3423–3430, 2004.

[31] T. Park and G. Casella, "The Bayesian lasso," *Journal of the American Statistical Association*, vol. 103, no. 482, pp. 681–686, 2008.

[32] C. Hans, "Bayesian lasso regression," *Biometrika*, vol. 96, no. 4, pp. 835–845, 2009.

[33] A. Armagan, D. Dunson, and J. Lee, "Generalized double Pareto shrinkage," *Statistica Sinica*, vol. 23, no. 1, p. 119, 2013.

[34] A. Bhattacharya, D. Pati, N. S. Pillai, and D. B. Dunson, "Dirichlet-laplace priors for optimal shrinkage," *Journal of the American Statistical Association*, pp. 00–00, sep 2014.

[35] E. I. George and R. E. McCulloch, "Approaches for bayesian variable selection," *Statistica sinica*, pp. 339–373, 1997.

[36] A. Bhattacharya and D. Dunson, "Sparse Bayesian infinite factor models," *Biometrika*, vol. 98, no. 2, pp. 291–306, 2011.

[37] D. Pati, A. Bhattacharya, N. S. Pillai, D. Dunson, *et al.*, "Posterior contraction in sparse Bayesian factor models for massive covariance matrices," *The Annals of Statistics*, vol. 42, no. 3, pp. 1102–1130, 2014.

[38] D. Durante, B. Scarpa, and D. B. Dunson, "Locally adaptive factor processes for multivariate time series.," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1493–1522, 2014.

[39] N. G. Polson, J. G. Scott, and J. Windle, "The Bayesian bridge," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 76, no. 4, pp. 713–733, 2014.

[40] G. H. Golub and C. F. van Loan, *Matrix Computations*. John Hopkins University Press, 3 ed., 1996.

[41] W. W. Hager, "Updating the inverse of a matrix," *SIAM review*, vol. 31, no. 2, pp. 221–239, 1989.

[42] J. Fan and J. Lv, "Sure independence screening for ultrahigh dimensional feature space," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 70, no. 5, pp. 849–911, 2008.

[43] B. Szabó, A. van der Vaart, J. van Zanten, *et al.*, "Frequentist coverage of adaptive nonparametric bayesian credible sets," *The Annals of Statistics*, vol. 43, no. 4, pp. 1391–1428, 2015.

[44] J. H. Albert and S. Chib, "Bayesian analysis of binary and polychotomous response data," *Journal of the American statistical Association*, vol. 88, no. 422, pp. 669–679, 1993.

[45] C. C. Holmes and L. Held, "Bayesian auxiliary variable models for binary and multinomial regression," *Bayesian Analysis*, vol. 1, no. 1, pp. 145–168, 2006.

[46] T. Anderson, "Multivariate statistical analysis," *Willey and Sons, New York, NY*, 1984.

[47] T. W. Anderson, "Estimating linear restrictions on regression coefficients for multivariate normal distributions," *The Annals of Mathematical Statistics*, pp. 327–351, 1951.

[48] A. J. Izenman, "Reduced-rank regression for the multivariate linear model," *Journal of Multivariate Analysis*, vol. 5, no. 2, pp. 248–264, 1975.

[49] R. Velu and G. C. Reinsel, *Multivariate reduced-rank regression: theory and applications*, vol. 136. Springer Science & Business Media, 2013.

[50] T. Anderson, "Specification and misspecification in reduced rank regression," *Sankhyā: The Indian Journal of Statistics, Series A*, pp. 193–205, 2002.

[51] M. Yuan, A. Ekici, Z. Lu, and R. Monteiro, "Dimension reduction and coefficient estimation in multivariate linear regression," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 69, no. 3, pp. 329–346, 2007.

[52] K. Chen, H. Dong, and K.-S. Chan, "Reduced rank regression via adaptive nuclear norm penalization," *Biometrika*, vol. 100, no. 4, pp. 901–920, 2013.

[53] F. Bunea, Y. She, and M. H. Wegkamp, "Optimal selection of reduced rank estimators of high-dimensional matrices," *The Annals of Statistics*, vol. 39, no. 2, pp. 1282–1309, 2011.

[54] F. Bunea, Y. She, M. H. Wegkamp, *et al.*, "Joint variable and rank selection for parsimonious estimation of high-dimensional matrices," *The Annals of Statistics*, vol. 40, no. 5, pp. 2359–2388, 2012.

[55] L. Chen and J. Z. Huang, "Sparse reduced-rank regression for simultaneous dimension reduction and variable selection," *Journal of the American Statistical Association*, vol. 107, no. 500, pp. 1533–1545, 2012.

[56] J. Geweke, "Bayesian reduced rank regression in econometrics," *Journal of econometrics*, vol. 75, no. 1, pp. 121–146, 1996.

[57] P. J. Green, "Reversible jump Markov chain Monte Carlo computation and Bayesian model determination," *Biometrika*, vol. 82, no. 4, pp. 711–732, 1995.

[58] Y. J. Lim and Y. W. Teh, "Variational Bayesian approach to movie rating prediction," in *Proceedings of KDD cup and workshop*, vol. 7, pp. 15–21, Citeseer, 2007.

[59] R. Salakhutdinov and A. Mnih, "Bayesian probabilistic matrix factorization using markov chain monte carlo," in *Proceedings of the 25th international conference on Machine learning*, pp. 880–887, ACM, 2008.

[60] S. D. Babacan, R. Molina, and A. K. Katsaggelos, "Variational Bayesian super resolution," *IEEE Transactions on Image Processing*, vol. 20, no. 4, pp. 984–999, 2011.

[61] P. Alquier, "Bayesian methods for low-rank matrix estimation: Short survey and theoretical study," in *Algorithmic Learning Theory*, pp. 309–323, Springer, 2013.

[62] J. Lucas, C. Carvalho, Q. Wang, A. Bild, J. Nevins, and M. West, "Sparse statistical modelling in gene expression genomics," *Bayesian Inference for Gene Expression and Proteomics*, vol. 1, pp. 0–1, 2006.

[63] H. Wang, "Sparse seemingly unrelated regression modelling: Applications in finance and econometrics," *Computational Statistics & Data Analysis*, vol. 54, no. 11, pp. 2866–2877, 2010.

[64] A. Bhadra and B. K. Mallick, "Joint high-dimensional bayesian variable and covariance selection with an application to eqtl analysis," *Biometrics*, vol. 69, no. 2, pp. 447–457, 2013.

[65] A. Bhattacharya, A. Chakraborty, and B. K. Mallick, "Fast sampling with gaussian scale mixture priors in high-dimensional regression," *Biometrika*, vol. 103, no. 4, pp. 985–991, 2016.

[66] H. D. Bondell and B. J. Reich, "Consistent high-dimensional Bayesian variable selection via penalized credible regions," *Journal of the American Statistical Association*, vol. 107, no. 500, pp. 1610–1624, 2012.

[67] S. Kundu, V. Baladandayuthapani, and B. K. Mallick, "Bayes regularized graphical model estimation in high dimensions," tech. rep., 2013.

[68] P. R. Hahn and C. M. Carvalho, "Decoupling shrinkage and selection in Bayesian linear models: a posterior summary perspective," *Journal of the American Statistical Association*, vol. 110, no. 509, pp. 435–448, 2015.

[69] J. S. Liu and Y. N. Wu, "Parameter expansion for data augmentation," *Journal of the American Statistical Association*, vol. 94, no. 448, pp. 1264–1274, 1999.

[70] P. Hoff, "Simulation of the matrix Bingham-von Mises-Fisher distribution, with applications to multivariate and relational data," *Journal of Computational and Graphical Statistics*, vol. 18, no. 2, pp. 438–456, 2009.

[71] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 68, no. 1, pp. 49–67, 2006.

[72] H. Zou, "The adaptive lasso and its oracle properties," *Journal of the American statistical association*, vol. 101, no. 476, pp. 1418–1429, 2006.

[73] H. Chun and S. Keleş, "Sparse partial least squares regression for simultaneous dimension reduction and variable selection," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 72, no. 1, pp. 3–25, 2010.

[74] M. M. Barbieri and J. O. Berger, "Optimal predictive model selection," *The Annals of Statistics*, vol. 32, no. 3, pp. 870–897, 2004.

[75] P. T. Spellman, G. Sherlock, M. Q. Zhang, V. R. Iyer, K. Anders, M. B. Eisen, P. O. Brown, D. Botstein, and B. Futcher, "Comprehensive identification of cell cycle–regulated genes of

the yeast saccharomyces cerevisiae by microarray hybridization," *Molecular biology of the cell*, vol. 9, no. 12, pp. 3273–3297, 1998.

[76] T. I. Lee, N. J. Rinaldi, F. Robert, D. T. Odom, Z. Bar-Joseph, G. K. Gerber, N. M. Hannett, C. T. Harbison, C. M. Thompson, I. Simon, *et al.*, "Transcriptional regulatory networks in saccharomyces cerevisiae," *Science*, vol. 298, no. 5594, pp. 799–804, 2002.

[77] S. Ghosal, J. K. Ghosh, and A. W. van der Vaart, "Convergence rates of posterior distributions," *The Annals of Statistics*, vol. 28, no. 2, pp. 500–531, 2000.

[78] W. Jiang *et al.*, "Bayesian variable selection for high dimensional generalized linear models: convergence rates of the fitted densities," *The Annals of Statistics*, vol. 35, no. 4, pp. 1487–1511, 2007.

[79] G. Leung and A. R. Barron, "Information theory and mixing least-squares regressions," *IEEE Transactions on Information Theory*, vol. 52, no. 8, pp. 3396–3410, 2006.

[80] A. R. Barron, "Information-theoretic characterization of bayes performance and the choice of priors in parametric and nonparametric problems," *Bayesian Statistics,6*, pp. 27–52, 1999.

[81] J. Friedman, T. Hastie, H. Höfling, and R. Tibshirani, "Pathwise coordinate optimization," *The Annals of Applied Statistics*, vol. 1, no. 2, pp. 302–332, 2007.

[82] W. Shen, S. T. Tokdar, and S. Ghosal, "Adaptive bayesian multivariate density estimation with dirichlet mixtures," *Biometrika*, vol. 100, no. 3, pp. 623–640, 2013.

[83] A. Bhattacharya, D. B. Dunson, D. Pati, and N. S. Pillai, "Sub-optimality of some continuous shrinkage priors," *Stochastic Processes and their Applications*, vol. 126, no. 12, pp. 3828 – 3842, 2016. In Memoriam: Evarist GinÃl'.

[84] R. Vershynin, "Introduction to the non-asymptotic analysis of random matrices," *arXiv preprint arXiv:1011.3027*, 2010.

[85] S. A. Van de Geer *et al.*, "High-dimensional generalized linear models and the lasso," *The Annals of Statistics*, vol. 36, no. 2, pp. 614–645, 2008.

[86] P. J. Bickel and E. Levina, "Regularized estimation of large covariance matrices," *The Annals of Statistics*, pp. 199–227, 2008.

[87] P. J. Bickel, E. Levina, *et al.*, "Covariance regularization by thresholding," *The Annals of Statistics*, vol. 36, no. 6, pp. 2577–2604, 2008.

[88] J. Fan, Y. Fan, and J. Lv, "High dimensional covariance matrix estimation using a factor model," *Journal of Econometrics*, vol. 147, no. 1, pp. 186–197, 2008.

[89] C.-H. Zhang and S. S. Zhang, "Confidence intervals for low dimensional parameters in high dimensional linear models," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 76, no. 1, pp. 217–242, 2014.

[90] A. Javanmard and A. Montanari, "Confidence intervals and hypothesis testing for high-dimensional regression," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 2869–2909, 2014.

[91] K. Khare and J. P. Hobert, "Geometric ergodicity of the bayesian lasso," *Electronic Journal of Statistics*, vol. 7, pp. 2150–2163, 2013.

[92] S. Pal and K. Khare, "Geometric ergodicity for bayesian shrinkage models," *Electronic Journal of Statistics*, vol. 8, no. 1, pp. 604–645, 2014.

[93] J. E. Johndrow, P. Orenstein, and A. Bhattacharya, "Scalable mcmc for bayes shrinkage priors," tech. rep., 2017.

[94] D. L. Donoho, I. M. Johnstone, J. C. Hoch, and A. S. Stern, "Maximum entropy and the nearly black object," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 41–81, 1992.

[95] S. van der Pas, B. Szabó, A. van der Vaart, *et al.*, "Adaptive posterior contraction rates for the horseshoe," *Electronic Journal of Statistics*, vol. 11, no. 2, pp. 3196–3225, 2017.

[96] L. Schwartz, "On bayes procedures," *Journal of Probability Theory and Related Areas*, vol. 4, no. 1, pp. 10–26, 1965.

[97] A. Chakraborty, A. Bhattacharya, and B. K. Mallick, "Bayesian sparse multiple regression for simultaneous rank reduction and variable selection," tech. rep., 2016.

[98] Q. Song and F. Liang, "Nearly optimal bayesian shrinkage for high dimensional regression," tech. rep., 2017.

[99] R. Martin, R. Mess, S. G. Walker, *et al.*, "Empirical bayes posterior concentration in sparse high-dimensional linear models," *Bernoulli*, vol. 23, no. 3, pp. 1822–1847, 2017.

[100] S. A. Ross, "The arbitrage theory of capital asset pricing," in *HANDBOOK OF THE FUNDAMENTALS OF FINANCIAL DECISION MAKING: Part I*, pp. 11–30, World Scientific, 2013.

[101] S. A. Ross, "The capital asset pricing model (capm), short-sale restrictions and related issues," *The Journal of Finance*, vol. 32, no. 1, pp. 177–183, 1977.

[102] C. M. Carvalho, J. Chang, J. E. Lucas, J. R. Nevins, Q. Wang, and M. West, "High-dimensional sparse factor modeling: applications in gene expression genomics," *Journal of the American Statistical Association*, vol. 103, no. 484, pp. 1438–1456, 2008.

APPENDIX A

FRACTIONAL VERSUS USUAL POSTERIOR

In this section, we provide some additional discussion regarding our adoption of the fractional posterior framework in the main document. We begin with a detailed discussion on the sufficient conditions required to establish posterior contraction rates for the usual posterior from [77] and contrast them with those of fractional posteriors [28]. For simplicity, we discuss the i.i.d. case although the discussion is broadly relevant beyond the i.i.d. setup. We set out with some notation. Suppose we observe n independent and identically distributed random variables $X_1, \ldots, X_n \mid P \sim P$ where $P \in \mathcal{P}$, a family of probability measures. Denote $L_n(P)$ as the likelihood for this data which we abbreviate and write as $X^{(n)}$. We treat P as our parameter of interest and define a prior Π_n for P.

Let $P_0 \in \mathcal{P}$ be the true data generating distribution. For a measurable set B, the posterior probability assigned to B is

$$\Pi_n(B \mid X^{(n)}) = \frac{\int_B L_n(P)\, \Pi_n(dP)}{\int_{\mathcal{P}} L_n(P)\, \Pi_n(dP)} \tag{A.1}$$

For $\alpha \in (0, 1)$, the α-fractional posterior $\Pi_{n,\alpha}(\cdot \mid Y)$ is,

$$\Pi_{n,\alpha}(B \mid X^{(n)}) = \frac{\int_B \{L_n(P)\}^\alpha \Pi_n(dP)}{\int_{\mathcal{P}} \{L_n(P)\}^\alpha \Pi_n(dP)}. \tag{A.2}$$

The fractional posterior is obtained upon raising the likelihood to a fractional power α and combining with the prior using Bayes's theorem.

Let p and p_0 be the density of P and P_0 respectively with respect to some measure μ and $p^{(n)}$ and $p_0^{(n)}$ be the corresponding joint densities. Suppose ϵ_n is a sequence such that $\epsilon_n \to 0$ and $n\epsilon_n^2 \to \infty$ as $n \to \infty$. Define $B_n = \{p : \int p_0^{(n)} \log p_0^{(n)}/p^{(n)} \leq n\epsilon_n^2, \int p_0^{(n)} \log^2 p_0^{(n)}/p^{(n)} \leq n\epsilon_n^2\}$. Given a metric ρ on $\mathcal{P}$ and $\delta > 0$, let $N(P^*, \rho, \delta)$ be the covering number of $P^* \subset \mathcal{P}$ [77]. For

sake of concreteness, we focus on the case where ρ is the Hellinger distance. We now state the sufficient conditions for $\Pi_n(\cdot \mid X^{(n)})$ to contract at rate ϵ_n at P_0 [77].

Theorem A.0.1 ([77]). *Suppose ϵ_n be as above. If, there exists $\mathcal{P}_n \subset \mathcal{P}$ and positive constants C_1, C_2 such that,*

1. $\log N(\mathcal{P}_n, h, \epsilon_n) \lesssim n\epsilon_n^2$,

2. $\Pi_n(\mathcal{P}_n^c) \leq e^{-C_1 n\epsilon_n^2}$, *and*

3. $\Pi_n(B_n) \geq e^{-C_3 n\epsilon_n^2}$,

then $\Pi_n\{p : h^2(p, p_0) > M\epsilon_n \mid X^{(n)}\} \to 0$ *in P_0-probability for a sufficiently large M.*

However, if we use the fractional posterior $\Pi_{n,\alpha}(\cdot \mid X^{(n)})$ for $\alpha \in (0, 1)$, then we have the following result from [28],

Theorem A.0.2 ([28]). *Suppose condition 3 from Theorem S1 is satisfied. Then* $\Pi_{n,\alpha}\{h^2(p, p_0) > M\epsilon_n \mid X^{(n)}\} \to 0$ *in P_0-probability.*

We refer the reader to [28] for a more precise statement of Theorem A.0.2. The main difference between Theorems A.0.1 and A.0.2 is that the same rate of convergence (upto constants) can be arrived at verifying fewer conditions. The construction of the sets $\mathcal{P}_n$, known as sieves, can be challenging for heavy-tailed priors such as the horseshoe. On the other hand, one only needs to verify the prior concentration bound $\Pi_n(B_n) \geq e^{-C_3 n\epsilon_n^2}$ to ensure contraction of the fractional posterior. This allows one to obtain theoretical justification in complicated high-dimensional models as in ours. To quote the authors of [28], '*the condition of exponentially decaying prior mass assigned to the complement of the sieve implies fairly strong restrictions on the prior tails and essentially rules out heavy-tailed prior distributions on hyperparameters. On the other hand, a much broader class of prior choices lead to provably optimal posterior behavior for the fractional posterior*'. That said, the proof of the technical results below illustrate that verifying the prior concentration condition alone can pose a stiff technical challenge.

We now aim to provide some intuition behind why the theory simplifies with the fractional posterior. Define $U_n = \{p : h^2(p, p_0) > M\epsilon_n\}$. From equation (R2) and (R3) in [28] U_n can be al-

ternatively defined as, $U_n = \{p : D_\alpha(p, p_0) > M^* \epsilon_n\}$, where the constant M^* can be derived from M by the equivalence relation Rényi divergences [28, equation (R3)]. The posterior probability assigned to the set U_n is then obtained by (A.1) and the fractional posterior probability assigned to U_n follows from (A.2). Thus after dividing the numerator and denominator by the appropriate power of $L_n(P_0)$ we get,

$$\Pi(U_n \mid X^{(n)}) = \frac{\displaystyle\int_{U_n} \frac{L_n(P)}{L_n(P_0)} \Pi_n(dP)}{\displaystyle\int_{\mathcal{P}} \frac{L_n(P)}{L_n(P_0)} \Pi_n(dP)}, \tag{A.3}$$

and

$$\Pi_{n,\alpha}(U_n \mid X^{(n)}) = \frac{\displaystyle\int_{U_n} \left\{ \frac{L_n(P)}{L_n(P_0)} \right\}^{\alpha} \Pi_n(dP)}{\displaystyle\int_{\mathcal{P}} \left\{ \frac{L_n(P)}{L_n(P_0)} \right\}^{\alpha} \Pi_n(dP)}. \tag{A.4}$$

Taking expectation of the numerator in (A.4) with respect to P_0 and applying Fubini's theorem to interchange the integrals yields $\int_{U_n} e^{-(1-\alpha)D_\alpha(p,p_0)} \, \Pi_n(dP)$ which by definition of U_n is small. The same operation for (A.3) leads to $\int_{U_n} \Pi_n(dP)$ which isn't necessarily small, needing the introduction of the sieves $\mathcal{P}_n$ in the analysis.

We conducted a small simulation study carried out to compare the results of Π_n and $\Pi_{n,\alpha}$ for different choices of α in the context of model (3) in the main document with priors defined in (4). We obtain virtually indistinguishable operating characteristics of the point estimates, further corroborating our theoretical study.

We end this section by recording a high probability risk bound for the Rényi loss [28] in the following corollary that is subsequently used.

Corollary A.0.3 ([28])**.** *Fix $\alpha \in (0, 1)$. Recall the definition of Rényi divergence between p and p_0 from the main document: $D_\alpha(p, p_0) = (\alpha - 1)^{-1} \log \int p^\alpha p_0^{1-\alpha} d\mu$. Under the conditions of*

Table A.1: Empirical results comparing $\hat{r}$, MSPE = $(nq)^{-1}\|XC - XC_0\|_F^2$ and MSE=$(pq)^{-1}\|C - C_0\|_F^2$ for different choices of the fractional power α. $\alpha = 1$ corresponds to the usual posterior. The data used in this table was generated in a similar manner as described in section 3.4 of this document.

		(p,q)											
		(200,30)				**(500,10)**				**(1000,12)**			
		Independent		**Correlated**		**Independent**		**Correlated**		**Independent**		**Correlated**	
α	**Measures**	BSML	SPLS	BSML	SPLS	BSML	SPLS	BSML	SPLS	BSML	SPLS	BSML	SPLS
	$\hat{r}$	3.0	7.9	3.0	9.4	3.0	9.7	3.0	8.8	3.2	9.4	3.4	8.9
1	MSE	3	14	5	15	3	7	5	30	3	50	3	38
	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
	$\hat{r}$	3.0		3.0		3.0		3.0		3.0		3.1	
0.5	MSE	1.9		2.7		1.9		3.9		1.2		1.4	
	MSPE	0.05		0.06		0.15		0.25		0.22		0.32	
	$\hat{r}$	3.1		3.0		3.0		2.9		2.9		3.0	
0.75	MSE	1.8		2.4		1.6		4.3		1.2		1.2	
	MSPE	0.08		0.07		0.16		0.22		0.32		0.31	
	$\hat{r}$	3.0		3.1		3.0		3.0		3.1		2.9	
0.95	MSE	2.1		2.9		1.5		3.6		1.5		1.5	
	MSPE	0.09		0.07		0.16		0.29		0.31		0.31	

Theorem A.0.2, for any $k \geq 1$,

$$\int \left\{ \frac{1}{n} D_\alpha(p, p_0) \right\}^k d\Pi_{n,\alpha}(\cdot \mid X^{(n)}) \leq K_1 (1-\alpha)^{-k} \epsilon_n^{2k},$$

with P_0*-probability at least* $1 - K_2/(n\epsilon_n^2)$*, where* K_1 *and* K_2 *are positive constants independent of* α.