



Topics in False Discovery Rate Control and Factor Analysis

Citation

Ma, Yucong. 2021. Topics in False Discovery Rate Control and Factor Analysis. Doctoral dissertation, Harvard University Graduate School of Arts and Sciences.

Permanent link

<https://nrs.harvard.edu/URN-3:HUL.INSTREPOS:37368197>

Terms of Use

This article was downloaded from Harvard University's DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at <http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA>

Share Your Story

The Harvard community has made this article openly available.
Please share how this access benefits you. [Submit a story](#).

[Accessibility](#)

HARVARD UNIVERSITY
Graduate School of Arts and Sciences

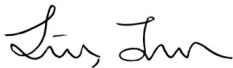


DISSERTATION ACCEPTANCE CERTIFICATE

The undersigned, appointed by the
Department of Statistics
have examined a dissertation entitled
Topics in False Discovery Rate Control and Factor Analysis

presented by Yucong Ma

candidate for the degree of Doctor of Philosophy and hereby
certify that it is worthy of acceptance.

Signature 

Typed name: Prof. Jun Liu

Signature 

Typed name: Prof. Tracy Ke

Signature **Subhabrata Sen**

Typed name: Prof. Subhabrata Sen

Signature _____

Typed name: Prof.

Signature _____

Typed name: Prof.

Date: March 5, 2021

Topics in False Discovery Rate Control and Factor Analysis

A DISSERTATION PRESENTED
BY
YUCONG MA
TO
THE DEPARTMENT OF STATISTICS

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
IN THE SUBJECT OF
STATISTICS

HARVARD UNIVERSITY
CAMBRIDGE, MASSACHUSETTS
MARCH 2021

©2021 – YUCONG MA
ALL RIGHTS RESERVED.

Topics in False Discovery Rate Control and Factor Analysis

ABSTRACT

This dissertation develops statistical theories and methodologies in the realm of false discovery rate (FDR) control and factor analysis. Both these topics are of great scientific importance in the field of social science, economics, and bioinformatics. The dissertation contains three self-contained chapters.

Chapter 1 studies how the key components (including symmetric statistics, ranking algorithm, design of fake variables, and the scheme of adding fake variables) of an FDR control method impact its power. We focus on two recent FDR control methods, the knockoff filter, and the Gaussian mirror, and develop a unified theoretical framework for power analyses under the rare/weak signal model. Our analyses lead to several noteworthy discoveries. First, the choice of the symmetric statistic in FDR control methods crucially affects the power. Second, when the components are designed properly, the operation of adding “noise” to achieve FDR control yields almost no loss of power compared with its prototype, at least for some special classes of designs. Third, a different FDR control method is preferred (in terms of power) under different sparsity levels and gram matrix designs. Our simulation studies nicely support these theoretical discoveries.

Chapter 2 studies the problem of estimating the number of spiked eigenvalues, K in a covariance matrix, or in other words identifying the number of factors in a factor model. We propose a novel approach for estimating K using the bulk eigenvalues of the sample covariance matrix. Our method imposes a working model on the residual covariance matrix, which is assumed to be a diagonal matrix whose entries are drawn from a gamma distribution. Under this model, the bulk eigenvalues are

asymptotically close to the quantiles of a fixed parametric distribution, which motivates us to propose a two-step method: the first step uses bulk eigenvalues to estimate parameters of this distribution, and the second step leverages these parameters to assist the estimation of K . We theoretically show the consistency of our estimator and we also propose a confidence interval estimate for K . Our extensive simulation studies show that the proposed method is robust and outperforms the existing methods in a range of scenarios. We finally apply the proposed method to the analysis of a lung cancer microarray data set and the 1000 Genomes data set.

Chapter 3 dives into the realm of the sparse Bayesian factor model and studies the posterior distribution inconsistency problem in the high dimensional regime, where the column-wise averaged nonzero element number in the loading matrix is larger than the number of observations. We analyze the inconsistency issue when using non-informative priors on the elements of the loading matrix. Namely, we show that using independent spike-and-slab prior on the elements of the loading matrix leads to a ‘magnitude inflation’ phenomenon for the posterior distribution of the loading matrix. Our theoretical analyses reveal the connection between posterior inconsistency and the assumption on the factors, which gives rise to a natural remedy—changing the normal factors (after scaling) to be uniform on the Stiefel manifold. Without losing any model interpretability, we propose to adopt this new orthonormal factor model in high dimensions (in place of the normal factor model) since it enjoys two major advantages. First, the posterior distribution is more robust against the choice of the prior distribution for elements of the loading matrix. Second, it leads to a significant efficiency gain in MCMC sampling. We verify these claims in both numerical studies and a real application to the AGEMAP data set.

Contents

0	INTRODUCTION	1
1	A POWER ANALYSIS OF FALSE DISCOVERY RATE CONTROL METHODS	5
1.1	Introduction	6
1.2	FDR control methods and criteria of power comparison	13
1.3	Power analysis of FDR control methods for orthogonal designs	20
1.4	Behavior of the prototypes for non-orthogonal designs	26
1.5	Power analysis of FDR control methods for non-orthogonal designs	30
1.6	The proof idea and geometric insight	43
1.7	Simulations	47
1.8	Discussions	53
2	ESTIMATING THE NUMBER OF SPIKES BY BULK EIGENVALUE MATCHING ANALYSIS	56
2.1	Introduction	57
2.2	BEMA for the standard spiked covariance model	62
2.3	BEMA for the general spiked covariance model	67
2.4	Theoretical properties	75
2.5	Simulation studies	83
2.6	Real applications	92
2.7	Discussion	98
3	ON POSTERIOR CONSISTENCY OF BAYESIAN FACTOR MODELS IN HIGH DIMENSIONS	101
3.1	Introduction	102
3.2	Bayesian sparse factor model and inference	107
3.3	The magnitude inflation phenomenon	111
3.4	Posterior dependence on the slab prior	114
3.5	Model modifications and posterior consistency	117
3.6	Numerical results	129
3.7	Dynamic exploration with application	135

3.8	Discussion	138
APPENDIX A SUPPLEMENTAL MATERIALS OF CHAPTER 1		140
A.1	Proofs	140
APPENDIX B SUPPLEMENTAL MATERIALS OF CHAPTER 2		206
B.1	GetQT algorithms	206
B.2	Proofs	213
B.3	Robustness of BEMA on real data	229
APPENDIX C SUPPLEMENTAL MATERIALS OF CHAPTER 3		231
C.1	Scaling group moves	231
C.2	The modified Ghosh-Dunson model	233
C.3	Proofs	235
C.4	Additional figures - the AGEMAP dataset	246
REFERENCES		254

TO MY PARENTS.

Acknowledgments

First, I would like to express my deep and sincere gratitude to my two research advisors, Professor Jun S. Liu and Professor Zheng Tracy Ke, for providing invaluable guidance throughout the researches. Their insightful understanding of statistics helps me shape my taste for research problems and develop professional skills. Beyond academics, they are respectable elder and friend to me, who provides thoughtful advice for my personal career. I cannot thank them enough for all that they have done for me. My sincere gratitude also goes to our collaborator Professor Xihong Lin, who always provides insightful and constructive suggestions to our research. I am honored to have Professor Subhabrata Sen on my thesis committee and I would like to thank him for giving valuable feedback.

In addition, I am thankful to all the faculty and administrative staff members of the Department of Statistics at Harvard University, for providing an excellent environment to pursue studies. Special thanks go to Joseph Blitzstein, Lucas Janson, Pierre Jacob, Mark Glickman, and Samuel Kou for teaching me the basics of advanced statistics and prepare me for researches.

I would also like to thank the senior students and my fellow students, especially Chenguang Dai, Sanqian Zhang, Wenshuo Wang, Han Yan, and Lu Zhang, for giving advice on my studies and career path. Talking to them has been a great pleasure and I really learned a lot from those discussions.

Last but not least, I dedicate my greatest thanks to my parents, for raising me and support me on all the choices I made. I wouldn't have made it this far if it hadn't been for you.

0

Introduction

With the increasing capacity of collecting and storing data, we have seen an explosion in dimensions of data sets statisticians are dealing with. In some financial or genetic data sets (e.g. high-frequency stock data, gene-expression measurements data), it is not uncommon to have the number of samples or the dimension of features to be at the order of 10^7 . With this amount of data, achieving statistical objectives with limited space and time has become the new challenge. A prevailing answer to this challenge is dimension reduction, which brought in the notion of *sparsity*.

Sparsity assumption plays a key role in high-dimensional statistical problems. For example, in a linear regression problem where the number of features exceeds the sample size. In this case, the ordinary least squares method is not applicable due to multicollinearity. But, by assuming that the response only depends on a small portion (sparsity) of the features, one can still establish proper linear models to model the response. Selecting the relevant features to include in the model is an important topic in modern statistics and has led to the developments of several branches of researches according to different variable selection objectives. In the first chapter, we study variable selection methods aiming at controlling the false discovery rate (FDR). More specifically, we study how the key components (including symmetric statistics, ranking algorithm, design of fake variables, and the scheme of adding fake variables) of an FDR control method impact its power. We focus on two recent FDR control methods, the knockoff filter (Barber & Candès, 2015) and the Gaussian mirror (Xing et al., 2019), and develop a unified theoretical framework for power analyses under the Rare/Weak signal model (Donoho & Jin, 2015). Existing literature focuses more on the analyses of the ability to control FDR while we acknowledge that the power study of these methods is of equal importance for practical use. Our study in chapter 1 aims at bridging this gap in theory and providing insight on the construction of an FDR control method in practice to boost power. The model setup and theoretical tools we used in chapter 1 are quite different from those in existing literature (Weinstein et al., 2017, 2020) for power analysis of FDR control methods, and we hope these can shed some light on related future theoretical developments.

Another example of assuming sparsity in high dimensions is the factor model. This model assumes that the observed data matrix is a perturbed version of a true data matrix that possesses a low dimensional matrix product representation. This is essentially imposing sparsity on the eigenvalues of the true data matrix. Though the model itself is rather simple, it possesses severe identifiability issues, making it difficult to pinpoint the factor dimensionality or the low dimensional decomposition. Factor analyses have a long history that dates back to 1900 and this realm stays active for almost

a century, for its wide use in biology, psychometrics, and finance. In this process, a large number of methods were proposed for the sole purpose of estimating the factor dimensionality. Prominent ones include the Kaiser’s criterion (Kaiser, 1960) and parallel analysis (Horn, 1965). These methods are later advanced to empirical Kaiser’s criterion (Braeken & Van Assen, 2017) and deterministic parallel analysis (Dobriban & Owen, 2019) utilizing the recent developments in random matrix theory. These methods, though having very nice theoretical properties, tend to fall apart in real applications due to a violation of certain modeling assumptions. In chapter 2, we propose a novel random matrix theory based approach for estimating the factor dimensionality. Our method imposes a working model for the residual covariance matrix, assuming that it is a diagonal matrix with entries drawn from a gamma distribution. Under this model, the bulk eigenvalues are asymptotically close to the quantiles of a fixed parametric distribution, which motivates us to propose a two-step method: the first step uses bulk eigenvalues to estimate parameters of this distribution, and the second step leverages these parameters to assist the estimation of factor dimensionality. We show that, besides having sharp theoretical guarantees, our model fits nicely with real data sets and our estimators are more robust against model misspecifications than existing alternatives.

In the last chapter, we switch to the Bayesian perspective. While Frequentists are seeking estimators with nice theoretical properties, Bayesians are trying to come up with priors that induce nice posterior consistency and low computational (sampling) cost (Ročková & George, 2016, Fruehwirth-Schnatter & Lopes, 2018). This turns out to be a subtle task for the high dimensional Bayesian factor model when the column-wise averaged nonzero element number in the loading matrix exceeds the number of observations. In such a scenario, common choices of prior setup for the loading matrix, e.g. the priors from (Bhattacharya & Dunson, 2011, Ročková & George, 2016), can easily lead to an inconsistent posterior distribution. Chapter 3 aims at analyzing this posterior inconsistency issue with the notion of using non-informative priors on the elements of the loading matrix. Problems with the use of diffuse priors in Bayesian inference when observation sample sizes are small relative to the num-

ber of parameters being estimated have been studied in the literature (Efron, 1973, Kass & Wasserman, 1996, Natarajan & McCulloch, 1998). In this chapter, we elaborate on this problem in the context of the factor model. We reveal connections between posterior inconsistency and the assumption on the factors, and suggest using orthonormal factors (after scaling) in place of normal factors in high dimensional settings. Without losing model interpretability, the orthonormal factor model is shown to be more robust against prior specification and leads to an efficiency gain in MCMC sampling.

1

A power analysis of False Discovery Rate

Control Methods

CONTRIBUTION This chapter is based on a paper [Ke et al. \(2020a\)](#) jointly with Prof. Zheng Tracy Ke and Prof. Jun S. Liu.

1.1 INTRODUCTION

We consider a linear regression model:

$$y = X\beta + z, \quad X = [X_1, X_2, \dots, X_n]' \in \mathbb{R}^{n \times p}, \quad z \sim N(0, \sigma^2 I_n). \quad (1.1)$$

Given a subset of selected variables $\hat{S} \subset \{1, 2, \dots, p\}$, the false discovery rate (FDR) is defined as

$$\mathbb{E} \left[\frac{\#\{j : \beta_j = 0, j \in \hat{S}\}}{\#\{j : j \in \hat{S}\} \vee 1} \right].$$

The control of FDR is a problem of great interest. When the design is orthogonal (i.e., $X'X$ is a diagonal matrix), the Benjamini-Hochberg procedure (Benjamini & Hochberg, 1995) can be employed to control FDR at a targeted level. When the design is non-orthogonal, the BH-procedure faces challenges, and several recent FDR control methods were proposed. Examples include but are not limited to the knockoff filter (Barber & Candès, 2015), model-X knockoff (Candès et al., 2018), Gaussian mirror (Xing et al., 2019), and multiple data splits (Dai et al., 2020). All these methods are shown to control FDR at a targeted level, but their power is less studied. This chapter aims to provide a theoretical understanding to the power of FDR control methods.

We introduce a unified framework that captures the key ideas behind recent FDR control methods. Starting from the seminal work of Barber & Candès (2015), this framework has been implicitly used in the literature, but it is the first time that we abstract it out:

- (a) There is a *ranking algorithm*, which assigns an importance metric to each variable.
- (b) An FDR control method creates a *tampered design matrix* by adding fake variables.
- (c) The tampered design and the response vector y are supplied to the ranking algorithm as input, and the output is converted to a (signed) importance metric for each original variable through

a symmetric statistic.

The three components, (a) ranking algorithm, (b) tampered design, and (c) symmetric statistic, need to coordinate so that the resulting importance metrics for null variables (i.e., $\beta_j = 0$) have symmetric distributions and the importance metrics for non-null variables (i.e., $\beta_j \neq 0$) are positive with high probability. Then, given any threshold $t > 0$, the number of false discoveries is estimated by counting the number of variables whose importance metric is below $-t$. As a result, one can mimic the BH procedure to control FDR at a targeted level.

The power of an FDR control method is essentially hinged on the quality of ranking variables by those importance metrics. In the aforementioned framework, each of the three components (a)-(c) has a significant impact on the resulting importance metrics and thus on the power of the FDR control method. The literature works have revealed a lot of insight on how to design these components to facilitate valid FDR control. However, there is very little understanding on how to design them so as to boost power. The main contribution of this chapter is to dissect and detail the impact of each component on the power. We discover that each of (a)-(c) can have a significant impact under some settings. Therefore, one has to be careful on the choice of these components in designing an FDR control method, and our theoretical results provide a useful guideline. Our study also helps answer a fundamental question: It is well known that adding noise often makes inference more difficult. The operation of adding fake variables to facilitate FDR control is essentially an operation of adding “noise.” Does it yield any loss of power, compared with variable selection methods that do not aim for FDR control? We find that the answer is complicated, depending on not only the choice of (a)-(c) but also model parameters such as sparsity, signal strength, and correlations among variables. For some particular model settings and particular choices of (a)-(c), we obtain encouraging answers where the operation of adding fake variables yields only a negligible power loss.

We focus our study primarily on two FDR control methods, the knockoff filter (Barber & Candès, 2015) and Gaussian mirror (Xing et al., 2019), but the analysis is readily extendable to other methods.

We chose these two methods as the object of study because they cover a variety of ideas in designing (a)-(c). For example, knockoff uses the solution path of Lasso to rank variables, while Gaussian mirror uses least-squares coefficients. Knockoff constructs the tampered design matrix by simultaneously adding p fake variables, while Gaussian mirror adds one fake variable at a time. Both methods adopt symmetric statistics including the signed maximum and the difference statistic. The study of these two methods allow us to explore quite a few different ideas in designing an FDR control method. We have also studied variants of these two methods by altering one or more component of (a)-(c). For example, we have considered the knockoff filter using least-squares as the ranking algorithm, and we have also investigated different ways of constructing fake variables in knockoff. For Gaussian mirror, we propose a de-randomized version of the method, and we also propose a hybrid of Gaussian mirror and knockoff by combining their construction of tampered design. We hope our results will shed light on power analysis of many other FDR control methods.

1.1.1 THE THEORETICAL FRAMEWORK AND RELATED LITERATURE

We study a challenging regime of “Rare and Weak signals” (Donoho & Jin, 2015, Jin & Ke, 2016), where for some constants $\vartheta \in (0, 1)$ and $r > 0$, we consider settings where

$$\text{number of nonzero } \beta_j \sim p^{1-\vartheta}, \quad \text{magnitude of nonzero } \beta_j \sim n^{-1/2} \sqrt{2r \log(p)}. \quad (1.2)$$

The two parameters, ϑ and r , characterize the signal rarity and signal weakness, respectively. Here, $n^{-1/2} \sqrt{\log(p)}$ is the minimax order for successful inference of the support of β (Genovese et al., 2012), and the constant factor r drives subtle phase transitions. When $n = 1$, the setting (1.2) has been commonly used in the literature of multiple testing (e.g., Donoho & Jin (2004), Jager & Wellner (2007), Cai et al. (2007), Hall & Jin (2010), Arias-Castro et al. (2011), Barnett et al. (2017)). Recently, this setting has been considered in the study of variable selection for sparse linear models (e.g., Ji & Jin

(2012), Jin et al. (2014), Ke et al. (2014)).

We study the power of FDR control methods under the above Rare/Weak signal setting. For any method, its power changes with the target FDR level q . Instead of fixing q , we derive a trade-off diagram between FDR and the true positive rate (TPR) as q varies. This trade-off diagram provides a full characterization of power, given any model parameters (\mathfrak{S}, r) . We also derive a phase diagram (Jin & Ke, 2016) for each FDR control method. The phase diagram is a partition of the two-dimensional space (\mathfrak{S}, r) into three regions, *region of no recovery (NR)*, *region of almost full recovery (AFR)*, and *region of exact recovery (ER)*, where the asymptotic behavior of the Hamming error, defined as the expected sum of false positives and false negatives, is different in different regions. The boundary between NR and AFR is related to the achievability of asymptotically full power under FDR control, and the boundary between AFR and ER is connected to the achievability of model selection consistency. The phase diagram is a visualization of power of an FDR control method for all (\mathfrak{S}, r) together.

Power analysis of FDR control methods is a small body of literature. Su et al. (2017) set up a framework for studying the trade-off between false positive rate and true positive rate across the lasso solution path. Weinstein et al. (2017) and Weinstein et al. (2020) extended this framework to find a trade-off for the knockoff filter, when the ranking algorithm is the Lasso and thresholded Lasso, respectively. These trade-off diagrams are for linear sparsity (i.e., the number of nonzero coefficients of β is a constant fraction of p), which is a limit of our Rare/Weak setting as $\mathfrak{S} \rightarrow 0$. Under linear sparsity, the phase transition happens when $|\beta_j| \asymp n^{-1/2}$, and the FDR takes constant values. In the current chapter, we consider a different sparsity framework in which the number of signals is much smaller than p . We thus need a higher signal strength at the individual coefficient level, and the phase transition happens when $|\beta_j| \asymp n^{-1/2} \sqrt{\log(p)}$. Note that the overall signal strength as characterized by $\|\beta\|$ in our framework is actually much smaller than that in the aforementioned work. The FDR is a negative power of p , and so we draw the trade-off diagram in the log scale. Additionally, these works only considered the uncorrelated design, but our framework can accommodate correlated designs.

For correlated designs, Liu & Rigollet (2019) investigated sufficient and necessary conditions on X such that the knockoff has a full power, but they do not give the explicit trade-off diagram; furthermore, what they studied in the paper is not the orthodox knockoff but a variant using de-biased Lasso as the ranking algorithm. Beyond linear sparsity, Fan et al. (2019) studied the power of model-X knockoff for arbitrary sparsity, but they required a stronger signal strength by assuming $|\beta_j| \gg n^{-1/2} \sqrt{\log(p)}$. In a similar setting, Javanmard & Javadi (2019) studied the power of using de-biased Lasso directly as an FDR control method. Our work differs from these literature because we study the Rare/Weak signal setting (1.2) and derive explicit FDR-TPR trade-off diagrams and phase diagrams.

In our analysis, we develop a new technical tool. It relates the rates of convergence of variable selection errors with the geometry of the “rejection region” induced by an FDR control method. Consequently, the analysis of FDR-TPR trade-off diagram and phase diagram reduces to (i) deriving the rejection region and (ii) studying its geometric properties. This new tool will be useful for studying other problems under the Rare/Weak signal setting.

1.1.2 MAIN DISCOVERIES

We give a high-level summary of the discoveries. We use phase diagram as the main criterion of power comparison because a single phase diagram covers the whole parameter range (in contrast, the FDR-TPR trade-off diagram is tied to a specified (ϑ, r)). We say two methods have the “same power” if their associated phase diagrams are the same, and we say one method has a “higher power” than another if the phase diagram of the latter is inferior to that of the former. The precise statements will be given in Sections 1.2-1.5.

As mentioned, we are interested in the role of the three components, (a) ranking algorithm, (b) tampered design, and (c) symmetric statistic.

ROLE OF COMPONENT (A) We use the ranking algorithm to define a *prototype* for each FDR control method. The prototype runs the ranking algorithm on the original design matrix to obtain importance metrics for variables and then applies an ideal threshold (practically infeasible) to control FDR at a targeted level. We discover that the power of an FDR control method is primarily determined by the power of its prototype. We focus on two methods, knockoff and Gaussian mirror. The prototype of knockoff is a ranking method based on the lasso solution path (called “Lasso-path”), and the prototype of Gaussian mirror is a method that ranks variables by least-squares coefficients (called “least-squares”). The power comparison between knockoff and Gaussian mirror is largely the power comparison between Lasso-path and least-squares. Which prototype has a higher power depends on correlations in the design as well as the sparsity level of regression coefficients. Typically, Lasso-path is better when ϑ is large (i.e., β is sparser), and least-squares is better when ϑ is small (i.e., β is less sparse). See Section 1.4.

ROLE OF COMPONENT (C) Two commonly used symmetric statistics in knockoff are the signed maximum and the difference. It appears that using the difference as the symmetric statistic yields a considerable power loss relative to its prototype, even in the orthogonal design. In contrast, using the signed maximum as the symmetric statistic can successfully prevent power loss for a class of designs. Barber & Candès (2015) commented on the signed maximum as “*a specific instance that we find to perform well empirically.*” Our result is a theoretical justification to their numerical observation. We also provide a geometric interpretation, which suggests that the signed maximum is indeed the “best” choice among all possible symmetric statistics. See Section 1.3.

ROLE OF COMPONENT (B) The construction of the *tampered design matrix* usually involves adding fake variables (i.e., “noise”). A natural concern is whether “adding noise” for the purpose of FDR control reduces power. We first consider orthogonal designs. We show that the phase diagrams of

knockoff and Gaussian mirror (using signed maximum as symmetric statistics) are the same as the optimal phase diagram. This suggests that “adding noise” to achieve FDR control yields negligible power loss for orthogonal designs. See Section 1.3.

We then consider non-orthogonal designs. For these designs, even the prototypes of knockoff and Gaussian mirror may have non-optimal power (Ke et al., 2014). Therefore, it makes more sense to compare the power of an FDR control method with its own prototype. The answer for Gaussian mirror is relatively clear. For a wide class of designs, we show that the Gaussian mirror has negligible power loss compared with its prototype, least-squares. See Section 1.5.

The study of knockoff is much more demanding because the Lasso solution path has no explicit form. To get tractable results, we restrict to a class of block-wise diagonal designs: In this design matrix, p variables are divided into $p/2$ pairs, where variables in distinct pairs are uncorrelated, and variables in the same pair have a correlation of $\rho \in (-1, 1)$. We show that there exists a constant $\rho_0 \approx -0.35$, such that: If $\rho \in (\rho_0, 1)$, knockoff and Lasso-path share the same phase diagram; if $\rho \in (-1, \rho_0)$, they have the same phase transitions only when \mathfrak{D} is appropriately large. The discrepancy of power between knockoff and Lasso-path can be mitigated by modifying the tampered design in knockoff. We consider a variant of knockoff, where the tampered design follows the construction in Liu & Rigollet (2019) (called conditional-independence knockoff). We show that the conditional-independence knockoff and Lasso-path share the same phase diagram for every $\rho \in (-1, 1)$.

Since the ranking algorithm in knockoff can be replaced by least-squares, we also make a direct comparison of knockoff and Gaussian mirror by fixing the ranking algorithm as least-squares. We find that the phase diagram of Gaussian mirror is better than that of knockoff, and the main reason is that Gaussian mirror adds 1 fake variable at a time while knockoff adds p fake variables simultaneously. It motivates us to propose a general principle of constructing fake variables that suits for the “one-at-a-time” scheme. We call the resulting FDR control method the “de-randomized Gaussian mirror.” It turns out that the fake variables in knockoff suit for one-at-a-time scheme, which gives rise to a

new FDR control method that is a hybrid of Gaussian mirror and knockoff. We show that this new method improves the brute-forth “knockoff plus least-squares” and attains the same phase diagram as its prototype for a broad class of designs. On the other hand, the one-at-a-time scheme is limited to using least-squares to rank, and it does not apply to the original “knockoff plus Lasso-path.” See Section 1.5.

1.1.3 ORGANIZATION

The remainder of this chapter is organized as follows. Section 1.2 introduces the Rare/Weak signal model and explains how to use it as a theoretical platform to study and compare FDR control methods. Sections 1.3-1.5 contain the main results, where Section 1.3 studies the power of FDR control methods for orthogonal designs, Section 1.4 investigates the prototypes of FDR control methods, and Section 1.5 studies the power of FDR control methods for non-orthogonal designs. Section 1.6 sketches the proof and explains the geometrical insight behind the proof. Section 1.7 contains simulation results, and Section 1.8 concludes with a short discussion. Detailed proofs are relegated to the Supplementary Material.

1.2 FDR CONTROL METHODS AND CRITERIA OF POWER COMPARISON

Consider a linear regression model, $y = X\beta + \varepsilon$, where $y \in \mathbb{R}^n$, $X = [X_1, X_2, \dots, X_n]' \in \mathbb{R}^{n \times p}$, and $\varepsilon \sim N(0, \sigma^2 I_n)$. Throughout this chapter, we fix $\sigma = 1$. The Gram matrix is

$$G = X'X \in \mathbb{R}^{p \times p}, \quad \text{where we assume } G_{jj} = 1, \text{ for all } 1 \leq j \leq p. \quad (1.3)$$

Here each column of X is normalized to have a unit ℓ^2 -norm. Such a normalization is common in the study of Rare/Weak setting but is different from the standard normalization where each column of X has an ℓ^2 -norm of \sqrt{n} . The β vector in our setting is actually the vector of $\sqrt{n}\beta$ in a standard

normalization. In this chapter, we only consider the setting that $n > p$ and that the design is non-random, but the results are extendable to the setting that $n < p$ and that the rows of X are iid drawn from a multivariate Gaussian distribution.

We adopt the Rare/Weak signal model (Donoho & Jin, 2004) to assume that β satisfies:

$$\beta_j \stackrel{iid}{\sim} (1 - \varepsilon_p)\nu_0 + \varepsilon_p\nu_{\tau_p}, \quad 1 \leq j \leq p, \quad (1.4)$$

where ν_a denotes a point mass at a . Here, $\varepsilon_p \in (0, 1)$ is the expected fraction of signals, and $\tau_p > 0$ is the signal strength. We let p be the driving asymptotic parameter and tie (ε_p, τ_p) with p through fixed constants $\vartheta \in (0, 1)$ and $r > 0$:

$$\varepsilon_p = p^{-\vartheta}, \quad \tau_p = \sqrt{2r \log(p)}. \quad (1.5)$$

The parameters, ϑ and r , characterize the signal rarity and the signal weakness, respectively.

1.2.1 THE KNOCKOFF FILTER AND GAUSSIAN MIRROR

The knockoff filter (Barber & Candès, 2015) creates a design matrix $\tilde{X} \in \mathbb{R}^{n \times p}$ such that $\tilde{X}'\tilde{X} = G$ and $X'\tilde{X} = G - \text{diag}(s)$, where $G = X'X$ and $\text{diag}(s)$ is a nonnegative diagonal matrix satisfying that $\text{diag}(s) \preceq 2G$. The j -th column of \tilde{X} is called a *knockoff* of variable j . Let $\hat{\beta}(\lambda) \in \mathbb{R}^{2p}$ be the solution of running Lasso on the expanded design matrix $[X, \tilde{X}]$:

$$\hat{\beta}(\lambda) = \underset{b}{\text{argmin}} \{ \|y - [X, \tilde{X}]b\|^2 / 2 + \lambda \|b\|_1 \}.$$

For each $1 \leq j \leq p$, let $Z_j = \sup\{\lambda > 0 : \hat{\beta}_j(\lambda) \neq 0\}$ and $\tilde{Z}_j = \sup\{\lambda > 0 : \hat{\beta}_{p+j}(\lambda) \neq 0\}$. The importance of variable j is measured by a *symmetric statistic*

$$W_j = f(Z_j, \tilde{Z}_j), \quad (1.6)$$

where $f(\cdot, \cdot)$ is a bivariate function satisfying $f(v, u) = -f(u, v)$. Here $\{W_j\}_{j=1}^p$ are (signed) importance metrics for variables. Under some regularity conditions, it can be shown that W_j has a symmetric distribution when $\beta_j = 0$ and that W_j is positive with high probability when $\beta_j \neq 0$. Hence, given a threshold $t > 0$, the number of false discoveries is estimated by $\#\{j : W_j < -t\}$, and the data-driven threshold to control FDR at q is

$$T_1(q) = \min \left\{ t > 0 : \frac{\#\{j : W_j < -t\}}{\#\{j : W_j > t\} \vee 1} \leq q \right\}.$$

This method falls into the framework we introduced in Section 1.1. The ranking algorithm uses Lasso solution path to assign an importance metric to each variable, the tampered design is the $n \times (2p)$ matrix $[X, \tilde{X}]$, and the symmetric statistic is defined in (1.6). The ultimate importance metrics W_j are obtained by first applying the ranking algorithm on the tampered design and then re-combining the output via the symmetric statistic.

The Gaussian mirror (Xing et al., 2019) creates two columns $x_j^\pm = x_j \pm c_j z_j$ for each variable j , where $z_j \sim N(0, I_n)$ is sampled independently from data and $c_j = \|(I_n - P_{-j})x_j\| / \|(I_n - P_{-j})z_j\|$, where P_{-j} is the projection matrix to the column space of X_{-j} . Let $\hat{\beta}_j^\pm$ be the ordinary least-squares coefficients of x_j^\pm by regressing y on

$$\tilde{X}^{(j)} = [x_1, \dots, x_{j-1}, x_j^+, x_j^-, x_{j+1}, \dots, x_p].$$

The importance of variable j is measured by the *mirror statistic*:

$$M_j = |\hat{\beta}_j^+ + \hat{\beta}_j^-| - |\hat{\beta}_j^+ - \hat{\beta}_j^-|. \quad (1.7)$$

The construction of x_j^\pm ensures that M_j has a symmetric distribution when $\beta_j = 0$ and that M_j is positive with high probability when $\beta_j \neq 0$. The data-driven threshold to control FDR at q is

$$T_2(q) = \min \left\{ t > 0 : \frac{\#\{j : M_j < -t\}}{\#\{j : M_j > t\} \vee 1} \leq q \right\}.$$

Again, this method follows the framework in Section 1.1. The ranking algorithm uses least-squares coefficients to rank variables, the tampered design is the $n \times (p+1)$ matrix $\tilde{X}^{(j)}$ for each $1 \leq j \leq p$, and the symmetric statistic is as in (1.7). Different from knockoff, Gaussian mirror adds 1 fake variable at a time. When applying the ranking algorithm to the tampered design, Gaussian mirror solves p linear models, each with $(p+1)$ variables, while knockoff solves 1 linear model with $2p$ variables.

1.2.2 THE FDR-TPR TRADE-OFF DIAGRAM AND THE PHASE DIAGRAM

Under the Rare/Weak signal model (1.4)-(1.5), we define two diagrams for characterizing the power of an FDR control method. Let I_j be the importance metric assigned to variable j by the FDR control method, and consider the set of selected variables at a threshold $\sqrt{2u \log(p)}$:

$$\hat{S}(u) = \{1 \leq j \leq p : I_j > \sqrt{2u \log(p)}\}.$$

Let $S = \{1 \leq j \leq p : \beta_j \neq 0\}$. Define $\text{FP}_p(u) = \mathbb{E}(|\hat{S}(u) \setminus S|)$, $\text{FN}_p(u) = \mathbb{E}(|S \setminus \hat{S}(u)|)$, and $\text{TP}_p(u) = \mathbb{E}(|S \cap \hat{S}(u)|)$, where the expectation is taken with respect to the randomness of both β

and y . Write $s_p = p\varepsilon_p$. Define

$$\text{Hamm}_p(u) = \text{FP}_p(u) + \text{FN}_p(u), \quad \text{FDR}_p(u) = \frac{\text{FP}_p(u)}{\text{FP}_p(u) + \text{TP}_p(u)}, \quad \text{TPR}_p(u) = \frac{\text{TP}_p(u)}{s_p}.$$

The first quantity is the expected Hamming selection error. The last two quantities are proxy of the false discovery rate and true positive rate, respectively.

Definition 1.2.1. Let L_p be a generic multi-log(p) term, which may change from occurrence to occurrence and satisfies that $L_p p^\delta \rightarrow \infty$ and $L_p p^{-\delta} \rightarrow 0$ as $p \rightarrow \infty$ for any $\delta > 0$.

In the Rare/Weak signal model, fixing an FDR control method and a class of designs of interest, $\text{FDR}_p(u)$ and $\text{TPR}_p(u)$ often have the form: For any fixed (\mathfrak{J}, r, u) , as $p \rightarrow \infty$,

$$\text{FDR}_p(u) = L_p p^{-g_{\text{FDR}}(u; \mathfrak{J}, r)}, \quad 1 - \text{TPR}_p(u) = L_p p^{-g_{\text{TPR}}(u; \mathfrak{J}, r)}, \quad (1.8)$$

where $g_{\text{FDR}}(\cdot; \mathfrak{J}, r)$ and $g_{\text{TPR}}(\cdot; \mathfrak{J}, r)$ are two fixed functions, determined by the FDR control method and the design class. We propose the FDR-TPR trade-off diagram as follows:

Definition 1.2.2 (FDR-TPR trade-off diagram). Given an FDR control method and a sequence of designs indexed by p , if $\text{FDR}_p(u)$ and $\text{TPR}_p(u)$ satisfy (1.8) under the Rare/weak signal model (1.4)-(1.5), then the FDR-TPR trade-off diagram associated with (\mathfrak{J}, r) is the plot with $g_{\text{FDR}}(u; \mathfrak{J}, r)$ in the y-axis and $g_{\text{TPR}}(u; \mathfrak{J}, r)$ in the x-axis, as u varies.

An FDR-TPR trade-off diagram is tied to a particular (\mathfrak{J}, r) . To compare the performance of two FDR control methods, we need to draw many curves for different values of (\mathfrak{J}, r) . Here we introduce another metric for characterizing the power of an FDR control method at all (\mathfrak{J}, r) simultaneously. Define $\text{Hamm}_p^* \equiv \min_u \{\text{FP}_p(u) + \text{FN}_p(u)\}$. This is the minimum expected Hamming selection error when the threshold u is chosen optimally. We will see that for each method and each class of

designs of interest in this chapter, there exists a fixed bivariate function $f_{\text{Hamm}}^*(\mathcal{D}, r)$ such that, for any fixed (\mathcal{D}, r) , as $p \rightarrow \infty$,

$$\text{Hamm}_p^* = L_p p^{f_{\text{Hamm}}^*(\mathcal{D}, r)}. \quad (1.9)$$

Definition 1.2.3 (Phase diagram). Given an FDR control method and a sequence of designs indexed by p , if Hamm_p^* satisfies (1.9), then the phase diagram is a partition of the space (\mathcal{D}, r) into three regions:

- Region of Exact Recovery (ER): $\{(\mathcal{D}, r) : f_{\text{Hamm}}^*(\mathcal{D}, r) < 0\}$.
- Region of Almost Full Recovery (AFR): $\{(\mathcal{D}, r) : 0 < f_{\text{Hamm}}^*(\mathcal{D}, r) < 1 - \mathcal{D}\}$.
- Region of No Recovery (NR): $\{(\mathcal{D}, r) : f_{\text{Hamm}}^*(\mathcal{D}, r) \geq 1 - \mathcal{D}\}$.

The curves separating different regions are called phase curves. We use $h_{\text{AR}}(\mathcal{D})$ to denote the curve between NR and AFR, and $h_{\text{ER}}(\mathcal{D})$ the curve between AFR and ER.

In the ER region, the expected Hamming error, Hamm_p^* , tends to zero. As a result, with an overwhelming probability, the support of β is exactly recovered. In the AFR region, Hamm_p^* does not tend to zero but is much smaller than $p\varepsilon_p$ (which is the expected number of signals). As a result, with an overwhelming probability, the majority of signals are correctly recovered. In the region of NR, Hamm_p^* is comparable with the number of signals, and variable selection fails. The phase diagram was introduced in the literature (Genovese et al., 2012, Ji & Jin, 2012) but has never been used to study FDR control methods.

We illustrate these definitions with an example where we apply the BH-procedure to the marginal regression coefficients to control FDR at a targeted level. In this example,

$$I_j = |x_j' y|, \quad 1 \leq j \leq p. \quad (1.10)$$

The following proposition is proved in the supplementary material. Throughout this chapter, we use a_+ to denote $\max\{a, 0\}$, for any $a \in \mathbb{R}$.

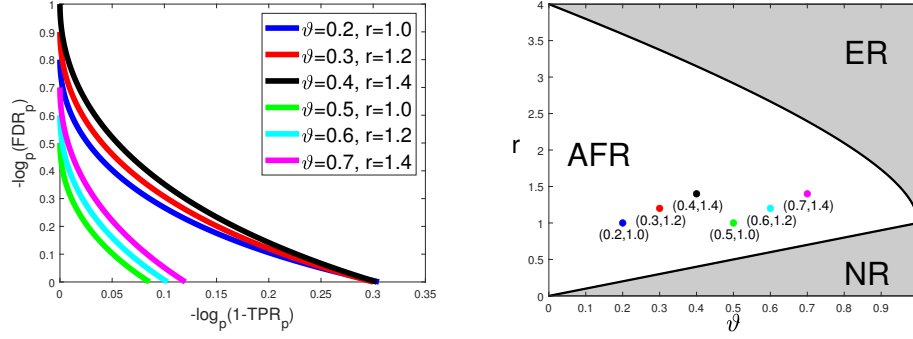


Figure 1.1: The FDR-TPR trade-off diagram (left) and the phase diagram (right) for the FDR control method in (1.10) under orthogonal designs. Each FDR-TPR trade-off diagram corresponds to one point in the phase diagram.

Proposition 1.2.1. Fix the FDR control method as in (1.10), and consider a sequence of orthogonal designs, that is, $X'X = I_p$.

- When $r > \vartheta$, the FDR-TPR trade-off diagram associated with (ϑ, r) is $g_{\text{FDR}}(\mathbf{u}; \vartheta, r) = (u - \vartheta)_+$ and $g_{\text{TPR}}(\mathbf{u}; \vartheta, r) = (\sqrt{r} - \sqrt{u})_+^2$.
- The phase diagram is such that $b_{\text{AR}}(\vartheta) = \vartheta$ and $b_{\text{ER}}(\vartheta) = (1 + \sqrt{1 - \vartheta})^2$.

These diagrams are shown in Figure 1.1.

Remark 1. The FDR-TPR trade-off diagram and the phase diagram are determined only by the importance metrics assigned to variables (i.e., the way variables are ranked). Although in many real applications feature ranking is often of the primary interest, another important aspect of an FDR control method is to derive a threshold so as to achieve the targeted FDR level q accurately. Thus, the power of an FDR-controlled feature selection method is affected not only by its ability of ranking the features properly, but also by its ability of estimating the FDR. We feel that, without a good ability in ranking features, a method may be of little interest to practitioners even if it can control the FDR well. It is desirable, however, to have a method that compromises with only a little loss of power in exchange of a precise FDR control. It is known that knockoff can control FDR precisely if certain conditions

about X are satisfied and Gaussian mirror can control FDR asymptotically. Thus, the power analysis in this chapter focuses only on the comparison of feature ranking abilities of different FDR control methods.

1.3 POWER ANALYSIS OF FDR CONTROL METHODS FOR ORTHOGONAL DESIGNS

Given an FDR control method that follows the unified framework in Section 1.1, we define its *prototype* as the method that assigns an importance metric to each variable by applying the ranking algorithm on the original design matrix X (in comparison, the FDR control method applies the ranking algorithm on the tampered design matrix and then re-combines the output through symmetric statistics). It is generally infeasible to estimate a proper threshold to control FDR based on the importance metrics given by the prototype. We use the prototype as a benchmark.

The solution of Lasso is defined by $\hat{\beta}^{\text{lasso}}(\lambda) = \operatorname{argmin}_b \{ \|y - Xb\|^2/2 + \lambda \|b\|_1 \}$. The prototype of knockoff assigns an importance metric to variable j as

$$W_j^* = \sup \{ \lambda > 0 : \hat{\beta}_j^{\text{lasso}}(\lambda) \neq 0 \}. \quad (1.11)$$

We call this method the *Lasso-path*. Let $\hat{\beta}^{\text{ols}} = \operatorname{argmin}_b \{ \|y - Xb\|^2 \}$ be the ordinary least squares estimator. The prototype of Gaussian mirror assigns an importance metric to variable j as

$$M_j^* = |\hat{\beta}_j^{\text{ols}}| = |e_j' G^{-1} X' y|. \quad (1.12)$$

We call this method the *least-squares*. In an orthogonal design, $X'X = I_p$. Both W_j^* and M_j^* reduce to the absolute marginal regression coefficient in (1.10). Therefore, we use the FDR-TPR trade-off diagram and the phase diagram in Figure 1.1 as the benchmark for the respective diagram of each FDR control method.

First, we study the knockoff filter. This method involves constructing a matrix \tilde{X} such that $\tilde{X}'\tilde{X} = G$ and $X'\tilde{X} = G - \text{diag}(s)$. We consider the form

$$\text{diag}(s) = (1 - a)I_p, \quad \text{where } -1 < a < 1. \quad (1.13)$$

The value of a controls the correlation between a variable and its own knockoff variable. Let Z_j and \tilde{Z}_j be the same as in (1.6). Two commonly-used symmetric statistics are:

$$W_j^{\text{sgm}} = (Z_j \vee \tilde{Z}_j) \cdot \begin{cases} +1, & \text{if } Z_j > \tilde{Z}_j \\ -1, & \text{if } Z_j \leq \tilde{Z}_j \end{cases}, \quad \text{and} \quad W_j^{\text{dif}} = Z_j - \tilde{Z}_j. \quad (1.14)$$

We call the first one the *signed maximum* statistic and the second one the *difference* statistic. The next theorem is proved in the supplementary material.

Theorem 1.3.1 (Knockoff, orthogonal designs). Consider a linear regression model where β satisfies Models (1.4)-(1.5). Suppose $n \geq 2p$ and $G = I_p$. We construct \tilde{X} in the knockoff filter as in (1.13), for a constant $a \in (-1, 1)$. For any constant $u > 0$, let $\text{FP}_p(u)$ and $\text{FN}_p(u)$ be the expected numbers of false positives and false negatives, by selecting variables with $W_j > \sqrt{2u \log(p)}$. When W_j is the signed maximum statistic in (1.14), as $p \rightarrow \infty$,

$$\text{FP}_p(u) = L_p p^{1-u}, \quad \text{FN}_p(u) = L_p p^{1-\vartheta - \min\left\{\frac{(1-|a|)r}{2}, (\sqrt{r}-\sqrt{u})_+^2\right\}}.$$

When W_j is the difference statistic in (1.14), as $p \rightarrow \infty$,

$$\text{FP}_p(u) = L_p p^{1-u}, \quad \text{FN}_p(u) = L_p p^{1-\vartheta - \frac{(1-|a|)}{2}(\sqrt{r}-\sqrt{u})_+^2}.$$

Corollary 1.3.1. In the same setting of Theorem 1.3.1, when $r > \vartheta$, the FDR-TPR trade-off diagram

of the knockoff filter associated with (ϑ, r) is given by

$$g_{\text{FDR}}(u; \vartheta, r) = (u - \vartheta)_+, \quad g_{\text{TPR}}(u) = \begin{cases} \min\left\{\frac{(1-|a|r)}{2}, (\sqrt{r} - \sqrt{u})_+^2\right\}, & \text{if } W_j = W_j^{\text{sgn}}, \\ \frac{(1-|a|)}{2}(\sqrt{r} - \sqrt{u})_+^2 & \text{if } W_j = W_j^{\text{dif}}. \end{cases}$$

The phase diagram of the knockoff filter is given by

$$b_{\text{AR}}(\vartheta) = \vartheta, \quad b_{\text{ER}}(\vartheta) = \begin{cases} \max\left\{\frac{2-2\vartheta}{1-|a|}, (1 + \sqrt{1-\vartheta})^2\right\}, & \text{if } W_j = W_j^{\text{sgn}}, \\ \left(1 + \sqrt{\frac{2-2\vartheta}{1-|a|}}\right)^2, & \text{if } W_j = W_j^{\text{dif}}. \end{cases}$$

The FDR-TPR trade-off diagram and the phase diagram are shown in Figure 1.2.

A noteworthy observation is that the value of a in the construction of the tampered design matrix affects the power. The best choice is $a = 0$, which means that a variable is uncorrelated with its own knockoff variable. Another noteworthy observation is that the symmetric statistic plays a crucial role. The signed maximum is strictly better than the difference. In the end of this section, we will provide geometric insight to explain that the signed maximum is (almost) the only best choice.

If we fix $a = 0$ in (1.13) and use the signed maximum as the symmetric statistic, the phase diagram of knockoff is the same as the phase diagram in Figure 1.1. This means that, using phase diagram as the criterion for power comparison, knockoff has no power loss relative to its prototype. On the hand, the FDR-TPR trade-off diagram is different from that in Figure 1.1. From Theorem 1.3.1, we see that $(1 - \text{TPR}_p) = \text{FN}_p/s_p \geq L_p p^{-r/2}$. Therefore, the FDR-TRP trade-off curve is truncated at $r/2$ in the x-axis. For large ϑ , the trade-off curve hits zero before the x-axis reaches $r/2$, and the truncation has no impact. However, for small ϑ , the trade-off curve has changed due to the truncation. See Figure 1.2.

Next, we study the Gaussian mirror. Let $\hat{\beta}_j^\pm$ be the same as in (1.7). The importance metric

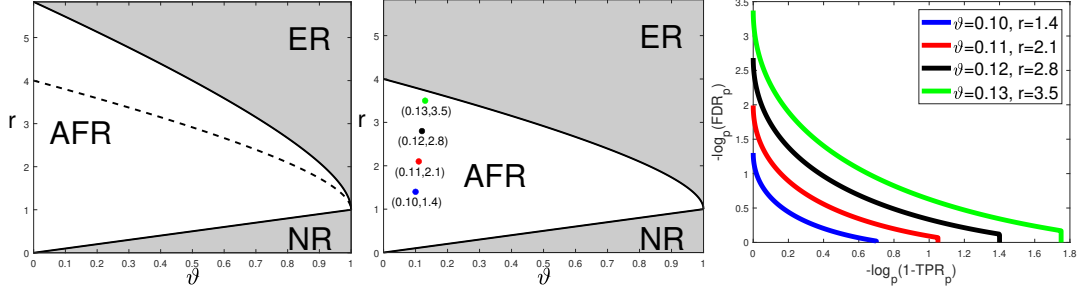


Figure 1.2: The power of knockoff ($a = 0$) and Gaussian mirror for orthogonal designs. The left and middle panels contain the variable selection phase diagrams, where the symmetric statistic is difference (left) and signed maximum (middle). The right panel contains the FDR-TPR trade-off diagram, where the symmetric statistic is signed maximum. Each FDR-TPR trade-off diagram corresponds to one point in the phase diagram in the middle panel.

assigned to variable j is the mirror statistic:

$$\mathcal{M}_j^{\text{dif}} = |\hat{\beta}_j^+ + \hat{\beta}_j^-| - |\hat{\beta}_j^+ - \hat{\beta}_j^-|. \quad (1.15)$$

It is reminiscent of the statistic $\mathcal{W}_j^{\text{dif}}$ in (1.14). Inspired by (1.14), we introduce a variant of the Gaussian mirror by replacing the mirror statistic by

$$\begin{aligned} \mathcal{M}_j^{\text{sgm}} &= (|\hat{\beta}_j^+ + \hat{\beta}_j^-| \vee |\hat{\beta}_j^+ - \hat{\beta}_j^-|) \cdot \begin{cases} +1, & \text{if } |\hat{\beta}_j^+ + \hat{\beta}_j^-| > |\hat{\beta}_j^+ - \hat{\beta}_j^-| \\ -1, & \text{if } |\hat{\beta}_j^+ + \hat{\beta}_j^-| \leq |\hat{\beta}_j^+ - \hat{\beta}_j^-| \end{cases} \\ &= (|\hat{\beta}_j^+| + |\hat{\beta}_j^-|) \cdot \text{sgn}(\hat{\beta}_j^+) \cdot \text{sgn}(\hat{\beta}_j^-). \end{aligned} \quad (1.16)$$

For this variant to be a valid FDR control method, we require that $\mathcal{M}_j^{\text{sgm}}$ has a symmetric distribution when $\beta_j = 0$. This can be verified easily. The following theorem is proved in the supplementary material.

Theorem 1.3.2 (Gaussian mirror, orthogonal designs). Consider a linear regression model where β satisfies Models (1.4)-(1.5). Suppose $n \geq p + p^\delta$ for a constant $\delta > 0$, and $G = I_p$. For any constant

$u > 0$, let $\text{FP}_p(u)$ and $\text{FN}_p(u)$ be the expected numbers of false positives and false negatives, by selecting all variables with $M_j > \sqrt{2u \log(p)}$, where M_j is the mirror statistic and the expectation here is taken with respect to the randomness of both y and z_1, z_2, \dots, z_p . When M_j is the difference statistic in (1.15), as $p \rightarrow \infty$,

$$\text{FP}_p(u) = L_p p^{1-u}, \quad \text{FN}_p(u) = L_p p^{1-\vartheta - \frac{1}{2}(\sqrt{r}-\sqrt{u})_+^2}.$$

When M_j is the signed maximum statistic in (1.16), as $p \rightarrow \infty$,

$$\text{FP}_p(u) = L_p p^{1-u}, \quad \text{FN}_p(u) = L_p p^{1-\vartheta - \min\left\{\frac{r}{2}, (\sqrt{r}-\sqrt{u})_+^2\right\}}.$$

Corollary 1.3.2. In the same setting of Theorem 1.3.2, the FDR-TPR trade-off diagram of the Gaussian mirror associated with (ϑ, r) is given by

$$g_{\text{FDR}}(u; \vartheta, r) = (u - \vartheta)_+, \quad g_{\text{TPR}}(u) = \begin{cases} \min\left\{\frac{r}{2}, (\sqrt{r}-\sqrt{u})_+^2\right\}, & \text{if } M_j = M_j^{\text{sgn}}, \\ \frac{1}{2}(\sqrt{r}-\sqrt{u})_+^2 & \text{if } M_j = M_j^{\text{dif}}. \end{cases}$$

The phase diagram of Gaussian mirror is given by

$$b_{\text{AR}}(\vartheta) = \vartheta, \quad b_{\text{ER}}(\vartheta) = \begin{cases} (1 + \sqrt{1 - \vartheta})^2, & \text{if } M_j = M_j^{\text{sgn}}, \\ (1 + \sqrt{2 - 2\vartheta})^2, & \text{if } M_j = M_j^{\text{dif}}. \end{cases}$$

Comparing Corollary 1.3.2 with Corollary 1.3.1, we find that Gaussian mirror and the knockoff with $a = 0$ (i.e., a variable is uncorrelated with its own knockoff variable) have the same FDR-TPR trade-off diagram and the same phase diagram when they both use signed maximum as the symmetric statistic. Similarly, they share the same phase diagram when they both use difference as the symmetric

statistic.

Last, we provide some geometric insight behind these results. Take knockoff for example. We abbreviate the knockoff using signed maximum and difference as symmetric statistic the *knockoff- sgm* and *knockoff- dif* , respectively. By default, we set $a = 0$ in (1.13). Under orthogonal designs, the ultimate importance metrics W_j can be written as $W_j = I(x'_j y, \tilde{x}'_j y)$, where x_j and \tilde{x}_j are the j th variable and its knockoff, and $I(\cdot, \cdot)$ is a fixed bivariate function. Define the “rejection region” as

$$\mathcal{R} = \left\{ (b_1, b_2) \in \mathbb{R}^2 : I\left(b_1 \sqrt{2 \log(p)}, b_2 \sqrt{2 \log(p)}\right) > \sqrt{2u \log(p)} \right\}.$$

Figure 1.3 shows the rejection region induced by knockoff- sgm , knockoff- dif , and their prototype (see (1.10)). Write $\hat{b}_1 = x'_j y / \sqrt{2 \log(p)}$ and $\hat{b}_2 = \tilde{x}'_j y / \sqrt{2 \log(p)}$. The random vector $(\hat{b}_1, \hat{b}_2)'$ follows a bivariate normal distribution with a covariance matrix $\frac{1}{\log(p)} I_2$, and a mean vector $(0, 0)'$ when $\beta_j = 0$ and $(\sqrt{r}, 0)'$ when $\beta_j = \tau_p$. By Lemma 1.6.1 (to be introduced in Section 1.6), the exponent in FP_p is determined by the Euclidean distance from $(0, 0)'$ to \mathcal{R} and the exponent in FN_p is determined by the Euclidean distance from $(\sqrt{r}, 0)'$ to \mathcal{R}^c . From Figure 1.3, it is clear that the difference statistic is inferior to the signed maximum statistic because the distance from $(\sqrt{r}, 0)'$ to \mathcal{R}^c is strictly smaller in the former.

The phase diagram of knockoff- sgm is the same as the phase diagram of the prototype. It suggests that signed maximum is already the “optimal” choice of symmetric statistic. Figure 1.3 also gives a geometric interpretation of why signed maximum is optimal. From (1.6) and that $(Z_j, \tilde{Z}_j) = (|x'_j y|, |\tilde{x}'_j y|)'$, we can derive necessary conditions for a subset \mathcal{R} to be an eligible rejection region (i.e., there exists a symmetric statistic whose induced rejection region is \mathcal{R}):

- (i) \mathcal{R} is symmetric with respect to both x-axis and y-axis.
- (ii) $\mathcal{R} \cap \mathcal{R}_\pm = \emptyset$, where \mathcal{R}_\pm is the reflection of \mathcal{R} with respect to the line $y = \pm x$.

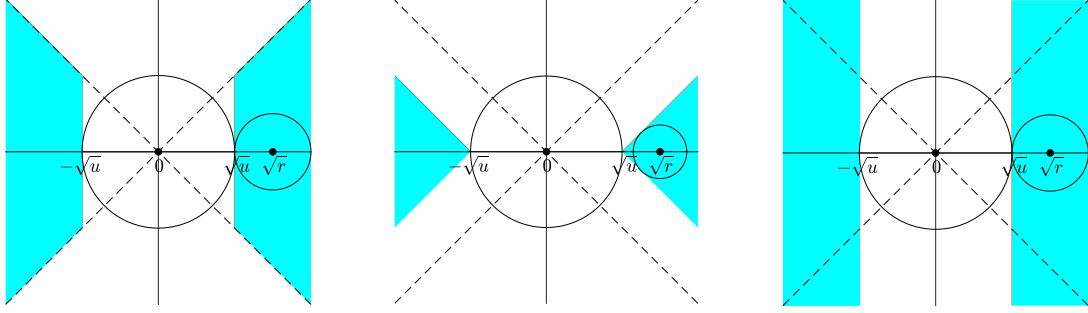


Figure 1.3: The rejection region of symmetric statistics (orthogonal design, $a = 0$ in the construction of knockoff). Left: the signed maximum statistic. Middle: the difference statistic. Right: the thresholding estimator in Section 1.2, which is used as a benchmark. In each plot, the x-axis is $x'_j y / \sqrt{2 \log(p)}$, and the y-axis: $\tilde{x}'_j y / \sqrt{2 \log(p)}$.

The rejection region \mathcal{R}_0 of the prototype (Figure 1.3, right panel) does not satisfy requirement (ii). The rejection region of knockoff-sgm (left panel) is a *minimal* modification of \mathcal{R}_0 to tailor to requirement (ii). From this perspective, it is almost impossible to find a symmetric statistic better than signed maximum.

1.4 BEHAVIOR OF THE PROTOTYPES FOR NON-ORTHOGONAL DESIGNS

The power of an FDR control method is related to (i) the power of its prototype and (ii) the difference of power between this method and its prototype. For orthogonal designs, the prototypes of knockoff and Gaussian mirror both reduce to the simple method in (1.10). However, for non-orthogonal designs, their prototypes can have different behaviors, which we study in this section. To save space, from now on, we only present the phase diagram. The FDR-TPR trade-off diagram can be easily derived from the expressions of $FP_p(u)$ and $FN_p(u)$, so we omit it.

We are often interested in a class of block-wise diagonal designs. For a fixed $\rho \in (-1, 1)$, Gram

matrix $G \in \mathbb{R}^{p \times p}$ satisfies that $G = \text{diag}(B, B, \dots, B, B_1)$, where

$$B = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}, \quad \text{and} \quad B_1 = \begin{cases} B, & \text{if } p \text{ is even,} \\ 1, & \text{if } p \text{ is odd.} \end{cases} \quad (1.17)$$

It is a theoretical idealization of the block-wise covariance structure in many real data (e.g., in genetics and bioinformatics). In this class of designs, the level of correlations is characterized by a single parameter ρ , so that it is possible to get a tractable form of the rate of convergence of variable selection errors.

First, we consider the prototype of Gaussian mirror. It uses least-squares coefficients to assign an importance metric M_j^* to variable j ; see (1.12). We call this method the least-squares. The following theorem is proved in the supplementary material.

Theorem 1.4.1 (Least-squares, general designs). Consider a linear regression model where β satisfies Models (1.4)-(1.5). Suppose $n \geq p$. For any constant $u > 0$, let $\text{FP}_p(u)$ and $\text{FN}_p(u)$ be the expected numbers of false positives and false negatives, by selecting variables with $M_j^* > \sqrt{2u \log(p)}$. Let $\omega_j > 0$ be the j -th diagonal element of the inverse of the Gram matrix (note that the Gram matrix has been normalized to have its diagonal elements equal to 1). Suppose $\omega_j \leq C_0$, for all $1 \leq j \leq p$, where $C_0 > 0$ is a constant. As $p \rightarrow \infty$,

$$\text{FP}_p(u) \leq L_p \sum_{j=1}^p p^{-\omega_j^{-1}u}, \quad \text{FN}_p(u) \leq L_p p^{-\vartheta} \sum_{j=1}^p p^{-\omega_j^{-1}(\sqrt{r}-\sqrt{u})_+^2}.$$

In the special case where G is the block-wise diagonal matrix as in (1.17) with a constant $\rho \in (-1, 1)$, as $p \rightarrow \infty$,

$$\text{FP}_p(u) = L_p p^{1-(1-\rho^2)u}, \quad \text{FN}_p(u) = L_p p^{1-\vartheta-(1-\rho^2)(\sqrt{r}-\sqrt{u})_+^2}.$$

Corollary 1.4.1. In the same setting of Theorem 1.4.1, consider a special case where G is the block-wise

diagonal matrix as in (1.17). The phase diagram of least-squares is given by

$$b_{\text{AR}}(\vartheta) = \frac{\vartheta}{1 - \rho^2}, \quad b_{\text{ER}}(\vartheta) = \frac{(1 + \sqrt{1 - \vartheta})^2}{1 - \rho^2}.$$

Figure 1.4 (left panel) shows the phase diagram for $|\rho| = 0.5$.

Next, we consider the prototype of knockoff. It utilizes the solution path of Lasso to assign an importance metric W_j^* to variable j ; see (1.11). We call it the Lasso-path. This method is difficult to characterize for a general design. We focus on the block-wise design (1.17).

Theorem 1.4.2 (Lasso-path, block-wise diagonal designs). Consider a linear regression model where β satisfies Models (1.4)-(1.5). Suppose $n \geq p$ and G is a block-wise diagonal matrix as in (1.17) with a constant $\rho \in (-1, 1)$. For any constant $u > 0$, let $\text{FP}_p(u)$ and $\text{FN}_p(u)$ be the expected numbers of false positives and false negatives, by selecting variables with $W_j^* > \sqrt{2u \log(p)}$. As $p \rightarrow \infty$,

$$\text{FP}_p(u) = L_p p^{1 - \min\{u, \vartheta + (\sqrt{u} - |\rho| \sqrt{r})^2 + (\xi_\rho \sqrt{r} - \eta_\rho \sqrt{u})_+^2 - (\sqrt{r} - \sqrt{u})_+^2\}},$$

and

$$\text{FN}_p(u) = \begin{cases} L_p p^{1 - \vartheta - \{(\sqrt{r} - \sqrt{u})_+ - [(1 - \xi_\rho) \sqrt{r} - (1 - \eta_\rho) \sqrt{u}]_+\}^2}, & \rho \geq 0, \\ L_p p^{1 - \min\{\vartheta + \{(\sqrt{r} - \sqrt{u})_+ - [(1 - \xi_\rho) \sqrt{r} - (1 - \eta_\rho) \sqrt{u}]_+\}^2, 2\vartheta + (\xi_\rho \sqrt{r} - \eta_\rho^{-1} \sqrt{u})_+^2\}}, & \rho < 0, \end{cases}$$

where $\xi_\rho = \sqrt{1 - \rho^2}$ and $\eta_\rho = \sqrt{(1 - |\rho|)/(1 + |\rho|)}$.

Corollary 1.4.2. In the same setting of Theorem 1.4.2, the phase diagram of Lasso-path is given by

$$b_{\text{AR}}(\vartheta) = \vartheta, \quad b_{\text{ER}}(\vartheta) = \begin{cases} \max\{b_1(\vartheta), b_2(\vartheta)\}, & \text{when } \rho \geq 0, \\ \max\{b_1(\vartheta), b_2(\vartheta), b_3(\vartheta)\}, & \text{when } \rho < 0, \end{cases}$$

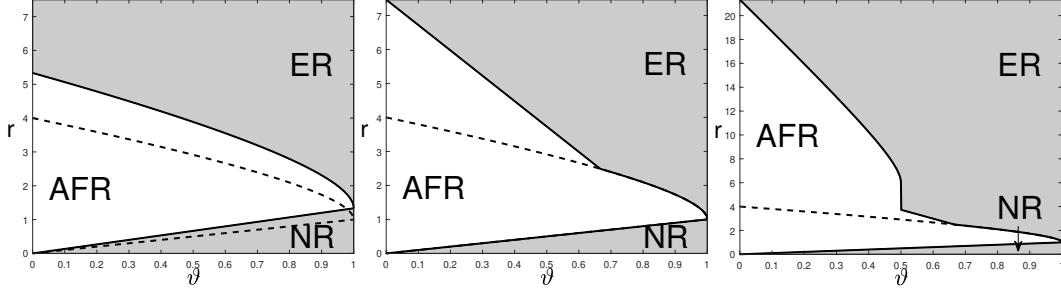


Figure 1.4: The phase diagrams for block-wise diagonal designs. Left: least-squares ($\rho = \pm 0.5$). Middle: Lasso-path ($\rho = 0.5$). Right: Lasso-path ($\rho = -0.5$). Least-squares and Lasso-path are the prototypes of Gaussian mirror and knockoff, respectively.

where $b_1(\vartheta) = (1 + \sqrt{1 - \vartheta})^2$, $b_2(\vartheta) = (1 + \sqrt{\frac{1 + \rho}{1 - \rho}})^2 (1 - \vartheta)$, and $b_3(\vartheta) = \frac{1}{(1 + \rho)^2} (\sqrt{\frac{1 + \rho}{1 - \rho}} \sqrt{1 - 2\vartheta} + \sqrt{\frac{1 - \rho}{1 + \rho}} \sqrt{1 - \vartheta})^2 \cdot 1\{\vartheta < 1/2\}$.

Figure 1.4 (middle and right panels) shows the phase diagrams for $\rho = \pm 0.5$.

We compare the two prototypes for block-wise diagonal designs.

- In terms of $b_{\text{AR}}(\vartheta)$, Lasso-path is always better than least-squares. To achieve Almost Full Recovery, Lasso-path only requires $r > \vartheta$, but least-squares requires $r > \vartheta / (1 - \rho^2)$.
- In terms of $b_{\text{ER}}(\vartheta)$, Lasso-path is better than least-squares when ϑ is relatively large (i.e., β is comparably sparser), and least-squares is better than Lasso-path when ϑ is relatively small (i.e., β is comparably denser).
- The sign of ρ also matters. For small ϑ , the advantage of least-squares over Lasso-path on $b_{\text{ER}}(\vartheta)$ is much more obvious when ρ is negative.

In Section 1.6, we will provide a geometric interpretation to the above statement. Here we give an intuitive explanation. We say a signal variable (i.e., $\beta_j \neq 0$) is ‘isolated’ if it is the only signal variable in the 2×2 block, and we say two signals are ‘nested’ if they are in the same 2×2 block. In the sparser regime (i.e., ϑ is large), least-squares has a disadvantage because it is inefficient in discovering

an ‘isolated’ signal. In the less sparse regime (i.e., ϑ is small), Lasso-path has a disadvantage because it suffers from signal cancellation when estimating a pair of ‘nested’ signals (‘signal cancellation’ means a signal variable has a weak marginal correlation with y due to the effect of other signals correlated with this one).

For broader design classes, similar phenomenons are observed empirically (Xing et al., 2019). In Section 1.7, we show simulations on various design classes, where the insight here continues to apply.

Remark 2. There is a duality between setting a negative ρ in the block-wise diagonal design and allowing for negative entries in β . We modify the Rare/Weak signal model to $\beta_j \stackrel{iid}{\sim} (1 - \varepsilon_p)\nu_0 + (\varepsilon_p/2)\nu_{\tau_p} + (\varepsilon_p/2)\nu_{-\tau_p}$, for $1 \leq j \leq p$. Under this model, by a similar proof, we can show that, for block-wise diagonal designs parametrized by ρ and any given method, the exponent in $\text{FP}_p(u)$ (or $\text{FN}_p(u)$) is the maximum of the two previous exponents in $\text{FP}_p(u)$ (or $\text{FN}_p(u)$) corresponding to $\pm|\rho|$. Consequently, the phase diagram is equal to the worse of the previous two phase diagrams associated with $\pm|\rho|$. With this being said, even for applications where the correlations are all positive, our study of a negative ρ is still useful, because it helps understand the case of allowing for positive and negative signs in β .

Remark 3. The phase diagram for Lasso-path is connected to the phase diagram for Lasso in Ji & Jin (2012) but is different in important ways. They considered using Lasso (with a proper tuning parameter λ) for variable selection, but we considered using the solution path of Lasso to rank variables. The results and the analysis are both different.

1.5 POWER ANALYSIS OF FDR CONTROL METHODS FOR NON-ORTHOGONAL DESIGNS

In Section 1.4, we investigate the prototypes of FDR control methods. In this section, we compare them with their prototypes. In light of the study in Section 1.3, we always use the signed maximum as the symmetric statistic.

1.5.1 RANKING BY LEAST-SQUARES

In this subsection, we study FDR control methods whose prototype is least-squares. The first method is Gaussian mirror.

Theorem 1.5.1 (Gaussian mirror, general designs). Consider a linear regression model where β satisfies Models (1.4)-(1.5). Suppose $n \geq p + p^\delta$, for a constant $\delta > 0$. For any constant $u > 0$, let $\text{FP}_p(u)$ and $\text{FN}_p(u)$ be the expected numbers of false positives and false negatives, by selecting variables with $M_j > \sqrt{2u \log(p)}$, where M_j is the signed maximum statistic in (1.16) and the expectation here is taken with respect to the randomness of y and z_1, z_2, \dots, z_p . Let $\omega_j > 0$ be the j -th diagonal of the inverse of the Gram matrix. Suppose $\omega_j \leq C_0$, for all $1 \leq j \leq p$, where $C_0 > 0$ is a constant. As $p \rightarrow \infty$,

$$\text{FP}_p(u) \leq L_p \sum_{j=1}^p p^{-\omega_j^{-1}u}, \quad \text{FN}_p(u) \leq L_p p^{-\vartheta} \sum_{j=1}^p p^{-\omega_j^{-1} \min\{(\sqrt{r}-\sqrt{u})_+^2, \frac{1}{2}r\}}.$$

In the special case where G is the block-wise diagonal matrix as in (1.17) with a constant $\rho \in (-1, 1)$, as $p \rightarrow \infty$,

$$\text{FP}_p(u) = L_p p^{1-(1-\rho^2)u}, \quad \text{FN}_p(u) = L_p p^{1-\vartheta-(1-\rho^2) \min\{(\sqrt{r}-\sqrt{u})_+^2, \frac{1}{2}r\}}.$$

Compare Theorem 1.5.1 with Theorem 1.4.1: The rate of convergence for $\text{FP}_p(u)$ is the same, and the rate of convergence for $\text{FN}_p(u)$ has a minor difference. This minor difference has no impact on the rate of convergence of $\text{FP}_p(u) + \text{FN}_p(u)$, and thus no impact on the phase diagram. The next corollary confirms that, for block-wise diagonal designs, the phase diagram of Gaussian mirror matches with that of its prototype.

Corollary 1.5.1. Under the same setting as Theorem 1.5.1, consider a special case where G is the block-

wise diagonal matrix as in (1.17). For Gaussian mirror, the phase curves are the same as those in Corollary 1.4.1.

The second method is knockoff-OLS. Knockoff can accommodate different ranking algorithms, not limited to Lasso-path. We use least-squares here. Same as before, let $\tilde{X} \in \mathbb{R}^{n \times p}$ be such that $\tilde{X}'\tilde{X} = G$ and $X'\tilde{X} = G - \text{diag}(s)$. Let $\hat{\beta}_j$ and $\tilde{\beta}_j$ be the respective least-squares coefficient of x_j and \tilde{x}_j by regressing y on $[X, \tilde{X}]$. Define $Z_j = |\hat{\beta}_j|$ and $\tilde{Z}_j = |\tilde{\beta}_j|$. The importance metric W_j is computed from (Z_j, \tilde{Z}_j) in the same way as W_j^{sgm} in (1.14).

Theorem 1.5.2 (Knockoff-OLS). Consider a linear regression model where β satisfies Models (1.4)-(1.5). Suppose $n \geq 2p$. We apply the knockoff filter and use least-squares as the ranking algorithm. For any constant $u > 0$, let $\text{FP}_p(u)$ and $\text{FN}_p(u)$ be the expected numbers of false positives and false negatives, by selecting variables with $W_j > \sqrt{2u \log(p)}$. Let $G^* = [X, \tilde{X}]'[X, \tilde{X}] \in \mathbb{R}^{2p \times 2p}$, and let $A_j \in \mathbb{R}^{2 \times 2}$ be the submatrix of $(G^*)^{-1}$ restricted to the j th and $(j+p)$ th rows and columns. Denote $\omega_{1j} = A_j(1, 1)$ and $\omega_{2j} = A_j(1, 2)$. Suppose $\omega_{1j} \leq C_0$, for all $1 \leq j \leq p$, where $C_0 > 0$ is a constant. As $p \rightarrow \infty$,

$$\text{FP}_p(u) \leq L_p \sum_{j=1}^p p^{-\omega_{1j}^{-1}u}, \quad \text{FN}_p(u) \leq L_p p^{-s} \sum_{j=1}^p p^{-\omega_{1j}^{-1} \min\{(\sqrt{r}-\sqrt{u})_+^2, \frac{\omega_{1j}}{\omega_{1j}+|\omega_{2j}|} \cdot \frac{1}{2}r\}}.$$

By Theorem 1.5.2 and elementary calculations, the phase diagram of knockoff-OLS is governed by the quantities ω_{1j} . In comparison, by Theorem 1.5.1, the phase diagram of Gaussian mirror is governed by the quantities ω_j . We compare ω_{1j} and ω_j . Recall that they are the j th diagonal elements of G^{-1} and $(G^*)^{-1}$, respectively. Since G is a principal submatrix of G^* , by elementary linear algebra,

$$\omega_j \leq \omega_{1j}.$$

The inequality is often strict, e.g., see Corollary 1.5.2 below. It suggests that the phase diagram of

Gaussian mirror is better than that of knockoff-OLS. We will show that such difference is primarily due to that Gaussian mirror uses a one-at-a-time scheme of adding fake variables.

The third method is a new FDR control method that can be viewed as a variant of Gaussian mirror by removing randomness in the tampered design. We call it “de-randomized Gaussian mirror.” This method creates a design matrix $\tilde{X} \in \mathbb{R}^{n \times p}$ and regresses y on

$$\tilde{X}^{(j)} = [x_1, \dots, x_{j-1}, x_j^+, x_j^-, x_{j+1}, \dots, x_p], \quad \text{where } x_j^\pm = x_j \pm \tilde{x}_j.$$

Let $\hat{\beta}_j^\pm$ be the least-square coefficients of x_j^\pm . The mirror statistic of variable j is defined by

$$\mathcal{M}_j = (|\hat{\beta}_j^+| + |\hat{\beta}_j^-|) \cdot \text{sgn}(\hat{\beta}_j^+) \cdot \text{sgn}(\hat{\beta}_j^-). \quad (1.18)$$

This is similar to $\mathcal{M}_j^{\text{sgm}}$ in (1.16). Given $\{\mathcal{M}_j\}_{j=1}^p$, we can micmic the procedure in Section 1.2.1 to find a data-driven threshold that controls FDR at a targeted level. The next lemma gives a sufficient condition on \tilde{X} such that the above method stays valid for FDR control.

Lemma 1.5.1. In a linear regression model $y = X\beta + \mathcal{N}(0, \sigma^2 I_n)$, let $P_{-j} \in \mathbb{R}^{n \times n}$ be the projection matrix to the column space of X_{-j} , $1 \leq j \leq p$. Suppose the following conditions are satisfied:

- $\|(I - P_{-j})\tilde{x}_j\| = \|(I - P_{-j})x_j\|$, for each $1 \leq j \leq p$.
- There exist constants $C > 0$ and $\delta \in (0, 2)$ such that, for the set of null features $\mathcal{T} = \{j : \beta_j \neq 0\}$, $\#\{(j, k) \in \mathcal{T}^2 : j \neq k, (x_j, \tilde{x}_j)'(I - P_{-j})(I - P_{-k})(x_k, \tilde{x}_k) \neq \mathbf{0}_{2 \times 2}\} \leq C|\mathcal{T}|^\delta$.

Then, the de-randomized Gaussian mirror yields asymptotically valid FDR control.

Here, $\|(I - P_{-j})\tilde{x}_j\| = \|(I - P_{-j})x_j\|$ is the key requirement. It guarantees that \mathcal{M}_j has a sym-

metric distribution when $\beta_j = 0$. The orthodox Gaussian mirror uses a random \tilde{X} :

$$\tilde{x}_j = \frac{\|(I_n - P_{-j})x_j\|}{\|(I_n - P_{-j})z_j\|} z_j, \quad \text{where } z_j \sim N(0, I_n) \text{ is independent of } X.$$

It automatically satisfies that $\|(I - P_{-j})\tilde{x}_j\| = \|(I - P_{-j})x_j\|$. Alternatively, we can always construct a non-random \tilde{X} to satisfy this equation. The next theorem characterizes the power of de-randomized Gaussian mirror:

Theorem 1.5.3 (De-randomized Gaussian mirror). Consider a linear regression model where β satisfies Models (1.4)-(1.5). Suppose $n \geq 2p$ and we are given a matrix $\tilde{X} \in \mathbb{R}^{n \times p}$ such that $\|(I - P_{-j})\tilde{x}_j\| = \|(I - P_{-j})x_j\|$ for all $1 \leq j \leq p$. We apply the de-randomized Gaussian mirror. For any constant $u > 0$, let $\text{FP}_p(u)$ and $\text{FN}_p(u)$ be the expected numbers of false positives and false negatives, by selecting variables with $M_j > \sqrt{2u \log(p)}$, where M_j is as in (1.18). Let $\tilde{G}^{(j)} = [x_1, \dots, x_j, \tilde{x}_j, \dots, x_p][x_1, \dots, x_j, \tilde{x}_j, \dots, x_p]' \in \mathbb{R}^{(p+1) \times (p+1)}$, and let $D_j \in \mathbb{R}^{2 \times 2}$ be the submatrix of $(\tilde{G}^{(j)})^{-1}$ restricted to the j th and $(j+1)$ th rows and columns. Denote $\sigma_{1j} = D_j(1, 1)$ and $\sigma_{2j} = D_j(1, 2)$. Suppose $\sigma_{1j} \leq C_0$, for all $1 \leq j \leq p$, where $C_0 > 0$ is a constant. As $p \rightarrow \infty$,

$$\text{FP}_p(u) \leq L_p \sum_{j=1}^p p^{-\sigma_{1j}^{-1} u}, \quad \text{FN}_p(u) \leq L_p p^{-s} \sum_{j=1}^p p^{-\sigma_{1j}^{-1} \min\{(\sqrt{r} - \sqrt{u})_+^2, \frac{\sigma_{1j}}{\sigma_{1j} + |\sigma_{2j}|} \cdot \frac{1}{2} r\}}.$$

There are many eligible choices of \tilde{X} . We are particularly interested in using the \tilde{X} from knockoff.

Re-write

$$\|(I - P_{-j})\tilde{x}_j\|^2 = \tilde{x}_j' \tilde{x}_j - \tilde{x}_j' X_{-j} (X_{-j}' X_{-j})^{-1} X_{-j}' \tilde{x}_j.$$

The \tilde{X} from knockoff satisfies that $\tilde{x}_j' \tilde{x}_j = x_j' x_j$ and $\tilde{x}_j' X_{-j} = x_j' X_{-j}$. It is easy to see that $\|(I - P_{-j})\tilde{x}_j\| =$

$\|(I - P_{-j})x_j\|$. We can thus use this \tilde{X} in de-randomized Gaussian mirror.* It gives rise to a “hybrid” of knockoff and Gaussian mirror.

Fixing \tilde{X} to be the matrix from knockoff, we compare least-squares, knockoff-OLS, and derandomized Gaussian mirror. By Theorems 1.4.1 and 1.5.2-1.5.3, their phase diagrams are governed by ω_j , ω_{1j} , and σ_{1j} , respectively. Note that $(\omega_j, \sigma_{1j}, \omega_{1j})$ are the respective j th diagonal element of G^{-1} , $(\tilde{G}^{(j)})^{-1}$ and $(G^*)^{-1}$. Since that G is a principal submatrix of $\tilde{G}^{(j)}$ and that $\tilde{G}^{(j)}$ is a principal submatrix of G^* , we immediately have

$$\omega_j \leq \sigma_{1j} \leq \omega_{1j}.$$

Therefore, with the same \tilde{X} , the phase diagram of knockoff-OLS is always no better than that of de-randomized Gaussian mirror. Now, it is clear that the advantage of Gaussian mirror over knockoff-OLS is essentially from the one-at-a-time scheme of incorporating fake variables. Given the same collection of fake variables, knockoff-OLS enrolls all of them simultaneously while de-randomized Gaussian mirror enrolls one at a time. The more variables included in a linear regression, the larger variance of an individual least-squares coefficient. This explains that adding 1 fake variable at a time is a better strategy.

Lemma 1.5.2. Given two matrices $X \in \mathbb{R}^{n \times p}$ and $\tilde{X} \in \mathbb{R}^{n \times p}$, let ω_j and σ_{1j} be the same as in Theorem 1.5.1 and Theorem 1.5.3. For each $1 \leq j \leq p$, if $x_j'(I - P_{-j})\tilde{x}_j = 0$, then $\sigma_{1j} = \omega_j$ and $\sigma_{2j} = 0$. Furthermore, if $x_j'(I - P_{-j})\tilde{x}_j = 0$ for all $1 \leq j \leq p$, then this choice of \tilde{X} minimizes both $\text{FP}_p(u)$ and $\text{FN}_p(u)$ of de-randomized Gaussian mirror, for any $u > 0$.

By Lemma 1.5.2, the best option of \tilde{X} is such that $x_j'(I - P_{-j})\tilde{x}_j = 0$, i.e., the projections of x_j and \tilde{x}_j onto the orthogonal complement of X_{-j} are mutually orthogonal. In the orthodox Gaussian

*We need some regularity conditions on X to ensure that the second bullet point of Lemma 1.5.1 is satisfied. When \tilde{X} is from knockoff, a sufficient condition is that the Gram matrix restricted to noise variables is a block-wise diagonal matrix, where the size of the largest block is $\leq Cp^{1-a}$ for some constants $a \in (0, 1)$ and $C > 0$.

mirror, $\tilde{x}_j \propto z_j$, where $z_j \sim N(0, I_n)$ is drawn independently from x_j and X_{-j} . It can be shown that $x_j'(I - P_{-j})\tilde{x}_j \approx 0$, as long as $n - p \geq p^\delta$, for any constant $\delta > 0$. This explains why the phase diagram of Gaussian mirror matches with that of least-squares. There are many possible ways of constructing a non-random \tilde{X} such that $x_j'(I - P_{-j})\tilde{x}_j = 0$. If we construct \tilde{X} from knockoff, we can use the choice of $\text{diag}(s)$ suggested by [Liu & Rigollet \(2019\)](#):

$$\text{diag}(s) = [\text{diag}(G^{-1})]^{-1}. \quad (1.19)$$

They showed that the resulting \tilde{X} satisfies $x_j'(I - P_{-j})\tilde{x}_j = 0$ [†] and called this construction the *conditional-independence knockoff*.[‡] By matrix inversion formula, an equivalent expression of $\text{diag}(s)$ is $s_j = \|x_j\|^2 - \|P_{-j}x_j\|^2$, which implies that the covariance between x_j and its knockoff should be $\|P_{-j}x_j\|^2$. We exemplify this idea on the block-wise diagonal designs parametrized by $\rho \in (-1, 1)$, where (1.19) reduces to $\text{diag}(s) = (1 - \rho^2)I_p$.

Corollary 1.5.2. Under the same setting of Theorems 1.5.2-1.5.3, consider a special case where G is the block-wise diagonal matrix as in (1.17). We construct \tilde{X} from knockoff with $\text{diag}(s) = (1 - \rho^2)I_p$. The phase diagram of knockoff-OLS is given by

$$b_{\text{AR}}(\vartheta) = \frac{\vartheta}{(1 - \rho^2)^2}, \quad b_{\text{ER}}(\vartheta) = \frac{(1 + \sqrt{1 - \vartheta})^2}{(1 - \rho^2)^2}.$$

The phase diagram of de-randomized Gaussian mirror is given by

$$b_{\text{AR}}(\vartheta) = \frac{\vartheta}{1 - \rho^2}, \quad b_{\text{ER}}(\vartheta) = \frac{(1 + \sqrt{1 - \vartheta})^2}{1 - \rho^2}.$$

[†]Their equation (9) shows that, if $x_j'(I - P_{-j})\tilde{x}_j = 0$ for every j , then $\text{diag}(s)$ has to equal to $[\text{diag}(G^{-1})]^{-1}$. In fact, the opposite is also true. See the remark in the end of the proof of Lemma 1.5.2.

[‡]It is not guaranteed that $\text{diag}(s) \preceq 2G$. If this is violated, some truncation on $\text{diag}(s)$ may be needed.

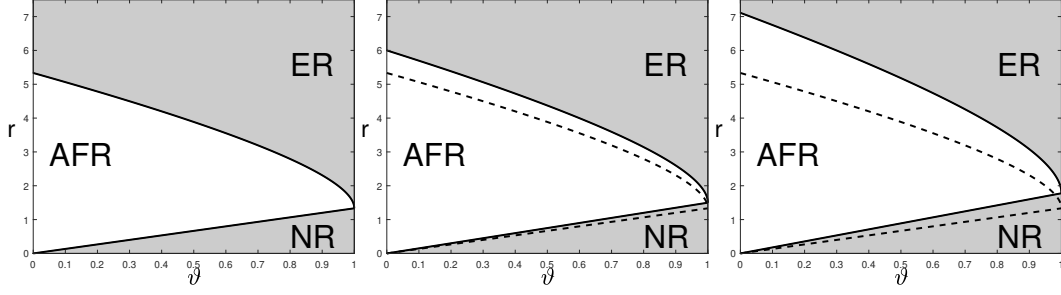


Figure 1.5: The phase diagrams of methods that use least-squares as the ranking algorithm (block-wise diagonal designs, $\rho = \pm 0.5$). Left: least-squares, Gaussian mirror, and de-randomized Gaussian mirror with CI-knockoff design (the three methods share the same phase diagram). Middle: de-randomized Gaussian mirror with SDP-knockoff design. Right: knockoff-OLS with CI-knockoff design.

Figure 1.5 shows the phase diagrams for $\rho = \pm 0.5$.

Remark 4. The main insight gained here is that the one-at-a-time scheme of incorporating fake variables (as in Gaussian mirror) yields a higher power than the p -at-a-time scheme (as in knockoff). However, we note that the one-at-a-time scheme is tied to using least-squares as the ranking algorithm. For a general ranking algorithm, the one-at-a-time scheme may not guarantee valid FDR control. In comparison, the p -at-a-time scheme is flexible to accommodate different ranking algorithms.

Remark 5. Another ranking algorithm that is closely related to least-squares is the debiased Lasso (see [Javanmard & Javadi \(2019\)](#) and references therein). The de-biased Lasso estimator is $\hat{\beta}^{dbLasso} = \hat{\beta} + \Omega X'(y - \hat{\beta})$, where $\hat{\beta}$ is the Lasso estimator and Ω is a matrix such that $\Omega \cdot \mathbb{E}[X'X] \approx I_p$. Under some regularity conditions, the asymptotic distribution of $\hat{\beta}_j^{dbLasso}$ is the same as that of $\hat{\beta}_j^{ols}$. Hence, the results in this subsection also shed light on the power of FDR control methods based on de-biased Lasso.

1.5.2 RANKING BY LASSO-PATH

In this subsection, we study FDR control methods whose prototype is Lasso-path. Since the solution path of Lasso has no tractable form, the analysis is much more demanding than that in Section 1.5.1.

We thereby restrict to the block-wise diagonal designs.

We consider knockoff. It involves choosing a diagonal matrix $\text{diag}(s)$. Two options are recommended in Barber & Candès (2015), the equi-correlated knockoff and the SDP knockoff. For block-wise diagonal designs as in (1.17), these two options are the same:

$$\text{diag}(s) = (1 - a)I_p, \quad \text{where } a = \begin{cases} 2|\rho| - 1, & |\rho| \geq 1/2, \\ 0, & |\rho| < 1/2. \end{cases} \quad (1.20)$$

When $|\rho| \geq 1/2$, the tampered design matrix $[X, \tilde{X}]$ is always singular. In this case, we can obtain the explicit rates of convergence of FP_p and FN_p .

Theorem 1.5.4 (Knockoff, block-wise diagonal designs, $|\rho| \geq 1/2$). Consider a linear regression model where β satisfies Models (1.4)-(1.5). Suppose $n \geq 2p$ and G is the block-wise diagonal matrix as in (1.17) with a constant ρ , where $|\rho| \geq 1/2$. We construct \tilde{X} in the knockoff filter with $\text{diag}(s)$ as in (1.20). For any constant $u > 0$, let $\text{FP}_p(u)$ and $\text{FN}_p(u)$ be the expected numbers of false positives and false negatives, by selecting variables with $W_j > \sqrt{2u \log(p)}$, where W_j is the signed maximum statistic in (1.14). As $p \rightarrow \infty$,

$$\text{FP}_p(u) = L_p p^{1 - \min\left\{u, \vartheta + (\sqrt{u} - |\rho|\sqrt{r})^2 + (\xi_\rho \sqrt{r} - \eta_\rho \sqrt{u})_+^2 - (\sqrt{r} - \sqrt{u})_+^2\right\}},$$

and for $\rho \geq 1/2$,

$$\text{FN}_p(u) = L_p p^{1 - \vartheta - \left\{(\sqrt{r} - \sqrt{u})_+ - [(1 - \xi_\rho)\sqrt{r} - (1 - \eta_\rho)\sqrt{u}]_+ - (\lambda_\rho \sqrt{r} - \eta_\rho \sqrt{u})_+\right\}^2},$$

and for $\rho \leq -1/2$,

$$\text{FN}_p(u) = L_p p^{1 - \min\left\{\vartheta + \left\{(\sqrt{r} - \sqrt{u})_+ - [(1 - \xi_\rho)\sqrt{r} - (1 - \eta_\rho)\sqrt{u}]_+ - (\lambda_\rho \sqrt{r} - \eta_\rho \sqrt{u})_+\right\}^2, 2\vartheta\right\}},$$

where $\xi_\rho = \sqrt{1-\rho^2}$, $\eta_\rho = \sqrt{(1-|\rho|)/(1+|\rho|)}$, and $\lambda_\rho = \sqrt{1-\rho^2} - \sqrt{1-|\rho|}$.

When $|\rho| < 1/2$, the tampered design matrix $[X, \tilde{X}]$ is non-singular. In this case, listing the separate forms of FP_p and FN_p is very tedious. We instead present the rate of convergence of $\text{FP}_p + \text{FN}_p$, which is sufficient for deriving the phase diagram.

Theorem 1.5.5 (Knockoff, block-wise diagonal designs, $|\rho| < 1/2$). Consider a linear regression model where β satisfies Models (1.4)-(1.5). Suppose $n \geq 2p$ and G is the block-wise diagonal matrix as in (1.17) with a constant ρ , where $|\rho| < 1/2$. We construct \tilde{X} in the knockoff filter with $\text{diag}(s)$ as in (1.20). For any constant $u > 0$, let $\text{FP}_p(u)$ and $\text{FN}_p(u)$ be the expected numbers of false positives and false negatives, by selecting variables with $W_j > \sqrt{2u \log(p)}$, where W_j is the signed maximum statistic in (1.14). As $p \rightarrow \infty$,

$$\text{FP}_p(u) + \text{FN}_p(u) = \begin{cases} L_p p^{1-f_{\text{Hamm}}^+(u, r, \vartheta)}, & 0 \leq \rho < 1/2, \\ L_p p^{1-\min\{f_{\text{Hamm}}^+(u, r, \vartheta), 2\vartheta + (\xi_\rho \sqrt{r} - \eta_\rho^{-1} \sqrt{u})_+^2, 2\vartheta + \frac{(1+2\rho)^2(1-\rho)}{2(1+\rho)} r\}}, & -1/2 < \rho < 0, \end{cases}$$

where

$$f_{\text{Hamm}}^+(u, r, \vartheta) = \min\{u, \vartheta + (\sqrt{u} - |\rho|\sqrt{r})^2 + ((\xi_\rho \sqrt{r} - \eta_\rho \sqrt{u})_+)^2 - ((\sqrt{r} - \sqrt{u})_+)^2, \\ \vartheta + [(\sqrt{r} - \sqrt{u})_+ - ((1 - \xi_\rho)\sqrt{r} - (1 - \eta_\rho)\sqrt{u})_+]^2\},$$

and ξ_ρ, η_ρ are the same as those in Theorem 1.5.4.

We combine Theorem 1.5.4 and Theorem 1.5.5 to obtain the phase diagram:

Corollary 1.5.3. In the same setting of Theorems 1.5.4-1.5.5, consider the knockoff filter with $\text{diag}(s)$

as in (1.20). Define

$$\rho_0 = \sqrt{2} - 1 - \sqrt{2 - \sqrt{2}} \quad (\text{note: } \rho_0 \approx -0.35).$$

The phase curve $b_{AR}(\mathcal{G}) = \mathcal{G}$. The phase curve $b_{ER}(\mathcal{G})$ has three cases:

- When $\rho \in [\rho_0, 1)$,

$$b_{ER}(\mathcal{G}) = b_{ER}^{LassoPath}(\mathcal{G}),$$

where $b_{ER}^{LassoPath}(\mathcal{G})$ is the phase curve in Corollary 1.4.2.

- When $\rho \in (-0.5, \rho_0)$,

$$b_{ER}(\mathcal{G}) = \max\{b_{ER}^{LassoPath}(\mathcal{G}), b_5(\mathcal{G})\}, \quad \text{where } b_5(\mathcal{G}) = \frac{2(1 - 2\mathcal{G})(1 + \rho)}{(1 + 2\rho)^2(1 - \rho)}.$$

- When $\rho \in (-1, -0.5]$,

$$b_{ER}(\mathcal{G}) = \begin{cases} b_{ER}^{LassoPath}(\mathcal{G}), & \mathcal{G} > 1/2, \\ \infty, & \mathcal{G} < 1/2. \end{cases}$$

Comparing Corollary 1.5.3 and Corollary 1.4.2, we observe that, when $\rho \in [\rho_0, 1)$, the phase diagram of knockoff is the same as that of Lasso-path. When $\rho \in (-1, \rho_0)$, the phase diagrams of two methods are different. Figure 1.6 shows the phase diagram of knockoff for different values of ρ . To see what causes the discrepancy of the phase diagram between knockoff and Lasso-path, we first look at the range of $\rho \in (-0.5, \rho_0)$. In this case, the construction in (1.20) guarantees that the j th knockoff is uncorrelated with the j th original variable. However, this knockoff is still highly correlated with the $(j + 1)$ th original variable. Suppose j is a true signal variable. Then, a true signal at $(j + 1)$ will increase

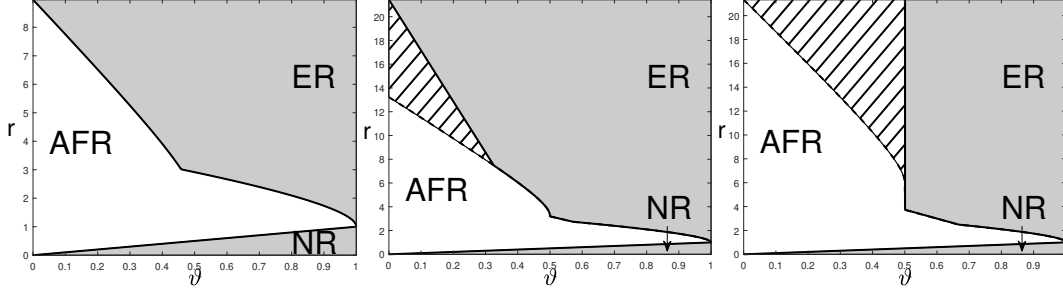


Figure 1.6: The phase diagrams of knockoff (SDP knockoff, symmetric statistic is signed maximum). The design is the block-wise diagonal design, where $\rho = -0.3$ (left), $\rho = -0.4$ (middle), and $\rho = -0.5$ (right), corresponding to the three cases in Corollary 1.5.3. The shadowed area is the Almost Full Recovery region for knockoff but Exact Recovery region for Lasso-path. If SDP-knockoff is replaced by CI-knockoff, then in each of three cases the phase diagram is the same as that of Lasso-path.

the absolute correlation between y and \tilde{x}_j but decrease the absolute correlation between y and x_j (since $\rho < 0$), making it more difficult for x_j to stand out.

We then look at the range of $\rho \in (-1, -0.5]$. In this range, the construction of knockoff variables changes to a different form (see (1.20)). This has a significant consequence on the phase curve $h_{ER}(\vartheta)$. To gain some insight, we look at a scenario of two ‘nested’ signals, i.e., $(\beta_j, \beta_{j+1}) = (\tau_p, \tau_p)$. By elementary calculation,

$$\mathbb{E}[x'_j y] = (1 + \rho)\tau_p, \quad \mathbb{E}[\tilde{x}'_j y] = \begin{cases} \rho\tau_p, & \text{when } -0.5 < \rho < 0, \\ -(1 + \rho)\tau_p, & \text{when } -1 < \rho \leq -0.5. \end{cases}$$

When $\rho \leq -0.5$, variable j and its knockoff have the same absolute correlation with y . Consequently, there is a non-diminishing probability that the true signal variable fails to dominate its knockoff variable, making it impossible to select j consistently. In the Rare/Weak signal model, ‘nested’ signals appear with a non-diminishing probability if $\vartheta < 1/2$. This explains why $h_{ER}(\vartheta) = \infty$ when $\rho \leq -0.5$ and $\vartheta < 1/2$.

The above issue can be resolved by modifying $\text{diag}(s)$. We take the conditional independence

knockoff in (1.19). For block-wise diagonal designs, it reduces to

$$\text{diag}(s) = (1 - \rho^2)I_p, \quad \text{for all } \rho \in (-1, 1). \quad (1.21)$$

We now revisit the scenario of two ‘nested’ signals, i.e., $(\beta_j, \beta_{j+1}) = (\tau_p, \tau_p)$. It is seen that

$$\mathbb{E}[x'_j y] = (1 + \rho)\tau_p, \quad \mathbb{E}[\tilde{x}'_j y] = \rho(1 + \rho)\tau_p.$$

The signal strength is always higher at the original variable than at its knockoff. We conclude that $h_{ER}(\vartheta) < \infty$ for all $\vartheta \in (0, 1)$. The next theorem gives the explicit rate of convergence of $\text{FP}_p + \text{FN}_p$ for this version of knockoff, which is proved in the supplementary material.

Theorem 1.5.6 (CI-Knockoff, block-wise diagonal designs). Consider a linear regression model where β satisfies Models (1.4)-(1.5). Suppose $n \geq 2p$ and G is the block-wise diagonal matrix as in (1.17) with a constant ρ . We construct \tilde{X} in the knockoff filter with $\text{diag}(s)$ as in (1.21). For any constant $u > 0$, let $\text{FP}_p(u)$ and $\text{FN}_p(u)$ be the expected numbers of false positives and false negatives, by selecting variables with $W_j > \sqrt{2u \log(p)}$, where W_j is the signed maximum statistic in (1.14). As $p \rightarrow \infty$,

$$\text{FP}_p(u) + \text{FN}_p(u) = \begin{cases} L_p p^{1-f_{\text{Hamm}}^+(u, r, \vartheta)}, & \rho \geq 0, \\ L_p p^{1-\min\{f_{\text{Hamm}}^+(u, r, \vartheta), 2\vartheta + (\frac{\xi}{\varrho} \sqrt{r} - \eta_\rho^{-1} \sqrt{u})_+^2\}}, & \rho < 0, \end{cases}$$

where $f_{\text{Hamm}}^+(u, r, \vartheta)$ is the same as that in Theorem 1.5.5.

It can be verified that the above rate of convergence for $\text{FP}_p(u) + \text{FN}_p(u)$ is the same as that in Theorem 1.4.2. We immediately know that CI-knockoff yields the same phase diagram as its prototype, Lasso-path.

Corollary 1.5.4. Under the same setting as Theorem 1.5.6, consider a special case where G is the block-wise diagonal matrix as in (1.17). For the conditional independence knockoff, the phase curves are the same as those in Corollary 1.4.2.

Our results show the advantage of CI-knockoff over SDP-knockoff for block-wise diagonal designs. It is an interesting question whether CI-knockoff can improve the phase diagram of SDP-knockoff for general designs. The theoretical study is extremely tedious. We instead investigate it numerically in Section 1.7.

1.6 THE PROOF IDEA AND GEOMETRIC INSIGHT

A key technical tool in the proof is the following lemma, which is proved in the supplementary material.

Lemma 1.6.1. Fix $d \geq 1$, a vector $\mu \in \mathbb{R}^d$, a covariance matrix $\Sigma \in \mathbb{R}^{d \times d}$, and an open set $S \subset \mathbb{R}^d$ such that $\mu \notin S$. Suppose $b \equiv \inf_{x \in S} \{(x - \mu)' \Sigma^{-1} (x - \mu)\} < \infty$. Consider a sequence of random vectors $X_p \in \mathbb{R}^d$, indexed by p , satisfying that

$$X_p | (\mu_p, \Sigma_p) \sim \mathcal{N}_d \left(\mu_p, \frac{1}{2 \log(p)} \Sigma_p \right),$$

where $\mu_p \in \mathbb{R}^d$ is a random vector and $\Sigma_p \in \mathbb{R}^{d \times d}$ is a random covariance matrix. As $p \rightarrow \infty$, suppose for any fixed $\gamma > 0$ and $L > 0$, $\mathbb{P}(\|\mu_p - \mu\| > \gamma) \leq p^{-L}$ and $\mathbb{P}(\|\Sigma_p - \Sigma\| > \gamma) \leq p^{-L}$. Then, as $p \rightarrow \infty$,

$$\mathbb{P}(X_p \in S) = L_p p^{-b}.$$

This lemma connects the rate of convergence of $\mathbb{P}(X_p \in S)$ with the geometric property of the set S . The exponent b is the “radius” of the largest ellipsoid that centers at μ and is fully contained in the complement of S .

We now illustrate how to use Lemma 1.6.1 to prove the claims in Sections 1.3-1.5. Take the proof of Theorem 1.4.1 for example. Consider the block-wise diagonal design in (1.17). Fix j and let W_j^* be as in (1.11). Write

$$\hat{b} = (x'_j y, x'_{j+1} y)' / \sqrt{2 \log(p)} \in \mathbb{R}^2. \quad (1.22)$$

It is not hard to see that W_j^* is purely determined by \hat{b} . Hence, the selection criteria $W_j^* > u$ can be equivalently written as $\hat{b} \in \mathcal{R}_u$, where \mathcal{R}_u is a subset in the two-dimensional space. We call it the “rejection region” of Lasso-path. The probabilities of a false positive and a false negative occurring at j are respectively

$$\mathbb{P}(\hat{b} \in \mathcal{R}_u, \beta_j = 0) \quad \text{and} \quad \mathbb{P}(\hat{b} \in \mathcal{R}_u^c, \beta_j = \tau_p).$$

Conditioning on β , the random vector \hat{b} has a bivariate normal distribution, whose mean is a constant vector and whose covariance matrix is $\frac{1}{2 \log(p)} B$, where B is the same as in (1.17). Applying Lemma 1.6.1, we reduce the proof into two steps:

- (i) Derive the rejection region \mathcal{R}_u .
- (ii) For each possible β with $\beta_j = 0$, obtain $b(\beta) \equiv \inf_{x \in \mathcal{R}_u} \{(x - \mu(\beta))' B^{-1} (x - \mu(\beta))\}$, and for each possible β with $\beta_j \neq 0$, obtain $b(\beta) \equiv \inf_{x \in \mathcal{R}_u^c} \{(x - \mu(\beta))' B^{-1} (x - \mu(\beta))\}$, where $\mu(\beta) \equiv \mathbb{E}[\hat{b} | \beta]$.

Both steps can be carried out by direct calculations.

We use a similar strategy to prove other theorems. The proof is sometimes complicated and tedious. For example, to analyze knockoff for block-wise diagonal designs, we have to consider the random vector $\hat{b} = (x'_j y, x'_{j+1} y, \tilde{x}'_j y, \tilde{x}'_{j+1} y)' / \sqrt{2 \log(p)} \in \mathbb{R}^4$. The proof requires deriving a 4-dimensional rejection region and calculating $b(\beta)$, for an arbitrary $\rho \in (-1, 1)$. The calculations are very tedious. To study Gaussian mirror, we have to deal with the randomness introduced by the algorithm. Let $\hat{b} = (x'_j y, \tilde{x}'_j y) / \sqrt{2 \log(p)}$, and let z_j be the random vector used to construct x_j^\pm in

Gaussian mirror. The covariance matrix of $\hat{b}|(\beta, z_j)$ depends on the realization of z_j , and we need to show that this matrix, scaled by $2 \log(p)$, converges to a fixed covariance matrix at a sufficiently fast rate.

As seen, our proof has a straightforward geometric visualization. We now use this geometric visualization to reveal some useful insight about the different performance of Lasso-path and least-squares. Their associated rejection regions in \mathbb{R}^2 are given by the following lemma. It is proved in the supplementary material.

Lemma 1.6.2. Consider a linear regression model, where the Gram matrix G is as in (1.17) with a constant $\rho \in (-1, 1)$. Let W_j^* and M_j^* be the same as in (1.11) and (1.12). Define

$$\begin{aligned} \mathcal{R}_u^{\text{path}}(\rho) &= \{(b_1, b_2) : b_1 - \rho b_2 > (1 - \rho)\sqrt{u}, b_1 > \sqrt{u}\} \\ &\quad \cup \{(b_1, b_2) : b_1 - \rho b_2 > (1 + \rho)\sqrt{u}\} \\ &\quad \cup \{(b_1, b_2) : b_1 - \rho b_2 < -(1 - \rho)\sqrt{u}, b_1 < -\sqrt{u}\} \\ &\quad \cup \{(b_1, b_2) : x - \rho y < -(1 + \rho)u\}, \quad \text{for } \rho \geq 0, \\ \mathcal{R}_u^{\text{path}}(\rho) &= \{(b_1, b_2) : (b_1, -b_2) \in \mathcal{R}_u^{\text{path}}(-\rho)\}, \quad \text{for } \rho < 0, \\ \mathcal{R}_u^{\text{ols}}(\rho) &= \{(b_1, b_2) : b_1 - \rho b_2 > (1 - \rho^2)\sqrt{u}\} \\ &\quad \cup \{(b_1, b_2) : b_1 - \rho b_2 < -(1 - \rho^2)\sqrt{u}\}. \end{aligned}$$

Let $\hat{b} = (x'_j y, x'_{j+1} y)' / \sqrt{2 \log(p)}$. Then, $W_j^* > \sqrt{2u \log(p)}$ if and only if $\hat{b} \in \mathcal{R}_u^{\text{path}}(\rho)$, and $M_j^* > u$ if and only if $\hat{b} \in \mathcal{R}_u^{\text{ols}}(\rho)$.

These rejection regions are shown in Figure 1.7. Their geometric properties are different for positive and negative ρ . Fix j . Let \hat{b} be as in (1.22), and write $\mu(\beta) = \mathbb{E}[\hat{b}|\beta]$.

- The rate of convergence of $\text{FP}_\rho(u)$ is determined by the largest ellipsoid that centers at $\mu(\beta)$ and is contained in \mathcal{R}_u^c . We call this ellipsoid the *FP-ellipsoid*.

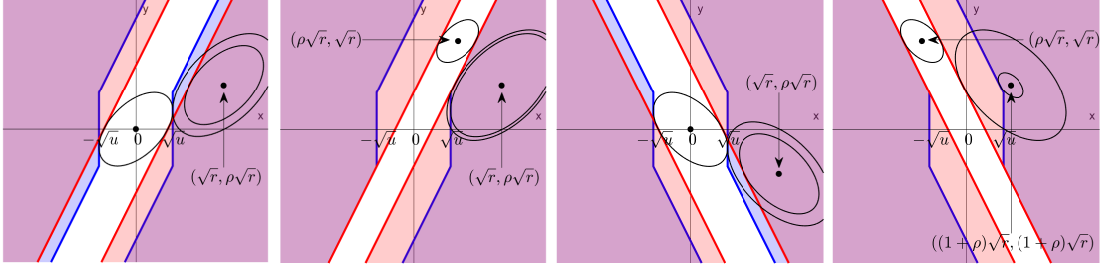


Figure 1.7: Rejection regions and ‘most-likely’ cases in block-wise diagonal designs (x-axis: $x'_j y / \sqrt{2 \log(p)}$; y-axis: $x'_{j+1} y / \sqrt{2 \log(p)}$). From left to right: (i) positive ρ and large ϑ , (ii) positive ρ and small ϑ , (iii) negative ρ and large ϑ , (iv) negative ρ and small ϑ . In each plot, the blue solid lines define rejection region of Lasso-path, and the red solid lines define rejection region of least-squares. For each method, FP_p is determined by the largest FP-ellipsoid in \mathcal{R}^c , and FN_p is determined by the largest FN-ellipsoid in \mathcal{R} , where the centers of these ellipsoids are determined by (β_j, β_{j+1}) in the ‘most-likely’ case. In each plot, the largest FP-ellipsoid is controlled to be the same for both Lasso-path and least-squares, and so the method with a larger FN-ellipsoid is better.

- The rate of convergence of $\text{FN}_p(u)$ is determined by the largest ellipsoid that centers at $\mu(\beta)$ and is contained in \mathcal{R}_u . We call this ellipsoid the *FN-ellipsoid*.

By direct calculations,

$$\mu(\beta) = (\beta_j + \rho\beta_{j+1}, \rho\beta_j + \beta_{j+1})' / \sqrt{2 \log(p)}.$$

It depends on β only through (β_j, β_{j+1}) . Under our model, (β_j, β_{j+1}) has 4 possible values to take: $\{(0, 0), (0, \tau_p), (\tau_p, 0), (\tau_p, \tau_p)\}$, where the first two correspond to a null at j and the last two correspond to a non-null at j . The probability of having a selection error at j thus splits into 4 terms, and which term is dominating depends on the values of ϑ and ρ . The realization of (β_j, β_{j+1}) that plays a dominating role is called the ‘most-likely’ case. For example, when ϑ is large (i.e., β is sparser), the most-likely case of a false positive occurring at j is when $(\beta_j, \beta_{j+1}) = (0, 0)$; when ϑ is small (i.e., β is less sparse), the most-likely case of a false positive is when $(\beta_j, \beta_{j+1}) = (0, \tau_p)$. The table below summarizes the ‘most-likely’ cases.

To see why Lasso-path and least-squares have different behavior, we visualize the ‘most-likely’ cases

Sparsity	Correlation	Error type	Most-likely case	Center of ellipsoid
large ϑ	positive/negative ρ	FP	$\beta_j = 0, \beta_{j+1} = 0$	$(0, 0)$
		FN	$\beta_j = \tau_p, \beta_{j+1} = 0$	$(\sqrt{r}, \rho\sqrt{r})$
small ϑ	positive ρ	FP	$\beta_j = 0, \beta_{j+1} = \tau_p$	$(\rho\sqrt{r}, \sqrt{r})$
		FN	$\beta_j = \tau_p, \beta_{j+1} = 0$	$(\sqrt{r}, \rho\sqrt{r})$
small ϑ	negative ρ	FP	$\beta_j = 0, \beta_{j+1} = \tau_p$	$(\rho\sqrt{r}, \sqrt{r})$
		FN	$\beta_j = \tau_p, \beta_{j+1} = \tau_p$	$((1 + \rho)\sqrt{r}, (1 + \rho)\sqrt{r})$

Table 1.1: The ‘most-likely’ cases and the corresponding ellipsoid center $\mathbb{E}[\hat{b}|\beta]$

for different (ρ, ϑ) in Figure 1.7. In each plot of Figure 1.7, we have coordinated the thresholds u in two methods so that the FP-ellipsoid is exactly the same. It suffices to compare the FN-ellipsoid: The method with a larger FN-ellipsoid has a faster rate of convergence on the Hamming error.

In Figure 1.7, it is clear that, when ϑ is large, the FN-ellipsoid of Lasso-path is larger; when ϑ is small, the FN-ellipsoid of least-squares is larger. This explains the different performance of two methods. Moreover, when ϑ is small, comparing the case of a positive ρ with the case of a negative ρ , we find that the difference between FN-ellipsoids of two methods are much more prominent in the case of a negative ρ . This explains why the sign of ρ matters.

1.7 SIMULATIONS

We use numerical experiments to support the theoretical results in Sections 1.3-1.5. In Experiments 1 and 2, we investigate orthogonal designs and the 2×2 block-wise diagonal designs, respectively. In Experiments 3-5, we investigate more design classes, including block-wise diagonal designs with larger blocks, factor models, exponentially decaying designs, and normalized Wishart designs. We consider four different ranking methods, Lasso-path (Lasso), least-squares (OLS), knockoff (KF) and Gaussian mirror (GM). For KF and GM, we use either signed maximum or difference as the symmetric statistic. For KF, we choose $\text{diag}(s) = \min\{1, 2\lambda_{\min}(G)\} \cdot I_p$, unless specified otherwise. It is called the equi-correlated knockoff (EC-KF), and is the same as the SDP-knockoff for orthogonal designs and the

2×2 block-wise diagonal designs. In Experiments 1-3, this is the only $\text{diag}(s)$ we use, and so we write EC-KF as KF for short. In Experiments 4-5, we also consider the conditional independence knockoff (CI-KF). For most experiments, fixing a parameter setting, we generate 200 data sets and record the averaged Hamming selection error among these 200 repetitions.

Experiment 1. We investigate the performance of different methods for orthogonal designs. Given $(n, p) = (2000, 1000)$, $\vartheta \in \{0.3, 0.5\}$ and r ranging on a grid from 0 to 6 with step size 0.2, we generate data y from $\mathcal{N}(X\beta, I_n)$ where X is an $n \times p$ matrix with unit length columns that are orthogonal to each other and β is generated from (1.4). We implemented Lasso and OLS, as well as KF and GM using both the signed maximum and difference as the symmetric statistic. Each method outputs p importance statistics, and we threshold these importance statistics at $\sqrt{2u^* \log(p)}$ where u^* minimizes $\text{FN}_p(u) + \text{FP}_p(u)$ in theory. The results are in Figure 1.8, where the y-axis is $\log_p(H_p/p)$, and H_p is the averaged Hamming selection error over 200 repetitions. For KF and GM, we also plot $\log_p(H_p^*/p)$ via solid lines, where H_p^* is $\text{FP}_p(u^*) + \text{FN}_p(u^*)$ excluding the multi- $\log(p)$ term L_p . It serves as a theoretical reference for H_p .

The theory in Section 1.3 suggests the following for orthogonal designs: (i) Regarding the choice of symmetric statistic, for both KF and GM, signed maximum outperforms difference. (ii) With signed maximum as the symmetric statistic, KF has a similar performance as Lasso, and GM has a similar performance as OLS. These theoretical results are perfectly validated by simulations (see Figure 1.8). We also notice that there is a discrepancy between the error curves and their theoretical reference curves. This is because we ignore the multi- $\log(p)$ term L_p in plotting the theoretical curves. Ignoring L_p causes an increase of $\asymp \log(\log(p))$ in the error curve, which is non-negligible for a moderately large p such as $p = 1000$. After taking L_p into account, the empirical and theoretical error curves are nicely aligned.

Experiment 2. We here consider the block-wise diagonal design with 2×2 blocks, where we

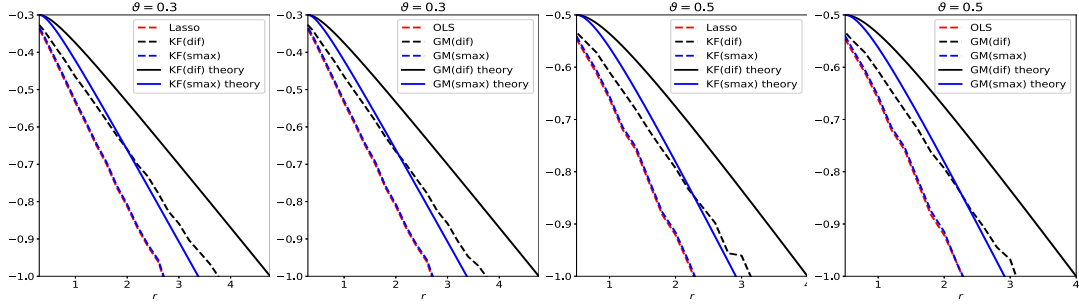


Figure 1.8: Experiment 1 (orthogonal designs). The y-axis is $\log_p(H_p/p)$, where H_p is the average Hamming error over 200 repetitions. The solid curves are the theoretical values from Section 1.3.

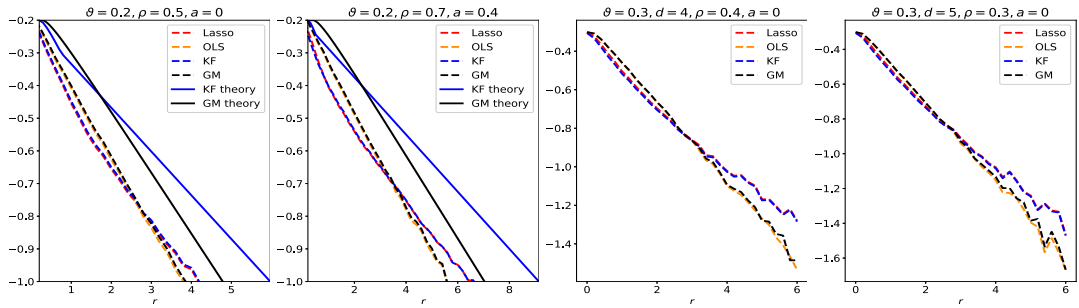


Figure 1.9: Experiments 2 and 3 (block-wise diagonal designs, d : block size, ρ : off-diagonal entries). The y-axis is $\log_p(H_p/p)$, where H_p is the average Hamming error. The parameter a controls the construction of knock-off. The solid curves are the theoretical values from Section 1.5.

take $\rho = 0.5$ and $\rho = 0.7$. In the data generation, we fix an $n \times p$ matrix X such that $X'X$ has the desirable form. We then generate (β, y) in the same way as before. For each ρ , we fix $(n, p, \vartheta) = (2000, 1000, 0.2)$, and let r range on a grid from 0 to 8 with a step size 0.2. For KF and GM, we now fix the symmetric statistic as signed maximum. In KF, the default choice of $\text{diag}(s)$ yields that $\text{diag}(s) = (1 - a)I_p$ with $a = 2\rho - 1$. The results are in the first two panels of Figure 1.9.

The theory in Section 1.5 suggests the following for block-wise diagonal designs: (i) GM has a similar performance as its prototype, OLS. (ii) Since the two values of ρ considered here are in $(\rho_0, 1)$, KF has a similar performance as its prototype, Lasso. The simulation results are consistent with these theoretical predictions. Additionally, from the theoretical reference curves, we can see that, for the current ϑ value, GM has a smaller Hamming error than that of KF when r is large, and the opposite

is true when r is small. The actual error curves exhibit the same phenomenon, confirming our theory even for moderate dimension p and sample size n .

Experiment 3. We further consider blockwise diagonal designs with larger-size blocks. Given $d \geq 2$ and p that is a multiple of d , we generate $X \in \mathbb{R}^{n \times p}$ such that $X'X$ is block-wise diagonal with $d \times d$ diagonal blocks, where the off-diagonal elements of each block are all equal to ρ . Other steps of the data generation are the same as in Experiment 2. We consider $(d, \rho) = (4, 0.4)$ and $(d, \rho) = (5, 0.3)$. For each choice of (d, ρ) , we set $(n, p, \vartheta) = (2000, 1000, 0.3)$ and let r range on a grid from 0 to 6 with a step size 0.2. We use signed maximum as symmetric statistic in KF and GM. For KF, we use the equi-correlated knockoff described above. The results are in the last two panels of Figure 1.9.

One noteworthy observation is that KF still has a similar performance as its own prototype, so does GM. Another observation is that GM outperforms KF when r is large, and KF slightly outperforms GM when r is small. While our theoretical results are only derived for $d = 2$, the simulations suggest that similar insight continues to apply when the block size gets larger.

Experiment 4. In Section 1.5.2, we studied variants of knockoff. The theory for 2×2 block-wise diagonal designs suggests that using CI-knockoff to construct \tilde{X} yields a higher power than using EC-knockoff (for 2×2 block-wise diagonal designs, EC-knockoff is the same as SDP-knockoff). In this experiment, we investigate whether using CI-knockoff still yields a power boost for other design classes. We consider 4 types of designs:

- *Factor models:* $X'X = (BB' + I_p)/2$, where B is a $p \times 2$ matrix whose j -th row is equal to $[\cos(\alpha_j), \sin(\alpha_j)]$ with $\{\alpha_j\}_{j=1, \dots, p}$ *iid* drawn from $\text{Uniform}[0, 2\pi]$;
- *Block diagonal:* Same as in Experiment 2, where $\rho = 0.5$.
- *Exponential decay:* The (i, j) -th element of $X'X$ is $0.6^{|i-j|}$, for $1 \leq i, j \leq p$.

- *Normalized Wishart*: $X'X$ is the sample correlation matrix of n iid samples of $N(0, I_p)$.

In the normalized Wishart design, the CI-knockoff in (1.19) may not satisfy $\text{diag}(s) \preceq 2G$. We modify it to $\text{diag}(s) = \alpha[\text{diag}(G^{-1})]^{-1}$, where α is the maximum value in $[0, 1]$ such that $\text{diag}(s) \preceq 2G$. For each design, we fix $(n, p) = (1000, 300)$, let ϑ take values in $\{0.2, 0.4\}$ and let r range on a grid from 0 to 6 with a step size 0.2. Different from previous experiments, we generate β from $\beta_j \stackrel{iid}{\sim} (1 - \varepsilon_p)\nu_0 + \frac{1}{2}\varepsilon_p\nu_{\tau_p} + \frac{1}{2}\varepsilon_p\nu_{-\tau_p}$, for $1 \leq j \leq p$. The motivation of using this model is to allow for negative entries in β . As mentioned in Remark 2 (see the end of Section 1.4), even when $X'X$ contains only nonnegative elements, this signal model can still reveal the effect of having negative correlations in the design. We compare two versions of knockoff, EC-knockoff and CI-knockoff, along with the prototype, Lasso. The results are in Figure 1.10.

For the 2×2 block-wise diagonal design, the simulations suggest that CI-KF significantly outperforms EC-KF, and that CI-KF has a similar performance as the prototype, Lasso. This is consistent with the theory in Section 1.5.2. CI-KF also yields a significant improvement over EC-KF in the factor design, and the two methods perform similarly in the exponentially decaying design and the normalized Wishart design. We notice that the Gram matrix of the normalized Wishart design has uniformly small off-diagonal entries for the current (n, p) , which is similar to the orthogonal design and explains why EC-KF and CI-KF do not have much difference. Combining these simulation results, we recommend CI-KF for practical use. Additionally, in some settings (e.g., factor design, $\vartheta = 0.4$; exponentially decaying design, $\vartheta = 0.2$), CI-KF even outperforms its prototype Lasso. One possible reason is that the ideal threshold we use is derived by ignoring the multi- $\log(p)$ term, but this term can have a non-negligible effect for a moderately large p . As a result, the Hamming error of Lasso presented here may be larger than the actual optimal one.

Experiment 5. In the previous experiments, we only examined the Hamming errors. In this experiment, we examine FDR and power separately. Fixing $(n, p, \vartheta, r) = (1000, 300, 0.2, 5)$, we

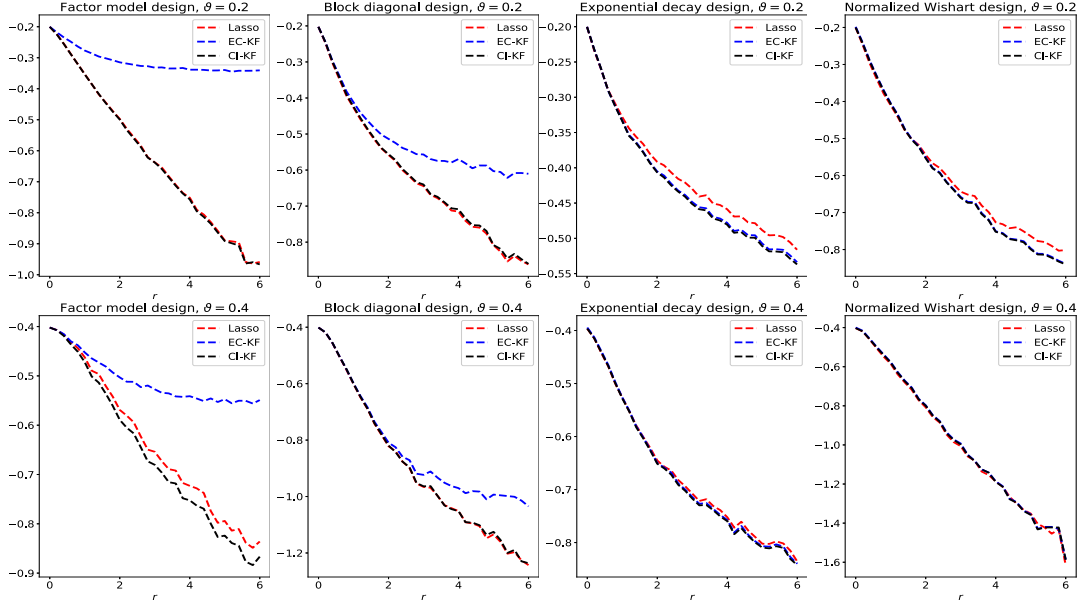


Figure 1.10: Experiment 4 (general designs). The y-axis is $\log_p(H_p/p)$, where H_p is the average Hamming error. We focus on comparing two constructions of knockoff's, EC-KF and CI-KF, and include Lasso as the benchmark.

generate data in a similar way as in Experiment 4, where the entries of β are *iid* drawn from $(1 - \varepsilon_p)\nu_0 + \frac{1}{2}\varepsilon_p\nu_{\tau_p} + \frac{1}{2}\varepsilon_p\nu_{-\tau_p}$. We consider (i) the 2×2 block-wise diagonal design, where the off-diagonal entries in each block are ρ , and (ii) the exponentially decaying design, where the (i, j) -th element of $X^T X$ is $\rho^{|i-j|}$. We let ρ range from 0.1 to 0.9, to cover the cases of weak, moderate, and strong correlations. We implement GM and two versions of knockoff, EC-KF and CI-KF (for all methods we use signed maximum as symmetric statistic). The prototypes, Lasso and OLS, are not included, as they do not aim for FDR control.

The results are shown in Figure 1.11. We set the targeted FDR level at 10%. The first and third panels show the boxplots of actual FDR. Except for the extreme case of $\rho = 0.9$ in the exponentially decaying design, all three methods yield satisfactory FDR control. The second and fourth panels show boxplots of the true positive rate (TPR). As ρ increases, the TPR of all methods has a considerable decrease. In comparison, GM has a higher TPR than two versions of knockoff in most scenarios.

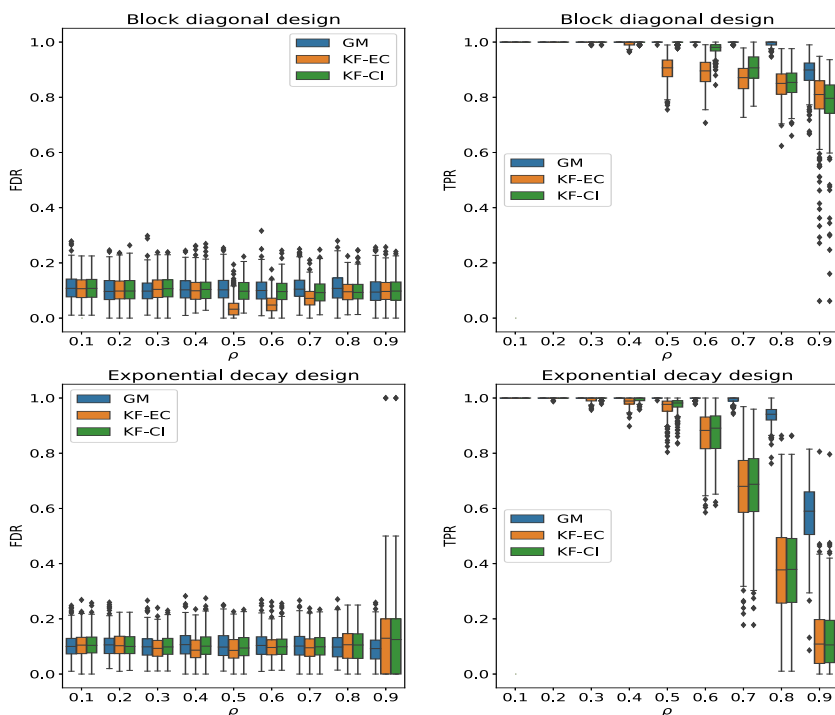


Figure 1.11: Experiment 5 (block-wise diagonal designs and exponential decayed designs). The FDR and TPR of two knockoff methods (with equi-correlated and conditional independence construction) as well as Gaussian mirror in 500 repetitions. The parameter ρ characterizes how far away is $X'X$ from I_p .

This difference is primarily caused by the difference of ranking algorithm. For the settings considered here, $\vartheta = 0.2$, and our theory in Section 1.4 suggests that least-squares is a better ranking algorithm than Lasso-path. Between two versions of knockoff, CI-KF has an advantage over EC-KF for $\rho \in \{0.4, 0.5, 0.6\}$.

1.8 DISCUSSIONS

How to maximize the power when controlling FDR is a problem of great interest. Most existing results on power analysis focus on one particular method. In this chapter, we introduce a unified framework that captures 3 key components in recent FDR control methods—(a) ranking algorithm, (b) tampered design, and (c) symmetric statistic, and our theoretical power analysis reveals the im-

part of each component. The results not only facilitate a deeper understanding of existing methods but also provide useful insights for developing new methods. We focus on the knockoff filter and Gaussian mirror as two illustrating examples, but they have covered different aspects of designing (a)-(c). Our analysis allows for comparison of different proposals of designing (a)-(c) and inspires improvements/variants/hybrid of two methods. It is unlikely to gain such insights from studying one particular FDR control method only.

We have several noteworthy discoveries: (i) The power of an FDR control method is primarily determined by the ranking algorithm; which ranking algorithm to use depends on the sparsity level and the design correlations. (ii) The choice of symmetric statistic affects the power; between the two common choices, the signed maximum is better than the difference. (iii) The tampered design can follow the p -at-a-time scheme (as in knockoff) or the one-at-a-time scheme (as in Gaussian mirror) in adding fake variables; the former is more flexible as it can accommodate different ranking algorithms, and the latter yields a higher power when the ranking algorithm is restricted to least-squares. (iv) The construction of fake variables also matters for power (e.g., SDP-knockoff versus CI-knockoff); it is sometimes beneficial to let a fake variable be properly correlated with the corresponding original variable.

Our analysis adopts a Rare/Weak signal model and uses the phase diagram and the FDR-TPR trade-off diagram to characterize the power of an FDR control method. These criteria are essentially measuring the quality of variable ranking. Good ranking is a necessary condition for simultaneously attaining a low FDR and a high power. This perspective is shared by other works on power analysis of FDR control methods, where it is common to measure the performance of variable ranking (via some trade-off diagram). Compared with works focusing on linear sparsity (e.g., [Su et al. \(2017\)](#), [Weinstein et al. \(2017\)](#)), our framework allows for a wide range of sparsity.

We focus on $p < n$ in theory. When $p > n$ and X_i 's are *iid* generated, most FDR control methods split samples, half for screening and half for FDR control, where on the second half sample the method

is the same as before. Therefore, the roles of three components are similar as before. Extending the results from $p < n$ to $p > n$ requires more complicated analysis but leads to very few new discoveries. For this reason, we only consider $p < n$ in this chapter.

There are several directions to extend current results. First, we focus on the regime where FDR and TPR converge to either 0 or 1 and characterize the rates of convergence. The more subtle regime where FDR and TPR converge to constants between 0 and 1 is not studied. We leave it to future work. Second, the study of knockoff here is only for block-wise diagonal designs. For general designs, it is very tedious to derive the precise phase diagram, but some cruder results may be less tedious to derive, such as an upper bound for the Hamming error. This kind of results will help shed more insights on how to construct the knockoff variables (e.g., how to choose $\text{diag}(s)$). Third, we only investigate Lasso-path or least-squares as options of ranking algorithms. It is interesting to study the power of FDR control methods based on other ranking algorithms, such as the marginal screening and iterative sure screening (Fan & Lv, 2008) and the covariance assisted screening (Ke et al., 2014, Ke & Yang, 2017). The covariance assisted screening was shown to yield optimal phase diagrams for a broad class of sparse designs; whether it can be developed into an FDR control method with “optimal” power remains unknown and is worth future study. Last, some FDR control methods may not fit exactly the unified framework here. For instance, the multiple data splits (Dai et al., 2020) is a method that controls FDR through data splitting. We can similarly assess its power using the Rare/Weak signal model and phase diagram, except that we need to assume the rows of X are *i.i.d.* generated. We leave such study to future work.

2

Estimating the number of Spikes by Bulk Eigenvalue Matching Analysis

CONTRIBUTION This chapter is based on a paper [Ke et al. \(2020b\)](#) jointly with Prof. Zheng Tracy Ke and Prof. Xihong Lin.

2.1 INTRODUCTION

The spiked covariance model (Johnstone, 2001) has been widely used to model the covariance structure of high-dimensional data. In this model, the population covariance matrix has K large eigenvalues, called *spiked eigenvalues*, where K is presumably much smaller than the dimension. Estimation of K is of great interest in practice, as it helps determination of the latent dimension of data. For example, in a clustering model with K_0 clusters (Jin et al., 2017), the pooled covariance matrix has $(K_0 - 1)$ spiked eigenvalues; therefore, an estimate of K tells the number of clusters. Similarly, in Genome-Wide Association Studies (GWAS), the number of spiked eigenvalues of a genetic covariance matrix reveals the number of ancestry groups in the study (Patterson et al., 2006). In high-dimensional covariance matrix estimation, K is often required as input for factor-based covariance estimation (Fan et al., 2013).

In this chapter, we assume the data vectors $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n \in \mathbb{R}^p$ are independently generated from a multivariate distribution with covariance matrix $\mathbf{\Sigma} \in \mathbb{R}^{p \times p}$, which has positive values $\mu_1 \geq \mu_2 \geq \dots \geq \mu_K$ and mutually orthogonal unit-norm vectors $\xi_1, \xi_2, \dots, \xi_K \in \mathbb{R}^p$ such that

$$\mathbf{\Sigma} = \sum_{k=1}^K \mu_k \xi_k \xi_k^\top + \mathbf{D}, \quad \text{where } \mathbf{D} = \text{diag}(\sigma_1^2, \sigma_2^2, \dots, \sigma_p^2). \quad (2.1)$$

Here, \mathbf{D} is called the residual covariance matrix. The goal is to estimate K from $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n$. We are primarily interested in the settings where K is finite and $p/n \rightarrow \gamma$, for a constant $\gamma > 0$. Throughout the chapter, we denote by $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p$ the eigenvalues of $\mathbf{\Sigma}$, and denote by $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \dots \geq \hat{\lambda}_{n \wedge p}$ the nonzero eigenvalues of the sample covariance matrix.

In the literature, there are several approaches for estimating K . The first is the information criterion approach, which finds \hat{K} that minimizes an objective of the form $L_n(K) + P_n(K)$, where $L_n(K)$ is a measure of goodness-of-fit and $P_n(K)$ is a penalty on K . An influential work is Bai & Ng (2002),

who let $L_n(K)$ be the sum of squared residuals after fitting a K -factor model and studied a few choices of the penalty function $L_n(K)$. Other examples include [Wax & Kailath \(1985\)](#), where $L_n(K)$ is a function of the arithmetic and geometric means of $(n-K)$ smallest eigenvalues. However, the information criterion approach requires the spiked eigenvalues to be sufficiently large. In [Bai & Ng \(2002\)](#), the spiked eigenvalues are at the order of p , which is much larger than the necessary order. It has been recognized that correct estimation of K is possible even when the spiked eigenvalues are at the constant order ([Baik et al., 2005](#)).

The second approach finds a big “gap” between eigenvalues of the sample covariance matrix. Recall that $\hat{\lambda}_k$ is the k th eigenvalue of the sample covariance matrix. [Onatski \(2009\)](#) introduced a test statistic, $\max_{K_0 < k \leq K_{\max}} (\hat{\lambda}_i - \hat{\lambda}_{i+1}) / (\hat{\lambda}_{i+1} - \hat{\lambda}_{i+2})$, for testing against the null hypothesis $K = K_0$ and then applied it sequentially to estimate K . [Cai et al. \(2020\)](#) proposed an iterative algorithm for estimating K that searches for a gap of $\gtrsim O(n^{-2/3})$ between eigenvalues. [Passemier & Yao \(2014\)](#) suggested estimating K by finding two consecutive gaps in eigenvalues. Such methods rely on sharp limiting distributions of the first K empirical eigenvalues, which theoretically requires a large magnitude of the spiked eigenvalues. Additionally, while utilizing eigengap is a neat idea in theory, its practical use faces challenges, since the actual eigengaps in many real data sets are slowly varying, without a clear cut.

The last approach estimates K by thresholding the empirical eigenvalues. For this approach, the key is to calculate a proper data-driven threshold. The threshold should reflect the “scaling” of the residual matrix \mathbf{D} . One idea is to first standardize the data matrix so that each variable has a unit variance and then use a scale-free threshold. Examples include the empirical Kaiser’s criterion ([Braeken & Van Assen, 2017](#)) and parallel analysis ([Horn, 1965](#)), where the scale-free threshold is determined by asymptotic behavior of the largest eigenvalue of sample covariance matrix associated with $\mathbf{X}_i \stackrel{iid}{\sim} N(0, \mathbf{I}_p)$. Another idea is to estimate \mathbf{D} by the diagonal of the sample covariance matrix and then calculate the threshold via a deterministic algorithm ([Dobriban, 2015](#)). The success of both ideas rely

on regularity conditions to ensure that the low-rank part in Model (2.1) has a negligible effect on the diagonal of Σ ; for example, the population eigenvalues cannot be enormously large and the population eigenvectors have to satisfy “delocalization” conditions. [Dobriban & Owen \(2019\)](#) improved the algorithm in [Dobriban \(2015\)](#) by a recursive procedure to remove leading eigenvalues and eigenvectors, but their method still requires some “delocalization” conditions on eigenvectors. Other related work includes [Onatski \(2010\)](#), which used a convex combination of $\hat{\lambda}_{K_{\max}+1}$ and $\hat{\lambda}_{2K_{\max}+1}$ as the threshold, where K_{\max} is a pre-specified upper bound of K , and [Fan et al. \(2020\)](#), which introduced an unbiased estimator for each of the first few eigenvalues of the population correlation matrix, and estimated K by thresholding these unbiased estimators at $1 + \sqrt{p/n}$.

To address the limitations of these methods, we propose a new estimator of K . Different from the existing work, our attention is largely focused on how to better utilize the *bulk* empirical eigenvalues in the estimation of K , especially those eigenvalues in the middle range:

$$\{\hat{\lambda}_k : \alpha(n \wedge p) \leq k \leq (1 - \alpha)(n \wedge p)\}, \quad \text{for some constant } \alpha \in (0, 1/2).$$

It is well-known in random matrix theory that these bulk eigenvalues are almost not affected by the low-rank part in Model (2.1) (e.g., see [Bloemendal et al. \(2016\)](#)). We can use these eigenvalues to gauge the “scaling” of \mathbf{D} and determine an appropriate threshold for top eigenvalues. To this end, we impose a working model on the diagonal matrix \mathbf{D} . Let $\text{Gamma}(a, b)$ denote the gamma distribution with shape parameter a and rate parameter b . Fixing $\sigma > 0$ and $\theta > 0$, we assume

$$\sigma_j^2 \stackrel{iid}{\sim} \text{Gamma}(\theta, \theta/\sigma^2), \quad 1 \leq j \leq p. \quad (2.2)$$

The mean and variance of $\text{Gamma}(\theta, \theta/\sigma^2)$ is σ^2 and σ^4/θ , respectively. As a result, the diagonal entries of \mathbf{D} are centered around σ^2 , where the level of dispersion is controlled by θ . As $\theta \rightarrow \infty$,

$\text{Gamma}(\theta, \theta/\sigma^2)$ converges to a point mass at σ^2 , and it yields $\mathbf{D} = \sigma^2 \mathbf{I}_p$. This case corresponds to the standard spiked covariance model which is frequently studied in the literature (Johnstone, 2001, Donoho et al., 2018). Combining Model (2.2) with Model (2.1), we now have a flexible spiked covariance model that includes the standard spiked covariance model as a special case.

Under Models (2.1)-(2.2), the empirical spectral distribution (ESD) converges to a limit, which is a fixed distribution with two parameters (σ^2, θ) (Silverstein, 2009). Since the empirical eigenvalues are nothing but quantiles of the ESD, we expect that all the bulk eigenvalues are asymptotically close to the corresponding quantiles of the limit of ESD. We thus estimate (σ^2, θ) by minimizing the sum of squared differences between bulk eigenvalues and quantiles of the limiting distribution. Once $(\hat{\sigma}^2, \hat{\theta})$ are available, we borrow the idea of parallel analysis (Horn, 1965) to decide a threshold for the top eigenvalues by Monte Carlo sampling. This gives rise to a new method for estimating K , which we call *bulk eigenvalue matching analysis (BEMA)*. Analogous to the orators' bema in Athens, our BEMA is a platform for gathering a large number of bulk eigenvalues and utilizing them efficiently in the estimation of K . Additional to the point estimator, we also propose a confidence interval for K .

Our method has an intuitive explanation in terms of a scree plot. Figure 2.1 shows the scree plot of a simulated example. There are multiple elbow points, and it is hard to decide where the true K is. The core idea of our method is to explore the “shape” of the scree plot in the middle range and fit it with a parametric curve; this curve is determined by the theoretical quantiles of the limit of ESD, governed by two parameters σ^2 and θ . Then, this curve can be extended to the left boundary of the scree plot to produce a threshold for top eigenvalues.

The goodness-of-fit check of Model (2.2) on real datasets can also be done via the scree plot. If the middle range of the scree plot can be well approximated by the estimated parametric curve, then it suggests that the model indeed fits the real data. In Section 2.6, we shall see that Model (2.2) is well suited to gene microarray data and GWAS data. We remark that assuming the diagonal entries of \mathbf{D} are generated from a fixed distribution is only a mild assumption. Similar conditions appear in the

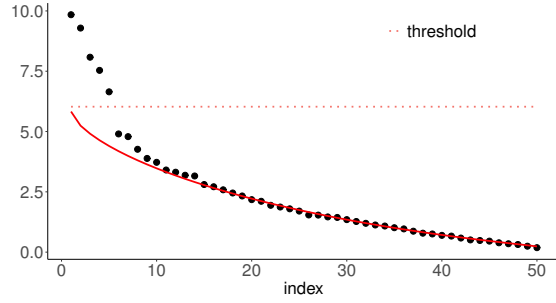


Figure 2.1: Illustration of BEMA via a scree plot. The red solid curve shows the quantiles of the theoretical limit of Empirical Spectral Distribution (ESD) under Models (2.1)-(2.2). It is a parametric curve with two parameters (σ^2, θ) , and by random matrix theory, it should fit the bulk eigenvalues well. BEMA first uses bulk eigenvalues to estimate (σ^2, θ) and then extends the estimated curve to the left boundary to get a threshold for top eigenvalues.

literature (often implicitly as regularity conditions in the theory); e.g., [Dobriban & Owen \(2019\)](#) and [Fan et al. \(2020\)](#) assume that the histogram of population eigenvalues of \mathbf{D} converges to a fixed limit. We make one step ahead by assuming that this fixed distribution is a gamma distribution. At the first glance, restricting to the gamma family seems restrictive, but Model (2.2) is in fact much more flexible than expected. With only two parameters (σ^2, θ) , it can accommodate various kinds of real data and even misspecified models (see Section 2.5).

The special case of $\theta = \infty$ is of independent interest. It corresponds to the standard spiked covariance model ([Johnstone, 2001](#)), where $\mathbf{D} = \sigma^2 \mathbf{I}_p$. This model has attracted a lot of attention ([Baik et al., 2005](#), [Paul, 2007](#), [Donoho et al., 2018](#)). In this special case, BEMA reduces to a simpler algorithm. We conduct theoretical analysis under this model. First, we give an explicit error bound for estimating σ^2 . This is connected to the robust estimation of σ^2 in the literature of reconstruction of spiked covariance matrices ([Donoho et al., 2018](#), [Shabalin & Nobel, 2013](#)). In our method, we obtain a new robust estimator of σ^2 as a byproduct, and we study it theoretically. Second, we prove the consistency of estimating K under minimal conditions. Our results impose no assumptions on the population eigenvectors ξ_1, \dots, ξ_K and only require the spiked eigenvalues $\lambda_1, \dots, \lambda_K$ to be larger than a constant. In comparison, literature works often either require some regularity conditions on

eigenvectors or need much larger spiked eigenvalues. We also provide theory for the general case of $\theta < \infty$, which has never been studied.

The remaining of this chapter is organized as follows: In Section 2.2, we describe BEMA for the standard spiked covariance model (i.e., $\theta = \infty$); in this case, the idea is easier to understand and the algorithm is simpler. In Section 2.3, we describe BEMA for the general case. Section 2.4 states the theoretical properties. Section 2.5 and Section 2.6 provide simulation study results and real data analysis, respectively. Section 2.7 concludes the chapter. Proofs are relegated to the Appendix.

2.2 BEMA FOR THE STANDARD SPIKED COVARIANCE MODEL

In this section, we consider the standard spiked covariance model (Johnstone, 2001), a special case of Models (2.1)-(2.2) with $\theta = \infty$. Since each σ_j^2 is equal to σ^2 , the model is re-written as

$$\Sigma = \sum_{k=1}^K \mu_k \xi_k \xi_k^\top + \sigma^2 \mathbf{I}_p. \quad (2.3)$$

The first K eigenvalues of Σ are $\lambda_k = \mu_k + \sigma^2$, and the remaining eigenvalues are σ^2 . The sample covariance matrix is $\mathbf{S} = \frac{1}{n} \sum_{i=1}^n (\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})^\top$, where $\bar{\mathbf{X}} = \frac{1}{n} \sum_{i=1}^n \mathbf{X}_i$. With probability 1, \mathbf{S} has $n \wedge p$ distinct nonzero eigenvalues (Uhlir, 1994), denoted as $\hat{\lambda}_1 > \hat{\lambda}_2 > \dots > \hat{\lambda}_{n \wedge p}$.

We first review some existing results about the asymptotic behavior of empirical eigenvalues.

Definition 2.2.1. Given a parameter $\gamma > 0$, the zero-excluded Machedenko-Pastur (MP) distribution is defined by the density

$$f_\gamma(x; \sigma^2) = \frac{1}{2\pi\sigma^2} \frac{1}{x(\gamma \wedge 1)} \sqrt{(x - \sigma^2 b_-)(\sigma^2 b_+ - x)} \cdot \mathbf{1}\{\sigma^2 b_- < x < \sigma^2 b_+\}, \quad (2.4)$$

where $b_\pm = (1 \pm \sqrt{\gamma})^2$. We let $F_\gamma(x; \sigma)$ denote its cumulative distribution function.

When $\gamma \leq 1$, this definition is the same as the classical MP law; when $\gamma > 1$, it excludes the point mass at zero in the classical MP law. The zero-excluded empirical spectral distribution (ESD) is given by $F_n(x) = \frac{1}{n \wedge p} \sum_{i=1}^{n \wedge p} \mathbf{1}\{\hat{\lambda}_i \leq x\}$. For convenience, we shall omit the word ‘zero-excluded’ and still call them MP and ESD.

When Σ satisfies (2.3), K is fixed and $p/n \rightarrow \gamma$ for a constant $\gamma \in (0, \infty)$, under mild regularity conditions, the following statements are true (Bloemendal et al., 2016):

- The ESD converges to the MP distribution with parameter γ ; more precisely, it holds that $\mathbb{E}[\sup_x |F_n(x) - F_\gamma(x)|] = O(n^{-1/2})$ (Götze et al., 2004).
- If $\mu_K \geq \sigma^2 \sqrt{\gamma} + n^{-1/3}$, the first K empirical eigenvalues are located outside the support of the MP distribution with high probability.

See Figure 2.2 for an illustration via simulated data ($n = 1000, p = 500$).

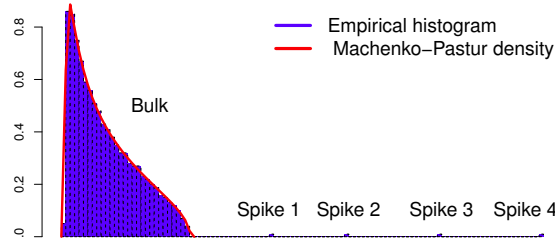


Figure 2.2: The asymptotic behavior of empirical eigenvalues. The histogram of bulk eigenvalues converges to an MP distribution, and K top eigenvalues are outside the support.

Inspired by the asymptotic behavior of empirical eigenvalues, we propose a two-step approach to estimating K . In the first step, we use bulk eigenvalues to fit an MP distribution. The density $f_\gamma(x; \sigma^2)$ in (2.4) has two parameters (γ, σ^2) , where γ can be approximated by $\gamma_n = p/n$. It reduces to considering $f_{\gamma_n}(x; \sigma^2)$, for all possible σ^2 . We aim to find $\hat{\sigma}^2$ such that $f_{\gamma_n}(x; \hat{\sigma}^2)$ is the best fit to the histogram of empirical eigenvalues. In the second step, we determine K by comparing top eigenvalues with the right boundary of the support of the estimated MP density, namely, $\hat{\sigma}^2(1 + \sqrt{\gamma_n})^2$.

Now, we describe the method in detail. First, consider the estimation of σ^2 . Fixing a constant $\alpha \in (0, 1/2)$, we take only a fraction of nonzero eigenvalues:

$$\{\hat{\lambda}_k : \alpha(n \wedge p) \leq k \leq (1 - \alpha)(n \wedge p)\}.$$

Since K is fixed and $n \wedge p \rightarrow \infty$, any α guarantees that the first K eigenvalues are excluded. The choice of α does not matter. We usually set $\alpha = 0.2$, so that 60% of the nonzero eigenvalues in the middle range are used. Write for short $\tilde{p} = n \wedge p$. By definition, $\hat{\lambda}_k$ is the (k/\tilde{p}) -upper-quantile of the ESD. Let $q_k = q_k(\gamma_n)$ denote the (k/\tilde{p}) -upper-quantile of the MP distribution associated with $\gamma = \gamma_n$ and $\sigma^2 = 1$, that is,

$$q_k \text{ is the unique value such that } \int_{q_k}^{(1+\sqrt{\gamma_n})^2} f_{\gamma_n}(x; 1) dx = k/\tilde{p}. \quad (2.5)$$

These q_k 's can be easily computed (e.g., via the R package *RMTstat*). For an MP distribution with a general σ^2 , its (k/\tilde{p}) -upper-quantile equals to $\sigma^2 q_k$. Since the ESD is asymptotically close to the MP distribution, we expect that

$$\hat{\lambda}_k \approx \sigma^2 \cdot q_k.$$

It motivates us to use $\{(q_k, \hat{\lambda}_k)\}_{\alpha\tilde{p} \leq k \leq (1-\alpha)\tilde{p}}$ to fit a line without intercept, and this can be done by a simple least-squares. The slope of this line is an estimator of σ^2 .

Next, we use $\hat{\sigma}^2$ to determine a threshold for the top eigenvalues. A natural choice of threshold is $\hat{\sigma}^2(1 + \sqrt{\gamma_n})^2$, but it has a considerable probability of over-estimating K . We slightly increase this threshold by taking an advantage of another result in random matrix theory. When $\mu_K > \sigma^2 \sqrt{\gamma}$, it is known that (Johnstone, 2001, Bloemendal et al., 2016)

$$\frac{\hat{\lambda}_{K+1} - \sigma^2(1 + \sqrt{\gamma_n})^2}{\sigma^2 n^{-\frac{2}{3}} \gamma_n^{-\frac{1}{6}} (1 + \sqrt{\gamma_n})^{\frac{4}{3}}} \xrightarrow{d} \text{type-I Tracy-Widom distribution.} \quad (2.6)$$

Algorithm 1. BEMA for the standard spiked covariance model.

Input: Nonzero eigenvalues $\hat{\lambda}_1, \dots, \hat{\lambda}_{n \wedge p}$, $\alpha \in (0, 1/2)$ and $\beta \in (0, 1)$.

Output: An estimate of K .

Step 1: Write $\tilde{p} = n \wedge p$. For each $\alpha\tilde{p} \leq k \leq (1 - \alpha)\tilde{p}$, obtain q_k , the (k/\tilde{p}) -upper-quantile of the MP distribution associated with $\sigma^2 = 1$ and $\gamma_n = p/n$. Compute

$$\hat{\sigma}^2 = \frac{\sum_{\alpha\tilde{p} \leq k \leq (1-\alpha)\tilde{p}} q_k \hat{\lambda}_k}{\sum_{\alpha\tilde{p} \leq k \leq (1-\alpha)\tilde{p}} q_k^2}.$$

Step 2: Obtain $t_{1-\beta}$, the $(1 - \beta)$ -quantile of Tracy-Widom distribution. Estimate K by

$$\hat{K} = \#\{1 \leq k \leq \tilde{p} : \hat{\lambda}_k > \hat{\sigma}^2 [(1 + \sqrt{\gamma_n})^2 + t_{1-\beta} \cdot n^{-\frac{2}{3}} \gamma_n^{-\frac{1}{6}} (1 + \sqrt{\gamma_n})^{\frac{4}{3}}]\}.$$

We propose thresholding the top eigenvalues at

$$\hat{T} = \hat{\sigma}^2 \left[(1 + \sqrt{\gamma_n})^2 + t_{1-\beta} \cdot n^{-\frac{2}{3}} \gamma_n^{-\frac{1}{6}} (1 + \sqrt{\gamma_n})^{\frac{4}{3}} \right],$$

where $t_{1-\beta}$ denotes the $(1 - \beta)$ -quantile of the Tracy-Widom distribution. Then, the probability of over-estimating K is controlled by β .

Algorithm 1 has two tuning parameters (α, β) . The output of the algorithm is insensitive to α if α is not too small, and we set $\alpha = 0.2$ by default. β controls the probability of over-estimating K and is specified by the user. In theory, the ideal choice of β should satisfy that $\beta \rightarrow 0$ at a properly slow rate (see Section 2.4). In practice, choosing a moderate β often yields the best finite-sample performance. Our numerical experiments suggest that $\beta = 0.1$ is a good choice for most settings.

A simulation example. We illustrate Algorithm 1 on a simulation example. Fix $(n, p, K) = (1000, 500, 10)$. We generate $\mathbf{X}_i \stackrel{iid}{\sim} N(0, \Sigma)$, where Σ is a diagonal matrix whose first K diagonals

equal to 5.4 and the remaining diagonals equal to $\sigma^2 = 2$. In the left panel of Figure 2.3, we plot $\hat{\lambda}_k$ versus q_k . Except for a few top eigenvalues, it fits well to a straight line crossing the origin. We use 300 bulk eigenvalues $\{\hat{\lambda}_k\}_{100 < k \leq 400}$ (the blue dots) to fit a regression line (the red dotted line). The slope of this line gives the estimate $\hat{\sigma}^2 = 2.04$. In the middle panel of Figure 2.3, we plot $\hat{\lambda}_k$ versus k . The red solid line is the curve of $\hat{\sigma}^2 q_k$ versus k . Although it is estimated using the blue dots only, we can extend this curve to the left boundary, which gives rise to the value $\hat{\sigma}^2(1 + \sqrt{\gamma_n})^2$. We then use this value and the Tracy-Widom distribution to calculate a threshold for the top eigenvalues. The estimator \hat{K} equals to the number of top eigenvalues that exceed this threshold. The right panel of Figure 2.3 is a zoom-in of the middle panel. As k gets smaller (e.g., $k < 50$), the eigenvalues stay above the fitted MP quantile curve. This is because these $\hat{\lambda}_k$ are influenced by the spiked eigenvalues of Σ . Such eigenvalues are already excluded in the estimation of σ^2 . The right panel can also be viewed as a scree plot. Finding the elbow point of the scree plot is a common ad-hoc method for estimating K . In this plot, the elbow points are $\{6, 7, 10, 11\}$, hard to decide the true K . In contrast, our method correctly picks $\hat{K} = 10$.

Remark 1 (*Connection to the robust estimation of σ^2*). As a byproduct, the BEMA algorithm yields a new estimator for σ^2 in the standard spiked covariance model, which can be useful for other problems such as reconstruction of spiked covariance matrices. [Gavish & Donoho \(2014\)](#) proposed a robust estimator of σ^2 , which is the ratio between the median of eigenvalues and the median of a standard MP distribution. Viewed in the Q-Q plot (left panel of Figure 2.3), their method is equivalent to using *a single point* to decide the slope. In comparison, our method uses a number of bulk eigenvalues to decide the slope and is thus more robust. [Kritchman & Nadler \(2009\)](#) proposed an estimator of σ^2 by solving a non-linear system of equations, and [Shabalin & Nobel \(2013\)](#) estimated σ^2 by minimizing the Kolmogorov-Smirnov distance between the ESD and its theoretical limit. In comparison, our estimator of σ^2 is from a simple least-squares and is much easier to compute. In Section 2.4, we also give an explicit error bound for our estimator.

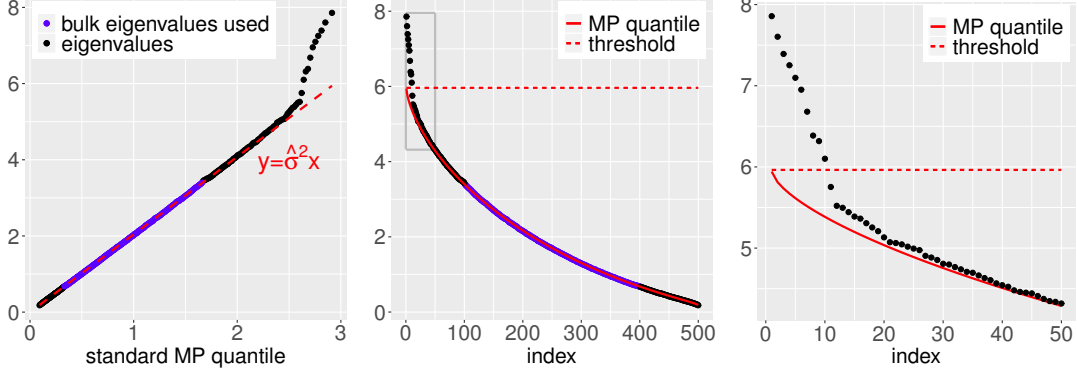


Figure 2.3: Illustration of BEMA for the standard spiked covariance model (simulated data, $n = 1000$, $p = 500$, $K = 10$). The left panel plots $\hat{\lambda}_k$ versus q_k , where q_k is the (k/\tilde{p}) -upper-quantile of the standard MP distribution. The dashed red line is the fitted regression line on bulk eigenvalues (blue dots), whose slope is an estimate of σ^2 . The middle panel plots $\hat{\lambda}_k$ versus k , which is the scree plot. The red solid curve is $\hat{\sigma}^2 q_k$ versus k . It fits the bulk eigenvalues (blue dots) very well. When this curve is extended to the left boundary, it hits $\hat{\sigma}^2(1 + \sqrt{\gamma_n})^2$. Our threshold for the top eigenvalues, which is the $(1 - \beta)$ -quantile of the Tracy-Widom distribution, is slightly larger than this value and shown by the dotted red line. The right panel zooms into the grey square area of the middle panel. It shows that 10 empirical eigenvalues exceeds the threshold, resulting in $\hat{K} = 10$.

2.3 BEMA FOR THE GENERAL SPIKED COVARIANCE MODEL

We now consider the general case where the residual covariance matrix can have unequal diagonal entries. We shall modify Algorithm 1 to accommodate this setting. Re-write Models (2.1)-(2.2) as

$$\Sigma = \sum_{k=1}^K \mu_k \xi_k \xi_k^\top + \text{diag}(\sigma_1^2, \sigma_2^2, \dots, \sigma_p^2), \quad \text{where } \sigma_k^2 \stackrel{iid}{\sim} \text{Gamma}(\theta, \theta/\sigma^2). \quad (2.7)$$

Same as before, let $\hat{\lambda}_1 > \hat{\lambda}_2 > \dots > \hat{\lambda}_{n \wedge p}$ denote the nonzero eigenvalues of the sample covariance matrix. Below, in Section 2.3.1, we first state some well-known results from random matrix theory and motivate our methodology idea. In Section 2.3.2, we formally introduce the BEMA algorithm. In Section 2.3.3, we provide an asymptotic confidence interval for K .

2.3.1 THE ASYMPTOTIC BEHAVIOR OF EMPIRICAL EIGENVALUES

Under Model (2.7), the asymptotic behavior of bulk eigenvalues and top eigenvalues exhibit some similarity to the case of standard spiked covariance model:

- The empirical spectral distribution (ESD) converges to a fixed limit.
- The first K empirical eigenvalues stand out of the bulk.

However, the precise statement is more sophisticated.

We first consider the ESD. When K is finite and $p/n \rightarrow \gamma$, the ESD converges to a distribution $F_\gamma(x; \sigma^2, \theta)$. This distribution is parametrized by (σ^2, θ) , but it does not have an explicit form. It is defined implicitly by an equation of its Stieltjes transform (Marcenko & Pastur, 1967). Let $H_{\sigma^2, \theta}(t)$ be the CDF of Gamma($\theta, \theta/\sigma^2$). For each $z \in \mathbb{C}^+$, there is a unique $m = m(z; \gamma, \sigma^2, \theta) \in \mathbb{C}^+$ such that

$$z = -\frac{1}{m} + \gamma \int \frac{t}{1+tm} dH_{\sigma^2, \theta}(t). \quad (2.8)$$

The density of $F_\gamma(x; \sigma^2, \theta)$, denoted by $f_\gamma(x; \sigma^2, \theta)$, satisfies that

$$f_\gamma(x; \sigma^2, \theta) = \lim_{y \rightarrow 0^+} \left\{ \frac{1}{\pi(\gamma \wedge 1)} \Im(m(x + iy; \gamma, \sigma^2, \theta)) \right\},^* \quad (2.9)$$

where $\Im(\cdot)$ denotes the imaginary part of a complex number.

We aim to estimate (σ^2, θ) by comparing the bulk eigenvalues with the corresponding quantiles of $F_\gamma(x; \sigma^2, \theta)$. In the special case of $\theta = \infty$, $F_\gamma(x; \sigma^2, \theta)$ reduces to the MP distribution. Therefore, we can compute its quantiles explicitly and estimate σ^2 by a simple least-squares. For the general case, we have to compute the quantiles of $F_\gamma(x; \sigma^2, \theta)$ numerically. There are two approaches, one is solving the

*The factor $1/(\gamma \wedge 1)$ is due to considering the zero-excluded ESD. If we consider the classical ESD, this factor should be $1/\gamma$.

density from equations (2.8)-(2.9) and then computing the quantiles, and the other is using Monte Carlo simulations. We will describe them in Section 2.3.2.

Next, we consider the top eigenvalues. It requires a precise definition of “standing out” of the bulk. We use the distribution of $\hat{\lambda}_{K+1}$ under Model (2.7) as a benchmark, i.e., $\hat{\lambda}_k$ needs to be much larger than a high-probability upper bound of $\hat{\lambda}_{K+1}$ in order to be called “standing out.” Fortunately, the behavior of $\hat{\lambda}_{K+1}$ has been studied in the literature of random matrix theory. We define the following null model, which is a special case of Model (2.7) with $K = 0$:

$$\Sigma = \text{diag}(\sigma_1^2, \sigma_2^2, \dots, \sigma_p^2), \quad \text{where } \sigma_k^2 \stackrel{iid}{\sim} \text{Gamma}(\theta, \theta/\sigma^2). \quad (2.10)$$

Let $\hat{\lambda}_1^*$ denote the largest eigenvalue of the sample covariance matrix under this null model. By eigenvalue sticking result (see Bloemendal et al. (2016), Knowles & Yin (2017) and a detailed discussion in Section 2.4.3), the distribution of $\hat{\lambda}_{K+1}$ is asymptotically close to the distribution of $\hat{\lambda}_1^*$. We now re-frame the statement that “the first K empirical eigenvalues stand out” as follows: Under some regularity conditions, each of $\hat{\lambda}_1, \dots, \hat{\lambda}_K$ is significantly larger than $\hat{\lambda}_1^*$ associated with Model (2.10).

We aim to threshold the top eigenvalues by the $(1 - \beta)$ -quantile of the distribution of $\hat{\lambda}_1^*$, where β controls the probability of over-estimating K . In the special case of $\theta = \infty$, the distribution of $\hat{\lambda}_1^*$ converges to a Tracy-Widom distribution, so that we have a closed-form expression for the threshold. In the general case, we calculate this threshold by Monte Carlo simulation, where we simulate data from the null model to approximate the distribution of $\hat{\lambda}_1^*$. We relegate the details to Section 2.3.2.

2.3.2 THE ALGORITHM OF ESTIMATING K

Same as before, the BEMA algorithm has two steps: Step 1 estimates (σ^2, θ) from bulk eigenvalues, and Step 2 calculates a threshold for the top eigenvalues.

Consider Step 1. Write $\tilde{p} = p \wedge n$ and $\gamma_n = p/n$. For a constant $\alpha \in (0, 1/2)$, we take the

$(1 - 2\alpha)$ -fraction of bulk eigenvalues in the middle range, i.e., $\{\hat{\lambda}_k : \alpha\tilde{p} \leq k \leq (1 - \alpha)\tilde{p}\}$. Each empirical eigenvalue $\hat{\lambda}_k$ is also the (k/\tilde{p}) -upper-quantile of the ESD. We recall that $F_{\gamma_n}(x; \sigma^2, \theta)$ is the theoretical limit of ESD as defined in (2.8)-(2.9). Let $\bar{F}_{\gamma_n}^{-1}(k/\tilde{p}; \sigma^2, \theta)$ denote the (k/\tilde{p}) -upper-quantile of this distribution. We expect to see

$$\hat{\lambda}_k \approx \bar{F}_{\gamma_n}^{-1}(k/\tilde{p}; \sigma^2, \theta).$$

It motivates the following estimator of (σ^2, θ) :

$$(\hat{\sigma}^2, \hat{\theta}) = \operatorname{argmin}_{(\sigma^2, \theta)} \left\{ \sum_{\alpha\tilde{p} \leq k \leq (1-\alpha)\tilde{p}} [\hat{\lambda}_k - \bar{F}_{\gamma_n}^{-1}(k/\tilde{p}; \sigma^2, \theta)]^2 \right\}. \quad (2.11)$$

We now describe how to solve (2.11). This is a two-dimensional optimization. As long as we can evaluate the objective function for arbitrary (σ^2, θ) , we can solve it via a simple grid search. To further simplify the objective, we first get rid of σ^2 and reduce it to an optimization on θ only. Note that $\Gamma(\theta, \theta/\sigma^2)$ is equivalent to $\sigma^2 \cdot \Gamma(\theta, \theta)$. We can deduce from (2.8)-(2.9) that a similar connection holds between $F_{\gamma_n}(x; \sigma^2, \theta)$ and $F_{\gamma_n}(x; 1, \theta)$. Then, their quantiles satisfy

$$\bar{F}_{\gamma_n}^{-1}(k/\tilde{p}; \sigma^2, \theta) = \sigma^2 \cdot \bar{F}_{\gamma_n}^{-1}(k/\tilde{p}; 1, \theta).$$

We re-write (2.11) as

$$\min_{\theta} H(\theta), \quad \text{where} \quad H(\theta) \equiv \min_{\sigma^2} \left\{ \sum_{\alpha\tilde{p} \leq k \leq (1-\alpha)\tilde{p}} [\hat{\lambda}_k - \sigma^2 \bar{F}_{\gamma_n}^{-1}(k/\tilde{p}; 1, \theta)]^2 \right\}.$$

As long as we can compute $\bar{F}_{\gamma_n}^{-1}(y; 1, \theta)$ for any $\theta > 0$ and $y \in [0, 1]$, we can obtain $H(\theta)$ for each θ by least squares regression of the $\hat{\lambda}_k$'s on the $\bar{F}_{\gamma_n}^{-1}(k/\tilde{p}; 1, \theta)$'s. Given $H(\theta)$, we can solve the optimization by a grid search on θ .

This is described in Step 1 of Algorithm 2. Suppose there is an available algorithm GetQT that computes $\bar{F}_{\gamma_n}^{-1}(y; 1, \theta)$ for any $\theta > 0$ and $y \in [0, 1]$. Fix a set of grid points $\{\theta_j\}_{j=1}^N$. For each θ_j , we first compute $\bar{F}_{\gamma_n}^{-1}(k/\tilde{p}; 1, \theta_j)$ for all $\alpha\tilde{p} \leq k \leq (1-\alpha)\tilde{p}$. Given θ_j , the value of σ^2 that minimizes (2.11) is obtained by regressing $\{\hat{\lambda}_k\}_{\alpha\tilde{p} \leq k \leq (1-\alpha)\tilde{p}}$ on $\{\bar{F}_{\gamma_n}^{-1}(k/\tilde{p}; 1, \theta_j)\}_{\alpha\tilde{p} \leq k \leq (1-\alpha)\tilde{p}}$ with a least-squares. Let $\hat{\sigma}^2(\theta_j)$ denote this optimal value of σ^2 , and let v_j denote the objective in (2.11) associated with $\{\theta_j, \hat{\sigma}^2(\theta_j)\}$. We select j^* so that v_j is minimized and set $\hat{\theta} = \theta_{j^*}$ and $\hat{\sigma}^2 = \hat{\sigma}^2(\theta_{j^*})$.

What remains is the design of an algorithm $\text{GetQT}(y, \gamma_n, \theta)$ to compute the y -upper-quantile of the distribution $F_{\gamma_n}(\cdot; 1, \theta)$ for arbitrary (θ, y) . We note that $F_{\gamma_n}(x; 1, \theta)$ only has an implicit definition through equations (2.8)-(2.9). In the appendix, we propose two algorithms that serve for this purpose:

- GetQT1 takes advantage of the fact that $F_{\gamma_n}(x; \sigma^2, \theta)$ is also the theoretical limit of the ESD of the null model (2.10). This algorithm simulates data from Model (2.10) with $\sigma^2 = 1$ to get the Monte Carlo approximation of the target quantile.
- GetQT2 first utilizes the definition (2.8)-(2.9) to solve the density $f_{\gamma_n}(x; 1, \theta)$ and then uses the density to compute quantiles.

The two GetQT algorithms have comparable numerical performance, but each has an advantage on running time in some cases; see the appendix for more discussions.

Consider Step 2. We estimate K by comparing each top eigenvalue with the $(1 - \beta)$ -quantile of the distribution of $\hat{\lambda}_1^*$ under the null model (2.10), with $(\hat{\sigma}^2, \hat{\theta})$ plugged in. The threshold is

$$\hat{T} = \left\{ \begin{array}{l} (1 - \beta)\text{-quantile of the distribution of } \hat{\lambda}_1^* \text{ under the null model} \\ \Sigma = \text{diag}(\sigma_1^2, \sigma_2^2, \dots, \sigma_p^2), \text{ where } \sigma_j^2 \stackrel{iid}{\sim} \text{Gamma}(\hat{\theta}, \hat{\theta}/\hat{\sigma}^2) \end{array} \right\}. \quad (2.12)$$

The \hat{T} here generalizes the threshold in Algorithm 1. The threshold in Algorithm 1 is a special case of \hat{T} at $\hat{\theta} = \infty$, which happens to have an explicit formula.

Algorithm 2. BEMA for the general spiked covariance model.

Input: Nonzero eigenvalues $\hat{\lambda}_1, \dots, \hat{\lambda}_{n \wedge p}$, $\alpha \in (0, 1/2)$, $\beta \in (0, 1)$, a grid of values $0 < \theta_1 < \theta_2 < \dots < \theta_N$, an algorithm `GetQT`, and an integer $M \geq 1$.

Output: An estimate of K .

Step 1: Write $\tilde{p} = n \wedge p$ and $\gamma_n = p/n$. For each $1 \leq j \leq N$:

- For each $\alpha\tilde{p} \leq k \leq (1 - \alpha)\tilde{p}$, run the algorithm `GetQT`($k/\tilde{p}, \gamma_n, \theta_j$) to obtain q_{kj} .
- Compute $\hat{\sigma}^2(\theta_j) = (\sum_{\alpha\tilde{p} \leq k \leq (1-\alpha)\tilde{p}} q_{kj} \hat{\lambda}_k) / (\sum_{\alpha\tilde{p} \leq k \leq (1-\alpha)\tilde{p}} q_{kj}^2)$.
- Let $v_j = \sum_{\alpha\tilde{p} \leq k \leq (1-\alpha)\tilde{p}} [\hat{\lambda}_k - \hat{\sigma}^2(\theta_j) \cdot q_{kj}]^2$.

Find $j^* = \operatorname{argmin}_{1 \leq j \leq N} v_j$. Let $\hat{\theta} = \theta_{j^*}$ and $\hat{\sigma}^2 = \hat{\sigma}^2(\theta_{j^*})$.

Step 2: For $1 \leq m \leq M$:

- Sample $d_j^* \sim \text{Gamma}(\hat{\theta}, \hat{\theta})$, independently for $1 \leq j \leq p$. Sample $X_i^*(j) \sim N(0, \hat{\sigma}^2 d_j^*)$, independently for $1 \leq i \leq n$ and $1 \leq j \leq p$.
- Compute the largest singular value of $n^{-1/2} X^*$, where $X^* = [X_1^*, X_2^*, \dots, X_n^*]^\top$. Let $\hat{\lambda}_{1(m)}^*$ be the square of this singular value.

Let \hat{T} be the $(1 - \beta)$ -quantile of $\{\hat{\lambda}_{1(m)}^*\}_{1 \leq m \leq M}$. Output $\hat{K} = \#\{1 \leq k \leq \tilde{p} : \hat{\lambda}_k > \hat{T}\}$.

We compute \hat{T} via Monte Carlo simulations. We first draw Σ from the null model in (2.12), and then draw the data matrix from multivariate normal distributions and compute the largest eigenvalue of the sample covariance matrix. By repeating these steps multiple times, we obtain the sampling distribution of $\hat{\lambda}_1^*$ in (2.12). This is described in Step 2 of Algorithm 2.

The BEMA algorithm has three tuning parameters (α, β, M) , where α controls the percentage of bulk eigenvalues used for estimating (σ^2, θ) and M is the number of Monte Carlo repetitions for approximating \hat{T} . The performance of the algorithm is insensitive to (α, M) (see Section 2.5). We set

$\alpha = 0.2$ and $M = 500$ by default. The parameter β controls the probability of over-estimating K . Theoretically, if the spiked eigenvalues are large enough, we should use a diminishing β (i.e., $\beta \rightarrow 0$ as $n \rightarrow \infty$) so that the probability of over-estimating K tends to zero. In practice, it often happens that the spiked eigenvalues are only moderately large. We thus need a moderate β to strike a balance between the probability of over-estimating K and the probability of under-estimating K . We leave it to the users to decide. It is analogous to the situation in false discovery rate control, where the users select the target false discovery rate. In our numerical experiments, we find that $\beta = 0.1$ is a good choice.

A Simulation Example. We illustrate Algorithm 2 using a simulation example. Fix $(n, p, K) = (1000, 200, 5)$ and $(\sigma^2, \theta) = (1, 10)$. We generate \mathbf{X}_i iid from $\mathcal{N}(0, \Sigma)$, where Σ satisfies model (2.7) with $\mu_k = 2.3$ for $1 \leq k \leq K$. The left panel of Figure 2.4 shows the plot of $\hat{\lambda}_k$ versus the MP quantiles q_k . It does not fit a line crossing the origin, suggesting that Algorithm 1 does not work for this general covariance model. The middle panel contains the plot of $\hat{\lambda}_k$ versus $\bar{F}_{\gamma_n}^{-1}(k/\tilde{p}; 1, \hat{\theta})$, where $\hat{\theta}$ is from Algorithm 2. Except for a few top eigenvalues, it fits very well a line crossing the origin, suggesting that Algorithm 2 is successful in this setting. The estimated parameters are $(\hat{\sigma}^2, \hat{\theta}) = (1.02, 10.39)$, which is close to the true values. The right panel contains the plot of $\hat{\lambda}_k$ versus k , and the fitted curve of $\hat{\sigma}^2 \cdot \bar{F}_{\gamma_n}^{-1}(k/\tilde{p}; 1, \hat{\theta})$ versus k (solid red line). The threshold \hat{T} is also shown by the dashed line. It yields $\hat{K} = 5$, which is the same as the ground truth.

Remark 2 (Connection to parallel analysis). Parallel analysis (Horn, 1965) is a popular method for estimating the number of spiked eigenvalues in real applications. It samples data from a *null covariance model* that has no spiked eigenvalues, and estimates K by comparing the distribution of top empirical eigenvalues on simulated data to those actually observed from the original data. The most common version of parallel analysis first standardizes the data matrix so that each variable has a unit variance and then uses $\Sigma = \mathbf{I}_p$ as the null model. Our algorithm has a similar spirit as parallel analysis, but we adopt a more sophisticated null covariance model, Model (2.10), and estimate parameters of this null

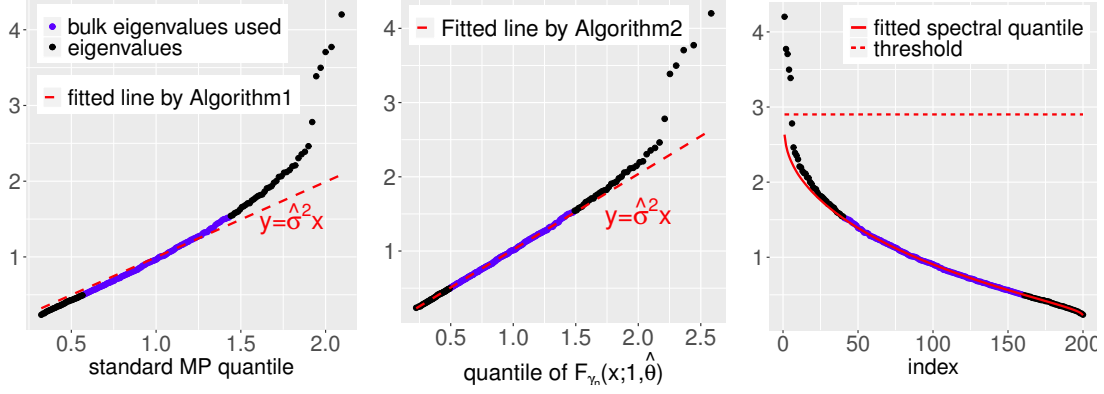


Figure 2.4: Illustration of BEMA for the general spiked covariance model. The left panel plots $\hat{\lambda}_k$ versus q_k , where the q_k 's are quantiles of the standard MP distribution. It fits the regression line poorly, suggesting that Algorithm 1 is no longer working for this general model. The middle panel plots $\hat{\lambda}_k$ versus $\bar{F}_{\gamma_n}^{-1}(x; 1, \hat{\theta})$, where $\hat{\theta}$ is an estimate of θ by Algorithm 2. The bulk eigenvalues (blue dots) fit the regression line very well. The right panel is the scree plot, where the red solid curve is $\bar{F}_{\gamma_n}^{-1}(x; \hat{\sigma}, \hat{\theta})$ versus k . A threshold (red dotted line) is given by the 90%-quantile of the distribution of $\hat{\lambda}_1^*$ from a null model; see (2.12). There are 5 empirical eigenvalues exceeding this threshold, which gives $\hat{K} = 5$.

model carefully from bulk eigenvalues.

Remark 3 (Memory use of BEMA). The input of BEMA includes nonzero eigenvalues of the sample covariance matrix. These eigenvalues can be computed by eigen-decomposition on either the $p \times p$ matrix $\mathbf{X}^\top \mathbf{X}$ or the $n \times n$ matrix $\mathbf{X}\mathbf{X}^\top$. Therefore, the memory use depends on the minimum of n and p . In many real applications, p is very large but n is relatively small, and BEMA is still implementable under even strict memory constraints.

2.3.3 A CONFIDENCE INTERVAL OF K

By varying β in Algorithm 2, we get different estimators of K , where the over-shooting probability is controlled at different levels. We use these estimators to construct a confidence interval for K .

Definition 2.3.1 (Confidence interval of K). Denote the output of Algorithm 2 by \hat{K}_β to indicate its dependence on β . Given any $\omega_0 \in (0, 1)$, we introduce the following $(1 - \omega_0)$ -confidence interval of

K as $[\hat{K}_{\omega_0/2}, \hat{K}_{1-\omega_0/2}]$.

We explain why the confidence interval is asymptotically valid. Let $\hat{T} = \hat{T}_\beta$ be the threshold in (2.12), and let $\hat{\lambda}_1^*$ be the largest eigenvalue of the sample covariance matrix when data are from the null model (2.10). We use \mathbb{P}_0 to denote the probability measure associated with Model (2.10). By definition of \hat{T}_β , $\mathbb{P}_0\{\hat{\lambda}_1^* \leq t\} \Big|_{t=\hat{T}_\beta} = 1 - \beta$. At the same time, the eigenvalue sticking result (Bloemendal et al., 2016, Knowles & Yin, 2017) states that, under some regularity conditions, the distribution of $\hat{\lambda}_{K+1}$ is asymptotically close to the distribution $\hat{\lambda}_1^*$. It follows that

$$\begin{aligned} \mathbb{P}\{\hat{K}_{\omega_0/2} > K\} &\leq \mathbb{P}\{\hat{\lambda}_{K+1} > \hat{T}_{\omega_0/2}\} \approx \mathbb{P}_0\{\hat{\lambda}_1^* > t\} \Big|_{t=\hat{T}_{\omega_0/2}} = \omega_0/2, \\ \mathbb{P}\{\hat{K}_{1-\omega_0/2} < K\} &\leq \mathbb{P}\{\hat{\lambda}_K \leq \hat{T}_{1-\omega_0/2}\} \leq \mathbb{P}\{\hat{\lambda}_{K+1} \leq \hat{T}_{1-\omega_0/2}\} \approx \mathbb{P}_0\{\hat{\lambda}_1^* \leq t\} \Big|_{t=\hat{T}_{1-\omega_0/2}} = \omega_0/2. \end{aligned}$$

2.4 THEORETICAL PROPERTIES

We study in this section the theoretical properties of the proposed BEMA method. In Section 2.4.1, we focus on the standard spiked covariance model ($\theta = \infty$), where we derive the error rate of $\hat{\sigma}^2$ and the consistency of \hat{K} . In Section 2.4.2, we study the general spiked covariance model ($\theta < \infty$). This setting is much more complicated. It connects to an unsolved problem in random matrix theory, that is, how to get sharp asymptotic theory for eigenvalues when the limiting spectrum of Σ is unbounded and has convex decay in the tail. Only partial results are known (Kwak et al., 2019). To overcome the technical difficulty, in our theoretical investigation, we approximate Model (2.7) by a proxy model where σ_j^2 are *iid* generated from a truncated Gamma distribution. Under this proxy model, we derive the rate of convergence for $(\hat{\sigma}^2, \hat{\theta})$ and the consistency of \hat{K} . In Section 2.4.3, we connect Model (2.7) to the proxy model and discuss the theory for Model (2.7).

Through this section, we assume $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n$ are generated as follows:

Assumption 2.4.1. Let $\mathbf{Y} = [\mathbf{Y}_1, \mathbf{Y}_2, \dots, \mathbf{Y}_n]^\top \in \mathbb{R}^{n \times p}$ be a random matrix with independent but not necessarily identically distributed entries, where $\mathbb{E}[\mathbf{Y}_i(j)] = 0$ and $\text{Var}(\mathbf{Y}_i(j)) = 1$, for $1 \leq i \leq n$, $1 \leq j \leq p$. Given $\sigma_1, \sigma_2, \dots, \sigma_p > 0$, $\mu_1 \geq \mu_2 \geq \dots \geq \mu_K > 0$, and orthonormal vectors $\xi_1, \xi_2, \dots, \xi_p \in \mathbb{R}^p$, let $\Sigma = \sum_{k=1}^r (\sigma_k^2 + \mu_k) \xi_k \xi_k^\top + \sum_{j=r+1}^p \sigma_j^2 \xi_j \xi_j^\top$. We assume $\mathbf{X}_i = \Sigma^{1/2} \mathbf{Y}_i$, for $1 \leq i \leq n$.

Under this assumption, each \mathbf{X}_i is a linear transform of a random vector \mathbf{Y}_i that has independent entries. This is stronger than assuming $\text{Cov}(\mathbf{X}_i) = \Sigma$ but is conventional in the literature.

Assumption 2.4.2. For each integer $m \geq 1$, there exists a universal constant $C_m > 0$ such that $\sup_{1 \leq i \leq n, 1 \leq j \leq p} \mathbb{E}[|\mathbf{Y}_i(j)|^m] \leq C_m$.

This assumption can be further relaxed. For example, we do not actually need the inequality to hold for every $m \geq 1$ but only for $1 \leq m \leq M$, where M is a properly large integer (Bloemendal et al., 2016, Knowles & Yin, 2017). We use the current assumption for convenience.

We will use the following notation frequently, which is conventional in random matrix theory:

Definition 2.4.1. Let U_n and V_n be two sequences of random variables indexed by n . We say that U_n is stochastically dominated by V_n , if for any $\varepsilon > 0$ and $s > 0$ there exists $N = N(\varepsilon, s)$ such that $\mathbb{P}(U_n > n^\varepsilon V_n) \leq n^{-s}$ for all $n \geq N$. We write $U_n \prec V_n$.

2.4.1 THE STANDARD SPIKED COVARIANCE MODEL

The standard spiked covariance model (Johnstone, 2001) assumes $\mathbf{D} = \sigma^2 \mathbf{I}_p$. In this case, BEMA simplifies to Algorithm 1. It outputs $\hat{\sigma}^2$ and \hat{K} . We first give an error bound on estimating σ^2 .

Theorem 2.4.1 (Estimation error of $\hat{\sigma}^2$). Suppose $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n$ satisfy Assumptions 2.4.1-2.4.2 with $\sigma_j^2 \equiv \sigma^2$. Suppose $K \geq 1$ is fixed and $p/n \rightarrow \gamma$ for a constant $\gamma > 0$. Let $\hat{\sigma}^2$ be the estimator of σ^2 by Algorithm 1, where the tuning parameter α is a constant in $(0, 1/2)$. Then, $|\hat{\sigma}^2 - \sigma^2| \prec n^{-1}$.

This result is connected to the robust estimation of σ^2 in a standard spiked covariance model (Gavish & Donoho, 2014, Kritchman & Nadler, 2009, Shabalin & Nobel, 2013). In these work, there are only consistency results available (Donoho et al., 2018) which say that $\hat{\sigma}^2 \rightarrow \sigma^2$ almost surely, but there are no explicit error rates. Using the recent advancement in random matrix theory on sharp large-deviation bounds for individual empirical eigenvalues (see Ke (2016) for a survey), we can leverage those results to obtain an explicit bound for $|\hat{\sigma}^2 - \sigma^2|$.

We then establish the consistency on estimating K .

Theorem 2.4.2 (Consistency of \hat{K}). Suppose $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n$ satisfy Assumptions 2.4.1-2.4.2 with $\sigma_j^2 \equiv \sigma^2$. Suppose $K \geq 1$ is fixed, $p/n \rightarrow \gamma \in (0, \infty)$, and $\mu_K \geq \sigma^2(\sqrt{\gamma} + \tau_n)$, where $\tau_n \gg n^{-1/3}$. Let \hat{K} be the estimator of K by Algorithm 1, where the tuning parameters are such that $\alpha \in (0, 1/2)$ is a constant and that $\beta \rightarrow 0$ at a properly slow rate. As $n \rightarrow \infty$, $\mathbb{P}\{\hat{K} = K\} = 1 - o(1)$.

We compare the conditions required for consistent estimation of K with those in other work. Let $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p$ denote the eigenvalues of Σ . In our model, $\lambda_k = \mu_k + \sigma^2$ for $1 \leq k \leq K$. The condition in Theorem 2.4.2 translates to

$$\lambda_K > \sigma^2(1 + \sqrt{\gamma} + \tau_n), \quad \tau_n \gg n^{-1/3}.$$

It is weaker than the conditions in Bai & Ng (2002) and Cai et al. (2020), where the former requires $\lambda_K \asymp p$ and the latter needs $\lambda_K \rightarrow \infty$. Our condition on λ_K matches with the critical phase transition threshold in Baik et al. (2005) and is hardly improvable. In fact, Fan et al. (2020) showed that if $\lambda_K \leq \sigma^2(1 + \sqrt{\gamma})$ then there exists no consistent estimator of K . Dobriban & Owen (2019) impose the same condition on λ_K , but they need stronger conditions on population eigenvectors. Their “de-localization” condition states as $\|\Xi\Lambda^{1/2}\|_\infty \rightarrow 0$, where $\Xi = [\xi_1, \dots, \xi_K]$, $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_K)$, and $\|\cdot\|_\infty$ is the maximum absolute row sum. It requires the eigenvectors to be incoherent (i.e., $\max_{1 \leq k \leq K} \|\xi_k\|_\infty$ is sufficiently small) and that the eigenvalues cannot be too large. Examples such

as equal-correlation matrices (i.e., $\Sigma(i, j) = a$, for all $i \neq j$, where $a \in (0, 1)$ is a constant) are excluded. We do not need such a de-localization condition.[†]

The proof of Theorem 2.4.2 is an application of the *eigenvalue sticking* theory (Bloemendal et al., 2016). It compares the distribution of empirical eigenvalues $\{\hat{\lambda}_k\}$ under the spiked covariance model with the distribution of empirical eigenvalues $\{\hat{\lambda}_k^*\}$ under the null model $\Sigma = \sigma^2 \mathbf{I}_p$. The claim is that the distribution of $\hat{\lambda}_{K+s}$ is asymptotically close to the distribution of $\hat{\lambda}_s^*$, for a wide range of s . We use this result to study the thresholding step in Algorithm 1.

2.4.2 THE TRUNCATED GAMMA-BASED GENERAL SPIKED COVARIANCE MODEL

The general spiked covariance model (2.7) assumes σ_j^2 are *iid* drawn from $\text{Gamma}(\theta, \theta/\sigma^2)$. It differs from the conventional settings in random matrix theory because Σ is not a deterministic matrix and because the limiting spectral density of Σ does not have a compact support. Unfortunately, there is no existed random matrix theory that deals with this setting directly (Bao, 2020). We thus approximate Model (2.2) by

$$\sigma_j^2 \stackrel{iid}{\sim} \text{TruncGamma}(\theta, \theta/\sigma^2, \sigma^2 T_1, \sigma^2 T_2), \quad 1 \leq j \leq p, \quad (2.13)$$

where $\text{TruncGamma}(\alpha, \beta, l, u)$ denotes the truncated Gamma distribution with rate and shape parameters α and β and truncations at l and u . When $(T_1, T_2) = (0, \infty)$, it reduces to Model (2.2). Given fixed $0 < T_1 < T_2 < \infty$, the limiting spectral density of Σ has a compact support, so that we can take advantage of the existing random matrix theory (Knowles & Yin, 2017, Ding, 2020). We first present the theory for Model (2.13) and then discuss how to extend it to $(T_1, T_2) = (0, \infty)$.

[†]We remark that the comparison is for the standard spiked covariance model only. For this model, our method has the weakest conditions for consistent estimation of K . On the other hand, other methods apply to some other settings, which are not considered in the comparison.

Fixing $0 < T_1 < T_2 < \infty$ and two intervals $\mathcal{J}_{\sigma^2} = [a, b] \subset (0, \infty)$ and $\mathcal{J}_{\theta} = [c, d] \in (0, \infty)$, let $\mathcal{Q}(T_1, T_2, \mathcal{J}_{\sigma^2}, \mathcal{J}_{\theta})$ be the family of distributions $\text{TruncGamma}(\theta, \theta/\sigma^2, \sigma^2 T_1, \sigma^2 T_2)$ satisfying that $\sigma^2 \in \mathcal{J}_{\sigma^2}$ and $\theta \in \mathcal{J}_{\theta}$. The following Lemma is a result of Theorem 3.12 and Example 2.9 in Knowles & Yin (2017), and its proof is omitted.

Lemma 2.4.1. Suppose $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n$ satisfy Assumptions 2.4.1-2.4.2 with σ_j^2 generated from Model (2.13). Suppose $K \geq 1$ is fixed and $p/n \rightarrow \gamma$ for a constant $\gamma \neq 1$. Suppose the truncated Gamma distribution in (2.13) is from the family $\mathcal{Q}(T_1, T_2, \mathcal{J}_{\sigma^2}, \mathcal{J}_{\theta})$, for fixed $(T_1, T_2, \mathcal{J}_{\sigma^2}, \mathcal{J}_{\theta})$. Let $H_{\sigma^2, \theta, T_1, T_2}(t)$ be the CDF of $\text{TruncGamma}(\theta, \theta/\sigma^2, \sigma^2 T_1, \sigma^2 T_2)$. Define a distribution $F_{\gamma_n}(\cdot; \sigma^2, \theta, T_1, T_2)$ in the same way as in (2.8)-(2.9), with $H_{\sigma^2, \theta}(t)$ replaced by $H_{\sigma^2, \theta, T_1, T_2}(t)$ and γ replaced by $\gamma_n = p/n$. Let $q_i \equiv \bar{F}_{\gamma_n}^{-1}(i/\tilde{p}; \sigma^2, \theta, T_1, T_2)$ be the (i/\tilde{p}) -upper-quantile of this distribution, where $\tilde{p} = n \wedge p$. As $n \rightarrow \infty$, for every $K < i \leq \tilde{p}$, we have $|\hat{\lambda}_i - q_i| \prec [i \wedge (\tilde{p} + 1 - i)]^{-1/3} n^{-2/3}$.

Given (T_1, T_2) , we estimate σ^2 and θ by

$$(\hat{\sigma}^2, \hat{\theta}) = \underset{(\sigma^2, \theta) \in \mathcal{J}_{\sigma^2} \times \mathcal{J}_{\theta}}{\text{argmin}} \left\{ \sum_{\alpha \tilde{p} \leq i \leq (1-\alpha)\tilde{p}} [\hat{\lambda}_i - \bar{F}_{\gamma_n}^{-1}(i/\tilde{p}; \sigma^2, \theta, T_1, T_2)]^2 \right\}. \quad (2.14)$$

It can be solved by a slight modification of Step 1 of Algorithm 2. We note that (2.13) is equivalent to $\sigma_j^2/\sigma^2 \stackrel{iid}{\sim} \text{TruncGamma}(\theta, \theta, T_1, T_2)$. Hence, the quantiles satisfy that $\bar{F}_{\gamma_n}^{-1}(i/\tilde{p}; \sigma^2, \theta, T_1, T_2) = \sigma^2 \cdot \bar{F}_{\gamma_n}^{-1}(i/\tilde{p}; 1, \theta, T_1, T_2)$. We first modify GetQT so that it outputs the quantiles of $F_{\gamma_n}(\cdot; 1, \theta, T_1, T_2)$ for any given θ . Next, we mimic Step 1 of Algorithm 2 to solve (2.14), where we run a least-squares for every θ and then optimize over θ via a grid search. The details are relegated to the Appendix.

Theorem 2.4.3 (Estimation error of $\hat{\sigma}^2$ and $\hat{\theta}$). Suppose the conditions of Lemma 2.4.1 hold, where $K, \gamma, T_1, T_2, \mathcal{J}_{\sigma^2}$, and \mathcal{J}_{θ} are fixed. Let

$$\Phi(\theta) = \Phi(\theta; T_1, T_2) = \frac{[\int_{T_1}^{T_2} x^{\theta+1} \exp(-\theta x) dx] [\int_{T_1}^{T_2} x^{\theta-1} \exp(-\theta x) dx]}{[\int_{T_1}^{T_2} x^{\theta} \exp(-\theta x) dx]^2}.$$

Suppose there exists a constant $\omega = \omega(T_1, T_2, \mathcal{J}_\theta)$ such that $\sup_{\theta \in \mathcal{J}_\theta} \Phi'(\theta) \leq -\omega$. Let $\hat{\sigma}^2$ and $\hat{\theta}$ be the estimators from (2.14), where the tuning parameter α satisfies $\alpha\tilde{p} > K$ and $\alpha\tilde{p} = O(n/\log(n))$. As $n \rightarrow \infty$, we have $|\hat{\sigma}^2 - \sigma^2| \prec n^{-1}$ and $|\hat{\theta} - \theta| \prec n^{-1}$.

Theorem 2.4.3 assumes $\sup_{\theta \in \mathcal{J}_\theta} \Phi'(\theta) \leq -\omega$ for some constant $\omega > 0$. It is a regularity condition on $(\mathcal{J}_\theta, T_1, T_2)$. The next lemma shows that this condition is mild.

Lemma 2.4.2. For any fixed $\mathcal{J}_\theta = [c, d]$ and $\omega < d^{-2}$, there exist constants $0 < T_1^* < T_2^* < \infty$ such that $\sup_{\theta \in \mathcal{J}_\theta} \Phi'(\theta; T_1, T_2) \leq -\omega$ holds for all $T_1 \leq T_1^*$ and $T_2 \geq T_2^*$.

With the estimates $\hat{\sigma}^2$ and $\hat{\theta}$, we then slightly modify Step 2 of Algorithm 2 by thresholding all the empirical eigenvalues at

$$\hat{T}_\beta = \left\{ \begin{array}{l} (1 - \beta)\text{-quantile of the distribution of } \hat{\lambda}_1^* \text{ under the null model} \\ \Sigma = \text{diag}(\sigma_1^2, \dots, \sigma_p^2), \text{ where } \sigma_j^2 \stackrel{iid}{\sim} \text{TruncGamma}(\hat{\theta}, \hat{\theta}/\hat{\sigma}^2, \hat{\sigma}^2 T_1, \hat{\sigma}^2 T_2) \end{array} \right\}. \quad (2.15)$$

This threshold can be computed via Monte Carlo simulations, similarly as in Step 2 of Algorithm 2. We estimate K by the number of empirical eigenvalues exceeding \hat{T} .

To establish the consistency of \hat{K} , we introduce the function

$$G(x) = -\frac{1}{x} + \gamma \int \frac{1}{t^{-1} + x} dH_{\sigma^2, \theta, T_1, T_2}(t). \quad (2.16)$$

By Example 2.9 of Knowles & Yin (2017), $G(x)$ has 2 critical points $0 > x_1^* > x_2^*$ (the definition of critical points can be found in Knowles & Yin (2017)), and the distribution $F_\gamma(\cdot; \sigma^2, \theta, T_1, T_2)$ defined in Lemma 2.4.1 has the support $[G(x_2^*), G(x_1^*)]$. The next theorem is proved in the appendix. It uses a result in Ding (2020) about the top empirical eigenvalues.

Theorem 2.4.4 (Consistency of \hat{K}). Suppose the conditions of Lemma 2.4.1 and Theorem 2.4.3 hold. Let x_1^* be the largest critical point of the function $G(x)$ in (2.16). We assume $-1/(T_1 + \mu_K) \geq$

$x_1^* + \tau,^\ddagger$ where $\tau > 0$ is a constant and T_1 is a truncation point in (2.13). Let $\hat{K} = \#\{1 \leq i \leq (n \wedge p) : \hat{\lambda}_i > \hat{T}_\beta\}$, where \hat{T}_β is as in (2.15) with $\beta \rightarrow 0$ at a properly slow rate. As $n \rightarrow \infty$, $\mathbb{P}\{\hat{K} = K\} = 1 - o(1)$.

2.4.3 REMARKS ON EXTENSION TO THE GAMMA-BASED GENERAL SPIKED COVARIANCE MATRIX

We now discuss extension of the theoretical results to the Gamma-based general spiked covariance model (2.7), which is an extreme case of Model (2.13) at $T_1 = 0$ and $T_2 = \infty$. As mentioned earlier, this setting is unconventional because the eigenvalues of Σ are stochastic and the support of the limiting spectral density of Σ is unbounded.

First, we discuss the estimation of (σ^2, θ) . The accuracy of $(\hat{\sigma}^2, \hat{\theta})$ depends on whether we have similar large deviation bounds to those in Lemma 2.4.1. Our conjecture is that the stochasticity and unboundedness of the spectrum of Σ has a negligible effect on the eigenvalues deep into the bulk. To see why, we note that the classical result about weak convergence of ESD (Marcenko & Pastur, 1967) does not need the limiting spectrum of Σ to have a compact support; therefore, the unboundedness is not an issue. The stochasticity is not an issue, either, because almost surely, the spectral distribution of Σ converges weakly to $\text{Gamma}(\theta, \theta/\sigma^2)$. We conclude that the weak convergence of ESD still holds. This further implies that the bulk eigenvalues still converge to the corresponding quantiles of the theoretical limit of ESD.

The open question is whether we have the rates of convergence as in Lemma 2.4.1. The stochasticity and unboundedness of the spectrum of Σ affect the rates of convergence of large eigenvalues. We thus do not expect Lemma 2.4.1 to hold for all i . Fortunately, the estimation of (σ^2, θ) in BEMA

[‡]In our model (see Assumption 2.4.1), the spiked eigenvalues of Σ are $\{\mu_k + \sigma_k\}_{1 \leq k \leq K}$. Therefore, $\mu_K + T_1$ is a lower bound of these spiked eigenvalues.

only involves bulk eigenvalues in the middle range, i.e., $\alpha\tilde{p} \leq i \leq (1 - \alpha)\tilde{p}$, where $\alpha \in (0, 1/2)$ is a constant. We conjecture that Lemma 2.4.1 continues to hold for these eigenvalues. If our conjecture is correct, then we can show similar results for $\hat{\sigma}^2$ and $\hat{\theta}$ as those in Theorem 2.4.3.

Next, we discuss the consistency of \hat{K} . The stochasticity and unboundedness of the spectrum of Σ together yields a significant change of the behavior of edge eigenvalues. This can be seen from a relevant setting in Kwak et al. (2019)— Σ is a diagonal matrix whose diagonal entries are *iid* drawn from a density $\rho(t) \propto (1 - t)^b f(t) \cdot 1\{l \leq t \leq 1\}$, where $b > 1$ and $l \in (0, 1)$ are constants and $f \in C^1([l, 1])$. This setting has no spike. They showed that the limiting distribution of the largest eigenvalue, $\hat{\lambda}_1^*$, is not a Tracy-Widom distribution; it is a Weibull distribution if $\gamma < \gamma_0$ and a Gaussian distribution if $\gamma > \gamma_0$, where γ_0 is a positive constant. Our model is even more complicated, where the Gamma density exhibits a similar convex decay on the right tail but has an unbounded support. We do not expect $\hat{\lambda}_1^*$ to follow a Tracy-Widom distribution any more.

However, this does not eliminate the consistency of \hat{K} . To prove consistency, we first need that the stochastic threshold (2.12) in BEMA well approximates the $(1 - \alpha)$ -upper-quantile of $\hat{\lambda}_1^*$, where $\hat{\lambda}_1^*$ is the largest eigenvalue of the null model with no spike. This follows from the nature of Monte Carlo simulations, no matter whether $\hat{\lambda}_1^*$ converges to a Tracy-Widom distribution. Furthermore, the implementation of (2.12) does not need any knowledge of the limiting distribution of $\hat{\lambda}_1^*$.

To prove consistency, we also need to show that, under Model (2.7), when μ_K is appropriately large, (i) the distribution of $\hat{\lambda}_{K+1}$ is asymptotically close to the distribution of $\hat{\lambda}_1^*$ in the null model (this is the “eigenvalue sticking” argument), and (ii) each of $\hat{\lambda}_1, \hat{\lambda}_2, \dots, \hat{\lambda}_K$ is significantly larger than the $(1 - \alpha)$ -upper-quantile of $\hat{\lambda}_1^*$. We conjecture that both (i)-(ii) are correct, provided that $\mu_K \gg \log(n)$. If our conjectures are correct, then we can obtain the consistency of \hat{K} as in Theorem 2.4.4, under the slightly stronger condition that $\mu_K \gg \log(n)$.

The rigorous proofs of our conjectures require re-development of several fundamental results in random matrix theory for Model (2.7), such as the local law on bulk eigenvalues and the limiting

behavior of edge eigenvalues (including the spiked and non-spiked ones). It is beyond the scope of this chapter, and we leave for future work.

2.5 SIMULATION STUDIES

We examine the performance of our methods in simulations. To differentiate between Algorithm 1 and Algorithm 2, we call the former BEMA0 and the latter BEMA. BEMA0 is a simplified version of BEMA, specifically designed for the standard spiked covariance model. The tuning parameters are fixed as $(\alpha, \beta) = (0.2, 0.1)$ for BEMA0 and $(\alpha, \beta, M) = (0.2, 0.1, 500)$ for BEMA when not particularly specified.

In Section 2.4.2, we also introduced a modification of BEMA using the truncated Gamma-based spiked mode for technical needs in our theoretical studies. We showed that this algorithm has desirable theoretical properties. It however requires two additional tuning parameters (T_1, T_2) . Our simulation studies (not reported here) show that the performance of the modified BEMA is similar to that of BEMA, when T_1 is appropriately small and T_2 is appropriately large. For this reason, we use BEMA, instead of the modified BEMA, in the following simulation studies.

We compare our methods with a few methods in the literature, including the deterministic parallel analysis (DDPA) from [Dobriban & Owen \(2019\)](#), the empirical Kaiser’s criterion (EKC) from [Braeken & Van Assen \(2017\)](#), the information criteria IC_{p1} (Bai&Ng) from [Bai & Ng \(2002\)](#) and the eigen-gap detection (Pass&Yao) from [Passemier & Yao \(2014\)](#).

SIMULATION 1. This experiment is for the standard spiked covariance model, where we investigate the performance of BEMA0 and the confidence interval for K as described in Section 2.3.3. We generate data from $\mathbf{X}_i \stackrel{iid}{\sim} N(0, \Sigma), 1 \leq i \leq n$, where Σ satisfies Model (2.3) with

$$\mu_1 = \mu_2 = \cdots = \mu_K = \rho \cdot \sigma^2 \sqrt{p/n}, \quad \text{for some } \rho > 0.$$

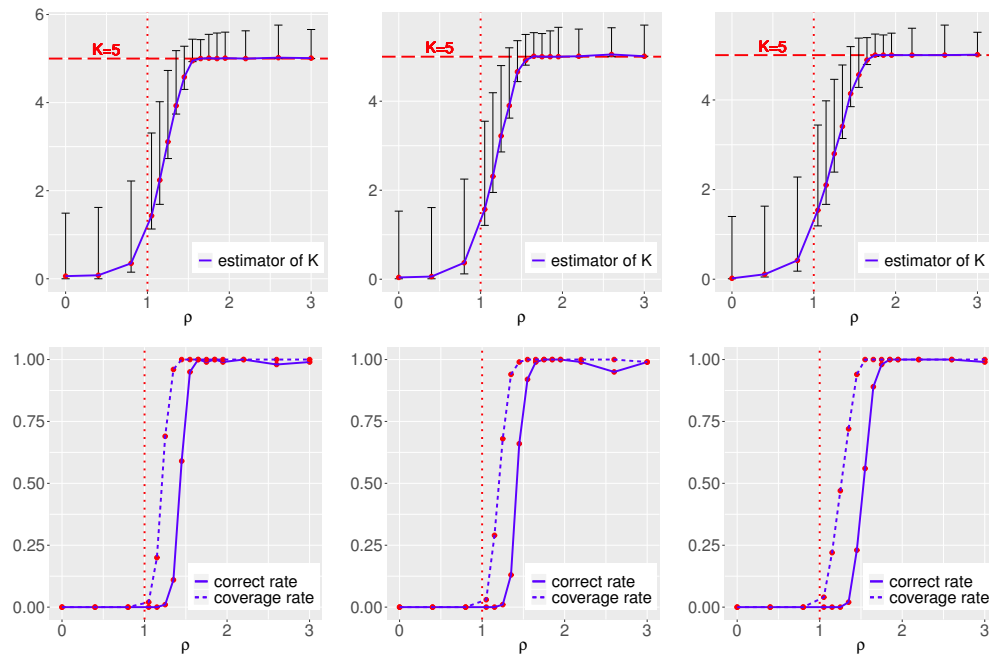


Figure 2.5: Simulation 1: The performance of BEMA0 in a standard spiked model. $K = 5$, and (n, p) take the value of $(10000, 1000)$, $(1500, 5000)$, and $(1500, 1500)$ (from left to right). The top three panels show the estimator \hat{K} along with the 95% confidence upper/lower bound, where each quantity is the average of 100 repetitions. The bottom three panels show the probability of correctly estimating K (correct rate) and the coverage probabilities of the 95% confidence intervals (coverage rates). In each panel, the x-axis is the value of ρ (see the text for definition), controlling the magnitude of spiked eigenvalues. Our theory states that BEMA0 gives a consistent estimator of K when ρ slightly exceeds 1. This is confirmed by these simulations.

The value of ρ controls the magnitude of spiked eigenvalues. $\rho \leq 1$ is the region where consistent estimation of K is impossible (Baik et al., 2005, Fan et al., 2020). We examine the performance of BEMA0 in the region of $\rho > 1$.

Fix $K = 5$ and $\sigma^2 = 1$. We consider three settings, where (n, p) are $(10000, 1000)$, $(1500, 5000)$, and $(1500, 1500)$, respectively. They cover different cases of size relationship between p and n . The eigenvector matrix \mathbf{Z} is drawn uniformly from the Stiefel manifold (which is the collection of all $p \times K$ matrices that have orthonormal columns). For each of the three settings, we vary the value of ρ and report the average of \hat{K} and upper/lower boundary of a 95% confidence interval, based on 100 repetitions; the results are in the top three panels of Figure 2.5. We also report the probability of correctly

estimating K (correct rate) and the coverage probability of the 95% confidence interval (coverage rate); see the bottom three panels of Figure 2.5.

It agrees with our theoretical understanding that $\rho = 1$ is the critical phase transition point. When ρ slightly departs from 1, the coverage rate starts to increase from 0% and quickly reaches the target of 95%. The increase of the correct rate is slightly slower, but it reaches 100% before $\rho = 1.5$, for all three settings. Our theory suggests that the correct rate is asymptotically 100% as long as $\rho > 1$, but in the finite-sample performance we need a larger ρ to attain a 100% correct rate. Furthermore, as ρ increases, the estimated \hat{K} increases from 0 to 5, with a sharp change at around $\rho = 1$. The length of the 95% confidence interval initially decreases with ρ and then stays almost constant.

SIMULATION 2. In this simulation, we compare BEMA0 and BEMA with other methods. We consider both the standard spiked covariance model (2.3) and the general spiked covariance model (2.7). BEMA0 and BEMA are designed for these two settings, respectively. We note that BEMA can also be applied to Model (2.3), which simply ignores the prior knowledge of equal diagonal in the residual covariance matrix. We thereby also include BEMA in the numerical comparison on the standard spiked covariance model.

Given $(n, p, K, \lambda, \theta)$, we generate data $\mathbf{X}_i \stackrel{iid}{\sim} N(0, \Sigma)$, $1 \leq i \leq n$, where Σ satisfies Model (2.7) with $\sigma^2 = 1$ and $\mu_k = \lambda$, for $1 \leq k \leq K$. The eigenvector matrix Ξ is drawn uniformly from the Stiefel manifold. We allow θ to take the value of ∞ ; when $\theta = \infty$, it indicates that Σ follows the standard spiked covariance model (2.3). We consider 8 different settings which cover a wide range of parameter values. The results are shown in Table 2.1, where the average \hat{K} and the probability of correctly estimating K (correct rate) are reported based on 500 repetitions.

We have a few observations. First, in the standard spiked covariance model ($\theta = \infty$, top four rows of Table 2.1), BEMA0 has the best performance. Interestingly, BEMA has nearly comparable performance. The reason is that the algorithm will automatically output a very large $\hat{\theta}$, so that the estimator

$(n, p, K, \lambda, \theta)$	BEMA0	BEMA	DDPA	EKC	Bai&Ng	Pass&Yao
(100, 500, 5, 9, ∞)	4.996 (99.6%)	4.982 (98.2%)	6.102 (41%)	5.552 (57.8%)	0 (0%)	4.904 (92%)
(100, 500, 5, 49, ∞)	5 (100%)	5 (100%)	6.328 (38%)	6.4 (27.4%)	5 (100%)	5.012 (98.8%)
(500, 100, 5, 1.5, ∞)	5 (100%)	4.93 (93.0%)	6.1 (45.6%)	5.016 (98.4%)	0 (0%)	2.784 (43.8%)
(500, 100, 5, 3, ∞)	5 (100%)	5 (100%)	5.92 (45.4%)	5.056 (94.4%)	0 (0%)	4.432 (84.4%)
(100, 500, 5, 15, 3)	–	5.182 (85.2%)	9.222 (20.8%)	5.974 (40.2%)	0.078 (0%)	5.292 (73.2%)
(100, 500, 5, 50, 3)	–	5.142 (88.4%)	9.214 (20.8%)	9.852 (8.6%)	5 (100%)	5.362 (70.4%)
(500, 100, 5, 4.5, 3)	–	4.748 (81.2%)	57.954 (25.4%)	5.588 (49.0%)	3.392 (39%)	7.624 (5%)
(500, 100, 5, 6, 3)	–	5.018 (98.2%)	43.734 (38.8%)	6.244 (18.4%)	5.002 (99.8%)	8.098 (4.2%)

Table 2.1: Simulation 2: Comparison of different methods in the standard/general spiked model. In these settings, all the spiked eigenvalues are equal to λ , and the eigenvectors are randomly generated from the Stiefel manifold. The top four rows ($\theta = \infty$) correspond to the standard spiked model, and the bottom four rows correspond to the general spiked model. The number in each cell is the average \hat{K} over 500 repetitions, and the number in brackets is the probability of correctly estimating K (correct rate).

is similar to that of knowing $\theta = \infty$. This suggests that we do not have to choose between BEMA0 and BEMA in practice. We can always use BEMA, even when the data come from the standard spiked covariance model. On the other hand, BEMA0 is conceptually simpler and computationally much faster, hence, it is still the better choice if we are confident that the standard spiked covariance model holds.

Second, in the general spiked covariance model (bottom four rows of Table 2.1), BEMA outperforms DDPA, EKC and Pass&Yao in all settings, and outperforms Bai&Ng in two out of four settings. BEMA is the only method whose correct rate is above 80% in *all settings*.

DDPA requires a delocalization condition. Let Ξ be the $p \times K$ matrix of eigenvectors, and let Λ be the diagonal matrix consisting of spiked eigenvalues. The delocalization condition is $\|\Xi\Lambda^{1/2}\|_\infty \rightarrow 0$. It prevents eigenvectors from having large entries. This condition is not satisfied here, explaining the unsatisfactory performance of DDPA. Bai&Ng requires that the spikes are sufficiently large. The larger p/n , the higher requirement of spikes. When $p/n = 5$ and $\lambda = 49$ or when $p/n = 0.2$ and $\lambda = 6$, Bai&Ng has a nearly 100% correct rate. However, as λ decreases, the correct rate drops very quickly. EKC uses a thresholding scheme that gives smaller thresholds to lower ranked eigenvalues (e.g., the threshold for $\hat{\lambda}_2$ is smaller than the threshold for $\hat{\lambda}_1$). This method often over-estimates K ,

(n, p, K, s_1, θ)	BEMA0	BEMA	DDPA	EKC	Bai&Ng	Pass&Yao
(100, 500, 1, 1, ∞)	0.988 (96%)	0.956 (95.2%)	1.086 (88.6%)	1.07 (88.8%)	0 (0%)	0.934 (91.8%)
(100, 500, 1, 3, ∞)	1.012 (98.8%)	1.008 (99.2%)	1.138 (87%)	1.146 (86.4%)	1 (100%)	1.036 (96.8%)
(500, 100, 1, 3, ∞)	1.020 (98%)	1 (100%)	1.152 (85.6%)	1.056 (94.4%)	0 (0%)	1.018 (98.2%)
(500, 100, 1, 6, ∞)	1.014 (98.6%)	1 (100%)	1.124 (88.6%)	1.12 (88%)	1 (100%)	1.014 (98.8%)
(100, 500, 1, 2, 10)	–	1.096 (90.6%)	1.2 (82.6%)	1.102 (90.4%)	0.388 (38.8%)	1.084 (92.6%)
(100, 500, 1, 6, 10)	–	1.104 (89.8%)	1.226 (79%)	1.608 (54.2%)	1 (100%)	1.054 (95%)
(500, 100, 1, 6, 3)	–	1.114 (89.2%)	1.062 (95.4%)	1.226 (78.2%)	1.008 (99.4%)	3.93 (6.2%)
(500, 100, 1, 12, 3)	–	1.124 (88.0%)	1.042 (97.4%)	3.782 (0.8%)	1.006 (99.4%)	3.672 (9.8%)

Table 2.2: Simulation 3: Comparison of different methods in the standard/general spiked model, when the eigenvectors are ‘delocalized’. Here, s_1 controls the magnitude of spiked eigenvalues, where $s_1^2(p/n)$ plays the role of λ in Simulation 2. The top four rows ($\theta = \infty$) correspond to the standard spiked model, and the bottom four rows correspond to the general spiked model. The number in each cell is the average \hat{K} , and the number in brackets is the probability of correctly estimating K (correct rate).

especially when all the spikes are large (e.g., Row 6 of Table 2.1). Pass&Yao is developed for the standard spiked model. It has an unsatisfactory performance in the general spiked model (bottom four rows of Table 2.1).

SIMULATION 3. In this simulation, we change the generation process of eigenvectors to satisfy the “delocalization condition” (Dobriban & Owen, 2019). This condition means $\|\Xi\Lambda^{1/2}\|_\infty$ is sufficiently small, where Ξ is the $p \times K$ matrix consisting of eigenvectors and Λ is the diagonal matrix consisting of spiked eigenvalues.

We adapt the simulation settings in Dobriban & Owen (2019) to our general spiked model. Given (n, p, K, θ) and $s_1, \dots, s_K > 0$, we generate $\mathbf{X}_i \stackrel{iid}{\sim} N(0, \Sigma)$, $1 \leq i \leq n$, where $\Sigma = \mathbf{B}\mathbf{B}^\top + \mathbf{D}$. The matrix $\mathbf{D} = \text{diag}(\sigma_1^2, \sigma_2^2, \dots, \sigma_p^2)$ is generated in the same way as in Model (2.7), and \mathbf{B} is a $p \times K$ matrix obtained by first generating a $p \times K$ matrix with independent $N(0, 1)$ entries and then re-normalizing each column to have an ℓ^2 -norm equal to $s_k \sqrt{p/n}$. Under this setting, the L_∞ -norm of each population eigenvector is only $O(p^{-1/2} \sqrt{\log(p)})$, so the “delocalization” condition is satisfied. We fix $K = 1$ and let (n, p, s_1, θ) vary. The results are shown in Table 2.2.

Compared with Simulation 2, the performance of DDPA is significantly better. BEMA0 and

BEMA continue to perform well, indicating that their performance is insensitive to the generating process of eigenvectors. This is consistent with our theoretic understanding. In Section 2.4, we have seen that the success of BEMA0 and BEMA requires no conditions on eigenvectors.

SIMULATION 4. In this simulation, we investigate the case of model misspecification. We still assume that Σ is a low-rank matrix plus a residual covariance matrix \mathbf{D} . However, we no longer let \mathbf{D} be a diagonal matrix. Below, we consider three misspecified models, where \mathbf{D} is a Toeplitz matrix, a block-wise diagonal matrix, and a sparse matrix, respectively.

- In the first model, $\mathbf{D}(i, j) = (1 + |i - j|)^{-t}$, for $1 \leq i, j \leq p$. Here, \mathbf{D} is a Toeplitz matrix with polynomial decays in the off-diagonal. The larger t , the closer to a diagonal matrix.
- In the second model, $\mathbf{D}(i, i) = 1$ for $1 \leq i \leq p$, and $\mathbf{D}(2j - 1, 2j) = \mathbf{D}(2j, 2j - 1) = b$ for $1 \leq j \leq p/2$. \mathbf{D} is a block-wise diagonal matrix which has many 2×2 diagonal blocks. The smaller b , the closer to a diagonal matrix.
- In the third model, $\mathbf{D}(i, i) = 1$ for $1 \leq i \leq p$, and $\mathbf{D}(i, j) = \mathbf{D}(j, i) \sim c \cdot \text{Bernoulli}(0.1)$ for $i \neq j$. The matrix \mathbf{D} has approximately $0.1p$ nonzero entries in each row. The smaller c , the closer to a diagonal matrix.

The low-rank part of Σ is generated in the same way as before: We let all μ_k equal to λ and let the eigenvector matrix Ξ be drawn uniformly from the Stiefel manifold, which allows Ξ to have orthonormal columns. Fix $(n, p, K) = (500, 100, 1)$. The results are shown in Table 2.3.

For each misspecified model, we consider two settings, where \mathbf{D} is closer to a diagonal matrix in the first setting (Rows 1,3,5 of Table 2.3) than in the second one (Rows 2,4,6 of Table 2.3). Every method performs better in the first case, suggesting that the diagonal assumption on \mathbf{D} is indeed critical. In comparison, BEMA is least sensitive to a non-diagonal \mathbf{D} . In Rows 2,4,6 of Table 2.3, the correct rate

λ	residual covariance	BEMA0	BEMA	DDPA	EKC	Bai&Ng	Pass&Yao
6	Toeplitz($\tau=4$)	1.104 (89.6%)	1 (100%)	1.422 (65.4%)	1.36 (67.4%)	1 (100%)	1.06 (94.8%)
3	Toeplitz($\tau=2$)	9.352 (0%)	1.12 (88.6%)	100 (0%)	15.148 (0%)	0 (0%)	2.46 (24.6%)
6	block diagonal($b=0.1$)	1.344 (66.8%)	1 (100%)	2.378 (31.6%)	1.854 (33.6%)	1 (100%)	1.038 (96.6%)
3	block diagonal($b=0.2$)	3.764 (0%)	1 (100%)	100 (0%)	6.602 (0%)	0 (0%)	1.12 (89.8%)
6	sparse($c=0.05$)	1.784 (30.2%)	1.016 (98.4%)	5.024 (9.4%)	2.474 (9.4%)	1 (100%)	1.084 (91.6%)
3	sparse($c=0.08$)	3.348 (0%)	1.036 (96.4%)	97.752 (0%)	5.18 (0%)	0 (0%)	1.58 (47.4%)

Table 2.3: Simulation 4: Comparison of different methods in three misspecified models, where the residual covariance matrix \mathbf{D} is a Toeplitz matrix, a block diagonal matrix, and a sparse matrix, respectively. $(n, p, K) = (500, 100, 1)$. The spiked eigenvalue is equal to λ . For each misspecified model, we consider two settings, where \mathbf{D} is closer to a diagonal matrix in the first setting (rows 1, 3, 5) than in the second setting (rows 2, 4, 6). The number in each cell is the average \hat{K} , and the number in brackets is the probability of correctly estimating K (correct rate).

of BEMA is still above 80%, while the correct rate of some other methods is only 0%. Pass&Yao is the second least sensitive to a non-diagonal \mathbf{D} .

To try to understand this phenomenon, we first note that one can always apply an orthogonal transformation to data vectors $\mathbf{X}_1, \dots, \mathbf{X}_n$, so that the post-transformation data follow a different spiked covariance model whose residual covariance matrix $\tilde{\mathbf{D}}$ is a diagonal matrix containing the eigenvalues of \mathbf{D} . This orthogonal transformation is unknown in practice. However, if a method uses the empirical eigenvalues *only*, it does not matter whether or not we know this orthogonal transformation, because any orthogonal transformation does not change eigenvalues of the sample covariance matrix and thus it does not change the estimator of K . It implies that, for methods that only use eigenvalues, we can treat the misspecified model as if \mathbf{D} is replaced by the diagonal matrix $\tilde{\mathbf{D}}$. Therefore, the surprising robustness of BEMA can be interpreted as the capability of the gamma model (2.2) in approximating the eigenvalue structure in \mathbf{D} . The flexibility of this gamma model comes from the parameter θ . In comparison, such strong robustness is not observed for BEMA0, where θ is fixed as ∞ .

The method of DDPA uses empirical eigenvectors in the procedure, thus, it is more sensitive to the diagonal assumption of \mathbf{D} . EKC uses eigenvalues only, but its thresholding scheme is too con-

distribution	(λ, θ)	BEMA0 (0.1)	BEMA0 (0.2)	BEMA0 (0.3)	BEMA (0.1)	BEMA (0.2)	BEMA (0.3)
Gaussian	$(1.5, \infty)$	5 (100%)	5 (100%)	5 (100%)	4.95 (95%)	4.93 (93%)	4.904 (90.4%)
Random sign	$(1.5, \infty)$	4.996 (99.6%)	4.996 (99.6%)	4.998 (99.8%)	4.972 (97.2%)	4.96 (96%)	4.94 (94%)
Laplace	$(1.5, \infty)$	4.998 (99.8%)	4.998 (99.8%)	4.998 (99.8%)	4.914 (91.4%)	4.9 (90%)	4.88 (88%)
Gaussian	$(4.5, 3)$	–	–	–	4.518 (69%)	4.748 (81.2%)	4.76 (81%)
Random sign	$(4.5, 3)$	–	–	–	4.678 (78.4%)	4.818 (85%)	4.9 (85.4%)
Laplace	$(4.5, 3)$	–	–	–	4.352 (56.8%)	4.634 (73.8%)	4.656 (74.8%)

Table 2.4: Simulation 5: The robustness of BEMA0 and BEMA under non-Gaussian data and different values of α . Data are generated from the factor model with Gaussian/random-sign/Laplace factors and noise. $K = 5$, and all the spiked eigenvalues are equal to λ . BEMA0 and BEMA are implemented with $\alpha \in \{0.1, 0.2, 0.3\}$ (denoted as BEMA0 (α)/BEMA (α) in the table). The number in each cell is the average \hat{K} , and the number in brackets is the probability of correctly estimating K (correct rate).

servative. In these misspecified models, some bulk empirical eigenvalues can get large; EKC gives too small thresholds to non-leading eigenvalues and yields over-estimation of K .

SIMULATION 5. In this simulation, we tested the robustness of our proposed methods against the choice of α and the distributional assumption on data generation. Fix $(n, p, K) = (500, 100, 5)$. We generate $\mathbf{X}_i = \mathbf{\Xi}\omega_i + \varepsilon_i$ where $\mathbf{\Xi} \in \mathbb{R}^{p \times K}$ is uniformly drawn from the Stiefel manifold, ω_i are *iid* drawn from a multivariate zero-mean distribution with covariance matrix $\lambda\mathbf{I}_K$, ε_i are *iid* drawn from a multivariate zero-mean distribution with covariance matrix \mathbf{D} , and \mathbf{D} is generated in the same way as in Model (2.7) with $\sigma^2 = 1$ and $\theta \in \{\infty, 3\}$. We consider three settings where the entries of ω_i and ε_i are Gaussian, random sign, or Laplace variables (centered and re-scaled to match the required variance), respectively. The results are in shown Table 2.4.

For the standard spiked covariance model (top 3 rows of Table 2.4), the results are very similar for different distributions. For the general spiked covariance model (bottom 3 rows of Table 2.4), the performance of BEMA increases/decreases when the data have lighter/heavier tails, but the difference is within a reasonable range. Our theory only requires a mild distributional assumption (Assumption 2.4.2), which is validated by this simulation.

The choice of α decides the fraction of bulk eigenvalues used to estimate (σ^2, θ) . The larger α , we

method	(λ, θ)	$\beta = 0.01$	$\beta = 0.05$	$\beta = 0.1$	$\beta = 0.2$	$\beta = 0.3$	$\beta = 0.5$
BEMA0	$(1.5, \infty)$	4.994 (99.4%)	5 (100%)	5 (100%)	5 (100%)	5 (100%)	5.006 (99.4%)
BEMA	$(1.5, \infty)$	4.712 (72.6%)	4.888 (89%)	4.93 (93%)	4.966 (96.6%)	4.982 (98.2%)	4.996 (99.6%)
BEMA	$(6, 3)$	4.734 (83.6%)	4.978 (97.2%)	5.018 (98.2%)	5.056 (94.4%)	5.082 (92%)	5.188 (83%)

Table 2.5: Simulation 6: The dependency of BEMA0 and BEMA upon the value of β . All the spiked eigenvalues are equal to λ , and the eigenvectors are randomly generated from the Stiefel manifold. BEMA0 and BEMA are implemented with $\beta \in \{0.01, 0.05, 0.1, 0.2, 0.3, 0.5\}$. The number in each cell is the average \hat{K} , and the number in brackets is the probability of correctly estimating K (correct rate).

restrict to a narrower range of eigenvalues deep into the bulk. The performance of BEMA is similar for $\alpha \in \{0.2, 0.3\}$ and slightly worse for $\alpha = 0.1$. In the asymptotic theory, α can be chosen as any constant, but for good finite-sample performance we need $(\tilde{p}\alpha - K)$ to be properly large, where $\tilde{p} = n \wedge p$. In practice, if \tilde{p} is extremely large, the choice of α has a negligible effect; if \tilde{p} is only moderately large, we recommend choosing a large α so that we are confident that $\tilde{p}\alpha$ is significantly larger than K .

SIMULATION 6. In this simulation, we briefly tested the dependency of our proposed methods upon the choice of β . Fix $(n, p, K) = (500, 100, 5)$. We generate data $\mathbf{X}_i \stackrel{iid}{\sim} N(0, \Sigma)$, $1 \leq i \leq n$, where Σ satisfies Model (2.7) with $\sigma^2 = 1$ and $\mu_k = \lambda$, for $1 \leq k \leq K$. The eigenvector matrix $\mathbf{\Xi}$ is drawn uniformly from the Stiefel manifold. The results are shown in Table 2.5.

The choice of β decides the probability of overestimating K asymptotically. Ideally, when the spike population eigenvalues are large enough, we can apply our methods with a very small β to tie down the probability of overestimation without concerning about underestimation. In the case where the spike population eigenvalues are only moderately large, we need to make a tradeoff between overestimation and underestimation. As we observe from Table 2.5, $\beta \in [0.05, 0.3]$ empirically gives a nice tradeoff for both BEMA and BEMA0. In practice, we recommend using $\beta = 0.1$ as a benchmark and one may adjust this value base on his tolerance towards overestimating/underestimating K .

2.6 REAL APPLICATIONS

We apply BEMA to two real datasets. We compare our method with EKC (Braeken & Van Assen, 2017), Bai&Ng (Bai & Ng, 2002), DDPA and its variants (Dobriban & Owen, 2019), and Pass&Yao (Passemier & Yao, 2014). DDPA has 3 versions: DPA is a deterministic implementation of parallel analysis (Horn, 1965); DDPA is an improvement of DPA aiming to resolve the issue of “eigenvalue shadowing,” that is, an extremely large spiked eigenvalue shadows the other spiked eigenvalues and causes an under-estimation of K ; DDPA+ is a robust version of DDPA recommended for real data analysis. We include all three versions in comparison.

2.6.1 THE LUNG CANCER DATA

The Lung Cancer dataset was collected and cleaned by Gordon et al. (2002). The original data set contains the expression data of 12,533 genes and 181 subjects. The subjects divide into two groups, the diseased group and the normal group. Jin & Wang (2016) processed this data set by removing genes that are not differentially expressed across subject groups and resulted in a new data matrix with $(p, n) = (251, 181)$. The selection of these 251 “influential genes” used no information of true groups, including the number of groups. We use this processed data matrix, because the original data matrix contains too many features (genes) that are irrelevant to the clustering structure, where no method gives meaningful results. It was argued in Jin & Wang (2016) that this data matrix follows a clustering model. As a result, the covariance matrix has $(K_0 - 1)$ spiked eigenvalues, where K_0 is the number of clusters. Here, the ground-truth is $K_0 = 2$, i.e., the true number of spiked eigenvalues is $K = 1$.

We apply BEMA with $(\alpha, \beta, M) = (0.2, 0.1, 500)$, i.e., 60% ($= 1 - 2\alpha$) of the bulk eigenvalues in the middle range are used to estimate model parameters, the probability of over-estimating K is controlled by 0.1, and 500 Monte Carlo samples are used to determine the ultimate threshold for eigenvalues. The BEMA algorithm outputs $(\hat{\theta}, \hat{\sigma}^2) = (0.288, 0.926)$. In Figure 2.6(a), we check

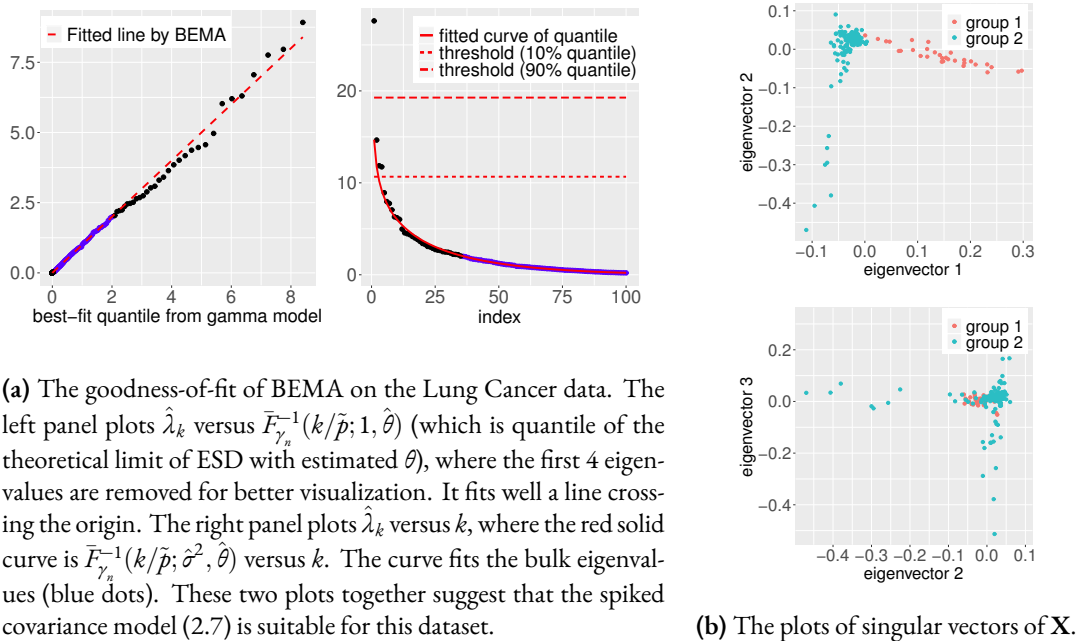
	BEMA	BEMA0	EKC	Bai&Ng	Pass&Yao	DDPA	DPA	DDPA+	truth
Lung Cancer Data	1	27	56	180	8	180	1	11	1
1000 Genomes Data	28	67	2503	4	28	85	20	4	25

Table 2.6: Comparison of different estimators of K using two real data sets: the lung cancer gene expression data and the 1000 Genome data of genome-wide common genetic variants. For BEMA and BEMA0, the choices of tuning parameters are described in the text. In the Appendix, we report the results with various choices of tuning parameters, which are very stable.

the goodness-of-fit. If the proposed spiked covariance model (2.7) is suited for the data, we expect to see $\hat{\lambda}_k \approx \hat{\sigma}^2 \cdot \bar{F}_{\gamma_n}^{-1}(k/\tilde{p}; 1, \hat{\theta})$, except for a few small k . The left panel of Figure 2.6(a) plots $\hat{\lambda}_k$ versus $\bar{F}_{\gamma_n}^{-1}(k/\tilde{p}; 1, \hat{\theta})$, suggesting a good fit to a line crossing the origin. The right panel contains the scree plot, i.e., $\hat{\lambda}_k$ versus k . We also plot the curve of $\bar{F}_{\gamma_n}^{-1}(k/\tilde{p}; \hat{\sigma}^2, \hat{\theta})$ versus k . This curve is a good fit to the scree plot in the middle range. These plots suggest that Model (2.7) is well-suited for this dataset.

The estimator of K by BEMA is $\hat{K} = 1$, which is exactly the same as the ground truth. This is the output of the algorithm by setting $\beta = 0.1$. Using the argument in Section 2.3.3, this is also a confidence lower bound for K . By setting $\beta = 0.9$ in the algorithm, we get a confidence upper bound which is 4. This gives an 80% confidence interval for K as $[1, 4]$. Figure 2.6(b) contains the scatter plots of the left singular vectors of X , colored by the true group label. The first singular vector clearly contains information for separating two groups, but other singular vectors also contain some information. This explains why the confidence upper bound is larger than 1.

The comparison with other methods is summarized in Table 2.6. The behavior of EKC is consistent with our observation in simulations. In this dataset, the eigenvalues of the residual covariance matrix vary widely (this can be seen from the estimated θ by BEMA, $\hat{\theta} = 0.288$, which is far from ∞), and EKC gives too small threshold to non-leading eigenvalues. The behavior of Bai&Ng is different from what we observe in simulations. Note that we have to use the effective p after the data processing by Jin & Wang (2016), where the dimension reduces from 12,533 to 251. As a result, the penalty in Bai&Ng is weaker than that in simulations, and so the method significantly over-estimates



(a) The goodness-of-fit of BEMA on the Lung Cancer data. The left panel plots $\hat{\lambda}_k$ versus $\bar{F}_{\gamma_n}^{-1}(k/\hat{p}; 1, \hat{\theta})$ (which is quantile of the theoretical limit of ESD with estimated θ), where the first 4 eigenvalues are removed for better visualization. It fits well a line crossing the origin. The right panel plots $\hat{\lambda}_k$ versus k , where the red solid curve is $\bar{F}_{\gamma_n}^{-1}(k/\hat{p}; \hat{\sigma}^2, \hat{\theta})$ versus k . The curve fits the bulk eigenvalues (blue dots). These two plots together suggest that the spiked covariance model (2.7) is suitable for this dataset.

(b) The plots of singular vectors of \mathbf{X} .

Figure 2.6: Results for the Lung Cancer data.

K . Pass&Yao also over-estimates K . Among DDPA and its variants, DPA performs the best. A possible reason is that DPA does not use empirical eigenvectors and is more stable than DDPA and DDPA+.

Different from all other methods, BEMA not only outputs an estimator of K but also yields a fitted model, $\text{Gamma}(\hat{\theta}, \hat{\theta}/\hat{\sigma}^2) = \text{Gamma}(0.288, 0.311)$, for eigenvalues of the residual covariance matrix. This can be useful for many other statistical inference tasks.

2.6.2 THE 1000 GENOMES DATA

The 1000 Genomes Phase 3 whole genome sequencing dataset (1000 Genomes Project Consortium, 2015) consists of the genotypes of 2504 subjects for over 84.4 million variants. We restrict the analysis to common variants with minor allele frequencies greater than 0.01. There are 26 self-reported ethnicity groups, coming from five super-populations: African (AFR), Ad Mixed American (AMR), East Asian (EAS), European (EUS), and South Asian (SAS).

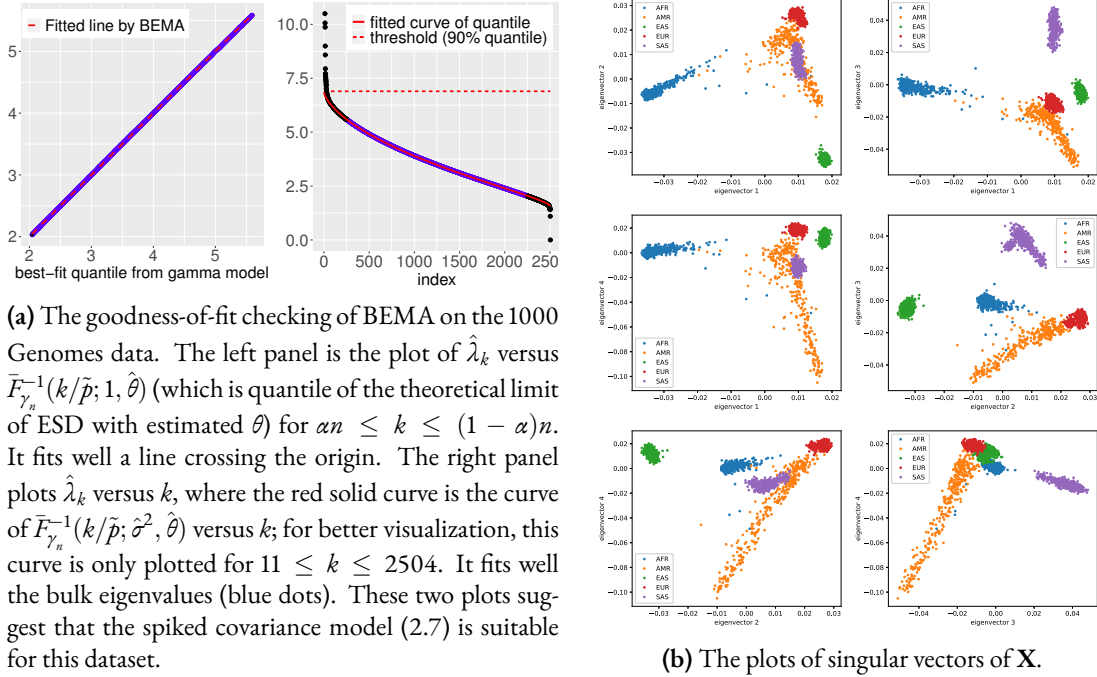


Figure 2.7: Results for analysis of the 1000 Genomes data .

In view of high linkage disequilibrium (LD) among some variants, which can distort the eigenvector and eigenvalue structure (Patterson et al., 2006), we first performed LD pruning. We used an independent pair-wise LD pruning, with window size 1000, step size 50 and a threshold 0.02 for R-squared. Restricting to LD pruned markers, we obtain a data matrix with $p = 24,248$ and $n = 2,504$. The number of spiked eigenvalues equals to the number of true ancestry groups minus one (Patterson et al., 2006). We treat the self-reported ethnicity groups as the ground truth, which gives $K = 25$.

We apply BEMA with $(\alpha, \beta, \mathcal{M}) = (0.1, 0.1, 500)$. First, we check the goodness-of-fit. BEMA outputs $(\hat{\theta}, \hat{\sigma}^2) = (4.256, 0.377)$. Figure 2.7(a) shows the Q-Q plot and the scree plot, with reference curves from the BEMA fitting. The meaning of these plots is the same as described in Section 2.6.1 and is also explained in the caption of this figure, which we do not repeat here. The conclusion is that

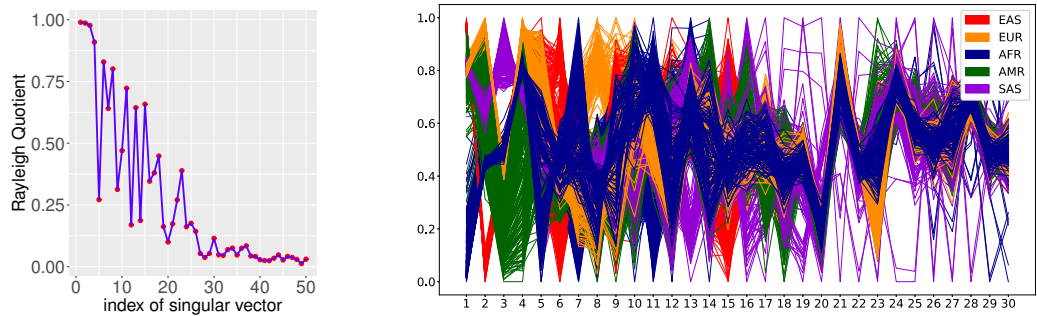
our proposed spiked covariance model (2.7) is an excellent fit to this dataset.

The estimated model for eigenvalues of the residual covariance matrix is $\text{Gamma}(\hat{\theta}, \hat{\theta}/\hat{\sigma}^2) = \text{Gamma}(4.256, 11.3)$. We note that the variance of the genotype on each SNP is $2q(1-q)$, where q is the null Minor Allele Frequency (MAF) of this SNP. We thus interpret the BEMA fitting as follows: After the ancestry effect is removed, the null MAFs q_j (on LD pruned SNPs) satisfy that $2q_j(1-q_j) \stackrel{iid}{\sim} \text{Gamma}(4.256, 11.3)$. The mean and standard deviation of this gamma distribution is 0.377 and 0.18, respectively.

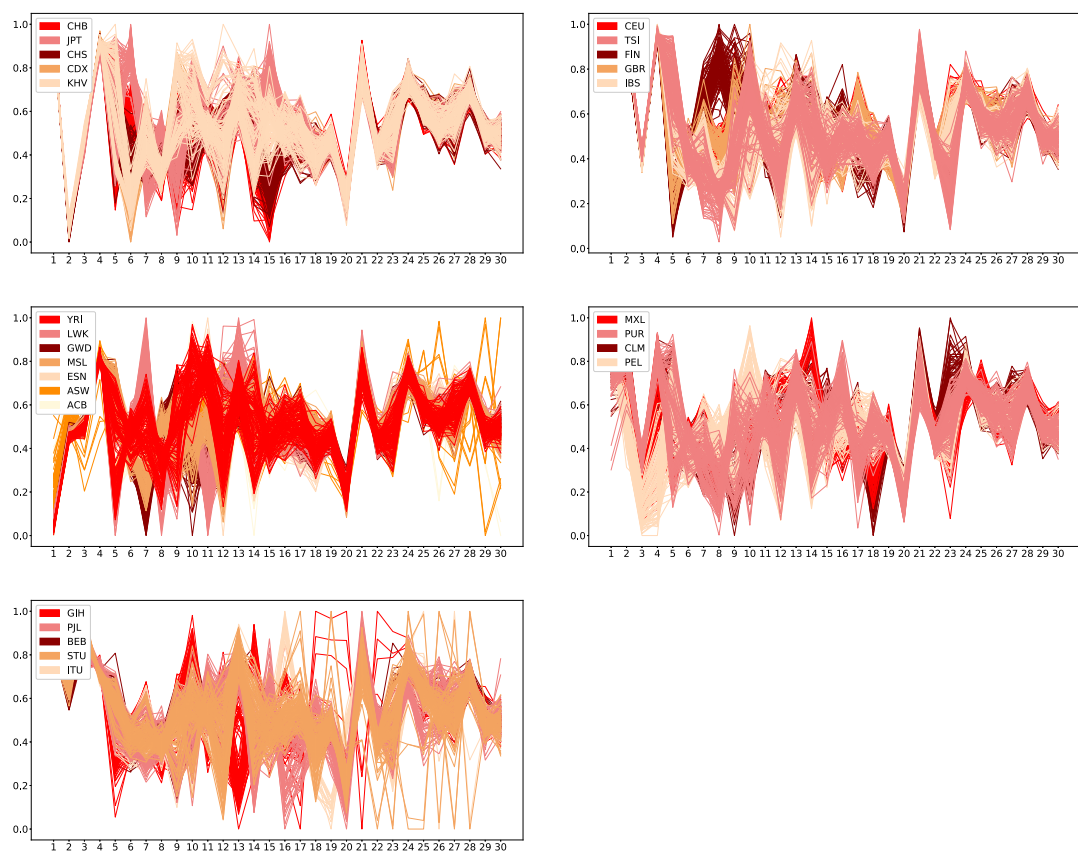
Next, we look at the estimation of K . The BEMA algorithm outputs $\hat{K} = 28$, which is very close to the ground truth $K = 25$. The 98% confidence interval of K is [27, 31].

A comparison with other methods is summarized in Table 2.6. EKC and DDPA significantly over-estimate K , and Bai&Ng and DDPA+ significantly under-estimate K . DPA gives $\hat{K} = 20$, which is relatively close to the ground truth. BEMA and Pass&Yao both give $\hat{K} = 28$, which is closest to the ground truth. Pass&Yao assumes that all σ_j^2 are equal. In this data set, BEMA estimates the standard deviation of σ_j^2 to be 0.18, which is relatively small. This explains why Pass&Yao also performs well.

Last, we validate the results by investigating the singular vectors of \mathbf{X} . We first measure the association between each singular vector and the true ethnicity labels by the Rayleigh quotient (Horn & Johnson, 2012). Let $\hat{\eta}_k \in \mathbb{R}^n$ be the k th left singular vector of the centralized data matrix. We treat its entries as n data points and compute the ratio of between-cluster-variance and within-cluster-variance, denoted as RQ_k . A larger RQ_k indicates that $\hat{\eta}_k$ is more correlated with the true ethnicity labels. Figure 2.8(a) plots RQ_k versus k . The first a few singular vectors have very high association with the ethnicity labels. These singular vectors capture the super population structure. The pairwise scatter plots of the first 4 singular vectors are contained in Figure 2.7(b), which show clearly that super populations are well separated on these singular vectors. Besides the first few singular vectors, the remaining singular vectors capture more of the sub-structure within each super population. Figure 2.8(b) is the parallel coordinate plot. In Figure 2.8(c), we re-generate parallel coordinate plots by



(a) The association between singular vectors of X and the true ethnicity labels. (b) The parallel coordinate plot of singular vectors, color-coded by five super-populations.



(c) The parallel coordinate plots of singular vectors for each super-population, color-coded by the ethnicity groups within each super-population. The five super-populations are EAS (top left), EUR (top right), AFR (middle left), AMR (middle right), and SAS (bottom left). The sub-population labels used in the legends of can be found in [1000 Genomes Project Consortium \(2015\)](#).

Figure 2.8: Interpretation of results for the 1000 Genomes data.

restricting to each super population. Within the super population AMR, there is still separation of ethnicity groups for k as large as 27. This explains why BEMA outputs a \hat{K} that is slightly larger than the ground truth.

2.7 DISCUSSION

We propose a new method for estimating the number of spiked eigenvalues in a large covariance matrix. The novelty of our method lies in a systematic approach to incorporating bulk eigenvalues in the estimation of K . Under a working model which assumes the diagonal entries of the residual covariance matrix are *iid* drawn from a Gamma distribution, we fit a parametric curve on bulk eigenvalues. The estimated parameters of this curve are then used to decide a threshold for top eigenvalues and produce an estimator of K . We study the theoretical properties of our method under a standard spiked covariance model, and show that our estimator requires weaker conditions for consistent estimation of K compared with the existing methods. We examine the performance of our method using both simulated data and two real data sets. Our empirical results show that the proposed method outperforms other competitors in a variety of scenarios.

Our approach is conceptually connected to the empirical null (Efron, 2004) in multiple testing. The empirical null imposes a working model (e.g., a normal distribution) on Z -scores of individual null hypotheses and estimates the parameters of this distribution from a large number of Z -scores. The fitted null model is then used to correct p -values and help identify the non-null hypotheses. Similarly, we impose a working model (i.e., a Gamma distribution) on non-spiked population eigenvalues and estimate the parameters of this distribution from a large number of bulk empirical eigenvalues. The fitted null model is then used to assist estimation of K . From this perspective, our method can be regarded as a *conceptual* application of the empirical null approach to eigenvalues. Meanwhile, our setting is much more complicated than that in multiple testing. The bulk eigenvalues are highly

correlated, and their marginal distribution has no explicit form. These impose great challenges on algorithm design and theoretical analysis.

For the theoretical study, we first analyze the special case of $\theta = \infty$. This corresponds to the well-known standard spiked covariance model (Johnstone, 2001), which has attracted many theoretical interests. Our theory contributes to this literature with an explicit error bound on estimating σ^2 and consistency theory on estimating K . The theoretical study for a general θ that corresponds to the setting of heterogeneous residual variances is of great interest but is technically challenging. Instead, we study a proxy model where the population eigenvalues are *iid* drawn from a truncated Gamma distribution. Under this model we derive error bounds for $(\hat{\sigma}^2, \hat{\theta})$ and prove the consistency of \hat{K} with mild conditions. The analysis uses advanced results in random matrix theory (Bloemendal et al., 2016, Knowles & Yin, 2017, Ding, 2020).

The method can be extended in multiple directions. Here we assume that the diagonal entries of the residual covariance matrix are from a Gamma distribution. It can be generalized to other parametric distributions. In Section 2.4.2, we have already seen a variant of our method by using a truncated Gamma distribution, which assumption helps eliminate extremely large variances for the residuals. We can also use a mixture of Gamma distributions to accommodate heterogeneous feature groups. Our main algorithm can be easily adapted to such cases. When the distribution family is unknown, we may combine our method with the techniques in nonparametric density estimation. The thresholding scheme in our method can also be modified. We currently apply a single threshold to all eigenvalues. Alternatively, we may use different thresholds for different eigenvalues. One proposal is to use the $(1 - \beta)$ -quantile of the distribution of $\hat{\lambda}_k^*$ in the null model (2.12) as a threshold for $\hat{\lambda}_k$. We leave these extensions to future work.

In the numerical experiments, our method exhibits robustness to model misspecification. It is suggested by Simulation 4 of Section 2.5 that our method continues to work when the residual covariance matrix is a Toeplitz matrix, or a block-wise diagonal matrix, or a sparse matrix. A theoretical

understanding to this phenomenon will be useful. As stated in Section 2.5, we have observed empirically that there always exist (σ^2, θ) such that the theoretical limit of ESD induced by the Gamma model (2.2) can accurately approximate the theoretical limit of ESD induced by a Toeplitz or block-wise diagonal or sparse covariance matrix. It remains an interesting question on how to justify it theoretically. We leave it to future work.

3

On Posterior Consistency of Bayesian Factor Models in High Dimensions

CONTRIBUTION This chapter is based on a paper [Ma & Liu \(2020\)](#) jointly with Prof. Jun S. Liu.

3.1 INTRODUCTION

Factor models have been widely adopted in social science, economics, bioinformatics, and many other fields that need interpretable dimension reduction for their data. They serve as a formal way to encode high-dimensional observations as a linear combination of a few latent factors plus idiosyncratic errors, which accommodate some intuitive interpretations and can sometimes be further validated by additional knowledge. In this chapter, we consider the following standard parametric formulation: each G -dimensional vector observation \mathbf{y}_i (e.g., daily returns of ~ 3000 U.S. stocks) is assumed to be linearly related to a K -dimensional vector of latent factors ω_i (e.g., 20 market factors) through a skinny tall factor loading matrix \mathbf{B} :

$$\mathbf{y}_i \mid \omega_i, \mathbf{B}, \boldsymbol{\Sigma} \stackrel{i.i.d.}{\sim} \mathcal{N}_G(\mathbf{B}\omega_i, \boldsymbol{\Sigma}), \quad i = 1, \dots, n, \quad (3.1)$$

and the idiosyncratic variance matrix $\boldsymbol{\Sigma}$ is assumed to be diagonal as in the literature. In matrix form, we denote the observations as $\mathbf{Y} = (\mathbf{y}_1, \dots, \mathbf{y}_n)$, which is a $G \times n$ matrix, and the factors as a $K \times n$ matrix $\boldsymbol{\Omega} = (\omega_1, \dots, \omega_n)$. The factors are usually assumed to be independently and normally distributed: $\omega_i \sim \mathcal{N}_K(0, \mathbf{I}_K)$.

People are often interested in estimating the $G \times K$ loading matrix \mathbf{B} in order to gain insight on the correlation structure of the observations. Marginalizing out ω_i , we obtain that $[\mathbf{y}_i \mid \mathbf{B}, \boldsymbol{\Sigma}] \sim \mathcal{N}_G(0, \mathbf{B}\mathbf{B}^T + \boldsymbol{\Sigma})$, implying that the loading matrix \mathbf{B} is only identifiable up to a right orthogonal transformation (rotationally invariant). It is thus rather difficult to pinpoint the factor loading matrix consistently, to determine the dimensionality of the latent factors, or to design efficient algorithms to conduct a proper full Bayesian analysis of the model.

In recent years, researchers begin to investigate the effects of sparsity assumptions on factor loadings, since a sparse loading matrix has a better interpretability and is easier to be identified. Consid-

erable progresses have been made in the realm of sparse Bayesian factor analysis, such as [Fruehwirth-Schnatter & Lopes \(2018\)](#) and [Ročková & George \(2016\)](#), which are two representatives of the approaches using hierarchical continuous or discrete spike-and-slab (SpSL) priors (i.e., a mixture of a concentrated distribution, which can be either continuous with a small variance or a point mass, and a diffuse distribution) to represent the sparsity of the factor loading matrix. The identifiability issues of sparse factor models are formally discussed in [Fruehwirth-Schnatter & Lopes \(2018\)](#), who also designed an efficient Markov chain Monte Carlo (MCMC) procedure to simulate from the posterior distribution of an over-parameterized sparse factor model under the discrete SpSL prior. [Ročková & George \(2016\)](#) proposed a sparse Bayesian factor analysis framework assuming independent (conditioned on the feature allocation) continuous SpSL priors on loading matrix’s elements, under which a fast posterior mode-detecting strategy is proposed.

Our work originates from a peculiar phenomena we observed when implementing a full Bayesian inference procedure for the factor model in (3.1) under the SpSL prior from [Ročková & George \(2016\)](#). Although the simulation studies of [Ročková & George \(2016\)](#) show a good consistency (up to trivial rotations) of the *maximum a posteriori* (MAP) estimation of the loading matrix in various large G and large n scenarios, we found that the corresponding Wald type consistency for the posterior distribution requires n diverging at a faster rate than s besides other numerical conditions on the true loading matrix that are generally required for justifying the posterior contraction ([Pati et al., 2014](#)). Here s is the average number of nonzero elements of each column of the loading matrix \mathbf{B} and is usually much smaller than G .

When $s \geq n$ but is still much smaller than G , we observed from simulations a ‘magnitude inflation’ phenomenon. That is, posterior samples of the loading matrix are inflated in the matrix norm compared to the data-generating loading matrix, and the extent of inflation is affected by the variance of the slab part of the SpSL prior —the more diffuse the slab prior we use the more inflation we observe. This $s \geq n$ setting is not unusual in practice. For example, the gene expression dataset analyzed

in Section 3.7 contains measures of mRNA expression levels of $G = 8932$ genes in 10 mice in four age periods ($n = 40$). Each factor may correspond to a pathway and s would be the average number of genes in each pathway, which can be much larger than n .

The reason for this inflation phenomena is not immediately obvious since the total number of observed quantities is $n \times G$, corresponding to n observed G -dimensional vectors \mathbf{y}_i , $i = 1, \dots, n$, which is often much larger than $s \times K$, the number of nonzero elements in the loading matrix. Consider a special case with $K = 1$, $G = s$, and $\Sigma = \mathbf{I}_G$ is known. Then ω_i for $i = 1, \dots, n$ is a scalar, and $\mathbf{B} = (b_1, \dots, b_G)^T$ is a G -dimensional vector. Thus, each component y_{ij} of \mathbf{y}_i can be written as

$$y_{ij} = \omega_i b_j + \varepsilon_{ij}, \quad \varepsilon_{ij} \stackrel{i.i.d.}{\sim} \mathcal{N}(0, 1).$$

Although the total number of unknown parameters in the model is $G + n$, the number of independent scalar observations y_{ij} is $n \times G$, much larger than $G + n$. The model is unidentifiable because $\omega_i \times b_j = (\omega_i/c) \times (b_j c)$ for any $c \neq 0$. Requiring that the $\omega_i \stackrel{i.i.d.}{\sim} \mathcal{N}(0, 1)$, $i = 1, \dots, n$, can indeed alleviate the identifiability issue, but is not enough to “tie down” the b_j ’s in the posterior distribution if there are too many of them, which manifests itself in the inflation phenomena. But how many is “too many”? In this simple example, there are “too many” if $G \geq n$ (Section 3.4). Our later theoretical analysis shows that, if s , the column average number of nonzero elements of \mathbf{B} , is no smaller than n , the inflation will provably happen, although we observed empirically that the inflation occurs when $s \sim n$. An apparent remedy revealed from the above intuition and our later analysis is to further restrict the ω_i ’s, such as requiring that $\sum_{i=1}^n \omega_i^2 = n$.

More generally speaking, due to a nearly non-identifiable structure of model (3.1), an overdose of independent diffuse priors on loading matrix elements dilutes the signal from the data. Problems with the use of diffuse priors in Bayesian inference when observation sample sizes are small relative to the number of parameters being estimated have been studied in the literature (Efron, 1973, Kass &

Wasserman, 1996, Natarajan & McCulloch, 1998). This problem for Bayesian factor analysis was also noted in Ghosh & Dunson (2009) and a practical solution was proposed without further theoretical investigations.

The Ghosh-Dunson model allows each factor to have an unknown variance that follows an inverse Gamma prior and imposes the standard Gaussian prior on the loading matrix’s elements. If one reallocates the variance of factors to the loading matrix side, this model is equivalent to reformatting the loading matrix as $\mathbf{B} = \mathbf{Q} \times \mathbf{D}$ with \mathbf{D} being a diagonal matrix, and assuming that *a priori* elements in \mathbf{Q} are *i.i.d.* standard Gaussian and diagonal elements of \mathbf{D} follow an inverse-Gamma distribution. When assigning non-informative priors to diagonal elements of \mathbf{D} , elements of \mathbf{B} can also marginally have non-informative priors. Consequently, this hierarchical prior construction resolves the magnitude inflation problem by reducing the number of diffuse parameters, which is achieved by imposing a dependency between the magnitudes of \mathbf{B} ’s elements within the same column through \mathbf{D} .

Some later work (e.g. Bhattacharya & Dunson (2011) and Legramanti et al. (2020)) all follows this loading matrix decomposition idea to induce dependencies among the magnitudes. However, informative priors are usually applied to \mathbf{D} . In the fixed p and $n \rightarrow \infty$ scenario, they develop posterior consistency results. But in the “Large s , Small n ” scenario, these informative priors on \mathbf{D} can be influential for the magnitude of the loading matrix sampled from its posterior distribution as we verified in simulations.

In this chapter, we study asymptotic behaviors of the posterior distributions when an independent SpSL prior is employed for elements of the loading matrix and a right-rotational invariant distribution is assumed on the factor matrix $\mathbf{\Omega}$ (i.e., $\mathbf{\Omega}$ and $\mathbf{\Omega}\mathbf{R}$ follows the same distribution for all $n \times n$ orthogonal matrix \mathbf{R} ; this is different from the left-rotational invariance that makes \mathbf{B} nonidentifiable), to thoroughly understand the inflation phenomena. All consistency and convergence concepts in our work are in the frequentist (repeated-sampling) sense. Take the loading matrix for example. If for any open neighborhood \mathcal{N} of an entry of the true loading matrix (the magnitude of entries are at

the constant order), the probability for a random draw from the posterior distribution of that entry to fall in \mathcal{N} , as a function of the data in the repeated sampling sense, converges to 1 almost surely as n and G go to infinity, we say that the posterior inference of the loading matrix is consistent, or simply that “*the posterior sample of the loading matrix converges to the truth.*”

We theoretically show that the observed inflation phenomena of the posterior distribution is due to the fact that the control of $\mathbf{\Omega}\mathbf{\Omega}^T/n$, or more specifically, the singular values of $\mathbf{\Omega}\mathbf{\Omega}^T/n$, is too weak under only the normality assumption on the factor matrix $\mathbf{\Omega}$. Our analysis also suggests that employing a stronger control over $\mathbf{\Omega}\mathbf{\Omega}^T/n$ can result in consistent posterior distribution for the loading matrix in Ročková & George (2016)’s framework under high dimensions. Consequently, we consider the \sqrt{n} -orthonormal factor model: let $\mathbf{\Omega}/\sqrt{n}$ be uniform on the Stiefel manifold $St(K, n)$, which is the set of all orthonormal K -frames in \mathbb{R}^n , or, equivalently, the first K rows of a $n \times n$ Haar-distributed random orthogonal matrix (there exists a unique right and left invariant Haar measure on the set of orthogonal matrices, see Meckes (2014)).

From the modelling perspective, whenever the data is generated from the normal factor model (3.1) where $\mathbf{Y} = \mathbf{B}\mathbf{\Omega} + \Delta$, it can also be viewed as generated by an \sqrt{n} -orthonormal factor model $\mathbf{Y} = (\mathbf{B}\mathbf{K}(\mathbf{\Omega})/\sqrt{n}) \times (\sqrt{n} \cdot \mathbf{V}(\mathbf{\Omega})) + \Delta$ with loading matrix being $(\mathbf{B}\mathbf{K}(\mathbf{\Omega})/\sqrt{n})$. Here $\mathbf{K}(\mathbf{\Omega})$ and $\mathbf{V}(\mathbf{\Omega})$ are from the LQ decomposition $\mathbf{\Omega} = \mathbf{K}(\mathbf{\Omega})\mathbf{V}(\mathbf{\Omega})$. The new loading matrix $(\mathbf{B}\mathbf{K}(\mathbf{\Omega})/\sqrt{n})$ inherits the same generalized lower triangular structure (Fruehwirth-Schnatter & Lopes, 2018) from \mathbf{B} (if it possesses any) and they are identical in the asymptotic sense as $n \rightarrow \infty$. Beside having the same model interpretability, we reveal in our work that the \sqrt{n} -orthonormal factor model enjoys two major advantages:

- (A) The posterior distribution is more robust against the choice of the prior distribution for elements of the loading matrix in the “Large s , Small n ” scenario. The posterior consistency can hold for a broader set of prior choices including the one from Ročková & George (2016).

(B) Gibbs samplers for the normal factor model can be easily adapted to handle \sqrt{n} -orthonormal factors by only modifying the conditional sampling step for $\mathbf{\Omega}$. This modification requires negligible computational cost, but leads to a significant efficiency gain in MCMC sampling.

For these reasons, in the high-dimensional “Large s , Small n ” scenario, we propose to use the \sqrt{n} -orthonormal factor model in place of the normal factor model when doing full Bayesian inference on the population covariance matrix. Our proposed Gibbs sampler provides encouraging results in both simulations and a real data example.

This chapter is structured as follows. Section 3.2 introduces the Bayesian factor model of [Ročková & George \(2016\)](#) and a corresponding basic Gibbs sampler. Under their framework, Section 3.3 illustrates by a synthetic example the ‘magnitude inflation’ phenomenon of the posterior samples of the loading matrix and its dependence upon the slab prior. Section 3.4 provides theoretical explanations for the phenomenon. Section 3.5 reveals the connection between the phenomenon and the factor modeling assumption, and proposes the \sqrt{n} -orthonormal factor model whose posterior consistency can be guaranteed. By revisiting the synthetic example, Section 3.6 numerically verifies the consistency and robustness against prior, and provides a comparison between our method and alternative approaches from [Ghosh & Dunson \(2009\)](#) and [Bhattacharya & Dunson \(2011\)](#). Section 3.7 presents a real-data application. Section 3.8 concludes with a short discussion.

3.2 BAYESIAN SPARSE FACTOR MODEL AND INFERENCE

3.2.1 PRIOR SETTINGS FOR LOADING COEFFICIENT SELECTION

In order to enhance model identifiability and interpretability, one often imposes a sparsity assumption for the loading matrix. Traditional approaches considered post-hoc rotations as well as regularization methods, see, e.g. [Kaiser \(1958\)](#) and [Carvalho et al. \(2008\)](#). By integrating these two paradigms, [Ročková & George \(2016\)](#) proposed a sparse Bayesian factor model framework along with a fast mode-

identifying PXL-EM algorithm. In their framework, the sparsity assumption on factor loading matrix is encoded through a hierarchical SpSL prior, and we mostly follow their framework in this chapter.

Let β_{jk} denote the $(j, k)^{th}$ element of the loading matrix \mathbf{B} . We assume that *a priori* the β_{jk} 's follow a SpSL prior and are mutually independent given the hyper-parameters. We introduce for each element a binary indicator γ_{jk} such that

$$p(\beta_{jk} | \gamma_{jk}, \lambda_0, \lambda_1) = (1 - \gamma_{jk})\psi(\beta_{jk} | \lambda_0) + \gamma_{jk}\psi(\beta_{jk} | \lambda_1), \quad \lambda_0 \gg \lambda_1 \quad (3.2)$$

where $\psi(\beta | \lambda) = \frac{\lambda}{2} \exp(-\lambda|\beta|)$ is a Laplace distribution, and

$$\gamma_{jk} | \theta_k \stackrel{ind}{\sim} \text{Bernoulli}(\theta_k) \quad \text{and} \quad \theta_k = \prod_{l=1}^k \nu_l, \quad \nu_l \stackrel{i.i.d.}{\sim} \text{Beta}(\alpha, 1). \quad (3.3)$$

We note that θ_k decreases with respect to k . We call $\Theta = (\theta_1, \dots, \theta_K)$ the *feature sparsity vector* and $\Gamma = (\gamma_{jk})_{G \times K}$ the *feature allocation matrix*. The idiosyncratic variance matrix Σ is assumed to be diagonal with elements σ_j^2 endowed with a conjugate prior: $\sigma_1^2, \dots, \sigma_G^2 \stackrel{i.i.d.}{\sim} \text{Inverse-Gamma}(\eta/2, \eta\varepsilon/2)$. When $K = \infty$, the foregoing setup leads to an infinite factor model, for which some weak consistency results of the posterior distribution are established for the fixed- p scenario (Ročková & George, 2016). In simulations, they adopted a truncated approximation to the infinite factor model by setting K to a pre-specified value larger than the true K in data generation. Throughout the chapter, we assume that K is a pre-specified finite value.

Ročková & George (2016) showed in simulations that the PXL-EM converges dramatically faster than the EM algorithm in finding the *maximum a posteriori* (MAP) estimator (i.e., $\hat{\mathbf{B}}, \hat{\Sigma}, \hat{\Theta}$ that maximizes $\pi(\mathbf{B}, \Sigma, \Theta | \mathbf{Y})$) and also demonstrated the consistency of MAP estimator in estimating the loading matrix under the ‘‘Large s, Small n’’ setting. However, converting their method into a full Bayesian inference procedure turns out to be more subtle and challenging.

3.2.2 A STANDARD GIBBS SAMPLING PROCEDURE

The full posterior distribution of the parameters, $(\mathbf{B}, \mathbf{\Omega}, \mathbf{\Sigma}, \mathbf{\Gamma}, \mathbf{\Theta})$, in a Bayes factor model can be written generically as

$$\pi(\mathbf{B}, \mathbf{\Omega}, \mathbf{\Sigma}, \mathbf{\Gamma}, \mathbf{\Theta} \mid \mathbf{Y}) \propto f(\mathbf{Y} \mid \mathbf{B}, \mathbf{\Omega}, \mathbf{\Sigma}) f(\mathbf{\Omega}) p(\mathbf{B} \mid \mathbf{\Gamma}) p(\mathbf{\Gamma} \mid \mathbf{\Theta}) p(\mathbf{\Theta}) p(\mathbf{\Sigma}), \quad (3.4)$$

where f denotes the likelihood, p denotes prior, $\mathbf{\Omega}$ denotes the $K \times n$ matrix with columns given by ω_i , $\mathbf{\Gamma}$ denotes the $G \times K$ matrix with entries given by γ_{jk} and $\mathbf{\Theta}$ denotes the K -dimensional feature sparsity vector formed by the θ_k 's. Here observation \mathbf{Y} represents a $G \times n$ matrix with columns \mathbf{y}_i .

A standard Gibbs sampler (Gelfand & Smith, 1990, Liu, 2008, Tanner & Wong, 1987) for sampling from the full posterior distribution (3.4) iteratively update each component according to the following conditional distributions:

- Update \mathbf{B} iteratively as

$$\pi(\beta_{jk} \mid \beta_{-jk}, \mathbf{\Omega}, \mathbf{\Gamma}, \mathbf{\Sigma}) \propto \exp(-a_{jk}\beta_{jk}^2 + b_{jk}\beta_{jk} - c_{jk}|\beta_{jk}|), \text{ all } j, k;$$

where $a_{jk} = \sum_{i=1}^n \omega_{ik}^2 / 2\sigma_j^2$, $b_{jk} = \sum_{i=1}^n \omega_{ik} (y_{ij} - \sum_{l \neq k} \beta_{jl} \omega_{il}) / \sigma_j^2$, $c_{jk} = \lambda_1 \gamma_{jk} + \lambda_0 (1 - \gamma_{jk})$.

This conditional density can be written as a mixture of two truncated normal density, and thus can be sampled efficiently.

- Update $\mathbf{\Omega}$ component by component independently:

$$\omega_i \mid \mathbf{B}, \mathbf{\Sigma} \sim \mathcal{N}_K((\mathbf{I}_K + \mathbf{B}^T \mathbf{\Sigma}^{-1} \mathbf{B})^{-1} \mathbf{B}^T \mathbf{\Sigma}^{-1} \mathbf{y}_i, (\mathbf{I}_K + \mathbf{B}^T \mathbf{\Sigma}^{-1} \mathbf{B})^{-1}), \quad i = 1, \dots, n.$$

- Update Γ component by component independently:

$$\gamma_{jk} \mid \mathbf{B}, \Theta \sim \text{Bern} \left(\frac{\lambda_1 \exp(-\lambda_1 |\beta_{jk}|) \theta_k}{\lambda_0 \exp(-\lambda_0 |\beta_{jk}|) (1 - \theta_k) + \lambda_1 \exp(-\lambda_1 |\beta_{jk}|) \theta_k} \right),$$

for $j = 1, \dots, G; k = 1, \dots, K$.

- Update Θ iteratively:

$$\theta_k \mid \Gamma, \theta_{-k} \sim \text{Trunc-Beta}(\theta_{k+1}, \theta_{k-1}; \tilde{\alpha}_k, \tilde{\beta}_k)$$

where $\theta_0 = 1, \theta_{K+1} = 0$ and

$$\begin{aligned} \tilde{\alpha}_k &= \begin{cases} \#\{\gamma_{jk} = 1, j = 1, \dots, G\}, & k < K \\ \#\{\gamma_{jk} = 1, j = 1, \dots, G\} + \alpha, & k = K \end{cases}, \\ \tilde{\beta}_k &= \#\{\gamma_{jk} = 0, j = 1, \dots, G\} + 1. \end{aligned}$$

Here $\text{Trunc-Beta}(a, b; \alpha, \beta)$ is the density proportional to $f_{\text{Beta}}(x; \alpha, \beta) I_{\{x \in [a, b]\}}$.

- Update Σ along its diagonal:

$$\sigma_j^2 \mid \mathbf{B}, \Omega \sim \text{Inverse-Gamma} \left(\frac{1}{2}(\eta + n), \frac{1}{2}(\eta\varepsilon + \sum_{i=1}^n (y_{ij} - \mathbf{B}_j^T \omega_i)^2) \right)$$

where \mathbf{B}_j^T represents the j -th row vector of \mathbf{B} .

Due to multimodality of the posterior distribution caused by the invariance of the likelihood function under matrix rotations (therefore only the sparsity prior can provide information to differentiate different modes) and the strong ties between the factor loading and common factors (thus making gaps among different modes very deep), the performance of this basic Gibbs sampler is very

sticky and can only explore the neighborhood of the initial values. By initializing the sampler at some estimated mode such as the MAP estimator from the PXL-EM algorithm, however, this sampler appears to be a reasonable tool for exploring the local posterior behavior around the MAP. Indeed, more dramatic global MCMC transition moves are needed in order to have a fully functional MCMC sampler (see Appendix C.1).

3.3 THE MAGNITUDE INFLATION PHENOMENON

3.3.1 A SYNTHETIC EXAMPLE

To illustrate the magnitude inflation phenomenon in high dimensional sparse factor models, we generate a dataset from model (3.1) similar to that of Ročková & George (2016), which consists of $n = 100$ observations, $G = 1956$ responses, and $K = 5$ factors drawn from $\mathcal{N}(0, \mathbf{I}_5)$. The true loading matrix is a block diagonal matrix as shown in the leftmost sub-figure of Figure 3.1, where black entries correspond to 1 and blank entries correspond to 0 (thus $s = 500 > n$). Σ_{true} is selected to be the identity matrix. With the synthetic dataset, we use the basic Gibbs sampler from section 2.2 with $\alpha = 1/G$, $\eta = \varepsilon = 1$, $\lambda_0 = 20$, $\lambda_1 \in \{0.001, 0.1\}$ and $K = 8$, to explore the posterior distribution.

Ten snapshots of heat-maps of $|\mathbf{B}|$ in a Gibbs trajectory of 100 iterations initialized at the true value is displayed in Figure 3.1, from which we can conclude that the direction of each column vector in the loading matrix is well preserved during Gibbs iterations, whereas the absolute value of every non-zero element increases over the iteration time and eventually stabilizes around a much larger value than the true one (about 4000 in our test setting with $\lambda_1 = 0.001$). As a demonstration of the inflation, Figure 3.2(a) displays the trace plot of $\log(|\beta_{1,1}|)$ with $\lambda = 0.001$ and 0.1, respectively, which also indicates the slow convergence of the basic Gibbs sampler using a small λ_1 . The degree of inflation is influenced by the ratio of the number of observations n over the average number of nonzero elements of each column in the true factor loading matrix, s , as well as the choice of independent slab priors. For

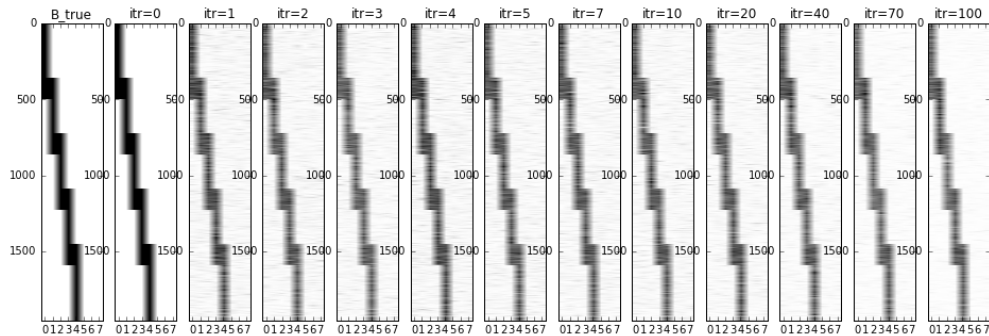


Figure 3.1: Heat-maps of $|\mathbf{B}|$ in 100 iterations from the basic Gibbs sampler. The black entries correspond to 1 and blank entries correspond to 0. The directions of the columns of the loading matrix are well preserved throughout the Gibbs iterations.

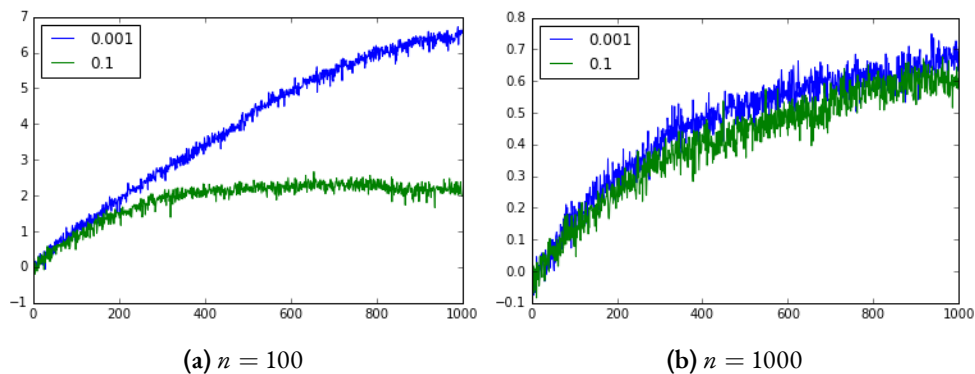


Figure 3.2: Trace plot of $\log(|\beta_{1,1}|)$ from Gibbs sampler with $n = 100, 1000$, and $\lambda_1 = 0.001, 0.1$. The sampler of $\beta_{1,1}$ stabilizes around a much larger value than the truth, 1. The inflation of samples is more severe when n is smaller or the variance of slab priors is larger.

example, when n is increased from 100 to 1000, the posterior samples of the loading matrix stabilize around somewhere much closer to the true loading matrix.

By adding some scaling group moves (Liu & Wu, 1999, Liu & Sabatti, 2000) to the basic Gibbs sampler (details can be found in Appendix C.1), which takes negligible computing time, we can greatly improve the convergence rate of the sampler, as demonstrated by contrasting Figure 3.2 with Figure 3.3, of which the latter shows the trace plot for $\log(|\beta_{1,1}|)$ of the modified Gibbs sampler under various slab priors, for the case with $n = 100$. Figure 3.3 shows that as λ_1 decreases from 0.5 to 0.001

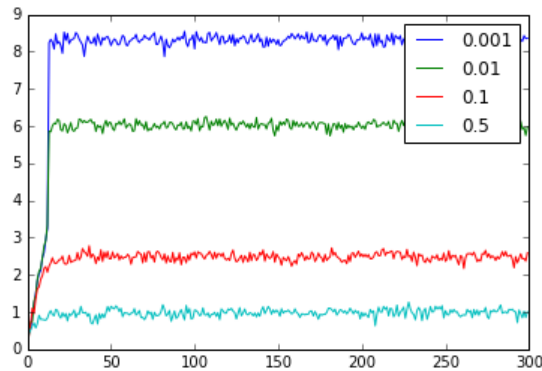


Figure 3.3: Trace plot of $\log(|\beta_{1,1}|)$ from the modified Gibbs sampler with $\lambda_1=0.001, 0.01, 0.1, 0.5$ for the case with $n = 100$. The modified Gibbs sampler has a much shorter burn-in process.

so that the slab part becomes more and more diffused, the posterior mean of $|\beta_{1,1}|$ increases from around 2.5 to around 4000. Heat-maps of the factor loading are similar to Figure 3.1 in all cases with $\lambda_1 \in \{0.001, 0.01, 0.1, 0.5\}$, which means that the direction of each column vector in the loading matrix remains roughly the same throughout Gibbs iterations.

3.3.2 MAGNITUDE INFLATION AND DIRECTION CONSISTENCY

Our numerical results revealed some perplexing consequences of using independent SpSL priors for a Bayesian factor model when $s \geq n$, which can be summarized as “magnitude inflation” and “direction consistency”. While the former means that the posterior draws of the loading matrix are inflated entry-wise compared with the true loading matrix with the inflation magnitude dependent on how diffuse the slab prior is, the latter says that the direction of columns of posterior samples of the loading matrix somehow still converges to the true direction as $n, s \rightarrow \infty$. Intuitively, when the number of independent slab priors employed grows at a faster rate than the number of observations, these priors will overwhelm the signal from data. The interesting observation is that the overdose of independent slab priors only dilutes the signal for the magnitude part in the loading matrix but has little impact on the identification of the column space. It is also worth mentioning, regardless of the occurrence of

“magnitude inflation”, the posterior distribution of the idiosyncratic variance matrix Σ still has a nice concentration around the truth.

The inflation problem is quite a concern in practice when people try to use these posterior samples of the loading matrix for estimating the observation covariance structure. The low rank part ($\mathbf{B}\mathbf{B}^T$) in the estimated covariance matrix is usually exaggerated to some extent depending on the selected slab prior. Traditional literature tends to ignore the inflation problem by treating it as a consequence of the lack of enough observations (i.e., n is too small compared to s) to guarantee posterior sample consistency. But this argument is inaccurate as we will show in next sections. Furthermore, we notice that, with the same amount of observations, the MAP estimator is rather precise in estimating the true loading matrix and directions of columns of the loading matrix are well captured by the posterior samples, provided that the structure of the true feature allocation matrix is known, as in the synthetic example. This suggests that the data provide sufficient information for recovering the true loading with the aid of knowing true feature allocation matrix. Thus, the magnitude inflation phenomena may be caused by some modeling issues. In the next two sections, we will provide some theoretical verifications for the magnitude inflation as well as a simple and provable remedy.

3.4 POSTERIOR DEPENDENCE ON THE SLAB PRIOR

It is generally recognized that in a Bayesian factor model using an improper flat prior on elements of the loading matrix can be dangerous, and will lead to an improper posterior distribution when $G \geq n$. This is in fact not very intuitive, so we illustrate this point with a very simple example with $K = 1$ factor, $n = 2$ observations, and independent noises. Let the two vector observations be \mathbf{y}_1 and \mathbf{y}_2 , each of G -dimensional. We can therefore write $\mathbf{y}_1 = \mathbf{v}_1 + \varepsilon_1$, and $\mathbf{y}_2 = \mathbf{v}_2 + \varepsilon_2$, with $\varepsilon_i \sim \mathcal{N}(0, \mathbf{I}_G)$, which is very much like the canonical Normal means problem, with only one additional requirement: $\mathbf{v}_1 = \omega_1 \mathbf{b}$ and $\mathbf{v}_2 = \omega_2 \mathbf{b}$. Here, the model assumes that the factor $\omega_j \sim \mathcal{N}(0, 1)$, and \mathbf{b} is a G -

dimensional loading matrix (vector). Thus, marginally we have $\mathbf{y}_i \sim \mathcal{N}(0, \mathbf{I}_G + \mathbf{b}\mathbf{b}^T)$, $i = 1, 2$.

A peculiar thing is that in the canonical Normal means problem, if we assign flat priors to \mathbf{v}_1 and \mathbf{v}_2 , their posterior distributions are simply $\mathcal{N}(\mathbf{y}_1, \mathbf{I}_G)$ and $\mathcal{N}(\mathbf{y}_2, \mathbf{I}_G)$, respectively, which are still proper although they yield inadmissible estimators for \mathbf{v}_1 and \mathbf{v}_2 when $G \geq 3$. However, with the factor model assumptions, which effectively reduce the number of parameters from $2G$ to G , the posterior distribution for \mathbf{b} becomes improper if we assign \mathbf{b} a flat prior and $G \geq 2$.

Mathematically equivalent phenomena occur even in the simple univariate Gaussian mean estimation: let $y \sim \mathcal{N}(\alpha\beta, 1)$. If we assume that $\alpha \sim \mathcal{N}(0, 1)$, then, when assuming a flat prior, the posterior distribution of β is proportional to $(\beta^2 + 1)^{-1/2} \exp\{-2(\beta^2 + 1)^{-1}y^2\}$, which is a non-integrable function, thus improper. But if we assume a proper prior on β , its posterior distribution becomes proper but its posterior variance relies heavily on its prior variance. A simple fix of the problem is to realize that we cannot identify both parameters simultaneously and have to let α take a fixed value. These phenomena also happen for the general factor models in certain settings, and our goal is to understand how these issues play out in high dimensional factor models and whether certain intuitive remedies work both theoretically and computationally for these more complex cases.

For the general factor model, we can similarly marginalize out the factor variables and derive the posterior distribution of the loading matrix under the flat prior:

$$\pi(\mathbf{B} \mid \mathbf{Y}, \mathbf{\Sigma}) \propto |\mathbf{B}\mathbf{B}^T + \mathbf{\Sigma}|^{-n/2} \exp\left\{-\frac{1}{2}\text{tr}\left[(\mathbf{B}\mathbf{B}^T + \mathbf{\Sigma})^{-1}\left(\sum_{i=1}^n \mathbf{y}_i \mathbf{y}_i^T\right)\right]\right\},$$

where the exponential term is both upper and lower bounded by some functions of \mathbf{Y} and $\mathbf{\Sigma}$. Term $|\mathbf{B}\mathbf{B}^T + \mathbf{\Sigma}|^{-n/2}$ is lower bounded by $(\|\mathbf{B}\|_F^2 + \lambda_{\max}(\mathbf{\Sigma}))^{-\frac{n \times K}{2}}$, where $\|\mathbf{B}\|_F$ represents the Frobenius norm of \mathbf{B} , and $\lambda_{\max}(\mathbf{\Sigma})$ denotes the largest eigenvalue of $\mathbf{\Sigma}$. When the dimension of \mathbf{B} , which is $G \times K$, is no smaller than $n \times K$, $\pi(\mathbf{B} \mid \mathbf{Y}, \mathbf{\Sigma})$ will integrate to infinity in the complement region of any bounded set in $\mathcal{R}^{G \times K}$, leading to an improper posterior distribution. If we impose a proper but

diffuse slab prior instead of the improper flat prior on elements of \mathbf{B} , the posterior distribution can still be very sensitive to the variance of slab prior, as seen in Figure 3.3.

To formalize this intuition for general Bayesian factor models, we provide the following theorem on the divergence of the posterior distribution of the loading matrix if we use a sequence of increasingly diffused “slab” priors. Note that for theorems in Section 3.4, we do not require Σ to be diagonal. To cover generic prior choices, we replace (3.2) with

$$p(\beta_{jk}|\gamma_{jk}) = (1 - \gamma_{jk})\psi(\beta_{jk}) + \gamma_{jk}\varphi(\beta_{jk}) \quad (3.5)$$

where ψ denotes the spike prior density and φ denotes the slab prior density.

Theorem 3.4.1. Let $\{\varphi_m\}_{m=1,\dots}$ be a sequence of densities such that $\lim_{m \rightarrow \infty} \varphi_m(\beta) = 0$ for every $\beta \in \mathcal{R}$ and there exists a constant $C \in (0, 1)$ such that $\varphi_m(\beta) > C \max_{\beta}(\varphi_m(\beta))$ holds for every β in some non-decreasing Borel sets S_m that converges to \mathcal{R} as $m \rightarrow \infty$. If $s = \|\Gamma\|_F^2/K \geq n$, then for any fixed finite-measure Borel set S , $\lim_{m \rightarrow \infty} P(\mathbf{B} \in S | \mathbf{Y}, \Sigma, \Gamma, m) = 0$, where $[\mathbf{B} | \mathbf{Y}, \Sigma, \Gamma, m]$ is based on the posterior distribution from model (3.1) with normally distributed factors and φ_m as the slab part in the SpSL prior on loading matrix elements.

Theorem 3.4.1 partially explains the magnitude inflation and the dependence of the inflation rate on the choice of the slab prior. Let S be any fixed $G \times K$ dimensional ball. The theorem implies that the probability of a posterior sample \mathbf{B} , conditional on $\mathbf{Y}, \Sigma, \Gamma, m$, having a matrix norm smaller than any constant goes to zero as we use a series of slab priors $\{\varphi_m\}_{m=1,2,\dots}$ that is increasingly diffused. In a general sense, it can also be understood as the convergence in distribution of $\mathbf{B} | \mathbf{Y}, \Sigma, \Gamma, m$ towards $\mathbf{B} | \mathbf{Y}, \Sigma, \Gamma, \infty$ (conditional posterior of B with flat slab prior), which is a point mass at infinity when $s \geq n$. For cases such that $\mathbf{B} | \mathbf{Y}, \Sigma, \Gamma, \infty$ is indeed proper, e.g., when $s \ll n$ or the assumed distribution on the factors is changed, we strictly have the convergence of $\mathbf{B} | \mathbf{Y}, \Sigma, \Gamma, m$ towards $\mathbf{B} | \mathbf{Y}, \Sigma, \Gamma, \infty$ in distribution as stated in the next theorem. Therefore, if the posterior distribution of the loading

matrix is proper under a flat slab prior and the Bayesian consistency is justified in this situation, we have approximately the same consistency when employing a reasonably diffuse slab prior.

Theorem 3.4.2. Consider model (3.1) without the normality assumption on factors. Let $\{\varphi_m\}_{m=1,\dots}$ be a sequence of prior densities maximized at 0 such that, $\forall \beta \in \mathbb{R}$, $\lim_{m \rightarrow \infty} \varphi_m(\beta) \varphi_m^{-1}(0) = 1$. Let $\pi(\mathbf{B}|\mathbf{Y}, \boldsymbol{\Sigma}, \boldsymbol{\Gamma}, m)$ denote the conditional posterior density of \mathbf{B} under a SpSL prior for its elements, with the spike density ψ and the slab density φ_m , and let $\pi(\mathbf{B}|\mathbf{Y}, \boldsymbol{\Sigma}, \boldsymbol{\Gamma}, \infty)$ be the one corresponding to the flat slab prior (this is appropriate since the indicator matrix $\boldsymbol{\Gamma}$ is conditioned on). If $\pi(\mathbf{B}|\mathbf{Y}, \boldsymbol{\Sigma}, \boldsymbol{\Gamma}, \infty)$ is integrable, then $\mathbf{B}|\mathbf{Y}, \boldsymbol{\Sigma}, \boldsymbol{\Gamma}, m$ converges to $\mathbf{B}|\mathbf{Y}, \boldsymbol{\Sigma}, \boldsymbol{\Gamma}, \infty$ in distribution as $m \rightarrow \infty$.

3.5 MODEL MODIFICATIONS AND POSTERIOR CONSISTENCY

To concentrate on the magnitude inflation and direction consistency problems, we study behaviors of the posterior distribution of the Bayesian factor model assuming that the diagonal idiosyncratic covariance matrix $\boldsymbol{\Sigma}$ and the true number of factors (for the basic factor model) or the true feature allocation matrix $\boldsymbol{\Gamma}$ (for the sparse factor model) are known. In contrast to the solution provided by Ghosh & Dunson (2009), which imposes dependency among the magnitudes of loading matrix elements within the same column through prior setup, we restrict ourselves to a special class of SpSL priors for loading matrix elements, which have a point mass at zero as the spike and a flat (limit of a sequence of increasingly diffused distributions) slab part. This is a natural choice for being non-informative and is always appropriate when considering the conditional posterior distributions given $\boldsymbol{\Gamma}$. We focus on studying the connection between posterior consistency and the factor assumption, and demonstrate why \sqrt{n} -orthonormal factor model is a natural choice under high dimensions.

Notations: Let H_n denote the Haar measure (i.e., uniform distribution) on the space of $n \times n$ orthogonal matrices and let m_n be the uniform measure on the Stiefel manifold $St(K, n)$. Let \mathbf{M}_i .

and $\mathbf{M}_{\cdot j}$ denote the i -th row and the j -th column of matrix \mathbf{M} , respectively, as column vectors, and let $\mathbf{M}_{i,j}$ denote the element at i -th row and j -th column of \mathbf{M} . $\mathbf{M}_{i_1:i_2}$ denotes the sub-matrix formed by row i_1 -th to i_2 and $\mathbf{M}_{i_1:i_2,j_1:j_2}$ denote the sub-matrix formed by rows i_1 -th to i_2 and columns j_1 to j_2 . Notation \mathbf{M}^\perp represents an orthogonal complement (not unique) of \mathbf{M} when \mathbf{M} is not a square matrix, $\mathcal{P}_{(\cdot)}$ represents the projection mapping towards the row vector space of a matrix and $\mathbf{P}_{(\cdot)}$ is the projection matrix of the mapping. Let $\lambda_{max}(\cdot)$ and $\lambda_{min}(\cdot)$ denote the largest and smallest singular values of a matrix, and let $\lambda_k(\cdot)$ denote the k -th largest singular values. The L_2 norm is denoted by $\|\cdot\|$, the Frobenius norm is denoted by $\|\cdot\|_F$ and the outer product is “ \otimes ”.

3.5.1 THE BASIC BAYESIAN FACTOR MODEL

We show the posterior consistency of the loading matrix by first studying the posterior consistency of the factor matrix $\mathbf{\Omega}$ (defined in section 2.2). It is easy to see that, with a flat prior on every element of \mathbf{B} , the posterior distribution of \mathbf{B} and $\mathbf{\Omega}$ can be written as:

$$\mathbf{B}_j | \mathbf{Y}, \mathbf{\Omega}, \mathbf{\Sigma} \stackrel{ind}{\sim} \mathcal{N}((\mathbf{\Omega}\mathbf{\Omega}^T)^{-1}\mathbf{\Omega}\mathbf{Y}_j, \sigma_j^2(\mathbf{\Omega}\mathbf{\Omega}^T)^{-1}) \quad (3.6)$$

$$\pi(d\mathbf{\Omega} | \mathbf{Y}, \mathbf{\Sigma}) \propto |\mathbf{\Omega}\mathbf{\Omega}^T|^{-G/2} \exp\left(\sum_{j=1}^G \frac{1}{2\sigma_j^2} \mathbf{Y}_j^T \mathbf{\Omega}^T (\mathbf{\Omega}\mathbf{\Omega}^T)^{-1} \mathbf{\Omega} \mathbf{Y}_j\right) p_{\mathbf{\Omega}}(d\mathbf{\Omega}) \quad (3.7)$$

where $p_{\mathbf{\Omega}}$ denotes the prior distribution of $\mathbf{\Omega}$ and “ $\stackrel{ind}{\sim}$ ” means that the \mathbf{B}_j ’s are mutually independent.

For this section, we no longer restrict the factors in $\mathbf{\Omega}$ to follow the standard Normal distribution, only requiring its distribution $p_{\mathbf{\Omega}}$ to satisfy the following two conditions: (a) $cov(\omega_i) = \mathbf{I}_K$, so as to keep the marginal covariance structure of \mathbf{Y} unchanged; (b) right rotational-invariant (i.e., $\mathbf{\Omega}$ and $\mathbf{\Omega}\mathbf{R}$ follow the same distribution $\forall n \times n$ orthogonal matrix \mathbf{R}). Two non-Gaussian examples are: (i) each row of $\mathbf{\Omega}$ follows independently a uniform distribution on the \sqrt{n} -radius sphere; (ii) $\mathbf{\Omega}/\sqrt{n}$ is

uniform on the Stiefel manifold $St(K, n)$, i.e., $\mathbf{\Omega}/\sqrt{n}$ is the first K rows of a Haar-distributed $n \times n$ orthogonal random matrix. A straightforward characterization of condition (b) can be made through the LQ decomposition (the transpose of the QR decomposition). Suppose the LQ decomposition of $\mathbf{\Omega} = \mathbf{K}(\mathbf{\Omega})\mathbf{V}(\mathbf{\Omega})$ is done by the Gram–Schmidt orthogonalization starting from the first row of $\mathbf{\Omega}$, resulting in a $K \times K$ lower triangular matrix $\mathbf{K}(\mathbf{\Omega})$ and a $K \times n$ orthonormal matrix $\mathbf{V}(\mathbf{\Omega})$. Then, requirement (b) enables us to generate $\mathbf{\Omega}$ from $p_{\mathbf{\Omega}}$ by generating a pair of $\mathbf{K}(\mathbf{\Omega})$ and $\mathbf{V}(\mathbf{\Omega})$ from two independent distributions—a marginal distribution on $\mathbf{K}(\mathbf{\Omega})$ (denoted as $p_{\mathbf{K}}$) and a uniform distribution on the Stiefel manifold $St(K, n)$ for $\mathbf{V}(\mathbf{\Omega})$.

Using the LQ decomposition, we can rewrite expression (3.7) as

$$\begin{aligned} \pi(d\mathbf{\Omega}|\mathbf{Y}, \mathbf{\Sigma}) &\propto \left(|\mathbf{K}(\mathbf{\Omega})\mathbf{K}(\mathbf{\Omega})^T|^{-G/2} p_{\mathbf{K}}(d\mathbf{K}(\mathbf{\Omega})) \right) \\ &\times \left(\exp\left(\sum_{j=1}^G \frac{1}{2\sigma_j^2} \|\mathcal{P}_{\mathbf{V}(\mathbf{\Omega})}(\mathbf{Y}_{j\cdot})\|^2 \right) m(d\mathbf{V}(\mathbf{\Omega})) \right) \end{aligned} \quad (3.8)$$

since $|\mathbf{\Omega}\mathbf{\Omega}^T| = |\mathbf{K}(\mathbf{\Omega})\mathbf{K}(\mathbf{\Omega})^T|$, and $\mathbf{Y}_{j\cdot}^T \mathbf{\Omega}^T (\mathbf{\Omega}\mathbf{\Omega}^T)^{-1} \mathbf{\Omega} \mathbf{Y}_{j\cdot}$ is the square of the length of $\mathbf{Y}_{j\cdot}$'s projection on the row space of $\mathbf{\Omega}$. Therefore, $\mathbf{K}(\mathbf{\Omega})$ and $\mathbf{V}(\mathbf{\Omega})$ are independent *a posteriori*, and

$$\pi(d\mathbf{K}(\mathbf{\Omega})|\mathbf{Y}, \mathbf{\Sigma}) \propto |\mathbf{K}(\mathbf{\Omega})\mathbf{K}(\mathbf{\Omega})^T|^{-G/2} p_{\mathbf{K}}(d\mathbf{K}(\mathbf{\Omega})) \quad (3.9)$$

$$\pi(d\mathbf{V}(\mathbf{\Omega})|\mathbf{Y}, \mathbf{\Sigma}) \propto \exp\left(\sum_{j=1}^G \frac{1}{2\sigma_j^2} \|\mathcal{P}_{\mathbf{V}(\mathbf{\Omega})}(\mathbf{Y}_{j\cdot})\|^2 \right) m(d\mathbf{V}(\mathbf{\Omega})). \quad (3.10)$$

Equation (3.9) implies that $\mathbf{K}(\mathbf{\Omega})$ may have an improper posterior distribution because the likelihood term $|\mathbf{K}(\mathbf{\Omega})\mathbf{K}(\mathbf{\Omega})^T|^{-G/2}$ creates “attractors” when the determinant of $\mathbf{K}(\mathbf{\Omega})\mathbf{K}(\mathbf{\Omega})^T$ is close to 0. Therefore, with large enough G , the right-hand side of (3.9) explodes to infinity fast enough around the attractors and becomes non-integrable, thus leading to an improper posterior distribution for

$\mathbf{K}(\boldsymbol{\Omega})$. In contrast, since $\exp\left(\sum_{j=1}^G \frac{1}{2\sigma_j^2} \|\mathcal{P}_{\mathbf{V}(\boldsymbol{\Omega})}(\mathbf{Y}_j)\|^2\right)$ is upper bounded by $\exp\left(\sum_{j=1}^G \frac{1}{2\sigma_j^2} \|\mathbf{Y}_j\|^2\right)$, the posterior distribution (3.10) for $\mathbf{V}(\boldsymbol{\Omega})$ is always proper, based on which we can further derive posterior consistency of the row vector space of $\boldsymbol{\Omega}$.

CONSISTENCY OF THE ROW VECTOR SPACE OF THE FACTOR MATRIX

The consistency of row vector space of $\boldsymbol{\Omega}$ is intuitive from (3.10) for the noiseless case (i.e., $\mathbf{Y} = \mathbf{B}_0\boldsymbol{\Omega}_0$), since the exponential term in (3.10) is uniquely maximized when the row vector spaces of $\boldsymbol{\Omega}$ and $\boldsymbol{\Omega}_0$ coincide. As in an annealing algorithm, the exponential term enforces the growing contraction towards the maximum point (where row spaces of $\boldsymbol{\Omega}$ and $\boldsymbol{\Omega}_0$ coincide) as G increases. On the other hand, the prior measure in a neighborhood of the row vector space of $\boldsymbol{\Omega}_0$ (defined as $p_{\boldsymbol{\Omega}}(\{\boldsymbol{\Omega} : \|\mathbf{V}(\boldsymbol{\Omega}_0)^\perp \mathbf{V}(\boldsymbol{\Omega})^T\|_F < \varepsilon\})$) gets more diffused as n grows. Therefore, in an asymptotic regime with $G, n \rightarrow \infty$, and under some mild conditions on the growing rate of G and n to ensure that the diffusion is slower than the contraction, the consistency of the row vector space of $\boldsymbol{\Omega}$ follows immediately as summarized below. Detailed proofs of the lemma and theorem can be found in Appendix C.3.

Lemma 3.5.1. Let $\mathbf{B}_{0,G}$ be a $G \times K$ matrix, $\boldsymbol{\Omega}_{0,n}$ be a $K \times n$ matrix, and $\boldsymbol{\Sigma}_G$ be a known $G \times G$ diagonal matrix. Suppose noiseless data generated as $\mathbf{Y} = \mathbf{B}_{0,G}\boldsymbol{\Omega}_{0,n}$ are given. We, however, model each column of \mathbf{Y} as mutually independent and $\mathbf{Y}_{\cdot i} \sim \mathcal{N}_G(\mathbf{B}\boldsymbol{\Omega}_{\cdot i}, \boldsymbol{\Sigma}_G)$, $i = 1, \dots, n$. With a flat prior on each of \mathbf{B} 's elements and a right-rotational invariant prior on $\boldsymbol{\Omega}$, we have the following inequality for the posterior distribution of $\boldsymbol{\Omega}$:

$$P(\|\mathbf{V}(\boldsymbol{\Omega}_{0,n})^\perp \mathbf{V}(\boldsymbol{\Omega})^T\|_F > \varepsilon | \mathbf{Y}, \boldsymbol{\Sigma}_G) \leq \left(1 + m_n(\{\mathbf{V} : \|\mathbf{V}_0 \mathbf{V}^T\|_F < \frac{\varepsilon}{L}\}) \times \exp\left(\frac{3}{8}\varepsilon^2 \lambda_{\min}(\boldsymbol{\Sigma}_G^{-1/2} \mathbf{B}_{0,G} \mathbf{K}(\boldsymbol{\Omega}_{0,n}))\right)\right)^{-1}$$

where $L = 2\lambda_{\max}(\boldsymbol{\Sigma}_G^{-1/2} \mathbf{B}_{0,G} \mathbf{K}(\boldsymbol{\Omega}_{0,n})) / \lambda_{\min}(\boldsymbol{\Sigma}_G^{-1/2} \mathbf{B}_{0,G} \mathbf{K}(\boldsymbol{\Omega}_{0,n}))$ and \mathbf{V}_0 is any fixed $K \times n$ orthonormal matrix.

Lemma 3.5.1 provides a probability bound between $\mathbf{V}(\boldsymbol{\Omega})$ sampled from the posterior distribution and $\mathbf{V}(\boldsymbol{\Omega}_{0,n})$ when there is no noise in the observation \mathbf{Y} . Since $\|\mathbf{V}(\boldsymbol{\Omega}_{0,n})^\perp \mathbf{V}(\boldsymbol{\Omega})^T\|_F^2$ equals to the sum of squared sine canonical angles between the row space of $\boldsymbol{\Omega}$ and $\boldsymbol{\Omega}_0$, lemma 3.5.1 implies the convergence of these canonical angles towards 0 as $n, G = s \rightarrow \infty$ (i.e. the Bayesian consistency of row vector space of $\boldsymbol{\Omega}$) when $-\log(m_n(\{\mathbf{V} : \|\mathbf{V}_0^\perp \mathbf{V}^T\|_F < \frac{\varepsilon}{L}\})) = o(\varepsilon^2 \lambda_{\min}(\boldsymbol{\Sigma}_G^{-1/2} \mathbf{B}_{0,G} \mathbf{K}(\boldsymbol{\Omega}_{0,n}))^2)$, which is the technical requirement that ensures the dilution is “covered up” by the contraction. Base on this lemma, we generalize the consistency of row vector space of $\boldsymbol{\Omega}$ to the noisy observation case under the “Large p(s), Small n” paradigm.

Definition 3.5.1. Let \mathbf{B}_0 be a countable array, or a bivariate function of the form $\mathbf{B}_0(j, k)$, with $j = 1, \dots, \infty$ and $k = 1, \dots, K$. Intuitively, this is an $\infty \times K$ matrix. We say that \mathbf{B}_0 is a regular infinite loading matrix if there are two universal constants $C_1, C_2 > 0$ such that, $\|(\mathbf{B}_0)_j\| \leq C_1$ and $\lambda_{\min}((\mathbf{B}_0)_{1:j})/\sqrt{j} \geq C_2$ for $j = 1, \dots, \infty$.

Theorem 3.5.1. Suppose \mathbf{B}_0 is a regular infinite loading matrix. Let $\boldsymbol{\Omega}_{0,n}$ be a $K \times n$ matrix with linear independent rows and let $\boldsymbol{\Sigma} = \text{diag}(\sigma_1^2, \dots)$ be a known infinite diagonal matrix in which $\sigma_j, \forall j$, is bounded below and above by constants $c_3 > 0$ and $c_4 < \infty$, respectively. Let \mathbf{Y} be an $\infty \times n$ matrix, whose j -th row is generated from $\mathcal{N}_n((\mathbf{B}_0)_j \boldsymbol{\Omega}_{0,n}, \sigma_j^2 \mathbf{I}_n)$, independently. For every fixed G , consider modeling the i -th column of $\mathbf{Y}_{1:G}$ by $\mathcal{N}_G(\mathbf{B}\boldsymbol{\Omega}_i, \boldsymbol{\Sigma}_G)$ for $i = 1, \dots, n$ with $\boldsymbol{\Sigma}_G = \text{diag}(\sigma_1^2, \dots, \sigma_G^2)$. With a flat prior on each of \mathbf{B} 's elements and a proper right-rotational invariant prior on $\boldsymbol{\Omega}$, we have, for a random draw $\boldsymbol{\Omega}$ from its posterior distribution, almost surely (with respect to the randomness in \mathbf{Y}) that

$$\|\mathbf{V}(\boldsymbol{\Omega}_{0,n})^\perp \mathbf{V}(\boldsymbol{\Omega})^T\|_F \mid \mathbf{Y}_{1:G}, \boldsymbol{\Sigma}_G \rightarrow 0 \text{ in probability as } G \rightarrow \infty.$$

POSTERIOR DISTRIBUTION OF THE LOADING MATRIX

From (3.10), it is clear that data only provide information on the row vector space of $\mathbf{V}(\boldsymbol{\Omega})$, the posterior distribution of $\mathbf{V}(\boldsymbol{\Omega})$ conditioned on its row vector space is uniform among all the $K \times n$ orthonormal matrices within the row space. Utilizing the posterior consistency of the row space provided by Theorem 3.5.1, we can approximate an $\mathbf{V}(\boldsymbol{\Omega})$ drawn from its posterior by another random variable of the form $\mathbf{O}\mathbf{V}(\boldsymbol{\Omega}_{0,n})$, where \mathbf{O} is a $K \times K$ uniform (Haar distributed) random orthogonal matrix (see Appendix C.3.5 for details).

Let $\mathbf{B}_{0,G}$ denotes the matrix formed by the first G rows of \mathbf{B}_0 . By plugging $\mathbf{V}(\boldsymbol{\Omega}) = \mathbf{O}\mathbf{V}(\boldsymbol{\Omega}_{0,n})$ into the matrix form of (3.6), which can be written as

$$\mathbf{B} \mid \mathbf{Y}, \boldsymbol{\Omega}, \boldsymbol{\Sigma} \sim \mathcal{N}_{K \times G}(\mathbf{Y}\boldsymbol{\Omega}^T(\boldsymbol{\Omega}\boldsymbol{\Omega}^T)^{-1}, (\boldsymbol{\Omega}\boldsymbol{\Omega}^T)^{-1} \otimes \boldsymbol{\Sigma}),$$

we obtain a decomposition for the posterior samples of $\mathbf{BK}(\boldsymbol{\Omega})/\sqrt{n}$ as:

$$\begin{aligned} \frac{1}{\sqrt{n}}\mathbf{BK}(\boldsymbol{\Omega}) \mid \mathbf{Y}, \boldsymbol{\Sigma} &\sim \mathbf{B}_{0,G}(\mathbf{K}(\boldsymbol{\Omega}_{0,n})/\sqrt{n})\mathbf{O}^T + ((\mathbf{Y} - \mathbf{B}_{0,G}\boldsymbol{\Omega}_{0,n})/\sqrt{n})\mathbf{V}(\boldsymbol{\Omega}_{0,n})^T\mathbf{O}^T \\ &+ \mathcal{N}_{G \times K}(0, \frac{1}{n}\mathbf{I}_K \otimes \boldsymbol{\Sigma}). \end{aligned} \quad (3.11)$$

For a considerable large n and normal true factor matrix $\boldsymbol{\Omega}_{0,n}$, $\mathbf{K}(\boldsymbol{\Omega}_{0,n})/\sqrt{n}$, as the Cholesky factor of $\boldsymbol{\Omega}_{0,n}\boldsymbol{\Omega}_{0,n}^T/n$, approaches the identity matrix, so the first term of the right hand side of (3.11) approaches $\mathbf{B}_{0,G}\mathbf{O}^T$. Meanwhile, the second term $((\mathbf{Y} - \mathbf{B}_{0,G}\boldsymbol{\Omega}_{0,n})/\sqrt{n})\mathbf{V}(\boldsymbol{\Omega}_{0,n})^T\mathbf{O}^T$ is the row projection of the idiosyncratic noise matrix $(\mathbf{Y} - \mathbf{B}_{0,G}\boldsymbol{\Omega}_{0,n})$ to a K dimensional space, divided by \sqrt{n} , which converges in probability to 0 entry-wise as $n \rightarrow \infty$. The third term is a centered normal (independent with \mathbf{O}) with variance shrinking to 0 as n increases. This implies that under $G = s \gg n \rightarrow \infty$ regime, posterior samples of $\mathbf{BK}(\boldsymbol{\Omega})/\sqrt{n}$ can be asymptotically expressed as the true loading matrix times an uniform random orthogonal matrix.

Factor assumption and consistency. Posterior distributions of \mathbf{B} and $\mathbf{K}(\boldsymbol{\Omega})$ are coupled. A “deflation” problem of $\mathbf{K}(\boldsymbol{\Omega})/\sqrt{n}$ occurs when the factors in $\boldsymbol{\Omega}$ are assumed to be normal and $n = O(G)$, in which case the posterior distribution of $\mathbf{K}(\boldsymbol{\Omega})/\sqrt{n}$ can be derived in closed form by the Bartlett decomposition as:

$$\begin{aligned} \frac{1}{\sqrt{n}}(\mathbf{K}(\boldsymbol{\Omega}))_{k,k}|\mathbf{Y}, \boldsymbol{\Sigma} &\sim \frac{1}{\sqrt{n}}\chi_{n-k+1-G}, \quad k = 1, \dots, K, \\ \frac{1}{\sqrt{n}}(\mathbf{K}(\boldsymbol{\Omega}))_{k',k}|\mathbf{Y}, \boldsymbol{\Sigma} &\sim \mathcal{N}\left(0, \frac{1}{n}\right), \quad 1 \leq k < k' \leq K, \end{aligned} \quad (3.12)$$

where χ_ν denotes the Chi distribution with ν degrees of freedom. Posterior samples of the loading matrix, therefore, have to be inflated correspondingly. Ideally, we desire the convergence of the posterior distribution of $\mathbf{K}(\boldsymbol{\Omega})/\sqrt{n}$ towards a point mass at the identity matrix to guarantee the posterior consistency (up to rotations) of the loading matrix, and can indeed achieve this by imposing a stronger control over the singular values of $\boldsymbol{\Omega}$ through the assumption on $p_{\boldsymbol{\Omega}}$. Such remedy is not unique. A particular simple strategy is to require that all factors are orthogonal and have equal norm, which implies that $\boldsymbol{\Omega}/\sqrt{n}$ is uniform in the Stiefel manifold $St(K, n)$. More discussions are deferred to Section 3.5.2.

3.5.2 SPARSE BAYESIAN FACTOR MODEL

With a special feature allocation design, $\mathbf{V}(\boldsymbol{\Omega})$ is identifiable so that the consistency of the row space of the factor matrix can be generalized to the consistency of $\mathbf{V}(\boldsymbol{\Omega})$. We impose a *generalized lower triangular structure* (Fruehwirth-Schnatter & Lopes, 2018) on the feature allocation matrix $\boldsymbol{\Gamma}$ to cope with the rotational invariance problem of the loading matrix. We call $\boldsymbol{\Gamma}$ a generalized lower triangular matrix if the row index of the top nonzero entry in the k -th column l_k (define $l_0 = 1, l_{K+1} = G + 1$) increases with k and $\gamma_{jk} = 1$ if and only if $j \geq l_k$. Under the flat SpSL prior (use a mixture of point mass at zero and flat distribution as prior) on entries of \mathbf{B} in the Sparse Bayesian factor model introduced

in section 3.2, we can derive the conditional distributions of \mathbf{B} and $\mathbf{\Omega}$: for $j = l_k, \dots, l_{k+1} - 1$,

$$\mathbf{B}_{j,1:k} | \mathbf{Y}, \mathbf{\Omega}, \mathbf{\Sigma}, \mathbf{\Gamma} \stackrel{ind}{\sim} \mathcal{N}((\mathbf{\Omega}_{1:k} \mathbf{\Omega}_{1:k}^T)^{-1} \mathbf{\Omega}_{1:k} \mathbf{Y}_j, \sigma_j^2 (\mathbf{\Omega}_{1:k} \mathbf{\Omega}_{1:k}^T)^{-1}), \quad (3.13)$$

$$\pi(d\mathbf{\Omega} | \mathbf{Y}, \mathbf{\Sigma}, \mathbf{\Gamma}) \propto \prod_{k=1}^K |\mathbf{\Omega}_{1:k} \mathbf{\Omega}_{1:k}^T|^{-(l_{k+1}-l_k)/2} \exp \left(\sum_{k=1}^K \sum_{j=l_k}^{l_{k+1}-1} \frac{1}{2\sigma_j^2} \|\mathcal{P}_{\mathbf{\Omega}_{1:k}}(\mathbf{Y}_j)\|^2 \right) p_{\mathbf{\Omega}}(d\mathbf{\Omega}), \quad (3.14)$$

where $\mathbf{B}_{j,1:k} = (\beta_{j1}, \beta_{j2}, \dots, \beta_{jk})^T$ and $p_{\mathbf{\Omega}}$ denotes the distribution assumed on $\mathbf{\Omega}$ such that condition (a) and (b) holds.

Given the LQ decomposition $\mathbf{\Omega} = \mathbf{K}(\mathbf{\Omega})\mathbf{V}(\mathbf{\Omega})$ and

$$\mathbf{\Omega}_{1:k} = \mathbf{K}(\mathbf{\Omega})_{1:k} \mathbf{V}(\mathbf{\Omega}) = \mathbf{K}(\mathbf{\Omega})_{1:k,1:k} \mathbf{V}(\mathbf{\Omega})_{1:k},$$

since $\mathbf{K}(\mathbf{\Omega})$ is lower triangular, $\mathbf{\Omega}_{1:k} \mathbf{\Omega}_{1:k}^T = \mathbf{K}(\mathbf{\Omega})_{1:k,1:k} \mathbf{K}(\mathbf{\Omega})_{1:k,1:k}^T$ is a function of $\mathbf{K}(\mathbf{\Omega})$. $\mathcal{P}_{\mathbf{\Omega}_{1:k}}(\mathbf{Y}_j)$ is the projection of \mathbf{Y}_j towards the row vector space of $\mathbf{\Omega}_{1:k}$, which is a function of $\mathbf{V}(\mathbf{\Omega})$. The adoption of the generalized lower triangular structure on feature allocation matrix ensures a separation in likelihood of (3.14) so that the determinant part is connected to $\mathbf{\Omega}$ only through $\mathbf{K}(\mathbf{\Omega})$ and the exponential part only through $\mathbf{V}(\mathbf{\Omega})$. We thus can derive that $\mathbf{K}(\mathbf{\Omega})$ and $\mathbf{V}(\mathbf{\Omega})$ are independent *a posteriori* and that:

$$\pi(d\mathbf{K}(\mathbf{\Omega}) | \mathbf{Y}, \mathbf{\Sigma}, \mathbf{\Gamma}) \propto \prod_{k=1}^K \mathbf{K}(\mathbf{\Omega})_{k,k}^{-(G-l_k+1)} p_K(d\mathbf{K}(\mathbf{\Omega})) \quad (3.15)$$

$$\pi(d\mathbf{V}(\mathbf{\Omega}) | \mathbf{Y}, \mathbf{\Sigma}, \mathbf{\Gamma}) \propto \exp \left(\sum_{k=1}^K \sum_{j=l_k}^{l_{k+1}-1} \frac{1}{2\sigma_j^2} \|\mathcal{P}_{\mathbf{V}(\mathbf{\Omega})_{1:k}}(\mathbf{Y}_j)\|^2 \right) m(d\mathbf{V}(\mathbf{\Omega})). \quad (3.16)$$

Expression (3.16) gives a proper posterior for $\mathbf{V}(\boldsymbol{\Omega})$, and for the noiseless case (i.e. $\mathbf{Y} = \mathbf{B}_0\boldsymbol{\Omega}_0$), the density is maximized when the row vector space of $\mathbf{V}(\boldsymbol{\Omega})_{1:k}$ and $\mathbf{V}(\boldsymbol{\Omega}_0)_{1:k}$ coincide for $k = 1, \dots, K$, based on which we can generalize theorem 5.2 to the consistency (up to sign permutations) of $\mathbf{V}(\boldsymbol{\Omega})$.

CONSISTENCY OF $\mathbf{V}(\boldsymbol{\Omega})$

Definition 3.5.2. Let \mathbf{B}_0 be an $\infty \times K$ matrix with nonzero rows and let Γ_0 be a binary matrix of the same shape. We call Γ_0 a generalized lower triangular feature allocation matrix of \mathbf{B}_0 if it satisfies

1. $\mathbb{I}_{(\mathbf{B}_0)_{j,k} \neq 0} \leq (\Gamma_0)_{j,k}$ holds for $j = 1, \dots, \infty, k = 1, \dots, K$, where \mathbb{I} is the indicator function;
2. $(\Gamma_0)_{j,k_1} \leq (\Gamma_0)_{j,k_2}$ holds for $j = 1, \dots, \infty, K \geq k_1 > k_2 \geq 1$.

Furthermore, for every fixed dimension G , let ψ_G denote the unique permutation of $(1, \dots, G)$, so that $\psi_G(j_1) < \psi_G(j_2)$ if and only if either (i) $(\sum_k \Gamma_{j_1,k}) < (\sum_k \Gamma_{j_2,k})$ or (ii) $(\sum_k \Gamma_{j_1,k}) = (\sum_k \Gamma_{j_2,k})$ but $j_1 < j_2$.

Definition 3.5.3. Let \mathbf{B}_0 be a $\infty \times K$ matrix with nonzero rows and let Γ_0 be a generalized lower triangular feature allocation matrix of \mathbf{B}_0 . The two $G \times K$ matrices $\mathbf{B}_{0,G}$ and $\Gamma_{0,G}$ are formed by permuting the first G rows of \mathbf{B}_0 and Γ_0 according to ψ_G (the j -th row of \mathbf{B}_0 is the $\psi_G(j)$ -th row of $\mathbf{B}_{0,G}$). Let $l_{0,k}$ be the row index of the top nonzero entry in the k -th column of the generalized lower triangular matrix $\Gamma_{0,G}$ (define $l_{0,0} = 1, l_{0,K+1} = G + 1$), and let $\mathbf{B}_{0,G}^{(k)}$ be the submatrix of $\mathbf{B}_{0,G}$ formed by rows indexed from $l_{0,k}$ to $l_{0,k+1} - 1$ and columns indexed from 1 to k . We call (\mathbf{B}_0, Γ_0) a regular infinite loading pair if there are two universal constants $C_1, C_2 > 0$ such that, $\|(\mathbf{B}_0)_j\| \leq C_1$ and $\min_k \lambda_{\min}(\mathbf{B}_{0,j}^{(k)}) / \sqrt{j} \geq C_2$ for $j = 1, \dots, \infty$.

Theorem 3.5.2. Let (\mathbf{B}_0, Γ_0) be a regular infinite loading pair with Γ_0 known, let $\boldsymbol{\Omega}_{0,n}$ be a $K \times n$ matrix with linearly independent rows, and let $\boldsymbol{\Sigma} = \text{diag}(\sigma_1^2, \dots)$ be a known infinite diagonal matrix

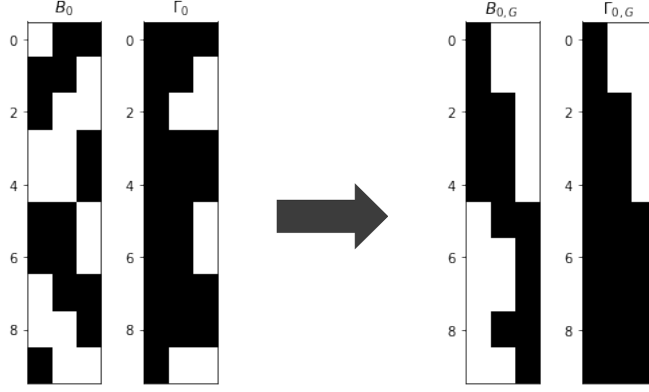


Figure 3.4: An example of \mathbf{B}_0, Γ_0 , and $\mathbf{B}_{0,G}, \Gamma_{0,G}$ after ψ_G permutation.

such that $C_3 \leq \sigma_j^2 \leq C_4$ holds for $j = 1, \dots$, with constants $C_3, C_4 > 0$. The j -th row of $\infty \times n$ matrix \mathbf{Y} is generated by $\mathcal{N}_n((\mathbf{B}_0)_j \cdot \boldsymbol{\Omega}_{0,n}, \sigma_j^2 \mathbf{I}_n)$. For every fixed G , let $\mathbf{Y}_{1:G}$ denote the matrix formed by permuting the first G rows of \mathbf{Y} according to ψ_G and consider modeling the i -th column of $\mathbf{Y}_{1:G}$ by $\mathcal{N}_G(\mathbf{B}\boldsymbol{\Omega}_{\cdot i}, \boldsymbol{\Sigma}_G)$ for $i = 1, \dots, n$ with $\boldsymbol{\Sigma}_G = \text{diag}(\sigma_{\psi_G^{-1}(1)}^2, \dots, \sigma_{\psi_G^{-1}(G)}^2)$. With a flat prior on each of \mathbf{B} 's non-zero element according to the feature allocation matrix $\Gamma_{0,G}$ and a prior on $\boldsymbol{\Omega}$ that is invariant under right orthogonal transformations, for a random draw $\boldsymbol{\Omega}$ from its posterior distribution, we have almost surely (with respect to the randomness in \mathbf{Y}) that

$$\|\mathbf{V}(\boldsymbol{\Omega}_{0,n})_{1:k}^\perp \mathbf{V}(\boldsymbol{\Omega})_{1:k}^T\|_F | \mathbf{Y}_{1:G}, \boldsymbol{\Sigma}_G, \Gamma_{0,G} \rightarrow 0,$$

for $k = 1, \dots, K$ as $G \rightarrow \infty$.

Theorem 3.5.2 is understood as the consistency (up to sign permutations) of $\mathbf{V}(\boldsymbol{\Omega})$ for fixed n and $G \asymp s \rightarrow \infty$, in the sense that $\|\mathbf{V}(\boldsymbol{\Omega}_{0,n})_{1:k}^\perp \mathbf{V}(\boldsymbol{\Omega})_{1:k}^T\|_F$ converges to 0 for all k , which implies that the canonical angles between the row space of $\mathbf{V}(\boldsymbol{\Omega}_{0,n})_{1:k}$ and that of $\mathbf{V}(\boldsymbol{\Omega})_{1:k}$ converge to 0 as $G \rightarrow \infty$. When these angles are all equal to 0, $\mathbf{V}(\boldsymbol{\Omega})$ differs from $\mathbf{V}(\boldsymbol{\Omega}_{0,n})$ only by a sign for each row. Since the data provides no information on the signs, in the asymptotic regime with $G \asymp s \gg n \rightarrow \infty$,

we can approximate $\mathbf{V}(\boldsymbol{\Omega})$ drawn from its posterior distribution by a random sign diagonal matrix \mathbf{S} , i.e., a diagonal matrix with *i.i.d.* random signs on the diagonal, times $\mathbf{V}(\boldsymbol{\Omega}_{0,n})$.

POSTERIOR SAMPLE CONSISTENCY

Recall that from Section 3.5.1, for the basic Bayesian factor model with $G = s \gg n \rightarrow \infty$, $\mathbf{BK}(\boldsymbol{\Omega})/\sqrt{n}$ drawn from the posterior distribution can be asymptotically represented as the true loading matrix times a uniform random orthogonal matrix. If the true feature allocation matrix is lower triangular, we have

$$\begin{aligned} \mathbf{B}^{(k)}\mathbf{K}(\boldsymbol{\Omega})_{1:k}/\sqrt{n} | \mathbf{Y}, \boldsymbol{\Omega}, \boldsymbol{\Sigma}_G, \boldsymbol{\Gamma}_{0,G} &\sim \mathbf{B}_{0,G}^{(k)}(\mathbf{K}(\boldsymbol{\Omega}_{0,n})_{1:k,1:k}/\sqrt{n})\mathbf{V}(\boldsymbol{\Omega}_{0,n})_{1:k}\mathbf{V}(\boldsymbol{\Omega})_{1:k}^T \\ &+ ((\mathbf{Y}_{l_k:l_{k+1}-1} - \mathbf{B}_{0,G}^{(k)}(\boldsymbol{\Omega}_{0,n})_{1:k})/\sqrt{n})\mathbf{V}(\boldsymbol{\Omega})_{1:k}^T \quad (3.17) \\ &+ \mathcal{N}_{(l_{k+1}-l_k) \times k}(\mathbf{0}, \frac{1}{n}\mathbf{I}_k \otimes \boldsymbol{\Sigma}_G^{(k)}), \end{aligned}$$

whose right hand side converges entry-wise in probability to $\mathbf{B}_{0,G}^{(k)}\mathbf{S}_{1:k,1:k}^T$ under the $G \asymp s \gg n \rightarrow \infty$ setting (by similar argument as in section 5.1.2). Note that $\mathbf{B}^{(k)}\mathbf{K}(\boldsymbol{\Omega})_{1:k} = \mathbf{B}_{l_k:l_{k+1}-1}\mathbf{K}(\boldsymbol{\Omega})$, we can therefore summarize the convergence of $\mathbf{B}^{(k)}\mathbf{K}(\boldsymbol{\Omega})_{1:k}/\sqrt{n}$ to derive the convergence of posterior samples of $\mathbf{BK}(\boldsymbol{\Omega})/\sqrt{n}$ towards $\mathbf{B}_{0,G}\mathbf{S}^T$.

The posterior sample consistency (up to sign permutations) of the loading matrix is immediate once we have $\mathbf{K}(\boldsymbol{\Omega})/\sqrt{n}$, or equivalently $\boldsymbol{\Omega}\boldsymbol{\Omega}^T/n$, from its posterior distribution converging in probability to the identity matrix. The density in (3.15) indicates that the posterior distribution of $\boldsymbol{\Omega}\boldsymbol{\Omega}^T/n$ is contributed by two terms: the determinant $\prod_{k=1}^K \mathbf{K}(\boldsymbol{\Omega})_{k,k}^{-(G-l_k+1)}$ and the model assumption represented by $p_{\boldsymbol{\Omega}}$. The determinant term creates singularities when $\mathbf{K}(\boldsymbol{\Omega})_{k,k} = 0$ and the order of these ‘‘poles’’ $\sim s$. When this term dominates, we observe the inflation phenomenon of posterior samples of the loading matrix. Meanwhile, the model assumption term can bound $\mathbf{K}(\boldsymbol{\Omega})$ away from these singularities by assigning little probability measure in their neighborhoods and also induces the convergence

of $\mathbf{\Omega}\mathbf{\Omega}^T/n$ towards the identity matrix (through requirement (a) introduced in section 3.5.1). Consequently, the posterior behavior of $\mathbf{\Omega}\mathbf{\Omega}^T/n$ is influenced by both the increasing rate of n, s and the choice of distribution $p_{\mathbf{\Omega}}$. Those $p_{\mathbf{\Omega}}$ that bounds away singularities with high probability and forces a fast convergence of $\mathbf{\Omega}\mathbf{\Omega}^T/n$ towards the identity matrix can allow a fast rate of s going to infinity comparing to n , to guarantee the posterior consistency of the loading matrix. A simple and effective choice is to adopt the \sqrt{n} -orthonormal factor model. That is, we assume *a priori* that $\mathbf{\Omega}/\sqrt{n}$ is uniform in the Stiefel manifold $St(K, n)$. With this choice, we have $\mathbf{\Omega}\mathbf{\Omega}^T/n = I_K$ and that the posterior sample consistency of \mathbf{B} naturally holds even when n has a rather slow growing rate compared with s .

Our analysis regarding the relation between the factor assumption and the magnitude problem is specific to the independent spike and slab prior setup. But we believe that the magnitude problem, meaning that the column-wise magnitude of the loading matrix sampled from its posterior distribution is very sensitive to its prior distribution, exists for general priors in the “Large s , Small n ” regime when the factors are only assumed to be normally distributed. When a more complicated prior is assigned on the loading matrix, the analysis becomes rather challenging and the magnitude problem may be expressed in other forms (as we will see in the simulations) rather than an ‘inflation’ (inflation is typical for using non-informative priors on loading matrix). The intuition we gain from the analysis is that the factor assumption crucially impacts the strength of posterior contraction (through data) of the magnitude of the loading matrix towards its true value, and using \sqrt{n} -orthonormal factors achieves the strongest contraction thus allows more flexibility in prior assignment of the loading matrix while maintaining the posterior consistency.

3.6 NUMERICAL RESULTS

3.6.1 MODIFICATION OF THE GIBBS SAMPLER

In Section 3.5, we justify the adoption of the \sqrt{n} -orthonormal factor model in the “Large s , Small n ” paradigm (i.e., the factor matrix $\mathbf{\Omega}$ scaled by $1/\sqrt{n}$ is uniform in the Stiefel manifold $St(K, n)$). To construct a Gibbs sampler under this new factor model and the prior setup in Section 3.2.1 (denoted as SpSL-orthonormal factor model), we only need to revise the conditional sampling step of $\mathbf{\Omega}|\mathbf{Y}, \mathbf{B}, \mathbf{\Sigma}$ in the basic Gibbs sampler described in Section 3.2.2.

Let $\mathbf{\Omega}_k$ denote the k -th row of the factor matrix and $\mathbf{\Omega}_{-k}$ denote the remaining rows, all as column vectors. The conditional distribution $\mathbf{\Omega}_k|\mathbf{Y}, \mathbf{\Omega}_{-k}, \mathbf{B}, \mathbf{\Sigma}$ is altered from a multivariate normal distribution to:

$$\pi(d\mathbf{\Omega}_k|\mathbf{Y}, \mathbf{\Omega}_{-k}, \mathbf{B}, \mathbf{\Sigma}) \propto f(\mathbf{\Omega}_k; \bar{\mathbf{\Omega}}_k, \bar{\sigma}_k^2 \mathbf{I}_n) \times p_{\mathbf{\Omega}_{-k}}(d\mathbf{\Omega}_k) \quad (3.18)$$

where $p_{\mathbf{\Omega}_{-k}}$ is the uniform measure on the centred \sqrt{n} -radius sphere in the orthogonal space of $\mathbf{\Omega}_{-k}$, and $f(\mathbf{\Omega}_k; \bar{\mathbf{\Omega}}_k, \bar{\sigma}_k^2 \mathbf{I}_n)$ is the multivariate normal density function with mean $\bar{\mathbf{\Omega}}_k$ and covariance matrix $\bar{\sigma}_k^2 \mathbf{I}_n$, with

$$\bar{\mathbf{\Omega}}_k = (\mathbf{B}_{\cdot k}^T \mathbf{\Sigma}^{-1} \mathbf{B}_{\cdot k})^{-1} (\mathbf{Y} - \sum_{t \neq k} \mathbf{B}_{\cdot t} \mathbf{\Omega}_t^T)^T \mathbf{\Sigma}^{-1} \mathbf{B}_{\cdot k}, \quad \bar{\sigma}_k^2 = (\mathbf{B}_{\cdot k}^T \mathbf{\Sigma}^{-1} \mathbf{B}_{\cdot k})^{-1}.$$

To sample from (3.18), we cut this \sqrt{n} -radius sphere by hyperplanes that are orthogonal to vector $\bar{\mathbf{\Omega}}_k$ and denote this collection of intersections of the sphere and hyperplanes as $\{S_d \mid d \in (-\sqrt{n}, \sqrt{n})\}$, where d is the Euclidean distance between the origin and the hyperplane. Essentially, $\{S_d\}$ are $(n-k)$ -dimensional spheres and every point in the same S_d has the same multivariate normal density $f(\cdot; \bar{\mathbf{\Omega}}_k, \bar{\sigma}_k^2 \mathbf{I}_n)$, so we can sample $\mathbf{\Omega}_k$ from (3.18) by first sampling d from its marginal distribution and then uni-

formly sample from sphere S_d given the sampled d . Using the area formula of sphere, we can deduce the marginal distribution for d as

$$\pi(d|\mathbf{Y}, \mathbf{\Omega}_{-k}, \mathbf{B}, \mathbf{\Sigma}) \propto (n - d^2)^{(n-K-2)/2} \exp(-\|\mathcal{P}_{\mathbf{\Omega}_{-k}^\perp}(\bar{\mathbf{\Omega}}_k)\|d/\bar{\sigma}_k^2) \quad (3.19)$$

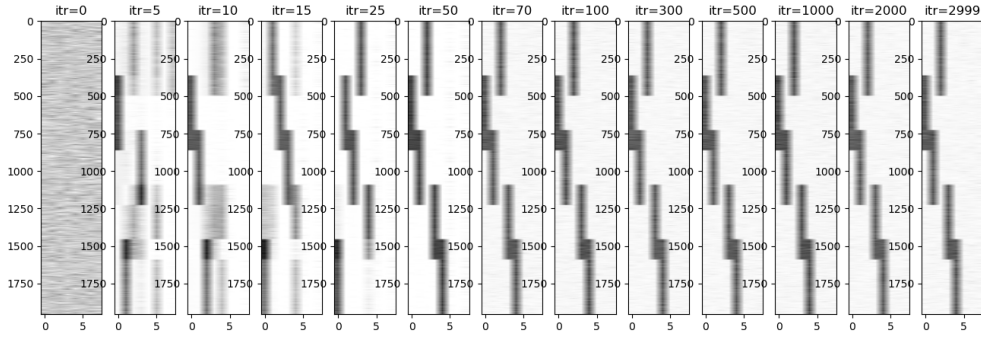
and sample from this unimodal distribution using the Metropolis algorithm. The additional computational cost brought by the model revision only comes from the Metropolis algorithm and is almost negligible.

3.6.2 COMPARISON WITH ALTERNATIVE APPROACHES

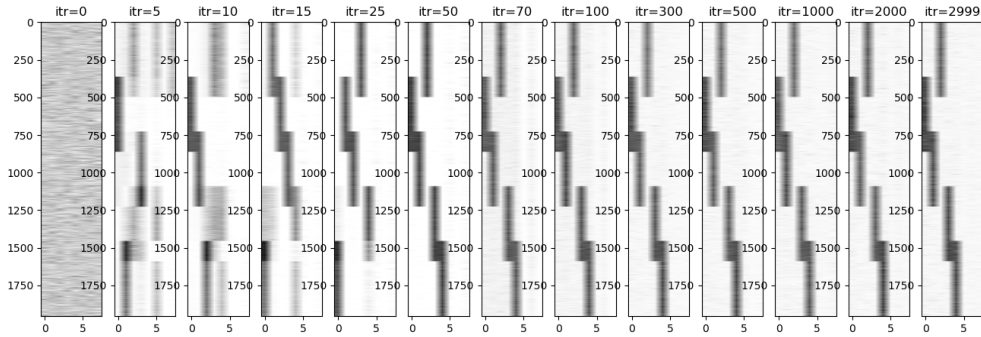
We now revisit the synthetic example in Section 3.3.1 to check the consistency of the posterior distribution of the loading matrix under the SpSL-orthonormal factor model and compare the MCMC performance with the sampler of two alternative approaches: a modified Ghosh-Dunson model (details provided in Appendix B) and the model from [Bhattacharya & Dunson \(2011\)](#) (applied with $\nu = 3, a_\sigma = 1, b_\sigma = 0.3, a_1, a_2 \sim \text{Gamma}(2, 1)$). The factor dimensionality K is fixed at 8 in all Gibbs samplers.

Figure 3.5 shows the heat map of $|\mathbf{B}|$ in 3000 iterations. We perform the PXL-EM algorithm for the first 50 iterations and then Gibbs sampling in all three approaches, respectively, for the next 2950 iterations. Figure 3.6 shows the posterior means of the nonzero elements of loading matrix obtained by averaging over 2500 posterior samples after burn-in. For the SpSL-orthonormal factor model, the posterior means are nicely centered around the true value 1. For the other two approaches, there exhibit some ‘twists’ in the column-wise magnitude of the posterior means. This is most obvious for the fifth (purple) column of panel (c) in Figure 3.6.

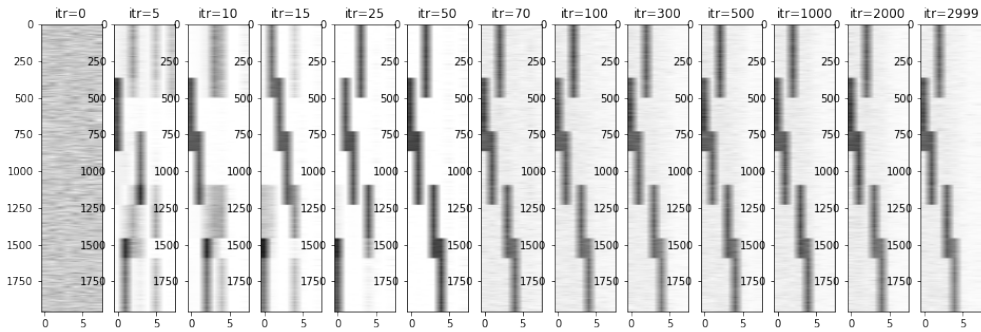
To understand the cause, we examine more closely the priors employed by the latter two approaches. These two approaches share the same idea of imposing dependency among the magnitudes



(a) The SpSL-orthonormal factor model



(b) The modified Ghosh-Dunson model



(c) The model from [Bhattacharya & Dunson \(2011\)](#)

Figure 3.5: Heat-maps of $|B|$ in 3000 iterations of Gibbs sampler using specified models.

of loading matrix elements within the same column via the decomposition $B = Q \times D$ as we explained in the introduction. However, they differ in the scheme of learning factor dimensionality: the modified Ghosh-Dunson model adopts a SpSL prior on elements of Q , an Indian buffet process on Θ and

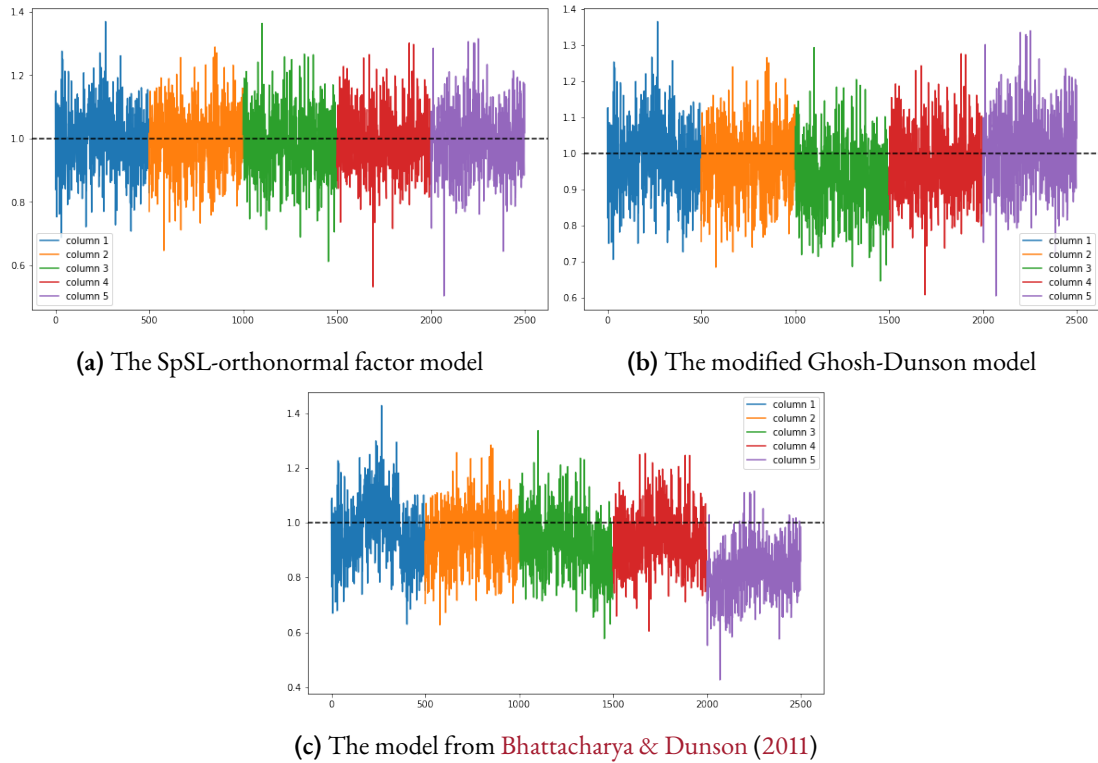


Figure 3.6: Posterior means of the nonzero elements of loading matrix under three factor models. Nonzero elements are sorted first by the column index and then by the row index, both in ascending order, e.g. the first 500 entries colored in blue correspond to the posterior means of $\beta_{365,1}, \dots, \beta_{864,1}$.

a diffuse prior on diagonals of \mathbf{D} ; whereas [Bhattacharya & Dunson \(2011\)](#) uses a continuous prior on elements of \mathbf{Q} and a shrinkage prior on \mathbf{D} . The former model learns factor dimensionality through the shrinkage on the feature sparsity vector Θ while the latter does so through the shrinkage on diagonals of \mathbf{D} . Under the “Large s , Small n ” regime, using an informative prior on \mathbf{D} can be influential for the posterior of the column-wise magnitude and results in the ‘twists’ in panel (c). As for the ‘twist’ in panel (b), we think it is caused by the high auto-correlation among the samples generated by the Gibbs sampler for the modified Ghosh-Dunson model, so that the sample mean estimator still has a large Monte Carlo error using 2500 Gibbs samples.

With a sufficient computation budget, Gibbs sampler for the modified Ghosh-Dunson model

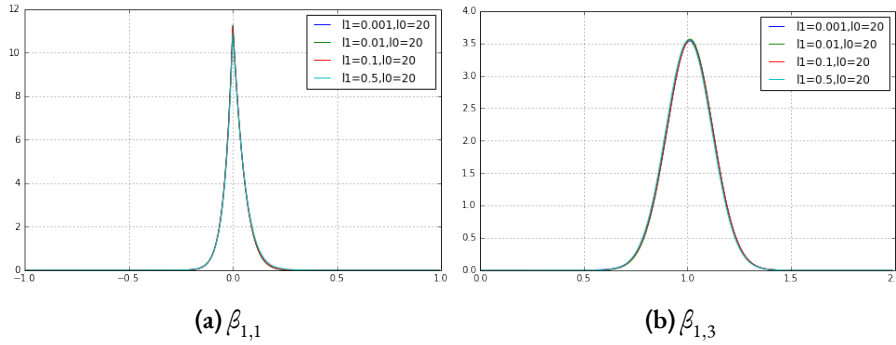


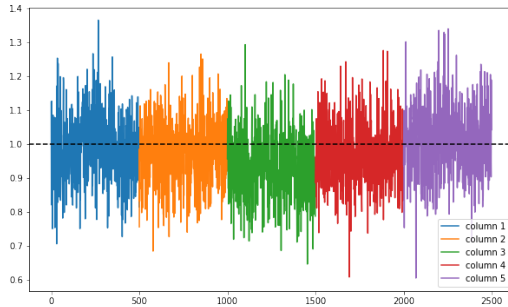
Figure 3.7: Posterior densities of (a) $\beta_{1,1}$, and (b) $\beta_{1,3}$, with $\lambda_1 \in \{0.001, 0.01, 0.1, 0.5\}$, under SpSL-orthonormal factor model. The posterior densities are robust against the choice of slab priors.

gives similar posterior results as our approach. This highlights another advantage of our approach—computational efficiency. When running the corresponding Gibbs sampler for 2500 rounds after burn-in, our approach with the SpSL-orthonormal factor model attained an average effective sample size (ESS) of 2758.8; whereas the ESS for the other two approaches are only 51.6 and 82.5, respectively, on average. The computation times per iteration of Gibbs sampling for the three methods are 2.8, 2.2, and 1.8 seconds, respectively. Besides computational aspect, although both the SpSL-orthonormal factor model and the modified Ghosh-Dunson model give very similar numerical results after appropriately adjusting tuning parameters of the priors, our analysis rigorously justifies the consistency of the former model, whereas a similar theoretical study of the latter model is still beyond our reach.

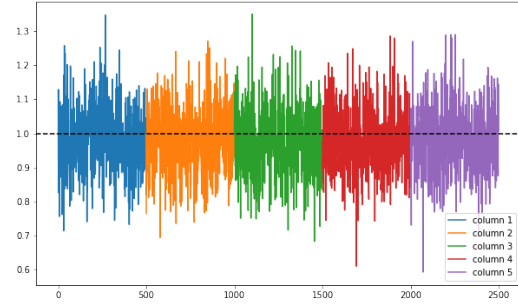
3.6.3 ROBUSTNESS AGAINST PRIOR SPECIFICATION

Under the SpSL-orthonormal factor model, Figure 3.7 illustrates the posterior density of $\beta_{1,1}$ and $\beta_{1,3}$ (estimated by averaging over the conditional posterior densities) using slab priors with ranging variances. We tested with $\lambda_0 = 20$, $\lambda_1 \in \{0.001, 0.01, 0.1, 0.5\}$ and the posterior distribution shows a great robustness against the choice of the slab prior.

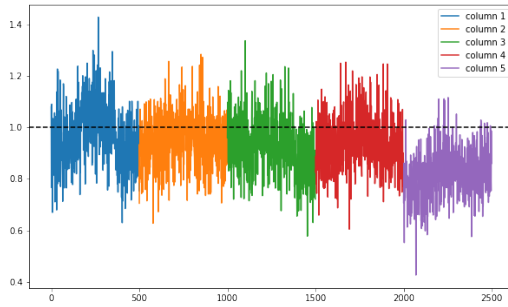
At the end of Section 3.5.2, we claim that restricting to \sqrt{n} -orthonormal factors grants more



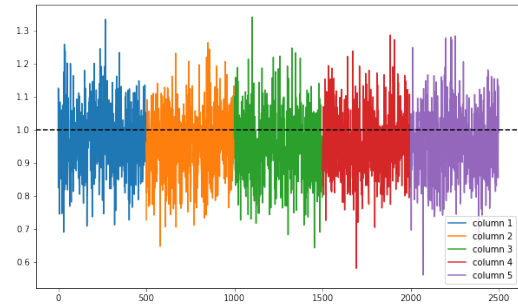
(a) The modified Ghosh-Dunson model with normal factors



(b) The modified Ghosh-Dunson model with \sqrt{n} -orthonormal factors



(c) The model of [Bhattacharya & Dunson \(2011\)](#) with normal factors



(d) The model of [Bhattacharya & Dunson \(2011\)](#) with \sqrt{n} -orthonormal factors

Figure 3.8: Posterior mean of the nonzero elements of loading matrix using specified models. Column-wise magnitudes of the loading matrix from posterior are balanced after changing from normal factors (left panels) to \sqrt{n} -orthonormal factors (right panels).

flexibility in prior assignment of the loading matrix while maintaining the posterior consistency. We verify this claim by applying the prior setups from Ghosh-Dunson model and [Bhattacharya & Dunson \(2011\)](#) to the \sqrt{n} -orthonormal factor model. Note that we only need to revise the conditional sampling step $\mathbf{\Omega}|\mathbf{Y}, \mathbf{B}, \mathbf{\Sigma}$ in the Gibbs samplers as we did in Section 3.6.1.

Figure 3.8 plots the posterior means of the nonzero elements of the loading matrix estimated by averaging over 2500 posterior samples. We observe that the ‘twists’ in column-wise magnitudes disappear after switching to the \sqrt{n} -orthonormal factor model. Furthermore, the average ESS increases significantly from 51.6 and 82.5 to 2667.9 and 2296.6, respectively, for the two approaches since the source of high auto-correlations—strong tie between the magnitudes of the loading matrix and the

factors, is removed by restricting the magnitude of factors to a specific value. Summary figures for credible intervals of the loading matrix elements of all implemented approaches are illustrated in the appendix of [Ma & Liu \(2020\)](#).

3.7 DYNAMIC EXPLORATION WITH APPLICATION

Although the \sqrt{n} -orthonormal factor model can be coupled with general prior assignments on the loading matrix, we focus on the setup from [Ročková & George \(2016\)](#) (i.e., the SpSL-orthonormal factor model), under which posterior consistency has a theoretical guarantee. When applying this framework to real data, the choice of the factor dimensionality K as well as the penalty parameters λ_0 and λ_1 (parameters in the spike and the slab parts, respectively) is crucial.

The application of our Gibbs sampler requires a successful implementation of the PXL-EM algorithm to search for a posterior mode that can serve to initialize the sampler. For the choice of K when applying the sampler, we make two recommendations: (i) use the estimated number of factors from PXL-EM as a plug-in estimator for K ; (ii) choose K to be sufficiently large initially and discard the useless factors (whose corresponding $\{\gamma_{jk}\}_{j=1, \dots, G}$ are all zero) in the sampling process, which is similar to the idea of choosing the number of factors adaptively from [Bhattacharya & Dunson \(2011\)](#). More precisely, we discard useless factors if there are any, and append a null factor whenever there is no useless factor remained. Though this adaptive approach also provides posterior samples for K , it is worth mentioning that the computational complexity of the Gibbs sampler scales linearly with the factor dimensionality K .

The penalty parameters determine the threshold for a loading matrix's element to follow either a spike or a slab prior. For the PXL-EM algorithm, Ročková and George proposed a dynamic posterior exploration process to help searching for the MAP in a sequence of prior settings as well as determining the appropriate value for these penalty parameters. Initially, they fix λ_1 at a small value and gradually

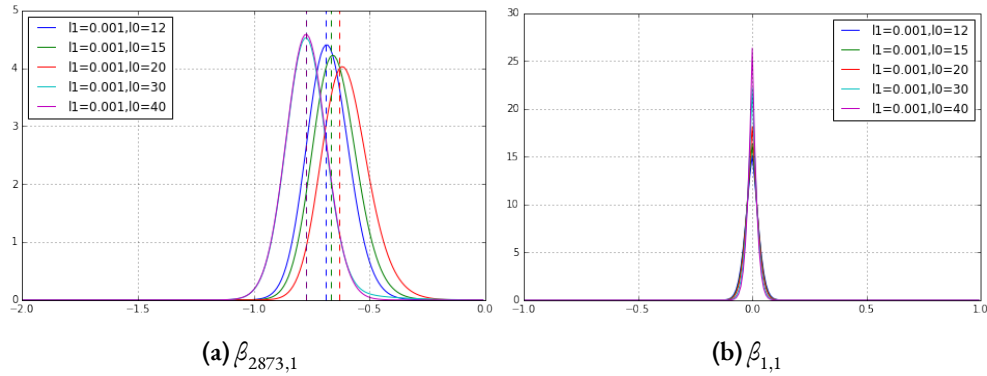


Figure 3.9: Posterior pdf of (a) $\beta_{2873,1}$ and (b) $\beta_{1,1}$ under SpSL-orthonormal factor model with increasing λ_0

increase λ_0 until the solution path is stabilized. The solution given by the PXL-EM under the final value of λ_0 approximates the MAP estimate under a flat and point mass mixture prior on loading matrix elements and is proposed as the estimator for parameters. The same procedure can be applied to the full posterior inference based on the SpSL-orthonormal factor model.

We observed a similar stabilization of the posterior distributions of every nonzero loading element when performing dynamic exploration for the SpSL-orthonormal factor model, which is illustrated in the application of our method to the cerebrum microarray data from AGEMAP (Atlas of Gene Expression in Mouse Aging Project) database of Zahn et al. (2007), which was analyzed by Ročková & George (2016) using their PXL-EM algorithm. For every mice individual in this dataset (5 males and 5 females, at four age periods), cerebrum microarray expression data from 8932 genes are recorded, observations $y_i, i = 1, \dots, 40$ for the factor model are taken to be the residuals of the expression values for each of the 8932 genes regressed on age and gender with an intercept.

We ran the posterior sampler initialized at the MAP detected by the PXL-EM algorithm with $\lambda_1 = 0.001, \alpha = 1/G$, and λ_0 gradually increasing in the sequence of 12,15,20,30,40. As the detected factor dimensionality by the PXL-EM algorithm is 1, we specify K to be 1 in our framework. Figure 3.9 demonstrates the evolution of the posterior density of $\beta_{2873,1}$ and $\beta_{1,1}$ as λ_0 changes.

The posterior distribution of $\beta_{1,1}$ centers at 0 and becomes more and more spiky as λ_0 increases.

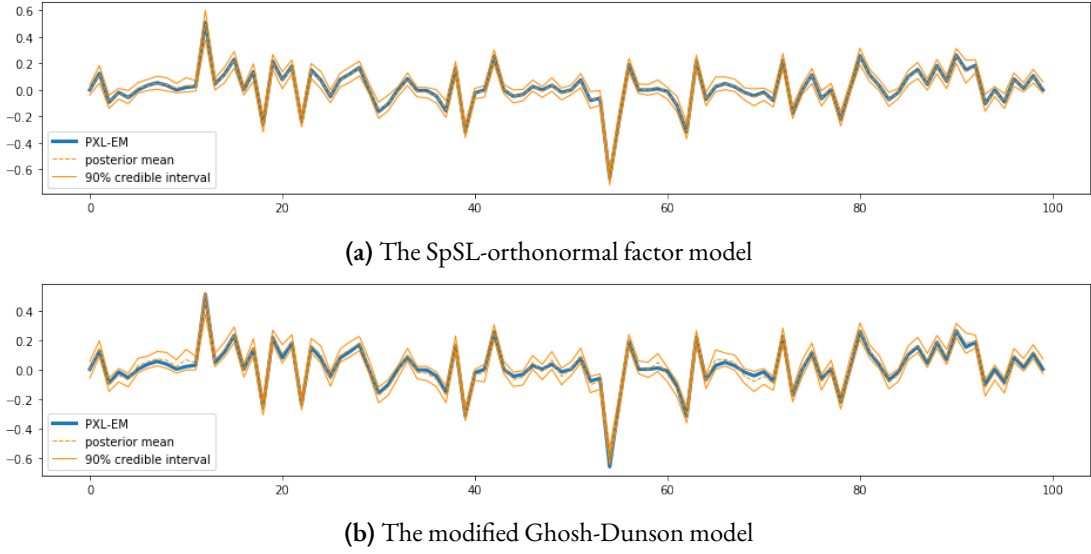


Figure 3.10: Posterior mean and credible interval of $\beta_{1,1}, \dots, \beta_{100,1}$ estimated from samples of specified model and the MAP estimate from PXL-EM algorithm

For the nonzero element $\beta_{2873,1}$, its posterior distributions resemble the normal distribution with a relative stable variance. The posterior mean of $\beta_{2873,1}$ first moves towards zero and then away and stabilizes. This change of direction is caused by the alteration of its slab indicator $\gamma_{2873,1}$ from 0 to 1 in posterior samples, in which case the posterior distribution of $\beta_{j,k}$ is only influenced by the slab, but not the spike prior. Vertical dotted lines are the MAP estimates, which are close to the posterior means. Having recognized that the stabilization of the MAP estimates and the posterior distributions occur almost simultaneously as λ_0 increases, in practice we can find the ideal pair of penalty parameters such that the posterior distribution is stabilized by looking for the stabilization of the MAP estimates instead of sampling from the posterior distribution with λ_0 on multiple levels. More summary and comparative figures of the posterior simulation are illustrated in Appendix C.4 with $\lambda_0 = 30$.

Figure 3.10 provides a comparison between the posterior results from the SpSL-orthonormal factor model ($\lambda_0 = 30$) and the modified Ghosh-Dunson model ($\lambda = 0.001, \lambda_0 = 200$) which shows that the two models give very similar posterior credible intervals (computed using 1000 posterior sam-

ples after burn-in) for the loading matrix, and both posterior means are also very close to the MAP estimator from PXL-EM algorithm. Additionally, the Gibbs sampler for the SpSL-orthonormal factor model results in a much larger ESS compared to that for the Ghosh-Dunson model (e.g., the ESS for $\beta_{55,1}$ are 905.0 and 42.7 for the two methods, respectively). We omit scientific interpretations of the inference results since our goal is only to verify that our procedure gives similar results as those in [Ročková & George \(2016\)](#) based on points estimates of the normal factor model, and to show how to conduct the full Bayesian analysis properly and efficiently for this dataset.

In summary, we can start our Bayesian inference for the SpSL-orthonormal factor model by first choosing a small λ_1 and a sequence of increasing λ_0 , denoted as $\{\lambda_0^{(t)}\}_{t=1,\dots}$. We then run the PXL-EM algorithm sequentially with λ_1 and $\lambda_0^{(t)}$ for $t = 1, \dots$, with parameters initialized at the MAP estimate found in the previous round. The process is terminated when the difference between the new MAP estimate and the one from the previous round is below a chosen threshold. Afterward, we run our Gibbs sampler under the SpSL-orthonormal factor model using the final pair of penalty parameters with \mathbf{B} , Σ , Θ and K initialized at the MAP estimate and Ω , Γ initialized with random draws from their domains.

3.8 DISCUSSION

A primary intention of our work is to provide an efficient posterior sampler for the Bayesian factor model in high dimensions and show its consistency. [Ročková & George \(2016\)](#)'s sparse Bayesian factor model framework serves as a promising starting point, for both its explicit encoding of the sparsity and its providing of a fast posterior mode finding algorithm. By analyzing the magnitude inflation problem of the posterior samples of the loading matrix under the prior setup of [Ročková & George \(2016\)](#), we propose the \sqrt{n} -orthonormal factor model as a practical remedy, which not only processes posterior consistency and robustness against prior settings, but also dramatically improves the

computational efficiency. Our work naturally bridges the gap between the point estimation based on posterior modes and the full Bayesian analysis under the SpSL factor modeling framework.

Besides our proposed solution, i.e., enforcing a common scale and orthogonality among the factors, [Bernardo et al. \(2003\)](#) and [Ghosh & Dunson \(2009\)](#) provided another perspective, which is to reduce the dimensionality of diffuse parameters in the prior to ensure that they do not overwhelm the data. Their approach allows the factors to have different variances, but restricts elements of the loading matrix to follow standard Gaussian *a priori*. In this chapter, we provide a further modification of their model by imposing a SpSL prior on the loading matrix's elements, which allows a greater flexibility in handling sparsity in high dimensions (details in Appendix C.2).

Using the prior from [Ročková & George \(2016\)](#), we are able to show theoretically that the adoption of a strict \sqrt{n} -orthonormal factor assumption can ensure posterior consistency. But this type of rigorous analysis for other models, including the Ghosh-Dunson model and its modification, still evades our vigorous attempts. Furthermore, in some follow up work, informative priors were assigned to the diffuse parameters in Ghosh-Dunson model, and it is unclear how these priors influence the posterior magnitude of the loading matrix generally. Interests for future exploration may be focused on the design of dependent priors for easy posterior sampling as well as the justification of posterior consistency when using such priors. The \sqrt{n} -orthonormal factor model itself is also interesting, since the posterior consistency under this model is empirically more robust against prior specification of the loading matrix in the high dimensional setting. It would be interesting to see a mathematical formulation of this empirical result in future works.



Supplemental Materials of Chapter 1

A.1 PROOFS

A.1.1 PROOF OF LEMMA 1.6.1

By definition of the multi-log(p) term, it suffices to show that, for every $\varepsilon > 0$, as $p \rightarrow \infty$,

$$p^{-\varepsilon+b} \mathbb{P}(X_p \in S) \rightarrow 0, \quad \text{and} \quad p^{\varepsilon+b} \mathbb{P}(X_p \in S) \rightarrow \infty. \quad (\text{A.1})$$

We introduce two sets \underline{S} and \bar{S} such that

$$\underline{S} \subset S \subset \bar{S}.$$

Define $m(x) = (x - \mu)' \Sigma^{-1} (x - \mu)$ for any $x \in \mathbb{R}^d$. By definition, $b = \inf_{x \in S} m(x)$. As a result, $m(x) \geq b$ for all $x \in S$. Define

$$\bar{S} = \{x \in \mathbb{R}^d : m(x) \geq b\}. \quad (\text{A.2})$$

Then, $S \subset \bar{S}$. Furthermore, since $m(x)$ is a quadratic function and $b = \inf_{x \in S} m(x)$, given any $\varepsilon > 0$, there exists $x_0 \in S$ such that

$$m(x_0) \leq b + \varepsilon/8. \quad (\text{A.3})$$

Note that (A.3) guarantees that $\|x_0 - \mu\|$ is bounded. For any $x \in S$ and $\|x - x_0\| \leq 1$,

$$\begin{aligned} |m(x) - m(x_0)| &\leq 2|(x - \mu)' \Sigma^{-1} (x - x_0)| + |(x - x_0)' \Sigma^{-1} (x - x_0)| \\ &\leq 2\|x - \mu\| \|\Sigma^{-1}\| \cdot \|x - x_0\| + \|\Sigma^{-1}\| \|x - x_0\|^2 \\ &\leq C_1 \|x - x_0\| + C_2 \|x - x_0\|^2, \end{aligned}$$

where C_1 and C_2 are positive constants that only depend on $(\mu, \Sigma, b, \varepsilon)$. It follows that there exists a constant $\delta_1 > 0$ such that

$$x \in S, \quad \|x - x_0\| \leq \delta_1 \quad \implies \quad |m(x) - m(x_0)| \leq \varepsilon/8. \quad (\text{A.4})$$

Additionally, since S is an open set and $x_0 \in S$, there exists $\delta_2 > 0$, such that

$$\{x \in \mathbb{R}^d : \|x - x_0\| \leq \delta_2\} \subset S.$$

Define

$$\underline{S} = \{x \in \mathbb{R}^d : \|x - x_0\| \leq \delta\}, \quad \text{where } \delta = \min\{\delta_1, \delta_2\}. \quad (\text{A.5})$$

It is easy to see that $\underline{S} \subset S$. Additionally, in light of (A.3) and (A.4),

$$m(x) \leq b + \varepsilon/4, \quad \text{for all } x \in \underline{S}. \quad (\text{A.6})$$

Since $\underline{S} \subset S \subset \bar{S}$, to show (A.1), it suffices to show that

$$p^{\varepsilon+b} \mathbb{P}(X_p \in \underline{S}) \rightarrow \infty \quad (\text{A.7})$$

and

$$p^{-\varepsilon+b} \mathbb{P}(X_p \in \bar{S}) \rightarrow 0. \quad (\text{A.8})$$

First, we show (A.7). Let $f_p(x)$ denote the density of $\mathcal{N}_d(\mu_p, \frac{1}{2\log(p)}\Sigma_p)$. Write $m_p(x) = (x - \mu_p)' \Sigma_p^{-1} (x - \mu_p)$. It is seen that

$$f_p(x) = \frac{[2\log(p)]^{d/2}}{(2\pi)^{d/2} |\det(\Sigma_p)|^{1/2}} \cdot p^{-m_p(x)}. \quad (\text{A.9})$$

By direct calculations,

$$\begin{aligned} \mathbb{P}(X_p \in \underline{S} \mid \mu_p, \Sigma_p) &= \frac{[2\log(p)]^{d/2}}{(2\pi)^{d/2} |\det(\Sigma_p)|^{1/2}} \int_{x \in \underline{S}} p^{-m_p(x)} dx \\ &\geq \frac{[\log(p)]^{d/2}}{\pi^{d/2} |\det(\Sigma_p)|^{1/2}} \cdot \text{Volume}(\underline{S}) \cdot p^{-\sup_{x \in \underline{S}} \{m_p(x)\}}. \end{aligned} \quad (\text{A.10})$$

The assumptions on (μ_p, Σ_p) imply that, for any constant $\gamma > 0$,

$$\lim_{p \rightarrow \infty} \mathbb{P}(\|\mu_p - \mu\| > \gamma \text{ or } \|\Sigma_p - \Sigma\| > \gamma) = 0.$$

Let E be the event that $\|\mu_p - \mu\| \leq \gamma_*$ and $\|\Sigma_p - \Sigma\| \leq \gamma_*$, for some γ_* to be decided. On this event, for any $x \in \underline{\mathcal{S}}$,

$$\begin{aligned}
|m(x) - m_p(x)| &\leq |(x - \mu)' \Sigma^{-1} (x - \mu) - (x - \mu)' \Sigma_p^{-1} (x - \mu)| \\
&\quad + |(x - \mu)' \Sigma_p^{-1} (x - \mu) - (x - \mu_p)' \Sigma_p^{-1} (x - \mu_p)| \\
&\leq |(x - \mu)' (\Sigma^{-1} - \Sigma_p^{-1}) (x - \mu)| + 2|(x - \mu)' \Sigma_p^{-1} (\mu - \mu_p)| \\
&\quad + (\mu - \mu_p)' \Sigma_p^{-1} (\mu - \mu_p) \\
&\leq \|x - \mu\|^2 \|\Sigma^{-1} - \Sigma_p^{-1}\| \cdot \|\Sigma_p - \Sigma\| + 2\|x - \mu\| \|\Sigma_p^{-1}\| \cdot \|\mu - \mu_p\| \\
&\quad + \|\Sigma_p^{-1}\| \cdot \|\mu - \mu_p\|^2 \\
&\leq C_3 \gamma_* + C_4 \gamma_*^2,
\end{aligned}$$

where C_3 and C_4 are positive constants that do not depend on γ_* , and in the last line we have used the fact that $\underline{\mathcal{S}}$ is a bounded set so that $\|x - \mu\|$ is bounded. It follows that we can choose an appropriately small γ_* such that

$$|m(x) - m_p(x)| \leq \varepsilon/4, \quad \text{for all } x \in \underline{\mathcal{S}}. \quad (\text{A.11})$$

Combining (A.11) with (A.6) gives

$$\sup_{x \in \underline{\mathcal{S}}} m_p(x) \leq b + \varepsilon/2, \quad \text{on the event } E.$$

Moreover, since $\underline{\mathcal{S}}$ is a ball with radius δ ,

$$\text{Volume}(\underline{\mathcal{S}}) = \delta^d \cdot \text{Volume}(B_d),$$

where B_d is the unit ball in \mathbb{R}^d , whose volume is a constant. We plug the above results into (A.10) and notice that $|\det(\Sigma_p)| \geq |\det(\Sigma)| - C_5\delta$ on the event E , for a constant $C_5 > 0$. It yields that, when (μ_p, Σ_p) satisfies the event E ,

$$\mathbb{P}(X_p \in \underline{\mathcal{S}} \mid \mu_p, \Sigma_p) \geq c_0 [\log(p)]^{d/2} \cdot p^{-(b+\varepsilon/2)}, \quad (\text{A.12})$$

for some constant $c_0 > 0$. It follows that

$$\mathbb{P}(X_p \in \underline{\mathcal{S}}) \geq \mathbb{P}(E) \cdot c_0 [\log(p)]^{d/2} p^{-(b+\varepsilon/2)}.$$

We plug it into the left hand side of (A.7) and note that $\mathbb{P}(E) \rightarrow 1$ as $p \rightarrow \infty$. This gives the desirable claim in (A.7).

Next, we show (A.8). We define a counterpart of the set $\bar{\mathcal{S}}$ by

$$\bar{\mathcal{S}}_p = \{x \in \mathbb{R}^d : m_p(x) \geq b\}.$$

Define $Y_p = \sqrt{2 \log(p)} \cdot \Sigma_p^{-1/2} (X_p - \mu_p)$. Then, $Y_p \sim \mathcal{N}_d(0, I_d)$ and

$$X_p \in \bar{\mathcal{S}}_p \quad \text{if and only if} \quad \|Y_p\|^2 \geq 2b \log(p).$$

The distribution of $\|Y_p\|^2$ is a χ_d^2 distribution, which does not depend on (μ_p, Σ_p) . We have

$$\begin{aligned} \mathbb{P}(X_p \in \bar{\mathcal{S}}_p) &= \mathbb{E}[\mathbb{P}(X_p \in \bar{\mathcal{S}}_p \mid \mu_p, \Sigma_p)] \\ &= \mathbb{E}[\mathbb{P}(\|Y_p\|^2 \geq 2b \log(p))] \\ &= \mathbb{P}(\chi_d^2 \geq 2b \log(p)). \end{aligned} \quad (\text{A.13})$$

For chi-square distribution, the tail probability has an explicit form:

$$\mathbb{P}(\chi_d^2 \geq 2b \log(p)) = \frac{\Gamma(d/2, b \log(p))}{\Gamma(d/2)},$$

where $\Gamma(s, x) \equiv \int_x^\infty t^{s-1} \exp(-t) dt$ is the upper incomplete gamma function and $\Gamma(s) \equiv \Gamma(s, 0)$ is the ordinary gamma function. By property of the upper incomplete gamma function,

$$\Gamma(s, x)/(x^{s-1} \exp(-x)) \rightarrow 1, \quad \text{as } x \rightarrow \infty.$$

It follows that

$$\frac{\Gamma(d/2, b \log(p))}{[b \log(p)]^{d/2-1} p^{-b}} \rightarrow 1, \quad \text{as } p \rightarrow \infty.$$

In particular, when p is sufficiently large, the left hand side is $\geq 1/2$. We plug these results into (A.13) to get

$$\mathbb{P}(X_p \in \bar{S}_p) \geq \frac{[b \log(p)]^{d/2-1}}{2\Gamma(d/2)} \cdot p^{-b}. \quad (\text{A.14})$$

It remains to study the difference caused by replacing \bar{S}_p by \bar{S} . Let

$$U_p = (\bar{S} \setminus \bar{S}_p) \cup (\bar{S}_p \setminus \bar{S}).$$

Then,

$$|\mathbb{P}(X_p \in \bar{S}) - \mathbb{P}(X_p \in \bar{S}_p)| \leq \mathbb{P}(X_p \in U_p). \quad (\text{A.15})$$

Similar to (A.10), we have

$$\begin{aligned} \mathbb{P}(X_p \in U_p \mid \mu_p, \Sigma_p) &= \frac{[2 \log(p)]^{d/2}}{(2\pi)^{d/2} |\det(\Sigma_p)|^{1/2}} \int_{x \in U_p} p^{-m_p(x)} dx \\ &\leq \frac{[\log(p)]^{d/2}}{\pi^{d/2} |\det(\Sigma_p)|^{1/2}} \cdot \text{Volume}(U_p) \cdot p^{-\inf_{x \in U_p} \{m_p(x)\}}. \end{aligned} \quad (\text{A.16})$$

For a constant $\gamma > 0$ to be decided, let F be the event that

$$\|\mu_p - \mu\| \leq \gamma, \quad \text{and} \quad \|\Sigma_p - \Sigma\| \leq \gamma. \quad (\text{A.17})$$

On this event, we study both $\text{Volume}(U_p)$ and $\inf_{x \in U_p} m_p(x)$. Re-write

$$U_p = (\bar{\mathcal{S}}^c \setminus \bar{\mathcal{S}}_p^c) \cup (\bar{\mathcal{S}}_p^c \setminus \bar{\mathcal{S}}^c).$$

By definition, $\bar{\mathcal{S}}^c = \{x \in \mathbb{R}^d : m(x) \leq b\} = \{x \in \mathbb{R}^d : \|\Sigma^{-1/2}(x - \mu)\| \leq \sqrt{b}\}$, and $\bar{\mathcal{S}}_p^c = \{x \in \mathbb{R}^d : \|\Sigma_p^{-1/2}(x - \mu_p)\| \leq \sqrt{b}\}$. On the event F , for any $x \in \bar{\mathcal{S}}_p^c$,

$$\begin{aligned} \|\Sigma^{-1/2}(x - \mu)\| &\leq \sqrt{b} + \|\Sigma^{-1/2}(x - \mu) - \Sigma_p^{-1/2}(x - \mu_p)\| \\ &\leq \sqrt{b} + \|\Sigma^{-1/2}(\mu_p - \mu)\| + \|(\Sigma^{-1/2} - \Sigma_p^{-1/2})(x - \mu_p)\| \\ &\leq \sqrt{b} + \|\Sigma^{-1/2}\| \cdot \|\mu_p - \mu\| + \|\Sigma^{1/2}\Sigma_p^{-1/2} - I_d\| \cdot \|\Sigma_p^{-1/2}(x - \mu_p)\| \\ &\leq \sqrt{b} + \|\Sigma^{-1/2}\| \cdot \|\mu_p - \mu\| + \sqrt{b} \cdot \|\Sigma^{1/2}\Sigma_p^{-1/2} - I_d\| \\ &\leq \sqrt{b} + C_5\gamma, \end{aligned}$$

for a constant $C_5 > 0$ that does not depend on γ . Choosing $\gamma < C_5^{-1}\sqrt{b}$, we have $\|\Sigma^{-1/2}(x - \mu)\| \leq 2\sqrt{b}$ for all $x \in \bar{\mathcal{S}}_p^c$. Additionally, by definition, $\|\Sigma^{-1/2}(x - \mu)\| \leq \sqrt{b}$ for all $x \in \bar{\mathcal{S}}^c$. Combining the above gives

$$U_p \subset (\bar{\mathcal{S}}^c \cup \bar{\mathcal{S}}_p^c) \subset \{x \in \mathbb{R}^d : \|\Sigma^{-1/2}(x - \mu)\| \leq 2\sqrt{b}\}.$$

Recall that B_d is the unit ball in \mathbb{R}^d . It follows immediately that

$$\text{Volume}(U_p) \leq (2\sqrt{b})^d \cdot \text{Volume}(B_d), \quad \text{on the event } F. \quad (\text{A.18})$$

At the same time, for any $x \in \bar{S}$, on the event F ,

$$\begin{aligned}
\|\Sigma_p^{-1/2}(x - \mu_p)\| &\geq \|\Sigma^{-1/2}(x - \mu)\| - \|\Sigma_p^{-1/2}(x - \mu_p) - \Sigma^{-1/2}(x - \mu)\| \\
&\geq \|\Sigma^{-1/2}(x - \mu)\| - \|\Sigma_p^{-1/2}(\mu_p - \mu)\| - \|(\Sigma^{-1/2} - \Sigma_p^{-1/2})(x - \mu)\| \\
&\geq \|\Sigma^{-1/2}(x - \mu)\| - \|\Sigma_p^{-1/2}\| \cdot \|\mu_p - \mu\| - \|\Sigma_p^{-1/2}\Sigma^{1/2} - I_d\| \cdot \|\Sigma^{-1/2}(x - \mu)\| \\
&= \|\Sigma^{-1/2}(x - \mu)\|(1 - \|\Sigma_p^{-1/2}\Sigma^{1/2} - I_d\|) - \|\Sigma_p^{-1/2}\| \cdot \|\mu_p - \mu\| \\
&\geq \|\Sigma^{-1/2}(x - \mu)\|(1 - C_6\gamma) - \|\Sigma^{-1/2}\|\gamma \\
&\geq \sqrt{b}(1 - C_6\gamma) - \|\Sigma^{-1/2}\|\gamma,
\end{aligned}$$

where $C_6 > 0$ is a constant that does not depend on γ and in the last line we have used the fact that $\|\Sigma^{-1/2}(x - \mu)\| \geq \sqrt{b}$ for $x \in \bar{S}$. We choose γ properly small so that $\sqrt{b}(1 - C_6\gamma) - \|\Sigma^{-1/2}\|\gamma \geq \sqrt{b - \varepsilon/2}$. It follows that

$$m_p(x) = \|\Sigma_p^{-1/2}(x - \mu_p)\|^2 \geq b - \varepsilon/2, \quad \text{for all } x \in \bar{S}. \quad (\text{A.19})$$

Additionally, the definition of \bar{S}_p already guarantees that $m_p(x) \geq b$ for all $x \in \bar{S}_p$. Consequently,

$$\inf_{x \in U_p} m_p(x) \geq \inf_{x \in \bar{S} \cup \bar{S}_p} \{m_p(x)\} \geq b - \varepsilon/2, \quad \text{on the event } F. \quad (\text{A.20})$$

We plug (A.18) and (A.20) into (A.16). It yields that, on the event F ,

$$\mathbb{P}(X_p \in U_p \mid \mu_p, \Sigma_p) \leq C_7[\log(p)]^{d/2} \cdot p^{-(b-\varepsilon/2)}, \quad (\text{A.21})$$

for a constant $C_7 > 0$. Then,

$$\mathbb{P}(X_p \in U_p) \leq \mathbb{P}(F) \cdot C_7[\log(p)]^{d/2} \cdot p^{-(b-\varepsilon/2)} + \mathbb{P}(F^c).$$

By our assumption, for any $\gamma > 0$ and $L > 0$, $\mathbb{P}(\|\mu_p - \mu\| > \gamma) \leq p^{-L}$ and $\mathbb{P}(\|\Sigma_p - \Sigma\| > \gamma) \leq p^{-L}$.

In particular, we can choose $L = b$. It gives

$$\mathbb{P}(F^c) \leq p^{-b}.$$

We combine the above results and plug them into (A.15). It follows that

$$|\mathbb{P}(X_p \in \bar{S}) - \mathbb{P}(X_p \in \bar{S}_p)| \leq C_7 [\log(p)]^{d/2} \cdot p^{-(b-\varepsilon/2)} + p^{-b}. \quad (\text{A.22})$$

Combining (A.14) and (A.22) gives

$$\mathbb{P}(X_p/ \in \bar{S}) \leq [1 + o(1)] \cdot C_7 [\log(p)]^{d/2} \cdot p^{-(b-\varepsilon/2)}.$$

This gives the claim in (A.8). The proof of this lemma is complete. \square

A.1.2 PROOF OF LEMMA 1.6.2

First, we study the least-squares. Note that $\hat{\beta}$ has an explicit solution: $\hat{\beta} = G^{-1}X^T y$. Since G is a block-wise diagonal matrix, we immediately have

$$\begin{bmatrix} \hat{\beta}_j \\ \hat{\beta}_{j+1} \end{bmatrix} = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}^{-1} \begin{bmatrix} x_j^T y \\ x_{j+1}^T y \end{bmatrix} = \frac{1}{1-\rho^2} \begin{bmatrix} x_j^T y - \rho x_{j+1}^T y \\ x_{j+1}^T y - \rho x_j^T y \end{bmatrix}.$$

Recall that $\tilde{y} = X^T y / \sqrt{2 \log(p)}$. Then, $|\hat{\beta}_j| > \sqrt{2u \log(p)}$ if and only if

$$\frac{1}{1-\rho^2} |\tilde{y}_j - \rho \tilde{y}_{j+1}| > \sqrt{u}.$$

It immediately gives the rejection region for least-squares.

Next, we study the Lasso-path. The lasso estimate $\hat{\beta}(\lambda)$ minimizes the objective

$$Q(b) = \frac{1}{2}\|y - Xb\|^2 + \lambda\|b\|_1 = \frac{1}{2}\|y\|^2 - y^T Xb + \frac{1}{2}b^T Gb + \lambda\|b\|_1.$$

When G is a block-wise diagonal matrix, the objective $Q(b)$ is separable, and we can optimize over each pair of (b_j, b_{j+1}) separately. It reduces to solving many bi-variate problems:

$$(\hat{\beta}_j(\lambda), \hat{\beta}_{j+1}(\lambda))^T = \operatorname{argmin}_b \left\{ \frac{1}{2}\|y - [x_j, x_{j+1}]b\|_2^2 + \lambda\|b\|_1 \right\}. \quad (\text{A.23})$$

Write $\hat{b} = (\hat{\beta}_j(\lambda), \hat{\beta}_{j+1}(\lambda))^T$ and let

$$B = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \quad \text{and} \quad b = \begin{bmatrix} x_j^T y \\ x_{j+1}^T y \end{bmatrix}.$$

Then, the optimization (A.23) can be written as

$$\hat{b} = \operatorname{argmin}_b \left\{ -b^T b + b^T B b / 2 + \lambda\|b\|_1 \right\}. \quad (\text{A.24})$$

Recall that W_j^* is the value of λ at which \hat{b}_1 becomes nonzero for the first time. Our goal is to find a region of (b_1, b_2) such that $W_j^* > t_p(u) \equiv \sqrt{2u \log(p)}$.

It suffices to consider the case of $\rho \geq 0$. To see this, we consider changing ρ to $-\rho$ in the matrix B . The objective remains unchanged if we also change b_1 to $-b_1$ and b_2 to $-b_2$. Note that the change of b_1 to $-b_1$ has no impact on W_j^* ; this means W_j^* is unchanged if we simultaneously flip the sign of ρ and b_1 . Consequently, once we know the rejection region for $\rho > 0$, we can immediately obtain that for $\rho < 0$ by a reflection of the region with respect to the x-axis.

Below, we fix $\rho \geq 0$. We first derive the explicit form of the whole solution path and then use it

to decide the rejection region. Taking sub-gradients of (A.23), we find that \hat{b} has to satisfy

$$\begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \begin{bmatrix} \hat{b}_1 \\ \hat{b}_2 \end{bmatrix} + \lambda \begin{bmatrix} \text{sgn}(\hat{b}_1) \\ \text{sgn}(\hat{b}_2) \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}, \quad (\text{A.25})$$

where $\text{sgn}(x) = 1$ if $x > 0$, $\text{sgn}(x) = -1$ if $x < 0$, and $\text{sgn}(x)$ can be equal to any value in $[-1, 1]$ if $x = 0$. Let $\lambda_1 > \lambda_2 > 0$ be the values at which variables enter the solution path. When $\lambda \in (\lambda_1, \infty)$, $\hat{b}_1 = 0$ and $\hat{b}_2 = 0$. Plugging them into (A.25) gives $\text{sgn}(\hat{b}_1) = \lambda^{-1}b_1$. The definition of $\text{sgn}(\hat{b}_1)$ implies that $|b_1| \leq \lambda$, for any $\lambda > \lambda_1$. We then have $|b_1| \leq \lambda_1$. Similarly, it is true that $|b_2| \leq \lambda_1$. It gives

$$\lambda_1 = \max\{|b_1|, |b_2|\}. \quad (\text{A.26})$$

We first assume $|b_1| > |b_2|$. By (A.25) and continuity of solution path, there exists a sufficiently small constant $\delta > 0$ such that, for $\lambda \in (\lambda_2 - \delta, \lambda_2)$, the following equation holds.

$$\begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \begin{bmatrix} \hat{b}_1(\lambda) \\ \hat{b}_2(\lambda) \end{bmatrix} + \lambda \begin{bmatrix} \text{sgn}(\hat{b}_1) \\ \text{sgn}(\hat{b}_2) \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}. \quad (\text{A.27})$$

The sign vector of \hat{b} for $\lambda \in (\lambda_2 - \delta, \lambda_2)$ has four different cases: $(1, 1)^T$, $(1, -1)^T$, $(-1, 1)^T$, $(-1, -1)^T$. For these four different cases, we can use (A.27) to solve \hat{b} . The solutions in four cases are respectively

$$\begin{aligned} & \frac{1}{1-\rho^2} \begin{bmatrix} (b_1 - \rho b_2) - (1-\rho)\lambda \\ (b_2 - \rho b_1) - (1-\rho)\lambda \end{bmatrix}, & \frac{1}{1-\rho^2} \begin{bmatrix} (b_1 - \rho b_2) - (1+\rho)\lambda \\ (b_2 - \rho b_1) + (1+\rho)\lambda \end{bmatrix}, \\ & \frac{1}{1-\rho^2} \begin{bmatrix} (b_1 - \rho b_2) + (1+\rho)\lambda \\ (b_2 - \rho b_1) - (1+\rho)\lambda \end{bmatrix}, & \frac{1}{1-\rho^2} \begin{bmatrix} (b_1 - \rho b_2) + (1-\rho)\lambda \\ (b_2 - \rho b_1) + (1-\rho)\lambda \end{bmatrix}. \end{aligned}$$

The solution \hat{b} has to match the sign assumption on \hat{b} . For each of the four cases, the requirement

becomes

- Case 1: $(b_1 - \rho b_2) - (1 - \rho)\lambda > 0$, $(b_2 - \rho b_1) - (1 - \rho)\lambda > 0$.
- Case 2: $(b_1 - \rho b_2) - (1 + \rho)\lambda > 0$, $(b_2 - \rho b_1) + (1 + \rho)\lambda < 0$.
- Case 3: $(b_1 - \rho b_2) + (1 + \rho)\lambda < 0$, $(b_2 - \rho b_1) - (1 + \rho)\lambda > 0$.
- Case 4: $(b_1 - \rho b_2) + (1 - \rho)\lambda < 0$, $(b_2 - \rho b_1) + (1 - \rho)\lambda < 0$.

Note that we have assumed $|b_1| > |b_2|$. Then, Case k is possible only in the region \mathcal{A}_k , where

$$\begin{aligned} \mathcal{A}_1 &= \{(b_1, b_2) : b_1 > 0, \rho b_1 < b_2 < b_1\}, & \mathcal{A}_2 &= \{(b_1, b_2) : b_1 > 0, -b_1 < b_2 < \rho b_1\}, \\ \mathcal{A}_3 &= \{(b_1, b_2) : b_1 < 0, \rho b_1 < b_2 < -b_1\}, & \mathcal{A}_4 &= \{(b_1, b_2) : b_1 < 0, b_1 < b_2 < \rho b_1\}. \end{aligned}$$

In each case, $\lambda_1 = |b_1|$. To get the value of λ_2 , we use the continuity of the solution path. It implies that $\hat{b}_2(\lambda) = 0$ at $\lambda = \lambda_2$. As a result, the value of λ_2 in Case k is

$$\lambda_2^{(1)} = \frac{b_2 - \rho b_1}{1 - \rho}, \quad \lambda_2^{(2)} = \frac{\rho b_1 - b_2}{1 + \rho}, \quad \lambda_2^{(3)} = \frac{b_2 - \rho b_1}{1 + \rho}, \quad \lambda_2^{(4)} = \frac{\rho b_1 - b_2}{1 - \rho}. \quad (\text{A.28})$$

It is easy to verify that $\lambda_2 < \lambda_1$ in each case. We also need to check that in the region \mathcal{A}_k , the KKT condition (A.25) can be satisfied with $\hat{b}_2 = 0$ for all $\lambda \in (\lambda_2^{(k)}, \lambda_1)$. For example, in Case 1, (A.25)

becomes

$$\begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \begin{bmatrix} \hat{b}_1 \\ 0 \end{bmatrix} + \lambda \begin{bmatrix} 1 \\ c \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}, \quad \text{for some } |c| \leq 1.$$

We can solve the equations to get $\hat{b}_1 = b_1 - \lambda$ and $\lambda c = b_2 - \rho \hat{b}_1 = (b_2 - \rho b_1) - \lambda$. It can be verified that $|(b_2 - \rho b_1) - \lambda| \leq \lambda$ for $(b_1, b_2) \in \mathcal{A}_1$ and $\lambda \in (\lambda_2^{(1)}, \lambda_1)$. The verification for other cases is similar and thus omitted. We then assume $|b_2| > |b_1|$. By symmetry, we will have the same result,

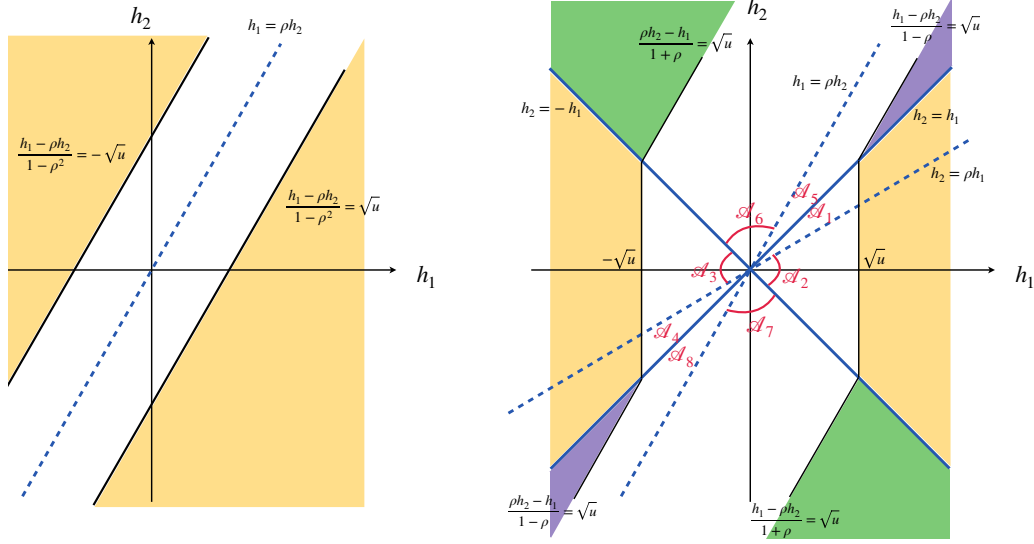


Figure A.1: The rejection region of least-squares (left) and Lasso-path (right). On the right panel, the regions \mathcal{A}_1 - \mathcal{A}_8 are the same as those defined in the proof. In the regions \mathcal{A}_1 - \mathcal{A}_4 , $M_j^* = |b_1|$, and the rejection region is colored by yellow. In the regions \mathcal{A}_5 and \mathcal{A}_8 , $M_j^* = |b_1 - \rho b_2|/(1 - \rho)$, and the rejection region is colored by purple. In the regions \mathcal{A}_6 and \mathcal{A}_7 , $M_j^* = |b_1 - \rho b_2|/(1 + \rho)$, and the rejection region is colored by green.

except that (b_1, b_2) are switched in the expression of \mathcal{A} and (λ_1, λ_2) . This gives the other four cases:

$$\begin{aligned} \mathcal{A}_5 &= \{(b_1, b_2) : b_2 > 0, \rho b_2 < b_1 < b_2\}, & \mathcal{A}_6 &= \{(b_1, b_2) : b_2 > 0, -b_2 < b_1 < \rho b_2\}, \\ \mathcal{A}_7 &= \{(b_1, b_2) : b_2 < 0, \rho b_2 < b_1 < -b_2\}, & \mathcal{A}_8 &= \{(b_1, b_2) : b_2 < 0, b_2 < b_1 < \rho b_2\}. \end{aligned}$$

In these four cases, we similarly have $\lambda_1 = |b_2|$ and

$$\lambda_2^{(5)} = \frac{b_1 - \rho b_2}{1 - \rho}, \quad \lambda_2^{(6)} = \frac{\rho b_2 - b_1}{1 + \rho}, \quad \lambda_2^{(7)} = \frac{b_1 - \rho b_2}{1 + \rho}, \quad \lambda_2^{(8)} = \frac{\rho b_2 - b_1}{1 - \rho}. \quad (\text{A.29})$$

These eight regions are shown in Figure A.1.

We then compute M_j^* and the associated rejection region. Note that $M_j^* = \lambda_1$ in Case 1-Case 4,

and $M_j^* = \lambda_2$ in Case 5-Case 8. It follows directly that

$$M_j^* = \begin{cases} |b_1|, & \text{if } (b_1, b_2) \in \mathcal{A}_1 \cup \mathcal{A}_2 \cup \mathcal{A}_3 \cup \mathcal{A}_4, \\ |b_1 - \rho b_2|/(1 - \rho), & \text{if } (b_1, b_2) \in \mathcal{A}_5 \cup \mathcal{A}_8, \\ |b_1 - \rho b_2|/(1 + \rho), & \text{if } (b_1, b_2) \in \mathcal{A}_6 \cup \mathcal{A}_7. \end{cases} \quad (\text{A.30})$$

As a result, the region $M_j^* > \sqrt{2u \log(p)}$ if and only if the vector $(x_j^T y, x_{j+1}^T y)/\sqrt{2 \log(p)}$ is in the following set:

$$\begin{aligned} \mathcal{R} = & [(\mathcal{A}_1 \cup \mathcal{A}_2 \cup \mathcal{A}_3 \cup \mathcal{A}_4) \cap \{|b_1| > \sqrt{u}\}] \\ & \cup [(\mathcal{A}_5 \cup \mathcal{A}_8) \cap \{|b_1 - \rho b_2| > (1 - \rho)\sqrt{u}\}] \\ & \cap [(\mathcal{A}_6 \cup \mathcal{A}_7) \cap \{|b_1 - \rho b_2| > (1 + \rho)\sqrt{u}\}]. \end{aligned}$$

In Figure A.1, the 3 subsets are colored by yellow, purple, and green, respectively. This gives the rejection region for Lasso-path. \square

A.1.3 PROOF OF THEOREM 1.3.1

By definition of $(\text{FP}_p, \text{FN}_p)$ and the Rare/Weak signal model (1.4)-(1.5), we have

$$\text{FP}_p = \sum_{j=1}^p (1 - \varepsilon_p) \mathbb{P}(W_j > t_p(u) | \beta_j = 0), \quad \text{FN}_p = \sum_{j=1}^p \varepsilon_p \mathbb{P}(W_j < t_p(u) | \beta_j = \tau_p), \quad (\text{A.31})$$

where $\varepsilon_p = p^{-\delta}$, $\tau_p = \sqrt{2r \log(p)}$, and $t_p(u) = \sqrt{2u \log(p)}$. Therefore, it suffices to study $\mathbb{P}(W_j > t_p(u) | \beta_j = 0)$ and $\mathbb{P}(W_j < t_p(u) | \beta_j = \tau_p)$.

Fix $1 \leq j \leq p$. The knockoff filter applies Lasso to the design matrix $[X, \tilde{X}]$. This design belongs to the block-wise diagonal design (1.17) with a dimension $2p$ and $\rho = a$. The variable j and its own

knockoff are in one block. Fix j and write

$$b_1 = x'_j y / \sqrt{2 \log(p)}, \quad \text{and} \quad b_2 = \tilde{x}'_j y / \sqrt{2 \log(p)}. \quad (\text{A.32})$$

It is easy to see that $(x'_j y, \tilde{x}'_j y)'$ follows a distribution $\mathcal{N}_2(0_2, \Sigma)$ when $\beta_j = 0$, and it follows a distribution $\mathcal{N}_2(\mu \sqrt{2 \log(p)}, \Sigma)$, when $\beta_j = \tau_p$, where

$$\mu = \begin{bmatrix} \sqrt{r} \\ a \sqrt{r} \end{bmatrix}, \quad \Sigma = \begin{bmatrix} 1 & a \\ a & 1 \end{bmatrix}.$$

Let \mathcal{R} be the region of (b_1, b_2) corresponding to the event that $\{W_j > t_p(u)\}$. It follows from Lemma 1.6.1 that

$$\begin{aligned} \mathbb{P}(W_j > t_p(u) | \beta_j = 0) &= L_p p^{-\inf_{b \in \mathcal{R}} \{b' \Sigma^{-1} b\}}, \\ \mathbb{P}(W_j < t_p(u) | \beta_j = \tau_p) &= L_p p^{-\inf_{b \in \mathcal{R}^c} \{(b - \mu)' \Sigma^{-1} (b - \mu)\}}. \end{aligned} \quad (\text{A.33})$$

Below, we first derive the rejection region \mathcal{R} , and then compute the exponents in (A.33).

Recall that Z_j and \tilde{Z}_j are the same as in (1.14). They are indeed the values of λ at which the variable j and its knockoff enter the solution path of a bivariate lasso as in (A.23). We can apply the solution path derived in the proof of Lemma 1.6.2, with $\rho = a$. Before we proceed to the proof, we argue that it suffices to consider the case of $a \geq 0$. If $a < 0$, we can simultaneously flip the signs of a and b_2 , so that the objective (A.23) remains unchanged; as a result, the values of (Z_j, \tilde{Z}_j) remain unchanged, so does the symmetric statistic W_j . It implies that, if we flip the sign of a , the rejection region is reflected with respect to the x-axis. At the same time, in light of the exponents in (A.33), we consider two ellipsoids

$$\mathcal{E}_{\text{FP}}(t) = \{b \in \mathbb{R}^2 : b' \Sigma^{-1} b \leq t\}, \quad \mathcal{E}_{\text{FN}}(t) = \{b \in \mathbb{R}^2 : (b - \mu)' \Sigma^{-1} (b - \mu)' \leq t\}. \quad (\text{A.34})$$

Similarly, if we simultaneously flip the signs of a and b_2 , these ellipsoids remain unchanged. It implies that, if we flip the sign of a , these ellipsoids are reflected with respect to the x -axis. Combining the above observations, we know that the exponents in (A.33) are unchanged with a sign flip of a , i.e., they only depend on $|a|$. We assume $a \geq 0$ without loss of generality.

Fix $a \geq 0$. Write $z = Z_j/\sqrt{2\log(p)}$ and $\tilde{z} = \tilde{Z}_j/\sqrt{2\log(p)}$. The symmetric statistics in (1.14) can be re-written as

$$W_j^{\text{sgm}} = (z \vee \tilde{z})\sqrt{2\log(p)} \cdot \begin{cases} +1, & \text{if } z > \tilde{z} \\ -1, & \text{if } z \leq \tilde{z} \end{cases}, \quad W_j^{\text{dif}} = (z - \tilde{z})\sqrt{2\log(p)}.$$

Recall that b_1 and b_2 are as in (A.32). Let $\lambda_1 > \lambda_2 > 0$ be the values of λ at which variables enter the solution path of a bivariate lasso. In the proof of Lemma 1.6.2, we have derived the formula of (λ_1, λ_2) ; see (A.28) and (A.29) (with ρ replaced by a). It follows that

$$(z, \tilde{z}) = \begin{cases} (\lambda_1, \lambda_2), & \text{in the regions } \mathcal{A}_1\text{-}\mathcal{A}_4, \\ (\lambda_2, \lambda_1), & \text{in the regions } \mathcal{A}_5\text{-}\mathcal{A}_8, \end{cases}$$

where regions $\mathcal{A}_1\text{-}\mathcal{A}_8$ are the same as those on the right panel of Figure A.1 (with ρ replaced by a). Plugging in (A.28) and (A.29) gives the following results:

- Region \mathcal{A}_1 : $z = b_1$, $\tilde{z} = \frac{b_2 - \rho b_1}{1 - a}$, $W_j^{\text{sgm}} = b_1\sqrt{2\log(p)}$, $W_j^{\text{dif}} = \frac{b_1 - b_2}{1 - a}\sqrt{2\log(p)}$.
- Region \mathcal{A}_2 : $z = b_1$, $\tilde{z} = \frac{\rho b_1 - b_2}{1 + a}$, $W_j^{\text{sgm}} = b_1\sqrt{2\log(p)}$, $W_j^{\text{dif}} = \frac{b_1 + b_2}{1 + a}\sqrt{2\log(p)}$.
- Region \mathcal{A}_3 : $z = -b_1$, $\tilde{z} = \frac{b_2 - \rho b_1}{1 + a}$, $W_j^{\text{sgm}} = -b_1\sqrt{2\log(p)}$, $W_j^{\text{dif}} = -\frac{b_1 + b_2}{1 + a}\sqrt{2\log(p)}$.
- Region \mathcal{A}_4 : $z = -b_1$, $\tilde{z} = \frac{\rho b_1 - b_2}{1 - a}$, $W_j^{\text{sgm}} = -b_1\sqrt{2\log(p)}$, $W_j^{\text{dif}} = \frac{b_2 - b_1}{1 - a}\sqrt{2\log(p)}$.
- Regions $\mathcal{A}_5\text{-}\mathcal{A}_8$: $|Z_j| < |\tilde{Z}_j|$, $W_j^{\text{sgm}} < 0$, $W_j^{\text{dif}} < 0$.

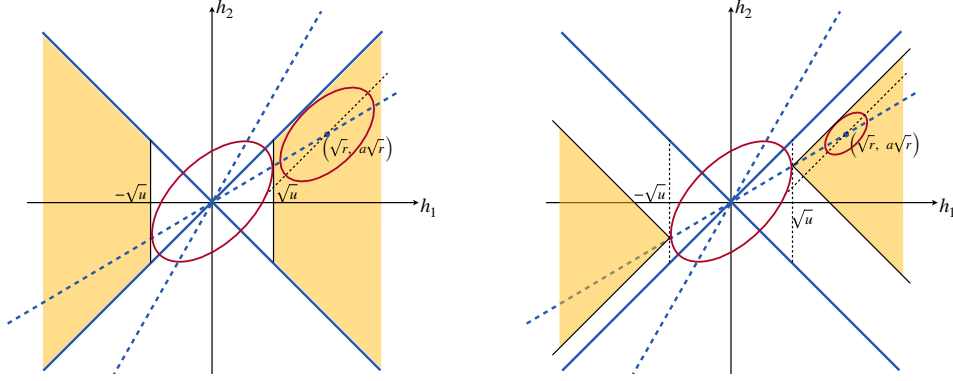


Figure A.2: The rejection region of knockoff in the orthogonal design, where the symmetric statistic is signed maximum (left) and difference (right). The rate of convergence of FP_p is captured by an ellipsoid centered at $(0, 0)$, and the rate of convergence of FN_p is captured by an ellipsoid centered at $(\sqrt{r}, a\sqrt{r})$.

The event that $W_j^{\text{sgm}} > \sqrt{2u \log(p)}$ corresponds to that (b_1, b_2) is in the region of

$$\begin{aligned} \mathcal{R}_u^{\text{sgm}} &= (\mathcal{A}_1 \cup \mathcal{A}_2 \cup \mathcal{A}_3 \cup \mathcal{A}_4) \cap \{|b_1| > \sqrt{u}\} \\ &= \{|b_1| > |b_2|, |b_1| > \sqrt{u}\}. \end{aligned} \quad (\text{A.35})$$

The event that $W_j^{\text{dif}} > \sqrt{2u \log(p)}$ corresponds to that (b_1, b_2) is in the region of

$$\begin{aligned} \mathcal{R}_u^{\text{dif}} &= (\mathcal{A}_1 \cap \{b_1 - b_2 > (1-a)\sqrt{u}\}) \cup (\mathcal{A}_2 \cap \{b_1 + b_2 > (1+a)\sqrt{u}\}) \\ &\quad \cup (\mathcal{A}_3 \cap \{b_1 + b_2 < -(1+a)\sqrt{u}\}) \cup (\mathcal{A}_4 \cap \{b_1 - b_2 < -(1-a)\sqrt{u}\}). \end{aligned} \quad (\text{A.36})$$

These two regions are shown in Figure A.2.

We are now ready to compute the exponents in (A.33). First, we compute $\inf_{b \in \mathcal{R}} \{b' \Sigma^{-1} b\}$. Let $\mathcal{E}_{\text{FP}}(t)$ be the same as in (A.34). Then,

$$\inf_{b \in \mathcal{R}} \{b' \Sigma^{-1} b\} = \sup \{t > 0 : \mathcal{E}_{\text{FP}}(t) \cap \mathcal{R} \neq \emptyset\}.$$

When the rejection region is $\mathcal{R}_u^{\text{sgm}}$, from Figure A.2, we can increase t until $\mathcal{E}_{\text{FP}}(t)$ intersects with the line of $b_1 = \pm\sqrt{u}$. For any b on the surface of this ellipsoid, the perpendicular vector of its tangent plane is proportional to $\Sigma^{-1}b$. When the ellipsoid intersects with the line of $b_1 = \pm\sqrt{u}$, the perpendicular vector should be proportional to $(1, 0)'$. Therefore, we need to find b such that

$$b_1 = \pm\sqrt{u}, \quad b'\Sigma^{-1}b = t, \quad \text{and} \quad \Sigma^{-1}b \propto (1, 0)'.$$

The second equation requires that $b_2 = ab_1$. Combining it with the first equation gives $b = (\pm\sqrt{u}, \pm a\sqrt{u})$. We then plug it into the second equation to obtain $t = u$. This gives

$$\inf_{b \in \mathcal{R}_u^{\text{sgm}}} \{b'\Sigma^{-1}b\} = u. \quad (\text{A.37})$$

When the rejection region is $\mathcal{R}_u^{\text{dif}}$, there are 3 possible cases:

- (i) The ellipsoid intersects with the line $b_1 - b_2 = (1 - a)\sqrt{u}$,
- (ii) The ellipsoid intersects with the line $b_1 + b_2 = (1 + a)\sqrt{u}$,
- (iii) The ellipsoid intersects with the point $b = (\sqrt{u}, a\sqrt{u})$.

In Case (i), we can compute the intersection point by solving b for $b_1 - b_2 = (1 - a)\sqrt{u}$ and $\Sigma^{-1}b \propto (1, -1)'$. The second relationship gives $b_2 = -b_1$. Together with the first relationship, we have $b = (\frac{1-a}{2}\sqrt{u}, \frac{1-a}{2}\sqrt{u})$. It is not in $\mathcal{R}_u^{\text{dif}}$. Similarly, for Case (ii), we can show that the intersection point is $b = (\frac{1+a}{2}\sqrt{u}, \frac{1+a}{2}\sqrt{u})$, which is not in $\mathcal{R}_u^{\text{dif}}$ either. The only possible case is Case (iii), where the intersection point is $(\sqrt{u}, a\sqrt{u})$ and the associated $t = b'\Sigma^{-1}b = u$. We have proved that

$$\inf_{b \in \mathcal{R}_u^{\text{dif}}} \{b'\Sigma^{-1}b\} = u. \quad (\text{A.38})$$

Next, we compute $\inf_{b \in \mathcal{R}^c} \{(b - \mu)' \Sigma^{-1} (b - \mu)\}$. Let $\mathcal{E}_{\text{FN}}(t)$ be the same as in (A.34). Then,

$$\inf_{b \in \mathcal{R}^c} \{(b - \mu)' \Sigma^{-1} (b - \mu)\} = \sup \{t > 0 : \mathcal{E}_{\text{FN}}(t) \cap \mathcal{R}^c \neq \emptyset\}.$$

Note that the center of the ellipsoid is $\mu = (\sqrt{r}, a\sqrt{r})$. When either $\mathcal{R} = \mathcal{R}_u^{\text{sgm}}$ or $\mathcal{R} = \mathcal{R}_u^{\text{dif}}, \mu \notin \mathcal{R}^c$ if and only if $r > u$. In other words, the above is well defined only if $r > u$. We now fix $r > u$. When the rejection region is $\mathcal{R}_u^{\text{sgm}}$, the ellipsoid intersects with either the line of $b_1 = \sqrt{u}$ or the line of $b_1 = b_2$. Since the perpendicular vector of the tangent plane of the ellipsoid at b is proportional to $\Sigma^{-1}(b - \mu)$, we can solve the intersection points from

$$\begin{cases} b_1 = \sqrt{u}, \\ \Sigma^{-1}(b - \mu) \propto (1, 0)', \end{cases} \quad \text{and} \quad \begin{cases} b_1 = b_2, \\ \Sigma^{-1}(b - \mu) \propto (1, -1)'. \end{cases}$$

By calculations, the two intersection points are $b = (\sqrt{u}, a\sqrt{u})$ and $b = (\frac{1+a}{2}\sqrt{r}, \frac{1+a}{2}\sqrt{r})$. The associated value of $(b - \mu)' \Sigma^{-1} (b - \mu)$ is $t = (\sqrt{r} - \sqrt{u})^2$ and $t = (1 - a)r/2$, respectively. When we increase the ellipsoid until it interacts with $(\mathcal{R}_u^{\text{sgm}})^c$, the corresponding t is the smaller of the above two values. This gives

$$\inf_{b \in (\mathcal{R}_u^{\text{sgm}})^c} \{(b - \mu)' \Sigma^{-1} (b - \mu)\} = \min \left\{ (\sqrt{r} - \sqrt{u})_+^2, \frac{1-a}{2}r \right\}. \quad (\text{A.39})$$

When the rejection region is $\mathcal{R}_u^{\text{dif}}$, the ellipsoid intersects with either the line of $b_1 - b_2 = (1 - a)\sqrt{u}$ or the line of $b_1 + b_2 = (1 + a)\sqrt{u}$. We can solve the intersection points from

$$\begin{cases} b_1 - b_2 = (1 - a)\sqrt{u}, \\ \Sigma^{-1}(b - \mu) \propto (1, -1)', \end{cases} \quad \text{and} \quad \begin{cases} b_1 + b_2 = (1 + a)\sqrt{u}, \\ \Sigma^{-1}(b - \mu) \propto (1, 1)'. \end{cases}$$

Solving these equations gives the two intersection points: $b = (\frac{1+a}{2}\sqrt{r} + \frac{1-a}{2}\sqrt{u}, \frac{1+a}{2}\sqrt{r} - \frac{1-a}{2}\sqrt{u})$ and $b = (\frac{1-a}{2}\sqrt{r} + \frac{1+a}{2}\sqrt{u}, -\frac{1-a}{2}\sqrt{r} + \frac{1+a}{2}\sqrt{u})$. The corresponding value of $(b - \mu)' \Sigma^{-1} (b - \mu)$ is $t = \frac{1-a}{2}(\sqrt{r} - \sqrt{u})^2$ and $t = \frac{1+a}{2}(\sqrt{r} - \sqrt{u})^2$, respectively. The smaller of these two values is $\frac{1-a}{2}(\sqrt{r} - \sqrt{u})^2$. We have proved that

$$\inf_{b \in (\mathcal{R}_u^{\text{dif}})^c} \{(b - \mu)' \Sigma^{-1} (b - \mu)\} = \frac{1-a}{2}(\sqrt{r} - \sqrt{u})_+^2. \quad (\text{A.40})$$

We plug (A.37)-(A.40) into (A.33), and we further plug it into (A.31). This gives the claim for $a \geq 0$. As we have argued, the results for $a < 0$ only requires replacing a by $|a|$. \square

A.1.4 PROOF OF THEOREM 1.3.2

This theorem is a special case of Theorem 1.5.1. The proof can be found there. \square

A.1.5 PROOF OF THEOREM 1.4.1

The least-squares estimator satisfies that $\hat{\beta} \sim \mathcal{N}_p(\beta, G^{-1})$. It gives $\hat{\beta}_j \sim \mathcal{N}(\beta_j, \omega_j)$. Applying Lemma 1.6.1 to $X_p = \hat{\beta}_j$ and $S = \{x \in \mathbb{R} : x \geq \sqrt{u}\}$, we have

$$\mathbb{P}(|\hat{\beta}_j| > t_p(u) | \beta_j = 0) = L_p p^{-\omega_j^{-1}u}, \quad \mathbb{P}(|\hat{\beta}_j| \leq t_p(u) | \beta_j = \tau_p) = L_p p^{-\omega_j^{-1}(\sqrt{r}-\sqrt{u})_+^2}.$$

It follows that

$$\begin{aligned} \text{FP}_p(u) &= \sum_{j=1}^p (1 - \varepsilon_p) \cdot \mathbb{P}(\mathcal{M}_j^* > t_p(u) | \beta_j = 0) = L_p \sum_{j=1}^p p^{-\omega_j^{-1}u}, \\ \text{FN}_p(u) &= \sum_{j=1}^p \varepsilon_p \cdot \mathbb{P}(\mathcal{M}_j^* < t_p(u) | \beta_j = \tau_p) = L_p p^{-s} \sum_{j=1}^p p^{-\omega_j^{-1}(\sqrt{r}-\sqrt{u})_+^2}. \end{aligned}$$

For the block-wise diagonal design (1.17), $\omega_j = (1 - \rho^2)^{-1}$ for all $1 \leq j \leq p - 1$. \square

A.1.6 PROOF OF THEOREM 1.4.2

Without loss of generality, we assume p is even. Then, for block-wise diagonal designs as in (1.17), the Lasso objective is separable. Therefore, for each W_j^* , it is not affected by any β_k outside the block. Additionally, by symmetry, the distribution of W_j^* is the same for all $1 \leq j \leq p$. It follows that

$$\begin{aligned} \text{FP}_p(u) &= L_p p \cdot \mathbb{P}\{W_j^* > t_p(u) \mid (\beta_j, \beta_{j+1}) = (0, 0)\} \\ &\quad + L_p p^{1-s} \cdot \mathbb{P}\{W_j^* > t_p(u) \mid (\beta_j, \beta_{j+1}) = (0, \tau_p)\}, \end{aligned} \quad (\text{A.41})$$

where j can be odd index. Similarly, we can derive that

$$\begin{aligned} \text{FN}_p(u) &= L_p p^{1-s} \cdot \mathbb{P}\{W_j^* < t_p(u) \mid (\beta_j, \beta_{j+1}) = (\tau_p, 0)\} \\ &\quad + L_p p^{1-2s} \cdot \mathbb{P}\{W_j^* < t_p(u) \mid (\beta_j, \beta_{j+1}) = (\tau_p, \tau_p)\}. \end{aligned} \quad (\text{A.42})$$

Fix variables $\{j, j+1\}$, and consider the random vector $\hat{b} = (x'_j y, x'_{j+1} y)' / \sqrt{\log(p)}$. Then,

$$\hat{b} \sim \mathcal{N}_2\left(\mu, \frac{1}{\log(p)} \Sigma\right), \quad \text{where } \Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}.$$

The vector μ is equal to

$$\mu^{(1)} \equiv \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad \mu^{(2)} \equiv \begin{bmatrix} \rho \sqrt{r} \\ \sqrt{r} \end{bmatrix}, \quad \mu^{(3)} \equiv \begin{bmatrix} \sqrt{r} \\ \rho \sqrt{r} \end{bmatrix}, \quad \mu^{(4)} \equiv \begin{bmatrix} (1+\rho)\sqrt{r} \\ (1+\rho)\sqrt{r} \end{bmatrix}, \quad (\text{A.43})$$

in the four cases where $(\beta_j, \beta_{j+1})'$ is $(0, 0)'$, $(0, \tau_p)'$, $(\tau_p, 0)'$, and $(\tau_p, \tau_p)'$, respectively. Let \mathcal{R}_u be the rejection region induced by Lasso-path, given explicitly in Lemma 1.6.2. By Lemma 1.6.1, the

probabilities in (A.41) and (A.42) are related to the following quantities:

$$\alpha_k = \begin{cases} \inf \inf_{b \in \mathcal{R}_u} \{(b - \mu^{(k)})' \Sigma^{-1} (b - \mu^{(k)})\}, & k = 1, 2, \\ \inf_{b \in \mathcal{R}_u^c} \{(b - \mu^{(k)})' \Sigma^{-1} (b - \mu^{(k)})\}, & k = 3, 4. \end{cases}$$

and plug it into (A.41) and (A.42). It gives

$$\text{FP}_\rho(u) = L_p p^{1 - \min\{\alpha_1, \vartheta + \alpha_2\}}, \quad \text{FN}_\rho(u) = L_p p^{1 - \min\{\vartheta + \alpha_3, 2\vartheta + \alpha_4\}}. \quad (\text{A.44})$$

It remains to compute the exponents α_1 - α_4 .

First, we consider the case that $\rho \geq 0$. The rejection region in Figure A.1 is defined by the following lines:

- Line 1: $b_1 - \rho b_2 = (1 - \rho)\sqrt{u}$.
- Line 2: $b_1 = \sqrt{u}$.
- Line 3: $b_1 - \rho b_2 = (1 + \rho)\sqrt{u}$.
- Line 4: $b_1 - \rho b_2 = -(1 - \rho)\sqrt{u}$.
- Line 5: $b_1 = -\sqrt{u}$.
- Line 6: $b_1 - \rho b_2 = -(1 + \rho)\sqrt{u}$.

Consider a general ellipsoid:

$$\mathcal{E}(t; \mu) = \{b \in \mathbb{R}^2 : (b - \mu)' \Sigma^{-1} (b - \mu) \leq t\}.$$

Given any line $b_1 + \rho b_2 = c$, as t increases, this ellipsoid eventually intersects with this line. The

intersection point is computed by the following equations:

$$b_1 + bb_2 = c, \quad \Sigma^{-1}(b - \mu) \propto (1, b)'$$

The second equation (it is indeed a linear equation on b) says that the perpendicular vector of the tangent plane is orthogonal to the line. Solving the above equations gives the intersection point and the value of t : As long as $b^2 \neq 1$, we have

$$b^* = \mu + \frac{c - (\mu_1 + b\mu_2)}{1 + b^2 + 2b\rho} \begin{bmatrix} 1 + b\rho \\ b + \rho \end{bmatrix}, \quad t^* = \frac{[c - (\mu_1 + b\mu_2)]^2}{1 + b^2 + 2b\rho}. \quad (\text{A.45})$$

Using the expressions of lines 1-6, we can obtain the corresponding t^* for 6 lines:

$$\begin{aligned} t_1^* &= \frac{[(1 - \rho)\sqrt{u} - (\mu_1 - \rho\mu_2)]^2}{1 - \rho^2}, & t_2^* &= (\sqrt{u} - \mu_1)^2, & t_3^* &= \frac{[(1 + \rho)\sqrt{u} - (\mu_1 - \rho\mu_2)]^2}{1 - \rho^2}, \\ t_4^* &= \frac{[(1 - \rho)\sqrt{u} + (\mu_1 - \rho\mu_2)]^2}{1 - \rho^2}, & t_5^* &= (\sqrt{u} + \mu_1)^2, & t_6^* &= \frac{[(1 + \rho)\sqrt{u} + (\mu_1 - \rho\mu_2)]^2}{1 - \rho^2}. \end{aligned}$$

We first look at the ellipsoid $\mathcal{E}(t; \mu^{(1)})$ and study when it intersects with \mathcal{R}_μ . Note that $\mu^{(1)} = (0, 0)'$.

The above t^* values become

$$t_2^* = t_5^* = u, \quad t_1^* = t_4^* = \frac{u}{1 + \rho}, \quad t_3^* = t_6^* = \frac{u}{1 - \rho}.$$

Therefore, as we increase t , this ellipsoid first intersects with line 1 and line 4. For line 1, the intersection point is $((1 - \rho)\sqrt{u}, 0)'$, but it is outside the rejection region (see Figure A.1); the situation for line 4 is similar. We then further increase t , and the ellipsoid intersects with line 2 and line 5, where the intersection point is $(\sqrt{u}, \rho\sqrt{u})'$; this point is indeed on the boundary of the rejection region. We

thus conclude that

$$\inf_{b \in \mathcal{R}_u} \{(b - \mu^{(1)})' \Sigma^{-1} (b - \mu^{(1)})\} = u. \quad (\text{A.46})$$

We then look at the the ellipsoid $\mathcal{E}(t; \mu^{(2)})$, with $\mu^{(2)} = (\rho\sqrt{r}, \sqrt{r})'$. The t^* values for 6 lines are:

$$t_1^* = t_4^* = \frac{1-\rho}{1+\rho}u, \quad t_2^* = (\sqrt{u} - \rho\sqrt{r})^2, \quad t_3^* = t_6^* = \frac{1+\rho}{1-\rho}u, \quad t_5^* = (\sqrt{u} + \rho\sqrt{r})^2.$$

The smallest t^* is among $\{t_1^*, t_2^*, t_4^*\}$. Since $\mu^{(2)}$ is in the positive orthant, the intersection point of the ellipsoid with line 4 must be outside the rejection region, so we further restrict to t_1^* and t_2^* . The ellipsoid intersects with line 1 at $(\rho\sqrt{r} + (1-\rho)\sqrt{u}, \sqrt{r})'$. This point is on the boundary of \mathcal{R}_u if and only if its second coordinate is $\geq \sqrt{u}$ (see Figure A.1), i.e., $u \leq r$. The ellipsoid intersects with line 2 at $(\sqrt{u}, \rho\sqrt{u} + (1-\rho^2)\sqrt{r})'$. This point is on the boundary of \mathcal{R}_u if and only if its second coordinate is $\leq \sqrt{u}$ (see Figure A.1), i.e., $u \geq (1+\rho)^2 r$. In the range of $r < u < (1+\rho)^2 r$, the ellipsoid intersects with \mathcal{R}_u at the corner point $(\sqrt{u}, \sqrt{u})'$, with the corresponding

$$t^* = r + \frac{2}{1+\rho}u - 2\sqrt{ru} = \begin{cases} \frac{1-\rho}{1+\rho}u + (\sqrt{u} - \sqrt{r})^2, \\ (\sqrt{u} - \rho\sqrt{r})^2 + \frac{1-\rho}{1+\rho}(\sqrt{u} - (1+\rho)\sqrt{r})^2. \end{cases}$$

This t^* has two equivalent expressions. Comparing them with t_1^* and t_2^* , we can see that the smallest t^* is a continuous function of u , given (ρ, r) . It follows that

$$\begin{aligned} & \inf_{b \in \mathcal{R}_u} \{(b - \mu^{(2)})' \Sigma^{-1} (b - \mu^{(2)})\} \\ &= \frac{1-\rho}{1+\rho}u + (\sqrt{u} - \sqrt{r})_+^2 - \frac{1-\rho}{1+\rho}(\sqrt{u} - (1+\rho)\sqrt{r})_+^2. \end{aligned} \quad (\text{A.47})$$

We plug (A.46) and (A.47) into (A.44). It gives the expression of $\text{FP}_p(u)$ for $\rho \geq 0$.

We then look at the ellipsoid $\mathcal{E}(t; \mu^{(3)})$, with $\mu^{(3)} = (\sqrt{r}, \rho\sqrt{r})'$. Note that we now investigate

its distance to the complement of \mathcal{R}_u . In order for $\mu^{(3)}$ to be outside \mathcal{R}_u^c (i.e., in the interior of \mathcal{R}_u), we require that $u < r$; furthermore, when $u < r$, the ellipsoid can only intersect with lines 1-2 (see Figure A.1). Using the formula of t^* in the equation below (A.45), we have

$$t_1^* = \frac{1-\rho}{1+\rho}((1+\rho)\sqrt{r} - \sqrt{u})^2, \quad t_2^* = (\sqrt{r} - \sqrt{u})^2.$$

By (A.45), the ellipsoid intersects with line 1 at $(\sqrt{r} - (1-\rho)[(1+\rho)\sqrt{r} - \sqrt{u}], \rho\sqrt{r})'$. To guarantee that this point is on the boundary of \mathcal{R}_u , we need its second coordinate to be $\geq \sqrt{u}$ (see Figure A.1), i.e., $u \leq \rho^2 r$; furthermore, when $u > \rho^2 r$, it can be easily seen from Figure A.1 that the ellipsoid must have already crossed line 2. By (A.45) again, the ellipsoid intersects with line 2 at $(\sqrt{u}, \rho\sqrt{u})'$. This point is always on the boundary of \mathcal{R}_u . It follows that

$$\inf_{b \in \mathcal{R}_u^c} \{(b - \mu^{(3)})' \Sigma^{-1} (b - \mu^{(3)})\} = \min \left\{ \frac{1-\rho}{1+\rho}((1+\rho)\sqrt{r} - \sqrt{u})^2, (\sqrt{r} - \sqrt{u})_+^2 \right\}. \quad (\text{A.48})$$

We then look at the ellipsoid $\mathcal{E}(t; \mu^{(4)})$, with $\mu^{(4)} = ((1+\rho)\sqrt{r}, (1+\rho)\sqrt{r})'$. It follows from Figure A.1 that $\mu^{(4)}$ is in the interior of the ellipsoid if and only if $(1+\rho)\sqrt{r} > \sqrt{u}$. We restrict to $(1+\rho)\sqrt{r} > \sqrt{u}$. Then, this ellipsoid can only touch lines 1-2 first. The t^* values are

$$t_1^* = \frac{1-\rho}{1+\rho}((1+\rho)\sqrt{r} - \sqrt{u})^2, \quad t_2^* = ((1+\rho)\sqrt{r} - \sqrt{u})^2.$$

Since $t_1^* < t_2^*$, the ellipsoid touches line 1 first, at the intersection point $((1-\rho)\sqrt{u} + \rho(1+\rho)\sqrt{r}, (1+\rho)\sqrt{r})'$. In order for this point to be on the boundary of \mathcal{R}_u , we need that its second coordinate is $\geq \sqrt{u}$, which translates to $\sqrt{u} \leq (1+\rho)\sqrt{r}$. This is always true when $r > u$ and $\rho > 0$. It follows that

$$\inf_{b \in \mathcal{R}_u^c} \{(b - \mu^{(4)})' \Sigma^{-1} (b - \mu^{(4)})\} = \frac{1-\rho}{1+\rho}((1+\rho)\sqrt{r} - \sqrt{u})_+^2. \quad (\text{A.49})$$

We plug (A.48) and (A.49) into (A.44). It gives the expression of $\text{FN}_p(u)$ for $\rho \geq 0$.

Next, we consider the case that $\rho < 0$. By Lemma 1.6.2, $\mathcal{R}_u(\rho)$ is a reflection of $\mathcal{R}_u(|\rho|)$ with respect to the x-axis. As a result, if we re-define $\hat{b} = (x'_j y, -x'_{j+1} y) / \sqrt{2 \log(p)}$, then the rejection region becomes $\mathcal{R}_u(|\rho|)$, which has the same shape as that in Figure A.1. At the same time, the distribution of \hat{b} becomes

$$\hat{b} \sim \mathcal{N}_2\left(\mu, \frac{1}{\log(p)} \Sigma\right), \quad \text{where } \Sigma = \begin{bmatrix} 1 & |\rho| \\ |\rho| & 1 \end{bmatrix}.$$

The vector μ is equal to

$$\mu^{(1)} \equiv \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad \mu^{(2)} \equiv \begin{bmatrix} -|\rho| \sqrt{r} \\ -\sqrt{r} \end{bmatrix}, \quad \mu^{(3)} \equiv \begin{bmatrix} \sqrt{r} \\ |\rho| \sqrt{r} \end{bmatrix}, \quad \mu^{(4)} \equiv \begin{bmatrix} (1 - |\rho|) \sqrt{r} \\ -(1 - |\rho|) \sqrt{r} \end{bmatrix}, \quad (\text{A.50})$$

when $(\beta_j, \beta_{j+1})'$ is $(0, 0)'$, $(0, \tau_p)'$, $(\tau_p, 0)'$, and $(\tau_p, \tau_p)'$, respectively. Therefore, the calculations are similar, except that the expressions of $\mu^{(1)}$ to $\mu^{(4)}$ have changed to (A.50).

Below, for a negative ρ , we calculate the exponents in (A.44) as follows: We pretend that $\rho > 0$ and calculate the exponents using the same \mathcal{R}_u and Σ as before, with $\mu^{(1)}$ to $\mu^{(4)}$ replaced by those in (A.50). Finally, we replace ρ by $|\rho|$ in all four exponents.

We now pretend that $\rho > 0$. Then, for each ellipsoid $\mathcal{E}(t; \mu^{(k)})$, its intersection point with a line $b_1 + b b_2 = c$ still obeys the formula in (A.45), and the corresponding t_* values associated with line 1-line 6 are still the same as those in the equation below (A.45) (but the vector μ has changed). Comparing (A.50) with (A.43), we notice that $\mu^{(1)}$ and $\mu^{(3)}$ are unchanged. Therefore, the expressions of exponents in (A.46) and (A.48) are still correct. The current $\mu^{(2)}$ is a sign flip (on both x-axis and y-axis) of the $\mu^{(2)}$ in (A.43); also, it can be seen from Figure A.1 that the rejection region remains unchanged subject to a sign flip. Therefore, the expression in (A.48) is also valid. We only need to re-calculate the exponent in (A.49). The current $\mu^{(4)}$ is in the 4-th orthant. It is in the interior of \mathcal{R}_u

only if $(1-\rho)\sqrt{r} > \sqrt{u}$, i.e., $u < (1-\rho)^2 r$. As we increase t , the ellipsoid $\mathcal{E}(t; \mu^{(4)})$ will first intersect with either line 2 or line 3. Using the formula of t^* in the equation below (A.45), we have

$$t_2^* = (\sqrt{u} - (1-\rho)\sqrt{r})^2, \quad t_3^* = \frac{1+\rho}{1-\rho} ((1-\rho)\sqrt{r} - \sqrt{u})^2.$$

While t_2^* is the smaller one, the intersection point of the ellipsoid with line 2 is $(\sqrt{u}, -(1-\rho)\sqrt{r})'$, which by Figure A.1 is in the interior of \mathcal{R}_u . Hence, the ellipsoid hits line 3 first. We conclude that

$$\inf_{b \in \mathcal{R}_u^c} \{(b - \mu^{(4)})' \Sigma^{-1} (b - \mu^{(4)})\} = \frac{1+\rho}{1-\rho} ((1-\rho)\sqrt{r} - \sqrt{u})_+^2. \quad (\text{A.51})$$

Finally, we plug (A.46), (A.47), (A.48) and (A.51) into (A.44), and then change ρ to $|\rho|$. This gives the expressions of $\text{FP}_\rho(u)$ and $\text{FN}_\rho(u)$ for a negative ρ . \square

A.1.7 PROOF OF THEOREM 1.5.1

By elementary properties of the least-squares estimator, $\hat{\beta}_j^\pm$ depends on β only through β_j . We thus have a decomposition of $\text{FP}_\rho(u)$ and $\text{FN}_\rho(u)$ similarly as in (A.31). It suffices to study $\mathbb{P}(M_j > t_p(u) | \beta_j = 0)$ and $\mathbb{P}(M_j < t_p(u) | \beta_j = \tau_p)$.

The statistic M_j is a function of $\hat{\beta}_j^\pm$, where $\hat{\beta}_j^\pm$ are the least-squares coefficients of x_j^\pm by regressing y on $\tilde{X}^{(j)} = [x_1, \dots, x_{j-1}, x_j^+, x_j^-, x_{j+1}, \dots, x_p]$, with $x_j^\pm = x_j \pm c_j z_j$. According to [Xing et al. \(2019\)](#), c_j is chosen such that

$$c_j = \sqrt{\frac{x_j'(I_n - P)x_j}{z_j'(I_n - P)z_j}} = \frac{\|(I_n - P)x_j\|}{\|(I_n - P)z_j\|}, \quad \text{where } P = X_{-j}(X_{-j}'X_{-j})^{-1}X_{-j}'. \quad (\text{A.52})$$

Using \tilde{x}_j^\pm , we can re-express the model of y as

$$y = \sum_{1 \leq k \leq p: k \neq j} \beta_k x_k + \frac{\beta_j}{2} \tilde{x}_j^+ + \frac{\beta_j}{2} \tilde{x}_j^- + \mathcal{N}(0, I_n).$$

Therefore, conditioning on (X, z_j) , $(\hat{\beta}_j^+, \hat{\beta}_j^-)'$ follows a bivariate normal distribution, whose mean vector is $(\beta_j/2, \beta_j/2)'$ and whose covariance matrix is a 2×2 block of \tilde{G}^{-1} , where $\tilde{G} = (\tilde{X}^{(j)})'(\tilde{X}^{(j)})$.

Recall that $G = X'X$. Write $\eta = z_j'X_{-j}$. By direct calculations,

$$\tilde{G} = \begin{bmatrix} \|x_j + c_j z_j\|^2 & \|x_j\|^2 - c_j^2 \|z_j\|^2 & G_{j,-j} + c_j \eta' \\ \|x_j\|^2 - c_j^2 \|z_j\|^2 & \|x_j - c_j z_j\|^2 & G_{j,-j} - c_j \eta' \\ \hline G_{-j,j} + c_j \eta & G_{-j,j} - c_j \eta & G_{-j,-j} \end{bmatrix} \equiv \begin{bmatrix} M & A \\ A' & G_{-j,-j} \end{bmatrix},$$

where we have re-arranged the order so that x_j^\pm are the first two variables. The matrix inversion formula implies that the top left 2×2 block of \tilde{G}^{-1} is equal to $(M - A'G_{-j,-j}^{-1}A)^{-1}$. By direct calculations,

$$M - A'G_{-j,-j}^{-1}A = \begin{bmatrix} \|v_j^+\|^2 & (v_j^+)'(v_j^-) \\ (v_j^-)'(v_j^+) & \|v_j^-\|^2 \end{bmatrix}, \quad \text{where } v_j^\pm = (I_n - P)(x_j \pm c_j z_j).$$

Write $x_j^* = (I_n - P)x_j$ and $z_j^* = (I_n - P)z_j$. The choice of c_j in (A.52) yields that $(v_j^+)'(v_j^-) = 0$ and that $v_j^\pm = \|x_j^*\| \cdot v_\pm^*$, where $v_\pm^* = (x_j^*/\|x_j^*\| \pm z_j^*/\|z_j^*\|)$. It follows that the top left 2×2 block of \tilde{G}^{-1} is

$$\begin{bmatrix} 1/(\|v_+^*\|^2 \|x_j^*\|^2) & \\ & 1/(\|v_-^*\|^2 \|x_j^*\|^2) \end{bmatrix}.$$

Using the definition of P in (A.52), we have $\|x_j^*\|^2 = x_j'(I_n - P)x_j = G_{jj} - G_{j,-j}G_{-j,-j}^{-1}G_{-j,j}$. Com-

binning it with the matrix inversion formula gives

$$\|x_j^*\|^2 = \omega_j^{-1}, \quad \text{where } \omega_j \text{ is the } j\text{-th diagonal of } G^{-1}. \quad (\text{A.53})$$

We have obtained that the distribution of $(\hat{\beta}_j^+, \hat{\beta}_j^-)'$ conditional on (X, z_j) is

$$\mathcal{N}_2\left(\left(\beta_j/2\right)\mathbf{1}_2, \Sigma_p\right), \quad \text{where } \Sigma_p = \omega_j \begin{bmatrix} \frac{1}{\|v_+^*\|^2} & \\ & \frac{1}{\|v_-^*\|^2} \end{bmatrix}. \quad (\text{A.54})$$

To apply Lemma 1.6.1, we further study Σ_p . Note that $\|v_\pm^*\|^2 = (x_j^*/\|x_j^*\| \pm z_j^*/\|z_j^*\|)^2 = 2 \pm 2\langle x_j^*/\|x_j^*\|, z_j^*/\|z_j^*\| \rangle$. Here $x_j^* = (I_n - P)x_j$ is a vector in the orthogonal space of the column space of X_{-j} , and z_j^* is the projection of $z_j \sim \mathcal{N}_n(0, I_n)$ into the same subspace. Since the distribution of z_j is spherically symmetric, we can assume that the orthogonal space of X_{-j} is spanned by the standard basis vectors $e_1, e_2, \dots, e_{n-p+1}$ and that $x_j^*/\|x_j^*\| = e_1$, without loss of generality. It follows that

$$\|v_\pm^*\|^2 \stackrel{(d)}{=} 2 \pm 2\xi_1/\|\xi\|, \quad \text{where } \xi \sim \mathcal{N}(0, I_{n-p+1}) \text{ and } \xi_1 \text{ is the first coordinate of } \xi.$$

Introduce $\Sigma = (\omega_j/2) \cdot I_2$. By direct calculations,

$$\|\Sigma_p - \Sigma\| \stackrel{(d)}{=} \frac{\omega_j}{2} \frac{|\xi_1|/\|\xi\|}{1 - |\xi_1|/\|\xi\|}. \quad (\text{A.55})$$

We aim to bound $\mathbb{P}(\|\Sigma_p - \Sigma\| > \gamma)$ for any $\gamma > 0$. Note that

$$\|\Sigma_p - \Sigma\| > \gamma \iff \frac{\xi_1^2}{\|\xi_{-1}\|^2} > \frac{4\omega_j^{-2}\gamma^2}{1 + 4\omega_j^{-1}\gamma} \equiv (\gamma^*)^2, \quad (\text{A.56})$$

where ξ_{-1} is the subvector of ξ excluding the first coordinate. Here $\xi_1 \sim \mathcal{N}(0, 1)$, $\|\xi_{-1}\|^2 \sim \chi_{n-p}^2$,

and they are independent of each other. Let E be the event that $\|\xi_{-1}\|^2 > (n-p)/2$.

$$\begin{aligned}
\mathbb{P}(|\xi_1|/\|\xi\| > \gamma^*) &= \mathbb{E}\left[\mathbb{P}\left(|\xi_1| > \gamma^*\|\xi_{-1}\| \mid \xi_{-1}\right)\right] \\
&\leq \mathbb{P}(E^c) + \mathbb{E}\left[I_E \cdot \frac{2}{\sqrt{2\pi}} \int_{\gamma^*\|\xi_{-1}\|}^{\infty} \exp(-x^2/2) dx\right] \\
&\leq \mathbb{P}(E^c) + \mathbb{E}\left[\frac{2}{\gamma^*\sqrt{(n-p)\pi}} \exp\left(-\frac{(\gamma^*)^2\|\xi_{-1}\|^2}{2}\right)\right] \\
&\leq \mathbb{P}(E^c) + \frac{2}{\gamma^*\sqrt{(n-p)\pi}} [1 + (\gamma^*)^2]^{-(n-p)/2}, \tag{A.57}
\end{aligned}$$

where in the third line we have used the well-known inequality of $\int_{\alpha}^{\infty} e^{-x^2/2} dx \leq \frac{1}{\alpha} e^{-\alpha^2/2}$, for any $\alpha > 0$, and in the last line we have used expression of the moment generating function of χ_{n-p}^2 . To bound $\mathbb{P}(E^c)$, we use a concentration inequality for chi-square distributions (it is an application the Bernstein's inequality for sub-exponential variables): If $W \sim \chi_k^2$, then

$$\mathbb{P}(|k^{-1}W - 1| > t) \leq 2 \exp(-kt^2/8), \quad \text{for any } t \in (0, 1).$$

We apply this inequality to get

$$\mathbb{P}(E) = \mathbb{P}\left(\frac{\|\xi_{-1}\|^2}{n-p} - 1 < -\frac{1}{2}\right) \leq 2 \exp\left(-\frac{n-p}{32}\right).$$

We plug it into (A.57) and then combine (A.57) with (A.56). It yields that

$$\mathbb{P}(\|\Sigma_p - \Sigma\| > \gamma) \leq 2(e^{-\frac{1}{32}})^{n-p} + \frac{\sqrt{\omega_j + 4\gamma}}{\gamma\sqrt{(n-p)\pi}} \left(1 + \frac{4\omega_j^{-2}\gamma^2}{1 + 4\omega_j^{-1}\gamma}\right)^{\frac{n-p}{2}} \leq p^{-L}, \tag{A.58}$$

as long as p is sufficiently large, for any fixed constant $L > 0$. Here we have used the assumption of $n-p \geq p^\delta$ and $\omega_j^{-1} \geq C_0^{-1}$ for constants $\delta > 0$ and $C_0 > 0$.

We apply Lemma 1.6.1 to the random vector $\hat{b} = (\hat{\beta}_j^+, \hat{\beta}_j^-)' / \sqrt{2 \log(p)}$. By (A.54),

$$\hat{b} | (\beta_j = 0) \sim \mathcal{N}_2 \left(0_2, \frac{1}{2 \log(p)} \Sigma \right), \quad \hat{b} | (\beta_j = \tau_p) \sim \mathcal{N}_2 \left(\mu, \frac{1}{2 \log(p)} \Sigma \right),$$

where $\Sigma = (\omega_j/2) \cdot I_2$ and $\mu = (\sqrt{r}, \sqrt{r})' / 2$. Together with (A.58), it is implied by Lemma 1.6.1 that

$$\begin{aligned} \mathbb{P}(M_j > t_p(u) | \beta_j = 0) &= L_p p^{-\inf_{b \in \mathcal{R}_u} \{b' \Sigma^{-1} b\}}, \\ \mathbb{P}(M_j < t_p(u) | \beta_j = \tau_p) &= L_p p^{-\inf_{b \in \mathcal{R}_u^c} \{(b-\mu)' \Sigma^{-1} (b-\mu)\}}, \end{aligned}$$

where \mathcal{R}_u is the collection of values of \hat{b} such that $M_j > \sqrt{2u \log(p)}$. Recall that

$$\frac{M_j^{\text{dif}}}{\sqrt{2 \log(p)}} = |\hat{b}_1 + \hat{b}_2| - |\hat{b}_1 - \hat{b}_2|, \quad \frac{M_j^{\text{sgm}}}{\sqrt{2 \log(p)}} = |\hat{b}_1 + \hat{b}_2| \cdot \text{sgn}(\hat{b}_1) \cdot \text{sgn}(\hat{b}_2).$$

The associated $\mathcal{R}_u^{\text{dif}}$ and $\mathcal{R}_u^{\text{sgm}}$ are shown in Figure A.3. These rejection regions do not depend on the design, but Σ depends on the design. By direct calculations,

$$\begin{aligned} \mathbb{P}(M_j > t_p(u) | \beta_j = 0) &= L_p p^{-\omega_j^{-1} u}, \\ \mathbb{P}(M_j > t_p(u) | \beta_j = \tau_p) &= \begin{cases} L_p p^{-\omega_j^{-1} \min\{(\sqrt{r}-\sqrt{u})_+^2, r/2\}}, & \text{if } M_j = M_j^{\text{sgm}}, \\ L_p p^{-(2\omega_j)^{-1} (\sqrt{r}-\sqrt{u})_+^2}, & \text{if } M_j = M_j^{\text{dif}}. \end{cases} \end{aligned}$$

The claim follows immediately. □

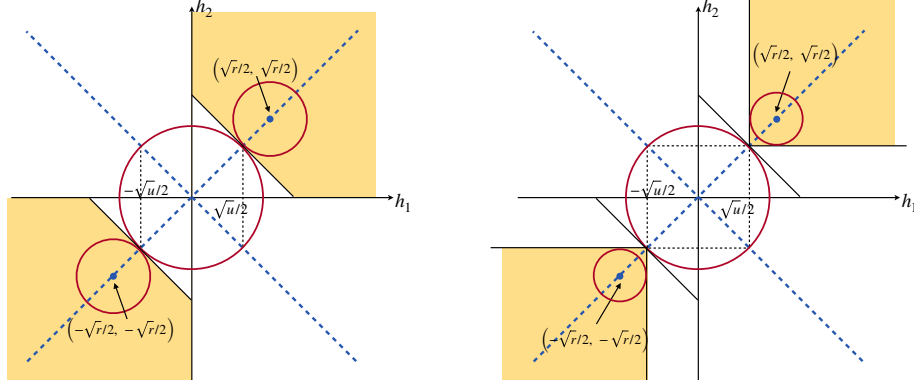


Figure A.3: The rejection region of Gaussian mirror in a general design, where the symmetric statistic is signed maximum (left) and difference (right). The rate of convergence of FP_p is captured by a ball centered at $(0, 0)$, and the rate of convergence of FN_p is captured by a ball centered at $(\sqrt{r}/2, \sqrt{r}/2)$.

A.1.8 PROOF OF THEOREM 1.5.2

By the property of least-square coefficients,

$$(\hat{\beta}_1, \dots, \hat{\beta}_p, \tilde{\beta}_1, \dots, \tilde{\beta}_p) \sim \mathcal{N}_{2p}((\beta_1, \dots, \beta_p, 0, \dots, 0), (G^*)^{-1}).$$

Consider the joint distribution of $\hat{\beta}_j$ and $\tilde{\beta}_j$ which are the regression coefficient of x_j and \tilde{x}_j , we know that $(\hat{\beta}_j, \tilde{\beta}_j) \sim \mathcal{N}_2((\beta_j, 0), A_j)$ where A_j has ω_{jj} as its diagonal element and ω_{2j} as its off-diagonal elements. Then theorem 1.5.2 is immediate from the following lemma:

Lemma A.1.1. If (Z_j, \tilde{Z}_j) follows $\mathcal{N}_2((\beta_j, 0)^T, \Sigma)$ with $\Sigma = ((\sigma_1, \sigma_2), (\sigma_2, \sigma_1))$, then

$$\mathbb{P}(|Z_j| > \sqrt{2u \log(p)}, |Z_j| \geq |\tilde{Z}_j| | \beta_j = 0) = L_p p^{-u/\sigma_1} \quad (\text{A.59})$$

and

$$\begin{aligned} & \mathbb{P}(|Z_j| \leq \sqrt{2u \log(p)} \text{ or } |Z_j| < |\tilde{Z}_j| | \beta_j = \sqrt{2r \log(p)}) \\ & = L_p p^{-\min\{(\sqrt{r}-\sqrt{u})_+^2/\sigma_1, r/(2 \max\{\sigma_1+\sigma_2, \sigma_1-\sigma_2\})\}}. \end{aligned} \quad (\text{A.60})$$

Next, we prove Lemma A.1.1. To compute the left hand side of (A.59), we only need to find the t such that ellipsoid $(x, y)\Sigma^{-1}(x, y)^T = t^2$ is tangent with $x = \pm\sqrt{2u\log(p)}$. This is because when we increase the radius of the ellipsoid, it must intersect with $x = \pm\sqrt{2u\log(p)}$ first amongst the boundaries of the region that pick variable j as a signal. When they intersect,

$$t^2 = \frac{1}{\sigma_1^2 - \sigma_2^2}(\sigma_1 x^2 - 2\sigma_2 xy + \sigma_1 y^2) = \frac{1}{\sigma_1^2 - \sigma_2^2} \left(\sigma_1 \left(y - \frac{\sigma_2}{\sigma_1} x \right)^2 + \left(\sigma_1 - \frac{\sigma_2^2}{\sigma_1} \right) x^2 \right) \geq \frac{2u\log(p)}{\sigma_1}.$$

When $t^2 = \frac{2u\log(p)}{\sigma_1}$, the tangent points are $(\pm\sqrt{2u\log(p)}, \pm\frac{\sigma_2}{\sigma_1}\sqrt{2u\log(p)})$. By Lemma 1.6.1, we verified (A.59).

For (A.60), when $r < u$, the center of the bi-variate normal is in the region of rejecting variable j as a signal thus the false positive rate is L_p . When $r > u$, we need to find the t such that ellipsoid $(x - \beta_j, y)\Sigma^{-1}(x - \beta_j, y)^T = t^2$ is tangent with either $x = \pm\sqrt{2u\log(p)}$ or $y = \pm x$. When the ellipsoid intersects with $x = \pm\sqrt{2u\log(p)}$,

$$t^2 = \frac{1}{\sigma_1^2 - \sigma_2^2} \left(\sigma_1 \left(y - \frac{\sigma_2}{\sigma_1} (x - \beta_j) \right)^2 + \left(\sigma_1 - \frac{\sigma_2^2}{\sigma_1} \right) (x - \beta_j)^2 \right) \geq \frac{2(\sqrt{u} - \sqrt{r})^2 \log(p)}{\sigma_1},$$

therefore, they are tangent at $(\pm\sqrt{2u\log(p)}, \frac{\sigma_2}{\sigma_1}(\pm\sqrt{2u\log(p)} - \beta_j))$ when $t^2 = \frac{2(\sqrt{u} - \sqrt{r})^2 \log(p)}{\sigma_1}$.

Meanwhile, since the long/short shaft of the ellipsoid are paralleled with $y = \pm x$, the tangent points of ellipsoid with $y = \pm x$ must be $(\beta_j/2, \beta_j/2)$ and $(\beta_j/2, -\beta_j/2)$, which gives $t^2 = \frac{r\log(p)}{\sigma_1 + \sigma_2}$ and $\frac{r\log(p)}{\sigma_1 - \sigma_2}$. From here we can conclude the "distance" between the center of the normal distribution and the region that reject variable j as a signal is

$$\min \left\{ \frac{2(\sqrt{r} - \sqrt{u})_+^2 \log(p)}{\sigma_1}, \frac{r\log(p)}{\sigma_1 + \sigma_2}, \frac{r\log(p)}{\sigma_1 - \sigma_2} \right\}.$$

By Lemma 1.6.1, we know

$$\mathbb{P}(|Z_j| \leq \sqrt{2u \log(p)} | \beta_j = \sqrt{2r \log(p)}) = L_p p^{-\min\{(\sqrt{r}-\sqrt{u})^2_+/\sigma_1, r/(2 \max\{\sigma_1+\sigma_2, \sigma_1-\sigma_2\})\}}.$$

□

A.1.9 PROOF OF LEMMA 1.5.1

By [Xing et al. \(2019\)](#), the Gaussian mirror framework can achieve asymptotically valid FDR control when the following two requirements are satisfied:

- The mirror statistics M_j is symmetrically distributed for any null feature j .
- There exist constants $C > 0$ and $\delta \in (0, 2)$ such that, for the set of null features $\mathcal{T} = \{j : \beta_j \neq 0\}$, $\sum_{j,k \in \mathcal{T}} \text{Cov}(\mathbb{I}(M_j \geq t), \mathbb{I}(M_k \geq t)) \leq C|\mathcal{T}|^\delta$ holds for $\forall t$.

$(\hat{\beta}_j^+ + \hat{\beta}_j^-)$ and $(\hat{\beta}_j^+ - \hat{\beta}_j^-)$ are, respectively, the regression coefficient of x_j and \tilde{x}_j when regressing y on $[x_1, \dots, x_{j-1}, x_j, \tilde{x}_j, x_{j+1}, \dots, x_p]$. Therefore, $(\hat{\beta}_j^+ + \hat{\beta}_j^-, \hat{\beta}_j^+ - \hat{\beta}_j^-) \sim \mathcal{N}_2((\beta_j, 0), D_j)$ where D_j is the inverse of gram matrix $\tilde{G}^{(j)} = [x_1, \dots, x_j, \tilde{x}_j, \dots, x_p][x_1, \dots, x_j, \tilde{x}_j, \dots, x_p]'$ restricted to the j th and $(j+1)$ th rows and columns. By the block matrix inversion formula, $D_j^{-1} = (x_j, \tilde{x}_j)^T(I - P_{-j})(x_j, \tilde{x}_j)$. Since $\|(I - P_{-j})\tilde{x}_j\| = \|(I - P_{-j})x_j\|$ holds for each $1 \leq j \leq p$, $D_j(1, 1) = D_j(2, 2)$. For any null feature j , $(\hat{\beta}_j^+ + \hat{\beta}_j^-, \hat{\beta}_j^+ - \hat{\beta}_j^-) \sim \mathcal{N}_2((0, 0), D_j)$. By construction, M_j is the signed maximum of $|\hat{\beta}_j^+ + \hat{\beta}_j^-|$ and $|\hat{\beta}_j^+ - \hat{\beta}_j^-|$, thus M_j is symmetrically distributed for any null feature j .

Secondly, we will show that $(x_j, \tilde{x}_j)'(I - P_{-j})(I - P_{-k})(x_k, \tilde{x}_k) = 0$ implies $\text{Cov}(\mathbb{I}(M_j \geq t), \mathbb{I}(M_k \geq t)) = 0$. This gives

$$\begin{aligned} \sum_{j,k \in \mathcal{T}} \text{Cov}(\mathbb{I}(M_j \geq t), \mathbb{I}(M_k \geq t)) &\leq \frac{1}{4} \times \#\{j \in \mathcal{T}, k \in \mathcal{T} | \text{Cov}(\mathbb{I}(M_j \geq t), \mathbb{I}(M_k \geq t)) \neq 0\} \\ &\leq \frac{1}{4} \times \#\{j \in \mathcal{T}, k \in \mathcal{T} | (x_j, \tilde{x}_j)'(I - P_{-j})(I - P_{-k})(x_k, \tilde{x}_k) \neq 0\}, \end{aligned}$$

so that condition 2 in Lemma 1.5.1 guarantees the second requirement at the beginning of our proof.

The regression coefficients when regressing y on $[x_j, \tilde{x}_j, x_1, \dots, x_{j-1}, x_{j+1}, \dots, x_p]$ is explicitly given by $((x_j, \tilde{x}_j, X_{-j})'(x_j, \tilde{x}_j, X_{-j}))^{-1}(x_j, \tilde{x}_j, X_{-j})'y$. We focus on the first two coordinates:

$$\begin{bmatrix} \hat{\beta}_j^+ + \hat{\beta}_j^- \\ \hat{\beta}_j^+ - \hat{\beta}_j^- \end{bmatrix} = \begin{bmatrix} D_j, & -D_j(x_j, \tilde{x}_j)'X_{-j}(X_j'X_{-j})^{-1} \end{bmatrix} (x_j, \tilde{x}_j, X_{-j})'y = D_j(x_j, \tilde{x}_j)'(I - P_{-j})y.$$

When $(x_j, \tilde{x}_j)'(I - P_{-j})(I - P_{-k})(x_k, \tilde{x}_k) = 0$,

$$\text{Cov}\left(\begin{bmatrix} \hat{\beta}_j^+ + \hat{\beta}_j^- \\ \hat{\beta}_j^+ - \hat{\beta}_j^- \end{bmatrix}, \begin{bmatrix} \hat{\beta}_k^+ + \hat{\beta}_k^- \\ \hat{\beta}_k^+ - \hat{\beta}_k^- \end{bmatrix}\right) = \sigma^2 D_j(x_j, \tilde{x}_j)'(I - P_{-j})(I - P_{-k})(x_k, \tilde{x}_k)D_k = 0,$$

thus $(\hat{\beta}_j^+ + \hat{\beta}_j^-, \hat{\beta}_j^+ - \hat{\beta}_j^-) \perp (\hat{\beta}_k^+ + \hat{\beta}_k^-, \hat{\beta}_k^+ - \hat{\beta}_k^-)$, which implies $M_j \perp M_k$ and $\text{Cov}(\mathbb{I}(M_j \geq t), \mathbb{I}(M_k \geq t)) = 0$. \square

A.1.10 PROOF OF THEOREM 1.5.3

Similar as in the proof of theorem 1.5.2, when regression Y on $[x_1, \dots, x_j, \tilde{x}_j, \dots, x_p]$, the regression coefficient of x_j and \tilde{x}_j are jointly normal distributed: $\mathcal{N}_2((\beta_j, 0), D_j)$ where D_j has σ_{1j} as its diagonal element and σ_{2j} as its off-diagonal elements. Theorem 1.5.3 immediately holds by Lemma A.1.1. \square

A.1.11 PROOF OF LEMMA 1.5.2

ω_j is the j th diagonal of the inverse of $X'X$, thus $\omega_j = (x_j'(I - P_{-j})x_j)^{-1}$.

σ_{1j} and σ_{2j} are the diagonal and off-diagonal of $D_j = ((x_j, \tilde{x}_j)^T(I - P_{-j})(x_j, \tilde{x}_j))^{-1}$. When $x_j'(I - P_{-j})\tilde{x}_j = 0$, D_j has its diagonal elements equal to $x_j'(I - P_{-j})x_j$ and off-diagonal elements equal to 0, so $\sigma_{1j} = \omega_j$ and $\sigma_{2j} = 0$. \square

Remark. Here we provide a proof for the footnote on page 17.

$$\begin{aligned}
\text{diag}(s) &= [\text{diag}(G^{-1})]^{-1} \\
\iff \text{diag}(\cdots, x_j'x_j - \tilde{x}_j'x_j, \cdots) &= [\text{diag}(\cdots, (x_j'(I - P_{-j})x_j)^{-1}, \cdots)]^{-1} \\
\iff x_j'x_j - \tilde{x}_j'x_j &= x_j'(I - P_{-j})x_j, \quad \forall j \\
\iff \tilde{x}_j'x_j - x_j'P_{-j}x_j &= 0, \quad \forall j \\
\iff \tilde{x}_j(I - P_{-j})x_j &= 0, \quad \forall j
\end{aligned}$$

□

A.1.12 PROOF OF THEOREM 1.5.4

We assume $\rho \geq 1/2$ throughout the proof. The calculation for the case where $\rho \leq -1/2$ is similar. By the design of the gram matrix $X^T X$ and the construction of the knockoff variables, we know Lasso regression problem with $2p$ variables can be reduced to $(p/2)$ independent four-variate Lasso regression problems:

$$(\hat{\beta}_j, \hat{\beta}_{j+1}, \hat{\beta}_{j+p}, \hat{\beta}_{j+p+1})(\lambda) = \underset{b}{\text{argmin}} \left\{ \frac{1}{2} \|y - (x_j, x_{j+1}, \tilde{x}_j, \tilde{x}_{j+1})b\|_2^2 + \lambda \|b\|_1 \right\} \quad (\text{A.61})$$

for $j = 1, 3, \dots, p-1$. By taking the sub-gradients of the objective function in (A.61), we know $(\hat{\beta}_j, \hat{\beta}_{j+1}, \hat{\beta}_{j+p}, \hat{\beta}_{j+p+1})$ should satisfy:

$$\begin{aligned}
(\hat{\beta}_j, \hat{\beta}_{j+1}, \hat{\beta}_{j+p}, \hat{\beta}_{j+p+1})G + \lambda(\text{sgn}(\hat{\beta}_j), \text{sgn}(\hat{\beta}_{j+1}), \text{sgn}(\hat{\beta}_{j+p}), \text{sgn}(\hat{\beta}_{j+p+1})) \\
= (y^T x_j, y^T x_{j+1}, y^T \tilde{x}_j, y^T \tilde{x}_{j+1})
\end{aligned} \quad (\text{A.62})$$

where $G = ((1, \rho, 2\rho - 1, \rho)^T, (\rho, 1, \rho, 2\rho - 1)^T, (2\rho - 1, \rho, 1, \rho)^T, (\rho, 2\rho - 1, \rho, 1)^T)$ and $\text{sgn}(x) = 1$ if $x > 0$; -1 if $x < 0$; any value in $[-1, 1]$ if $x = 0$. We have choose the correlation between a true

variable and its knockoff to be $2\rho - 1$, which is the smallest value such that $(X, \tilde{X})^T(X, \tilde{X})$ is semi-positive definite. In this case, G is degenerated and has rank 3. As λ is decreasing from infinity, we recognize that the first two variables (assume these two features are linear independent) entering the model will not leave before the third variable enters the model, which is obviously true from the close form solution of the bi-variate Lasso problem. We then show that the first two variables enter the Lasso path, individually. Furthermore, if the first two variables are a true variable and its knockoff variable, then the third and fourth variable enter the Lasso path simultaneously.

Since $(y^T x_j, y^T x_{j+1}, y^T \tilde{x}_j, y^T \tilde{x}_{j+1})^T \sim \mathcal{N}(G(\beta_j, \beta_{j+1}, 0, 0)^T, G)$ is a degenerated normal random variable, we reparametrize it as $(m + d_1, m + d_2, m - d_1, m - d_2)$ with $(m, d_1, d_2)^T \sim \mathcal{N}((\rho\beta_j + \rho\beta_{j+1}, (1 - \rho)\beta_j, (1 - \rho)\beta_{j+1})^T, \text{diag}(\rho, 1 - \rho, 1 - \rho))$. We intend to give the Lasso solution path (or Z_j, \tilde{Z}_j) as a function of m, d_1 and d_2 . We only present the result in the case where $d_1 > d_2 > 0$. Results from other cases are immediate by permuting the rows in equation set (A.62) and transforming to the $d_1 > d_2 > 0$ case. Lasso solution path are obtained by the KKT condition (A.62) and summarized in the table below.

range of m	λ_1	sign_1	λ_2	sign_2	λ_3	sign_3
$(-\infty, \frac{\rho}{1-\rho}(d_2 - d_1))$	$-m + d_1$	$(0, 0, 0^-, 0)$	$-m - \frac{\rho}{1-\rho}d_1 + \frac{1}{1-\rho}d_2$	$(0, 0, -, 0^-)$		
$(\frac{\rho}{1-\rho}(d_2 - d_1), 0)$	$-m + d_1$	$(0, 0, 0^-, 0)$	$\frac{1-\rho}{\rho}m + d_1$	$(0^+, 0, -, 0)$	d_2	$(+, 0^+, -, 0^-)$
$(0, \frac{\rho}{1-\rho}(d_1 - d_2))$	$m + d_1$	$(0^+, 0, 0, 0)$	$\frac{\rho-1}{\rho}m + d_1$	$(+, 0, 0^-, 0)$	d_2	$(+, 0^+, -, 0^-)$
$(\frac{\rho}{1-\rho}(d_1 - d_2), \infty)$	$m + d_1$	$(0^+, 0, 0, 0)$	$m - \frac{\rho}{1-\rho}d_1 + \frac{1}{1-\rho}d_2$	$(+, 0^+, 0, 0)$		

Table A.1: Summary of solution path of the Lasso problem (A.61). λ_i record the critical value of λ where a new variable enters the model and sign_i records the sign and the limiting behavior of $(\hat{\beta}_j, \hat{\beta}_{j+\rho})$ as $\lambda \rightarrow \lambda_i^-$. Value of λ_3 is omitted in row 1 and 4 since it will not affect the value of W_j and W_{j+1} .

Here we explain the third row of the table as an example, $b_1 = (\varepsilon, 0, 0, 0)^T$ is a solution of the KKT condition (A.62) when $\lambda = m + d_1 - \varepsilon$ for $\varepsilon \in (0, \frac{m}{\rho}]$, so sign_1 is expressed as $(0^+, 0, 0, 0)$. By property of the Lasso solution, if b_1 and b_2 are both Lasso solutions, then $G(b_1 - b_2) = 0$ and $\|b_1\|_1 = \|b_2\|_1$. $G(b_1 - b_2) = 0$ implies $b_1 - b_2 = \delta \times (1, -1, 1, -1)^T$ for some $\delta \neq 0$. Therefore, $b_2 = (\varepsilon - \delta, \delta, -\delta, \delta)^T$ and $\|b_2\|_1 \geq \|b_1\|_1 + 2|\delta|$. This means the Lasso solution is unique with

$\lambda = m + d_1 - \varepsilon$ and variable 1 is the only one entering the model when λ gets below λ_1 . When $\lambda = \frac{\rho-1}{\rho}m + d_1 - \varepsilon$ for $\varepsilon \in (0, \frac{\rho-1}{\rho}m + d_1 - d_2]$, $b_1 = (\frac{m}{\rho} + \frac{\varepsilon}{2-2\rho}, 0, -\frac{\varepsilon}{2-2\rho}, 0)^T$ is a solution of the KKT conditions. If there is another Lasso solution b_2 , then $b_2 = (\frac{m}{\rho} + \frac{\varepsilon}{2-2\rho} - \delta, \delta, -\frac{\varepsilon}{2-2\rho} - \delta, \delta)^T$ and $\|b_2\|_1 \geq \|b_1\|_1 + 2|\delta|$. So b_2 does not exist and variable 3 is the only one entering the model when λ gets below λ_2 . When $\lambda = d_2 - \varepsilon$ for sufficient small positive ε , $b_1 = (\frac{m}{2\rho} + \frac{d_1}{2-2\rho}, \frac{\varepsilon}{2-2\rho}, \frac{m}{2\rho} - \frac{d_1}{2-2\rho}, -\frac{\varepsilon}{2-2\rho})^T$ satisfies the KKT condition, thus variable 2 and 4 enters the model simultaneously. At this point, the Lasso solution is not unique and all solutions can be expressed as $b_1 - \delta \times (1, -1, 1, -1)^T$ with $\delta \in [-\frac{\varepsilon}{2-2\rho}, \frac{\varepsilon}{2-2\rho}]$. Other rows from the table can be analyzed similarly.

Table A.1 implicitly expresses $Z_j, Z_{j+1}, \tilde{Z}_j$ and \tilde{Z}_{j+1} as a function of d_1, d_2 and m . By examining all possible ordinal relationship of d_1, d_2 and 0, we record the region in the space of (d_1, d_2, m) such that $\hat{\beta}_j(u) > 0$ and denote it as $R(u)$. $R(u)$ is the union of 4 disjoint sub-regions $\{R_i(u)\}_{i=1, \dots, 4}$, defined as following:

$$\begin{aligned}
R_1(u) = & \{(x, y, z) : x > 0, y > 0, x > y, z > 0, x + z > T\} \\
& \cup \frac{1}{2} \{(x, y, z) : x > 0, y > 0, x < y, z < 0, z > x - y, x > T\} \\
& \cup \frac{1}{2} \{(x, y, z) : x > 0, y > 0, x < y, z > 0, z < \frac{\rho}{1-\rho}(y-x), x > T\} \\
& \cup \{(x, y, z) : x > 0, y > 0, x < y, z > 0, z > \max(\frac{\rho}{1-\rho}(y-x), T + \frac{\rho}{1-\rho}y - \frac{1}{1-\rho}x)\},
\end{aligned} \tag{A.63}$$

$R_2(u) = \{(x, y, z) : (-x, y, -z) \in R_1(u)\}$, $R_3(u) = \{(x, y, z) : (x, -y, z) \in R_1(u)\}$ and $R_4(u) = \{(x, y, z) : (-x, -y, -z) \in R_1(u)\}$, where $T = \sqrt{2u \log(p)}$ and the $\frac{1}{2}$ ahead of a certain region means when (d_1, d_2, m) is in this region, $\hat{\beta}_j(u) > 0$ happens with $1/2$ probability. Let the four disjoint regions that composes $R_1(u)$ in (A.63) be denoted by $R_{1,j}(u)$ for $j = 1, \dots, 4$. We can similarly define

$R_{i,j}(u)$ for $i = 2, 3, 4$. By Lemma 1, as $p \rightarrow \infty$,

$$\begin{aligned}
\mathbb{P}(\beta_j = 0, \hat{\beta}_j(u) \neq 0) &= \mathbb{P}(\hat{\beta}_j(u) \neq 0 | \beta_j = 0, \beta_{j+1} = 0) \times \mathbb{P}(\beta_j = 0, \beta_{j+1} = 0) \\
&\quad + \mathbb{P}(\hat{\beta}_j(u) \neq 0 | \beta_j = 0, \beta_{j+1} = \tau_p) \times \mathbb{P}(\beta_j = 0, \beta_{j+1} = \tau_p) \\
&= L_p p^{-\inf_{R(u)} [(x^2/\rho + x^2/(1-\rho) + y^2/(1-\rho))/(2 \log(p))]} \\
&\quad + L_p p^{-\mathcal{S} - \inf_{R(u)} [(z - \rho\tau_p)^2/\rho + x^2/(1-\rho) + (y - (1-\rho)\tau_p)^2/(1-\rho)]/(2 \log(p))},
\end{aligned} \tag{A.64}$$

$$\begin{aligned}
\mathbb{P}(\beta_j \neq 0, \hat{\beta}_j(u) = 0) &= \mathbb{P}(\hat{\beta}_j(u) = 0 | \beta_j = \tau_p, \beta_{j+1} = 0) \times \mathbb{P}(\beta_j = \tau_p, \beta_{j+1} = 0) \\
&\quad + \mathbb{P}(\hat{\beta}_j(u) = 0 | \beta_j = \tau_p, \beta_{j+1} = \tau_p) \times \mathbb{P}(\beta_j = \tau_p, \beta_{j+1} = \tau_p) \\
&= L_p p^{-\mathcal{S} - \inf_{R(u)^c} [(z - \rho\tau_p)^2/\rho + (x - (1-\rho)\tau_p)^2/(1-\rho) + y^2/(1-\rho)]/(2 \log(p))} \\
&\quad + L_p p^{-2\mathcal{S} - \inf_{R(u)^c} [(z - 2\rho\tau_p)^2/\rho + (x - (1-\rho)\tau_p)^2/(1-\rho) + (y - (1-\rho)\tau_p)^2/(1-\rho)]/(2 \log(p))}.
\end{aligned} \tag{A.65}$$

Define the ρ -distance function of two sets A and B in \mathbb{R}^3 as

$$d_\rho(A, B) = \inf_{a \in A, b \in B} [(a_1 - b_1)^2/(1-\rho) + (a_2 - b_2)^2/(1-\rho) + (a_3 - b_3)^2/\rho]$$

where a_k, b_k denote the k -th coordinate of vector a and b . An immediate property of the ρ -distance function would be

$$d_\rho(\cup_{i=1, \dots, M} A_i, \cup_{j=1, \dots, N} B_j) = \min_{i,j} d_\rho(A_i, B_j).$$

Utilizing the symmetry of the regions, we can compute the region distances involved in (A.64)

and (A.65) explicitly. Take the second exponent in (A.64) as an example, it can be simplified as

$$\begin{aligned}
& -\mathcal{J} - d_\rho(R(u), \{(0, (1-\rho)\tau_p, \rho\tau_p)\}) / (2 \log(p)) \\
&= -\mathcal{J} - d_\rho(R_1(u) \cup R_2(u) \cup R_3(u) \cup R_4(u), \{(0, (1-\rho)\tau_p, \rho\tau_p)\}) / (2 \log(p)) \\
&= -\mathcal{J} - d_\rho(R_{1,1}(u) \cup R_{1,3}(u) \cup R_{1,4}(u) \cup R_{2,2}(u), \{(0, (1-\rho)\tau_p, \rho\tau_p)\}) / (2 \log(p)).
\end{aligned}$$

Define $\tilde{R}_{1,2}(u) = \{(x, y, z) : x > 0, y > 0, z > 0, x < y, x > T, z < y - x\}$, $\tilde{R}_{1,3}(u) = \{(x, y, z) : x > 0, y > 0, z > 0, x < y, x > T\}$ and $\tilde{R}_{1,4}(u) = \{(x, y, z) : x > 0, y > 0, z > 0, x < y, x < T, z > T + \frac{\rho}{1-\rho}y - \frac{1}{1-\rho}x\}$. Then $\tilde{R}_{1,2}(u) \subset \tilde{R}_{1,3}(u)$ and $R_{1,3}(u) \cup R_{1,4}(u) = \tilde{R}_{1,3}(u) \cup \tilde{R}_{1,4}(u)$. Since $\tilde{R}_{1,2}(u)$ and $R_{2,2}(u)$ are symmetric about the plane $x = 0$, we know

$$d_\rho(R_{2,2}(u), \{(0, (1-\rho)\tau_p, \rho\tau_p)\}) = d_\rho(\tilde{R}_{1,2}(u), \{(0, (1-\rho)\tau_p, \rho\tau_p)\}).$$

Therefore,

$$\begin{aligned}
& d_\rho(R(u), \{(0, (1-\rho)\tau_p, \rho\tau_p)\}) \\
&= \min\{d_\rho(R_{1,1}(u), \{(0, (1-\rho)\tau_p, \rho\tau_p)\}), d_\rho(\tilde{R}_{1,3}(u), \{(0, (1-\rho)\tau_p, \rho\tau_p)\}), \\
&\quad d_\rho(\tilde{R}_{1,4}(u), \{(0, (1-\rho)\tau_p, \rho\tau_p)\})\} \\
&= \min\left\{\frac{1-\rho}{2} \times \tau_p^2 + \frac{2}{1+\rho} \times [(T - (1+\rho)\tau_p/2)_+]^2 - \frac{1-\rho}{1+\rho} \times [(T - (1+\rho)\tau_p)_+]^2, \right. \\
&\quad \left. \frac{1}{1-\rho} \times T^2 + \frac{1}{1-\rho} \times [(T - (1-\rho)\tau_p)_+]^2, \frac{1-\rho}{1+\rho} \times T^2 + ((T - \tau_p)_+)^2\right\} \\
&= (T - \rho\tau_p)^2 + (\xi_\rho \tau_p - \eta_\rho T)_+^2 - (\tau_p - T)_+^2,
\end{aligned}$$

where $\xi_\rho = \sqrt{1-\rho^2}$ and $\eta_\rho = \sqrt{(1-\rho)/(1+\rho)}$.

Let $\tau_p = 0$, we know $d_\rho(R(u), \{(0, 0, 0)\}) = T^2$. By (A.64) we immediately have

$$\mathbb{P}(\beta_j = 0, \hat{\beta}_j(u) \neq 0) = L_p p^{-\min\left\{u, \vartheta + (\sqrt{u} - \rho\sqrt{r})^2 + (\xi_\rho\sqrt{r} - \eta_\rho\sqrt{u})_+^2 - (\sqrt{r} - \sqrt{u})_+^2\right\}}. \quad (\text{A.66})$$

We can see the false positive rate is exactly the same when using the Lasso filter and the Knockoff filter when $\rho > 0$. For $\rho \geq 1/2$, we can similarly compute $d_\rho(R(u)^C, \{((1 - \rho)\tau_p, 0, \rho\tau_p)\})$ to be

$$[(\tau_p - T)_+ - ((1 - \xi_\rho)\tau_p - (1 - \eta_\rho)T)_+ - (\lambda_\rho\tau_p - \eta_\rho T)_+]^2,$$

and $d_\rho(R(u)^C, \{((1 - \rho)\tau_p, (1 - \rho)\tau_p, 2\rho\tau_p)\})$ to be

$$[(\xi_\rho\tau_p - \eta_\rho T)_+ - (\lambda_\rho\tau_p - \eta_\rho T)_+]^2,$$

where $\xi_\rho = \sqrt{1 - \rho^2}$, $\eta_\rho = \sqrt{(1 - \rho)/(1 + \rho)}$, and $\lambda_\rho = \sqrt{1 - \rho^2} - \sqrt{1 - \rho}$.

Plug these results in to (A.65), we have

$$\mathbb{P}(\beta_j \neq 0, \hat{\beta}_j(u) = 0) = L_p p^{-\vartheta - \left\{(\sqrt{r} - \sqrt{u})_+ - [(1 - \xi_\rho)\sqrt{r} - (1 - \eta_\rho)\sqrt{u}]_+ - (\lambda_\rho\sqrt{r} - \eta_\rho\sqrt{u})_+\right\}^2}. \quad (\text{A.67})$$

From here we have prove the result for $\rho \geq 1/2$ case.

In the case where $\rho \leq -1/2$, the exponent of false negative rate is additionally lower bounded by -2ϑ . One can verify the rate given in the theorem through similar calculations. This is somehow more straight forwards since in the case where $\beta_j = \beta_{j+1} = \tau$, $(y^T x_j, y^T x_{j+1}, y^T \tilde{x}_j, y^T \tilde{x}_{j+1})^T \sim \mathcal{N}((1 + \rho)\tau \cdot (1, 1, -1, -1)^T, G)$, meaning there is no way to distinguish the true variable from its knockoff variable. \square

A.1.13 PROOF OF THEOREM 1.5.5

In the following proofs, we only consider $\rho \geq 0$ case, since $\rho < 0$ case can be transformed to the positive $|\rho|$ case by flipping the sign of either β_j or β_{j+1} for $j = 1, 3, \dots, p-1$. By the block diagonal structure of the gram matrix, the Lasso problem with $2p$ features can be reduced to $(p/2)$ independent four-variate Lasso regression problems:

$$\hat{b}(\lambda) = \operatorname{argmin}_b \left\{ \frac{1}{2} \|y - (x_j, x_{j+1}, \tilde{x}_j, \tilde{x}_{j+1})b\|_2^2 + \lambda \|b\|_1 \right\} \quad (\text{A.68})$$

for $j = 1, 3, \dots, p-1$. Before we turn to the proof of the theorem, we first analysis the solution path of the following four-variate Lasso problem:

$$\hat{b} = \operatorname{argmin}_b \left\{ -b^T b + b^T B b / 2 + \lambda \|b\|_1 \right\}. \quad (\text{A.69})$$

with $B = ((1, \rho, a, \rho)^T, (\rho, 1, \rho, a)^T, (a, \rho, 1, \rho)^T, (\rho, a, \rho, 1)^T)$ and $a \in [2|\rho| - 1, 1]$. By taking the sub-gradients, we know \hat{b} should satisfy

$$B \hat{b} + \lambda \operatorname{sgn}(\hat{b}) = b. \quad (\text{A.70})$$

Let \hat{b}_i and b_i denotes the i -th coordinate of \hat{b} and b . Let $\lambda_1 > \lambda_2 > \lambda_3 > \lambda_4$ be the values at which variables enter the solution path. As discussed in the proof of Lemma 1.6.2, $\lambda_1 = \max\{|b_1|, |b_2|, |b_3|, |b_4|\}$. Without loss of generality, assume $\lambda_1 = |b_1|$ and variable 1 is the first variable entering the model in solution path. We know for one variate Lasso problem, the only feature will not leave the model after its entry as λ is decreasing. So in the four-variate Lasso (A.69), variable 1 will stay in the model

until the second variable enters the model. Consider three bi-variate Lasso problems ($k = 2, 3, 4$):

$$\hat{b}^{(k)} = \operatorname{argmin}_{b^{(k)}} \left\{ -(b^{(k)})^T b^{(k)} + (b^{(k)})^T B^{(k)} b^{(k)} / 2 + \lambda \|b^{(k)}\|_1 \right\} \quad (\text{A.71})$$

with

$$B^{(2)} = B^{(4)} = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \quad \text{and} \quad B^{(3)} = \begin{bmatrix} 1 & a \\ a & 1 \end{bmatrix},$$

$b^{(2)} = (b_1, b_2)$, $b^{(3)} = (b_1, b_3)$ and $b^{(4)} = (b_1, b_4)$. Now, we claim $\lambda_2 = \max_k \{\lambda_2^{(k)}\}$ where $\lambda_2^{(k)}$ is the value at which the second variables enter the solution path in the k -th bi-variate Lasso problems. Suppose $\lambda_2^{(i)} > \lambda_2^{(k)}$ for $i \neq k \in \{2, 3, 4\}$, when $\lambda \in [\lambda_2^{(i)}, \lambda_1]$, we know the KKT condition (A.70) is satisfied with $b_2 = b_3 = b_4 = 0$ by looking at the KKT conditions of the bi-variate Lasso problems. When $\lambda \in [\lambda_2^{(i)} - \varepsilon, \lambda_2^{(i)})$, a second variable i must have entered the four-variate Lasso path, since the objective function of (A.69) is smaller when including variable 1 and i than including variable 1 alone (this is because the second variable have entered the model in the i -th bi-variate Lasso path when $\lambda \in [\lambda_2^{(i)} - \varepsilon, \lambda_2^{(i)})$). We are ready to prove the theorem now, using what we have shown regarding λ_1 and λ_2 . We next compute the false positive rate and false negative rate given $(\beta_j, \beta_{j+1}) = (0, 0), (0, \tau_p), (\tau_p, 0), (\tau_p, \tau_p), (-\tau_p, \tau_p)$ by deriving upper and lower bounds for those rates.

We first establish some notations. For the four-variate Lasso problem (A.68), let A_i denotes the event that variable i is the first one entering the model, A_{i_1, i_2} denotes the event that variable i_1 and i_2 are the first two entering the model (ignoring the order between i_1 and i_2) and $A_{i_1 \rightarrow i_2}$ denotes the event that variable i_1 is the first one and variable i_2 is the second one entering the model. Let L_{i_1, i_2} denote the bi-variate Lasso problem with y as the response and x_{i_1}, x_{i_2} as the variables. Let $b \equiv (y^T x_j, y^T x_{j+1}, y^T \tilde{x}_j, y^T \tilde{x}_{j+1})$, then $b \sim \mathcal{N}(\mu, G)$ with $\mu = G(\beta_j, \beta_{j+1}, 0, 0)^T$ and $G = ((1, \rho, 0, \rho)^T, (\rho, 1, \rho, 0)^T, (0, \rho, 1, \rho)^T, (\rho, 0, \rho, 1)^T)$. When not causing any confusing, we write t_p in place of $t_p(u)$ for simplicity.

- When $(\beta_j, \beta_{j+1}) = (0, 0)$,

$$\mathbb{P}\{W_j > t_p | (\beta_j, \beta_{j+1}) = (0, 0)\} = L_p p^{-u}. \quad (\text{A.72})$$

To derive a lower bound for $\mathbb{P}\{W_j > t_p | (\beta_j, \beta_{j+1}) = (0, 0)\}$, we look for a point in the region (or on the boundary of the region) that choose variable j as a signal and apply Lemma 1.6.1. The point we choose is $p_1 = (t_p, \rho t_p, 0, \rho t_p)^T$ where $t_p = \sqrt{2u \log(p)}$. It's obvious that when $b = p_1$, variable j is the first one entering the Lasso path. Though $b = p_1$ is in the rejection region, it is also on the boundary of the region that choose variable j as a signal because slight increasing the first coordinate will result in variable j being selected. Since $b \sim \mathcal{N}(\mu_1, G)$ with $\mu_1 = \mathbf{0}$, by Lemma 1.6.1,

$$\mathbb{P}\{W_j > t_p | (\beta_j, \beta_{j+1}) = (0, 0)\} \geq L_p p^{-(p_1 - \mu_1)^T G^{-1} (p_1 - \mu_1) / 2 \log(p)} = L_p p^{-u}.$$

The upper bound is straight forward by considering the first variable- i entering the model and notice that $W_i \sim \mathcal{N}(0, 1)$:

$$\begin{aligned} \mathbb{P}\{W_j > t_p | (\beta_j, \beta_{j+1}) = (0, 0)\} &= \sum_i \mathbb{P}\{W_j > t_p, A_i | (\beta_j, \beta_{j+1}) = (0, 0)\} \\ &\leq \sum_i \mathbb{P}\{W_i > t_p | (\beta_j, \beta_{j+1}) = (0, 0)\} = L_p p^{-u}. \end{aligned} \quad (\text{A.73})$$

- When $(\beta_j, \beta_{j+1}) = (0, \tau_p)$,

$$\mathbb{P}\{W_j > t_p | (\beta_j, \beta_{j+1}) = (0, \tau_p)\} \geq L_p p^{-(\sqrt{u} - \rho \sqrt{r})^2 - (\xi_p \sqrt{r} - \eta_p \sqrt{u})_+^2 + (\sqrt{r} - \sqrt{u})_+^2}, \quad (\text{A.74})$$

$$\mathbb{P}\{W_j > t_p, A | (\beta_j, \beta_{j+1}) = (0, \tau_p)\} \leq L_p p^{-u} \quad (\text{A.75})$$

for $A = A_{j+p+1 \rightarrow j}, A_{j+1, j+p+1}$ and

$$\mathbb{P}\{W_j > t_p, A | (\beta_j, \beta_{j+1}) = (0, \tau_p)\} \leq L_p p^{-(\sqrt{u} - \rho\sqrt{r})^2 - (\xi_\rho \sqrt{r} - \eta_\rho \sqrt{u})_+^2 + (\sqrt{r} - \sqrt{u})_+^2} \quad (\text{A.76})$$

for $A = A_{j, j+1}, A_{j, j+p}, A_{j \rightarrow j+p+1}$.

This time we choose

$$p_2^T = \begin{cases} (t_p, \rho t_p + (1 - \rho^2)\tau_p, \rho\tau_p, \rho t_p - \rho^2\tau_p), & (1 + \rho)\tau_p \leq t_p, \\ (t_p, t_p, \frac{\rho}{1+\rho}t_p, \frac{\rho}{1+\rho}t_p), & \tau_p \leq t_p < (1 + \rho)\tau_p, \\ (t_p + \rho(\tau_p - t_p), \tau_p, \rho(\tau_p - t_p) + \frac{\rho}{1+\rho}t_p, \frac{\rho}{1+\rho}t_p), & t_p < \tau_p. \end{cases}$$

When $h = p_2$ and $t_p \geq \tau_p$, variable j is the first variable entering the four-variate Lasso path with $W_j = t_p$; when $h = p_2$ and $t_p < \tau_p$, variable $j + 1$ is the first and j is the second variable entering the Lasso path with $W_j = t_p$ and $W_{j+1} = \tau_p$. $h = p_2$ is on the boundary of the region that chooses variable j as a signal. Since $h \sim \mathcal{N}(\mu_2, G)$ with $\mu_2 = (\rho\tau_p, \tau_p, \rho\tau_p, 0)^T$, by Lemma 1.6.1,

$$\begin{aligned} \mathbb{P}\{W_j > t_p | (\beta_j, \beta_{j+1}) = (0, \tau_p)\} &\geq L_p p^{-(p_2 - \mu_2)^T G^{-1} (p_2 - \mu_2) / 2 \log(p)} \\ &= L_p p^{-(\sqrt{u} - \rho\sqrt{r})^2 - (\xi_\rho \sqrt{r} - \eta_\rho \sqrt{u})_+^2 + (\sqrt{r} - \sqrt{u})_+^2}. \end{aligned}$$

When $A_{j+1, j+p+1}$ occurs, since by our argument on λ_1 and λ_2 , Z_{j+1} and Z_{j+p+1} are the λ value at which the variables enter the solution path in the bi-variate Lasso problem $L_{j+1, j+p+1}$. Therefore, $Z_{j+1} = |y^T x_{j+1}|, Z_{j+p+1} = |y^T \tilde{x}_{j+1}|$. We notice that $Z_{j+p+1} > Z_j > t_p$ and

marginally $y^T \tilde{x}_{j+1} \sim \mathcal{N}(0, 1)$, so

$$\begin{aligned} & \mathbb{P}\{W_j > t_p, A_{j+1, j+p+1} | (\beta_j, \beta_{j+1}) = (0, \tau_p)\} \\ & \leq \mathbb{P}\{|y^T \tilde{x}_{j+1}| > t_p | (\beta_j, \beta_{j+1}) = (0, \tau_p)\} = L_p p^{-u}. \end{aligned}$$

Above inequality also holds for $A_{j+p+1, j}$ since if variable $j+p+1$ is the first entering the Lasso path, then we must have $|y^T \tilde{x}_{j+1}| = Z_{j+p+1} > Z_j > t_p$.

When any one of $A_{j, j+1}, A_{j, j+p}, A_{j \rightarrow j+p+1}$ occurs, it implies in the bi-variate Lasso problem $L_{j, j+1}$, the largest λ such that variable 1 enters the model for the first time is equal to W_j , thus larger than t_p . In other words, if variable j is a false positive using Knockoff for variable selection, then it is also a false positive when using bi-variate Lasso $L_{j, j+1}$. This means $\mathbb{P}\{W_j > t_p, A | (\beta_j, \beta_{j+1}) = (0, \tau_p)\}$ is upper bounded by the corresponding false positive rate of Lasso, which is $L_p p^{-(\sqrt{u}-\rho\sqrt{r})^2 - (\xi_\rho \sqrt{r} - \eta_\rho \sqrt{u})_+^2 + (\sqrt{r}-\sqrt{u})_+^2}$, for $A = A_{j, j+1}, A_{j, j+p}, A_{j \rightarrow j+p+1}$.

Since $A_{j+1, j+p}$ and $A_{j+p, j+p+1}$ can never occur when $W_j > 0$, (A.75) and (A.76) implies

$$\mathbb{P}\{W_j > t_p | (\beta_j, \beta_{j+1}) = (0, \tau_p)\} \leq L_p p^{-\min\{u, (\sqrt{u}-\rho\sqrt{r})^2 + (\xi_\rho \sqrt{r} - \eta_\rho \sqrt{u})_+^2 - (\sqrt{r}-\sqrt{u})_+^2\}}. \quad (\text{A.77})$$

Further coupled with (A.72) and (A.74), we have

$$\mathbb{P}\{W_j > t_p, \beta_j = 0\} = L_p p^{-\min\{u, \vartheta + (\sqrt{u}-\rho\sqrt{r})^2 + (\xi_\rho \sqrt{r} - \eta_\rho \sqrt{u})_+^2 - (\sqrt{r}-\sqrt{u})_+^2\}}. \quad (\text{A.78})$$

- When $(\beta_j, \beta_{j+1}) = (\tau_p, 0)$,

$$\mathbb{P}\{W_j \leq t_p | (\beta_j, \beta_{j+1}) = (\tau_p, 0)\} \geq L_p p^{-[(\sqrt{r}-\sqrt{u})_+]^2}, \quad (\text{A.79})$$

and

$$\mathbb{P}\{W_j \leq t_p | (\beta_j, \beta_{j+1}) = (\tau_p, 0)\} \leq L_p p^{\mathfrak{D} - f_{\text{Hamm}}^+(u, r, \mathfrak{D})}. \quad (\text{A.80})$$

Let $p_3 = (t_p, \rho t_p, 0, \rho t_p)^T$. When $h = p_3$, variable j is the first variable entering the Lasso path and p_3 is in the region of rejecting variable j as a signal. Since $h \sim \mathcal{N}(\mu_3, G)$ with $\mu_3 = (\tau_p, \rho \tau_p, 0, \rho \tau_p)^T$, by Lemma 1.6.1,

$$\begin{aligned} \mathbb{P}\{W_j \leq t_p | (\beta_j, \beta_{j+1}) = (\tau_p, 0)\} &\geq L_p p^{-(p_3 - \mu_3)^T G^{-1} (p_3 - \mu_3) / 2 \log(p)} \\ &= L_p p^{-[(\sqrt{r} - \sqrt{u})_+]^2}. \end{aligned}$$

Before we prove (A.80), we first analysis $f_{\text{Hamm}}^+(u, r, \mathfrak{D})$. By simply calculation, we find the optimal value of u that maximize $f_{\text{Hamm}}^+(u, r, \mathfrak{D})$ given r, \mathfrak{D} is

$$u^* = \begin{cases} \frac{1+\rho}{(\sqrt{1+\rho} + \sqrt{1-\rho})^2} r, & \mathfrak{D} \leq \frac{2\rho}{(\sqrt{1+\rho} + \sqrt{1-\rho})^2} r, \\ \frac{(r+\mathfrak{D})^2}{4r}, & \frac{2\rho}{(\sqrt{1+\rho} + \sqrt{1-\rho})^2} r \leq \mathfrak{D} < r, \\ \mathfrak{D}, & r < \mathfrak{D}. \end{cases}$$

This implies $u^* \geq \frac{1+\rho}{(\sqrt{1+\rho} + \sqrt{1-\rho})^2} r$ regardless of the relationship of \mathfrak{D} and r . Consider r, \mathfrak{D} as fixed, $f_{\text{Hamm}}^+(r, u, \mathfrak{D})$ as a function of u is monotonically non-decreasing in $[0, u^*]$ and monotonically non-increasing in $[u^*, \infty)$. $f_{\text{Hamm}}^+(r, \mathfrak{D}) = \mathfrak{D} + [(\sqrt{r} - \sqrt{u})_+ - ((1 - \xi_\rho)\sqrt{r} - (1 - \eta_\rho)\sqrt{u})_+]^2$ if and only if $u > u^*$. Since $(1 - \xi_\rho)\sqrt{r} - (1 - \eta_\rho)\sqrt{u^*} < 0$, $(1 - \xi_\rho)\sqrt{r} - (1 - \eta_\rho)\sqrt{u} < 0$ for all $u > u^*$, which implies $f_{\text{Hamm}}^+(r, \mathfrak{D}) = \mathfrak{D} + [(\sqrt{r} - \sqrt{u})_+]^2$ when $u > u^*$. Therefore,

$$\begin{aligned} f_{\text{Hamm}}^+(r, u, \mathfrak{D}) &= \min\{u, \mathfrak{D} + (\sqrt{u} - |\rho|\sqrt{r})^2 + ((\xi_\rho\sqrt{r} - \eta_\rho\sqrt{u})_+)^2 - ((\sqrt{r} - \sqrt{u})_+)^2, \\ &\quad \mathfrak{D} + [(\sqrt{r} - \sqrt{u})_+]^2\}. \end{aligned}$$

Now, we show that (A.80) holds for $u \geq u^*$. This would implies (A.80) for all $u \geq 0$, since the false negative rate $\mathbb{P}\{W_j \leq t_p(u) | (\beta_j, \beta_{j+1}) = (\tau_p, 0)\}$ is monotone non-decreasing with u , so for $u < u^*$, $\mathbb{P}\{W_j \leq t_p(u) | (\beta_j, \beta_{j+1}) = (\tau_p, 0)\} \leq \mathbb{P}\{W_j \leq t_p(u^*) | (\beta_j, \beta_{j+1}) = (\tau_p, 0)\} \leq L_p p^{\mathfrak{g} - f_{\text{Hamm}}^+(r, u^*, \mathfrak{g})} \leq L_p p^{\mathfrak{g} - f_{\text{Hamm}}^+(r, u, \mathfrak{g})}$.

Assume $u \geq u^*$, so $u \geq \frac{1+\rho}{(\sqrt{1+\rho} + \sqrt{1-\rho})^2} r$ and

$$-[(\sqrt{r} - \sqrt{u})_+]^2 \geq -\left(\frac{\sqrt{1-\rho}}{\sqrt{1+\rho} + \sqrt{1-\rho}}\right)^2 r \geq -(2 - \sqrt{3})(1-\rho)r \geq -\frac{1-\rho}{2}r \geq -\frac{1}{2}r. \quad (\text{A.81})$$

We next prove (A.80) by showing that

$$\mathbb{P}\{W_j \leq t_p, A | (\beta_j, \beta_{j+1}) = (\tau_p, 0)\} \leq L_p p^{-[(\sqrt{r} - \sqrt{u})_+]^2} \quad (\text{A.82})$$

holds for $A = A_j, A_{j+1}, A_{j+p}, A_{j+p+1}$ and $u \geq u^*$. Respectively,

$$\begin{aligned} \mathbb{P}\{W_j \leq t_p, A_j | (\beta_j, \beta_{j+1}) = (\tau_p, 0)\} &\leq \mathbb{P}\{|y^T x_j| \leq t_p | (\beta_j, \beta_{j+1}) = (\tau_p, 0)\} \\ &= L_p p^{-[(\sqrt{r} - \sqrt{u})_+]^2}, \end{aligned}$$

and by symmetry and (A.81),

$$\begin{aligned} \mathbb{P}\{W_j \leq t_p, A_{j+1} | (\beta_j, \beta_{j+1}) = (\tau_p, 0)\} &= \mathbb{P}\{W_j \leq t_p, A_{j+p+1} | (\beta_j, \beta_{j+1}) = (\tau_p, 0)\} \\ &\leq \mathbb{P}\{|y^T x_j| \leq |y^T x_{j+p+1}| | (\beta_j, \beta_{j+1}) = (\tau_p, 0)\} \leq L_p p^{-\frac{1-\rho}{2}r} \leq L_p p^{-[(\sqrt{r} - \sqrt{u})_+]^2}, \end{aligned}$$

$$\begin{aligned} \mathbb{P}\{W_j \leq t_p, A_{j+p} | (\beta_j, \beta_{j+1}) = (\tau_p, 0)\} &\leq \mathbb{P}\{|y^T x_j| \leq |y^T x_{j+p}| | (\beta_j, \beta_{j+1}) = (\tau_p, 0)\} \\ &\leq L_p p^{-\frac{1}{2}r} \leq L_p p^{-[(\sqrt{r} - \sqrt{u})_+]^2}. \end{aligned}$$

(A.80) is immediate by $[(\sqrt{r} - \sqrt{u})_+]^2 \geq f_{\text{Hamm}}^+(r, u, \mathfrak{g}) - \mathfrak{g}$.

- When $(\beta_j, \beta_{j+1}) = (\tau_p, \tau_p)$,

$$\mathbb{P}\{W_j \leq t_p | (\beta_j, \beta_{j+1}) = (\tau_p, \tau_p)\} \leq L_p \mathfrak{p}^{\mathfrak{s} - f_{\text{Hamm}}^+(u, r, \mathfrak{s})}. \quad (\text{A.83})$$

More precisely, we will prove that

$$\mathbb{P}\{W_j \leq t_p | (\beta_j, \beta_{j+1}) = (\tau_p, \tau_p)\} \leq L_p \mathfrak{p}^{-[(\sqrt{r} - \sqrt{u})_+]^2} \quad (\text{A.84})$$

holds for $u \geq u^*$, thus implies (A.83). We prove (A.84) by showing

$$\mathbb{P}\{W_j \leq t_p, A | (\beta_j, \beta_{j+1}) = (\tau_p, \tau_p)\} \leq L_p \mathfrak{p}^{-[(\sqrt{r} - \sqrt{u})_+]^2} \quad (\text{A.85})$$

holds for $A = A_j, A_{j+1 \rightarrow j}, A_{j+1 \rightarrow j+p}, A_{j+1 \rightarrow j+p+1}, A_{j+p}, A_{j+p+1}$ and $u \geq u^*$, which cover all possibilities. Respectively,

$$\begin{aligned} \mathbb{P}\{W_j \leq t_p, A_j | (\beta_j, \beta_{j+1}) = (\tau_p, \tau_p)\} &\leq \mathbb{P}\{|y^T x_j| \leq t_p | (\beta_j, \beta_{j+1}) = (\tau_p, \tau_p)\} \\ &\leq L_p \mathfrak{p}^{-[(1+\rho)\sqrt{r} - \sqrt{u}]_+^2} \leq L_p \mathfrak{p}^{-[(\sqrt{r} - \sqrt{u})_+]^2}, \end{aligned}$$

$$\begin{aligned} \mathbb{P}\{W_j \leq t_p, A_{j+p} | (\beta_j, \beta_{j+1}) = (\tau_p, \tau_p)\} &\leq \mathbb{P}\{|y^T x_j| \leq |y^T \tilde{x}_j| | (\beta_j, \beta_{j+1}) = (\tau_p, \tau_p)\} \\ &\leq L_p \mathfrak{p}^{-\frac{1}{2}r} \leq L_p \mathfrak{p}^{-[(\sqrt{r} - \sqrt{u})_+]^2}, \end{aligned}$$

$$\begin{aligned} \mathbb{P}\{W_j \leq t_p, A_{j+p+1} | (\beta_j, \beta_{j+1}) = (\tau_p, \tau_p)\} &\leq \mathbb{P}\{|y^T x_j| \leq |y^T \tilde{x}_{j+1}| | (\beta_j, \beta_{j+1}) = (\tau_p, \tau_p)\} \\ &\leq L_p \mathfrak{p}^{-\frac{1}{2(1-\rho)}r} \leq L_p \mathfrak{p}^{-[(\sqrt{r} - \sqrt{u})_+]^2}. \end{aligned}$$

When $A_{j+1 \rightarrow j}$ occurs, the bi-variate Lasso problem $L_{j,j+1}$ shares the same λ_1 and λ_2 with the four-variate Lasso problem. So variable j is a false negative when doing variable selection using the bi-variate Lasso $L_{j,j+1}$ given $W_j \leq t_p$, which implies $\mathbb{P}\{W_j \leq t_p, A_{j+1 \rightarrow j} | (\beta_j, \beta_{j+1}) =$

(τ_p, τ_p) is upper bounded by the corresponding false negative rate of Lasso, which is $L_p \cdot p^{-(\xi_p \sqrt{r} - \eta_p \sqrt{u})_+^2} \leq L_p p^{-[(\sqrt{r} - \sqrt{u})_+]^2}$. The last inequality is equivalent to

$$(1 - \sqrt{1 - \rho^2})\sqrt{r} \leq \left(1 - \sqrt{\frac{1 - \rho}{1 + \rho}}\right)\sqrt{u}.$$

By (A.81), the right hand side is no smaller than \sqrt{r} , thus no smaller than the left hand side.

When $A_{j+1 \rightarrow j+p}$ occurs, we know variable $j + p$ instead of variable j is the second one entering the Lasso path. This means the λ_2 (the λ value when the second variable entering Lasso path) of the bi-variate Lasso problem $L_{j+1, j+p}$ is larger than the λ_2 of the bi-variate Lasso problem $L_{j, j+1}$. Since we have derived the explicit expression of λ_2 in bi-variate Lasso problems, when $y^T x_{j+1} \geq 0$, we must have

$$\max\left\{\frac{y^T x_j - \rho y^T x_{j+1}}{1 - \rho}, \frac{y^T x_j - \rho y^T x_{j+1}}{-1 - \rho}\right\} < \max\left\{\frac{y^T \tilde{x}_j - \rho y^T x_{j+1}}{1 - \rho}, \frac{y^T \tilde{x}_j - \rho y^T x_{j+1}}{-1 - \rho}\right\}.$$

Therefore, $A_{j+1 \rightarrow j+p}$ implies one the three following events must occur:

$$y^T x_{j+1} < 0, \frac{y^T x_j - \rho y^T x_{j+1}}{1 - \rho} < \frac{y^T \tilde{x}_j - \rho y^T x_{j+1}}{1 - \rho}, \frac{y^T x_j - \rho y^T x_{j+1}}{1 - \rho} < \frac{y^T \tilde{x}_j - \rho y^T x_{j+1}}{-1 - \rho}$$

The probability of these three events given $(\beta_j, \beta_{j+1}) = (\tau_p, \tau_p)$ are $L_p p^{-(1+\rho)^2 r}$, $L_p p^{-\frac{r}{2}}$ and $L_p p^{-\frac{(1+2\rho)^2(1-\rho)}{2(1+\rho)} r}$, all of which are upper bounded by $L_p p^{-[(\sqrt{r} - \sqrt{u})_+]^2}$ when $u \geq u^*$.

When $A_{j+1 \rightarrow j+p+1}$ occurs, the λ_2 of the bi-variate Lasso problem $L_{j+1, j+p+1}$ is larger than the λ_2 of the bi-variate Lasso problem $L_{j, j+1}$. When $y^T x_{j+1} \geq 0$, we must have

$$\max\left\{\frac{y^T x_j - \rho y^T x_{j+1}}{1 - \rho}, \frac{y^T x_j - \rho y^T x_{j+1}}{-1 - \rho}\right\} < |y^T \tilde{x}_{j+1}|.$$

Therefore, $A_{j+1 \rightarrow j+p+1}$ implies one the three following events must occur:

$$y^T x_{j+1} < 0, \frac{y^T x_j - \rho y^T x_{j+1}}{1 - \rho} < y^T \tilde{x}_{j+1}, \frac{y^T x_j - \rho y^T x_{j+1}}{1 - \rho} < -y^T \tilde{x}_{j+1}.$$

Respectively, the probability of these three events are $L_p p^{-(1+\rho)^2 r}$, $L_p p^{-\frac{r}{2}}$ and $L_p p^{-\frac{(1+2\rho)^2(1-\rho)}{2(1+\rho)} r}$, all of which are upper bounded by $L_p p^{-[(\sqrt{r} - \sqrt{u})_+]^2}$ when $u \geq u^*$. From here we have verified (A.85), thus implies (A.83).

From (A.78) and (A.79), we have

$$\mathbb{P}\{W_j > t_p, \beta_j = 0\} + \mathbb{P}\{W_j \leq t_p, \beta_j = \tau_p\} \geq L_p p^{-f_{\text{Hamm}}^+(r, u, \vartheta)}. \quad (\text{A.86})$$

$$\begin{aligned} \mathbb{P}\{W_j \leq t_p, \beta_j = \tau_p\} &= p^{-\vartheta} \times \mathbb{P}\{W_j \leq t_p | (\beta_j, \beta_{j+1}) = (\tau_p, 0)\} \\ &\quad + p^{-2\vartheta} \times \mathbb{P}\{W_j \leq t_p | (\beta_j, \beta_{j+1}) = (\tau_p, \tau_p)\} \\ &\leq L_p p^{-f_{\text{Hamm}}^+(r, u, \vartheta)} \end{aligned} \quad (\text{A.87})$$

Since (A.78) also implies $\mathbb{P}\{W_j > t_p, \beta_j = 0\} \leq L_p p^{-f_{\text{Hamm}}^+(r, u, \vartheta)}$, we know

$$\mathbb{P}\{W_j > t_p, \beta_j = 0\} + \mathbb{P}\{W_j \leq t_p, \beta_j = \tau_p\} = L_p p^{-f_{\text{Hamm}}^+(r, u, \vartheta)}. \quad (\text{A.88})$$

- When $(\beta_j, \beta_{j+1}) = (-\tau_p, \tau_p)$,

$$\mathbb{P}\{W_j \leq t_p | (\beta_j, \beta_{j+1}) = (-\tau_p, \tau_p)\} \geq L_p p^{-((\xi_\rho \sqrt{r} - \eta_\rho^{-1} \sqrt{u})_+)^2}, \quad (\text{A.89})$$

$$\mathbb{P}\{W_j \leq t_p | (\beta_j, \beta_{j+1}) = (-\tau_p, \tau_p)\} \geq L_p p^{-\frac{(1-2\rho)^2(1+\rho)}{2(1-\rho)}r}, \quad (\text{A.90})$$

and

$$\mathbb{P}\{W_j \leq t_p | (\beta_j, \beta_{j+1}) = (-\tau_p, \tau_p)\} \leq L_p p^{-\min\{((\xi_\rho \sqrt{r} - \eta_\rho^{-1} \sqrt{u})_+)^2, \frac{(1-2\rho)^2(1+\rho)}{2(1-\rho)}r, -\vartheta + f_{\text{Hamm}}^+(u, r, \vartheta)\}}. \quad (\text{A.91})$$

Let

$$p_4^T = \begin{cases} (-(1-\rho)\tau_p, (1-\rho)\tau_p, \rho\tau_p, -\rho\tau_p), & (1-\rho)\tau_p \leq t_p, \\ (\rho(1-\rho)\tau_p - (1+\rho)t_p, (1-\rho)\tau_p, \rho(1-\rho)\tau_p + \frac{\rho^2}{1-\rho}t_p, -\frac{\rho}{1-\rho}t_p), & (1-\rho)\tau_p > t_p. \end{cases}$$

When $h = p_4$ and $(1-\rho)\tau_p \leq t_p$, variable j is the first variable entering the Lasso path with $W_j = (1-\rho)\tau_p \leq t_p$; when $h = p_4$ and $(1-\rho)\tau_p > t_p$, $j+1$ is the first and j is the second variable entering the Lasso path with $W_j = t_p$. Regardless of the relationship between τ_p and t_p , $h = p_4$ is always in the region of rejecting j as a signal. Since $h \sim \mathcal{N}(\mu_4, G)$ with $\mu_4 = (-(1-\rho)\tau_p, (1-\rho)\tau_p, \rho\tau_p, -\rho\tau_p)^T$, by Lemma 1.6.1,

$$\begin{aligned} \mathbb{P}\{W_j \leq t_p | (\beta_j, \beta_{j+1}) = (-\tau_p, \tau_p)\} &\geq L_p p^{-(p_4 - \mu_4)^T G^{-1} (p_4 - \mu_4) / 2 \log(p)} \\ &= L_p p^{-((\xi_\rho \sqrt{r} - \eta_\rho^{-1} \sqrt{u})_+)^2}. \end{aligned}$$

Let $p_5^T = \left(\frac{4\rho^3 - 2\rho^2 + \rho - 1}{2(1-\rho)}\tau_p, (1-\rho)\tau_p, \frac{1+2\rho-4\rho^2}{2}\tau_p, -\frac{\rho^2}{1-\rho}\tau_p \right)$. When $h = p_5$, variable $j+1$ is the first one entering the Lasso path with $W_{j+1} = (1-\rho)\tau_p$, if we slightly increase the value of the third coordinate of p_5 , then it falls in the region of rejecting j as a signal since variable $j+p$ is the second variable entering the Lasso path. This implies $h = p_5$ in on the boundary

of the region that rejects j as a signal, by Lemma 1.6.1,

$$\begin{aligned}\mathbb{P}\{W_j \leq t_p \mid (\beta_j, \beta_{j+1}) = (-\tau_p, \tau_p)\} &\geq L_p \mathfrak{p}^{-(p_5 - \mu_4)^T G^{-1} (p_5 - \mu_4) / 2 \log(\mathfrak{p})} \\ &= L_p \mathfrak{p}^{-\frac{(1-2\rho)^2(1+\rho)}{2(1-\rho)} r}.\end{aligned}$$

Next, we show that

$$\mathbb{P}\{W_j \leq t_p, A \mid (\beta_j, \beta_{j+1}) = (-\tau_p, \tau_p)\} \leq L_p \mathfrak{p}^{-\min\{((\xi_\rho \sqrt{r} - \eta_\rho^{-1} \sqrt{u})_+)^2, \frac{(1-2\rho)^2(1+\rho)}{2(1-\rho)} r, -\mathfrak{g} + f_{\text{Hamm}}^+(u, r, \mathfrak{g})\}}. \quad (\text{A.92})$$

holds for $A = A_j, A_{j+1 \rightarrow j}, A_{j+1 \rightarrow j+p}, A_{j+1 \rightarrow j+p+1}, A_{j+p}, A_{j+p+1}$, which cover all possibilities.

When $A = A_j$ or $A_{j+1 \rightarrow j}$ occurs, as previously discussed, variable j is a false negative when doing variable selection using the bi-variate Lasso $L_{j,j+1}$ given $W_j \leq t_p$, which implies $\mathbb{P}\{W_j \leq t_p, A \mid (\beta_j, \beta_{j+1}) = (-\tau_p, \tau_p)\}$ is upper bounded by the corresponding false negative rate of Lasso, which is $L_p \mathfrak{p}^{-((\xi_\rho \sqrt{r} - \eta_\rho^{-1} \sqrt{u})_+)^2}$.

When $A_{j+1 \rightarrow j+p}$ occurs, the λ_2 of the bi-variate Lasso problem $L_{j+1,j+p}$ is larger than the λ_2 of the bi-variate Lasso problem $L_{j,j+1}$. When $y^T x_{j+1} \geq 0$, we must have

$$\max\left\{\frac{y^T x_j - \rho y^T x_{j+1}}{1 - \rho}, \frac{y^T x_j - \rho y^T x_{j+1}}{-1 - \rho}\right\} < \max\left\{\frac{y^T \tilde{x}_j - \rho y^T x_{j+1}}{1 - \rho}, \frac{y^T \tilde{x}_j - \rho y^T x_{j+1}}{-1 - \rho}\right\}.$$

Therefore, $A_{j+1 \rightarrow j+p}$ implies one of the three following events must occur:

$$y^T x_{j+1} < 0, \frac{y^T x_j - \rho y^T x_{j+1}}{-1 - \rho} < \frac{y^T \tilde{x}_j - \rho y^T x_{j+1}}{1 - \rho}, \frac{y^T x_j - \rho y^T x_{j+1}}{-1 - \rho} < \frac{y^T \tilde{x}_j - \rho y^T x_{j+1}}{-1 - \rho}$$

The probability of these three events are $L_p \mathfrak{p}^{-(1-\rho)^2 r}$, $L_p \mathfrak{p}^{-\frac{(1-2\rho)^2(1+\rho)}{2(1-\rho)} r}$ and $L_p \mathfrak{p}^{-\frac{r}{2}}$, all of which are upper bounded by $L_p \mathfrak{p}^{-\min\{\frac{(1-2\rho)^2(1+\rho)}{2(1-\rho)} r, -\mathfrak{g} + f_{\text{Hamm}}^+(u, r, \mathfrak{g})\}}$.

When $A_{j+1 \rightarrow j+p+1}$ occurs, the λ_2 of the bi-variate Lasso problem $L_{j+1,j+p+1}$ is larger than the

λ_2 of the bi-variate Lasso problem $L_{j,j+1}$. When $y^T x_{j+1} \geq 0$, we must have

$$\max\left\{\frac{y^T x_j - \rho y^T x_{j+1}}{1 - \rho}, \frac{y^T x_j - \rho y^T x_{j+1}}{-1 - \rho}\right\} < |y^T \tilde{x}_{j+1}|.$$

Therefore, $A_{j+1 \rightarrow j+p+1}$ implies one of the three following events must occur:

$$y^T x_{j+1} < 0, \frac{y^T x_j - \rho y^T x_{j+1}}{-1 - \rho} < y^T \tilde{x}_{j+1}, \frac{y^T x_j - \rho y^T x_{j+1}}{-1 - \rho} < -y^T \tilde{x}_{j+1}.$$

The probability of these three events are $L_p \mathfrak{p}^{-(1-\rho)^2 r}$, $L_p \mathfrak{p}^{-\frac{r}{2}}$ and $L_p \mathfrak{p}^{-\frac{(1-2\rho)^2(1+\rho)}{2(1-\rho)} r}$, all of which are upper bounded by $L_p \mathfrak{p}^{-\min\{\frac{(1-2\rho)^2(1+\rho)}{2(1-\rho)} r, -\mathfrak{g} + \mathfrak{f}_{\text{Hamm}}^+(u, r, \mathfrak{g})\}}$.

When A_{j+p} occurs, then $|y^T \tilde{x}_j| > |y^T x_j|$ and $|y^T \tilde{x}_j| > |y^T x_{j+1}|$. If $y^T \tilde{x}_j > 0$, we further have $(y^T \tilde{x}_j - y^T x_{j+1}) + \frac{1}{2\rho+1}(y^T \tilde{x}_j + y^T x_j) > 0$; if $y^T \tilde{x}_j \leq 0$, we further have $y^T \tilde{x}_j + y^T x_{j+1} < 0$.

Therefore,

$$\begin{aligned} & \mathbb{P}\{W_j \leq t_p, A_{j+p} | (\beta_j, \beta_{j+1}) = (-\tau_p, \tau_p)\} \\ & \leq \mathbb{P}\left\{(y^T \tilde{x}_j - y^T x_{j+1}) + \frac{1}{2\rho+1}(y^T \tilde{x}_j + y^T x_j) > 0 | (\beta_j, \beta_{j+1}) = (-\tau_p, \tau_p)\right\} \\ & \quad + \mathbb{P}\{y^T \tilde{x}_j + y^T x_{j+1} < 0 | (\beta_j, \beta_{j+1}) = (-\tau_p, \tau_p)\} \\ & \leq L_p \mathfrak{p}^{-\frac{2(1-2\rho)^2(1+\rho)}{3-4\rho^2} r} + L_p \mathfrak{p}^{-\frac{r}{2}} \leq L_p \mathfrak{p}^{-\min\{\frac{(1-2\rho)^2(1+\rho)}{2(1-\rho)} r, -\mathfrak{g} + \mathfrak{f}_{\text{Hamm}}^+(u, r, \mathfrak{g})\}}. \end{aligned}$$

For $\mathcal{A} = A_{j+p+1}$, (A.92) is immediate due to the symmetry between variable $j+p$ and $j+p+1$.

Now consider the case where β_j takes value in $\{0, -\tau_p\}$ and β_{j+1} takes value in $\{0, \tau_p\}$, this corresponds to the $\rho < 0$ case (we flipped the sign of ρ and β_j simultaneously). By (A.72),

(A.74), (A.79), (A.89) and (A.90), we know

$$\begin{aligned} & \mathbb{P}\{W_j > t_p, \beta_j = 0\} + \mathbb{P}\{W_j \leq t_p, \beta_j = -\tau_p\} \\ & \geq L_p p^{-\min\{c_{\text{Hamm}}^+(u, r, \vartheta), 2\vartheta + ((\xi_\rho \sqrt{r} - \eta_\rho^{-1} \sqrt{u})_+)^2, 2\vartheta + \frac{(1-2|\rho|)^2(1+|\rho|)}{2(1-|\rho|)}r\}}. \end{aligned} \quad (\text{A.93})$$

Meanwhile, (A.72), (A.75), (A.76), (A.80) and (A.91) gives

$$\begin{aligned} & \mathbb{P}\{W_j > t_p, \beta_j = 0\} + \mathbb{P}\{W_j \leq t_p, \beta_j = -\tau_p\} \\ & \leq L_p p^{-\min\{c_{\text{Hamm}}^+(u, r, \vartheta), 2\vartheta + ((\xi_\rho \sqrt{r} - \eta_\rho^{-1} \sqrt{u})_+)^2, 2\vartheta + \frac{(1-2|\rho|)^2(1+|\rho|)}{2(1-|\rho|)}r\}}. \end{aligned} \quad (\text{A.94})$$

Therefore,

$$\begin{aligned} & \mathbb{P}\{W_j > t_p, \beta_j = 0\} + \mathbb{P}\{W_j \leq t_p, \beta_j = -\tau_p\} \\ & = L_p p^{-\min\{c_{\text{Hamm}}^+(u, r, \vartheta), 2\vartheta + ((\xi_\rho \sqrt{r} - \eta_\rho^{-1} \sqrt{u})_+)^2, 2\vartheta + \frac{(1-2|\rho|)^2(1+|\rho|)}{2(1-|\rho|)}r\}}. \end{aligned} \quad (\text{A.95})$$

(A.88) and (A.95) complete the proof for Theorem 1.5.5. \square

A.1.14 PROOF OF THEOREM 1.5.6

The only difference of the conditional knockoff from the Equal-correlated knockoff construction is that $x_j^T \tilde{x}_j$ is changed from 0 to ρ^2 for $j = 1, \dots, p$. Therefore,

$$G = ((1, \rho, \rho^2, \rho)^T, (\rho, 1, \rho, \rho^2)^T, (\rho^2, \rho, 1, \rho)^T, (\rho, \rho^2, \rho, 1)^T)$$

is the new gram matrix for the four-variate Lassos (A.68). We follow the same notations and workflow from the previous proof.

- When $(\beta_j, \beta_{j+1}) = (0, 0)$,

$$\mathbb{P}\{W_j > t_p | (\beta_j, \beta_{j+1}) = (0, 0)\} = L_p p^{-u}. \quad (\text{A.96})$$

Let $p_1 = (t_p, \rho t_p, \rho^2 t_p, \rho t_p)^T$ where $t_p = \sqrt{2u \log(p)}$. When $h = p_1$, variable j is the first one entering the Lasso path. Though $h = p_1$ is in the rejection region, it is also on the boundary of the region that choose variable j as a signal. Since $h \sim \mathcal{N}(\mu_1, G)$ with $\mu_1 = \mathbf{0}$, by Lemma 1.6.1,

$$\mathbb{P}\{W_j > t_p | (\beta_j, \beta_{j+1}) = (0, 0)\} \geq L_p p^{-(p_1 - \mu_1)^T G^{-1} (p_1 - \mu_1) / 2 \log(p)} = L_p p^{-u}.$$

The upper bound is derived exactly the same as (A.73).

- When $(\beta_j, \beta_{j+1}) = (0, \tau_p)$,

$$\mathbb{P}\{W_j > t_p | (\beta_j, \beta_{j+1}) = (0, \tau_p)\} = L_p p^{-(\sqrt{u} - \rho\sqrt{r})^2 - (\xi_\rho \sqrt{r} - \eta_\rho \sqrt{u})_+^2 + (\sqrt{r} - \sqrt{u})_+^2}. \quad (\text{A.97})$$

This time we choose

$$p_2^T = \begin{cases} (t_p, \rho t_p + (1 - \rho^2)\tau_p, \rho^2 t_p + \rho(1 - \rho^2)\tau_p, \rho t_p)^T, & (1 + \rho)\tau_p \leq t_p, \\ (t_p, t_p, \rho t_p, \rho t_p)^T, & \tau_p \leq t_p < (1 + \rho)\tau_p, \\ ((1 - \rho)t_p + \rho\tau_p, \tau_p, \rho\tau_p, \rho(1 - \rho)t_p + \rho^2\tau_p)^T, & t_p < \tau_p. \end{cases}$$

When $h = p_2$ and $t_p \geq \tau_p$, variable j is the first variable entering the four-variate Lasso path with $W_j = t_p$; when $h = p_2$ and $t_p < \tau_p$, variable $j + 1$ is the first and j is the second variable entering the Lasso path with $W_j = t_p$ and $W_{j+1} = \tau_p$. $h = p_2$ is on the boundary of the region that chooses variable j as a signal. Since $h \sim \mathcal{N}(\mu_2, G)$ with $\mu_2 = (\rho\tau_p, \tau_p, \rho\tau_p, \rho^2\tau_p)^T$,

by Lemma 1.6.1,

$$\begin{aligned}\mathbb{P}\{W_j > t_p | (\beta_j, \beta_{j+1}) = (0, \tau_p)\} &\geq L_p \mathfrak{p}^{-(p_2 - \mu_2)^T G^{-1} (p_2 - \mu_2) / 2 \log(\mathfrak{p})} \\ &= L_p \mathfrak{p}^{-(\sqrt{u} - \rho \sqrt{r})^2 - (\xi_\rho \sqrt{r} - \eta_\rho \sqrt{u})_+^2 + (\sqrt{r} - \sqrt{u})_+^2}.\end{aligned}\quad (\text{A.98})$$

Next we show that

$$\mathbb{P}\{W_j > t_p, A | (\beta_j, \beta_{j+1}) = (0, \tau_p)\} \leq L_p \mathfrak{p}^{-(\sqrt{u} - \rho \sqrt{r})^2 - (\xi_\rho \sqrt{r} - \eta_\rho \sqrt{u})_+^2 + (\sqrt{r} - \sqrt{u})_+^2} \quad (\text{A.99})$$

holds for $A = A_{j,j+1}, A_{j,j+p}, A_{j \rightarrow j+p+1}, A_{j+p+1 \rightarrow j}, A_{j+1,j+p+1}$, which covers all possibilities.

When any one of $A_{j,j+1}, A_{j,j+p}, A_{j \rightarrow j+p+1}$ occurs, same as for EC-knockoff, it implies if variable j is a false positive using Knockoff for variable selection, then it is also a false positive when using bi-variate Lasso $L_{j,j+1}$. So $\mathbb{P}\{W_j > t_p, A | (\beta_j, \beta_{j+1}) = (0, \tau_p)\}$ is upper bounded by the corresponding false positive rate of Lasso, which is $L_p \mathfrak{p}^{-(\sqrt{u} - \rho \sqrt{r})^2 - (\xi_\rho \sqrt{r} - \eta_\rho \sqrt{u})_+^2 + (\sqrt{r} - \sqrt{u})_+^2}$, for $A = A_{j,j+1}, A_{j,j+p}, A_{j \rightarrow j+p+1}$.

When $A = A_{j+p+1 \rightarrow j}, j+p+1$ is the first variable entering the model in the four-variate Lasso problem, thus it's also the first variable entering the model in the bi-variate Lasso problem $L_{j+1,j+p+1}$ and $L_{j,j+p+1}$. Variable $j+p+1$ gets picked up as a signal in $L_{j+1,j+p+1}$ implies

$$\begin{aligned}\mathbb{P}\{W_j > t_p, A_{j+p+1 \rightarrow j} | (\beta_j, \beta_{j+1}) = (0, \tau_p)\} &\leq L_p \mathfrak{p}^{-(\sqrt{u} - |\rho^2| \sqrt{r})^2 - (\xi_{\rho^2} \sqrt{r} - \eta_{\rho^2} \sqrt{u})_+^2 + (\sqrt{r} - \sqrt{u})_+^2} \\ &\leq L_p \mathfrak{p}^{-(\sqrt{u} - |\rho| \sqrt{r})^2 - (\xi_\rho \sqrt{r} - \eta_\rho \sqrt{u})_+^2 + (\sqrt{r} - \sqrt{u})_+^2}\end{aligned}$$

when $u \geq (1 + \rho)^2 r$ or $u \leq (1 + \rho^2)^2 r$.

Now consider bi-variate Lasso problem $L_{j,j+p+1}$ given $(1 + \rho^2)^2 r < u < (1 + \rho)^2 r$. Variable $j, j+p+1$ both get picked up as signals with $j+p+1$ entering the model first given $W_j > t_p$. This implies $(y^T x_j, y^T \tilde{x}_{j+1})$ falls in the purple or green region of the right panel of Figure A.1.

Marginally, $(y^T x_j, y^T \tilde{x}_{j+1}) \sim \mathcal{N}((\rho \tau_p, \rho^2 \tau_p)^T, [(1, \rho), (\rho, 1)])$. The point in purple or green region that has the smallest ellipsoid distance to $(\rho \tau_p, \rho^2 \tau_p)^T$ is (t_p, t_p) when $(1 + \rho^2)^2 r < u < (1 + \rho)^2 r$, thus by Lemma 1.6.1,

$$\begin{aligned} \mathbb{P}\{W_j > t_p, A_{j+p+1 \rightarrow j} | (\beta_j, \beta_{j+1}) = (0, \tau_p)\} &\leq L_p \mathfrak{p}^{-(\sqrt{u} - \rho \sqrt{r})^2 - \frac{1 - \rho}{1 + \rho} u} \\ &\leq L_p \mathfrak{p}^{-r + 2\sqrt{ru} - \frac{2}{1 + \rho} u} \\ &= L_p \mathfrak{p}^{-(\sqrt{u} - |\rho| \sqrt{r})^2 - (\xi_\rho \sqrt{r} - \eta_\rho \sqrt{u})_+^2 + (\sqrt{r} - \sqrt{u})_+^2} \end{aligned}$$

for $u \in ((1 + \rho^2)^2 r, (1 + \rho)^2 r)$, which completes the proof of (A.99) for $A = A_{j+p+1 \rightarrow j}$.

When $A_{j+1, j+p+1}$ occurs, consider the bi-variate Lasso problem $L_{j+1, j+p+1}$. In this bi-variate Lasso problem, $\{\lambda_1, \lambda_2\} = \{Z_{j+1}, Z_{j+p+1}\}$, both of which are larger than W_j . Thus in this bi-variate Lasso problem, both variables will be picked up as signals given $W_j > t_p$. So $(y^T x_{j+1}, y^T \tilde{x}_{j+1}) / \sqrt{2 \log(p)}$ falls in one of the four regions in the right panel of Figure A.1 (with $x_{j+1}^T \tilde{x}_{j+1} = \rho^2$ instead of ρ): the purple region, the mirror of purple region against $x = y$, the green region and the mirror of green region against $x = -y$. Since $(y^T x_{j+1}, y^T \tilde{x}_{j+1}) \sim \mathcal{N}((\tau_p, \rho^2 \tau_p)^T, [(1, \rho^2), (\rho^2, 1)])$. By Lemma 1.6.1, we need to find the point in those regions that has the smallest ellipsoid distance to the center $(\tau_p, \rho^2 \tau_p)^T$. When $\tau_p \leq t_p$, this critical point is $(y^T x_{j+1}, y^T \tilde{x}_{j+1}) = (t_p, t_p)$; when $\tau_p > t_p$, this critical point is $(y^T x_{j+1}, y^T \tilde{x}_{j+1}) = (\tau_p, t_p + \rho(\tau_p - t_p))$. So Lemma 1.6.1 gives the probability for λ_1 and λ_2 in $L_{j+1, j+p+1}$ to be both larger than t_p is

$$L_p \mathfrak{p}^{-(\sqrt{u} - \sqrt{r})_+^2 - \frac{1 - \rho^2}{1 + \rho^2} u} \leq L_p \mathfrak{p}^{-(\sqrt{u} - |\rho| \sqrt{r})^2 - (\xi_\rho \sqrt{r} - \eta_\rho \sqrt{u})_+^2 + (\sqrt{r} - \sqrt{u})_+^2}.$$

Since $A_{j+1, j+p+1} \cap \{W_j > t_p\}$ implies $\{\lambda_1 > t_p\} \cap \{\lambda_2 > t_p\}$ in $L_{j+1, j+p+1}$, we know

$$\mathbb{P}\{W_j > t_p, A_{j+1, j+p+1} | (\beta_j, \beta_{j+1}) = (0, \tau_p)\} \leq L_p \mathfrak{p}^{-(\sqrt{u} - |\rho| \sqrt{r})^2 - (\xi_\rho \sqrt{r} - \eta_\rho \sqrt{u})_+^2 + (\sqrt{r} - \sqrt{u})_+^2}.$$

Now, we have verified (A.99). Further coupled with (A.98), we have (A.97).

- When $(\beta_j, \beta_{j+1}) = (\tau_p, 0)$,

$$\mathbb{P}\{W_j \leq t_p | (\beta_j, \beta_{j+1}) = (\tau_p, 0)\} \geq L_p \hat{p}^{-[(\sqrt{r}-\sqrt{u})_+]^2}, \quad (\text{A.100})$$

and

$$\mathbb{P}\{W_j \leq t_p | (\beta_j, \beta_{j+1}) = (\tau_p, 0)\} \leq L_p \hat{p}^{\mathfrak{g} - f_{\text{Hamm}}^+(u, r, \mathfrak{g})}. \quad (\text{A.101})$$

Let $p_3 = (t_p, \rho t_p, \rho^2 t_p, \rho t_p)^T$. when $b = p_3$, variable j is the first variable entering the Lasso path and p_3 is in the region of rejecting variable j as a signal. Since $b \sim \mathcal{N}(\mu_3, G)$ with $\mu_3 = (\tau_p, \rho \tau_p, \rho^2 \tau_p, \rho \tau_p)^T$, by Lemma 1.6.1,

$$\begin{aligned} \mathbb{P}\{W_j \leq t_p | (\beta_j, \beta_{j+1}) = (\tau_p, 0)\} &\geq L_p \hat{p}^{-(p_3 - \mu_3)^T G^{-1} (p_3 - \mu_3) / 2 \log(p)} \\ &= L_p \hat{p}^{-[(\sqrt{r}-\sqrt{u})_+]^2}. \end{aligned}$$

Now, we show that (A.101) holds for $u \geq u^*$, which implies (A.101) for all $u \geq 0$ as discussed in the proof of EC-knockoff. We prove (A.101) by showing that

$$\mathbb{P}\{W_j \leq t_p, A | (\beta_j, \beta_{j+1}) = (\tau_p, 0)\} \leq L_p \hat{p}^{-[(\sqrt{r}-\sqrt{u})_+]^2} \quad (\text{A.102})$$

holds for $A = A_j, A_{j+1}, A_{j+p}, A_{j+p+1}$ given $u \geq u^*$. Respectively,

$$\begin{aligned} \mathbb{P}\{W_j \leq t_p, A_j | (\beta_j, \beta_{j+1}) = (\tau_p, 0)\} &\leq \mathbb{P}\{|y^T x_j| \leq t_p | (\beta_j, \beta_{j+1}) = (\tau_p, 0)\} \\ &= L_p \hat{p}^{-[(\sqrt{r}-\sqrt{u})_+]^2}, \end{aligned}$$

and by symmetry and (A.81),

$$\begin{aligned} \mathbb{P}\{W_j \leq t_p, A_{j+1} | (\beta_j, \beta_{j+1}) = (\tau_p, 0)\} &= \mathbb{P}\{W_j \leq t_p, A_{j+p+1} | (\beta_j, \beta_{j+1}) = (\tau_p, 0)\} \\ &\leq \mathbb{P}\{|y^T x_j| \leq |y^T x_{j+p+1}| | (\beta_j, \beta_{j+1}) = (\tau_p, 0)\} \leq L_p \mathfrak{p}^{-\frac{1-\rho}{2}r} \leq L_p \mathfrak{p}^{-[(\sqrt{r}-\sqrt{u})_+]^2}, \end{aligned}$$

$$\begin{aligned} \mathbb{P}\{W_j \leq t_p, A_{j+p} | (\beta_j, \beta_{j+1}) = (\tau_p, 0)\} &\leq \mathbb{P}\{|y^T x_j| \leq |y^T x_{j+p}| | (\beta_j, \beta_{j+1}) = (\tau_p, 0)\} \\ &\leq L_p \mathfrak{p}^{-\frac{1-\rho^2}{2}r} \leq L_p \mathfrak{p}^{-[(\sqrt{r}-\sqrt{u})_+]^2}. \end{aligned}$$

(A.101) is immediate by $[(\sqrt{r}-\sqrt{u})_+]^2 \geq f_{\text{Hamm}}^+(r, u, \mathfrak{D}) - \mathfrak{D}$.

- When $(\beta_j, \beta_{j+1}) = (\tau_p, \tau_p)$,

$$\mathbb{P}\{W_j \leq t_p | (\beta_j, \beta_{j+1}) = (\tau_p, \tau_p)\} \leq L_p \mathfrak{p}^{\mathfrak{D} - f_{\text{Hamm}}^+(u, r, \mathfrak{D})}. \quad (\text{A.103})$$

We prove (A.103) by showing

$$\mathbb{P}\{W_j \leq t_p, A | (\beta_j, \beta_{j+1}) = (\tau_p, \tau_p)\} \leq L_p \mathfrak{p}^{-[(\sqrt{r}-\sqrt{u})_+]^2} \quad (\text{A.104})$$

holds for $A = A_j, A_{j+1 \rightarrow j}, A_{j+1 \rightarrow j+p}, A_{j+1 \rightarrow j+p+1}, A_{j+p}, A_{j+p+1}$ given $u \geq u^*$, which cover all possibilities. Respectively,

$$\begin{aligned} \mathbb{P}\{W_j \leq t_p, A_j | (\beta_j, \beta_{j+1}) = (\tau_p, \tau_p)\} &\leq \mathbb{P}\{|y^T x_j| \leq t_p | (\beta_j, \beta_{j+1}) = (\tau_p, \tau_p)\} \\ &\leq L_p \mathfrak{p}^{-[(1+\rho)\sqrt{r}-\sqrt{u}]^2} \leq L_p \mathfrak{p}^{-[(\sqrt{r}-\sqrt{u})_+]^2}, \end{aligned}$$

$$\begin{aligned} \mathbb{P}\{W_j \leq t_p, A_{j+p} | (\beta_j, \beta_{j+1}) = (\tau_p, \tau_p)\} &\leq \mathbb{P}\{|y^T x_j| \leq |y^T \tilde{x}_j| | (\beta_j, \beta_{j+1}) = (\tau_p, \tau_p)\} \\ &\leq L_p \mathfrak{p}^{-\frac{1-\rho^2}{2}r} \leq L_p \mathfrak{p}^{-[(\sqrt{r}-\sqrt{u})_+]^2}, \end{aligned}$$

$$\begin{aligned}\mathbb{P}\{W_j \leq t_p, A_{j+p+1} | (\beta_j, \beta_{j+1}) = (\tau_p, \tau_p)\} &\leq \mathbb{P}\{|y^T x_j| \leq |y^T \tilde{x}_{j+1}| | (\beta_j, \beta_{j+1}) = (\tau_p, \tau_p)\} \\ &\leq L_p p^{-\frac{(1-\rho)(1+\rho)^2}{2} r} \leq L_p p^{-[(\sqrt{r}-\sqrt{u})_+]^2}.\end{aligned}$$

When $A_{j+1 \rightarrow j}$ occurs, the bi-variate Lasso problem $L_{j,j+1}$ has variable j is a false negative given $W_j \leq t_p$, which implies $\mathbb{P}\{W_j \leq t_p, A_{j+1 \rightarrow j} | (\beta_j, \beta_{j+1}) = (\tau_p, \tau_p)\}$ is upper bounded by the corresponding false negative rate of Lasso, which is $L_p p^{-\frac{(\xi_\rho \sqrt{r} - \eta_\rho \sqrt{u})_+^2}{2}} \leq L_p p^{-[(\sqrt{r}-\sqrt{u})_+]^2}$ for $u \geq u^*$.

When $A_{j+1 \rightarrow j+p}$ occurs, we know variable $j+p$ instead of variable j is the second one entering the Lasso path. This means the λ_2 (the λ value when the second variable entering Lasso path) of the bi-variate Lasso problem $L_{j+1,j+p}$ is larger than the λ_2 of the bi-variate Lasso problem $L_{j,j+1}$. When $y^T x_{j+1} \geq 0$, we must have

$$\max\left\{\frac{y^T x_j - \rho y^T x_{j+1}}{1-\rho}, \frac{y^T x_j - \rho y^T x_{j+1}}{-1-\rho}\right\} < \max\left\{\frac{y^T \tilde{x}_j - \rho y^T x_{j+1}}{1-\rho}, \frac{y^T \tilde{x}_j - \rho y^T x_{j+1}}{-1-\rho}\right\}.$$

Therefore, $A_{j+1 \rightarrow j+p}$ implies one the three following events must occur:

$$y^T x_{j+1} < 0, \frac{y^T x_j - \rho y^T x_{j+1}}{1-\rho} < \frac{y^T \tilde{x}_j - \rho y^T x_{j+1}}{1-\rho}, \frac{y^T x_j - \rho y^T x_{j+1}}{1-\rho} < \frac{y^T \tilde{x}_j - \rho y^T x_{j+1}}{-1-\rho}$$

The probability of these three events given $(\beta_j, \beta_{j+1}) = (\tau_p, \tau_p)$ are $L_p p^{-(1+\rho)^2 r}$, $L_p p^{-\frac{1-\rho^2}{2} r}$ and $L_p p^{-\frac{(1+\rho)^3(1-\rho)}{2(1+\rho^2)} r}$, all of which are upper bounded by $L_p p^{-[(\sqrt{r}-\sqrt{u})_+]^2}$ when $u \geq u^*$.

When $A_{j+1 \rightarrow j+p+1}$ occurs, the λ_2 of the bi-variate Lasso problem $L_{j+1,j+p+1}$ is larger than the λ_2 of the bi-variate Lasso problem $L_{j,j+1}$. When $y^T x_{j+1} \geq 0$, we must have

$$\max\left\{\frac{y^T x_j - \rho y^T x_{j+1}}{1-\rho}, \frac{y^T x_j - \rho y^T x_{j+1}}{-1-\rho}\right\} < \max\left\{\frac{y^T \tilde{x}_{j+1} - \rho^2 y^T x_{j+1}}{1-\rho^2}, \frac{y^T \tilde{x}_{j+1} - \rho^2 y^T x_{j+1}}{-1-\rho^2}\right\}.$$

Therefore, $A_{j+1 \rightarrow j+p+1}$ implies one the three following events must occur:

$$y^T x_{j+1} < 0, \frac{y^T x_j - \rho y^T x_{j+1}}{1 - \rho} < \frac{y^T \tilde{x}_{j+1} - \rho^2 y^T x_{j+1}}{1 - \rho^2}, \frac{y^T x_j - \rho y^T x_{j+1}}{1 - \rho} < -\frac{y^T \tilde{x}_{j+1} - \rho^2 y^T x_{j+1}}{-1 - \rho^2}.$$

Respectively, the probability of these three events are $L_p p^{-(1+\rho)^2 r}$, $L_p p^{-\frac{1-\rho^2}{2} r}$ and $L_p p^{-\frac{(1+\rho)^3(1-\rho)}{2(1+\rho^2)} r}$, all of which are upper bounded by $L_p p^{-[(\sqrt{r}-\sqrt{u})_+]^2}$ when $u \geq u^*$. From here we have verified (A.104), thus implies (A.103).

From (A.96), (A.97), (A.100), (A.101) and (A.103), we have

$$\mathbb{P}\{W_j > t_p, \beta_j = 0\} + \mathbb{P}\{W_j \leq t_p, \beta_j = \tau_p\} = L_p p^{-f_{\text{Hamm}}^+(r; u, \vartheta)}, \quad (\text{A.105})$$

which completes the proof for positive ρ .

- When $(\beta_j, \beta_{j+1}) = (-\tau_p, \tau_p)$,

$$\mathbb{P}\{W_j \leq t_p | (\beta_j, \beta_{j+1}) = (-\tau_p, \tau_p)\} \geq L_p p^{-((\xi_\rho \sqrt{r} - \eta_\rho^{-1} \sqrt{u})_+)^2} \quad (\text{A.106})$$

and

$$\mathbb{P}\{W_j \leq t_p | (\beta_j, \beta_{j+1}) = (-\tau_p, \tau_p)\} \leq L_p p^{-\min\{((\xi_\rho \sqrt{r} - \eta_\rho^{-1} \sqrt{u})_+)^2, \frac{(1-\rho)^3(1+\rho)}{2(1+\rho^2)} r\}}. \quad (\text{A.107})$$

Let

$$p_4^T = \begin{cases} (-(1-\rho)\tau_p, (1-\rho)\tau_p, \rho(1-\rho)\tau_p, -\rho(1-\rho)\tau_p), & (1-\rho)\tau_p \leq t_p, \\ (\rho(1-\rho)\tau_p - (1+\rho)t_p, (1-\rho)\tau_p, \rho(1-\rho)\tau_p, \rho^2(1-\rho)\tau_p - \rho(1+\rho)t_p), & (1-\rho)\tau_p > t_p. \end{cases}$$

When $h = p_4$ and $(1-\rho)\tau_p \leq t_p$, variable j is the first variable entering the Lasso path with

$W_j = (1 - \rho)\tau_p \leq t_p$; when $b = p_4$ and $(1 - \rho)\tau_p > t_p$, $j + 1$ is the first and j is the second variable entering the Lasso path with $W_j = t_p$. Regardless of the relationship between τ_p and t_p , $b = p_4$ is always in the region of rejecting j as a signal. Since $b \sim \mathcal{N}(\mu_4, G)$ with $\mu_4 = (-(1 - \rho)\tau_p, (1 - \rho)\tau_p, \rho(1 - \rho)\tau_p, -\rho(1 - \rho)\tau_p)^T$, by Lemma 1.6.1,

$$\begin{aligned} \mathbb{P}\{W_j \leq t_p \mid (\beta_j, \beta_{j+1}) = (-\tau_p, \tau_p)\} &\geq L_p p^{-(p_4 - \mu_4)^T G^{-1} (p_4 - \mu_4) / 2 \log(p)} \\ &= L_p p^{-((\xi_\rho \sqrt{r} - \eta_\rho^{-1} \sqrt{u})_+)^2}. \end{aligned}$$

Next, we show that

$$\mathbb{P}\{W_j \leq t_p, A \mid (\beta_j, \beta_{j+1}) = (-\tau_p, \tau_p)\} \leq L_p p^{-\min\{((\xi_\rho \sqrt{r} - \eta_\rho^{-1} \sqrt{u})_+)^2, \frac{(1-\rho)^3(1+\rho)}{2(1+\rho^2)}r\}}. \quad (\text{A.108})$$

holds for $A = A_j, A_{j+1 \rightarrow j}, A_{j+1 \rightarrow j+p}, A_{j+1 \rightarrow j+p+1}, A_{j+p}, A_{j+p+1}$, which cover all possibilities.

When $A = A_j$ or $A_{j+1 \rightarrow j}$ occurs, as previously discussed, variable j is a false negative in the bi-variate Lasso $L_{j,j+1}$ given $W_j \leq t_p$, which implies $\mathbb{P}\{W_j \leq t_p, A \mid (\beta_j, \beta_{j+1}) = (-\tau_p, \tau_p)\}$ is upper bounded by the corresponding false negative rate of Lasso, which is $L_p p^{-((\xi_\rho \sqrt{r} - \eta_\rho^{-1} \sqrt{u})_+)^2}$.

When $A_{j+1 \rightarrow j+p}$ occurs, the λ_2 of the bi-variate Lasso problem $L_{j+1,j+p}$ is larger than the λ_2 of the bi-variate Lasso problem $L_{j,j+1}$. When $y^T x_{j+1} \geq 0$, we must have

$$\max\left\{\frac{y^T x_j - \rho y^T x_{j+1}}{1 - \rho}, \frac{y^T x_j - \rho y^T x_{j+1}}{-1 - \rho}\right\} < \max\left\{\frac{y^T \tilde{x}_j - \rho y^T x_{j+1}}{1 - \rho}, \frac{y^T \tilde{x}_j - \rho y^T x_{j+1}}{-1 - \rho}\right\}.$$

Therefore, $A_{j+1 \rightarrow j+p}$ implies one of the three following events must occur:

$$y^T x_{j+1} + y^T \tilde{x}_j < 0, \frac{y^T x_j - \rho y^T x_{j+1}}{-1 - \rho} < \frac{y^T \tilde{x}_j - \rho y^T x_{j+1}}{1 - \rho}, \frac{y^T x_j - \rho y^T x_{j+1}}{-1 - \rho} < \frac{y^T \tilde{x}_j - \rho y^T x_{j+1}}{-1 - \rho}$$

The probability of these three events are $L_p p^{-\frac{(1+\rho)(1-\rho)^2}{2}r}$, $L_p p^{-\frac{(1-\rho)^3(1+\rho)}{2(1+\rho^2)}r}$ and $L_p p^{-\frac{1-\rho^2}{2}r}$, all of

which are upper bounded by $L_p \mathfrak{p}^{-\frac{(1-\rho)^3(1+\rho)}{2(1+\rho^2)} r}$.

When $A_{j+1 \rightarrow j+p+1}$ occurs, the λ_2 of the bi-variate Lasso problem $L_{j+1, j+p+1}$ is larger than the λ_2 of the bi-variate Lasso problem $L_{j, j+1}$. When $y^T x_{j+1} \geq 0$, we must have

$$\max\left\{\frac{y^T x_j - \rho y^T x_{j+1}}{1 - \rho}, \frac{y^T x_j - \rho y^T x_{j+1}}{-1 - \rho}\right\} < \max\left\{\frac{y^T \tilde{x}_{j+1} - \rho^2 y^T x_{j+1}}{1 - \rho^2}, \frac{y^T \tilde{x}_{j+1} - \rho^2 y^T x_{j+1}}{-1 - \rho^2}\right\}.$$

Therefore, $A_{j+1 \rightarrow j+p+1}$ implies one of the three following events must occur:

$$y^T x_{j+1} + y^T \tilde{x}_j < 0, \frac{y^T x_j - \rho y^T x_{j+1}}{-1 - \rho} < \frac{y^T \tilde{x}_{j+1} - \rho^2 y^T x_{j+1}}{1 - \rho^2}, \frac{y^T x_j - \rho y^T x_{j+1}}{-1 - \rho} < \frac{y^T \tilde{x}_{j+1} - \rho^2 y^T x_{j+1}}{-1 - \rho^2}.$$

The probability of these three events are $L_p \mathfrak{p}^{-\frac{(1+\rho)(1-\rho)^2}{2} r}$, $L_p \mathfrak{p}^{-\frac{1-\rho^2}{2} r}$, $L_p \mathfrak{p}^{-\frac{(1-\rho)^3(1+\rho)}{2(1+\rho^2)} r}$, all of which are upper bounded by $L_p \mathfrak{p}^{-\frac{(1-\rho)^3(1+\rho)}{2(1+\rho^2)} r}$.

When A_{j+p} occurs, if $y^T \tilde{x}_j < 0$, then $y^T x_{j+1} + y^T \tilde{x}_j \leq 0$, which happens with probability $L_p \mathfrak{p}^{-\frac{(1+\rho)(1-\rho)^2}{2} r} \leq L_p \mathfrak{p}^{-\frac{(1-\rho)^3(1+\rho)}{2(1+\rho^2)} r}$. If $y^T \tilde{x}_j \geq 0$, then $y^T \tilde{x}_j + \frac{1-\rho}{2} y^T x_j - \frac{1+\rho}{2} y^T x_{j+1} \geq 0$, which happens with probability $L_p \mathfrak{p}^{-\frac{2(1-\rho)^3}{3+\rho^2} r} \leq L_p \mathfrak{p}^{-\frac{(1-\rho)^3(1+\rho)}{2(1+\rho^2)} r}$. Therefore, (A.108) holds for A_{j+p} and also for A_{j+p+1} due to symmetry. We thus complete the proof for (A.108).

Now consider the case where β_j takes value in $\{0, -\tau_p\}$ and β_{j+1} takes value in $\{0, \tau_p\}$, this corresponds to the $\rho < 0$ case (we flipped the sign of ρ and β_j simultaneously). By (A.96), (A.97), (A.100) and (A.106), we know

$$\begin{aligned} & \mathbb{P}\{W_j > t_p, \beta_j = 0\} + \mathbb{P}\{W_j \leq t_p, \beta_j = -\tau_p\} \\ & \geq L_p \mathfrak{p}^{-\min\{f_{\text{Hamm}}^+(u, r, \vartheta), 2\vartheta + ((\xi_\rho \sqrt{r} - \eta_\rho^{-1} \sqrt{u})_+)^2\}}. \end{aligned} \tag{A.109}$$

Meanwhile, (A.96), (A.97), (A.101) and (A.107) gives

$$\begin{aligned} & \mathbb{P}\{W_j > t_p, \beta_j = 0\} + \mathbb{P}\{W_j \leq t_p, \beta_j = -\tau_p\} \\ & \leq L_p p^{-\min\{f_{\text{Hamm}}^+(u, r, \vartheta), 2\vartheta + ((\xi_\rho \sqrt{r} - \eta_\rho^{-1} \sqrt{u})_+)^2, 2\vartheta + \frac{(1-|\rho|)^3(1+|\rho|)}{2(1+|\rho|^2)}r\}}. \end{aligned} \quad (\text{A.110})$$

The proof is complete once we show that

$$\min\{f_{\text{Hamm}}^+(u, r, \vartheta), 2\vartheta + ((\xi_\rho \sqrt{r} - \eta_\rho^{-1} \sqrt{u})_+)^2\} \leq 2\vartheta + \frac{(1-|\rho|)^3(1+|\rho|)}{2(1+|\rho|^2)}r. \quad (\text{A.111})$$

Otherwise, there exists a tuple of $(\vartheta, r, \rho, u, r)$ such that

$$2\vartheta + \frac{(1-|\rho|)^3(1+|\rho|)}{2(1+|\rho|^2)}r < 2\vartheta + ((\xi_\rho \sqrt{r} - \eta_\rho^{-1} \sqrt{u})_+)^2 \quad (\text{A.112})$$

and

$$2\vartheta + \frac{(1-|\rho|)^3(1+|\rho|)}{2(1+|\rho|^2)}r < \vartheta + (\sqrt{u} - |\rho|\sqrt{r})^2 + ((\xi_\rho \sqrt{r} - \eta_\rho \sqrt{u})_+)^2 - ((\sqrt{r} - \sqrt{u})_+)^2 \quad (\text{A.113})$$

are satisfied simultaneously.

By (A.112), $\xi_\rho \sqrt{r} - \eta_\rho^{-1} \sqrt{u} > 0$, which implies $(1-|\rho|)\sqrt{r} > \sqrt{u}$. Therefore, the right hand side of (A.113) simplifies to $\vartheta + \frac{1-|\rho|}{1+|\rho|}u$. By (A.113), we know

$$\frac{(1-|\rho|)^3(1+|\rho|)}{2(1+|\rho|^2)}r \leq \vartheta + \frac{(1-|\rho|)^3(1+|\rho|)}{2(1+|\rho|^2)}r < \frac{1-|\rho|}{1+|\rho|}u.$$

Plug this into the right hand side of (A.112), we have

$$\begin{aligned}
2\mathfrak{g} + \frac{(1 - |\rho|)^3(1 + |\rho|)}{2(1 + \rho^2)} r &< 2\mathfrak{g} + ((\xi_\rho \sqrt{r} - \eta_\rho^{-1} \sqrt{u})_+)^2 \\
&\leq 2\mathfrak{g} + \left(\sqrt{1 - \rho^2} - \sqrt{\frac{(1 - |\rho|)(1 + |\rho|)^3}{2(1 + \rho^2)}} \right)^2 r,
\end{aligned} \tag{A.114}$$

which can only be true when $\rho^2 > 1$. By reductio, we proved (A.111). \square

B

Supplemental Materials of Chapter 2

B.1 GetQT ALGORITHMS

We present details of the GetQT algorithms used in BEMA. Under the general spiked covariance model (2.7), the empirical spectral distribution (ESD) converges to a fixed distribution $F_\gamma(x; \sigma^2, \theta)$. Write $\gamma_n = p/n$. The purpose of the algorithm $\text{GetQT}(\gamma, \gamma_n, \theta)$ is as follows: Fixing $\sigma = 1$, given any $\theta > 0$ and $y \in [0, 1]$, it outputs the y -upper-quantile of the distribution $F_{\gamma_n}(x; 1, \theta)$.

B.1.1 THE MONTE CARLO SIMULATION ALGORITHM GetQT1

As explained in Section 2.3.1, $F_{\gamma_n}(\cdot; 1, \theta)$ is also the theoretical limit of the ESD under the following null covariance model:

$$\Sigma = \text{diag}(\sigma_1^2, \dots, \sigma_p^2), \quad \text{where } \sigma_k^2 \stackrel{iid}{\sim} \text{Gamma}(\theta, \theta). \quad (\text{B.1})$$

We can simulate data from (B.1) and use its ESD as a numerical approximation to $F_{\gamma_n}(\cdot; 1, \theta)$.

Write $\tilde{p} = \min\{n, p\}$ and $y = k/\tilde{p}$. When the population covariance matrix satisfies (B.1), the k th eigenvalue of the sample covariance matrix, $\hat{\lambda}_k$, is asymptotically close to the y -upper-quantile of $F_{\gamma_n}(\cdot; 1, \theta)$. We thereby use the mean of $\hat{\lambda}_k$, obtained by sampling the data matrix multiple times, to estimate the desired quantile. We note that model (B.1) only specifies how to sample Σ , but it does not specify how to sample \mathbf{X}_i 's. Due to universality theory of eigenvalues (Knowles & Yin, 2017, Section 3.3), the choice of distribution of \mathbf{X}_i 's does not matter. For convenience, we sample \mathbf{X}_i 's from multivariate normal distributions. See Algorithm 3.

Algorithm 3. GetQT1.

Input: n, p, θ, k , and an integer B .

Output: An estimate of the (k/\tilde{p}) -upper-quantile of $F_{\gamma_n}(\cdot; 1, \theta)$.

1. For $b = 1, 2, \dots, B$, repeat: First generate $\Sigma^{(b)}$ from (B.1), and then generate $\mathbf{X}_i^{(b)} \stackrel{iid}{\sim} N(0, \Sigma^{(b)})$, $1 \leq i \leq n$. Write $\mathbf{X}^{(b)} = [\mathbf{X}_1^{(b)}, \dots, \mathbf{X}_n^{(b)}]^\top \in \mathbb{R}^{n \times p}$. Construct the sample covariance matrix $\mathbf{S}^{(b)} = (1/n)(\mathbf{X}^{(b)})^\top \mathbf{X}^{(b)}$ and obtain its k th eigenvalue $\hat{\lambda}_k^{(b)}$.
2. Output $\frac{1}{B} \sum_{b=1}^B \hat{\lambda}_k^{(b)}$ as the estimated (k/\tilde{p}) -upper-quantile.

In the practical implementation, we use the following strategies to further reduce computation

time and memory use: (i) When n is smaller than p , we no longer construct the $p \times p$ covariance matrix $\mathbf{S}^{(b)}$. Instead, we construct an $n \times n$ matrix $(1/n)\mathbf{X}^{(b)}(\mathbf{X}^{(b)})^\top$. This matrix shares the same nonzero eigenvalues as $\mathbf{S}^{(b)}$ but requires much less memory in eigen-decomposition. This strategy is especially useful for genomic data, where n is typically much smaller than p . (ii) In the main algorithm, Algorithm 2, GetQT1 is applied multiple times to compute the (k/\tilde{p}) -upper-quantile for a collection of k . We let the sampling step, Step 1 above, be shared across different values of k : For each $b = 1, 2, \dots, B$, we obtain and store $\hat{\lambda}_k^{(b)}$ for all values of k ; next, in Step 2, we output the estimated (k/\tilde{p}) -upper-quantile simultaneously for all values of k . This strategy can significantly reduce the actual running time.

B.1.2 THE DETERMINISTIC ALGORITHM GetQT2

This algorithm directly uses the definition of $F_{\gamma_n}(\cdot; \cdot, 1, \theta)$. Let $H_\theta(t)$ be the CDF of $\text{Gamma}(\theta, \theta)$. Given a positive sequence ξ_n such that $\xi_n \rightarrow 0$ as $n \rightarrow \infty$, let $m_n(y) = m_n(y, \xi_n, \gamma_n, \theta) \in \mathbb{C}^+$ be the unique solution to the equation

$$y + i\xi_n = -\frac{1}{m_n} + \gamma_n \int \frac{t}{1 + tm_n} dH_\theta(t). \quad (\text{B.2})$$

Then, the density of $F_{\gamma_n}(\cdot; 1, \theta)$, denoted by $f_{\gamma_n}(y; 1, \theta)$, is approximated by

$$\hat{f}_{\gamma_n}^*(y; 1, \theta) = \frac{1}{\pi(\gamma_n \wedge 1)} \Im(m_n(y, \xi_n, \gamma_n, \theta)), \quad (\text{B.3})$$

where $\Im(\cdot)$ denotes the imaginary part of a complex number. The choice of ξ_n needs to satisfy $\xi_n \gg n^{-1}$, in order to guarantee that the approximation is not governed by stochastic fluctuations (Knowles & Yin, 2017). We choose $\xi_n = n^{-2/3}$ for convenience.

The above motivates a three-step algorithm.

1. Fix a grid $y_1 < y_2 < \dots < y_N$. Solve equation (B.2) to obtain $m_n(y_j)$ for $1 \leq j \leq N$.
2. Use equation (B.3) to obtain $\hat{f}_{\gamma_n}^*(y_j; 1, \theta)$, for $1 \leq j \leq N$. Obtain the whole density curve $\hat{f}_{\gamma_n}(y; 1, \theta)$ by linear interpolation.
3. Find q such that $\int_q^{(1+\sqrt{\gamma_n})^2} \hat{f}_{\gamma_n}(z; 1, \theta) dz = y$. Output q as the estimated y -upper-quantile.

Step 2 is straightforward. Step 3 is also easy to implement, since $\hat{f}_{\gamma_n}(y; 1, \theta)$ is a piece-wise linear function. Below, we describe Step 1 with more details.

In Step 1, fix y and write $m = a + bi$, where $i = \sqrt{-1}$, and $a \in \mathbb{R}$ and $b \in \mathbb{R}^+$ are the real and imaginary parts of m , respectively. We aim to find (a, b) so that m solves the complex equation (B.2). Pretending that $\xi_n = 0$, the equation (B.2) can be re-written as a set of real equations: *

$$\begin{cases} y = \gamma_n \int \frac{t}{1+2at+(a^2+b^2)t^2} dH_\theta(t), \\ \frac{1}{a^2+b^2} = \gamma_n \int \frac{t^2}{1+2at+(a^2+b^2)t^2} dH_\theta(t), \end{cases} \iff \begin{cases} 2ay = \gamma_n \int \frac{2at}{1+2at+(a^2+b^2)t^2} dH_\theta(t), \\ 1 = \gamma_n \int \frac{(a^2+b^2)t^2}{1+2at+(a^2+b^2)t^2} dH_\theta(t). \end{cases}$$

First, by combining the above equations with $\gamma_n = \gamma_n \int \frac{1+2at+(a^2+b^2)t^2}{1+2at+(a^2+b^2)t^2} dH_\theta(t)$, we have

$$\gamma_n - 1 - 2ay = \gamma_n \int \frac{1}{1+2at+(a^2+b^2)t^2} dH_\theta(t) > 0.$$

It yields that $a < (\gamma_n - 1)/2y$. Second, by Cauchy-Schwarz inequality, $[\int \frac{t}{1+2at+(a^2+b^2)t^2} dH_\theta(t)]^2 \leq \int \frac{1}{1+2at+(a^2+b^2)t^2} dH_\theta(t) \cdot \int \frac{t^2}{1+2at+(a^2+b^2)t^2} dH_\theta(t)$. It follows that

$$y^2 \leq (\gamma_n - 1 - 2ay) \cdot \frac{1}{a^2 + b^2}.$$

*The second equation is obtained by letting the imaginary part of both hand sides of (B.2) be equal. The first equation is obtained by letting the real part of both hand sides of (B.2) be equal and then substituting $\frac{a}{a^2+b^2}$ by a times the second equation.

Algorithm 4. GetQT2.

Input: n, p, θ , and $y \in [0, 1]$.

Output: An estimate of the y -upper-quantile of $F_{\gamma_n}(\cdot; 1, \theta)$.

Step 1: Write $\tilde{p} = n \wedge p$ and $\gamma_n = p/n$. Fix a grid $y_1 < y_2 < \dots < y_{N-1} < y_N$. For each $1 \leq j \leq N$, compute $\hat{m}_n(y)$ as follows:

- For a tuning parameter $\delta > 0$, construct the set of grid points in $\mathbb{R} \times \mathbb{R}^+$:

$$S_{y, \gamma_n, \delta} = \{(a, b) : a = k\delta, b = \ell\delta, k, \ell \in \mathbb{Z}, (a - 1/y_j)^2 + b^2 \leq \gamma_n/y_j^2, \\ a < (\gamma_n - 1)/2y_j\}.$$

- For each $(a, b) \in S_{y, \gamma_n, \delta}$ and $\xi_n = n^{-2/3}$, compute

$$\Delta(a, b) = \left| y + i\xi_n + \frac{1}{m} - \gamma_n \int \frac{t}{1 + tm} dH_\theta(t) \right|,$$

where $H_\theta(t)$ is the CDF of $\text{Gamma}(\theta, \theta)$. The integral above can be computed via standard Monte Carlo approximation (by sampling data from $\text{Gamma}(\theta, \theta)$).

- Find $(\hat{a}, \hat{b}) = \operatorname{argmin}_{(a, b) \in S_{y, \gamma_n, \delta}} \Delta(a, b)$. Let $\hat{m}(y) = \hat{a} + \hat{b}i$.

Step 2: Let $\hat{f}_{\gamma_n}(y_j; 1, \theta) = \frac{1}{\pi(\gamma_n \wedge 1)} \Im(\hat{m}(y))$, for $1 \leq j \leq N$. For any $y_{j-1} < z < y_j$, let

$$\hat{f}_{\gamma_n}(z; 1, \theta) = \frac{y_j - z}{y_j - y_{j-1}} \hat{f}_{\gamma_n}(y_{j-1}; 1, \theta) + \frac{z - y_{j-1}}{y_j - y_{j-1}} \hat{f}_{\gamma_n}(y_j; 1, \theta).$$

Step 3: Find q such that $\int_q^{(1+\sqrt{\gamma_n})^2} \hat{f}_{\gamma_n}(z; 1, \theta) = y$. Output q as the estimated y -upper-quantile.

Re-arranging the terms gives $(a - 1/y)^2 + b^2 \leq \gamma_n/y^2$. So far, we have obtained a feasible set of (a, b)

for the solution of (B.2) when $\xi_n = 0$:

$$\mathcal{S}_{y,\gamma_n} = \{(a, b) : (a - 1/y)^2 + b^2 \leq \gamma_n/y^2, a < (\gamma_n - 1)/2y\}. \quad (\text{B.4})$$

Since ξ_n is very close to 0, we use the same feasible set when solving (B.2). Observing that \mathcal{S}_{y,γ_n} is bounded, we solve equation (B.2) by a grid search on this feasible set. See Algorithm 4.

B.1.3 COMPARISON

We compare the performance of two GetQT algorithms on a numerical example where $(n, p, \theta) = (10000, 1000, 1)$. The results are in Figure B.1. To generate this figure, first, we simulate eigenvalues $\{\hat{\lambda}_k^{(b)}\}_{1 \leq k \leq p, 1 \leq b \leq B}$ as in Step 1 of GetQT1, where $B = 20$, and plot the histogram of eigenvalues. Next, we plot the estimated density $\hat{f}_{\gamma_n}(y; 1, \theta)$ from GetQT2 (tuning parameter is $\delta = 0.05$). The estimated density fits the histogram well, suggesting that the steps in GetQT2 for estimating $f_{\gamma_n}(y; 1, \theta)$ are successful. Furthermore, the estimated quantiles from two algorithms are very close to each other.

In terms of numerical performance, the two GetQT algorithms are similar. We now discuss the computing time. The main computational cost of GetQT1 comes from computing the eigenvalues of $\mathbf{S}^{(b)}$ at each iteration. As we have mentioned in the end of Section B.1.1, if $p < n$, we conduct eigen-decomposition on an $p \times p$ matrix; if $n < p$, we conduct eigen-decomposition on an $n \times n$ matrix. Therefore, as long as $\min\{n, p\}$ is not too large, GetQT1 is fast.

Compared with GetQT1, the advantage of GetQT2 is that it does not need to compute any eigen-decomposition. As a result, when $\min\{n, p\}$ is large, GetQT2 is much faster than GetQT1 (and GetQT2 also requires less memory use). The computational cost of GetQT2 is proportional to the number of grid points in the algorithm, governed by the tuning parameter δ . Sometimes, we need to choose δ sufficiently small to guarantee the accuracy of computing $\hat{m}(y, \gamma_n, \theta)$, which significantly increases the cost of grid search. Our experience suggests that GetQT2 is faster than GetQT1 only in the case that

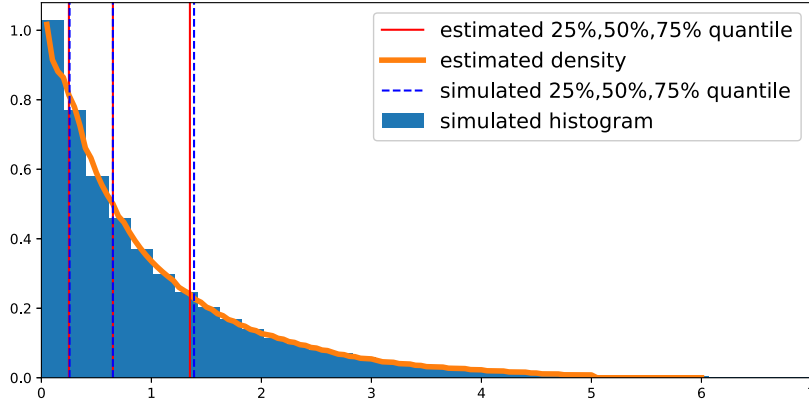


Figure B.1: Comparison of two GetQT algorithms. The simulated histogram is from GetQT1, and the density curve is estimated by GetQT2.

$\min\{n, p\}$ is larger than 10^4 .

B.1.4 MODIFICATIONS UNDER MODEL (2.13)

Section 2.4.2 introduces Model (2.13), as a proxy of Model (2.2), to facilitate the theoretical analysis. In Model (2.13), the diagonal entries of \mathbf{D} are *iid* generated from a truncated Gamma distribution. In Section 2.4.2, we described how to adapt Algorithm 2 to this setting, where the key is to modify GetQT so that it can compute the γ -upper-quantile of the distribution $F_\gamma(\cdot; 1, \theta, T_1, T_2)$, for any given γ and (θ, T_1, T_2) .

To modify GetQT1, we note that $F_{\gamma_n}(\cdot; 1, \theta, T_1, T_2)$ is the theoretical limit of the ESD under the null covariance model:

$$\Sigma = \text{diag}(\sigma_1^2, \dots, \sigma_p^2), \quad \text{where } \sigma_k^2 \stackrel{iid}{\sim} \text{TruncGamma}(\theta, \theta, T_1, T_2). \quad (\text{B.5})$$

We can simulate data from (B.5) and use its ESD as a numerical approximation to $F_{\gamma_n}(\cdot; 1, \theta, T_1, T_2)$. In Algorithm 3, we only need to modify Step 1 so that $\Sigma^{(b)}$ is generated from (B.5).

To modify GetQT2, we solve (B.2) with $H_\theta(t)$ replaced by $H_{\theta, T_1, T_2}(t)$, where $H_{\theta, T_1, T_2}(\cdot)$ is the

CDF of $\text{TruncGamma}(\theta, \theta, T_1, T_2)$. We note that the feasible set in (B.4) is derived without using the explicit form of $H_\theta(t)$, so it continues to apply. In Algorithm 4, we only need to modify the definition of $\Delta(a, b)$ to

$$\Delta(a, b) = \left| y + i\xi_n + \frac{1}{m} - \gamma_n \int \frac{t}{1+tm} dH_{\theta, T_1, T_2}(t) \right|,$$

and the other steps remain the same.

B.2 PROOFS

B.2.1 PROOF OF THEOREM 2.4.1

Let $z_k = \hat{\lambda}_k - \sigma^2 q_k$, for all $1 \leq k \leq \tilde{p}$. It follows that

$$\hat{\sigma}^2 = \frac{\sum_{\alpha\tilde{p} \leq k \leq (1-\alpha)\tilde{p}} q_k (\sigma^2 q_k + z_k)}{\sum_{\alpha\tilde{p} \leq k \leq (1-\alpha)\tilde{p}} q_k^2} = \sigma^2 + \frac{\sum_{\alpha\tilde{p} \leq k \leq (1-\alpha)\tilde{p}} q_k z_k}{\sum_{\alpha\tilde{p} \leq k \leq (1-\alpha)\tilde{p}} q_k^2}.$$

It follows that

$$|\hat{\sigma}^2 - \sigma^2| \leq \underbrace{\frac{\sum_{\alpha\tilde{p} \leq k \leq (1-\alpha)\tilde{p}} |q_k|}{\sum_{\alpha\tilde{p} \leq k \leq (1-\alpha)\tilde{p}} q_k^2}}_{\equiv B_{n,p}(\alpha)} \times \max_{\alpha\tilde{p} \leq k \leq (1-\alpha)\tilde{p}} |z_k|.$$

We recall that q_k is the (k/\tilde{p}) -upper-quantile of a standard Machedenko-Pastur distribution associated with $\gamma_n = p/n$. Note that $p/n \rightarrow \gamma$ and $\alpha \leq k/\tilde{p} \leq 1 - \alpha$, where $\gamma > 0$ and $\alpha \in (0, 1/2)$ are constants. It follows immediately that there is a constant $C_1 = C_1(\alpha, \gamma)$ such that $B_{n,p}(\alpha) \leq C_1$. As a result,

$$|\hat{\sigma}^2 - \sigma^2| \leq C_1 \max_{\alpha\tilde{p} \leq k \leq (1-\alpha)\tilde{p}} |\hat{\lambda}_k - \sigma^2 q_k|. \quad (\text{B.6})$$

We bound the right hand side of (B.6). By Assumption 2.4.1, the data vectors $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n$ are obtained from a random matrix $\mathbf{Y} = [\mathbf{Y}_1, \mathbf{Y}_2, \dots, \mathbf{Y}_n]^\top \in \mathbb{R}^{n \times p}$, where the entries of \mathbf{Y} are

independent variables with zero mean and unit variance. Given \mathbf{Y} , define $\mathbf{X}_1^*, \mathbf{X}_2^*, \dots, \mathbf{X}_n^*$ by

$$\mathbf{X}_i^*(j) = \sigma \cdot \mathbf{Y}_i(j), \quad 1 \leq i \leq n, 1 \leq j \leq p.$$

Then, $\mathbf{X}_1^*, \dots, \mathbf{X}_n^*$ follow a “null” model that is similar to the factor model in Assumption 2.4.1 but corresponds to $K = 0$. Let \mathbf{S}^* be the sample covariance matrix of $\mathbf{X}_1^*, \dots, \mathbf{X}_n^*$. Then, \mathbf{S}^* serves as a reference matrix for \mathbf{S} . The *eigenvalue sticking* result says that eigenvalues of \mathbf{S} “stick” to eigenvalues of the reference matrix. The precise statement is as follows: Let $\hat{\lambda}_1^* > \hat{\lambda}_2^* > \dots > \hat{\lambda}_{\tilde{p}}^*$ be the nonzero eigenvalues of \mathbf{S}^* . When the entries of \mathbf{Y} satisfy the regularity conditions stated in Theorem 2.4.1, by Theorem 2.7 of [Bloemendal et al. \(2016\)](#), there is a constant $C_2 = C_2(\alpha, \gamma, \sigma^2)$ such that, for any $\varepsilon > 0$ and $s > 0$,

$$\mathbb{P}\left\{\max_{(\alpha/2)\tilde{p} \leq j \leq (1-\alpha/2)\tilde{p}} |\hat{\lambda}_{j+K_1} - \hat{\lambda}_j^*| > C_2 n^{-(1-\varepsilon)}\right\} \leq n^{-s}, \quad (\text{B.7})$$

where K_1 is the total number of spiked eigenvalues in Model (2.3) such that $\lambda_k = \sigma^2(\sqrt{\gamma} + \tau_k)$ for some $\tau_k \geq n^{-1/3}$. It remains to study $\hat{\lambda}_j^*$. Its large deviation bound can be found in [Pillai & Yin \(2014\)](#) (also, see Theorem 3.3 of [Ke \(2016\)](#)). There is a constant $C_3 = C_3(\alpha, \gamma, \sigma^2) > 0$ such that, for any $\varepsilon > 0$ and $s > 0$,

$$\mathbb{P}\left\{\max_{(\alpha/2)\tilde{p} \leq j \leq (1-\alpha/2)\tilde{p}} |\hat{\lambda}_j^* - \sigma^2 q_j| > C_3 n^{-(1-\varepsilon)}\right\} \leq n^{-s}. \quad (\text{B.8})$$

Furthermore, since $K_1 \leq K$ and K is fixed, there is a constant $C_4 = C_4(\gamma, K)$ such that

$$\max_{(\alpha/2)\tilde{p} \leq j \leq (1-\alpha/2)\tilde{p}} |q_j - q_{j+K_1}| \leq C_4 n^{-1}. \quad (\text{B.9})$$

Combining (B.7)-(B.9) gives that, for any $\varepsilon > 0$ and $s > 0$,

$$\mathbb{P}\left\{\max_{(\alpha/2)\bar{p} \leq j \leq (1-\alpha)\bar{p}} |\hat{\lambda}_{j+K_1} - \sigma^2 q_{j+K_1}| > Cn^{-(1-\varepsilon)}\right\} \leq n^{-s}.$$

We plug it into (B.6). The claim follows immediately. \square

B.2.2 PROOF OF THEOREM 2.4.2

Denote by $T_{n,p}(\hat{\sigma}^2, \beta_n)$ the threshold used in Algorithm 1. It satisfies that

$$T_{n,p}(\hat{\sigma}^2, \beta_n) = \hat{\sigma}^2[(1 + \sqrt{\gamma_n})^2 + \omega_n], \quad \text{where } \omega_n = O(n^{-2/3}t_{1-\beta_n}). \quad (\text{B.10})$$

Here, $t_{1-\beta_n}$ is the $(1 - \beta_n)$ -quantile of Tracy-Widom distribution. Note that $\tau_n \gg n^{-1/3}$. We can choose $\beta_n \rightarrow \infty$ appropriately slow such that $1 \ll t_{1-\beta_n} \ll n^{2/3} \min\{\tau_n^2, 1\}$. It follows that

$$n^{-2/3} \ll \omega_n \ll \min\{\tau_n^2, 1\}. \quad (\text{B.11})$$

First, we derive a lower bound for $\hat{\lambda}_K$ and show that $\hat{K} \geq K$ with probability $1 - o(1)$. Recall that λ_k denotes the k th largest eigenvalue of Σ . In view of Model (2.3), it is true that $\lambda_k = \mu_k + \sigma^2$ for $1 \leq k \leq K$ and $\lambda_k = \sigma^2$, for $K < k \leq p$. Introduce

$$\lambda_k^* = \lambda_k \left(1 + \frac{\gamma_n}{\lambda_k/\sigma^2 - 1}\right), \quad 1 \leq k \leq K.$$

Write $\delta_k = \lambda_k/\sigma^2 - 1$, for $k = 1, 2, \dots, K$. Let $g(t) = (1+t)(1+\gamma_n/t)$. Then,

$$\lambda_k^* = \sigma^2 \cdot g(\delta_k), \quad 1 \leq k \leq K.$$

The function g satisfies that $g(\sqrt{\gamma_n}) = (1 + \sqrt{\gamma_n})^2$ and $g'(t) \geq 1 - \sqrt{\gamma_n}/t$. Hence, it is monotone increasing in $(\sqrt{\gamma_n}, \infty)$. For any $\tau > 0$ and $t > \sqrt{\gamma_n} + \tau$, we have $g(t) \geq g(\sqrt{\gamma_n}) + g'(\sqrt{\gamma_n} + \tau) \cdot \tau \geq (1 + \sqrt{\gamma_n})^2 + \tau^2/(\sqrt{\gamma_n} + \tau)$. It follows that

$$\lambda_K^* \geq \sigma^2 \left[(1 + \sqrt{\gamma_n})^2 + \frac{\delta_K^2}{\sqrt{\gamma_n} + \delta_K} \right]. \quad (\text{B.12})$$

At the same time, by Theorem 2.3 of [Bloemendal et al. \(2016\)](#), with probability $1 - o(1)$,

$$|\hat{\lambda}_K - \lambda_K^*| \leq C_2 \sigma^2 n^{-1/2} \begin{cases} \delta_K^{1/2}, & \text{if } \delta_K < 1, \\ 1 + \delta_K/(1 + \sqrt{\gamma_n}), & \text{if } \delta_K \geq 1, \end{cases} \quad (\text{B.13})$$

for a constant $C_2 > 0$. If $\delta_K \geq 1$, then (B.12) implies $\lambda_K^* - \sigma^2(1 + \sqrt{\gamma_n})^2 \geq C_3 \sigma^2 \delta_K$, for a constant $C_3 > 0$, and (B.13) yields that $|\hat{\lambda}_K - \lambda_K^*| \leq C_2 \sigma^2 (1 + \delta_K) n^{-1/2}$. It follows that

$$\hat{\lambda}_K - \sigma^2(1 + \sqrt{\gamma_n})^2 \geq (C_3/2) \cdot \sigma^2 \delta_K \geq (C_3/2) \cdot \sigma^2.$$

If $\delta_K < 1$, then (B.12) yields that $\lambda_K^* - \sigma^2(1 + \sqrt{\gamma_n})^2 \geq C_4 \sigma^2 \delta_K^2$, for a constant $C_4 > 0$, and (B.13) yields that $|\hat{\lambda}_K - \lambda_K^*| \leq C_2 \sigma^2 \delta_K^{1/2} n^{-1/2}$. It follows that

$$\hat{\lambda}_K - \sigma^2(1 + \sqrt{\gamma_n})^2 \geq C_4 \sigma^2 \delta_K^2 - \frac{C_2 \sigma^2 \delta_K^2}{\sqrt{n \delta_K^3}} \geq (C_4/2) \cdot \sigma^2 \delta_K^2,$$

where the last inequality is because $\delta_K \geq \tau_n \gg n^{-1/3}$. We combine the two cases and note that $\delta_K \geq \tau_n$. It gives that

$$\mathbb{P} \left\{ \hat{\lambda}_K \geq \sigma^2 \left[(1 + \sqrt{\gamma_n})^2 + C \min\{\tau_n^2, 1\} \right] \right\} = 1 - o(1).$$

Furthermore, by Theorem 2.4.1, $|\hat{\sigma}^2 - \sigma^2| \prec n^{-1} \ll \min\{\tau_n^2, 1\}$. Hence, we can replace σ^2 by $\hat{\sigma}^2$ in the above equation, i.e.,

$$\mathbb{P}\left\{\hat{\lambda}_K \geq \hat{\sigma}^2 [(1 + \sqrt{\gamma_n})^2 + C \min\{\tau_n^2, 1\}]\right\} = 1 - o(1). \quad (\text{B.14})$$

We compare $\hat{\lambda}_K$ with the threshold in (B.10). Since $\omega_n \ll \min\{\tau_n^2, 1\}$, it is implied from (B.14) that $\hat{\lambda}_K$ exceeds this threshold with probability $1 - o(1)$. Therefore,

$$\mathbb{P}\left\{\hat{K} \geq K\right\} = 1 - o(1).$$

Next, we derive an upper bound for $\hat{\lambda}_{K+1}$ and show that $\hat{K} \leq K$ with probability $1 - o(1)$. We apply Theorem 2.3 of [Bloemendal et al. \(2016\)](#) again: For any $\varepsilon > 0$ and $s > 0$,

$$\mathbb{P}\left\{\hat{\lambda}_{K+1} - \sigma^2(1 + \sqrt{\gamma_n})^2 \leq \sigma^2 n^{-(2/3-\varepsilon)}\right\} = 1 - o(1). \quad (\text{B.15})$$

Since $\omega_n \gg n^{-2/3}$, we can take ε arbitrarily small to make $n^{-(2/3-\varepsilon)} \leq \omega_n/2$. We also apply the large deviation bound for $\hat{\sigma}^2$ in Theorem 2.4.1 to replace σ^2 by $\hat{\sigma}^2$. It follows immediately that

$$\mathbb{P}\left\{\hat{\lambda}_{K+1} \leq \hat{\sigma}^2 [(1 + \sqrt{\gamma_n})^2 + \omega_n/2]\right\} = 1 - o(1). \quad (\text{B.16})$$

We compare $\hat{\lambda}_{K+1}$ with the threshold in (B.10). It is seen that $\hat{\lambda}_{K+1}$ is below this threshold with probability $1 - o(1)$. Therefore,

$$\mathbb{P}\left\{\hat{K} \leq K\right\} = 1 - o(1).$$

The claim follows immediately. \square

B.2.3 PROOF OF THEOREM 2.4.3

Throughout this proof, we let C be a generic constant, whose meaning may vary from occurrence to occurrence. Let $F_\gamma(\cdot; \sigma^2, \theta, T_1, T_2)$ be the theoretical limit of ESD, whose definition is given in Lemma 2.4.1. We replace γ by $\gamma_n = p/n$ in this definition, write $\bar{F}_{\gamma_n} = 1 - F_{\gamma_n}$ and let $q_i(\sigma^2, \theta) = \bar{F}_{\gamma_n}^{-1}(\gamma; \sigma^2, \theta, T_1, T_2)$ denote the (i/\tilde{p}) -upper-quantile of this distribution, where $\tilde{p} = n \wedge p$. We use (σ_0^2, θ_0) to denote the true parameters. Write $s_n = \lceil \alpha \tilde{p} \rceil$ and

$$\hat{R}(\sigma^2, \theta) = \sum_{s_n \leq i \leq \tilde{p} - s_n} [\hat{\lambda}_i - q_i(\sigma^2, \theta)]^2, \quad R(\sigma^2, \theta) = \sum_{s_n \leq i \leq \tilde{p} - s_n} [q_i(\sigma_0^2, \theta_0) - q_i(\sigma^2, \theta)]^2.$$

Let $\Delta = \sum_{s_n \leq i \leq \tilde{p} - s_n} |\hat{\lambda}_i - q_i(\sigma_0^2, \theta_0)|^2$. By direct calculations and Cauchy-Schwarz inequality,

$$\begin{aligned} |\hat{R}(\sigma^2, \theta) - R(\sigma^2, \theta)| &\leq 2 \sum_{s_n \leq i \leq \tilde{p} - s_n} |q_i(\sigma_0^2, \theta_0) - q_i(\sigma^2, \theta)| \cdot |\hat{\lambda}_i - q_i(\sigma_0^2, \theta_0)| \\ &\quad + \sum_{s_n \leq i \leq \tilde{p} - s_n} |\hat{\lambda}_i - q_i(\sigma_0^2, \theta_0)|^2 \\ &\leq 2\sqrt{R(\sigma^2, \theta)}\sqrt{\Delta} + \Delta. \end{aligned}$$

It follows that $\hat{R}(\sigma^2, \theta) \leq R(\sigma^2, \theta) + 2\sqrt{R(\sigma^2, \theta)}\sqrt{\Delta} + \Delta = (\sqrt{R(\sigma^2, \theta)} + \sqrt{\Delta})^2$. In the above inequality, we can switch $\hat{R}(\sigma^2, \theta)$ and $R(\sigma^2, \theta)$ and similarly derive that $R(\sigma^2, \theta) \leq (\sqrt{\hat{R}(\sigma^2, \theta)} + \sqrt{\Delta})^2$. As a result,

$$\left| \sqrt{\hat{R}(\sigma^2, \theta)} - \sqrt{R(\sigma^2, \theta)} \right| \leq \sqrt{\Delta}. \quad (\text{B.17})$$

We now bound Δ . By Lemma 2.4.1, for all $K < i \leq \tilde{p}$,

$$|\hat{\lambda}_i - q_i(\sigma_0^2, \theta_0)| \prec [i \wedge (\tilde{p} + 1 - i)]^{-1/3} n^{-2/3}.$$

We note that the stochastic dominance in Lemma 2.4.1 can be made ‘uniform’ over i ; i.e., the integer $N(\varepsilon, s)$ in Definition 2.4.1 is shared by all $K < i \leq \tilde{p}$ (Knowles & Yin, 2017). Hence, summing over i preserves ‘stochastic dominance.’ Additionally, $\sum_{i=s_n}^{\tilde{p}/2} i^{-2/3} n^{-4/3} \leq Cn^{-1} \left[\frac{1}{\tilde{p}} \sum_{i=s_n}^{\tilde{p}/2} (i/\tilde{p})^{-2/3} \right] \leq Cn^{-1} \int_{s_n/n}^{1/2} x^{-2/3} dx \leq Cn^{-1}$. Combining the above arguments gives

$$\begin{aligned} \sum_{s_n \leq i \leq \tilde{p}-s_n} |\hat{\lambda}_i - q_i(\sigma_0^2, \theta_0)|^2 &\prec \sum_{s_n \leq i \leq \tilde{p}-s_n} [i \wedge (\tilde{p} + 1 - i)]^{-2/3} n^{-4/3} \\ &\prec \sum_{s_n \leq i \leq \tilde{p}/2} i^{-2/3} n^{-4/3} \prec n^{-1}. \end{aligned}$$

This gives $\Delta \prec n^{-1}$. We plug it into (B.17) to get

$$\left| \sqrt{\hat{R}(\sigma^2, \theta)} - \sqrt{R(\sigma^2, \theta)} \right| \prec n^{-1/2}. \quad (\text{B.18})$$

Since Δ does not depend on (σ^2, θ) , the ‘stochastic dominance’ here is uniform for all $(\sigma^2, \theta) \in \mathcal{J}_{\sigma^2} \times \mathcal{J}_{\theta}$. We claim that there exists a constant $c_0 > 0$ such that for any (σ^2, θ) in $\mathcal{J}_{\sigma^2} \times \mathcal{J}_{\theta}$,

$$R(\sigma^2, \theta) \geq c_0 n \cdot [(\sigma^2 - \sigma_0^2)^2 + (\theta - \theta_0)^2]. \quad (\text{B.19})$$

Note that $R(\sigma_0^2, \theta_0) = 0$. Combining it with (B.18)-(B.19) gives

$$\sqrt{\hat{R}(\sigma_0^2, \theta_0)} \prec n^{-1/2}, \quad \sqrt{c_0 n} \sqrt{(\hat{\sigma}^2 - \sigma_0^2)^2 + (\hat{\theta} - \theta_0)^2} \leq \sqrt{\hat{R}(\hat{\sigma}^2, \hat{\theta})} + O_{\prec}(n^{-1/2}),$$

where a random variable is $O_{\prec}(b_n)$ if its absolute value is $\prec b_n$. Since $(\hat{\sigma}^2, \hat{\theta})$ minimizes $\hat{R}(\sigma^2, \theta)$, we have $\hat{R}(\hat{\sigma}^2, \hat{\theta}) \leq \hat{R}(\sigma_0^2, \theta_0) \prec n^{-1}$. It follows that

$$\sqrt{(\hat{\sigma}^2 - \sigma_0^2)^2 + (\hat{\theta} - \theta_0)^2} \prec n^{-1}.$$

This proves the claim.

What remains is to show (B.19). Define the quantile function $b_{\sigma^2, \theta}(\alpha) = \bar{F}_{\gamma_n}^{-1}(\alpha; \sigma^2, \theta, T_1, T_2)$. Then, $q_i(\sigma^2, \theta) = b_{\sigma^2, \theta}(i/\tilde{p})$. We can re-write

$$R(\sigma^2, \theta) = \sum_{i=s_n}^{\tilde{p}-s_n} [b_{\sigma^2, \theta}(i/\tilde{p}) - b_{\sigma_0^2, \theta_0}(i/\tilde{p})]^2.$$

Introduce $R^*(\sigma^2, \theta) = \tilde{p} \int_0^1 [b_{\sigma^2, \theta}(\alpha) - b_{\sigma_0^2, \theta_0}(\alpha)]^2 d\alpha$. Then, $\tilde{p}^{-1}R(\sigma^2, \theta)$ is the Riemann approximation of the integral $\tilde{p}^{-1}R^*(\sigma^2, \theta)$. Note that $s_n/\tilde{p} = o(1)$. Furthermore, $b_{\sigma^2, \theta}(\alpha)$ is uniformly square integrable for $(\sigma^2, \theta) \in \mathcal{J}_{\sigma^2} \times \mathcal{J}_{\theta}$ (the proof is very similar to the analysis of C_2 below; we thus omit it). Hence, the Riemann approximation error is negligible. Particularly, there exists a constant $c_1 \in (0, 1)$ such that

$$R(\sigma^2, \theta) \geq c_1 \cdot R^*(\sigma^2, \theta). \quad (\text{B.20})$$

It suffices to study $R^*(\sigma^2, \theta)$. The next lemma is proved in Section B.2.6.

Lemma B.2.1. Let $F(x)$ be a distribution on $(0, \infty)$ with a continuous density $f(x)$. Let $\bar{F}(x) = 1 - F(x)$, $b_F(\alpha) = \bar{F}^{-1}(\alpha)$, and $\mu_m(f) = \int x^m f(x) dx$, $m \geq 1$. For another distribution $G(x)$ on $(0, \infty)$ with a continuous density $g(x)$, we define $\bar{G}(x)$, $b_G(\alpha)$, and $\mu_m(g)$ similarly. Suppose $\int x^2 |\bar{F}(x) - \bar{G}(x)| dx < \infty$. Let $\check{g}(x, y) = \max_{z \in [x, y] \cup [y, x]} g(z)$ for $x, y \in (0, \infty)$. We assume that $C_1 \equiv \int_0^1 \left[\frac{\check{g}(b_F(\alpha), b_G(\alpha))}{f(b_F(\alpha))} \right]^2 d\alpha < \infty$ and $C_2 \equiv \int_0^1 \left[\frac{b_F(\alpha) \check{g}(b_F(\alpha), b_G(\alpha))}{f(b_F(\alpha))} \right]^2 d\alpha < \infty$. Then,

$$\int_0^1 [b_G(\alpha) - b_F(\alpha)]^2 d\alpha \geq \frac{|\mu_1(f) - \mu_1(g)|^2}{4C_1}, \quad \int_0^1 [b_G(\alpha) - b_F(\alpha)]^2 d\alpha \geq \frac{|\mu_2(f) - \mu_2(g)|^2}{4C_2}.$$

We apply Lemma B.2.1 to $F(\cdot) = F_{\gamma_n}(\cdot; \sigma_0^2, \theta_0, T_1, T_2)$ and $G(\cdot) = F_{\gamma_n}(\cdot; \sigma^2, \theta, T_1, T_2)$. Define

$$\mu_1(\sigma^2, \theta) = \int x dF_{\gamma_n}(x; \sigma^2, \theta, T_1, T_2), \quad \mu_2(\sigma^2, \theta) = \int x^2 dF_{\gamma_n}(x; \sigma^2, \theta, T_1, T_2).$$

We now show that the quantities C_1, C_2 in Lemma B.2.1 are uniformly upper bounded by constants for all $(\sigma^2, \theta) \in \mathcal{J}_\sigma^2 \times \mathcal{J}_\theta$. We only study C_2 , and the analysis of C_1 is similar. By Knowles & Yin (2017), Ding (2020), the support of $F_{\gamma_n}(\cdot; \sigma^2, \theta, T_1, T_2)$ is in a compact subset of $(0, \infty)$, and the density is upper bounded by a constant; these constants are uniform for $(\sigma^2, \theta) \in \mathcal{J}_\sigma^2 \times \mathcal{J}_\theta$. It follows that

$$C_2 \leq C \int_0^1 \left[\frac{1}{f(b_F(\alpha))} \right]^2 d\alpha = \int \frac{1}{f^2(x)} f(x) dx = \int \frac{1}{f(x)} dx.$$

Here we have used a change of variable $x = b_F(\alpha)$, where $\alpha = 1 - F(x)$ and $d\alpha = f(x)dx$. We then apply Theorem 3.3 of Ji (2020). Note that $F(\cdot) = F_{\gamma_n}(\cdot; \sigma_0^2, \theta_0, T_1, T_2)$ is the free multiplicative convolution between a truncated Gamma distribution and the standard MP distribution. These two distributions are compactly supported and have power law behavior on left/right ends. The conditions in Theorem 3.3 of Ji (2020) are satisfied for $t_\pm^\mu = 0$ (truncated Gamma) and $t_\pm^\nu = 1/2$ (MP law). By that theorem, the density of $F(\cdot)$ has a square-root decay at the left/right edge: Let $[b^-, b^+]$ be the support of $F(\cdot)$; then, $C^{-1} \leq f(x)/\sqrt{(x-b^-)(b^+-x)} \leq C$ for $x \in [b^-, b^+]$. It yields that

$$C_2 \leq \int_{b^-}^{b^+} \frac{C}{\sqrt{(x-b^-)(b^+-x)}} dx = O(1).$$

We have verified that C_1 and C_2 in Lemma B.2.1 are uniformly upper bounded. As a result,

$$R^*(\sigma^2, \theta) \geq Cn \left(|\mu_1(\sigma^2, \theta) - \mu_1(\sigma_0^2, \theta_0)|^2 + |\mu_2(\sigma^2, \theta) - \mu_2(\sigma_0^2, \theta_0)|^2 \right). \quad (\text{B.21})$$

Below, we study $\mu_1(\sigma^2, \theta)$ and $\mu_2(\sigma^2, \theta)$. Note that $\text{Gamma}(\theta, \theta/\sigma^2, \sigma^2 T_1, \sigma^2 T_2)$ is equivalent to $\sigma^2 \cdot \text{Gamma}(\theta, \theta, T_1, T_2)$. Then, the distributions $F_{\gamma_n}(\cdot; \sigma^2, \theta, T_1, T_2)$ and $F_{\gamma_n}(\cdot; 1, \theta, T_1, T_2)$ also have such a connection. This implies $\mu_1(\sigma^2, \theta) = \sigma^2 \cdot \mu_1(1, \theta)$ and $\mu_2(\sigma^2, \theta) = \sigma^4 \cdot \mu_2(1, \theta)$. Define

$$\kappa(\theta) = \mu_2(\sigma^2, \theta) / [\mu_1(\sigma^2, \theta)]^2.$$

Consider a mapping M from \mathbb{R}^2 to \mathbb{R}^2 , where $M(x, y) = (x, y/x^2)$. It maps $(\mu_1(\sigma^2, \theta), \mu_2(\sigma^2, \theta))$ to $(\mu_1(\sigma^2, \sigma^2), \kappa(\theta))$. The Jacobian matrix is

$$\begin{bmatrix} 1 & 0 \\ -2y/x^3 & 1/x^2 \end{bmatrix}.$$

When $(\sigma^2, \theta) \in \mathcal{J}_{\sigma^2} \times \mathcal{J}_{\theta}$, the vector $(\mu_1(\sigma^2, \theta), \mu_2(\sigma^2, \theta))$ is in a compact set. The spectral norm of Jacobian is uniformly upper bounded. It follows that

$$\begin{aligned} & |\mu_1(\sigma^2, \theta) - \mu_1(\sigma_0^2, \theta_0)|^2 + |\mu_2(\sigma^2, \theta) - \mu_2(\sigma_0^2, \theta_0)|^2 \\ & \geq C \left(|\mu_1(\sigma^2, \theta) - \mu_1(\sigma_0^2, \theta_0)|^2 + |\kappa(\theta) - \kappa(\theta_0)|^2 \right). \end{aligned} \quad (\text{B.22})$$

We then study $\mu_1(\sigma^2, \theta)$ and $\kappa(\theta)$. Denote by $\hat{F}(\cdot; \sigma^2, \theta, T_1, T_2)$ the ESD when (σ^2, θ) are true parameters. Write $\hat{\mu}_1(\sigma^2, \theta) = \int x d\hat{F}(x; \sigma^2, \theta, T_1, T_2)$ and $\hat{\mu}_2(\sigma^2, \theta) = \int x^2 d\hat{F}(x; \sigma^2, \theta, T_1, T_2)$. The converges of ESD to its theoretical limit yields that $|\hat{\mu}_1(\sigma^2, \theta) - \mu_1(\sigma^2, \theta)| \rightarrow 0$ and $|\hat{\mu}_2(\sigma^2, \theta) - \mu_2(\sigma^2, \theta)| \rightarrow 0$ in probability. In fact, we have a stronger result (Knowles & Yin, 2017):

$$|\mathbb{E}[\hat{\mu}_1(\sigma^2, \theta)] - \mu_1(\sigma^2, \theta)| \prec n^{-1}, \quad |\mathbb{E}[\hat{\mu}_2(\sigma^2, \theta)] - \mu_2(\sigma^2, \theta)| \prec n^{-1}. \quad (\text{B.23})$$

Here the expectation is with respect to the null model (i.e., $K = 0$) with true parameters (σ^2, θ) . The left hand sides above are non-stochastic quantities, and “ $\prec n^{-1}$ ” is interpreted as “ $\leq n^{-1+\varepsilon}$ for any $\varepsilon > 0$.” Since $\mu_1(\sigma^2, \theta)$ and $\mu_2(\sigma^2, \theta)$ are uniformly upper/lower bounded, it follows that

$$\left| \hat{\kappa}(\theta) - \frac{\mathbb{E}[\hat{\mu}_2(\sigma^2, \theta)]}{(\mathbb{E}[\hat{\mu}_1(\sigma^2, \theta)])^2} \right| \prec n^{-1}. \quad (\text{B.24})$$

By definition, we can also write $\hat{\mu}_1 = \frac{1}{\tilde{p}} \sum_{i=1}^{\tilde{p}} \hat{\lambda}_i = \frac{1}{\tilde{p}} \text{tr}(\mathbf{S})$ and $\hat{\mu}_2 = \frac{1}{\tilde{p}} \sum_{i=1}^{\tilde{p}} \hat{\lambda}_i^2 = \frac{1}{\tilde{p}} \|\mathbf{S}\|_F^2$, where

$\mathbf{S} = \frac{1}{n} \mathbf{X}^\top \mathbf{X}$ is the sample covariance matrix under the null model of $K = 0$. By Assumption 2.4.1, $\mathbf{X} = \mathbf{Y} \boldsymbol{\Sigma}^{1/2}$, where \mathbf{Y} contains *iid* zero-mean, unit variance entries. Note that our purpose here is to approximate the moments of the theoretical limit of ESD, and we are flexible to choose the eigenvectors in $\boldsymbol{\Sigma}$. We choose $\boldsymbol{\xi}_k$ as the k th standard basis, and so $\boldsymbol{\Sigma} = \text{diag}(\sigma_1^2, \sigma_2^2, \dots, \sigma_p^2)$. By direct calculations,

$$\begin{aligned} \mathbb{E}[\hat{\mu}_1(\sigma^2, \theta)] &= \frac{1}{n\tilde{p}} \mathbb{E} \left[\sum_{j=1}^p \left(\sum_{i=1}^n \sigma_j^2 Y_{ij}^2 \right) \right] = (\gamma_n \vee 1) \cdot \mathbb{E}[\sigma_1^2], \\ \mathbb{E}[\hat{\mu}_2(\sigma^2, \theta)] &= \frac{1}{n^2\tilde{p}} \mathbb{E} \left[\sum_{j=1}^p \left(\sum_{i=1}^n \sigma_j^2 Y_{ij}^2 \right)^2 + \sum_{1 \leq j \neq \ell \leq p} \left(\sum_{i=1}^n \sigma_j \sigma_\ell Y_{ij} Y_{i\ell} \right)^2 \right] \\ &= \frac{1}{n^2\tilde{p}} \left[np \mathbb{E}[\sigma_1^4] \mathbb{E}[Y_{11}^4] + pn(n-1) \mathbb{E}[\sigma_1^4] + p(p-1)n (\mathbb{E}[\sigma_1^2])^2 \right] \\ &= O(n^{-1}) + (\gamma_n \vee 1) \cdot \mathbb{E}[\sigma_1^4] + \gamma_n (\gamma_n \vee 1) \cdot (\mathbb{E}[\sigma_1^2])^2. \end{aligned}$$

Note that $\sigma_1^2/\sigma^2 \sim \text{Gamma}(\theta, \theta, T_1, T_2)$. The density of $\text{Gamma}(\theta, \theta, T_1, T_2)$ is equal to $x^{\theta-1} e^{-\theta x} \cdot (\int_{T_1}^{T_2} z^{\theta-1} e^{-\theta z} dz)^{-1}$. We immediately have

$$\begin{aligned} \mathbb{E}[\hat{\mu}_1(\sigma^2, \theta)] &= (\gamma_n \vee 1) \sigma^2 \cdot \frac{\int_{T_1}^{T_2} x^\theta \exp(-\theta x) dx}{\int_{T_1}^{T_2} x^{\theta-1} \exp(-\theta x) dx} \\ \mathbb{E}[\hat{\mu}_2(\sigma^2, \theta)] &= O\left(\frac{1}{n}\right) + (\gamma_n \vee 1) \frac{\sigma^4 \int_{T_1}^{T_2} x^{\theta+1} \exp(-\theta x) dx}{\int_{T_1}^{T_2} x^{\theta-1} \exp(-\theta x) dx} + \gamma_n (\gamma_n \vee 1) \frac{\sigma^4 \left[\int_{T_1}^{T_2} x^\theta \exp(-\theta x) dx \right]^2}{\left[\int_{T_1}^{T_2} x^{\theta-1} \exp(-\theta x) dx \right]^2}. \end{aligned}$$

Define $\Psi(\theta) = \Psi(\theta; T_1, T_2) \equiv (\int_{T_1}^{T_2} x^\theta e^{-\theta x} dx) / (\int_{T_1}^{T_2} x^{\theta-1} e^{-\theta x} dx)$. Let $\Phi(\theta)$ be the same as in the statement of this theorem. We plug the above equations into (B.23)-(B.24) to get

$$\begin{aligned} \mu_1(\sigma^2, \theta) &= (\gamma_n \vee 1) \sigma^2 \cdot \Psi(\theta) + O_{\prec}(n^{-1}), \\ \kappa(\theta) &= \frac{1}{(\gamma_n \vee 1)} \cdot \Phi(\theta) + \frac{\gamma_n}{(\gamma_n \vee 1)} + O_{\prec}(n^{-1}). \end{aligned} \tag{B.25}$$

Consider the mapping from (σ^2, θ) to $(\mu_1(\sigma^2, \theta), \kappa(\theta))$. The Jacobian matrix is

$$J = (\gamma_n \vee 1) \begin{bmatrix} \Psi(\theta) & \sigma^2 \cdot \Psi'(\theta) \\ 0 & \frac{1}{(\gamma_n \vee 1)^2} \cdot \Phi'(\theta) \end{bmatrix} + O_{\prec}(n^{-1}).$$

First, since \mathcal{J}_θ is a bounded set, $\Psi(\theta)$, $\Psi'(\theta)$ and $\Phi'(\theta)$ are uniformly upper bounded by constants. Second, we have $\Psi(\theta) > 0$ in a fixed compact set \mathcal{J}_θ . As a result, $\Psi(\theta)$ must be uniformly lower bounded by a constant. Last, the assumption says that $\inf_{\theta \in \mathcal{J}_\theta} |\Phi'(\theta)| \geq \omega$, for a constant $\omega > 0$. Combining these arguments with the formula of the inverse of a 2×2 matrix, we have $\|J^{-1}\| \leq C$. It follows that

$$\begin{aligned} & |\mu_1(\sigma^2, \theta) - \mu_1(\sigma_0^2, \theta_0)|^2 + |\kappa(\theta) - \kappa(\theta_0)|^2 \\ & \geq C(|\sigma^2 - \sigma_0^2|^2 + |\theta - \theta_0|^2). \end{aligned} \tag{B.26}$$

We plug (B.26) into (B.22), and then into (B.21), and then combine it with (B.20). It gives (B.19). □

B.2.4 PROOF OF LEMMA 2.4.2

Write

$$J_1(\theta) = \left(\int_{t_1}^{t_2} x^{\theta+1} \exp(-\theta x) dx \right) \left(\int_{t_1}^{t_2} x^{\theta-1} \exp(-\theta x) dx \right), \quad J_2(\theta) = \left(\int_{t_1}^{t_2} x^\theta \exp(-\theta x) dx \right)^2.$$

Then $\Psi(\theta) = J_1(\theta)/J_2(\theta)$ and

$$\Psi'(\theta) = \frac{J_1'(\theta)J_2(\theta) - J_1(\theta)J_2'(\theta)}{J_2(\theta)^2}. \tag{B.27}$$

By direct calculations,

$$\begin{aligned} \mathcal{J}'_1(\theta) = & \left(\int_{t_1}^{t_2} \log(x)x^{\theta+1} \exp(-\theta x) dx - \int_{t_1}^{t_2} x^{\theta+2} \exp(-\theta x) dx \right) \left(\int_{t_1}^{t_2} x^{\theta-1} \exp(-\theta x) dx \right) \\ & + \left(\int_{t_1}^{t_2} \log(x)x^{\theta-1} \exp(-\theta x) dx - \int_{t_1}^{t_2} x^{\theta} \exp(-\theta x) dx \right) \left(\int_{t_1}^{t_2} x^{\theta+1} \exp(-\theta x) dx \right), \end{aligned}$$

$$\mathcal{J}'_2(\theta) = 2 \left(\int_{t_1}^{t_2} x^{\theta} \exp(-\theta x) dx \right) \left(\int_{t_1}^{t_2} \log(x)x^{\theta} \exp(-\theta x) dx - \int_{t_1}^{t_2} x^{\theta+1} \exp(-\theta x) dx \right).$$

Let $L(\alpha, \theta; t_1, t_2)$ denote $\int_{t_1}^{t_2} \log(x)x^{\alpha} \exp(-\theta x) dx$ and $I(\alpha, \theta; t_1, t_2)$ denote $\int_{t_1}^{t_2} x^{\alpha} \exp(-\theta x) dx$. When not causing any confusion, we write them as $L(\alpha)$ and $I(\alpha)$. Then

$$\mathcal{J}_1(\theta) = I(\theta+1) \times I(\theta-1), \quad \mathcal{J}_2(\theta) = I(\theta)^2$$

$$\mathcal{J}'_1(\theta) = (L(\theta+1) - I(\theta+2)) \times I(\theta-1) + (L(\theta-1) - I(\theta)) \times I(\theta+1)$$

$$\mathcal{J}'_2(\theta) = 2(L(\theta) - I(\theta+1)) \times I(\theta)$$

Plugging them into (B.27), we have

$$\Psi'(\theta) = \frac{I(\theta+1)I(\theta-1)}{I(\theta)^2} \left(\left(\frac{L(\theta+1)}{I(\theta+1)} + \frac{L(\theta-1)}{I(\theta-1)} - 2 \frac{L(\theta)}{I(\theta)} \right) - \left(\frac{I(\theta+2)}{I(\theta+1)} + \frac{I(\theta)}{I(\theta-1)} - 2 \frac{I(\theta+1)}{I(\theta)} \right) \right).$$

Recall that we are interested in $\theta \in \mathcal{J}_{\theta} = [c, d]$. For $\alpha \in [c-1, d+2]$ and $\theta \in [c, d]$,

$$\int_0^{\infty} \log(x)x^{\alpha} \exp(-\theta x) dx - L(\alpha, \theta; t_1, t_2) = \int_0^{t_1} \log(x)x^{\alpha} \exp(-\theta x) dx + \int_{t_2}^{\infty} \log(x)x^{\alpha} \exp(-\theta x) dx,$$

$$\left| \int_0^{t_1} \log(x)x^{\alpha} \exp(-\theta x) dx \right| \leq \int_0^{t_1} (-\log(x))x^{\alpha-1} \exp(-cx) dx \rightarrow 0, \quad \text{as } t_1 \rightarrow 0,$$

$$\left| \int_{t_2}^{\infty} \log(x)x^{\alpha} \exp(-\theta x) dx \right| \leq \int_{t_2}^{\infty} \log(x)x^{d+2} \exp(-cx) dx \rightarrow 0, \quad \text{as } t_2 \rightarrow \infty.$$

This implies for $\alpha \in [c-1, d+2]$, $\theta \in [c, d]$, as $(t_1, t_2) \rightarrow (0, \infty)$, $L(\alpha, \theta; t_1, t_2)$ uniformly converges to $L_0(\alpha, \theta) = \int_0^\infty \log(x)x^\alpha \exp(-\theta x)dx$. By a similar argument, we can show that $I(\alpha, \theta; t_1, t_2)$ uniformly converges to $I_0(\alpha, \theta) = \int_0^\infty x^\alpha \exp(-\theta x)dx$. From the uniform convergence and the fact that $I_0(\alpha, \theta)$ is lower bounded by a common positive constant when $\alpha \in [c-1, d+2]$, $\theta \in [c, d]$, we know that as $(t_1, t_2) \rightarrow (0, \infty)$ we have $\Psi'(\theta)$ uniformly converges to

$$\frac{I_0(\theta+1)I_0(\theta-1)}{I_0(\theta)^2} \left(\left(\frac{L_0(\theta+1)}{I_0(\theta+1)} + \frac{L_0(\theta-1)}{I_0(\theta-1)} - 2\frac{L_0(\theta)}{I_0(\theta)} \right) - \left(\frac{I_0(\theta+2)}{I_0(\theta+1)} + \frac{I_0(\theta)}{I_0(\theta-1)} - 2\frac{I_0(\theta+1)}{I_0(\theta)} \right) \right),$$

for all $\theta \in [c, d]$. Here, $L_0(\alpha)$ and $I_0(\alpha)$ are short for $L_0(\alpha, \theta)$ and $I_0(\alpha, \theta)$. Let $Z \sim \text{Gamma}(\alpha, \theta)$ and let ψ denote the digamma function. By properties of the Gamma distribution,

$$\frac{I_0(\alpha, \theta)}{I_0(\alpha-1, \theta)} = \mathbb{E}(Z) = \frac{\alpha}{\theta}, \quad \frac{L_0(\alpha-1, \theta)}{I_0(\alpha-1, \theta)} = \mathbb{E}(\log(Z)) = \psi(\alpha) - \log(\theta).$$

Therefore, $\Psi'(\theta)$ uniformly converges to

$$\frac{\theta+1}{\theta} \left(\left(\psi(\theta+2) + \psi(\theta) - 2\psi(\theta+1) \right) - \left(\frac{\theta+2}{\theta} + \frac{\theta}{\theta} - 2 \times \frac{\theta+1}{\theta} \right) \right) = \frac{\theta+1}{\theta} \left(\frac{1}{\theta+1} - \frac{1}{\theta} \right) = -\frac{1}{\theta^2}.$$

The first equation uses the recurrence relation of digamma function. By the uniform convergence, for any $\delta > 0$ there exists $0 < T_1^* < T_2^* < \infty$ such that $\sup_{\theta \in [c, d]} |\Psi'(\theta) - (-\frac{1}{\theta^2})| \leq \delta$. The claim follows by choosing $\delta = 1/d^2 - \omega$. \square

B.2.5 PROOF OF THEOREM 2.4.4

Let $d_j = \sigma_j^2 + \mu_j$ for $1 \leq k \leq K$ and $d_j = \sigma_j^2$ for $K+1 \leq j \leq p$. Then, d_1, d_2, \dots, d_p are all the eigenvalues of Σ . Define

$$\hat{G}(x) = -\frac{1}{x} + \frac{\gamma}{p} \sum_{j=1}^p \frac{1}{x + \sigma_j^{-2}}. \quad (\text{B.28})$$

By Lemma 2.2 and Condition 3.6 of Ding (2020), this function $\hat{G}(x)$ has 2 critical points $0 > \hat{x}_1 > \hat{x}_2$; furthermore, conditioning on Σ , the ESD converges to a limit whose support is $[\hat{G}(\hat{x}_2), \hat{G}(\hat{x}_1)]$. We apply Theorem 3.2 of Ding (2020). Using the first claim there, if $-1/d_k \geq \hat{x}_1 + n^{1/3}$ for each $1 \leq k \leq K$, then

$$|\hat{\lambda}_k - \hat{G}(-1/d_k)| \prec n^{-1/2}(-1/d_k - \hat{x}_1)^{1/2}, \quad 1 \leq k \leq K.$$

Using the second claim there,

$$|\hat{\lambda}_{K+1} - \hat{G}(\hat{x}_1)| \prec n^{-2/3}.$$

The above “stochastic dominance” arguments are conditioning on Σ . Under Model (2.13) for Σ , $\hat{G}(x)$ converges weakly to $G(x)$ defined in (2.16), and the critical points (\hat{x}_1, \hat{x}_2) also converge to (x_1^*, x_2^*) , the critical points of $G(x)$, almost surely. Replacing $\hat{G}(\cdot)$ and \hat{x}_1 by $G(\cdot)$ and x_1^* in the above inequalities has a negligible effect (e.g., see Example 3.9 of Ding (2020)). It follows that

$$\max_{1 \leq k \leq K} |\hat{\lambda}_k - G(-1/d_k)| \prec n^{-1/2}, \quad |\hat{\lambda}_{K+1} - G(x_1^*)| \prec n^{-2/3}.$$

Note that $d_k = \sigma_k^2 + \mu_k \geq \mu_K + T_1$. The assumption of $-1/(T_1 + \mu_K) \geq x_1^* + \tau$ guarantees that $G(-1/d_k) \geq G(-1/(T_1 + \mu_K)) \geq G(x_1^* + \tau) \geq G(x_1^*) + c$, where $c > 0$ is a constant. Therefore,

$$\min_{1 \leq k \leq K} \{\hat{\lambda}_k\} - G(x_1^*) \geq c + O_{\prec}(n^{-1/2}), \quad \hat{\lambda}_{K+1} - G(x_1^*) \prec n^{-2/3}, \quad (\text{B.29})$$

where $O_{\prec}(b_n)$ means the absolute value is $\prec b_n$.

The estimator \hat{K} is obtained by thresholding the empirical eigenvalues at \hat{T}_{β} as in (2.15). Let $\hat{\lambda}_1^* = \hat{\lambda}_1^*(\sigma^2, \theta)$ be the largest empirical eigenvalue under the null model ($K = 0$) with parameters

(σ^2, θ) . Applying Theorem 3.2 of [Ding \(2020\)](#) again, for the same x_1^* as above,

$$|\hat{\lambda}_1^*(\sigma^2, \theta) - G(x_1^*)| \prec n^{-2/3}.$$

In Theorem 2.4.3, we have shown $|\hat{\sigma}^2 - \sigma^2| \prec n^{-1}$ and $|\hat{\theta} - \theta| \prec n^{-1}$. Now, let \hat{x}_1^* be the largest critical point of $G(x)$ in (2.16), except that (σ^2, θ) is replaced by $(\hat{\sigma}^2, \hat{\theta})$. Then, we have $|G(\hat{x}_1^*) - G(x_1^*)| = O(\sqrt{|\hat{\sigma}^2 - \sigma^2|^2 + |\hat{\theta} - \theta|^2}) \prec n^{-1}$ and $|\hat{\lambda}_1^*(\hat{\sigma}^2, \hat{\theta}) - G(\hat{x}_1^*)| \prec n^{-2/3}$. Combining these claims gives

$$|\hat{\lambda}_1^*(\hat{\sigma}^2, \hat{\theta}) - G(x_1^*)| \prec n^{-2/3}.$$

Note that \hat{T}_β is the $(1 - \beta)$ -quantile of $\hat{\lambda}_1^*(\hat{\sigma}^2, \hat{\theta})$ (it means the quantile of $\hat{\lambda}_1^*(\sigma^2, \theta)$ evaluated at $(\sigma^2, \theta) = (\hat{\sigma}^2, \hat{\theta})$). The above inequality implies that there exists $\beta \rightarrow 0$ properly slow such that

$$n^{-2/3} \ll \hat{T}_\beta - G(x_1^*) \ll 1. \quad (\text{B.30})$$

It follows from (B.29) and (B.30) that $\hat{K} = K$. □

B.2.6 PROOF OF LEMMA B.2.1

We only show the second inequality. The proof of the first inequality is similar and thus omitted.

Note that $f(x) - g(x)$ is the derivative of $\bar{G}(x) - \bar{F}(x)$. Using integration by part, we have

$$\mu_2(f) - \mu_2(g) = \int x^2 [f(x) - g(x)] dx = 2 \int x [\bar{F}(x) - \bar{G}(x)] dx. \quad (\text{B.31})$$

We consider a change of variable from x to $\alpha = \bar{F}(x)$. Note that $x = h_F(\alpha)$. It follows that

$$\begin{aligned} \int x[\bar{F}(x) - \bar{G}(x)]dx &= \int_0^1 h_F(\alpha)[\alpha - \bar{G}(h_F(\alpha))]h'_F(\alpha)d\alpha \\ &= \int_0^1 h_F(\alpha)[\bar{G}(h_G(\alpha)) - \bar{G}(h_F(\alpha))]h'_F(\alpha)d\alpha. \end{aligned}$$

By mean value theorem, there is x^* between $h_F(\alpha)$ and $h_G(\alpha)$ such that $\bar{G}(h_G(\alpha)) - \bar{G}(h_F(\alpha)) = -g(x^*)[h_G(\alpha) - h_F(\alpha)]$. Recall that $\check{g}(x, y) = \max_{z \in [x, y] \cup [y, x]} g(z)$. It follows that $|\bar{G}(h_G(\alpha)) - \bar{G}(h_F(\alpha))| \leq \check{g}(h_F(\alpha), h_G(\alpha)) \cdot |h_G(\alpha) - h_F(\alpha)|$. We plug it into the above equation to get

$$\left| \int x[\bar{F}(x) - \bar{G}(x)]dx \right| \leq \int_0^1 |h_G(\alpha) - h_F(\alpha)| \cdot |h_F(\alpha) \check{g}(h_F(\alpha), h_G(\alpha)) h'_F(\alpha)| d\alpha.$$

Since $h_F(\cdot) = \bar{F}^{-1}$, we have $h'_F(\alpha) = -1/f(h_F(\alpha))$. It follows that

$$\begin{aligned} \left| \int x[\bar{F}(x) - \bar{G}(x)]dx \right| &\leq \int_0^1 |h_G(\alpha) - h_F(\alpha)| \cdot \frac{h_F(\alpha) \cdot \check{g}(h_F(\alpha), h_G(\alpha))}{f(h_F(\alpha))} d\alpha \\ &\leq \sqrt{\int_0^1 |h_G(\alpha) - h_F(\alpha)|^2 d\alpha} \sqrt{\int_0^1 \left[\frac{h_F(\alpha) \cdot \check{g}(h_F(\alpha), h_G(\alpha))}{f(h_F(\alpha))} \right]^2 d\alpha} \\ &\leq \sqrt{\int_0^1 |h_G(\alpha) - h_F(\alpha)|^2 d\alpha} \cdot \sqrt{C_2}. \end{aligned} \tag{B.32}$$

Combining (B.31)-(B.32) gives the claim. □

B.3 ROBUSTNESS OF BEMA ON REAL DATA

For the two real data sets in Section 2.6, we apply BEMA with different values of α . The results are presented in the tables below. Both the point estimator and the confidence interval are very stable as long as α is in a reasonable range.

	BEMA (0.1)	BEMA (0.2)	BEMA (0.3)	BEMA (0.4)
$\hat{\theta}$	0.343	0.288	0.281	0.270
$\hat{\sigma}^2$	0.869	0.926	0.949	1
$\hat{K}(\beta = 0.1)$	1	1	1	1
90% quantile	16.074	19.231	20.261	21.944
10% quantile	9.379	10.872	11.186	12.098
confidence interval	[1,4]	[1,4]	[1,4]	[1,2]

Table B.1: Lung Cancer data. BEMA is applied with $\alpha \in \{0.1, 0.2, 0.3, 0.4\}$ (denoted as BEMA (α) in the table). The quantiles are from Gamma($\hat{\theta}, \hat{\theta}/\hat{\sigma}^2$), and they are used to construct the 80% confidence interval.

	BEMA (0.1)	BEMA (0.2)	BEMA (0.3)	BEMA (0.4)
$\hat{\theta}$	4.256	4.239	4.198	4.261
$\hat{\sigma}^2$	0.3779	0.3780	0.3782	0.3783
$\hat{K}(\beta = 0.1)$	28	28	28	28
90% quantile	6.895	6.899	6.909	6.903
10% quantile	6.822	6.829	6.838	6.831
confidence interval	[28,30]	[28,30]	[28,29]	[28,30]

Table B.2: 1000 Genomes data. BEMA is applied with $\alpha \in \{0.1, 0.2, 0.3, 0.4\}$ (denoted as BEMA (α) in the table). The quantiles are from Gamma($\hat{\theta}, \hat{\theta}/\hat{\sigma}^2$), and they are used to construct the 80% confidence interval.



Supplemental Materials of Chapter 3

C.1 SCALING GROUP MOVES

To see how the posterior distribution of the loading matrix is influenced by the SpSL prior, we need to observe the sample behavior at equilibrium with different priors. Due to the strong ties between the loading matrix and the latent factors, samples are inflating slowly along the basic Gibbs sampling iterations, which demonstrates the slow mixing behavior of the Gibbs sampler.

A promising way to improve Markov Chain Monte Carlo (MCMC) convergence is to add a group move into the sampler. [Liu & Sabatti \(2000\)](#) proposed “generalized Gibbs sampling”, which can be seen as a generalization of [Liu & Wu \(1999\)](#) for conditional sampling along the trajectories of any designed transformation group. By taking advantage of the model structure and proposing a group trajectory that can cross various significant local modes, this group move can dramatically improve the MCMC convergence. The following theorem from [Liu & Sabatti \(2000\)](#) characterizes how a group move should be conducted.

Theorem C.1.1. (Liu and Sabatti(2000)) Let π be an arbitrary distribution on a space \mathcal{Z} , and suppose $t_\alpha(z) : \mathcal{Z} \rightarrow \mathcal{Z}$ is a transformation parameterized by $\alpha \in \mathcal{A}$. Assume there is a group structure on both \mathcal{A} and the transformation family, and a left-Haar measure H on \mathcal{A} . If z follows distribution π and α is drawn from

$$\pi(\alpha|z) \propto \pi(t_\alpha(z)) \left| \frac{\partial t_\alpha(z)}{\partial z} \right| H(d\alpha), \quad (\text{C.1})$$

then $t_\alpha(z)$ follows distribution π .

If π in Theorem C.1.1. is the full posterior distribution, then t_α generated by the conditional distribution (C.1) gives a transformation that preserves the target distribution π . We can add this transformation after each round of Gibbs sampling to improve convergence. To design group moves that can move the loading matrix and factors jointly in the synthetic example, we consider the following group of scale transformations for $k = 1, \dots, K$:

$$t_{\alpha_k}(\beta_{1k}, \dots, \beta_{Gk}, \omega_{1k}, \dots, \omega_{nk}) = \left(\alpha_k \beta_{1k}, \dots, \alpha_k \beta_{Gk}, \frac{1}{\alpha_k} \omega_{1k}, \dots, \frac{1}{\alpha_k} \omega_{nk} \right),$$

and draw α_k sequentially from:

$$p(d\alpha_k) \propto \prod_{j=1}^G ((1 - \gamma_{jk}) \psi(\alpha_k \beta_{jk} | \lambda_0) + \gamma_{jk} \psi(\alpha_k \beta_{jk} | \lambda_1)) \times \prod_{i=1}^n \exp\left(-\frac{\omega_{ik}^2}{2\alpha_k^2}\right) \times \alpha_k^{G-n-1} d\alpha_k$$

We design these group moves to rescale each column since we observe a synchronous inflation within every column during Gibbs sampling and changes of magnitude are encumbered due to the strong connection between factors and loading. These scaling group moves are cheap to implement since the conditional distribution of α_k is a univariate and unimodal distribution. More delicate moves such as linear restructuring (corresponding to ‘rotate’ the loading in PXL-EM) $t_A(B, \Omega) : B, \Omega \rightarrow BA, A^{-1}\Omega$ can be difficult to implement in practice.

C.2 THE MODIFIED GHOSH-DUNSON MODEL

Since the magnitude inflation is associated with the overdose of independent slab priors on the loading matrix, an immediate counter measure would be to control the number of slab priors used. Ghosh & Dunson (2009) proposed to use an inverse gamma prior for the variance of the normal factors and impose the standard Gaussian prior on elements of the loading matrix, which will be called the Ghosh-Dunson model. Here we propose a modified Ghosh-Dunson model by relocating the variance parameters of the factors to the loading matrix and imposing a SpSL prior on its elements:

$$\begin{aligned}
\text{Model: } & \mathbf{y}_i \mid \omega_i, \mathbf{B}, \boldsymbol{\Sigma} \stackrel{i.i.d.}{\sim} \mathcal{N}_G(\mathbf{B}\omega_i, \boldsymbol{\Sigma}), \quad \omega_i \stackrel{i.i.d.}{\sim} \mathcal{N}_K(0, \mathbf{I}_K) \\
\text{Priors: } & \beta_{jk} = q_{jk}r_k, \quad p(r_k \mid \lambda) = \psi(r_k \mid \lambda); \\
& p(q_{jk} \mid \gamma_{jk}, \lambda_0, \lambda_1) = (1 - \gamma_{jk})\psi(q_{jk} \mid \lambda_0) + \gamma_{jk}\psi(q_{jk} \mid \lambda_1), \quad \lambda_0 \gg \lambda_1; \\
& \gamma_{jk} \mid \theta_k \sim \text{Bernoulli}(\theta_k) \text{ independently}; \\
& \theta_k = \prod_{l=1}^k \nu_l, \quad \nu_l \stackrel{i.i.d.}{\sim} \text{Beta}(\alpha, 1); \\
& \sigma_j^2 \stackrel{i.i.d.}{\sim} \text{Inverse-Gamma}(\eta/2, \eta\varepsilon/2).
\end{aligned} \tag{C.2}$$

where β_{jk} denote the $(j, k)^{th}$ element of \mathbf{B} and $\psi(\cdot \mid \lambda)$ is the normal density with precision λ . We chose λ_0 large and $\lambda_1 = 1$.

In this framework, each loading element β_{jk} is expressed as the product of a column-wise magnitude parameter r_k and the ‘normalized’ loading q_{jk} . Ghosh & Dunson (2009)’s original model corresponds to assuming $\gamma_{jk} \equiv \theta_k \equiv 1$, i.e., a normal instead of mixture normal prior for the q_{jk} . We impose a diffuse normal prior on the r_k ’s and a SpSL prior on q_{jk} . With this dependent prior specification, the number of the “slab parameters” is greatly reduced (all elements in each column of \mathbf{B} share a common “slab parameter” r_k), while marginally the prior on each β_{jk} is the same as that of the independent SpSL prior. This prior setup on the loading matrix is similar to the one in the hierarchical linear model in Jia & Xu (2007) where $\mathbf{\Omega}$ is prescribed, and the prior setup on \mathbf{B} establishes connections between rows of the loading matrix to prevent the degeneration of the original model to multiple independent linear regressions. However, the hierarchical linear model is not subject to the inflation problem even if completely independent priors are imposed on the loading matrix since $\mathbf{\Omega}$ is already prescribed.

Although the dependent slab prior specification is an effective way for resolving the posterior inflation problem, the justification of the posterior consistency is rather difficult under this framework. We simply provide some numerical results in Section 3.6 and 3.7 to compare the posterior distribution based on the modified Ghosh-Dunson model (C.2) with that resulting from our strategy of imposing the \sqrt{n} -orthonormal factor assumption. The simulations are performed with $\alpha = 1/G$, $\eta = \varepsilon = 1$, $\lambda = 0.001$, $\lambda_0 = 200$, $\lambda_1 = 1$ and $K = 8$ (in Section 3.6) / 1 (in Section 3.7) using a Gibbs sampler starting from the MAP identified by the PXL-EM algorithm.

C.3 PROOFS

C.3.1 PROOF OF THEOREM 3.4.1

Proof. Let β_1 be the vector formed by the β_{jk} 's with their corresponding $\gamma_{jk} = 1$ and let β_0 be the vector formed by β_{jk} 's with their corresponding $\gamma_{jk} = 0$.

$$\begin{aligned} \pi(\mathbf{B}|\mathbf{Y}, \boldsymbol{\Sigma}, \boldsymbol{\Gamma}, m) \propto f_m(\beta_1, \beta_0) &\equiv \prod_{\{j,k:\gamma_{jk}=1\}} \varphi_m(\beta_{jk}) \prod_{\{j,k:\gamma_{jk}=0\}} \psi(\beta_{jk}) \\ &\times |\mathbf{B}\mathbf{B}^T + \boldsymbol{\Sigma}|^{-\frac{n}{2}} \exp \left\{ -\frac{1}{2} \text{tr} \left[(\mathbf{B}\mathbf{B}^T + \boldsymbol{\Sigma})^{-1} \left(\sum_{i=1}^n \mathbf{y}_i \mathbf{y}_i^T \right) \right] \right\} \end{aligned} \quad (\text{C.3})$$

Let $\lambda_1(\mathcal{M}) \geq \dots \geq \lambda_G(\mathcal{M})$ denote the eigenvalues of a matrix \mathcal{M} and let $\mu_1 \geq \dots \geq \mu_G$ be the eigenvalues of $\mathbf{B}\mathbf{B}^T + \boldsymbol{\Sigma}$. According to Weyl's inequality,

$$\lambda_j(\mathbf{B}\mathbf{B}^T) + \lambda_1(\boldsymbol{\Sigma}) \geq \mu_j \geq \lambda_j(\boldsymbol{\Sigma}), \quad j = 1, \dots, G,$$

we have

$$\begin{aligned} \sum_{j=1}^G \frac{\lambda_j(\mathbf{Y}\mathbf{Y}^T)}{\lambda_j(\mathbf{B}\mathbf{B}^T) + \lambda_1(\boldsymbol{\Sigma})} &\leq \sum_{j=1}^G \frac{\lambda_j(\mathbf{Y}\mathbf{Y}^T)}{\mu_j} \leq \text{tr} \left[(\mathbf{B}\mathbf{B}^T + \boldsymbol{\Sigma})^{-1} \left(\sum_{i=1}^n \mathbf{y}_i \mathbf{y}_i^T \right) \right] \\ &\leq \sum_{j=1}^G \frac{\lambda_j(\mathbf{Y}\mathbf{Y}^T)}{\mu_{G+1-j}} \leq \sum_{j=1}^G \frac{\lambda_j(\mathbf{Y}\mathbf{Y}^T)}{\lambda_{G+1-j}(\boldsymbol{\Sigma})} \end{aligned} \quad (\text{C.4})$$

Note that $\lambda_j(\mathbf{B}) = 0$ for $j > K$, so we have:

$$\sum_{j=K+1}^G \frac{\lambda_j(\mathbf{Y}\mathbf{Y}^T)}{\lambda_1(\boldsymbol{\Sigma})} \leq \text{tr} \left[(\mathbf{B}\mathbf{B}^T + \boldsymbol{\Sigma})^{-1} \left(\sum_{i=1}^n \mathbf{y}_i \mathbf{y}_i^T \right) \right] \leq \sum_{j=1}^G \frac{\lambda_j(\mathbf{Y}\mathbf{Y}^T)}{\lambda_{G+1-j}(\boldsymbol{\Sigma})} \quad (\text{C.5})$$

According to the Minkowski determinant theorem, $|\mathbf{B}\mathbf{B}^T + \boldsymbol{\Sigma}| \geq |\boldsymbol{\Sigma}|$. Furthermore,

$$|\mathbf{B}\mathbf{B}^T + \boldsymbol{\Sigma}| = \prod_{j=1}^G \mu_j \leq \prod_{j=1}^G (\lambda_j(\mathbf{B}\mathbf{B}^T) + \lambda_1(\boldsymbol{\Sigma})) \leq (\lambda_1(\boldsymbol{\Sigma}))^{G-K} \prod_{j=1}^K (\|\mathbf{B}\mathbf{B}^T\|_F + \lambda_1(\boldsymbol{\Sigma})).$$

Combining this with (C.5), we have

$$|\mathbf{B}\mathbf{B}^T + \boldsymbol{\Sigma}|^{-n/2} \exp \left\{ -\frac{1}{2} \text{tr} \left[(\mathbf{B}\mathbf{B}^T + \boldsymbol{\Sigma})^{-1} \left(\sum_{i=1}^n \mathbf{y}_i \mathbf{y}_i^T \right) \right] \right\} \leq |\boldsymbol{\Sigma}|^{-n/2} \exp \left(-\frac{1}{2} \sum_{j=K+1}^G \frac{\lambda_j(\mathbf{Y}\mathbf{Y}^T)}{\lambda_1(\boldsymbol{\Sigma})} \right) \quad (\text{C.6})$$

and

$$\begin{aligned} & |\mathbf{B}\mathbf{B}^T + \boldsymbol{\Sigma}|^{-n/2} \exp \left\{ -\frac{1}{2} \text{tr} \left[(\mathbf{B}\mathbf{B}^T + \boldsymbol{\Sigma})^{-1} \left(\sum_{i=1}^n \mathbf{y}_i \mathbf{y}_i^T \right) \right] \right\} \\ & \geq (\lambda_1(\boldsymbol{\Sigma}))^{-n(G-K)/2} \prod_{j=1}^K (\|\mathbf{B}\|_F^2 + \lambda_1(\boldsymbol{\Sigma}))^{-n/2} \exp \left(-\frac{1}{2} \sum_{j=1}^G \frac{\lambda_j(\mathbf{Y}\mathbf{Y}^T)}{\lambda_{G+1-j}(\boldsymbol{\Sigma})} \right). \end{aligned} \quad (\text{C.7})$$

Therefore,

$$\begin{aligned} & \int_{\mathbf{B} \in \mathcal{S}} f_m(\beta_1, \beta_0) d\mathbf{B} \\ & \leq \int_{\mathbf{B} \in \mathcal{S}} d\mathbf{B} |\boldsymbol{\Sigma}|^{-n/2} \exp \left(-\frac{1}{2} \sum_{j=K+1}^G \frac{\lambda_j(\mathbf{Y}\mathbf{Y}^T)}{\lambda_1(\boldsymbol{\Sigma})} \right) (\max_{\beta}(\varphi_m(\beta)))^{\#\{\gamma_{jk}=1\}} (\max_{\beta}(\psi(\beta)))^{\#\{\gamma_{jk}=0\}} \\ & = C_1 (\max_{\beta}(\varphi_m(\beta)))^{\#\{\gamma_{jk}=1\}} \end{aligned} \quad (\text{C.8})$$

For a constant $R > 0$,

$$\begin{aligned}
& \int_{|\beta_0| \leq R} \int_{\beta_1 \in S_m^{\#\{\gamma_{jk}=1\}}} f_m(\beta_1, \beta_0) d\beta_1 d\beta_0 \\
& \geq \int_{|\beta_0| \leq R} \int_{\beta_1 \in S_m^{\#\{\gamma_{jk}=1\}}} \prod_{j=1}^K (\|\mathbf{B}\|_F^2 + \lambda_1(\boldsymbol{\Sigma}))^{-n/2} d\beta_1 d\beta_0 \\
& \times (\lambda_1(\boldsymbol{\Sigma}))^{-n(G-K)/2} \exp\left(-\frac{1}{2} \sum_{j=1}^G \frac{\lambda_j(\mathbf{Y}\mathbf{Y}^T)}{\lambda_{G+1-j}(\boldsymbol{\Sigma})}\right) (C \max_{\beta}(\varphi_m(\beta)))^{\#\{\gamma_{jk}=1\}} (\min_{\beta < R}(\psi(\beta)))^{\#\{\gamma_{jk}=0\}} \\
& \geq C_2 (\max_{\beta}(\varphi_m(\beta)))^{\#\{\gamma_{jk}=1\}} \int_{|\beta_0| \leq R} \int_{\beta_1 \in S_m^{\#\{\gamma_{jk}=1\}}} \prod_{j=1}^K (\|\mathbf{B}\|_F^2 + \lambda_1(\boldsymbol{\Sigma}))^{-n/2} d\beta_1 d\beta_0 \\
& \rightarrow C_2 (\max_{\beta}(\varphi_m(\beta)))^{\#\{\gamma_{jk}=1\}} \int_{|\beta_0| \leq R} \int_{\beta_1 \in \mathcal{R}^{\#\{\gamma_{jk}=1\}}} \prod_{j=1}^K (\|\mathbf{B}\|_F^2 + \lambda_1(\boldsymbol{\Sigma}))^{-n/2} d\beta_1 d\beta_0
\end{aligned} \tag{C.9}$$

as $m \rightarrow \infty$ following the monotone convergence theorem. We also know that

$$\begin{aligned}
& \int_{|\beta_0| \leq R} \int_{\beta_1 \in \mathcal{R}^{\#\{\gamma_{jk}=1\}}} \prod_{j=1}^K (\|\mathbf{B}\|_F^2 + \lambda_1(\boldsymbol{\Sigma}))^{-n/2} d\beta_1 d\beta_0 \\
& \geq \int_{|\beta_0| \leq R} \int_{\beta_1 \in \mathcal{R}^{\#\{\gamma_{jk}=1\}}} \prod_{j=1}^K (\|\beta_1\|^2 + R^2 + \lambda_1(\boldsymbol{\Sigma}))^{-n/2} d\beta_1 d\beta_0 \\
& = \left(\int_{|\beta_0| \leq R} d\beta_0 \right) \int_{\beta_1 \in \mathcal{R}^{\#\{\gamma_{jk}=1\}}} \prod_{j=1}^K (\|\beta_1\|^2 + R^2 + \lambda_1(\boldsymbol{\Sigma}))^{-n/2} \|\beta_1\|^{\#\{\gamma_{jk}=1\}-1} d\|\beta_1\| d(\gamma(\beta_1))
\end{aligned} \tag{C.10}$$

from the polar coordinate transformation, of which the last term goes to infinity since $\#\{\gamma_{jk} = 1\} \geq n \times K$. Taken together, we have shown that

$$\lim_{m \rightarrow \infty} \frac{\int_{\mathbf{B} \in \mathcal{S}} f_m(\beta_1, \beta_0) d\mathbf{B}}{\int_{|\beta_0| \leq R} \int_{\beta_1 \in S_m^{\#\{\gamma_{jk}=1\}}} f_m(\beta_1, \beta_0) d\beta_1 d\beta_0} = 0, \tag{C.11}$$

which implies the theorem. \square

C.3.2 PROOF OF THEOREM 3.4.2

Proof. By marginalizing out $\mathbf{\Omega}$ from the full posterior distribution, we know that:

$$\pi(\mathbf{B}|\mathbf{Y}, \mathbf{\Sigma}, \mathbf{\Gamma}, m) \propto \int f(\mathbf{Y}|\mathbf{B}, \mathbf{\Omega}, \mathbf{\Sigma})f(\mathbf{\Omega})d\mathbf{\Omega} \prod_{\{jk:\gamma_{jk}=1\}} \frac{\varphi_m(\beta_{jk})}{\varphi_m(0)} \prod_{\{jk:\gamma_{jk}=0\}} \psi(\beta_{jk}) = \pi_m^\mu(\mathbf{B}) \quad (\text{C.12})$$

$$\pi(\mathbf{B}|\mathbf{Y}, \mathbf{\Sigma}, \mathbf{\Gamma}, \infty) \propto \int f(\mathbf{Y}|\mathbf{B}, \mathbf{\Omega}, \mathbf{\Sigma})f(\mathbf{\Omega})d\mathbf{\Omega} \prod_{\{jk:\gamma_{jk}=0\}} \psi(\beta_{jk}) = \pi_\infty^\mu(\mathbf{B}) \quad (\text{C.13})$$

For any Borel set S , $\int_S \pi_m^\mu(\mathbf{B})d\mathbf{B} \leq \int_S \pi_\infty^\mu(\mathbf{B})d\mathbf{B} < \infty$, by the dominant convergence theorem we have:

$$\lim_{m \rightarrow \infty} \int_S \pi_m^\mu(\mathbf{B})d\mathbf{B} = \int_S \pi_\infty^\mu(\mathbf{B})d\mathbf{B}, \quad (\text{C.14})$$

$$\lim_{m \rightarrow \infty} \int_S \pi_m^\mu(\mathbf{B})d\mathbf{B} / \int_{\mathcal{R}^{G \times K}} \pi_m^\mu(\mathbf{B})d\mathbf{B} = \int_S \pi_\infty^\mu(\mathbf{B})d\mathbf{B} / \int_{\mathcal{R}^{G \times K}} \pi_\infty^\mu(\mathbf{B})d\mathbf{B}. \quad (\text{C.15})$$

This means that $\mathbf{B}|\mathbf{Y}, \mathbf{\Sigma}, \mathbf{\Gamma}, m$ converges to $\mathbf{B}|\mathbf{Y}, \mathbf{\Sigma}, \mathbf{\Gamma}, \infty$ in distribution as $m \rightarrow \infty$. \square

C.3.3 PROOF OF LEMMA 3.5.1

Proof. For $\varepsilon > 0$ and $L > 0$,

$$P(\|\mathbf{V}(\mathbf{\Omega}_{0,n})^\perp \mathbf{V}(\mathbf{\Omega})^T\|_F > \varepsilon | \mathbf{Y}, \mathbf{\Sigma}_G) \leq 1 / \left(1 + \frac{P(\|\mathbf{V}(\mathbf{\Omega}_{0,n})^\perp \mathbf{V}(\mathbf{\Omega})^T\|_F < \varepsilon / L | \mathbf{Y}, \mathbf{\Sigma}_G)}{P(\|\mathbf{V}(\mathbf{\Omega}_{0,n})^\perp \mathbf{V}(\mathbf{\Omega})^T\|_F > \varepsilon | \mathbf{Y}, \mathbf{\Sigma}_G)} \right). \quad (\text{C.16})$$

From

$$\pi(d\mathbf{V}(\mathbf{\Omega})|\mathbf{Y}, \mathbf{\Sigma}) \propto \exp\left(\sum_{j=1}^G \frac{1}{2\sigma_j^2} \|\mathcal{P}_{\mathbf{V}(\mathbf{\Omega})}(\mathbf{Y}_j)\|^2\right) m(d\mathbf{V}(\mathbf{\Omega})), \quad (\text{C.17})$$

we can compute

$$\begin{aligned}
& P(\|\mathbf{V}(\boldsymbol{\Omega}_{0,n})^\perp \mathbf{V}(\boldsymbol{\Omega})^T\|_F < \varepsilon/L | \mathbf{Y}, \boldsymbol{\Sigma}_G) \\
&= C \int_{\{\mathbf{V}(\boldsymbol{\Omega}) : \|\mathbf{V}(\boldsymbol{\Omega}_{0,n})^\perp \mathbf{V}(\boldsymbol{\Omega})^T\|_F < \varepsilon/L\}} \exp\left(-\sum_{j=1}^G \frac{1}{2\sigma_j^2} \|\mathcal{P}_{\mathbf{V}(\boldsymbol{\Omega})^\perp}(\mathbf{Y}_j)\|^2\right) m(d\mathbf{V}(\boldsymbol{\Omega})) \\
&= C \int_{\{\mathbf{V}(\boldsymbol{\Omega}) : \|\mathbf{V}(\boldsymbol{\Omega}_{0,n})^\perp \mathbf{V}(\boldsymbol{\Omega})^T\|_F < \varepsilon/L\}} \exp\left(-\frac{1}{2} \|\mathbf{V}(\boldsymbol{\Omega})^\perp \mathbf{V}(\boldsymbol{\Omega}_{0,n})^T \mathbf{K}(\boldsymbol{\Omega}_{0,n})^T \mathbf{B}_{0,G}^T \boldsymbol{\Sigma}_G^{-1/2}\|_F^2\right) m(d\mathbf{V}(\boldsymbol{\Omega})) \\
&\geq C \int_{\{\mathbf{V}(\boldsymbol{\Omega}) : \|\mathbf{V}(\boldsymbol{\Omega}_{0,n})^\perp \mathbf{V}(\boldsymbol{\Omega})^T\|_F < \varepsilon/L\}} \exp\left(-\frac{1}{2} \|\mathbf{V}(\boldsymbol{\Omega})^\perp \mathbf{V}(\boldsymbol{\Omega}_{0,n})^T\|_F^2 \lambda_{\max}(\boldsymbol{\Sigma}_G^{-1/2} \mathbf{B}_{0,G} \mathbf{K}(\boldsymbol{\Omega}_{0,n}))^2\right) m(d\mathbf{V}(\boldsymbol{\Omega})) \\
&\geq C m_n(\{\mathbf{V} : \|\mathbf{V}_0 \mathbf{V}^T\|_F < \frac{\varepsilon}{L}\}) \times \exp\left(-\frac{1}{2} \frac{\varepsilon^2}{L^2} \lambda_{\max}(\boldsymbol{\Sigma}_G^{-1/2} \mathbf{B}_{0,G} \mathbf{K}(\boldsymbol{\Omega}_{0,n}))^2\right),
\end{aligned} \tag{C.18}$$

where \mathbf{V}_0 is a $K \times n$ orthonormal matrix. Similarly, we can derive

$$\begin{aligned}
& P(\|\mathbf{V}(\boldsymbol{\Omega}_{0,n})^\perp \mathbf{V}(\boldsymbol{\Omega})^T\|_F > \varepsilon | \mathbf{Y}, \boldsymbol{\Sigma}_G) \\
&\leq C m_n(\{\mathbf{V} : \|\mathbf{V}_0 \mathbf{V}^T\|_F > \varepsilon\}) \exp\left(-\frac{1}{2} \varepsilon^2 \lambda_{\min}(\boldsymbol{\Sigma}_G^{-1/2} \mathbf{B}_{0,G} \mathbf{K}(\boldsymbol{\Omega}_{0,n}))^2\right).
\end{aligned} \tag{C.19}$$

Inserting (C.18) and (C.19) to (C.16), we complete the proof. \square

C.3.4 PROOF OF THEOREM 3.5.1

Proof. First, we show a strong uniform law of large number that:

$$\lim_{G \rightarrow \infty} \sup_{\boldsymbol{\Omega}} \left| \frac{1}{G} \sum_{j=1}^G \frac{1}{2\sigma_j^2} \|\mathcal{P}_{\mathbf{V}(\boldsymbol{\Omega})}(\mathbf{Y}_j)\|^2 - \frac{1}{G} \sum_{j=1}^G \frac{1}{2\sigma_j^2} \mathbb{E} \|\mathcal{P}_{\mathbf{V}(\boldsymbol{\Omega})}(\mathbf{Y}_j)\|^2 \right| = 0 \text{ a.s.} \tag{C.20}$$

Define the inner part of the absolute value on left-hand side of (C.20) as $D_G(\mathbf{\Omega}, \mathbf{Y})$. We know for $\mathbf{\Omega}$ and $\mathbf{\Omega}_1$,

$$\begin{aligned} \left| \|\mathcal{P}_{\mathbf{V}(\mathbf{\Omega})}(\mathbf{Y}_j)\|^2 - \|\mathcal{P}_{\mathbf{V}(\mathbf{\Omega}_1)}(\mathbf{Y}_j)\|^2 \right| &= \mathbf{Y}_j^T (\mathbf{P}_{\mathbf{V}(\mathbf{\Omega})} - \mathbf{P}_{\mathbf{V}(\mathbf{\Omega}_1)}) \mathbf{Y}_j \\ &\leq 2\sqrt{K(n-K)} \|\mathbf{V}(\mathbf{\Omega}_1)^\perp \mathbf{V}(\mathbf{\Omega})^T\|_F \|\mathbf{Y}_j\|^2 \end{aligned} \quad (\text{C.21})$$

and

$$\left| \mathbb{E} \|\mathcal{P}_{\mathbf{V}(\mathbf{\Omega})}(\mathbf{Y}_j)\|^2 - \mathbb{E} \|\mathcal{P}_{\mathbf{V}(\mathbf{\Omega}_1)}(\mathbf{Y}_j)\|^2 \right| \leq 2\sqrt{K(n-K)} \|\mathbf{V}(\mathbf{\Omega}_1)^\perp \mathbf{V}(\mathbf{\Omega})^T\|_F \|\mathbf{\Omega}_{0,n}^T(\mathbf{B}_0)_j\|^2. \quad (\text{C.22})$$

Thus

$$\begin{aligned} |D_G(\mathbf{\Omega}, \mathbf{Y}) - D_G(\mathbf{\Omega}_1, \mathbf{Y})| &\leq 2\sqrt{K(n-K)} \|\mathbf{V}(\mathbf{\Omega}_1)^\perp \mathbf{V}(\mathbf{\Omega})^T\|_F \\ &\quad \times \left(\frac{1}{G} \sum_{j=1}^G \frac{1}{2\sigma_j^2} \|\mathbf{Y}_j\|^2 + \frac{1}{G} \sum_{j=1}^G \frac{1}{2\sigma_j^2} \|\mathbf{\Omega}_{0,n}^T(\mathbf{B}_0)_j\|^2 \right). \end{aligned} \quad (\text{C.23})$$

In order to apply the Kolmogorov's strong law of large number, we check the variance of $\frac{1}{2\sigma_j^2} \|\mathcal{P}_{\mathbf{V}(\mathbf{\Omega}_1)}(\mathbf{Y}_j)\|^2$ and $\frac{1}{2\sigma_j^2} \|\mathbf{Y}_j\|^2$:

$$\text{Var}\left(\frac{1}{2\sigma_j^2} \|\mathcal{P}_{\mathbf{V}(\mathbf{\Omega}_1)}(\mathbf{Y}_j)\|^2\right) = \frac{1}{\sigma_j^2} \|\mathbf{V}(\mathbf{\Omega}_1) \mathbf{\Omega}_{0,n}^T(\mathbf{B}_0)_j\|^2 + K/2 \quad (\text{C.24})$$

$$\text{Var}\left(\frac{1}{2\sigma_j^2} \|\mathbf{Y}_j\|^2\right) = \frac{1}{\sigma_j^2} \|\mathbf{\Omega}_{0,n}^T(\mathbf{B}_0)_j\|^2 + n/2. \quad (\text{C.25})$$

Both of them are uniformly upper bounded with respect to j . So by Kolmogorov's strong law, we have for every fixed $\mathbf{\Omega}_1$, $D_G(\mathbf{\Omega}_1, \mathbf{Y})$ is almost surely converging to 0 as $G \rightarrow \infty$ and

$$\frac{1}{G} \sum_{j=1}^G \frac{1}{2\sigma_j^2} \|\mathbf{Y}_j\|^2 - \frac{1}{G} \sum_{j=1}^G \left(\frac{1}{2\sigma_j^2} \|\mathbf{\Omega}_{0,n}^T(\mathbf{B}_0)_j\|^2 + n/2 \right) \rightarrow 0 \text{ a.s.} \quad (\text{C.26})$$

For a fixed $\varepsilon > 0$, define a neighborhood $U_{\mathbf{V}(\mathbf{\Omega}_1)}$ for every $\mathbf{V}(\mathbf{\Omega}_1)$,

$$U_{\mathbf{V}(\mathbf{\Omega}_1)} = \left\{ \mathbf{V} : \|\mathbf{V}(\mathbf{\Omega}_1)^\perp \mathbf{V}^T\|_F < \frac{\varepsilon}{4\sqrt{K(n-K)}} \left(\|\mathbf{\Omega}_{0,n}\|_F^2 \max_j \left\| \frac{(\mathbf{B}_0)_j}{\sigma_j} \right\|^2 + n/2 + \varepsilon \right)^{-1}, \right. \\ \left. \mathbf{V} \text{ is an orthonormal } K\text{-frames in } \mathbb{R}^n \right\} \quad (\text{C.27})$$

Let \mathcal{V} denote the Stiefel manifold $St(K, n)$, then there exists $\mathbf{\Omega}_1, \mathbf{\Omega}_2, \dots, \mathbf{\Omega}_m$ such that $\mathcal{V} = \bigcup_{t=1}^m U_{\mathbf{V}(\mathbf{\Omega}_t)}$.

For $t = 1, \dots, m$, $D_G(\mathbf{\Omega}_t, \mathbf{Y}) \rightarrow 0$ almost surely, let \mathcal{Y} denotes the realizations of \mathbf{Y} such that $D_G(\mathbf{\Omega}_t, \mathbf{Y}) \rightarrow 0$ for all t and

$$\frac{1}{G} \sum_{j=1}^G \frac{1}{2\sigma_j^2} \|\mathbf{Y}_j\|^2 - \frac{1}{G} \sum_{j=1}^G \left(\frac{1}{2\sigma_j^2} \|\mathbf{\Omega}_{0,n}^T(\mathbf{B}_0)_j\|^2 + n/2 \right) \rightarrow 0.$$

By definition $P(\mathcal{Y}) = 1$, for a realization \mathbf{y} in \mathcal{Y} there exist G_0, G_1, \dots, G_m such that

$$\frac{1}{G} \sum_{j=1}^G \frac{1}{2\sigma_j^2} \|\mathbf{y}_j\|^2 - \frac{1}{G} \sum_{j=1}^G \left(\frac{1}{2\sigma_j^2} \|\mathbf{\Omega}_{0,n}^T(\mathbf{B}_0)_j\|^2 + n/2 \right) < \varepsilon/2, \text{ for } G > G_0 \quad (\text{C.28})$$

$$D_G(\mathbf{\Omega}_t, \mathbf{y}) < \varepsilon/2, \text{ for } G > G_t, t = 1, \dots, m \quad (\text{C.29})$$

When $G > \max_t \{G_t\}$, for any $\mathbf{\Omega}$, there exists $\mathbf{\Omega}_{t_0}$ such that $\mathbf{V}(\mathbf{\Omega}) \in U_{\mathbf{V}(\mathbf{\Omega}_{t_0})}$, by (C.23), (C.28) and (C.29):

$$|D_G(\mathbf{\Omega}, \mathbf{y})| \leq |D_G(\mathbf{\Omega}, \mathbf{y}) - D_G(\mathbf{\Omega}_{t_0}, \mathbf{y})| + |D_G(\mathbf{\Omega}_{t_0}, \mathbf{y})| \leq \varepsilon \quad (\text{C.30})$$

From here we have proved (C.20).

Combined with lemma 5.1, we know when $G > \max_t \{G_t\}$,

$$\begin{aligned} & P(\|\mathbf{V}(\mathbf{\Omega}_{0,n})^\perp \mathbf{V}(\mathbf{\Omega})^T\|_F < \tilde{\varepsilon}/L | \mathbf{y}, \mathbf{\Sigma}_G) \\ &= C \int_{\{\mathbf{V}(\mathbf{\Omega}) : \|\mathbf{V}(\mathbf{\Omega}_{0,n})^\perp \mathbf{V}(\mathbf{\Omega})^T\|_F < \tilde{\varepsilon}/L\}} \exp\left(\sum_{j=1}^G \frac{1}{2\sigma_j^2} \|\mathcal{P}_{\mathbf{V}(\mathbf{\Omega})}(\mathbf{y}_j)\|^2\right) m(d\mathbf{V}(\mathbf{\Omega})) \\ &= \tilde{C} \int_{\{\mathbf{V}(\mathbf{\Omega}) : \|\mathbf{V}(\mathbf{\Omega}_{0,n})^\perp \mathbf{V}(\mathbf{\Omega})^T\|_F < \tilde{\varepsilon}/L\}} \exp\left(G D_G(\mathbf{\Omega}, \mathbf{y}) \right. \\ &\quad \left. - \frac{1}{2} \|\mathbf{V}(\mathbf{\Omega})^\perp \mathbf{V}(\mathbf{\Omega}_{0,n})^T \mathbf{K}(\mathbf{\Omega}_{0,n})^T \mathbf{B}_{0,G}^T \mathbf{\Sigma}_G^{-1/2}\|_F^2\right) m(d\mathbf{V}(\mathbf{\Omega})) \\ &\geq \tilde{C} m_n(\{\mathbf{V} : \|\mathbf{V}_0 \mathbf{V}^T\|_F < \frac{\tilde{\varepsilon}}{L}\}) \exp\left(-\frac{1}{2} \frac{\tilde{\varepsilon}^2}{L^2} \lambda_{\max}(\mathbf{\Sigma}_G^{-1/2} \mathbf{B}_{0,G} \mathbf{K}(\mathbf{\Omega}_{0,n}))^2 - G\varepsilon\right), \end{aligned} \quad (\text{C.31})$$

and on the other hand,

$$\begin{aligned} & P(\|\mathbf{V}(\mathbf{\Omega}_{0,n})^\perp \mathbf{V}(\mathbf{\Omega})^T\|_F > \tilde{\varepsilon} | \mathbf{y}, \mathbf{\Sigma}_G) \\ &\leq \tilde{C} m_n(\{\mathbf{V} : \|\mathbf{V}_0 \mathbf{V}^T\|_F > \tilde{\varepsilon}\}) \exp\left(-\frac{1}{2} \tilde{\varepsilon}^2 \lambda_{\min}(\mathbf{\Sigma}_G^{-1/2} \mathbf{B}_{0,G} \mathbf{K}(\mathbf{\Omega}_{0,n}))^2 + G\varepsilon\right). \end{aligned} \quad (\text{C.32})$$

Therefore we have

$$\begin{aligned} \frac{P(\|\mathbf{V}(\mathbf{\Omega}_{0,n})^\perp \mathbf{V}(\mathbf{\Omega})^T\|_F < \tilde{\varepsilon}/L | \mathbf{y}, \mathbf{\Sigma}_G)}{P(\|\mathbf{V}(\mathbf{\Omega}_{0,n})^\perp \mathbf{V}(\mathbf{\Omega})^T\|_F > \tilde{\varepsilon} | \mathbf{y}, \mathbf{\Sigma}_G)} &\geq m_n(\{\mathbf{V} : \|\mathbf{V}_0 \mathbf{V}^T\|_F < \frac{\tilde{\varepsilon}}{L}\}) \\ &\quad \times \exp\left(\frac{3}{8} \tilde{\varepsilon}^2 \lambda_{\min}(\mathbf{\Sigma}_G^{-1/2} \mathbf{B}_{0,G} \mathbf{K}(\mathbf{\Omega}_{0,n}))^2 - 2G\varepsilon\right) \end{aligned} \quad (\text{C.33})$$

Since $\lambda_{\min}(\mathbf{B}_{0,G})/\sqrt{G}$ is lower bounded, $\lambda_{\min}(\mathbf{\Sigma}_G^{-1/2} \mathbf{B}_{0,G} \mathbf{K}(\mathbf{\Omega}_{0,n}))/\sqrt{G}$ is also lower bounded. Se-

lect ε such that

$$\varepsilon \leq \frac{1}{8} \tilde{\varepsilon}^2 \left(\lambda_{\min}(\boldsymbol{\Sigma}_G^{-1/2} \mathbf{B}_{0,G} \mathbf{K}(\boldsymbol{\Omega}_{0,n})) / \sqrt{G} \right)^2,$$

then the right hand side of (C.33) is no smaller than

$$m_n(\{\mathbf{V} : \|\mathbf{V}_0 \mathbf{V}^T\|_F < \frac{\tilde{\varepsilon}}{L}\}) \times \exp\left(\frac{1}{8} \tilde{\varepsilon}^2 \lambda_{\min}(\boldsymbol{\Sigma}_G^{-1/2} \mathbf{B}_{0,G} \mathbf{K}(\boldsymbol{\Omega}_{0,n}))^2\right).$$

which goes to infinity by the lower boundedness of $\lambda_{\min}(\mathbf{B}_{0,G}) / \sqrt{G}$.

Thus $\|\mathbf{V}(\boldsymbol{\Omega}_{0,n})^\perp \mathbf{V}(\boldsymbol{\Omega})^T\|_F | \mathbf{y}, \boldsymbol{\Sigma} \rightarrow 0$ in probability for every \mathbf{y} in \mathcal{Y} which leads to the conclusion. □

The spirit of this proof is essentially the same as that of the classical Bayesian consistency theorem, but is involved with infinite-dimensional potential data. In theorem 3.5.1, we made the assumption that the L_2 norm of rows of \mathbf{B}_0 are upper bounded due to the proof, which restricted ourselves to the case where all singular values of $\mathbf{B}_{0,G}$ are increasing at the order of \sqrt{G} . This condition can be satisfied when rows of \mathbf{B}_0 are i.i.d from an underlying distribution p_B :

$$\lambda_k(\mathbf{B}_{0,G}) / \sqrt{G} = \sqrt{\lambda_k(\mathbf{B}_{0,G}^T \mathbf{B}_{0,G} / G)} \rightarrow \sqrt{\lambda_k(E_{p_B}(\mathbf{B}_j \mathbf{B}_j^T))}, \quad G \rightarrow \infty \text{ a.s.}$$

C.3.5 REMARK OF SECTION 3.5.1

From [Cai et al. \(2018\)](#), for every pair of $\mathbf{V}(\boldsymbol{\Omega}_{0,n})$ and $\mathbf{V}(\boldsymbol{\Omega})$ there exists an orthogonal matrix \mathbf{W} such that $\|\mathbf{V}(\boldsymbol{\Omega}) - \mathbf{W} \mathbf{V}(\boldsymbol{\Omega}_{0,n})\|_F \leq \sqrt{2} \|\sin(\angle(\mathbf{V}(\boldsymbol{\Omega}_{0,n}), \mathbf{V}(\boldsymbol{\Omega})))\|_F$ where $\angle(\mathbf{V}(\boldsymbol{\Omega}_{0,n}), \mathbf{V}(\boldsymbol{\Omega}))$ denotes the diagonal matrix formed by canonical angles between row spaces of $\boldsymbol{\Omega}_{0,n}$ and $\boldsymbol{\Omega}$. For fixed n and $G = s \rightarrow \infty$, using the shrinkage of canonical angles between row spaces from Theorem 3.5.1, there exists a orthogonal random matrix \mathbf{W} such that $\|\mathbf{V}(\boldsymbol{\Omega}) - \mathbf{W} \mathbf{V}(\boldsymbol{\Omega}_{0,n})\|_F | \mathbf{Y}, \boldsymbol{\Sigma} \rightarrow 0$

in probability as $G \rightarrow \infty$. The posterior distribution of $\mathbf{V}(\boldsymbol{\Omega})$ conditioned on the row vector space of $\boldsymbol{\Omega}$ is actually an uniform distribution on all the orthonormal basis within since the density in (3.10) involves $\mathbf{V}(\boldsymbol{\Omega})$ only through the row vector space. Therefore $\mathbf{V}(\boldsymbol{\Omega})|\mathbf{Y}, \boldsymbol{\Sigma} \sim \mathbf{O}_1\mathbf{V}(\boldsymbol{\Omega})|\mathbf{Y}, \boldsymbol{\Sigma} \sim \mathbf{O}_1(\mathbf{W}\mathbf{V}(\boldsymbol{\Omega}_{0,n}) + (\mathbf{V}(\boldsymbol{\Omega}) - \mathbf{W}\mathbf{V}(\boldsymbol{\Omega}_{0,n})))|\mathbf{Y}, \boldsymbol{\Sigma}$ for an independent uniform random orthogonal matrix \mathbf{O}_1 . Since $\|\mathbf{O}_1(\mathbf{V}(\boldsymbol{\Omega}) - \mathbf{W}\mathbf{V}(\boldsymbol{\Omega}_{0,n}))\|_F|\mathbf{Y}, \boldsymbol{\Sigma} \rightarrow 0$, the posterior sample of $\mathbf{V}(\boldsymbol{\Omega})$ can be asymptotically express as $\mathbf{O}\mathbf{V}(\boldsymbol{\Omega}_{0,n})$ where $\mathbf{O} = \mathbf{O}_1\mathbf{W}$ is an independent uniform random orthogonal matrix, i.e., $\mathbf{V}(\boldsymbol{\Omega})$ differs $\mathbf{O}\mathbf{V}(\boldsymbol{\Omega}_{0,n})$ by a matrix that has Frobenius norm converging to 0 under the asymptotic regime of Theorem 3.5.1.

C.3.6 PROOF OF THEOREM 3.5.2

Theorem 3.5.2 is an immediate result of the following lemma and the proof of Theorem 3.5.1.

Lemma C.3.1. Let (\mathbf{B}_0, Γ_0) be a regular infinite loading pair with Γ_0 known, $\boldsymbol{\Omega}_0$ be a $K \times \infty$ matrix and $\boldsymbol{\Sigma} = \text{diag}(\sigma_1^2, \dots)$ be a known infinite diagonal matrix. Define $\boldsymbol{\Sigma}_G = \text{diag}(\sigma_{\pi^{-1}(1)}^2, \dots, \sigma_{\pi^{-1}(G)}^2)$ and $\boldsymbol{\Sigma}_G^{(k)} = \text{diag}(\sigma_{\pi^{-1}(l_{0,k})}^2, \dots, \sigma_{\pi^{-1}(l_{0,k+1}-1)}^2)$. $\boldsymbol{\Omega}_{0,n}$ denotes the matrix formed by the first n columns of $\boldsymbol{\Omega}$. Suppose there exists an $\varepsilon > 0$ such that the following holds for the increasing pair $(n, G) = \{(n_t, G_t)\}_{t=1, \dots}$

1. $\min_{k'} \lambda_{\min}((\boldsymbol{\Sigma}_G^{(k')})^{-1/2} \mathbf{B}_{0,G}^{(k')} \mathbf{K}(\boldsymbol{\Omega}_{0,n})_{1:k'}) \rightarrow \infty$ as $t \rightarrow \infty$.
2. Let \mathbf{V}_0 be any fixed $K \times n$ orthonormal matrix,

$$-\log(m_n \left(\bigcap_{k=1}^K \left\{ \mathbf{V} : \|(\mathbf{V}_0)_{1:k}^\perp \mathbf{V}_{1:k}^T\|_F < \frac{\varepsilon \min_{k'} \lambda_{\min}((\boldsymbol{\Sigma}_G^{(k')})^{-1/2} \mathbf{B}_{0,G}^{(k')} \mathbf{K}(\boldsymbol{\Omega}_{0,n})_{1:k'})}{\lambda_{\max}((\boldsymbol{\Sigma}_G^{(k)})^{-1/2} \mathbf{B}_{0,G}^{(k)} \mathbf{K}(\boldsymbol{\Omega}_{0,n})_{1:k})} \right\} \right))$$

$$= o(\varepsilon^2 \min_{k'} \lambda_{\min}((\boldsymbol{\Sigma}_G^{(k')})^{-1/2} \mathbf{B}_{0,G}^{(k')} \mathbf{K}(\boldsymbol{\Omega}_{0,n})_{1:k'})^2) \text{ as } t \rightarrow \infty.$$

Let $\mathbf{Y} = \mathbf{B}_{0,G} \boldsymbol{\Omega}_{0,n}$ and model $\mathbf{Y}_{\cdot i}$ with $\mathcal{N}_G(\mathbf{B}\boldsymbol{\Omega}_{\cdot i}, \boldsymbol{\Sigma}_G)$ for $i = 1, \dots, n$. Impose a point mass and flat mixture prior on entries of \mathbf{B} according to the feature allocation matrix $\Gamma_{0,G}$ and assume a

distribution on $\mathbf{\Omega}$ that is invariant under right orthogonal transformations, then for a random draw $\mathbf{\Omega}$ from its posterior distribution,

$$P\left(\bigcup_{k=1}^K \{\mathbf{V} : \|\mathbf{V}(\mathbf{\Omega}_{0,n})_{1:k}^\perp \mathbf{V}(\mathbf{\Omega})_{1:k}^T\|_F > \sqrt{K+1}\varepsilon\} | \mathbf{Y}, \mathbf{\Sigma}_G, \mathbf{\Gamma}_{0,G}\right) \rightarrow 0$$

as $t \rightarrow \infty$.

Proof. We know that for $f(n, G) = \varepsilon \min_{k'} \lambda_{\min}((\mathbf{\Sigma}_G^{(k')})^{-1/2} \mathbf{B}_{0,G}^{(k')} \mathbf{K}(\mathbf{\Omega}_{0,n})_{1:k'})$:

1'. $f(n, G)$ goes to infinity.

2'. Let \mathbf{V}_0 be a fixed $K \times n$ orthonormal matrix,

$$-\log(m_n(\bigcap_{k=1}^K \left\{ \mathbf{V} : \|(\mathbf{V}_0)_{1:k}^\perp \mathbf{V}_{1:k}^T\|_F < \frac{f(n, G)}{\lambda_{\max}((\mathbf{\Sigma}_G^{(k)})^{-1/2} \mathbf{B}_{0,G}^{(k)} \mathbf{K}(\mathbf{\Omega}_{0,n})_{1:k})} \right\})) = o(f(n, G)^2).$$

Define two disjoint set S_1 and S_2 as following

$$S_1 = \bigcap_{k=1}^K \left\{ \mathbf{V} : \|\mathbf{V}(\mathbf{\Omega}_{0,n})_{1:k}^\perp \mathbf{V}_{1:k}^T\|_F < \frac{f(n, G)}{\lambda_{\max}((\mathbf{\Sigma}_G^{(k)})^{-1/2} \mathbf{B}_{0,G}^{(k)} \mathbf{K}(\mathbf{\Omega}_{0,n})_{1:k})} \right\}$$

$$S_2 = \bigcup_{k=1}^K \left\{ \mathbf{V} : \|\mathbf{V}(\mathbf{\Omega}_{0,n})_{1:k}^\perp \mathbf{V}_{1:k}^T\|_F > \frac{\sqrt{K+1}f(n, G)}{\lambda_{\min}((\mathbf{\Sigma}_G^{(k)})^{-1/2} \mathbf{B}_{0,G}^{(k)} \mathbf{K}(\mathbf{\Omega}_{0,n})_{1:k})} \right\}$$

Similar as (C.18), we can compute:

$$\begin{aligned} & P(\mathbf{V}(\mathbf{\Omega}) \in S_1 | \mathbf{Y}, \mathbf{\Sigma}_G, \mathbf{\Gamma}_{0,G}) \\ &= C \int_{S_1} \exp\left(-\frac{1}{2} \sum_{k=1}^K \|\mathbf{V}(\mathbf{\Omega})_{1:k}^\perp \mathbf{V}(\mathbf{\Omega}_{0,n})_{1:k}^T \mathbf{K}(\mathbf{\Omega}_{0,n})_{1:k}^T (\mathbf{B}_{0,G}^{(k)})^T (\mathbf{\Sigma}_G^{(k)})^{-1/2}\|_F^2\right) m(d\mathbf{V}(\mathbf{\Omega})) \quad (\text{C.34}) \\ &\geq C \cdot m_n(S_1) \exp\left(-\frac{K}{2} f(n, G)^2\right) \end{aligned}$$

$$\begin{aligned}
& P(\mathbf{V}(\boldsymbol{\Omega}) \in S_2 | \mathbf{Y}, \boldsymbol{\Sigma}_G, \boldsymbol{\Gamma}_{0,G}) \\
&= C \int_{S_2} \exp\left(-\frac{1}{2} \sum_{k=1}^K \|\mathbf{V}(\boldsymbol{\Omega})_{1:k}^\perp \mathbf{V}(\boldsymbol{\Omega}_{0,n})_{1:k}^T \mathbf{K}(\boldsymbol{\Omega}_{0,n})_{1:k}^T (\mathbf{B}_{0,G}^{(k)})^T (\boldsymbol{\Sigma}_G^{(k)})^{-1/2}\|_F^2\right) m(d\mathbf{V}(\boldsymbol{\Omega})) \quad (\text{C.35}) \\
&\leq C \cdot m_n(S_2) \exp\left(-\frac{K+1}{2} f(n, G)^2\right)
\end{aligned}$$

Combine (C.34) and (C.35), we have:

$$\frac{P(\mathbf{V}(\boldsymbol{\Omega}) \in S_1 | \mathbf{Y}, \boldsymbol{\Sigma}_G, \boldsymbol{\Gamma}_{0,G})}{P(\mathbf{V}(\boldsymbol{\Omega}) \in S_2 | \mathbf{Y}, \boldsymbol{\Sigma}_G, \boldsymbol{\Gamma}_{0,G})} \geq m_n(S_1) \exp\left(\frac{1}{2} f(n, G)^2\right) \quad (\text{C.36})$$

From condition 2', the right hand side goes to infinity for the increasing pair $(n, G) = \{(n_t, G_t)\}_{t=1, \dots}$ as $t \rightarrow \infty$, thus

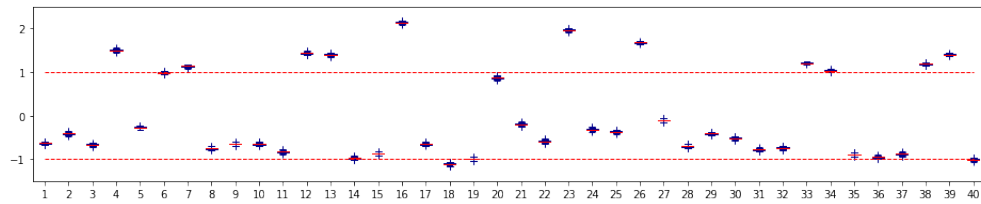
$$P(\mathbf{V}(\boldsymbol{\Omega}) \in S_2 | \mathbf{Y}, \boldsymbol{\Sigma}_G, \boldsymbol{\Gamma}_{0,G}) \rightarrow 0.$$

Therefore,

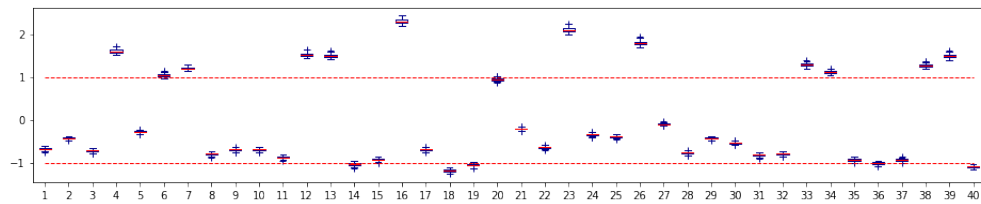
$$P\left(\bigcup_{k=1}^K \{\mathbf{V} : \|\mathbf{V}(\boldsymbol{\Omega}_{0,n})_{1:k}^\perp \mathbf{V}(\boldsymbol{\Omega})_{1:k}^T\|_F > \sqrt{K+1}\varepsilon\} | \mathbf{Y}, \boldsymbol{\Sigma}_G, \boldsymbol{\Gamma}_{0,G}\right) \rightarrow 0 \text{ as } t \rightarrow \infty.$$

□

C.4 ADDITIONAL FIGURES - THE AGEMAP DATASET

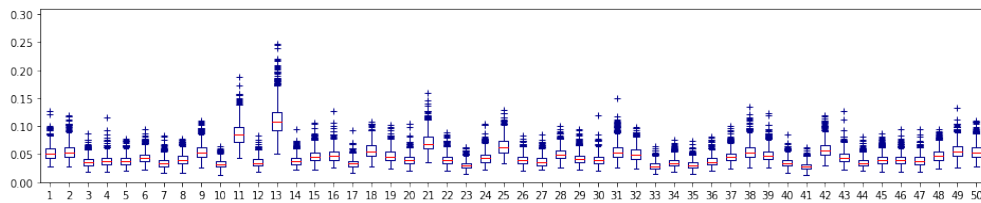


(a) The SpSL-orthonormal factor model

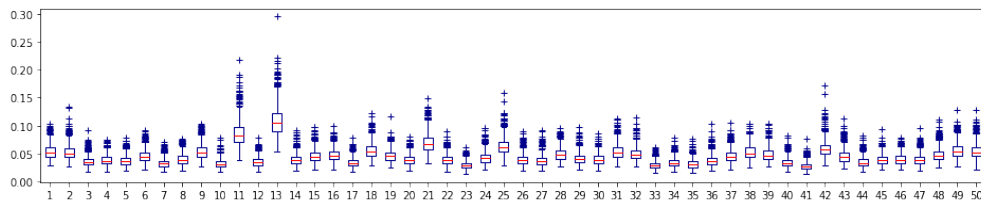


(b) The modified Ghosh-Dunson model

Figure C.1: Boxplots of posterior samples of the latent factors under specified models.

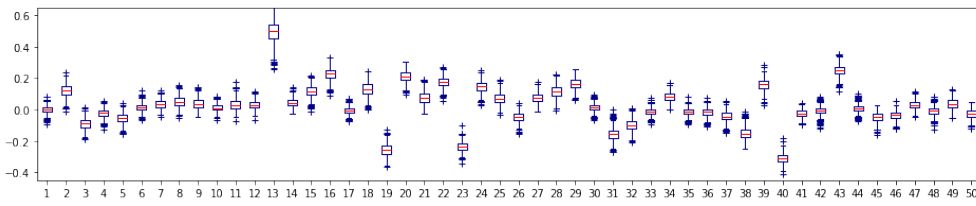


(a) The SpSL-orthonormal factor model

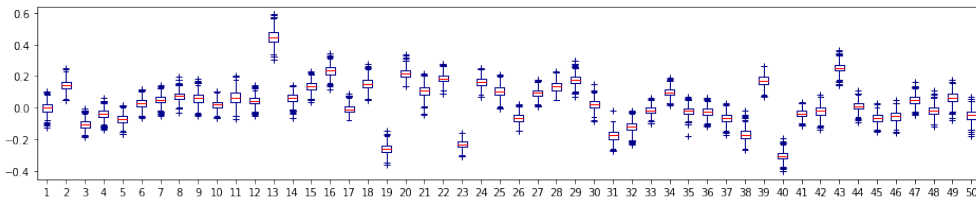


(b) The modified Ghosh-Dunson model

Figure C.2: Boxplots of posterior samples of the first 50 entries of idiosyncratic variances under specified models.



(a) The SpSL-orthonormal factor model



(b) The modified Ghosh-Dunson model

Figure C.3: Boxplots of posterior samples of the first 50 entries of the loading vector under specified models.

References

- 1000 Genomes Project Consortium, A. (2015). A global reference for human genetic variation. *Nature*, 526(7571), 68.
- Arias-Castro, E., Candès, E. J., & Plan, Y. (2011). Global testing under sparse alternatives: Anova, multiple comparisons and the higher criticism. *The Annals of Statistics*, 39(5), 2533–2556.
- Bai, J. & Ng, S. (2002). Determining the number of factors in approximate factor models. *Econometrica*, 70(1), 191–221.
- Baik, J., Ben Arous, G., & Peche, S. (2005). Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *The Annals of Probability*, 33(5), 1643–1697.
- Bao, Z. (2020). Personal communications.
- Barber, R. F. & Candès, E. J. (2015). Controlling the false discovery rate via knockoffs. *The Annals of Statistics*, 43(5), 2055–2085.
- Barnett, I., Mukherjee, R., & Lin, X. (2017). The generalized higher criticism for testing snp-set effects in genetic association studies. *Journal of the American Statistical Association*, 112(517), 64–76.
- Benjamini, Y. & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1), 289–300.
- Bernardo, J., Bayarri, M., Berger, J., Dawid, A., Heckerman, D., Smith, A., & West, M. (2003). Bayesian factor regression models in the “large p, small n” paradigm. *Bayesian statistics*, 7, 733–742.
- Bhattacharya, A. & Dunson, D. B. (2011). Sparse bayesian infinite factor models. *Biometrika*, 98(2), 291.
- Bloemendal, A., Knowles, A., Yau, H.-T., & Yin, J. (2016). On the principal components of sample covariance matrices. *Probability Theory and Related Fields*, 164(1-2), 459–552.
- Braeken, J. & Van Assen, M. A. (2017). An empirical kaiser criterion. *Psychological Methods*, 22(3), 450.

- Cai, T. T., Han, X., & Pan, G. (2020). Limiting laws for divergent spiked eigenvalues and largest nonspiked eigenvalue of sample covariance matrices. *Annals of Statistics*, 48(3), 1255–1280.
- Cai, T. T., Jin, J., & Low, M. G. (2007). Estimation and confidence sets for sparse normal mixtures. *The Annals of Statistics*, 35(6), 2421–2449.
- Cai, T. T., Zhang, A., et al. (2018). Rate-optimal perturbation bounds for singular subspaces with applications to high-dimensional statistics. *The Annals of Statistics*, 46(1), 60–89.
- Candes, E., Fan, Y., Janson, L., & Lv, J. (2018). Panning for gold: model-x?knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(3), 551–577.
- Carvalho, C. M., Chang, J., Lucas, J. E., Nevins, J. R., Wang, Q., & West, M. (2008). High-dimensional sparse factor modeling: applications in gene expression genomics. *Journal of the American Statistical Association*, 103(484), 1438–1456.
- Dai, C., Lin, B., Xing, X., & Liu, J. S. (2020). False discovery rate control via data splitting. *arXiv preprint arXiv:2002.08542*.
- Ding, X. (2020). Spiked sample covariance matrices with possibly multiple bulk components. *Random Matrices: Theory and Applications*, (pp. 2150014).
- Dobriban, E. (2015). Efficient computation of limit spectra of sample covariance matrices. *Random Matrices: Theory and Applications*, 4(04), 1550019.
- Dobriban, E. & Owen, A. B. (2019). Deterministic parallel analysis: an improved method for selecting factors and principal components. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81(1), 163–183.
- Donoho, D. & Jin, J. (2004). Higher criticism for detecting sparse heterogeneous mixtures. *The Annals of Statistics*, 32(3), 962–994.
- Donoho, D. & Jin, J. (2015). Special invited paper: Higher criticism for large-scale inference, especially for rare and weak effects. *Statistical Science*, (pp. 1–25).
- Donoho, D. L., Gavish, M., & Johnstone, I. M. (2018). Optimal shrinkage of eigenvalues in the spiked covariance model. *The Annals of Statistics*, 46(4), 1742.
- Efron, B. (2004). Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *Journal of the American Statistical Association*, 99(465), 96–104.
- Efron, B. E. (1973). Discussion of “marginalization paradoxes in bayesian and structural inference”. *Journal of the Royal Statistical Society*.

- Fan, J., Guo, J., & Zheng, S. (2020). Estimating number of factors by adjusted eigenvalues thresholding. *Journal of the American Statistical Association*, (pp. 1–10).
- Fan, J., Liao, Y., & Mincheva, M. (2013). Large covariance estimation by thresholding principal orthogonal complements. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(4), 603–680.
- Fan, J. & Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5), 849–911.
- Fan, Y., Demirkaya, E., Li, G., & Lv, J. (2019). Rank: large-scale inference with graphical nonlinear knockoffs. *Journal of the American Statistical Association*, (pp. 1–43).
- Fruehwirth-Schnatter, S. & Lopes, H. F. (2018). Sparse bayesian factor analysis when the number of factors is unknown. *arXiv preprint arXiv:1804.04231*.
- Gavish, M. & Donoho, D. L. (2014). The optimal hard threshold for singular values is $4/\sqrt{3}$. *IEEE Transactions on Information Theory*, 60(8), 5040–5053.
- Gelfand, A. E. & Smith, A. F. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American statistical association*, 85(410), 398–409.
- Genovese, C. R., Jin, J., Wasserman, L., & Yao, Z. (2012). A comparison of the lasso and marginal regression. *Journal of Machine Learning Research*, 13(Jun), 2107–2143.
- Ghosh, J. & Dunson, D. B. (2009). Default prior distributions and efficient posterior computation in bayesian factor analysis. *Journal of Computational and Graphical Statistics*, 18(2), 306–320.
- Gordon, G. J., Jensen, R. V., Hsiao, L.-L., Gullans, S. R., Blumenstock, J. E., Ramaswamy, S., Richards, W. G., Sugarbaker, D. J., & Bueno, R. (2002). Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma. *Cancer Research*, 62(17), 4963–4967.
- Götze, F., Tikhomirov, A., et al. (2004). Rate of convergence in probability to the marchenko-pastur law. *Bernoulli*, 10(3), 503–548.
- Hall, P. & Jin, J. (2010). Innovated higher criticism for detecting sparse signals in correlated noise. *The Annals of Statistics*, 38(3), 1686–1732.
- Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, 30(2), 179–185.
- Horn, R. A. & Johnson, C. R. (2012). *Matrix analysis*. Cambridge university press.
- Jager, L. & Wellner, J. A. (2007). Goodness-of-fit tests via phi-divergences. *The Annals of Statistics*, 35(5), 2018–2053.

- Javanmard, A. & Javadi, H. (2019). False discovery rate control via debiased lasso. *Electronic Journal of Statistics*, 13(1), 1212–1253.
- Ji, H. C. (2020). Regularity properties of free multiplicative convolution on the positive line. *International Mathematics Research Notices*.
- Ji, P. & Jin, J. (2012). UPS delivers optimal phase diagram in high-dimensional variable selection. *The Annals of Statistics*, 40(1), 73–103.
- Jia, Z. & Xu, S. (2007). Mapping quantitative trait loci for expression abundance. *Genetics*, 176(1), 611–623.
- Jin, J. & Ke, Z. T. (2016). Rare and weak effects in large-scale inference: methods and phase diagrams. *Statistica Sinica*, (pp. 1–34).
- Jin, J., Ke, Z. T., & Wang, W. (2017). Phase transitions for high dimensional clustering and related problems. *The Annals of Statistics*, 45(5), 2151–2189.
- Jin, J. & Wang, W. (2016). Influential features PCA for high dimensional clustering. *The Annals of Statistics*, 44(6), 2323–2359.
- Jin, J., Zhang, C.-H., & Zhang, Q. (2014). Optimality of graphlet screening in high dimensional variable selection. *The Journal of Machine Learning Research*, 15(1), 2723–2772.
- Johnstone, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *The Annals of Statistics*, 29(2), 295–327.
- Kaiser, H. F. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, 23(3), 187–200.
- Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement*, 20(1), 141–151.
- Kass, R. E. & Wasserman, L. (1996). The selection of prior distributions by formal rules. *Journal of the American Statistical Association*, 91(435), 1343–1370.
- Ke, T., Jin, J., & Fan, J. (2014). Covariance assisted screening and estimation. *The Annals of statistics*, 42(6), 2202.
- Ke, Z. T. (2016). Detecting rare and weak spikes in large covariance matrices. *arXiv preprint arXiv:1609.00883*.
- Ke, Z. T., Liu, J. S., & Ma, Y. (2020a). Power of fdr control methods: The impact of ranking algorithm, tampered design, and symmetric statistic. *arXiv preprint arXiv:2010.08132*.
- Ke, Z. T., Ma, Y., & Lin, X. (2020b). Estimation of the number of spiked eigenvalues in a covariance matrix by bulk eigenvalue matching analysis. *arXiv preprint arXiv:2006.00436*.

- Ke, Z. T. & Yang, F. (2017). Covariate assisted variable ranking. *arXiv preprint arXiv:1705.10370*.
- Knowles, A. & Yin, J. (2017). Anisotropic local laws for random matrices. *Probability Theory and Related Fields*, 169(1-2), 257–352.
- Kritchman, S. & Nadler, B. (2009). Non-parametric detection of the number of signals: Hypothesis testing and random matrix theory. *IEEE Transactions on Signal Processing*, 57(10), 3930–3941.
- Kwak, J., Lee, J. O., & Park, J. (2019). Extremal eigenvalues of sample covariance matrices with general population. *arXiv preprint arXiv:1908.07444*.
- Legramanti, S., Durante, D., & Dunson, D. B. (2020). Bayesian cumulative shrinkage for infinite factorizations. *Biometrika*, 107(3), 745–752.
- Liu, J. & Rigollet, P. (2019). Power analysis of knockoff filters for correlated designs. In *Advances in Neural Information Processing Systems* (pp. 15420–15429).
- Liu, J. S. (2008). *Monte Carlo strategies in scientific computing*. Springer Science & Business Media.
- Liu, J. S. & Sabatti, C. (2000). Generalised gibbs sampler and multigrid monte carlo for bayesian computation. *Biometrika*, 87(2), 353–369.
- Liu, J. S. & Wu, Y. N. (1999). Parameter expansion for data augmentation. *Publications of the American Statistical Association*, 94(448), 1264–1274.
- Ma, Y. & Liu, J. S. (2020). On posterior consistency of bayesian factor models in high dimensions. *arXiv preprint arXiv:2006.01055*.
- Marcenko, V. A. & Pastur, L. A. (1967). Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, 1(4), 457–483.
- Meckes, E. (2014). Concentration of measure and the compact classical matrix groups.
- Natarajan, R. & McCulloch, C. E. (1998). Gibbs sampling with diffuse proper priors: A valid approach to data-driven inference? *Journal of Computational and Graphical Statistics*, 7(3), 267–277.
- Onatski, A. (2009). Testing hypotheses about the number of factors in large factor models. *Econometrica*, 77(5), 1447–1479.
- Onatski, A. (2010). Determining the number of factors from empirical distribution of eigenvalues. *The Review of Economics and Statistics*, 92(4), 1004–1016.
- Passemier, D. & Yao, J. (2014). Estimation of the number of spikes, possibly equal, in the high-dimensional case. *Journal of Multivariate Analysis*, 127, 173–183.
- Pati, D., Bhattacharya, A., Pillai, N. S., Dunson, D., et al. (2014). Posterior contraction in sparse bayesian factor models for massive covariance matrices. *The Annals of Statistics*, 42(3), 1102–1130.

- Patterson, N., Price, A. L., & Reich, D. (2006). Population structure and eigenanalysis. *PLoS Genetics*, 2(12), e190.
- Paul, D. (2007). Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statistica Sinica*, 17(4), 1617–1642.
- Pillai, N. S. & Yin, J. (2014). Universality of covariance matrices. *The Annals of Applied Probability*, 24(3), 935–1001.
- Ročková, V. & George, E. I. (2016). Fast bayesian factor analysis via automatic rotations to sparsity. *Journal of the American Statistical Association*, 111(516), 1608–1622.
- Shabalin, A. A. & Nobel, A. B. (2013). Reconstruction of a low-rank matrix in the presence of gaussian noise. *Journal of Multivariate Analysis*, 118, 67–76.
- Silverstein, J. W. (2009). The stieltjes transform and its role in eigenvalue behavior of large dimensional random matrices. *Random Matrix Theory and Its Applications. Lect. Notes Ser. Inst. Math. Sci. Natl. Univ. Singap*, 18, 1–25.
- Su, W., Bogdan, M., & Candes, E. (2017). False discoveries occur early on the lasso path. *The Annals of statistics*, 45(5), 2133–2150.
- Tanner, M. A. & Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American statistical Association*, 82(398), 528–540.
- Uhlig, H. (1994). On singular wishart and singular multivariate beta distributions. *The Annals of Statistics*, 22, 395–405.
- Wax, M. & Kailath, T. (1985). Detection of signals by information theoretic criteria. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 33(2), 387–392.
- Weinstein, A., Barber, R., & Candes, E. (2017). A power and prediction analysis for knockoffs with lasso statistics. *arXiv preprint arXiv:1712.06465*.
- Weinstein, A., Su, W. J., Bogdan, M., Barber, R. F., & Candès, E. J. (2020). A power analysis for knockoffs with the lasso coefficient-difference statistic. *arXiv preprint arXiv:2007.15346*.
- Xing, X., Zhao, Z., & Liu, J. S. (2019). Controlling false discovery rate using gaussian mirrors. *arXiv preprint arXiv:1911.09761*.
- Zahn, J. M., Poosala, S., Owen, A. B., Ingram, D. K., Lustig, A., Carter, A., Weeraratna, A. T., Taub, D. D., Gorospe, M., Mazan-Mamczarz, K., et al. (2007). Agemap: a gene expression database for aging in mice. *PLoS genetics*, 3(11), e201.