# Model-Based Penalized Regression

Maryclare Griffin

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2018

Reading Committee:

Peter Hoff, Chair

Daniela Witten

Mathias Drton

Program Authorized to Offer Degree:
Statistics

University of Washington

**Abstract**

Model-Based Penalized Regression

Maryclare Griffin

Chair of the Supervisory Committee:
Professor Peter Hoff
Duke University, Department of Statistical Science

This thesis contains three chapters that consider penalized regression from a model-based perspective, interpreting penalties as assumed prior distributions for unknown regression coefficients. In the first chapter, we show that treating an $\ell_1$ penalty as a prior can facilitate the choice of tuning parameters when standard methods for choosing the tuning parameters are not available, and when it is necessary to choose multiple tuning parameters simultaneously. In the second chapter, we consider a possible drawback of treating penalties as models, specifically possible misspecification. We introduce an easy-to-compute moment-based misspecification test for the Laplace prior, argue that the risk of misspecification calls for consideration of a larger class of $\ell_q$ penalties and corresponding prior distributions, and define easy-to-compute moment-based unknown prior parameters that yield improved estimation of the unknown regression coefficients in simulations. In the third chapter, we introduce structured shrinkage priors for dependent regression coefficients which generalize popular independent shrinkage priors. These can be useful in various applied settings where many regression coefficients are not only expected to be nearly or exactly equal to zero, but also structured.

# TABLE OF CONTENTS

# LIST OF FIGURES

# ACKNOWLEDGMENTS

To Peter, for teaching me to be creative in my research. Thank you for giving me the freedom to explore my various ideas, both good and bad, and for believing that I might eventually have some good ones. I am very grateful to you for pointing me to the problems that led to this thesis and for paying for me to move to Duke so I could continue to drop into your office with questions and work through problems on the whiteboard.

To my committee members. Thank you to Mathias for encouraging me to apply to the University of Washington to study statistics and for giving me advice on most of the major career decisions I've made since 2012, Daniela for teaching a phenomenal course on how to be an academic statistician and for offering pep talks as I navigated the publishing process and Betz for introducing me to the fascinating world of statistical methods for infectious disease research and for continuing to be supportive even when my research interests took me in another direction.

To my collaborators on my first statistical publication. Thank you to Elena for inviting me to participate in a research project before I even arrived on campus and guiding it to completion, Krista for having been and continuing to be a consistently wonderful mentor and Mark for frequently providing positive feedback and encouraging guidance.

To the many fantastic teachers I have had over the years. Thank you to Mrs. Donaghue for teaching advanced math classes in middle school, Mr. Garr for introducing me to some of my favorite authors, Mrs. Petti and Mrs. McFarland for making me a better writer, Mrs. Leonard for going above and beyond as math teacher and Professor Heckman for valuing my work as a research assistant and continuing to encourage me long afterwards.

To my fiancé, for supporting me, especially when I decided it would be best for me to

move across the country to finish this degree. Thank you for picking up the phone every time I would call, coming to visit Durham for weeks at a time and not breaking up with me when I spontaneously adopted a second troubled dog.

To my loving and perfectly weird family. To my dad for raising me to believe I could do anything, my mom for teaching me to be curious and inventive, my sister Norah for keeping me humble by mocking me mercilessly, Grandma Griffin for her care and praise, Grandma Carney for her wit and energy and Grandpa Carney for his kindness and generosity.

To my friends, in no particular order! To Kelly for always being an eager partner in crime, James and Evelyn for always knowing where to go dancing, Kamala for working beside me all through college, Cecilia for being an exceptional letter writer, Jack, Ricky and Robby for making analysis more bearable, Andy for being my most reliable ally in the battle for base `R` graphics, Lily for introducing me to my favorite pens, Cathy, Paul, Evan, Becky and Bobby for listening to me panic about math tests between dives, Rebecca for collaborating with me on a spreadsheet of inspiring lady statisticians, Karsten for dragging me up Mailbox Peak, Kyle for educating me on the merits of Taco Bell, Alden for putting up with Baxter in the office and on several alpine lake hikes, Sam for being the best co-GSR, Jessica for keeping me up to date on internet slang, Theresa, Leigh, Laina and Bailey for giving me advice and being great role models, Hannah, Corinne and Nilanjana for letting me give them advice and becoming good friends, Matt, Ashley and Angélica for making Durham feel like home, Felipé, Mauricio, Andee, Carrie and Victor for always making time for lunch, Mike, Liz, Lindsay, Jake, Abbas, Steph, Jody and Phil for letting me have a desk in your social office.

To future young statisticians. I read several theses as I prepared to write my own. In the event that you come across mine, I'd like you to know that I never felt like writing this material came naturally. Everything presented here was nearly thrown away on several occasions and I often wanted to quit. The first time I started to feel confident in and proud of my work was only several months ago, and I still only feel that way part of the time.

# DEDICATION

To my dogs. To Baxter, for being by my side from the beginning, and to Hope, for trying to make me into a morning person.

# Chapter 1
# INTRODUCTION

The material in this thesis deals with penalized regression and its Bayesian counterpart, regression under a prior on the regression coefficients. The relationship between penalized regression and Bayesian regression has long been recognized (Tibshirani, 1996). Let $\boldsymbol{y}$ be an observed $n \times 1$ response, let $\boldsymbol{X}$ be an observed $n \times p$ matrix of covariates and let $\boldsymbol{\beta}$ be a $p \times 1$ vector of unknown regression coefficients. In the linear regression setting, Lasso penalized regression solves

$$\text{minimize}_{\boldsymbol{\beta}} \left\{ ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2 + \lambda \, ||\boldsymbol{\beta}||_1 \right\}.$$

This is equivalent to computing the posterior mode of $\boldsymbol{\beta}$ under the model

$$\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\beta} \sim \text{normal}\left(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}_n\right), \; \beta_j \sim \text{Laplace}\left(0, 2\sigma^2/\lambda\right),$$

where $\boldsymbol{I}_n$ is an $n \times n$ identity matrix and $\sigma^2 > 0$ and $\lambda > 0$ are scale parameters. More generally, let $h\left(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{\phi}\right)$ be an objective function depending on nuisance parameters $\boldsymbol{\phi}$ that relates the unknown regression coefficients $\boldsymbol{\beta}$ and covariates $\boldsymbol{X}$ to the observed response $\boldsymbol{y}$, and let $\boldsymbol{f}\left(\boldsymbol{\beta}|\boldsymbol{\theta}\right)$ be a penalty function that depends on tuning parameters $\boldsymbol{\theta}$. Any penalized regression estimate obtained by solving

$$\text{minimize}_{\boldsymbol{\beta}} \left\{ h\left(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{y}, \boldsymbol{\phi}\right) + f\left(\boldsymbol{\beta}|\boldsymbol{\theta}\right) \right\}$$

corresponds to the posterior mode of $\boldsymbol{\beta}$ under the model

$$p_{\boldsymbol{y}|\boldsymbol{\beta}}\left(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{\phi}\right) \propto \exp\left\{-h\left(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{y}, \boldsymbol{\phi}\right)\right\}, \; p_{\boldsymbol{\beta}}\left(\boldsymbol{\beta}|\boldsymbol{\theta}\right) \propto \exp\left\{-f\left(\boldsymbol{\beta}|\boldsymbol{\theta}\right)\right\},$$

where $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ can be interpreted as distributional parameters for the distributions of the data and the coefficients.

Despite the fact that this correspondence between penalties and priors is well established, the penalized regression and Bayesian literatures remain fairly divided. Within the penalized regression literature, penalties are rarely treated as implying a specific model for the data and regression coefficients. Rather, the penalty is interpreted as a tool for producing estimates of regression coefficients with certain desirable properties, e.g. sparsity, and nuisance and tuning parameters tend to be chosen to optimize predictive performance. This approach can be advantageous in that it is relatively simple to implement, outperforms many alternatives in terms of prediction and scales well to accommodate massive data. Within the Bayesian literature, researchers either take an empirical Bayes approach, estimating $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ from the marginal likelihood of the data, or a fully Bayes approach, assuming prior distributions for $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$. Fully Bayes approaches are more popular, because they can be used to account for uncertainty in the estimation of $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ and because estimating $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ the marginal likelihood of the data as a function of $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ can be prohibitively computationally demanding under many common penalized regression models. Letting $p_{\boldsymbol{\phi}}(\boldsymbol{\phi})$ and $p_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ refer to the prior densities of $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$, this yields the model

$$p_{\boldsymbol{y}|\boldsymbol{\beta}}(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{\beta}) \propto \int_{-\infty}^{\infty} \exp\left\{-h\left(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{\beta},\boldsymbol{y},\boldsymbol{\phi}\right)\right\} p_{\boldsymbol{\phi}}(\boldsymbol{\phi})\, d\boldsymbol{\phi},$$
$$p_{\boldsymbol{\beta}}(\boldsymbol{\beta}|\boldsymbol{\theta}) \propto \int_{-\infty}^{\infty} \exp\left\{-f\left(\boldsymbol{\beta}|\boldsymbol{\theta}\right)\right\} p_{\boldsymbol{\theta}}(\boldsymbol{\theta})\, d\boldsymbol{\theta}.$$

Under this model, the posterior mode can be more computationally challenging to compute because the correspondingly penalty is a function of an intractable integral. Accordingly, researchers in the Bayesian literature tend to use posterior means or medians as estimators of $\boldsymbol{\beta}$, which are easier to compute but do not have the same desirable properties of the posterior mode for fixed values of $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$, e.g. elementwise sparsity. Even when $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ are treated as fixed and empirical Bayes approaches are used, the Bayesian literature still tends to discourage use of the posterior mode as an estimator of $\boldsymbol{\beta}$ and encourage use of the posterior mean or median instead (Park and Casella, 2008; Hans, 2009). This is motivated by decision-theoretic concerns, as the posterior mean and median are known to minimize squared and absolute error loss, respectively. However it still results in estimators of $\boldsymbol{\beta}$ that

may not have the same desirable properties as the posterior mode, e.g. sparsity.

To our knowledge relatively little research has occupied the middle ground, i.e. has interpreted penalties as priors without assuming prior distributions for unknown distributional parameters or discarding the posterior mode as a viable estimator. Doing so affords us the ability to *estimate* tuning parameters/unknown distributional parameters, which can be difficult to specify when they are numerous, data are limited or inference on the unknown regression parameters as opposed to prediction is the primary goal. Furthermore, treating the use of a penalty as equivalent to assuming a specific prior distribution allows us to draw on a vast decision theory literature that can be used to justify the use of various posterior summaries as estimators depending on the inferential goals of a specific problem (Bernardo and Smith, 2000; Hans, 2009). This thesis helps to fill this gap in the literature. Although we will use Bayesian terminology throughout this thesis, the ideas discussed here are not strictly Bayesian in scope. Rather, they relate to the long literature on smoothing and mixed models, in which regression coefficients are modeled to encourage various desirable properties, distributional parameters are estimated from the marginal likelihood of the data, and posterior summaries are used as estimators (Wand, 2003).

The first chapter in this thesis, "Lasso ANOVA Decompositions for Matrix and Tensor Data," focuses on the use of Lasso penalized ANOVA for matrices and higher order arrays. In the matrix case, we consider Lasso penalized linear regression for $\boldsymbol{y} = \text{vec}\,(\boldsymbol{Y})$, where $\boldsymbol{Y}$ is a $p_1 \times p_2$ matrix and $\boldsymbol{X} = \left[ \begin{array}{cccc} \mathbf{1}_{p_1 p_2} & \mathbf{1}_{p_2} \otimes \boldsymbol{I}_{p_1} & \boldsymbol{I}_{p_2} \otimes \mathbf{1}_{p_1} & \boldsymbol{I}_{p_1 p_2} \end{array} \right]$, where $\mathbf{1}_{p_1 p_2}$ is a $p_1 p_2 \times 1$ vector of ones. The unknown regression coefficients, $\boldsymbol{\beta}$, can be decomposed into a grand mean, row, column and element-wise components $\boldsymbol{\beta} = \left[ \begin{array}{cccc} \mu & \boldsymbol{a}' & \boldsymbol{b}' & \text{vec}\,(\boldsymbol{C})' \end{array} \right]'$, where only elements of $\boldsymbol{C}$ are subjected to a Lasso penalty. This problem poses the following challenge: every element of $\boldsymbol{Y}$ corresponds to a unique element of $\boldsymbol{C}$, so cross validation is actually infeasible. We tackle this problem by treating the Lasso penalty as a Laplace prior and recognizing that differences in normal and Laplace tail behavior, as measured by kurtosis, allow for the construction of moment-based estimators of the noise variance and Lasso tuning parameter. In the process, we explore how treating the Lasso penalty as a

prior is uniquely vulnerable to model misspecification, specifically of the prior distribution of the regression coefficients. We observe evidence that suggests a Laplace prior will tend to *underpenalize* the unknown regression coefficients when the empirical distribution of the unobserved true regression coefficients is heavier than Laplace tailed and will *overpenalize* the unknown regression coefficients when the empirical distribution of the unobserved true regression coefficients is lighter than Laplace tailed.

Motivated by the issue of prior specification, the second chapter in this thesis "Testing Sparsity-Inducing Penalties," explores the choice of prior for more general linear regression problems. We introduce a moment-based test of whether or not the tail behavior of ordinary least squares or ridge estimates of the unknown regression coefficients is consistent with what would be observed if the true unobserved regression coefficients were Laplace distributed, as well as moment-based estimators of the prior parameters of a broader class of priors which corresponds to $\ell_q$ penalties. We demonstrate that implementing such a test and adaptively specifying the prior in the event of rejection can improve estimation of the regression coefficients, resolving the over- and underpenalization problems associated with treating the Lasso penalty as a Laplace prior distribution.

The third chapter in this thesis, "Structured Shrinkage Priors," considers an additional kind of misspecification. In the first two chapters, we considered *independent* prior distributions for the regression coefficients, which correspond to *separable* penalties which can be decomposed. These priors do not encourage *structure* among the regression coefficients. However, in many high dimensional regression settings the regression coefficients may have some known structure a priori, e.g. the regression coefficients may be ordered in space or correspond to a vectorized matrix or tensor of regression coefficients. Accordingly, in this last chapter we develop structured shrinkage priors that generalize multivariate normal, Laplace, exponential power and normal-gamma priors. These priors allow for the regression coefficients to be correlated a priori without sacrificing sparsity and shrinkage. The primary challenges in working with these structured shrinkage priors are computational, as the corresponding penalty is an intractable $p$-dimensional integral and the full conditional

distributions that are needed to simulate from the full conditional distribution of the regression coefficients are not necessarily standard distributions. We overcome these issues using a flexible elliptical slice sampling procedure, and demonstrate that these priors can be used to introduce structure while preserving sparsity of the corresponding penalized estimate given by the posterior mode.

# Chapter 2

# LASSO ANOVA DECOMPOSITIONS FOR MATRIX AND TENSOR DATA

## *2.1 Introduction*

Researchers are often interested in estimating the entries of an unknown $n \times p$ mean matrix $\boldsymbol{M}$ given a single noisy realization, $\boldsymbol{Y} = \boldsymbol{M} + \boldsymbol{Z}$, where the entries of $\boldsymbol{Z}$ are assumed to be independent, identically distributed mean zero normal random variables with unknown variance $\sigma_z^2$. Consider a noisy matrix $\boldsymbol{Y}$ of gene expression measurements for $p$ different genes and $n$ tumors. Researchers may be interested in which tumors have unique gene expression profiles and which genes are differentially expressed across different tumors.

This is challenging because no replicates are observed; each unknown $m_{ij}$ corresponds to a single observation $y_{ij}$. As a result, the maximum likelihood estimate $\boldsymbol{Y}$ has high variability. Accordingly, simplifying assumptions that reduce the dimensionality of $\boldsymbol{M}$ are often made. Many such assumptions relate to a two-way ANOVA decomposition of $\boldsymbol{M}$:

$$\boldsymbol{M} = \mu \mathbf{1}_n \mathbf{1}_p' + \boldsymbol{a} \mathbf{1}_p' + \mathbf{1}_n \boldsymbol{b}' + \boldsymbol{C}, \tag{2.1}$$

where $\mu$ is an unknown grand mean, $\boldsymbol{a}$ is an $n \times 1$ vector of unknown row effects, $\boldsymbol{b}$ is a $p \times 1$ vector of unknown column effects, $\boldsymbol{C}$ is a matrix of elementwise "interaction" effects and $\mathbf{1}_n$ and $\mathbf{1}_p$ are $n \times 1$ and $p \times 1$ vectors of ones, respectively. In the absence of replicates, implicitly assuming $\boldsymbol{C} = \mathbf{0}$ is common. This reduces the number of freely varying unknown parameters, from $np$ to $n + p$, but is also unlikely to be appropriate in practice.

In the settings we consider, it is likely that at least some elements of $\boldsymbol{C}$ are nonzero, e.g. some tumor-gene combinations or treatment pairs in a factorial design may have large interaction effects that are not explained by a strictly additive model. If $\boldsymbol{M}$ is approximately

additive in the sense that large deviations from additivity are rare, then $\boldsymbol{C}$ is sparse and estimation of $\boldsymbol{M}$ may be improved by penalizing elements of $\boldsymbol{C}$:

$$\text{minimize}_{\mu,\boldsymbol{a},\boldsymbol{b},\boldsymbol{C}} \left\{ \frac{1}{2\sigma_z^2} \left|\left|\text{vec}\left\{\boldsymbol{Y} - \left(\mu \mathbf{1}_n \mathbf{1}_p' + \boldsymbol{a}\mathbf{1}_p' + \mathbf{1}_n \boldsymbol{b}' + \boldsymbol{C}\right)\right\}\right|\right|_2^2 + \lambda_c \left|\left|\text{vec}\left(\boldsymbol{C}\right)\right|\right|_1 \right\}. \quad (2.2)$$

The $\ell_1$ penalty induces sparsity among the estimated entries of $\boldsymbol{C}$, and solving this penalized regression problem yields unique estimates of $\boldsymbol{M}$ and $\boldsymbol{C}$. Elements of $\boldsymbol{C}$ can be interpreted as interactions insofar as they indicate deviation from a strictly additive model.

This is a departure from the literature, in which assumptions on the rank of $\boldsymbol{C}$ are often made. Some set $\mu = 0$, $\boldsymbol{a} = \mathbf{0}$ and $\boldsymbol{b} = \mathbf{0}$ and assume $\boldsymbol{C}$ is low rank (Fazel, 2002; Hoff, 2007; Candes et al., 2013; Josse and Sardy, 2016), while others treat $\mu$, $\boldsymbol{a}$ and $\boldsymbol{b}$ as unknown, apply the standard ANOVA zero-sum constraints for identifiability, $\mathbf{1}_n'\boldsymbol{a} = \mathbf{1}_p'\boldsymbol{b} = 0$ and $\boldsymbol{C}'\mathbf{1}_n = \mathbf{0}$ and $\boldsymbol{C}\mathbf{1}_p = \mathbf{0}$, and assume that $\boldsymbol{C}$ is low rank (Gollob, 1968; Johnson and Graybill, 1972; Mandel, 1971; Goodman and Haberman, 1990). We refer to the latter as an additive-plus-low-rank model for $\boldsymbol{M}$. Assuming that $\boldsymbol{C}$ is low rank implies that elements of $\boldsymbol{C}$ are multiplicative in row and column factors, i.e. if $\boldsymbol{C}$ is rank $R$ then $c_{ij} = \sum_{r=1}^{R} \lambda_r u_{r,i} v_{r,j}$.

Although useful in many settings, assuming low rank $\boldsymbol{C}$ has two main limitations. First, in the presence of unknown noise variance, $\sigma_z^2$, existing methods for choosing the rank can be computationally expensive for large matrices (Hoff, 2007), require strong assumptions such as known $\sigma_z^2$ (Candes et al., 2013), or rely on approximations to account for unknown $\sigma_z^2$ that may not always perform well in practice (Josse and Sardy, 2016). Second, even when the rank can be chosen well, these methods conflate the presence of elementwise effects of scientific interest with the presence of multiplicative effects. While this may be plausible in many settings, it is easy to imagine scenarios in which a low rank $\boldsymbol{C}$ may fail to capture elementwise effects of scientific interest. For instance, if $\boldsymbol{Y}$ were an $n \times n$ square matrix and all $c_{ii}$ were large while all $c_{ij}$, $i \neq j$ were equal to zero, a rank $n$ estimate of $\boldsymbol{C}$ would be needed.

That said, solving Equation (2.2) requires specification of $\lambda_c$ and $\sigma_z^2$. One approach might be to recognize that we can rewrite Equation (2.2) to depend on a single parameter

$\eta = \lambda_c \sigma_z^2$ and specify $\eta$ using cross-validation. However, cross-validation is not appropriate for this problem. Consider a subset $\boldsymbol{Y}_1$ of $\boldsymbol{Y}$ with at least two elements from each row and column of $\boldsymbol{Y}$ and let $\boldsymbol{Y}_2$ denote the remaining entries. For a fixed value of $\eta$, we can obtain estimates of $\mu$, $\boldsymbol{a}$, $\boldsymbol{b}$, and elements $\boldsymbol{C}_1$ of $\boldsymbol{C}$ corresponding to $\boldsymbol{Y}_1$. However, computing out-of-sample predictions for $\boldsymbol{Y}_2$ requires estimates of the elements $\boldsymbol{C}_2$ of $\boldsymbol{C}$ corresponding to $\boldsymbol{Y}_2$. Cross-validation cannot be performed without making additional assumptions that relate the values of $\mu$, $\boldsymbol{a}$, $\boldsymbol{b}$ and $\boldsymbol{C}_1$ to $\boldsymbol{C}_2$. Another approach involves viewing the $\ell_1$ penalty on $\boldsymbol{C}$ as a Laplace prior distribution, in which case $\lambda_c$ and $\sigma_z^2$ can be interpreted as nuisance parameters. However, obtaining empirical Bayes estimates of $\lambda_c$ and $\sigma_z^2$ via maximum marginal likelihood may be prohibitively computationally demanding (Figueiredo, 2003; Park and Casella, 2008).

In this chapter we present moment-based empirical Bayes estimators of the nuisance parameters $\lambda_c$ and $\sigma_z^2$ that are easy to compute, consistent and independent of assumptions made regarding $\boldsymbol{a}$ and $\boldsymbol{b}$. Moment-based estimators can be sensitive to outliers, however we consider settings in which $n$ or $p$ are very large and moments can be estimated well. As our approach to estimating $\lambda_c$ and $\sigma_z^2$ uses the Laplace prior interpretation of the $\ell_1$ penalty, we refer to estimation of $\boldsymbol{M}$ via optimization of Equation (2.2) using these nuisance parameter estimators as LANOVA penalization, and we refer to the estimate $\widehat{\boldsymbol{M}}$ as the LANOVA estimate.

The chapter proceeds as follows: In Section 2.2, we introduce moment-based estimators for $\lambda_c$ and $\sigma_z^2$, show that they are consistent as *either* the number of rows *or* columns of $\boldsymbol{Y}$ goes to infinity and show that their efficiency is comparable to that of asymptotically efficient marginal maximum likelihood estimators (MMLEs). In Section 2.3, we discuss estimation of $\boldsymbol{M}$ via Equation (2.2) given estimates of $\lambda_c$ and $\sigma_z^2$ and introduce a test of whether or not elements of $\boldsymbol{C}$ are heavy-tailed, which allows us to avoid LANOVA penalization in settings where it is especially inappropriate. We also investigate the performance of LANOVA estimates of $\boldsymbol{M}$ relative to strictly additive estimates, strictly non-additive estimates and additive-plus-low-rank estimates and examine robustness to misspecification of the distribution of elements of $\boldsymbol{C}$. In Section 2.4, we extend LANOVA penalization to

include penalization of lower-order mean parameters $\boldsymbol{a}$ and $\boldsymbol{b}$ and also to apply to the case where $\boldsymbol{Y}$ is a $K$-way tensor. In Section 2.5, we apply LANOVA penalization to a matrix of gene expression measurements, a three-way tensor of fMRI data and a three-way tensor of wheat infection data. In Section 2.6 we discuss extensions, specifically multilinear regression models and opportunities that arise in the presence of replicates.

### *2.2 LANOVA Nuisance Parameter Estimation*

Consider the following statistical model for deviations of $\boldsymbol{Y}$ from a strictly additive model:

$$\begin{aligned} &\boldsymbol{Y} = \mu \mathbf{1}_n \mathbf{1}_p' + \boldsymbol{a}\mathbf{1}_p' + \mathbf{1}_n \boldsymbol{b}' + \boldsymbol{C} + \boldsymbol{Z} \\ &\boldsymbol{C} = \{c_{ij}\} \sim \text{i.i.d. Laplace}\left(0, \lambda_c^{-1}\right), \quad \boldsymbol{Z} = \{z_{ij}\} \sim \text{i.i.d. } N\left(0, \sigma_z^2\right). \end{aligned} \tag{2.3}$$

The posterior mode of $\mu$, $\boldsymbol{a}$, $\boldsymbol{b}$ and $\boldsymbol{C}$ under this Laplace prior for $\boldsymbol{C}$ and flat priors for $\mu$, $\boldsymbol{a}$ and $\boldsymbol{b}$ corresponds to the solution of the LANOVA penalization problem given by Equation (2.2).

We construct estimators of $\lambda_c$ and $\sigma_z^2$ as follows. Letting $\boldsymbol{H}_k = \boldsymbol{I}_k - \mathbf{1}_k\mathbf{1}_k/k$ be the $k \times k$ centering matrix, we define $\boldsymbol{R} = \boldsymbol{H}_n \boldsymbol{Y} \boldsymbol{H}_p$. $\boldsymbol{R}$ depends on $\boldsymbol{C}$ and $\boldsymbol{Z}$ alone, specifically $\boldsymbol{R} = \boldsymbol{H}_n (\boldsymbol{C} + \boldsymbol{Z}) \boldsymbol{H}_p$. We construct estimators of $\lambda_c$ and $\sigma_z^2$ from $\boldsymbol{R}$ by leveraging the difference between Laplace and normal tail behavior as measured by fourth order moments. The fourth order central moment of any random variable $x$ with mean $\mu_x$ and variance $\sigma_x^2$ can be expressed as $\mathbb{E}\left[(x-\mu_x)^4\right] = (\kappa + 3)\,\sigma_x^4$, where $\kappa$ is interpreted as the *excess kurtosis* of the distribution of $x$ relative to a normal distribution. A normally distributed variable has excess kurtosis equal to 0, whereas a Laplace distributed random variable has excess kurtosis equal to 3. It follows that the second and fourth order central moments of elements of $\boldsymbol{C}+\boldsymbol{Z}$ are $\mathbb{E}\left[(c_{ij}+z_{ij})^2\right] = \sigma_c^2 + \sigma_z^2$ and $\mathbb{E}\left[(c_{ij}+z_{ij})^4\right] = 3\sigma_c^4 + 3\left(\sigma_c^2 + \sigma_z^2\right)^2$, respectively, where $\sigma_c^2 = 2/\lambda_c^2$ is the variance of a Laplace$(\lambda_c^{-1})$ random variable. Given values of $\mathbb{E}\left[(c_{ij}+z_{ij})^2\right]$ and $\mathbb{E}\left[(c_{ij}+z_{ij})^4\right]$, we see that $\sigma_c^2$ and $\sigma_z^2$, and accordingly $\lambda_c$, can easily be recovered.

We do not observe $\boldsymbol{C} + \boldsymbol{Z}$ directly, but we can use the second and fourth order sample moments of $\boldsymbol{R}$, an estimate of $\boldsymbol{C} + \boldsymbol{Z}$, given by $\overline{r}^{(2)} = \frac{1}{np}\sum_{i=1}^{n}\sum_{j=1}^{p} r_{ij}^2$ and $\overline{r}^{(4)} =$

$\frac{1}{np}\sum_{i=1}^{n}\sum_{j=1}^{p} r_{ij}^4$, respectively, to separately estimate $\sigma_c^2$ and $\sigma_z^2$. These estimators are:

$$\begin{aligned}\widehat{\sigma}_c^4 =& \left\{\frac{n^3p^3}{(n-1)(n^2-3n+3)(p-1)(p^2-3p+3)}\right\}\left\{\overline{r}^{(4)}/3-\left(\overline{r}^{(2)}\right)^2\right\} \qquad (2.4)\\ \widehat{\sigma}_c^2 =& \sqrt{\widehat{\sigma}_c^4}, \quad \widehat{\sigma}_z^2 = \left\{\frac{np}{(n-1)(p-1)}\right\}\overline{r}^{(2)} - \widehat{\sigma}_c^2.\end{aligned}$$

An estimator of $\lambda_c$ is then given by $\widehat{\lambda}_c = \sqrt{2/\widehat{\sigma}_c^2}$.

The estimator $\widehat{\sigma}_c^4$ is biased. It is possible to obtain an unbiased estimator for $\sigma_c^4$, however the unbiased estimator will not be consistent as $n \to \infty$ with $p$ fixed or $p \to \infty$ with $n$ fixed. Because these estimators depend on higher-order terms which can be very sensitive to outliers, it is desirable to have consistency as either the number of rows or columns grows. Accordingly, we prefer the biased estimator and examine its bias in the following proposition.

**Proposition 2.2.1** *Under the model given by Equation* (2.3),

$$\begin{aligned}\mathbb{E}\left[\widehat{\sigma}_c^4\right] - \sigma_c^4 =& -\left\{\frac{n^3p^3}{(n-1)(n^2-3n+3)(p-1)(p^2-3p+3)}\right\}\left[\left\{\frac{3(n-1)^2(p-1)^2}{n^3p^3}\right\}\sigma_c^4 + \right.\\ & \left.\left\{\frac{2(n-1)(p-1)}{n^2p^2}\right\}\left(\sigma_c^2+\sigma_z^2\right)^2\right].\end{aligned}$$

A proof of this proposition and all other results presented in this chapter are given in an appendix. The bias is always negative and accordingly, yields overpenalization of $\boldsymbol{C}$. When both $n$ and $p$ are small, $\widehat{\sigma}_c^4$ tends to underestimate $\sigma_c^4$. Recalling that $\sigma_c^4$ is inversely related to $\lambda_c$, this reflects a tendency to overpenalize and accordingly overshrink elements of $\boldsymbol{C}$ when both $n$ and $p$ are small. This is not undesirable, in that it reflects a tendency to prefer the simple additive model when few data are available. We also observe that the bias depends not only on both $\sigma_c^2$ and $\sigma_z^2$. Holding $n$, $p$ and $\sigma_c^2$ fixed, we will overestimate $\lambda_c$ more when $\sigma_z^2$ is larger. Again, this is not undesirable, in that it reflects a tendency to prefer the simple additive model when the data are very noisy. Last, we see that the bias is $O(1/np)$, i.e. the bias approaches zero as *either* the number of rows *or* the number of columns increases. The large sample behavior of our estimators of the nuisance parameters is similar.

**Proposition 2.2.2** *Under the model given by Equation* (2.3), $\widehat{\sigma}_c^4 \xrightarrow{p} \sigma_c^4$, $\widehat{\sigma}_c^2 \xrightarrow{p} \sigma_c^2$, $\widehat{\lambda}_c \xrightarrow{p} \lambda_c$ *and* $\widehat{\sigma}_z^2 \xrightarrow{p} \sigma_z^2$ *as* $n \to \infty$ *with* $p$ *fixed,* $p \to \infty$ *with* $n$ *fixed, or* $n, p \to \infty$.

Although these nuisance parameter estimators are easy to compute and consistent as $n \to \infty$ or $p \to \infty$, they are not maximum likelihood estimators and may not be asymptotically efficient even as $n \to \infty$ and $p \to \infty$. Accordingly, we compare the asymptotic efficiency of our estimator $\widehat{\sigma}_c^2$ to that of the corresponding asymptotically efficient marginal maximum likelihood estimator (MMLE) denoted by $\widetilde{\sigma}_c^2$ as $n$ and $p \to \infty$. As noted in the Introduction, obtaining $\widetilde{\sigma}_c^2$ is computationally demanding because maximizing the marginal likelihood of the data requires a Gibbs-within-EM algorithm that is slow to converge (Park and Casella, 2008). Fortunately, computing the asymptotic variance of $\widetilde{\sigma}_c^2$ is simpler than computing $\widetilde{\sigma}_c^2$ itself. The asymptotic variance of $\widetilde{\sigma}_c^2$ is given by the Cramér-Rao lower bound for $\sigma_c^2$, which can be computed numerically from the density of the sum of Laplace and normally distributed variables (Nadarajah, 2006; Díaz-Francés and Montoya, 2008). The asymptotic variance of $\widehat{\sigma}_c^2$ is straightforward to compute as $\sqrt{np}\,(\widehat{\sigma}_c^4 - \sigma_c^4)$ converges in distribution to a moment estimator of $\sigma_c^4$. We note that the asymptotic variance of $\widehat{\lambda}_c$ is similarly straightforward to compute; both asymptotic variances are given in the appendix. Letting $\mathbb{V}\left[\widetilde{\sigma}_c^2\right]$ and $\mathbb{V}\left[\widehat{\sigma}_c^2\right]$ refer to the variances of the estimators $\widetilde{\sigma}_c^2$ and $\widehat{\sigma}_c^2$, we plot the asymptotic relative efficiency $\mathbb{V}\left[\widetilde{\sigma}_c^2\right] / \mathbb{V}\left[\widehat{\sigma}_c^2\right]$ over values of $\sigma_c^2, \sigma_z^2 \in [0, 1]$ in Figure 2.1. Note that the relative efficiency of $\widehat{\sigma}_c^2$ compared to $\widetilde{\sigma}_c^2$ also reflects the relative efficiency of our estimators $\widehat{\lambda}_c$ and $\widehat{\sigma}_z^2$ compared to the MMLEs $\widetilde{\lambda}_c$ and $\widetilde{\sigma}_c^2$, respectively, because both are simple functions of $\widehat{\sigma}_c^2$.

When $\sigma_c^2$ is small relative to $\sigma_z^2$, the MMLE $\widetilde{\sigma}_c^2$ tends to be slightly more efficient. When $\sigma_c^2$ is large relative to $\sigma_z^2$, $\widetilde{\sigma}_c^2$ tends to be much more efficient. However, in such cases the interactions will not be heavily penalized and LANOVA penalization will not tend to yield a simplified, nearly additive estimate of $\boldsymbol{M}$. Put another way, Figure 2.1 indicates that $\widehat{\lambda}_c$ and $\widehat{\sigma}_z^2$ will be nearly as efficient as the corresponding MMLEs when LANOVA penalization is useful for producing a simplified, nearly additive estimate of $\boldsymbol{M}$ with sparse interactions. We also note that because they are moment-based, our estimators may be more robust to

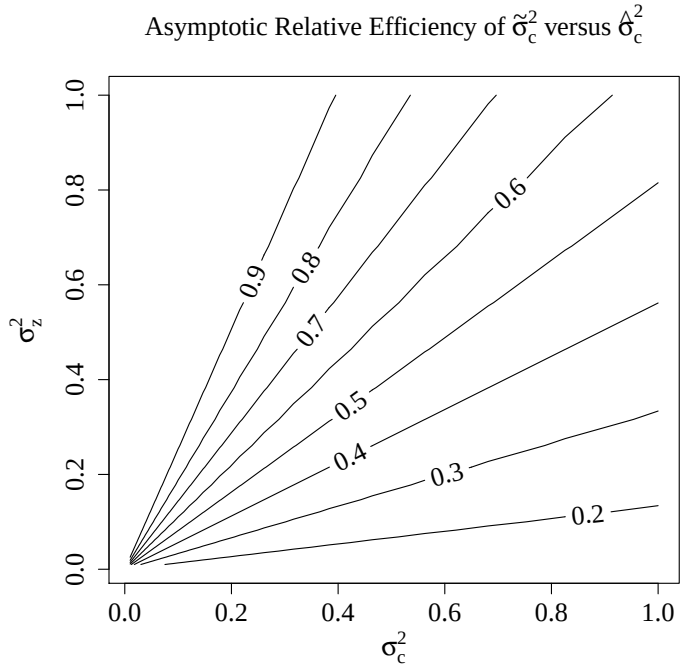



Figure 2.1: Asymptotic relative efficiency $\mathbb{V}[\widetilde{\sigma}_c^2]/\mathbb{V}[\widehat{\sigma}_c^2]$ of the MMLE $\widetilde{\sigma}_c^2$ versus our moment-based estimator $\widehat{\sigma}_c^2$ as a function of the true variances $\sigma_c^2$ and $\sigma_z^2$.

misspecification of the distribution of elements of $\boldsymbol{C}$ and $\boldsymbol{Z}$ than the MMLEs.

## *2.3 Mean Estimation, Interpretation, Model Checking and Robustness*

### *2.3.1 Mean Estimation*

In practice, our nuisance parameter estimators are not guaranteed to be nonnegative and two special cases can arise. When $\widehat{\sigma}_c^4 < 0$, we set $\widehat{\sigma}_c^2 = 0$ and $\widehat{\boldsymbol{C}} = \boldsymbol{0}$ and set $\widehat{\boldsymbol{M}} = \widehat{\boldsymbol{M}}_{ADD}$, where $\widehat{\boldsymbol{M}}_{ADD} = (\boldsymbol{I}_n - \boldsymbol{H}_n)\,\boldsymbol{Y}\,(\boldsymbol{I}_p - \boldsymbol{H}_p)$ is the strictly additive estimate. When $\widehat{\sigma}_z^2 < 0$, we reset $\widehat{\sigma}_z^2 = 0$ and set $\widehat{\boldsymbol{M}} = \widehat{\boldsymbol{M}}_{MLE}$, where $\widehat{\boldsymbol{M}}_{MLE} = \boldsymbol{Y}$ is the strictly non-additive estimate. Neither special case prohibits estimation of $\boldsymbol{M}$.

We assess how often these special cases arise via a small simulation study. Setting $\sigma_z^2 = 1$,

$n = p = 25$, $\mu = 0$, $\boldsymbol{a} = \boldsymbol{0}$ and $\boldsymbol{b} = \boldsymbol{0}$, we simulate $10,000$ realizations of $\boldsymbol{Y} = \boldsymbol{C} + \boldsymbol{Z}$ under the model given by Equation (2.3) for each value of $\sigma_c^2 \in \{1/2, 1, 3/2\}$. We obtain $\widehat{\sigma}_c^2 \leq 0$ in 13.7%, 1.64% and 0.02% of simulations for $\sigma_c^2$ equal to 1/2, 1 and 3/2, respectively. This means that when the magnitude of elements of $\boldsymbol{C}$ is smaller, we are more likely to obtain a strictly additive estimate of $\boldsymbol{M}$. We do not obtain $\widehat{\sigma}_z^2 = 0$ in any simulations.

When $\widehat{\sigma}_c^2 > 0$ and $\widehat{\sigma}_z^2 > 0$, we can obtain an estimate of $\boldsymbol{M}$ from Equation (2.2) using block coordinate descent. Setting $\widehat{\boldsymbol{C}}^0 = \boldsymbol{H}_n \boldsymbol{Y} \boldsymbol{H}_p$ and $k = 1$, our block coordinate descent algorithm iterates the following until the objective function Equation (2.2) converges:

- Set $\widehat{\mu}^k = \boldsymbol{1}_n'(\boldsymbol{Y} - \widehat{\boldsymbol{C}}^{k-1})\boldsymbol{1}_p/np$, $\widehat{\boldsymbol{a}}^k = \boldsymbol{H}_n'(\boldsymbol{Y} - \widehat{\boldsymbol{C}}^{k-1})\boldsymbol{1}_p/p$, $\widehat{\boldsymbol{b}}^k = \boldsymbol{H}_p(\boldsymbol{Y} - \widehat{\boldsymbol{C}}^{k-1})'\boldsymbol{1}_n/n$ and $\boldsymbol{R}^k = \boldsymbol{Y} - \widehat{\mu}^k \boldsymbol{1}_n \boldsymbol{1}_p' - \widehat{\boldsymbol{a}}^k \boldsymbol{1}_p' - \boldsymbol{1}_n(\widehat{\boldsymbol{b}}^k)'$;

- Set $\widehat{\boldsymbol{C}}^k = \text{sign}(\boldsymbol{R}^k)(|\boldsymbol{R}^k| - \widehat{\lambda}_c \widehat{\sigma}_z^2)_+$, where $\widehat{\lambda}_c = \sqrt{2/\widehat{\sigma}_c^2}\,\text{sign}(\cdot)$ and the soft-thresholding function $(\cdot)_+$ are applied elementwise. Set $k = k + 1$.

### *2.3.2 Interpretation*

The nonzero entries of $\widehat{\boldsymbol{C}}$ correspond to the $r$ largest residuals from fitting a strictly additive model with $\boldsymbol{C} = \boldsymbol{0}$, where $r$ is determined by $\widehat{\lambda}_c$ and $\widehat{\sigma}_z^2$. Elements of $\widehat{\boldsymbol{C}}$ can be interpreted as interactions insofar as they indicate deviation from a strictly additive model for $\boldsymbol{M}$. However, because we do impose the standard ANOVA zero-sum constraints, we cannot interpret elements of $\widehat{\boldsymbol{C}}$ directly as population average interaction effects, i.e. $\widehat{c}_{ij} \neq \mathbb{E}[y_{ij}] - \frac{1}{p}\sum_{j=1}^{p}\mathbb{E}[y_{ij}] - \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[y_{ij}] + \frac{1}{np}\sum_{i=1}^{n}\sum_{j=1}^{p}\mathbb{E}[y_{ij}]$. For the same reason, $\mu$, $\boldsymbol{a}$ and $\boldsymbol{b}$ cannot be interpreted as the grand mean and population average main effects. To obtain estimates that have the standard population average interpretation, we recommend performing a two-way ANOVA decomposition of $\widehat{\boldsymbol{M}}$. In the appendix, we show that the grand mean and population average main effects obtained via ANOVA decomposition of $\widehat{\boldsymbol{M}}$ are identical to those obtained by performing an ANOVA decomposition of $\boldsymbol{Y}$, whereas the population average interaction effects obtained via ANOVA decomposition of $\widehat{\boldsymbol{M}}$ will differ from those

obtained via ANOVA decomposition of $\boldsymbol{Y}$ and may include many entries that are nearly equal to zero.

### *2.3.3 Testing*

LANOVA penalization assumes the distribution of entries of $\boldsymbol{C}$ have tail behavior consistent with a Laplace distribution. It is natural to ask if this assumption is appropriate but it is difficult to test it, because $\boldsymbol{C}$ and $\boldsymbol{Z}$ enter into the observed data through their sum $\boldsymbol{C}+\boldsymbol{Z}$. Accordingly, we suggest a test of the more general assumption that elements of $\boldsymbol{C}$ are heavy-tailed. This allows us to rule out LANOVA penalization when it is especially inappropriate, i.e. when the data suggest elements of $\boldsymbol{C}$ are normal tailed. When the distribution of elements of $\boldsymbol{C}$ is heavy-tailed, the distribution of elements of $\boldsymbol{C}+\boldsymbol{Z}$ will also be heavy-tailed and will have strictly positive excess kurtosis. In contrast, when elements of $\boldsymbol{C}$ are either all zero or have a distribution with normal tails, elements of $\boldsymbol{C}+\boldsymbol{Z}$ have excess kurtosis equal to exactly zero. We construct a test of the null hypothesis $H_0$: $c_{ij}+z_{ij} \sim$ i.i.d. $N\left(0, \sigma_c^2+\sigma_z^2\right)$, which encompasses the cases in which $\boldsymbol{C}=\boldsymbol{0}$ or elements of $\boldsymbol{C}$ are normally distributed. Conveniently, the test statistic is a simple function of $\widehat{\sigma}_c^2$, and $\widehat{\sigma}_z^2$ and can be computed at little additional computational cost. We can also think of this as a test of deconvolvability of $\boldsymbol{C}+\boldsymbol{Z}$ or whether or not the variances $\sigma_c^2$ and $\sigma_z^2$ can be separately identified, where the null hypothesis is that deconvolution of $\boldsymbol{C}+\boldsymbol{Z}$ is not possible and the variances $\sigma_c^2$ and $\sigma_z^2$ cannot be separately identified.

**Proposition 2.3.1** *For* $\boldsymbol{Y}=\mu \mathbf{1}_n \mathbf{1}_p'+\boldsymbol{a} \mathbf{1}_p'+\mathbf{1}_n \boldsymbol{b}'+\boldsymbol{C}+\boldsymbol{Z}$, *as* $n$ *and* $p \to \infty$ *an asymptotically level-*$\alpha$ *test of* $H_0$*:* $c_{ij}+z_{ij} \sim$ *i.i.d.* $N\left(0, \sigma_c^2+\sigma_z^2\right)$ *is obtained by rejecting* $H_0$ *when*

$$\sqrt{np}\left\{\frac{\widehat{\sigma}_c^4}{\sqrt{\frac{8}{3}\left(\widehat{\sigma}_c^2+\widehat{\sigma}_z^2\right)^2}}\right\}>z_{1-\alpha},$$

*where* $z_{1-\alpha}$ *denotes the* $1-\alpha$ *quantile of the standard normal distribution.*

This test gives us power against the alternative where elements $\boldsymbol{C}$ are heavy-tailed and LANOVA penalization may be appropriate.

Because this is an approximate test, we assess its level in finite samples in a small simulation study. Setting $\sigma_z^2 = 1$, $n = p$, $\mu = 0$, $\boldsymbol{a} = \mathbf{0}$ and $\boldsymbol{b} = \mathbf{0}$, we simulate $10,000$ realizations of $\boldsymbol{Y} = \boldsymbol{C} + \boldsymbol{Z}$ under $H_0$ for each value of $n \in \{25, 100\}$ and $\sigma_c^2 \in \{1/2, 1, 3/2\}$. When $n = p = 25$, the test rejects at a slightly higher rate than the nominal level. It rejects in 7.98%, 7.65% and 8.66% of simulations for $\sigma_c^2$ equal to 1/2, 1 and 3/2, respectively. When $n = p = 100$, the test nearly achieves the desired level. It rejects in 6.13%, 5.60% and 6.00% of simulations for $\sigma_c^2$ equal to 1/2, 1 and 3/2, respectively. We compute the approximate power of this test under two heavy-tailed distributions for elements of $\boldsymbol{C}$: the Laplace distribution assumed for LANOVA penalization and a Bernoulli-normal spike-and-slab distribution.

**Proposition 2.3.2** *Assume that elements of $\boldsymbol{C}$ are independent, identically distributed mean zero Laplace random variables with variance $\sigma_c^2$ and let $\phi^2 = \sigma_c^2/\sigma_z^2$. Then as $n$ and $p \to \infty$, the asymptotic power of the test given by Proposition 2.3.1 is:*

$$1 - \Phi\left[\frac{z_{1-\alpha} - \sqrt{\frac{3np}{8}}\left(\frac{\phi^2}{\phi^2+1}\right)^2}{\sqrt{1 + \left\{\frac{68\phi^8+36\phi^6+9\phi^4}{(1+\phi^2)^4}\right\}}}\right].$$

The power depends on the variances $\sigma_c^2$ and $\sigma_z^2$ only through their ratio $\phi^2$. It is plotted for for $\alpha = 0.05$, $\phi^2 \in [0, 2]$ and $np = \{100, 200, \ldots, 1000\}$ in Figure 2.2. The power of the test is increasing in $\phi^2$ and increasing more quickly when $np$ is larger and more data are available.

Now we consider the power for Bernoulli-normal spike-and-slab distributed elements of $\boldsymbol{C}$.

**Proposition 2.3.3** *Assume that elements of $\boldsymbol{C}$ are independent, identically distributed Bernoulli-normal random variables. An element of $\boldsymbol{C}$ is exactly equal to zero with probability $1 - \pi_c$, and normally distributed with mean zero and variance $\tau_c^2$ otherwise. Letting $\phi^2 = \tau_c^2/\sigma_z^2$, as $n$ and $p \to \infty$, the asymptotic power of the test given by Proposition 2.3.1 is:*

$$1 - \Phi\left[\frac{z_{1-\alpha} - \pi_c(1-\pi_c)\left\{\sqrt{\frac{3np}{8}}\left(\frac{\phi^2}{\pi_c\phi^2+1}\right)^2\right\}}{\sqrt{1 + \pi_c(1-\pi_c)\left\{\frac{(20\pi_c^2-28\pi_c+35)\phi^8+16(5-\pi_c)\phi^6+72\phi^4}{8(\pi_c\phi^2+1)^4}\right\}}}\right].$$

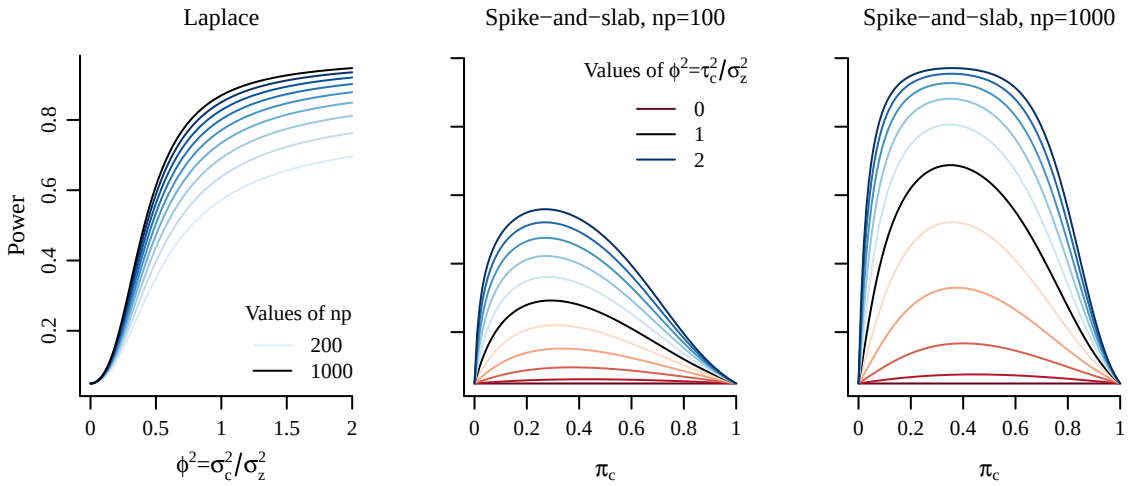



Figure 2.2: Approximate power of the test described in Proposition 2.3.1.

The approximate power depends on the variances of the nonzero effects $\tau_c^2$ and the noise $\sigma_z^2$ only through their ratio $\phi^2$. It is plotted for $\alpha = 0.05$, $\pi_c \in [0, 1]$, $\phi^2 \in \{0, 0.2, \ldots, 2\}$ and $np = \{100, 1000\}$ in Figure 2.2. The approximate power is always increasing in $\phi^2$ and $np$. For fixed $\phi^2$ and $np$, power diminishes as the probability of an element of $\boldsymbol{C}$ being nonzero $\pi_c$ approaches 0 or 1 and $\boldsymbol{C} + \boldsymbol{Z}$ becomes more normally distributed. Overall, the test is more powerful when estimating $\boldsymbol{C}$ separately from $\boldsymbol{Z}$ is more valuable, e.g. when elements of $\boldsymbol{C}$ are large in magnitude relative to the noise and when many entries of $\boldsymbol{C}$ are exactly zero.

### *2.3.4 Robustness*

If the true model is not the LANOVA model and elements of $\boldsymbol{C}$ are drawn from a different heavy-tailed distribution, it is natural to ask how our empirical Bayes estimates $\widehat{\sigma}_c^2$ and $\widehat{\sigma}_z^2$ and our estimate $\widehat{\boldsymbol{M}}$ perform. We find that the excess kurtosis $\kappa$ of the "true" distribution of elements of $\boldsymbol{C}$ determines our ability to estimate $\sigma_c^2$ separately from $\sigma_z^2$.

**Proposition 2.3.4** *Under the model* $\boldsymbol{Y} = \mu \mathbf{1}_n \mathbf{1}_p' + \boldsymbol{a}\mathbf{1}_p' + \mathbf{1}_n \boldsymbol{b}' + \boldsymbol{C} + \boldsymbol{Z}$*, where elements of* $\boldsymbol{C}$ *are independent, identically distributed draws from a mean zero, symmetric distribution with*

*variance* $\sigma_c^2$*, excess kurtosis* $\kappa$ *and finite eighth moment and elements of* $\boldsymbol{Z}$ *are normally distributed with mean zero and variance* $\sigma_z^2$*, we have that* $\widehat{\sigma}_c^2 \xrightarrow{p} \sqrt{\kappa/3}\sigma_c^2$ *and* $\widehat{\sigma}_z^2 \xrightarrow{p} \sigma_z^2 + \left(1 - \sqrt{\kappa/3}\right)\sigma_c^2$ *as* $n \to \infty$ *with* $p$ *fixed,* $p \to \infty$ *with* $n$ *fixed, or* $n$ *and* $p \to \infty$*.*

Proposition 2.3.4 indicates that we underestimate $\sigma_c^2$ when elements of $\boldsymbol{C}$ are lighter-than-Laplace tailed and we overestimate $\sigma_c^2$ when elements of $\boldsymbol{C}$ are heavier-than-Laplace tailed. To see how this affects estimation of $\boldsymbol{M}$, we consider Bernoulli-normal spike-and-slab distributed elements of $\boldsymbol{C}$ and compare the risk of the LANOVA estimate $\widehat{\boldsymbol{M}}$ to the risk of the maximum likelihood estimate $\widehat{\boldsymbol{M}}_{MLE}$, the risk of the strictly additive estimate $\widehat{\boldsymbol{M}}_{ADD}$ and the risk of additive-plus-low rank estimates of $\boldsymbol{M}$ denoted by $\widehat{\boldsymbol{M}}_{LOW,1}$ and $\widehat{\boldsymbol{M}}_{LOW,5}$ which assume rank-1 and rank-5 $\boldsymbol{C}$, respectively. Additive-plus-low-rank estimates are computed according to Johnson and Graybill (1972). We compute Monte Carlo estimates of the risks setting $n = p = 25$, $\mu = 0$, $\boldsymbol{a} = \boldsymbol{0}$, $\boldsymbol{b} = \boldsymbol{0}$ and $\sigma_z^2 = 1$ and varying $\tau_c^2 = \{1/2, 1, 2\}$ and $\pi_c = \{0, 0.1, \ldots, 0.9, 1\}$. For each value of $\tau_c^2$, the variance of nonzero elements of $\boldsymbol{C}$, and $\pi_c$, the probability any element of $\boldsymbol{C}$ is nonzero, the Monte Carlo estimate is based on 500 simulated $\boldsymbol{Y}$. As $\boldsymbol{M}$ is a function of $\mu$, $\boldsymbol{a}$, $\boldsymbol{b}$ and $\boldsymbol{C}$, the performance $\widehat{\boldsymbol{M}}$ reflects the performance of $\widehat{\mu}$, $\widehat{\boldsymbol{a}}$, $\widehat{\boldsymbol{b}}$ and $\widehat{\boldsymbol{C}}$ indirectly.

The results indicate generally favorable performance of the LANOVA estimate $\widehat{\boldsymbol{M}}$ for Bernoulli-normal $\boldsymbol{C}$. Recalling that $\widehat{\boldsymbol{M}}_{ADD}$ is strictly additive and $\widehat{\boldsymbol{M}}_{MLE}$ is strictly non-additive, the LANOVA estimate $\widehat{\boldsymbol{M}}$ outperforms $\widehat{\boldsymbol{M}}_{ADD}$ when $\pi_c$ is larger and $\boldsymbol{M}$ has fewer additive elements and outperforms $\widehat{\boldsymbol{M}}_{MLE}$ when $\pi_c$ is smaller and $\boldsymbol{M}$ has more additive elements. The LANOVA estimate $\widehat{\boldsymbol{M}}$ outperforms additive-plus-low-rank estimates $\widehat{\boldsymbol{M}}_{LOW,1}$ and $\widehat{\boldsymbol{M}}_{LOW,5}$ when $\pi_c$ is smaller and more elements of $\boldsymbol{C}$ are exactly equal to zero. The LANOVA estimate $\widehat{\boldsymbol{M}}$ performs worse than $\widehat{\boldsymbol{M}}_{ADD}$ and $\widehat{\boldsymbol{M}}_{MLE}$ when $\pi_c \approx 0$ or $\pi_c \approx 1$ and $\boldsymbol{M}$ is nearly strictly additive and nearly strictly non-additive, respectively.

Recalling Proposition 2.3.4, note that the LANOVA estimate $\widehat{\boldsymbol{M}}$ performs best relative to $\widehat{\boldsymbol{M}}_{ADD}$ when $\pi_c \approx 0.5$. When $\pi_c = 0.5$, the excess kurtosis of the Bernoulli-normal spike-and-slab distribution $\kappa = 3\left(1 - \pi_c\right)/\pi_c$ matches that of the Laplace distribution. This suggests

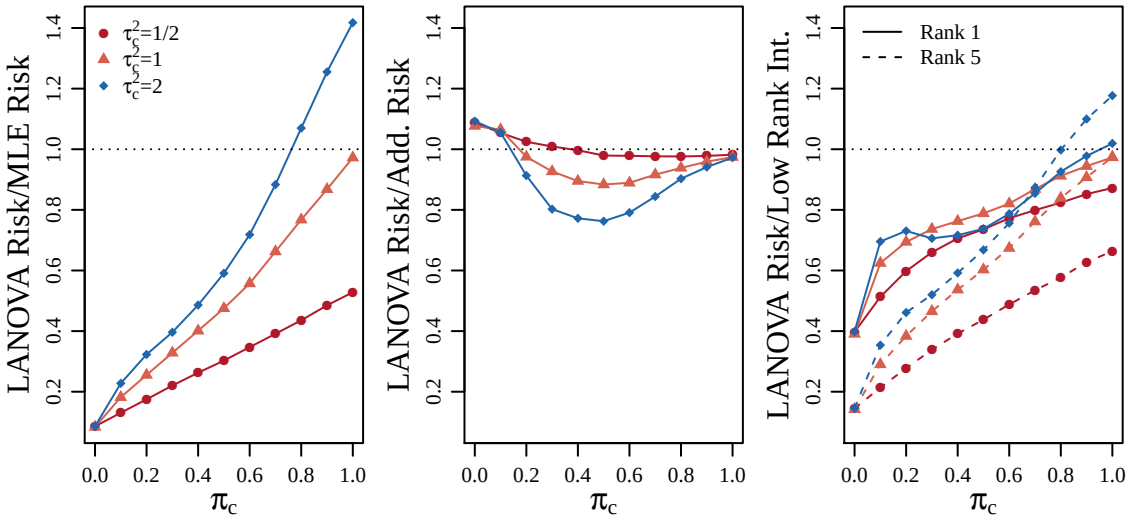



Figure 2.3: Monte Carlo approximations of relative risk of LANOVA estimate $\widehat{\boldsymbol{M}}$ versus the MLE $\widehat{\boldsymbol{M}}_{MLE}$, the strictly additive estimate of $\widehat{\boldsymbol{M}}_{ADD}$ and additive-plus-low rank. Estimates $\widehat{\boldsymbol{M}}_{LOW,1}$ and $\widehat{\boldsymbol{M}}_{LOW,5}$ are based on rank-1 and rank-5 estimates of $\boldsymbol{C}$.

that biased estimation of $\widehat{\sigma}_c^2$ and $\widehat{\sigma}_z^2$ yields poorer LANOVA estimates. Proposition 2.3.4 suggests a correction of multiplying $\widehat{\sigma}_c^2$ by $\sqrt{3/\kappa}$ and subtracting $(1-\sqrt{\kappa/3})\sqrt{3/\kappa}\widehat{\sigma}_c^2$ from $\widehat{\sigma}_z^2$. However because excess kurtosis $\kappa$ is not a readily interpretable quantity, specifying a more appropriate value of $\kappa$ *a priori* may be difficult. However if a Bernoulli-normal distribution for elements of $\boldsymbol{C}$ is plausible, a more appropriate value of $\pi_c$ can be used to specify a more appropriate value of $\kappa$. If the new value of $\pi_c$ is close to the "true" proportion of nonzero elements of $\boldsymbol{C}$, this may improve estimation of $\sigma_c^2$ and $\sigma_z^2$ and accordingly, $\boldsymbol{M}$.

## 2.4 Extensions

### 2.4.1 Penalizing Lower-Order Parameters

When $\boldsymbol{Y}$ has many rows or columns, it may be reasonable to believe that many elements of $\boldsymbol{a}$ or $\boldsymbol{b}$ are exactly zero. A natural extension of Equation (2.2) is given by

$$\text{minimize}_{\mu,\boldsymbol{a},\boldsymbol{b},\boldsymbol{C}} \frac{1}{2\sigma_z^2} \left|\left|\text{vec}\left(\boldsymbol{Y}-\boldsymbol{M}\right)\right|\right|_2^2 + \lambda_a \left|\left|\boldsymbol{a}\right|\right|_1 + \lambda_b \left|\left|\boldsymbol{b}\right|\right|_1 + \lambda_c \left|\left|\text{vec}\left(\boldsymbol{C}\right)\right|\right|_1, \tag{2.5}$$

where we still have $\boldsymbol{M} = \boldsymbol{1}_n\boldsymbol{1}_p'\mu + \boldsymbol{a}\boldsymbol{1}_p' + \boldsymbol{1}_n\boldsymbol{b}' + \boldsymbol{C}$. Again, using the posterior mode interpretation of Equation (2.5), we can estimate $\sigma_a^2$ and $\sigma_b^2$ from the observed data, $\boldsymbol{Y}$:

$$\widehat{\sigma}_a^2 = \frac{1}{n-1}\sum_{i=1}^{n}\breve{a}_i^2 - \frac{n}{(n-1)(p-1)}\overline{r}^{(2)}, \quad \widehat{\sigma}_b^2 = \frac{1}{p-1}\sum_{j=1}^{p}\breve{b}_j^2 - \frac{p}{(n-1)(p-1)}\overline{r}^{(2)}.$$

where $\breve{\boldsymbol{a}} = \boldsymbol{H}_n\boldsymbol{Y}\boldsymbol{1}_p/p$ and $\breve{\boldsymbol{b}} = \boldsymbol{H}_p\boldsymbol{Y}'\boldsymbol{1}_n/n$ are OLS estimates for $\boldsymbol{a}$ and $\boldsymbol{b}$. The estimators $\widehat{\lambda}_a = \sqrt{2/\widehat{\sigma}_a^2}$ and $\widehat{\lambda}_b = \sqrt{2/\widehat{\sigma}_b^2}$ can be shown to be consistent for $\lambda_a$ and $\lambda_b$ as $n \to \infty$ and $p \to \infty$, respectively. Because $\widehat{\lambda}_c$ and $\widehat{\sigma}_z^2$ do not depend on of $\boldsymbol{a}$ and $\boldsymbol{b}$, our estimators for $\lambda_c$ and $\sigma_z^2$ are unchanged. A block coordinate descent algorithm for solving Equation (2.5) is given in the appendix. With respect to interpretation, population average row and column main effects can be obtained via ANOVA decomposition of $\widehat{\boldsymbol{M}}$.

### 2.4.2 Tensor Data

LANOVA penalization can be extended to a $p_1 \times p_2 \times \cdots \times p_K$ $K$-mode tensor $\boldsymbol{Y}$. We consider:

$$\begin{aligned} \text{vec}\left(\boldsymbol{Y}\right) &= \boldsymbol{W}\boldsymbol{\beta} + \text{vec}\left(\boldsymbol{C}\right) + \text{vec}\left(\boldsymbol{Z}\right) \\ \boldsymbol{C} = \{c_{i_1\ldots i_K}\} &\sim \text{i.i.d. Laplace}\left(0, \lambda_c^{-1}\right), \quad \boldsymbol{Z} = \{z_{i_1\ldots i_K}\} \sim \text{i.i.d. N}\left(0, \sigma_z^2\right), \end{aligned} \tag{2.6}$$

where $\text{vec}\left(\boldsymbol{Y}\right)$ is the $\prod_{k=1}^{K} p_k$ vectorization of the $K$-mode tensor $\boldsymbol{Y}$ with "lower" indices moving "faster" and $\boldsymbol{W}$ and $\boldsymbol{\beta}$ are the design matrix and unknown mean parameters corresponding to a $K$-way ANOVA decomposition treating the $K$ modes of $\boldsymbol{Y}$ as factors. The

matrix $\boldsymbol{W} = [\boldsymbol{W}_1, \ldots, \boldsymbol{W}_{2^K-1}]$ is obtained by concatenating the $2^K - 1$ unique matrices of the form $\boldsymbol{W}_l = (\boldsymbol{W}_{l,1} \otimes \cdots \otimes \boldsymbol{W}_{l,K})$, where each $\boldsymbol{W}_{l,k}$ is equal to either $\boldsymbol{I}_{p_k}$ or $\boldsymbol{1}_{p_k}$, excluding the identity matrix, $\boldsymbol{I}_{p_K} \otimes \cdots \otimes \boldsymbol{I}_{p_1}$. As in the matrix case, approaches that assume a low rank $\boldsymbol{C}$ are common (van Eeuwijk and Kroonenberg, 1998; Gerard and Hoff, 2017). We penalize elements of the highest order mean term $\boldsymbol{C}$ for which no replicates are observed. In the three-way tensor case, the first part of Equation (2.6) refers to the following decomposition:

$$y_{ijk} = \mu + a_i + b_j + d_k + e_{ij} + f_{ik} + g_{jk} + c_{ijk} + z_{ijk}. \tag{2.7}$$

Estimates of $\sigma_z^2$ and $\lambda_c$ are constructed from $\text{vec}\,(\boldsymbol{R}) = (\boldsymbol{H}_K \otimes \cdots \otimes \boldsymbol{H}_1)\,\text{vec}\,(\boldsymbol{Y})$, where $\boldsymbol{H}_k = \boldsymbol{I}_{p_k} - \boldsymbol{1}_{p_k}\boldsymbol{1}_{p_k}/p_k$ is the $p_k \times p_k$ centering matrix and '$\otimes$' is the Kronecker product. As in the matrix case, $\text{vec}\,(\boldsymbol{R})$ is independent of the lower-order unknown mean parameters $\boldsymbol{\beta}$, i.e. $\text{vec}\,(\boldsymbol{R}) = (\boldsymbol{H}_K \otimes \cdots \otimes \boldsymbol{H}_1)\,\text{vec}\,(\boldsymbol{C} + \boldsymbol{Z})$. Our estimates of $\sigma_z^2$ and $\lambda_c$ are still functions of the second and fourth sample moments of $\boldsymbol{R}$: $\overline{r}^{(2)} = \frac{1}{p}\sum_{i=1}^{p} r_i^2$ and $\overline{r}^{(4)} = \frac{1}{p}\sum_{i=1}^{p} r_i^4$, where $p = \prod_{k=1}^{K} p_k$. We extend our empirical Bayes estimators as follows:

$$\begin{aligned} \widehat{\sigma}_c^4 &= \left\{ \prod_{k=1}^{K} \frac{p_k^3}{(p_k - 1)\left(p_k^2 - 3p_k + 3\right)} \right\} \left\{ \overline{r}^{(4)}/3 - \left(\overline{r}^{(2)}\right)^2 \right\} \\ \widehat{\sigma}_c^2 &= \sqrt{\widehat{\sigma}_c^4}, \quad \widehat{\sigma}_z^2 = \left( \prod_{k=1}^{K} \frac{p_k}{p_k - 1} \right) \overline{r}^{(2)} - \widehat{\sigma}_c^2, \end{aligned} \tag{2.8}$$

where $\widehat{\lambda}_c = \sqrt{2/\widehat{\sigma}_c^2}$. As in the matrix case, we can compute the bias of $\widehat{\sigma}_c^4$.

**Proposition 2.4.1** *Under the model given by Equation* (2.6)*,*

$$\begin{aligned} \mathbb{E}\left[\widehat{\sigma}_c^4\right] - \sigma_c^4 = -\left\{ \prod_{k=1}^{K} \frac{p_k^3}{(p_k - 1)\left(p_k^2 - 3p_k + 3\right)} \right\} \Bigg[ \left\{ 3\prod_{k=1}^{K} \frac{(p_k - 1)^2}{p_k^3} \right\} \sigma_c^4 + \\ \left( 2\prod_{k=1}^{K} \frac{p_k - 1}{p_k^2} \right) \left(\sigma_c^2 + \sigma_z^2\right)^2 \Bigg]. \end{aligned}$$

Interpretation of this result is analogous to the matrix case. We tend to prefer the simpler model with $\text{vec}\,(\boldsymbol{C}) = \boldsymbol{0}$ over a more complicated model with nonzero elements of $\text{vec}\,(\boldsymbol{C})$ when few data are available or when the data are very noisy. Additionally, $\mathbb{E}\left[\widehat{\sigma}_c^4\right] - \sigma_c^4 =$

$O(1/p)$, i.e. the bias of $\widehat{\sigma}_c^4$ diminishes as the number of levels of *any* mode increases. We also assess the large-sample performance of our empirical Bayes estimators in the $K$-way tensor case.

**Proposition 2.4.2** *Under the model given by Equation* (2.6), $\widehat{\sigma}_c^4 \xrightarrow{p} \sigma_c^4$, $\widehat{\sigma}_c^2 \xrightarrow{p} \sigma_c^2$, $\widehat{\lambda}_c \xrightarrow{p} \lambda_c$ *and* $\widehat{\sigma}_z^2 \xrightarrow{p} \sigma_z^2$ *as* $p_{k'} \to \infty$ *with* $p_k$, $k \neq k'$, *fixed or* $p_1, \ldots, p_K \to \infty$.

A block coordinate descent algorithm for estimating the unknown mean parameters is given in the appendix. Results for testing the appropriateness of assuming heavy-tailed $\boldsymbol{C}$ and robustness carry over to $K$-way tensors. $K$-way tensor analogues to Propositions 2.3.1-2.3.4, where we replace $np$ with $p$ and assume all $p_1, \ldots, p_K \to \infty$, are shown to hold in the appendix. Lastly we can also extend LANOVA penalization for tensor data to penalize lower-order mean parameters. Because tensor-variate $\boldsymbol{Y}$ include even more lower-order mean parameters, penalizing lower-order parameters is especially useful. We give nuisance parameter estimators for penalizing lower-order parameters in the three-way case in the appendix.

### *2.5 Numerical Examples*

**Brain Tumor Data:** We consider a $356 \times 43$ matrix of gene expression measurements for 356 genes and 43 brain tumors. The 43 brain tumors include 24 glioblastomas and 19 oligodendroglial tumors, which include 5 astrocytomas, 8 oligodendrogliomas and 6 mixed oligoastrocytomas. This data is contained in the `denoiseR` package for `R` (Josse et al., 2016), and it has been used to identify genes associated with glioblastomas versus oligodendroglial tumors (Bredel et al., 2005; de Tayrac et al., 2009). We focus on comparison to de Tayrac et al. (2009), who used a variation of principal components analysis of $\boldsymbol{Y}$ to identify differentially expressed genes and groups of tumors which is similar to using an additive-plus-low-rank estimate of $\boldsymbol{M}$. Unlike pairwise test-based methods which require prespecified tumor groupings, LANOVA penalization and additive-plus-low-rank estimates can be used to examine differential expression both within and across types of brain tumors.

Differential expression *within* types of brain tumors in particular is of recent interest (Bleeker et al., 2012).

We apply LANOVA penalization with penalized interaction effects and unpenalized main effects. The test given by Proposition 2.3.1 supports a non-additive estimate of $\boldsymbol{M}$; we obtain a test statistic of 18.45 and reject the null hypothesis of normally distributed elementwise variability at level $\alpha = 0.05$ with $p < 10^{-5}$. We estimate that $11,188$ elements of $\boldsymbol{C}$ (73%) are exactly equal to zero, i.e. most genes are not differentially expressed. Figure 2.4 shows $\widehat{\boldsymbol{C}}$ and a subset containing fifty genes with the lowest gene-by-tumor sparsity rates.

The results of LANOVA penalization are consistent with those of de Tayrac et al. (2009). We observe that 49% and 56% of the elements of $\widehat{\boldsymbol{C}}$ involving the genes ASPA and PDPN are nonzero. Examination of $\widehat{\boldsymbol{M}}$ indicates overexpression of these genes among glioblastomas relative to oligodendroglial tumors, as observed in de Tayrac et al. (2009).

LANOVA penalization yields additional results that are consistent with the wider literature. The gene DLL3 has the highest rate of gene-by-tumor interactions at 74% and tends to be underexpressed in glioblastomas. This is consistent with findings of overexpression of DLL3 in brain tumors with better prognoses (Bleeker et al., 2012). The KLRC genes KLRC1, KLRC2 and KLRC3.1 all have very high rates of gene-by-tumor interactions at 72%, 70% and 60%. Ducray et al. (2008) has found evidence for differential KLRC expression across glioma subtypes. LANOVA penalization also indicates that several brain tumors have unique gene expression profiles. Glioblastomas 3, 4 and 30 have rates of nonzero gene-by-tumor interactions exceeding 50% and similar gene expression profiles. Specifically, we observe overexpression of FCGR2B and HMOX1 and underexpression of RTN3 for gliomastomas 3, 4 and 30. Overexpression of FCGR2B or HMOX1 is associated with poorer prognosis (Zhang et al., 2016; Ghosh et al., 2016), and RTN3 is differentially expressed across subgroups of glioblastoma that differ with respect to prognosis (Cosset et al., 2017). This suggests that glioblastomas 3, 4 and 30 may correspond to an especially aggressive subtype.

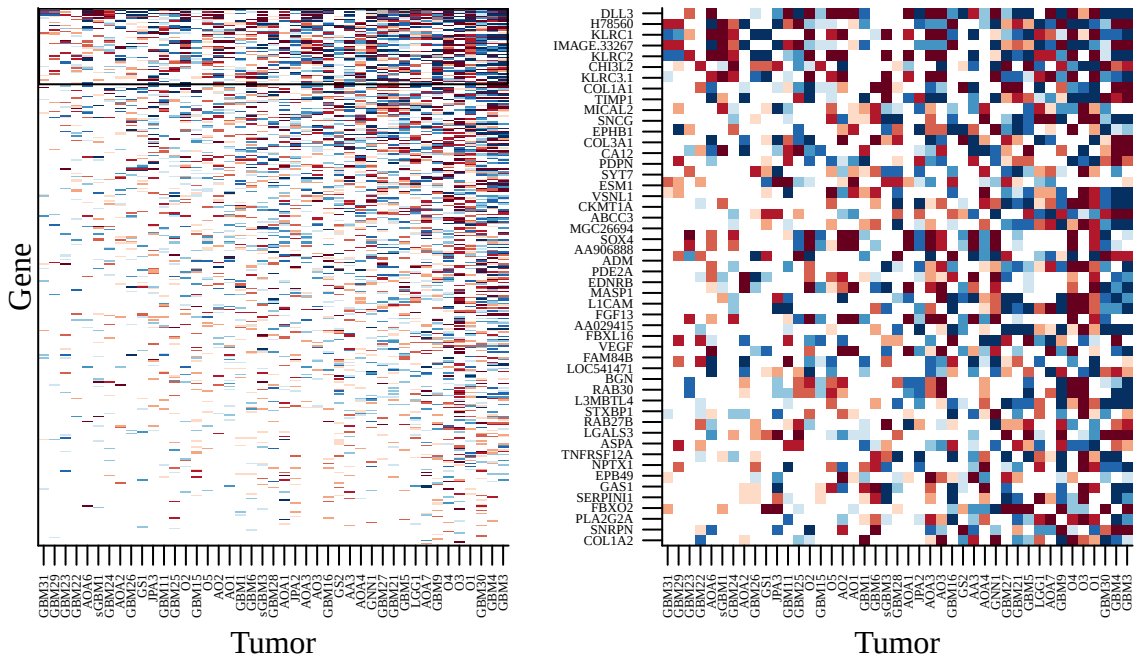



Figure 2.4: Elements of the $356 \times 43$ matrix $\widehat{\boldsymbol{C}}$. The first panel shows the entire matrix $\widehat{\boldsymbol{C}}$ with rows (genes) and columns (tumors) sorted in decreasing order of the row and column sparsity rates. The second panel zooms in on the rows of $\widehat{\boldsymbol{C}}$ (genes) marked in the first panel, which correspond to the fifty rows (genes) with the lowest sparsity rates. Colors correspond to positive (red) versus negative (blue) values of $\widehat{\boldsymbol{C}}$ and darkness corresponds to magnitude.

**fMRI Data:** Second, we consider a tensor of fMRI data which appeared in Mitchell et al. (2004). During each of 36 tasks, fMRI activations were measured at 55 time points and $4,698$ locations (voxels). Accordingly, the data can be represented as a $36 \times 55 \times 4,698$ three-way tensor. Because the data are so high dimensional, many methods of analysis are prohibitively computationally burdensome. Accordingly, parcellation approaches that reduce the spatial resolution of fMRI data by grouping voxels into spatially contiguous groups are common, however, the choice of a specific parcellation can be difficult (Thirion et al., 2014). Instead,

we propose LANOVA penalization as an exploratory method to identify relevant dimensions of spatial variation that should be accounted for in a subsequent analysis.

The test given by Proposition 2.3.1 supports a non-additive estimate of $\boldsymbol{M}$; we obtain a test statistic of 298.87 and reject the null hypothesis of normally distributed elementwise variability at level $\alpha = 0.05$ with $p < 10^{-5}$. Having found support for the use of a non-additive estimate of $\boldsymbol{M}$, we also penalize lower-order mean parameters as $\boldsymbol{Y}$ is high dimensional and sparsity of lower-order mean parameters could result in substantial dimension reduction and improved interpretability. The LANOVA estimate has $1,751,179$ nonzero parameters, a small fraction of the $9,302,040$ parameters needed to represent the raw data, $\boldsymbol{Y}$ (18.83%).

Recalling that we are primarily interested in spatial variation, we examine estimated task-by-location interactions $\widehat{\boldsymbol{F}}$ and task-by-time-by-location elementwise interactions $\widehat{\boldsymbol{C}}$, as defined in Equation (2.7). Figure 2.5 shows the percent of nonzero entries $\widehat{\boldsymbol{F}}$ and $\widehat{\boldsymbol{C}}$ at each location. At each location, the proportion of nonzero entries of $\widehat{\boldsymbol{F}}$ is much larger than the proportion of nonzero entries of $\widehat{\boldsymbol{C}}$. This suggests that much of the spatial variation of activations by task can be attributed to an overall level change in activation over the duration of the task, as opposed to time-specific changes in activation. In a subsequent analysis, it may be reasonable to ignore task-by-time-by-location interactions.

By examining the percent of nonzero entries of $\widehat{\boldsymbol{F}}$ by location, we can get a sense of which locations correspond to level changes in fMRI activity response by task. There is evidence for an overall level change in response to at least some tasks for all locations; the minimum percent of nonzero entries of $\widehat{\boldsymbol{F}}$ per location is 33%. However, voxels in the parietal region, the calcarine fissure and the right- and left-dorsolateral prefrontal cortex have particularly high proportions of nonzero entries of $\widehat{\boldsymbol{F}}$, suggesting that overall activation in these regions is particularly responsive to tasks. By examining the percent of nonzero entries of $\widehat{\boldsymbol{C}}$ by location, we can get a sense of which locations correspond to time-specific differential activity by task over time. We see that nonzero entries of $\widehat{\boldsymbol{C}}$ are concentrated among voxels in the upper supplementary motor area, the calcarine fissure and the left- and right-temporal lobes. In this way, we can use LANOVA estimates to identify subsets of

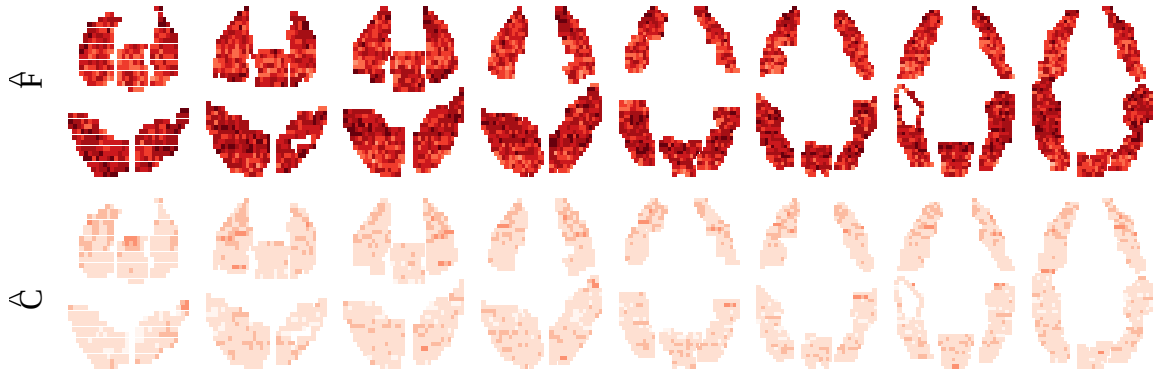



Figure 2.5: Percent of nonzero entries of $\widehat{\boldsymbol{F}}$ and $\widehat{\boldsymbol{C}}$ by location, where $\widehat{\boldsymbol{F}}$ and $\widehat{\boldsymbol{C}}$ estimate $\boldsymbol{F}$ and $\boldsymbol{C}$ as defined in Equation (2.7). Entries of $\boldsymbol{F}$ index task-by-location interaction terms and entries of $\boldsymbol{C}$ index task-by-time-by-location elementwise interaction terms. Darker colors indicate higher percentages.

relevant voxels that should be included in a subsequent analysis.

**Fusarium Data:** Last, we consider the problem of checking for nonzero three-way interactions in experimental data without replicates. The data is a $20 \times 7 \times 4$ three-way tensor containing severity of disease incidence ratings for 20 varieties of wheat infected with 7 strains of Fusarium head blight over 4 years, from 1990-1993, that appeared in van Eeuwijk and Kroonenberg (1998). There is scientific reason to believe that several nonzero three-way variety-by-strain-by-year interactions are present. van Eeuwijk and Kroonenberg (1998) examined these interactions using a rank-1 tensor model for $\boldsymbol{C}$ using a single multiplicative component for $\boldsymbol{C}$ with $c_{ijk} = \alpha_i \gamma_j \delta_k$, however as noted in the Introduction if the true interactions are sparse a low rank model may not be sufficient even if few nonzero interactions are present.

As in van Eeuwijk and Kroonenberg (1998), we transform the severity ratings to the logit scale before estimating LANOVA parameters. The test given by Proposition 2.3.1 supports a non-additive estimate of $\boldsymbol{M}$; we obtain a test statistic of 3.99 and reject the null hypothesis

of normally distributed elementwise variability at level $\alpha = 0.05$ with $p = 3.34 \times 10^{-5}$. We obtain 87 nonzero entries of $\widehat{\boldsymbol{C}}$ (16%).

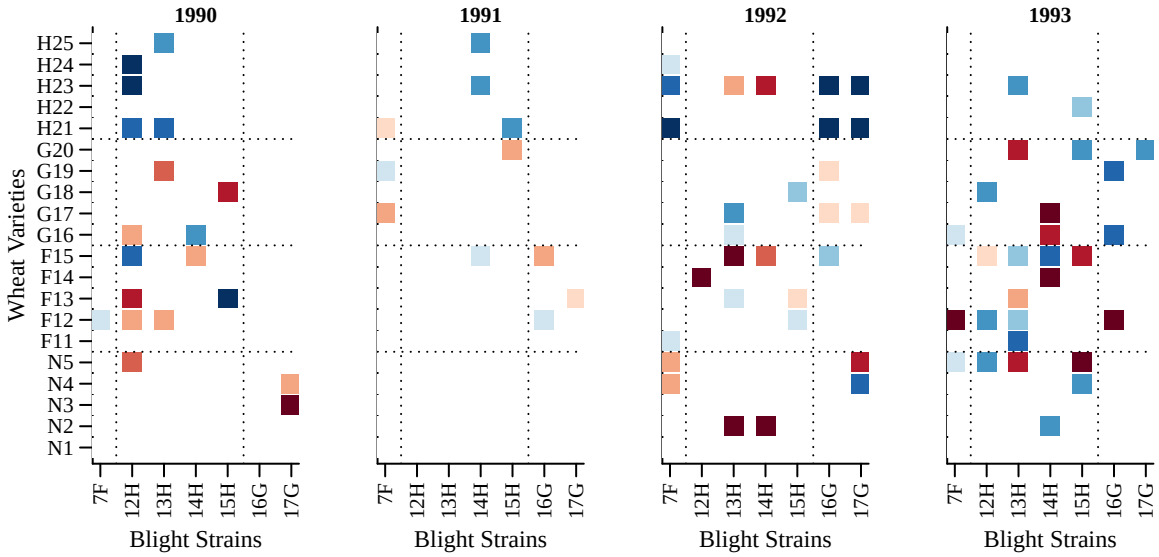



Figure 2.6: Entries of $\widehat{\boldsymbol{C}}$. Red (blue) points indicate positive (negative) nonzero entries of $\widehat{\boldsymbol{C}}$ and darker colors correspond to larger magnitudes.

Figure 2.6 shows nonzero elements of $\widehat{\boldsymbol{C}}$, with dashed lines separating groups of wheat and blight by country of origin: Hungary (H), Germany (G), France (F) or the Netherlands (N). We can interpret elements of $\widehat{\boldsymbol{C}}$ as evidence for variety-by-year-by-strain interactions that cannot be expressed as additive in variety-by-year, year-by-strain and variety-by-strain effects. Like van Eeuwijk and Kroonenberg (1998), we observe large three-way interactions in 1992, during which there was a disturbance in the storage of blight strains. Specifically, we observe interactions involving Dutch variety 2, the only variety with no infections at all in 1992, and interactions between Hungarian varieties 21 and 23 and foreign blight strains, which despite the storage disturbance were still able to cause infection in these two Hungarian varieties alone. We also observe evidence for "real" three way interactions for varieties exposed to Hungarian strain 12 in 1991 and French varieties exposed to Hungarian strains

in 1993 as well as other interactions. We do not have enough information about the data to assess whether or not these estimated interactions are related to features of the study or known patterns in the behavior of certain varieties and strains, however they suggest further investigation of these varieties and strains may be warranted.

### *2.6 Discussion*

This chapter demonstrates the use the common Lasso penalty and the corresponding Laplace prior distribution for estimating elementwise effects of scientific interest in the absence of replicates. Our procedure, LANOVA penalization, can also be interpreted as assessing evidence for nonadditivity. We show that our nuisance parameter estimators are consistent, explore their behavior when assumptions are violated and demonstrate that the corresponding mean parameter estimates tend to perform favorably relative to strictly additive, strictly non-additive and additive-plus-low-rank estimates when elements of $\boldsymbol{C}$ are Bernoulli-normal distributed. We emphasize that LANOVA penalization is computationally simple. The nuisance parameter estimators are easy to compute for arbitrarily large matrices, and estimates of $\boldsymbol{M}$ can be computed using a fast block coordinate descent algorithm that exploits the structure of the problem. We also extend LANOVA penalization to penalize lower-order mean parameters and apply to tensors. Finally, we show that LANOVA estimates can be used to examine gene-by-tumor interactions using microarray data, to perform exploratory analysis of spatial variation in activation response to tasks over time in high dimensional fMRI data and to assess evidence for "real" elementwise interaction effects in experimental data. To conclude, we discuss several limitations and extensions.

One limitation is that we assume heavy-tailed elementwise variation is of scientific interest and should be incorporated into the mean $\boldsymbol{M}$, whereas normally distributed elementwise variation is spurious noise $\boldsymbol{Z}$. If the noise is heavy-tailed, it may erroneously be incorporated into the estimate of $\boldsymbol{C}$. Similar issues arise with low rank models for $\boldsymbol{C}$, insofar as systematic correlated noise can erroneously be incorporated into the estimate of $\boldsymbol{C}$. In general, any method that aims to separate elementwise variation into components that are of

scientific interest and spurious noise requires strong assumptions that must be considered in the context of the problem. Another limitation is that the results of the simulation study assessing the performance of the LANOVA estimate $\widehat{\boldsymbol{M}}$ do not necessarily suggest favorable performance of the LANOVA estimate in all settings. First, we only consider Bernoulli-normal elements of $\boldsymbol{C}$ where each element $c_{ij}$ is exactly equal to 0 with probability $1 - \pi_c$ and normally distributed with mean zero and variance $\tau_c^2$ otherwise. Although we expect the LANOVA estimate to perform well when elements of $\boldsymbol{C}$ are heavy tailed, we do not expect the LANOVA estimate to perform well when elements of $\boldsymbol{C}$ are light-tailed. Second, this scenario considers $\boldsymbol{C}$ with independent, identically distributed elements. In some settings, it may be reasonable to expect dependence across elements of $\boldsymbol{C}$ and additive-plus-low-rank estimates may perform better.

The methods presented in this chapter could be extended in several ways. Although we use our estimators of $\lambda_c$ and $\sigma_z^2$ in posterior mode estimation, the same estimators could also be used to simulate from the posterior distribution using a Gibbs sampler. The output of a Gibbs sampler could be used to construct credible intervals for elements of $\boldsymbol{M}$, which would be one way of addressing uncertainty. We may also want to account for additional uncertainty induced by estimating $\lambda_c$ and $\sigma_z^2$. To address this, a fully Bayesian approach with prior distributions set for $\lambda_c$ and $\sigma_z^2$ could be taken at the cost of losing a sparse estimate of $\boldsymbol{C}$. Our empirical Bayes nuisance parameter estimators could be used to set parameters of prior distributions for $\lambda_c$ or $\sigma_z^2$. Last, LANOVA penalization for matrices is a specific case of the more general bilinear regression model, where we assume $\boldsymbol{Y} = \boldsymbol{AW} + \boldsymbol{BX} + \boldsymbol{C} + \boldsymbol{Z}$, given known $\boldsymbol{W}$ and $\boldsymbol{X}$. The same logic used in this chapter could be extended to this context and a more general multilinear context for tensor $\boldsymbol{Y}$.

Finally, our intuition combined with the results of Section 2.3 suggests that the Laplace distributional assumptions used in this chapter are likely to be violated in many settings. Whereas the strength of the Laplace distributional assumption for elements of $\boldsymbol{C}$ can be justified by the need to make *some* assumptions in the absence of replicate measurements, it cannot be similarly justified for lower-order mean parameters. This leads us to one last

extension: improved distribution specification for unknown mean parameters in the presence of replicate measurements. In future work, we will consider scenarios where we have enough data to estimate the variances of unknown mean parameters and noise from second order moments alone and can use fourth order moments to test the appropriateness of a specific distribution for unknown mean parameters or choose a better one.

# Chapter 3

# TESTING SPARSITY-INDUCING PENALTIES

## *3.1 Introduction*

Lasso estimators are ubiquitous in linear regression due to their desirable properties and computational feasibility, as they can be used to produce sparse estimates of regression coefficients without sacrificing convexity of the estimation problem (Tibshirani, 2011). The lasso estimator solves minimize$_{\boldsymbol{\beta}} ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2 + \lambda ||\boldsymbol{\beta}||_1$, where $\boldsymbol{y}$ is an $n \times 1$ vector of responses, $\boldsymbol{X}$ is an $n \times p$ design matrix and the value $\lambda > 0$ determines the relative importance of the penalty $||\boldsymbol{\beta}||_1$ compared to the model fit $||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2$ in estimating $\boldsymbol{\beta}$. It has long been recognized that the lasso estimator corresponds to the posterior mode when $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{z}$ and elements of $\boldsymbol{z}$ and $\boldsymbol{\beta}$ are independent normal and Laplace random variables, respectively (Tibshirani, 1996; Figueiredo, 2003). The Laplace prior interpretation is popular in part because sampling from the full posterior distribution using a Gibbs sampler is computationally feasible (Park and Casella, 2008). This allows computation of alternative posterior summaries, e.g. the posterior mean, median and quantiles, which can be used to obtain point and interval estimates of $\boldsymbol{\beta}$. Furthermore, the prior interpretation of a penalty yields decision theoretic justifications for using the corresponding penalized estimate or alternative posterior summaries to estimate $\boldsymbol{\beta}$ (Hans, 2009).

However, many researchers have found that the lasso estimator may perform suboptimally compared to other penalized estimators if the true value of $\boldsymbol{\beta}$ is highly sparse or not sparse at all (Fan and Li, 2001; Leeb and Pötscher, 2008). Analogously, posterior summaries under a Laplace prior have been found to be suboptimal compared to posterior summaries under other priors, depending on the empirical distribution of true values of the elements of $\boldsymbol{\beta}$ (Griffin and Brown, 2010; Carvalho et al., 2010; Castillo et al., 2015; Bhattacharya et al.,

2015; van der Pas et al., 2017). As a result, although a lasso estimator or a Laplace prior may be a reasonable default choice in high-dimensional problems due to its blend of desirable properties and computational feasibility, it would be useful to have a data-driven means to assess its appropriateness, and choose a more appropriate prior or penalty if suggested by the data.

One approach is to embed the lasso estimator in a larger class of penalized estimators and determine if the lasso estimator is a reasonable choice within this class. Consider the class of $\ell_q$ penalized estimators, which solve minimize$_{\boldsymbol{\beta}} ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2 + \lambda ||\boldsymbol{\beta}||_q^q$ for some $q > 0$, where $||\boldsymbol{\beta}||_q^q = \sum_{j=1}^p |\beta_j|^q$. This includes the ridge estimator given by $q = 2$, which has long been known to have desirable shrinkage properties (Hoerl and Kennard, 1970). The $\ell_q$ class includes the class of bridge estimators described by Frank and Friedman (1993) and accordingly, penalties that can outperform $\ell_1$ penalties when the true value of $\boldsymbol{\beta}$ is highly sparse, at the cost of losing convexity of the estimation problem (Huang et al., 2008; Mazumder et al., 2011; Marjanovic and Solo, 2014). Approaches to evaluating the appropriateness of $q = 1$ from this perspective might include cross validation, generalized cross validation or unbiased risk estimate minimization over values of $\lambda$ and $q$ simultaneously. However, these procedures can be challenging to perform over a two dimensional grid. Another approach is to assume the more flexible class of exponential power prior distributions which includes the Laplace prior as a special case (Subbotin, 1923; Box and Tiao, 1973):

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{z}, \quad \beta_1, \ldots, \beta_p \overset{i.i.d.}{\sim} EP(\tau, q), \quad \boldsymbol{z} \sim N\left(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n\right), \tag{3.1}$$

where $EP$ is the exponential power distribution with unknown shape parameter $q > 0$ and $\tau^2$ and $\sigma^2$ are the unknown variances of the regression coefficients $\boldsymbol{\beta}$ and error $\boldsymbol{z}$. The corresponding posterior mode of $\boldsymbol{\beta}$ is an $\ell_q$-penalized estimate with $\lambda = \tau^{-q}(\frac{\Gamma(3/q)}{\Gamma(1/q)})^{q/2}$. Posterior simulation under an exponential power prior can be more computationally demanding, however the corresponding posterior summaries may outperform those based on Laplace priors for highly sparse or non-sparse $\boldsymbol{\beta}$. Fully Bayesian inference could proceed by assuming priors for the error variance $\sigma^2$, $\lambda$ and $q$ at the expense of losing a possibly sparse and compu-

tationally tractable posterior mode (Polson et al., 2014). Accordingly, we could evaluate the appropriateness of a Laplace prior by computing a Bayes factor. However, specifying reasonable priors for $\lambda$ and $q$ that yield a proper posterior distribution is difficult in practice (Fabrizi and Trivisano, 2010; Salazar et al., 2012).

Alternatively, a likelihood ratio or score test of the null hypothesis $H : q = 1$ against the alternative hypothesis, $K : q \neq 1$ under the model given by (3.1) could be constructed. However, constructing a likelihood ratio test would require prohibitively computationally demanding maximum likelihood estimation of $\tau^2$, $\sigma^2$ and $q$ under the alternative. Both likelihood ratio and score tests would also require derivation of the distribution of the test statistic under the null, which is challenging due to marginal dependence of $\boldsymbol{y}$ induced by assuming a prior distribution for $\boldsymbol{\beta}$. Importantly, all of these approaches except the construction of a score test share the disadvantage of requiring penalized estimation or posterior simulation for $q \neq 1$ be performed regardless of whether or not $q = 1$ is deemed appropriate, which negates any computational advantages offered by assuming $q = 1$.

In this chapter we consider the Laplace and exponential power prior interpretations of the lasso and $\ell_q$ penalties. We propose fast and easy-to-implement procedures for testing the appropriateness of a Laplace prior ($q = 1$) and estimating $q$ in the event of rejection. In Section 3.2 we describe our testing procedure, which rejects the Laplace prior if an estimate of the kurtosis of the elements of $\boldsymbol{\beta}$ exceeds a particular threshold. The threshold is chosen so that the test rejects with probability approximately equal to $\alpha$, on average across datasets and Laplace-distributed coefficient vectors $\boldsymbol{\beta}$. We evaluate the performance of the approximation and the power of the testing procedure numerically via simulation. In Section 3.3 we introduce moment-based empirical Bayes estimates of $q$ and the variances $\sigma^2$ and $\tau^2$ of the error and the regression coefficients. We also propose a two-stage adaptive procedure for estimating $\boldsymbol{\beta}$. If the testing procedure fails to reject the null, the adaptive estimation procedure defaults to an estimate computed under a Laplace prior. Otherwise, we estimate $\boldsymbol{\beta}$ under an exponential power prior using an estimated value of $q$. We show via simulation that the adaptive estimation procedure outperforms estimators based on a Laplace prior when

elements of $\boldsymbol{\beta}$ have an exponential power distribution with $q \neq 1$ and performs similarly to estimators based on a Laplace prior when elements of $\boldsymbol{\beta}$ have a Laplace distribution. In Section 3.4, we demonstrate that the adaptive procedure also improves estimation of sparse $\boldsymbol{\beta}$ when elements of $\boldsymbol{\beta}$ have a spike-and-slab distribution. In Section 3.5, we apply the testing and estimation procedures to several datasets commonly used in the penalized regression literature. A discussion follows in Section 3.6.

### *3.2 Testing the Laplace Prior*

Our approach to testing the appropriateness of the Laplace prior treats the Laplace prior as a special case of the larger class of exponential power distributions (Subbotin, 1923; Box and Tiao, 1973). This class includes the normal and Laplace distributions. The exponential power density is given by

$$p\left(\beta_j|\tau^2, q\right) = \left(\frac{q}{2\tau}\right)\sqrt{\frac{\Gamma\left(3/q\right)}{\Gamma\left(1/q\right)^3}}\exp\left\{-\left(\frac{\Gamma\left(3/q\right)}{\Gamma\left(1/q\right)}\right)^{q/2}\left|\frac{\beta_j}{\tau}\right|^q\right\}, \tag{3.2}$$

where $q > 0$ is an unknown shape parameter and $\tau^2$ is the variance. The first panel of Figure 3.1 plots exponential power densities for different values of $q$ with the variance fixed at $\tau^2 = 1$. Because the exponential power distributions have simple distribution functions with easy to compute moments and can accommodate a wide range of tail behaviors, they quickly became popular as alternative error distributions (Subbotin, 1923; Diananda, 1949; Box, 1953; Box and Tiao, 1973).

When an exponential power prior is assumed for $\boldsymbol{\beta}$, we can understand how the choice of $q$ provides flexible penalization by examining the mode thresholding function. The mode thresholding function relates the OLS estimate for a simplified problem with a single standardized covariate to the posterior mode of $\boldsymbol{\beta}$. Let $\boldsymbol{x}$ be a standardized $n \times 1$ covariate vector with $||\boldsymbol{x}||_2^2 = 1$, let $\beta$ be a scalar and let $\hat{\beta}_{ols} = \boldsymbol{x}^\top \boldsymbol{y}$. The mode thresholding function is given by:

$$\arg\min_\beta \frac{1}{2\sigma^2}\left(\hat{\beta}_{ols} - \beta\right)^2 + \left(\frac{\Gamma\left(3/q\right)}{\Gamma\left(1/q\right)}\right)^{q/2}\left|\frac{\beta}{\tau}\right|^q.$$

This function is not generally available in closed form but can be computed numerically, even when $q < 1$ and the mode thresholding problem is non-convex (Marjanovic and Solo, 2014). The second panel of Figure 3.1 shows the mode thresholding function for $\sigma^2 = \tau^2 = 1$.

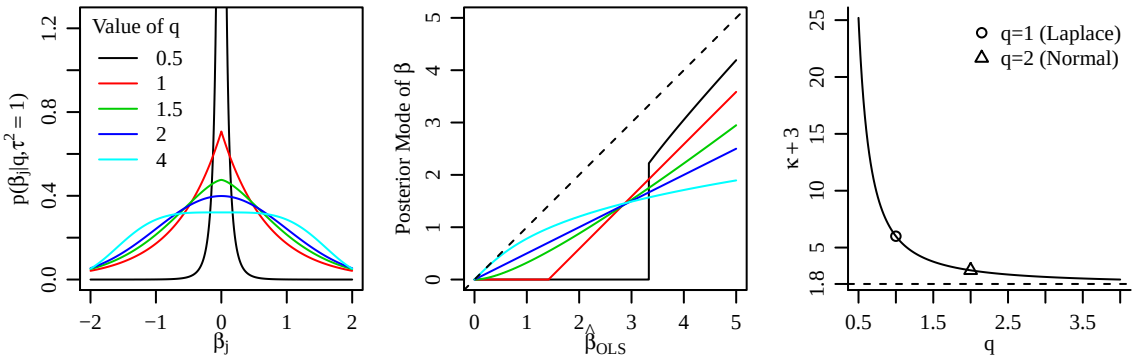



Figure 3.1: The first panel shows exponential power densities for fixed variance $\tau^2 = 1$ and varying values of the shape parameter $q$. The second panel shows the mode thresholding function for $\sigma^2 = \tau^2 = 1$ and the values of $q$ considered in the first panel. The third panel shows the relationship between the kurtosis of the exponential power distribution and the shape parameter, $q$.

Within the class of exponential power priors, the relationship between the shape parameter $q$ and kurtosis $\mathbb{E}\left[\beta_j^4\right] / \mathbb{E}\left[\beta_j^2\right]^2$ is one-to-one and given by:

$$\kappa + 3 = \Gamma\left(5/q\right) \Gamma\left(1/q\right) / \Gamma\left(3/q\right)^2, \tag{3.3}$$

where $\kappa$ refers to the excess kurtosis relative to a normal distribution. We plot kurtosis as a function of $q$ in the third panel of Figure 3.1. Accordingly, if $\boldsymbol{\beta}$ were observed we could naively construct a test statistic based on the empirical kurtosis of the elements of $\boldsymbol{\beta}$. We define the test statistic $\psi(\boldsymbol{\beta}) = m_4\left(\boldsymbol{\beta}\right) / m_2\left(\boldsymbol{\beta}\right)^2$, where $m_2\left(\boldsymbol{\beta}\right) = \frac{1}{p}\sum_{j=1}^{p} \beta_p^2$ and $m_4\left(\boldsymbol{\beta}\right) = \frac{1}{p}\sum_{j=1}^{p} \beta_p^4$ are the second and fourth empirical moments of $\boldsymbol{\beta}$. The test statistic $\psi(\boldsymbol{\beta})$ is the empirical kurtosis of the elements of the vector $\boldsymbol{\beta}$. An exact level-$\alpha$ test of $H$ could be performed by comparing the test statistic to the $\alpha/2$ and $1-\alpha/2$ quantiles of the distribution of $\psi(\boldsymbol{\beta})$ under

the null. Because the distribution of $\psi(\boldsymbol{\beta})$ under an exponential power prior depends only on $q$ and not $\tau^2$, we can obtain Monte Carlo estimates of the $\alpha/2$ and $1-\alpha/2$ quantiles $\psi_{\alpha/2}$ and $\psi_{1-\alpha/2}$ by simulating entries of $\boldsymbol{\beta}^*$ from any Laplace distribution and and computing $\psi(\boldsymbol{\beta}^*)$. As this test is available *only* when $\boldsymbol{\beta}$ is observed, we refer to this as the oracle test.

In practice, $\boldsymbol{\beta}$ is not observed. However when $n > p$ and $\boldsymbol{X}^\top\boldsymbol{X}$ is full rank, the OLS estimator $\hat{\boldsymbol{\beta}}_{ols} = \left(\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}\boldsymbol{X}^\top\boldsymbol{y}$ is available. As a surrogate for $\psi(\boldsymbol{\beta})$, we can use $\psi(\hat{\boldsymbol{\beta}}_{ols})$ as a test statistic. If $n >> p$, then $\hat{\boldsymbol{\beta}}_{ols} \approx \boldsymbol{\beta}$ conditional on $\boldsymbol{\beta}$. It follows that $\psi\left(\hat{\boldsymbol{\beta}}_{ols}\right) \overset{d}{\approx} \psi(\boldsymbol{\beta})$ when treating $\boldsymbol{\beta}$ as random.

**Proposition 3.2.1** *Under normality of the errors* $\boldsymbol{z}$ *as assumed in* (3.1),

$$\mathbb{E}\left[(\psi(\hat{\boldsymbol{\beta}}_{ols}) - \psi(\boldsymbol{\beta}))^2|\boldsymbol{\beta}\right] \leq 16\sigma^2\left(\frac{m_6(\boldsymbol{\beta})}{m_2(\boldsymbol{\beta})^4}\right) tr((\boldsymbol{X}^\top\boldsymbol{X})^{-1})/p + o\left(\sigma^2 tr((\boldsymbol{X}^\top\boldsymbol{X})^{-1})/p\right) \quad (3.4)$$

*where* $m_6(\boldsymbol{\beta}) = \frac{1}{p}\sum_{j=1}^{p}\beta_j^6$.

Details are provided in the appendix. Accordingly, when $\text{tr}((\boldsymbol{X}^\top\boldsymbol{X})^{-1})/p$ is small an approximate level-$\alpha$ test of $H$ is obtained by rejecting $H$ when $\psi\left(\hat{\boldsymbol{\beta}}_{ols}\right) \notin \left(\psi_{\alpha/2}, \psi_{1-\alpha/2}\right)$.

Although the behavior of the OLS estimator is well understood, we introduce some additional notation to help explain when $\text{tr}((\boldsymbol{X}^\top\boldsymbol{X}))^{-1}$ is likely to be small for large $n$. Let $\boldsymbol{V}$ be a diagonal matrix with diagonal elements $\sqrt{\text{diag}\left(\boldsymbol{X}^\top\boldsymbol{X}\right)}$ and $\boldsymbol{C}$ be the "correlation" matrix corresponding to $\boldsymbol{X}^\top\boldsymbol{X}$, such that $\boldsymbol{X}^\top\boldsymbol{X} = \boldsymbol{V}\boldsymbol{C}\boldsymbol{V}$. Let $\eta_j$ refer to eigenvalues of $\boldsymbol{C}$. The eigenvalues $\eta_j$ indicate the overall collinearity of $\boldsymbol{X}$. When columns of $\boldsymbol{X}$ are orthogonal, $\eta_1 = \cdots = \eta_p = 1$, whereas when $\boldsymbol{X}$ is highly collinear the smallest values of $\eta_j$ may be very close or exactly equal to 0. Applying Theorem 3.4 of Styan (1973) we can write

$$\text{tr}((\boldsymbol{X}^\top\boldsymbol{X})^{-1})/p \leq \max_j\left(\frac{1}{||\boldsymbol{x}_j||_2^2}\right)\max_j\left(\frac{1}{\eta_j^2}\right).$$

We can see that as long as $||\boldsymbol{x}_j||_2^2$ are large, which will tend to be the case when $n$ is large, and eigenvalues of $\eta_j$ are not very small, i.e. $\boldsymbol{X}$ is not too highly collinear, $\text{tr}((\boldsymbol{X}^\top\boldsymbol{X})^{-1})/p$ will be small enough to justify using $\psi(\hat{\boldsymbol{\beta}}_{ols})$ as a surrogate for $\psi(\boldsymbol{\beta})$.

However, penalized regression is often considered when $n < p$ or $\boldsymbol{X}$ is highly collinear. When $n < p$, the OLS estimator is not unique and so neither is $\psi(\hat{\boldsymbol{\beta}}_{ols})$. When $n \geq p$ but $\boldsymbol{X}$ is highly collinear, i.e. some columns of $\boldsymbol{X}$ are strongly correlated with others and $\text{tr}((\boldsymbol{X}^\top \boldsymbol{X})^{-1})/p$ may not be small even for large values of $n$. When columns of $\boldsymbol{X}$ have been centered and standardized to have norm $n$ according to standard practice, we can easily see that the quantity $\text{tr}((\boldsymbol{X}^\top \boldsymbol{X})^{-1})/p = \frac{1}{np}\sum_{j=1}^{p} \frac{1}{\eta_j}$ will "blow up" if any values of $\eta_j$ are very close to or exactly equal to zero and quantiles of $\psi(\boldsymbol{\beta})$ will poorly approximate quantiles of $\psi(\hat{\boldsymbol{\beta}}_{ols})$.

Fortunately, we can construct a modified test using a ridge estimate of $\boldsymbol{\beta}$, $\hat{\boldsymbol{\beta}}_\delta = \boldsymbol{V}^{-1}(\boldsymbol{C} + \delta^2 \boldsymbol{I}_p)^{-1}\boldsymbol{V}^{-1}\boldsymbol{X}^\top \boldsymbol{y}$, where $\delta \geq 0$ is a nonnegative constant. Ridge estimators reduce variance at the cost of yielding a biased estimate of $\boldsymbol{\beta}$, $\mathbb{E}[\hat{\boldsymbol{\beta}}_\delta|\boldsymbol{\beta}] = \boldsymbol{V}^{-1}\left(\boldsymbol{C} + \delta^2 \boldsymbol{I}_p\right)^{-1}\boldsymbol{C}\boldsymbol{V}\boldsymbol{\beta}$. However letting $\boldsymbol{\beta}_\delta = \mathbb{E}[\hat{\boldsymbol{\beta}}_\delta|\boldsymbol{\beta}]$, the distribution of $\psi\left(\boldsymbol{\beta}_\delta\right)$ under an exponential power prior still only depends on $q$ and not $\tau^2$. Accordingly, if $\boldsymbol{\beta}_\delta$ were observed we could perform an exact level-$\alpha$ test of $H$ by comparing $\psi\left(\boldsymbol{\beta}_\delta\right)$ to Monte Carlo estimates of the $\alpha/2$ and $1-\alpha/2$ quantiles $\psi_{\delta,\alpha/2}$ and $\psi_{\delta,1-\alpha/2}$ obtained by simulating $\boldsymbol{\beta}^*$ from any Laplace distribution and computing $\psi(\boldsymbol{\beta}^*_\delta = \psi\left(\boldsymbol{V}^{-1}\left(\boldsymbol{C} + \delta^2 \boldsymbol{I}_p\right)^{-1}\boldsymbol{C}\boldsymbol{V}\boldsymbol{\beta}^*\right)$. In practice, we can use $\psi(\hat{\boldsymbol{\beta}}_\delta)$ as a surrogate for $\psi\left(\boldsymbol{\beta}_\delta\right)$ to obtain an approximate level-$\alpha$ test.

**Proposition 3.2.2** *Let* $\boldsymbol{\Sigma}_\delta = \boldsymbol{V}^{-1}\left(\boldsymbol{C} + \delta^2 \boldsymbol{I}_p\right)^{-1}\boldsymbol{C}\left(\boldsymbol{C} + \delta^2 \boldsymbol{I}_p\right)^{-1}\boldsymbol{V}^{-1}$. *Under normality of the errors* $\boldsymbol{z}$ *as assumed in the model given by* (3.1),

$$\mathbb{E}\left[(\psi(\hat{\boldsymbol{\beta}}_\delta) - \psi(\boldsymbol{\beta}_\delta))^2|\boldsymbol{\beta}\right] \leq 16\sigma^2\left(\frac{m_6\left(\boldsymbol{\beta}_\delta\right)}{m_2(\boldsymbol{\beta}_\delta)^4}\right) tr\left(\boldsymbol{\Sigma}_\delta\right)/p + o\left(\sigma^2 tr\left(\boldsymbol{\Sigma}_\delta\right)/p\right), \tag{3.5}$$

*where* $m_k\left(\boldsymbol{\beta}_\delta\right) = \frac{1}{p}\sum_{j=1}^{p}\beta_j^k$.

Details are provided in the appendix. It follows that when $\text{tr}\left(\boldsymbol{\Sigma}_\delta\right)/p$ is small, an approximate level-$\alpha$ test of $H$ is obtained by rejecting $H$ when $\psi(\hat{\boldsymbol{\beta}}_\delta) \notin \left(\psi_{\delta,\alpha/2}, \psi_{\delta,1-\alpha/2}\right)$.

As the performance of this test depends on $\text{tr}\left(\boldsymbol{\Sigma}_\delta\right)/p$, it depends not only on $\boldsymbol{X}$ but also on $\delta^2$. Again applying Theorem 3.4 of Styan (1973), we can write

$$\text{tr}\left(\boldsymbol{\Sigma}_\delta\right)/p \leq \max_j\left(\frac{1}{||\boldsymbol{x}_j||_2^2}\right)\max_j\left(\frac{\eta_j}{\left(\eta_j + \delta^2\right)^2}\right).$$

The first term depends only on the design matrix, $\boldsymbol{X}$. As long as $||\boldsymbol{x}_j||_2^2$ are large, which again is likely to be the case for large $n$, the first term will be small. The second term depends on $\delta^2$ through the eigenvalue ratios $\frac{\eta_j}{(\eta_j+\delta^2)^2}$. Heuristically, the eigenvalue ratios are decreasing in $\delta^2$ and setting $\delta^2$ to be very large would ensure that $\mathrm{tr}\left(\boldsymbol{\Sigma}_\delta\right)/p$ is very close to 0 and that $\psi(\hat{\boldsymbol{\beta}}_\delta)$ performs well as a surrogate for $\psi(\boldsymbol{\beta}_\delta)$. However, increasing $\delta^2$ also reduces the power of the test as it forces the ridge estimate closer to the zero vector. To ensure that the eigenvalue ratios do not "blow up" while retaining as much power as possible we recommend setting $\delta^2 = \left(1 - \mathrm{minimize}_j \eta_j\right)_+$, where $\eta_1, \ldots, \eta_p$ are the eigenvalues of $\boldsymbol{C}$. When the columns of $\boldsymbol{X}$ are standardized to have norm $n$, $\mathrm{tr}\left(\boldsymbol{\Sigma}_\delta\right)/p = \frac{1}{np}\sum_{i=1}^{p} \frac{\eta_j}{(\delta^2+\eta_j)^2}$. Accordingly with $\delta^2$ set to $\delta^2 = \left(1 - \mathrm{minimize}_j \eta_j\right)_+$, we can at least ensure that $\mathrm{tr}\left(\boldsymbol{\Sigma}_\delta\right)/p \leq \frac{1}{n}$.

The tests based on $\psi(\hat{\boldsymbol{\beta}}_{ols})$ and $\psi(\hat{\boldsymbol{\beta}}_\delta)$ have several good features. First, the approximate distributions of the test statistics $\psi(\hat{\boldsymbol{\beta}}_{ols})$ and $\psi(\hat{\boldsymbol{\beta}}_\delta)$ do not depend on the values of the unknown parameters $\tau^2$ or $\sigma^2$, and so their approximate null distributions may be simulated easily. Second, both test statistics are easy and quick-to-compute even for very high dimensional data. Third, both test statistics are invariant to rescaling of $\boldsymbol{y}$ or $\boldsymbol{X}$ by a constant.

We examine the performance of the tests with a simulation study. We simulate parameters and data according to the model (3.1). When simulating data and parameters, we set $\sigma^2 = \tau^2 = 1$ and consider $p \in \{25, 50, 75, 100\}$, $n \in \{50, 100, 200\}$ and $q \in \{0.1, \ldots, 2\}$. Because the OLS and ridge test statistics are invariant to rescaling of $\boldsymbol{y}$ by a constant, the simulation results depend only on $\tau^2/\sigma^2$, and in this case reflect the performance of the tests when $\tau^2 = \sigma^2$. For each combination of $p$, $n$ and $q$, we simulate $1,000$ values of $\boldsymbol{X}$ and $\boldsymbol{\beta}$, drawing entries of $\boldsymbol{X}$ independently from the standard normal distribution. When $n > p$, we use the OLS test statistic $\psi(\hat{\boldsymbol{\beta}}_{ols})$. When $n \leq p$, we use the ridge test statistic $\psi(\hat{\boldsymbol{\beta}}_\delta)$. Figure 3.2 shows the power of the level-0.05 tests, i.e. the proportion of simulated datasets for which we reject $H$ at level-0.05 as a function of $q$ and $n$. When $q = 1$, this gives the level of the test. The last panel shows the power of the oracle test based on $\psi(\boldsymbol{\beta})$.

The simulation results shown in Figure 3.2 indicate that the tests will perform well relative

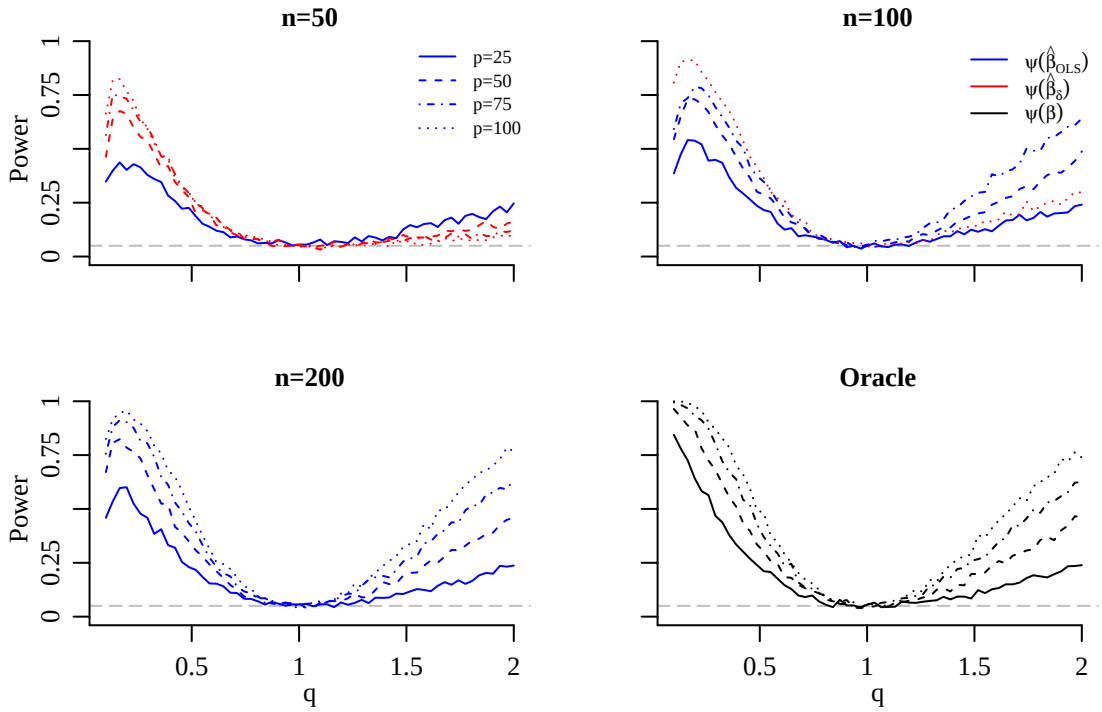



Figure 3.2: Power and Type-I error of level-0.05 tests for data simulated from model (3.1) with exponential power distributed $\boldsymbol{\beta}$ and $\sigma^2 = \tau^2 = 1$. A horizontal dashed gray line is given at 0.05.

to the oracle test for this range of values of $n$ and $p$. The power of the test is increasing in $p$, as this in a sense represents our sample size for evaluating the distribution of $\boldsymbol{\beta}$. The power of the test is also increasing in $q$ moves away from $q = 1$, i.e. as the empirical distribution of the elements of $\boldsymbol{\beta}$ becomes less similar to a Laplace distribution. As we might expect given that the ridge estimator corresponds to an estimator of $\boldsymbol{\beta}$ under a normal prior with variance $1/\delta^2$, using the modified ridge-based test results in a reduction of power especially against alternatives with $q > 1$. Interestingly, we see that the power of the test is not symmetric with respect to how far the true value of $q$ is from 1. This is due to the fact that kurtosis is changing more slowly as a function of $q$ as $q$ increases, as can be seen in Figure 3.1. We also

observe a dip in power for very small values of $q$ when $\psi(\hat{\boldsymbol{\beta}}_{ols})$ or $\psi(\hat{\boldsymbol{\beta}}_{\delta})$ are used. This can be explained by examining the bias of $m_2(\hat{\boldsymbol{\beta}}_{ols})$ which appears in the denominator of $\psi(\hat{\boldsymbol{\beta}}_{ols})$

$$\mathbb{E}\left[m_2(\hat{\boldsymbol{\beta}}_{ols})\right] = m_2(\boldsymbol{\beta}) + \sigma^2 \text{tr}((\boldsymbol{X}^\top \boldsymbol{X})^{-1})/p.$$

When $q$ is very small and $p$ is less than or equal to 100, most elements of $\boldsymbol{\beta}$ will be very close to zero with high probability. For instance, when $q = 0.1$ and $\tau^2 = 1$, $\Pr(|\beta_j| \geq 0.1) \approx 0.08$. When most elements of $\boldsymbol{\beta}$ are very close to 0, $m_2(\boldsymbol{\beta})$ will be small and $m_2(\hat{\boldsymbol{\beta}}_{ols})$ will be dominated by the error incurred by estimating $\boldsymbol{\beta}$. The behavior of $\psi(\hat{\boldsymbol{\beta}}_{\delta})$ at small values of $q$ can be explained analogously.

### *3.3 Adaptive Estimation of $\boldsymbol{\beta}$*

Rejection of the null hypothesis implies that the empirical distribution of the unobserved entries of $\boldsymbol{\beta}$ does not resemble a Laplace distribution. This suggests a two-stage adaptive procedure for estimating $\boldsymbol{\beta}$ that first tests the appropriateness of the Laplace prior and estimates $\boldsymbol{\beta}$ under a Laplace prior if the test accepts and estimates $\boldsymbol{\beta}$ under an exponential power prior otherwise. This procedure requires estimates of $\tau^2$ and $\sigma^2$ if the test accepts and $\tau^2$, $\sigma^2$ and $q$ if the test rejects, as well as procedures for computing the posterior mode or simulating from the posterior distribution of $\boldsymbol{\beta}$ under an exponential power prior. We do not specify which posterior summary should be used to estimate $\boldsymbol{\beta}$ in general. It is well known that different posterior summaries minimize different loss functions (Hans, 2009), and we view the choice of posterior summary as problem-specific.

We consider empirical Bayes estimation of $q$, $\tau^2$ and $\sigma^2$. Estimating these parameters by maximizing the marginal likelihood of the data $\int p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\beta}, \sigma^2)\, p(\boldsymbol{\beta}|\tau^2, q)\, d\boldsymbol{\beta}$ is difficult because the integral is not available in closed form for arbitrary values of $q$. The problem is not amenable to a Gibbs-within-EM algorithm for maximizing over $\sigma^2$, $\tau^2$ and $q$ jointly and Gibbs-within-EM algorithms to obtain maximum marginal likelihood estimates of $\tau^2$ and $\sigma^2$ for fixed $q$ are computationally intensive and tend to be slow to converge (Roy and Chakraborty, 2016). As a result, we consider easy and quick to compute moment-based

empirical Bayes estimators of $\sigma^2$, $\tau^2$ and $q$. As moment estimators, they are more robust to misspecification of the prior and residual distributions than likelihood-based alternatives. Conveniently, the estimators for $\tau^2$ and $\sigma^2$ do not depend on $q$. This yields simple and interpretable comparisons of estimates of $\boldsymbol{\beta}$ computed under Laplace versus exponential power priors.

### *3.3.1 Estimation of q*

The test statistics $\psi(\hat{\boldsymbol{\beta}}_{ols})$ and $\psi(\hat{\boldsymbol{\beta}}_{\delta})$ can be used to estimate $q$. In the previous section, we demonstrated that an approximate test of $H$ could be obtained by using $\psi(\hat{\boldsymbol{\beta}}_{ols})$ as a surrogate for $\psi(\boldsymbol{\beta})$ when $\text{tr}((\boldsymbol{X}^\top\boldsymbol{X})^{-1})/p$ is small. Recall that the quantity $\psi(\boldsymbol{\beta})$ is the empirical kurtosis of $\boldsymbol{\beta}$ and is defined as a function of the second and fourth empirical moments of $\boldsymbol{\beta}$, $m_2(\boldsymbol{\beta})$ and $m_4(\boldsymbol{\beta})$ . As $m_2(\boldsymbol{\beta}) \xrightarrow{p} \mathbb{E}[\beta_j^2]$ and $m_4(\boldsymbol{\beta}) \xrightarrow{p} \mathbb{E}[\beta_j^4]$ as $p \to \infty$, it follows from the continuous mapping theorem that $\psi(\boldsymbol{\beta}) \xrightarrow{p} \kappa + 3$ as $p \to \infty$, where $\kappa + 3$ is the kurtosis of the distribution of elements of $\boldsymbol{\beta}$. Accordingly, we can use $\psi(\hat{\boldsymbol{\beta}}_{ols})$ directly as an estimator of the kurtosis $\kappa + 3$ when $\text{tr}((\boldsymbol{X}^\top\boldsymbol{X})^{-1})$ is small *and* $p$ is large.

When the ridge-based test statistic $\psi(\hat{\boldsymbol{\beta}}_{\delta})$ is used, estimation of $\kappa$ is less straightforward. Even if $m_2(\boldsymbol{\beta}_{\delta}) \xrightarrow{p} \mathbb{E}[m_2(\boldsymbol{\beta}_{\delta})]$ and $m_4(\boldsymbol{\beta}_{\delta}) \xrightarrow{p} \mathbb{E}[m_4(\boldsymbol{\beta}_{\delta})]$ as $p \to \infty$, the continuous mapping theorem implies $\psi(\boldsymbol{\beta}_{\delta}) \xrightarrow{p} (\gamma(\kappa + 3) + \omega)/\alpha^2$ as $p \to \infty$, where $\alpha = \text{tr}(\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{D}^2\boldsymbol{X}^\top\boldsymbol{X})/p$, $\gamma = \sum_{j=1}^{p}\sum_{k=1}^{p}(\boldsymbol{D}\boldsymbol{X}^\top\boldsymbol{X})_{jk}^4/p$, $\omega = 3((\sum_{j=1}^{p}(\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{D}^2\boldsymbol{X}^\top\boldsymbol{X})_{jj}^2)/p - \gamma)$ and $\boldsymbol{D} = \boldsymbol{V}^{-1}(\boldsymbol{C} + \delta^2\boldsymbol{I}_p)^{-1}\boldsymbol{V}^{-1}$. This suggests the follow bias correction

$$\widehat{\kappa + 3} = \left(\frac{\alpha^2}{\gamma}\right)\left(\psi(\hat{\boldsymbol{\beta}}_{\delta}) - \frac{\omega}{\alpha^2}\right). \tag{3.6}$$

Additional details are provided in the appendix.

Given an estimate of $\kappa + 3$, we estimate $q$ from (3.3), $\kappa + 3 = \Gamma(5/q)\Gamma(1/q)/\Gamma(3/q)^2$ using Newton's method.

### *3.3.2 Estimation of $\sigma^2$ and $\tau^2$*

Under the model given by (3.1), the marginal mean and variance of the data $\boldsymbol{y}$ are given by $\mathbb{E}[\boldsymbol{y}] = 0$ and $\mathbb{V}[\boldsymbol{y}] = \boldsymbol{X}\boldsymbol{X}^\top\tau^2 + \sigma^2\boldsymbol{I}_n$. We can estimate $\tau^2$ and $\sigma^2$ by solving:

$$\text{minimize}_{\tau^2,\sigma^2} \log\left(\left|\boldsymbol{X}\boldsymbol{X}^\top\tau^2 + \boldsymbol{I}_n\sigma^2\right|\right) + \text{tr}\left(\boldsymbol{y}\boldsymbol{y}^\top\left(\boldsymbol{X}\boldsymbol{X}^\top\tau^2 + \boldsymbol{I}_n\sigma^2\right)^{-1}\right). \tag{3.7}$$

Intuitively, this provides moment-based estimates of $\tau^2$ and $\sigma^2$ by minimizing a loss function relating the empirical variance $\boldsymbol{y}\boldsymbol{y}^\top$ to the variance $\boldsymbol{X}\boldsymbol{X}^\top\tau^2 + \sigma^2\boldsymbol{I}_n$ under the model (3.1), while requiring positive definiteness of $\boldsymbol{X}\boldsymbol{X}^\top\tau^2+\sigma^2\boldsymbol{I}_n$. Hoff and Yu (2017) demonstrate that these estimates will be consistent for $\tau^2$ and $\sigma^2$ as $n$ and $p \to \infty$ even if the distribution of $\boldsymbol{\beta}$ is not normal. Solving (3.7) has been treated thoroughly in the random effects literature (Demidenko, 2013). We caution that when $n < p$, the solution to (3.7) can lie on the boundary of the parameter space at $\sigma^2 = 0$.

### *3.3.3 Estimation of $\boldsymbol{\beta}$ Given $\tau^2$, $\sigma^2$ and $q$*

Given $\tau^2$, $\sigma^2$ and $q$, we can compute the posterior mode of $\boldsymbol{\beta}$ using a coordinate descent algorithm that utilizes the mode thresholding function depicted in Figure 3.1. Fu (1998) provided coordinate descent algorithms for $q \geq 1$ and Marjanovic and Solo (2014) gave a coordinate descent algorithm for $q < 1$ that is guaranteed to converge to a local minimum under certain conditions on $\boldsymbol{X}$. Details of the coordinate descent algorithm are given in the appendix. We note that when $q < 1$, the posterior mode optimization problem is not convex and the mode may not be unique.

Alternative posterior summaries, e.g. the posterior mean or median of $\boldsymbol{\beta}$ under the model given by (3.1) can be approximated using a Gibbs sampler that simulates from the posterior distribution of $\boldsymbol{\beta}$. For any value of $q > 0$ there is a uniform scale-mixture representation of the exponential power distribution (Walker and Guttiérez-Pena, 1999). If $\beta_j$ has an exponential power distribution, we can write $\beta_j|\gamma_j \sim \text{uniform}(-\Delta_j, \Delta_j)$, where $\Delta_j = \gamma_j^{1/q}\sqrt{\left(\frac{\Gamma(1/q)}{\Gamma(3/q)}\right)\left(\frac{\tau^2}{2}\right)}$ and $\gamma_j \sim \text{gamma}\left(\text{shape} = 1 + 1/q, \text{rate} = 2^{-q/2}\right)$. To our knowledge, this representation has not been used to construct a Gibbs sampler when an exponential power prior is assumed

for regression coefficients corresponding to an arbitrary design matrix $\boldsymbol{X}$. Using this representation, the full conditional distribution for $\boldsymbol{\beta}$ given $\boldsymbol{\gamma}$ is a truncated multivariate normal distribution and the full conditional distributions for elements of $\boldsymbol{\gamma}$ given $\boldsymbol{\beta}$ are independent translated exponential distributions. Full conditional distributions are given in the appendix.

### *3.3.4 Simulation Results*

We assess the performance of the adaptive procedure via simulation. We simulate data from (3.1) with $\tau^2 = \sigma^2 = 1$, $p = 100$, $n \in \{100, 200\}$ and $q \in \{0.25, 1, 4\}$ and entries of $\boldsymbol{X}$ drawn from a standard normal distribution. For each pair of values of $n$ and $q$, we simulate 100 values of $\boldsymbol{y}$ from (3.1). When $n > p$, we use 100 different design matrices, $\boldsymbol{X}$, whereas when $n \leq p$ we fix the design matrix $\boldsymbol{X}$ so that some matrix calculations involving $\boldsymbol{X}$ can be precomputed. As noted previously, when $n \leq p$ the solution to the variance component estimation problem (3.7) can lie on the boundary of the parameter space at $\sigma^2 = 0$. For the purposes of this simulation study, we require simulated datasets yield $\hat{\sigma}^2 = 0$. We use posterior means of $\boldsymbol{\beta}$ in the adaptive estimation procedure as opposed to other posterior summaries because the posterior mean minimizes posterior squared error loss and accordingly allows for straightforward performance comparisons (Hans, 2009). We approximate posterior means from $10,500$ simulations from the posterior distribution using the Gibbs sampler described in Section 3.3.3, discarding the first 500 iterations as burn-in. In general, the sampler mixes better with larger $q$ and $n$. The smallest effective sample sizes for $n = 100$ and $n = 200$ are 53 and 222, respectively.

For $n = 100$, we reject the null hypothesis that $q = 1$ at level $\alpha = 0.05$ in 100%, 1% and 62% of the simulations when $q = 0.25$, $q = 1$ and $q = 4$. Analogously, when $n = 200$ we reject the null hypothesis that $q = 1$ at level $\alpha = 0.05$ in 93%, 5% and 100% of simulations when $q = 0.25$, $q = 1$ and $q = 4$. These rejection rates are roughly as expected given the results of the simulation study of the testing procedure given that we only perform 100 simulations for each value of $n$ and $q$. Figure 3.3 shows mean squared error (MSE) for estimating $\boldsymbol{\beta}$ using the adaptive procedure plotted against the mean squared error for estimating $\boldsymbol{\beta}$ under

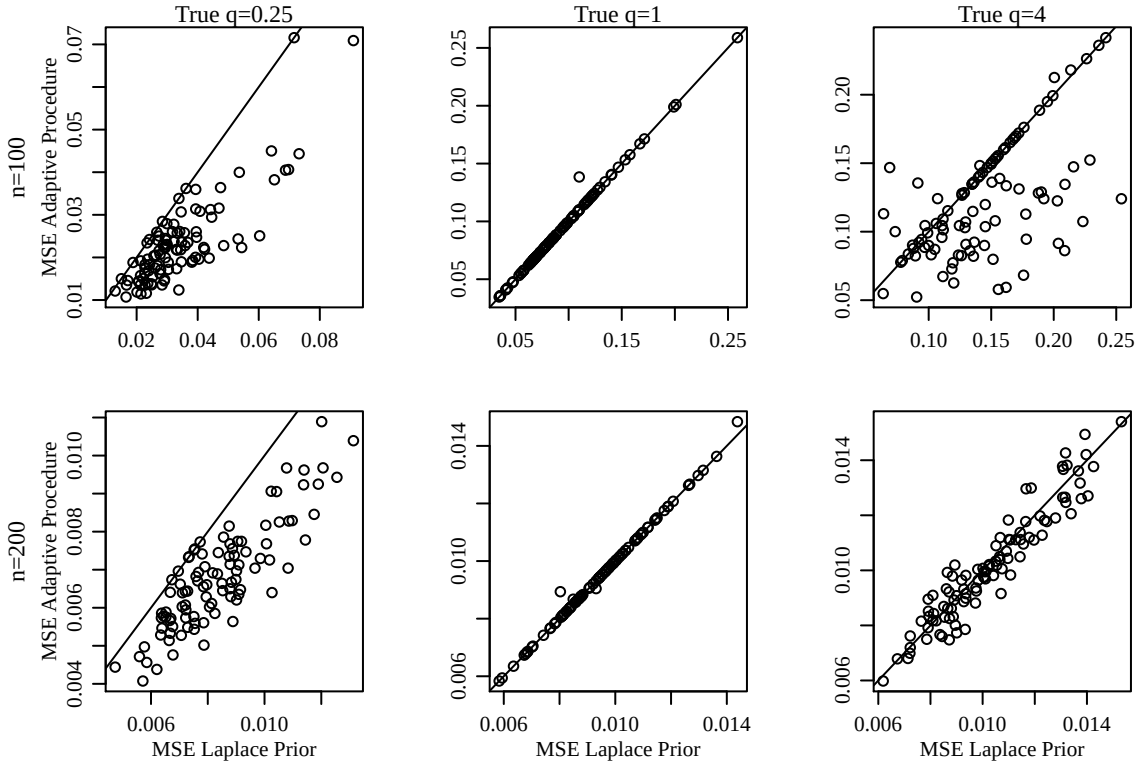



Figure 3.3: Adaptive estimation procedure versus Laplace prior performance for data simulated from model (3.1) with exponential power distributed $\boldsymbol{\beta}$ and $\sigma^2 = \tau^2 = 1$.

a Laplace prior. Histograms of estimates of $\sigma^2$, $\tau^2$ and $q$ are given in the appendix.

We see that the adaptive procedure yields substantial improvements when the true value of $q$ is small, almost no loss when the true value of $q$ is in fact equal to 1 and some small improvements and little loss when the true value of $q$ is large. Smaller improvements when the true value of $q$ is large are likely due to the fact that the estimates of $q$ are more variable when $q$ is larger. We note that incorporating testing into the adaptive procedure is important to these performance results. Recall that both tests of $H$ based on $\psi(\hat{\boldsymbol{\beta}}_{ols})$ and $\psi(\hat{\boldsymbol{\beta}}_{\delta})$ have low power when $p$ is relatively small, i.e. when little information about the features of the distribution of $\boldsymbol{\beta}$ is observed. Accordingly, incorporating testing into the estimation procedure protects us against losses in performance that could result from imprecise estimation of $q$.

### *3.4 Relationship to Estimating Sparse $\boldsymbol{\beta}$*

The lasso penalty/Laplace prior is often used when $\boldsymbol{\beta}$ is believed to be sparse, i.e. many elements of $\boldsymbol{\beta}$ are believed to be equal to exactly zero. Accordingly, we repeat the testing and estimation simulation studies performed in the previous sections for Bernoulli-normal spike-and-slab distributed $\boldsymbol{\beta}$ where $\beta_j$ is exactly equal to zero with probability $1-\pi$ and drawn from a $N\left(0, \tau^2/\pi\right)$ distribution otherwise. This parametrization ensures that elements of $\boldsymbol{\beta}$ have variance $\tau^2$. The kurtosis of this distribution is given by $\kappa + 3 = 3/\pi$ and when $\pi = 0.5$, the kurtosis of this distribution matches that of a Laplace distribution. We repeat the testing simulation study in Section 3.2 for a range of values of $\pi$ instead of $q$ and show the results in Figure 3.4.

As expected, our tests tend not to reject $H : q = 1$ when the kurtosis of the spike-and-slab distribution is similar to the kurtosis of the Laplace prior at $\pi = 0.5$. Importantly, our tests reject a Laplace prior when $\pi$ is small and $\boldsymbol{\beta}$ is very sparse. This suggests that the adaptive procedure for estimating $\boldsymbol{\beta}$ might yield performance improvements even when elements of $\boldsymbol{\beta}$ do not have an exponential power distribution. We repeat the estimation simulations described in Section 3.3.4 for $\pi \in \{0.1, 0.5, 0.9\}$ and show the results in Figure 3.5. Again,

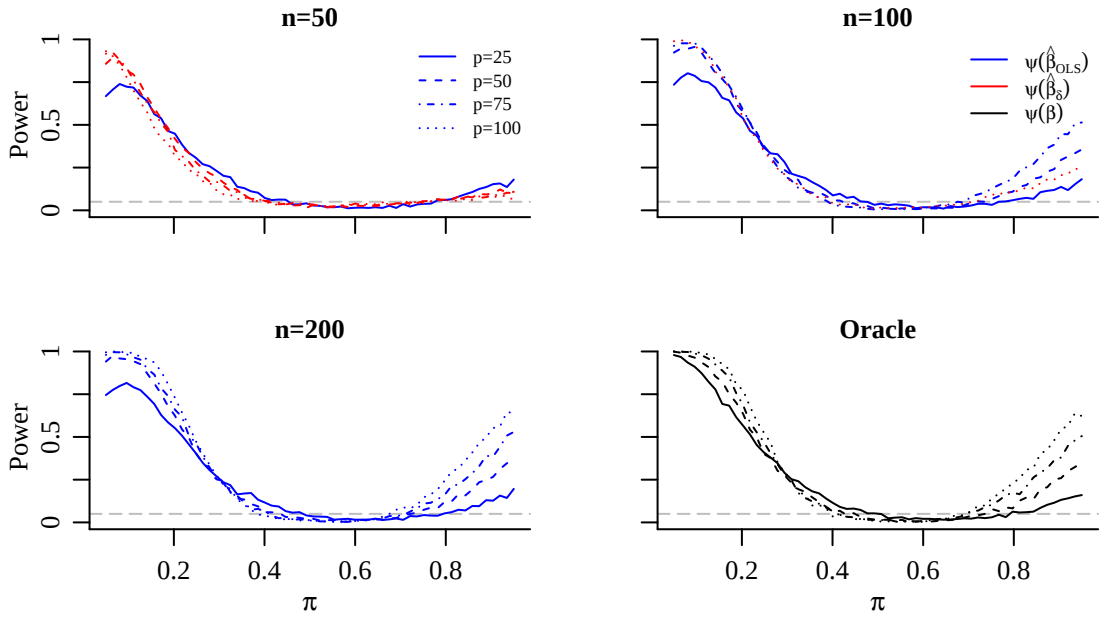



Figure 3.4: Power of level-0.05 tests for data simulated from a linear regression model with standard normal errors and Bernoulli-normal regression coefficients with sparsity rate $1 - \pi$ and unit variance. A horizontal dashed gray line is given at 0.05.

histograms of estimates of $\sigma^2$, $\tau^2$ and $q$ are given in the appendix. Before discussing the results of the simulations, we note that as we might expect based on the previous simulations the sampler mixes better with larger values of $\pi$ and $n$. The smallest effective sample sizes for $n = 100$ and $n = 200$ are 70 and 253, respectively.

With spike-and-slab distributed $\boldsymbol{\beta}$, the adaptive procedure still outperforms estimates based on the Laplace prior in the majority of simulations. We see substantial performance gains from the adaptive procedure when $\pi = 0.1$ for both $n = 100$ and $n = 200$. Again, we observe some losses in performance when $\pi = 0.9$, i.e. when the kurtosis is relatively low and estimates of $q$ are more variable. Again, we emphasize that incorporating the test into the adaptive procedure does play an important role in its performance. When $\pi = 0.9$ and results of a test of $H$ are ignored, the mean squared error for estimating $\boldsymbol{\beta}$ using $q = \hat{q}$ exceeds

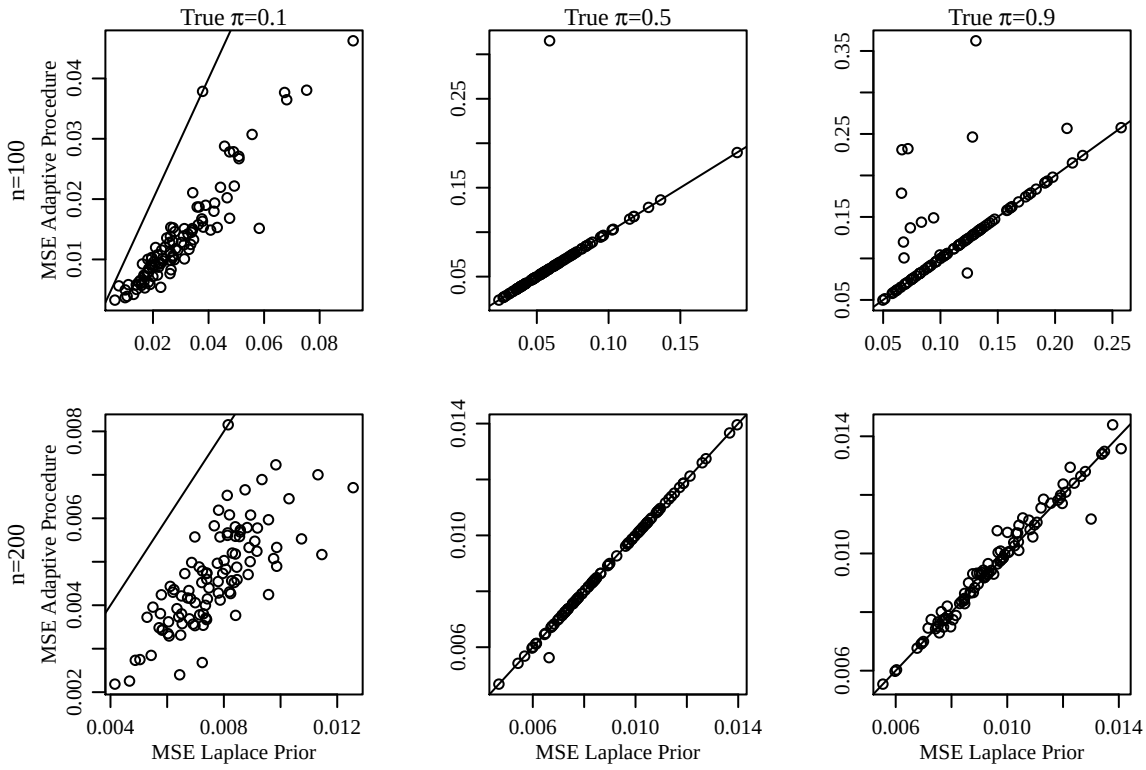



Figure 3.5: Adaptive estimation procedure versus Laplace prior performance for data simulated from a linear regression model with standard normal errors and Bernoulli-normal regression coefficients with sparsity rate $1 - \pi$ and unit variance.

the mean squared error using a Laplace prior in 56% and 67% of simulations when $n = 100$ and $n = 200$, respectively. When the exponential power prior is only used when a test of $H$ rejects and a Laplace prior is used otherwise, this drops to 12% and 34%, respectively.

### *3.5 Applications*

We apply the methods described in this chapter to four datasets that have appeared previously in the penalized regression literature: the diabetes data, the Boston housing data, motif data and glucose data (Efron et al., 2004; Park and Casella, 2008; Polson et al., 2014; Bühlmann and van de Geer, 2011; Priami and Morine, 2015). The diabetes data featured in Efron et al. (2004), Park and Casella (2008) and Polson et al. (2014) contains a quantitative measure of diabetes progression for $n = 442$ patients $\boldsymbol{y}$ and ten covariates: age, sex, body mass index, average blood pressure and six blood serum measurements. A design matrix $\boldsymbol{X}$ is obtained from the ten original covariates, $\binom{10}{2}$ pairwise interactions and 9 quadratic terms yielding $p = 64$. In the Boston housing data (Polson et al., 2014), the response vector $\boldsymbol{y}$ is the median house price for $n = 506$ Boston census tracts and the design matrix is made up of 13 measurements of census tract characteristics and all $\binom{13}{2}$ squared terms and pairwise interactions, yielding $p = 104$. The motif data featured in Bühlmann and van de Geer (2011) contains measurements of protein binding intensity $\boldsymbol{y}$ at $n = 287$ regions of the DNA sequence and $p = 195$ covariates $\boldsymbol{X}$ made up of measurements of motif abundance for $p$ motifs at each region. The glucose data from Priami and Morine (2015) contains measurements of blood glucose concentration $\boldsymbol{y}$ for $n = 68$ subjects belonging to several families with complete data on $p = 72$ covariates, which include various metabolite measurements along with several health indicators. We subtract an overall mean and family-specific group means off of the response and the design matrix containing the 72 covariates to be used for regression.

For all four data sets, we centered and standardized the response $\boldsymbol{y}$ and the columns of the design matrix $\boldsymbol{X}$ by subtracting off their means and dividing by their standard deviations. We use the ridge-based test for all four data sets because either $n < p$ or the design matrix is highly collinear and $\boldsymbol{X}^\top \boldsymbol{X}$ has condition number less than $10^{-5}$. As in the simulations shown

previously, we perform level-$\alpha = 0.05$ tests and approximate the corresponding quantiles $\psi_{0.025}$ and $\psi_{0.975}$ by simulating $1,000,000$ draws from the approximate distribution of the test statistic under the null. Table 3.1 summarizes the features of the data and the test results.

| **Dataset** | $n$ | $p$ | $\psi_{\delta,0.025}$ | $\psi_{\delta,0.975}$ | $\psi(\hat{\boldsymbol{\beta}}_\delta)$ | $\Pr\left(\psi(\boldsymbol{\beta}_\delta) \leq \psi(\hat{\boldsymbol{\beta}}_\delta) \mid q = 1\right)$ |
|---|---|---|---|---|---|---|
| Diabetes | 422 | 64 | 2.31 | 7.68 | 10.36 | 0.993 |
| Boston Housing | 506 | 104 | 1.97 | 7.59 | 6.57 | 0.959 |
| Motif | 287 | 195 | 2.87 | 10.34 | 5.77 | 0.748 |
| Glucose | 68 | 72 | 2.31 | 7.05 | 9.38 | 0.993 |

Table 3.1: Results of testing the appropriateness of a Laplace prior for four datasets.

We reject the null hypothesis that a Laplace prior is appropriate for the diabetes and glucose data sets. For these two data sets, we estimate $\sigma^2$, $\tau^2$ and $q$ and compute the posterior modes and means of $\boldsymbol{\beta}$ under exponential power and Laplace priors. When computing the posterior modes, we address nonconvexity when $q < 1$ by repeating the coordinate descent algorithm for 100 randomly selected starting values and saving the estimate that gives the greatest posterior likelihood. Again, we caution that a unique posterior mode may not exist when $\boldsymbol{X}$ is so highly collinear or not full rank. We approximate posterior means using $1,000,500$ draws from each posterior distribution, discarding the first 500 iterations as burn-in and thinning the remaining $1,000,000$ samples by a factor of 20.

Table 3.2 summarizes the variance and shape parameter estimates, mode sparsity rates and effective sample sizes for both datasets and priors. For both data sets, estimates of the shape parameter $\hat{q}$ are less than 1, suggesting that an even heavier tailed prior is more appropriate. Accordingly, the exponential power prior yields a sparser posterior mode than the Laplace prior. Mixing of the Gibbs samplers used to approximate the posterior means is better when a Laplace prior is used.

| | Par. Ests. | | | Mode Sparsity | | Min. ESS | |
|---|---|---|---|---|---|---|---|
| **Dataset** | $\hat{\sigma}^2$ | $\hat{\tau}^2$ | $\hat{q}$ | $L$ | $EP$ | $L$ | $EP$ |
| Diabetes | 0.4708 | 0.0071 | 0.5505 | 50.0% | 87.5% | 21,988 | 3,976 |
| Glucose | 0.4460 | 0.0077 | 0.5509 | 80.6% | 95.8% | 5,545 | 782 |

Table 3.2: Variance and shape parameter estimates, posterior mode sparsity rates and minimum effective sample sizes of posterior samples under Laplace (L) and exponential power (EP) priors.

Figure 3.6 compares posterior modes, means and selected marginal distributions under Laplace and exponential power priors for $\boldsymbol{\beta}$. Examining the posterior modes, we observe not only higher sparsity rates but also less shrinkage of nonzero values when the exponential power prior is used. We observe similar but less stark differences when comparing posterior means across both priors. We also compare the marginal posterior distributions for several elements of $\boldsymbol{\beta}$, chosen to demonstrate how using an exponential power prior affects inference for these datasets. In the right four panels of Figure 3.6, we see that using the exponential power prior can cause the mode of the marginal posterior distribution to change locations or can introduce bimodality of the marginal posterior distribution. Overall, we gain more interpretable estimates of $\boldsymbol{\beta}$ with fewer large entries by using a more appropriate prior.

### *3.6 Discussion*

In this chapter, we have introduced a simple procedure for testing the null hypothesis that a Laplace prior is appropriate by assessing whether or not the kurtosis of the distribution of unknown regression coefficients matches that of a Laplace distribution. We also introduce two-step adaptive estimation procedure for $\boldsymbol{\beta}$ that uses an exponential power prior for $\boldsymbol{\beta}$ if a Laplace prior is rejected. We show that our testing and estimation procedures perform well for the kinds of values of $n$ and $p$ we might encounter in practice both when elements of $\boldsymbol{\beta}$

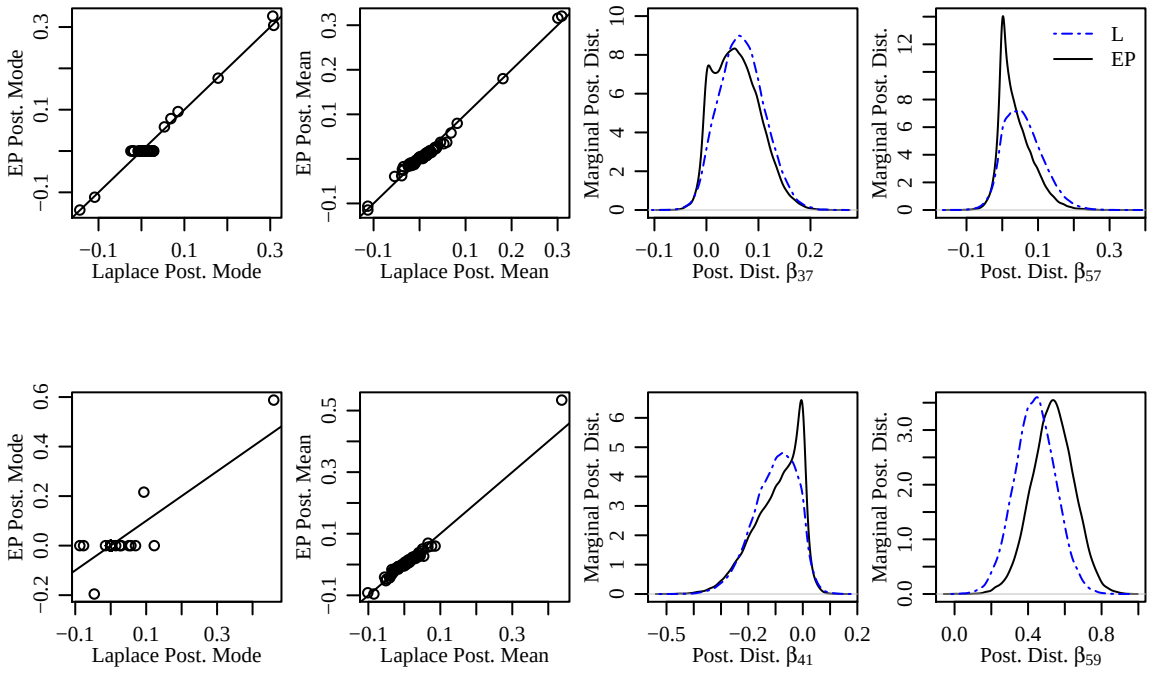



Figure 3.6: Posterior modes, means and selected marginal distributions under exponential power (EP) priors and Laplace (L) priors of $\boldsymbol{\beta}$ for diabetes and glucose datasets.

have an exponential power distribution and when they are sparse with a Bernoulli-normal spike-and-slab distribution. We have demonstrated that the appropriateness of a Laplace prior for estimating Bernoulli-normal spike-and-slab $\boldsymbol{\beta}$ depends on the sparsity rate and that estimates based on a Laplace prior can be suboptimal when we expect that $\boldsymbol{\beta}$ follow a spike-and-slab distribution with a high sparsity rate. As dependence of kurtosis on the sparsity rate is not limited to the Bernoulli-normal spike-and-slab distribution but rather extends to any spike-and-slab distribution where the slab is a mean zero distribution with finite fourth moments, we expect that the performance improvements we observe might persist for more general sparsely distributed $\boldsymbol{\beta}$. This complements the existing statistical literature on the suboptimality of the Laplace prior (Griffin and Brown, 2010; Carvalho et al., 2010; Polson

et al., 2014; Bhattacharya et al., 2015).

This chapter focuses specifically on misspecification of the Laplace prior for the regression coefficients in a linear model with normal errors. Because the derivation of the approximate level-$\alpha$ test follows from the existence of a consistent estimator of $\boldsymbol{\beta}$ or a known linear function of $\boldsymbol{\beta}$, the methods described in this chapter can be extended to include linear models with elliptically contoured errors and generalized linear models. The methods described in this chapter could also be extended to include the construction of a confidence interval either for the kurtosis of the distribution of the elements of $\boldsymbol{\beta}$ or the exponential power shape parameter $q$. Furthermore, the methods can be generalized to test the null hypothesis that elements of $\boldsymbol{\beta}$ have an exponential power distribution with $q = \tilde{q} > 0$ or to test a null hypothesis that elements of $\boldsymbol{\beta}$ have a different symmetric, mean zero distribution as long as this different distribution can be characterized by its kurtosis and is easy to simulate from, e.g. the normal-gamma distribution given by Griffin and Brown (2010) or the Dirichlet-Laplace distribution given by Bhattacharya et al. (2015).

Throughout this chapter we have conflated heavy tails (high kurtosis) with peakedness of the density of $\boldsymbol{\beta}$. However, it is not generally true that "peakedness" must increase with tail weight (Westfall, 2014). This is important because sparsity of the posterior mode specifically arises from the "peakedness" of the prior on $\boldsymbol{\beta}$. Accordingly, three parameter distributions like the generalized $t$-distribution given by Choy and Chan (2008) that allow kurtosis and "peakedness" to vary separately may be useful alternative priors for $\boldsymbol{\beta}$.

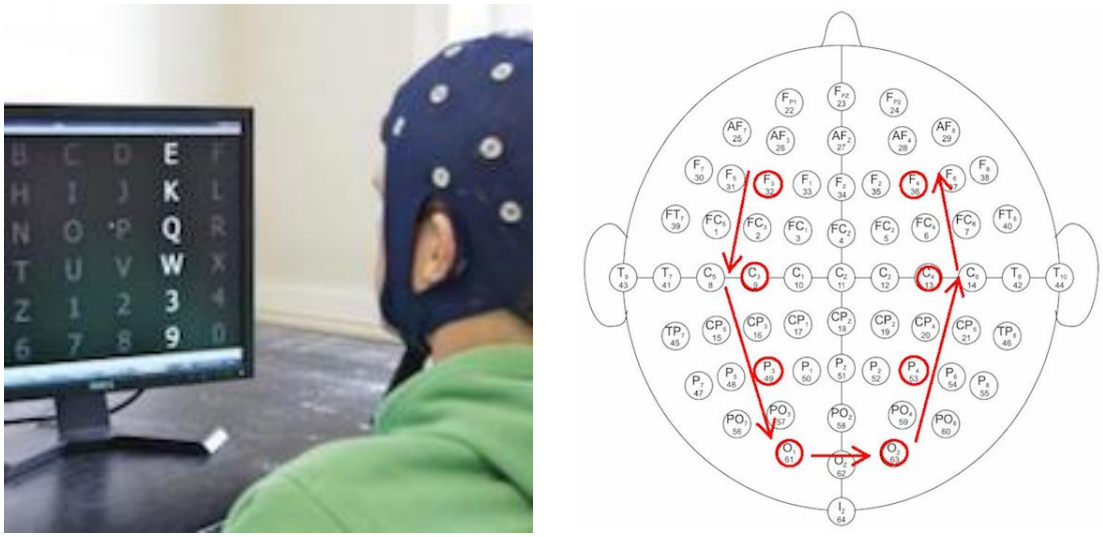
# Chapter 4

# STRUCTURED SHRINKAGE PRIORS

## *4.1 Introduction*

Consider the problem of estimating regression coefficients for a generalized linear model, where we observe an $n \times 1$ vector of responses $\boldsymbol{y}$ and an $n \times p$ matrix of regressors $\boldsymbol{X}$ and are interested in estimating a $p \times 1$ vector of regression coefficients $\boldsymbol{\beta}$ which are related to $\boldsymbol{y}$ via a known link function $\mathbb{E}[\boldsymbol{y}] = g(\boldsymbol{X}\boldsymbol{\beta})$ and a probabilistic model for $\boldsymbol{y}$ given $\mathbb{E}[\boldsymbol{y}]$ and unknown distributional parameters $\boldsymbol{\phi}$. Letting $\exp\{-h(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{\phi})\}$ refer to the likelihood of the data $\boldsymbol{y}$ given $\boldsymbol{X}\boldsymbol{\beta}$ and $\boldsymbol{\phi}$, maximum likelihood estimates of $\boldsymbol{\beta}$ given $\boldsymbol{\phi}$ are obtained by minimizing $h(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{\phi})$ over $\boldsymbol{\beta}$. This is a standard, well known problem (McCullagh and Nelder, 1989). However, when the the data is high dimensional, i.e. number of covariates $p$ is large, especially relative to the number of responses $n$, inference and prediction can be challenging. Standard maximum likelihood estimates of $\boldsymbol{\beta}$ may have prohibitively large variance or fail to be unique. In such high dimensional settings, a common approach is to assume a prior distribution for $\boldsymbol{\beta}$. For instance, a mean-zero multivariate normal prior with covariance matrix $\boldsymbol{\Sigma}$, $\boldsymbol{\beta} \sim \text{normal}(\boldsymbol{0}, \boldsymbol{\Sigma})$ might be chosen to encourage a *structured* estimate of $\boldsymbol{\beta}$, or a mean-zero independent Laplace prior with variance $\sigma^2$, $\beta_j \overset{i.i.d.}{\sim} \text{Laplace}\left(0, \sigma/\sqrt{2}\right)$, might be chosen to encourage a *sparse* or *nearly sparse* estimate of $\boldsymbol{\beta}$ depending on which posterior summary is used to estimate $\boldsymbol{\beta}$. Using the posterior mode to estimate $\boldsymbol{\beta}$ corresponds to penalized estimation of $\boldsymbol{\beta}$.

In this paper, we consider problems where the covariate vectors $\boldsymbol{x}_i$, and accordingly the regression coefficients $\boldsymbol{\beta}$, can be interpreted as a vectorized $K$-mode matrix or tensor with dimensions $p_1 \times \ldots p_K$. As such, $\boldsymbol{\beta} = \text{vec}(\boldsymbol{B})$ is *structured*. Because such problems are often very high dimensional with large $p = \prod_{k=1}^{K} p_k$ relative to $n$, we focus specifically on



(a) (b)

Figure 4.1: Panel (a) reprinted from Bruneel (2018) shows a subject using the P300 speller. Panel (b) shows the locations of EEG sensors on the skull reprinted and modified from Sharma (2013), with sensors included in our analysis highlighted in red. Arrows indicate the order of the channels appear in the data.

settings where we also expect $\boldsymbol{B}$ to be *sparse* or *nearly sparse*. We motivate this with a problem that arises in the development of brain-computer interfaces. Brain-computer interfaces (BCIs) are used to detect changes in subjects' cognitive state from contemporaneous electroencephalography (EEG) measurements, which can be collected non-invasively at high temporal resolution (Makeig et al., 2012; Wolpaw and Wolpaw, 2012). We consider the P300 speller, a specific BCI device which is designed to detect when a subject is viewing a specified target letter, as depicted in use in Figure 4.1a. We consider a subset of data collected from fourteen subjects using a P300 speller and a gGAMMA.sys EEG device Forney et al. (2013). For an individual subject, we observe 240 indicators $y_i$ for whether or not the subject was viewing a specified target letter during trial $i$ and $\boldsymbol{x}_i$ is a vectorized $208 \times 8$ matrix of EEG

measurements from $p_1 = 208$ time points and $p_2 = 8$ channels, which correspond to the eight distinct physical locations on the top of the skull shown in Figure 4.1b. In this paper, we set aside the last 100 trials to assess predictive performance and use a subset of $n = 140$ trials for exploratory analysis and estimation. This data is clearly high-dimensional; $\boldsymbol{X}$ has $p_1 p_2 = 1,664$ columns which vastly outnumber the $n = 140$ observations.

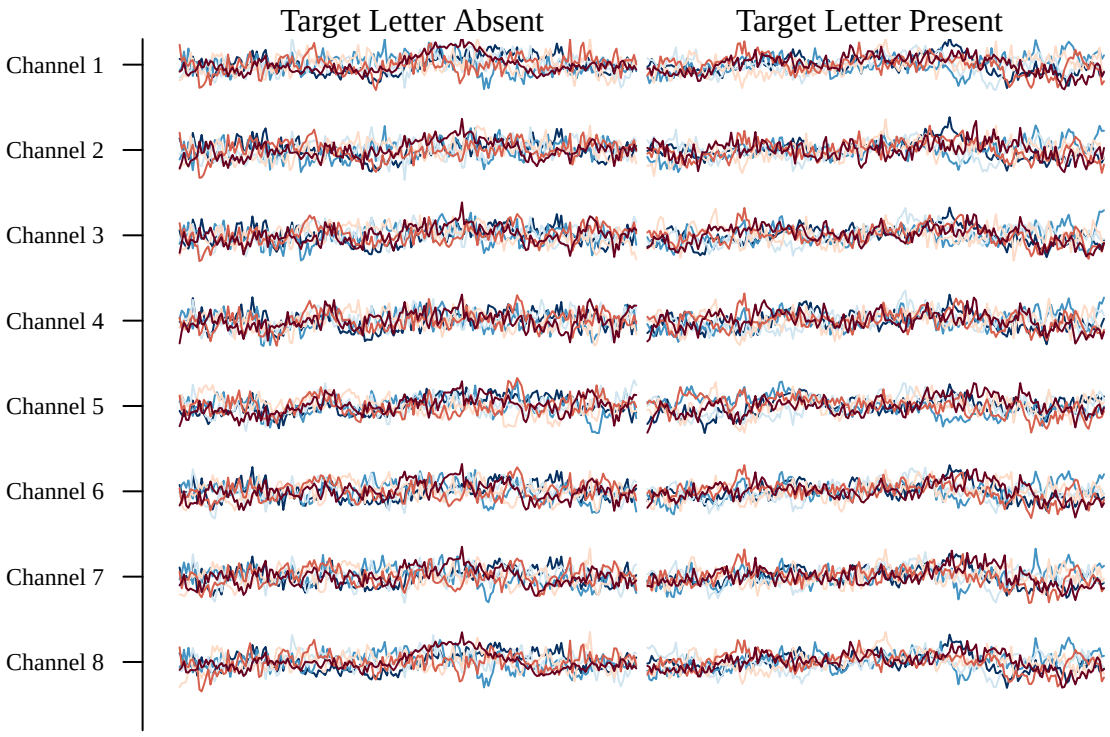



Figure 4.2: A subset of single-subject P300 speller data. Lines represents trials, i.e. rows $\boldsymbol{x}_i$. Trials are plotted separately by whether or not the target letter was being shown during the trial.

Figure 4.2 shows data from a single subject and a subset of trials. Scientifically, we expect to observe a P300 wave during trials when the target letter is present, which is characterized by a sharp rise and then dip before returning to equilibrium. We expect that the wave will begin shortly after the target letter is shown, and will be observed earlier and more clearly

on some channels than others. That said, EEG data is notoriously noisy and the P300 wave is difficult to observe in the data. Fortunately, the scientific context suggests that it is reasonable to assume that the unobserved regression coefficients $\boldsymbol{B}$ should be *sparse*, as only elements of $\boldsymbol{B}$ that correspond to time points where the P300 wave occurs are expected to be nonzero, and *structured*, as elements of $\boldsymbol{B}$ corresponding to consecutive time points or neighboring channels are expected to be similar.

To assess whether or not these assumptions might be reasonable, we perform exploratory analysis of a ridge estimate $\hat{\boldsymbol{\beta}}_R = \text{vec}\left(\hat{\boldsymbol{B}}_R\right)$ of logistic regression coefficients relating $n = 140$ indicators of the presence of the target letter $\boldsymbol{y}$ to a $n \times p = 140 \times 1,664$ design matrix of EEG measurements $\boldsymbol{X}$ which encode the presence or absence of the P300 wave. The first panel of Figure 4.3, which shows values of elements of $\hat{\boldsymbol{B}}_R$, indicates that the assumption of structure is reasonable. Elements of $\hat{\boldsymbol{B}}_R$ corresponding to consecutive time points or neighboring channels tend to be relatively more similar to each other. The second panel of Figure 4.3 shows a histogram of elements of $\hat{\boldsymbol{B}}_R$ compared to a mean zero normal density with variance $\frac{1}{p_1 p_2}\sum_{i}^{p_1}\sum_{i_2=1}^{p_2} b_{R,i_1 i_2}^2$. We observe some evidence of possible sparsity of $\boldsymbol{B}$, as slightly more elements of $\hat{\boldsymbol{B}}_R$ are nearly equal to zero than would be expected if $\hat{\boldsymbol{B}}_R$ were normally distributed with the same mean and variance. In the third panel of Figure 4.3, we examine similarity of elements of $\hat{\boldsymbol{B}}_R$ corresponding to consecutive time points and observe clear evidence of temporal structure. Lastly, in the fourth panel of Figure 4.3 we see clear evidence of positive correlations of varying strengths across channels.

Unfortunately, few prior distributions are able to encourage *both* structure *and* near or exact sparsity. This is easiest to see by considering the normal scale-mixture representations of various common prior distributions for $\boldsymbol{\beta}$. Letting '$\circ$' be the Hadamard elementwise product, $\boldsymbol{z} \sim \text{normal}(\boldsymbol{0}, \boldsymbol{\Omega})$ and $\boldsymbol{s}$ be a vector of stochastic "scales" that are independent of $\boldsymbol{z}$, a prior distribution for $\boldsymbol{\beta}$ has a normal scale-mixture representation if there exists a density $p\left(\boldsymbol{s}|\boldsymbol{\theta}\right)$ such that $\boldsymbol{\beta} \overset{d}{=} \boldsymbol{s} \circ \boldsymbol{z}$. There is an extensive literature on normal scale-mixture representations of various distributions (Andrews and Mallows, 1974; West, 1987), especially normal scale-mixture representations of common shrinkage priors (Polson and Scott, 2010).

Structure can be encouraged through the choice of $\boldsymbol{\Omega}$, and shrinkage can be encouraged through the choice of the distribution of the scales $\boldsymbol{s}$ and the scale distribution parameters $\boldsymbol{\theta}$.

Independent shrinkage priors that correspond to separable penalties are obtained by setting $\boldsymbol{\Omega} = \omega^2 \boldsymbol{I}_p$. The Laplace prior is obtained by assuming independent and identical exponential distributions for the squared scales $s_j^2$ and exponential power priors which correspond to $\ell_q$ penalties when $0 < q < 2$ are obtained by assuming that squared scales $s_j^2$ have polynomially tilted positive $\alpha = q/2$ stable distributions (Figueiredo, 2003; Devroye, 2009; Polson et al., 2014).

Elliptically contoured priors, including the prior that corresponds to a group lasso penalty on $\boldsymbol{\beta}$ or the multivariate Laplace prior used by van Gerven et al. (2009) are obtained by assuming that elements of $\boldsymbol{s}$ are constant with $\boldsymbol{s} = s\mathbf{1}_p$. These priors have computational advantages, as the corresponding marginal prior distributions for $\boldsymbol{\beta}$ and penalties depend on $\boldsymbol{\beta}$ strictly through $\boldsymbol{\beta}'\boldsymbol{\Omega}^{-1}\boldsymbol{\beta}$. However as these priors only have a single shrinkage factor $s$, they shrink all elements of $\boldsymbol{\beta}$ jointly and posterior modes based on these priors will only produce sparse estimates of $\boldsymbol{\beta}$ for which all entries are exactly equal to zero. Furthermore, they do not generalize their independent counterparts, e.g. the multivariate Laplace prior with $\boldsymbol{\Omega} = \omega^2 \boldsymbol{I}_p$ which corresponds to the group lasso penalty is *not* equivalent to the independent Laplace prior which corresponds to the lasso penalty. While this may be appropriate in some settings, it will not be appropriate in settings where we are confident that only *some* elements of $\boldsymbol{\beta}$ are nearly or exactly sparse but unsure about the presence of structure.

Naively, structure and exact or near sparsity could be simultaneously encouraged by assuming a sparsity inducing prior for elements of $\boldsymbol{s}$ and allowing $\boldsymbol{\Omega}$ to be non-diagonal. However, this tends to yield intractable marginal prior distributions for $\boldsymbol{\beta}$ that correspond to integral penalties $-\log\left(\int p\left(\boldsymbol{\beta}|\boldsymbol{\Omega}, \boldsymbol{s}\right) d\boldsymbol{s}\right)$ that require computationally demanding and non-standard Markov Chain Monte Carlo (MCMC) based inference on $\boldsymbol{\beta}$. As a result, the development of such priors has been largely ignored. An exception is Finegold and Drton (2011), which develops a multivariate $t$-distribution by assuming that that squared scales $s_j^2$

are independent and identically inverse-gamma distributed and $\boldsymbol{\Omega}$ is non-diagonal.

Instead, several structured shrinkage priors have been constructed to encourage structure and exact or near sparsity while retaining computational tractability by assuming that $\boldsymbol{\Omega} = \omega^2 \boldsymbol{I}_p$ but elements of $\boldsymbol{s}$ are dependent (van Gerven et al., 2010; Kalli and Griffin, 2014; Wu et al., 2014; Zhao et al., 2016; Kowal et al., 2017). van Gerven et al. (2010) introduces structured auxiliary variables $\boldsymbol{u}, \boldsymbol{v} \sim \text{normal}\left(\boldsymbol{0}, \boldsymbol{\Xi}\right)$ and assumes $s_j^2 = u_j^2 + v_j^2$, Wu et al. (2014) assumes a hierarchical prior for elements $s_j^2$ that reflects a pre-specified group structure of $\boldsymbol{\beta}$, Zhao et al. (2016) introduces an auxiliary Gaussian process with a squared exponential kernel $\boldsymbol{u} \sim \text{normal}\left(b\boldsymbol{1}_p, \boldsymbol{K}\left(\rho, l\right)\right)$ and assumes $s_j^2 = \exp\left\{u_j\right\}$ and Kalli and Griffin (2014) and Kowal et al. (2017) assume that $\log\left(s_j^2\right)$ are distributed according to an autoregressive process with parameter $\phi$ where $\log\left(s_{j+1}^2\right) = \phi \log\left(s_j^2\right) + \eta_j$. Kalli and Griffin (2014) assume that $\eta_j$ are independent and $\exp\left\{\eta_j\right\}$ are identically distributed according to a gamma distribution and Kowal et al. (2017) assume that $\eta_j$ are independent and identically distributed according to a $Z$-distribution. These priors all share the same drawback, which is that although they introduce structure via the distribution of elements of $\boldsymbol{s}$, the prior marginal variance-covariance matrix of $\boldsymbol{\beta}$ is diagonal, i.e. elements of $\boldsymbol{\beta}$ are uncorrelated a priori.

Some other priors that have been constructed to encourage structure and exact or near sparsity treat elements of $\boldsymbol{\Omega}$ as random. As described in Kyung et al. (2010), the prior distribution that yields the fused lasso penalty $\lambda_1 \left|\left|\boldsymbol{\beta}\right|\right|_1 + \lambda_2 \sum_{i-j=1} \left|\beta_j - \beta_i\right|$ can be obtained by assuming $s_1^{-2} = r_1^{-2} + t_{12}^{-2}$, $s_j^{-2} = r_j^{-2} + t_{j-1,j}^{-2} + t_{j,j+1}^{-2}$ for $1 < j < p$ and $s_p^{-2} = r_p^{-2} + t_{p-1,p}^{-2}$, where $r_j^2$ are independent and identically exponentially distributed with rate parameter $\lambda_r$ and $t_{ij}^2$ are independently and identically exponentially distributed with rate parameter $\lambda_t$, and setting $\omega^{jj} = 1$ and $\omega^{ij} = -t_{ij}^{-2}$ if $\left|i - j\right| = 1$ and $\omega^{ij} = 0$ otherwise, where $\omega^{ij}$ refers to the element in the $i$-th row and $j$-th column of $\boldsymbol{\Omega}^{-1}$. Priors that correspond to more general structured penalties of the form $\boldsymbol{\beta}'\boldsymbol{Q}^{-1}\boldsymbol{\beta} + \lambda \left|\left|\boldsymbol{\beta}\right|\right|_1$ where $\boldsymbol{Q}^{-1}$ is positive semidefinite can be obtained similarly by letting $\boldsymbol{A} = \left(\boldsymbol{Q}^{-1} + \text{diag}\left\{\boldsymbol{r}^{-2}\right\}\right)^{-1}$, where $r_j^2$ are independently distributed according to an exponential distribution with rate $\lambda$, and setting $\boldsymbol{s}^2 = \text{diag}\left\{\boldsymbol{A}\right\}$ and $\boldsymbol{\Omega}$ to the correlation matrix of $\boldsymbol{A}$. These priors are certainly useful and are arguably the

most reasonable set of existing priors for encouraging structure and near or exact sparsity. However, priors like the prior corresponding to the fused lasso penalty become much more difficult to work with when the structure is not pre-specified, i.e. orderings or groupings of elements of $\boldsymbol{\beta}$ are not known a priori, and can suffer from computational inefficiencies as the number of latent variables increases with the complexity of the structure of $\boldsymbol{\beta}$ if posterior summaries other than the posterior mode are of interest or maximum marginal likelihood estimates of the unknown prior distribution parameters are to be obtained. Furthermore, the unknown parameters for priors corresponding to the more general structured penalty or more structured generalizations of the fused lasso penalty can be difficult to set or estimate as they become higher dimensional, especially because they do not relate straightforwardly to prior moments of $\boldsymbol{\beta}$. One last class of relevant priors are the structured normal-gamma priors introduced in Griffin and Brown (2012b) and Griffin and Brown (2012a), which are obtained by introducing structure via a $p \times p$ matrix $\boldsymbol{C}$ and setting $s_j = c_{jj} r_j$ and $\left(\boldsymbol{\Omega}^{1/2}\right)_{ij} = c_{ij} r_j / \left(c_{ii} r_i\right)$, where $r_j^2$ are independent and identically gamma distributed. Like the priors previously discussed, estimation or specification of $\boldsymbol{C}$ can be challenging. Furthermore, many of the desirable properties of this prior are specific to the gamma distribution assumption for $r_j^2$.

Ideally, we would like to work with a class of structured shrinkage priors that (i) can encourage general forms of structure by allowing for correlations across elements of $\boldsymbol{\beta}$ a priori, (ii) can encourage near or exact sparsity of elements of $\boldsymbol{\beta}$, (iii) yield computationally tractable posterior mode and posterior distribution simulation problems and (iv) span the range of common structured and sparsity inducing priors by generalizing the multivariate normal prior as well as the independent Laplace prior. We construct such a class of priors by revisiting the naive approach where we assume $\boldsymbol{\beta} \stackrel{d}{=} \boldsymbol{s} \circ \boldsymbol{z}$ and non-diagonal $\boldsymbol{\Omega}$. These priors are easy to interpret from a data generating perspective and have easy-to-compute moments, specifically $\boldsymbol{\Sigma} = \mathbb{V}\left[\boldsymbol{\beta}\right] = \mathbb{E}\left[\boldsymbol{s}\boldsymbol{s}'\right] \circ \Omega$. We demonstrate that elliptical slice sampling can be used to overcome the computational issues that have previously made use of such prior distributions challenging, regardless of the distribution assumed for elements of $\boldsymbol{s}$ (Murray

et al., 2010). In Section 4.2, we introduce three novel priors, the structured normal product (SPN), structured normal-gamma (SNG) and structured power/bridge (SPB) shrinkage prior distributions for penalized regression. The SPN and SNG priors generalize the normal-gamma prior described in Caron and Doucet (2008) and Griffin and Brown (2010), and the SPB prior generalizes the power/bridge prior (Frank and Friedman, 1993; Polson et al., 2014). Figure 4.4 shows how the novel structured shrinkage priors we introduce relate to Laplace and normal prior distributions. We discuss the properties of these priors in Section 4.2. In Section 4.3, we describe posterior inference as it pertains to performing Gibbs-within-EM estimation of posterior modes and simulating from the full posterior distribution. Because problems that benefit from prior distributions that encourage both structure and exact or near sparsity are often very high dimensional, i.e. have large $n$ or $p = \prod_{k=1}^{K} p_k$, we show how elliptical slice sampling can be used to avoid large matrix crossproducts and inversions for arbitrary log likelihoods $h\left(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{\phi}\right)$. In Section 4.4, we consider the unknown variance components and discuss maximum marginal likelihood empirical Bayes estimation, moment-based empirical Bayes estimation and fully Bayes inference. In Section 4.5, we apply the shrinkage priors considered to the P300 speller data depicted in Figures 4.2 and 4.3.

## *4.2 Three Structured Shrinkage Priors and Their Properties*

### *4.2.1 Prior Distributions*

**Structured Product Normal (SPN) Prior:** The structured product normal (SPN) prior is obtained by assuming that $\boldsymbol{s} \sim \text{normal}(\boldsymbol{0}, \boldsymbol{\Psi})$. Originally discussed as a sparsity inducing penalty in Hoff (2016), the SPN prior is uniquely computationally simple to work with as the full conditional distributions of $\boldsymbol{z}$ and $\boldsymbol{s}$ are both multivariate normal distributions. The SPN prior is also appealing intuitively as it allows elements of $\boldsymbol{s}$ to be correlated a priori, in contrast to the SNG and SPB priors which assume elements of $\boldsymbol{s}$ are independent. Incidentally, the SPN prior generalizes a special case of the independent normal gamma prior (Griffin and Brown, 2010). When $\boldsymbol{\Psi}, \boldsymbol{\Omega} \propto \boldsymbol{I}_p$, the SPN prior corresponds to the independent

normal-gamma prior with $s_j^2 \sim \text{gamma}(\text{shape} = 1/2, \text{rate} = 1/2)$.

The SPN prior has $2p + p\,(p-1)$ unknown variance parameters, of which $p + p\,(p-1)\,/2$ are separately identified. Letting $\boldsymbol{C}_\Omega$ and $\boldsymbol{C}_\Psi$ be the correlation matrices corresponding to $\boldsymbol{\Omega}$ and $\boldsymbol{\Psi}$, we can write $\boldsymbol{\Omega} = \text{diag}\left\{\sqrt{\boldsymbol{\omega}}\right\}\boldsymbol{C}_\Omega\text{diag}\left\{\sqrt{\boldsymbol{\omega}}\right\}$ and $\boldsymbol{\Psi} = \text{diag}\left\{\sqrt{\boldsymbol{\psi}}\right\}\boldsymbol{C}_\Psi\text{diag}\left\{\sqrt{\boldsymbol{\psi}}\right\}$. We can rewrite the SPN prior for $\boldsymbol{\beta}$ as

$$\boldsymbol{\beta} \overset{d}{=} \left(\sqrt{\boldsymbol{\omega} \circ \boldsymbol{\psi}}\right)\left(\boldsymbol{z}/\sqrt{\boldsymbol{\omega}}\right)\left(\boldsymbol{s}/\sqrt{\boldsymbol{\psi}}\right),$$

where $\sqrt{}$ and $/$ are applied elementwise. We can see that the variances $\boldsymbol{\omega}$ and $\boldsymbol{\psi}$ are not separately identified from the data, because the first term only depends on the variances through their product, and the distributions of the second two terms depend only on $\boldsymbol{C}_\Omega$ and $\boldsymbol{C}_\Psi$. Off-diagonal elements of $\boldsymbol{C}_{\boldsymbol{\Omega}}$ and $\boldsymbol{C}_\Psi$ *are* identified from the data, as they determine fourth-order cross moments of elements of $\boldsymbol{\beta}$. Recalling that we have defined $\boldsymbol{\Sigma} = \boldsymbol{\Omega} \circ \boldsymbol{\Psi}$ and letting $\boldsymbol{C}_\Sigma$ be the correlation matrix corresponding to $\boldsymbol{\Sigma}$, we have

$$\frac{\mathbb{E}\left[\beta_j^2\beta_k^2\right]}{\sigma_{jj}\sigma_{kk}} = 1 + 2c_{\Omega,jk}^2 + 2c_{\Psi,jk}^2 + 4c_{\Sigma,jk}^2 \qquad j \neq k$$

$$\frac{\mathbb{E}\left[\beta_j^2\beta_k\beta_l\right]}{\sigma_{jj}\sqrt{\sigma_{kk}\sigma_{ll}}} = c_{\Sigma,kl} + 4c_{\Sigma,jk}c_{\Sigma,jl} + 2c_{\Omega,kl}c_{\Psi,jk}c_{\Psi,jl} + 2c_{\Psi,kl}c_{\Omega,jk}c_{\Omega,jl} \qquad j \neq k,\ j \neq l,\ k \neq l.$$

We can see that these fourth-order cross moments depend on the values of *both* $c_{\Omega,jk}$ and $c_{\Psi,jk}$, in addition to their product $c_{\Sigma,jk}$. As such, both unknown parameters are separately identifiable. However, whether or not the individual values of $c_{\Omega,jk}$ and $c_{\Psi,jk}$ correspond to correlations of elements of $\boldsymbol{z}$ or $\boldsymbol{s}$ is not known, as we can write:

$$\boldsymbol{\beta} \overset{d}{=} \left(\sqrt{\boldsymbol{\omega} \circ \boldsymbol{\psi}}\right)\left(\boldsymbol{C}_\Psi^{1/2}\boldsymbol{C}_\Omega^{-1/2}\boldsymbol{z}/\sqrt{\boldsymbol{\omega}}\right)\left(\boldsymbol{C}_\Omega^{1/2}\boldsymbol{C}_\Psi^{-1/2}\boldsymbol{s}/\sqrt{\boldsymbol{\psi}}\right).$$

To reduce the number of freely varying parameters and to facilitate identifiability of $\boldsymbol{C}_\Omega$ and $\boldsymbol{C}_\Psi$ from second order moments of $\boldsymbol{\beta}$ alone, we define a special case of the SPN prior which requires that $|c_{\Omega,ij}| = |c_{\Psi,ij}|$. We refer to this as the symmetric SPN (sSPN) prior.

**Structured Normal-Gamma (SNG) Prior:** The structured normal-gamma (SNG) prior is a structured generalization of the normal-gamma prior of Griffin and Brown (2010) obtained by assuming that $s_j^2$ are independent gamma random variables. Because the scales

of each $s_j^2$ are not separately identifiable from the variances of elements of $\boldsymbol{z}$, we assume $s_j^2 \sim \text{gamma}(\text{shape} = c, \text{rate} = c)$ such that $\mathbb{E}\left[s_j^2\right] = 1$ without loss of generality, where $c$ is treated as fixed and known. When $\boldsymbol{\Omega} \propto \boldsymbol{I}_p$ and $c = 1$, the normal-gamma prior reduces to an independent Laplace prior. Accordingly, we can think of the special case of the SNG prior when $c = 1$ and arbitrary $\boldsymbol{\Omega}$ as a novel multivariate Laplace prior.

**Structured Power/Bridge (SPB) Prior:** The structured power/bridge (SPB) prior generalizes the bridge or exponential power prior discussed in Polson et al. (2014). This prior is obtained by assuming that $s_j^2$ are independently distributed according to a polynomially tilted positive $\alpha$-stable distribution with index of stability $\alpha = q/2$, where $q$ is treated as fixed and known (Devroye, 2009; Polson et al., 2014). Again, the scales of each $s_j^2$ are not separately identifiable from the variances of elements of $\boldsymbol{z}$, we parametrize the polynomially tilted positive $\alpha$-stable distribution to ensure that $\mathbb{E}\left[s_j^2\right] = 1$. When $q = 1$, the SPB prior also generalizes the independent Laplace prior and accordingly, coincides with the SNG prior when $c = 1$.

### *4.2.2 Properties*

*Joint Marginal Prior Distributions*

Under the SPN, SNG and SPB priors, the marginal prior distributions $p_{\boldsymbol{\beta}}\left(\boldsymbol{\beta}|\boldsymbol{\Omega}, \boldsymbol{\theta}\right)$ correspond to intractable integrals $\int p_{\boldsymbol{\beta}}\left(\boldsymbol{\beta}|\boldsymbol{s}, \boldsymbol{\Omega}\right) p\left(\boldsymbol{s}|\boldsymbol{\theta}\right) d\boldsymbol{s}$. As a result, it is challenging to study the properties of these prior distributions and the penalties functions they correspond to directly. Fortunately, we can simulate from the marginal prior distributions easily. We simulate from four SNG priors with $c \in \{0.3, 0.5, 1, 10\}$, the sSPN prior, four SPB priors with $q = \{0.65, 0.78, 1, 1.75\}$ and an SPN prior with unit variances and $\rho_{\boldsymbol{\Omega}} = \rho^{0.1}$ and $\rho_{\boldsymbol{\Psi}}^{0.9}$. We note that values of $c$ and $q$ are chosen to correspond to equal tail weight as measured by kurtosis, e.g. the SNG prior with $c = 0.3$ has the same kurtosis as the SPB prior with $q = 0.65$ and the SNG prior with $c = 0.5$ has the same kurtosis as the SPB prior with $q = 0.78$.

Figure 4.5 shows contour plots of the log marginal prior distributions $\log\left(\int p_{\boldsymbol{\beta}}\left(\boldsymbol{\beta}|\boldsymbol{s},\boldsymbol{\Omega}\right)p\left(\boldsymbol{s}|\boldsymbol{\theta}\right)d\boldsymbol{s}\right)$ when $\boldsymbol{\beta}\in\mathbb{R}^2$ with variance-covariance matrix $\boldsymbol{\Sigma}=(1-\rho)\,\boldsymbol{I}_2+\rho\mathbf{1}_2\mathbf{1}_2'$ and correlation $\rho=0.5$. These contours can also be interpreted as contours of a penalty function. The contours have several interesting features. First, they encourage similar estimates of $\beta_1$ and $\beta_2$ by pushing contours away from the origin when $\beta_1$ and $\beta_2$ have the same sign and pushing the contours towards the origin when $\beta_1$ and $\beta_2$ have opposite signs. The SPN, SNG with $c\leq 1$ and SPB with $q\leq 1$ priors encourage possibly sparse estimates of $\boldsymbol{\beta}$ by retaining discontinuities of the log marginal prior on the axes. Interestingly, the contours do not necessarily keep the same shape as the value of log-likelihood changes. Contours closer to the origin are more similar to their independent counterparts, with relatively more encouragement of sparsity than structure, whereas contours farther from the origin tend to encourage relatively more structure and less sparsity. This is especially evident under the SPN prior with $\rho_\omega=\rho^{0.1}$. We note that it is clear from the contours that these priors are *not* log-concave when $\rho\neq 0$ and accordingly correspond to *non-convex* penalties.

In Figure 4.5, we also observe clear spikes that encourage sparse solutions under the SNG prior when $c\leq 0.5$ and under the SPN prior. These reflect the presence of an infinite spike or pole of the joint marginal prior $p_{\boldsymbol{\beta}}\left(\boldsymbol{b}|\boldsymbol{\Omega},\boldsymbol{\theta}\right)$ when at least one element of $\boldsymbol{b}$ is equal to 0.

**Proposition 4.2.1** *If $b_j=0$ for any $j\in\{1,\ldots,p\}$, then $p_{\boldsymbol{\beta}}(\boldsymbol{b}|\boldsymbol{\Psi},\boldsymbol{\Omega})=+\infty$.*

The behavior of the marginal joint prior distribution under the SNG prior $p_{\boldsymbol{\beta}}(\boldsymbol{\beta}|c,\boldsymbol{\Omega})$ is similar when $c<1/2$.

**Proposition 4.2.2** *If $c<1/2$ and $b_j=0$ for any $j\in\{1,\ldots,p\}$, then $p_{\boldsymbol{\beta}}(\boldsymbol{b}|c,\boldsymbol{\Omega})=+\infty$.*

Proofs of these propositions are given in the appendix. The presence of an infinite spike or pole at $b_j=0$ has been demonstrated for several independent priors including the SPN and SNG priors' independent counterparts and has been viewed in the literature as a sufficient condition for the recovery of sparse signals (Carvalho et al., 2010; Griffin and Brown, 2010; Bhattacharya et al., 2015). That said, the presence of an infinite spike or pole at $b_j=0$ also

makes the already challenging problem of computing the posterior mode intractable, as the log marginal prior is not only nonconvex but also has infinitely many modes.

*Range of* $\boldsymbol{\Sigma}$

The SPN, SNG and SPB priors differ in the amount of structure that they can be used to characterize because the marginal variance $\boldsymbol{\Sigma} = \mathbb{E}[\boldsymbol{s}\boldsymbol{s}'] \circ \boldsymbol{\Omega}$ must be positive semidefinite. The behavior of $\mathbb{E}[\boldsymbol{s}\boldsymbol{s}']$ determines the amount of structure a prior can characterize. Under the SPN prior, $\boldsymbol{\Psi}$ can be chosen freely and accordingly $\mathbb{E}[\boldsymbol{s}\boldsymbol{s}']$ can take on any value. Because it is possible to find at least one pair of positive semidefinite covariance matrices $\boldsymbol{\Omega}$ and $\boldsymbol{\Psi}$ that satisfy $\boldsymbol{\Sigma} = \boldsymbol{\Omega} \circ \boldsymbol{\Psi}$ (Styan, 1973), the SPN prior can be used to characterize any structure that can be represented by a positive semidefinite matrix $\boldsymbol{\Sigma}$.

The sSPN prior is less flexible. Although it is challenging to describe the specific properties of covariance matrices $\boldsymbol{\Sigma}$ for which $\boldsymbol{\Omega}$ and $\boldsymbol{\Psi}$ satisfying $|\omega_{ij}| = |\psi_{ij}|$ are positive semidefinite, it is straightforward to find a positive semidefinite matrix $\boldsymbol{\Sigma}$ for which $\boldsymbol{\Omega}$ or $\boldsymbol{\Psi}$ satisfying $|\omega_{ij}| = |\psi_{ij}|$ are not positive definite.

The SNG and SPB priors are also limited to characterizing a restricted range of structure. Diagonal elements of $\mathbb{E}[\boldsymbol{s}\boldsymbol{s}']$ are equal to 1, whereas off-diagonal elements are equal to $\mathbb{E}[s_j]^2 < \mathbb{E}[s_j^2] = 1$. Thus, $\mathbb{E}[\boldsymbol{s}\boldsymbol{s}']$ shrinks off-diagonal elements of $\boldsymbol{\Omega}$, reducing dependence. Similar behavior was observed for alternative multivariate $t$-distributions in Finegold and Drton (2011). When $\boldsymbol{\beta} \in \mathbb{R}^2$, we can explicitly calculate the maximum marginal prior correlation $\rho$ under the SNG and SPB priors as a function of $c$ or $q$. Under the SNG prior, the maximum correlation is equal to $c^{-1}(\Gamma(c + 1/2)/\Gamma(c))^2$ and under the SPB prior, the maximum correlation is equal to $(\pi/2)(\Gamma(2/q)/\sqrt{\Gamma(1/q)\Gamma(3/q)})^2$. When $q = c = 1$ and both priors are equivalent, the maximum correlation is equal to $\pi/4 \approx 0.79$.

We plot the maximum correlation as a function of kurtosis under both priors in Figure 4.6. We observe greater reductions in the maximum correlation when the tails are heavier, and under the SNG prior relative to a SPB prior with equal kurtosis. The restricted range of $\boldsymbol{\Sigma}$ under the SNG and SPB priors is similar to the restricted range of the variance-covariance

matrix of the alternative multivariate $t$-distribution introduced in Finegold and Drton (2011). Intuitively, the restricted range of $\boldsymbol{\Sigma}$ relates to the conflict that arises between the properties of the marginal density $p_{\boldsymbol{\beta}}(\boldsymbol{\beta}|\boldsymbol{\Omega}, \boldsymbol{\theta})$ needed to preserve elementwise near or exact sparsity, specifically concentration of the density along the axes, and the properties of the marginal density needed to encourage structure, specifically concentration of the density along a line when $p = 2$.

*Copulas*

We can visualize the dependence structures induced by all three priors by estimating the copulas for $\boldsymbol{\beta} \in \mathbb{R}^2$ with unit marginal variances and correlation $\rho = 0.5$. Figure 4.7 compares estimates of the copulas of the same SPN, SNG and SPB priors considered previously. All estimates by simulating $1,000,000$ values from the corresponding prior, transforming simulated values of $\beta_1$ and $\beta_2$ into percentiles $u_1$ and $u_2$, and computing a kernel bivariate density estimate of the percentiles.

The behavior of the SPN copula depends on how $\rho$ is factored into $\rho_{\boldsymbol{\Omega}}$ and $\rho_{\boldsymbol{\Psi}}$. Under the sSPN prior with $\rho_{\boldsymbol{\Omega}} = \rho_{\boldsymbol{\Psi}}$, the sSNG copula indicates that when elements of $\boldsymbol{\beta}$ are positively correlated, the sSPN prior concentrates around values of $\boldsymbol{\beta}$ that are jointly very small or exactly equal to zero and around values of $\boldsymbol{\beta}$ that have the same sign and are very large in magnitude. When we set $\rho_{\boldsymbol{\Omega}} = \rho^{0.1}$ and $\rho_{\boldsymbol{\Psi}} = \rho^{0.9}$, the SPN prior concentrates more strongly around values of $\boldsymbol{\beta}$ that are jointly very small or exactly equal to zero and around values of $\boldsymbol{\beta}$ that that have the same sign and are very large in magnitude. This parametrization of the SPN prior also concentrates slightly less around values of $\boldsymbol{\beta}$ along the axes, where one element is very small or exactly zero and the other is not. Last, this parametrization of the SPN prior concentrates more around values of $\boldsymbol{\beta}$ that are opposite in sign but similar in magnitude. Arguably, this could be considered undesirable behavior and reflects the strangeness of a data generating process that assumes $\boldsymbol{\beta}$ is made up of one strongly correlated component and one very weakly correlated component, both of which can take any value on $\mathbb{R}$. Such a prior model for $\boldsymbol{\beta}$ is unlikely to make sense in many real

applications.

The behavior of the SNG and SPB copulas is generally similar when the kurtosis is held constant under both priors, and depends on how similar the prior is to multivariate normal distribution. Both display increasingly strong orthant dependence as the priors become less normal with heavier tails, i.e. as $c$ or $q \to 0$. This means that as $c$ or $q \to 0$, the priors concentrate more strongly around values of $\boldsymbol{\beta}$ that have the same sign. Compared to the SPN priors, the SNG and SPB priors also concentrate more around the axes where at a single element of $\boldsymbol{\beta}$ is close to or exactly equal to zero. The SNG prior displays especially strong orthant dependence and appears to converge to a uniform distribution over the positive and negative orthants as $c \to 0$, which likely results from the presence of an infinite spike or pole of the marginal prior density when any element of $\boldsymbol{\beta}$ is equal to zero under the SNG prior. Like the concentration of the SPN prior with $\rho_{\boldsymbol{\Omega}} = \rho^{0.1}$ around values of $\boldsymbol{\beta}$ that have opposite signs but similar magnitudes, convergence of the SNG prior to a uniform distribution on the positive and negative orthants as $c \to 0$ may be undesirable behavior and may not make sense in many real applications. The softer contours of the SPB prior likely correspond to more plausible models for $\boldsymbol{\beta}$ in many settings.

*Univariate Marginal Prior Distributions*

For all three priors, the univariate marginal distributions of elements of $\boldsymbol{\beta}$ belong to the same family as the unstructured independent priors that they generalize. We show this in the appendix, although it is straightforward to see from the stochastic representations of the priors.

We also show that under the SPN and SNG priors with $c \leq 1/2$, the univariate marginal prior distributions an infinite spike or pole of the joint marginal prior $p_{\boldsymbol{\beta}}(\boldsymbol{b}|\boldsymbol{\Omega}, \boldsymbol{\theta})$ when at least one element of $\boldsymbol{b}$ is equal to 0.

**Proposition 4.2.3** *For any* $j \in \{1, \ldots, p\}$, $p_{\beta_j}(0|\boldsymbol{\Psi}, \boldsymbol{\Omega}) = +\infty$.

Likewise, the marginal distribution of a single element of $\beta_j$ has a pole at 0 under the

SNG prior when $c < 1/2$.

**Proposition 4.2.4** *If* $c < 1/2$, $p_{\beta_j}(0|c, \boldsymbol{\Omega}) = +\infty$ *for any* $j \in \{1, \ldots, p\}$.

This suggests that marginal posterior means and medians computed under the SPN and SNG priors will possibly be very small, although not exactly equal to zero.

*Univariate Conditional Prior Distributions*

Because the SPN, SNG and SPB priors can all be represented as normal scale-mixtures, it is straightforward to simulate from the univariate conditional prior distributions, $p\left(\beta_j|\boldsymbol{\beta}_{-j}|\boldsymbol{\Omega}, \boldsymbol{\theta}\right)$. Figure 4.8 shows estimated univariate conditional prior distributions $p\left(\beta_1|\boldsymbol{\beta}_2|\boldsymbol{\Omega}, \boldsymbol{\theta}\right)$, again for $\boldsymbol{\beta} \in \mathbb{R}^2$ with unit marginal prior variance and marginal prior correlation $\rho = 0.5$. Under all ten of the structured priors, introducing correlation between $\beta_1$ and $\beta_2$ can help us estimate $\beta_1$ if the true values of $\beta_1$ and $\beta_2$ are in fact similar. Under all ten of the structured priors, the conditional prior distribution of $\beta_1$ given $\beta_2$ becomes more strongly peaked at 0 when $\beta_2 = 0$ than under an independent prior and the conditional prior distribution of $\beta_1$ given $\beta_2$ shifts its mass towards 2 when $\beta_2 = 2$. Interestingly, these trends are most pronounced under the SPN prior with $\rho_{\boldsymbol{\Omega}} = \rho^{0.1}$. Furthermore, under the sSPN prior and the SNG priors with $c \leq 0.5$, the mode of the conditional prior distribution of $\beta_1$ given $\beta_2$ stays at 0 even when $\beta_2 = 2$, which suggests that the introduction of a marginal prior correlation of $\rho = 0.5$ between $\beta_1$ and $\beta_2$ results in relatively more dependence near the origin than in the tails.

*Mode Thresholding Functions*

One last perspective on the properties of the SPN, SNG and SPB priors can be gained by examining the mode thresholding function under a linear model for $\boldsymbol{y}$, which relates the OLS estimates of $\boldsymbol{\beta}$ given by $\hat{\boldsymbol{\beta}}_{OLS}$ to the posterior mode of $\beta_1$ given by

$$\hat{\boldsymbol{\beta}}_1 = \text{argmin}_{\beta_1} \frac{1}{2\phi^2} \left|\left|\hat{\boldsymbol{\beta}}_{OLS} - \boldsymbol{\beta}\right|\right|_2^2 + \log\left(\int p_{\boldsymbol{\beta}|\boldsymbol{s}}\left(\boldsymbol{\beta}|\boldsymbol{s}, \boldsymbol{\Omega}\right) p_{\boldsymbol{s}}\left(\boldsymbol{s}|\boldsymbol{\theta}\right) d\boldsymbol{s}\right), \tag{4.1}$$

for fixed values of the noise variance $\phi^2$, prior parameters $\boldsymbol{\Omega}$ and $\boldsymbol{\theta}$ and $\hat{\boldsymbol{\beta}}$. The mode-thresholding function is not available in closed form but can be approximated using a Gibbs-within-EM described in the following section.

Figure 4.9 shows approximate mode-thresholding functions for $\boldsymbol{\beta} \in \mathbb{R}^2$ with unit marginal prior variances, marginal prior correlation $\rho = 0.5$, noise variance $\phi^2 = 0.1$, $\hat{\beta}_{OLS,2} \in \{-0.5, 0, 0.5, 1\}$ and $\hat{\beta}_{OLS,1} \in (0, 1)$, computed from $1,001,000$ Gibbs sampler iterations, with the first $1,000$ samples discarded as burn-in. Examination of the mode-thresholding functions confirms that the SPN prior, SNG prior with $c \leq 1$ and SPB prior with $c \leq 1$ can yield possible sparse posterior mode estimates of $\boldsymbol{\beta}$, and that the introduction of structure encourages or discourages sparse estimates $\hat{\beta}_1$ depending not only on the observed value of $\hat{\beta}_{OLS,1}$ but also the observed value of $\hat{\beta}_{OLS,2}$. There are a few interesting trends that relate to previously identified properties of the SPN, SNG and SPB priors. We observe that both SPN priors shrink $\hat{\beta}_1$ towards 0 more aggressively when $\hat{\beta}_{OLS,2} = -0.5$ than when $\hat{\beta}_{OLS,2} = 0$, which reflects the tendency of the SPN priors to encourage not only estimates of $\boldsymbol{\beta}$ with similar signs but also similar magnitudes. For all four sparsity inducing priors, the value of $\hat{\beta}_{OLS,2}$ appears to affect the estimate $\hat{\beta}_1$ more when $\hat{\beta}_{OLS,1}$ is smaller, which reflects the general tendency of all of the Laplace- and heavier-than-Laplace tailed priors to display relatively stronger dependence at the origin than in the tails.

### *4.3 Computation Under Structured Shrinkage Priors*

The posterior mode of $\boldsymbol{\beta}$ maximizes the integral $\int p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{\phi})\, p(\boldsymbol{\beta}|\boldsymbol{\Omega}, \boldsymbol{s})\, p(\boldsymbol{s}|\boldsymbol{\theta})\, d\boldsymbol{s}$ over $\boldsymbol{\beta}$. As noted previously, we can also think of the posterior mode as a penalized estimate of $\boldsymbol{\beta}$, where the penalty is given by $-\log\left(\int p(\boldsymbol{\beta}|\boldsymbol{\Omega}, \boldsymbol{s})\, p(\boldsymbol{s}|\boldsymbol{\theta})\, d\boldsymbol{s}\right)$. This integral is generally intractable when $\boldsymbol{\Omega}$ is not diagonal, but can be maximized using an MCMC within Expectation Maximization (EM) algorithm (Dempster et al., 1977). Given an initial value $\boldsymbol{\beta}^{(0)}$, this algorithm proceeds by iterating the following until $\left|\left|\boldsymbol{\beta}^{(i+1)} - \boldsymbol{\beta}^{(i)}\right|\right|_2^2$ converges:

- using MCMC to simulate $M$ draws $\boldsymbol{s}^{(1)}, \ldots, \boldsymbol{s}^{(M)}$ from the full conditional distribution

of $\boldsymbol{s}$ given $\boldsymbol{\beta}^{(i)}$ and $\boldsymbol{\theta}$, set $\widehat{\mathbb{E}}\left[\left(\frac{\mathbf{1}}{\boldsymbol{s}}\right)\left(\frac{\mathbf{1}}{\boldsymbol{s}}\right)'|\boldsymbol{\beta}^{(i)},\boldsymbol{\theta}\right] = \frac{1}{M}\sum_{j=1}^{M}\left(\frac{\mathbf{1}}{\boldsymbol{s}^{(j)}}\right)\left(\frac{\mathbf{1}}{\boldsymbol{s}^{(j)}}\right)'$;

- set $\boldsymbol{\beta}^{(i+1)} = \text{argmin}_{\boldsymbol{\beta}} h\left(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{\beta},\boldsymbol{\phi}\right) + \frac{1}{2}\boldsymbol{\beta}'\left(\boldsymbol{\Omega}^{-1}\circ\widehat{\mathbb{E}}\left[\left(\frac{\mathbf{1}}{\boldsymbol{s}}\right)\left(\frac{\mathbf{1}}{\boldsymbol{s}}\right)'|\boldsymbol{\beta}^{(i)},\boldsymbol{\theta}\right]\right)\boldsymbol{\beta}$.

Alternative posterior summaries can be obtained by simulating $M$ values $\boldsymbol{\beta}^{(1)},\ldots,\boldsymbol{\beta}^{(M)}$ from the joint posterior distribution of $(\boldsymbol{\beta},\boldsymbol{s})$ given $\boldsymbol{y}$, $\boldsymbol{X}$, $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ given an initial value $\boldsymbol{\beta}^{(0)}$ by iteratively simulating $\boldsymbol{s}^{(j)}$ from the full conditional distribution of $\boldsymbol{s}$ given $\boldsymbol{\beta}^{(j-1)}$ and $\boldsymbol{\theta}$ and simulating $\boldsymbol{\beta}^{(j)}$ from the full conditional distribution of $\boldsymbol{\beta}$ given $\boldsymbol{y}$, $\boldsymbol{X}$, $\boldsymbol{s}^{(j)}$ and $\boldsymbol{\phi}$. The posterior mean of $\boldsymbol{\beta}$ can be obtained by computing the sample mean $\frac{1}{M}\sum_{j=1}^{M}\boldsymbol{\beta}^{(j)}$, and alternative posterior summaries can be obtained by computing the corresponding sample quantities for $\boldsymbol{\beta}^{(1)},\ldots,\boldsymbol{\beta}^{(M)}$.

We rely on elliptical slice sampling to simulate from both full conditional distributions (Murray et al., 2010). Elliptical slice sampling can be used in cases where a density for a multivariate random variable $p_{\boldsymbol{z}}\left(\boldsymbol{z}\right)$ can be factored into a normal component and a remainder as follows

$$p_{\boldsymbol{z}}\left(\boldsymbol{z}\right) \propto_{\boldsymbol{z}} \exp\left\{-\frac{1}{2}\left(\boldsymbol{z}-\boldsymbol{m}\right)'\boldsymbol{V}^{-1}\left(\boldsymbol{z}-\boldsymbol{m}\right)\right\} f\left(\boldsymbol{z},\boldsymbol{\eta}\right), \tag{4.2}$$

where $\boldsymbol{m}$ and $\boldsymbol{V}$ are known and $f\left(\boldsymbol{\eta},\boldsymbol{z}\right)$ is a known arbitrary function of $\boldsymbol{z}$ and some known quantity $\boldsymbol{\eta}$. Elliptical slice sampling leverages the fact that $\boldsymbol{z} \overset{d}{=} \boldsymbol{z}_1\text{sin}\left(\zeta\right) + \boldsymbol{z}_2\text{cos}\left(\zeta\right) + \boldsymbol{m}$ when $\boldsymbol{z} \sim \text{normal}\left(\boldsymbol{m},\boldsymbol{V}\right)$ and $\boldsymbol{z}_1,\boldsymbol{z}_2 \sim \text{normal}\left(\boldsymbol{0},\boldsymbol{V}\right)$, regardless of what distribution is assumed for $\zeta$. It follows that simulating $\boldsymbol{z}$ according to (4.2) is equivalent to simulating $(\zeta,\boldsymbol{z}_1,\boldsymbol{z}_2)$ from according to

$$\begin{aligned} p_{\boldsymbol{z}_1,\boldsymbol{z}_2,\zeta}\left(\boldsymbol{z}_1,\boldsymbol{z}_2,\zeta\right) \propto_{\boldsymbol{z}_1,\boldsymbol{z}_2,\zeta} &\exp\left\{-\frac{1}{2}\left(\boldsymbol{z}_1'\boldsymbol{V}^{-1}\boldsymbol{z}_1 + \boldsymbol{z}_2\boldsymbol{V}^{-1}\boldsymbol{z}_2\right)\right\} \\ &\times f_{\boldsymbol{z}}\left(\boldsymbol{z}_1\text{sin}\left(\zeta\right) + \boldsymbol{z}_2\text{cos}\left(\zeta\right) + \boldsymbol{m},\boldsymbol{\eta}\right), \end{aligned} \tag{4.3}$$

where $\zeta \in [0,2\pi]$ and setting $\boldsymbol{z} = \boldsymbol{z}_1\text{sin}\left(\zeta\right) + \boldsymbol{z}_2\text{cos}\left(\zeta\right) + \boldsymbol{m}$. Importantly, simulating from (4.3) can be performed in two straightforward steps. Given initial values $\zeta^{(0)}$ and $\boldsymbol{z}^{(0)}$, we iteratively:

- simulate $\boldsymbol{\epsilon}^{(i)} \sim \text{normal}\left(\mathbf{0}, \boldsymbol{V}\right)$ and set $\boldsymbol{z}_1^{(i)}$ and $\boldsymbol{z}_2^{(i)}$

$$\left( \begin{array}{cc} \boldsymbol{z}_1^{(i)} & \boldsymbol{z}_2^{(i)} \end{array} \right) = \left( \begin{array}{cc} \boldsymbol{z}^{(i-1)} - \boldsymbol{m} & \boldsymbol{\epsilon}^{(i)} \end{array} \right) \left( \begin{array}{cc} \sin\left(\zeta^{(i-1)}\right) & \cos\left(\zeta^{(i-1)}\right) \\ \cos\left(\zeta^{(i-1)}\right) & -\sin\left(\zeta^{(i-1)}\right) \end{array} \right),$$

  i.e. by rotating $\boldsymbol{\epsilon}^{(i)}$ and $\boldsymbol{z}^{(i-1)} - \boldsymbol{m}$;

- simulate $\zeta^{(i)}$ given $\boldsymbol{z}_1^{(i)}$ and $\boldsymbol{z}_1^{(i)}$.

Naively, we can think of elliptical slice sampling as iteratively proposing new simulated values of $\boldsymbol{z}$ centered at $\boldsymbol{m}$, not unlike random-walk Metropolis-Hastings, but replacing the choice of acceptance or rejection with rotation. This is computationally convenient, as it reduces the very difficult problem of simulating $\boldsymbol{z}$ from a nonstandard multivariate distribution to the much simpler problem of simulating a multivariate normal vector and a scalar from a nonstandard univariate distribution on a bounded support, which can be achieved quickly and efficiently using univariate slice sampling as described in the appendix (Neal, 2003). Furthermore, computational benefits may arise even when (4.2) is a standard distribution if $\boldsymbol{z}$ is high dimensional and $\boldsymbol{V}$ is structured in some way that facilitates simulation, e.g. $\boldsymbol{V}$ is diagonal, block-diagonal or separable.

### *4.3.1 Simulating from the Full Conditional Distribution of $\boldsymbol{\beta}$*

For computational reasons, we simulate from $p_{\boldsymbol{\beta}|\boldsymbol{s},\boldsymbol{y}}\left(\boldsymbol{\beta}|\boldsymbol{s}, \boldsymbol{X}, \boldsymbol{y}, \boldsymbol{\beta}, \boldsymbol{\Omega}, \boldsymbol{\phi}\right)$ by performing a change of variables to $\boldsymbol{u} = \boldsymbol{\Omega}^{-1/2}\text{diag}\left\{\boldsymbol{s}\right\}^{-1}\boldsymbol{\beta}$ and simulating from $p_{\boldsymbol{u}|\boldsymbol{y},\tilde{\boldsymbol{X}}}\left(\boldsymbol{u}|\tilde{\boldsymbol{X}}, \boldsymbol{y}, \boldsymbol{\phi}\right)$ where $\tilde{\boldsymbol{X}} = \boldsymbol{X}\text{diag}\left\{\boldsymbol{s}\right\}\boldsymbol{\Omega}^{1/2}$ and then setting $\boldsymbol{\beta} = \text{diag}\left\{\boldsymbol{s}\right\}\boldsymbol{\Omega}^{1/2}\boldsymbol{u}$. The full conditional distribution of $\boldsymbol{u}$ given the data and other unknown parameters is proportional to

$$\exp\left\{-h\left(\boldsymbol{y}|\tilde{\boldsymbol{X}}, \boldsymbol{v}, \boldsymbol{\phi}\right)\right\}\exp\left\{-\frac{1}{2}\boldsymbol{u}'\boldsymbol{u}\right\}.$$

As written, elliptical slice sampling can be used to simulate from the full conditional distribution. However, it may mix slowly if the posterior mean of $\boldsymbol{u}$ is not close to the prior mean $\mathbf{0}$ and if the posterior variance is not close to $\boldsymbol{I}_p$. To improve mixing, we pre- and

post-multiply by the kernel of a multivariate normal density with mean $\tilde{\boldsymbol{u}}$ and variance $\tilde{\boldsymbol{V}}$, where neither $\tilde{\boldsymbol{u}}$ nor $\tilde{\boldsymbol{V}}$ depend on $\boldsymbol{u}$,

$$\underbrace{\exp\left\{-\frac{1}{2}\left(\boldsymbol{u}-\tilde{\boldsymbol{u}}\right)'\tilde{\boldsymbol{V}}^{-1}\left(\boldsymbol{u}-\tilde{\boldsymbol{u}}\right)\right\}}_{(*)} \tag{4.4}$$

$$\times \underbrace{\exp\left\{-h\left(\boldsymbol{y}|\tilde{\boldsymbol{X}},\boldsymbol{u},\boldsymbol{\phi}\right)-\frac{1}{2}\left(\boldsymbol{u}'\boldsymbol{u}+\left(\boldsymbol{u}-\tilde{\boldsymbol{u}}\right)'\tilde{\boldsymbol{V}}^{-1}\left(\boldsymbol{u}-\tilde{\boldsymbol{u}}\right)\right)\right\}}_{(**)},$$

and use the normal kernel $(*)$ as the normal component for elliptical slice sampling and $(**)$ as the remainder. We can improve mixing by choosing $\tilde{\boldsymbol{u}}$ and $\tilde{\boldsymbol{V}}$ to approximate the mean and variance of the full conditional distribution of $\boldsymbol{u}$. A natural choice for approximating the the mean and variance of the full conditional distribution is the Laplace approximation, where the mean is the mode of $p_{\boldsymbol{u}|\boldsymbol{s},\boldsymbol{y},\tilde{\boldsymbol{X}}}\left(\boldsymbol{u}|\boldsymbol{y},\tilde{\boldsymbol{X}},\boldsymbol{\phi}\right)$,

$$\tilde{\boldsymbol{u}} = \text{argmin}_{\boldsymbol{u}} h\left(\boldsymbol{y}|\tilde{\boldsymbol{X}},\boldsymbol{u},\boldsymbol{\phi}\right)+\frac{1}{2}\boldsymbol{u}'\boldsymbol{u},$$

and the variance as a function of $\tilde{\boldsymbol{u}}$ is

$$\tilde{\boldsymbol{V}}\left(\tilde{\boldsymbol{u}}\right) = \left(\frac{\partial^2}{\partial\tilde{\boldsymbol{u}}\partial\tilde{\boldsymbol{u}}'}h\left(\boldsymbol{y}|\tilde{\boldsymbol{X}},\tilde{\boldsymbol{u}},\boldsymbol{\phi}\right)+\boldsymbol{I}_p\right)^{-1}.$$

The mean can be computed easily for many common choices of $h\left(\boldsymbol{y}|\tilde{\boldsymbol{X}},\boldsymbol{u},\boldsymbol{\phi}\right)$, e.g. via coordinate descent, even when $p$ is very large. However when $p$ is very large, as will be the case in many scenarios where structured shrinkage priors are useful, the matrix inversion required to compute $\tilde{\boldsymbol{V}}\left(\tilde{\boldsymbol{u}}\right)$ may not be possible. Furthermore, when $\boldsymbol{X}$ is poorly conditioned and $p$ is large, coordinate descent may converge prohibitively slowly. Fortunately, we only require that $\tilde{\boldsymbol{u}}$ and $\tilde{\boldsymbol{V}}$ do not depend past values of $\boldsymbol{u}$. As a result, we can address feasibility problems that arise in high dimensions by using an approximate posterior mode $\tilde{\boldsymbol{u}}$, e.g. from a handful of iterations of coordinate descent initialized at the previous value of $\tilde{\boldsymbol{u}}$, and by using a diagonal approximation of $\tilde{\boldsymbol{V}}$. We find that the computationally simple approximation $\tilde{\boldsymbol{V}} = \boldsymbol{I}_p$ works well, because

$$\left(\frac{\partial^2}{\partial\tilde{\boldsymbol{u}}\partial\tilde{\boldsymbol{u}}'}h\left(\boldsymbol{y}|\tilde{\boldsymbol{X}},\tilde{\boldsymbol{u}},\boldsymbol{\phi}\right)+\boldsymbol{I}_p\right)^{-1} = \boldsymbol{I}_p - \left(\boldsymbol{I}_p+\frac{\partial^2}{\partial\tilde{\boldsymbol{u}}\partial\tilde{\boldsymbol{u}}'}h\left(\boldsymbol{y}|\tilde{\boldsymbol{X}},\tilde{\boldsymbol{u}},\boldsymbol{\phi}\right)\right)^{-1}\frac{\partial^2}{\partial\tilde{\boldsymbol{u}}\partial\tilde{\boldsymbol{u}}'}h\left(\boldsymbol{y}|\tilde{\boldsymbol{X}},\tilde{\boldsymbol{u}},\boldsymbol{\phi}\right)$$

and accordingly diag $\left\{\tilde{\boldsymbol{V}}\left(\tilde{\boldsymbol{u}}\right)\right\} \leq \mathbf{1}_p$. Using this approach, we can simulate from the full conditional distribution $p_{\boldsymbol{\beta}|\boldsymbol{s},\boldsymbol{y}}\left(\boldsymbol{\beta}|\boldsymbol{s},\boldsymbol{X},\boldsymbol{y},\boldsymbol{\beta},\boldsymbol{\Omega},\boldsymbol{\phi}\right)$ quickly using a handful of matrix operations that scale well when $n$ and $p$ are large, especially when rows of $\boldsymbol{X}$ can be represented as arrays and $\boldsymbol{\Omega}$ is separable with components that correspond to variability along dimensions of $\boldsymbol{X}$.

### *4.3.2 Simulating from the Full Conditional Distribution of* $\boldsymbol{s}$

Given a prior $p_{s_j}\left(s_j|\boldsymbol{\theta}\right)$, the full conditional distribution of $\boldsymbol{s}$ can be written as proportional to

$$\underbrace{\exp\left\{-\frac{1}{2}\sum_{j=1}^{p} s_j^{-2}\beta_j^2\omega^{jj}\right\}}_{(*)}\underbrace{\left(\prod_{j=1}^{p}\left(s_j^{-2}\right)^{\frac{1}{2}-1}\right)}_{(**)} \times\left(\prod_{j=1}^{p} s_j^{-2}p_{s_j}\left(s_j|\boldsymbol{\theta}\right)\exp\left\{-\frac{1}{2}\sum_{j'\neq j} s_j^{-1}s_{j'}^{-1}\beta_j\beta_{j'}\omega^{jj'}\right\}\right), \tag{4.5}$$

where $\omega^{ij}$ refer to the element in the $i$-th row and $j$-th column of $\boldsymbol{\Omega}^{-1}$ The choices $p_{s_j}\left(s_j|\boldsymbol{\theta}\right)$ that yield the SPN, SNG and SPB models do not yield standard distributions that are easy to simulate from when $\boldsymbol{\Omega}$ is not a diagonal matrix. However, after performing a change of variables we can use elliptical slice sampling to simulate from the full conditional distribution of $\boldsymbol{s}$ for arbitrary priors on $\boldsymbol{s}$. The procedure varies depending on whether elements $s_j$ are defined on $\mathbb{R}$, as is the case under the SPN prior, or $\mathbb{R}_{++} = (0, +\infty)$, as is the case under the SNG and SPB priors.

#### *Elements $s_j$ Defined on $\mathbb{R}$*

When elements $s_j$ are defined on $\mathbb{R}$, we note that the term $(*)$ in (4.5) is the kernel of the product of $p$ mean zero normal densities in $s_j^{-1}$. Accordingly, we perform the change of variables $s_j = r_j^{-1}$, yielding

$$\exp\left\{-\frac{1}{2}\sum_{j=1}^{p} r_j^2\beta_j^2\omega^{jj}\right\}\left(\prod_{j=1}^{p}\left(r_j\right)^{\frac{1}{2}-1}\right)\left(\prod_{j=1}^{p} p_{s_j}\left(r_j^{-1}|\boldsymbol{\theta}\right)\exp\left\{-\frac{1}{2}\sum_{j'\neq j} r_j r_{j'}\beta_j\beta_{j'}\omega^{jj'}\right\}\right).$$

It follows that simulating from the full conditional distribution of $\boldsymbol{s}$ is equivalent to simulating from the full conditional distribution of $\boldsymbol{r}$ and setting $s_j = 1/r_j$. Because the first term is the kernel of a multivariate normal density in $\boldsymbol{r}$, elliptical slice sampling can again be used.

*Elements $s_j$ Defined on $\mathbb{R}_{++}$*

When elements $s_j$ are defined on $\mathbb{R}_{++}$, we note that the product of the terms $(*)$ and $(**)$ in (4.5) is the kernel of the product of $p$ scaled chi-square density in $s_j^{-2}$ with one degree of freedom. Accordingly, we perform the change of variables $s_j^{-2} = t_j$, yielding

$$\exp\left\{-\frac{1}{2}\sum_{j=1}^{p} t_j \beta_j^2 \omega^{jj}\right\}\left(\prod_{j=1}^{p} t_j^{\frac{1}{2}-1}\right)\left(\prod_{j=1}^{p} t_j^{-\frac{1}{2}} p_{s_j}\left(t_j^{-\frac{1}{2}}|\boldsymbol{\theta}\right)\exp\left\{-\frac{1}{2}\sum_{j'\neq j}\sqrt{t_j t_{j'}}\beta_j\beta_{j'}\omega^{jj'}\right\}\right).$$

As a scaled chi-square random variable with one degree of freedom, $t_j \overset{d}{=} r_j^2$ where $r_j \overset{i.i.d.}{\sim}$ normal $\left(0, 1/\left(\beta_j^2\omega^{jj}\right)\right)$. It follows that simulating from the full conditional distribution of $\boldsymbol{s}$ is equivalent to simulating from the full conditional distribution of $\boldsymbol{r}$,

$$\exp\left\{-\frac{1}{2}\sum_{j=1}^{p} r_j^2 \beta_j^2 \omega^{jj}\right\}\left(\prod_{j=1}^{p} |r_j|^{-1} p_{s_j}\left(|r_j|^{-1}|\boldsymbol{\theta}\right)\exp\left\{-\frac{1}{2}\sum_{j'\neq j}|r_j r_{j'}|\,\beta_j\beta_{j'}\omega^{jj'}\right\}\right),$$

and setting $s_j = 1/r_j$. Because the first term is the kernel of a multivariate normal density in $\boldsymbol{r}$, elliptical slice sampling can be used.

*Simulating from the Full Conditional of $\boldsymbol{\delta}$ Under the SPB Prior*

In the case of the SPB prior, an additional computational challenge arises because the polynomially tilted positive $\alpha$-stable density $p_{s_j}\left(s_j|\boldsymbol{\theta}\right)$ is not available in closed form. Fortunately, a polynomially tilted positive $\alpha$-stable density that can be represented as a rate mixture of generalized gamma random variables (Devroye, 2009):

$$s_j^2 \overset{d}{=} \frac{\Gamma\left(\frac{1}{2\alpha}\right)\xi_j^{\frac{1-\alpha}{\alpha}}}{2\Gamma\left(\frac{3}{2\alpha}\right)},\ \xi_j \overset{i.i.d.}{\sim} \text{gamma}\left(\text{shape} = \frac{1+\alpha}{2\alpha}, \text{rate} = f\left(\delta_j|\alpha\right)\right) \text{ and}$$
$$p_{\delta_j}\left(\delta_j|\alpha\right) \propto f\left(\delta_j|\alpha\right)^{\frac{\alpha-1}{2\alpha}},$$

where $f(\delta_j|\alpha) = \sin(\alpha\delta_j)^{\frac{\alpha}{1-\alpha}} \sin((1-\alpha)\delta_j) / \sin(\delta_j)^{\frac{1}{1-\alpha}}$ and $\delta_j \in (0, \pi)$. As a result, it is also necessary to simulate from the full conditional of $\boldsymbol{\delta}$, which is given by

$$p(\delta_j|\xi_j, q) \propto f(\delta_j|\alpha) \exp\{-f(\delta_j|\alpha)\xi_j\}.$$

This is not a standard density, however it is straightforward to simulate from it using univariate slice sampling techniques because $\delta_j \in (0, \pi)$. Details on the derivation of this full conditional distribution and the implementation of the univariate slice sampling step are provided in the appendix.

## *4.4 Variance Component Estimation*

### *4.4.1 Maximum Marginal Likelihood*

In the previous section, we presented a general approach for simulating from the posterior distribution of $\boldsymbol{\beta}$ under the SPN, SNG and SPB priors. As a result, we can perform maximum marginal likelihood estimation (MMLE) of the unknown variance components $\boldsymbol{\Omega}$ and, in the case of the SPN prior, $\boldsymbol{\Psi}$, using an Gibbs-within-EM algorithm. This is easiest to approach by writing the likelihood as an integral over $\boldsymbol{s}$ and $\boldsymbol{z}$, where $\boldsymbol{\beta} = \boldsymbol{s} \circ \boldsymbol{z}$. We have:

$$\int \exp\{-h(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{s} \circ \boldsymbol{z}, \boldsymbol{\phi})\} \left( \frac{1}{\sqrt{2\pi|\boldsymbol{\Omega}|}} \exp\left\{ -\frac{1}{2} \boldsymbol{z}' \boldsymbol{\Omega}^{-1} \boldsymbol{z} \right\} \right) p(\boldsymbol{s}|\boldsymbol{\theta}) \, d\boldsymbol{z} d\boldsymbol{s}.$$

Given an initial value $\boldsymbol{\Omega}^{(0)}$, Gibbs-within-EM estimates of $\boldsymbol{\Omega}$ can be obtained by iterating the following until $\left|\left|\boldsymbol{\Omega}^{(i+1)} - \boldsymbol{\Omega}^{(i)}\right|\right|_F$ converges:

- using MCMC to simulate $M$ draws $\left(\boldsymbol{z}^{(1)}, s^{(1)}\right), \ldots, \left(\boldsymbol{z}^{(M)}, \boldsymbol{s}^{(M)}\right)$ from the joint posterior distribution of $(\boldsymbol{z}, \boldsymbol{s})$ given $\boldsymbol{\Omega}^{(i)}$, $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$, set $\widehat{\mathbb{E}}\left[\boldsymbol{z}\boldsymbol{z}'|\boldsymbol{\Omega}^{(i)}, \boldsymbol{X}, \boldsymbol{y}, \boldsymbol{\phi}, \boldsymbol{\theta}\right] = \frac{1}{M} \sum_{j=1}^{M} \boldsymbol{z}^{(j)} \left(\boldsymbol{z}^{(j)}\right)'$;
- set $\boldsymbol{\Omega}^{(i+1)} = \text{argmin}_{\boldsymbol{\Omega}>0} \log(|\boldsymbol{\Omega}|) + \frac{1}{2}\text{tr}\left(\boldsymbol{\Omega}^{-1} \widehat{\mathbb{E}}\left[\boldsymbol{z}\boldsymbol{z}'|\boldsymbol{\Omega}^{(i)}, \boldsymbol{X}, \boldsymbol{y}, \boldsymbol{\phi}, \boldsymbol{\theta}\right]\right)$.

In practice, we will rarely assume $\boldsymbol{\Omega}$ is unstructured, as doing so requires estimating $p + p(p-1)/2$ unknown parameters from a single observation of a $p \times 1$ vector. Instead, we

will parametrize $\boldsymbol{\Omega}$ as a function of lower-dimensional parameters, e.g. as a function of a single autoregressive parameter $\rho$ or as a function of several separable covariance matrices $\boldsymbol{\Omega} = \boldsymbol{\Omega}_K \otimes \cdots \otimes \boldsymbol{\Omega}_1$ which correspond to variation along different modes of the matrix or tensor of regression coefficients $\boldsymbol{B}$. However, the general approach to maximum marginal likelihood estimation of $\boldsymbol{\Omega}$ remains the same, with minimzation over $\boldsymbol{\Omega}$ in the second step replaced by minimization over the lower-dimensional components.

### *4.4.2 Method of Moments*

In practice, maximum marginal likelihood estimation of variance components can converge prohibitively slowly (Roy and Chakraborty, 2016). Furthermore, it requires repeated simulation from the posterior distribution of $\boldsymbol{\beta}$ for different values of the unknown parameters $\boldsymbol{\phi}$, $\boldsymbol{\Omega}$ and in the case of the SPN prior $\boldsymbol{\Psi}$, which can be prohibitively computationally demanding when the data are high dimensional. Fortunately, method of moments type estimates of the unknown variance components can be obtained under the SNG and SPB priors for fixed $c$ and $q$ respectively, and under the SPN prior when the symmetric formulation is used. As noted in the Introduction, the prior moments are easy to compute under all three priors

$$\mathbb{E}\left[\boldsymbol{\beta}\right] = \mathbb{E}\left[\boldsymbol{s}\right] \circ \mathbb{E}\left[\boldsymbol{z}\right] \text{ and } \boldsymbol{\Sigma} = \mathbb{E}\left[\boldsymbol{s}\boldsymbol{s}'\right] \circ \mathbb{E}\left[\boldsymbol{z}\boldsymbol{z}'\right].$$

Furthermore, under the sSPN, SNG and SPB priors, the unknown parameters are determined by second order moments of $\boldsymbol{\beta}$. When $y$ is linearly related to $\boldsymbol{X}\boldsymbol{\beta}$, a positive semidefinite estimate of $\boldsymbol{\Sigma}$ can be obtained using methods from Perry (2017). When a generalized linear model with a nonlinear relationship between $\boldsymbol{X}\boldsymbol{\beta}$ and $\boldsymbol{y}$ is used, we are not able to take such a simple approach. However, the ease of computing prior moments under the SPN, SNG and SPB priors suggests that an estimate of $\boldsymbol{\Sigma}$ could be computed approximately either under a multivariate normal prior for $\boldsymbol{\beta}$ using maximum marginal likelihood or using an EM algorithm with a Laplace approximation to perform the expectation step. Henceforth, we refer to the former as normal Gibbs-within-EM estimates and the latter as normal LA-within-EM estimates.

Under the sSPN prior, estimates of $\boldsymbol{\Omega}$ and $\boldsymbol{\Psi}$ can be obtained from an estimate $\hat{\boldsymbol{\Sigma}}$ by projecting $\sqrt{\left|\hat{\boldsymbol{\Sigma}}\right|}$ and sign $\left\{\hat{\boldsymbol{\Sigma}}\right\}\sqrt{\left|\hat{\boldsymbol{\Sigma}}\right|}$ onto the positive semi-definite cone, where $\sqrt{\cdot}$, $|\cdot|$ and sign $\{\cdot\}$ are applied elementwise. Under the SNG and SPB priors, an estimate of $\boldsymbol{\Omega}$ can be obtained from an estimate $\hat{\boldsymbol{\Sigma}}$ by projecting $\hat{\boldsymbol{\Sigma}}/\left(\left(1-\mathbb{E}\left[s_j\right]^2\right)\boldsymbol{I}_p+\mathbb{E}\left[s_j\right]^2\mathbf{1}_p\mathbf{1}_p'\right)$ onto the positive semi-definite cone, where '/' is applied elementwise, $\mathbb{E}\left[s_j\right]^2=c^{-1}(\Gamma(c+1/2)/\Gamma(c))^2$ under the SNG prior and $\mathbb{E}\left[s_j\right]^2=(\pi/2)(\Gamma(2/q)/\sqrt{\Gamma(1/q)\Gamma(3/q)})^2$ under the SPB prior.

### *4.4.3 Fully Bayes Inference*

Alternatively, fully Bayes inference with prior distributions over unknown prior parameters could be performed. For all three priors, a conjugate inverse-Wishart prior for $\boldsymbol{\Omega}$ is a natural choice. For the SPN prior, a conjugate inverse-Wishart prior for $\boldsymbol{\Psi}$ is likewise a natural choice.

## *4.5 Application*

We assume $\boldsymbol{\Sigma}$ is separable, i.e. $\boldsymbol{\Sigma}=\boldsymbol{\Sigma}_2\otimes\boldsymbol{\Sigma}_1$, where $\boldsymbol{\Sigma}_2$ is a $8\times 8$ covariance matrix that characterizes the relationships of regression coefficients $\boldsymbol{B}$ across channels and $\boldsymbol{\Sigma}_1$ is a $208\times 208$ covariance matrix that characterizes the relationship of regression coefficients $\boldsymbol{B}$ over time. Because we only observe a single realization of $\boldsymbol{B}$, we assume that $\boldsymbol{\Sigma}_1$ is has an autoregressive structure of order one. Because the scales of $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ are not separately identifiable, we assume that $\boldsymbol{\Sigma}_1$ can be written as a function of a single autoregressive parameter $\rho$ with $\sigma_{1,ij}=\rho^{|i-j|}$ without loss of generality. We include an intercept $\gamma$ in the linear model and assume an improper uniform prior $\gamma\sim\text{uniform}\left(-\infty,\infty\right)$. We perform empirical Bayes inference, with fixed $\boldsymbol{\rho}$ and $\boldsymbol{\Sigma}_2$ estimated from the data, as well as fully Bayes inference with prior distributions assumed for $\boldsymbol{\Omega}$ and, in the case of the SPN prior, $\boldsymbol{\Psi}$. We note that autoregressive structure is preserved when the elementwise product of two autoregressive matrices is computed, i.e. if $\boldsymbol{\Omega}_1$ and $\boldsymbol{\Psi}_1$ have autoregressive structure of order one with parameters $\rho_{\boldsymbol{\Omega}}$ and $\rho_{\boldsymbol{\Psi}}$, then $\boldsymbol{\Sigma}_1$ has autoregressive structure of order one with parameter $\rho=\rho_{\boldsymbol{\Omega}}\rho_{\boldsymbol{\Psi}}$. In contrast, autoregressive and separable structures are not preserved

by multiplying an exchangeable matrix, $\mathbb{E}[\boldsymbol{s}\boldsymbol{s}']$, and an autoregressive or separable matrix.

### *4.5.1 Empirical Bayes Estimation*

Empirical Bayes results are based on estimates of $\rho$ and $\boldsymbol{\Sigma}_2$ obtained under using the normal LA-within-EM approach described in Section 4.4.2. We make this decision because maximum marginal likelihood estimation is challenging for such high dimensional data, in this case with $p = 1,644$. The empirical Bayes estimates suggest strong autocorrelation of elements of $\boldsymbol{B}$ over time with $\hat{\rho} = 0.43$. Figure 4.10 shows the estimated variances and correlations of $\hat{\boldsymbol{\Sigma}}_2$, which likewise suggests heterogeneity of $\boldsymbol{B}$ magnitudes across channels and strong positive and negative correlations across channels. We compare estimates of $\boldsymbol{\beta}$ obtained under structured shrinkage priors to estimates of $\boldsymbol{\beta}$ obtained under the corresponding independent priors, with $\rho = 0$ and $\boldsymbol{\Sigma}_2 = \sigma \boldsymbol{I}_{p_2}$. We estimate $\hat{\sigma} = 4.18 \times 10^{-4}$ using the same LA-within-EM approach. In this setting, we found that the Pólya-Gamma latent variable of Polson et al. (2013) approach outperformed elliptical slice sampling for simulating from $p_{\boldsymbol{\beta}|\boldsymbol{s},\boldsymbol{y}}(\boldsymbol{\beta}|\boldsymbol{s}, \boldsymbol{X}, \boldsymbol{y}, \boldsymbol{\phi}, \boldsymbol{\Omega})$ in terms of speed and effective sample size. Accordingly, the results presented were computed using Pólya-Gamma latent variables.

The Gibbs samplers based on independent priors appear to mix relatively well, with a minimum effective sample size for $\gamma$ and $\boldsymbol{\beta}$ of $1,211$ based on $5,000$ samples from the posterior. The Gibbs samplers based on structured priors mix much less well, with minimum expected sample sizes of $3,367$ under a multivariate normal prior, $5,000$ under the SPN prior, 196.49, $2,739.58$, 16.34 and 60.45 under the SNG priors with $c = 0.3$, $c = 0.5$, $c = 1$ and $c = 10$, respectively, and 5000.00, $3,901.81$, $5,000.00$ and 25.89 under the SPB prior with $q = 0.65$, $q = 0.78$, $q = 1$ and $q = 1.75$, respectively. Poor mixing is especially evident when comparing posterior medians under the SNG prior with $c = 1$ and the SPB prior with $q = 1$, which should be identical.

Figure 4.11 shows posterior medians for $\boldsymbol{B}$ under independent and corresponding structured priors. Unsurprisingly given that we estimate a smaller overall variance under an independent prior, we observe that entries of $\boldsymbol{B}$ are shrink more aggressively under the in-

dependent prior. We also observe little structure under the independent priors. Under the structured priors, we observe clear and strong structure of $\boldsymbol{B}$ over time and across channels under all of the structured priors. The posterior medians under the structured priors are easier to interpret, as they indicate that the behavior of EEG measurements at several distinct time points on a subset of channels is relevant to prediction.

Figure 4.12 shows ROC curves for using the posterior median fitted values to classify 100 held out trials for a subset of priors. Although not shown here, we note that in all cases the structured shrinkage priors outperform their independent counterparts. We observe that for this data, several of the structured shrinkage priors slightly outperform the multivariate SNO (multivariate normal) prior, indicating that there can be classification benefits from using a more complex structured shrinkage prior.

### *4.5.2 Fully Bayes Estimation*

Fully Bayes results are based on assuming $\boldsymbol{\Omega}_2^{-1} \sim \text{Wishart}\,(10, \boldsymbol{I}_8)$, $\rho_{\boldsymbol{\Omega}} \sim \text{uniform}\,(-1, 1)$ and, when using the SPN prior, $\boldsymbol{\Psi}_2^{-1} \sim \text{Wishart}\,(10, \boldsymbol{I}_8)$, $\rho_{\boldsymbol{\Psi}} \sim \text{uniform}\,(-1, 1)$.

Regarding simulation from $p_{\boldsymbol{\beta}|\boldsymbol{s},\boldsymbol{y}}\,(\boldsymbol{\beta}|\boldsymbol{s}, \boldsymbol{X}, \boldsymbol{y}, \boldsymbol{\Omega}, \boldsymbol{\psi})$, we find that elliptical slice sampling is more stable than the Pólya-Gamma latent variable approach in this case, although neither performs well for data on this scale Polson et al. (2013). We compute $\tilde{\boldsymbol{u}}$ using a single cycle of block coordinate descent each iteration starting from the previous value of $\tilde{\boldsymbol{u}}$, with elements of $\boldsymbol{\beta}$ split into approximately 67 groups of 25 elements per group. We use a weak convergence threshold of $\frac{1}{q_k}\left|\left|\boldsymbol{\beta}_k^{(i+1)} - \boldsymbol{\beta}_k^{(i)}\right|\right|_1 < 10^{-2}$ for each coordinate descent iteration, where $k$ indexes the block, $q_k$ is the number of elements in block $k$ and $i$ indexes Newton-Raphson iterations within each block and we permit a maximum of 100 Newton-Raphson iterations per cycle for each block of elements of $\boldsymbol{\beta}$. We assume $\tilde{\boldsymbol{V}} = \boldsymbol{I}_p$. We simulate from the full conditional distributions of $\rho_{\boldsymbol{\Omega}}$ and $\rho_{\boldsymbol{\Psi}}$ using univariate slice sampling as described in the appendix.

Unsurprisingly given the choices of priors for $\boldsymbol{\Omega}_2$ and $\boldsymbol{\Psi}$ centered at $\boldsymbol{I}_{p_2}$ and the priors for $\rho_{\boldsymbol{\Omega}}$ and $\rho_{\boldsymbol{\Psi}}$ centered at 0, we estimate larger overall variances and less structure than in the empirical Bayes setting. Posterior medians of the overall variances range from 0.20 to

0.62 and posterior medians of the marginal autoregressive parameter $\rho$ rangr from $-0.005$ to 0.008.

Figure 4.13 shows posterior median estimates of elements of $\boldsymbol{B}$ under a multivariate normal prior (SNO) and nine choices of structured shrinkage priors.They are less structured than the empirical Bayes posterior medians, however they nonetheless show some evidence of similarity across channels and over time. The Gibbs samplers do not mix well, with minimum effective sample sizes for $\gamma$, $\boldsymbol{\beta}$ and elements of $\boldsymbol{\Sigma}$ of of 1.44 under a multivariate normal prior, 3.31 under the SPN prior, 3, 1.32, 1.33 and 1.07 under the SNG priors with $c = 0.3$, $c = 0.5$, $c = 1$ and $c = 10$, respectively, and 2.14, 1.34, 1.65 and 3.27 under the SPB prior with $q = 0.65$, $q = 0.78$, $q = 1$ and $q = 1.75$, respectively, based on $51,000$ samples from the posterior. Poor mixing is especially evident form the lack of correspondence between posterior medians under the SNG and SPB priors with $c = 1$ and $q = 1$, which are identical priors, and under the SNO prior and SNG priors with $c = 10$ and $q = 1.75$, which are very similar priors. Despite the poor mixing, we note that the posterior medians under the SNG priors conform to our prior expectations insofar as they all display similar structure but increasing shrinkage of individual elements of $\boldsymbol{\beta}$ as $c \to 0$.

Figure 4.14 shows ROC curves for using the posterior median fitted values to classify 100 held out trials for a subset of priors. We observe that for this data, the SNO (multivariate normal) prior tends to perform best, suggesting that the fully Bayes implementation of the structured shrinkage priors does not improve performance. This could be due to poor mixing under the SPN and SNG priors, poor choice of priors on $\boldsymbol{\Omega}$, $\boldsymbol{\Psi}$, $\rho_{\boldsymbol{\Omega}}$ and $\rho_{\boldsymbol{\Psi}}$, or too poor fit of the more complex structured shrinkage priors for this data.

### *4.6 Discussion*

We introduce three novel structured shrinkage priors for regression coefficients in generalized linear models, the SPN, SNG and SPB priors. We carefully explore and compare properties of these prior distributions and show that they can encourage both shrinkage and near or exact sparsity and structure, which can be difficult to simultaneously model using existing prior

distributions. We provide a parsimonious and general approach to posterior mode estimation and full posterior distribution simulation based on univariate and elliptical slice sampling that not only allows for straightforward simulation of elements of $\boldsymbol{s}$ from nonstandard full conditional distributions but also is tailored to especially high dimensional problems where $n$ or $p$ are large, as it avoids large matrix inversions and crossproducts in the simulation from the full conditional distribution of $\boldsymbol{\beta}$. Last, we consider the problem of estimating logistic regression coefficients in a P300 speller prediction problem, in which we are interested in relating indicators of whether or not a subject is viewing a specific target letter $\boldsymbol{y}$ to a multivariate time series of contemporaneous EEG measurements. We demonstrate how all three prior distributions can improve interpretability of estimated logistic regression coefficients and out of sample prediction relative to multivariate normal or independent priors.

This material has several additional applications. Throughout, we have focused on the development of structured shrinkage prior distributions for regression coefficients in generalized linear models, however the same distributions could be used to model correlated and heavy-tailed errors as in Finegold and Drton (2011), as alternatives to Gaussian process models or to construct shrinkage priors for other quantities, e.g. covariance matrices (Daniels and Pourahmadi, 2002). Also, the key insight that allows us to construct a tractable Gibbs sampler for simulating from $p_{\boldsymbol{s}|\boldsymbol{\beta}}(\boldsymbol{s}|\boldsymbol{\beta},\boldsymbol{\theta})$ under the SPN, SNG and SPB priors can be to perform posterior inference under other novel structured generalizations of *any* shrinkage priors that have normal scale-mixture representations. This includes the popular horseshoe prior, as well as others (Polson and Scott, 2010; Bhattacharya et al., 2015). This follows from the fact that a change of variables facilitates the use of elliptical slice sampling for simulating from the full conditional distribution $p_{\boldsymbol{s}|\boldsymbol{\beta}}(\boldsymbol{s}|\boldsymbol{\beta},\boldsymbol{\theta})$ for *arbitrary* priors $p_{\boldsymbol{s}}(\boldsymbol{s}|\boldsymbol{\theta})$. Last, when $\boldsymbol{X}$ is full rank the material presented in this paper could be used to develop novel $g$-priors by setting $\boldsymbol{\Omega} \propto (\boldsymbol{X}'\boldsymbol{X})^{-1}$ (Zellner, 1986). Such priors might be useful for addressing problems that arise when the Laplace prior or lasso penalty are used when columns of $\boldsymbol{X}$ are correlated.

One limitation of this work is that the approach we develop for simulating from $p_{\boldsymbol{r}|\boldsymbol{\beta}}(\boldsymbol{r}|\boldsymbol{\beta},\boldsymbol{\theta})$

can mix very slowly. This is especially evident under the SPB prior when a fully Bayes approach is taken. A natural next step would be to approach simulating from $p_{\boldsymbol{r}|\boldsymbol{\beta}}(\boldsymbol{r}|\boldsymbol{\beta},\boldsymbol{\theta})$ as we approach simulating from $p_{\boldsymbol{\beta}|\boldsymbol{s}},(\boldsymbol{\beta}|\boldsymbol{s},\boldsymbol{y},\boldsymbol{X},\boldsymbol{\Omega},\boldsymbol{\phi})$, i.e. pre- and post-multiply the full conditional distribution of $\boldsymbol{r}$ by the kernel of a multivariate normal density that is either exactly equal to or an approximation of the Laplace approximation of the full conditional distribution $p_{\boldsymbol{r}|\boldsymbol{\beta}}(\boldsymbol{r}|\boldsymbol{\beta},\boldsymbol{\theta})$. This may be more feasible for some prior distributions for $\boldsymbol{s}$ than others, and would also require the use of algorithms tailored to the specific choice of prior distribution for $\boldsymbol{s}$. Another limitation is that we treat $c$ and $q$ as fixed under the SNG and SPB priors, respectively. In future work, we might consider maximum marginal likelihood estimation of or assuming prior distributions for $c$ and $q$. However, we have observed especially poor mixing especially poor mixing when $c$ and $q$ are not fixed a priori. Accordingly, if prior distributions were assumed for $c$ and $q$, an improved approach to simulating from $p_{\boldsymbol{r}|\boldsymbol{\beta}}(\boldsymbol{r}|\boldsymbol{\beta},\boldsymbol{\theta})$ would be even more necessary. One more limitation is that the use of *general* Gibbs sampling procedures that apply to *any* generalized linear model for $\boldsymbol{y}$ and *any* scale prior $p_{\boldsymbol{s}}(\boldsymbol{s}|\boldsymbol{\theta})$ does result in a loss of computational efficiency when specialized algorithms are available. We observe that using Pólya-Gamma latent variables as described in Polson et al. (2013) can produce larger effective sample sizes per iteration of the Gibbs sampler when a logistic regression model is assumed for $\boldsymbol{y}$. Additionally, when the full conditional distribution $p_{\boldsymbol{\beta}}(\boldsymbol{\beta}|\boldsymbol{y},\boldsymbol{X},\boldsymbol{\phi})$ is quadratic in $\boldsymbol{\beta}$, specialized algorithms for simulating from the full posterior distribution under the SPN and SNG priors are available. Under the SPN prior, simulating from the full conditionals $p_{\boldsymbol{z}|\boldsymbol{s}}(\boldsymbol{z}|\boldsymbol{s},\boldsymbol{y},\boldsymbol{X},\boldsymbol{\Omega})$ and $p_{\boldsymbol{s}|\boldsymbol{z}}(\boldsymbol{s}|\boldsymbol{z},\boldsymbol{y},\boldsymbol{X},\boldsymbol{\Psi})$ and under the SNG prior, importance sampling from $p_{s_j|\boldsymbol{s}_{-j}}(s_j|\boldsymbol{s}_{-j},\boldsymbol{z},\boldsymbol{X},\boldsymbol{y},\boldsymbol{\phi},\boldsymbol{\theta})$ as in Finegold and Drton (2011) or Gibbs sampling from $p_{s_j|\boldsymbol{s}_{-j}}(s_j|\boldsymbol{s}_{-j},\boldsymbol{\beta},\boldsymbol{\Omega})$ using the auxiliary variable method developed by Damien et al. (1999) as described in the appendix can produce larger effective sample sizes per iteration of the Gibbs sampler. That said, both of these specialized algorithms also require more time per iteration than the more general slice sampling approach we present when $n$ or $p$ is very large even when elements of $\boldsymbol{\beta}$, $\boldsymbol{z}$ or $\boldsymbol{s}$ are simulated in blocks or elementwise.

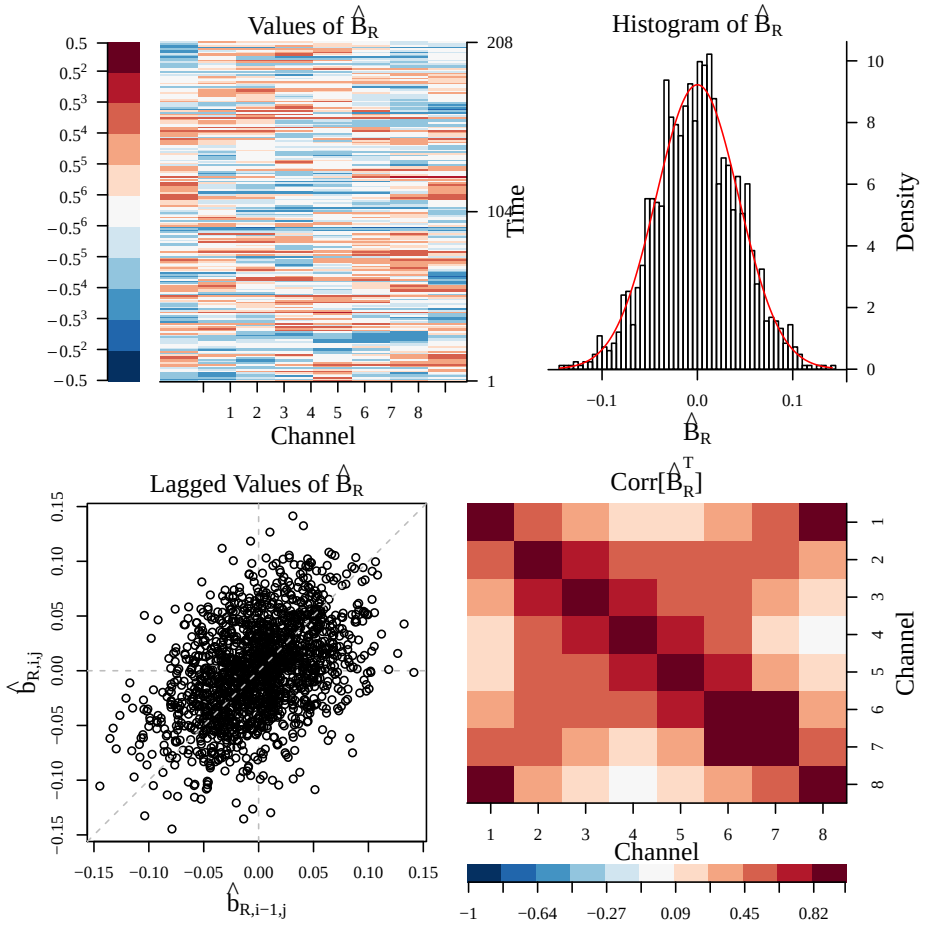



Figure 4.3: The first panel shows the values of a $208 \times 8$ ridge estimate $\hat{\boldsymbol{B}}_R$ for a single subject based on $n = 140$ trials which minimizes $h(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{\phi}) + \frac{1}{2}\boldsymbol{\beta}'\boldsymbol{\beta}$. The second panel shows a histogram of elements of $\hat{\boldsymbol{B}}_R$ plotted against a normal density with the same mean and variance as elements of $\hat{\boldsymbol{B}}_R$. The third panel plots pairs of elements of $\hat{\boldsymbol{B}}_R$ that correspond to consecutive time points $(\hat{b}_{R,ij}, \hat{b}_{R,(i+1)j})$ against a gray dashed 45° line. The last panel shows correlations of elements of $\hat{\boldsymbol{B}}_R$ across channels.

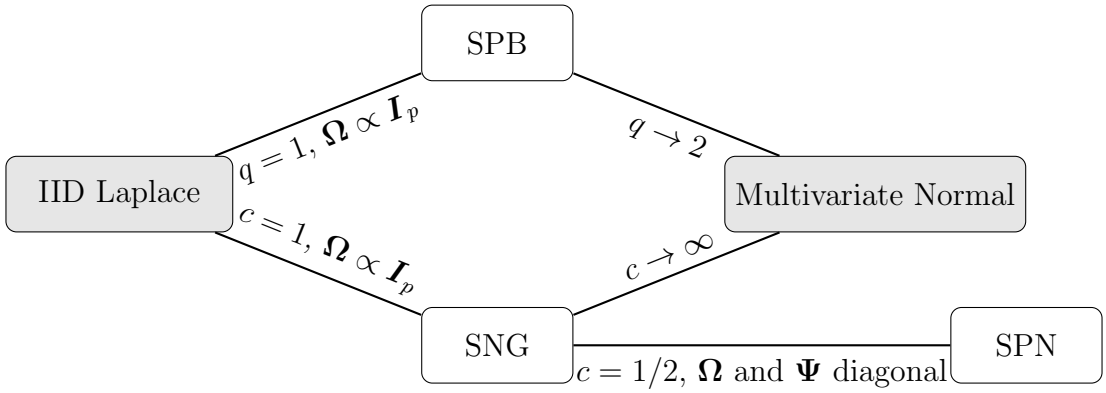



Figure 4.4: Relationships between structured product normal (SPN), structured normal-gamma (SNG) and structured power/bridge (SPB) shrinkage priors and independent, identically distributed (IID) Laplace and multivariate normal priors.

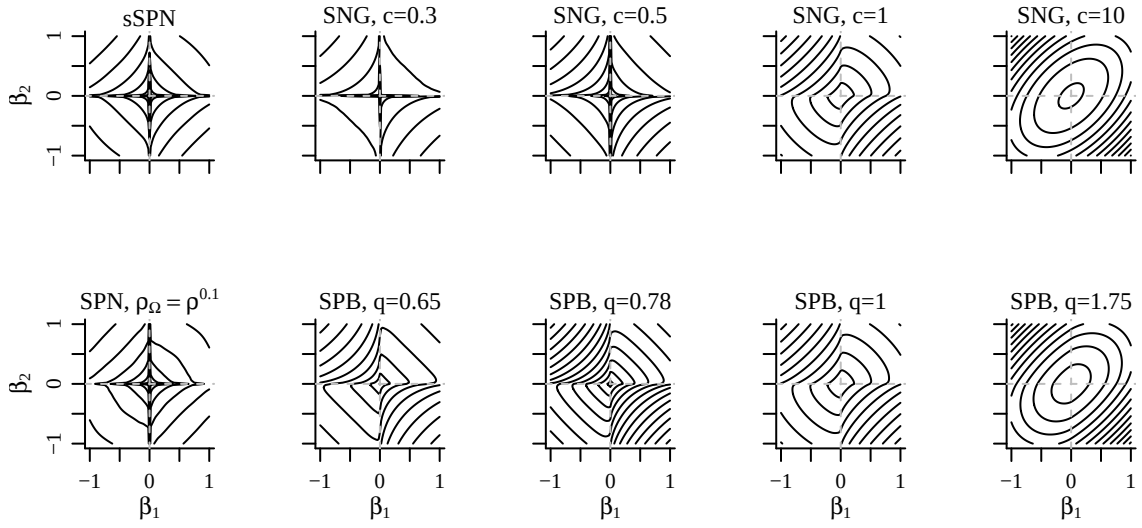



Figure 4.5: Log-likelihood contours for $\boldsymbol{\beta} \in \mathbb{R}^2$ with unit marginal prior variance and marginal prior correlation $\rho = 0.5$.

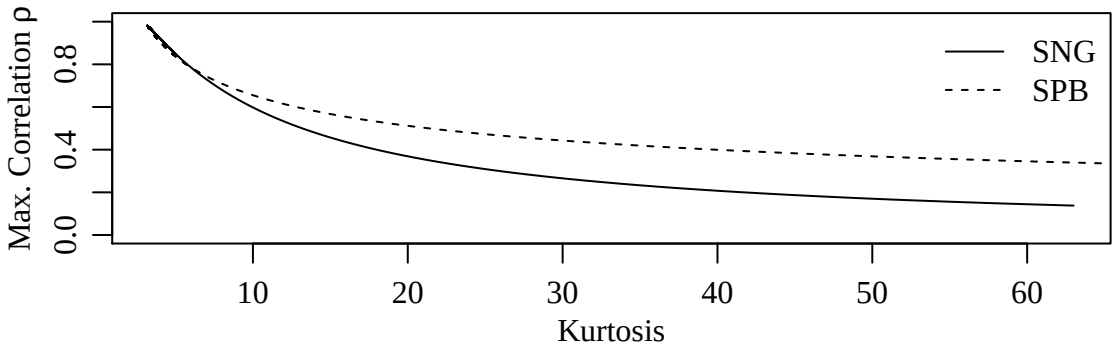



Figure 4.6: Maximum marginal prior correlation $\rho$ for $\boldsymbol{\beta} \in \mathbb{R}^2$ as a function of kurtosis.

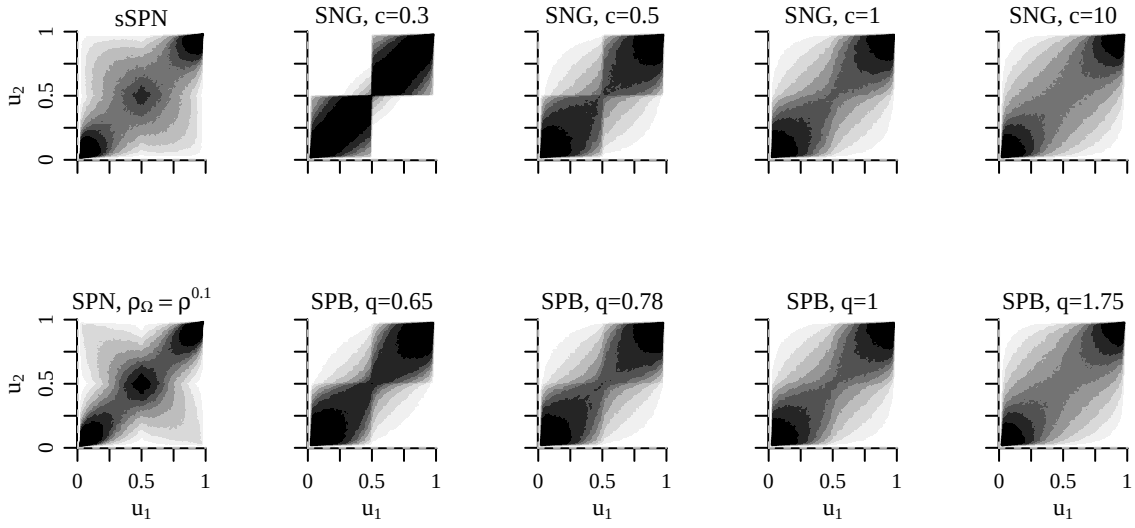



Figure 4.7: Copula estimates for $\boldsymbol{\beta} \in \mathbb{R}^2$ with unit marginal prior variances and marginal prior correlation $\rho = 0.5$.

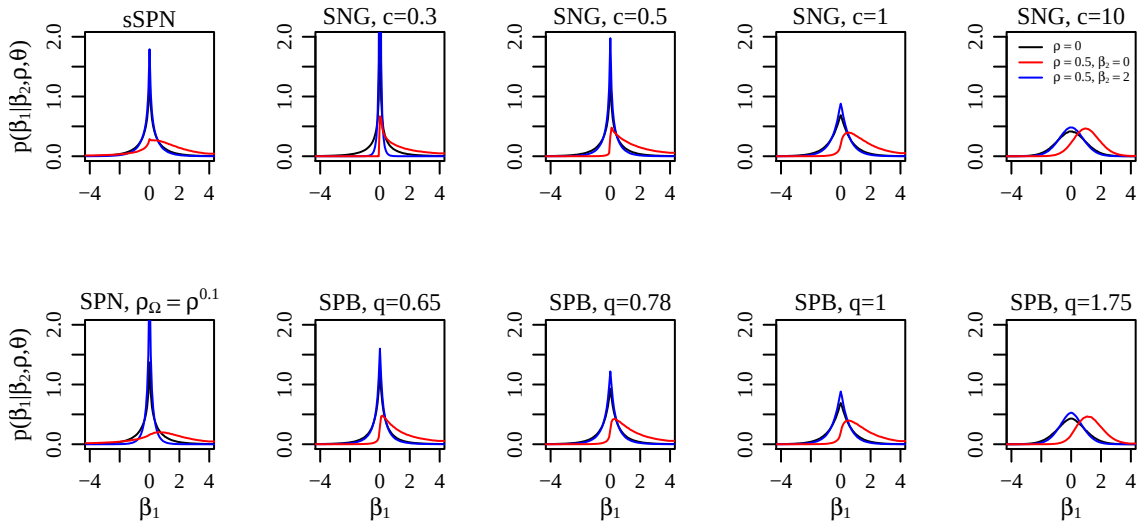



Figure 4.8: Univariate conditional prior distributions $p_{\beta_1|\beta_2}\left(\beta_1|\beta_2,\rho,\boldsymbol{\theta}\right)$ for $\boldsymbol{\beta}\in\mathbb{R}^2$ with unit marginal prior variances, marginal prior correlation $\rho\in\{0,0.5\}$ and $\beta_2\in\{0,2\}$.

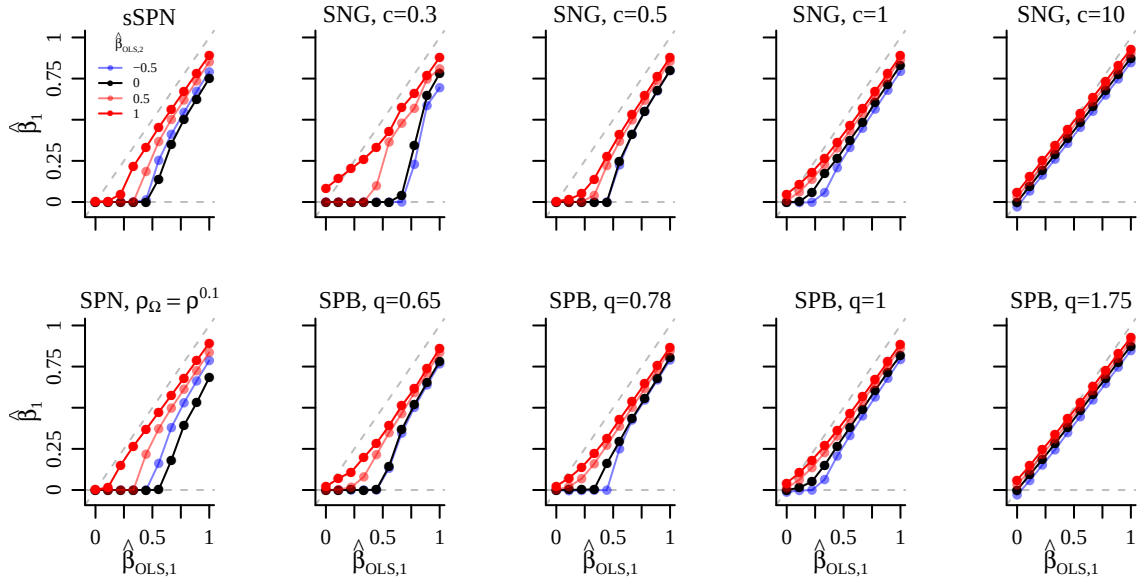



Figure 4.9: Approximate mode-thresholding functions computed according to (4.1) for $\boldsymbol{\beta}\in\mathbb{R}^2$ with unit marginal prior variances, marginal prior correlation $\rho=0.5$, noise variance $\phi^2=0.1$, $\hat{\beta}_{OLS,2}\in\{-0.5,0,0.5,1\}$ and $\hat{\beta}_{OLS,1}\in(0,1)$.

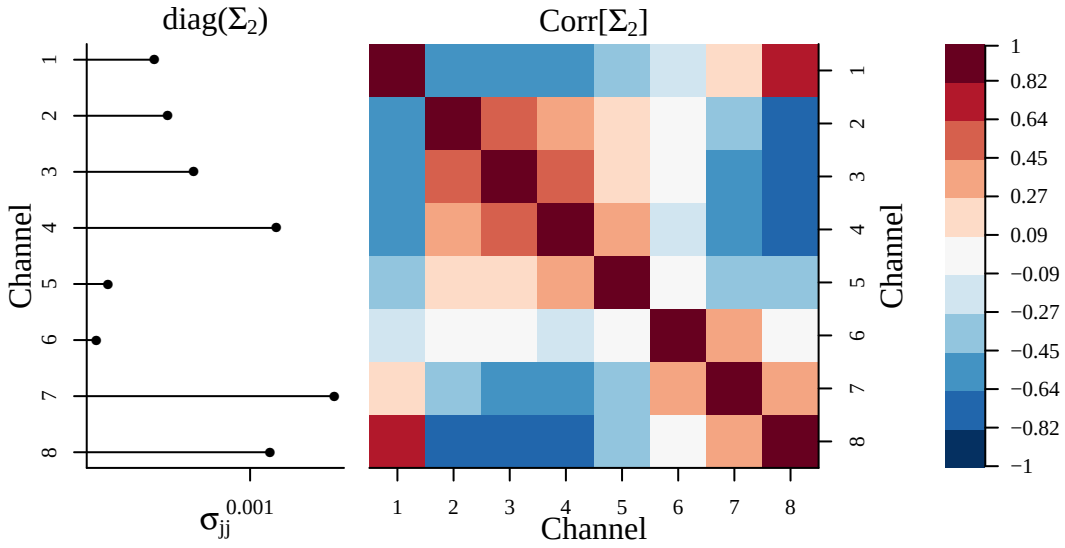



Figure 4.10: Normal LA-within-EM estimates of variances and correlations of $\mathbf{\Sigma}_2$.

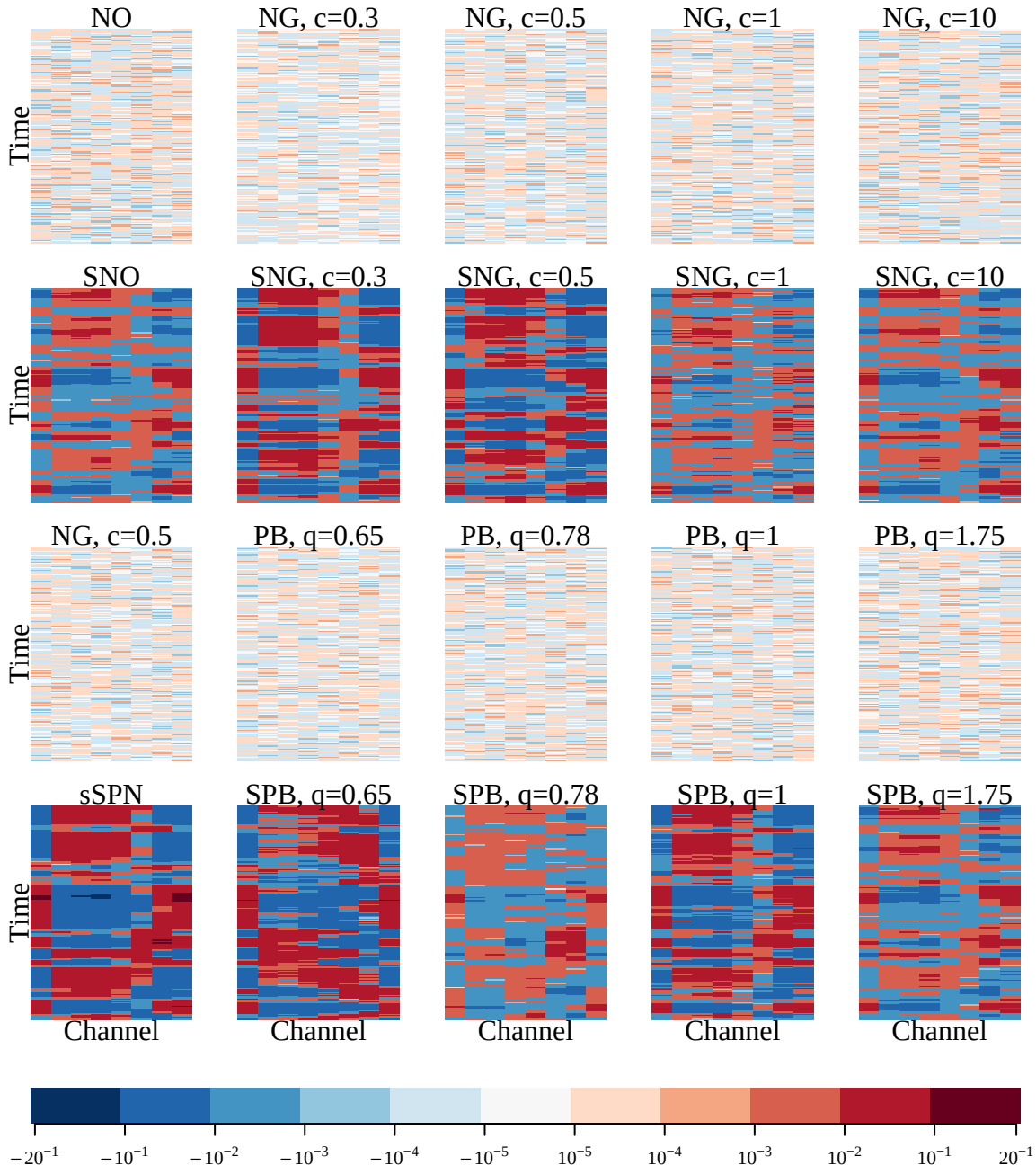



Figure 4.11: Approximate posterior medians of empirical Bayes estimates of $\boldsymbol{B}$ under independent and corresponding structured priors.

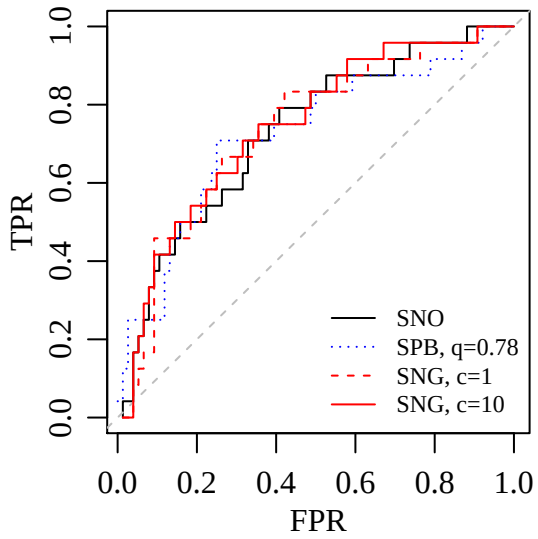



Figure 4.12: ROC curves for held out test data using empirical Bayes approximate posterior medians of the fitted values.

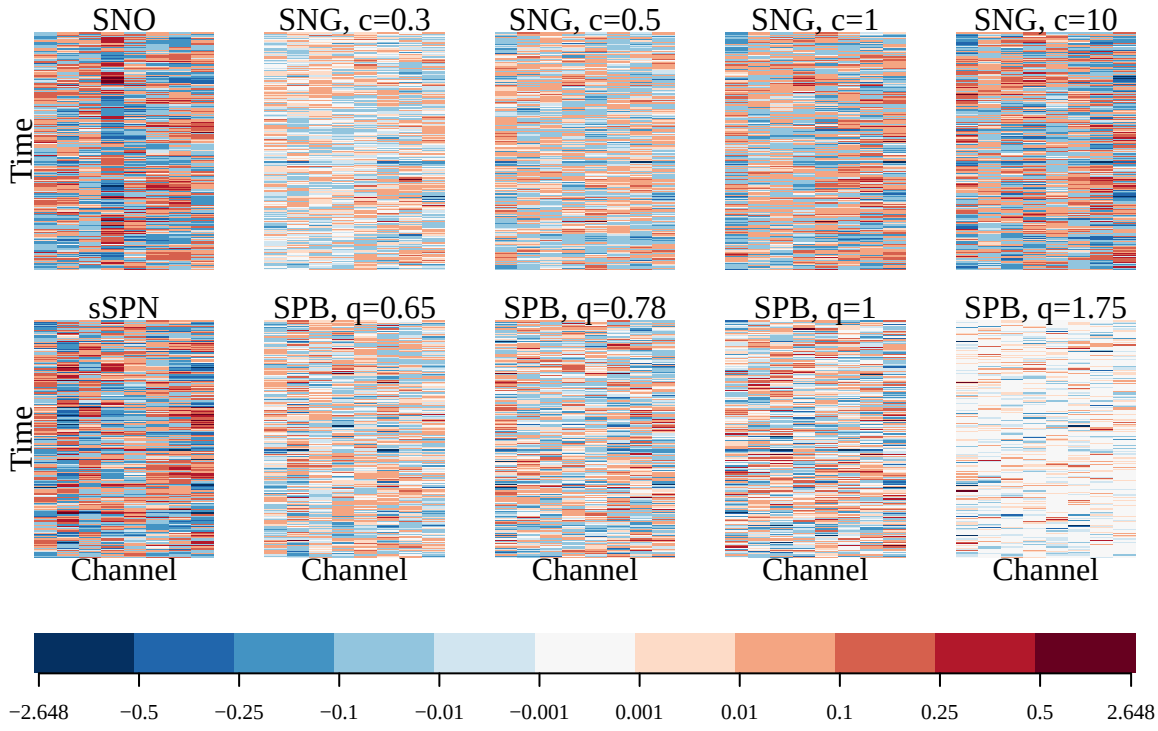



Figure 4.13: Approximate posterior medians of fully Bayes estimates of $\boldsymbol{B}$.

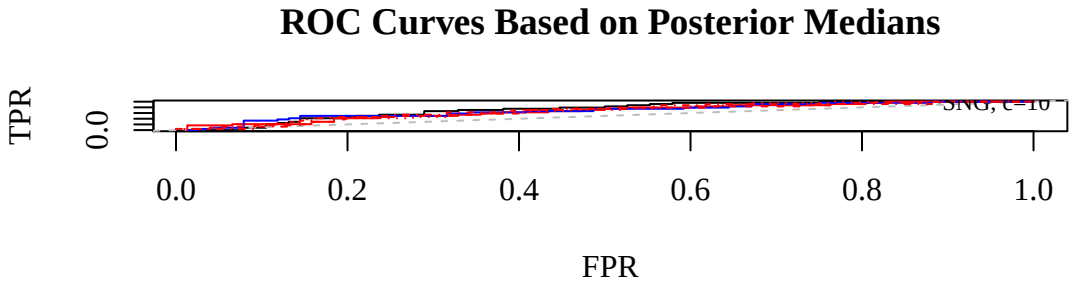



Figure 4.14: ROC curves for held out test data using fully Bayes approximate posterior medians of fitted values.

# Chapter 5

# CONCLUSIONS AND FUTURE WORK

This thesis has considered advantages and disadvantages of interpreting penalized regression problems from a model-based perspective. We show that treating penalties as priors facilitates estimation of unknown prior or tuning parameters in settings where other approaches either fail or are impractical, enables joint estimation of many unknown prior parameters simultaneously, and aids in the construction of novel, interpretable priors with desirable properties. We also show that treating penalties as priors without carefully considering whether or not a certain prior is appropriate can be perilous. We encourage the use of broader classes of priors to protect against poor prior specification.

In the first two chapters, we consider problems that require estimation of a one-to-three dimensional vector of unknown likelihood and prior distribution parameters for which the maximum marginal likelihood estimates are not computationally feasible to obtain. We are are able to develop moment-based approaches to these problems that are easy and quick to compute. However, such moment-based approaches are likely less efficient than likelihood-based approaches as well as challenging to extend to settings where the response $\boldsymbol{y}$ is related nonlinearly to $\boldsymbol{\beta}$, e.g. where generalized linear models for $\boldsymbol{y}$ are used or where $\boldsymbol{\beta} = \text{vec}\,(\boldsymbol{B})$ is the vectorization of a matrix for which a rank shrinkage prior is assumed. Doss (2010) provides an MCMC assisted approach to the approximation of intractable likelihood surfaces that could be applied to obtain maximum marginal likelihood estimates of the unknown likelihood and prior distribution parameters. Regarding the second chapter, "Testing Sparsity-Inducing Penalties," maximum marginal likelihood estimation of the unknown likelihood and prior distribution parameters could be used to construct a likelihood ratio test of the null hypothesis that a Laplace prior for regression coefficients $\boldsymbol{\beta}$ is appropriate if the distribution

of the likelihood test statistic under the null could be derived.

Material in the second chapter, "Testing Sparsity-Inducing Penalties," suggests the development of pathwise coordinate descent algorithms for $\ell_q$ penalties. Previous work has found that pathwise coordinate descent algorithms can be useful for solving penalized regression problems under other penalties that span $\ell_0$ to $\ell_1$ penalties, but has also argued that the family of $\ell_q$ penalties are not amenable to pathwise coordinate descent algorithms (Mazumder et al., 2011). Observations we make in "Testing Sparsity-Inducing Penalties" suggest that the properties that Mazumder et al. (2011) argued make $\ell_q$ penalties ill suited to pathwise coordinate descent are not inherent to $\ell_q$ penalties but rather are a consequence of the standard parametrization used in the optimization literature, $\lambda \left|\left|\boldsymbol{\beta}\right|\right|_q^q$. As a result, we may conduct a small additional project that shows how the parametrization of the $\ell_q$ penalty used in 'Testing Sparsity-Inducing Penalties" $2\left(\frac{\Gamma(3/q)}{\Gamma(1/q)\tau^2}\right)^{q/2}\left|\left|\boldsymbol{\beta}\right|\right|_q^q$ can be used to construct pathwise coordinate descent algorithms for fast computation of $\ell_q$ penalized estimates and entire regularization surfaces.

The third chapter, "Structured Shrinkage Priors," suggests several avenues for future work. Because a change of variables facilitates the use of elliptical slice sampling for simulating from the full conditional distribution $p_{\boldsymbol{s}|\boldsymbol{\beta}}\left(\boldsymbol{s}|\boldsymbol{\beta},\boldsymbol{\theta}\right)$ for *arbitrary* priors $p_{\boldsymbol{s}}\left(\boldsymbol{s}|\boldsymbol{\theta}\right)$, one natural extension might be to consider the construction of novel correlated priors for $\boldsymbol{s}$ that generalize the independent normal and polynomially tilted priors used to generate SNG and SPB priors. The main challenge would be in defining density functions for $\boldsymbol{s}$ that correspond to integrable multivariate distributions with reasonable properties. The slice sampling approach we take for simulating from the full conditional distribution $p_{\boldsymbol{s}|\boldsymbol{\beta},\boldsymbol{\theta}}\left(\boldsymbol{s}|\boldsymbol{\beta},\boldsymbol{\theta}\right)$ could be used even more generally to simulate from full conditional distributions of correlation matrices and elements on the Stiefel or Grassman manifolds in a generic way. Consider the problem of simulating from a probability distribution over $p \times p$ correlation matrices $\boldsymbol{C}$ with density $p_{\boldsymbol{C}}\left(\boldsymbol{C}\right)$, which might arise in data fusion type problems where $p \times p$ sample covariance matrices $\boldsymbol{S}_1, \ldots, \boldsymbol{S}_n$ are observed, each of which measures relationships between the same $p$ units measured using different devices or at different times. We might want to model

each covariance matrix $\boldsymbol{S}_i$ as a function of its variances $\boldsymbol{v}_i$ and correlations $\boldsymbol{C}_i$, as they are interpretable quantities that we may have access to prior information about, however such an approach may require simulation from $p_{\boldsymbol{C}_i}(\boldsymbol{C}_i|\boldsymbol{v}_i)$. In general, it is challenging from distributions over correlation matrices because the distributions of freely varying elements $c_{ij}$ are often nonstandard *and* constrained as $\boldsymbol{C}$ is necessarily positive semidefinite. However, it is always possible to factor $p_{\boldsymbol{C}}(\boldsymbol{C})$ as

$$p_{\boldsymbol{C}}(\boldsymbol{C}) = \underbrace{|\boldsymbol{C}|^{(n-p-1)/2}}_{(*)} p_{\boldsymbol{C}}(\boldsymbol{C}) / |\boldsymbol{C}|^{(n-p-1)/2}$$

where $n \geq p+1$ and $(*)$ corresponds to the kernel of the joint distribution of the elements of the $p \times p$ correlation matrix of $n$ vectors $\boldsymbol{x}_i \overset{i.i.d.}{\sim} \text{normal}(\boldsymbol{0}, \boldsymbol{I}_p)$. Accordingly, simulating values of $\boldsymbol{C}$ according to $p_{\boldsymbol{C}}(\boldsymbol{C})$ is equivalent to simulating the unconstrained values of an $n \times p$ matrix $\boldsymbol{X}$ according to

$$\underbrace{\exp\left\{-\frac{1}{2}\text{vec}(\boldsymbol{X})' \text{vec}(\boldsymbol{X})\right\}}_{(*)} p_{\boldsymbol{C}}(\boldsymbol{C}(\boldsymbol{X})) / \det(\boldsymbol{C}(\boldsymbol{X}))^{(n-p-1)/2},$$

where $\boldsymbol{C}(\boldsymbol{X})$ is the $p \times p$ sample correlation matrix of $\boldsymbol{X}$ and $(*)$ corresponds to the kernel $np$ independent standard normal densities.

One last avenue for future work related to "Structured Shrinkage Priors" is to reexamine priors that correspond to the fused lasso and other structured penalties that are in use and revisit the question of whether or not the unknown prior parameters can be related to moments of $\boldsymbol{\beta}$. It would be valuable to have a computationally simple and fast approach to estimating unknown prior parameters under these priors because computation of posterior summaries under these priors is well studied and straightforward, whereas computation under the structured shrinkage priors we introduce is challenging both because they correspond to nonconvex posterior mode optimization problems and because they require MCMC based inference. This is tricky because under priors corresponding to the fused lasso and other structured penalties, entries of $\boldsymbol{\Omega}$ are functions of random variables that enter linearly into $\boldsymbol{\Omega}^{-1}$ but not $\boldsymbol{\Omega}$. However, possible ease of computing $\mathbb{E}\left[\boldsymbol{\Omega}^{-1} \circ \left(\boldsymbol{s}^{-1} \circ (\boldsymbol{s}^{-1})'\right)\right]$ under these

priors suggests that we might be able to overcome this difficulty and relate moments of $\boldsymbol{\beta}$ to the unknown prior parameters.

# BIBLIOGRAPHY

Andrews, D. F. and C. L. Mallows (1974). Scale Mixtures of Normal Distributions. *Journal of the Royal Statistical Society. Series B 36*(1), 99–102.

Bernardo, J. and A. F. M. Smith (2000). *Bayesian Theory*. Chichester: Wiley.

Bhattacharya, A., D. Pati, N. S. Pillai, and D. B. Dunson (2015). Dirichlet Laplace Priors for Optimal Shrinkage. *Journal of the American Statistical Association 110*(512), 1479–1490.

Bleeker, F. E., R. J. Molenaar, and S. Leenstra (2012). Recent advances in the molecular understanding of glioblastoma. *Journal of Neuro-Oncology 108*(1), 11–27.

Box, G. E. P. (1953). A Note on Regions for Tests of Kurtosis. *Biometrika 40*(3/4), 465–468.

Box, G. E. P. and G. C. Tiao (1973). Bayesian Assessment of Assumptions. In *Bayesian Inference in Statistical Analysis*, Chapter 3, pp. 149–202. Reading, MA: Addison-Wesley Pub. Co.

Bredel, M., C. Bredel, D. Juric, G. R. Harsh, H. Vogel, L. D. Recht, and B. I. Sikic (2005). Functional Network Analysis Reveals Extended Gliomagenesis Pathway Maps and Three Novel MYC-Interacting Genes in Human Gliomas. *Cancer Research 65*(19), 8679–8689.

Bruneel, B. (2018). DE DUIKERKLOK EN DE P300-SPELLER.

Bühlmann, P. and S. van de Geer (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Heidelberg: Springer.

Candes, E. J., C. A. Sing-Long, and J. D. Trzasko (2013). Unbiased Risk Estimates for Singular Value Thresholding and Spectral Estimators. *IEEE Transactions on Signal Processing 61*(19), 4643–4657.

Caron, F. and A. Doucet (2008). Sparse Bayesian Nonparametric Regression. *International Conference on Machine Learning*, 88–95.

Carvalho, C. M., N. G. Polson, and J. G. Scott (2010). The horseshoe estimator for sparse signals. *Biometrika 97*(2), 465–480.

Castillo, I., J. Schmidt-Hieber, and A. W. Van Der Vaart (2015). Bayesian linear regression with sparse priors. *Annals of Statistics 43*(5), 1986–2018.

Choy, S. T. B. and J. S. K. Chan (2008). Scale Mixtures Distributions in Statistical Modelling. *Australian and New Zealand Journal of Statistics 50*(2), 135–146.

Cosset, É., S. Ilmjärv, V. Dutoit, K. Elliott, T. von Schalscha, M. F. Camargo, A. Reiss, T. Moroishi, L. Seguin, G. Gomez, J. S. Moo, O. Preynat-Seauve, K. H. Krause, H. Chneiweiss, J. N. Sarkaria, K. L. Guan, P. Y. Dietrich, S. M. Weis, P. S. Mischel, and D. A. Cheresh (2017). Glut3 Addiction Is a Druggable Vulnerability for a Molecularly Defined Subpopulation of Glioblastoma. *Cancer Cell 32*, 856–868.

Damien, P., J. Wakefield, and S. Walker (1999). Gibbs Sampling for Bayesian Non-Conjugate and Hierarchical Models by Using Auxiliary Variables. *Journal of the Royal Statistical Society. Series B (Statistical Methodology) 61*(2), 331–344.

Daniels, M. J. and M. Pourahmadi (2002). Bayesian analysis of covariance matrices and dynamic models for longitudinal data. *Biometrika 89*(3), 553–566.

de Tayrac, M., S. Lê, M. Aubry, J. Mosser, and F. Husson (2009). Simultaneous analysis of distinct Omics data sets with integration of biological knowledge: Multiple Factor Analysis approach. *BMC Genomics 10*(1), 32.

Demidenko, E. (2013). *Mixed Models: Theory and Applications with R* (Second ed.). Hoboken, New Jersey: John Wiley and Sons, Inc.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B 39*(1), 1–38.

Devroye, L. (2009). Random variate generation for exponentially and polynomially tilted stable distributions. *ACM Transactions on Modeling and Computer Simulation 19*(4), 1–20.

Diananda, P. H. (1949). Note on some properties of maximum likelihood estimates. *Mathematical Proceedings of the Cambridge Philosophical Society 45*(04), 536–544.

Díaz-Francés, E. and J. A. Montoya (2008). Correction to On the linear combination of normal and Laplace random variables, by Nadarajah, S., Computational Statistics, 2006, 21, 6371. *Computational Statistics 23*(4), 661–666.

Doss, H. (2010). Estimation of Large Families of Bayes Factors from Markov Chain Output. *Statistica Sinica 20*, 537–560.

Ducray, F., A. Idbaih, A. de Reyniès, I. Bièche, J. Thillet, K. Mokhtari, S. Lair, Y. Marie, S. Paris, M. Vidaud, K. Hoang-Xuan, O. Delattre, J. Y. Delattre, and M. Sanson (2008). Anaplastic oligodendrogliomas with 1p19q codeletion have a proneural gene expression profile. *Molecular Cancer 7*, 1–17.

Efron, B., T. Hastie, I. Johnstone, and R. Tibshirani (2004). Least Angle Regression. *The Annals of Statistics 32*(2), 407–499.

Fabrizi, E. and C. Trivisano (2010). Robust linear mixed models for Small Area Estimation. *Journal of Statistical Planning and Inference 140*(2), 433–443.

Fan, J. and R. Li (2001). Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties. *Journal of the American Statistical Association 96*(456), 1348–1360.

Fazel, M. (2002). *Matrix Rank Minimization with Applications*. Ph. d. thesis, Stanford University.

Figueiredo, M. A. T. (2003). Adaptive Sparseness for Supervised Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence 25*(9), 1150–1159.

Finegold, M. and M. Drton (2011). Robust graphical modeling of gene networks using classical and alternative t-distributions. *Annals of Applied Statistics 5*(2 A), 1057–1080.

Forney, E., C. Anderson, P. Davies, W. Gavin, B. Taylor, and M. Roll (2013). A Comparison of EEG Systems for Use in P300 Spellers by Users With Motor Impairments in Real-World Environments. *Proceedings of the Fifth International Brain-Computer Interface Meeting*.

Frank, I. E. and J. H. Friedman (1993). A Statistical View of Some Chemometrics Regression Tools. *Technometrics 35*(2), 109.

Fu, W. J. (1998). Penalized Regressions: The Bridge Versus the Lasso. *Journal of Computational and Graphical Statistics 7*(3), 397–416.

Gerard, D. and P. Hoff (2017). Adaptive Higher-order Spectral Estimators. *Electronic Journal of Statistics 11*(2), 3703–3737.

Ghosh, D., I. V. Ulasov, L. Chen, L. E. Harkins, K. Wallenborg, P. Hothi, S. Rostad, L. Hood, and C. S. Cobbs (2016). TGFb-Responsive HMOX1 Expression Is Associated with Stemness and Invasion in Glioblastoma Multiforme. *Stem Cells 34*(9), 2276–2289.

Gollob, H. F. (1968). A statistical model which combines features of factor analytic and analysis of variance techniques. *Psychometrika 33*(1), 73–115.

Goodman, L. A. and S. J. Haberman (1990). The Analysis of Nonadditivity in Two-Way Analysis of Variance. *Journal of the American Statistical Association 85*(409), 139–145.

Griffin, J. E. and P. J. Brown (2010). Inference with normal-gamma prior distributions in regression problems. *Bayesian Analysis 5*(1), 171–188.

Griffin, J. E. and P. J. Brown (2012a). Competing Sparsity: A hierarchical prior for sparse regression with grouped effects.

Griffin, J. E. and P. J. Brown (2012b). Structuring shrinkage: some correlated priors for regression. *Biometrika 99*(2), 481–487.

Hans, C. (2009). Bayesian lasso regression. *Biometrika 96*(4), 835–845.

Hoerl, A. E. and R. W. Kennard (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics 12*(1), 55.

Hoff, P. D. (2007). Model averaging and dimension selection for the singular value decomposition. *Journal of the American Statistical Association 102*(478), 674–685.

Hoff, P. D. (2016). Equivariant and Scale-Free Tucker Decomposition Models. *Bayesian Analysis 11*(3), 627–648.

Hoff, P. D. and C. Yu (2017). Exact adaptive confidence intervals for linear regression coefficients. *arXiv: 1705.08331v2*.

Huang, J., J. L. Horowitz, and S. Ma (2008). Asymptotic Properties of Bridge Estimators in Sparse High-Dimensional Regression Models. *The Annals of Statistics 36*(2), 587–613.

Johnson, D. E. and F. A. Graybill (1972). An Analysis of a Two-Way Model with Interaction and No Replication. *Journal of the American Statistical Association 67*(340), 862–868.

Josse, J. and S. Sardy (2016). Adaptive shrinkage of singular values. *Statistics and Computing 26*(3), 715–724.

Josse, J., S. Sardy, and S. Wager (2016). *denoiseR: A Package for Low Rank Matrix Estimation*. R package version 1.0.

Kalli, M. and Griffin (2014). Time-varying sparsity in dynamic regression models. *Journal of Econometrics 178*, 779–793.

Konig, R. H., H. Neudecker, and T. Wansbeek (1992). Unbiased Estimation of Fourth-Order Matrix Moments. *Linear Algebra and Its Applications 160*, 163–174.

Kowal, D. R., D. S. Matteson, and D. Ruppert (2017). Dynamic Shrinkage Processes.

Kyung, M., J. Gill, M. Ghosh, and G. Casella (2010). Penalized Regression, Standard Errors, and Bayesian Lassos. *Bayesian Analysis 5*(2), 369–412.

Leeb, H. and B. M. Pötscher (2008). Sparse estimators and the oracle property, or the return of Hodges' estimator. *Journal of Econometrics 142*(1), 201–211.

Makeig, S., C. Kothe, T. Mullen, N. Bigdely-Shamlo, Z. Zhang, and K. Kreutz-Delgado (2012). Evolving Signal Processing for Brain Computer Interfaces. *Proceedings of the IEEE 100*(Special Centennial Issue), 1567–1584.

Mandel, J. (1971). A New Analysis of Variance Model for Non-Additive Data. *Technometrics 13*(1), 1–18.

Marjanovic, G. and V. Solo (2014). lq Sparsity Penalized Linear Regression with Cyclic Descent. *IEEE Transactions on Signal Processing 62*(6), 1464–1475.

Mazumder, R., J. H. Friedman, and T. Hastie (2011). SparseNet: Coordinate Descent With Nonconvex Penalties. *Journal of the American Statistical Association 106*(495), 1125–1138.

McCullagh, P. and J. A. Nelder (1989). *Generalized Linear Models* (2nd ed.). London: Chapman & Hall.

Mitchell, T. M., R. Hutchinson, R. S. Niculescu, and X. Wang (2004). Learning to Decode Cognitive States from Brain Images. *Machine Learning 57*(January), 145–175.

Murray, I., R. P. Adams, and D. J. C. Mackay (2010). Elliptical slice sampling. *Journal of Machine Learning Research: WC&P 9*, 541–548.

Nadarajah, S. (2006). On the linear combination of normal and Laplace random variables. *Computational Statistics 21*(1), 63–71.

Neal, R. M. (2003). Slice Sampling. *Annals of Statistics 31*(3), 758–767.

Park, T. and G. Casella (2008). The Bayesian Lasso. *Journal of the American Statistical Association 103*(482), 681–686.

Perry, P. O. (2017). Fast moment-based estimation for hierarchical models. *Journal of the Royal Statistical Society. Series B 79*(1), 267–291.

Polson, N. G. and J. G. Scott (2010). Shrink Globally, Act Locally: Bayesian Sparsity and Regularization. In J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, and M. West (Eds.), *Bayesian Statistics 9*, pp. 501–538. Oxford University Press.

Polson, N. G., J. G. Scott, and J. Windle (2013). Bayesian inference for logistic models using Pólya-Gamma latent variables. *Journal of the American Statistical Association 108*(504), 1339–1349.

Polson, N. G., J. G. Scott, and J. Windle (2014). The Bayesian bridge. *Journal of the Royal Statistical Society. Series B: Statistical Methodology 76*(4), 713–733.

Priami, C. and M. J. Morine (2015). *Analysis of Biological Systems*. London: Imperial College Press.

Robert, C. P. (1995). Simulation of truncated normal variables. *Statistics and Computing 5*, 121–125.

Rodriguez-Yam, G., R. A. Davis, and L. L. Scharf (2004). Efficient Gibbs Sampling of Truncated Multivariate Normal with Application to Constrained Linear Regression.

Roy, V. and S. Chakraborty (2016). Selection of Tuning Parameters, Solution Paths and Standard Errors for Bayesian Lassos. *Bayesian Analysis TBA*(TBA), 1–26.

Salazar, E., M. A. R. Ferreira, and H. S. Migon (2012). Objective Bayesian analysis for exponential power regression models. *Sankhya: The Indian Journal of Statistics, Series B 74*(1), 107–125.

Sharma, N. (2013). *Single-Trial P300 Classification with LDA and Neural Networks*. Ph. D. thesis, Colorado State University.

Styan, P. H. (1973). Hadamard Products and Multivariate Statistical Analysis. *Linear Algebra and Its Applications 6*, 217–240.

Subbotin, M. T. (1923). On the Law of Frequency of Error. *Matematicheskii Sbornik 31*(2), 296–301.

Thirion, B., G. Varoquaux, E. Dohmatob, and J. B. Poline (2014). Which fMRI clustering gives good brain parcellations? *Frontiers in Neuroscience 8*(8 JUL), 1–13.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 58*(1), 267–288.

Tibshirani, R. (2011). Regression Shrinkage and Selection Via the Lasso: a retrospective. *Journal of the Royal Statistical Society. Series B: Statistical Methodology 73*(3), 273–282.

Tibshirani, R. J. (2013). The lasso problem and uniqueness. *Electronic Journal of Statistics 7*(1), 1456–1490.

van der Pas, S., B. Szabó, and A. W. van der Vaart (2017). Adaptive posterior contraction rates for the horseshoe. *Electronic Journal of Statistics 11*(2).

Van der Vaart, A. W. (2000). *Asymptotic Statistics*. Cambridge University Press.

van Eeuwijk, F. A. and P. M. Kroonenberg (1998). Multiplicative Models for Interaction in Three-Way ANOVA, with Applications to Plant Breeding. *Biometrics 54*(4), 1315–1333.

van Gerven, M., B. Cseke, R. Oostenveld, and T. Heskes (2009). Bayesian Source Localization with the Multivariate Laplace Prior. *Advances in Neural Information Processing Systems 22*, 1–9.

van Gerven, M. A. J., B. Cseke, F. P. de Lange, and T. Heskes (2010). Efficient Bayesian multivariate fMRI analysis using a sparsifying spatio-temporal prior. *Neuroimage 50*, 150–161.

Walker, S. G. and E. Guttiérez-Pena (1999). Robustifying Bayesian Procedures. *Bayesian Statistics 6*, 685–710.

Wand, M. P. (2003). Smoothing and mixed models. *Computational Statistics 18*(2), 223–249.

West, M. (1987). On Scale Mixtures of Normal Distributions. *Biometrika 74*(3), 646–648.

Westfall, P. H. (2014). Kurtosis as Peakedness, 1905 - 2014. R.I.P. *The American Statistician 68*(3), 191–195.

Wolpaw, J. and E. W. Wolpaw (2012). *Brain-Computer Interfaces: Principles and Practice*. USA: Oxford University Press.

Wu, A., M. Park, O. Koyejo, and J. W. Pillow (2014). Sparse Bayesian structure learning with dependent relevance determination prior. *Advances in Neural Information Processing Systems*, 1628–1636.

Zellner, A. (1986). On Assessing Prior Distributions and Bayesian Regression Analysis with g Prior Distributions. In P. K. Goel and Z. A. (Eds.), *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*, pp. 233–243. New York: Elsevier.

Zhang, C., J. Li, H. Wang, and S. Wei Song (2016). Identification of a five B cell-associated gene prognostic and predictive signature for advanced glioma patients harboring immunosuppressive subtype preference. *Oncotarget 7*(45), 73971–73983.

Zhao, S., C. Gao, S. Mukherjee, and Engel (2016). Bayesian group factor analysis with structured sparsity. *Journal of Machine Learning Research 17*, 1–47.

# Appendix A

# SUPPLEMENT TO "LASSO ANOVA DECOMPOSITIONS FOR MATRIX AND TENSOR DATA"

## *A.1 Uniqueness of $\hat{M}$, $\hat{A}$ and $\hat{C}$*

As this is a standard lasso regression problem, uniqueness of $\hat{\boldsymbol{M}}$ is well known (Tibshirani, 2013). Suppose we have two unique solutions for $\boldsymbol{A}$: $\hat{\boldsymbol{A}}$ and $\tilde{\boldsymbol{A}}$, where $\hat{\boldsymbol{A}} \neq \tilde{\boldsymbol{A}}$. Then it must be the case that $\tilde{\boldsymbol{A}} = \hat{\boldsymbol{A}} + \boldsymbol{\alpha}\mathbf{1}_p' + \mathbf{1}_n\boldsymbol{\beta}'$, where $\boldsymbol{\alpha} \neq \mathbf{0}$ and/or $\boldsymbol{\beta} \neq \mathbf{0}$, i.e. $\boldsymbol{\alpha}\mathbf{1}_p' + \mathbf{1}_n\boldsymbol{\beta}' \neq 0$. Let $\hat{\boldsymbol{C}}$ and $\tilde{\boldsymbol{C}}$ be the corresponding solutions for $\boldsymbol{C}$. By uniqueness of $\hat{\boldsymbol{M}}$, we have:

$$\hat{\boldsymbol{A}} + \hat{\boldsymbol{C}} = \tilde{\boldsymbol{A}} + \tilde{\boldsymbol{C}} = \hat{\boldsymbol{M}}.$$

It follows that $\hat{\boldsymbol{C}} = \hat{\boldsymbol{M}} - \hat{\boldsymbol{A}}$ and $\tilde{\boldsymbol{C}} = \hat{\boldsymbol{M}} - \tilde{\boldsymbol{A}}$. Substituting our expression for $\tilde{\boldsymbol{A}}$ in terms of $\hat{\boldsymbol{A}}$ into $\tilde{\boldsymbol{C}}$ yields $\tilde{\boldsymbol{C}} = \hat{\boldsymbol{M}} - \hat{\boldsymbol{A}} - \boldsymbol{\alpha}\mathbf{1}_p' - \mathbf{1}_n\boldsymbol{\beta}'$. Then:

$$\begin{aligned}
\left|\left|\text{vec}\left(\tilde{\boldsymbol{C}}\right)\right|\right|_1 &= \left|\left|\text{vec}\left(\hat{\boldsymbol{M}} - \hat{\boldsymbol{A}} - \boldsymbol{\alpha}\mathbf{1}_p' - \mathbf{1}_n\boldsymbol{\beta}'\right)\right|\right|_1 \\
&\leq \left|\left|\text{vec}\left(\hat{\boldsymbol{M}} - \hat{\boldsymbol{A}}\right)\right|\right|_1 + \left|\left|\text{vec}\left(\boldsymbol{\alpha}\mathbf{1}_p' + \mathbf{1}_n\boldsymbol{\beta}'\right)\right|\right|_1 \\
&= \left|\left|\text{vec}\left(\hat{\boldsymbol{C}}\right)\right|\right|_1 + \left|\left|\text{vec}\left(\boldsymbol{\alpha}\mathbf{1}_p' + \mathbf{1}_n\boldsymbol{\beta}'\right)\right|\right|_1.
\end{aligned}$$

It follows that $\left|\left|\text{vec}\left(\tilde{\boldsymbol{C}}\right)\right|\right|_1 < \left|\left|\text{vec}\left(\hat{\boldsymbol{C}}\right)\right|\right|_1$ and

$$\frac{1}{2\sigma_z^2}\left|\left|\text{vec}\left(\boldsymbol{Y} - \hat{\boldsymbol{M}}\right)\right|\right|_2^2 + \lambda_c\left|\left|\text{vec}\left(\tilde{\boldsymbol{C}}\right)\right|\right|_1 < \frac{1}{2\sigma_z^2}\left|\left|\text{vec}\left(\boldsymbol{Y} - \hat{\boldsymbol{M}}\right)\right|\right|_2^2 + \lambda_c\left|\left|\text{vec}\left(\hat{\boldsymbol{C}}\right)\right|\right|_1.$$

This contradicts our claim that the $\hat{\boldsymbol{A}}$ and $\hat{\boldsymbol{C}}$ solve the penalized regression problem. It follows that $\hat{\boldsymbol{A}}$ is unique and as $\hat{\boldsymbol{C}}$ is uniquely determined by $\hat{\boldsymbol{M}}$ and $\hat{\boldsymbol{A}}$, it follows that $\hat{\boldsymbol{C}}$ is unique as well.

### *A.2 Equivalence of ANOVA Decomposition of $\boldsymbol{Y}$ and $\hat{\boldsymbol{M}}$*

We can rewrite the penalized problem as follows:

$$\text{minimize}_{\mu,\boldsymbol{a},\boldsymbol{b},\boldsymbol{C}} \frac{1}{2\sigma_z^2} \left|\left|\text{vec}\left(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\theta}\right)\right|\right|_2^2 + \lambda_c \left|\left|\text{vec}\left(\boldsymbol{C}\right)\right|\right|_1,$$

where $\boldsymbol{\theta}' = \left[\begin{array}{cccc} \mu & \boldsymbol{a}' & \boldsymbol{b}' & \text{vec}\left(\boldsymbol{C}\right)' \end{array}\right]$ and

$$\boldsymbol{X} = \left[\begin{array}{cccc} \mathbf{1}_p \otimes \mathbf{1}_n & \mathbf{1}_p \otimes \boldsymbol{I}_n & \boldsymbol{I}_p \otimes \mathbf{1}_n & \boldsymbol{I}_p \otimes \boldsymbol{I}_n \end{array}\right]. \tag{A.1}$$

A solution, $\hat{\boldsymbol{\theta}} = \left[\begin{array}{cccc} \hat{\mu} & \hat{\boldsymbol{a}} & \hat{\boldsymbol{b}} & \text{vec}\left(\hat{\boldsymbol{C}}\right) \end{array}\right]'$ satisfies the KKT conditions:

$$\boldsymbol{X}'\left(\text{vec}\left(\boldsymbol{Y} - \hat{\boldsymbol{M}}\right)\right) = \left[\begin{array}{c} 0 \cdot \mathbf{1}_{1+n+p} \\ \lambda_c \text{vec}\left(\boldsymbol{\Gamma}\right) \end{array}\right],$$

where $\gamma_{ij} \in \begin{cases} \{\text{sign}\left(\hat{c}_{ij}\right)\} & \text{if } \hat{c}_{ij} = 0 \\ [-1, 1] & \text{if } \hat{c}_{ij} \neq 0 \end{cases}$.

Using Equation (A.1), the first $1 + n + p$ equalities give:

$$\begin{aligned} \bar{y}_{\cdot\cdot} &= \bar{\hat{m}}_{\cdot\cdot} \\ \bar{y}_{i\cdot} &= \bar{\hat{m}}_{i\cdot},\ i = 1, \ldots, n \\ \bar{y}_{\cdot j} &= \bar{\hat{m}}_{\cdot j},\ j = 1, \ldots, p, \end{aligned}$$

where $\bar{y}_{\cdot\cdot} = \frac{1}{np}\sum_{i=1}^{n}\sum_{j=1}^{p} y_{ij}$, $\bar{y}_{i\cdot} = \frac{1}{p}\sum_{j=1}^{p} y_{ij}$ and $\bar{y}_{\cdot j} = \frac{1}{n}\sum_{i=1}^{n} y_{ij}$ and $\bar{\hat{m}}_{\cdot\cdot}$, $\bar{\hat{m}}_{i\cdot}$ and $\bar{\hat{m}}_{\cdot j}$ are defined accordingly. This means that the two-way ANOVA decompositions fits based on $\boldsymbol{Y}$ and $\hat{\boldsymbol{M}}$ will be the same.

### *A.3 Proof of Proposition 2.2.1*

To prove this proposition, we need to compute $\mathbb{E}\left[\bar{r}^{(4)}\right]$ and $\mathbb{E}\left[\left(\bar{r}^{(2)}\right)^2\right]$. We use the following result from Konig et al. (1992):

$$\mathbb{E}\left[\left(\boldsymbol{x}'\boldsymbol{S}\boldsymbol{x}\right)^2\right] = \alpha\mathbb{E}\left[x_i^4\right] + \left(\theta + 2\beta\right)\mathbb{E}\left[x_i^2 x_j^2\right], \tag{A.2}$$

for a mean zero random vector $\boldsymbol{x}$ of independent elements and symmetric fixed matrix $\boldsymbol{S}$, where $\alpha = \sum_i s_{ii}^2$, $\theta = \sum_{i \neq j} s_{ii} s_{jj}$ and $\beta = \sum_{i \neq j} s_{ij}^2$.

First, we rewrite $\mathbb{E}\left[\left(\bar{r}^{(2)}\right)^2\right]$ to resemble Equation (A.2):

$$\left(\bar{r}^{(2)}\right)^2 = \frac{1}{n^2 p^2}\left(\operatorname{vec}\left(\boldsymbol{C}+\boldsymbol{Z}\right)'\left(\boldsymbol{H}_p \otimes \boldsymbol{H}_n\right)\left(\operatorname{vec}\left(\boldsymbol{C}+\boldsymbol{Z}\right)\right)\right)^2.$$

We compute $\alpha$, $\theta$ and $\beta$ below:

$$\begin{aligned}
\alpha &= (np)\left(\frac{(n-1)(p-1)}{np}\right)^2 \\
&= \frac{(n-1)^2(p-1)^2}{np} \\
\theta &= np(np-1)\left(\frac{(n-1)(p-1)}{np}\right)^2 \\
&= \frac{(np-1)(n-1)^2(p-1)^2}{np} \\
\beta &= np(p-1)\left(\frac{n-1}{np}\right)^2 + n(n-1)p\left(\frac{p-1}{np}\right)^2 + n(n-1)p(p-1)\left(\frac{1}{np}\right)^2 \\
&= \frac{(n-1)^2(p-1) + (n-1)(p-1)^2 + (n-1)(p-1)}{np}
\end{aligned}$$

Recalling $\kappa$ is defined such that $\mathbb{E}\left[c_{ij}^4\right] = (\kappa+3)\sigma_c^4$ and plugging these expressions for $\alpha$, $\beta$ and $\theta$ into Equation (A.2) yields:

$$\begin{aligned}
\mathbb{E}\left[\left(\bar{r}^{(2)}\right)^2\right] =& \frac{(n-1)^2(p-1)^2}{n^3p^3}\left(\kappa\sigma_c^4 + 3\left(\sigma_c^2+\sigma_z^2\right)^2\right) + \\
& \frac{(np-1)(n-1)^2(p-1)^2}{n^3p^3}\left(\sigma_c^2+\sigma_z^2\right)^2 + \\
& 2\left(\frac{(n-1)^2(p-1) + (n-1)(p-1)^2 + (n-1)(p-1)}{n^3p^3}\right)\left(\sigma_c^2+\sigma_z^2\right)^2 \\
=& \kappa\left(\frac{(n-1)^2(p-1)^2}{n^3p^3}\right)\sigma_c^4 + \left(\frac{(n-1)(p-1)}{np}\right)^2\left(\sigma_c^2+\sigma_z^2\right)^2 + \\
& 2\left(\frac{(n-1)(p-1)}{n^2p^2}\right)\left(\sigma_c^2+\sigma_z^2\right)^2.
\end{aligned} \tag{A.3}$$

Now we compute $\bar{r}^{(4)}$. Letting $\boldsymbol{E}_{kk}$ be an $np \times np$ matrix with the $k$-th diagonal element equal to 1 and all other elements equal to 0. We can rewrite $\bar{r}^{(4)}$ so each term in resembles

Equation (A.2):

$$\begin{aligned}\bar{r}^{(4)} &= \frac{1}{np}\sum_{k=1}^{np}\left(\text{vec}\left(\boldsymbol{R}\right)'\boldsymbol{E}_{kk}\text{vec}\left(\boldsymbol{R}\right)\right)^2 \\ &= \frac{1}{np}\sum_{k=1}^{np}\left(\text{vec}\left(\boldsymbol{C}+\boldsymbol{Z}\right)'\left(\boldsymbol{H}_p\otimes\boldsymbol{H}_n\right)\boldsymbol{E}_{kk}\left(\boldsymbol{H}_p\otimes\boldsymbol{H}_n\right)\text{vec}\left(\boldsymbol{C}+\boldsymbol{Z}\right)\right)^2.\end{aligned}$$

The matrices $\left(\boldsymbol{H}_p\otimes\boldsymbol{H}_n\right)\boldsymbol{E}_{kk}\left(\boldsymbol{H}_p\otimes\boldsymbol{H}_n\right)$ take the place of $\boldsymbol{S}$ in Equation (A.2).

From Equation (A.2), we have to compute the sum of squared diagonal elements, denoted by $\alpha_k$, the sum of pairwise products of distinct diagonal elements, denoted by $\theta_k$, and the sum of squared off-diagonal elements, denoted by $\beta_k$, for each of these matrices. Before going further, we show that $\theta_k = \beta_k$. Letting $\boldsymbol{e}_k$ is an $np\times 1$ vector with the $k$-th element equal to 1 and all other elements equal to 0,

$$\begin{aligned}\alpha_k+\beta_k &= \text{tr}\left(\left(\boldsymbol{H}_p\otimes\boldsymbol{H}_n\right)\boldsymbol{E}_{kk}\left(\boldsymbol{H}_p\otimes\boldsymbol{H}_n\right)\left(\boldsymbol{H}_p\otimes\boldsymbol{H}_n\right)\boldsymbol{E}_{kk}\left(\boldsymbol{H}_p\otimes\boldsymbol{H}_n\right)\right) \\ &= \text{tr}\left(\left(\boldsymbol{H}_p\otimes\boldsymbol{H}_n\right)\boldsymbol{e}_k\boldsymbol{e}_k'\left(\boldsymbol{H}_p\otimes\boldsymbol{H}_n\right)\boldsymbol{e}_k\boldsymbol{e}_k'\right) \\ &= \left(\boldsymbol{e}_k'\left(\boldsymbol{H}_p\otimes\boldsymbol{H}_n\right)\boldsymbol{e}_k\right)^2 \\ &= \left(\frac{\left(n-1\right)\left(p-1\right)}{np}\right)^2, \\ \alpha_k+\theta_k &= \text{tr}\left(\left(\boldsymbol{H}_p\otimes\boldsymbol{H}_n\right)\boldsymbol{E}_{kk}\left(\boldsymbol{H}_p\otimes\boldsymbol{H}_n\right)\otimes\left(\boldsymbol{H}_p\otimes\boldsymbol{H}_n\right)\boldsymbol{E}_{kk}\left(\boldsymbol{H}_p\otimes\boldsymbol{H}_n\right)\right) \\ &= \text{tr}\left(\left(\boldsymbol{H}_p\otimes\boldsymbol{H}_n\right)\boldsymbol{E}_{kk}\left(\boldsymbol{H}_p\otimes\boldsymbol{H}_n\right)\right)^2 \\ &= \text{tr}\left(\left(\boldsymbol{H}_p\otimes\boldsymbol{H}_n\right)\boldsymbol{e}_k\boldsymbol{e}_k'\right)^2 \\ &= \left(\boldsymbol{e}_k'\left(\boldsymbol{H}_p\otimes\boldsymbol{H}_n\right)\boldsymbol{e}_k\right)^2 \\ &= \left(\frac{\left(n-1\right)\left(p-1\right)}{np}\right)^2.\end{aligned}$$

It follows that $\beta_k = \theta_k$.

This means we *only* need to compute the sum of squared diagonal elements and the sum of pairwise products of distinct diagonal elements of $\left(\boldsymbol{H}_p\otimes\boldsymbol{H}_n\right)\boldsymbol{E}_{kk}\left(\boldsymbol{H}_p\otimes\boldsymbol{H}_n\right)$.

Each matrix, $\left(\boldsymbol{H}_p\otimes\boldsymbol{H}_n\right)\boldsymbol{E}_{kk}\left(\boldsymbol{H}_p\otimes\boldsymbol{H}_n\right)$, is constructed by multiplying the $k$-th column and the $k$-th row of $\left(\boldsymbol{H}_p\otimes\boldsymbol{H}_n\right)$. By symmetry of $\left(\boldsymbol{H}_p\otimes\boldsymbol{H}_n\right)$, this is the same as multiplying

the $k$-th column, denoted by $(\boldsymbol{H}_p \otimes \boldsymbol{H}_n)_k$, by its transpose,

$$(\boldsymbol{H}_p \otimes \boldsymbol{H}_n)\, \boldsymbol{E}_{kk}\, (\boldsymbol{H}_p \otimes \boldsymbol{H}_n) = (\boldsymbol{H}_p \otimes \boldsymbol{H}_n)_k\, (\boldsymbol{H}_p \otimes \boldsymbol{H}_n)_k' .$$

Regardless of $k$, each column $(\boldsymbol{H}_p \otimes \boldsymbol{H}_n)_k$ has the following elements:

- One element equal to $\frac{(n-1)(p-1)}{np}$;
- $(p-1)$ elements equal to $-\frac{(n-1)}{np}$;
- $(n-1)$ elements equal to $-\frac{(p-1)}{np}$
- $(n-1)\,(p-1)$ elements equal to $\frac{1}{np}$.

The diagonal elements of $(\boldsymbol{H}_p \otimes \boldsymbol{H}_n)\, \boldsymbol{E}_{kk}\, (\boldsymbol{H}_p \otimes \boldsymbol{H}_n)$ will be given by the squared elements of $(\boldsymbol{H}_p \otimes \boldsymbol{H}_n)_k$. Note that the only feature of the diagonal elements of $(\boldsymbol{H}_p \otimes \boldsymbol{H}_n)\, \boldsymbol{E}_{kk}\, (\boldsymbol{H}_p \otimes \boldsymbol{H}_n)$ that depends on $k$ is the order of these elements. Accordingly, the sum of squared diagonal elements, $\alpha_k$, and the sum of pairwise products of distinct

diagonal elements, $\theta_k$, will *not* depend on $k$. Dropping the subscript, we compute $\alpha$ and $\theta$:

$$\begin{aligned}
\alpha =& \left(\frac{(n-1)(p-1)}{np}\right)^4 + (p-1)\left(\frac{n-1}{np}\right)^4 + (n-1)\left(\frac{p-1}{np}\right)^4 + (n-1)(p-1)\left(\frac{1}{np}\right)^4 \\
=& \frac{(n-1)(p-1)}{n^4p^4}\left((n-1)^3(p-1)^3 + (n-1)^3 + (p-1)^3 + 1\right) \\
=& \frac{(n-1)(n^2-3n+3)(p-1)(p^2-3p+3)}{n^4p^4} \\
\theta =& (p-1)\left(\frac{(n-1)^4(p-1)^2}{n^4p^4}\right) + (n-1)\left(\frac{(n-1)^2(p-1)^4}{n^4p^4}\right) + (n-1)(p-1)\left(\frac{(n-1)^2(p-1)^2}{n^4p^4}\right) + \\
& (p-1)(p-2)\left(\frac{(n-1)^4}{n^4p^4}\right) + (n-1)(p-1)\left(\frac{(n-1)^2(p-1)^2}{n^4p^4}\right) + (n-1)(p-1)^2\left(\frac{(n-1)^2}{n^4p^4}\right) + \\
& (n-1)(n-2)\left(\frac{(p-1)^4}{n^4p^4}\right) + (n-1)^2(p-1)\left(\frac{(p-1)^2}{n^4p^4}\right) + \\
& (n-1)(p-1)\left((n-1)(p-1)-1\right)\left(\frac{1}{n^4p^4}\right) \\
=& \frac{1}{n^4p^4}\left((p-1)^3(n-1)^4 + (p-1)^4(n-1)^3 + (p-1)^3(n-1)^3\right. \\
& + (p-1)^3(n-1)^4 + (p-1)(p-2)(n-1)^4 + (p-1)^3(n-1)^3 + (p-1)^2(n-1)^3 + \\
& (p-1)^4(n-1)^3 + (p-1)^3(n-1)^3 + (n-1)(n-2)(p-1)^4 + (p-1)^3(n-1)^2 + \\
& \left.(n-1)^3(p-1)^3 + (n-1)^3(p-1)^2 + (n-1)^2(p-1)^3 + (n-1)(p-1)\left((n-1)(p-1)-1\right)\right) \\
=& \frac{1}{n^4p^4}\left(2(p-1)^3(n-1)^4 + 2(p-1)^4(n-1)^3 + 4(p-1)^3(n-1)^3 + \right. \\
& (p-1)(p-2)(n-1)^4 + 2(p-1)^2(n-1)^3 + (n-1)(n-2)(p-1)^4 + 2(p-1)^3(n-1)^2 + \\
& \left.(n-1)(p-1)\left((n-1)(p-1)-1\right)\right) \\
=& \frac{(n-1)(p-1)(2np^2+2n^2p-3p^2-3n^2-8np+9p+9n-9)}{n^3p^3} \\
=& \left(\frac{(n-1)(p-1)}{np}\right)^2 - \frac{(n-1)(n^2-3n+3)(p-1)(p^2-3p+3)}{n^3p^3}.
\end{aligned}$$

Plugging these expressions into Equation (A.2) yields:

$$\mathbb{E}\left[\bar{r}^{(4)}\right] = \kappa\left(\frac{(n-1)\left(n^2-3n+3\right)(p-1)\left(p^2-3p+3\right)}{n^3p^3}\right)\sigma_c^4 + 3\left(\frac{(n-1)(p-1)}{np}\right)^2\left(\sigma_c^2+\sigma_z^2\right)^2. \tag{A.4}$$

Plugging in the excess kurtosis, $\kappa = 3$, of Laplace distributed $c_{ij}$, Proposition 2.2.1 follows straightforwardly from the definition of $\hat{\sigma}_c^4$ and Equations (A.3) and (A.4).

### *A.4 Proof of Proposition 2.2.2*

We start by showing consistency of our empirical Bayes estimators as $p \to \infty$ with $n$ fixed. The proof of consistency of our empirical Bayes estimators as $n \to \infty$ with $p$ fixed is analogous.

Recalling that our empirical Bayes estimators are functions of $\bar{r}^{(2)}$ and $\bar{r}^{(4)}$, we start by proving that $\bar{r}^{(2)} \xrightarrow{p} \left(\frac{n-1}{n}\right)\left(\sigma_c^2+\sigma_z^2\right)$ as $p \to \infty$ for fixed $n$. First, we compute $\mathbb{E}\left[\bar{r}^{(2)}\right]$.

$$\begin{aligned}
\mathbb{E}\left[\bar{r}^{(2)}\right] &= \frac{1}{np}\mathbb{E}\left[\operatorname{vec}(\boldsymbol{R})'\operatorname{vec}(\boldsymbol{R})\right] \\
&= \frac{1}{np}\mathbb{E}\left[\operatorname{vec}(\boldsymbol{C}+\boldsymbol{Z})'\left(\boldsymbol{H}_p\otimes\boldsymbol{H}_n\right)\operatorname{vec}(\boldsymbol{C}+\boldsymbol{Z})\right] \\
&= \frac{1}{np}\mathbb{E}\left[\operatorname{vec}(\boldsymbol{C}+\boldsymbol{Z})'\left(\boldsymbol{H}_p\otimes\boldsymbol{H}_n\right)\operatorname{vec}(\boldsymbol{C}+\boldsymbol{Z})\right] \\
&= \frac{1}{np}\operatorname{tr}\left(\left(\boldsymbol{H}_p\otimes\boldsymbol{H}_n\right)\mathbb{E}\left[\operatorname{vec}(\boldsymbol{C}+\boldsymbol{Z})\operatorname{vec}(\boldsymbol{C}+\boldsymbol{Z})'\right]\right) \\
&= \frac{1}{np}\operatorname{tr}\left(\boldsymbol{H}_p\otimes\boldsymbol{H}_n\right)\left(\sigma_c^2+\sigma_z^2\right) \\
&= \frac{(n-1)(p-1)}{np}\left(\sigma_c^2+\sigma_z^2\right).
\end{aligned} \tag{A.5}$$

For fixed $n$, as $p \to \infty$, $\mathbb{E}\left[\bar{r}^{(2)}\right] = O(1)\left(\frac{n-1}{n}\right)\left(\sigma_c^2+\sigma_z^2\right)$.

We can compute $\mathbb{V}\left[\bar{r}^{(2)}\right]$ explicitly using Equation (A.3).

$$\mathbb{V}\left[\bar{r}^{(2)}\right] = \kappa\left(\frac{(n-1)^2(p-1)^2}{n^3p^3}\right)\sigma_c^4 + 2\left(\frac{(n-1)(p-1)}{n^2p^2}\right)\left(\sigma_c^2+\sigma_z^2\right)^2. \tag{A.6}$$

For fixed $n$, as $p \to \infty$, $\mathbb{V}\left[\bar{r}^{(2)}\right] = O\left(\frac{1}{p}\right)$. It follows that $\bar{r}^{(2)} \xrightarrow{p} \left(\frac{n-1}{n}\right)(\sigma_c^2 + \sigma_z^2)$ as $p \to \infty$ for fixed $n$.

Now we show $\bar{r}^{(4)} \xrightarrow{p} \kappa\left(\frac{(n-1)\left(n^2-3n+3\right)}{n^3}\right)\sigma_c^4 + 3\left(\frac{n-1}{n}\right)^2(\sigma_c^2+\sigma_z^2)^2$ as $p \to \infty$ for fixed $n$.

From Equation (A.4), we have $\mathbb{E}\left[\bar{r}^{(4)}\right] = \left(O(1)\,\kappa\frac{(n-1)\left(n^2-3n+3\right)}{n^3}\right)\sigma_c^4 + O(1)\,3\left(\frac{n-1}{n}\right)^2(\sigma_c^2+\sigma_z^2)^2$ as $p \to \infty$ for fixed $n$. Now we evaluate the order of $\mathbb{V}\left[\bar{r}^{(4)}\right]$, starting by evaluating the order of $\mathbb{E}\left[\left(\bar{r}^{(4)}\right)^2\right]$. For convenience, we let $x_{ij} = c_{ij} + z_{ij}$ and use the following way of writing elements of $\boldsymbol{R}$: $r_{ij} = x_{ij} - \bar{x}_{i\cdot} - \bar{x}_{\cdot j} + \bar{x}_{\cdot\cdot}$, where $\bar{x}_{i\cdot} = \frac{1}{p}\sum_{j=1}^{p} x_{ij}$, $\bar{x}_{\cdot j} = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$ and $\bar{x}_{\cdot\cdot} = \frac{1}{np}\sum_{i=1}^{n}\sum_{j=1}^{p} x_{ij}$.

$$
\begin{aligned}
\mathbb{E}\left[\left(\bar{r}^{(4)}\right)^2\right] =& \mathbb{E}\left[\left(\frac{1}{np}\sum_{i=1}^{n}\sum_{j=1}^{p} r_{ij}^2\right)^2\right] \\
=& \frac{1}{n^2p^2}\mathbb{E}\left[\sum_{i=1}^{n}\sum_{j=1}^{p} r_{ij}^8 + \sum_{i'=1}^{n}\sum_{j'=p,i'j'\neq ij} r_{ij}^4 r_{i'j'}^4\right] \\
=& \frac{1}{n^2p^2}\mathbb{E}\left[\sum_{i=1}^{n}\sum_{j=1}^{p}\left(x_{ij} - \bar{x}_{\cdot j} + O_p\left(\frac{1}{\sqrt{p}}\right)\right)^8\right] + \\
& \frac{1}{n^2p^2}\mathbb{E}\left[\sum_{i=1}^{n}\sum_{j=1}^{p}\sum_{i'\neq i}^{n}\left(x_{ij} - \bar{x}_{\cdot j} + O_p\left(\frac{1}{\sqrt{p}}\right)\right)^4\left(x_{i'j'} - \bar{x}_{\cdot j'} + O_p\left(\frac{1}{\sqrt{p}}\right)\right)^4\right] + \\
& \frac{1}{n^2p^2}\mathbb{E}\left[\sum_{i=1}^{n}\sum_{j=1}^{p}\sum_{i'=1}^{n}\sum_{j'\neq j}\left(x_{ij} - \bar{x}_{\cdot j} + O_p\left(\frac{1}{\sqrt{p}}\right)\right)^4\left(x_{i'j'} - \bar{x}_{\cdot j'} + O_p\left(\frac{1}{\sqrt{p}}\right)\right)^4\right] \\
=& \frac{1}{np}\mathbb{E}\left[x_{11}^8\right] + O\left(\frac{1}{p}\right) + \\
& \frac{1}{n^2p^2}\sum_{i=1}^{n}\sum_{j=1}^{p}\sum_{i'=1}^{n}\mathbb{E}\left[\left(x_{ij} - \bar{x}_{\cdot j}\right)^4\left(x_{i'j} - \bar{x}_{\cdot j}\right)^4\right] + O\left(\frac{1}{\sqrt{p}}\right) + \\
& \frac{1}{n^2p^2}\sum_{i=1}^{n}\sum_{j=1}^{p}\sum_{i'=1}^{n}\sum_{j'\neq j}\mathbb{E}\left[\left(x_{ij} - \bar{x}_{\cdot j}\right)^4\right]\mathbb{E}\left[\left(x_{i'j'} - \bar{x}_{\cdot j'}\right)^4\right] + O\left(\frac{1}{p^2}\right) \\
=& \frac{1}{p}\underbrace{\mathbb{E}\left[\left(x_{11} - \bar{x}_{\cdot 1}\right)^4\left(x_{21} - \bar{x}_{\cdot 1}\right)^4\right]}_{*} + \left(\frac{p-1}{p}\right)\mathbb{E}\left[\left(x_{11} - \bar{x}_{\cdot 1}\right)^4\right]\mathbb{E}\left[\left(x_{12} - \bar{x}_{\cdot 2}\right)^4\right] + O\left(\frac{1}{\sqrt{p}}\right) \\
=& \left(\frac{p-1}{p}\right)\mathbb{E}\left[\left(x_{11} - \bar{x}_{\cdot 1}\right)^4\right]^2 + O\left(\frac{1}{\sqrt{p}}\right).
\end{aligned}
$$

The term marked by $*$ is constant for fixed $n$, it is a degree eight polynomial of $n$ independent elements of $\boldsymbol{X}$, with coefficients that depend on $n$ alone.

Now we need to compute the remaining unknown term,

$$\begin{aligned}
\mathbb{E}\left[(x_{11}-\bar{x}_{\cdot 1})^4\right] &= \mathbb{E}\left[\left(\frac{n-1}{n}x_{11}-\frac{1}{n}\sum_{i=2}^{n}x_{i1}\right)^4\right] \\
&= \left(\frac{n-1}{n}\right)^4\mathbb{E}\left[x_{11}^4\right]+6\left(\frac{n-1}{n}\right)^2\mathbb{E}\left[x_{11}^2\right]\mathbb{E}\left[\left(\frac{1}{n}\sum_{i=2}^{n}x_{i1}\right)^2\right]+\mathbb{E}\left[\left(\frac{1}{n}\sum_{i=2}^{n}x_{i1}\right)^4\right] \\
&= \left(\frac{n-1}{n}\right)^4\mathbb{E}\left[x_{11}^4\right]+\frac{6\left(n-1\right)^3}{n^4}\mathbb{E}\left[x_{11}^2\right]^2+\frac{n-1}{n^4}\mathbb{E}\left[x_{11}^4\right]+\frac{3\left(n-1\right)\left(n-2\right)}{n^4}\mathbb{E}\left[x_{11}^2\right]^2 \\
&= \frac{\left(n-1\right)\left(\left(n-1\right)^3+1\right)}{n^4}\mathbb{E}\left[x_{11}^4\right]+\frac{\left(n-1\right)\left(6\left(n-1\right)^2+3\left(n-2\right)\right)}{n^4}\mathbb{E}\left[x_{11}^2\right]^2 \\
&= \frac{\left(n-1\right)\left(n^2-3n+3\right)}{n^3}\mathbb{E}\left[x_{11}^4\right]+\frac{3\left(n-1\right)\left(2n-3\right)}{n^3}\mathbb{E}\left[x_{11}^2\right]^2 \\
&= \kappa\frac{\left(n-1\right)\left(n^2-3n+3\right)}{n^3}\sigma_c^4+3\left(\frac{n-1}{n}\right)^2\left(\sigma_c^2+\sigma_z^2\right)^2.
\end{aligned}$$

Combining this with our expression for $\mathbb{E}\left[\left(\bar{r}^{(4)}\right)^2\right]$, we have $\mathbb{V}\left[\left(\bar{r}^{(4)}\right)^2\right]=O\left(\frac{1}{\sqrt{p}}\right)$ as $p\to\infty$ for fixed $n$. It follows that $\bar{r}^{(4)}\xrightarrow{p}\kappa\left(\frac{\left(n-1\right)\left(n^2-3n+3\right)}{n^3}\right)\sigma_c^4+3\left(\frac{n-1}{n}\right)^2\left(\sigma_c^2+\sigma_z^2\right)^2$ as $p\to\infty$ for fixed $n$.

Plugging in the excess kurtosis, $\kappa=3$, of Laplace distributed $c_{ij}$, combining $\bar{r}^{(4)}\xrightarrow{p}3\left(\frac{\left(n-1\right)\left(n^2-3n+3\right)}{n^3}\right)\sigma_c^4+3\left(\frac{n-1}{n}\right)^2\left(\sigma_c^2+\sigma_z^2\right)^2$ as $p\to\infty$ for fixed $n$ with $\bar{r}^{(2)}\xrightarrow{p}\left(\frac{n-1}{n}\right)\left(\sigma_c^2+\sigma_z^2\right)$ as $p\to\infty$ for fixed $n$, and applying the continuous mapping theorem yields $\hat{\sigma}_c^4\xrightarrow{p}\sigma_c^4$, $\hat{\sigma}_c^2\xrightarrow{p}\sigma_c^2$, $\hat{\lambda}_c\xrightarrow{p}\lambda_c$ and $\hat{\sigma}_z^2\xrightarrow{p}\sigma_z^2$ as $p\to\infty$ for fixed $n$.

Now we show consistency of our empirical Bayes estimators as $n,p\to\infty$. It follows directly from Equations (A.5) and (A.6) that $\bar{r}^{(2)}\xrightarrow{p}\sigma_c^2+\sigma_z^2$ as $n,p\to\infty$. Regarding $\bar{r}^{(4)}$, we again let $x_{ij}=c_{ij}+z_{ij}$ and use the following decomposition:

$$r_{ij}=x_{ij}-\bar{x}_{i\cdot}-\bar{x}_{\cdot j}+\bar{x}_{\cdot\cdot}$$

For convenience, we also let $n=\pi_1 q$, $p=\pi_2 q$, where $0<\pi_1<1$ and $0<\pi_2<1$ are fixed and $\pi_1+\pi_2=1$. Then the scenario $n,p\to\infty$ is equivalent to $q\to\infty$. We have:

$$
\begin{aligned}
\bar{r}^{(4)} =& \frac{1}{\pi_1 \pi_2 q^2} \sum_{i=1}^{\pi_1 q} \sum_{j=1}^{\pi_2 q} \left(x_{ij} - \bar{x}_{i\cdot} - \bar{x}_{\cdot j} + \bar{x}_{\cdot\cdot}\right)^4 \\
=& \frac{1}{\pi_1 \pi_2 q^2} \sum_{i=1}^{\pi_1 q} \sum_{j=1}^{\pi_2 q} \left(x_{ij} + O_p\left(\frac{1}{\sqrt{q}}\right)\right)^4 \\
=& \frac{1}{\pi_1 \pi_2 q^2} \sum_{i=1}^{\pi_1 q} \sum_{j=1}^{\pi_2 q} x_{ij}^4 + x_{ij}^3 O_p\left(\frac{1}{\sqrt{q}}\right) + x_{ij}^2 O_p\left(\frac{1}{q}\right) + x_{ij} O_p\left(\frac{1}{\sqrt{q}^3}\right) + O_p\left(\frac{1}{q^2}\right) \\
=& \underbrace{\bar{x}^{(4)} - \left(\kappa \sigma_c^4 + 3\left(\sigma_c^2 + \sigma_z^2\right)^2\right)}_{O_p\left(\frac{1}{q}\right)} + \left(\kappa \sigma_c^4 + 3\left(\sigma_c^2 + \sigma_z^2\right)^2\right) + O_p\left(\frac{1}{q}\right) \\
=& \kappa \sigma_c^4 + 3\left(\sigma_c^2 + \sigma_z^2\right)^2 + O_p\left(\frac{1}{q}\right).
\end{aligned} \tag{A.7}
$$

It follows that $\bar{r}^{(4)} \xrightarrow{p} \kappa \sigma_c^4 + 3\left(\sigma_c^2 + \sigma_z^2\right)^2$ as $q \to \infty$, i.e. as $n, p \to \infty$.

Plugging in the excess kurtosis, $\kappa = 3$, of Laplace distributed $c_{ij}$, combining $\bar{r}^{(4)} \xrightarrow{p} 3\sigma_c^4 + 3\left(\sigma_c^2 + \sigma_z^2\right)^2$ as $n, p \to \infty$ with $\bar{r}^{(2)} \xrightarrow{p} \sigma_c^2 + \sigma_z^2$ as $n, p \to \infty$ and applying the continuous mapping theorem yields $\hat{\sigma}_c^4 \xrightarrow{p} \sigma_c^4$, $\hat{\sigma}_c^2 \xrightarrow{p} \sigma_c^2$, $\hat{\lambda}_c \xrightarrow{p} \lambda_c$ and $\hat{\sigma}_z^2 \xrightarrow{p} \sigma_z^2$ as $n, p \to \infty$.

### *A.5 Derivation of Asymptotic Variance of $\hat{\sigma}_c^2$*

Letting $x_{ij} = c_{ij} + z_{ij}$, we will use the following decomposition:

$$r_{ij} = x_{ij} - \bar{x}_{i\cdot} - \bar{x}_{\cdot j} + \bar{x}_{\cdot\cdot}.$$

Without loss of generality, we let $n = \pi_1 q$, $p = \pi_2 q$, where $0 < \pi_1 < 1$ and $0 < \pi_2 < 1$ are fixed and $\pi_1 + \pi_2 = 1$. Then the scenario $n, p \to \infty$ is equivalent to $q \to \infty$.

For $\bar{r}^{(2)}$ we have:

$$
\begin{aligned}
\bar{r}^{(2)} =& \frac{1}{\pi_1\pi_2 q^2}\sum_{i=1}^{\pi_1 q}\sum_{j=1}^{\pi_2 q}\left(x_{ij}-\bar{x}_{i\cdot}-\bar{x}_{\cdot j}+\bar{x}_{\cdot\cdot}\right)^2 \\
=& \frac{1}{\pi_1\pi_2 q^2}\sum_{i=1}^{\pi_1 q}\sum_{j=1}^{\pi_2 q}\left(x_{ij}-\bar{x}_{i\cdot}-\bar{x}_{\cdot j}+O_p\left(\frac{1}{q}\right)\right)^2 \\
=& \frac{1}{\pi_1\pi_2 q^2}\sum_{i=1}^{\pi_1 q}\sum_{j=1}^{\pi_2 q}\left(x_{ij}-\bar{x}_{i\cdot}-\bar{x}_{\cdot j}\right)^2+\left(x_{ij}-\bar{x}_{i\cdot}-\bar{x}_{\cdot j}\right)O_p\left(\frac{1}{q}\right)+O_p\left(\frac{1}{q^2}\right) \\
=& \frac{1}{\pi_1\pi_2 q^2}\sum_{i=1}^{\pi_1 q}\sum_{j=1}^{\pi_2 q}\left(x_{ij}-\bar{x}_{i\cdot}-\bar{x}_{\cdot j}\right)^2+O_p\left(\frac{1}{q^2}\right) \\
=& \frac{1}{\pi_1\pi_2 q^2}\sum_{i=1}^{\pi_1 q}\sum_{j=1}^{\pi_2 q}x_{ij}^2+\bar{x}_{i\cdot}^2+\bar{x}_{\cdot j}^2-2x_{ij}\bar{x}_{i\cdot}-2x_{ij}\bar{x}_{\cdot j}+2\bar{x}_{i\cdot}\bar{x}_{\cdot j}+O_p\left(\frac{1}{q^2}\right)
\end{aligned}
$$

We find the order of most of the terms below:

$$
\begin{aligned}
\frac{1}{\pi_1\pi_2 q^2}\sum_{i=1}^{\pi_1 q}\sum_{j=1}^{\pi_2 q}x_{ij}\bar{x}_{i\cdot} &= \left(\frac{1}{\pi_1\pi_2 q^2}\sum_{i=1}^{\pi_1 q}\sum_{j=1}^{\pi_2 q}x_{ij}\right)O_p\left(\frac{1}{\sqrt{q}}\right) \\
&= O_p\left(\frac{1}{\sqrt{q}^3}\right) \\
\frac{1}{\pi_1\pi_2 q^2}\sum_{i=1}^{\pi_1 q}\sum_{j=1}^{\pi_2 q}x_{ij}\bar{x}_{\cdot j} &= O_p\left(\frac{1}{\sqrt{q}^3}\right) \\
\frac{1}{\pi_1\pi_2 q^2}\sum_{i=1}^{\pi_1 q}\sum_{j=1}^{\pi_2 q}\bar{x}_{i\cdot}\bar{x}_{\cdot j} &= \left(\frac{1}{\pi_1 q}\sum_{i=1}^{\pi_1 q}\bar{x}_{i\cdot}\right)\left(\frac{1}{\pi_2 q}\sum_{j=1}^{\pi_2 q}\bar{x}_{\cdot j}\right) \\
&= \bar{x}_{\cdot\cdot}^2 \\
&= O_p\left(\frac{1}{q^2}\right).
\end{aligned}
$$

Plugging these expressions into the equation for $\bar{r}^{(2)}$ gives:

$$
\bar{r}^{(2)} = \bar{x}^{(2)}+\frac{1}{\pi_1 q}\sum_{i=1}^{\pi_1 q}\bar{x}_{i\cdot}^2+\frac{1}{\pi_2 q}\sum_{j=1}^{\pi_2 q}\bar{x}_{\cdot j}^2+O_p\left(\frac{1}{\sqrt{q}^3}\right).
$$

Now we compute the means and find the orders of the variances of the middle two terms, starting with $\frac{1}{\pi_1 q}\sum_{i=1}^{\pi_1 q}\bar{x}_{i\cdot}^2$.

$$\begin{aligned}\mathbb{E}\left[\frac{1}{\pi_1 q}\sum_{i=1}^{\pi_1 q}\bar{x}_{i\cdot}^2\right] &= \mathbb{E}\left[\bar{x}_{i\cdot}^2\right] \\ &= \frac{1}{\pi_2^2 q^2}\mathbb{E}\left[\left(\sum_{j=1}^{\pi_2 q} x_{ij}\right)^2\right] \\ &= \frac{1}{\pi_2^2 q^2}\left(\pi_2 q\mathbb{E}\left[x_{ij}^2\right]\right) \\ &= \frac{\mathbb{E}\left[x_{ij}^2\right]}{\pi_2 q}.\end{aligned}$$

Note that $\frac{1}{\pi_1 q}\sum_{i=1}^{\pi_1 q}\bar{x}_{i\cdot}^2$ is a mean over independent variables, $\bar{x}_{i\cdot}^2$. Accordingly, $\mathbb{V}\left[\frac{1}{\pi_1 q}\sum_{i=1}^{\pi_1 q}\bar{x}_{i\cdot}^2\right] = \frac{1}{\pi_1 q}\mathbb{V}\left[\bar{x}_{i\cdot}^2\right]$. Now we find the order of $\mathbb{V}\left[\bar{x}_{i\cdot}^2\right]$:

$$\begin{aligned}\mathbb{V}\left[\bar{x}_{i\cdot}^2\right] &= \mathbb{E}\left[\bar{x}_{i\cdot}^4\right] - \mathbb{E}\left[x_{i\cdot}^2\right]^2 \\ &= \frac{1}{\pi_2^4 q^4}\mathbb{E}\left[\left(\sum_{j=1}^{\pi_2 q} x_{ij}\right)^4\right] + O\left(\frac{1}{q^2}\right) \\ &= \frac{1}{\pi_2^4 q^4}\left(\pi_2 q\mathbb{E}\left[x_{ij}^4\right] + 3\pi_2 q\left(\pi_2 q - 1\right)\mathbb{E}\left[x_{ij}^2\right]^2\right) + O\left(\frac{1}{q^2}\right) \\ &= O\left(\frac{1}{q^2}\right).\end{aligned}$$

Then $\mathbb{V}\left[\frac{1}{\pi_1 q}\sum_{i=1}^{\pi_1 q}\bar{x}_{i\cdot}^2\right] = O\left(\frac{1}{q^3}\right)$ and $\frac{1}{\pi_1 q}\sum_{i=1}^{\pi_1 q}\bar{x}_{i\cdot}^2 - \frac{\mathbb{E}\left[x_{ij}^2\right]}{\pi_2 q} = O_p\left(\frac{1}{\sqrt{q^3}}\right)$. The same logic yields $\frac{1}{\pi_2 q}\sum_{j=1}^{\pi_2 q}\bar{x}_{\cdot j}^2 - \frac{\mathbb{E}\left[x_{ij}^2\right]}{\pi_1 q} = O_p\left(\frac{1}{\sqrt{q^3}}\right)$. Plugging this into the equation for $\bar{r}^{(2)}$ gives:

$$\begin{aligned}\bar{r}^{(2)} =& \bar{x}^{(2)} + \frac{\mathbb{E}\left[x_{ij}^2\right]}{\pi_1 q} + \frac{\mathbb{E}\left[x_{ij}^2\right]}{\pi_2 q} + O_p\left(\frac{1}{\sqrt{q^3}}\right) \\ =& \bar{x}^{(2)} + \frac{\mathbb{E}\left[x_{ij}^2\right]}{\pi_1\pi_2 q} + O_p\left(\frac{1}{\sqrt{q^3}}\right)\end{aligned}$$

For $\bar{r}^{(4)}$ we have:

$$
\begin{aligned}
\bar{r}^{(4)} =& \frac{1}{\pi_1\pi_2 q^2}\sum_{i=1}^{\pi_1 q}\sum_{j=1}^{\pi_2 q}\left(x_{ij}-\bar{x}_{i\cdot}-\bar{x}_{\cdot j}+\bar{x}_{\cdot\cdot}\right)^4 \\
=& \frac{1}{\pi_1\pi_2 q^2}\sum_{i=1}^{\pi_1 q}\sum_{j=1}^{\pi_2 q}\left(x_{ij}-\bar{x}_{i\cdot}-\bar{x}_{\cdot j}+O_p\left(\frac{1}{q}\right)\right)^4 \\
=& \frac{1}{\pi_1\pi_2 q^2}\sum_{i=1}^{\pi_1 q}\sum_{j=1}^{\pi_2 q}\left(x_{ij}-\bar{x}_{i\cdot}-\bar{x}_{\cdot j}\right)^4+\left(x_{ij}-\bar{x}_{i\cdot}-\bar{x}_{\cdot j}\right)^3 O_p\left(\frac{1}{q}\right)+ \\
& \left(x_{ij}-\bar{x}_{i\cdot}-\bar{x}_{\cdot j}\right)^2 O_p\left(\frac{1}{q^2}\right)+\left(x_{ij}-\bar{x}_{i\cdot}-\bar{x}_{\cdot j}\right)O_p\left(\frac{1}{q^3}\right)+O_p\left(\frac{1}{q^4}\right) \\
=& \frac{1}{\pi_1\pi_2 q^2}\sum_{i=1}^{\pi_1 q}\sum_{j=1}^{\pi_2 q}\left(x_{ij}-\bar{x}_{i\cdot}-\bar{x}_{\cdot j}\right)^4+O_p\left(\frac{1}{q^2}\right)+O_p\left(\frac{1}{q^2}\right)+O_p\left(\frac{1}{q^4}\right)+O_p\left(\frac{1}{q^4}\right) \\
=& \frac{1}{\pi_1\pi_2 q^2}\sum_{i=1}^{\pi_1 q}\sum_{j=1}^{\pi_2 q}\left(x_{ij}-\bar{x}_{i\cdot}-\bar{x}_{\cdot j}\right)^4+O_p\left(\frac{1}{q^2}\right) \\
=& \frac{1}{\pi_1\pi_2 q^2}\sum_{i=1}^{\pi_1 q}\sum_{j=1}^{\pi_2 q}x_{ij}^4+\bar{x}_{i\cdot}^4+\bar{x}_{\cdot j}^4+4x_{ij}^3\bar{x}_{i\cdot}+4x_{ij}\bar{x}_{i\cdot}^3+4x_{ij}^3\bar{x}_{\cdot j}+4x_{ij}\bar{x}_{\cdot j}^3+4\bar{x}_{i\cdot}^3\bar{x}_{\cdot j}+4\bar{x}_{i\cdot}\bar{x}_{\cdot j}^3+ \\
& 6x_{ij}^2\bar{x}_{i\cdot}^2+6x_{ij}^2\bar{x}_{\cdot j}^2+6\bar{x}_{i\cdot}^2\bar{x}_{\cdot j}^2+12x_{ij}^2\bar{x}_{i\cdot}\bar{x}_{\cdot j}+12x_{ij}\bar{x}_{i\cdot}^2\bar{x}_{\cdot j}+12x_{ij}\bar{x}_{i\cdot}\bar{x}_{\cdot j}^2+O_p\left(\frac{1}{q^2}\right)
\end{aligned}
$$

As with $\bar{r}^{(2)}$, we can find the order of most of these terms:

$$\begin{aligned}
\frac{1}{\pi_1\pi_2 q^2}\sum_{i=1}^{\pi_1 q}\sum_{j=1}^{\pi_2 q}\bar{x}_{i\cdot}^4 &= O_p\left(\frac{1}{q^2}\right)\\
\frac{1}{\pi_1\pi_2 q^2}\sum_{i=1}^{\pi_1 q}\sum_{j=1}^{\pi_2 q}\bar{x}_{\cdot j}^4 &= O_p\left(\frac{1}{q^2}\right)\\
\frac{1}{\pi_1\pi_2 q^2}\sum_{i=1}^{\pi_1 q}\sum_{j=1}^{\pi_2 q}x_{ij}^3\bar{x}_{i\cdot} &= \frac{1}{\pi_1\pi_2 q^2}\sum_{i=1}^{\pi_1 q}\sum_{j=1}^{\pi_2 q}x_{ij}^3 O_p\left(\frac{1}{\sqrt{q}}\right) = O_p\left(\frac{1}{\sqrt{q}^3}\right)\\
\frac{1}{\pi_1\pi_2 q^2}\sum_{i=1}^{\pi_1 q}\sum_{j=1}^{\pi_2 q}x_{ij}^3\bar{x}_{\cdot j} &= O_p\left(\frac{1}{\sqrt{q}^3}\right)\\
\frac{1}{\pi_1\pi_2 q^2}\sum_{i=1}^{\pi_1 q}\sum_{j=1}^{\pi_2 q}x_{ij}\bar{x}_{i\cdot}^3 &= \frac{1}{\pi_1\pi_2 q^2}\sum_{i=1}^{\pi_1 q}\sum_{j=1}^{\pi_2 q}x_{ij} O_p\left(\frac{1}{\sqrt{q}^3}\right) = O_p\left(\frac{1}{\sqrt{q}^5}\right)\\
\frac{1}{\pi_1\pi_2 q^2}\sum_{i=1}^{\pi_1 q}\sum_{j=1}^{\pi_2 q}x_{ij}\bar{x}_{\cdot j}^3 &= O_p\left(\frac{1}{\sqrt{q}^5}\right)\\
\frac{1}{\pi_1\pi_2 q^2}\sum_{i=1}^{\pi_1 q}\sum_{j=1}^{\pi_2 q}\bar{x}_{i\cdot}\bar{x}_{\cdot j}^3 &= O_p\left(\frac{1}{\sqrt{q}}\right)O_p\left(\frac{1}{\sqrt{q}^3}\right) = O_p\left(\frac{1}{q^2}\right)\\
\frac{1}{\pi_1\pi_2 q^2}\sum_{i=1}^{\pi_1 q}\sum_{j=1}^{\pi_2 q}\bar{x}_{i\cdot}^3\bar{x}_{\cdot j} &= O_p\left(\frac{1}{q^2}\right)\\
\frac{1}{\pi_1\pi_2 q^2}\sum_{i=1}^{\pi_1 q}\sum_{j=1}^{\pi_2 q}\bar{x}_{i\cdot}^2\bar{x}_{\cdot j}^2 &= O_p\left(\frac{1}{q}\right)O_p\left(\frac{1}{q}\right) = O_p\left(\frac{1}{q^2}\right)\\
\frac{1}{\pi_1\pi_2 q^2}\sum_{i=1}^{\pi_1 q}\sum_{j=1}^{\pi_2 q}x_{ij}\bar{x}_{i\cdot}^2\bar{x}_{\cdot j} &= O_p\left(\frac{1}{q}\right)O_p\left(\frac{1}{q}\right)O_p\left(\frac{1}{\sqrt{q}}\right) = O_p\left(\frac{1}{\sqrt{q}^5}\right)\\
\frac{1}{\pi_1\pi_2 q^2}\sum_{i=1}^{\pi_1 q}\sum_{j=1}^{\pi_2 q}x_{ij}\bar{x}_{i\cdot}\bar{x}_{\cdot j}^2 &= O_p\left(\frac{1}{\sqrt{q}^5}\right).
\end{aligned}$$

Plugging these expressions into the equation for $\bar{r}^{(2)}$ gives:

$$\bar{r}^{(4)} = \bar{x}^{(4)} + \frac{6}{\pi_1\pi_2 q^2}\sum_{i=1}^{\pi_1 q}\sum_{j=1}^{\pi_2 q}x_{ij}^2\bar{x}_{i\cdot}^2 + \frac{6}{\pi_1\pi_2 q^2}\sum_{i=1}^{\pi_1 q}\sum_{j=1}^{\pi_2 q}x_{ij}^2\bar{x}_{\cdot j}^2 + \frac{12}{\pi_1\pi_2 q^2}\sum_{i=1}^{\pi_1 q}\sum_{j=1}^{\pi_2 q}x_{ij}^2\bar{x}_{i\cdot}\bar{x}_{\cdot j} + O_p\left(\frac{1}{\sqrt{q}^3}\right).$$

Now we compute the means and find the orders of the variances of the remaining terms, starting with $\frac{1}{\pi_1\pi_2 q^2}\sum_{i=1}^{\pi_1 q}\sum_{j=1}^{\pi_2 q}x_{ij}^2\bar{x}_{i\cdot}^2$:

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{\pi_1\pi_2 q^2}\sum_{i=1}^{\pi_1 q}\sum_{j=1}^{\pi_2 q} x_{ij}^2\bar{x}_{i\cdot}^2\right] =& \frac{1}{\pi_1 q}\sum_{i=1}^{\pi_1 q}\mathbb{E}\left[\bar{x}_{i\cdot}^2\left(\frac{1}{\pi_2 q}\sum_{j=1}^{\pi_2 q} x_{ij}^2\right)\right] \\
=& \frac{1}{\pi_1 q}\sum_{i=1}^{\pi_1 q}\mathbb{E}\left[\left(\frac{1}{\pi_2 q}\sum_{j=1}^{\pi_2 q} x_{ij}\right)^2\left(\frac{1}{\pi_2 q}\sum_{j=1}^{\pi_2 q} x_{ij}^2\right)\right] \\
=& \frac{1}{\pi_1 q}\sum_{i=1}^{\pi_1 q}\left(\frac{1}{\pi_2^3 q^3}\mathbb{E}\left[\left(\sum_{j=1}^{\pi_2 q} x_{ij}\right)^2\left(\sum_{j=1}^{\pi_2 q} x_{ij}^2\right)\right]\right) \\
=& \frac{1}{\pi_1 q}\sum_{i=1}^{\pi_1 q}\left(\frac{1}{\pi_2^3 q^3}\left(\pi_2 q\mathbb{E}\left[x_{ij}^4\right] + \pi_2 q\left(\pi_2 q - 1\right)\mathbb{E}\left[x_{ij}^2\right]^2\right)\right) \\
=& \frac{\mathbb{E}\left[x_{ij}^4\right] + \left(\pi_2 q - 1\right)\mathbb{E}\left[x_{ij}^2\right]^2}{\pi_2^2 q^2} \\
=& O\left(\frac{1}{q^2}\right) + \frac{\mathbb{E}\left[x_{ij}^2\right]^2}{\pi_2 q}.
\end{aligned}$$

Note that this term is the mean of $\pi_1 q$ independent, identically distributed terms, so to find order of its variance we just need to find the order of the variance of a single one of the terms.

$$\begin{aligned}
\mathbb{V}\left[\bar{x}_{i\cdot}^2\left(\frac{1}{\pi_2 q}\sum_{j=1}^{\pi_2 q} x_{ij}^2\right)\right] &= \mathbb{E}\left[\left(\bar{x}_{i\cdot}^2\left(\frac{1}{\pi_2 q}\sum_{j=1}^{\pi_2 q} x_{ij}^2\right)\right)^2\right] - \mathbb{E}\left[\bar{x}_{i\cdot}^2\left(\frac{1}{\pi_2 q}\sum_{j=1}^{\pi_2 q} x_{ij}^2\right)\right]^2 \\
&= \frac{1}{\pi_2^6 q^6}\mathbb{E}\left[\left(\sum_{j=1}^{\pi_2 q} x_{ij}\right)^4\left(\sum_{j=1}^{\pi_2 q} x_{ij}^2\right)^2\right] + O\left(\frac{1}{q^2}\right) \\
&= \frac{1}{\pi_2^6 q^6}\Big(\pi_2 q\mathbb{E}\left[x_{ij}^8\right] + \pi_2 q\left(\pi_2 q - 1\right)\mathbb{E}\left[x_{ij}^4\right]^2 + 2\pi_2 q\left(\pi_2 q - 1\right)\mathbb{E}\left[x_{ij}^6\right]\mathbb{E}\left[x_{ij}^2\right] + \\
&\quad 2\pi_2 q\left(\pi_2 q - 1\right)\left(\pi_2 q - 2\right)\mathbb{E}\left[x_{ij}^4\right]\mathbb{E}\left[x_{ij}^2\right]^2 + 6\pi_2 q\left(\pi_2 q - 1\right)\mathbb{E}\left[x_{ij}^6\right]\mathbb{E}\left[x_{ij}^2\right] + \\
&\quad 12\pi_2 q\left(\pi_2 q - 1\right)\mathbb{E}\left[x_{ij}^4\right]\mathbb{E}\left[x_{ij}^4\right] + 12\pi_2 q\left(\pi_2 q - 1\right)\left(\pi_2 q - 2\right)\left(\pi_2 q - 3\right)\mathbb{E}\left[x_{ij}^2\right]^4\Big) + \\
&\quad O\left(\frac{1}{q^2}\right) = O\left(\frac{1}{q^2}\right).
\end{aligned}$$

It follows that $\mathbb{V}\left[\frac{1}{\pi_1\pi_2 q^2}\sum_{i=1}^{\pi_1 q}\sum_{j=1}^{\pi_2 q} x_{ij}^2\bar{x}_{i\cdot}^2\right] = O\left(\frac{1}{q^3}\right)$ and

$$\frac{1}{\pi_1\pi_2 q^2}\sum_{i=1}^{\pi_1 q}\sum_{j=1}^{\pi_2 q} x_{ij}^2\bar{x}_{i\cdot}^2 - \mathbb{E}\left[\frac{1}{\pi_1\pi_2 q^2}\sum_{i=1}^{\pi_1 q}\sum_{j=1}^{\pi_2 q} x_{ij}^2\bar{x}_{i\cdot}^2\right] = O_p\left(\frac{1}{\sqrt{q}^3}\right).$$

By the same logic, $\mathbb{E}\left[\frac{1}{\pi_1\pi_2 q^2}\sum_{i=1}^{\pi_1 q}\sum_{j=1}^{\pi_2 q} x_{ij}^2\bar{x}_{\cdot j}^2\right] = O\left(\frac{1}{q^2}\right) + \frac{\mathbb{E}\left[x_{ij}^2\right]^2}{\pi_1 q}$ and

$$\frac{1}{\pi_1\pi_2 q^2}\sum_{i=1}^{\pi_1 q}\sum_{j=1}^{\pi_2 q} x_{ij}^2\bar{x}_{\cdot j}^2 - \mathbb{E}\left[\frac{1}{\pi_1\pi_2 q^2}\sum_{i=1}^{\pi_1 q}\sum_{j=1}^{\pi_2 q} x_{ij}^2\bar{x}_{\cdot j}^2\right] = O_p\left(\frac{1}{\sqrt{q}^3}\right).$$

Now we just have to compute the mean and order of the variance of one last term, $\frac{1}{\pi_1\pi_2 q^2}\sum_{i=1}^{\pi_1 q}\sum_{j=1}^{\pi_2 q} x_{ij}^2\bar{x}_{i\cdot}\bar{x}_{\cdot j}$.

$$\begin{aligned}\mathbb{E}\left[\frac{1}{\pi_1\pi_2 q^2}\sum_{i=1}^{\pi_1 q}\sum_{j=1}^{\pi_2 q} x_{ij}^2\bar{x}_{i\cdot}\bar{x}_{\cdot j}\right] &= \left(\frac{1}{\pi_1^2\pi_2^2 q^4}\right)\sum_{i=1}^{\pi_1 q}\sum_{j=1}^{\pi_2 q}\mathbb{E}\left[x_{ij}^2\left(\sum_{j'=1}^{\pi_2 q} x_{ij'}\right)\left(\sum_{i'=1}^{\pi_1 q} x_{i'j}\right)\right] \\ &= \frac{\pi_1\pi_2 q^2\mathbb{E}\left[x_{ij}^4\right]}{\pi_1^2\pi_2^2 q^4} = \frac{\mathbb{E}\left[x_{ij}^4\right]}{\pi_1\pi_2 q^2} = O\left(\frac{1}{q^2}\right).\end{aligned}$$

Because this term cannot be written as an average over independent, identically distributed random variables, computing the variance in this case is a little bit trickier. We have

$$\begin{aligned}
\mathbb{V}\left[\frac{1}{\pi_1\pi_2 q^2}\sum_{i=1}^{\pi_1 q}\sum_{j=1}^{\pi_2 q} x_{ij}^2\bar{x}_{i\cdot}\bar{x}_{\cdot j}\right] =& \frac{1}{\pi_1^2\pi_2^2 q^4}\left(\sum_{i=1}^{\pi_1 q}\sum_{j=1}^{\pi_2 q}\mathbb{V}\left[x_{ij}^2\bar{x}_{i\cdot}\bar{x}_{\cdot j}\right] + 2\sum_{i=1}^{\pi_1 q}\sum_{j=1}^{\pi_2 q}\sum_{j'=1,j'\neq j}^{\pi_2 q}\mathrm{Cov}\left[x_{ij}^2\bar{x}_{i\cdot}\bar{x}_{\cdot j}, x_{ij'}^2\bar{x}_{i\cdot}\bar{x}_{\cdot j'}\right] + \right. \\
& 2\sum_{i=1}^{\pi_1 q}\sum_{j=1}^{\pi_2 q}\sum_{i'=1,i'\neq i}^{\pi_1 q}\mathrm{Cov}\left[x_{ij}^2\bar{x}_{i\cdot}\bar{x}_{\cdot j}, x_{i'j}^2\bar{x}_{i'\cdot}\bar{x}_{\cdot j}\right] + \\
& \left. 2\sum_{i=1}^{\pi_1 q}\sum_{j=1}^{\pi_2 q}\sum_{i'=1,i'\neq i}^{\pi_1 q}\sum_{j'=1,j'\neq j}^{\pi_2 q}\mathrm{Cov}\left[x_{ij}^2\bar{x}_{i\cdot}\bar{x}_{\cdot j}, x_{i'j'}^2\bar{x}_{i'\cdot}\bar{x}_{\cdot j'}\right]\right) \\
=& \frac{1}{\pi_1^2\pi_2^2 q^4}\left(\pi_1\pi_2 q^2\mathbb{V}\left[x_{ij}^2\bar{x}_{i\cdot}\bar{x}_{\cdot j}\right] + 2\pi_1\pi_2 q^2\left(\pi_2 q-1\right)\mathrm{Cov}\left[x_{ij}^2\bar{x}_{i\cdot}\bar{x}_{\cdot j}, x_{ij'}^2\bar{x}_{i\cdot}\bar{x}_{\cdot j'}\right] + \right. \\
& 2\pi_1\pi_2 q^2\left(\pi_2 q-1\right)\mathrm{Cov}\left[x_{ij}^2\bar{x}_{i\cdot}\bar{x}_{\cdot j}, x_{i'j}^2\bar{x}_{i'\cdot}\bar{x}_{\cdot j}\right] + \\
& \left. 2\pi_1\pi_2 q^2\left(\pi_1\pi_2 q^2-\pi_1 q-\pi_2 q+1\right)\mathrm{Cov}\left[x_{ij}^2\bar{x}_{i\cdot}\bar{x}_{\cdot j}, x_{i'j'}^2\bar{x}_{i'\cdot}\bar{x}_{\cdot j'}\right]\right) \\
=& \frac{1}{\pi_1\pi_2 q^2}\mathbb{V}\left[x_{ij}^2\bar{x}_{i\cdot}\bar{x}_{\cdot j}\right] + \frac{2\left(\pi_2 q-1\right)}{\pi_1\pi_2 q^2}\mathrm{Cov}\left[x_{ij}^2\bar{x}_{i\cdot}\bar{x}_{\cdot j}, x_{ij'}^2\bar{x}_{i\cdot}\bar{x}_{\cdot j'}\right] + \\
& \frac{2\left(\pi_2 q-1\right)}{\pi_1\pi_2 q^2}\mathrm{Cov}\left[x_{ij}^2\bar{x}_{i\cdot}\bar{x}_{\cdot j}, x_{i'j}^2\bar{x}_{i'\cdot}\bar{x}_{\cdot j}\right] + \\
& \frac{2\left(\pi_1\pi_2 q^2-\pi_1 q-\pi_2 q+1\right)}{\pi_1\pi_2 q^2}\mathrm{Cov}\left[x_{ij}^2\bar{x}_{i\cdot}\bar{x}_{\cdot j}, x_{i'j'}^2\bar{x}_{i'\cdot}\bar{x}_{\cdot j'}\right] \\
=& O\left(\frac{1}{q^2}\right)\mathbb{V}\left[x_{ij}^2\bar{x}_{i\cdot}\bar{x}_{\cdot j}\right] + O\left(\frac{1}{q}\right)\mathrm{Cov}\left[x_{ij}^2\bar{x}_{i\cdot}\bar{x}_{\cdot j}, x_{ij'}^2\bar{x}_{i\cdot}\bar{x}_{\cdot j'}\right] + \\
& O\left(\frac{1}{q}\right)\mathrm{Cov}\left[x_{ij}^2\bar{x}_{i\cdot}\bar{x}_{\cdot j}, x_{i'j}^2\bar{x}_{i'\cdot}\bar{x}_{\cdot j}\right] + O\left(1\right)\mathrm{Cov}\left[x_{ij}^2\bar{x}_{i\cdot}\bar{x}_{\cdot j}, x_{i'j'}^2\bar{x}_{i'\cdot}\bar{x}_{\cdot j'}\right].
\end{aligned}$$

Examining the first term we have:

$$\begin{aligned}
\mathbb{V}\left[x_{ij}^2\bar{x}_{i\cdot}\bar{x}_{\cdot j}\right] &= \mathbb{E}\left[x_{ij}^4\bar{x}_{i\cdot}^2\bar{x}_{\cdot j}^2\right] - \mathbb{E}\left[x_{ij}^2\bar{x}_{i\cdot}\bar{x}_{\cdot j}\right]^2 \\
&= \left(\frac{1}{\pi_1^2\pi_2^2 q^4}\right)\mathbb{E}\left[x_{ij}^4\left(\sum_{j'=1}^{\pi_2 q}x_{ij'}\right)^2\left(\sum_{i'=1}^{\pi_1 q}x_{i'j}\right)^2\right] + O\left(\frac{1}{q^4}\right) \\
&= \left(\frac{\mathbb{E}\left[x_{ij}^8\right] + \pi_1 q\left(\pi_2 q-1\right)\mathbb{E}\left[x_{ij}^4\right]\mathbb{E}\left[x_{ij}^2\right]^2}{\pi_1^2\pi_2^2 q^4}\right) + O\left(\frac{1}{q^4}\right) = O\left(\frac{1}{q^2}\right).
\end{aligned}$$

Then examining the second term we have:

$$\begin{aligned}\text{Cov}\left[x_{ij}^2\bar{x}_{i\cdot}\bar{x}_{\cdot j}, x_{ij'}^2\bar{x}_{i\cdot}\bar{x}_{\cdot j'}\right] =& \mathbb{E}\left[x_{ij}^2x_{ij'}^2\bar{x}_{i\cdot}^2\bar{x}_{\cdot j}\bar{x}_{\cdot j'}\right] - \mathbb{E}\left[x_{ij}^2\bar{x}_{i\cdot}\bar{x}_{\cdot j}\right]^2 \\ =& \frac{1}{\pi_1^2\pi_2^2q^4}\mathbb{E}\left[x_{ij}^2x_{ij'}^2\left(\sum_{j''}^{\pi_2 q}x_{ij''}\right)^2\left(\sum_{i'=1}^{\pi_1 q}x_{i'j}\right)\left(\sum_{i''=1}^{\pi_1 q}x_{i''j}\right)\right] + O\left(\frac{1}{q^4}\right) \\ =& \frac{2\mathbb{E}\left[x_{ij}^4x_{ij'}^4\right]}{\pi_1^2\pi_2^2q^4} + O\left(\frac{1}{q^4}\right) = O\left(\frac{1}{q^4}\right).\end{aligned}$$

The same logic yields $\text{Cov}\left[x_{ij}^2\bar{x}_{i\cdot}\bar{x}_{\cdot j}, x_{ij'}^2\bar{x}_{i\cdot}\bar{x}_{\cdot j'}\right] = O\left(\frac{1}{q^4}\right)$.

Then examining the last term yields:

$$\begin{aligned}\text{Cov}\left[x_{ij}^2\bar{x}_{i\cdot}\bar{x}_{\cdot j}, x_{i'j'}^2\bar{x}_{i'\cdot}\bar{x}_{\cdot j'}\right] &= \mathbb{E}\left[x_{ij}^2x_{i'j'}^2\bar{x}_{i\cdot}\bar{x}_{i'\cdot}\bar{x}_{\cdot j}\bar{x}_{\cdot j'}\right] - \mathbb{E}\left[x_{ij}\bar{x}_{i\cdot}\bar{x}_{\cdot j}\right]^2 \\ &= \mathbb{E}\left[x_{ij}^2\bar{x}_{i\cdot}\bar{x}_{\cdot j}\right]\mathbb{E}\left[x_{i'j'}^2\bar{x}_{i'\cdot}\bar{x}_{\cdot j'}\right] - \mathbb{E}\left[x_{ij}\bar{x}_{i\cdot}\bar{x}_{\cdot j}\right]^2 = 0.\end{aligned}$$

Putting the pieces together, we get:

$$\mathbb{V}\left[\frac{1}{\pi_1\pi_2q^2}\sum_{i=1}^{\pi_1 q}\sum_{j=1}^{\pi_2 q}x_{ij}^2\bar{x}_{i\cdot}\bar{x}_{\cdot j}\right] = O\left(\frac{1}{q^2}\right)O\left(\frac{1}{q^2}\right) + O\left(\frac{1}{q}\right)O\left(\frac{1}{q^4}\right) + O\left(\frac{1}{q}\right)O\left(\frac{1}{q^4}\right) = O\left(\frac{1}{q^4}\right).$$

It follows that

$$\frac{1}{\pi_1\pi_2q^2}\sum_{i=1}^{\pi_1 q}\sum_{j=1}^{\pi_2 q}x_{ij}^2\bar{x}_{i\cdot}\bar{x}_{\cdot j} - \mathbb{E}\left[\frac{1}{\pi_1\pi_2q^2}\sum_{i=1}^{\pi_1 q}\sum_{j=1}^{\pi_2 q}x_{ij}^2\bar{x}_{i\cdot}\bar{x}_{\cdot j}\right] = O_p\left(\frac{1}{q^2}\right).$$

Substituting these expressions into our last equation for $\bar{r}^{(4)}$, we get:

$$\begin{aligned}
\bar{r}^{(4)} =& \bar{x}^{(4)} + 6\left(\frac{1}{\pi_1\pi_2 q^2}\sum_{i=1}^{\pi_1 q}\sum_{j=1}^{\pi_2 q} x_{ij}^2\bar{x}_{i\cdot}^2 - \mathbb{E}\left[\frac{1}{\pi_1\pi_2 q^2}\sum_{i=1}^{\pi_1 q}\sum_{j=1}^{\pi_2 q} x_{ij}^2\bar{x}_{i\cdot}^2\right]\right) + \mathbb{E}\left[\frac{1}{\pi_1\pi_2 q^2}\sum_{i=1}^{\pi_1 q}\sum_{j=1}^{\pi_2 q} x_{ij}^2\bar{x}_{i\cdot}^2\right] + \\
& 6\left(\frac{1}{\pi_1\pi_2 q^2}\sum_{i=1}^{\pi_1 q}\sum_{j=1}^{\pi_2 q} x_{ij}^2\bar{x}_{\cdot j}^2 - \mathbb{E}\left[\frac{1}{\pi_1\pi_2 q^2}\sum_{i=1}^{\pi_1 q}\sum_{j=1}^{\pi_2 q} x_{ij}^2\bar{x}_{\cdot j}^2\right]\right) + 6\mathbb{E}\left[\frac{1}{\pi_1\pi_2 q^2}\sum_{i=1}^{\pi_1 q}\sum_{j=1}^{\pi_2 q} x_{ij}^2\bar{x}_{\cdot j}^2\right] + \\
& 12\left(\frac{1}{\pi_1\pi_2 q^2}\sum_{i=1}^{\pi_1 q}\sum_{j=1}^{\pi_2 q} x_{ij}^2\bar{x}_{i\cdot}\bar{x}_{\cdot j} - \mathbb{E}\left[\frac{1}{\pi_1\pi_2 q^2}\sum_{i=1}^{\pi_1 q}\sum_{j=1}^{\pi_2 q} x_{ij}^2\bar{x}_{i\cdot}\bar{x}_{\cdot j}\right]\right) + 12\mathbb{E}\left[\frac{1}{\pi_1\pi_2 q^2}\sum_{i=1}^{\pi_1 q}\sum_{j=1}^{\pi_2 q} x_{ij}^2\bar{x}_{i\cdot}\bar{x}_{\cdot j}\right] + \\
& O_p\left(\frac{1}{\sqrt{q}^3}\right) \\
=& \bar{x}^{(4)} + O_p\left(\frac{1}{\sqrt{q}^3}\right) + O\left(\frac{1}{q^2}\right) + \frac{6\mathbb{E}\left[x_{ij}^2\right]^2}{\pi_1 q} + \\
& O_p\left(\frac{1}{\sqrt{q}^3}\right) + O\left(\frac{1}{q^2}\right) + \frac{6\mathbb{E}\left[x_{ij}^2\right]^2}{\pi_2 q} + O_p\left(\frac{1}{q^2}\right) + O\left(\frac{1}{q^2}\right) + O_p\left(\frac{1}{\sqrt{q}^3}\right) \\
=& \bar{x}^{(4)} + \frac{6\mathbb{E}\left[x_{ij}^2\right]^2}{\pi_1\pi_2 q} + O\left(\frac{1}{q^2}\right) + O_p\left(\frac{1}{\sqrt{q}^3}\right).
\end{aligned}$$

Now we plug these expressions for $\bar{r}^{(2)}$ and $\bar{r}^{(4)}$ into our equation for $\hat{\sigma}_c^4$:

$$\begin{aligned}
\hat{\sigma}_c^4 =& O(1)\left(\bar{r}^{(4)}/3 - \left(\bar{r}^{(2)}\right)^2\right) \\
=& O(1)\left(\bar{x}^{(4)}/3 + \frac{2\mathbb{E}\left[x_{ij}^2\right]^2}{\pi_1\pi_2 q} + O\left(\frac{1}{q^2}\right) + O_p\left(\frac{1}{\sqrt{q}^3}\right) - \left(\bar{x}^{(2)} + \frac{\mathbb{E}\left[x_{ij}^2\right]}{\pi_1\pi_2 q} + O_p\left(\frac{1}{\sqrt{q}^3}\right)\right)^2\right) \\
=& O(1)\left(\bar{x}^{(4)}/3 - \left(\bar{x}^{(2)}\right)^2 + \frac{2\mathbb{E}\left[x_{ij}^2\right]^2}{\pi_1\pi_2 q} + O\left(\frac{1}{q^2}\right) + O_p\left(\frac{1}{\sqrt{q}^3}\right) - \right. \\
& \left. \underbrace{\left(\frac{\mathbb{E}\left[x_{ij}^2\right]}{\pi_1\pi_2 q}\right)^2}_{O\left(\frac{1}{q^2}\right)} + O_p\left(\frac{1}{q^3}\right) + O_p\left(\frac{1}{\sqrt{q}^3}\right) + O_p\left(\frac{1}{\sqrt{q}^5}\right) - 2\left(\bar{x}^{(2)}\right)\left(\frac{\mathbb{E}\left[x_{ij}^2\right]}{\pi_1\pi_2 q}\right)\right) \\
=& O(1)\left(\bar{x}^{(4)}/3 - \left(\bar{x}^{(2)}\right)^2 + \frac{2\mathbb{E}\left[x_{ij}^2\right]^2}{\pi_1\pi_2 q} - 2\underbrace{\left(\bar{x}^{(2)} - \mathbb{E}\left[x_{ij}^2\right]\right)}_{O_p\left(\frac{1}{q}\right)}\underbrace{\left(\frac{\mathbb{E}\left[x_{ij}^2\right]}{\pi_1\pi_2 q}\right)}_{O\left(\frac{1}{q}\right)} - 2\left(\frac{\mathbb{E}\left[x_{ij}^2\right]^2}{\pi_1\pi_2 q}\right) + \right. \\
& \left. O\left(\frac{1}{q^2}\right) + O_p\left(\frac{1}{\sqrt{q}^3}\right)\right) \\
=& O(1)\left(\bar{x}^{(4)}/3 - \left(\bar{x}^{(2)}\right)^2 + O\left(\frac{1}{q^2}\right) + O_p\left(\frac{1}{\sqrt{q}^3}\right)\right).
\end{aligned}$$

Recalling that $\kappa$ is defined such that $\mathbb{E}\left[x_{ij}^4\right]/3 - \mathbb{E}\left[x_{ij}^2\right]^2 = \kappa\sigma^4/3$, we have

$$\sqrt{\pi_1\pi_2 q^2}\left(\hat{\sigma}_c^4 - \kappa\sigma_c^4/3\right) \xrightarrow{d} \sqrt{\pi_1\pi_2 q^2}\left(\bar{x}^{(4)}/3 - \left(\bar{x}^2\right)^2 - \kappa\sigma_c^4/3\right)$$

as $q \to \infty$.

But what is the distribution of the right hand side? As $\bar{x}^{(2)}$ and $\bar{x}^{(4)}$ are sample moments of independent, identically distributed random variables $x_{ij}$, we know that as $q \to \infty$:

$$\sqrt{\pi_1\pi_2 q^2}\begin{pmatrix} \bar{x}^{(2)} - \mathbb{E}\left[x_{ij}^2\right] \\ \bar{x}^{(4)} - \mathbb{E}\left[x_{ij}^4\right] \end{pmatrix} \xrightarrow{d} N\left(\mathbf{0}, \begin{pmatrix} \mathbb{V}\left[x_{ij}^2\right] & \mathrm{Cov}\left[x_{ij}^2, x_{ij}^4\right] \\ \mathrm{Cov}\left[x_{ij}^2, x_{ij}^4\right] & \mathbb{V}\left[x_{ij}^4\right] \end{pmatrix}\right).$$

Let $f(y,z) = z/3 - y^2$ and note that $f\left(\mathbb{E}\left[x_{ij}^2\right], \mathbb{E}\left[x_{ij}^4\right]\right) = \kappa\sigma_c^4/3$ and

$$\nabla f\left(\mathbb{E}\left[x_{ij}^2\right], \mathbb{E}\left[x_{ij}^4\right]\right) = \begin{pmatrix} -2\mathbb{E}\left[x_{ij}^2\right] \\ 1/3 \end{pmatrix}.$$

Applying the delta method yields:

$$\sqrt{\pi_1\pi_2 q^2}\left(\bar{x}^{(4)}/3 - \left(\bar{x}^{(2)}\right)^2 - \kappa\sigma_c^4/3\right) \xrightarrow{d} \tag{A.8}$$
$$N\left(0, 4\mathbb{E}\left[x_{ij}^2\right]^2 \mathbb{V}\left[x_{ij}^2\right] - 4\mathbb{E}\left[x_{ij}^2\right] \mathrm{Cov}\left[x_{ij}^2, x_{ij}^4\right]/3 + \mathbb{V}\left[x_{ij}^4\right]/9\right)$$

as $q \to \infty$.

Up to this point, we have just used the assumption that $x_{ij} = c_{ij} + z_{ij}$, where $z_{ij}$ is normal and both $c_{ij}$ and $z_{ij}$ are independent, identically distributed from symmetric, mean zero distributions. This will be convenient later, when we need to find the asymptotic distribution of $\hat{\sigma}_c^4$ assuming (a) $c_{ij}$ are normally distributed and/or exactly equal to zero and (b) assuming $c_{ij}$ are drawn from a Bernoulli-normal spike-and-slab prior. However here, we are interested in finding the asymptotic variance of $\hat{\sigma}_c^4$ assuming $c_{ij}$ are Laplace distributed with variance $\sigma_c^2$. In this case, we have:

$$\begin{aligned}
\mathbb{E}\left[x_{ij}^2\right] &= \sigma_c^2 + \sigma_z^2 \\
\mathbb{E}\left[x_{ij}^4\right] &= 3\sigma_c^4 + 3\left(\sigma_c^2 + \sigma_z^2\right)^2 \\
\mathbb{V}\left[x_{ij}^2\right] &= 5\sigma_c^4 + 4\sigma_c^2\sigma_z^2 + 2\sigma_z^4 \\
\mathrm{Cov}\left[x_{ij}^2, x_{ij}^4\right] &= 6(14\sigma_c^6 + 13\sigma_c^4\sigma_z^2 + 6\sigma_c^2\sigma_z^4 + 2\sigma_z^6) \\
\mathbb{V}\left[x_{ij}^4\right] &= 12(207\sigma_c^8 + 204\sigma_c^6\sigma_z^2 + 99\sigma_c^4\sigma_z^4 + 32\sigma_c^2\sigma^6 + 8\sigma_z^8).
\end{aligned}$$

Substituting these in to our expression for the asymptotic variance of $\hat{\sigma}_c^4$ and simplifying yields and substituting $n$ and $p$ back in for $\pi_1 q$ and $\pi_2 q$, we have

$$\sqrt{np}\left(\hat{\sigma}_c^4 - \sigma_c^4\right) \xrightarrow{d} N\left(0, 8(69\sigma_c^8 + 42\sigma_c^6\sigma_z^2 + 15\sigma_c^4\sigma^4 + 4\sigma_c^2\sigma_z^6 + \sigma_z^8)/3\right) \tag{A.9}$$

as $n, p \to \infty$.

### *A.6 Derivation of Asymptotic Variance of $\hat{\lambda}_c$*

We can obtain the asymptotic variance of $\hat{\lambda}_c = \sqrt{2}/(\hat{\sigma}_c^4)^{1/4}$ from the asymptotic distribution of $\hat{\sigma}_c^2$ via the delta method. The asymptotic variance of $\hat{\lambda}_c$ is given by:

$$\left(-\frac{\sqrt{2}}{4\sigma_c^5}\right)^2 \left(\frac{8}{3}\right)\left(69\sigma_c^8 + 42\sigma_c^6\sigma_z^2 + 15\sigma_c^4\sigma^4 + 4\sigma_c^2\sigma_z^6 + \sigma_z^8\right)/np = \left(\frac{1}{3\sigma_c^2}\right)\left(69 + 42\psi^2 + 15\psi^4 + 4\psi^6 + \psi^8\right)/np,$$

where $\phi^2 = \sigma_z^2/\sigma_c^2$.

### *A.7 Derivation of Asymptotic Joint Distribution of $\hat{\sigma}_c^2$ and $\hat{\sigma}_z^2$*

Again, we let $x_{ij} = c_{ij} + z_{ij}$. Using the derivation given in Section 5, we have:

$$\sqrt{np}\left(\begin{array}{c}\bar{r}^2 \\ \bar{r}^4\end{array} - \left(\begin{array}{c}\sigma_c^2 + \sigma_z^2 \\ 3\sigma_c^4 + 3\left(\sigma_c^2 + \sigma_z^2\right)\end{array}\right)\right) \xrightarrow{d} \sqrt{np}\left(\begin{array}{c}\bar{x}^2 \\ \bar{x}^4\end{array} - \left(\begin{array}{c}\sigma_c^2 + \sigma_z^2 \\ 3\sigma_c^4 + 3\left(\sigma_c^2 + \sigma_z^2\right)\end{array}\right)\right)$$

as $n, p \to \infty$.

Applying the delta method and letting $\psi^2 = \sigma_z^2/\sigma_c^2$ yields:

$$\sqrt{np}\left(\begin{array}{c}\hat{\sigma}_c^2 \\ \hat{\sigma}_z^2\end{array} - \left(\begin{array}{c}\sigma_c^2 \\ \sigma_z^2\end{array}\right)\right) \xrightarrow{d} N\left(\mathbf{0}, \frac{\sigma_c^4}{3}\mathbf{\Sigma}\right),$$

where

$$\begin{aligned}\sigma_{11} &= 2\left(69 + 42\psi^2 + 15\psi^4 + 4\psi^6 + \psi^8\right) \\ \sigma_{12} &= -\left(111 + 72\psi^2 + 30\psi^4 + 8\psi^6 + 2\psi^8\right) \\ \sigma_{22} &= 99 + 72\psi^2 + 36\psi^4 + 8\psi^6 + 2\psi^8.\end{aligned}$$

Note that $\mathbf{\Sigma}$ only depends on $\psi^2$, the ratio of $\sigma_z^2$ to $\sigma_c^2$.

Letting $\boldsymbol{z} = \sqrt{\frac{3np}{\sigma_c^4}}\left(\begin{array}{c}\hat{\sigma}_c^2 \\ \hat{\sigma}_z^2\end{array} - \left(\begin{array}{c}\sigma_c^2 \\ \sigma_z^2\end{array}\right)\right)$ and noting that $\boldsymbol{z} \xrightarrow{d} N\left(\mathbf{0}, \mathbf{\Sigma}\right)$ as $n, p \to \infty$,

$$\begin{aligned}1 - \Pr\left(\left(\begin{array}{c}\hat{\sigma}_c^2 \\ \hat{\sigma}_z^2\end{array}\right) \geq \mathbf{0}\right) &= 1 - \Pr\left(\boldsymbol{z} \geq -\frac{1}{\sigma_c^2}\left(\begin{array}{c}\sigma_c^2 \\ \sigma_z^2\end{array}\right)\right) \\ &= 1 - \Pr\left(\boldsymbol{z} \geq -\sqrt{3np}\left(\begin{array}{c}1 \\ \psi^2\end{array}\right)\right).\end{aligned}$$

This probability depends only on the dimensions $n$ and $p$ and the ratio of the variances $\psi$.

### *A.8 Proof of Proposition 2.3.1*

The proof of Proposition 2.3.1 follows directly from the derivation of the asymptotic distribution of $\hat{\sigma}_c^4$ given in the previous section. To prove Proposition 2.3.1, we need to find the asymptotic distribution of $\hat{\sigma}_c^4$ under the null hypothesis, $H_0$: $c_{ij} + z_{ij} \sim$ i.i.d. $\mathrm{N}\left(0, \sigma_c^2 + \sigma_z^2\right)$. Under this null hypothesis, $c_{ij} + z_{ij} \stackrel{d}{=} \tilde{c}_{ij} + \tilde{z}_{ij}$, where $\tilde{c}_{ij}$ is Laplace distributed with 0 and $\tilde{z}_{ij}$ is normally distributed with variance $\sigma_c^2 + \sigma_z^2$. Applying equation (A.9), yields:

$$\sqrt{np}\hat{\sigma}_c^4 \stackrel{d}{\to} N\left(0, 8\left(\sigma_c^2 + \sigma_z^2\right)^4/3\right)$$

as $n, p \to \infty$.

However when we are testing the null hypothesis, $\sigma_c^2 + \sigma_z^2$ is unknown. Fortunately, $\left(\hat{\sigma}_c^2 + \hat{\sigma}_z^2\right)^4 = \left(\left(\frac{np}{(n-1)(p-1)}\right)\bar{r}^{(2)}\right)^4 \stackrel{p}{\to} \sigma_c^2 + \sigma_z^2$ as $n, p \to \infty$.

It follows directly from Equations (A.5) and (A.6) that $\hat{\sigma}_c^2 + \hat{\sigma}_z^2 \stackrel{p}{\to} \sigma_c^2 + \sigma_z^2$ as $n, p \to \infty$, and by the continuous mapping theorem $\left(\hat{\sigma}_c^2 + \hat{\sigma}_z^2\right)^4 \stackrel{p}{\to} \left(\sigma_c^2 + \sigma_z^2\right)^4$ as $n, p \to \infty$. Accordingly,

$$\sqrt{np}\left(\frac{\hat{\sigma}_c^4}{\sqrt{\frac{8}{3}}\left(\hat{\sigma}_c^2 + \hat{\sigma}_z^2\right)^2}\right) \stackrel{d}{\to} N\left(0, 1\right)$$

as $n, p \to \infty$ and under $H_0$, $\Pr\left(\sqrt{np}\left(\frac{\hat{\sigma}_c^4}{\sqrt{\frac{8}{3}}(\hat{\sigma}_c^2+\hat{\sigma}_z^2)^2}\right) > z_{1-\alpha}\right) \to \alpha$ as $n, p \to \infty$.

### *A.9 Proof of Proposition 2.3.2*

Combining Equation (A.9) and consistency of $\hat{\sigma}_c^2 + \hat{\sigma}_z^2$ for $\sigma_c^2 + \sigma_z^2$ as $n, p \to \infty$ that follows from Equations (A.5) and (A.6), we have:

$$\sqrt{np}\left(\frac{\hat{\sigma}_c^4}{\sqrt{\frac{8}{3}}\left(\hat{\sigma}_c^2 + \hat{\sigma}_z^2\right)^2} - \frac{\sigma_c^4}{\sqrt{\frac{8}{3}}\left(\sigma_c^2 + \sigma_z^2\right)^2}\right) \stackrel{d}{\to} N\left(0, \frac{69\sigma_c^8 + 42\sigma_c^6\sigma_z^2 + 15\sigma_c^4\sigma^4 + 4\sigma_c^2\sigma_z^6 + \sigma_z^8}{\left(\sigma_c^4 + \sigma_z^2\right)^4}\right).$$

We can simplify the mean and variance and write them as a function of $\phi^2 = \sigma_c^2/\sigma_z^2$:

$$\begin{aligned}
\frac{\sigma_c^4}{\sqrt{\frac{8}{3}}\left(\sigma_c^2+\sigma_z^2\right)^2} &= \sqrt{\frac{3}{8}}\left(\frac{\phi^4}{\left(\phi^2+1\right)^2}\right) \\
\frac{69\sigma_c^8+42\sigma_c^6\sigma_z^2+15\sigma_c^4\sigma^4+4\sigma_c^2\sigma_z^6+\sigma_z^8}{\left(\sigma_c^4+\sigma_z^2\right)^4} &= \frac{68\sigma_c^8+36\sigma_c^6\sigma_z^2+9\sigma_c^4\sigma^4+\left(\sigma_c^2+\sigma_c^2\right)^4}{\left(\sigma_c^4+\sigma_z^2\right)^4} \\
&= 1+\frac{68\sigma_c^8+36\sigma_c^6\sigma_z^2+9\sigma_c^4\sigma^4}{\left(\sigma_c^4+\sigma_z^2\right)^4} \\
&= 1+\frac{68\phi^8+36\phi^6+9\phi^4}{\left(\phi^4+1\right)^4}.
\end{aligned}$$

It follows that:

$$\Pr\left(\sqrt{np}\left(\frac{\hat{\sigma}_c^4}{\sqrt{\frac{8}{3}}\left(\hat{\sigma}_c^2+\hat{\sigma}_z^2\right)^2}\right) > z_{1-\alpha}\right) \to 1-\Phi\left(\frac{z_{1-\alpha}-\sqrt{\frac{3np}{8}}\left(\frac{\phi^4}{(\phi^2+1)^2}\right)}{\sqrt{1+\frac{68\phi^8+36\phi^6+9\phi^4}{(\phi^4+1)^4}}}\right)$$

as $n, p \to \infty$.

### *A.10 Proof of Proposition 2.3.3*

To prove Proposition 2.3.3, we first need to compute the unknown quantities that appear in Equation (A.8):

$$\begin{aligned}
\mathbb{E}\left[x_{ij}^2\right] =& \pi_c\tau_c^2+\sigma_z^2 \\
\mathbb{E}\left[x_{ij}^4\right] =& 3\left(\pi_c\tau_c^4+2\pi_c\tau_c^2\sigma_z^2+\sigma_z^4\right) \\
\mathbb{E}\left[x_{ij}^6\right] =& \mathbb{E}\left[c_{ij}^6\right]+15\mathbb{E}\left[c_{ij}^4\right]\mathbb{E}\left[z_{ij}^2\right]+15\mathbb{E}\left[c_{ij}^2\right]\mathbb{E}\left[z_{ij}^4\right]+\mathbb{E}\left[z_{ij}^6\right] \\
=& 15\left(\pi_c\tau_c^6+3\pi_c\tau_c^4\sigma_z^2+3\pi_c\tau_c^2\sigma_z^4+\sigma_z^6\right) \\
\mathbb{E}\left[x_{ij}^8\right] =& \mathbb{E}\left[c_{ij}^8\right]+28\mathbb{E}\left[c_{ij}^6\right]\mathbb{E}\left[z_{ij}^2\right]+70\mathbb{E}\left[c_{ij}^4\right]\mathbb{E}\left[z_{ij}^4\right]+28\mathbb{E}\left[c_{ij}^2\right]\mathbb{E}\left[z_{ij}^6\right]+\mathbb{E}\left[z_{ij}^8\right] \\
=& 105\left(\pi_c\tau_c^8+4\pi_c\tau_c^6\sigma_z^2+6\pi_c\tau_c^4\sigma_z^4+4\pi_c\tau_c^2\sigma_z^6+\sigma_z^8\right) \\
\mathbb{V}\left[x_{ij}^2\right] =& \left(3-\pi_c\right)\pi_c\tau_c^4+4\pi_c\tau_c^2\sigma_z^2+2\sigma_z^2 \\
\mathrm{Cov}\left[x_{ij}^2, x_{ij}^4\right] =& 12\sigma_z^6+36\pi_c\sigma_z^4\tau_c^2+42\pi_c\sigma_z^2\tau_c^4-6\pi_c^2\sigma_z^2\tau_c^4+15\pi_c\tau_c^6-3\pi_c^2\tau_c^6 \\
\mathbb{V}\left[x_{ij}^4\right] =& 96\sigma_z^8+384\pi_c\sigma_z^6\tau_c^2+612\pi_c\sigma_z^4\tau_c^4-36\pi_c^2\sigma_z^4\tau_c^4+420\pi_c\sigma_z^2\tau_c^6-36\pi_c^2\sigma_z^2\tau_c^6+105\pi_c\tau_c^8-9\pi_c^2\tau^8.
\end{aligned}$$

Equation (A.8) yields the following variance:

$$\frac{1}{3}\left(8\sigma_z^8 + 32\pi_c\sigma_z^6\tau_c^2 + 24\left(3-\pi_c\right)\pi_c\sigma_z^4\tau_c^4 + 16\pi_c\left(5-6\pi_c+3\pi_c^2\right)\sigma_z^2\tau_c^6 + \pi_c\left(35-63\pi_c+48\pi_c^2-12\pi_c^3\right)\tau_c^8\right).$$

This can be simplified slightly to:

$$\frac{1}{3}\left(8\left(\pi_c\tau_c^2+\sigma_c^2\right)^4 + \pi_c\left(1-\pi_c\right)\left(72\tau_c^4\sigma_z^4 + 16\left(5-\pi_c\right)\tau_c^6\sigma_z^2 + \left(35-28\pi_c+20\right)\tau_c^8\right)\right)$$

Now combining this with consistency of $\hat{\sigma}_c^2+\hat{\sigma}_z^2$ for $\sigma_c^2+\sigma_z^2 = \pi_c\tau_c^2+\sigma_z^2$ as $n,p\to\infty$ that follows directly from from Equations (A.5) and (A.6), we have:

$$\sqrt{np}\left(\frac{\hat{\sigma}_c^4}{\sqrt{\frac{8}{3}}\left(\hat{\sigma}_c^2+\hat{\sigma}_z^2\right)^2} - \frac{\pi_c\left(1-\pi_c\right)\tau_c^4}{\sqrt{\frac{8}{3}}\left(\pi_c\tau_c^2+\sigma_z^2\right)^2}\right) \xrightarrow{d}$$
$$N\left(0, 1+\frac{\pi_c\left(1-\pi_c\right)\left(72\tau_c^4\sigma_z^4 + 16\left(5-\pi_c\right)\tau_c^6\sigma_z^2 + \left(35-28\pi_c+20\right)\tau_c^8\right)}{8\left(\pi_c\tau_c^2+\sigma_z^2\right)^4}\right).$$

We can simplify the mean and variance and write them as a function of $\phi^2 = \tau_c^2/\sigma_z^2$:

$$\frac{\pi_c\left(1-\pi_c\right)\tau_c^4}{\sqrt{\frac{8}{3}}\left(\pi_c\tau_c^2+\sigma_z^2\right)^2} = \frac{\pi_c\left(1-\pi_c\right)\phi^4}{\sqrt{\frac{8}{3}}\left(\pi_c\phi^2+1\right)^2}$$
$$1+\frac{\pi_c\left(1-\pi_c\right)\left(72\tau_c^4\sigma_z^4 + 16\left(5-\pi_c\right)\tau_c^6\sigma_z^2 + \left(35-28\pi_c+20\right)\tau_c^8\right)}{8\left(\pi_c\tau_c^2+\sigma_z^2\right)^4} =$$
$$1+\frac{\pi_c\left(1-\pi_c\right)\left(72\phi^4 + 16\left(5-\pi_c\right)\phi^6 + \left(35-28\pi_c+20\right)\phi^8\right)}{8\left(\pi_c\phi^2+1\right)^4}$$

It follows that:

$$\Pr\left(\sqrt{np}\left(\frac{\hat{\sigma}_c^4}{\sqrt{\frac{8}{3}}\left(\hat{\sigma}_c^2+\hat{\sigma}_z^2\right)^2}\right) > z_{1-\alpha}\right) \to 1-\Phi\left(\frac{z_{1-\alpha} - \sqrt{\frac{3np}{8}}\frac{\pi_c(1-\pi_c)\phi^4}{\left(\pi_c\phi^2+1\right)^2}}{\sqrt{1+\frac{\pi_c(1-\pi_c)(72\phi^4+16(5-\pi_c)\phi^6+(35-28\pi_c+20\pi_c^2)\phi^8)}{8(\pi_c\phi^2+1)^4}}}\right)$$

as $n,p\to\infty$.

### *A.11 Proof of Proposition 2.3.4*

First, we consider the case where $p\to\infty$ with $n$ fixed. The case where $p\to\infty$ with $n$ fixed is analogous. In the proof of Proposition 2.2.2, we show that $\bar{r}^{(2)} \xrightarrow{p} \left(\frac{n-1}{n}\right)\left(\sigma_c^2+\sigma_z^2\right)$ and

$\bar{r}^{(4)} \xrightarrow{p} \kappa \left( \frac{(n-1)\left(n^2-3n+3\right)}{n^3} \right) \sigma_c^4 + 3\left(\frac{n-1}{n}\right)^2 \left(\sigma_c^2 + \sigma_z^2\right)^2$ as $p \to \infty$ with $n$ fixed. Applying the continous mapping theorem yields $\hat{\sigma}_c^2 \xrightarrow{p} \sqrt{\kappa/3}\sigma_c^2$ and $\hat{\sigma}_z^2 \xrightarrow{p} \sigma_z^2 + \left(1 - \sqrt{\kappa/3}\right)\sigma_z^2$ as $p \to \infty$ with $n$ fixed.

Now we consider the case where $n, p \to \infty$. In the proof of Proposition 2.2.2, we show that $\bar{r}^{(2)} \xrightarrow{p} \sigma_c^2 + \sigma_z^2$ and $\bar{r}^{(4)} \xrightarrow{p} \kappa\sigma_c^4 + 3\left(\sigma_c^2 + \sigma_z^2\right)^2$ as $n, p \to \infty$. Applying the continous mapping theorem yields $\hat{\sigma}_c^2 \xrightarrow{p} \sqrt{\kappa/3}\sigma_c^2$ and $\hat{\sigma}_z^2 \xrightarrow{p} \sigma_z^2 + \left(1 - \sqrt{\kappa/3}\right)\sigma_z^2$ as $n, p \to \infty$.

### *A.12 Negative Variance Parameter Estimates*

| % $\hat{\sigma}_z^2 \leq 0$ | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $\tau_c^2$ \ $\pi_c$ | 0 | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 | 1 |
| 0.5 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 1 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 1.5 | 0.00 | 0.00 | 0.20 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| % $\hat{\sigma}_c^2 \leq 0$ | | | | | | | | | | | |
| $\tau_c^2$ \ $\pi_c$ | 0 | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 | 1 |
| 0.5 | 56.40 | 47.40 | 40.20 | 33.60 | 35.40 | 37.00 | 36.40 | 45.00 | 44.60 | 48.60 | 52.20 |
| 1 | 56.60 | 29.00 | 20.80 | 12.80 | 13.00 | 13.40 | 15.60 | 28.20 | 29.80 | 40.80 | 55.80 |
| 1.5 | 52.60 | 12.00 | 4.20 | 2.80 | 3.60 | 4.80 | 6.60 | 10.40 | 22.20 | 35.20 | 55.60 |

Table A.1: Rates of negative estimates of $\sigma_c^2$ and $\sigma_z^2$ across simulations.

Table A.1 shows the rates of negative estimates of $\sigma_c^2$ and $\sigma_z^2$ across simulations using Bernoulli-normal spike-and-slab distributed elements of $\boldsymbol{C}$, setting $n = p = 25$, $\mu = 0$, $\boldsymbol{a} = \boldsymbol{0}$, $\boldsymbol{b} = \boldsymbol{0}$ and $\sigma_z^2 = 1$ and varying $\tau_c^2 = \{1/2, 1, 3/2\}$ and $\pi_c = \{0, 0.1, \ldots, 0.9, 1\}$. For each pair

of values of $\tau_c^2$ and $\pi_c$ 500 values of $\boldsymbol{Y}$ are simulated.

### *A.13 Block Coordinate Descent for LANOVA Penalization for Matrices with Lower-Order Mean Parameters Penalized*

We can use block coordinate descent to obtain estimates of the unknown mean parameters by iterating the following procedure until the objective function

$$\frac{1}{2\hat{\sigma}_z^2}\left|\left|\text{vec}\left(\boldsymbol{Y}-\boldsymbol{M}\right)\right|\right|_2^2+\hat{\lambda}_a\left|\left|\boldsymbol{a}\right|\right|_1+\hat{\lambda}_b\left|\left|\boldsymbol{b}\right|\right|_1+\hat{\lambda}_c\left|\left|\text{vec}\left(\boldsymbol{C}\right)\right|\right|_1$$

converges, starting with $\hat{\boldsymbol{a}}^{(0)}=\boldsymbol{H}_n\boldsymbol{Y}\boldsymbol{1}_p/p$, $\hat{\boldsymbol{b}}^{(0)}=\boldsymbol{H}_p\boldsymbol{Y}'\boldsymbol{1}_n/n$ and $\hat{\boldsymbol{C}}^{(0)}=\boldsymbol{H}_n\boldsymbol{Y}\boldsymbol{H}_p$ and $k=1$:

- Set $\hat{\mu}^{(k)}=\boldsymbol{1}_n'\left(\boldsymbol{Y}-\hat{\boldsymbol{a}}^{(k-1)}\boldsymbol{1}_p'-\boldsymbol{1}_n\left(\hat{\boldsymbol{b}}^{(k-1)}\right)'-\hat{\boldsymbol{C}}^{(k-1)}\right)\boldsymbol{1}_p/\left(np\right)$;
- Set $\boldsymbol{R}_a^{(k)}=\boldsymbol{Y}-\hat{\mu}^{(k)}\boldsymbol{1}_n\boldsymbol{1}_p'-\boldsymbol{1}_n\left(\hat{\boldsymbol{b}}^{(k-1)}\right)'-\hat{\boldsymbol{C}}^{(k-1)}$ and set $\hat{\boldsymbol{a}}^{(k)}=\text{sign}\left(\boldsymbol{H}_n\boldsymbol{R}_a^{(k)}\boldsymbol{1}_p/p\right)\left(\left|\boldsymbol{H}_n\boldsymbol{R}_a^{(k)}\boldsymbol{1}_p/p\right|-\hat{\lambda}_a\hat{\sigma}_z^2/p\right)_+$;
- Set $\boldsymbol{R}_b^{(k)}=\boldsymbol{Y}-\hat{\mu}^{(k)}\boldsymbol{1}_n\boldsymbol{1}_p'-\hat{\boldsymbol{a}}^{(k)}\boldsymbol{1}_p'-\hat{\boldsymbol{C}}^{(k-1)}$ and set $\hat{\boldsymbol{b}}^{(k)}=\text{sign}\left(\boldsymbol{H}_p\left(\boldsymbol{R}_b^{(k)}\right)'\boldsymbol{1}_n/n\right)\left(\left|\boldsymbol{H}_n\boldsymbol{R}_a^{(k)}\boldsymbol{1}_p/p\right|-\hat{\lambda}_b\hat{\sigma}_z^2/n\right)_+$;
- Set $\boldsymbol{R}_c^{(k)}=\boldsymbol{Y}-\hat{\mu}^{(k)}\boldsymbol{1}_n\boldsymbol{1}_p'-\hat{\boldsymbol{a}}^{(k)}\boldsymbol{1}_p'-\boldsymbol{1}_n\hat{\boldsymbol{b}}^{(k)}$ and set $\hat{\boldsymbol{C}}^{(k)}=\text{sign}\left(\boldsymbol{R}_c^{(k)}\right)\left(\left|\boldsymbol{R}_c^{(k)}\right|-\hat{\lambda}_c\hat{\sigma}_z^2\right)_+$.

### *A.14 Proof of Proposition 2.4.1*

We prove Proposition 2.4.1 by computing $\mathbb{E}\left[\left(\bar{r}^{(2)}\right)^2\right]$ and $\mathbb{E}\left[\bar{r}^{(4)}\right]$. Starting with $\mathbb{E}\left[\left(\bar{r}^{(2)}\right)^2\right]$, we again use the result of Konig et al. (1992) given by Equation (A.2), which yields:

$$\begin{aligned}\mathbb{E}\left[\left(\bar{r}^{(2)}\right)^2\right]=&\mathbb{E}\left[\left(\text{vec}\left(\boldsymbol{C}+\boldsymbol{Z}\right)'\boldsymbol{H}^{(K)}\text{vec}\left(\boldsymbol{C}+\boldsymbol{Z}\right)\right)^2\right] \\ =&\alpha_K\left(\kappa\sigma_c^4+3\left(\sigma_c^2+\sigma_z^2\right)^2\right)+\left(\theta_K+2\beta_K\right)\left(\sigma_c^2+\sigma_z^2\right)^2,\end{aligned} \tag{A.10}$$

where $\alpha_K = \sum_{j=1}^{p}\left(h_{jj}^{(K)}\right)^2$, $\beta_K = \sum_{j=1}^{p}\sum_{l=1,l\neq j}^{p} h_{jj}^{(K)}h_{ll}^{(K)}$ and $\theta_K = \sum_{j=1}^{p}\sum_{l=1,l\neq j}^{p}\left(h_{jl}^{(K)}\right)^2$ and $\kappa$ is the excess kurtosis for the distribution of the elements of $\boldsymbol{C}$, which in this case where elements of $\boldsymbol{C}$ are assumed to be Laplace distributed is $\kappa = 3$.

From the definition of $\boldsymbol{H}^{(K)}$ it is straightforward to see that all of the diagonal elements of $\boldsymbol{H}^{(K)}$ are identical and equal to $\prod_{k=1}^{K}\frac{p_k-1}{p_k}$. Therefore, it is straightforward to compute $\alpha_K$ and $\theta_K$.

$$\begin{aligned}
\alpha_K =& \prod_{k=1}^{K} p_K\left(\frac{p_K-1}{p_K}\right)^2 \\
=& \prod_{k=1}^{K}\frac{(p_k-1)^2}{p_k} \\
\theta_K =& \prod_{k=1}^{K} p_k\left(\prod_{k=1}^{K} p_k - 1\right)\prod_{k=1}^{K}\left(\frac{p_k-1}{p_k}\right)^2 \\
=& \left(\prod_{k=1}^{K} p_k - 1\right)\prod_{k=1}^{K}\frac{(p_k-1)^2}{p_k} \\
=& \prod_{k=1}^{K}(p_k-1)^2 - \prod_{k=1}^{K}\frac{(p_k-1)^2}{p_k}
\end{aligned}$$

Computing $\beta_K$ is a little bit trickier, however we can rewrite $\boldsymbol{H}^{(K)} = \boldsymbol{H}_K \otimes \boldsymbol{H}^{(K-1)}$ and obtain a formula for $\beta_K$ as a function of $\alpha_{K-1}$ and $\beta_{K-1}$:

$$\begin{aligned}
\beta_K =& p_K(p_K-1)\left(\frac{1}{p_K^2}\right)(\alpha_{K-1}+\beta_{K-1}) + p_K\left(\frac{p_K-1}{p_K}\right)^2\beta_{K-1} \\
=& \left(\frac{p_K-1}{p_K}\right)(\alpha_{K-1}+\beta_{K-1}) + \frac{(p_K-1)^2}{p_K}\beta_{K-1} \\
=& \left(\frac{p_K-1}{p_K}\right)\alpha_{K-1} + (p_K-1)\beta_{K-1}
\end{aligned}$$

To eliminate $\beta_{K-1}$ from this expression, we note that $\alpha_K + \beta_K = \operatorname{tr}\left(\boldsymbol{H}^{(K)}\boldsymbol{H}^{(K)}\right) =$

$\text{tr}\left(\boldsymbol{H}^{(K)}\right) = \left(\prod_{k=1}^{K} p_k\right)\left(\prod_{k=1}^{K} \frac{p_k-1}{p_k}\right) = \prod_{k=1}^{K}(p_k-1)$. Adding and subtracting $(p_K-1)\,\alpha_{K-1}$ yields:

$$\begin{aligned}
\beta_K &= \left(\frac{p_K-1}{p_K} - (p_K-1)\right)\alpha_{K-1} + (p_K-1)\left(\alpha_{K-1}+\beta_{K-1}\right) \\
&= \left(\frac{p_K-1}{p_K} - (p_K-1)\right)\alpha_{K-1} + (p_K-1)\prod_{k=1}^{K-1}(p_k-1) \\
&= \left(\frac{2p_K-1-p_K^2}{p_K}\right)\prod_{k=1}^{K-1}\frac{(p_k-1)^2}{p_k} + \prod_{k=1}^{K}(p_k-1) \\
&= -\underbrace{\prod_{k=1}^{K}\frac{(p_k-1)^2}{p_k}}_{\alpha_K} + \prod_{k=1}^{K}(p_k-1)\,.
\end{aligned}$$

Plugging these terms in to Equation (A.10) yields:

$$\begin{aligned}
\mathbb{E}\left[\left(\bar{r}^{(2)}\right)^2\right] =& \prod_{k=1}^{K}\frac{(p_k-1)^2}{p_k}\left(\kappa\sigma_c^4 + 3\left(\sigma_c^2+\sigma_z^2\right)^2\right) + \\
& \left(\prod_{k=1}^{K}(p_k-1)^2 - \prod_{k=1}^{K}\frac{(p_k-1)^2}{p_k} + \right. \\
& \left. 2\left(-\prod_{k=1}^{K}\frac{(p_k-1)^2}{p_k} + \prod_{k=1}^{K}(p_k-1)\right)\right)\left(\sigma_c^2+\sigma_z^2\right)^2 \\
=& \prod_{k=1}^{K}\frac{(p_k-1)^2}{p_k}\kappa\sigma_c^4 + \left(\prod_{k=1}^{K}(p_k-1)^2 + \prod_{k=1}^{K}2(p_k-1)\right)\left(\sigma_c^2+\sigma_z^2\right)^2.
\end{aligned} \tag{A.11}$$

Now we need to compute $\mathbb{E}\left[\bar{r}^{(4)}\right]$. We start by rewriting the expression,

$$\begin{aligned}
\mathbb{E}\left[\bar{r}^{(4)}\right] &= \frac{1}{\prod_{k=1}^{K} p_k}\sum_{k=1}^{K}\sum_{i_k=1}^{p_k}\left(\text{vec}\left(\boldsymbol{R}\right)\boldsymbol{Q}^{(i_1,\ldots,i_K)}\text{vec}\left(\boldsymbol{R}\right)\right)^2 \\
&= \frac{1}{\prod_{k=1}^{K} p_k}\sum_{i_1=1}^{p_1}\cdots\sum_{i_K=1}^{p_K}\left(\text{vec}\left(\boldsymbol{C}+\boldsymbol{Z}\right)'\boldsymbol{H}^{(K)}\boldsymbol{Q}^{(i_1,\ldots,i_K)}\boldsymbol{H}^{(K)}\text{vec}\left(\boldsymbol{C}+\boldsymbol{Z}\right)\right)^2
\end{aligned}$$

where $\boldsymbol{Q}^{(i_1\ldots i_K)} = \left(\boldsymbol{e}_{i_K}\boldsymbol{e}_{i_K}'\right)\otimes\cdots\otimes\left(\boldsymbol{e}_{i_1}\boldsymbol{e}_{i_1}'\right)$, where $\boldsymbol{e}_{i_k}$ is a $p_k\times 1$ vector with entry $i_k$ equal to 1 and all other entries equal to zero.

Again, we can use Equation (A.2):

$$\begin{aligned}\mathbb{E}\left[\bar{r}^{(4)}\right] =& \frac{1}{\prod_{k=1}^{K} p_k} \sum_{i_1}^{p_1} \cdots \sum_{i_K=1}^{p_K} \left( \sum_{j=1}^{\prod_{k=1}^{K} p_k} \left( \boldsymbol{H}^{(K)} \boldsymbol{Q}^{(i_1,\ldots,i_K)} \boldsymbol{H}^{(K)} \right)_{jj}^2 \left( \kappa \sigma_c^4 + 3 \left( \sigma_c^2 + \sigma_z^2 \right)^2 \right) + \right. \\ & \sum_{j=1}^{\prod_{k=1}^{K} p_k} \sum_{l=1, l \neq j}^{\prod_{k=1}^{K} p_k} \left( \left( \boldsymbol{H}^{(K)} \boldsymbol{Q}^{(i_1,\ldots,i_K)} \boldsymbol{H}^{(K)} \right)_{jj} \left( \boldsymbol{H}^{(K)} \boldsymbol{Q}^{(i_1,\ldots,i_K)} \boldsymbol{H}^{(K)} \right)_{ll} + \right. \\ & \left. \left. 2 \left( \boldsymbol{H}^{(K)} \boldsymbol{Q}^{(i_1,\ldots,i_K)} \boldsymbol{H}^{(K)} \right)_{jl}^2 \right) \left( \sigma_c^2 + \sigma_z^2 \right)^2 \right).\end{aligned}$$

First, to simplify the problem, we show that

$$\sum_{j=1}^{\prod_{k=1}^{K} p_k} \sum_{l=1, l \neq j}^{\prod_{k=1}^{K} p_k} \left( \boldsymbol{H}^{(K)} \boldsymbol{Q}^{(i_1,\ldots,i_K)} \boldsymbol{H}^{(K)} \right)_{jj} \left( \boldsymbol{H}^{(K)} \boldsymbol{Q}^{(i_1,\ldots,i_K)} \boldsymbol{H}^{(K)} \right)_{ll} = \sum_{j=1}^{\prod_{k=1}^{K} p_k} \sum_{l=1, l \neq j}^{\prod_{k=1}^{K} p_k} \left( \boldsymbol{H}^{(K)} \boldsymbol{Q}^{(i_1,\ldots,i_K)} \boldsymbol{H}^{(K)} \right)_{jl}^2.$$

First, we compute the following:

$$\begin{aligned}\sum_{j=1}^{\prod_{k=1}^{K} p_k} \left( \boldsymbol{H}^{(K)} \boldsymbol{Q}^{(i_1,\ldots,i_K)} \boldsymbol{H}^{(K)} \right)_{jj}^2 + \sum_{j=1}^{\prod_{k=1}^{K} p_k} \sum_{l=1, l \neq j}^{\prod_{k=1}^{K} p_k} \left( \boldsymbol{H}^{(K)} \boldsymbol{Q}^{(i_1,\ldots,i_K)} \boldsymbol{H}^{(K)} \right)_{jj} \left( \boldsymbol{H}^{(K)} \boldsymbol{Q}^{(i_1,\ldots,i_K)} \boldsymbol{H}^{(K)} \right)_{ll} = \\ \operatorname{tr}\left( \left( \boldsymbol{H}^{(K)} \boldsymbol{Q}^{(i_1,\ldots,i_K)} \boldsymbol{H}^{(K)} \right) \otimes \left( \boldsymbol{H}^{(K)} \boldsymbol{Q}^{(i_1,\ldots,i_K)} \boldsymbol{H}^{(K)} \right) \right) = \\ \operatorname{tr}\left( \boldsymbol{H}^{(K)} \boldsymbol{Q}^{(i_1,\ldots,i_K)} \boldsymbol{H}^{(K)} \right)^2 = \\ \operatorname{tr}\left( \boldsymbol{Q}^{(i_1,\ldots,i_K)} \boldsymbol{H}^{(K)} \right)^2 = \\ \prod_{k=1}^{K} \operatorname{tr}\left( \boldsymbol{e}_{ik} \boldsymbol{e}_{ik}' \left( \boldsymbol{I}_{p_k} - \boldsymbol{1}_{p_k} \boldsymbol{1}_{p_k}' / p_k \right) \right)^2 = \\ \prod_{k=1}^{K} \operatorname{tr}\left( \left( \boldsymbol{e}_{ik}' - \boldsymbol{1}_{p_k}' / p_k \right) \boldsymbol{e}_{ik} \right)^2 = \\ \prod_{k=1}^{K} \left( \frac{p_k - 1}{p_k} \right)^2.\end{aligned}$$

Now we compute:

$$\sum_{j=1}^{\prod_{k=1}^{K} p_k} \left( \boldsymbol{H}^{(K)} \boldsymbol{Q}^{(i_1,\ldots,i_K)} \boldsymbol{H}^{(K)} \right)_{jj}^2 + \sum_{j=1}^{\prod_{k=1}^{K} p_k} \sum_{l=1, l\neq j}^{\prod_{k=1}^{K} p_k} \left( \boldsymbol{H}^{(K)} \boldsymbol{Q}^{(i_1,\ldots,i_K)} \boldsymbol{H}^{(K)} \right)_{jl}^2 =$$

$$\operatorname{tr}\left( \boldsymbol{H}^{(K)} \boldsymbol{Q}^{(i_1,\ldots,i_K)} \boldsymbol{H}^{(K)} \boldsymbol{H}^{(K)} \boldsymbol{Q}^{(i_1,\ldots,i_K)} \boldsymbol{H}^{(K)} \right) =$$

$$\operatorname{tr}\left( \boldsymbol{H}^{(K)} \boldsymbol{Q}^{(i_1,\ldots,i_K)} \boldsymbol{H}^{(K)} \boldsymbol{Q}^{(i_1,\ldots,i_K)} \right) =$$

$$\prod_{k=1}^{K} \operatorname{tr}\left( \left( \boldsymbol{I}_{p_k} - \mathbf{1}_{p_k} \mathbf{1}_{p_k}'/p_k \right) \boldsymbol{e}_{ik} \boldsymbol{e}_{ik}' \left( \boldsymbol{I}_{p_k} - \mathbf{1}_{p_k} \mathbf{1}_{p_k}'/p_k \right) \boldsymbol{e}_{ik} \boldsymbol{e}_{ik}' \right) =$$

$$\prod_{k=1}^{K} \operatorname{tr}\left( \boldsymbol{e}_{ik}' \left( \boldsymbol{e}_{ik} - \mathbf{1}_{p_k}/p_k \right) \left( \boldsymbol{e}_{ik}' - \mathbf{1}_{p_k}'/p_k \right) \boldsymbol{e}_{ik} \right) =$$

$$\prod_{k=1}^{K} \left( \frac{p_k - 1}{p_k} \right)^2.$$

It follows that

$$\sum_{j=1}^{\prod_{k=1}^{K} p_k} \sum_{l=1, l\neq j}^{\prod_{k=1}^{K} p_k} \left( \boldsymbol{H}^{(K)} \boldsymbol{Q}^{(i_1,\ldots,i_K)} \boldsymbol{H}^{(K)} \right)_{jj} \left( \boldsymbol{H}^{(K)} \boldsymbol{Q}^{(i_1,\ldots,i_K)} \boldsymbol{H}^{(K)} \right)_{ll} = \sum_{j=1}^{\prod_{k=1}^{K} p_k} \sum_{l=1, l\neq j}^{\prod_{k=1}^{K} p_k} \left( \boldsymbol{H}^{(K)} \boldsymbol{Q}^{(i_1,\ldots,i_K)} \boldsymbol{H}^{(K)} \right)_{jl}^2.$$

Now, to determine if $\sum_{j=1}^{\prod_{k=1}^{K} p_k} \left( \boldsymbol{H}^{(K)} \boldsymbol{Q}^{(i_1,\ldots,i_K)} \boldsymbol{H}^{(K)} \right)_{jj}^2$ and $\sum_{j=1}^{\prod_{k=1}^{K} p_k} \sum_{l=1, l\neq j}^{\prod_{k=1}^{K} p_k} \left( \boldsymbol{H}^{(K)} \boldsymbol{Q}^{(i_1,\ldots,i_K)} \boldsymbol{H}^{(K)} \right)_{jl}^2$ depend on the values of $i_1, \ldots, i_K$, we examine the structure of each $\boldsymbol{H}^{(K)} \boldsymbol{Q}^{(i_1,\ldots,i_K)} \boldsymbol{H}^{(K)}$.

$$\begin{aligned} \boldsymbol{H}^{(K)} \boldsymbol{Q}^{(i_1,\ldots,i_K)} \boldsymbol{H}^{(K)} &= \left( \boldsymbol{H}_K \otimes \cdots \otimes \boldsymbol{H}_1 \right) \left( \left( \boldsymbol{e}_{i_K} \boldsymbol{e}_{i_K}' \right) \otimes \cdots \otimes \left( \boldsymbol{e}_{i_1} \boldsymbol{e}_{i_1}' \right) \right) \left( \boldsymbol{H}_K \otimes \cdots \otimes \boldsymbol{H}_1 \right) \\ &= \left( \boldsymbol{H}_K \boldsymbol{e}_{i_K} \boldsymbol{e}_{i_K}' \boldsymbol{H}_K \right) \otimes \cdots \otimes \left( \boldsymbol{H}_1 \boldsymbol{e}_{i_1} \boldsymbol{e}_{i_1}' \boldsymbol{H}_1 \right). \end{aligned}$$

Each matrix $\boldsymbol{H}_k \boldsymbol{e}_{i_k} \boldsymbol{e}_{i_k}' \boldsymbol{H}_k$ is formed by multiplying the $i_k$-th elements of each row of $\boldsymbol{H}_k$ by the $i_k$-th elements of each column of $\boldsymbol{H}_k$, yielding a matrix with:

- one diagonal element equal to $\left( \frac{p_k - 1}{p_k} \right)^2$;
- $p_k - 1$ diagonal elements equal to $\frac{1}{p_k^2}$;

- $2\left(p_k-1\right)$ off-diagonal elements equal to $-\left(\frac{p_k-1}{p_k^2}\right)$;
- $p_k\left(p_k-1\right)-2\left(p_k-1\right)$ elements equal to $\frac{1}{p_k^2}$.

Accordingly, the specific $i_k$ used to construct the matrix $\boldsymbol{H}_k\boldsymbol{e}_{i_k}\boldsymbol{e}_{i_k}'\boldsymbol{H}_k$ determines the relative locations of the diagonal elements and the relative locations of the off-diagonal elements, but not the composition and values of the off-diagonal elements and diagonal elements. As a result, the sum of squared diagonal elements, the sum of pairwise product of different diagonal elements and the sum of squared off-diagonal elements for each submatrix will will *not* depend on the specific $\boldsymbol{e}_{i_k}$ used to construct the matrix.

Likewise, because the diagonal elements of $\boldsymbol{H}^{(K)}\boldsymbol{Q}^{(i_1,\ldots,i_K)}\boldsymbol{H}^{(K)}$ are given by the products of diagonal elements of the components, $\boldsymbol{H}_k\boldsymbol{e}_{i_k}\boldsymbol{e}_{i_k}'\boldsymbol{H}_k$, and the off-diagonal elements of $\boldsymbol{H}^{(K)}\boldsymbol{Q}^{(i_1,\ldots,i_K)}\boldsymbol{H}^{(K)}$ are given by the products of diagonal and off-diagonal or off-diagonal alone elements of the components, $\boldsymbol{H}_k\boldsymbol{e}_{i_k}\boldsymbol{e}_{i_k}'\boldsymbol{H}_k$, the sum of squared diagonal elements and the sum of squared off diagonal elements of $\boldsymbol{H}^{(K)}\boldsymbol{Q}^{(i_1,\ldots,i_K)}\boldsymbol{H}^{(K)}$ will not depend on the values of $i_1,\ldots,i_K$.

Defining $\boldsymbol{S}^{(K)} = \boldsymbol{H}^{(K)}\boldsymbol{Q}^{(i_1,\ldots,i_K)}\boldsymbol{H}^{(K)}$ and setting $\alpha_K = \sum_{j=1}^{p}\left(s_{jj}^{(K)}\right)^2$, $\theta_K = \sum_{j=1}^{p}\sum_{l=1,l\neq j}^{p} s_{jj}^{(K)}s_{ll}$ and $\beta_K = \sum_{j=1}^{p}\sum_{l=1,k\neq j}^{p}\left(s_{jl}^{(K)}\right)^2$, applying Equation (4.1) from Konig et al. (1992) yields: For reasons discussed below, we drop the $(i_1,\ldots,i_K)$ superscript on $\boldsymbol{S}$ because the sums of squared entries of $\boldsymbol{S}$ and sums of products of diagonal entries of $\boldsymbol{S}$ that we are interested in computing do not depend on them and it simplifies notation. We explain why this is the case later on.

Now we simplify the notation. Let $\boldsymbol{S}^{(K)} = \boldsymbol{H}^{(K)}\boldsymbol{Q}^{(i_1,\ldots,i_K)}\boldsymbol{H}^{(K)}$, dropping the $i_1,\ldots,i_K$ indices because the functions of $\boldsymbol{S}^{(K)}$ we are interested in do not depend on them. Define $\alpha_K = \sum_{j=1}^{\prod_{k=1}^{K}p_k}\left(s_{ll}^{(K)}\right)^2$ and $\beta_K = \sum_{j=1}^{\prod_{k=1}^{K}p_k}\sum_{l=1,l\neq j}^{\prod_{k=1}^{K}p_k}\left(s_{jl}^{(K)}\right)^2$. Note that we can simplify our expression for $\mathbb{E}\left[\bar{r}^{(4)}\right]$ from earlier as follows:

$$\mathbb{E}\left[\bar{r}^{(4)}\right] = \alpha_K\left(\kappa\sigma_c^4 + 3\left(\sigma_c^2+\sigma_z^2\right)^2\right) + 3\beta_K\left(\sigma_c^2+\sigma_z^2\right)^2. \tag{A.12}$$

To help us compute $\alpha_K$ and $\beta_K$, we can expand and rewrite $\boldsymbol{S}^{(K)}$ as follows to obtain a recursive formula for $\boldsymbol{S}^{(K)}$ given $\boldsymbol{S}^{(K-1)}$, which will :

$$\begin{aligned}
\boldsymbol{S}^{(K)} &= (\boldsymbol{H}_K \otimes \cdots \otimes \boldsymbol{H}_1) \left( \left( \boldsymbol{e}_{i_K} \boldsymbol{e}'_{i_K} \right) \otimes \cdots \otimes \left( \boldsymbol{e}_{i_1} \boldsymbol{e}'_{i_1} \right) \right) (\boldsymbol{H}_K \otimes \cdots \otimes \boldsymbol{H}_1) \\
&= \left( \boldsymbol{H}_K \boldsymbol{e}_{i_1} \boldsymbol{e}'_{i_1} \boldsymbol{H}_K \right) \otimes \left( (\boldsymbol{H}_{K-1} \otimes \ldots \boldsymbol{H}_1) \left( \left( \boldsymbol{e}_{i_{K-1}} \boldsymbol{e}'_{i_{K-1}} \right) \otimes \cdots \otimes \left( \boldsymbol{e}_{i_1} \boldsymbol{e}'_{i_1} \right) \right) (\boldsymbol{H}_{K-1} \otimes \ldots \boldsymbol{H}_1) \right) \\
&= \left( \boldsymbol{H}_K \boldsymbol{e}_{i_1} \boldsymbol{e}'_{i_1} \boldsymbol{H}_K \right) \otimes \boldsymbol{S}^{(K-1)}.
\end{aligned}$$

We can compute $\alpha_K$ recursively from this formula; the sum of sqaured diagonal elements of $\boldsymbol{S}^{(K)}$ will equal the sum of each diagonal element of $\boldsymbol{H}_K \boldsymbol{e}_{i_1} \boldsymbol{e}'_{i_1} \boldsymbol{H}_K$ squared multiplied by the sum of squared diagonal elements of $\boldsymbol{S}^{(K-1)}$, $\alpha_{K-1}$:

$$\begin{aligned}
\alpha_K &= \left( \frac{(p_K - 1)^4}{p_K^4} + \frac{(p_K - 1)}{p_K^4} \right) \alpha_{K-1} \\
&= \frac{(p_K - 1)}{p_K^4} \left( p_K^3 - 3p_K^2 + 3p_K \right) \alpha_{K-1} \\
&= \frac{(p_K - 1)}{p_K^3} \left( p_K^2 - 3p_K + 3 \right) \alpha_{K-1} \\
&= \prod_{k=1}^{K} \frac{(p_k - 1)\left( p_k^2 - 3p_k + 3 \right)}{p_k^3}
\end{aligned}$$

Now we can compute $\beta_K$ in a similar way. The sum of squared off-diagonal elements of $\boldsymbol{S}^{(K)}$ will be given by the sum of each diagonal element of $\boldsymbol{H}_K \boldsymbol{e}_{i_1} \boldsymbol{e}'_{i_1} \boldsymbol{H}_K$ squared multiplied by the sum of squared off diagonal elements of $\boldsymbol{S}^{(K-1)}$, $\beta_{K-1}$, plus the sum of each off-diagonal element of $\boldsymbol{H}_K \boldsymbol{e}_{i_1} \boldsymbol{e}'_{i_1} \boldsymbol{H}_K$ squared multiplied by the sum of squared elements of $\boldsymbol{S}^{(K-1)}$, $\alpha_{K-1} + \beta_{K-1}$.

$$
\begin{aligned}
\beta_K = &\left(\left(\frac{p_K-1}{p_K}\right)^4 + \frac{(p_K-1)}{p_K^4}\right)\beta_{K-1} + \\
&\left(\frac{2\left(p_K-1\right)^3}{p_K^4} + \frac{p_K\left(p_K-1\right)-2\left(p_K-1\right)}{p_K^4}\right)\left(\alpha_{K-1}+\beta_{K-1}\right) \\
=&\left(\frac{p_K-1}{p_K^4}\right)\left(\left(p_K-1\right)^3+1+2\left(p_K-1\right)^2+p_K-2\right)\beta_{K-1} + \\
&\left(\frac{p_K-1}{p_K^4}\right)\left(2\left(p_K-1\right)^2+p_K-2\right)\alpha_{K-1} \\
=&\left(\frac{p_K-1}{p_K^4}\right)\left(p_K^3-3p_K^2+3p_K+1+2p_K^2-4p_K+2+p_K-2\right)\beta_{K-1} + \\
&\left(\frac{p_K-1}{p_K^4}\right)\left(2p_K^2-4p_K+2+p_K-2\right)\alpha_{K-1} \\
=&\left(\frac{p_K-1}{p_K^3}\right)\left(\left(p_K^2-p_K\right)\left(\alpha_{K-1}+\beta_{K-1}\right)+\left(-p_K^2+3p_K-3\right)\alpha_{K-1}\right)
\end{aligned}
$$

Noting that $\alpha_{K-1}+\beta_{K-1} = \operatorname{tr}\left(\boldsymbol{S}^{(K-1)}\boldsymbol{S}^{(K-1)}\right)$, which we computed earlier as equal to $\prod_{k=1}^{K-1}\left(\frac{p_k-1}{p_k}\right)^2$, we can eliminate $\beta_{K-1}$ from this expression, yielding:

$$
\begin{aligned}
\beta_K =& \left(\frac{p_K-1}{p_K}\right)^2\prod_{k=1}^{K-1}\left(\frac{p_k-1}{p_k}\right)^2 - \alpha_K \\
=&\prod_{k=1}^{K}\left(\frac{p_k-1}{p_k}\right)^2 - \alpha_K \\
=&\prod_{k=1}^{K}\left(\frac{p_k-1}{p_k}\right)^2 - \prod_{k=1}^{K}\frac{\left(p_k-1\right)\left(p_k^2-3p_k+3\right)}{p_k^3}.
\end{aligned}
$$

Substituting these quantities into Equation (A.12) yields:

$$
\mathbb{E}\left[\bar{r}^{(4)}\right] = \prod_{k=1}^{K}\frac{\left(p_k-1\right)\left(p_k^2-3p_k+3\right)}{p_k^3}\left(\kappa\sigma_c^4\right) + \prod_{k=1}^{K}\left(\frac{p_k-1}{p_k}\right)^2 3\left(\sigma_c^2+\sigma_z^2\right)^2. \tag{A.13}
$$

Proposition 2.4.1 follows directly from Equations (A.11) and (A.13) and plugging in $\kappa = 3$.

### *A.15 Proof of Proposition 2.4.2*

First, we consider the scenario where a single $p_{k'} \to \infty$, with all other $p_k$, $k \neq k'$ held fixed. Without loss of generality, we consider $p_K \to \infty$ and assume $p_k$, $k < K$, fixed. First, we show consistency of $\bar{r}^{(2)}$, which involves first computing $\mathbb{E}\left[\bar{r}^{(2)}\right]$ and then showing that $\mathbb{V}\left[\bar{r}^{(2)}\right] = O\left(\frac{1}{p_K}\right)$.

We compute $\mathbb{E}\left[\bar{r}^{(2)}\right]$ as follows, where $\boldsymbol{H}_k = \boldsymbol{I}_{p_k} - \mathbf{1}_{p_k}\mathbf{1}_{p_k}'/p_k$ and $\boldsymbol{H}^{(K)} = \boldsymbol{H}_K \otimes \cdots \otimes \boldsymbol{H}_1$:

$$\begin{aligned}
\mathbb{E}\left[\bar{r}^{(2)}\right] =& \frac{1}{\prod_{k=1}^{K} p_k}\mathbb{E}\left[\mathrm{vec}\left(\boldsymbol{Y}\right)'\boldsymbol{H}^{(K)}\mathrm{vec}\left(\boldsymbol{Y}\right)\right] \\
=& \frac{1}{\prod_{k=1}^{K} p_k}\mathbb{E}\left[\mathrm{vec}\left(\boldsymbol{C}+\boldsymbol{Z}\right)'\boldsymbol{H}^{(K)}\mathrm{vec}\left(\boldsymbol{C}+\boldsymbol{Z}\right)\right] \\
=& \frac{1}{\prod_{k=1}^{K} p_k}\mathrm{tr}\left(\boldsymbol{H}^{(K)}\right)\left(\sigma_c^2+\sigma_z^2\right) \\
=& \prod_{k=1}^{K}\left(\frac{p_k-1}{p_k}\right)\left(\sigma_c^2+\sigma_z^2\right).
\end{aligned} \tag{A.14}$$

Now we turn to computing $\mathbb{V}\left[\bar{r}^{(2)}\right]$, which we write as:

$$\begin{aligned}
\mathbb{V}\left[\bar{r}^{(2)}\right] =& \mathbb{E}\left[\left(\bar{r}^{(2)}\right)^2\right] - \mathbb{E}\left[\bar{r}^{(2)}\right]^2 \\
=& \frac{1}{\prod_{k=1}^{K} p_k^2}\mathbb{E}\left[\left(\mathrm{vec}\left(\boldsymbol{C}+\boldsymbol{Z}\right)'\boldsymbol{H}^{(K)}\mathrm{vec}\left(\boldsymbol{C}+\boldsymbol{Z}\right)\right)^2\right] - \prod_{k=1}^{K}\left(\frac{p_k-1}{p_k}\right)^2\left(\sigma_c^2+\sigma_z^2\right)^2.
\end{aligned} \tag{A.15}$$

Substituting the expression for $\mathbb{E}\left[\left(\bar{r}^{(2)}\right)^2\right]$ given by Equation (A.11) into the expression for the variance of $\bar{r}^{(2)}$ yields:

$$\begin{aligned}
\mathbb{V}\left[\bar{r}^{(2)}\right] =& \prod_{k=1}^{K}\frac{\left(p_k-1\right)^2}{p_k^3}\kappa\sigma_c^4 + \prod_{k=1}^{K}\frac{2\left(p_k-1\right)}{p_k^2}\left(\sigma_c^2+\sigma_z^2\right)^2 \\
=& O\left(\frac{1}{p_K}\right).
\end{aligned}$$

As a result, $\bar{r}^{(2)} \xrightarrow{p} \prod_{k=1}^{K-1}\left(\frac{p_k-1}{p_k}\right)\left(\sigma_c^2+\sigma_z^2\right)$ as $p_K \to \infty$ and $p_k$, $k < K$ fixed. More generally, for fixed $k'$, $\bar{r}^{(2)} \xrightarrow{p} \prod_{k=1,k\neq k'}^{K}\left(\frac{p_k-1}{p_k}\right)\left(\sigma_c^2+\sigma_z^2\right)$ as $p_{k'} \to \infty$ and $p_k$, $k \neq k'$ fixed.

Now we show consistency of $\bar{r}^{(4)}$, which involves using the expression for $\mathbb{E}\left[\bar{r}^{(2)}\right]$ given by Equation (A.13) and showing that $\mathbb{V}\left[\bar{r}^{(4)}\right] = O\left(\frac{1}{p_K}\right)$. First, we drop the $(K)$ superscript

from $\boldsymbol{H}^{(K)}$ and let $p = \prod_{k=1}^{K} p_k$ for simplicity. We decompose $\boldsymbol{H}$: $\boldsymbol{H} = \boldsymbol{F} - \frac{1}{p_K}\boldsymbol{G}$, where $\boldsymbol{F} = \boldsymbol{I}_{p_K} \otimes \boldsymbol{H}_{K-1} \otimes \cdots \otimes \boldsymbol{H}_1$ and $\boldsymbol{G} = \left(\mathbf{1}_{p_K}\mathbf{1}'_{p_K}\right) \otimes \boldsymbol{H}_{K-1} \otimes \cdots \otimes \boldsymbol{H}_1$. Accordingly, we have $\text{vec}\left(\boldsymbol{R}\right) = \left(\boldsymbol{F} - \frac{1}{p_k}\boldsymbol{G}\right)\text{vec}\left(\boldsymbol{Y}\right)$.

Let $\boldsymbol{f}_j$ and $\boldsymbol{g}_j$ refer to the $j$-th rows of $\boldsymbol{F}$ and $\boldsymbol{G}$, accordingly. Then the $j$-th element of $\text{vec}\left(\boldsymbol{R}\right)$, $r_j = \boldsymbol{f}'_j\text{vec}\left(\boldsymbol{Y}\right) - \boldsymbol{g}'_j\text{vec}\left(\boldsymbol{Y}\right)/p_K$. The operation $\boldsymbol{g}'_j\text{vec}\left(\boldsymbol{Y}\right)/p_K$ takes means of elements of $\boldsymbol{Y}$ over the $K$-th mode, so $\boldsymbol{g}'_j\text{vec}\left(\boldsymbol{Y}\right)/p_K = O_p\left(\frac{1}{\sqrt{p_K}}\right)$. Then:

$$\begin{aligned}\mathbb{V}\left[\bar{r}^{(4)}\right] =& \mathbb{V}\left[\frac{1}{p}\sum_{j=1}^{p} r_j^4\right] \\ =& \mathbb{E}\left[\left(\frac{1}{p}\sum_{j=1}^{p}\left(\boldsymbol{f}'_j\text{vec}\left(\boldsymbol{Y}\right) + O_p\left(\frac{1}{\sqrt{p_K}}\right)\right)^4\right)^2\right] - \left(\mathbb{E}\left[\frac{1}{p}\sum_{j=1}^{p} r_j^4\right]\right)^2\end{aligned} \tag{A.16}$$

Expanding the first term, we have:

$$\begin{aligned}&\mathbb{E}\left[\left(\frac{1}{p}\sum_{j=1}^{p}\left(\boldsymbol{f}'_j\text{vec}\left(\boldsymbol{Y}\right) + O_p\left(\frac{1}{\sqrt{p_K}}\right)\right)^4\right)^2\right] = \\ &\frac{1}{p^2}\sum_{j=1}^{p}\mathbb{E}\left[\left(\boldsymbol{f}'_j\text{vec}\left(\boldsymbol{Y}\right) + O_p\left(\frac{1}{\sqrt{p_K}}\right)\right)^8\right] + \\ &\frac{1}{p^2}\sum_{j=1}^{p}\sum_{j'=1,j'\neq j}^{p}\mathbb{E}\left[\left(\boldsymbol{f}'_j\text{vec}\left(\boldsymbol{Y}\right) + O_p\left(\frac{1}{\sqrt{p_K}}\right)\right)^4\left(\boldsymbol{f}'_{j'}\text{vec}\left(\boldsymbol{Y}\right) + O_p\left(\frac{1}{\sqrt{p_K}}\right)\right)^4\right].\end{aligned}$$

Now let's consider the structure of each $\boldsymbol{f}_j$, recalling that $\boldsymbol{F} = \boldsymbol{I}_{p_K} \otimes \boldsymbol{H}_{K-1} \otimes \cdots \otimes \boldsymbol{H}_1$. Each $\boldsymbol{f}_j$ has $p$ entries, but only $\prod_{k=1}^{K-1} p_k$ entries are nonzero. If $\text{vec}\left(\boldsymbol{Y}\right)_j$ belongs to the $i_j$-th level of the $K$-th mode of $\boldsymbol{Y}$, the nonzero entries correspond to elements of $\text{vec}\left(\boldsymbol{Y}\right)$ that also belong to the $i_j$-th level of the $K$-th mode of $\boldsymbol{Y}$. This means that each product, $\boldsymbol{f}'_j\text{vec}\left(\boldsymbol{Y}\right)$, includes only the $\prod_{k=1}^{K-1} p_k$ entries of $\text{vec}\left(\boldsymbol{Y}\right)$ that belong to the $i_j$-th level of the $K$-th mode of $\boldsymbol{Y}$.

Because the number of elements of $\boldsymbol{Y}$ in each $\boldsymbol{f}'_j\text{vec}\left(\boldsymbol{Y}\right)$, the entries of any $\boldsymbol{f}_j$, and the eighth moment of each element of $\boldsymbol{Y}$ are finite and do not depend at all on $p_K$ and are fixed when $p_1, \ldots, p_{K-1}$ are fixed, each $\mathbb{E}\left[\left(\boldsymbol{f}'_j\text{vec}\left(\boldsymbol{Y}\right)\right)^a\right]$ for $a = 1, \ldots, 8$ are finite and fixed for

fixed $p_1, \ldots, p_{K-1}$. As a result, $\mathbb{E}\left[\left(\boldsymbol{f}_j'\text{vec}\left(\boldsymbol{Y}\right)\right)^a\right] O_p\left(\frac{1}{\sqrt{p_K}}\right) = O_p\left(\frac{1}{\sqrt{p_K}}\right)$. Additionally, by exchangeability of the elements of $\boldsymbol{Y}$ and structure of the matrix $\boldsymbol{F}$ (each row of $\boldsymbol{F}$ contains the same elements in a different order), $\mathbb{E}\left[\left(\boldsymbol{f}_j'\text{vec}\left(\boldsymbol{Y}\right)\right)^a\right] = \mathbb{E}\left[\left(\boldsymbol{f}_1'\text{vec}\left(\boldsymbol{Y}\right)\right)^a\right]$ for all $j$ and $\frac{1}{p^2}\sum_{j=1}^{p}\mathbb{E}\left[\left(\boldsymbol{f}_j'\text{vec}\left(\boldsymbol{Y}\right)\right)^8\right] = \frac{1}{p}\mathbb{E}\left[\left(\boldsymbol{f}_1'\text{vec}\left(\boldsymbol{Y}\right)\right)^8\right] = O\left(\frac{1}{p_1}\right)$. Then we can simplify the first term as follows:

$$\mathbb{E}\left[\left(\frac{1}{p}\sum_{j=1}^{p}\left(\boldsymbol{f}_j'\text{vec}\left(\boldsymbol{Y}\right) + O_p\left(\frac{1}{\sqrt{p_K}}\right)\right)^4\right)^2\right] =$$
$$\frac{1}{p^2}\sum_{j=1}^{p}\sum_{j'=1,j'\neq j}^{p}\mathbb{E}\left[\left(\boldsymbol{f}_j'\text{vec}\left(\boldsymbol{Y}\right)\right)^4\left(\boldsymbol{f}_{j'}'\text{vec}\left(\boldsymbol{Y}\right)\right)^4\right] + O_p\left(\frac{1}{\sqrt{p_K}}\right).$$

Now, we consider how the $\boldsymbol{f}_j'\text{vec}\left(\boldsymbol{Y}\right)$ and $\boldsymbol{f}_{j'}'\text{vec}\left(\boldsymbol{Y}\right)$ relate when $j \neq j'$, specifically with respect to which elements of $\boldsymbol{Y}$ are contained in each and if there is any overlap. We define $i_j$ as the level of the $K$-th mode that the element $\text{vec}\left(\boldsymbol{Y}\right)_j$ belongs to. For a fixed $j$, let $\mathcal{J}_j$ denote the set of $\prod_{k=1}^{K-1} p_k - 1$ indices that correspond to other entries of $\boldsymbol{Y}$ belonging to the $i_j$-th level of the $K$-th mode of $\boldsymbol{Y}$, besides $\text{vec}\left(\boldsymbol{Y}\right)_j$. Then if $j' \notin \mathcal{J}_j$, $\boldsymbol{f}_{j'}'\text{vec}\left(\boldsymbol{Y}\right)$ does not contain any of the same elements of $\boldsymbol{Y}$ as $\boldsymbol{f}_j'\text{vec}\left(\boldsymbol{Y}\right)$, so we can write:

$$\mathbb{E}\left[\left(\frac{1}{p}\sum_{j=1}^{p}\left(\boldsymbol{f}_j'\text{vec}\left(\boldsymbol{Y}\right) + O_p\left(\frac{1}{\sqrt{p_K}}\right)\right)^4\right)^2\right] =$$
$$\frac{1}{p^2}\sum_{j=1}^{p}\sum_{j'\in\mathcal{J}_j}\mathbb{E}\left[\left(\boldsymbol{f}_j'\text{vec}\left(\boldsymbol{Y}\right)\right)^4\left(\boldsymbol{f}_{j'}'\text{vec}\left(\boldsymbol{Y}\right)\right)^4\right] + \frac{1}{p^2}\sum_{j=1}^{p}\sum_{j'\notin\mathcal{J}_j}\mathbb{E}\left[\left(\boldsymbol{f}_j'\text{vec}\left(\boldsymbol{Y}\right)\right)^4\right]\mathbb{E}\left[\left(\boldsymbol{f}_{j'}'\text{vec}\left(\boldsymbol{Y}\right)\right)^4\right] +$$
$$O_p\left(\frac{1}{\sqrt{p_K}}\right) =$$
$$\frac{1}{p^2}\sum_{j=1}^{p}\sum_{j'\in\mathcal{J}_j}\mathbb{E}\left[\left(\boldsymbol{f}_j'\text{vec}\left(\boldsymbol{Y}\right)\right)^4\left(\boldsymbol{f}_{j'}'\text{vec}\left(\boldsymbol{Y}\right)\right)^4\right] + \frac{(p_K-1)\prod_{k=1}^{K-1}p_k}{p}\mathbb{E}\left[\left(\boldsymbol{f}_1'\text{vec}\left(\boldsymbol{Y}\right)\right)^4\right]^2 + O_p\left(\frac{1}{\sqrt{p_K}}\right)$$
$$\frac{1}{p^2}\sum_{j=1}^{p}\sum_{j'\in\mathcal{J}_j}\mathbb{E}\left[\left(\boldsymbol{f}_j'\text{vec}\left(\boldsymbol{Y}\right)\right)^4\left(\boldsymbol{f}_{j'}'\text{vec}\left(\boldsymbol{Y}\right)\right)^4\right] + \mathbb{E}\left[\left(\boldsymbol{f}_1'\text{vec}\left(\boldsymbol{Y}\right)\right)^4\right]^2 + O_p\left(\frac{1}{\sqrt{p_K}}\right)$$

Note that by the same arguments, we can expand the second term of Equation (A.16)

similarly:

$$\left(\mathbb{E}\left[\frac{1}{p}\sum_{j=1}^{p} r_j^4\right]\right)^2 = \left(\mathbb{E}\left[\boldsymbol{f}_1'\text{vec}\left(\boldsymbol{Y}\right)\right]\right)^2 + O\left(\frac{1}{p_K}\right).$$

Then we have

$$\mathbb{V}\left[\bar{r}^{(4)}\right] = \frac{1}{p^2}\sum_{j=1}^{p}\sum_{j'\in\mathcal{J}_j}\mathbb{E}\left[\left(\boldsymbol{f}_j'\text{vec}\left(\boldsymbol{Y}\right)\right)^4\left(\boldsymbol{f}_{j'}'\text{vec}\left(\boldsymbol{Y}\right)\right)^4\right] + O_p\left(\frac{1}{\sqrt{p_K}}\right).$$

The last step is to address the first term in the above expression. Recall that for fixed $j$, $\left(\boldsymbol{f}_{j'}'\text{vec}\left(\boldsymbol{Y}\right)\right)^4$ for $j' \in \mathcal{J}_j$ contains the same $\prod_{k=1}^{K-1} p_k$ elements of $\boldsymbol{Y}$ as $\left(\boldsymbol{f}_j'\text{vec}\left(\boldsymbol{Y}\right)\right)^4$, with coefficients that do not depend on $p_K$. Therefore, $\left(\boldsymbol{f}_j'\text{vec}\left(\boldsymbol{Y}\right)\right)^4\left(\boldsymbol{f}_{j'}'\text{vec}\left(\boldsymbol{Y}\right)\right)^4$ is a polynomial in $\prod_{k=1}^{K-1} p_k$ elements of $\boldsymbol{Y}$ which does not depend on $p_K$ at all, therefore $\mathbb{E}\left[\left(\boldsymbol{f}_j'\text{vec}\left(\boldsymbol{Y}\right)\right)^4\left(\boldsymbol{f}_{j'}'\text{vec}\left(\boldsymbol{Y}\right)\right)^4\right]$ is fixed and finite. It may not be the case that $\mathbb{E}\left[\left(\boldsymbol{f}_j'\text{vec}\left(\boldsymbol{Y}\right)\right)^4\left(\boldsymbol{f}_{j'}'\text{vec}\left(\boldsymbol{Y}\right)\right)^4\right]$ is equal for any $j$,$j' \in \mathcal{J}_j$, but a maximum, which again does not depend on $p_K$ and is fixed for fixed $p_1, \ldots, p_{K-1}$ exists. Let $C = \text{minimize}_{j,j'\in\mathcal{J}_j}\mathbb{E}\left[\left(\boldsymbol{f}_j'\text{vec}\left(\boldsymbol{Y}\right)\right)^4\left(\boldsymbol{f}_{j'}'\text{vec}\left(\boldsymbol{Y}\right)\right)^4\right]$. Then

$$\begin{aligned}\frac{1}{p^2}\sum_{j=1}^{p}\sum_{j'\in\mathcal{J}_j}\mathbb{E}\left[\left(\boldsymbol{f}_j'\text{vec}\left(\boldsymbol{Y}\right)\right)^4\left(\boldsymbol{f}_{j'}'\text{vec}\left(\boldsymbol{Y}\right)\right)^4\right] &\leq \frac{p\left(\prod_{k=1}^{K-1} p_k - 1\right)}{p^2}C \\ &= \left(\frac{1}{p_K} + \frac{1}{p}\right)C = O\left(\frac{1}{p_K}\right).\end{aligned}$$

Accordingly, $\mathbb{V}\left[\bar{r}^{(4)}\right] = O_p\left(\frac{1}{\sqrt{p_K}}\right)$. Then

$$\bar{r}^{(4)} \xrightarrow{p} \prod_{k=1}^{K-1}\frac{\left(p_k-1\right)\left(p_k^2-3p_k+3\right)}{p_k^3}\left(\kappa\sigma_c^4\right) + \prod_{k=1}^{K-1}\left(\frac{p_k-1}{p_k}\right)^2 3\left(\sigma_c^2+\sigma_z^2\right)^2$$

as $p_K \to \infty$ when $p_k$, $k < K$ fixed.

Combining this result with $\bar{r}^{(2)} \xrightarrow{p} \prod_{k=1}^{K-1}\left(\frac{p_k-1}{p_k}\right)\left(\sigma_c^2+\sigma_z^2\right)$ as $p_K \to \infty$ when $p_k$, $k < K$ fixed, plugging in $\kappa = 3$ and applying the continuous mapping theorem yields $\hat{\sigma}_c^4 \xrightarrow{p} \sigma_c^4$, $\hat{\sigma}_c^2 \xrightarrow{p} \sigma_c^2$, $\hat{\lambda}_c \xrightarrow{p} \lambda_c$ and $\hat{\sigma}_z^2 \xrightarrow{p} \sigma_z^2$.

Now we consider the scenario where $p_1, \ldots, p_K \to \infty$. It follows directly from Equations (A.14) and (A.15) that $\bar{r}^{(2)} \xrightarrow{p} \sigma_c^2 + \sigma_z^2$ as $p_1, \ldots, p_K \to \infty$. Next, we need to examine $\bar{r}^{(4)}$. For convenience, we also let $p_k = \pi_k q$, where $0 < \pi_k < 1$ are fixed and $\sum_{k=1}^{K} \pi_k = 1$. We also let $\pi = \prod_{k=1}^{K} \pi_k$ Then the scenario $p_1, \ldots, p_K \to \infty$ is equivalent to $q \to \infty$. We again let $x_{i_1 \ldots i_K} = c_{i_1 \ldots i_K} + z_{i_1 \ldots i_K}$. In this case, we do not give an explicit decomposition of elements of $\boldsymbol{R}$ into elements of $\boldsymbol{X}$, rather we note that each element of $\boldsymbol{R}$ is a sum of the corresponding element of $\boldsymbol{X}$ and averages over terms of $\boldsymbol{X}$ that belong to at least one of the same modes. The averages with the largest variances are given by averages over the levels of just one mode, e.g. $\bar{x}_{\cdot i_2 \ldots i_K} = \frac{1}{\pi_1 q} \sum_{i_1=1}^{\pi_1 q} x_{i_1 \ldots i_K} = O_p\left(\frac{1}{\sqrt{q}}\right)$. Using the same logic as Equation (A.7) we can write:

$$\begin{aligned}
\bar{r}^{(4)} &= \frac{1}{\pi q^K} \sum_{k=1}^{K} \sum_{i_k=1}^{\pi_k q} \left( x_{i_1 \ldots i_K} + O_p\left(\frac{1}{\sqrt{q}}\right) \right)^4 \\
&= \bar{x}^{(4)} + O_p\left(\frac{1}{q}\right) \\
&= \kappa \sigma_c^4 + 3\left(\sigma_c^2 + \sigma_z^2\right)^2 .
\end{aligned}$$

It follows that $\bar{r}^{(4)} \xrightarrow{p} \kappa \sigma_c^4 + 3\left(\sigma_c^2 + \sigma_z^2\right)^2$ as $q \to \infty$, i.e. as $p_1, \ldots, p_K \to \infty$.

Plugging in the excess kurtosis, $\kappa = 3$, of Laplace distributed $c_{ij}$, combining $\bar{r}^{(4)} \xrightarrow{p} 3\sigma_c^4 + 3\left(\sigma_c^2 + \sigma_z^2\right)^2$ as $p_1, \ldots, p_K \to \infty$ with $\bar{r}^{(2)} \xrightarrow{p} \sigma_c^2 + \sigma_z^2$ as $p_1, \ldots, p_K \to \infty$ and applying the continuous mapping theorem yields $\hat{\sigma}_c^4 \xrightarrow{p} \sigma_c^4$, $\hat{\sigma}_c^2 \xrightarrow{p} \sigma_c^2$, $\hat{\lambda}_c \xrightarrow{p} \lambda_c$ and $\hat{\sigma}_z^2 \xrightarrow{p} \sigma_z^2$ as $p_1, \ldots, p_K \to \infty$.

### *A.16 Notes on Extending Propositions 2.3.1-2.3.3 to Tensor Data*

Extending Propositions 2.3.1-2.3.3 to tensor data amounts to showing

$$\sqrt{p}\left(\hat{\sigma}_c^4 - \kappa \sigma_c^4/3\right) \xrightarrow{d} \sqrt{p}\left(\bar{x}^{(4)}/3 - \left(\bar{x}^{(2)}\right) - \kappa \sigma_c^4/3\right),$$

where $x_{i_1 \ldots i_K} = c_{i_1 \ldots i_K} + z_{i_1 \ldots i_K}$, as $p_1, \ldots, p_K \to \infty$. We do not go into great detail showing this as we did in the matrix case because all the same logic used to prove $\bar{r}^{(2)} \xrightarrow{d} \bar{x}^{(2)}$ and $\bar{r}^{(4)} \xrightarrow{d} \bar{x}^{(4)}$ as $n, p \to \infty$ in the matrix case can be applied.

For convenience, we also let $p_k = \pi_k q$, where $0 < \pi_k < 1$ are fixed and $\sum_{k=1}^K \pi_k = 1$. We also let $\pi = \prod_{k=1}^K \pi_k$ Then the scenario $p_1, \ldots, p_K \to \infty$ is equivalent to $q \to \infty$. As in the previous proof, we do not give an explicit decomposition of elements of $\boldsymbol{R}$ into elements of $\boldsymbol{X}$, rather we note that each element of $\boldsymbol{R}$ is a sum of the corresponding element of $\boldsymbol{X}$ and averages over terms of $\boldsymbol{X}$ that belong to at least one of the same modes. The averages with the largest variances are given by averages over the levels of just one mode, e.g. $\bar{x}_{\cdot i_2 \ldots i_K} = \frac{1}{\pi_1 q} \sum_{i_1=1}^{\pi_1 q} x_{i_1 \ldots i_K} = O_p\left(\frac{1}{\sqrt{q}}\right)$, and the averages with the second largest variances are given by averages over the levels of two modes, e.g. $\bar{x}_{\cdot\cdot i_3 \ldots i_K} = \frac{1}{\pi_1 \pi_2 q^2} \sum_{i_1=1}^{\pi_1 q} \sum_{i_2=1}^{\pi_2 q} x_{i_1 i_2 \ldots i_K} = O_p\left(\frac{1}{q}\right)$. For ease of notation, we'll write the average of elements of $\boldsymbol{X}$ over level $i_k$ of the $k$-th mode as $\bar{x}_{-i_k}$. We can write:

$$r_{i_1 \ldots i_K} = x_{i_1 \ldots i_K} - \sum_{k=1}^K \bar{x}_{-i_k} + O_p\left(\frac{1}{q}\right).$$

For $\bar{r}^{(2)}$ we have:

$$\begin{aligned}
\bar{r}^{(2)} =& \frac{1}{\pi q^K} \sum_{k=1}^K \sum_{i_k=1}^{\pi_k q} \left( x_{i_1 \ldots i_K} - \sum_{k=1}^K \bar{x}_{-i_k} + O_p\left(\frac{1}{q}\right) \right)^2 \\
=& \frac{1}{\pi q^K} \sum_{k=1}^K \sum_{i_k=1}^{\pi_k q} \left( x_{i_1 \ldots i_K} - \sum_{k=1}^K \bar{x}_{-i_k} \right)^2 + \left( x_{i_1 \ldots i_K} - \sum_{k=1}^K \bar{x}_{-i_k} \right) O_p\left(\frac{1}{q}\right) + O_p\left(\frac{1}{q^2}\right) \\
=& \frac{1}{\pi q^K} \sum_{k=1}^K \sum_{i_k=1}^{\pi_k q} \left( x_{i_1 \ldots i_K} - \sum_{k=1}^K \bar{x}_{-i_k} \right)^2 + O_p\left(\frac{1}{q^2}\right) \\
=& \frac{1}{\pi q^K} \sum_{k=1}^K \sum_{i_k=1}^{\pi_k q} x_{i_1 \ldots i_K}^2 + \sum_{k=1}^K \bar{x}_{-i_k}^2 - 2 \sum_{k=1}^K x_{i_1 \ldots i_K} \bar{x}_{-i_k} + 2 \sum_{k=1}^K \sum_{k'=1, k' \neq k}^K \bar{x}_{-i_k} \bar{x}_{-i'_k} + O_p\left(\frac{1}{q^2}\right) \\
=& \bar{x}^{(2)} + \frac{\mathbb{E}\left[x_{i_1 \ldots i_K}^2\right]}{\pi q} + O_p\left(\frac{1}{\sqrt{q}^3}\right),
\end{aligned}$$

where the last line follows from using the same logic as the matrix case.

For $\bar{r}^{(4)}$ we have:

$$\begin{aligned}
\bar{r}^{(4)} =& \frac{1}{\pi q^K} \sum_{k=1}^{K} \sum_{i_k=1}^{\pi_k q} \left( x_{i_1 \ldots i_K} - \sum_{k=1}^{K} \bar{x}_{-i_k} + O_p\left(\frac{1}{q}\right) \right)^4 \\
=& \frac{1}{\pi q^K} \sum_{k=1}^{K} \sum_{i_k=1}^{\pi_k q} \left( x_{i_1 \ldots i_K} - \sum_{k=1}^{K} \bar{x}_{-i_k} \right)^4 + \left( x_{i_1 \ldots i_K} - \sum_{k=1}^{K} \bar{x}_{-i_k} \right)^3 O_p\left(\frac{1}{q}\right) + \\
& \left( x_{i_1 \ldots i_K} - \sum_{k=1}^{K} \bar{x}_{-i_k} \right)^2 O_p\left(\frac{1}{q^2}\right) + \left( x_{i_1 \ldots i_K} - \sum_{k=1}^{K} \bar{x}_{-i_k} \right) O_p\left(\frac{1}{q^3}\right) + O_p\left(\frac{1}{q^4}\right) \\
=& \frac{1}{\pi q^K} \sum_{k=1}^{K} \sum_{i_k=1}^{\pi_k q} \left( x_{i_1 \ldots i_K} - \sum_{k=1}^{K} \bar{x}_{-i_k} \right)^4 + O_p\left(\frac{1}{q^2}\right) \\
=& \frac{1}{\pi q^K} \sum_{k=1}^{K} \sum_{i_k=1}^{\pi_k q} x_{i_1 \ldots i_K}^2 + \sum_{k=1}^{K} \bar{x}_{-i_k}^2 - 2 \sum_{k=1}^{K} x_{i_1 \ldots i_K} \bar{x}_{-i_k} + 2 \sum_{k=1}^{K} \sum_{k'=1, k' \neq k}^{K} \bar{x}_{-i_k} \bar{x}_{-i'_k} + O_p\left(\frac{1}{q^2}\right) \\
=& \bar{x}^{(4)} + \frac{6\mathbb{E}\left[x_{i_1 \ldots i_K}^2\right]}{\pi q} + O\left(\frac{1}{q^2}\right) + O_p\left(\frac{1}{\sqrt{q}^3}\right),
\end{aligned}$$

where again the last line follows from using the same logic as the matrix case.

Now we plug these expressions for $\bar{r}^{(2)}$ and $\bar{r}^{(4)}$ into our equation for $\hat{\sigma}_c^4$:

$$\begin{aligned}
\hat{\sigma}_c^4 =& O(1) \left( \bar{r}^{(4)}/3 - \left(\bar{r}^{(2)}\right)^2 \right) \\
=& O(1) \left( \bar{x}^{(4)}/3 - \left(\bar{x}^{(2)}\right)^2 + O\left(\frac{1}{q^2}\right) + O_p\left(\frac{1}{\sqrt{q}^3}\right) \right).
\end{aligned}$$

Recalling that $\kappa$ is defined such that $\mathbb{E}\left[x_{ij}^4\right]/3 - \mathbb{E}\left[x_{ij}^2\right]^2 = \kappa\sigma^4/3$, we have

$$\sqrt{\pi q^K} \left( \hat{\sigma}_c^4 - \kappa\sigma_c^4/3 \right) \xrightarrow{d} \sqrt{\pi q^K} \left( \bar{x}^{(4)}/3 - \left(\bar{x}^2\right)^2 - \kappa\sigma_c^4/3 \right)$$

as $q \to \infty$, i.e.

$$\sqrt{p} \left( \hat{\sigma}_c^4 - \kappa\sigma_c^4/3 \right) \xrightarrow{d} \sqrt{p} \left( \bar{x}^{(4)}/3 - \left(\bar{x}^2\right)^2 - \kappa\sigma_c^4/3 \right).$$

as $p_1, \ldots, p_K \to \infty$.

### *A.17 Block Coordinate Descent for LANOVA Penalization for Three-Way Tensor Data*

We can use block coordinate descent to obtain estimates of the unknown mean parameters as by iterating the following procedure until the objective function,

$$\frac{1}{2\hat{\sigma}_c^2}\left|\left|\text{vec}\left(\boldsymbol{Y}\right)-\boldsymbol{X}\boldsymbol{\beta}-\text{vec}\left(\boldsymbol{C}\right)\right|\right|_2^2+\hat{\lambda}_c\left|\left|\text{vec}\left(\boldsymbol{C}\right)\right|\right|_1$$

converges, starting with $\text{vec}\left(\hat{\boldsymbol{C}}^{(0)}\right)=\left(\boldsymbol{H}_{p_K}\otimes\cdots\otimes\boldsymbol{H}_{p_1}\right)\text{vec}\left(\boldsymbol{Y}\right)$ and $k=1$:

- Set $\hat{\boldsymbol{\beta}}^{(k)}=\text{argmin}_{\boldsymbol{\beta}}\left|\left|\text{vec}\left(\boldsymbol{Y}\right)-\text{vec}\left(\hat{\boldsymbol{C}}^{(k-1)}\right)-\boldsymbol{W}\boldsymbol{\beta}\right|\right|_2^2$;
- Set $\text{vec}\left(\hat{\boldsymbol{C}}^{(k)}\right)=\text{sign}\left(\text{vec}\left(\boldsymbol{Y}\right)-\boldsymbol{W}\hat{\boldsymbol{\beta}}^{(k)}\right)\left(\left|\text{vec}\left(\boldsymbol{Y}\right)-\boldsymbol{X}\hat{\boldsymbol{\beta}}^{(k)}\right|-\hat{\lambda}_c\hat{\sigma}_z^2\right)_+$.

The step of setting $\hat{\boldsymbol{\beta}}^{(k)}$ is simpler than it appears, because as we have observed in the matrix case, the unpenalized regression problem is the equivalent to fitting a $K$-way ANOVA decomposition to $\boldsymbol{Y}-\hat{\boldsymbol{C}}^{(k-1)}$.

### *A.18 Lower-Order Mean Parameter Variance Estimators for Three-Way Tensors*

First we define the following unpenalized OLS estimates:

$$\begin{aligned}
\check{\boldsymbol{a}}&=\left(\boldsymbol{1}_{p_3}/p_3\otimes\boldsymbol{1}_{p_2}/p_2\otimes\boldsymbol{H}_{p_1}\right)\text{vec}\left(\boldsymbol{Y}\right)\\
\check{\boldsymbol{b}}&=\left(\boldsymbol{1}_{p_3}/p_3\otimes\boldsymbol{H}_{p_2}\otimes\boldsymbol{1}_{p_1}/p_1\right)\text{vec}\left(\boldsymbol{Y}\right)\\
\check{\boldsymbol{d}}&=\left(\boldsymbol{H}_{p_3}\otimes\boldsymbol{1}_{p_2}/p_2\otimes\boldsymbol{1}_{p_1}/p_1\right)\text{vec}\left(\boldsymbol{Y}\right)\\
\text{vec}\left(\check{\boldsymbol{E}}\right)&=\left(\boldsymbol{1}_{p_3}/p_3\otimes\boldsymbol{H}_{p_2}\otimes\boldsymbol{H}_{p_1}\right)\text{vec}\left(\boldsymbol{Y}\right)\\
\text{vec}\left(\check{\boldsymbol{F}}\right)&=\left(\boldsymbol{H}_{p_3}\otimes\boldsymbol{1}_{p_2}/p_2\otimes\boldsymbol{H}_{p_1}\right)\text{vec}\left(\boldsymbol{Y}\right)\\
\text{vec}\left(\check{\boldsymbol{G}}\right)&=\left(\boldsymbol{H}_{p_3}\otimes\boldsymbol{H}_{p_2}\otimes\boldsymbol{1}_{p_1}/p_1\right)\text{vec}\left(\boldsymbol{Y}\right).
\end{aligned}$$

Now we define the following estimators of lower order-mean parameter variances:

$$\begin{aligned}
\hat{\sigma}_a^2 &= \frac{p_1}{p_1-1}\left(\bar{\breve{a}}^{(2)} - \frac{1}{p_2-1}\bar{\breve{e}}^{(2)} - \frac{1}{p_3-1}\bar{\breve{f}}^{(2)} + \frac{1}{(p_2-1)(p_3-1)}\bar{\breve{r}}^{(2)}\right) \\
\hat{\sigma}_b^2 &= \frac{p_2}{p_2-1}\left(\bar{\breve{b}}^{(2)} - \frac{1}{p_1-1}\bar{\breve{e}}^{(2)} - \frac{1}{p_3-1}\bar{\breve{g}}^{(2)} + \frac{1}{(p_1-1)(p_3-1)}\bar{\breve{r}}^{(2)}\right) \\
\hat{\sigma}_d^2 &= \frac{p_3}{p_3-1}\left(\bar{\breve{d}}^{(2)} - \frac{1}{p_1-1}\bar{\breve{f}}^{(2)} - \frac{1}{p_2-1}\bar{\breve{g}}^{(2)} + \frac{1}{(p_1-1)(p_2-1)}\bar{\breve{r}}^{(2)}\right) \\
\hat{\sigma}_e^2 &= \frac{p_1p_2}{(p_1-1)(p_2-1)}\left(\bar{\breve{e}}^{(2)} - \frac{1}{p_3-1}\bar{\breve{r}}^{(2)}\right) \\
\hat{\sigma}_f^2 &= \frac{p_1p_3}{(p_1-1)(p_3-1)}\left(\bar{\breve{f}}^{(2)} - \frac{1}{p_2-1}\bar{\breve{r}}^{(2)}\right) \\
\hat{\sigma}_g^2 &= \frac{p_2p_3}{(p_2-1)(p_3-1)}\left(\bar{\breve{g}}^{(2)} - \frac{1}{p_1-1}\bar{\breve{r}}^{(2)}\right).
\end{aligned}$$

where $\bar{\breve{a}}^{(2)} = \frac{1}{n}\sum_{i=1}^{n}\breve{a}_i^2$, $\bar{\breve{d}}^{(2)} = \frac{1}{np_2}\sum_{i=1}^{n}\sum_{j=1}^{p_2}\breve{d}_{ij}^2$ and $\bar{\breve{b}}^{(2)}$, $\bar{\breve{c}}^{(2)}$, $\bar{\breve{e}}^{(2)}$ and $\bar{\breve{f}}^{(2)}$ are defined accordingly and $\hat{\sigma}_a^2$ estimates $\sigma_a^2 = 2/\lambda_a^2$, $\sigma_b^2$, $\sigma_c^2$, $\sigma_d^2$, $\sigma_e^2$ and $\sigma_f^2$. These estimators are unbiased and consistent under certain asymptotic scenarios; all of the estimators are consistent as $p_1, \ldots, p_K \to \infty$, however only $\hat{\sigma}_a^2$, $\hat{\sigma}_e^2$ and $\hat{\sigma}_f^2$ are consistent as $p_1 \to \infty$ with $p_2$, $p_3$ fixed.

# Appendix B

# SUPPLEMENT TO "TESTING SPARSITY-INDUCING PENALTIES"

## *B.1 Proof of Propositions 3.2.1 and 3.2.2*

Proposition 3.2.1 is a special case of Proposition 3.2.2 with $\delta^2 = 0$. Accordingly, a proof of 3.2.2 suffices for both.

Let $\hat{\boldsymbol{\beta}}_\delta = \boldsymbol{V}^{-1}(\boldsymbol{C} + \delta^2 \boldsymbol{I}_p)^{-1}\boldsymbol{V}^{-1}\boldsymbol{X}^\top \boldsymbol{y}$, where $\delta \geq 0$ is a nonzero constant, where $\boldsymbol{V}$ is a diagonal matrix with diagonal elements $\sqrt{\text{diag}\left(\boldsymbol{X}^\top \boldsymbol{X}\right)}$ and $\boldsymbol{C}$ is a "correlation" matrix corresponding to $\boldsymbol{X}^\top \boldsymbol{X}$, such that $\boldsymbol{X}^\top \boldsymbol{X} = \boldsymbol{V}\boldsymbol{C}\boldsymbol{V}$. Define $\boldsymbol{\beta}_\delta = \boldsymbol{V}^{-1}\left(\boldsymbol{C} + \delta^2 \boldsymbol{I}_p\right)^{-1}\boldsymbol{C}\boldsymbol{V}\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}_\delta = \boldsymbol{V}^{-1}\left(\boldsymbol{C} + \delta^2 \boldsymbol{I}_p\right)^{-1}\boldsymbol{C}\left(\boldsymbol{C} + \delta^2 \boldsymbol{I}_p\right)^{-1}\boldsymbol{V}^{-1}$. Under normality of the errors $\boldsymbol{z}$ as assumed in the model given by

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{z}, \quad \beta_1, \ldots, \beta_p \overset{i.i.d.}{\sim} EP\left(\tau, q\right), \quad \boldsymbol{z} \sim N\left(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n\right) \tag{B.1}$$

where $EP$ is the exponential power distribution with unknown shape parameter $q$, $\tau^2$ and $\sigma^2$ are the unknown variances of the regression coefficients $\boldsymbol{\beta}$ and noise $\boldsymbol{z}$, $\hat{\boldsymbol{\beta}}_\delta | \boldsymbol{\beta} \sim N\left(\boldsymbol{\beta}_\delta, \sigma^2 \boldsymbol{\Sigma}_\delta\right)$.

Following Van der Vaart (2000), the approximate distribution of $\psi\left(\hat{\boldsymbol{\beta}}_\delta\right)$ conditional on $\boldsymbol{\beta}$, can be obtained via a Taylor expansion/the delta method:

$$\psi\left(\hat{\boldsymbol{\beta}}_\delta\right) = \psi\left(\boldsymbol{\beta}_\delta\right) + \nabla\psi\left(\boldsymbol{\beta}_\delta\right)'\left(\hat{\boldsymbol{\beta}}_\delta - \boldsymbol{\beta}_\delta\right) + r\left(\hat{\boldsymbol{\beta}}_\delta, \boldsymbol{\beta}_\delta\right), \tag{B.2}$$

where elements of $\nabla\psi\left(\boldsymbol{\beta}_\delta\right)$ are given by

$$\frac{\partial}{\partial \beta_j}\frac{\frac{1}{p}\sum_{j'=1}^p \beta_j^4}{\left(\frac{1}{p}\sum_{j'=1}^p \beta_j^2\right)^2} = \frac{4}{p}\left(\frac{\beta_j^3\left(\frac{1}{p}\sum_{j'=1}^p \beta_j^2\right) - \beta_j\left(\frac{1}{p}\sum_{j'=1}^p \beta_j^4\right)}{\left(\frac{1}{p}\sum_{j'=1}^p \beta_j^2\right)^3}\right)$$

and $r\left(\hat{\boldsymbol{\beta}}_\delta - \boldsymbol{\beta}_\delta\right)$ is a remainder term that depends on $\boldsymbol{\beta}_\delta - \boldsymbol{\beta}_\delta$ and higher-order derivatives of the function $\psi\left(\cdot\right)$ evaluated at $\boldsymbol{\beta}_\delta$. Because higher order derivatives of the function $\psi\left(\cdot\right)$

will all involve the constant $\frac{1}{p}$ multiplied by a function of $\boldsymbol{\beta}_\delta$ and $p$ and and $\hat{\boldsymbol{\beta}}_\delta - \boldsymbol{\beta}_\delta$,

$$r\left(\hat{\boldsymbol{\beta}}_\delta, \boldsymbol{\beta}_\delta\right) = o_p\left(\frac{1}{p}\left|\left|\hat{\boldsymbol{\beta}}_\delta - \boldsymbol{\beta}_\delta\right|\right|_2\right) = o_p\left(\sigma^2 \text{tr}(\boldsymbol{\Sigma}_\delta)/p\right)$$

Returning to (B.2) and letting $\boldsymbol{\beta}_\delta^3$ refer to $\boldsymbol{\beta}_\delta$ with elements raised to the third power

$$\begin{aligned}\mathbb{E}[(\psi(\hat{\boldsymbol{\beta}}_\delta) - \psi(\boldsymbol{\beta}_\delta))^2|\boldsymbol{\beta}] =& \frac{16\sigma^2}{p^2 m_2\left(\boldsymbol{\beta}_\delta\right)^6}\left(m_2\left(\boldsymbol{\beta}_\delta\right)^2 \boldsymbol{\beta}_\delta^{3\top}\boldsymbol{\Sigma}_\delta\boldsymbol{\beta}_\delta^3 - m_4\left(\boldsymbol{\beta}_\delta\right)^2 \boldsymbol{\beta}_\delta^\top\boldsymbol{\Sigma}_\delta\boldsymbol{\beta}_\delta\right) + \\ & o\left(\sigma^2 \text{tr}\left(\boldsymbol{\Sigma}_\delta\right)/p\right).\end{aligned}$$

Letting $m_6\left(\boldsymbol{\beta}_\delta\right) = \frac{1}{p}\sum_{j=1}^p \beta_{\delta,j}^6$, it follows that

$$\begin{aligned}\mathbb{E}[(\psi(\hat{\boldsymbol{\beta}}_\delta) - \psi(\boldsymbol{\beta}_\delta))^2|\boldsymbol{\beta}] &\leq \frac{16\sigma^2}{p^2 m_2(\boldsymbol{\beta}_\delta)^4}(\boldsymbol{\beta}_\delta^{3\top}\boldsymbol{\Sigma}_\delta\boldsymbol{\beta}_\delta^3) + o\left(\sigma^2 \text{tr}\left(\boldsymbol{\Sigma}_\delta\right)\right) \\ &\leq 16\sigma^2\left(\frac{m_6\left(\boldsymbol{\beta}_\delta\right)}{m_2\left(\boldsymbol{\beta}_\delta\right)^4}\right)\text{tr}\left(\boldsymbol{\Sigma}_\delta\right)/p + o\left(\sigma^2 \text{tr}\left(\boldsymbol{\Sigma}_\delta\right)/p\right).\end{aligned}$$

### *B.2 Bias of* $\psi\left(\boldsymbol{\beta}_\delta\right)$

A first order approximation of the distribution of $\psi\left(\boldsymbol{\beta}_\delta\right)$ is given by:

$$\begin{aligned}\psi\left(\boldsymbol{\beta}_\delta\right) =& m_4\left(\boldsymbol{\beta}_\delta\right)/m_2\left(\boldsymbol{\beta}_\delta\right)^2 \approx \mathbb{E}\left[m_4\left(\boldsymbol{\beta}_\delta\right)\right]/\mathbb{E}\left[m_2\left(\boldsymbol{\beta}_\delta\right)\right]^2 + \\ & \nabla f\left(\mathbb{E}\left[m_2\left(\boldsymbol{\beta}_\delta\right)\right], \mathbb{E}\left[m_4\left(\boldsymbol{\beta}_\delta\right)\right]\right)' \boldsymbol{m},\end{aligned}$$

where $\boldsymbol{m} = \begin{pmatrix} m_2\left(\boldsymbol{\beta}_\delta\right) - \mathbb{E}\left[m_2\left(\boldsymbol{\beta}_\delta\right)\right] \\ m_4\left(\boldsymbol{\beta}_\delta\right) - \mathbb{E}\left[m_4\left(\boldsymbol{\beta}_\delta\right)\right] \end{pmatrix}$ and $f\left(m_2\left(\boldsymbol{\beta}_\delta\right), m_4\left(\boldsymbol{\beta}_\delta\right)\right) = m_4\left(\boldsymbol{\beta}_\delta\right)/m_2\left(\boldsymbol{\beta}_\delta\right)^2$. We can see that when $m_2(\boldsymbol{\beta}_\delta) \xrightarrow{p} \mathbb{E}[m_2(\boldsymbol{\beta}_\delta)]$ and $m_4(\boldsymbol{\beta}_\delta) \xrightarrow{p} \mathbb{E}[m_4(\boldsymbol{\beta}_\delta)]$, applying the continuous mapping theorem yields that approximate the mean of $\psi(\boldsymbol{\beta}_\delta)$ will be $\mathbb{E}\left[m_4\left(\boldsymbol{\beta}_\delta\right)\right]/\mathbb{E}\left[m_2\left(\boldsymbol{\beta}_\delta\right)\right]^2$.

We can compute $\mathbb{E}[m_2(\boldsymbol{\beta}_\delta)]$ and $\mathbb{E}[m_4(\boldsymbol{\beta}_\delta)]$ as follows:

$$\begin{aligned}\mathbb{E}\left[m_2\left(\boldsymbol{\beta}_\delta\right)\right] =& \frac{1}{p}\mathbb{E}\left[\text{tr}\left(\boldsymbol{\beta}_\delta^\top\boldsymbol{\beta}_\delta\right)\right] \\ =& \frac{1}{p}\mathbb{E}\left[\text{tr}\left(\boldsymbol{V}\boldsymbol{C}\left(\boldsymbol{C}+\delta^2\boldsymbol{I}_p\right)^{-1}\boldsymbol{V}^{-1}\boldsymbol{V}^{-1}\left(\boldsymbol{C}+\delta^2\boldsymbol{I}_p\right)^{-1}\boldsymbol{C}\boldsymbol{V}\boldsymbol{\beta}\boldsymbol{\beta}^\top\right)\right] \\ =& \frac{\tau^2}{p}\text{tr}\left(\boldsymbol{V}\boldsymbol{C}\left(\boldsymbol{C}+\delta^2\boldsymbol{I}_p\right)^{-1}\boldsymbol{V}^{-1}\boldsymbol{V}^{-1}\left(\boldsymbol{C}+\delta^2\boldsymbol{I}_p\right)^{-1}\boldsymbol{C}\boldsymbol{V}\right) \\ =& \frac{\tau^2}{p}\text{tr}\left(\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{D}^2\boldsymbol{X}^\top\boldsymbol{X}\right),\end{aligned}$$

where $\boldsymbol{D} = \boldsymbol{V}^{-1}\left(\boldsymbol{C} + \delta^2 \boldsymbol{I}_p\right)^{-1}\boldsymbol{V}^{-1}$.

Letting $\boldsymbol{A} = \boldsymbol{D}\boldsymbol{X}^\top\boldsymbol{X}$ with rows $\boldsymbol{a}_j$,

$$\begin{aligned}\mathbb{E}\left[m_4\left(\boldsymbol{\beta}_\delta\right)\right] =& \frac{1}{p}\sum_{j=1}^{p}\mathbb{E}\left[\left(\boldsymbol{a}_j^\top\boldsymbol{\beta}\right)^4\right] \\ =& \frac{1}{p}\sum_{j=1}^{p}\sum_{k=1}^{p}a_{jk}^4\mathbb{E}\left[\beta_j^4\right] + 3\sum_{k\neq k}^{p}a_{jk}^2a_{jk'}^2\mathbb{E}\left[\beta_j^2\right]^2 \\ =& \frac{\tau^4}{p}\sum_{j=1}^{p}\sum_{k=1}^{p}a_{jk}^4\left(\kappa+3\right) + 3\sum_{k\neq k}^{p}a_{jk}^2a_{jk'}^2\end{aligned}$$

Then letting $\alpha = \mathrm{tr}(\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{D}^2\boldsymbol{X}^\top\boldsymbol{X})/p$, $\gamma = \sum_{j=1}^{p}\sum_{k=1}^{p}(\boldsymbol{D}\boldsymbol{X}^\top\boldsymbol{X})_{jk}^4/p$ and $\omega = 3((\sum_{j=1}^{p}(\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{D}^2\boldsymbol{X}^\top\boldsymbol{X})_{jj}^2)/p - \gamma)$ we have

$$\mathbb{E}\left[m_4\left(\boldsymbol{\beta}_\delta\right)\right]/\mathbb{E}\left[m_2\left(\boldsymbol{\beta}_\delta\right)\right]^2 = p\left(\gamma\left(\kappa+3\right)+\omega\right)/\alpha^2.$$

### *B.3 Coordinate Descent Algorithm/Mode Thresholding Function*

Let $\boldsymbol{x}_j$ refer to the $j$-th column of $\boldsymbol{X}$, $\boldsymbol{X}_{-j}$ refer to the matrix $\boldsymbol{X}$ with the $j$-th column removed and $\boldsymbol{\beta}_{-j}$ refer to the vector $\boldsymbol{\beta}$ with the $j$-th element removed. The solution to the problem

$$\mathrm{minimize}_{\boldsymbol{\beta}}\left|\left|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right|\right|_2^2 + \lambda\left|\left|\boldsymbol{\beta}\right|\right|_q^q$$

can be computed by iteratively solving

$$\mathrm{minimize}_{\beta_j}\boldsymbol{x}_j^\top\boldsymbol{x}_j\beta_j^2 - 2\beta_j\boldsymbol{x}_j'\left(\boldsymbol{y} - \boldsymbol{X}_{-j}\boldsymbol{\beta}_{-j}\right) + \lambda\left|\beta_j\right|^q$$

for $j = 1, \ldots, p$ until convergence of the objective function is obtained.

This requires repeatedly solving the following (which corresponds to the mode thresholding problem when $z = \hat{\beta}_{ols}$)

$$\mathrm{minimize}_{\beta}\frac{1}{2}\left(\beta - z\right)^2 + \lambda\left|\beta\right|^q.$$

When $q \leq 1$, the solution is computed using the method described in Marjanovic and Solo (2014). When $q > 1$, the solution is computed as follows. Like in Marjanovic and Solo (2014), we assume that the optimal value of $\gamma$ will have the same sign as $z$. This allows us to reduce the original problem to the simpler problem:

$$\text{minimize}_{\gamma>0} \frac{1}{2} (\gamma - |z|)^2 + \lambda \gamma^q. \tag{B.3}$$

The optimal $\gamma$ solves:

$$\gamma + \lambda q \gamma^{q-1} = |z| .$$

$\gamma + \lambda q \gamma^{q-1}$ is monotonically increasing for $\gamma > 0$. There are various ways to solve this problem, we did so $\gamma$ using bisection. Because $\gamma + \lambda q \gamma^{q-1} \geq 0$ and $\gamma + \lambda q \gamma^{q-1} \geq \gamma$ for all $\gamma \geq 0$, we can be sure that the optimal value of $\gamma$ is in the interval $[0, |z|]$. Taking $l^{(0)} = 0$, $u^{(0)} = |z|$ and $\gamma^{(0)} = l^{(0)} + \left(u^{(0)} - l^{(0)}\right)/2$ and setting $k = 0$, the bisection algorithm is as follows:

- Set $z^{(k)} = \gamma^{(k)} + \lambda q \left(\gamma^{(k)}\right)^{q-1}$.
- Compare $z^{(k)}$ to $|z|$.
  - If $z^{(k)} = |z|$, take $l^{(k+1)} = u^{(k+1)} = \gamma^{(k)}$.
  - If $z^{(k)} > |z|$, keep $l^{(k+1)} = l^{(k)}$ and set $u^{(k+1)} = \gamma^{(k)}$.
  - If $z^{(k)} < |z|$, keep $u^{(k+1)} = u^{(k)}$ and set $l^{(k+1)} = \gamma^{(k)}$.
- Set $\gamma^{(k+1)} = l^{(k+1)} + \left(u^{(k+1)} - l^{(k+1)}\right)/2$.
- If $u^{(k+1)} = l^{(k+1)}$ set the optimal value of $\beta$ to $\text{sign}(z)\, \gamma^{(k+1)}$.
- If $u^{(k+1)} \neq l^{(k+1)}$, repeat these steps.

### *B.4 Full Conditional Distributions for Gibbs Sampling*

The full conditional distribution of $\boldsymbol{\beta}$ given $\boldsymbol{\gamma}$ is given by:

$$p\left(\boldsymbol{\beta}|\boldsymbol{X},\boldsymbol{y},\boldsymbol{\gamma}\right) \propto_{\boldsymbol{\beta}} \exp\left\{-\frac{1}{2\sigma^2}\left(\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}+2\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{y}\right)\right\}\mathbb{1}_{-\boldsymbol{\Delta}<\boldsymbol{\beta}<\boldsymbol{\Delta}},$$

where $\Delta_j = \left(\sqrt{\frac{\Gamma(1/q)}{\Gamma(3/q)}}\right)\left(\frac{\tau\gamma_j^{1/q}}{\sqrt{2}}\right)$. This is a truncated multivariate normal distribution. To draw samples from $p\left(\boldsymbol{\beta}|\boldsymbol{X},\boldsymbol{y},\boldsymbol{\gamma}\right)$, we use the method described in Rodriguez-Yam et al. (2004) apply a change of variables to transform the problem of sampling from a $p$-variate truncated multivariate normal distribution to $p$ univariate truncated normal distributions. We sample from each univariate truncated normal distribution using Robert (1995). Although this is not the most modern approach, it was the most stable method we considered.

For convenience, define $\eta_j = \left(\sqrt{\frac{\Gamma(3/q)}{\Gamma(1/q)}}\right)\left(\frac{\sqrt{2}|\beta|_j}{\tau}\right)$. Then the full conditional distribution for each $\gamma_j$ is given by:

$$p\left(\gamma_j|\beta_j\right) \propto_{\gamma_j} \exp\left\{-2^{-\frac{q}{2}}\gamma_j\right\}\mathbb{1}_{\eta_j^q<\gamma_j}.$$

This is a translated exponential distribution, from which sampling is straightforward.

## *B.5 Estimates of $\tau^2$, $\sigma^2$ and $q$ for EP Distributed $\beta$*

Figure B.1 shows histograms of the estimates of $\tau^2$, $\sigma^2$ and $q$ corresponding to the simulations considered in Figure 3 of the main text.

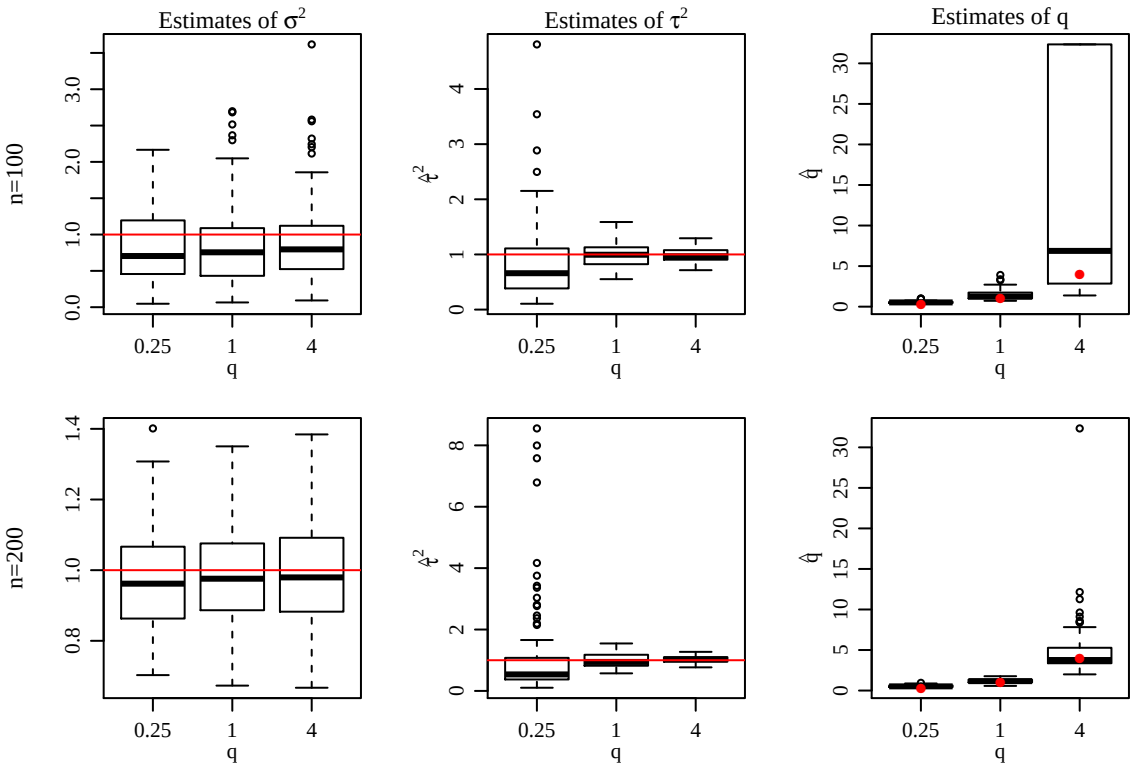



Figure B.1: Performance of estimators of $\tau^2$, $\sigma^2$ and $q$ for exponential power distributed $\boldsymbol{\beta}$. True values printed in red.

### *B.6 Estimates of $\tau^2$, $\sigma^2$ and $q$ for Spike-and-Slab Distributed $\beta$*

Figure B.2 shows histograms of the estimates of $\tau^2$, $\sigma^2$ and $q$ corresponding to the simulations considered in Figure 5 of the main text.

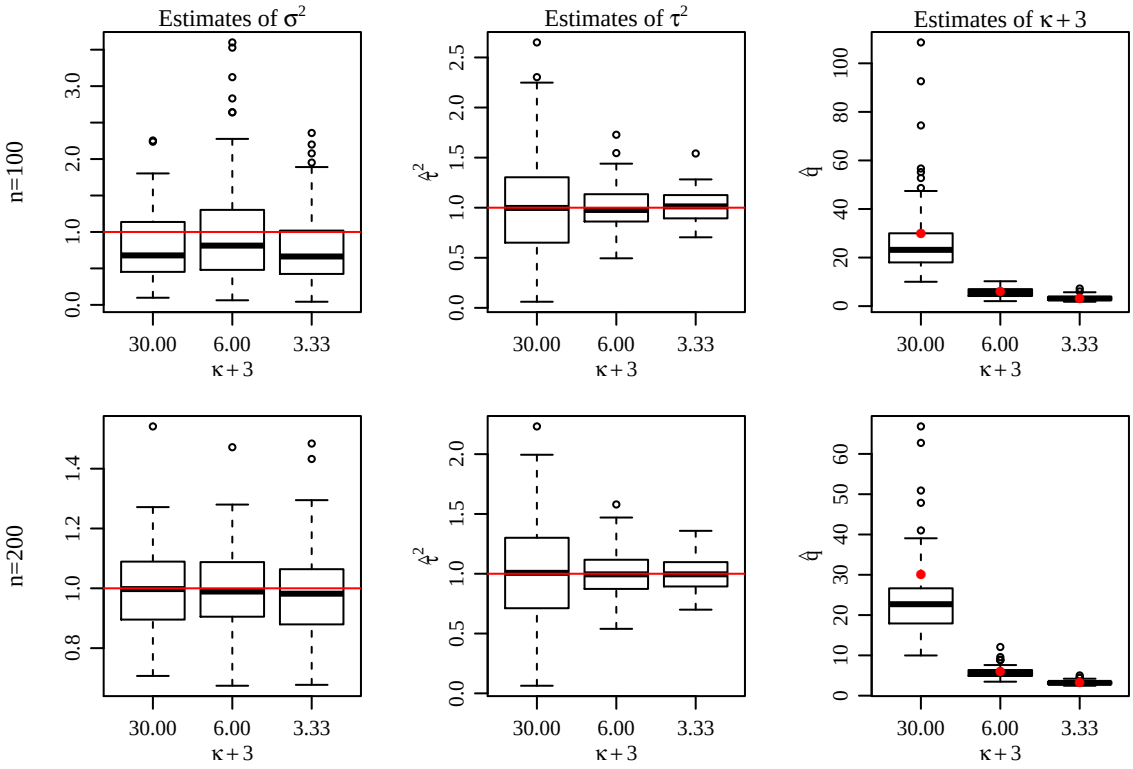



Figure B.2: Performance of estimators of $\tau^2$, $\sigma^2$ and $\kappa+3$ for Bernoulli-normal spike-and-slab distributed $\boldsymbol{\beta}$. True values printed in red.

# Appendix C

# SUPPLEMENT TO "STRUCTURED SHRINKAGE PRIORS"

## *C.1 Univariate Marginal Distributions*

Intuitively, it is clear from the stochastic representation that the marginal distributions are the same as the corresponding univariate shrinkage prior. We can show this directly as follows. The joint marginal prior distribution of $\boldsymbol{\beta}$ is

$$p_{\boldsymbol{\beta}}(\boldsymbol{\beta}) = \int p(\boldsymbol{s})\, p(\boldsymbol{\beta}/\boldsymbol{s}) \left( \prod_{j=1}^{p} \frac{1}{|s_j|} \right) ds_1 \dots ds_p$$

Then $p_{\beta_1}(\beta_1)$ is given by

$$\begin{aligned} p_{\beta_1}(\beta_1) &= \int p(\beta_1/s_1)\, p\left(\boldsymbol{\beta}_{-1}/\boldsymbol{s}_{-1}|\beta_1/s_1\right) \left( \prod_{j=1}^{p} \frac{p(s_j)}{|s_j|} \right) ds_1 \dots ds_p d\beta_2 \dots d\beta_p \\ &= \int p(\beta_1/s_1)\, p(s_1)/|s_1| \underbrace{\left( \int p\left(\boldsymbol{\beta}_{-1}/\boldsymbol{s}_{-1}|\beta_1/s_1\right) \left( \prod_{j=2}^{p} \frac{p(s_j)}{|s_j|} \right) ds_2 \dots ds_p d\beta_2 \dots d\beta_p \right)}_{(*)} ds_1. \end{aligned}$$

The term $(*)$ is equal to $\int p\left(\boldsymbol{\beta}_{-1}|\beta_1/s_1\right) d\beta_2 \dots d\beta_p$. This is the integral of a density, and accordingly $(*) = 1$ and $p_{\beta_1} = \int p(\beta_1/s_1)\, p(s_1)/|s_1|\, ds_1$.

## *C.2 Proofs of Propositions 2.1 and 2.3*

First, we prove the following lemma:

**Lemma C.2.1** *For $\alpha > 0$ and $\gamma \in \mathbb{R}$, $\int_{-\infty}^{\infty} \frac{1}{|s|} exp\{-\alpha(s^2 - \gamma s)\} ds = +\infty$.*

First let's consider this integral when $\gamma > 0$:

$$\begin{aligned} \int_{-\infty}^{\infty} \frac{1}{|s|} \exp\left\{-\alpha\left(s^2 - \gamma s\right)\right\} ds = &\int_{-\infty}^{0} -\frac{1}{s} \exp\{-\alpha s\}^{s-\gamma} ds + \\ &\int_{0}^{\gamma} \frac{1}{s} \exp\{-\alpha s\}^{s-\gamma} ds + \int_{\gamma}^{\infty} \frac{1}{s} \exp\{-\alpha s\}^{s-\gamma} ds. \end{aligned}$$

Because the integrand is nonnegative for all $s$, if *any* of these terms evaluate to $+\infty$, the entire integral evaluates to $+\infty$. Let's examine the middle integral. Note that over this range, $s-\gamma \leq 0$, $\exp\{-\alpha s\} \leq 1$ and accordingly, $\exp\{-\alpha s\}^{s-\gamma} \geq 1$.

$$\begin{aligned}\int_0^{\gamma} \frac{1}{s}\exp\{-\alpha s\}^{s-\gamma}\, ds &\geq \int_0^{\gamma} \frac{1}{s} ds \\ &= (\ln(\gamma) - \lim_{a\to 0+}\ln(a)) = +\infty.\end{aligned}$$

Now let's consider the same integral when $\gamma < 0$:

$$\begin{aligned}\int_{-\infty}^{\infty} \frac{1}{|s|}\exp\left\{-\alpha\left(s^2-\gamma s\right)\right\} ds = &\int_{-\infty}^{\gamma} -\frac{1}{s}\exp\{-\alpha s\}^{s-\gamma}\, ds + \\ &\int_{\gamma}^{0} -\frac{1}{s}\exp\{-\alpha s\}^{s-\gamma}\, ds + \int_0^{\infty} \frac{1}{s}\exp\{-\alpha s\}^{s-\gamma}\, ds.\end{aligned}$$

Again, if *any* of these terms evaluate to $+\infty$, the entire integral evaluates to $+\infty$. Again, let's consider the middle term. Over this interval, $\exp\{-\alpha s\} \geq 1$ and $s-\gamma \geq 0$. It follows that $\exp\{-\alpha s\}^{s-\gamma} \geq 1$ and:

$$\begin{aligned}\int_{\gamma}^{0} -\frac{1}{s}\exp\{-\alpha s\}^{s-\gamma}\, ds &\geq \int_{\gamma}^{0} -\frac{1}{s} ds \\ &= \int_0^{-\gamma} \frac{1}{s} ds = +\infty.\end{aligned}$$

Now we'll consider one last case where $\gamma = 0$:

$$\int_{-\infty}^{\infty} \frac{1}{|s|}\exp\left\{-\alpha s^2\right\} ds = \int_{-\infty}^{0} -\frac{1}{s}\exp\left\{-\alpha s^2\right\} ds + \int_0^{1/\sqrt{\alpha}} \frac{1}{s}\exp\left\{-\alpha s^2\right\} ds + \int_{1/\sqrt{\alpha}}^{\infty} \frac{1}{s}\exp\left\{-\alpha s^2\right\} ds.$$

As in the previous cases, *any* of these terms evaluate to $+\infty$, the entire integral evaluates to $+\infty$. Examining the middle term one last time, we have:

$$\begin{aligned}\int_0^{1/\sqrt{\alpha}} \frac{1}{s}\exp\left\{-\alpha s^2\right\} ds &\geq \int_0^{1/\sqrt{\alpha}} \frac{1}{s}\exp\left\{-\sqrt{\alpha} s\right\} ds \\ &\geq \int_0^{1/\sqrt{\alpha}} \frac{1-\sqrt{\alpha} s}{s} ds \\ &= \int_0^{1/\sqrt{\alpha}} \frac{1}{s} ds - \int_0^{1/\sqrt{\alpha}} \alpha ds \\ &= \ln\left(1/\sqrt{\alpha}\right) - \lim_{a\to 0+}\ln(a) - \sqrt{\alpha} = +\infty.\end{aligned}$$

**Proof of Proposition 2.1:** Now we can evaluate the marginal density $p_{\boldsymbol{\beta}}(\boldsymbol{b}|\boldsymbol{\Omega},\boldsymbol{\Psi})$ when $b_j = 0$ for some $j \in \{1,\ldots,p\}$. Without loss of generality, set $\beta_1 = 0$. Letting $\boldsymbol{b}_{-1}$ and $\boldsymbol{s}_{-1}$ refer to the vectors $\boldsymbol{b}$ and $\boldsymbol{s}$ each with the first element removed, $(\boldsymbol{\Omega}^{-1})_{-1,-1}$ be the matrix $\boldsymbol{\Omega}^{-1}$ with the first row and column removed, $(\boldsymbol{\Psi}^{-1})_{-1,-1}$ be the matrix $\boldsymbol{\Psi}^{-1}$ with the first row and column removed and $(\boldsymbol{\Psi}^{-1})_{-1,1}$ be the first column of $\boldsymbol{\Psi}^{-1}$ excluding the first element. For $\boldsymbol{b} = (0, \boldsymbol{b}_{-1})$,

$$
\begin{aligned}
p_{\boldsymbol{\beta}}\left((0,\boldsymbol{b}_{-1})\,|\boldsymbol{\Psi},\boldsymbol{\Omega}\right) &\propto \int \frac{1}{\prod_{i=1}^{p}|s_i|}\exp\left\{-\frac{1}{2}\left(\boldsymbol{b}'\mathrm{diag}\left\{\mathbf{1}/\boldsymbol{s}\right\}\boldsymbol{\Omega}^{-1}\mathrm{diag}\left\{\mathbf{1}/\boldsymbol{s}\right\}\boldsymbol{b}+\boldsymbol{s}'\boldsymbol{\Psi}^{-1}\boldsymbol{s}\right)\right\}d\boldsymbol{s} \\
&= \int \frac{1}{\prod_{2=1}^{p}|s_i|}\exp\left\{-\frac{1}{2}\left(\boldsymbol{b}'_{-1}\mathrm{diag}\left\{\mathbf{1}/\boldsymbol{s}_{-1}\right\}\left(\boldsymbol{\Omega}^{-1}\right)_{-1,-1}\mathrm{diag}\left\{\mathbf{1}/\boldsymbol{s}_{-1}\right\}\boldsymbol{b}_{-1}+\boldsymbol{s}'_{-1}\left(\boldsymbol{\Psi}^{-1}\right)_{-1,-1}\boldsymbol{s}_{-1}\right)\right\} \\
&\quad \underbrace{\int_{-\infty}^{\infty}\frac{1}{|s_1|}\exp\left\{-\frac{1}{2}\left(s_1^2\left(\boldsymbol{\Psi}^{-1}\right)_{11}-2s_1\left(\boldsymbol{\Psi}^{-1}\right)'_{-1,1}\boldsymbol{s}_{-1}\right)\right\}ds_1}_{(*)}\,d\boldsymbol{s}_{-1}.
\end{aligned}
$$

Applying Lemma C.2.1, the term denoted by $(*)$ evaluates to $+\infty$ for every value of $\boldsymbol{s}_{-1}$.

**Proof of Proposition 2.3:** Proposition 2.3 follows from Proposition 2.1 and Lemma C.2.1. Again, letting $\beta_1 = 0$, for $\boldsymbol{\beta} = (0, \boldsymbol{\beta}_{-1})$,

$$
\begin{aligned}
p_{\beta_1}\left(0|\boldsymbol{\Psi},\boldsymbol{\Omega}\right) &\propto \int \frac{1}{\prod_{i=1}^{p}|s_i|}\exp\left\{-\frac{1}{2}\left(\boldsymbol{\beta}'\mathrm{diag}\left\{\mathbf{1}/\boldsymbol{s}\right\}\boldsymbol{\Omega}^{-1}\mathrm{diag}\left\{\mathbf{1}/\boldsymbol{s}\right\}\boldsymbol{\beta}+\boldsymbol{s}'\boldsymbol{\Psi}^{-1}\boldsymbol{s}\right)\right\}d\boldsymbol{s}d\boldsymbol{\beta}_{-1} \\
&= \int \frac{1}{\prod_{2=1}^{p}|s_i|}\exp\left\{-\frac{1}{2}\left(\boldsymbol{\beta}'_{-1}\mathrm{diag}\left\{\mathbf{1}/\boldsymbol{s}_{-1}\right\}\left(\boldsymbol{\Omega}^{-1}\right)_{-1,-1}\mathrm{diag}\left\{\mathbf{1}/\boldsymbol{s}_{-1}\right\}\boldsymbol{\beta}_{-1}+\boldsymbol{s}'_{-1}\left(\boldsymbol{\Psi}^{-1}\right)_{-1,-1}\boldsymbol{s}_{-1}\right)\right\} \\
&\quad \underbrace{\int_{-\infty}^{\infty}\frac{1}{|s_1|}\exp\left\{-\frac{1}{2}\left(s_1^2\left(\boldsymbol{\Psi}^{-1}\right)_{11}-2s_1\left(\boldsymbol{\Psi}^{-1}\right)'_{-1,1}\boldsymbol{s}_{-1}\right)\right\}ds_1}_{(*)}\,d\boldsymbol{s}_{-1}d\boldsymbol{\beta}_{-1}.
\end{aligned}
$$

Again, $(*)$ evaluates to $+\infty$ for any value of $\boldsymbol{s}_{-1}$ and does not depend at all on $\boldsymbol{\beta}_{-1}$.

### *C.3 Expectation of $s_j$ when $s_j^2 \sim$ gamma$(c,c)$*

When $s_j^2 \sim \text{gamma}(c,c)$, we have:

$$
\begin{aligned}
\int_0^{\infty} s_j\frac{c^c}{\Gamma(c)}\left(s_j^2\right)^{c-1}\exp\left\{-cs_j^2\right\}ds_j^2 &= \frac{c^c}{\Gamma(c)}\int_0^{\infty}\left(s_j^2\right)^{c+1/2-1}\exp\left\{-cs_j^2\right\}ds_j^2 \\
&= \left(\frac{c^c\Gamma(c+1/2)}{c^{c+1/2}\Gamma(c)}\right) \\
&= c^{-1/2}\left(\Gamma(c+1/2)/\Gamma(c)\right)
\end{aligned}
$$

### *C.4 Maximum correlation for bivariate β under SNG prior*

Consider $\mathbf{\Omega} = (1-\omega)\,\boldsymbol{I}_2 + \omega \mathbf{1}_2 \mathbf{1}_2'$. When $s_j^2 \sim \text{gamma}(c,c)$, we have:

$$\mathbf{\Sigma} = \begin{pmatrix} 1 & \omega c^{-1}\left(\Gamma\left(c+1/2\right)/\Gamma\left(c\right)\right)^2 \\ \omega c^{-1}\left(\Gamma\left(c+1/2\right)/\Gamma\left(c\right)\right)^2 & 1 \end{pmatrix},$$

where $|\omega| \leq 1$ by positive semidefiniteness of $\mathbf{\Omega}$. It follows that for fixed $c$, the largest possible value of $|\rho|$ is $\left(\Gamma\left(c+1/2\right)/\Gamma\left(c\right)\right)^2/c$.

### *C.5 Kurtosis of $\beta_j$ Under an Unstructured SNG Prior*

When $s_j^2 \sim \text{gamma}(c,c)$, we have:

$$\begin{aligned}
\int_0^\infty s_j^4 \frac{c^c}{\Gamma\left(c\right)}\left(s_j^2\right)^{c-1}\exp\left\{-cs_j^2\right\} ds_j^2 &= \frac{c^c}{\Gamma\left(c\right)}\int_0^\infty \left(s_j^2\right)^{c+2-1}\exp\left\{-cs_j^2\right\} ds_j^2 \\
&= \left(\frac{c^c\Gamma\left(c+2\right)}{c^{c+2}\Gamma\left(c\right)}\right) \\
&= c^{-2}\left(\Gamma\left(c+2\right)/\Gamma\left(c\right)\right) \\
&= \left(c+1\right)/c.
\end{aligned}$$

A standard normal random variable has $\mathbb{E}\left[z_j^4\right] = 3$. It follows that

$$\mathbb{E}\left[\beta_j^4\right]/\mathbb{E}\left[\beta_j^2\right]^2 = 3\left(c+1\right)/c.$$

### *C.6 Expectation of $s$ under a unit variance stable prior*

This is a little challenging because working with the stable distribution directly is difficult. However, given knowledge of normal moments and the marginal distribution of $\beta$, we can "back out" $\mathbb{E}\left[s\right]$. Let $\beta = sz$, where $z$ is a standard normal random variable and $1/s^2$ has an $\alpha$-stable distribution on the positive real line. When the stable prior is parametrized to yield $\mathbb{E}\left[\beta^2\right] = 1$, we have:

$$p_\beta\left(\beta|q\right) = \left(\frac{q}{2}\right)\sqrt{\frac{\Gamma\left(3/q\right)}{\Gamma\left(1/q\right)^3}}\exp\left\{-\left(\frac{\Gamma\left(3/q\right)}{\Gamma\left(1/q\right)}\right)^{q/2}|\beta|^q\right\}.$$

Now, to determine $\mathbb{E}[s]$, note that $\mathbb{E}[|\beta|] = \mathbb{E}[s]\,\mathbb{E}[|z|]$. We have

$$\begin{aligned}\mathbb{E}[|\beta|] &= 2\int_0^\infty \beta p_\beta(\beta|q)\,d\beta \\ &= q\sqrt{\frac{\Gamma(3/q)}{\Gamma(1/q)^3}}\int_0^\infty \beta\exp\left\{-\left(\frac{\Gamma(3/q)}{\Gamma(1/q)}\right)^{q/2}\beta^q\right\}d\beta \\ &= \left(\sqrt{\frac{\Gamma(3/q)}{\Gamma(1/q)^3}}\right)\left(\frac{\Gamma(2/q)\,\Gamma(1/q)}{\Gamma(3/q)}\right) \\ &= \frac{\Gamma(2/q)}{\sqrt{\Gamma(1/q)\,\Gamma(3/q)}}.\end{aligned}$$

According to Wikipedia, $\mathbb{E}[|z|] = \sqrt{2/\pi}$. It follows that

$$\mathbb{E}[s] = \sqrt{\frac{\pi}{2}}\left(\frac{\Gamma(2/q)}{\sqrt{\Gamma(1/q)\,\Gamma(3/q)}}\right).$$

### *C.7 Maximum correlation for bivariate $\beta$ under SPB prior*

Consider $\boldsymbol{\Omega} = (1-\omega)\,\boldsymbol{I}_2 + \omega \mathbf{1}_2\mathbf{1}_2'$. We have:

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & \omega\left(\frac{\pi}{2}\right)\left(\frac{\Gamma(2/q)}{\sqrt{\Gamma(1/q)\Gamma(3/q)}}\right)^2 \\ \omega\left(\frac{\pi}{2}\right)\left(\frac{\Gamma(2/q)}{\sqrt{\Gamma(1/q)\Gamma(3/q)}}\right)^2 & 1 \end{pmatrix},$$

where $|\omega| \leq 1$ by positive semidefiniteness of $\boldsymbol{\Omega}$. It follows that for fixed $q$, the largest possible value of $|\rho|$ is $\left(\frac{\pi}{2}\right)\left(\frac{\Gamma(2/q)}{\sqrt{\Gamma(1/q)\Gamma(3/q)}}\right)^2$.

### *C.8 Proofs of Propositions 2.2 and 2.4*

First, we prove the following lemma:

**Lemma C.8.1** *For $0 < c < 1/2$, $\int_0^\infty (s^2)^{c-1/2-1} exp\{-cs^2\}\,ds^2 = +\infty$.*

We can break the integral into two nonnegative components, as follows:

$$\int_0^\infty (s^2)^{c-1/2-1}\exp\{-cs^2\}\,ds^2 = \int_0^{1/c} (s^2)^{c-1/2-1}\exp\{-cs^2\}\,ds^2 + \int_{1/c}^\infty (s^2)^{c-1/2-1}\exp\{-cs^2\}\,ds^2.$$

Now let's examine the first component. When $s^2 < 1/c$, $\exp\{-cs^2\} \geq 1 - cs^2$ and

$$\begin{aligned}
\int_0^{1/c} (s^2)^{c-1/2-1} \exp\left\{-cs^2\right\} ds^2 &\geq \int_0^{1/c} (s^2)^{c-1/2-1} \left(1 - cs^2\right) ds^2 \\
&= \int_0^{1/c} (s^2)^{c-1/2-1} - c \int_0^{1/c} (s^2)^{c-1/2} ds^2 \\
&= \frac{(1/c)^{c-1/2}}{c-1/2} - \frac{1}{c-1/2} \lim_{a\to 0} a^{c-1/2} - c\frac{(1/c)^{c-1/2+1}}{c-1/2+1} + \\
&\quad \frac{c}{c-1/2+1} \underbrace{\lim_{a\to 0} a^{c-1/2+1}}_{=0 \text{ for } c>0} \\
&= \begin{cases} +\infty & 0 < c < 1/2 \\ \frac{(1/c)^{c-1/2}}{c-1/2} - \frac{1}{c-1/2}\mathbb{1}_{\{c=1/2\}} - \frac{(1/c)^{c-1/2}}{c-1/2+1} & c \geq 1/2 \end{cases}
\end{aligned}$$

**Proof of Proposition 2.2**

Now we can evaluate the marginal density $p_{\boldsymbol{\beta}}(\boldsymbol{b}|c, \boldsymbol{\Omega})$ when $b_j = 0$ for some $j \in \{1, \ldots, p\}$. Without loss of generality, set $\beta_1 = 0$. Letting $\boldsymbol{b}_{-1}$ and $\boldsymbol{s}_{-1}$ refer to the vectors $\boldsymbol{b}$ and $\boldsymbol{s}$ each with the first element removed and $(\boldsymbol{\Omega}^{-1})_{-1,-1}$ be the matrix $\boldsymbol{\Omega}^{-1}$ with the first row and column removed. For $\boldsymbol{b} = (0, \boldsymbol{b}_{-1})$,

$$\begin{aligned}
p_{\boldsymbol{\beta}}\left((0, \boldsymbol{b}_{-1})\,|c, \boldsymbol{\Omega}\right) &\propto \int \frac{\left(s_i^2\right)^{c-1}}{\prod_{i=1}^p s_i} \exp\left\{-\frac{1}{2}\left(\boldsymbol{b}' \text{diag}\left\{\mathbf{1}/\boldsymbol{s}\right\} \boldsymbol{\Omega}^{-1} \text{diag}\left\{\mathbf{1}/\boldsymbol{s}\right\} \boldsymbol{b}\right) - \sum_{j=1}^p cs_j^2\right\} d\boldsymbol{s} \\
&= \int \left(\prod_{j=2}^p \left(s_j^2\right)^{c-1/2-1}\right) \exp\left\{-\frac{1}{2}\left(\boldsymbol{b}_{-1}' \text{diag}\left\{\mathbf{1}/\boldsymbol{s}_{-1}\right\} \left(\boldsymbol{\Omega}^{-1}\right)_{-1,-1} \text{diag}\left\{\mathbf{1}/\boldsymbol{s}_{-1}\right\} \boldsymbol{b}_{-1}\right) - \sum_{j=2}^p cs_j^2\right\} \\
&\quad \underbrace{\int_0^\infty \left(s_1^2\right)^{c-1/2-1} \exp\left\{-cs_1^2\right\} ds_1}_{(*)}\, d\boldsymbol{s}_{-1}.
\end{aligned}$$

Applying Lemma C.8.1, the term denoted by $(*)$ evaluates to $+\infty$ for every value of $\boldsymbol{s}_{-1}$.

**Proof of Proposition 2.4**

Proposition 2.4 follows from Proposition 2.2 and Lemma C.8.1. For $\boldsymbol{\beta} = (0, \boldsymbol{\beta}_{-1})$,

$$p_{\beta_1}(0|c, \boldsymbol{\Omega}) \propto \int \frac{\left(s_i^2\right)^{c-1}}{\prod_{i=1}^p s_i} \exp\left\{-\frac{1}{2}\left(\boldsymbol{\beta}' \text{diag}\left\{\mathbf{1}/\boldsymbol{s}\right\} \boldsymbol{\Omega}^{-1} \text{diag}\left\{\mathbf{1}/\boldsymbol{s}\right\} \boldsymbol{\beta}\right) - \sum_{j=1}^p c s_j^2\right\} d\boldsymbol{s} d\boldsymbol{\beta}_{-1}$$

$$= \int \left(\prod_{j=2}^p \left(s_j^2\right)^{c-1/2-1}\right) \exp\left\{-\frac{1}{2}\left(\boldsymbol{\beta}_{-1}' \text{diag}\left\{\mathbf{1}/\boldsymbol{s}_{-1}\right\} \left(\boldsymbol{\Omega}^{-1}\right)_{-1,-1} \text{diag}\left\{\mathbf{1}/\boldsymbol{s}_{-1}\right\} \boldsymbol{\beta}_{-1}\right) - \sum_{j=2}^p c s_j^2\right\}$$
$$\underbrace{\int_0^\infty \left(s_1^2\right)^{c-1/2-1} \exp\left\{-c s_1^2\right\} ds_1}_{(*)} d\boldsymbol{s}_{-1} d\boldsymbol{\beta}_{-1}.$$

Again, $(*)$ evaluates to $+\infty$ for any value of $\boldsymbol{s}_{-1}$ and does not depend at all on $\boldsymbol{\beta}_{-1}$.

### *C.9 Alternative Approaches to Posterior Computation*

Gibbs sampling is very straightforward when the SPN prior is used, as the full conditional distributions $p_{\boldsymbol{s}|\boldsymbol{z}}(\boldsymbol{s}|\boldsymbol{z}, \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\Psi}, \sigma^2)$ and $p_{\boldsymbol{z}|\boldsymbol{s}}(\boldsymbol{z}|\boldsymbol{s}, \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\Omega}, \sigma^2)$ are multivariate normal distributions. Exact full conditional distributions are given in the appendix.

Posterior inference for the SNG and SPB priors is more challenging than posterior inference for the SPN prior, as there is not an obvious stochastic representation of $\boldsymbol{\beta}$ that yields tractable full conditional distributions for elements of $\boldsymbol{s}$. Under the SNG and SPB priors we have $\boldsymbol{\beta}|\boldsymbol{s} \sim \text{normal}(\mathbf{0}, \text{diag}\{\boldsymbol{s}\}\boldsymbol{\Omega}\text{diag}\{\boldsymbol{s}\})$, so the full conditional distribution of $\boldsymbol{\beta}$ is a multivariate normal distribution. Letting $p_{s_j}(s_j|\theta)$ refer to the prior density of $\boldsymbol{s}$ under an SPB or SNG prior and letting $\theta$ refer to the corresponding prior parameter, $c$ or $q$, the full conditional distribution of $\boldsymbol{s}$ is

$$p_{\boldsymbol{s}}(\boldsymbol{s}|\boldsymbol{\beta}, \boldsymbol{\Omega}, \theta) \propto \underbrace{\exp\left\{-\frac{1}{2}\sum_{j=1}^p \sum_{j'=1, j' \neq j}^p \frac{\beta_j \beta_{j'}}{\omega^{jj'} s_j s_{j'}}\right\}}_{(*)} \underbrace{\prod_{j=1}^p \left(\frac{p_{s_j}(s_j|\theta)}{s_j}\right) \exp\left\{-\frac{1}{2}\left(\frac{1}{s_j^2}\left(\frac{\beta_j^2}{\omega^{jj}}\right)\right)\right\}}_{(**)},$$

where $\omega^{jj}$ is the entry in the $i$-th row and $j$-th column of of $\boldsymbol{\Omega}^{-1}$. This is not a standard distribution under either the SNG or SPB priors, either for the entire vector $\boldsymbol{s}$ or for each element of $s_j$ conditional on the remaining elements $\boldsymbol{s}_{-j}$. However, the term $(*)$ is a simple function of elements of $\boldsymbol{s}$ that is monotonic for in each $s_j$ conditional on the remaining

elements $\boldsymbol{s}_{-j}$, and the term $(**)$ is a product of $p$ densities that are straightforward to simulate from. Under the SNG prior, $(**)$ is a product of $p$ generalized inverse gaussian densities and under the SPB prior, $(**)$ is a product of $p$ exponentially tilted positive $\alpha$-stable densities. Accordingly, the auxiliary variable method introduced by Damien et al. (1999) can be used. We define a new variable $u > 0$, such that

$$p_{\boldsymbol{s}^2,u}\left(\boldsymbol{s}^2, u|\boldsymbol{\beta}, \boldsymbol{\Omega}, \theta\right) \propto \mathbb{1}_{\left\{u<\exp\left\{-\frac{1}{2}\sum_{j=1}^{p}\sum_{j'=1,j'\neq j}^{p}\frac{\beta_j\beta_{j'}}{\omega^{jj'}s_j s_{j'}}\right\}\right\}}\prod_{j=1}^{p}\left(\frac{p_{s_j}\left(s_j|\theta\right)}{s_j}\right)\exp\left\{-\frac{1}{2}\left(\frac{1}{s_j^2}\left(\frac{\beta_j^2}{\omega^{jj}}\right)\right)\right\}.$$

We can simulate from the full conditional distribution $p\left(\boldsymbol{s}|\boldsymbol{\beta}, \boldsymbol{\Omega}, c\right)$ by drawing a value of $u$ at the current value of $\boldsymbol{s}$ from a uniform distribution on $(0, \exp\{-\frac{1}{2}\sum_{j=1}^{p}\sum_{j'=1,j'\neq j}^{p}\frac{\beta_j\beta_{j'}}{\omega^{jj'}s_j s_{j'}}\})$ and then drawing each value of $s_j$ at the current value of $u$ and remaining values $\boldsymbol{s}_{-j}$ from a truncated generalized inverse Gaussian distribution or exponentially tilted positive $\alpha$-stable distribution that satisfies the constraint $u < \exp\{-\frac{1}{2}\sum_{j=1}^{p}\sum_{j'=1,j'\neq j}^{p}\frac{\beta_j\beta_{j'}}{\omega^{jj'}s_j s_{j'}}\}$. Importantly, this constraint can be rewritten as a bound on each $s_j$ at the current value of $u$ and remaining values $\boldsymbol{s}_{-j}^2$ as follows:

$$\underbrace{s_j\left(\log\left(u\right) + \frac{1}{2}\sum_{j'=1,j'\neq j}^{p}\sum_{j''=1,j''\neq j,j'}^{p}\frac{\beta_{j'}\beta_{j''}}{\omega^{j'j''}s_{j'}s_{j''}}\right)}_{a_j} < \underbrace{-\frac{1}{2}\sum_{j'=1,j'\neq j}^{p}\frac{\beta_j\beta_{j'}}{\omega^{jj'}s_{j'}}}_{d_j}.$$

The constraint on $s_j$ is $s_j < d_j/a_j$ if $a_j > 0$ and $s_j > d_j/a_j$ otherwise. The ability to obtain explicit constraints on each $s_j$ is advantageous because it can yield efficiency gains if a fast method for simulating from a left- or right-truncated generalized inverse Gaussian distribution or exponentially tilted positive $\alpha$-stable distribution is available. We are able to construct such a method for simulating from truncated generalized inverse Gaussian distributions and describe it in the appendix. Sampling from truncated exponentially tilted positive $\alpha$-stable distributions is more challenging, as the positive $\alpha$-stable density does not have a closed form. For this paper, we take the naive approach of simulating from a truncated exponentially tilted positive $\alpha$-stable distribution until the constraint is satisfied.

### *C.10 Univarate Slice Sampling*

In several places in this paper, we use a univariate slice sampling algorithm described in Neal (2003) to simulate values of a random variable $x \in [a, b]$ with density proportional to $\exp\{g(x)\}$. Given a previous value $\tilde{x}$, we can simulate a new value $\tilde{x}'$ from a full conditional using univariate slice sampling as follows:

1. Draw $e \sim \exp(1)$.

2. Draw $d \sim \text{uniform}(a, b)$.

   - If $g(\tilde{x}) - e \leq g(d)$, set $\tilde{x}' = d$.
   - If $g(\tilde{x}) - e > g(d)$:
     (a) If $d < \tilde{x}$, set $a = d$, else if $d \geq \tilde{x}$ set $b = d$.
     (b) Return to 3.

### *C.11 Simulation from Full Conditional Distribution for* $\delta_j$

When using the SPB prior, we have $s_j^2 \overset{d}{=} \Gamma(1/(2\alpha))\, \xi_j^{\frac{1-\alpha}{\alpha}} / (2\Gamma(3/(2\alpha)))$, where $\xi_j \sim \text{gamma}((1+\alpha)/(2\alpha), f(\delta_j|\alpha))$ and $p_{\delta_j}(\delta_j|\alpha) \propto f(\delta_j|\alpha)^{\frac{\alpha-1}{2\alpha}}$ on $(0, \pi)$. Suppose $s_j^2$ is fixed. Then $\xi_j = \left(2\Gamma(3/(2\alpha))\, s_j^2/\Gamma(1/(2\alpha))\right)^{\frac{\alpha}{1-\alpha}}$. Then we can write the full conditional distribution of $\delta_j$ as

$$\begin{aligned} p(\delta_j|\xi_j, q) \propto & f(\delta_j|\alpha)^{\frac{1+\alpha}{2\alpha}} \exp\{-f(\delta_j|\alpha)\,\xi_j\}\, f(\delta_j|\alpha)^{\frac{\alpha-1}{2\alpha}} \\ = & f(\delta_j|\alpha) \exp\{-f(\delta_j|\alpha)\,\xi_j\}. \end{aligned}$$

Apply the univariate slice sampling algorithm described in Section C.10 using $g(x) = \log(f(x|\alpha)) - f(x|\alpha)\,\xi_j$ and initial values $a = 0$ and $b = \pi$.

### *C.12 Simulation from Full Conditional Distribution for* $\rho$

In the real data application, I consider the setting where $\boldsymbol{B}$ and $\boldsymbol{S}$ are $p_1 \times p_2$ matrices and $\boldsymbol{\beta} = \text{vec}(\boldsymbol{B}/\boldsymbol{S})$ has covariance matrix $\mathbb{V}[\boldsymbol{\beta}/\boldsymbol{s}] = \boldsymbol{\Omega}_2 \otimes \boldsymbol{\Omega}_1$, where $\boldsymbol{\Omega}_1$ is depends on a single

autoregressive parameter, $\rho$, with $\omega_{1,ij} = \rho^{|i-j|}$. The full conditional distribution of $\rho$ is:

$$\begin{aligned} p_\rho(\rho|-) &\propto |\boldsymbol{\Omega}_1|^{-\frac{p_2}{2}} \exp\left\{-\frac{1}{2}\text{tr}\left(\boldsymbol{\Omega}_1^{-1}(\boldsymbol{B}/\boldsymbol{S})\boldsymbol{\Omega}_2^{-1}(\boldsymbol{B}/\boldsymbol{S})\right)\right\} \\ &= \left(1-\rho^2\right)^{-\frac{(p_1-1)p_2}{2}} \exp\left\{-\frac{1}{2}\text{tr}\left(\boldsymbol{\Omega}_1^{-1}(\boldsymbol{B}/\boldsymbol{S})\boldsymbol{\Omega}_2^{-1}(\boldsymbol{B}/\boldsymbol{S})\right)\right\} \pi_\rho(\rho), \end{aligned}$$

where elements of $\boldsymbol{\Omega}_1^{-1}$ are given by

$$\begin{aligned} \omega_1^{11} &= \omega_1^{pp} = \left(1-\rho^2\right)^{-1} \\ \omega_1^{jj} &= \left(1-\rho^2\right)^{-1}\left(1+\rho^2\right) && j \neq 1, j \neq p \\ \omega_1^{jk} &= -\rho\left(1-\rho^2\right)^{-1} && |j-k| = 1 \end{aligned}$$

and $\pi_\rho(\rho)$ is the density of an assumed prior distribution for $\rho$. We assume a uniform prior on $(-1,1)$, i.e. beta$(1,1)$ prior for $(\rho+1)/2$.

### *C.13 Prior Conditional Distribution of $s_j^2|\delta_j$ and $s_j|\delta_j$ under SPB Prior*

When using the SPB prior, we have $s_j^2 \overset{d}{=} \Gamma(1/(2\alpha))\,\xi_j^{\frac{1-\alpha}{\alpha}}/(2\Gamma(3/(2\alpha)))$, where $\xi_j \sim$ gamma$((1+\alpha)/(2\alpha), f(\delta_j|\alpha))$ and $p_{\delta_j}(\delta_j|\alpha) \propto f(\delta_j|\alpha)^{\frac{\alpha-1}{2\alpha}}$ on $(0,\pi)$. We want to compute the prior conditional distribution $\pi_{s_j^2}\left(s_j^2|\delta_j,\alpha\right)$ for use in the elliptical slice sampler for $\boldsymbol{s}$. We have:

$$\begin{aligned} s_j^2 \overset{d}{=}& \Gamma(1/(2\alpha))\,\xi_j^{\frac{1-\alpha}{\alpha}}/(2\Gamma(3/(2\alpha))) \\ =& \left(\left(\Gamma(1/(2\alpha))/(2\Gamma(3/(2\alpha)))\right)^{\frac{\alpha}{1-\alpha}}\xi_j\right)^{\frac{1-\alpha}{\alpha}} \\ \overset{d}{=}& \tilde{\xi}_j^{\frac{1-\alpha}{\alpha}},\ \tilde{\xi}_j \sim \text{gamma}\left(\frac{1+\alpha}{2\alpha}, \left(\frac{2\Gamma\left(\frac{3}{2\alpha}\right)}{\Gamma\left(\frac{1}{2\alpha}\right)}\right)^{\frac{\alpha}{1-\alpha}} f(\delta_j|\alpha)\right) \end{aligned}$$

We can find the density of $s_j^2$ via change of variables. We have

$$\tilde{\xi}_j = \left(s_j^2\right)^{\frac{\alpha}{1-\alpha}},\ d\tilde{\xi}_j = \left(\frac{\alpha}{1-\alpha}\right)\left(s_j^2\right)^{\frac{\alpha}{1-\alpha}-1} ds_j^2.$$

Then the prior conditional distribution is

$$\pi_{s_j^2}\left(s_j^2|\delta_j,\alpha\right) \propto \left(\left(s_j^2\right)^{\frac{\alpha}{1-\alpha}}\right)^{\frac{1+\alpha}{2\alpha}-1}\left(s_j^2\right)^{\frac{\alpha}{1-\alpha}-1}\exp\left\{-\left(\frac{2\Gamma\left(\frac{3}{2\alpha}\right)s_j^2}{\Gamma\left(\frac{1}{2\alpha}\right)}\right)^{\frac{\alpha}{1-\alpha}} f\left(\delta_j|\alpha\right)\right\}$$
$$= \left(s_j^2\right)^{\frac{1+\alpha}{2(1-\alpha)}-1}\exp\left\{-\left(\frac{2\Gamma\left(\frac{3}{2\alpha}\right)s_j^2}{\Gamma\left(\frac{1}{2\alpha}\right)}\right)^{\frac{\alpha}{1-\alpha}} f\left(\delta_j|\alpha\right)\right\}.$$

In practice, we end up working with $s_j$, so once more

$$\pi_{s_j}\left(s_j|\delta_j,\alpha\right) \propto s_j^{\frac{1+\alpha}{1-\alpha}-1}\exp\left\{-\left(\frac{2\Gamma\left(\frac{3}{2\alpha}\right)s_j^2}{\Gamma\left(\frac{1}{2\alpha}\right)}\right)^{\frac{\alpha}{1-\alpha}} f\left(\delta_j|\alpha\right)\right\}.$$

### *C.14 Prior Conditional Distribution of $s_j$ under SNG Prior*

Under the SNG prior, we have

$$\pi_{s_j^2}\left(s_j^2|c\right) \propto \left(s_j^2\right)^{c-1}\exp\left\{-s_j^2 c\right\}.$$

In practice, we end up working with $s_j$, which has the density

$$\pi_{s_j}\left(s_j|c\right) \propto \left(s_j\right)^{2c-1}\exp\left\{-s_j^2 c\right\}.$$

# VITA

<Insert information about myself here.>