UMBC

DISSERTATION APPROVAL SHEET

Title of Dissertation: Statistical Inference on High Dimensional Normal Mean Under Linear Inequality Constraints and Efficient Integration of Data in Meta-Analysis

Name of Candidate: Neha Agarwala

Doctor of Philosophy, 2022

Graduate Program: Statistics

Dissertation and Abstract Approved:

Anindya Roy

Anindya Roy

Department of Mathematics and Statistics

Dept of Statistics, Seoul National University

3/29/2022 | 3:11:04 PM EDT

Junyong Park

Junyong Park

Professor

Department of Statistics, Seoul National Univ

3/29/2022 | 4:45:11 오후 EDT

NOTE: *The Approval Sheet with the original signature must accompany the thesis or dissertation. No terminal punctuation is to be used.

ABSTRACT

Title of dissertation: Statistical Inference on High Dimensional Normal Mean Under Linear Inequality Constraints and Efficient Integration of Data in Meta-Analysis

Neha Agarwala, Doctor of Philosophy, 2022

Dissertation directed by: Professor Anindya Roy
Department of Mathematics and Statistics
University of Maryland, Baltimore County

Professor Junyong Park
Department of Statistics
Seoul National University

In this dissertation, we provide a framework for incorporating linear inequality parameter constraints in estimation and hypothesis testing involving high dimensional normal means. Modern statistical problems often involve such linear inequality constraints on model parameters. Ignoring natural parameter constraints usually results in less efficient statistical procedures. To this end, we define a notion of 'sparsity' for such restricted sets using lower-dimensional features (Chapter 2). We allow our framework to be flexible so that the number of restrictions may be higher than the number of parameters. We show that the proposed notion of sparsity agrees with the usual notion of sparsity in the unrestricted case and proves the validity of the proposed definition as a measure of sparsity. The proposed sparsity measure also allows us to generalize popular priors for sparse vector estimation to

the constrained case. We also explore the properties of some of these priors for the non-negativity restrictions (Chapter 1). Along with Bayesian estimation of the constrained mean, we also consider the classical one-sided normal mean testing problem where the null hypothesis of a zero mean vector is tested against the alternative that all the components are non-negative and at least one is positive (Chapter 3). It is unlikely for a single test to perform equally well for dense and sparse parameter configuration in high dimension. We develop a computationally efficient omnibus test with reasonable power for the entire spectrum of alternatives. Finally, we propose a meta-analysis approach for combining treatment effects across aggregate data (AD) and individual patient data (IPD) under a generalized linear model structure (Chapter 4). Often for some studies with AD, the associated IPD may be available, albeit at some extra effort or cost to the analyst. For many different models, design constraints under which the AD estimators are the IPD estimators, and hence fully efficient, are known. For such models, we advocate a selection procedure that chooses AD studies over IPD studies to force least departure from design constraints using the proposed combination method and hence ensures an efficient combined AD and IPD estimator.

Statistical Inference on High Dimensional Normal Mean Under Linear Inequality Constraints and Efficient Integration of Data in Meta-Analysis

by

Neha Agarwala

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, Baltimore County in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2022

Advisory Committee:
Professor Anindya Roy, Chair/Advisor
Professor Junyong Park, Co-Advisor
Professor Thomas Mathew
Professor Yaakov Malinovsky
Professor Jinglai Shen
Dr. Emanuel Ben-David

Dedication

To my Ma, Papa, Tauji and Prosenjit for their endless love, support and encouragement.

Acknowledgments

As my Ph.D marathon comes to an end, I want to express my most profound appreciation to the people who have contributed in one way or another and supported me in the last five years.

First, I want to extend my sincerest gratitude to my advisors, Dr. Anindya Roy and Dr. Junyong Park, for making my Ph.D. journey enjoyable and satisfactory. Dr. Roy is the epitome of brilliance and kindness. When the world, including me, found itself amidst a formidable pandemic and its unprecedented challenges, Dr. Roy made the transition to the virtual world of meetings and discussions completely seamless. His unwavering support, guidance, and immense knowledge in this field have made Ph.D. an inspiring experience for me. He is a great mentor who has always been available to discuss research and encouraged me to learn things beyond my Ph.D. work. He has nurtured my curiosity and help me develop as a successful researcher. I would also like to thank Dr. Park for the difficult questions during our regular meetings. His insights and scientific approach have genuinely helped me to complete my thesis.

Besides my advisors, I would like to thank Dr. Thomas Mathew, Dr. Yaakov Malinovsky, Dr. Jinglai Shen, and Dr. Emanuel Ben-David for serving as my thesis committee members and their suggestions. I want to pay special regard to Dr. Bimal Sinha and Dr. Yi Huang for their constant encouragement and advice. I want to take this opportunity to thank my other collaborators outside UMBC, Dr. Nilanjan Chatterjee, Dr. Jin Jin, Dr. Elena Rantou, Dr. Sungwoo Choi, Dr.

Ciprian Crainiceanu, and Dr. Andrew Leurox for introducing me to some exciting research in Biostatistics during the last few summers. Working closely on public health issues, which affect the entire humanity, was a motivating and enriching learning experience for me.

Next, I want to take a moment to acknowledge the members of the department administration, finance, and information system, who were a crucial part of this journey. In particular, I want to thank Janet for smoothly steering me through any required paperwork and for her prompt response to my many emails; Marshall and Maggie for assisting with any administrative-financial concerns; Boris for helping me with any cluster-related issues. I wish to thank my fellow Ph.D. mates, Michael, Naadesri, Ryan, Gaurab, Vahid, and Reetam, for joining me in the much-needed coffee breaks and our countless discussions. They have made my time at our office fun and educative.

Ph.D. is a marathon with its fair share of challenges. So, it goes without saying that I owe my Ph.D. to my family and friends, who are my pillars of strength and have made the ride easier and enjoyable. I want to start by thanking my friends in Baltimore who were responsible for the fun part of the journey. Thank you, Anindya, Debangan, Sayan, Soumya, Arunima, Ayan, Diptavo, Sayantan, and Prosenjit, for the myriad of weekend hangouts and trips, library sessions, cooking and eating events. My heartfelt gratitude to Prakriti, Goonj, and Antardeep for always being one text away and cheering me up. I am also thankful to my friends Apurv, Gargi, Sourav, and Indrayudh for visiting and staying connected. Some other friends who deserve special mention are Kunal, Ipshita, Parichoy, Tushita,

Roshni, Sayantan, Arijit, Debadrita, Sayak, and Disha, Ikbal, Sumit, Shatabdi, Rahul, Gaurush, Apurv among others.

I am really thankful to Prosenjit, my invaluable partner in crime, for being a constant source of inspiration. Thank you for keeping me safe from grief and failures and providing a life's worth of adventures and memories.

Lastly, I am grateful to my dear family for their unconditional love and support - my parents, Ranju and Dilip, for always believing in me; my brother Pankaj for staying home with me whenever I would visit India despite his very busy schedule and my sister Monika for being my support system. Lastly, I owe it to my eldest uncle, Shankar, for having the vision of sending us outside our small town for higher studies and my cousin Gautam for laying the path of learning for us. This thesis would not have been possible without all of you.

Finally, my apologies if I have inadvertently left someone out. Thank you all.

Contents

List of Tables

List of Figures

Chapter 0: Introduction

We first consider the classical normal mean estimation problem where the mean vector lies in a non-negative orthant for a n dimensional independent normal observation with common variance in chapter 1. We study the behavior and risk properties of Bayesian estimators under two popular priors, the horseshoe prior and Strawderman-Berger prior, originally developed in the unrestricted mean vector estimation regime and then restrict the distribution of prior to satisfy the parameter constraint. The performance of posterior mean based on the horseshoe prior and the posterior mean and posterior median based on Strawderman-Berger prior is compared with the maximum likelihood estimator, numerically under different sparsity configurations.

We then consider estimation of parameters restricted to a closed polyhedral cone. A polyhedral cone is defined as the solution set of a homogeneous set of linear inequalities (chapter 2). A special case is when the number of restrictions maybe higher than the number of parameters. One such situation arise in estimation of monotone curve using a non parametric approach e.g. splines. In these high dimensional problems one usually seeks a 'sparse' solution. We define a notion of sparsity for such conic restrictions using lower dimensional facets of the cone.

We propose weak and strong sparse estimators of the constrained parameter vector by invoking a continuous shrinkage prior or a spike-slab prior through the higher dimensional non-negative orthant representation of the polyhedral cone.

Next, we focus on constrained hypothesis testing problems for gaussian high dimensional mean vectors in chapter 3. Given a n dimensional normal distribution with covariance Σ, we consider the classical one-sided normal mean testing problem that all the components of the mean are zero against the alternative that all the components are non-negative and at least one is positive. It is unlikely for a single test to perform equally well for dense and sparse parameter configuration in such high dimension. Our goal is to develop a computationally efficient omnibus test for the entire spectrum of alternatives.

In chapter 4, we take a turn and look at a meta-analysis problem. Often both Aggregate Data (AD) studies and Individual Participant Data (IPD) studies are available for specific treatments. Combining these two sources of data could improve the overall meta-analytic estimates of treatment effects. Moreover, often for some studies with AD, the associated IPD maybe available, albeit at some extra effort or cost to the analyst. We propose a method for combining treatment effects across trials when the response is from the exponential family of distribution and hence a generalized linear model structure can be used. We consider the case when treatment effects are fixed and common across studies. Using the proposed combination method, we evaluate the wisdom of choosing AD when IPD is available by studying the relative efficiency of analyzing all IPD studies versus combining various percentages of AD and IPD studies. For many different models, design constraints

under which the AD estimators are the IPD estimators, and hence fully efficient, are known. For such models we advocate a selection procedure that chooses AD studies over IPD studies in a manner that force least departure from design constraints and hence ensures an efficient combined AD and IPD estimator.

Chapter 1: Horseshoe and Strawderman-Berger Estimator for Constrained Normal Means

1.1 Introduction

Traditional statistical theory has mostly focused on methods developed for large samples and a small number of features. The modern scientific world, however, is moving fast towards the regime of high dimensional data. In the high dimensional setting, often one deals with the case when only few variables are relevant. Thus, it has become increasingly important to identify true signals as the data tends to be sparse. Probably the most common of such high dimensional sparse estimation problems is estimation of the mean of a normal distribution when sample size is small compared to the dimension. It is the proverbial needle in a haystack problem that has received much attention in the literature. The setting of the problem is simple. Given data $y_1, \ldots, y_n$, arising independently from the model

$$y_i|\mu_i, \sigma^2 \sim N(\mu_i, \sigma^2),$$

one wishes to estimate the entire vector $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)$. Of course, given that there are only n independent observations for $(n + 1)$ unknown parameters, additional assumptions are needed for meaningful estimation of the mean vector. Usually

some level of sparsity is assumed for the true mean vector. Both Bayesian and frequentist estimators have been developed for this problem, the most well known being the shrinkage estimators starting with James and Stein (1961), thresholding estimators starting with Donoho and Johnston (1994), penalized estimators such as Lasso (Tibshirani, 1996), SCAD (Fan and Li, 2001) and many other variants of them [1–4].

In the Bayesian setting, popular approaches include using the spike and slab priors and continuous shrinkage priors for sparse mean estimation. Formulation of sparse mean vector scenarios as a combination of two regimes where the mean values are zero or arising from a measure which allows for possibly large values naturally leads to a mixture prior of the form

$$p(\mu) = p\delta_0 + (1-p)g(\mu).$$

The point mass p as $\mu = 0$ is the *spike* and the probability density $g(\cdot)$ allowing μ to take possibly large non-zero values is the *slab*. Mitchell and Beauchamp (1988) considered it in the context of variable selection in Gaussian regression [5]. Since then such priors have gained popularity in many contexts including variable selection, covariance matrix estimation, false discovery rate estimation. Many authors have advocated the use of such point mixture priors for normal mean estimation. Strawderman-Berger (SB) prior (Strawderman and Berger, 1996) explicitly considered in this article is an example of such a spike-and-slab prior in a hierarchical setting where the hyper-parameters governing the slab $g(\cdot)$ are allowed to change according to some prior for each μ_i [6]. Specifically, they propose the following

model

$$\begin{aligned} \mu_i|\tau,\lambda_i &\sim N(0,\tau^2\lambda_i^2), \\ p(\lambda_i) &\propto \lambda_i(1+\lambda_i^2)^{1/2}, \\ p(\tau) &\sim C[\sigma,\sigma]I(\tau>\sigma) \end{aligned}$$

where $C[a,b]$ is the Cauchy density with location and scale equal to a and b, respectively. A version of the spike-slab prior considered recently is the non-local prior recommended by Johnson and Rossell (2010, 2012) where the slab is well separated from the spike at zero [7, 8].

Another class of priors considered for sparse estimation of mean are the shrinkage priors or the global-local priors. Park and Casella (2008) proposed a scale mixture of Gaussian prior that they called the *Bayesian Lasso* [9]. However, these priors do not have sufficient prior mass near zero to work well in the very sparse regime. Carvalho *et al.* (2010) proposed the *horseshoe* (HS) prior defined as

$$\begin{aligned} \mu_i|\tau,\lambda_i &\sim N(0,\tau^2\lambda_i^2), \\ p(\lambda_i) &\propto C[0,1]_+, \\ p(\tau) &\sim \sigma C[0,1]_+ \end{aligned}$$

where $C[0,1]_+$ is the half-Cauchy density, the standard Cauchy truncated to the positive half [10]. The horseshoe prior has only one component as opposed to the two separate components of the spike-and-slab priors but overcomes the deficiency of the Bayesian Lasso in sparse regime by allowing infinite prior density at zero. Being a single component prior, horseshoe type priors are computationally less demanding

than the spike-slab priors.

While full Bayesian analysis is possible, empirical Bayes solutions have also been discussed for the two component mixture priors such as Strawderman-Berger and the single component shrinkage priors such as horseshoe. Empirical Bayes solutions for high-dimensional sparse mean estimation have been also looked at in the literature; see Johnston and Silverman (2004), Brown and Greenshtein (2009) [11, 12].

Often one has prior knowledge on the range of possible values for the mean parameter, such as the parameter is non-negative. We focus on the high dimensional normal means estimation problem when the dimension is large and the mean vector is assumed to be sparse. Let $\boldsymbol{y} = (y_1, \ldots, y_n)' \sim N(\boldsymbol{\mu}, \sigma^2 I)$ where the parameter of interest $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)'$ is assumed to belong to the non-negative orthant $\mathbb{R}^n_+$ where $\mu_i \geq 0$ for all i. For the rest of the manuscript, we denote $\mathbb{R}^n_+$ by $\mathcal{K}_+$.

We start with comparing the performance of sparsity generating spike-and-slab priors such as Strawderman-Berger and shrinkage priors such as horseshoe, when the priors are defined in terms of scale mixtures of truncated normal instead of normal. This straightforward generalization is probably not optimal, particularly if the conic geometry is very different from that of the entire space. However, given its special importance in the applications, we will first focus on the positive orthant. The geometry of the positive orthant is very similar to the unrestricted linear space, but there are subtle differences in estimation due to the constraint and that is what we explore via numerical investigation. In Sections 1.2.1 and 1.2.2, we discuss the Bayes estimators for the Strawderman-Berger and the horseshoe priors

when they are extended to the convex cone case. In Section 1.3 we present results of a numerical experiment comparing the performance of posterior quantities obtained using different priors along with that of the maximum likelihood estimator (MLE) projected to the convex cone followed by some real data analysis in section 1.4. We end with some discussions in Section 1.5.

1.2 Sparse Priors for the Non-negative Orthant

For the horseshoe prior and the Strawderman-Berger prior for the non-negative orthant, we simply replace the normal prior for μ_i with normal truncated to the positive half. To judge the performance of the estimators under different priors in the constrained case, we set forth a list of desirable properties. These are analogous to desirable properties in a sparse mean estimator in the unrestricted case, except adapted to the constrained mean case. For example, one would want the estimators for μ_i to provide considerable shrinkage for small to moderate y_i whereas to leave y_i nearly unperturbed for large positive y_i. In the constrained case, for negative y_i one would expect the estimated mean to be nearly zero, if not exactly zero. The maximum likelihood estimator for μ_i is exactly zero whenever y_i is negative.

1.2.1 Horseshoe prior

The extension for the horseshoe to the positive orthant considered here is then

$$\begin{aligned} \mu_i|\tau,\lambda_i &\sim N(0,\tau^2\lambda_i^2)_+, \\ \lambda_i &\sim C(0,1)_+. \\ \pi(\sigma) &\propto \frac{1}{\sigma}, \\ \tau|\sigma &\sim C(0,\sigma)_+. \end{aligned} \tag{1.1}$$

where $N(\theta,v)_+$ represent a $N(\theta,v)$ truncated from below at 0 and $C(0,1)_+$ represent a standard half-Cauchy distribution on the positive reals. The prior on σ is the standard Jeffrey's and that on τ is the standard half-Cauchy prior with scale equal to σ. One could estimate σ and τ using an Empirical Bayes approach. However, here we use a full Bayesian framework. Carvalho *et al* (2010) described λ_i as the local shrinkage parameter and τ the global shrinkage parameter. For the positive orthant, the horseshoe prior that we are considering is essentially a scale mixture of truncated normals, scale being a function of a common variance component, τ and an individual variance component, λ_i for each μ_i. Conditional on σ, τ and λ_i's, $\mu_i|\boldsymbol{y}$ are independently distributed as

$$\mu_i|\lambda_i,\tau,\sigma,\boldsymbol{y} \sim N(m_i,s_i^2)_+$$

where $m_i = s_i^2\frac{y_i}{\sigma^2}$ and $s_i^2 = [\frac{1}{\sigma^2}+\frac{1}{\tau^2\lambda_i^2}]^{-1}$. Then, we have

$$E(\mu_i|\lambda_i,\tau,\sigma,\boldsymbol{y}) = m_i + \frac{\phi(\frac{-m_i}{s_i})}{1-\Phi(\frac{-m_i}{s_i})}s_i. \tag{1.2}$$

The Bayes estimator of μ_i is given by

$$\hat{\mu}_i = E(\mu_i|\boldsymbol{y}) = E_{\lambda_i,\tau,\sigma|\boldsymbol{y}} E(\mu_i|\lambda_i,\tau,\sigma,\boldsymbol{y}).$$

From the bounds on the Mill's ratio for the standard normal, we know that for $t > 0$,

$$t \;<\; \frac{\phi(t)}{1-\Phi(t)} \;<\; \frac{1+t^2}{t}. \tag{1.3}$$

This implies $E(\mu_i|\lambda_i,\tau,\sigma,\boldsymbol{y}) > 0$ for all y. Also, for $y_i < 0$, $E(\mu_i|\lambda_i,\tau,\sigma,\boldsymbol{y}) < \sigma^2|y_i|^{-1}$. Moreover, for large positive y_i, $E(\mu_i|\lambda_i,\tau,\sigma,\boldsymbol{y}) \approx [1 - \frac{\tau^2\lambda_i^2}{\sigma^2+\tau^2\lambda_i^2}]y_i$. Hence, summarizing we have

Result 1: For the horseshoe prior for the constrained case when the true mean is restricted to the non-negative orthant:

1. $E(\mu_i|\boldsymbol{y}) > 0$ for all $\boldsymbol{y}$.

2. For $y_i < 0$, $E(\mu_i|\boldsymbol{y}) = O(|y_i|^{-1})$.

3. For large positive y_i, $E(\mu_i|\boldsymbol{y}) \approx \mathrm{E}([1 - \frac{\tau^2\lambda_i^2}{\sigma^2+\tau^2\lambda_i^2}]|\lambda_i,\tau,\sigma,\boldsymbol{y})y_i$.

Thus, the posterior mean of μ_i acts as a shrinkage estimator and its behavior is similar to what observed in the unrestricted case.

1.2.2 Strawderman-Berger Prior

The extension of Strawderman-Berger prior for the non-negative orthant puts a truncated normal distribution in place of the usual normal distribution.

$$\begin{aligned} \pi(\mu_i) &= p\delta_o + (1-p)\ N(0, \tau^2\lambda_i^2)_+, \\ \pi(\lambda_i) &\propto \lambda_i(1+\lambda_i^2)^{\frac{3}{2}}, \\ p &\sim \text{Unif}(0,1). \\ \tau|\sigma &\sim C(\sigma,\sigma)\ 1(\tau \geq \sigma), \\ \pi(\sigma) &\propto \frac{1}{\sigma}....eq:SBprior \end{aligned} \tag{1.4}$$

where similar to horseshoe a Jeffrey's prior is specified for σ and the prior on τ is a truncated Cauchy prior with location and scale both equal to σ bounded below at σ. Conditional on $\lambda_i, \tau, p, \sigma$, the posterior distribution of μ_i is a mixture distribution

$$\pi(\mu_i|\lambda_i, \tau, p, \sigma, \boldsymbol{y}) = c(\theta_i, y_i)\ \delta_o + \big(1 - c(\theta_i, y_i)\big)\ N(m_i, s_i^2)_+ \tag{1.5}$$

where

$$c(\theta_i, y_i) = \frac{\frac{p}{\sigma}\phi(\frac{y_i}{\sigma})}{\frac{p}{\sigma}\phi(\frac{y_i}{\sigma}) + \frac{2(1-p)}{l_i}\phi(\frac{y_i}{l_i})\Phi(\frac{m_i}{s_i})}$$

is the posterior probability of $\mu_i = 0$ which acts as local shrinkage, $\theta_i = \{\lambda_i, \tau, \sigma, p\}$ and $l_i^2 = \sigma^2 + \lambda_i^2\tau^2$ for $i = 1, \ldots, n$. Then, we have

$$E(\mu_i|\boldsymbol{\lambda}, \tau, \sigma, p, \boldsymbol{y}) = \big(1 - c(\theta_i, y_i)\big)\left(m_i + \frac{\phi\big(\frac{-m_i}{s_i}\big)}{\Phi\big(\frac{m_i}{s_i}\big)}s_i\right).$$

The Bayes estimator for μ_i is the posterior mean, $E(\mu_i|\boldsymbol{y}) = E_{\boldsymbol{\lambda},\tau,p|\boldsymbol{y}}E(\mu_i|\boldsymbol{\lambda}, \tau, \sigma, p, \boldsymbol{y})$.

Result 2: The following results hold for the posterior mean computed based on the

Strawderman-Berger prior in the constrained case:

1. $\hat{\mu}_i > 0$ since $E(\mu_i|\lambda_i, \tau, p, \sigma, \boldsymbol{y}) > 0$ using the inequality in (1.3).

2. $\hat{\mu}_i$ is non-decreasing in y_i.

3. For large positive y_i, $E(\mu_i|\boldsymbol{y}) \approx \mathrm{E}\big[\big(1 - c(\theta_i, y_i)\big)\big(1 - \frac{\tau^2\lambda_i^2}{\sigma^2+\tau^2\lambda_i^2}\big)|\lambda_i, \tau, \sigma, \boldsymbol{y}\big] y_i$.

Proof. *$\hat{\mu}_i$ is non-decreasing in y_i.*

For notational simplicity, we denote μ_i by μ and y_i by y. Without loss of generality, let us assume $\sigma = 1$.

$$y|\mu \sim N(\mu, 1), \quad \mu \sim g(\mu)$$

$$g(\mu) = \pi\delta_o + (1-\pi)g_1(\mu)$$

An estimator of μ is then given by

$$\begin{aligned} l(y) = E(\mu|y) &= \frac{\int \mu\phi(y-\mu)g(\mu)}{\int \phi(y-\mu)g(\mu)} \\ &= \frac{(1-\pi)\int \mu\phi(y-\mu)g_1(\mu)d\mu}{\pi\phi(y) + (1-\pi)\int \phi(y-\mu)g_1(\mu)d\mu} \\ &= \frac{\int \mu\phi(\mu)g_1(\mu)e^{\mu y}d\mu}{\frac{\pi}{1-\pi} + \int \phi(\mu)g_1(\mu)e^{\mu y}d\mu} \\ &= \frac{a(y)}{b(y)} \end{aligned}$$

where $a(y) = \int \mu\ \phi(\mu)\ g_1(\mu)\ e^{\mu y}\ d\mu$ and $b(y) = \frac{\pi}{1-\pi} + \int \phi(\mu)\ g_1(\mu)\ e^{\mu y}\ d\mu$.

Then $a'(y) = \int \mu^2\ \phi(\mu)\ g_1(\mu)\ e^{\mu y}\ d\mu$ and $b'(y) = a(y)$

$$\begin{aligned} l'(y) &= \frac{b(y)a'(y) - a(y)b'(y)}{b^2(y)} \\ &= \frac{(\frac{\pi}{1-\pi} + \int \phi(\mu)\ g_1(\mu)\ e^{\mu y}\ d\mu)(\int \mu^2\ \phi(\mu)\ g_1(\mu)\ e^{\mu y}\ d\mu) - (\int \mu\ \phi(\mu)\ g_1(\mu)\ e^{\mu y}\ d\mu)^2}{b^2(y)} \\ &= \frac{\frac{\pi}{1-\pi}\int \mu^2 f^*(\mu)\ d\mu + q(y)\int \mu^2 f^*(\mu)\ d\mu) - (\int \mu f^*(\mu)\ d\mu)^2}{\big(\frac{\pi}{1-\pi} + q(y)\big)^2} \end{aligned}$$

where $f^*(\mu) = \phi(\mu)\ g_1(\mu)\ e^{\mu y}$ and $q(y) = \int f^*(\mu)d\mu$. Therefore $l'(y)$ reduces to

$$\begin{aligned} l'(y) &= \frac{\frac{\pi}{1-\pi}\frac{1}{q(y)}\int \mu^2\ \frac{f^*(\mu)}{q(y)}\ d\mu + \int \mu^2\ \frac{f^*(\mu)}{q(y)}\ d\mu) - (\int \mu\ \frac{f^*(\mu)}{q(y)}\ d\mu)^2}{\left(\frac{\pi}{1-\pi}\frac{1}{q(y)} + 1\right)^2} \\ &= \frac{\frac{\pi}{(1-\pi)q(y)}E(\mu^2) + V(\mu)}{\left(\frac{\pi}{1-\pi}\frac{1}{q(y)} + 1\right)^2}\ \geq 0\ \forall y \end{aligned}$$

Hence $l(y)$ is non-decreasing function of y for any $g_1(\mu)$ defined on positive μ.

□

In the two component model, the posterior mean could be computed in a manner similar to that computed for the horseshoe type prior. However, for the spike-and-slab type prior, it is more interesting to look at the component-wise posterior median. For the posterior median, we use the estimator,

$$\hat{\mu}_{iM}(\mu_i|\boldsymbol{y}) = F_i^{-1}(1/2)$$

where

$$F_i(t) = E_{\boldsymbol{\theta}|\boldsymbol{y}}[P(\mu_i \leq t|\theta_i, y_i)]$$

and

$$P(\mu_i \leq t|\theta_i, y_i) = c(\theta_i, y_i) + \big(1 - c(\theta_i, y_i)\big)\ \Phi^{-1}\left(\frac{m_i}{s_i}\right)\ \left[\Phi\left(\frac{t - m_i}{s_i}\right) - \Phi\left(-\frac{m_i}{s_i}\right)\right].$$

A more specific form of the posterior median is

$$\hat{\mu}_{iM}(\mu_i|\boldsymbol{y}) = \begin{cases} 0 & \text{if } E_{\boldsymbol{\theta}|\boldsymbol{y}}\big(c(\theta_i, y_i)\big) \geq 0.5, \\ \inf\{x \geq 0 : F_i(x) \geq 0.5\} & \text{if } E_{\boldsymbol{\theta}|\boldsymbol{y}}\big(c(\theta_i, y_i)\big) < 0.5. \end{cases} \tag{1.6}$$

Thus, for an additive loss

$$L(\boldsymbol{\mu}, \hat{\boldsymbol{\mu}}) = \sum |\mu_i - \hat{\mu}_i|,$$

it makes sense to look at the component-wise posterior median, $\hat{\boldsymbol{\mu}}_M$.

One could also look at the Empirical Bayes estimator of the median which is the expectation of the posterior median expression with respect to $p(\boldsymbol{y}|\boldsymbol{\theta})$. Let

$$G_i(t|\theta_i, y_i) = P(\mu_i \leq t|\theta_i, y_i).$$

Then the expression for the median is

$$\tilde{\mu}_i(\theta_i, y_i) = 1[c(\theta_i, y_i) \leq 0.5]\ G^{-1}_{\theta_i, y_i}\left(\frac{\frac{1}{2} - c(\theta_i, y_i)}{1 - c(\theta_i, y_i)}\right). \tag{1.7}$$

One could show that the posterior median defined in (1.7) is a continuous shrinkage soft thresholding rule.

Result 3: For the Strawderman-Berger prior for the normal mean when the true mean is constrained to the non-negative orthant, the component-wise posterior median in (1.7) satisfies the following properties for a given value of the hyperparameter $\boldsymbol{\theta}$.

1. The posterior $p(\mu_i|\theta_i, y_i)$ is stochastically increasing in y_i and hence the posterior median of μ_i is a monotonically increasing in y_i for each value of the hyperparameter.

2. For each y_i, there exists $T(\theta_i)$ such that $\tilde{\mu}_i(\theta_i, y_i) = 0$ iff $y_i < T(\theta_i)$.

Proof. The posterior $p(\mu_i|\theta_i, y_i)$ is stochastically increasing in y_i and hence the posterior median of μ_i is a monotonically increasing in y_i for a given value of the hyperparameter $\boldsymbol{\theta}$.

For notational simplicity, we denote μ_i by μ and y_i by y. Without loss of generality,

let us assume $\sigma = \tau = \lambda_i = 1$ for $i = 1, \ldots, n$. From the expression for the posterior of μ, we have,

$$\pi(\mu|y) = c(p, y)\delta_o(\mu) + (1 - c(p, y))f(\mu|y)$$

where

$$\begin{aligned} f(\mu|y) &= h(y)^{-1}e^{\mu y}e^{-\frac{1}{2}\mu^2}g(\mu) \\ h(y) &= \int_0^\infty e^{\mu y}e^{-\frac{1}{2}\mu^2}g(\mu)d\mu \\ c(p, y) &= [1 + \frac{(1-p)}{p}h(y)]^{-1} \end{aligned}$$

To show that $\pi(\mu|y)$ is stochastically increasing (SI) in y, it is enough to show $f(\mu|y)$ is SI in y since $c(p, y)$ decreases with decrease in y.

Let $\mu_1 < \mu_2$ and $y_1 < y_2$. Then,

$$\frac{f(\mu_1|y_1)f(\mu_2|y_2)}{f(\mu_2|y_1)f(\mu_1|y_2)} = e^{(\mu_2-\mu_1)(y_2-y_1)} \geq 1$$

Thus,

$$f(\mu_1|y_1)f(\mu_2|y_2) \geq f(\mu_1|y_2)f(\mu_2|y_1)$$

Multiplying both sides by $\pi(y_1)\pi(y_2)$ where $\pi(y)$ is the marginal of y, we have,

$$f(\mu_1, y_1)f(\mu_2, y_2) \geq f(\mu_1, y_2)f(\mu_2, y_1)$$

Hence, $f(\mu, y)$ is Totally Positive of order 2 (TP_2). Hence, μ and y are SI in each other (Theorem 6.1, Dharmadhikari and Joag-Dev 1988) [13]. □

Proof. For each y_i, these exists $T(\theta_i)$ such that $\tilde{\mu}_i(\theta_i, y_i) = 0$ iff $y_i < T(\theta_i)$.

Since $c(\theta_i, y_i)$ is monotonically decreasing in y_i and

$$\begin{aligned} \lim_{y_i \to -\infty} c(\theta_i, y_i) &= 1 \\ \lim_{y_i \to \infty} c(\theta_i, y_i) &= 0 \end{aligned}$$

For each θ_i, $\exists \;\; \tilde{\mu}_i(\theta_i, y_i) = 0 \iff y_i < T(\theta_i)$. □

1.3 Posterior Computation and Numerical Results

Let $\boldsymbol{\theta}$ denote the set of all hyper-parameters. We use $\boldsymbol{\theta}$ interchangeably for horseshoe and Strawderman-Berger prior where $\boldsymbol{\theta} = \{\boldsymbol{\lambda}, \tau, \sigma\}$ for the former and $\boldsymbol{\theta} = \{\boldsymbol{\lambda}, \tau, p, \sigma\}$ for the later. We use a Metropolis within Gibbs algorithm to generate random samples from the marginal posterior distribution, $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ and thus compute posterior summaries for the posterior $\pi(\mu_i|\boldsymbol{y})$ by averaging the value of the hyperparameters over the randomly generated sample of $\boldsymbol{\theta}$. For the posterior mean $E(\mu_i|\boldsymbol{y})$ we use the estimator

$$E(\mu_i|\boldsymbol{y}) = L^{-1} \sum_{l=1}^{L} \mathrm{E}(\mu_i|\boldsymbol{\theta}_l, \boldsymbol{y}),$$

where $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_L$ are samples from $\pi(\boldsymbol{\theta}|\boldsymbol{y})$. For the posterior median, $Med(\mu_i|\boldsymbol{y})$, we use the estimator

$$Med(\mu_i|\boldsymbol{y}) = \hat{F}_i^{-1}(1/2),$$

where $\hat{F}_i(t) = L^{-1} \sum_{l=1}^{L} P(\mu_i \leq t|\boldsymbol{y}, \boldsymbol{\theta}_l)$.

The conditional marginal of $\boldsymbol{y}$ can be factorized as

$$\pi(\boldsymbol{y}|\boldsymbol{\theta}) = \prod_{i=1}^{n} \pi(y_i|\lambda_i, \tau, \sigma, p),$$

where $\pi(y_i|\lambda_i, \tau, \sigma) = \frac{1}{l_i}\phi(y_i/\lambda_i)\ \Phi(m_i/s_i)$ for the horseshoe prior and $\pi(y_i|\lambda_i, \tau, p, \sigma) = \frac{p}{\sigma}\phi(y_i/\sigma) + \frac{2(1-p)}{l_i}\ \phi(y_i/l_i)\ \Phi\Big(\frac{m_i}{s_i}\Big)$ for the Strawderman-Berger prior. The distribution of y_i conditional on the hyperparameters is Skew-Normal for the horseshoe prior and a mixture distribution of Normals for $\mu_i = 0$ and Skew-Normal for $\mu_i > 0$.

Hence the for the Gibbs sampling algorithm, the full conditionals are

1. $\pi(\lambda_i|\tau, \sigma, \boldsymbol{y}) \propto \pi(y_i|\lambda_i, \tau, \sigma)\ \pi(\lambda_i), \quad i = 1, \ldots, n$

2. $\pi(\tau, \sigma|\boldsymbol{\lambda}, \boldsymbol{y}) \propto \pi(\boldsymbol{y}|\boldsymbol{\lambda}, \tau, \sigma)\ \pi(\tau|\sigma)\ \pi(\sigma).$

For the Strawderman-Berger prior we have in addition,

1. $\pi(p|\boldsymbol{\lambda}, \tau, \sigma, \boldsymbol{y}) \propto \pi(\boldsymbol{y}|\boldsymbol{\lambda}, \tau, \sigma, p)\ \pi(p).$

The one-dimensional conditionals can be sampled using a standard Metropolis step.

1.3.1 Simulation results

We compare the performances of the Strawderman-Berger estimators, horseshoe estimator and Maximum Likelihood Estimator (MLE) under different degrees of sparsity. The MLE when $\boldsymbol{\mu} \in \mathcal{K} = \mathbb{R}^n_+$ for $\Sigma = \sigma^2 I$ is simply the projection of $\boldsymbol{y}$ onto the non-negative orthant i.e. $\hat{\mu}_i = \max(y_i, 0)$. For a general polyhedral cones, with Σ other than $\sigma^2 I$, the MLE is not straightforward to compute. We analyze the

risk properties of the estimators when the mean vector is simulated under strongly sparse signals and weakly sparse signals. For each of the sparsity level, we further consider two scenarios described below.

Strong sparsity: We use a discrete mixture model to generate exact zero entries for the mean vector using the model below:

$$\begin{aligned} y_i|\mu_i, \sigma^2 &\sim N(\mu_i, \sigma^2), \\ \pi(\mu_i) &= p\delta_o + (1-p)\ G(\alpha, \beta), \end{aligned}$$

where α is taken to be 5, β is 0.5 and 80% of the mean vector has exact zero entries. The major concentration of μ_i's is at 0 with an average concentration of $\mu_i > 0$ at 10 with variance 20. Two possible values of σ are considered: $\sigma = 1$ and $\sigma = 3$. The separation between y_i's at $\mu_i = 0$ and $\mu_i > 0$ is more prominent for $\sigma = 1$ than $\sigma = 3$.

Weak sparsity: For weakly sparse signals, we generate μ_i which decays according to the power law but none of its components are exactly zero. For this, we consider

$$\begin{aligned} y_i|\mu_i, \sigma^2 &\sim N(\mu_i, \sigma^2), \\ \mu_i|\eta, \alpha &\sim \text{Unif}(0, \eta c_i), \\ \eta &\sim \text{Ex}(2), \\ \alpha &\sim \text{Unif}(a, b), \end{aligned}$$

where $c_i = (n/i)^{1/\alpha}$ for $i = 1, \ldots, n$. For simulation purposes, $\sigma = 1$ is chosen and two possible scenarios of $\alpha \sim \text{Unif}(a, b)$ are considered: $a = 0.5, b = 1$ and $a = 1, b = 2$. The first scenario yields relatively large mean entries than the second scenario

depending on the randomly generated values of η and α. When $\alpha \sim \text{Unif}(1, 2)$, one can expect the concentration around 0 to be more dense than when $\alpha \sim \text{Unif}(0.5, 1)$ depending on the speed of decay, α.

For each of the scenarios, we simulate 1000 data sets from the corresponding model of dimension $n = 300$ using MCMC with 50000 runs and a burn-in period of 10000. The convergence is assessed using the standard MCMC diagnostic checks and all chains seem to converge. We report the median risk under squared error loss and absolute error loss along with the average risk ratios between the estimators in Table 1.1 and Table 1.2.

Figure 1.1 shows the plots for MLE estimates, posterior mean under horseshoe prior and posterior mean and posterior median under Strawderman-Berger prior for a single realization generated under strongly sparse signals with the variance set to $\sigma = 1$ or $\sigma = 3$. The dimension of the mean vector is 300. Figure 1.2 presents the same under weakly sparse signals for the two scenarios when $\alpha \sim \text{Unif}(0.5, 1)$ and $\alpha \sim \text{Unif}(1, 2)$.

From Figure 1.1 and Figure 1.2, we see that the posterior mean for the horseshoe provides shrinkage near zero, but it is still significantly positive even when the realized y is considerably negative. This is particularly undesirable in the constrained case when the true mean is known to be non-negative. From Result 1, we know that for negative y, the horseshoe estimator decays as $O(|y|^{-1})$. This induces considerable bias. The posterior mean under the Strawderman-Berger prior shrinks more than the horseshoe posterior mean estimator. However, for large positive y the horseshoe estimator seems to perform better, and shrinks less than the posterior

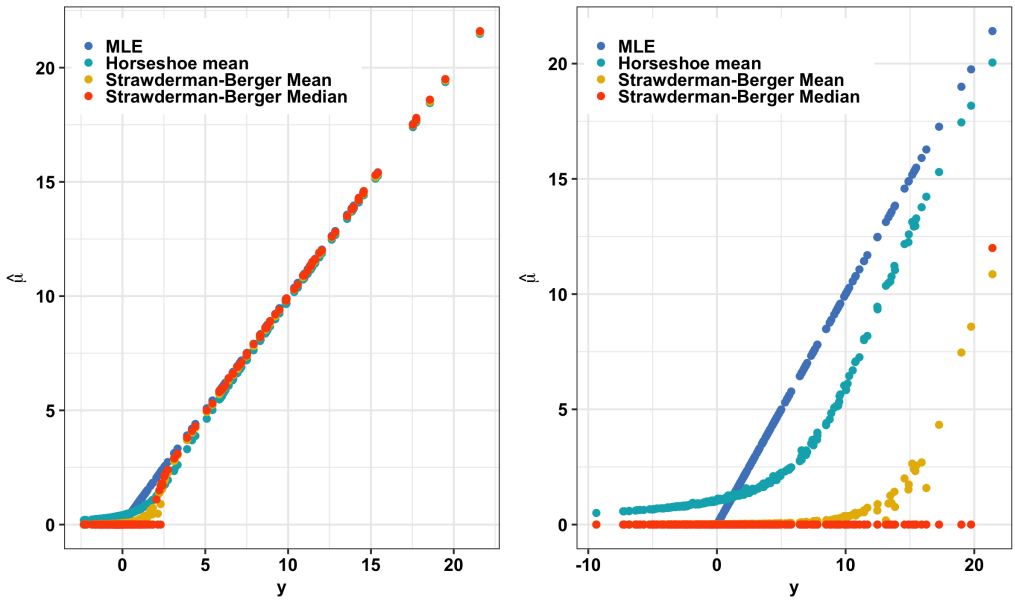

Figure 1.1: Plots of $\hat{\boldsymbol{\mu}}$ versus $\boldsymbol{y}$ under strong sparsity with $\sigma = 1$ (left) and $\sigma = 3$ (right)

mean under the Strawderman-Berger prior.

The posterior median for Strawderman-Berger prior, as expected from the results in Result 3, provides a soft thresholding estimator that is truncated to zero below the truncation point $T(y, p)$ and provides continuous shrinkage for y above the truncation point. All estimators are monotonic in y and the shrinkage factor tends to one as y tends to infinity, thereby satisfying the requirement to not perturbing the big realized values of y.

Table 1.1 shows that the risk performance of Strawderman-Berger posterior median and posterior mean is better than the MLE and horseshoe posterior mean both in terms of squared error loss and absolute loss for the strong sparsity case. In particular, the horseshoe posterior mean has at least 50% more risk than both

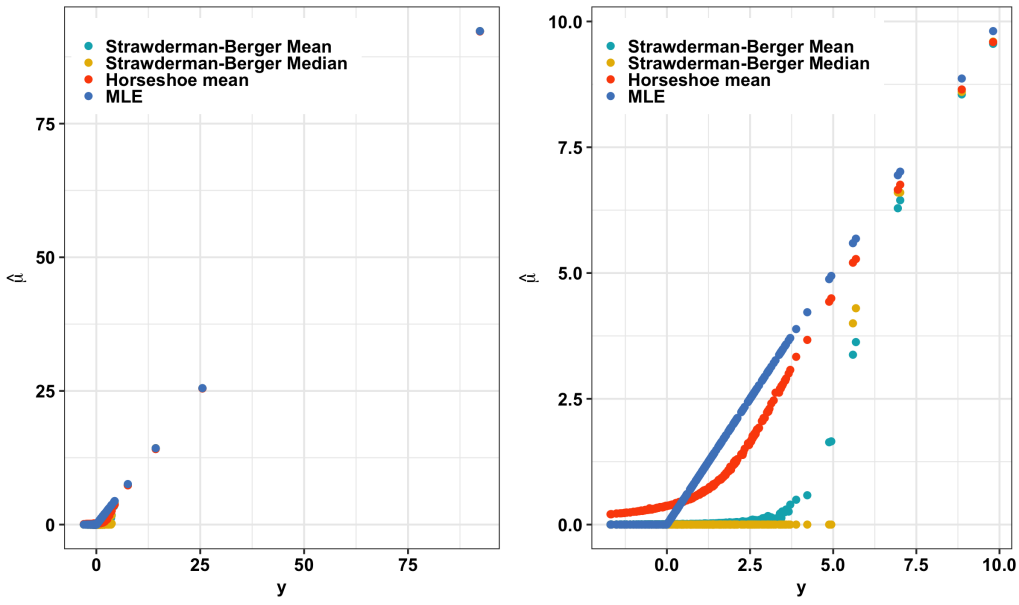

Figure 1.2: Plots of $\hat{\mu}$ versus y under weak sparsity where $\alpha \sim U(0.5, 1)$ (left) and $\alpha \sim U(1, 2)$ (right)

the Strawderman-Berger posterior mean and posterior median. However, the risk for horseshoe posterior mean under squared error loss is $20\% - 35\%$ less than the Strawderman-Berger estimators when $\sigma = 3$.

From Table 1.2, we see that the risk of horseshoe posterior mean is consistently less than that of MLE and Strawderman-Berger posterior mean and posterior median. Specifically, horseshoe posterior mean has of $6\% - 40\%$ more risk than the Strawderman-Berger estimators. However, when $\alpha \sim U(1, 2)$, horseshoe estimator has 63% more risk than the Strawderman-Berger posterior mean and approximately 411% more risk than SB posterior median, although the median squared error risk is less for horseshoe than the other estimators.

Table 1.1: Risk under squared error loss and absolute error loss for strongly sparse signals in two scenarios: $\sigma = 1$ and $\sigma = 3$. The diagonal components are median sum of squared error and absolute error. The off diagonal components are average error ratios of estimator in row by estimator in column.

		$\sigma = 1$				$\sigma = 3$			
		MLE	HS	SB Mean	SB Median	MLE	HS	SB Mean	SB Median
Square Error Loss	MLE	171	1.39	2.23	2.33	1598	1.19	0.98	0.77
	HS		131	1.6	1.67		1361	0.81	0.64
	SB Mean			82	1.04			1636	0.78
	SB Median				78				2129
Absolute Error Loss	MLE	143	0.92	1.97	2.6	428	0.95	1.43	1.42
	HS		156	2.13	2.8		452	1.5	1.49
	SB Mean			73	1.32			295	0.98
	SB Median				56				299

Table 1.2: Risk under squared error loss and absolute error loss for weakly sparse signals in two scenarios: $\alpha \sim U(0.5, 1)$ and $\alpha \sim U(1, 2)$. The diagonal components are median sum of squared error and absolute error. The off diagonal components are average error ratios of estimator in row by estimator in column.

		$\alpha \sim U(0.5, 1)$				$\alpha \sim U(1, 2)$			
		MLE	HS	SB Mean	SB Median	MLE	HS	SB Mean	SB Median
Square Error Loss	MLE	200.24	2.68	2.91	2.65	179	15.7	128.67	400
	HS		122.52	0.73	0.6		63	1.63	5.11
	SB Mean			185.7	0.81			128	0.92
	SB Median				235.8				136
Absolute Error Loss	MLE	181.8	1.6	1.75	1.63	166	3.09	4.32	4.8
	HS		136.39	0.94	0.85		86	0.91	0.93
	SB Mean			162.67	0.89			128	0.95
	SB Median				186.75				134

1.4 Real Data Analysis

We studied the performance of the estimators using the childhood acute lymphoblastic leukemia (ALL) data set (GSE412) which includes gene expression information for 110 childhood acute lymphoblastic leukemia samples before and after treatment [14]. From the originally measured 12625 probe sets, genes that were not present in at least one sample were removed to obtain 8280 genes. After cleaning the data, we selected 250 genes for 50 pediatric newly diagnosed children for our analysis. Our goal is to estimate the standardized difference between post-treatment mean, $\boldsymbol{\theta_2}$ and pre-treatment mean, $\boldsymbol{\theta_1}$ regardless of the type of treatment used i.e. $\boldsymbol{\mu} = \frac{\boldsymbol{\theta_2}-\boldsymbol{\theta_1}}{\sigma}$. For illustration purposes, we assume up-regulation of gene expression level in ALL cells so that $\boldsymbol{\mu} \in \mathbb{R}_+^n$. We further assumed that the gene expression levels are uncorrelated and have same variance. The observed data is the standardized difference of the average post-therapy and pre-therapy gene expression.

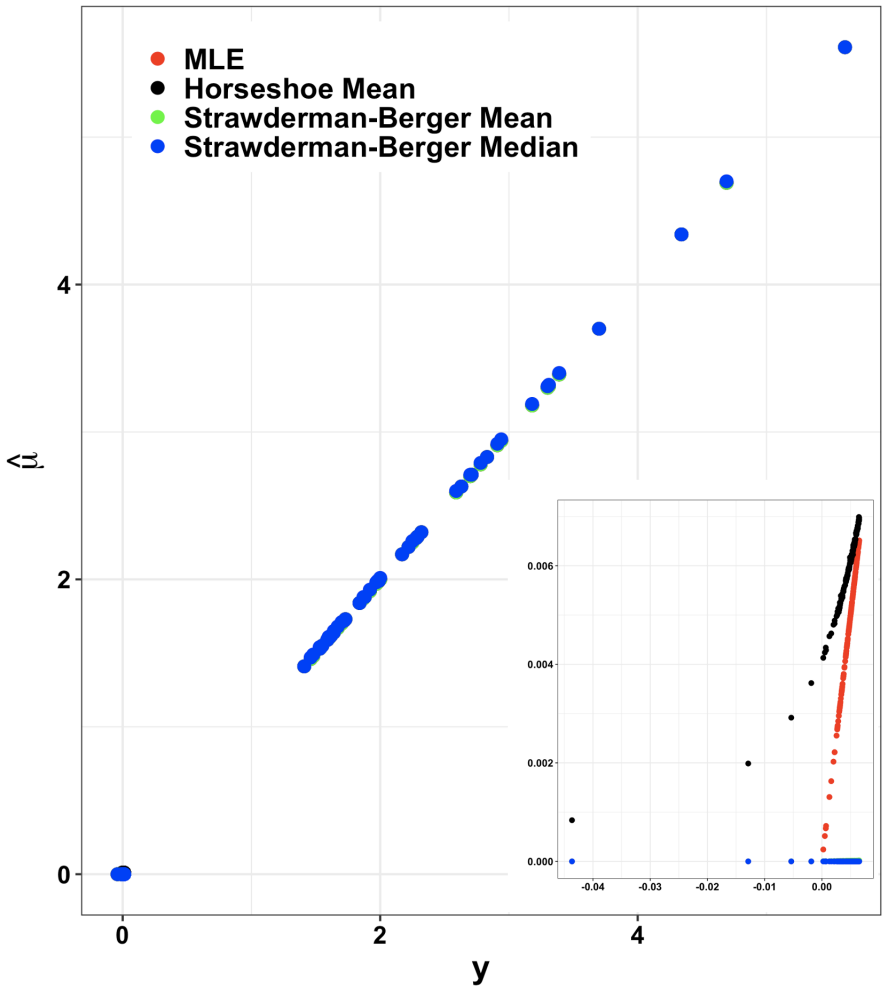

Figure 1.3: Plot of $\hat{\mu}$ versus y for the standardized difference in post-treatment and pre-treatment gene expression for 250 genes.

The estimated MLE, posterior means and posterior median are shown in Figure 1.3. The summary results for the observed data and the estimates are presented in Table 1.3. We noticed that Horseshoe posterior mean is always positive whereas the MLE is 0 for negative y's. While the Strawderman-Berger posterior median is exactly 0 for $y < 1.4$, the SB posterior mean is close to 0 for these values of y. All four estimates perform similarly for larger values of y.

Table 1.3: Summary statistics for the estimates under Horseshoe (HS) and Strawderman-Berger prior.

	$\boldsymbol{y}$	MLE	HS	SB	SB
			Posterior Mean	Posterior Mean	Posterior Median
Minimum	-0.0437	0.0000	0.0008	0.0000	0.0000
1st Quartile	0.0043	0.0043	0.0058	0.0000	0.0000
Median	0.0054	0.0054	0.0063	0.0000	0.0000
Mean	0.4618	0.4620	0.4630	0.4583	0.4592
3rd Quartile	0.0064	0.0064	0.0069	0.0000	0.0000
Maximum	5.6094	5.6094	5.6093	5.6094	5.6100

1.5 Discussion

In our simulation studies, we compared the performance of horseshoe posterior mean, Strawderman-Berger posterior mean and posterior median for strongly sparse signals and weakly sparse signals. While the posterior mean for both horseshoe and Strawderman-Berger prior are shrinkage estimators, MLE and Strawderman-Berger posterior median are truncation based estimators with exact zeros for small signals. When the true sparsity regime is strong sparsity, then truncation type estimators maybe preferred. The non-negative constraint does impact the relative performance of the mean and median estimators. It can be shown that the posterior mean under priors considered here are smooth differentiable functions of the observed value.

Hence it cannot be expected to capture the threshold like behavior present in the strongly sparse regime.

The numerical studies for the non-negative orthant are presented for the horseshoe prior and the Strawderman-Berger prior. It would be interesting to consider other scale mixture distributions, similar to Bayesian lasso, with hard thresholding properties for non-negative mean vectors. Another interesting domain is the discrete mixture models where the mixing kernel for the positive means could be chosen in a more flexible manner, belonging to flexible families on the non-negative orthant, e.g. product of gamma densities where heavy tailed priors are used for the hyperparamters. While the scope of this paper is limited to non-negative orthant which has many popular applications, one can think of exploring some of these priors to a general closed convex polyhedral cones. Moreover, the observations maybe allowed to be correlated with a known low-dimensional correlation structure. Hence we explore a more general set up in the next chapter.

Chapter 2: Characterization and estimation of high dimensional sparse regression parameters under linear inequality constraints

2.1 Introduction

In this chapter, we consider Bayesian estimation of possibly high dimensional parameter that are known to be restricted to a pointed closed convex polyhedral cone. We develop everything in the backdrop of normal mean estimation problem where the mean vector is constrained to a convex polyhedral cone but the concepts and the prior probability distributions developed here generalize easily to other models. Often, in constrained problems, the restricted models have to be embedded in higher dimensional models where the parameter space is unrestricted or at least more amenable to standard estimation methods. Thus, model complexity can be high in constrained problems even if the dimension of observations is not. In such situation some form of low dimensional formulation of the problem is required for making statistical inference possible without demanding a large sample size. The embedding to a higher dimensional space provides a parameterization of the model. For successful inference over a 'low dimensional' set of parameters the embedding needs to be an identifiable parameterization over that set. This property of the

embedding is not easily guaranteed. We look at the restriction of parameters to a pointed full-dimensional closed convex cone defined by a set of linear inequalities

$$\mathbb{C} = \{\boldsymbol{\mu} \in \mathbb{R}^n : \boldsymbol{A}\boldsymbol{\mu} \geq 0\} \tag{2.1}$$

where $\boldsymbol{A}$ is some fixed $m \times n$ matrix. Since the cone is the intersection of finitely many half-spaces, it is a polyhedral cone. We consider the natural embedding of the cone using its minimal set of generators and consider its restriction to lower dimensional faces of the cone. We show that ascribing sparsity on the parameters of the embedding is not sufficient to have identifiable representation of the lower dimensional parameter vectors.

The main contribution of this chapter is an identifiable parameterization of vectors lying in lower dimensional subsets of the cone described in terms of the minimal generators representations. We define such vectors lying on the lower dimensional faces as 'sparse' vector because the notion of sparsity agrees with the usual notion of sparsity when the cone is an orthant. Then, using the proposed definition of sparsity we defined flexible prior distributions that are either fully or nearly fully supported on the set of 'sparse' vectors and allows one to carry on Bayesian inference under sparsity and conic constraints.

Suppose $\boldsymbol{y} = (y_1, \ldots, y_n)' \sim N(\boldsymbol{\mu}, \sigma^2\boldsymbol{I})$ where the parameter of interest $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)' \in \mathbb{C}$. We assume that $\mathbb{C}$ has non-zero interior volume with respect to the n dimensional Lebesgue measure. We consider a general framework where $\boldsymbol{\mu}$ is constrained to a proper polyhedral cone $\boldsymbol{A}\boldsymbol{\mu} \geq \boldsymbol{0}$. A proper polyhedral cone is a closed convex full polyhedral cone that is pointed. A pointed cone is one that does

not contain any non-trivial subspace and it is full or full-dimensional if the dual cone is pointed. We assume the cone is pointed (acute) and irreducible, i.e., the $m \times n$ $(m \geq n)$ matrix describing the linear inequalities, $\boldsymbol{A}$, is full column rank, and the rows are conically independent in the sense that there are no non-negative linear combinations, other than the trivial combination, of the rows that gives the zero vector.

The importance of linear inequality constraints in the practice of statistics is two fold. First, linear constraints arise extensively in shape restricted inference including, but not limited to, monotonicity, concavity or convexity. Such restrictions can be imposed directly on the mean function parameter or they can be modelled non-parametrically to obtain a flexible and smooth estimate. For example, if our goal is to fit a function, f to the data $(x_1, y_1), \ldots, (x_n, y_n)$, so that

$$y_i = f(x_i) + \epsilon_i$$

where f is assumed to have some restrictions and $E(\boldsymbol{\epsilon}) = \mathbf{0}$ and $\text{cov}(\boldsymbol{\epsilon}) = \sigma^2 \boldsymbol{I}$ then assuming a parametric approach, the mean function f is the parameter $\boldsymbol{\mu}$ with $A\boldsymbol{\mu} \geq \mathbf{0}$.

Second, the linear inequalities constraint framework can be used to extend estimation of $\boldsymbol{\mu}$ in the non-negative orthant when the covariance matrix is a general positive definite matrix $\boldsymbol{\Sigma}$. Consider the model $\boldsymbol{y}|\boldsymbol{\mu} \sim N(\boldsymbol{\mu}, \sigma^2\boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma}$ is completely known. A standard approach to dealing with general $\boldsymbol{\Sigma}$ matrix is to transform the observations to $\boldsymbol{z} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{y}$ so that $\boldsymbol{z}|\boldsymbol{\theta} \sim N(\boldsymbol{\theta}, \sigma^2\boldsymbol{I})$ where $\boldsymbol{\theta} = \Sigma^{-1/2}\boldsymbol{\mu}$. However, the transformed mean $\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\mu}$ need not remain in the positive orthant unless

$\boldsymbol{\Sigma}$ is such that the square root $\boldsymbol{\Sigma}^{-1/2}$ is a *positive operator*, i.e., a matrix that leaves the cone unchanged. In the case of the positive orthant that would mean $\boldsymbol{\Sigma}^{-1/2}$ is a non-negative matrix e.g. $\boldsymbol{\Sigma}$ is an M-matrix with an inverse that admits a positive square-root. Hence one could reduce the problem of estimating $\boldsymbol{\mu}$ where $\boldsymbol{\mu} \geq \mathbf{0}$ to estimating $\boldsymbol{\theta}$ where $\boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta} \geq \mathbf{0}$.

Of course, one could combine these two problems and consider the bigger problem of linear inequality constraints for a general $\boldsymbol{\Sigma}$, i.e., estimate $\boldsymbol{\mu}$ where $\boldsymbol{y}|\boldsymbol{\mu} \sim N(\boldsymbol{\mu}, \sigma^2\boldsymbol{\Sigma})$ with $\boldsymbol{A\mu} \geq \mathbf{0}$. The problem can be transformed by taking $\boldsymbol{z} = \Sigma^{-1/2}\boldsymbol{y}$ so that $\boldsymbol{z}|\boldsymbol{\theta} \sim N(\boldsymbol{\theta}, \sigma^2\boldsymbol{I})$ where $\boldsymbol{\theta} = \Sigma^{-1/2}\boldsymbol{\mu}$. Hence, the estimation of $\boldsymbol{\mu}$ where $A\boldsymbol{\mu} \geq \mathbf{0}$ is condensed to estimating $\boldsymbol{\theta}$ where $\boldsymbol{A}\Sigma^{1/2}\boldsymbol{\theta} \geq \mathbf{0}$.

One way of estimating such a parameter is to first obtain an unrestricted estimate of the parameter and then truncate it so that the estimate lies in the constrained parameter space. Intuitively, the performance of the estimator is expected to be much better if such constraint conditions are incorporated in the estimation process. Hence the idea here is to incorporate the linear inequality restrictions into the model and in the inferential procedures.

From a frequentist estimation point of view this is a standard 'cone projection' problem of finding $\boldsymbol{\mu} \in \mathbb{C}$ such that it minimizes $||\boldsymbol{y} - \boldsymbol{\mu}||^2$. The cone projection problem is a special case of quadratic programming which involves finding $\boldsymbol{\theta}$ such that it minimizes $\boldsymbol{\theta}^T\boldsymbol{Q}\boldsymbol{\theta} - 2c^T\boldsymbol{\theta}$ over $\mathbb{C}$. When $\boldsymbol{Q}$ is positive definite, the objective function has a unique minimum and the solution reduces to finding the projection of a general Euclidean vector to the convex cone [15, 16]. Several algorithms have been studied in the literature to address the cone projection problem by Dykstra (1983),

Karmarkar (1984), Fraser and Massam (1989) among others [17–24]. A detailed account of the numerical stability and computational cost of the projection algorithms has been studied by Dimiccoli (2016) [25]. Constrained inference problems have been discussed in details in Sen and Silvapulle (2001) [26]. Polyhedral cone constraints or equivalently linear inequalities arise extensively in shape restricted inference. There are many papers on estimation of regression function under shape restrictions which are special cases of the conic restriction problem. In the Bayesian set up, Danaher *et al.* (2012) provided an example of Bayesian estimation of normal mean when the mean is constrained to a convex polytope [27]. In this paper, we look at more than just estimating the constrained parameter vector.

One of the most interesting questions that naturally arise in the context of closed convex polyhedral cone restrictions is how to specify sparsity in constrained spaces such as $\mathbb{C}$. We provide a novel characterization of "sparse" parameters restricted to a polyhedral cone in Section 2.2. The notion of 'sparsity' defined here conforms with the general definition in the unrestricted case or the case of the orthant. In Section 2.3 we define priors where the bulk of the support is on the sparse vectors. Such priors would facilitate sparse signal extraction under general convex polyhedral cone restrictions.

2.2 Sparsity on Closed Convex Polyhedral Cones

To begin with, we provide some background on the geometry of cone and produce examples of three dimensional cone to understand and set the ideas. For

any cone $\mathbb{C}$, let us denote its dimension by $\dim(\mathbb{C})$. A polyhedral cone is formed by the intersection of finitely many half spaces that contain the origin, i.e., for a matrix $\boldsymbol{A} \in \mathbb{R}^{m \times n}$, we define

$$\mathbb{C} = \{\boldsymbol{\mu} \in \mathbb{R}^n : \boldsymbol{A}\boldsymbol{\mu} \geq 0\} \tag{2.2}$$

to be a polyhedral cone with $\dim(\mathbb{C}) = n$. The halfspace representation of the cone containing the origin is called the facet representation or H-representation and the matrix A forming the set of linear inequalities is called the representation matrix. The face of a cone is a lower dimensional feature formed by the intersection of the cone with a supporting hyperplane. In particular, we focus on vertex, extreme ray and facet that are faces of a cone, each lying in different dimension. A vertex is a face of dimension 0, an extreme ray is a face of dimension 1 and a facet is a face of dimension $\dim(\mathbb{C}) - 1$.

We use the primal-dual representation of the cone to define 'sparsity'. Using Minkowski's theorem, a polyhedral cone (2.2) can also be represented using a finite set of vectors called generators or extreme rays. That is, for any $\boldsymbol{A}_{m \times n}$, there exists a generating matrix $\boldsymbol{\Delta}_{n \times d}$ such that

$$\mathbb{C} \;\; = \;\; \{\boldsymbol{\mu} \in \mathbb{R}^n : \boldsymbol{\mu} = \boldsymbol{\Delta}\boldsymbol{b} = \sum_{j=1}^{d} b_j \boldsymbol{\delta}_j, \;\; b_j \geq 0\} \tag{2.3}$$

where the columns $\boldsymbol{\delta}_j$ are the generators of the cone. This representation of a polyhedral cone is called the vertex representation or V-representation. The converse of the Minkowski's theorem is the Weyl's theorem for a polyhedral cone which states the existence of a representation matrix given a generating matrix. The generators are called *minimal* if they are conically independent, i.e., there is no positive linear

combination of the generators that equals the origin vector. For the rest of this chapter, we assume $\boldsymbol{A}$ to be a *irreducible* matrix meaning that rows of $\boldsymbol{A}$ are conically independent. If $\boldsymbol{A}$ is full row rank, then it is irreducible. We also assume that the $rank(A) = n$. The resulting cone is called an *acute* cone and the set of extreme rays is its minimal generating system. In that case, d is the minimal number of extreme rays forming the skeleton of the cone.

Remark 1. *The parameterization of a cone $\mathbb{C}$ in terms of $\boldsymbol{b}$ in its vertex representation is not a proper parameterization in the sense for each vector $\boldsymbol{\mu} \in \mathbb{C}$ there could be multiple $\boldsymbol{b}$ such that $\boldsymbol{\Delta b} = \boldsymbol{\mu}$ even when the cone is irreducible and acute. Thus, the vector $\boldsymbol{b}$ is not generally identifiable from the vector $\boldsymbol{\mu}$. Only when $m = n = d$ and the cone is irreducible and acute, in which case $\Delta = \boldsymbol{A}^{-1}$ is non-singular and the parameterization is a bijection between the cone and the non-negative orthant.*

Figure (2.1) shows an example of a polyhedral cone in $\mathbb{R}^3$ ($n = 3$) formed by $m = 6$ homogeneous linear inequalities. There are $m = 6$ hyperplanes intersecting with the cone and hence the number of facets is 6. Also, it turns out the number of extreme rays in $\mathbb{R}^3$ is equal to the number of facets. However, it is not true in general and d can be substantially larger than m which leads us to the next part.

Since there are two descriptions of a polyhedral cone, the pair $(\boldsymbol{A}, \boldsymbol{\Delta})$ is said be the Double description (DD) pair [28]. Switching between the two descriptions is called the representation conversion problem. Given the facet representation, the problem of finding the set of minimal extreme rays is called the extreme ray enumeration problem. Similarly, finding the irreducible representation from the

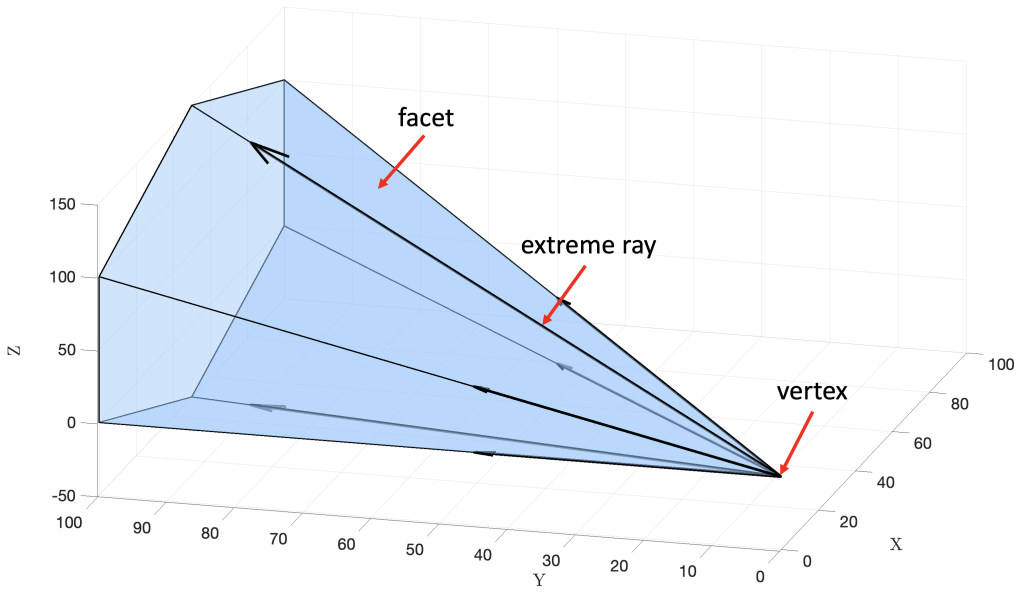

Figure 2.1: An example of polyhedral cone in $\mathbb{R}^3$ with $m = 6$ homogeneous linear inequalities and 6 extreme rays.

vertex representation is called facet enumeration problem. When A is full row rank, the extreme rays are given by the columns of $\boldsymbol{\Delta} = \boldsymbol{A}^T(\boldsymbol{A}\boldsymbol{A}^T)^{-1}$ and $d = m$ [16]. When $\boldsymbol{A}$ is not full row rank, the number of extreme rays may be substantially larger than m. In that case, the extreme rays of the cone can be obtained using proposition 1 from Meyer (1999) [29].

There have been many variations and modifications of the Double Description (DD) method to move back and forth between the two representations, right from the primitive DD method to standard DD method [28, 30–32]. We use the R package "rcdd" by K. Fakuda, a R interface for cddlib which is a C-implementation of the DD method of Motzkin et al. [31, 33].

The methodology proposed here depends on the idea of describing points on the boundary of the cone or describing a points with proximity to the boundary of

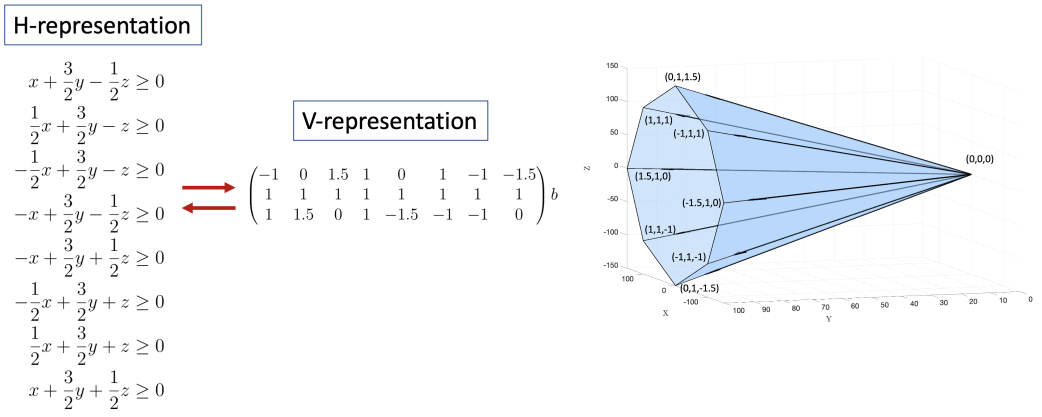

Figure 2.2: An illustration of the H-representation (left) and V-representation (center) for a irreducible polyhedral cone (right) in $\mathbb{R}^3$ with $n = 3, m = 8, d = 8$.

the cone when the point is in the interior. To this end, we need to use the adjacency graph with the list of adjacent extreme rays of the cone.

Definition 1. *For an acute cone $\mathbb{C} = \{\boldsymbol{\mu} : \boldsymbol{\mu} = \boldsymbol{\Delta b}\}$, two extreme rays $\boldsymbol{\delta}_i$ and $\boldsymbol{\delta}_j$ are adjacent if the minimal face containing both rays does not contain any other extreme rays of the cone.*

Two well-known tests for verifying the adjacency of extreme rays of a cone are the algebraic test and combinatorial test [28]. Given the adjacency relation one can define the *adjacency graph* of the cone. Let the columns of matrix Δ be given by $\boldsymbol{\delta}_j, j = 1, \ldots, d$. Then, $\{\boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_d\}$ correspond to a set of nodes in $V = \{1, \ldots, d\}$. Then the edge set E is defined through adjacency, i.e., each pair of adjacent extreme rays, i and j correspond to a edge in the graph network. The edge set E can be written as the union of the edge set for each node, i.e., $E = \{E_1, E_2, \ldots, E_d\}$ where

E_i denote the set of adjacent extreme rays corresponding to the $\boldsymbol{\delta}_i$. Then $G = (V, E)$ forms an undirected graph. The degree of a node of a graph is the number of edges that are incident to the node. We denote the degree of the node i by $deg(\boldsymbol{\delta}_i)$. Then $|E_i| = deg(\boldsymbol{\delta}_i) + 1$.

To illustrate the geometry of polyhedral cones in 3D, consider the following example with $n = 3, m = 8, d = 8$ from Figure 2.2. The corresponding adjacency graph is shown in Figure 2.3. For instance, $\boldsymbol{\delta}_1$ is an extreme ray, which is is adjacent to $\boldsymbol{\delta}_2$ and $\boldsymbol{\delta}_8$. Hence in the corresponding adjacency graph, node 1 is connected to node 2 and node 8. In this case, each extreme ray is connected to two other extreme rays. So $deg(\boldsymbol{\delta}_i) = 2$ and $|E_i| = 3 \ \ \forall i$.

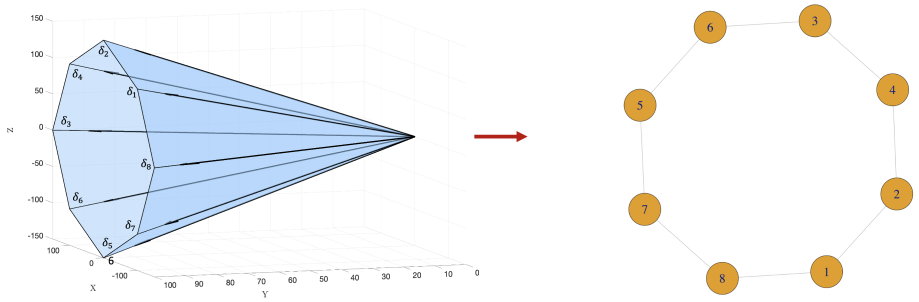

Figure 2.3: The graph network for the cone from Figure 2.2.

For high dimension with $n > 3$, the adjacency graph can become quite complicated with varying degree. A simple example for a polyhedral cone in $\mathbb{R}^4$ with $m = 7, d = 8$ is illustrated below with degree varying between 3 and 4.

When $m = n$, the number of minimal generators is the same as the dimension, and the adjacency graph is a *complete* graph. We will use the adjacency network to describe the notion of sparsity as well as the proposed priors.

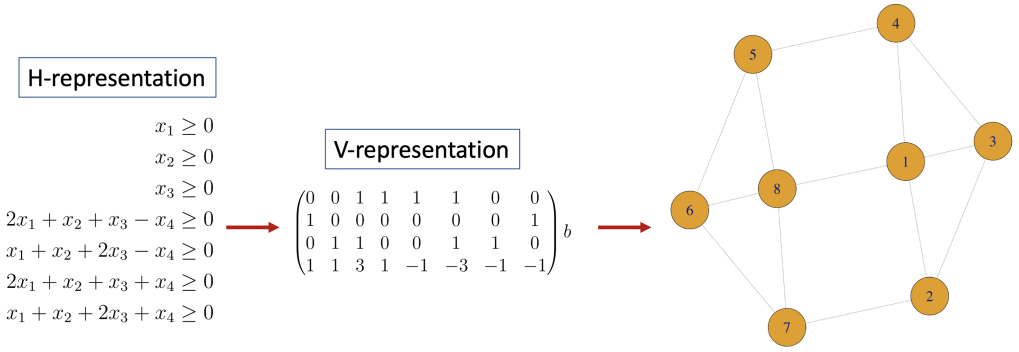

Figure 2.4: An illustration of the H-representation (left) and V-representation (center) for a irreducible polyhedral cone in $\mathbb{R}^4$ with its adjacency graph (right).

When there are no restrictions, a sparse vector is a vector that has a large number of zeros (or for a weaker notion of sparsity the vector has a large number of entries that are negligible). For non-negative orthant, the same definition applies, except the non-zero entries are required to be positive. Thus, the sparse vectors are the one which lie on (or close to) one of the lower dimensional faces of the orthant. Following the description of sparsity in the orthant, we define a sparse vector to be any vector lying on or near a lower dimensional face. Since any $\boldsymbol{x} \in \mathcal{K}$ can be represented as $\boldsymbol{x} = \Delta \boldsymbol{b}$, extrinsic 'sparsity' can be defined as $\boldsymbol{x}$ being specified by smaller number of lower dimensional features. In other words, $\boldsymbol{x}$ is sparse when $\boldsymbol{b}$ is a sparse vector. The idea is to map the vector $\boldsymbol{x}$ in $\mathbb{R}^n$ to a non-negative orthant in $\mathbb{R}^d$, use the definition of sparsity in the orthant and then use the inverse map to lift the notion of sparsity back to the polyhedral cone. The dimension d in which the vector $\boldsymbol{x}$ is being embedded is either equal to or larger than the original dimension

n.

For a non-negative orthant, the canonical vectors are the minimal generators and the usual definition of sparsity is that the vector can be written as a conic combination of a few of the full set of generators. Such vectors will lie on the boundary of the orthant, on a lower dimensional face of the orthant to be precise. It seems natural to use a similar definition of sparsity in the general case, i.e., vectors that lie on lower dimensional faces of the cone. The minimal two dimensional faces are the conic hull of pair of adjacent generator. Thus, to restrict the vector to the lower dimensional faces one can work with adjacent generators. However, simply generating a vector as a conic combination of a set of adjacent rays is not enough to guarantee that the vector lies on a lower dimensional face. It seems that the notion sparsity is more nuanced. For the vector to occupy a lower dimensional face, the sets of generators must form a *clique* or in other words the sub-adjacency graph corresponding to the set of generators used to define a sparse vector must be complete. This will ensure that the notion of sparsity is an identifiable notion in the sense that a sparse vector cannot have a non-sparse representation.

Recall that a clique, W, of an undirected graph $G = (V, E)$ is a subset of vertices, $W \subseteq V$ such that every two distinct nodes are connected by an edge. That is, a clique of a graph is an induced subgraph that is complete. A maximum clique of a graph, G, is a clique w such that $w \bigcup \{v\}$ in not a clique for any $v \in V \backslash w$. Then we have the following definition of a 'sparse' vector in a closed convex polyhedral cone.

Definition 2. *Let* $\mathbb{C} = \{\boldsymbol{\mu} \in \mathbb{R}^n : \boldsymbol{\Delta b}\}$ *be the vertex representation of a closed convex polyhedral cone* $\mathbb{C}$ *where the columns of* $\boldsymbol{\Delta}$ *form a set of* d *minimal generators of* $\mathbb{C}$*. Let* $G = (V, E)$ *be the adjacency graph of* $\mathbb{C}$*. Then* $\boldsymbol{\mu} = \boldsymbol{\Delta b} \in \mathbb{C}$ *is sparse iff the subgraph corresponding to* $i : b_i > 0$ *is a clique.*

The following result proves that the above definition is 'proper' in the sense for a sparse vector there cannot be a non-sparse representation.

Theorem 1. *Suppose* $\boldsymbol{\mu} \in \mathbb{C}$ *has a vertex representation* $\boldsymbol{\mu} = \boldsymbol{\Delta b}$ *such that the set of nodes* $\mathcal{I} = \{i : b_i > 0\}$ *forms a clique. Then in any vertex representation of* $\boldsymbol{\mu} = \boldsymbol{\Delta\beta}$ *we have* $\beta_i = 0$ *for all* $i \in \{1, \ldots, d\} \backslash \mathcal{I}$*.*

Proof. We will use method of induction to prove the result. From the definition of adjacency, the result is obvious true when the size of the clique is $k = 2$. Now suppose it is true a positive integer $k > 2$. Let $\boldsymbol{\mu} = \sum_{i=1}^{k+1} b_i \boldsymbol{\delta}_i$ be a vertex representation of a vector $\boldsymbol{\mu}$ where without loss of generality we assume that the nodes $\{1, \ldots, k+1\}$ form a clique. Suppose there is another representation of $\boldsymbol{\mu}$ as

$$\boldsymbol{\mu} = \sum_{i=1}^{k+1} \beta_i \boldsymbol{\delta}_i + \sum_{i=k+2}^{d} \beta_i \boldsymbol{\delta}_i.$$

Then $\mathbf{0} = \sum_{i=1}^{k+1} (\beta_i - b_i)\boldsymbol{\delta}_i + \sum_{i=k+2}^{d} \beta_i \boldsymbol{\delta}_i$. Consider two cases.

case1: $(\beta_i - b_i) \geq 0, \forall i$. In this case a non-negative linear combination of the columns of $\boldsymbol{\Delta}$ is zero which contradicts minimality of the generators.

case2: $(\beta_i - b_i) < 0$, for some i. Let $\mathcal{J} = \{i : (\beta_i - b_i) < 0\}$. Then

$$\boldsymbol{x} = \sum_{i \in \mathcal{J}} (b_i - \beta_i)\boldsymbol{\delta}_i = \sum_{i \in \{1,\ldots,k+1\} \backslash \mathcal{J}} (\beta_i - b_i)\boldsymbol{\delta}_i + \sum_{i=k+2}^{d} \beta_i \boldsymbol{\delta}_i.$$

Thus, the vector $\boldsymbol{x}$ has two representations one of which is based on a clique since any sub-clique of a clique is also a clique. Since $|\mathcal{J}| \leq k$, this contradicts the assumption unless $\beta_i = b_i$ for $i = 1, \ldots, k+1$ and $\beta_i = 0$ for $i = (k+1), \ldots, d$. This completes the proof. □

2.3 Sparse Priors for Closed Convex Polyhedral cone

To define probability measures on the cone that is supported mostly on lower dimensional sets, one could simply specify any sparse prior that are used in the unrestricted case as a prior on $\boldsymbol{b}$ in the vertex representation and invoke a prior on $\boldsymbol{\mu}$. Such a prior indeed works as a sparse prior on the cone provided the adjacency graph is a complete graph, as in the case of the positive orthant.

Thus, for the case when $d = n$, and hence the adjacency graph is a complete graph one could use popular sparse priors such as the continuous shrinkage priors like Horseshoe priors [10] or spike-and-slab priors like the Strawderman-Berger prior [6], where the continuous part is taken to be a density on the first orthant such as product of normal densities truncated to the positive half. Specifically, one could define priors on $\boldsymbol{b}$ as the Horseshoe prior

$$\begin{aligned} b_i|\tau, \lambda_i &\sim N(0, \tau^2\lambda_i^2)_+, \\ \lambda_i &\sim C(0,1)_+, \\ \tau|\sigma &\sim C(0,\sigma)_+, \\ \pi(\sigma) &\propto \tfrac{1}{\sigma} \end{aligned} \tag{2.4}$$

or as the Strawderman-Berger prior

$$\begin{aligned}
\pi(b_i) &= p\delta_o + (1-p)\ N(0, \tau^2\lambda_i^2)_+ \\
\pi(\lambda_i) &\propto \lambda_i(1+\lambda_i^2)^{\frac{3}{2}}, \\
p &\sim \text{Unif}(0,1), \\
\tau|\sigma &\sim C(\sigma,\sigma)\ 1(\tau \geq \sigma), \\
\pi(\sigma) &\propto \frac{1}{\sigma}
\end{aligned}$$

We use $\pi_{HS}(\boldsymbol{b})$ and $\pi_{SB}(\boldsymbol{b})$ to refer to the Horseshoe prior in 2.4 and the Strawderman-Berger Prior in 2.5, respectively. One could also specify other priors such as Bayesian lasso [3] like prior on $\boldsymbol{b}$. However, when the adjacency graph is not complete, simply demanding that the vector $\boldsymbol{b}$ is sparse does not ensure that the resulting $\boldsymbol{\mu}$ vector is near a lower dimensional face. To guarantee sparsity, it is important to specify which of the components of $\boldsymbol{b}$ are zero. For instance, consider the above example of the 3D cone with eight extreme rays and suppose only b_4 and b_8 are the only positive entries in $\boldsymbol{b}$. The resulting vector will lie on the 2D cone generated by the vectors $\boldsymbol{\delta}_4$ and $\boldsymbol{\delta}_8$. Points on this set can be far away from any of the faces and can have many equivalent dense representation (Remark 1) where none of the entries in $\boldsymbol{b}$ is zero or small. Hence, the vector will not be sparse according to the notion described above and general sparse prior on $\boldsymbol{b}$ may still put substantial mass in the dense interior of the cone.

It is evident from the definition of sparsity that one could simply restrict to the clique-lattice of the adjacency graph and work with the maximal cliques to define priors that will be supported only on sparse vectors. To define such a

probability measure, an obvious choice would be to define a Markov Random Field (MRF), specifically the Gibbs distribution describing the clique probabilities and then conditional on the clique defining a prior on the entries of $\boldsymbol{b}$ within the clique. We briefly review the Markov-Gibbs equivalence in the context of an undirected graph. Suppose $\{X_v : v \in V\}$ be a stochastic process with X_v taking values in S_v. Suppose further the joint distribution of the variables is $Q\{\boldsymbol{x}\} = P\{X_v = x_v \text{ for } v \in V\}$ where $\boldsymbol{x} = (x_1, \ldots, x_d)$ and $x_i \in S_i$.

Definition 3. *The probability distribution Q is called a Gibbs distribution for the graph if it can be written in the form*

$$Q\{\boldsymbol{x}\} = \prod_{S \in W} \phi_S(\boldsymbol{x})$$

where W is the set of cliques for G and ϕ_S is a positive function (also referred to as clique potential function) that depends on $\boldsymbol{x}$ only through $\{x_v : v \in S\}$. The definition is equivalent if maximal cliques are used instead of just cliques.

An MRF is characterized by its local property (the Markovianity) whereas a Gibbs Random Field (GRF) is characterized by its global property (the Gibbs distribution). The Hammersley-Clifford theorem establishes the equivalence of these two types of properties. The theorem asserts that the process $\{X_v : v \in V\}$ is a Markov Random field if and only if the corresponding Q is a Gibbs distribution. The practical value of the theorem is that it provides a simple way to parametrize the joint probability by specifying the clique potential functions. In other words, the theorem tells us it suffices to search over Gibbs distribution.

Given a particular maximal clique, define the sparsity of a vector in the usual sense by generating the vector using possibly sparse coefficients on the generators belonging to the clique. This procedure agrees with the usual method of selecting sparse vectors on the orthant or $\mathbb{R}^n$ where the generators are the canonical vectors and all the extreme rays together for the unique maximal clique.

Specifically we recommend the following class of sparse prior on $\mathbb{C}$. Let $\mathcal{W}$ be the set of maximal cliques of the adjacency graph of $\mathbb{C}$.

$$\begin{aligned} \boldsymbol{b}|w &\sim \pi(\boldsymbol{b}_w) \\ w &\sim \pi_{\mathcal{W}}(w) \end{aligned} \tag{2.5}$$

where given a clique $w \in \mathcal{W}$, $\boldsymbol{b}_w$ is the subvector of $\boldsymbol{b}$ constructed with entries of $\boldsymbol{b}$ whose indices belong to w, $\pi(\boldsymbol{b}_w)$ is a 'sparse' prior, such as the Horseshoe prior or the Strawderman-Berger prior on $\boldsymbol{b}_w$ in appropriate dimension, and $\pi_{\mathcal{W}}(w)$ is an MRF on $\mathcal{W}$. The priors $\pi(\cdot)$ and $\pi_{\mathcal{W}}(\cdot)$ an have their own hyper-parameters and hyperpriors can be specified accordingly.

In order to have a prior that is fully supported but has most of the support on the sparse vectors one could add a mixture component including the full set of extreme rays

$$\begin{aligned} \boldsymbol{b}|\rho, w &\sim \rho\pi^0(\boldsymbol{b}) + (1-\rho)\pi(\boldsymbol{b}_w) \\ w|\rho &\sim \rho I(w = V) + (1-\rho)\pi_{\mathcal{W}}(w) \\ \rho &\sim Bernoulli(\phi) \end{aligned} \tag{2.6}$$

where $\pi^0(\cdot)$ is a sparse prior on the interior of the positive orthant, $\mathbb{R}^n_+$, and the

Bernoulli parameter ϕ is either a pre-specified small probability or a prior can be specified on ϕ.

2.3.1 Prior with adjacency on $\boldsymbol{b}$

A 'weaker' notion of sparsity will be to allow for mass to be spread near the boundary of the cone instead of being only supported on the boundary. In high dimension, the probability for most fully supported measures on the entire cone will concentrate on or near the boundary and hence so will the posterior. However, how the prior is specified will have an impact on the recovery rate of the sparse sets.

Instead of restricting to cliques, one could choose priors that are supported on a cone generated by a single adjacency set. While not guaranteed, such priors would emphasize vectors where most of the coefficients in $\boldsymbol{b}$ are small in any representation of the vector. Of course the idea of small or negligible coefficients has to be formalized but in general this would mean $b_j < \epsilon, j \notin E_i$ for a given adjacency set E_i and for some pre-specified small value $\epsilon > 0$. Unfortunately, even when only a few coefficients within an adjacency are set to positive values, the resulting vector may still have equivalent representations that are very dense. If the prior specified on the elements of $\boldsymbol{b}$ within an adjacency set is sufficiently sparse, with high prior probability the generated vectors would be near one of the boundary sets, i.e., the minimum distance of the point to the boundary will be small.

To this end we define 'weakly sparse' priors that are fully supported on a closed convex polyhedral cone $\mathbb{C}$ and with most or all of its mass supported on or near the

boundary. To formally define this, let

$$S(\boldsymbol{\mu}) = \{\boldsymbol{\theta} \in \mathbb{R}^d : \boldsymbol{\mu} = \Delta\boldsymbol{\theta}, \boldsymbol{\theta} \geq 0\}. \tag{2.7}$$

Then we have the following definition for a weakly sparse vector.

Definition 4. *Let $\boldsymbol{\mu} \in \mathbb{C}$ where $\mathbb{C}$ is a closed convex polyhedral cone with vertex representation given by $\mathbb{C} = \{\boldsymbol{x} \in \mathbb{R}^n : \boldsymbol{x} = \boldsymbol{\Delta b}$ for some $\boldsymbol{b} \in \mathbb{R}^d_+\}$. Then $\boldsymbol{\mu}$ is weakly sparse if $\exists \boldsymbol{b} \in S(\boldsymbol{\mu})$ such that $\{i : b_i > 0\}$ corresponds to an adjacent set of an extreme ray in the adjacency graph of $\mathbb{C}$ where $S(\boldsymbol{\mu})$ is defined in (2.7).*

We propose adjacency prior based generalization of the Horseshoe or Strawderman-Berger priors as

$$\begin{aligned} \pi_1, \ldots, \pi_d &\sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_d) \\ u &\sim \text{Multinomial}(1, \pi_1, \ldots, \pi_d) \\ \boldsymbol{b}|u &\sim \pi(\boldsymbol{b}_{E_u}) \end{aligned} \tag{2.8}$$

where $\pi(\boldsymbol{b}_{E_u})$ can be $\pi_{HS}(\boldsymbol{b}_{E_u})$ or $\pi_{SB}(\boldsymbol{b}_{E_u})$. This is different from using a prior like modified lasso such as fused lasso (Tibshirani and Saunders, 2005) type selection, since we select only 1 adjacency set to stay on the surface whereas in fused lasso several clusters maybe selected and hence the results vectors may have dense representations.

2.4 Numerical Results

2.4.1 Distribution of points in a 3D cone

Figure 2.5 shows the distribution on 10000 points drawn from the Horseshoe kind prior on polyhedral cone. While most points lie near the face of the cone including the vertex, there are still many points in the interior of the cone. The 2D contour has been plotted by considering equal volume of circular cones inside the polyhedral cone and then calculating the relative frequency of 10000 points. The points very close to the vertex are included in the outermost region since they are anyway sparse for being close to the vertex. From the 2D contour, it is clear that there is a heavy positive mass in the interior most circle.

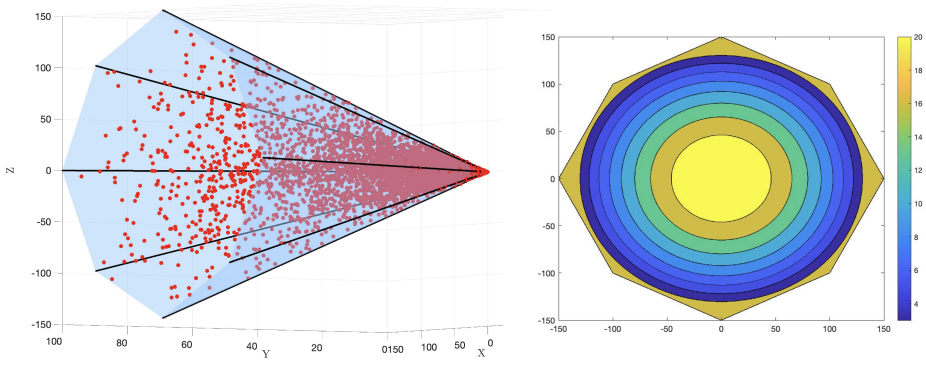

Figure 2.5: Plot showing points inside a 3D polyhedral cone by invoking a Horseshoe prior on b (left) and 2D contour of the cone showing the concentration of 10000 such points (right).

Figure 2.6 and 2.7 presents the points inside the 3D polyhedral cone and the reciprocal 2D contour for Horseshoe prior on adjacent set and on maximal clique,

respectively. The figures show some positive mass and no positive mass in the interior of the cone for the two cases. All points are either closer to or lie exactly on the lower dimension features be it vertex, extreme rays or facet.

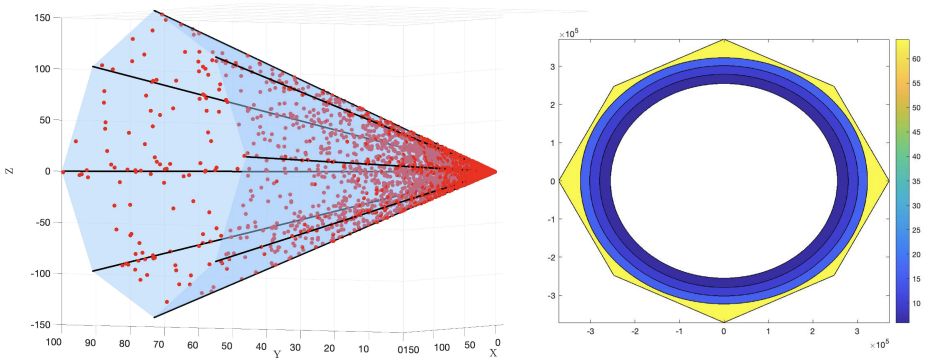

Figure 2.6: Plot showing points inside a 3D polyhedral cone by incorporating adjacency of extreme rays (left) and 2D contour of the cone showing the concentration of 10000 such points (right).

2.4.2 *max-min* distance of points from facet

In this numerical study, we consider polyhedral cone in different dimensions and simulate $R = 100000$ points using the three different priors discussed in the previous section. For a fair comparison, for each of the adjacent set E_j chosen, we select randomly $|E_j|$ rays so that $\pi(u) \sim \frac{1}{\binom{d}{|E_j|}}$ and the rest of the prior specifications are same as in horseshoe prior with adjacency(2.8). We also compare the results when the prior is defined using maximal clique set. We report the scaled max-min distance where the maximum is over R number of repetitions and minimum is considered with respect to the point's distance from the m facets. That is, we see

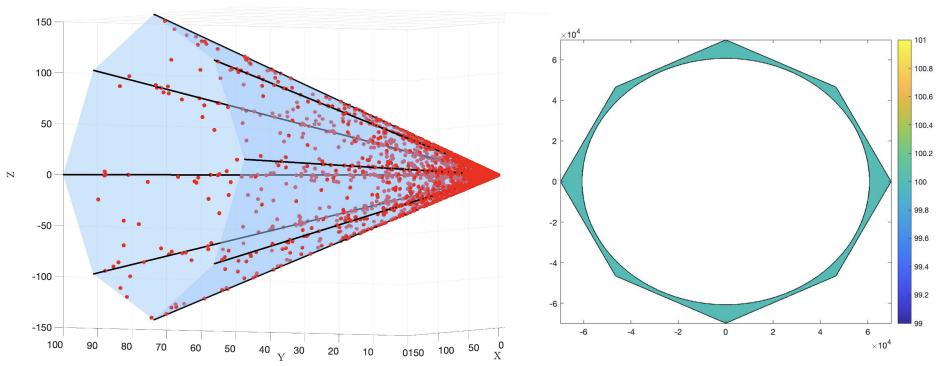

Figure 2.7: Plot showing points inside a 3D polyhedral cone by incorporating maximal cliques of extreme rays (left) and 2D contour of the cone showing the concentration of 10000 such points (right).

n	m	d	d_{ij}	d^r_{ij}	d^a_{ij}	d^c_{ij}
3	6	6	3.081	3.081	0.891	0
8	11	16	1.096	0.924	0.309	0
10	13	20	0.818	0.376	0.159	0

Table 2.1: Max-min distance of points from different priors

which hyperplane of the cone it is closest to.

Let $d_{ij} = \max_{i=1:R} \min_{j=1:m} \text{distance}_{ij}$. Table 2.1 reports the max-min distance for Horseshoe prior d_{ij}, Horseshoe prior on a randomly selected set d^r_{ij}, Horseshoe prior with adjacency d^a_{ij} and Horseshoe prior on cliques d^c_{ij}. As expected, the points always lie of the boundary for cliques and they are closer to the boundary for adjacent sets rather than for a randomly selected set of the same size.

2.5 Application

We discuss two examples in details. For the positive isotonic function estimation, we explain both the parametric and non-parametric approach. For the bell-shaped function, we show results for the parametric approach and additionally discuss how the non-parametric fit can be obtained.

2.5.1 Positive Isotonic Function

We consider the mean function $f(x) = \exp(x)$ over the interval $[-2, 2]$, a positive isotonic function so that A is $n \times n$ matrix with $A_{1,1} = 1$, $A_{i,i} = -1$, $A_{i,i+1} = 1$ for $i = 2, \ldots, n$. Hence the Bayes estimator $\hat{\boldsymbol{\mu}}$ is obtained using MCMC by invoking a Horseshoe kind prior and Strawderman-Berger kind prior on $\boldsymbol{b}$ based on the model

$$\boldsymbol{\mu} = \Delta \boldsymbol{b}.$$

Figure 2.8 shows the plot of the estimators for the priors along with the MLE.

In the non-parametric approach, we model $f(x) = \boldsymbol{\Psi}(x)\boldsymbol{\beta}$ where $\boldsymbol{\Psi}(x)$ is the p dimensional basis function at x. This will produce a flexible and smooth estimate depending on the choice of p. To enforce the monotonicity of $f(x)$, we consider a set of fine grid points $t_1 < \cdots < t_m$ over the range of x and construct A such that the i^{th} row of A is the derivative of the basis functions $\boldsymbol{\Psi}'(x)$ evaluated at t_i. These constraints are then applied on the coefficients parameter such that $A\boldsymbol{\beta} \geq \mathbf{0}$ where A is a $m \times p$ matrix. Specifically, we consider cubic B-splines with no intercept

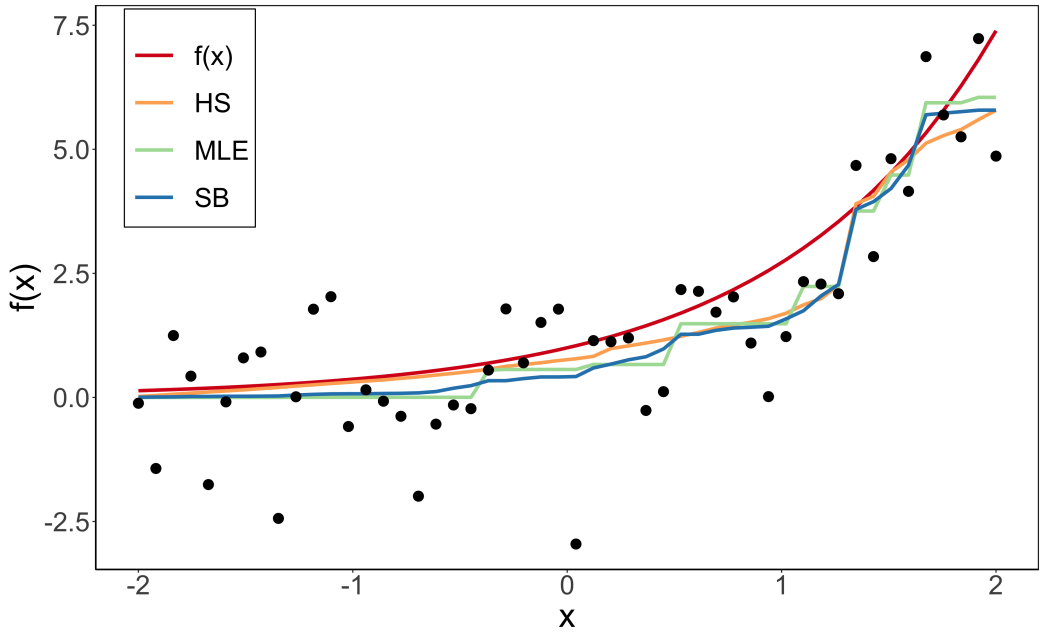

Figure 2.8: Bayes estimates for Horseshoe prior (HS), Strawderman-Berger prior (SB) and MLE for $n = 50$ points from $f(x) = \exp(x)$.

and $k = 3$ equidistant internal knots so that $p = 6$ [34]. We consider $m = 8$ equidistant grid points and since the number of constraints is greater than the number of parameters, the number of extreme rays $d = 18$ is greater than p. Similar to the parametric approach, $\hat{\boldsymbol{f}}$ is obtained using MCMC by invoking priors on $\boldsymbol{b}$ through the model

$$\boldsymbol{f} = \boldsymbol{\Psi\beta} = \boldsymbol{\Psi}\Delta\boldsymbol{b} = \tilde{\Delta}\boldsymbol{b}.$$

Figure 2.9 presents the results from all four priors, the Horseshoe kind estimator and Strawderman Berger kind estimator as well as the priors incorporating adjacency. As expected, all four the estimators are smoother compared to ones obtained by parametric approach. For the priors incorporating adjacency, Figure 2.10 demonstrates the d estimates based solely on one of the d adjacency sets for Horseshoe

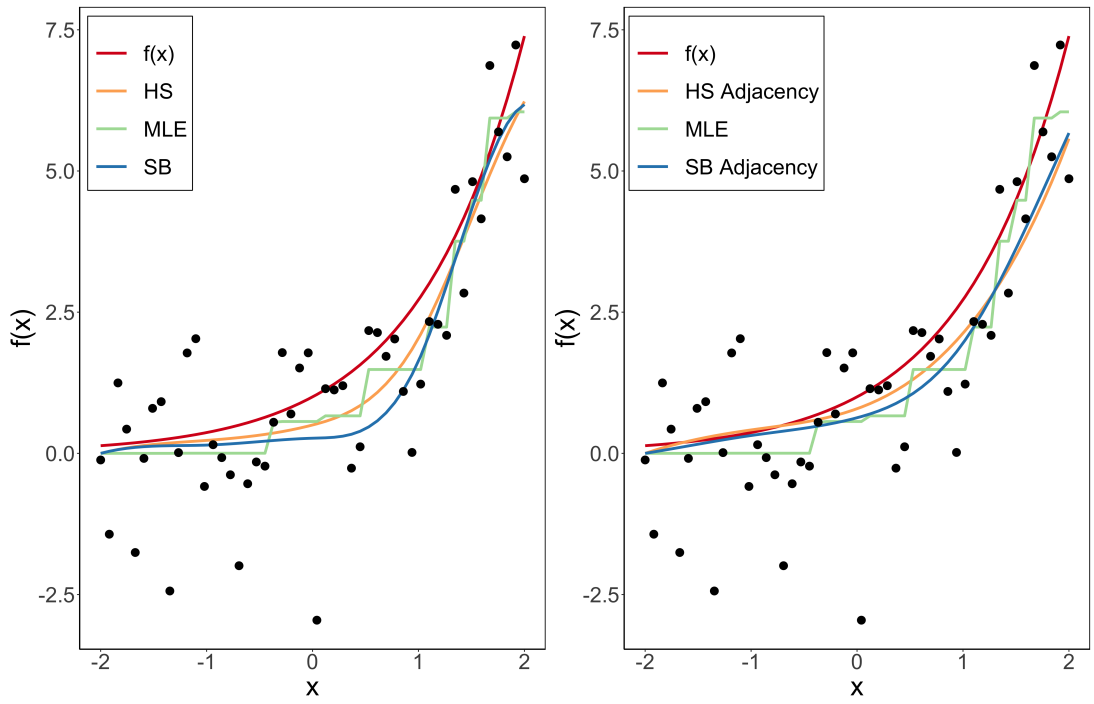

Figure 2.9: Bayes estimates using cubic B-spline with 3 internal knots for $n = 50$ points from $f(x) = \exp(x)$. Horseshoe prior (HS) and Strawderman-Berger prior (SB) (left) and Horseshoe prior (HS adjacency) and Strawderman-Berger prior with adjacency (SB adjacency) (right).

kind prior (left) and for Strawderman-Berger kind prior (right). The final estimates for the priors invoking adjacency are an average of these d estimators presented in the right panel of Figure 2.9 since all these adjacent sets appear with almost equal frequency in the mcmc chains.

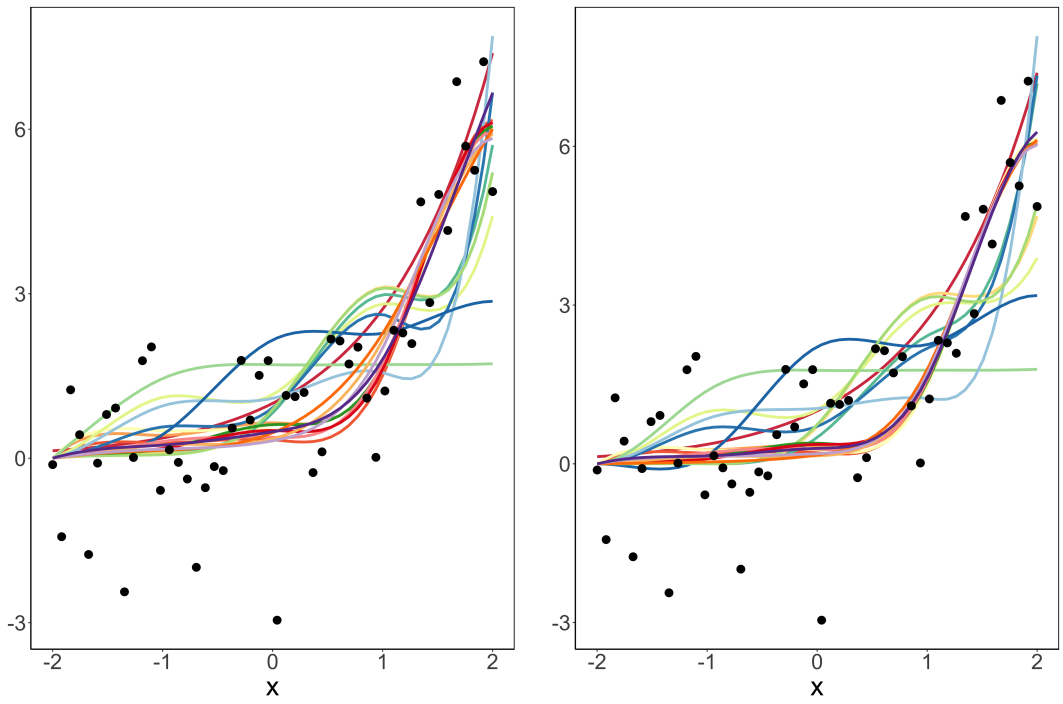

Figure 2.10: 18 estimates from the each of the $d = 18$ adjacency set for Horseshoe prior with adjacency (left) and Strawderman-Berger prior with adjacency (right) for $n = 50$ points from $f(x) = \exp(x)$.

2.5.2 Bell-shaped Function

In this example, we consider estimation of a symmetric bell-shaped curve. Given the inflection points k_1 and k_2, A is a $(n + 2) \times n$ matrix based on the constraints that the function is positive, increasing on the left, convex, concave, convex and then decreasing at the right. We consider the true mean function $f(x)$ to be a normal density scaled to have large values, i.e., $f(x) = 50\ \frac{1}{\sigma}\phi\left(\frac{x}{\sigma}\right)$ for $n = 40$ points over x in $[-2, 2]$. The estimated mean functions are obtained by invoking priors on $\boldsymbol{b}$ using the model $\boldsymbol{f} = \Delta\boldsymbol{b}$ where the number of extreme rays d becomes super large and is equal to 2551 for $n = 40$. The results are shown in Figure To-

be-inserted. Similar to the MLE, both simple Strawderman-Berger prior and the simple Horseshoe prior are piece-wise linear functions (left in Figure To-be-inserted). Figure 2.12 provides 18 estimates out of $d = 2551$ estimates one for each of the d extreme sets for the priors incorporating adjacency. Since, each of these sets appear almost equally in the mcmc, we take an average of the 2551 estimates to obtain the final estimate for both the priors using adjacency.

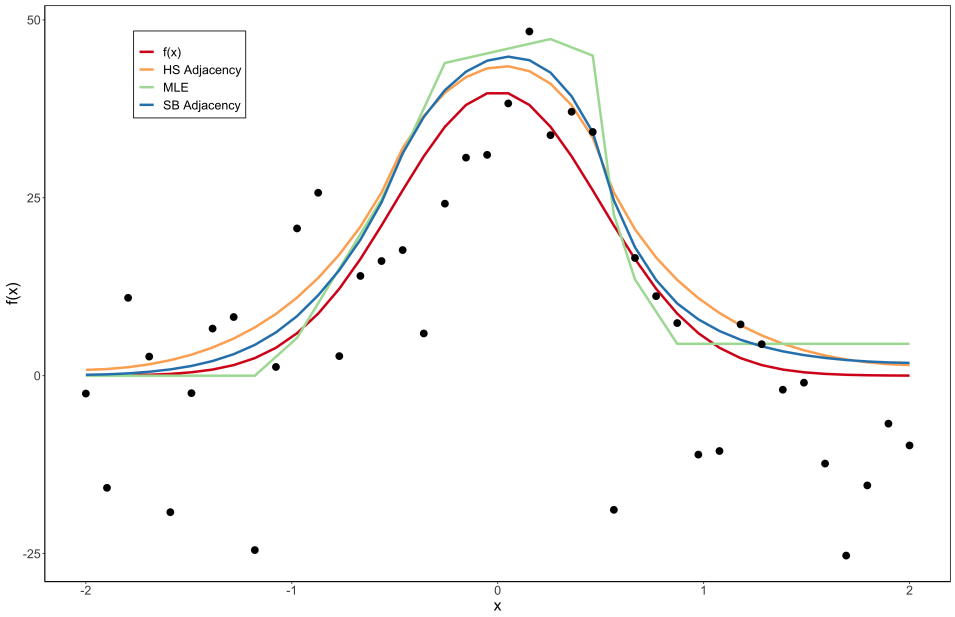

Figure 2.11: Bayes estimates for Horseshoe prior and Strawderman-Berger prior with adjacency for $n = 40$ points from $f(x) = 50\ \frac{1}{\sigma}\phi\big(\frac{x}{\sigma}\big)$.

2.6 Discussion

In this chapter, we have have introduced new priors on high-dimensional closed convex cone where most of the mass is on lower dimensional sets on the boundary. The priors facilitate Bayesian estimation of constrained priors. While the motivating

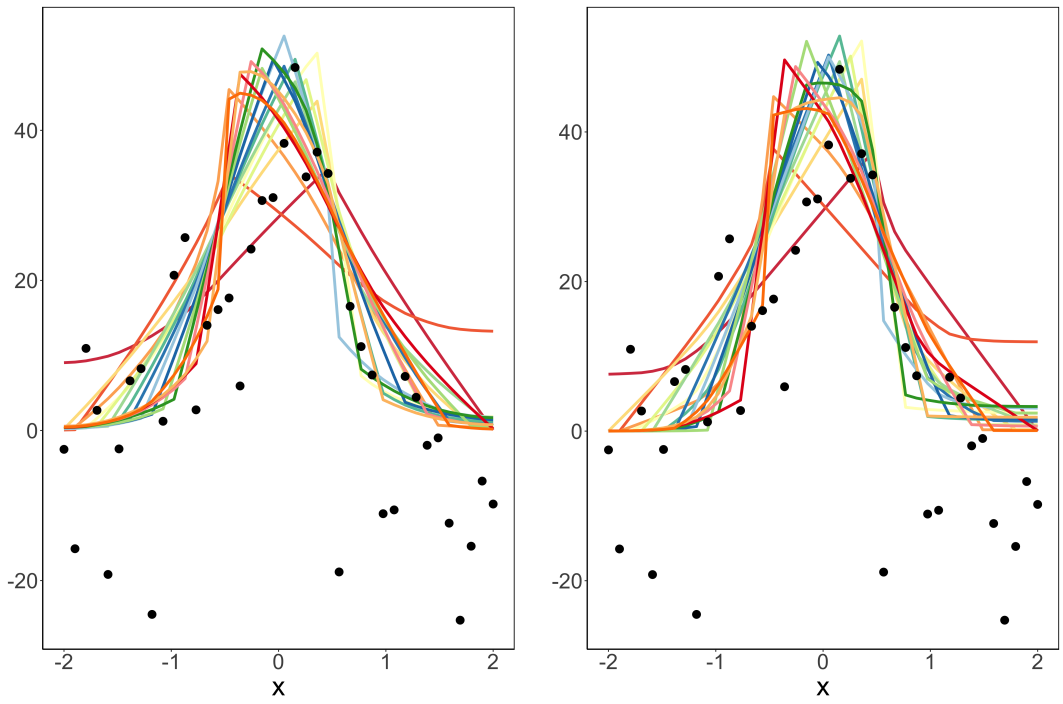

Figure 2.12: 18 **out of** $d = 2551$ **estimates from the each of the** 18 **adjacency set for Horseshoe prior with adjacency (left) and Strawderman-Berger prior with adjacency (right) for** $n = 40$ **points from** $f(x) = 50\ \frac{1}{\sigma}\phi\big(\frac{x}{\sigma}\big)$**.**

example is estimation of a constrained normal mean vector, the application of non-parametric estimation of shape-restricted functions show that the priors can easily applied to a regression model. In fact, it can be used for inference for any parameter vector with linear inequality constraints. For now, we have shown applications with inequality restrictions on the parameters but the notion of sparsity is related to having several of the inequalities reducing to equality in the true value of the parameter. While in the present set up these equality constraints are not necessarily binding, many examples where equality constraints are present as hard constraints in addition to inequality constraints can be also be incorporated in the proposed

method. Another interesting application of our work is testing for $H_0 : \boldsymbol{A\mu} = \mathbf{0}$ versus $H_1 : \boldsymbol{A\mu} \geq \mathbf{0}$ using Bayesian model comparison. When $\boldsymbol{A} = I$, the problem reduces to testing origin against non-negative orthant and the Likelihood Ratio Test is much easier to compute than for a general $\boldsymbol{A}$. The projection of the data vector to the cone will also lie on one of the lower dimensional faces and is the maximum likelihood estimator. In general the projection maybe hard to compute, but in principle the Bayesian posterior should concentrate around the Euclidean projection. Bayesian recovery results for the true clique and posterior concentration results need ot be investigated.

Chapter 3: Gaussian mean testing under linear inequality constraints in high dimension

3.1 Introduction

Hypothesis testing is one of the most important tools in the arsenal of modern empirical scientists that is used to support new scientific claims. Given that parameter restrictions are natural in most scientific settings, constrained hypothesis testing procedures are important in a wide range of scientific applications such as in genetic studies, social network analysis, drug testing. Some common examples constrained hypothesis are constraints of non-negativity or monotonicity in treatment effects, linear inequality constraints of regression parameters.

Of the different testing problems that arise in modern applications, testing for mean (vector) of a normal distribution is probably the most common. Several constrained testing problems can be reduced to testing constraints on the mean of a normal population. In this dissertation we investigate testing for normal mean under the restriction that it belongs to a convex polyhedral cone. The novelty in our approach is how the procedures handle high dimensional problems with possibly sparse configurations of parameters of interest. Specifically, suppose $\mathbf{X} \sim \mathbf{N}_n(\boldsymbol{\mu}, \Sigma)$

with n possibly much larger than those studied in traditional constrained mean testing problems. The general problem of testing in convex cones can be formulated as

$$\mathcal{H}_0 : \boldsymbol{\mu} \in \mathcal{M} \quad vs. \quad \mathcal{H}_1 : \boldsymbol{\mu} \in \mathcal{C} \backslash \mathcal{M}. \tag{3.1}$$

where $\mathbb{C}$ is a closed convex cone. The null set $\mathcal{M}$ could be a linear subspace contained in $\mathbb{C}$ e.g. the singleton set consisting of the apex of the cone, usually the origin or it could be the boundary of the cone. In the second case, the alternative set is the open interior of the cone $\mathbb{C}$.

As a popular and specific example of the testing problem, consider the one-sided testing problem for a multi-dimensional normal mean. Let the convex cone be the positive orthant in $\mathbb{R}^n$, denoted by $\mathcal{K}$. Then the null set could be the singleton $\mathcal{M} = \{0\}$ or it could be consisting of the boundary of the orthant $\partial\mathcal{K}$, i.e. faces of the orthant where at least one of the coordinates is zero. Using the convention "$\boldsymbol{\mu} \geq (\leq)\mathbf{0}$" to imply that "all coordinates of $\boldsymbol{\mu}$ are non-negative (non-positive) with at least one strictly positive (negative) coordinate" and "$\boldsymbol{\mu} > (<)\mathbf{0}$" to indicate that all coordinates of $\boldsymbol{\mu}$ are strictly positive (negative), the two basic forms of mean testing problem are

1. Singleton null set:

$$\mathcal{H}_0 : \boldsymbol{\mu} = \mathbf{0} \quad vs. \quad \mathcal{H}_1 : \boldsymbol{\mu} \geq \mathbf{0}. \tag{3.2}$$

2. Composite null set:

$$\mathcal{H}_0 : \boldsymbol{\mu} \leq \mathbf{0} \quad vs. \quad \mathcal{H}_1 : \boldsymbol{\mu} > \mathbf{0}. \tag{3.3}$$

In the second testing problem (3.3), the null hypothesis consists of the boundary of the positive orthant and is a composite null hypothesis. This is a more involved problem with a long history [35–38] but even in moderate dimension there are no tests that perform uniformly better than other.

In this dissertation we focus on the first kind of testing problem (3.2) which is a relatively simpler problem among the two scenarios, albeit the complications due to high dimension of the parameters. Traditionally, the problem (3.2) has been addressed for small to moderate dimensional mean vectors. There are recent work on applying high-dimensional tools to the one-sided testing problem. However, the tests mostly do well in specific regimes of sparseness, either when the alternative values are 'sparse' vectors or they are 'dense' vectors. The notions of sparseness and denseness of the vectors are traditionally related to how many zeros there are in the vectors. However, when the vectors are constrained to a closed set, the concept of sparsity can be more nuanced and will be made more specific later. In general, accommodating prior information about the parameter in the form of constraints into the testing framework will provide better inference. For example, suppose $X \sim N(\mu, 1)$ and consider testing the following two hypotheses: (1) $H_0 : \mu = 0$ vs $H_1 : \mu \neq 0$ and (2) $H_0 : \mu = 0$ vs $H_1 : \mu > 0$. For the non-directional alternative, an appropriate test is to reject H_0 when $|\bar{x}|$ is relatively large whereas for the directional alternatives in the positive orthant, it seems reasonable to reject H_0 when $\bar{x}$ is relatively large. Under the assumed normal model, $\bar{x} \sim N(\mu, 1/n)$ and the critical value of the test based on $\sqrt{n}|\bar{x}|$ is bigger than that based on $\sqrt{n}\bar{x}$. For values of μ close to zero, the test based on $\sqrt{n}|\bar{x}|$ will reject less often than the other

one and hence loose power under the alternative. In high dimensional setting, when the alternative positive vector may have only a few large positive values, the average mean vector can be very close to zero, resulting in unreasonably low power for the non-directional test statistic. These two aspects of the testing problem, namely, the restriction $\boldsymbol{\mu} \in \mathbb{C}$ and also $\boldsymbol{\mu}$ being possibly of very high dimension, make it challenging.

The motivation of the present work is that we want to develop a test that have reasonable power for most alternatives when the alternative is constrained to a convex polyhedral cone. While in the original problem (3.2) the alternative cone is the positive orthant $\mathcal{K}$, we still need to consider the situation of general convex polyhedral cone as the alternative. This is because for a general positive-semidefinite covariance matrix, the problem can be converted to the identity covariance matrix scenario under an alternative of a general polyhedral cone. Thus, we formally state the general conic alternative problems under the specific normal mean testing setting:

1. Singleton null set: $\mathcal{M} = \mathbf{0}$.

$$\mathcal{H}_0 : \boldsymbol{\mu} = \mathbf{0} \quad vs. \quad \mathcal{H}_1 : \boldsymbol{\mu} \in \mathbb{C}\backslash\{\mathbf{0}\}. \tag{3.4}$$

2. Composite null set: $\mathcal{M} = \partial\mathbb{C}$.

$$\mathcal{H}_0 : \boldsymbol{\mu} \in \partial\mathbb{C} \quad vs. \quad \mathcal{H}_1 : \boldsymbol{\mu} \in \mathbb{C}^0. \tag{3.5}$$

where $\mathcal{C}^0$ is the open interior of $\mathcal{C}$. As mentioned before, we pursue only the testing problem (3.4) in this dissertation under a high dimensional setting. We want to

develop an omnibus test to detect if $\boldsymbol{\mu} \in \mathbb{C}\backslash\{\mathbf{0}\}$ with reasonable power over the entire set.

We mention a few application areas where situations like these might arise in high dimensional settings:

- Consider an matched pair design to determine if a new treatment is effective in gene expression data. Let $\mu_1, \ldots, \mu_n$ denote the n dimensional difference in mean responses between the treatment and placebo. Suppose further it is known that the new treatment is expected to be at least as good as the placebo. The inference problem for concluding that the new treatment is better than the placebo can be phrased as:

$$\mathcal{H}_0 : \mu_1 = \cdots = \mu_p = 0 \quad vs. \quad \mathcal{H}_1 : \mu_i \geq 0 \text{ for } i = 1, \ldots, n \text{ and } ||\boldsymbol{\mu}||^2 > 0.$$

- The second example is motivated from Lee *et al.* (2008) [39]. The data arise from a randomized study where measurements of tidal volume, that is, the volume of gas exchanged during each ventilated breath, are taken on a number of individuals subject to interventions that may induce panic attacks. The focus of this study was to develop a procedure to determine if one mean curve dominates the other in a high dimensional setting. In order words, prior to the study the study investigators had an ordered mean hypothesis of the type $f_1(t) \geq f_2(t)$ for all t (and $f_1(t) > f_2(t)$ for at least one t), with f_i the mean curve for group $i = 1, 2$ and with groups 1 and 2 defined by the two interventions. The null hypothesis would be that the mean curves are identical at all time points, i.e., $f_1(t) = f_2(t)$ for all t vs $f_1(t) \geq f_2(t)$ with

strict inequality for at least one time point.

The literature on multivariate one-sided tests dates back to Bartholomew (1959, 1961) and Kudo (1963) who studied the likelihood ratio test for non-negative orthant assuming a known covariance matrix [40–42]. Later Perlman (1969) extended his work to the case of unknown covariance matrix [43]. For global alternative $\boldsymbol{\mu} \neq 0$, LRT has the asymptotic most stringent and best average power properties. However the (asymptotic) optimality properties of the LRT have yet to be fully understood for restricted alternatives. Furthermore, for a general covariance matrix, LRT is difficult to implement for practical purposes because of its computational burden which is discussed in details later. Among other the popular tests, Shi and Kudo (1987) gave a method for obtaining the optimal linear test (OLT) statistic, also referred to as most stringent and somewhere most powerful (MSSMP) test [44]. The OLT statistic has a normal distribution and is advantageous for computing critical level and power. In addition, OLT is most powerful for alternatives in a certain direction. Tsai and Sen (1993) showed that OLT is uniformly locally more powerful than the LRT [45]. Tang et al (1989) proposed an approximate LRT for non-negative orthant that is computationally more feasible [46]. A simple and practical one-sided test proposed by Follmann rejects $\mathcal{H}_0$ when $\boldsymbol{X}^T\Sigma^{-1}\boldsymbol{X}$ exceeds its 2α critical value and $\sum_{i=1}^{n} X_i > 0$ where $\boldsymbol{X}^T\Sigma^{-1}\boldsymbol{X}$ has a chi-squared distribution with n degrees of freedom [47]. The idea is to draw more power from not rejecting if the data is negative on average. However, all of these tests are designed for low dimensional problems.

With rapid advancement in technology and computerization, data are collected at faster rates and at higher resolutions with staggeringly high number of features. For example, healthcare data is notorious for having large number of variables. Some other fields where high dimensional data is prevalent include high genomics, frequency trading, brain imaging. In high dimensional setting, often one deals with the case when only few variables are relevant. Thus it has become increasingly important to identify true signals as the data tends to be sparse. Sparse-signal detection is even more difficult when there is no prior knowledge on sparsity.

One of the most popular estimator in this setting is Lasso estimator which performs both variable selection and regularization [3]. Another prominent approach for sparse-signal detection is horseshoe estimator whose major strengths are adaptivity to unknown sparsity as well as unknown signal-to-noise ratio, robustness to large, outlying signals and multiplicity control [10]. Javanmard and Lee (2019) have developed a framework for testing very general hypotheses regarding the model parameters assuming sparsity and approximate sparsity structure [48]. More recent works include Yu *et al.* (2019) in which they provided an two-step algorithm to the testing problem (2.2) under sparsity assumption [49]. Wei *et al.* (2018) considered hypothesis testing on whether the parameters lie on some closed convex cone [50]. They provided a sharp characterization of the generalized likelihood ratio test (GLRT) testing radius up to a universal multiplicative constant in terms of the geometric structure of the underlying convex cones. They further showed that GLRT is optimal in testing alternatives of positive orthants in the sense that there is no other test that can discriminate between the null and the alternative for smaller

separations.

Among the available tests, GLRT is the most natural choice as an omnibus test and hence we study it in details in the next section.

3.2 Generalized Likelihood Ratio Test (GLRT)

The likelihood ratio test (LRT) or sometimes referred to as the Generalized Likelihood Ratio Test (GLRT), defined as the negative two times the logarithm of the ratio of the likelihood at the null value to the likelihood at the maximum likelihood estimator (MLE), is the most obvious omnibus test that one can use in the testing problem (3.2). The GLRT can be thought of as the Neyman Pearson statistic where the alternative value is being estimated by the MLE. The GLRT is known to have better overall performance in a variety of testing scenarios and often reduces to finite sample optimal tests when such tests exist. However, the optimality property are usually found in scalar parameter cases where concepts such as monotone likelihood ratio that guarantees optimality of tests under composite hypothesis may be available. In the multi-parameter testing the reliance on GLRT is even more stark due to its large sample optimality, particularly in the setting of multivariate normal distribution (Hsieh, 1979) [51].

The GLRT statistic for the testing problem (3.4) based on data $\boldsymbol{X} \sim N(\boldsymbol{\mu}, \boldsymbol{I})$ when $\mathbb{C}$ is a polyhedral cone is

$$T = ||P_{\mathcal{C}}\boldsymbol{X}||^2$$

where $P_{\mathbb{C}}\boldsymbol{X}$ is the least-squares projection of $\boldsymbol{X}$ onto $\mathbb{C}$ defined as

$$P_{\mathbb{C}}\boldsymbol{X} = \underset{\boldsymbol{\mu}\in\mathbb{C}}{\operatorname{argmin}} \|\boldsymbol{X} - \boldsymbol{\mu}\|_2^2 \tag{3.6}$$

It is also the MLE of $\boldsymbol{\mu}$ under the normal model. The test statistic T is called the Chi-bar-squared ($\overline{\chi^2}$) test statistic and the distribution of the likelihood-ratio statistic under the null hypothesis has been worked out; see Sen and Silvapulle p.63-83 [26]. The null distribution is not exactly a chi-squared distribution, but instead has the form

$$pr[T \geq t] = \sum_{m=0}^{k} P(m, \mathbb{C}) P[\chi_m^2 \geq t].$$

where χ_m^2 is a standard chi-squared variable with m degrees of freedom. The quantity $P(m, \mathbb{C})$ is the probability that the maximum-likelihood estimate of $\boldsymbol{\mu}$ under the alternative hypothesis belongs to one of the m-dimensional faces of $\mathbb{C}$ under the assumption that the null hypothesis is true. This distribution is called a *chi-bar-squared* distribution.

For illustration consider the GLRT when $n = 2$. Suppose $\mathbf{X} \sim \mathbf{N}_2(\boldsymbol{\mu}, \mathbf{I})$ where $\boldsymbol{\mu} = (\mu_1, \mu_2)^T$. We want to find the MLE based on one observation on $\mathbf{X}$ and subject to the constraint $\boldsymbol{\mu} \in \mathcal{K}$ where $\mathcal{K}$ is the positive orthant in $\mathbb{R}^2$ defined as $\{(\mu_1, \mu_2)^T : \mu_1 \geq 0, \mu_2 \geq 0\}$. The likelihood is

$$L(\boldsymbol{\mu}|\boldsymbol{X}) = \frac{1}{(2\pi)^{n/2}} \, e^{-\frac{1}{2}||\boldsymbol{X}-\boldsymbol{\mu}||^2}$$

and $-2l(\boldsymbol{\mu})$ is equal to the squared distance between $\mathbf{X}$ and $\boldsymbol{\mu}$. Thus, maximizing the likelihood is the same as the constrained least squares problem (3.6) and the

MLE is the projection of $\boldsymbol{X}$ onto $\mathcal{K}$ (figure 3.1) and is given by

$$\hat{\boldsymbol{\mu}} = \| P_{\mathcal{K}}\mathbf{X} \|_{=} \begin{cases} (x_1, x_2), & \text{for } x_1 > 0, x_1 > 0, \\ (x_1, 0), & \text{for } x_1 > 0, x_1 < 0, \\ (0, x_2), & \text{for } x_1 < 0, x_1 > 0, \\ (0, 0), & \text{for } x_1 < 0, x_1 < 0 \end{cases}$$

where $P_{\mathcal{K}}\boldsymbol{X}$ the least-squares projection of X onto $\mathcal{K}$. The LRT for testing $\mathcal{H}_0 : \boldsymbol{\mu} =$

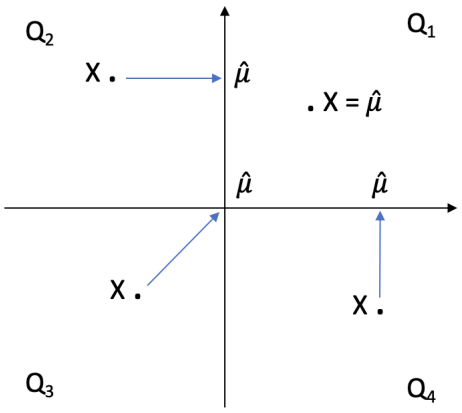

Figure 3.1: Projection of $\boldsymbol{X}$ onto $\mathcal{K}$

$\mathbf{0}$ *vs.* $\mathcal{H}_1 : \boldsymbol{\mu} \geq \mathbf{0}$ is

$$LRT = \|\boldsymbol{X}\|^2 - (\|\boldsymbol{X} - \hat{\boldsymbol{\mu}}\|^2 = \|\hat{\boldsymbol{\mu}}\|^2.$$

This implies

$$\text{LRT} = \begin{cases} x_1^2 + x_2^2, & \text{for } x_1 > 0, x_1 > 0, \\ x_1^2, & \text{for } x_1 > 0, x_1 < 0, \\ x_2^2, & \text{for } x_1 < 0, x_1 > 0, \\ 0, & \text{for } x_1 < 0, x_1 < 0. \end{cases}$$

The LRT $= \|P_{\mathcal{K}}\boldsymbol{X}\|^2$ is called the chi-bar-squared $\overline{\chi^2}$ test statistic. The distribution of $X_1^2 + X_2^2$, under $\mathcal{H}_0$ is no longer χ_2^2 because of the constraints $X_1 > 0, X_2 > 0$. However using circular symmetry of $N(0, I)$, we can say that the direction and length of X are statistically independent. Also for $N(0, I)$, $P(X_1 > 0, X_2 > 0)$ depends on the angle of the corresponding cone made at its vertex. Thus, under $\mathcal{H}_0$,

$$\text{LRT} = \begin{cases} x_1^2 + x_2^2, & \text{given } x_1 > 0, x_1 > 0 \sim \chi_2^2, \\ x_1^2, & \text{given } x_1 > 0, x_1 < 0 \sim \chi_1^2, \\ x_2^2, & \text{given } x_1 < 0, x_1 > 0 \sim \chi_1^2, \\ 0, & \text{given } x_1 < 0, x_2 < 0 \sim \chi_0^2. \end{cases}$$

and the distribution of LRT under $\mathcal{H}_0$ is

$$\begin{aligned} P(\text{LRT} \leq c) &= \sum P(\text{LRT} \leq c \;\&\; \mathbf{X} \in \mathbf{Q}_i) \\ &= \sum P(\text{LRT} \leq c \mid \mathbf{X} \in \mathbf{Q}_i) P(\mathbf{X} \in \mathbf{Q}_i) \\ &= \frac{1}{4} + \frac{1}{2} P(\chi_1^2 \leq c) + \frac{1}{4} P(\chi_2^2 \leq c). \end{aligned} \tag{3.7}$$

where $\mathbf{Q}_1, \mathbf{Q}_2, \mathbf{Q}_3, \mathbf{Q}_4$ denote the four quadrants in the two-dimensional plane. Under $\mathcal{H}_0$, LRT is in fact a mixture of χ^2 distributions with weights being the probability that $\boldsymbol{X}$ falls in the corresponding cone $\mathbf{Q}_i$. The power of the Chi-bar test statistic is

$$\begin{aligned} P_{\mathcal{H}_1}(LRT > c) &= \sum P_{\mathcal{H}_1}(LRT > c \;\&\; \mathbf{X} \in \mathbf{Q}_i) \\ &= \sum P_{\mathcal{H}_1}(LRT > c \mid \mathbf{X} \in \mathbf{Q}_i) P(\mathbf{X} \in \mathbf{Q}_i). \end{aligned} \tag{3.8}$$

Under $\mathcal{H}_1, (X_1^2 + X_2^2)|(X_1 > 0, X_2 > 0)$ is not χ^2 anymore and power is computed numerically. In general, T reduces to $\sum_{i=}^{n} X_i^2 I(X_i > 0)$ where n is the dimension.

For the simplest case illustrated above the form and the null distribution of the GLRT can be computed exactly. The GLRT continues to enjoy computational tractability even in higher dimension when the cone of interest $\mathbb{C}$ is the positive orthant $\mathcal{K}$. However, for a general polyhedral cone exact derivation of the projection may not be feasible which leads us to the next part.

Consider a general polyhedral cone $\mathbb{C}$ defined as an intersection of hyperplanes

$$\mathbb{C} = \{\boldsymbol{\mu} \in \mathbb{R}^n : \boldsymbol{A\mu} \geq \mathbf{0}\} \tag{3.9}$$

where the inequality is defined as element-wise and A is an $m \times n$ matrix. For finitely generated polyhedral cones such as (3.9), each vector in the cone can be written as a positive linear combination of a finite number of its extreme rays. Assuming A is full row rank, a minimal set of extreme rays which can generate the entire cone through positive linear combination is given by the columns of $A^T(AA^T)^{-1}$. The polar cone of $\mathbb{C}$ is then given by

$$\mathbb{C}^0 = \{\boldsymbol{x} \in \mathbb{R}^n : \langle \boldsymbol{x}, \boldsymbol{a}_i \rangle \leq 0, i = 1, \ldots, n\}$$

where $\boldsymbol{a}_i$ is the ith column of $A^T(AA^T)^{-1}$ and $\langle \cdot, \cdot \rangle$ is the usual inner product in $\mathbb{R}^n$. The polar cone is then finitely generated by the $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_m$, the columns of $B = -A^T$. For any vector $\boldsymbol{X} \in \mathbb{R}^n$, Moreau's decomposition theorem guarantees a unique decomposition

$$\boldsymbol{X} = \sum_{i=1}^{n} \alpha_i \boldsymbol{a}_i + \sum_{j=1}^{m} \beta_j \boldsymbol{b}_j$$

where $\alpha_1, \ldots, \alpha_n, \beta_1, \ldots, \beta_m$ are all nonnegative [52]. Then the projection of $\boldsymbol{X}$ to $\mathbb{C}$ is

$$P_{\mathbb{C}}(\boldsymbol{X}) = \sum_{i=1}^{n} \alpha_i \boldsymbol{a}_i.$$

For the case where the cone of interest is $\mathcal{K}$, the positive orthant, the matrix A is the n dimensional identity matrix and hence the columns $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n$ are the canonical vectors and the columns $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_n$ are the negative of the canonical representation. Hence, the coefficients are $\alpha_i = sgn(X_i)$ and $\beta_j = -sgn(X_j)$. Thus, the projection to the n dimensional first orthant of a genral vector $\boldsymbol{X}$ is $\boldsymbol{X}_+$, the vector with ith entry as $\max\{X_i, 0\}$. While conceptually simple, finding the positive linear combination for evaluating the projection is N-P hard. Thus extensions of the GLRT beyond the positive orthant often becomes prohibitive.

As mentioned earlier, even if the original test is $\boldsymbol{\mu} = \boldsymbol{0}$ versus $\boldsymbol{\mu} \geq 0$, if the the model is $\boldsymbol{X} \sim N(\boldsymbol{\mu}, \Sigma)$ for a general covariance matrix Σ with inverse Ω, the problem is generally reduced by the transformation $A\boldsymbol{X} \sim N(A\boldsymbol{\mu}, \boldsymbol{I})$ where A is such that $A\Sigma A^T = \boldsymbol{I}$ and the new mean $\boldsymbol{\theta} = A\boldsymbol{\mu}$ belongs to the cone

$$\mathbb{C} = \{\boldsymbol{\theta} \in \mathbb{R}^n : A^{-1}\boldsymbol{\theta} \geq \boldsymbol{0}\}. \tag{3.10}$$

To compute the GLRT we need to compute the projection of $A\boldsymbol{X}$ on to the general polyhedral cone $\mathbb{C}$ and tabulate the distribution of the test statistic under H_0 which is computationally infeasible. This makes the GLRT not appealing when the covariance is not the identity matrix. In such case alternative test or good approximations of the GLRT are tests of choice. Moreover, even in the case when the covariance is identity, the power of the GLRT may be extremely low when the alternative is very

sparse and the dimension is very high. For these reasons we will seek alternative common sense test for the constrained mean in a high dimensional setting.

3.3 Proposed test with positivity restriction

The simplest possible scenario for the testing problem (3.4) is when $\boldsymbol{X} \sim N(\boldsymbol{\mu}, \boldsymbol{I})$ and $\mathbb{C} = \mathcal{K}$, the positive orthant. Then the testing problem is reduced to

$$\mathcal{H}_0 : \boldsymbol{\mu} = \mathbf{0} \quad vs. \quad \mathcal{H}_1 : \boldsymbol{\mu} \geq \mathbf{0}. \tag{3.11}$$

where $\boldsymbol{\mu} \geq \mathbf{0}$ mean all components of $\boldsymbol{\mu}$ are non-negative with at least one positive component. As mentioned before, in this scenario the GLRT is available in closed form and its null distribution can be numerically simulated to obtain the rejection regions. Our goal will be to propose alternative tests which are also in closed form, whose null distribution can be evaluated with relative and the tests have two important properties:

- The power properties of the test are comparable to be that of GLRT in the case of positive orthant restriction with identity covariance matrix.
- The test has easily natural generalization to the case when the restriction cone is a general polyhedral cone, i.e. a case where the GLRT is not easy to compute.

In the supplement provided at the end of this chapter, we propose and compare several different omnibus test for the case where the covariance is the identity matrix. Based on our investigation of the battery of tests, we propose to use the following

test for the testing problem (3.2) in a high dimensional setting. The test is chosen because it has reasonable power properties and can be easily extended to the case when the covariance Σ is not a multiple of the identity matrix.

Let $\boldsymbol{X} = (X_1, \ldots, X_n)'$ be a single observation from the normal model and let $X_{1:n} < \cdots < X_{n:n}$ denote the ordered values of $X_1, \ldots, X_n$. Let k_n be a sequence of positive integers satisfying

Assumption A: For every n, $1 < k_n < n$ and $k_n \to \infty$ and there exists $0 < M < \infty$ such that $k_n\sqrt{\log n}/\sqrt{n} < M$ for all large values of n.

Define

$$M_n = (2k_n)^{-1/2} c(k_n/n)^{-1} \Big(\sum_{i=1}^{k_n} X_{n-i+1,n} - \mu_n(k_n) \Big) \tag{3.12}$$

where $c(s) = s^{-1} \int_{1-s}^{1} (1-u) d\Phi^{-1}(u)$ is slowly varying as $s \downarrow 0$ and $\mu_n(k_n) = n \int_{1-k_n/n}^{1} \Phi^{-1}(u) du$. Also define

$$S_n = n^{-1/2} \sum_{i=1}^{n-k_n} \Big(X_{i:n} - \mu_{i:n} \Big). \tag{3.13}$$

where $\mu_{i:n} = E(Z_{i:n})$, the expectation of the ith order statistics in a random sample of size n from $N(0,1)$. Then we propose using the test

$$T_n = \max\{S_n, M_n\} \tag{3.14}$$

for testing (3.2).

The following result provides conditions for universal consistency against all alternatives in $\mathcal{K}$.

Theorem 2. *Let $\boldsymbol{X} = (X_1, \ldots, X_n)'$ be a random vector defined on a rich enough probability space $(\Omega, \mathcal{B}, \mathcal{P})$, having the joint distribution $N_n(\boldsymbol{\mu}, \boldsymbol{I})$. Consider the*

testing problem (3.2). *Let* T_n *be the test statistic defined in* (3.14) *which rejects the null hypothesis if* $T_n > c_\alpha$ *where* $c_\alpha = z_{1-\sqrt{1-\alpha}}$ *and* z_α *is the upper* $\alpha-$*percentile of the standard normal distribution. Assume* k_n *satisfies Assumption A. As* $n \to \infty$, *let* $\{\boldsymbol{\mu}_n\}$ *be a sequence of alternatives such that* $\boldsymbol{\mu}_n \geq \mathbf{0}$ *for all* n *and* $\|\boldsymbol{\mu}_n\|_1 = r_n = O(n^{1/2}a(n))$ *for some positive sequence* $a(n)$ *with* $(a(n)\sqrt{n})/(k_n \log n) \to \infty$ *as* $n \to \infty$. *Then*

$$P_{\boldsymbol{\mu}=\mathbf{0}}(T_n > c_\alpha) \to \alpha \tag{3.15}$$

and

$$P_{\boldsymbol{\mu}_n}(T_n > c_\alpha) \to 1. \tag{3.16}$$

Proof. Since under the null $X_1, \ldots, X_n$ are iid $N(0,1)$ which belongs to the domain of attraction for the Gumbel law, by Lo (2016)

$$M_n \xrightarrow{d} N(0,1) \text{ as } n \to \infty.$$

Also, $S_n = n^{1/2}\bar{X} - \frac{2k_n^{1/2}c(k_n/n)}{n^{1/2}}M_n + o(1)$ where $\bar{X} = n^{-1}\sum_{i=1}^n X_i$. Thus, by Assumption A and the definition of $c(.)$, $S_n \xrightarrow{d} N(0,1)$ under the null hypothesis. Next we show asymptotic independence of S_n and M_n. Define

$$\sigma^2(s) = \int_{1-s}^1 \int_{1-s}^1 (min(u,v) - uv) d\Phi^{-1}(u) d\Phi^{-1}(v).$$

Following Csorgo and Mason (1985) and Lo (2016) there exist a sequence of Brownian bridges $B_1, B_2, \ldots$ such that

$$M_n = -(2k_n/n)^{-1/2}c(k_n/n)^{-1}\int_{1-k_n/n}^1 B_n(s)dQ(s) + o_P(1).$$

Rewriting $\sum_{i=1}^n X_{i:n}$ as the sum of extremes, M_n and sum of non-extremes, S_n,

$$S_n \;=\; n^{-1/2}\Big(\sum_{i=1}^n X_{i:n} - \mu_n(n)\Big) - (2k_n/n)^{1/2}c(k_n/n)M_n$$

The central statistics $n^{-1/2}\Big(\sum_{i=1}^{n} X_{i:n} - \mu_n(n)\Big)$ are independent of the extreme order statistics; see [53]. Thus, by the Brownian bridge representation of M_n, $\text{cov}(M_n, S_n) = (2k_n/n)^{1/2} c(k_n/n) \frac{\sigma^2(k_n/n)}{(2k_n/n)c(k_n/n)^2}$ where $\sigma^2(k_n/n)$ is the variance of $R_n = \Big(-\int_{1-k_n/n}^{1} B_n(s)dQ(s)\Big)$. Hence,

$$\begin{aligned} \text{cov}(M_n, S_n) &= (2k_n/n)^{1/2} c(k_n/n) \frac{\sigma^2(k_n/n)}{(2k_n/n)c(k_n/n)^2} \\ &= (2k_n/n)^{1/2} c(k_n/n) \to 0 \text{ as } n \to \infty \end{aligned}$$

which follows from the result $\sigma^2(s)/(2sc^2(s)) \to 1$ as $s \downarrow 0$. (Lemma 6; Lo, 2016). Thus, under null T_n is asymptotically distributed as the maximum of two independent standard normal variables. Hence, (3.15) follows. Now consider a fixed alternative sequence $\boldsymbol{\mu}_n^*$ such that $\|\boldsymbol{\mu}_n^*\|_1 = r_n$. Under the alternative, one can show that the distribution of T_n is stochastically bounded below by that of

$$\tilde{T}_n = \max\{\tilde{T}_{1n} + \Delta_{1n}, \tilde{T}_{2n} + \Delta_{2n}\}$$

where $\tilde{T}_{1n}$ and $\tilde{T}_{1n}$ are bounded in probability, and Δ_{1n} and Δ_{2n} are fixed sequences. Since $\max\{a+b, c+d\} \geq \min\{a, c\} + \max\{b, d\}$,

$$\begin{aligned} P_{\boldsymbol{\mu}_n}(T_n > c_\alpha) &\geq P_{\boldsymbol{\mu}_n}(\min\{\tilde{T}_{1n}, \tilde{T}_{2n}\} + \max\{\Delta_{1n}, \Delta_{2n}\} > c_\alpha) \\ &= P_{\boldsymbol{\mu}_n}(\min\{\tilde{T}_{1n}, \tilde{T}_{2n}\} > c_\alpha - \max\{\Delta_{1n}, \Delta_{2n}\}). \end{aligned}$$

Since $\min\{\tilde{T}_{1n}, \tilde{T}_{2n}\}$ is bounded in probability showing that $\max\{\Delta_{1n}, \Delta_{2n}\} \to \infty$ will complete the proof the proof. To derive the expressions for $\tilde{T}_{1n}$, $\tilde{T}_{2n}$ and Δ_{1n}, Δ_{2n} we first introduce a notation. Let $X_i = Z_i + \mu_i$ where Z_i will be referred to as the normal part and μ_i will be referred to as the mean part of X_i, respectively. Let

$\boldsymbol{\mu}_n^* = (\mu_1, \ldots, \mu_n)$ and let $\mu_{(i)} < \cdots < \mu_{(n)}$ be the order values of μ and similarly $Z_{1:n} < \cdots < Z_{n:n}$ be the ordered values of Z. Let $X_{i:n} = Z_{j_i:n} + \mu_{(l_i)}$, where $\mathbf{j}(\omega) := \mathbf{j} = (j_1, \ldots, j_n)$ and $\mathbf{l}(\omega) := \mathbf{l} = (l_1, \ldots, l_n)$ are two random permutations of $(1, \ldots, n)$ denoting the position of the associated normal part and the mean part of $X_{i:n}$ in their respective ordered sequences.

Consider

$$\begin{aligned} n^{1/2} S_n &= \sum_{i=1}^{n-k_n} (X_{i:n} - E(Z_{i:n})) \\ &= \sum_{i=1}^{n-k_n} (Z_{j_i:n} + \mu_{(l_i)} - E(Z_{i:n})) \\ &\geq \sum_{i=1}^{n-k_n} (Z_{i:n} + \mu_{(i)} - E(Z_{i:n})). \end{aligned}$$

Therefore $S_n \geq n^{-1/2} \sum_{i=1}^{n-k_n} (Z_{i:n} - E(Z_{i:n})) + n^{-1/2} \sum_{i=1}^{n-k_n} \mu_{(i)} = T_{1n} + \Delta_{1n}$. By assumption, $\sum_{i=1}^{n-k_n} \mu_{(i)}$ i.e. either $O(n^{1/2})$ or $O(n^{1/2} b(n))$ where $b(n) \to \infty$ as $\nu \to \infty$. Consider the case where $\sum_{i=1}^{n-k_n} \mu_{(i)} = O(n^{1/2} b(n))$. Then $\Delta_{1n} = n^{-1/2} \sum_{i=1}^{n-k_n} \mu_{(i)} = O(b(n)) \to \infty$ as $n \to \infty$. Now consider the case $\sum_{i=1}^{n-k_n} \mu_{(i)} = O(n^{1/2})$. In that case, by assumption, $\sum_{i=n-k_n+1}^{n} \mu_{(i)} = O(n^{1/2} a(n))$. Let $\delta > 0$ be given. Then for each n, define $\Omega_n \subseteq \Omega$ by

$$\Omega_n = \{\omega \in \Omega : \mu_{(l_i)} \geq \mu_{(i)} - 2\sqrt{-\log(\delta)},\ \forall i \in \{n - k_n + 1, \ldots, n\}\}.$$

Then by Lemma 1 we have

$$\begin{aligned} P(\Omega_n) &= 1 - P(\bigcup_{i=n-k_n+1}^{n} \{\mu_{(l_i)} + 2\sqrt{-\log(\delta)} < \mu_{(i)}\}) \\ &\geq 1 - \sum_{i=n-k_n+1}^{n} P(\{\mu_{(l_i)} + 2\sqrt{-\log(\delta)} < \mu_{(i)}\}) \\ &\geq 1 - k_n\delta. \end{aligned}$$

Denoting $(2k_n)^{1/2}c(k_n/n)$ by ζ_n, for $\omega \in \Omega_n$ we have

$$\begin{aligned} \zeta_n M_n(\omega) &= \sum_{i=n-k_n+1}^{n} (X_{i:n} - E(Z_{i:n})) \\ &= \sum_{i=n-k_n+1}^{n} (Z_{j_i:n} + \mu_{(l_i)} - E(Z_{i:n})) \\ &\geq \sum_{i=1}^{k_n} Z_{i:n} + \sum_{i=n-k_n+1}^{n} \mu_{(l_i)} - \sum_{i=n-k_n+1}^{n} E(Z_{i:n}) \\ &= \sum_{i=1}^{k_n} (Z_{i:n} - E(Z_{i:n})) + \sum_{i=n-k_n+1}^{n} \mu_{(l_i)} - \sum_{i=n-k_n+1}^{n} 2E(Z_{i:n}) \\ &\geq \sum_{i=1}^{k_n} (Z_{i:n} - E(Z_{i:n})) + \sum_{i=n-k_n+1}^{n} \mu_{(i)} \\ &\quad -2k_n\sqrt{-\log(\delta)} - \sum_{i=n-k_n+1}^{n} 2E(Z_{i:n}) \\ &\geq \sum_{i=1}^{k_n} (Z_{i:n} - E(Z_{i:n})) + \sum_{i=n-k_n+1}^{n} \mu_{(i)} \\ &\quad -2k_n\sqrt{-\log(\delta)} - 2k_n\sqrt{2\log n}. \end{aligned}$$

Choose $\delta := \delta_n$ such that $k_n\delta_n \to 0$. Then we have for $\omega \in \Omega_n$

$$\begin{aligned} M_n &\geq \tilde{T}_{2n} + \zeta_n^{-1} \sum_{i=n-k_n+1}^{n} \mu_{(i)} - 2\zeta_n^{-1}k_n\sqrt{-\log(\delta_n)} - 2\zeta_n^{-1}k_n\sqrt{\log n} \\ &= \tilde{T}_{2n} + \Delta_{2n}. \end{aligned}$$

Here $\tilde{T}_{2n} = \zeta_n^{-1}\sum_{i=1}^{k_n}(Z_{i:n} - E(Z_{i:n}))$ which due to symmetry of the standard normal

distribution converges to $N(0,1)$ [54]. Also, by Assumption A,

$$\begin{aligned}\Delta_{2n} &= \zeta_n^{-1} \sum_{i=n-k_n+1}^{n} \mu_{(i)} - 2\zeta_n^{-1} k_n \sqrt{-\log(\delta_n)} - 2\zeta_n^{-1} k_n \sqrt{\log n} \\ &= \eta_n + o(\eta_n)\end{aligned}$$

where $\eta_n = (n/k_n)^{1/2} a(n)/c(k_n/n)$. Since $c(\cdot)$ is a slowly varying function at zero, by assumption A we have $\eta_n \to \infty$ as $n \to \infty$.

Then letting $A_n = \{\tilde{T}_{1n} \wedge \tilde{T}_{2n} > c_\alpha - \Delta_{1n} \vee \Delta_{2n}\}$,

$$\begin{aligned}P_{\boldsymbol{\mu}_n}(T_n > c_\alpha) &\geq P_{\boldsymbol{\mu}_n}(A_n) \\ &= P(\Omega_n \bigcap A_n) + P(\Omega_n^c \bigcap A_n).\end{aligned}$$

Since $P(\Omega_n) \geq 1 - \delta_n \to 1$ and on Ω_n, $P(A_n) \to 1$, we have the result. □

Lemma 1. *If $X_1 \sim N(\mu_1, 1)$ and $X_2 \sim N(\mu_2, 1)$ are independent and $\mu_2 > \mu_1 + 2\sqrt{-\log(\delta)}$ for some $\delta > 0$ then $P(X_1 > X_2) \leq \delta$.*

Proof. Note that $P(X_1 > X_2) = \Phi(\Delta)$ where $\Delta = (\mu_2 - \mu_1)/\sqrt{2}$. The proof follows immediately noting that $\Phi(x) \leq \phi(x)/x$ for $x > 0$ where ϕ and Φ are the pdf and cdf of the standard normal. □

Corollary 1. *Consider the set up of Theorem 2. Suppse it is known that $\boldsymbol{\mu}$ is s_n-sparse, i.e. if $S \subset \{1, 2, \ldots, n\}$ is such that $\mu_i \neq 0$ iff $i \in S$, then the number of entries in S, $\#\{S\}$ is $O(s_n)$. Then if we choose $k_n = s_n$ and $\|\boldsymbol{\mu}\|_1 = a(n) s_n \log(n/s_n)$ for some increasing sequence $a(n)$ with $a(n) \to \infty$ when $n \to \infty$, we have*

$$P_{\boldsymbol{\mu}_n}(T_n > c_\alpha) \to 1.$$

The proof follows immediately from that of Theorem 2 and is omitted. Consistency against sparse alternatives is achieved if the norm grows faster than $s_n \log(n/s_n)$ which is the minimax risk rate for estimation of $\boldsymbol{\mu}$ [55, 56]. This intuitively makes sense since if the alternative is within the minimax radius, there is still positive probability (asymptotically) of not distinguishing it from the null value.

When $\Sigma = I$ one could also incorporate the LRT in the definition of the test statistics, e.g. $T_n = \max\{S_n, M_n, L_n\}$ where L_n is the LRT properly scaled and centered. This provides slight power enhancement in the regimes where $\boldsymbol{\mu}$ has moderate level of sparsity as neither S_n or M_n are particularly powerful in such a scenario.

3.4 General polyhedral cone

Suppose $\boldsymbol{X} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and let $\boldsymbol{\Sigma}^{-1} = \boldsymbol{\Omega}$ is fixed and known. In the previous section, we proposed a test statistic for $\boldsymbol{\mu} \in \mathcal{K}$ and $\boldsymbol{\Sigma} = I$. Under a linear transformation, $\boldsymbol{Z} = \boldsymbol{A}\boldsymbol{X} \sim N_d(\boldsymbol{A}\boldsymbol{\mu}, I)$ where A is such that $\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T = \boldsymbol{I}$ and $\boldsymbol{A}^T\boldsymbol{A} = \Omega$. Let us further denote $\boldsymbol{\theta} = \boldsymbol{A}\boldsymbol{\mu}$. Let us represent the transformed cone of parameter vectors by $\mathcal{K} = \{\boldsymbol{\theta} : \boldsymbol{A}^{-1}\boldsymbol{\theta} \geq 0\} = \{\boldsymbol{\theta} = \boldsymbol{A}\boldsymbol{\mu}, \boldsymbol{\mu} \geq 0\}$ which is not equal to $\mathcal{K}$ and hence the test statistic developed for non-negative orthant may no longer be powerful if directly applied on the transformed data $\boldsymbol{Z}$. Moreover, it could be that the linear inequality constraints specified on the mean vector implies that the mean lies in a general polyhedral cone. Suppose the model is $\boldsymbol{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\boldsymbol{R}\boldsymbol{\mu} \geq \boldsymbol{0}$ for some $d \times n$ matrix $\boldsymbol{R}$. We assume that rank of $\boldsymbol{R}$ is $d \leq n$. Then the problem is

equivalent to $\boldsymbol{Z} \sim N(\boldsymbol{\nu}, \boldsymbol{R\Sigma R}^T)$ where $\boldsymbol{Z} = \boldsymbol{RX}$ and the transformed mean $\boldsymbol{\nu} \in \mathcal{K}$. Thus, we consider the problem $\boldsymbol{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where $\boldsymbol{\mu} \in \mathcal{K}$.

For the non-negative orthant, under $\boldsymbol{\Sigma} = \boldsymbol{\Omega} = \boldsymbol{I}$, ordered values $X_{i:n}$ are equal to the ordered values of $\max_{1 \le i \le n} \langle \boldsymbol{X}, \tilde{\boldsymbol{e}}_i \rangle$ where $\tilde{\boldsymbol{e}}_i = \boldsymbol{e}_I / \|\boldsymbol{e}_i\| = \boldsymbol{e}_i$ and $(\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n)$ are the minimal set of extreme rays for $\mathcal{K}$. Thus $\sum_{n=k_n+1}^{n} X_{i:n}$ could be thought of as a statistic that captures *most alignment* with the the sparse vectors (the unit norm extreme rays) by finding the generators with k_n smallest angles with the data vector. Thus, $\sum_{n=k_n+1}^{n} X_{i:n}$ is expected to provide the required power when the true mean is near one of the sparse extreme rays. In the case of a general cone, the idea of sparsity is more nuanced. But as shown in [57], for the case when the number of minimal generators is equal to the dimension, sparsity can be thought of as the vectors lying on or near the extreme rays.

Now consider the transformed problem in the general case $\boldsymbol{AX} \sim N(\boldsymbol{\theta}, \boldsymbol{I})$ where $\boldsymbol{A}^T\boldsymbol{A} = \boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$ and the mean belongs to $\mathbb{C} = \{\boldsymbol{\theta} \in \mathbb{R}^n : \boldsymbol{A}^{-1}\boldsymbol{\theta} \ge \boldsymbol{0}\}$. Then the minimal generators of $\mathbb{C}$ are the columns of $\boldsymbol{A}$. Thus, the angles between the normalized extreme rays and the transformed data are

$$Y_i = \langle \|\boldsymbol{A}_{*i}\|^{-1} \boldsymbol{A}_{*i}, \boldsymbol{AX} \rangle,$$

the ith entry in the vector $\boldsymbol{Y} = \boldsymbol{\Delta}_{\boldsymbol{\omega}}^{-1} \boldsymbol{\Omega X}$. Here $\boldsymbol{\Delta}_{\boldsymbol{\omega}}$ is a diagonal matrix with ith diagonal equal to $\|\boldsymbol{A}_{*i}\| = \sqrt{\omega_{ii}}$ and ω_{ii} are the diagonals of $\boldsymbol{\Omega}$ and $\boldsymbol{A}_{*i}$ is the ith column of $\boldsymbol{A}$. Thus, the statistic

$$M_n(\boldsymbol{\Omega}) = (k_n)^{-1/2} c (k_n/n)^{-1} \sum_{n=k_n+1}^{n} (Y_{i:n} - \mu_{i:n}(\boldsymbol{\Omega})) \tag{3.17}$$

can be used in place of M_n in (3.12) where $Y_{i:n}$ are the order statistics for $\boldsymbol{Y}$ and

$\mu_{i:n}(\boldsymbol{\Omega})$ are the expected values of $Y_{i:n}$ when $\boldsymbol{\mu} = \mathbf{0}$. On a similar vein, the most dense vector for the non-negative orthant is obtained from the normalized sum of all normalized canonical vectors $(\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n)$. In the general case the interior direction against which the Neyman-Pearson test statistic for a dense mean vector is constructed is given by $\boldsymbol{A}\boldsymbol{\Delta}_{\boldsymbol{\omega}}^{-1}\mathbf{1}$. Thus, the main statistic will be of the form $\langle \boldsymbol{A}\boldsymbol{\Delta}_{\boldsymbol{\omega}}^{-1}\mathbf{1}, \boldsymbol{A}\boldsymbol{X}\rangle = \mathbf{1}^T\boldsymbol{Y}$. As before, to maintain asymptotic independence the statistic is constructed based on the bottom $(n - k_n)$ ordered values of the transformed data vector. hence, the test corresponding to S_n

$$S_n(\boldsymbol{\Omega}) = n^{-1/2} \sum_{i=1}^{n-k_n} (Y_{i:n} - E(Y_{i:n})) \tag{3.18}$$

For the general cone problem we recommend using the test statistics

$$T_n(\boldsymbol{\Omega}) = \max\{S_n(\boldsymbol{\Omega}), M_n(\boldsymbol{\Omega})\} \tag{3.19}$$

Note that the statistics in [58] is exactly equal to a scaled version of $Y_{n:n}^2$. In [58], the test was based on two sided hypothesis. For one sided hypothesis it seems reasonable to work with $Y_{n:n}$. Thus, our test can be thought of as a modification of the test proposed in [58] but adapted for the one-sided testing problem. Other example where the precision matrix transformation is used instead of a square root transformation can be found in [59].

When Σ is unknown, we have n observational vectors $\mathbf{X}_i \sim N(\boldsymbol{\mu}, \Sigma)$ for $1 \leq i \leq m$ where n is the dimensionality of vector and m is sample size. We consider an estimator of $\hat{\Sigma}^{-1}$ assuming possibly some sparse structure on precision matrix using a lasso penalty (L1) penalty [60–62]. Once a positive definite sparse covariance matrix estimate is available, we apply the test for the known $\boldsymbol{\Sigma}$ with $\boldsymbol{\Sigma} = \hat{\boldsymbol{\Sigma}}$.

3.5 Minimum distortion mapping

There are generally three approaches to deal with inadequacy of the GLRT due to intractability in the general cone constraint case and due to high dimension. One is to rely on tests other than GLRT in the case when $\Sigma = I$ that have reasonable performance in terms of minimal power over the entire first orthant and can be easily generalized and computed in the case when $\Sigma \neq I$. The generalization of the proposed test given in Section 3.3 falls under this category. Another approach is to find a transformation $\boldsymbol{AX}$ such that the cone $\mathbb{C}$ is not too different from the first orthant $\mathcal{K}$ and then apply the GLRT or other suitable tests developed for the orthant case. A third approach would be to approximate the GLRT under general conic constraint and apply the approximate test. We do not pursue this last option.

For a general Σ, [46] suggested a square-root transformation, $\boldsymbol{Z} = \boldsymbol{AX} \sim N(\boldsymbol{A\mu}, \boldsymbol{I})$ the center of $\mathcal{K} = \{\boldsymbol{\mu} | \mu_i \geq 0\}$ is equal to that of $\boldsymbol{A}\mathcal{K} = \{\boldsymbol{A\mu} | \boldsymbol{\mu} \in \mathcal{K}\}$. With this transformed $\boldsymbol{Z}$, [46] proposed an approximate GLRT which is $P(g(\boldsymbol{Z}) \geq c) = \sum_{i=0}^{n} \binom{n}{i} / 2^n P(\chi_i^2 \geq c)$ where $g(\boldsymbol{Z}) = \sum_{i=1}^{n} (\max(Z_i, 0))^2$. When $n \to \infty$, the computation of probability can become involved. The cone $\boldsymbol{A}\mathcal{K}$ is approximated by $\mathcal{K}$, but it is not clear how good this approximation is in high dimension.

A case when the methodology developed for the positive orthant can be directly applied to the general $\boldsymbol{\Sigma}$ problem is when the transformation $\boldsymbol{A}(\mathcal{K}) = \{\boldsymbol{A\mu} : \boldsymbol{\mu} \in \mathcal{K}\}$ leaves the positive orthant invariant. It is easy to see that the linear transformation will leave the cone unchanged iff the matrix square root $\boldsymbol{A}$ is non-negative entries, i.e., all entries of $\boldsymbol{A}$ are non-negative. Matrices that admit non-negative square

roots are called *completely positive matrices*. Thus, $\boldsymbol{\Omega}$ needs to be a completely positive matrix. However, completely positive matrices need not have square matrices as square roots. A matrix $\boldsymbol{R}$ is completely positive iff there exists a non-negative matrix $\boldsymbol{B}$ such that $\boldsymbol{R} = \boldsymbol{B}\boldsymbol{B}^T$. The number of least possible columns of $\boldsymbol{B}$ in the factorization is called the $cp - rank$ of $\boldsymbol{R}$. Thus, in the present problem, the invariance $\boldsymbol{A}(\mathcal{K}) = \mathcal{K}$ occurs iff $\boldsymbol{\Omega}$ is a completely positive matrix with $cp - rank$ equal to n. Determining whether a matrix is completely positive and then determining the $cp - rank$ is generally computationally intractable. Moreover, the class of $n \times n$ covariance matrices $\boldsymbol{\Sigma}$ such that the corresponding precision matrices $\boldsymbol{\Omega}$ belong to the class of rank n completely positive matrices is a rather restrictive class.

In general, there may not be a square root of $\boldsymbol{\Omega}$ that leaves the positive orthant unchanged and hence the proposed tests may not perform well for some alternatives. We propose using a square root of $\boldsymbol{\Omega}$ that induces *minimum distortion* to the positive orthant, thereby ensuring that the tests based on the positive orthant are still valid for most of the alternatives.

The first idea is to choose a transformation which will preserve some of the basic features of the testing problem that will still allow consistency of the tests. For example, if an alternative $\boldsymbol{\mu}$ is transformed to a new mean vector $\boldsymbol{\theta}$ which retains sufficiently many positive entries, the proposed test would still be consistent under the same framework as that of the orthant.

Let the set of square roots of $\boldsymbol{\Omega}$ be given by

$$\mathcal{S}(\boldsymbol{\Omega}) = \{\boldsymbol{Q}\boldsymbol{A}^+ : \boldsymbol{Q} \in \mathcal{O}(n)\}$$

where $\boldsymbol{A}^+$ is the unique symmetric square root of $\boldsymbol{\Omega}$ and $\mathcal{O}(n)$ is the group of $n \times n$ orthogonal matrices. Suppose $\mathcal{K}_{\boldsymbol{\Omega}} = \{\boldsymbol{A}^+\boldsymbol{\mu} : \boldsymbol{\mu} \geq 0, \boldsymbol{A}^T\boldsymbol{A} = \boldsymbol{\Omega}\}$ be the transformed cone of mean vectors under the symmetric square root transformation. Then the possible transformed cones are

$$\mathcal{K}(\boldsymbol{Q}) = \boldsymbol{Q}\mathcal{K}(\boldsymbol{\Omega}) = \{\boldsymbol{Q}\boldsymbol{A}^+\boldsymbol{\mu} : \boldsymbol{\mu} \in \mathcal{K}\}.$$

One idea related to leave the orthant least changed under the square root transformation would to look at the intrinsic volumes of the intersection $\mathcal{K} \bigcap \boldsymbol{Q}\mathcal{K}(\boldsymbol{\Omega})$ and maximize it over the rotation group $\mathcal{O}(n)$ [63]. The cone $\boldsymbol{Q}\mathcal{K}(\boldsymbol{\Omega})$ obtained under the linear transformation is also closed convex polyhedral cone and so is the intersection and hence intrinsic volume formula apply for the intersection as well. Computation of intrinsic volume is non-trivial and often approximated using probabilistic arguments. One possibility is to maximize a summary measure like the *statistical dimension* [64] of the intersection

$$\delta_{\boldsymbol{Q}} := \delta(\mathcal{K} \bigcap \boldsymbol{Q}\mathcal{K}(\boldsymbol{\Omega})) = \mathrm{E}(\|\Pi_{\mathcal{K} \bigcap \boldsymbol{Q}\mathcal{K}(\boldsymbol{\Omega})}(\boldsymbol{Z})\|^2)$$

where E denotes expectation, $\boldsymbol{Z}$ is an n dimensional normal random variables with mean $\boldsymbol{0}$ and variance $\boldsymbol{I}$ and $\Pi_{\mathbb{C}}$ is the projector of Euclidean vectors to the cone $\mathbb{C}$. Then an optimum $\boldsymbol{Q}$ would be one that maximizes $\delta_{\boldsymbol{Q}}$. However, this seems computationally challenging and we do not pursue it here.

Another idea would be to find a square root of $\boldsymbol{\Omega}$ that produces least distortion to the positive orthant. Thus if $\mathcal{D}(\mathbb{C}_1, \mathbb{C}_2)$ denote the distortion measure between two pointed finitely generated convex polyhedral cones in $\mathbb{R}^n$, then we want to

minimize $\mathcal{D}(\mathcal{K}, \mathcal{K}_{\boldsymbol{\Omega},\boldsymbol{Q}})$ over choices of $\boldsymbol{Q}$. Formally given such a distortion measure we want to find the square root $\boldsymbol{Q}^*\boldsymbol{A}^+$ where

$$\boldsymbol{Q}^* = \underset{\boldsymbol{Q}\in\mathcal{O}(n)}{\operatorname{argmin}}\, \mathcal{D}(\mathcal{K}, \mathcal{K}_{\boldsymbol{\Omega},\boldsymbol{Q}}).$$

While the test we are proposing is an omnibus test, it maybe in a given problem certain alternatives have to emphasized more. By the construction of the test, we are already emphasizing the sparse region, and hence in the case of a general covariance matrix, we would like to leave some of the sparse vectors relatively unchanged under the square root transformation. To this end, we define a distortion measure but we define it for a general finitely generated closed convex polyhedral cone $\mathbb{C}$ that is pointed. Let

$$\mathcal{D}(\mathbb{C}, \mathbb{C}_{\boldsymbol{\Omega},\boldsymbol{Q}}) := \mathcal{D}(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_n; \boldsymbol{Q}) = \sum_{i=1}^{n} \| \frac{\boldsymbol{Q}\boldsymbol{A}^+\boldsymbol{\mu}_i}{\sqrt{\boldsymbol{\mu}_i^T\boldsymbol{\Omega}\boldsymbol{\mu}_i}} - \boldsymbol{\mu}_i\|_2^2 \tag{3.20}$$

where $\mathbb{C}_{\boldsymbol{\Omega},\boldsymbol{Q}} = \{\boldsymbol{Q}\boldsymbol{A}^+\boldsymbol{\mu} : \boldsymbol{\mu} \in \mathbb{C}\}$ and the set of vectors to be left relatively unchanged is a specified set of unit norm vectors in $\mathbb{C}$. The next result shows that set of $\boldsymbol{\Omega}$ that allows $\mathcal{D}$ to be zero are limited.

Proposition 1. *Let the set of vectors $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_n$, with $\|\boldsymbol{\mu}_i\| = 1$ for each i, be a conically independent set of generators for $\mathbb{C}$. If there exist $\boldsymbol{Q}$ such that $\mathcal{D}(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_n; \boldsymbol{Q}) = 0$ then $\mathbb{C}(\boldsymbol{Q}) = \mathbb{C}$. In the orthant case with $\mathbb{C} = \mathcal{K}$, this would imply that $\boldsymbol{\Omega}$ is a completely positive matrix.*

Proof. Suppose $\mathcal{D}(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_n) = 0$. Then $\boldsymbol{Q}\boldsymbol{A}^+\boldsymbol{\mu}_i = \lambda_i\boldsymbol{\mu}_i$ for each i where $\lambda_i = \sqrt{\boldsymbol{\mu}_i^T\boldsymbol{\Omega}\boldsymbol{\mu}_i} > 0$. If $\boldsymbol{x} \in \mathbb{C}$, then by assumption $\boldsymbol{x} = \sum_{j=1}^n \beta_j\boldsymbol{\mu}_j$ where $\beta_j \geq 0$. Then $\boldsymbol{Q}\boldsymbol{A}^+\boldsymbol{x} = \sum_{j=1}^n \beta_j\boldsymbol{Q}\boldsymbol{A}^+\boldsymbol{\mu}_j = \sum_{j=1}^n \beta_j\lambda_j\boldsymbol{\mu}_j \in \mathbb{C}$. Conversely, if $\boldsymbol{x} = \sum_{j=1}^n \beta_j\boldsymbol{\mu}_j$ is a

vector in $\mathbb{C}$, then $\boldsymbol{x} = \boldsymbol{Q}\boldsymbol{A}^{+}\boldsymbol{\theta}$ where $\boldsymbol{\theta} = \sum_{j=1}^{n} \beta_j \lambda_j^{-1} \boldsymbol{\mu}_j \in \mathbb{C}$. Clearly, in this case $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_n$ are the eigenvectors of $\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{A}^{+}$ with positive eigenvalues. Hence, in the positive orthant case that would make $\boldsymbol{\Omega}$ has to be completely positive. □

The requirement of completely positive is too stringent. but one could still use the measure $\mathcal{D}$ for general $\boldsymbol{\Sigma}$ and obtain orthogonal matrices $\boldsymbol{Q}$ such that the transformed cone is close to $\mathbb{C}$.

We use a fixed set of linearly independent vectors of varying sparsity and find $\boldsymbol{Q}$ that optimizes the distortion measure with respect that set. Such a $\boldsymbol{Q}$ is unique and can be obtained in closed form. We will work a slightly different setting where the diagonals of the precision matrix $\boldsymbol{\Omega}$ are one, or in other words $\boldsymbol{\Omega}$ is a correlation matrix. This would make the two problems same in the precision scale. Note that, if $\boldsymbol{X} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\boldsymbol{\Delta}_\omega$ is a diagonal matrix with entries equal to $\sqrt{\omega_{ii}}$ then the testing problem in the re-scaled model $\boldsymbol{Y} = \boldsymbol{\Delta}_\omega \boldsymbol{X} \sim N_n(\boldsymbol{\Delta}_\omega \boldsymbol{\mu}, \boldsymbol{\Delta}_\omega \boldsymbol{\Sigma} \boldsymbol{\Delta}_\omega)$ remains invariant since the new mean vector $\boldsymbol{\nu} = \boldsymbol{\Delta}_\omega \boldsymbol{\mu}$ is non-negative (zero) iff $\boldsymbol{\mu}$ is non-negative (zero). Here the diagonals of the precision matrix $\boldsymbol{\Delta}_\omega^{-1} \boldsymbol{\Sigma}^{-1} \boldsymbol{\Delta}_\omega^{-}1$ are unity. The following proposition gives a closed form solution for the optimal square root.

Proposition 2. *Let $\boldsymbol{\Omega}$ be the positive definite precision matrix with unit diagonals. Let $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_n$ be a set of linearly independent unit norm vectors in $\mathcal{K}$. Define*

$$A_* = \underset{\Omega = A'A}{\arg\min} \sum_{i=1}^{n} || \frac{\boldsymbol{A}\boldsymbol{\mu}_i}{\sqrt{\boldsymbol{\mu}_i^T \boldsymbol{\Omega} \boldsymbol{\mu}_i}} - \boldsymbol{\mu}_i ||^2.$$

Then $\boldsymbol{A}^* = \boldsymbol{Q}^* \boldsymbol{A}^{+}$ where $\boldsymbol{A}^{+}$ is the symmetric square root of $\boldsymbol{\Omega}$, $\boldsymbol{Q}^* = \boldsymbol{V}\boldsymbol{U}^T$ and $\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{V}^T$ is the singular value decomposition of $\boldsymbol{A}^{+}\big(\sum_{i=1}^{n} (\boldsymbol{\mu}_i^T \boldsymbol{\Omega} \boldsymbol{\mu}_i)^{-1/2} \boldsymbol{\mu}_i \boldsymbol{\mu}_i^T\big)$.

Proof.

$$\begin{aligned}
\boldsymbol{A}^* &= \underset{\boldsymbol{\Omega}=\boldsymbol{A}'\boldsymbol{A}}{\arg\min} \sum_{i=1}^{m} || \frac{\boldsymbol{A}\boldsymbol{\mu}_i}{\sqrt{\boldsymbol{\mu}_i^T \boldsymbol{\Omega} \boldsymbol{\mu}_i}} - \mu_i ||^2 \\
&= \underset{\boldsymbol{Q}\in O(n)}{\arg\min} \sum_{i=1}^{m} || \frac{\boldsymbol{Q}\boldsymbol{A}^+\boldsymbol{\mu}_i}{\sqrt{\boldsymbol{\mu}_i^T \boldsymbol{\Omega} \boldsymbol{\mu}_i}} - \mu_i ||^2 \\
&= \underset{\boldsymbol{Q}\in O(n)}{\arg\min} \left\{ \mathrm{tr}(1 + \mathrm{tr}(\sum_{i=1}^{n} \boldsymbol{\mu}_i \boldsymbol{\mu}_i^T) - 2\mathrm{tr}(\boldsymbol{Q}\boldsymbol{A}^+\boldsymbol{M}) \right\}
\end{aligned}$$

where $M = \sum_{i=1}^{n} \frac{\boldsymbol{\mu}_i \boldsymbol{\mu}_i^T}{\sqrt{\boldsymbol{\mu}_i^T \boldsymbol{\Omega} \boldsymbol{\mu}_i}}$. Let $W = U\Lambda V^T$ be the singular value decomposition of $\boldsymbol{A}^+\boldsymbol{M}$. Then

$$\begin{aligned}
\boldsymbol{A}^* &= \underset{\boldsymbol{Q}\in O(n)}{\mathrm{argmax}}\ \mathrm{tr}(\boldsymbol{Q}\boldsymbol{A}^+\boldsymbol{M}) \\
&= \underset{\boldsymbol{Q}\in O(n)}{\mathrm{argmax}}\ \mathrm{tr}(\boldsymbol{Q}\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{V}^T) \\
&= \underset{\boldsymbol{Q}\in O(n)}{\mathrm{argmax}}\ \mathrm{tr}(\boldsymbol{V}^T\boldsymbol{Q}\boldsymbol{U}\boldsymbol{\Lambda}).
\end{aligned}$$

Since $\boldsymbol{U}^T\boldsymbol{Q}\boldsymbol{V}$ is an orthogonal matrix with singular values equal to one, by the von Neumann trace inequality the maximum is attained when $\boldsymbol{U}^T\boldsymbol{Q}\boldsymbol{V} = \boldsymbol{I}$ or $\boldsymbol{Q} = \boldsymbol{U}\boldsymbol{V}^T$.), Thus, $\boldsymbol{A}^* = \boldsymbol{Q}^*\boldsymbol{A}^+ = \boldsymbol{V}\boldsymbol{U}^T\boldsymbol{A}^+$. □

One choice for the basis vectors could be the set of generators of $\mathcal{K}$, i.e. the canonical vectors. The optimum $\boldsymbol{A}^*$ in that case turns out to be the symmetric square root.

Proposition 3. *If $\boldsymbol{\mu}_i = \boldsymbol{e}_i$, the ith canonical vector in (2), then the optimum solution is $\boldsymbol{A}^* = \boldsymbol{A}^+$.*

The proof follows easily noting that in this case $\boldsymbol{M} = \boldsymbol{I}$ as the diagonals of $\boldsymbol{\Omega}$ are all one.

Although the transformation based on $\boldsymbol{A}$ is computationally efficient, depending on $\boldsymbol{\Omega}$, there may be still significant disagreement between $\mathcal{K}(\boldsymbol{Q}^*)$ and $\mathcal{K}$ in high

dimension. While the boundaries are preserved under the linear map, the sparse canonical vectors $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n$ are no longer the extreme rays but rather gets mapped to those of $\mathcal{K}(\boldsymbol{Q}^*)$. Hence another approach will be to modify the test using the extreme rays of the transformed cone.

3.6 Simulation Studies

In this section, we compare the performance of our test statistics with existing test statistics like GLRT and Neyman-Pearson (NP). The NP test statistic for any fixed $\boldsymbol{\mu_1^T} \in \mathcal{H}_1$ is $\frac{\mathbf{X}^\top \Sigma^{-1} \boldsymbol{\mu_1}}{\sqrt{\boldsymbol{\mu_1}^\top \Sigma^{-1} \boldsymbol{\mu_1}}}$ and it is most powerful for testing $\mathcal{H}_0 : \boldsymbol{\mu} = \mathbf{0}$ $vs.$ $\mathcal{H}_1 : \boldsymbol{\mu} = \boldsymbol{\mu_1}$ where $\boldsymbol{\mu_1} \in \mathcal{K}$. We compute the NP test statistic at the center of the non-negative orthant which is the most dense alternative with all μ_i equal. The NP at $\boldsymbol{\mu_1} = \mu \mathbf{1}$ is simply $\frac{\mathbf{X}^\top \Sigma^{-1} \mathbf{1}}{\sqrt{\mathbf{1}^\top \Sigma^{-1} \mathbf{1}}}$ and the corresponding power of a size-α test is $1 - \Phi\left(z_\alpha - \frac{\mathbf{X}^\top \Sigma^{-1} \mathbf{1}}{\sqrt{\mathbf{1}^\top \Sigma^{-1} \mathbf{1}}} \right)$ where z_α is the upper-α point of a standard normal distribution [65]. The GLRT has been discussed in details in section 3.2.

For our simulation purposes we consider dimension $n = 500$ to investigate the power of the test statistics at 5% level of significance as we change the sparsity level keeping the $||\boldsymbol{\mu}||_2$ norm fixed. We generate 500 such $\boldsymbol{\mu}$ under alternative where $\boldsymbol{\mu}_k$ has k non-zero components i.e. $\boldsymbol{\mu}_k = (\underbrace{\tau, \ldots, \tau}_{\text{k times}}, 0, \ldots, 0)^T$, $\tau > 0$ such that $||\boldsymbol{\mu}_k||_2 = \sqrt{n}/4$. The most sparse case would be $\boldsymbol{\mu} = (\sqrt{n}/4, 0, ..., 0)^T$ and the most dense case would be $\boldsymbol{\mu} = \left(\frac{1}{4}, \frac{1}{4}, ..., \frac{1}{4}\right)^T$. We use HD to refer to the test statistic $\max\{S_n, M_n, L_n\}$, MnSn to refer to $\max M_n, S_n$ and FM to refer to the Follmann test statistic.

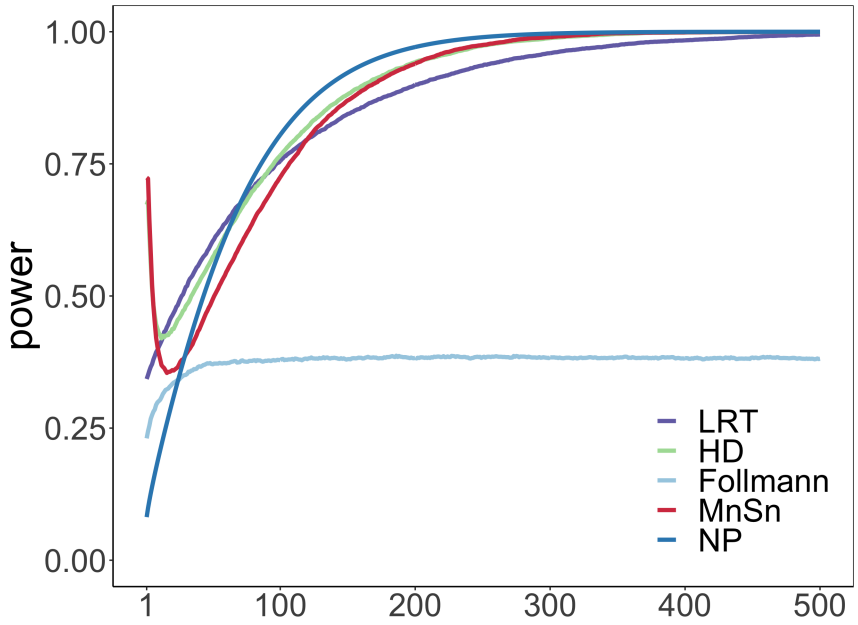

Figure 3.2: Power curve for our proposed test statistic (HD), LRT and NP for decreasing sparsity level when $\Sigma = I$.

When $\Sigma = I$, the power curve for HD, MnSn, LRT, NP and the Follmann test statistic are shown in Figure 3.2. We see that the NP performs well for the dense configuration of $\boldsymbol{\mu}$ but the power drops drastically for sparse alternatives. The LRT performs similar to the NP at the dense case, it performs better than the NP for the sparse signals. While HD performs close the NP for dense configuration and close to the LRT for sparse configuration, the power is substantially higher for ultra sparse signals. While HD performs better than $\max M_n, S_n$ in this case, for general Σ it performs only marginally better. Table 3.1 reports the minimum and average power along with the false positive rate (FPR) across the 500 alternatives. HD performs better than all the three test statistics in terms of the minimum power and average power. The corresponding local power is presented in Figure 3.3. The HD and

MnSn has better local power characteristics than others in the sparse regime but exhibit similar local power behaviour under dense regime.

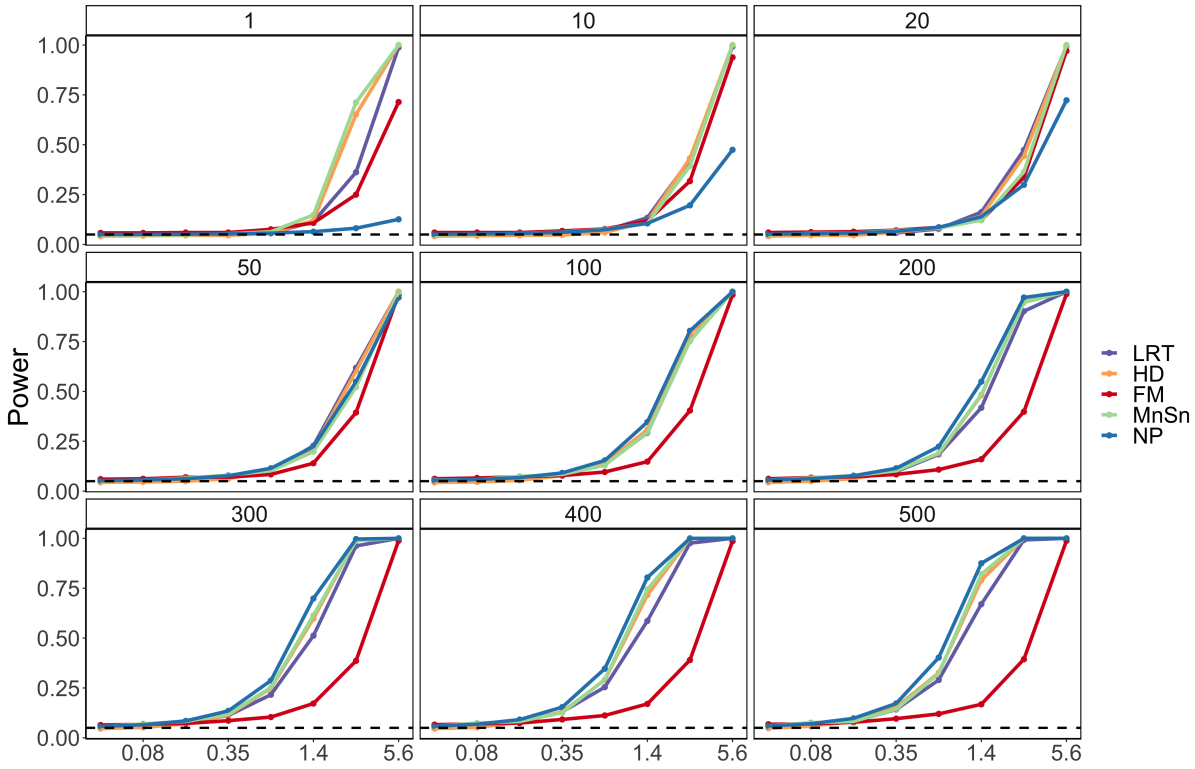

Figure 3.3: Local power for $\boldsymbol{\mu}$ under different sparsity levels represented in 9 panels when $\Sigma = I$. The numbers on top of each panel represent $\mathcal{L}_0$ norm. For instance, panel 1 has only one non-zero component, panel 2 has 10 non-zero components and so on. The x-axes has increasing $||\boldsymbol{\mu}|| = \frac{\sqrt{n}}{2^k}$ for $k = 1, \ldots, 8$.

For $\Sigma \neq I$, we fix the generalized norm $\boldsymbol{\mu}^T \Sigma^{-1} \boldsymbol{\mu}$ and three different Σ specifications are considered: (a) An AR(1) process with $\rho = 0.7$ (b) A tridiagonal banded matrix with bandwidth 100 (c) A random matrix with 90% randomly selected off-

Power	NP	LRT	HD	MnSn	Follmann
Minimum	0.082	0.342	0.42	0.35	0.231
Average	0.883	0.866	0.890	0.87	0.377
FPR	0.05	0.051	0.05	0.05	0.05

Table 3.1: Minimum power and average power when $\Sigma = I$

diagonal entries set to be non-zero. Figures 3.4, 3.5 and 3.6 display the power curve for different test statistics under different transformations. The left panel corresponds to the Ω transformation applied to the data and the statistic $\max(M_n, S_n)$ computed based on the transformed data. However, the actual LRT (left panel) is computationally intensive and hence we replace it with some sort of pretend LRT which is projection on $\mathcal{K}$ after applying suitable transformations such as A_0 and A_* (middle and right panel). Similarly, $\max(M_n, S_n)$ (middle and right panel) is obtained from A_0 and A_* transformation on the data and the corresponding HD is computed based on $\max(M_n, S_n)$ and projection on $\mathcal{K}$.

Table 3.2 reports the minimum and average power along with the false positive rate (FPR) across the 500 alternatives. For AR(1), $\max(M_n, S_n)$ performs reasonably well for dense and extreme sparse alternatives but the power dips substantially for sparse configuration under Ω transformation. Hence we bring in HD test statistic, which performs better in terms of the minimum power under both A_* and A_0 transformations. However, when Σ is banded, combining $\max(M_n, S_n)$ with projection on $\mathcal{K}$ raises the minimum power marginally. Finally, we present the worst

case scenario performance through a randomly generated Σ. While all test statistics perform equally inadequately in terms of the minimum power barring the ones under A_0 transformations, the opposite is true when average power is considered.

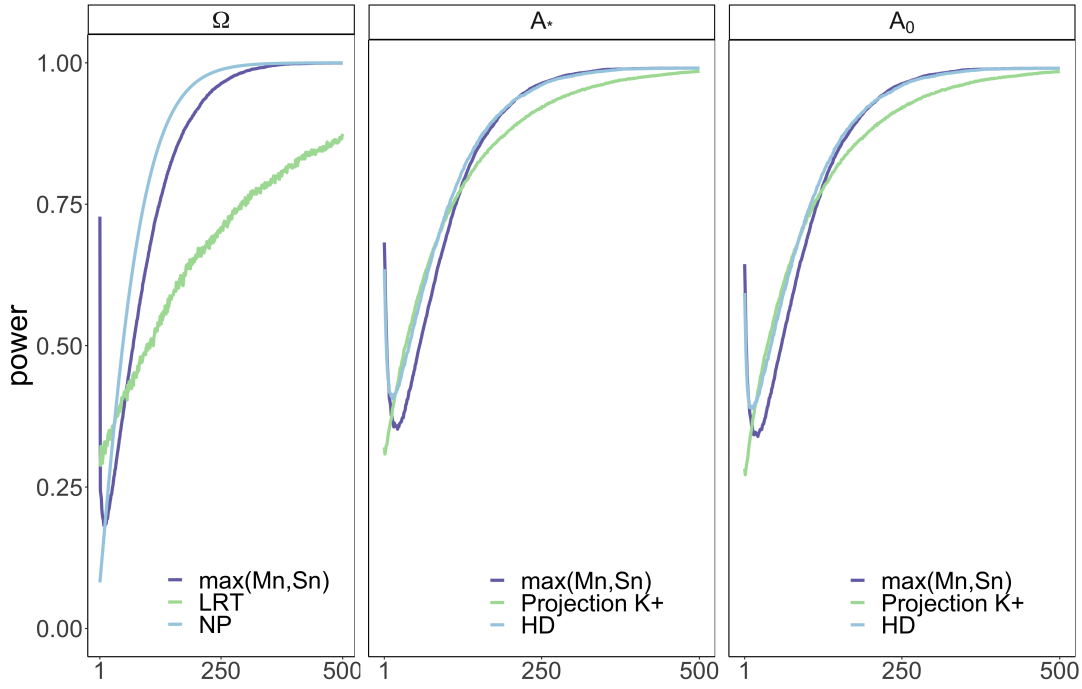

Figure 3.4: Power curve under different transformations for decreasing sparsity level when $\Sigma = AR(1)$ with $\rho = 0.7$.

3.7 Discussion

We have outlined an approach for testing whether the mean of a normal vector is zero versus alternatives where the mean is restricted to non-negative but not equal to zero. The methodology proposes an omnibus test which has reasonable power properties over the entire set of alternatives under a high-dimensional setting where the sample size is growing faster than the dimension. For specific sequence of sparse

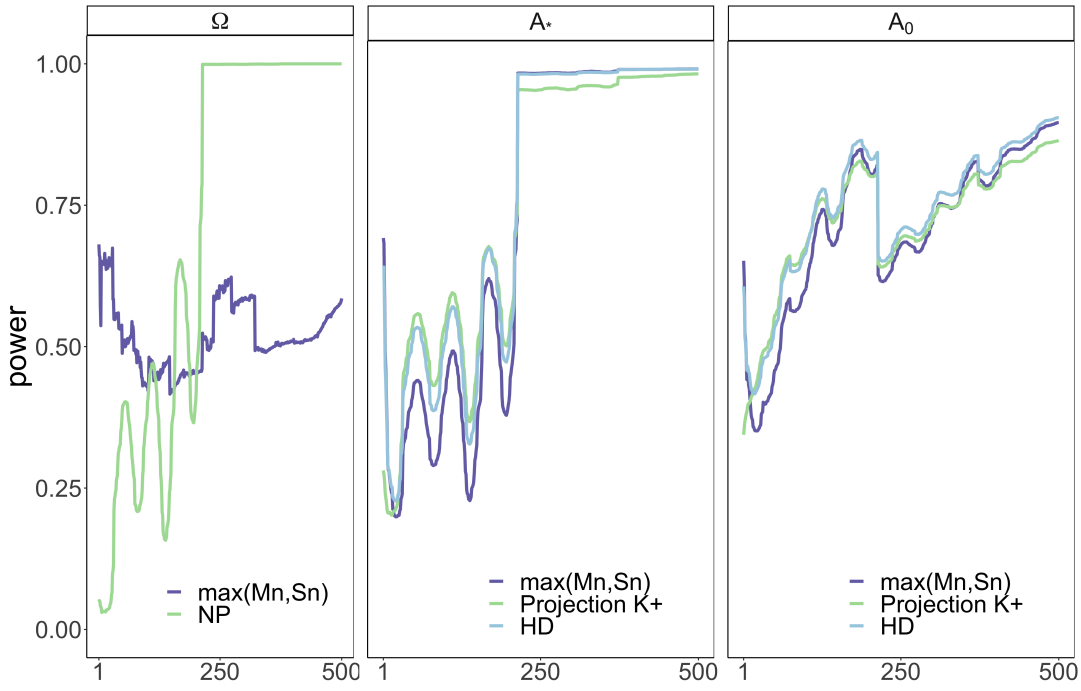

Figure 3.5: Power curve under different transformations for decreasing sparsity level when Σ is banded.

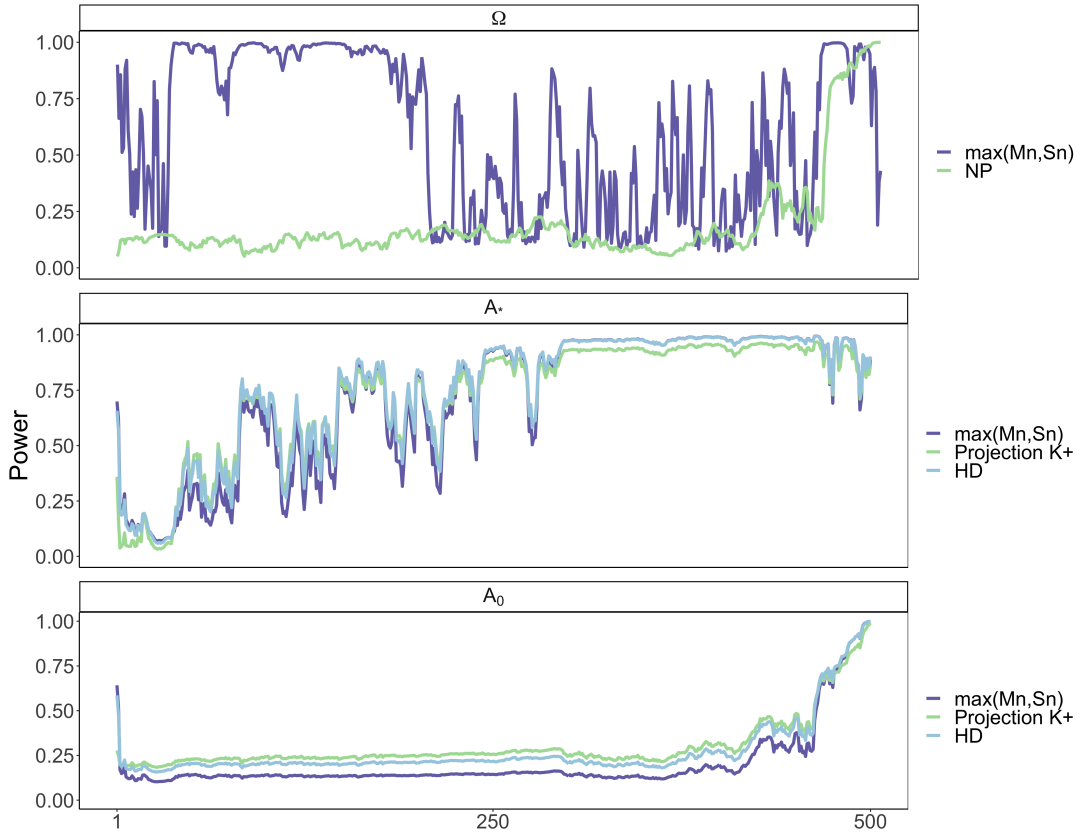

Figure 3.6: Power curve under different transformations for decreasing sparsity level Σ is random with 90% non-zero entries.

Σ	Power	Ω			A_*			A_0		
		NP	LRT	$\max(M_n, S_n)$	HD	Projection on $\mathcal{K}$	$\max(M_n, S_n)$	HD	Projection on $\mathcal{K}$	$\max(M_n, S_n)$
AR(1)	Minimum	0.081	0.286	0.182	0.410	0.312	0.356	0.393	0.275	0.343
	Average	0.875	0.666	0.834	0.877	0.854	0.863	0.874	0.851	0.860
	FPR	0.05	0.046	0.048	0.05	0.046	0.051	0.049	0.047	0.051
Banded	Minimum	0.030	-	0.416	0.230	0.204	0.201	0.421	0.348	0.355
	Average	0.718	-	0.520	0.777	0.771	0.745	0.749	0.730	0.717
	FPR	0.05	-	0.051	0.049	0.050	0.049	0.048	0.048	0.046
Random	Minimum	0.050	-	0.074	0.059	0.032	0.069	0.157	0.184	0.103
	Average	0.195	-	0.685	0.742	0.718	0.713	0.273	0.302	0.206
	FPR	0.05	-	0.053	0.051	0.048	0.05	0.049	0.048	0.05

Table 3.2: Minimum and average power for general Σ

Notes. FPR denote the false positive rate or probability of Type−I error.

alternatives, where the individual coordinates are tending to zero but the norm is growing slower than the dimension, the restriction on the sample size with respect to the dimension can be relaxed and consistency of the test can be shown under situations where sample size is much smaller than the dimension. The approach is novel in the sense it provide a viable approach for testing under arbitrary covariance structure that is computationally feasible and statistically efficient.

Supplementary result: Other omnibus tests for positive orthant

In this supplement we present several different omnibus tests considered for the positive orthant case with covariance matrix equal to the identity matrix.

In high dimensional testing problem with first orthant alternative, it is unlikely that a single statistic performs equally well for the dense and sparse parameter configuration. For vectors in the first orthant we define sparsity in terms of how many entries of the vector are 'close' to zero. Of course, this notion of weak sparsity as oppose to strong sparsity which is defined in terms of number of zero entries in the vector, depends on how one defines closeness to zero. Qualitatively speaking, a sparse vector would be where only a few entries are large in magnitude and the rest are small. As a prelimiary investigation, investigate a slew of tests for their performance in the one sided testing problem and then select and extend the tests with reasonable power against all alternatives and can be generalized to the case of general cones. In addition to the GLRT we consider the following tests.

3.7.1 Test Statistic 1

The guiding principles for proposing a candidate test is combining tests that perform best in specific alternative regimes (e.g. for dense alternatives versus sparse or ultra sparse alternatives) The first set of tests in subsection 3.7.1 are all hybrid tests combining linear tests with tests consisting of top order statistics. Note that

under the null model the data vector is simply a random sample from the standard normal distribution and hence finding null distribution of at least the critical points for any test that we define is not hard. The quantiles of the null distribution can be empirically computed in real time provided the test is not hard to evaluate.

Consider testing the null hypothesis $\boldsymbol{\mu} = \mathbf{0}$ against a dense alternative and a very sparse alternative. The Neyman-Pearson (NP) test against a unit norm dense vector $\boldsymbol{\mu} = n^{-1/2}\mathbf{1}$ where $\mathbf{1}$ is the n dimensional vectors of ones, is $S_n = \langle \boldsymbol{X}, \boldsymbol{\mu} \rangle = \frac{\sum_{i=1}^n X_i}{\sqrt{n}}$. Similarly the NP test against the ith extreme ray of the cone $\mathcal{K}$ is $\langle \boldsymbol{X}, \boldsymbol{e}_i \rangle = X_i$. Thus, the $X_{(n)} = \max_{1 \le i \le n} X_i$ would give the most detection power against the sparsest of alternatives while $\sqrt{n}\bar{\boldsymbol{X}}$ provides power against dense alternatives. We look at combining S_n and $X_{(n)}$ but since the distribution under the null have different scales for the two statistics, we combine the standardized version of the two. Under the null S_n has a standard normal distribution while $X_{(n)}$ normalized by $\sqrt{2 \log n}$ is asymptotically distributed as a Gumbel random variable. We consider the two statistics

$$S_n = \frac{\sum_{i=1}^n x_i}{\sqrt{n}}$$

$$M_n = \sqrt{2 \log n} X_{(n)} - 2 \log n + 1/2 \log \log n + 1/2 \log(4\pi).$$

where the centering and scaling of $X_{(n)}$ convergence to the Gumbel distribution. Moreover, M_n and S_n are asymptotically independent. Therefore for $\alpha \in [0, 1]$, we consider the test statistic

$$T(\alpha) = \alpha \frac{M_n - \beta}{v} + (1 - \alpha) S_n = \alpha G_n + (1 - \alpha) S_n. \tag{3.21}$$

where $G_n = \frac{M_n - \beta}{v}$. It is known that M_n is asymptotically distributed as a Gumbel

random variable with mean, β equal to Euler's constant and the variance, v equal to $\frac{\pi^2}{6}$ [66] [67]. Since α is unknown, we define a test statistic

$$\mathcal{P} = \min_{0 \leq \alpha \leq 1} p_\alpha.$$

where p_α is the p-value corresponding to $T(\alpha)$. We propose the following two ways to select α.

Selection of α through grid points

Take a set of grid points for α such that $0 < \alpha_0 < \alpha_1 < \cdots < \alpha_m = 1$ and compute use the test statistic $\mathcal{P} = \min_{1 \leq k \leq m} p_{\alpha_k}$. Under H_0, $P\big[M_n \leq z\big] \to G(z) = e^{-e^{-z}}$, the standard Gumbel Distribution.

Let $q_k(t)$ be the tth quantile of T_{α_k} under H_0. For a level-α test, we want to

find p such that $P_{H_o}(\mathcal{P} < p) \leq 0.05$. Therefore,

$$
\begin{aligned}
0.05 &= P_{H_o}(\mathcal{P} < p) \\
&= P_{H_o}\Big(\min_{1\leq k\leq m} P_{\alpha_k} < p \Big) \\
&= 1 - P_{H_o}\Big(p_{\alpha_k} \geq p \quad \forall\ 1 \leq k \leq m \Big) \\
&= 1 - P_{H_o}\Big(T_{\alpha_k} \leq q_k(1-p) \quad \forall\ 1 \leq k \leq m \Big) \\
&= 1 - P_{H_o}\left(S_n \leq q_1(1-p), G_n \leq \frac{q_k(1-p) - (1-\alpha_k)S_n}{\alpha_k} \ \forall\ 2 \leq k \leq m \right) \\
&= 1 - P_{H_o}\left(S_n \leq q_1(1-p), M_n \leq \min_{2\leq k\leq m} \left\{ \beta + \nu \frac{q_k(1-p) - (1-\alpha_k)S_n}{\alpha_k} \right\} \right) \\
&= 1 - \int_{-\infty}^{q_1(1-p)} \int_{-\infty}^{\min_{2\leq k\leq K}(a_k - b_k y)} f_{S_n, M_n}(y, z)\ dz\ dy \\
&\approx 1 - \int_{-\infty}^{q_1(1-p)} \int_{-\infty}^{min(a_k - b_k y)} f_{S_n}(y) f_{M_n}(z)\ dz\ dy \\
&= 1 - \int_{-\infty}^{q_1(1-p)} f_{S_n}(y) \int_{-\infty}^{min(a_k - b_k y)} f_{M_n}(z)\ dz\ dy \\
&= 1 - \int_{-\infty}^{q_1(1-p)} f_{S_n}(y)\ F_{M_n}(min(a_k - b_k y))\ dy \\
&= 1 - \int_{-\infty}^{q_1(1-p)} \phi(y)\ e^{-e^{-min(a_k - b_k y)}}\ dy \\
&= 1 - \sum_{k=2}^{m} \int_{A_k}^{B_k} \phi(y)\ e^{-e^{-(a_k - b_k y)}}\ dy.
\end{aligned}
$$

where a_k and b_k are intercept and slope for each line $y_k(x) = a_k - b_k x$, with $a_k = \beta + \nu q_k(1-p)/\alpha_k$ and $b_k = -\nu(1-\alpha_k)/\alpha_k$. Let $[A_k, B_k]$ denote the interval such

that $y_x(x) < y_j(x)$ for all$x \in [A_k, B_k]$ and $j \neq k$. Define $B_m = q_1(1-p)$. Then

$$P_{H_o}(\mathcal{P} < p) = 1 - \Bigg(\int_{-\infty}^{x_1} \phi(y)\ e^{-e^{-a_6}}\ dy + \int_{x_1}^{x_2} \phi(y)\ e^{-e^{-(a_5+b_5 y)}}\ dy + \int_{x_2}^{x_3} \phi(y)\ e^{-e^{-(a_4+b_4 y)}}\ dy + \int_{x_3}^{x_4} \phi(y)\ e^{-e^{-(a_3+b_3 y)}}\ dy + \int_{x_4}^{q_1(1-p)} \phi(y)\ e^{-e^{-(a_2+b_2 y)}}\ dy \Bigg). \quad (3.22)$$

The quantity $q_k(1-p)$ is obtained numerically from the convolution of S_n and M_n which are asymptotically independent. Thus, we want to find c such that $F_{T_{\alpha_k}}(c) \geq 1-p$ under H_o. Therefore,

$$\begin{aligned} 1 - p &= P_{H_o}(T_{\alpha_k} \leq t) \\ &= \begin{cases} F_{S_n}(t), & \text{for } \alpha_k = 0, \\ F_{M_n}(t), & \text{for } \alpha_k = 1, \\ \int F_{M_n}\left(\frac{t-(1-\alpha_k)s}{\alpha_k}\right)\ f_{S_n}(s)\ ds & \text{for } 0 < \alpha_k < 1. \end{cases} \end{aligned} \quad (3.23)$$

Let p^* be the cut-off for a size-α test. The power of the test is then computed numerically as follows: for a given set of $\alpha_k, k = 1, ..., m$, we compute $T_{\alpha_k} = t_{obs}$ and the corresponding p value

$$p_{\alpha_k} = P_{H_o}(T_{\alpha_k} > t_{obs}) = \begin{cases} 1 - F_{S_n}(t_{obs}), & \text{for } \alpha_k = 0, \\ 1 - F_{M_n}(\beta + \nu t_{obs}), & \text{for } \alpha_k = 1, \\ 1 - \int F_{M_n}\left(\beta + \nu \frac{t_{obs}-(1-\alpha_k)s}{\alpha_k}\right)\ f_{S_n}(s)\ ds & \text{for } 0 < \alpha_k < 1. \end{cases}$$

We then compute our test statistic $\min_{1 \leq k \leq K} p_{\alpha_k}$ and compute the power numerically.

Selection of α with the strongest signal

Another way of determining α is finding $\alpha^* \in [0,1]$ that maximizes the test statistic $T^*(\alpha) = \frac{\alpha G_n + (1-\alpha)S_n}{\sqrt{\alpha^2+(1-\alpha)^2}}$. Then the test statistic is

$$T = \max_{0\leq\alpha\leq 1} T^*(\alpha).$$

Using the second derivative test, we get

$$\alpha^* = \begin{cases} \frac{G_n}{S_n+G_n}, & \text{for } G_n > 0, S_n > 0 \text{ or } G_n < 0, S_n < 0, \\ 1, & \text{for } G_n > 0, S_n < 0, \\ 0, & \text{for } G_n < 0, S_n > 0. \end{cases}$$

Therefore, the test statistic is

$$T = \begin{cases} \sqrt{S_n^2 + G_n^2}, & \text{for } \ G_n > 0, S_n > 0, \\ -\sqrt{S_n^2 + G_n^2}, & \text{for } \ G_n < 0, S_n < 0, \\ G_n, & \text{for } G_n > 0, S_n < 0, \\ S_n, & \text{for } G_n < 0, S_n > 0. \end{cases}$$

We want to find $t \in \mathbb{R}$ such that $P_{H_o}(T > t) \leq 0.05$.

$$\begin{aligned} 0.05 = P_{H_o}(T > t) = P_{H_o}\Big(\sqrt{S_n^2 + G_n^2} > t, G_n > 0, S_n > 0\Big) + \\ P_{H_o}\Big(- \sqrt{S_n^2 + G_n^2} > t, G_n < 0, S_n < 0\Big) + \\ P_{H_o}\Big(G_n > t, G_n > 0, S_n < 0\Big) + \\ P_{H_o}\Big(S_n > t, G_n < 0, S_n > 0\Big) \\ = p_1 + p_2 + p_3 + p_4. \end{aligned} \tag{3.24}$$

where

$$
\begin{aligned}
p_1 &= \int_0^{|t|} P_{H_0}\big[M_n > \beta + v\sqrt{t^2 - s^2}\big]\phi_{S_n}(s)ds, \\
p_2 &= \int_{|t|}^{0} P_{H_0}\big[\beta - v\sqrt{t^2 - s^2} \le M_n \le \beta\big]\phi_{S_n}(s)ds, \\
p_3 &= P_{H_0}[S_n < 0]\ P_{H_0}[M_n > \max(\beta, \beta + tv)], \\
p_4 &= P_{H_0}[S_n > \max(t, 0)]\ P_{H_0}[M_n < \beta].
\end{aligned}
\tag{3.25}
$$

Let t^* be the cut-off for a level-α test. Then the power of the test is computed numerically.

3.7.2 Test Statistic 2

Whereas S_n in the test statistic T_1 is essentially the Neyman Pearson test for the dense alternative $\boldsymbol{\mu} = \mathbf{1}$ and $\max_i x_i$ is expected to perform well in the sparsest of signals $\boldsymbol{\mu} = (1, 0, ..., 0)$, they cannot capture the entire spectrum of alternatives. Therefore we propose a test that is more adapted to different degree of sparsity. For $\gamma \in [-1, 1]$, we define

$$
\begin{aligned}
T(\gamma) &= \sum_{i=1}^{n} x_i I(x_i > \text{sign}(\gamma)\sqrt{2|\gamma| \log n}) \\
L(\gamma) &= \frac{T(\gamma) - E_{H_0}(T(\gamma))}{\sqrt{V_{H_0}(T(\gamma))}}.
\end{aligned}
\tag{3.26}
$$

For $\gamma \in [-1, 1]$, define $\gamma^* = \text{sign}(\gamma)\sqrt{2|\gamma| \log n}$.

$$
\begin{aligned}
E_{H_0}(T(\gamma)) &= n\ \phi(\gamma^*) \\
V_{H_0}(T(\gamma)) &= n\Big(\gamma^*\ \phi(\gamma^*) + \Phi(-\gamma^*) - \phi^2(\gamma^*)\Big).
\end{aligned}
\tag{3.27}
$$

For a finite set of m grid points $(\gamma_1, \gamma_2, \ldots, \gamma_m) \in [-1, 1]$, it is easy to see that $L = (L(\gamma_1), L(\gamma_2), \ldots, L(\gamma_m))$ has asymptotic multivariate normal distribution and

is independent of max x_i. Let p_γ be the corresponding p value. We define our test statistic as

$$\mathcal{P} = \min_{\gamma \in \Gamma} p_\gamma. \tag{3.28}$$

where $\Gamma = \{\gamma_1, \gamma_2, \ldots, \gamma_m, \infty\}$ and define $L(\infty) = \frac{X_{(n)} - a_n}{b_n}$. Under H_0,

$$Pr\left[\frac{X_{(n)} - a_n}{b_n} \leq z\right] \to G(z) = e^{-e^{-z}}. \tag{3.29}$$

the standard Gumbel Distribution where $a_n = \sqrt{2\log n} - \frac{\log\log n + \log 4\pi}{2\sqrt{2\log n}}$ is the location and $b_n = \frac{1}{\sqrt{2\log n}}$ the scale. As γ increases, $L(\gamma)$ takes care of sparse signals while $L(\gamma_1)$ focuses on dense case.

Using asymptotic multivariate normality of $L(\gamma_1, \gamma_2, \ldots, \gamma_m)$ and the fact that it is independent of $X_{(n)}$, we can compute $P(\mathcal{P} < p)$ numerically. For a level-α test, we want p such that $P_{H_o}(\mathcal{P} < p) \leq 0.05$. Therefore,

$$\begin{aligned}
0.05 &= P_{H_o}(\mathcal{P} < p) \\
&= P_{H_o}\Big(\min_{\gamma \in \Gamma} p_\gamma < p\Big) \\
&= 1 - P_{H_o}\Big(p_\gamma \geq p \;\forall\; \gamma \in \Gamma\Big) \\
&= 1 - P_{H_o}\Big(L(\gamma) \leq q_\gamma(1-p) \;\forall\; \gamma \in \Gamma\Big) \\
&= 1 - P_{H_o}\Big(L(\gamma_k) \leq q_{\gamma_k}(1-p) \;\forall\; \gamma = 1(1)m\Big)\; P_{H_o}\Big(L(\infty) \leq q_\infty(1-p)\Big).
\end{aligned} \tag{3.30}$$

where $q_{\gamma_k}(t) = \Phi^{-1}(t)$ is the tth quantile of $L(\gamma_k)$ and $q_\infty(t) = G^{-1}(t)$ is the tth quantile of $L(\infty)$ under H_o. Under H_0, $(L(\gamma_1), \ldots, L(\gamma_m))^T \sim \mathrm{N}_m(\mathbf{0}, \Sigma)$ where $\mathbf{0} = (0, \ldots, 0)^T$ with the diagonals of Σ equal to 1 and the off-diagonal terms are

given by

$$\sigma_{ij} = \frac{n}{\sqrt{v_i v_j}}\Big(\sqrt{2\max(\gamma_i,\gamma_j)\log n}\ \phi\big(\sqrt{2\max(\gamma_i,\gamma_j)\log n}\big)\Big) + \Phi\Big(-\sqrt{2\max(\gamma_i,\gamma_j)\log n}\Big) - \phi\big(\sqrt{2\gamma_i \log n}\big)\Big)\phi\big(\sqrt{2\gamma_j \log n}\big)\Big).$$

Let p^* be the cut-off for a size-α test. The power of the test is computed numerically as

$$Power = P_{H_1}\Big(\min_{\gamma\in\Gamma} p_\gamma < p^*\Big). \tag{3.31}$$

As mentioned before, the maximum order statistic performs well as a test statistics only for the sparsest of cases. To improve the performance at other sparse parameter configuration, we look at the sum of a fixed number of top order statistics.

A modification of the test statistic (2.14) is considering the standardized $T(\gamma)$ for $(\gamma_1, \gamma_2, \ldots, \gamma_m) \in [0,1]$ and standardized sum of r extreme order statistics. Let us define for $r \in \mathcal{N} = \{1, ..., n\}$

$$\begin{aligned} M_n(r) &= \sum_{i=1}^{r} X_{(n-i+1)} \\ G_n(r) &= \frac{M_n(r) - E_{H_0}(M_n(r))}{\sqrt{\widehat{Var_{H_0}(M_n(r))}}}. \end{aligned} \tag{3.32}$$

We define our test statistic as

$$T = \max_{\gamma\in\Gamma r\in\mathcal{N}} \{L(\gamma), G_n(r)\}. \tag{3.33}$$

We use Monte-Carlo to obtain $E_{H_0}(M_n(r))$ and $E_{H_0}(L(\gamma))$ for this test statistic. We will investigate the asymptotic distribution of T. We intend to use the idea presented in Csorgo *et al.* (1991) to find the asymptotic distribution of extreme sums [68]. For

now, we find the cut-off for T under H_0 numerically as well as compute the power numerically.

A test that will adapt to the level of sparsity can be constructed using the top order statistics. Specifically, if π is the proportion of nonzero elements in $\boldsymbol{\mu}$, then analogous to $M_n(r)$ we one could define $M_n(\hat{r})$ where $\hat{r} = max(\lfloor \hat{\pi} n \rfloor, 1)$ and $\hat{\pi}$ is the empirical estimator of π. The asymptotic distribution theory in Csorgo (1991) has to be extended for this case. However, intuitively the estimator is appealing for the high-dimensional cases because it is adapted to removing noise components.

3.7.3 Test Statistic 3

Given that the Neyman Pearson for the dense alternative is a linear statistic and the LRT which performs better for the sparse case, involves squared sample values, we investigate statistics with powers in the range from linear to quadratic. Recently, Kock and Preinerstorfer (2021) have established consistency of p-norm based tests for a wide range of alternatives against the zero mean null in a high dimensional setting.

For $\zeta \in [1, 2]$, we define

$$T(\zeta) = \sum_{i=1}^{n} x_i^{\zeta} I(x_i > 0)$$

$$L(\zeta) = \frac{T(\zeta) - E_{H_0}(T(\zeta))}{\sqrt{\widehat{Var_{H_0}(T(\zeta))}}}.$$

where

$$E_{H_0}(T(\zeta)) = n\Gamma\Big(\frac{\zeta+1}{2}\Big)\frac{2^{\frac{\zeta-1}{2}}}{\sqrt{2\pi}}$$

$$V_{H_0}(T(\zeta)) = n\left[\frac{2^{\zeta-\frac{1}{2}}}{\sqrt{2\pi}}\Gamma\Big(\zeta+\frac{1}{2}\Big) - \left(\Gamma\Big(\frac{\zeta+1}{2}\Big)\frac{2^{\frac{\zeta-1}{2}}}{\sqrt{2\pi}}\right)^2\right].$$

Then $L = (L(\zeta_1), L(\zeta_2), \ldots, L(\zeta_m))$ has asymptotic multivariate normal distribution and is independent of $\max_i x_i$. If p_ζ is the corresponding p value, then the test statistic is

$$\mathcal{P} = \min_{\zeta\in\zeta} p_\zeta$$

where $\zeta = \{\zeta_1, \zeta_2, \ldots, \zeta_m, \infty\}$ is a finite set. Define $L(\infty) = \frac{\max_i x_i - a_n}{b_n}$ where $a_n = \sqrt{2\log n} - \frac{\log\log n + \log 4\pi}{2\sqrt{2\log n}}$ and $b_n = \frac{1}{\sqrt{2\log n}}$. As ζ increases, $L(\zeta)$ takes care of sparse signals while $L(1)$ focuses on the dense cases. Using independence of T and $\max_i x_i$ and asymptotic multivariate normality of T, we can compute $P(\mathcal{T} > t)$ numerically. For a level-α test, we want p such that $P_{H_o}(\mathcal{P} < p) \leq 0.05$. Therefore,

$$\begin{aligned} 0.05 &= P_{H_o}(\mathcal{P} < p) \\ &= 1 - P_{H_o}\Big(L(\zeta_k) \leq q_{\zeta_k}(1-p)\ \forall\ \zeta = 1(1)m\Big)\ P_{H_o}\Big(L(\infty) \leq q_\infty(1-p)\Big). \end{aligned} \tag{3.34}$$

where $q_{\zeta_k}(t) = \Phi^{-1}(t)$ is the tth quantile of $L(\zeta_k)$ and $q_\infty(t) = G^{-1}(t)$ is the tth quantile of $L(\infty)$ under H_o using the distributional assumption in equation (3.29). Under H_0, $(L(\zeta_1), \ldots, L(\zeta_m))^T \sim \mathrm{N}_m(\mathbf{0}, \Sigma)$ where $\mathbf{0} = (0, \ldots, 0)^T$ with the diagonals of Σ equal to 1 and the off-diagonal terms are given by

$$\sigma_{ij} = n\left[\frac{2^{\frac{\zeta_1+\zeta_2+1}{2}}}{\sqrt{2\pi}}\Gamma\Big(\frac{\zeta_1+\zeta_2+1}{2}\Big) - \Gamma\Big(\frac{\zeta_1+1}{2}\Big)\frac{2^{\frac{\zeta_1-1}{2}}}{\sqrt{2\pi}}\Gamma\Big(\frac{\zeta_1+1}{2}\Big)\frac{2^{\frac{\zeta_1-1}{2}}}{\sqrt{2\pi}}\right].$$

Let p^* be the cut-off for a size-α test. The power of the test is computed numerically similarly as test statistic 2 in section 3.7.2.

As before we can modify the test statistic above to consider the standardized $T(\zeta)$ for $(\zeta_1, \zeta_2, \ldots, \zeta_m) \in [1, 2]$ and standardized sum of r extreme order statistics. Let us define for $r \in \mathcal{N} = \{1, ..., n\}$

$$\begin{aligned} M_n(r) &= \sum_{i=1}^{r} X_{(n-i+1)} \\ G_n(r) &= \frac{M_n(r) - E_{H_0}(M_n(r))}{\sqrt{\widehat{Var_{H_0}(M_n(r))}}}. \end{aligned} \tag{3.35}$$

We define our test statistic as

$$T = \max_{\zeta \in \zeta r \in \mathcal{N}} \{L(\zeta), G_n(r)\}. \tag{3.36}$$

We use Monte-Carlo to obtain $E_{H_0}(M_n(r))$ and $E_{H_0}(L(\zeta))$ for this test statistic. We will investigate the asymptotic distribution of test statistic T. For now, we find the cut-off for T under H_0 numerically as well as compute the power numerically.

Figure 3.7 presents a comparison of the power performance for different classes of tests from section 3.7. Among the possible tests, the power performance of test statistic 4 is relatively better with respect to the minimum power criteria. Hence, we pursue our proposed high dimensional omnibus test statistic (HD) for further analysis.

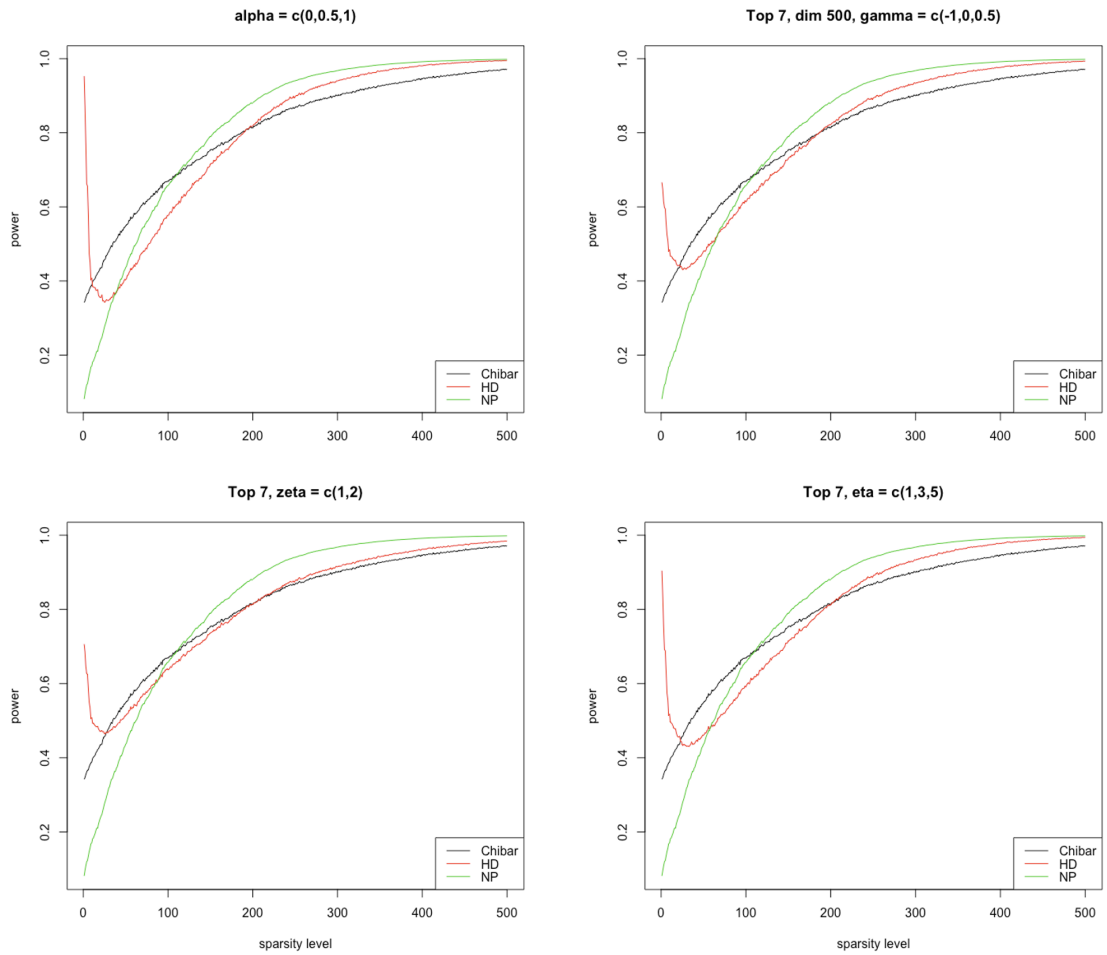

Figure 3.7: Power curve for the possible test statistics 1, 2, 3 and 4 clockwise from topleft along with LRT and NP for decreasing sparsity level when $\Sigma = I$.

Chapter 4: Efficient Integration of Aggregate Data and Individual Patient Data in One-Way Mixed Models

4.1 Introduction

Meta-analysis of individual participant data (IPD) is the gold standard statistical approach in systemic reviews of randomized clinical trials. [69] However, if IPD studies systematically differ from studies without access to IPD, synthesizing information solely based on IPD studies may lead to data availability bias or reviewer selection bias. [70] In many cases, IPD may not always be publicly available and access to IPD may be restricted due to limited resources. If only summary data are available on quantities of interest, meta-analysis of aggregate data (AD) approaches are the only recourse an investigator has for combining information about variables of interest across different studies. In addition to AD studies, if some IPD studies are available, combining these two levels of data could improve the overall meta-analysis estimates, compared to utilizing AD studies alone. [71, 72] Substitution of AD with IPD is generally advocated '*despite the extra cost, time and complexity required to obtain and manage raw data*'. [73] However, in certain situations the meta-analysis based exclusively on aggregate data (AD-MA) may yield estimators

that are comparable to those based solely on individual participant data (IPD-MA) and hence substitution of AD with IPD may not be worthwhile due to the extra cost.

In this paper, we discuss how to aggregate information from IPD studies and AD studies for a generic model and obtain a combined estimate for the parameter of interest that is cost effective yet efficient. We denote the combined estimator based on individual participant data and aggregate data meta-analysis as IPD-AD-MA. We explore many such combinations and investigate trade-offs between efficiency gains from substituting AD with IPD versus the cost of obtaining IPD studies when it is not easily available. The focus of the paper is to analyze whether it matters which combination of IPD and AD are selected in obtaining the IPD-AD-MA estimator. We then propose a novel selection algorithm for selecting which AD to substitute with IPD for maximum efficiency gain.

One common application of combining studies is testing the effectiveness of new treatments or medical interventions in randomized clinical trials (RCTs). Typically, RCTs tend to differ in study design and are often conducted across different centers, which may result in conflicting evidence. Thus, meta-analysis approaches have become the norm to integrate the results by borrowing power across different trials and provide an all-inclusive conclusion. This paper aims to exploit the design of the trials. Combination of results from IPD and AD studies work well when the model across the studies is simple with a common parameter of interest, e.g. treatment effect, and without disparate covariates. Due to randomization, RCTs on a common treatment provide ideal setup for exploring combination designs of

IPD-AD-MA.

Consider k independent studies with the jth study resulting in the estimated effect size $\hat{\beta}_j$, an estimate of the population effect size β_j for $j = 1, \ldots, k$. Suppose $\widehat{v(\hat{\beta}_j)}$ is the estimated variance of $\hat{\beta}_j$. When studies are homogeneous i.e, $\beta_j = \beta$ for all j, β is estimated efficiently by a fixed effects model, using weighted combination of $\hat{\beta}_j$ with the weights taken to be the reciprocal of $\widehat{v(\hat{\beta}_j)}$. When studies are heterogeneous, a random effects model is considered to account for the between study variances, τ^2. The meta-analysis estimate of β is a weighted mean of $\widehat{v(\hat{\beta}_j)}$ where the weights for each study is given by the reciprocal of $\widehat{v(\hat{\beta}_j)} + \tau^2$. Some widely used estimators for τ^2 are based on the Cochran's homogeneity test statistic.[74–76] Several other estimators for the heterogeneity variance are available.[77–83] Many papers like van Houwelingen *et al* (2002), Ritz *et al* (2008) have investigated the multivariate extension of these meta-analysis models.[84, 85]

The meta-analysis literature is rich with efficiency comparison of AD-MA estimators with the estimators obtained from full data based on different models. For treatment vs. control comparison with continuous outcome, Olkin and Sampson (1998) showed AD-MA estimator is equivalent to IPD-MA estimator if there is no study-by-treatment interactions and variances are constant across trials.[86] Mathew and Nordstrom (1999) further showed that this equivalence holds even if the error variances are different across trials.[87] The performance of IPD-MA estimator has been found to be similar but not identical to AD-MA estimator empirically (e.g. Whitehead, 2002, Ch. 5).[88] For a more general linear model with fixed treatment and random trial effect, Mathew and Nordstorm, 2010 provided conditions under

which AD analysis and IPD analysis coincide.[89] For all commonly used parametric and semi-parametric models, Lin and Zeng (2010) showed that IPD-MA estimator has no gain in efficiency over AD-MA estimator asymptotically in the context of fixed effects model and also provided the condition for their equality.[90] Liu *et al* (2015) introduced a meta-analysis approach for heterogeneous studies by combining the confidence density functions derived from the summary statistics of individual studies.[91] Doneal *et al* (2015) compared the performance of IPD-MA and AD-MA using different estimation procedures in generalized linear mixed model for binary outcomes.[92]

One potential limitation in standard meta-analysis approach is the requirement of a common set of parameters across studies. Different studies often tend to include different sets of covariates. For meta-analysis of IPD, Jackson *et al* proposed a method to estimate the fully adjusted effect across studies with different set of confounders.[93] Recently, Kundu *et al* developed an extension of meta-analysis method for fixed-effects model to combine information from studies with disparate covariate information.[94]

An extensive efficacy analysis of the one-stage and two-stage statistical methods for combining IPD and AD in meta-analysis for continuous outcome was explored by Riley *et al* among others.[95–98] In the simple situation of a fixed effects model with a single continuous outcome and covariate, Yamaguchi *et al* proposed a method to reconstruct the missing IPD for AD trials by a Bayesian sampling procedure and use the mixture of simulated IPD and collected IPD for an IPD meta-analysis.[99] Over the past decade, meta-analysis methods for mixture data

have also been developed for dichotomous outcomes and time-to-event data, some based on reconstruction of IPD.[97, 100, 101] Other popular methods for integrating binary data is random-effects mixed treatment comparison (MTC) models and likelihood based approaches.[102–106]

In section 4.2, we start with the case of a continuous response following a linear model and provide the combined treatment effects across trials when the treatment effect is fixed and common across trials while the trial effect is random. Assuming the observations within and between the studies are independent, we investigate the loss of efficiency from using combined estimator with various percentages of AD and IPD studies. When treatments are fixed and trial effects are random, Mathew and Nordstorm (2010) derived the necessary and sufficient condition for the IPD-MA estimator to coincide with AD-MA estimator for a general within trial covariance matrix. [89] The condition for equality requires that the fraction of observations corresponding to any given treatment to be same across trials. In practice, it is more likely to have studies with differential allocation to treatments. For such models, we study the relative efficiency of an estimator based on combining IPD and AD studies, denoted by IPD-AD-MA to the IPD-MA estimator under systematic departures from the same allocation proportion condition. We further propose a method to select the IPDs among the available studies so as to get the maximum efficiency in terms of the combined estimator.

In section 4.3, we propose a method of combining information across IPD and AD studies for a multidimensional parameter in a generalized linear mixed model (GLMM) and study the performance of the combined estimator empirically. This is a

more general setup where the covariate or response may be categorical or continuous and the common parameter of interest can be multidimensional. In addition, the random effects may not necessarily be independent. As a special case, we consider a logistic model with a similar setup to the LMM framework. For this case, we derive a relative efficiency expression and use the expression to propose efficient selection of IPD when combining IPD and AD studies for the IPD-AD-MA estimator.

We use a real data example to illustrate efficient selection of IPD when synthesizing information on the common parameter of interest across IPD and AD studies for a linear mixed model and a logistic model with random study effect. For all our analyses, we assume that all studies or trials are independent, which is a general and common assumption in the meta-analysis literature.

4.2 Efficient Aggregation in Linear Mixed Model (LMM)

Consider that there is one continuous outcome of interest and assume that there are two groups, namely treatment(T) group and control(C) group for all k independent studies. Let y_{ji} be the response of the ith participant in study j and x_{ji} is the corresponding treatment indicator. Suppose further that, n_{jT} and n_{jC} be the number of persons in the treatment group and the control group, respectively, for jth study with $n_{jT} + n_{jC} = n_j$. Suppose, for k_1 studies, IPD are available and for the remaining $k - k_1 = k_2$ studies, we have access to only AD. Let S_1 and S_2 denote the set of IPD studies and AD studies, respectively where $S_1 \cup S_2 = S$, S being the set with all studies. The model is

$$y_{ji} = \alpha_j + \beta x_{ji} + \epsilon_{ji}, \quad j = 1, ..., k,$$
$$\epsilon_{ji} \sim N(0, \sigma_j^2), \quad \alpha_j \sim N(\alpha, \sigma_\alpha^2), \tag{4.1}$$

where α_j and ϵ_{ji} are assumed independent. Our parameter of interest is the common treatment effect β.

4.2.1 Aggregation

For model (4.1), α acts as a nuisance parameter common across studies. Thus, the AD-MA estimator and the IPD-MA estimator does not necessarily coincide although the parameter of interest is common across studies (Mathew and Nordstorm 2010).[89] In fact, the IPD-MA estimator, which is the Best Linear Unbiased Estimator (BLUE), is more efficient than the AD-MA estimator for any finite and fixed k and n_j. [89] The two estimators coincide if and only if the vectors $(n_{jT}/n_j, n_{j2}/n_j)$ are all equal for $j = 1, ..., k$. Asymptotically, the AD-MA estimator has the same efficiency as the IPD-MA estimator.

For the k_1 studies with access to IPD, $\mathbf{y}_j$ is normal with mean and covariance given by

$$E(\boldsymbol{y_j}|X_j) = \alpha \mathbf{1}_{\boldsymbol{n_j}} + \beta X_j$$

$$Cov(\boldsymbol{y_j}|X_j) = H_j = \sigma_\alpha^2 \mathbf{1}_{\boldsymbol{n_j}} \mathbf{1}_{\boldsymbol{n_j}}^{\boldsymbol{T}} + \sigma_j^2 I_{n_j}.$$

For the k_2 AD studies, the maximum likelihood estimates (MLE), $\hat{\beta}_j$, and their

estimated variances, $\widehat{v(\hat{\beta}_j)}$ are available. The model for AD study is

$$\hat{\beta}_j \sim N(\beta, \widehat{v(\hat{\beta}_j)}), \quad j = 1, \ldots, k_2$$

$$v(\hat{\beta}_j) = \left(n_j \pi_j (1 - \pi_j)\right)^{-1} \sigma_j^2$$

where $\pi_j = n_{jT}/n_j$ is the proportion of treatment in study j.

Integration of IPD and AD uses the standard weighted combination approach with weights being inversely proportional to the variance, where the variance for AD part is simply the variance of MLE whereas for the IPD, it is the variance-covariance matrix of the marginal distribution of the data. With the above model, the combined estimator of $\boldsymbol{\theta} = (\alpha, \beta)'$ and the variance of the combined estimator are

$$\begin{aligned} \hat{\boldsymbol{\theta}}_{IPD-AD-MA} &= (U^T \Sigma^{-1} U)^{-1} U^T \Sigma^{-1} Y^*, \\ Cov(\hat{\boldsymbol{\theta}}_{IPD-AD-MA}) &= (U^T \Sigma^{-1} U)^{-1}, \end{aligned} \tag{4.2}$$

where $Y^* = (\mathbf{y}_1', \ldots, \mathbf{y}_{k_1}', \hat{\beta}_1, \ldots, \hat{\beta}_{k_2})'$ and

$$U = \begin{pmatrix} \mathbf{1_{n_1}} & X_1 \\ \cdots & \cdots \\ \mathbf{1_{n_{k_1}}} & X_{k_1} \\ 0 & \mathbf{1_{k_2}} \end{pmatrix}, \ \Sigma = \begin{pmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{pmatrix}, \ \Sigma_1 = \begin{pmatrix} H_1 & & \\ & \ddots & \\ & & H_{k_1} \end{pmatrix}, \ \Sigma_2 = \begin{pmatrix} v(\hat{\beta}_1) & & \\ & \ddots & \\ & & v(\hat{\beta}_{k_2}) \end{pmatrix}.$$

Since the estimator in (4.2) is unbiased, for analyzing efficiency of the estimator, we obtain the expression for its variance. To derive the variance of the combined estimator for the treatment effect, we assume σ_j^2 and σ_α^2 are known. However, both the variance components are estimated from the data in section 4.2.3.4 to reflect

practical considerations. The variance is given by

$$v(\hat{\beta}_{IPD-AD-MA}) = \Bigg[\sum_{j \in S_1} \frac{n_j \pi_j (1 + n_j (1-\pi_j) b_j)}{\sigma_j^2 a_j} + \sum_{j \in S_2} \frac{n_j \pi_j (1-\pi_j)}{\sigma_j^2} - \frac{(\sum_{j \in S_1} \frac{n_j \pi_j}{\sigma_j^2 a_j})^2}{\sum_{j \in S_1} \frac{n_j}{\sigma_j^2 a_j}} \Bigg]^{-1} \tag{4.3}$$

where $a_j = 1 + n_j b_j,\ b_j = \frac{\sigma_\alpha^2}{\sigma_j^2}$ & $\pi_j = \frac{n_{jT}}{n_j}$.

The gold standard for the combined analysis is having access to all IPD studies. Thus, we compare the efficiency of $\hat{\beta}_{IPD-AD-MA}$ with the "best" estimator, which is the maximum likelihood estimate (MLE) based on IPD from all studies. The variance of the "best" estimator of β, also known as the minimum variance unbiased estimator is

$$v(\hat{\beta}_{IPD-MA}) = \Bigg[\sum_{j=1}^{k} \frac{n_j \pi_j (1 + n_j (1-\pi_j) b_j)}{\sigma_j^2 a_j} - \frac{\big(\sum_{j=1}^{k} \frac{n_j \pi_j}{\sigma_j^2 a_j}\big)^2}{\sum_{j=1}^{k} \frac{n_j}{\sigma_j^2 a_j}} \Bigg]^{-1}.$$

The variance of the combined estimator of β, assuming only AD results are available for all k studies, is

$$v(\hat{\beta}_{AD-MA}) = \Bigg[\sum_{j=1}^{k} \frac{n_j \pi_j (1-\pi_j)}{\sigma_j^2} \Bigg]^{-1}.$$

For the balanced homoscedastic case, when $n_j = n$ and $\sigma_j^2 = \sigma^2$, the variance expression for the combined estimator simplifies to

$$v(\hat{\beta}_{IPD-AD-MA}) = \frac{\sigma^2}{n} \Bigg[\frac{nb}{a} \sum_{j \in S_1} \pi_j (1-\pi_j) + \sum_{j \in S_2} \pi_i (1-\pi_i) + \frac{1}{a k_1} \Big(\sum_{j \in S_1} \pi_j \Big) \Big(\sum_{j \in S_1} (1-\pi_j) \Big) \Bigg]^{-1}$$

where $a = 1 + nb,\ b = \frac{\sigma_\alpha^2}{\sigma^2}$. Similarly, the variance expression for the all IPD estimator reduces to

$$v(\hat{\beta}_{IPD-MA}) = \sigma^2 \Bigg[\frac{n^2 b}{a} \sum_{j=1}^{k} \pi_j (1-\pi_j) + \frac{n}{ak} \Big(\sum_{j=1}^{k} \pi_j \Big) \Big(\sum_{j=1}^{k} (1-\pi_j) \Big) \Bigg]^{-1}.$$

The relative efficiency of the combined estimator with respect to the estimator with all IPD studies is $RE(\hat{\beta}_{IPD-AD-MA}) = \frac{v(\hat{\beta}_{IPD-MA})}{v(\hat{\beta}_{IPD-AD-MA})}$ which for the simple case when $n_j = n$ and $\sigma_j^2 = \sigma^2$ is given by

$$\frac{\left[\frac{n^2 b}{a}\sum_{j\in S_1}\pi_j(1-\pi_j) + n\sum_{j\in S_2}\pi_j(1-\pi_j) + \frac{n}{ak_1}\Big(\sum_{j\in S_1}\pi_j\Big)\Big(\sum_{j\in S_1}(1-\pi_j)\Big)\right]}{\left[\frac{n^2 b}{a}\sum_{j=1}^{k}\pi_j(1-\pi_j) + \frac{n}{ak}\Big(\sum_{j=1}^{k}\pi_j\Big)\Big(\sum_{j=1}^{k}(1-\pi_j)\Big)\right]}. \tag{4.4}$$

We will use the above expression for relative efficiency for simulation purposes in section 4.2.3.

4.2.2 Selection

The main question that we address in this paper is whether, while substituting IPD studies with the corresponding AD results or vise versa, it matters which studies are selected. If yes, how can we minimize the loss in efficiency occurring from using AD results instead of the IPD results for a study?

The variance of the combined estimator in equation (4.3) can be simplified to

$$v(\hat{\beta}_{IPD-AD-MA}) = \left[\sum_{j\in S_1}\frac{n_j}{\sigma_j^2 a_j}(\pi_j - \tilde{\pi}_{S_1})^2 + \sum_{j=1}^{k}\frac{n_j\pi_j(1-\pi_j)}{\sigma_j^2}\right]^{-1} \tag{4.5}$$

where $\tilde{\pi}_{S_1} = \frac{\sum_{j\in S_1}\frac{\pi_j n_j}{\sigma_j^2 a_j}}{\sum_{j\in S_1}\frac{n_j}{\sigma_j^2 a_j}}$.

The expression for variance in equation (4.5) does not involve the response $\boldsymbol{y}$ and depends only on π_j, n_j and σ_j. This gives us a way of selecting k_1 IPD studies among the k studies to optimize the efficiency of combined estimator. In other words, selection of IPD studies depends on choosing the best subset S_1 that

maximizes $\sum_{j\in S_1} \frac{n_j}{\sigma_j^2 a_j}(\pi_j - \tilde{\pi}_{S_1})^2$. Essentially this is a combinatorial optimization problem which can be formulated as follows:

Selection Problem: Given $k > 1$ and $1 \leq k_1 \leq k$, and $\mathcal{P}_k$, the power set of $\{1, \ldots, k\}$, find

$$A_{opt} = \underset{A\in\mathcal{P}_k:|A|=k_1}{\operatorname{argmax}} \sum_{j\in A} v_j(u_j - \bar{u}_A)^2 \tag{4.6}$$

where $(u_1, v_1), \ldots, (u_k, v_k)$ with $u_j > 0$, $v_j > 0$, are prespecified constants with $\bar{u}_A = \frac{\sum_{j\in A} v_j u_j}{\sum_{j\in A} u_j}$ for any $A \in \mathcal{P}_k$.

For the balanced homoscedastic case, the variance further reduces to

$$v(\hat{\beta}_{IPD-AD-MA}) = \frac{\sigma^2}{n}\left[\sum_{j\in S_1} \frac{1}{a}(\pi_j - \tilde{\pi}_{S_1})^2 + \sum_{j\in S} \pi_j(1-\pi_j)\right]^{-1}.$$

For this special case, an exact algorithm exists for finding the optimum set, A_{opt} of IPD studies. Firstly we arrange the studies in increasing order of their π_j values. We include studies with extreme π_j values in the IPD set, alternating between the two ends to choose a total of k_1 IPD studies.

This is consistent with the result from Mathew and Nordstorm (2010) where the IPD-MA estimator coincides with AD-MA estimator when the fraction of observations corresponding to any given treatment is same across trials in a linear model with fixed treatments and random trial effects. When combining IPD and AD studies, we propose allocating studies with similar proportion of treatments to the AD set and studies with widely varying proportion of treatments to the IPD set.

Algorithm 1 Sequential Selection Algorithm: SSA

Require: $k_1 \geq 2$

$\mathcal{S} \leftarrow 1 : k$

$(m, n) \leftarrow \underset{p,q \in \mathcal{S}}{\arg\max}\ (u[p] - u[q])^2/[1/v[p] + 1/v[q])$

$A_{SSA} \leftarrow \{m, n\}$

$B \leftarrow \mathcal{S} \backslash A_{SSA}$

$D \leftarrow v[m] + v[n]$

$M \leftarrow (v[m] * u[m] + v[n] * u[n])/D$

while $|A_{SSA}| \neq k_1$ **do**

$l \leftarrow \underset{r \in B}{\arg\max}\ v[r] * (u[r] - M)^2/(D + v[r])$

$A_{SSA} \leftarrow A_{SSA} \bigcup \{l\}$

$B \leftarrow B \backslash \{l\}$

$M \leftarrow (D * M + v[l] * u[l])/(D + v[l])$

$D \leftarrow D + v[l]$

end while

For a more general case when n_j or σ_j may not be equal for all j, we propose an approximate sequential algorithm for the selection problem stated in (4.6). The sequential selection algorithm, SSA is a forward selection algorithm that involves choosing the first two studies so that the objective function is maximized. Then, we select the next $k_1 - 2$ studies sequentially using the weighted mean based on the preceding selected studies. It sidesteps the computationally intensive task of calculating the weighted mean for every subset S_1 with cardinality k_1. Two vectors, u and v are fed as an input to the SSA algorithm and the output is the set A_{SSA} of

k_1 studies. The performance of the algorithm is assessed using simulation studies and the results are presented in section 4.2.3.3.

4.2.3 Simulation

Assuming the variance components are known without uncertainty, the variance expression in (4.5) does not involve any data and hence the results in the following sections do not involve any simulation. The discussion on estimation of the variance components and its sensitivity to the selection algorithm is covered in section 4.2.3.4.

For linear mixed effects model, we present two scenarios to show the importance of selecting appropriate studies for constructing the combined estimator. For simplicity, we consider both k and n to be small and equal to 10 and $\sigma^2 = 2.5, \sigma_\alpha^2 = 0.025$.

4.2.3.1 Uniform distribution of proportion of treatment

For this scenario, we use the uniform distribution to randomly generate $k = 10$ proportion of treatment over the interval $[0, 1]$ as provided in Table 4.1:

Table 4.1: Uniform distribution of proportion of treatment across studies

study	1	2	3	4	5	6	7	8	9	10
proportion	0.1	0.2	0.3	0.3	0.3	0.5	0.6	0.6	0.8	0.8

Figure 4.1 shows that if the desired relative efficiency is 0.9, we could achieve that with many possible combinations of 60% IPD and 40% AD studies. However

60% IPD may not be cost effective. But if we choose the right combination of 40% IPD and 60% AD, we can have a RE of 0.956 in this case. One such combination which maximizes the RE for 40% IPD is study no 1, 2, 9 and 10 for IPD with treatment proportion 0.1, 0.2, 0.8 and 0.8, respectively, and AD for the rest of the studies.

Figure 4.2 shows the maximum and minimum relative efficiency that can be achieved for different combinations of the IPD and AD studies. In the figure, the numbers at the top and bottom of the line plot indicate the IPD studies which achieve the maximum and minimum relative efficiency, respectively. For example, among all combinations with 20% IPD and 80% AD, the combination which yields maximum relative efficiency includes study 1 and 10 for the IPD part and rest of the studies for the AD part. The minimum efficiency is achieved with the combination of study 3 and 4 for the IPD part and the rest for the AD part.

4.2.3.2 Bathtub distribution of proportion of treatment

An example of unbalanced proportion of treatment is given in Table 4.2. The importance of selecting the best subset of IPD studies for the combined estimate is more clearly seen in this case. The AD-MA estimator has a relative efficiency of 0.44 compared to the IPD-MA estimator. From Figure 4.3, we see that it is hard to obtain a relative efficiency of 0.9 even with 80% IPD and 20% AD.

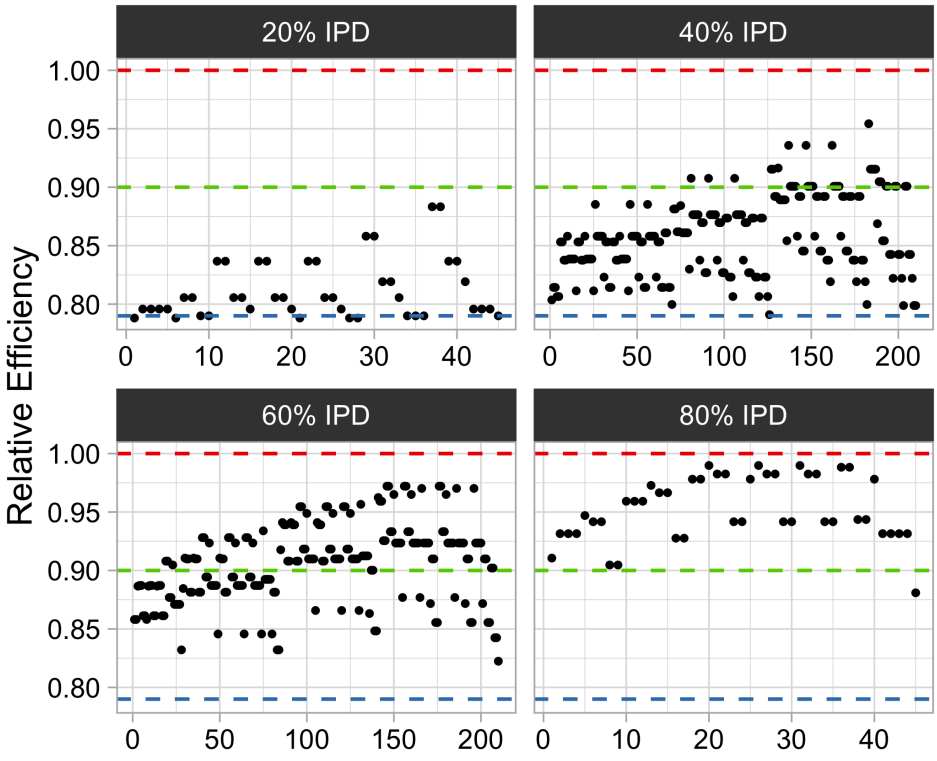

Figure 4.1: The relative efficiency of all 45 possible combinations for each of 20% IPD and 80% IPD, 210 possible combinations for each of 40% IPD and 60% IPD are plotted. Red dashed line represents the RE for IPD-MA estimator which is 1, blue dashed line is RE for AD-MA estimator which is 0.79 and the green dashed line represents the desired RE, say 0.9, for example.

Table 4.2: Unbalanced distribution of proportion of treatment across studies

study	1	2	3	4	5	6	7	8	9	10
proportion	0.1	0.1	0.1	0.1	0.2	0.8	0.9	0.9	0.9	0.9

The situation can worsen for severely unbalanced distribution of proportion of treatment. However, the loss in efficiency is an issue only with fixed small sample

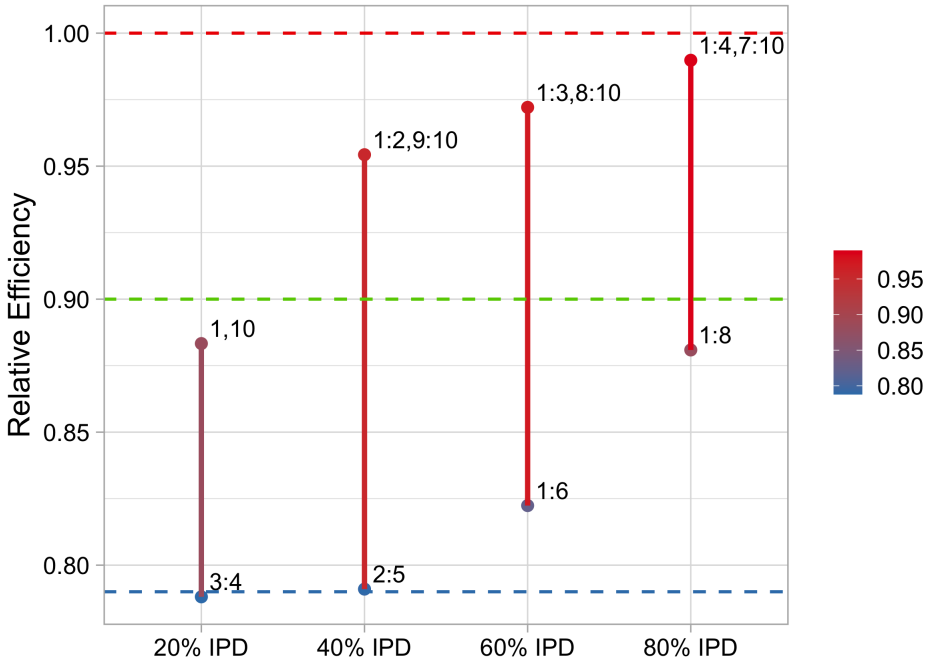

Figure 4.2: Plot showing the maximum and minimum relative efficiency for each percentage of IPD studies and the numbers at the top and bottom of the line plot indicate the IPD studies which achieves the maximum and minimum relative efficiency, respectively.

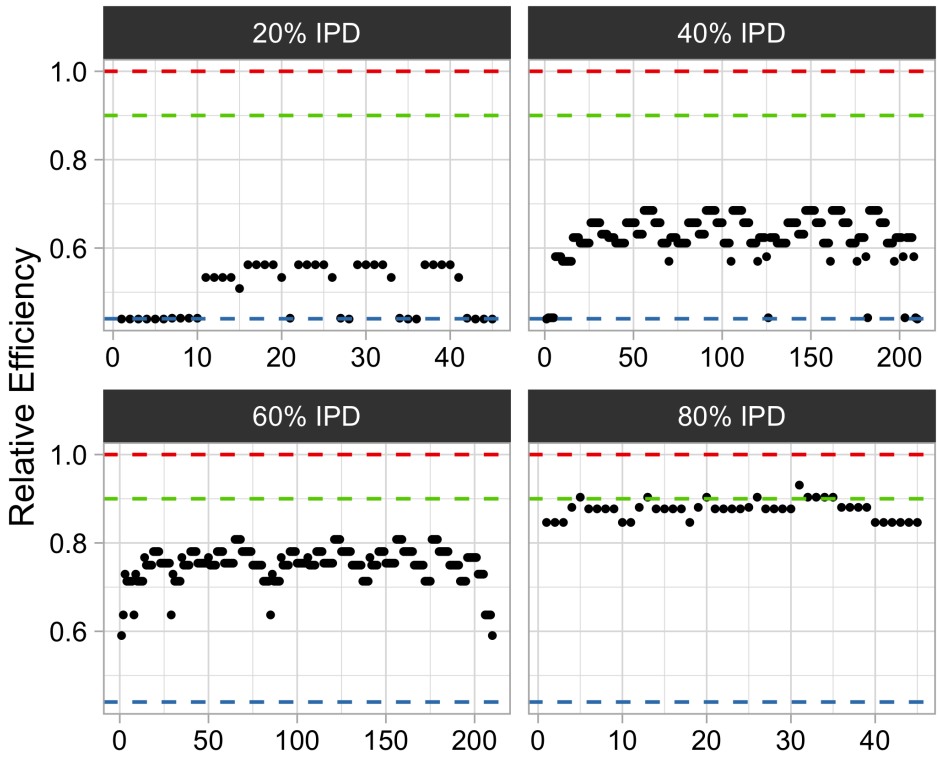

Figure 4.3: The relative efficiency of all 45 possible combinations for each of 20% IPD and 80% IPD, 210 possible combinations for each of 40% IPD and 60% IPD are plotted. Red dashed line represents the RE for IPD-MA estimator which is 1, blue dashed line is RE for AD-MA estimator which is 0.79 and the green dashed line represents the desired RE, say 0.9, for example.

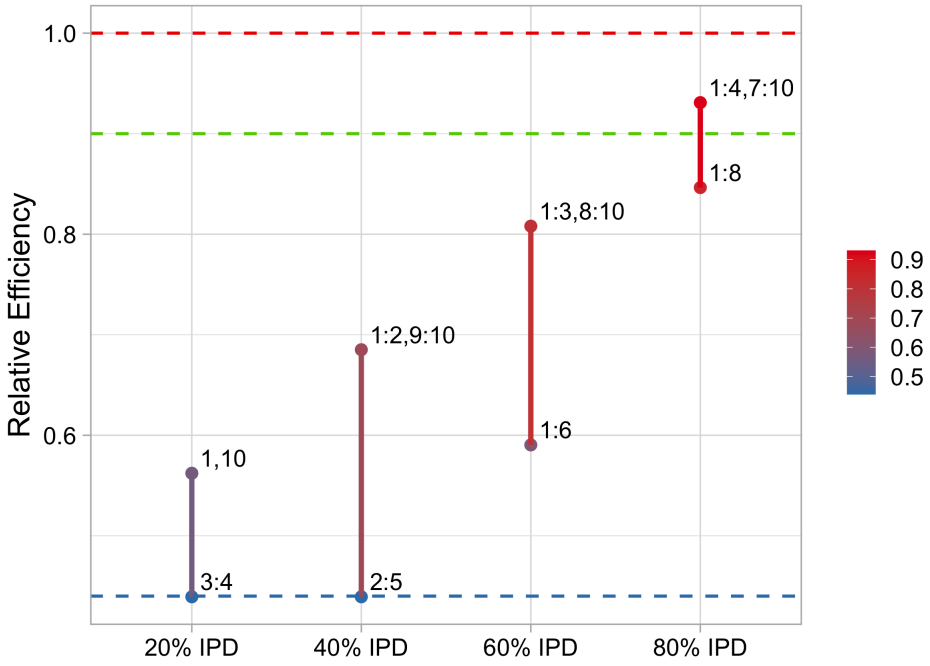

Figure 4.4: Plot showing the maximum and minimum relative efficiency for each percentage of IPD studies and the numbers at the top and bottom of the line plot indicate the IPD studies which achieves the maximum and minimum relative efficiency, respectively.

size studies. Asymptotically, the AD-MA combined estimator are quite efficient compared to IPD-MA combined estimator. Hence, the selection of IPD in data integration is useful for when combining treatment effect across randomized clinical trials (RCTs) with small cohort size. In such cases, one can obtain maximum efficiency of the combined estimator for a given number of IPD studies using the selection algorithm.

4.2.3.3 Selection of IPD studies through sequential algorithm

To assess the performance of sequential algorithm when n_j or σ_j are not equal, we consider $k = 30$ studies each with $n = 10$ and $\sigma_\alpha^2 = 0.025$. We simulate the σ_j's and π_j's as following:

$$\pi \sim 0.5 \text{ Beta}(\alpha = 2, \beta = 9) + 0.5 \text{ Beta}(\alpha = 9, \beta = 2),$$

$$\sigma^2 \sim \text{Inv-Gamma}(\alpha = 2, \beta = 5).$$

For $k_1 = 2, \ldots, 10$, we first obtain the optimal set with k_1 IPD studies for which the variance is minimized and then record the number of times the optimal set matches with the set obtained through the sequential algorithm. Table 4.3 presents the count of the matches for 100 simulations and the mean ratio of variance for the set from sequential algorithm to the variance of the optimal set. For larger k_1, the number of combinations is large and hence finding the optimal set can be time consuming and computationally expensive. The sequential algorithm, on the other hand, is quite fast and performs relatively well.

Table 4.3: Count of matches for a set of size k_1 out of $k = 10$ studies

	k_1								
Count Match	10	9	8	7	6	5	4	3	2
0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	2	0
2	0	0	0	0	0	0	2	8	100
3	0	0	0	0	0	3	15	90	
4	0	0	0	0	0	14	83		
5	0	0	0	0	7	83			
6	0	0	0	3	93				
7	0	0	6	97					
8	0	6	94						
9	9	94							
10	91								
Mean Ratio	1	1	1	1	1	1.001	1.001	1.001	1

4.2.3.4 Variance component estimation

In this section, we discuss about the estimation of the variance components σ_j^2 and σ_α^2 and its sensitivity to the sequential selection algorithm. When $\sigma_j^2 \neq \sigma_l^2$ for some j, l, σ_j^2 is estimated using the estimated variance of the treatment effect and the number of participants for the two groups in each study. One can also test for homogeneity of study variances and obtain a pooled estimate of σ^2 with the assumption $\sigma_j^2 = \sigma^2 \ \forall j$. Furthermore, σ_α^2 is estimated using some pilot IPD studies, that are accessible without any difficulty or effort.

To study the sensitivity of the variance component estimation to the selection algorithm, two different scenarios are considered: $\sigma_j^2 = \sigma^2 \ \forall j$ and $\sigma_j^2 \neq \sigma_l^2$ for some j, l. For each scenario, we generate 10000 data sets from the linear mixed effects model in (4.1) where $\alpha = 0.5, \beta = 1.5, \sigma_\alpha^2 = 0.025$ with $k = 10$ studies with $n_j = 50$ participants and proportion of treatment as provided in table 4.2. The σ_j's are simulated as:

$$\sigma^2 \sim \text{Inv-Gamma}(\alpha = 2, \beta = 5).$$

For each simulated data set, we find the optimal set of $k_1 = 5$ IPD studies when the σ_j's and σ_α's are known without uncertainty. Next, we derive the estimates of the variance components and obtain the set of $k_1 = 5$ IPD studies using the sequential algorithm based on these estimates. Table 4.4 reports the count of exact matches between the two sets. For both the cases, we found that the selection algorithm is robust to the estimation of these parameters.

Table 4.4: Count of matches for a set of size $k_1 = 5$ out of $k = 10$ studies

Count Match	$\sigma_j^2 = \sigma^2$	$\sigma_j^2 \neq \sigma^2$
0	0	0
1	0	0
2	0	0
3	0	386
4	607	3636
5	9393	5974

4.2.4 World Values Survey: Efficient Recombination

We illustrate the significance of selection of IPD studies when combining information across studies using data from the World Values Survey (WVS) (https://www.worldvaluessurvey.org/wvs.jsp). The WVS is a large consortium of social survey data from around 100 regions across 6 waves (1981-1984, 1990-1994, 1995-1998, 1999-2004, 2005-2009 and 2010-2017). The data consist of scores based on questionnaire covering a broad range of topics such as economic development, democratization, religion, gender equality, social capital, and subjective well-being.

For our analysis, we consider a simple linear mixed effects model based on the data for wave 6 in USA which contains 2232 participants. The dependent variable is life satisfaction score which ranges from 1 (completely dissatisfied) to 10 (completely satisfied). Based on financial satisfaction score, the predictor is a binary variable taking the value 1 for scores more than 5 and 0 otherwise. We consider category 1

as treatment and 0 as control for the predictor. We focus on individuals who are single and also account for heterogeneity due to age, sex, ethnicity and education in the study population through stratification. The variable age is categorized into intervals of 20 years as 0-20, 20-40, 40-60, 60-80 and more than 80. Ethnicity has 5 categories: Non-Hispanic white, Non-Hispanic Black, Hispanic, Non-Hispanic more than 2 races and other races while sex is categorized into male and female groups. The variable education has many categories starting with no formal education to doctorate degree. The strata are formed by considering different combinations of age, sex, ethnicity and education. We remove strata that have either only 1 or only 0 as response, resulting in treatment proportions to be 1 or 0. We also remove strata with number of participants (stratum size) less than 3. The total number of strata after exclusion is 30 with strata sizes ranging from 3 to 24 and the proportion of treatment ranging from 0.1 to 0.89. Each strata will be considered as a study and meta analysis will be in combining the strata specific results to obtain overall population results.

We assume that we have access to the AD for each study, i.e., the estimate of the treatment effect and the estimated standard error along with the number of participants in the treatment group and the control group are available for each study. The objective is to assess whether getting IPD is useful and determine which IPDs should be combined with AD to obtain maximum efficiency. We randomly sample 5 pilot studies among the 30 total studies and estimate $\sigma_\alpha^2 = 0.144$ using model (4.1). First, we consider the case when σ_j^2's are not same. The estimated variance for an all-IPD estimator of the treatment effect, $\hat{\beta}_{IPA-MA}$ is 0.033 whereas

for an estimator based on all-AD estimator, $\hat{\beta}_{AD-MA}$ the estimated variance is 0.038 resulting in an approximately 11% loss in efficiency. However, if we include 5 best subset IPD studies selected using the algorithm we proposed, the estimated variance for the combined estimator, $\hat{\beta}_{IPD-AD-MA}$ is 0.035 and hence we have a 7% gain in efficiency compared to $\hat{\beta}_{AD-MA}$.

Under the assumption $\sigma_j^2 = \sigma^2$, there is an approximate 12% loss in efficiency for the AD-MA estimator compared to the IPD-MA estimator. If we include the 5 best chosen IPD studies, there is a 9% gain in efficiency for the IPD-AD-MA estimator relative to the AD-MA estimator. Figure 4.5 shows which 5 studies are selected using the sequential algorithm. Both assumptions yield nearly the same set of best 5 IPD studies which reinforces that the selection algorithm is not sensitive to variance component estimation. Table 4.5 summarizes the pooled estimates and the standard errors for various scenarios.

Table 4.5: Pooled estimate and standard error of the treatment effect for IPD-MA estimator, IPD-AD-MA estimator with 5 best subset IPD and the AD-MA estimator

	$\sigma_j^2 = \sigma^2$			$\sigma_j^2 \neq \sigma^2$	
IPD Trials	$\hat{\boldsymbol{\beta}}$	**se**$(\hat{\boldsymbol{\beta}})$	**IPD Trials**	$\hat{\boldsymbol{\beta}}$	**se**$(\hat{\boldsymbol{\beta}})$
None	1.624	0.221	None	1.603	0.195
1:3, 29:30	1.762	0.207	1:2, 28:30	1.746	0.182
1:30	1.784	0.205	1:30	1.774	0.179

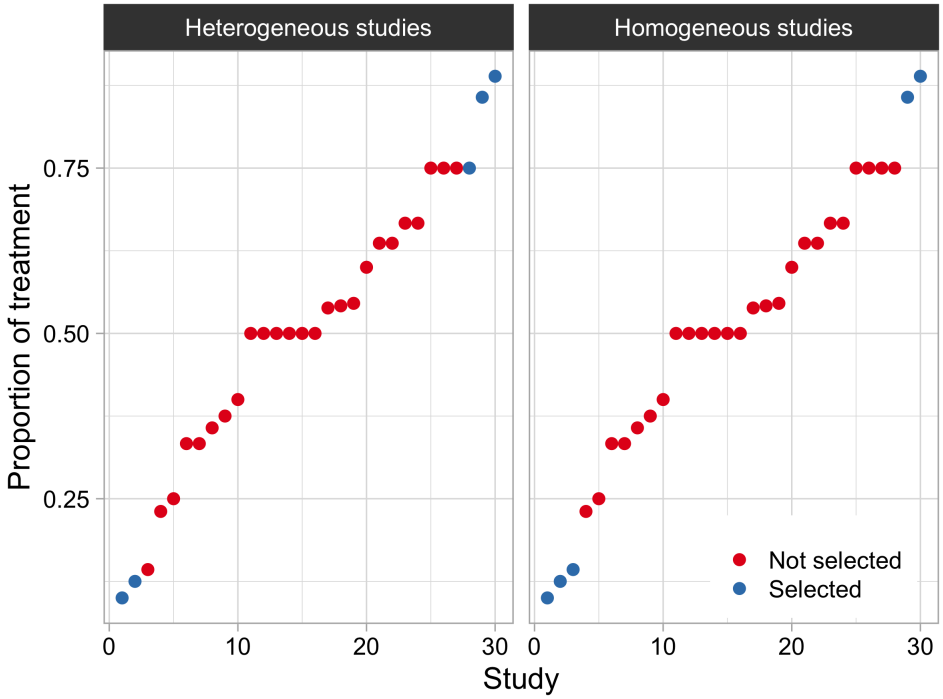

Figure 4.5: Plot showing the studies that are selected against the proportion of treatments when the study variances are homogeneous versus when they are heterogeneous.

4.3 Aggregation in Generalized Linear Mixed Model

For studies with non-Gaussian data, generalized linear models (GLM) form a large class, including popular models for binary response and count response. To aggregate several GLM studies with the same parameters of interest, one could use a meta analysis approach similar to that in the linear model case. We consider a generalized linear mixed model (GLMM) setup which allows for random effect terms for the study effect. We use a simple aggregation framework which will allow us to derive an efficiency expression for the treatment effect estimator similar to that in the LMM. We could then use the efficiency expression, written in terms of the treatment allocations to decide which AD could be replaced by IPD for maximum gain in efficiency.

We present the setup in a slightly more general form with several fixed effects and random effects terms. Suppose, for $j = 1, \ldots, k_1$ studies in set S_1, we have access to the IPD data whereas for k_2 studies in set S_2, only the summary statistics for the parameter of interest are available where $S_1 \cup S_2 = S$ is the total set of studies with $|S| = k$.

The IPD data $(Y_{ji}, X_{ji}|\beta_j, \alpha_j) \sim f_j(Y_{ji}, X_{ji}; \beta_j, \alpha_j)$ where f is a probability density belonging to the regular exponential family. We assume a random effect model for β_j and α_j, where both β_j and α_j may be multidimensional. We further assume a joint multivariate gaussian model for the random effects allowing for non-zero correlation between the β_j and α_j where β is the parameter vector of interest

and the vector α includes all nuisance parameters:

$$\begin{pmatrix} \beta_j \\ \alpha_j \end{pmatrix} \Big| \begin{pmatrix} \beta \\ \alpha \end{pmatrix} \sim MVN\left(\begin{pmatrix} \beta \\ \alpha \end{pmatrix}, \begin{pmatrix} \Sigma_{\beta\beta} & \Sigma_{\beta\alpha} \\ \Sigma_{\alpha\beta} & \Sigma_{\alpha\alpha} \end{pmatrix} \right)$$

with MVN denoting the multivariate normal density. This is a general model and can be reduced to specific mixed effects models or fixed effects model by constraining the corresponding components in the covariance matrix to be zero. This setup allows having additional covariates and the parameter corresponding to the additional covariates can be included in α_j.

For the k_2 AD studies, we have the MLE estimates of the main parameter vector with their estimated covariance matrices, $(\hat{\beta}_j, \widehat{V(\hat{\beta}_j)})$. The model for random effects is given by

$$\hat{\beta}_j | \beta_j \sim N(\beta_j, \widehat{V(\hat{\beta}_j)})$$

$$\beta_j | \beta \sim N(\beta, \Sigma_\beta).$$

We will use a composite likelihood where the likelihood from the studies with IPD and the likelihood from the studies with AD are simply multiplied together.[107] For the studies with only summary statistics available, we consider an approximate likelihood based on the asymptotic normal model for the maximum likelihood estimators (MLE). For the studies with IPD, we can write the full GLMM likelihood. But given the complexity of the GLMM likelihood involving integrals we use an approximate likelihood based on the Laplace approximation. We estimate the common parameter of interest using profile likelihood approach applied to the composite likelihood.

The full (log) likelihood function is obtained by combining IPD studies and

AD studies as:

$$L(\beta,\alpha)=\prod_{j\in S_1}L_j(\beta,\alpha)\prod_{j\in S_2}L_j(\beta),$$

$$\log L(\beta,\alpha)=l(\beta,\alpha)=\sum_{j\in S_1}l_j(\beta,\alpha)+\sum_{j\in S_2}l_j(\beta).$$

For each $k\in S_1$, the IPD part of log-likelihood, $l_j(\beta,\alpha)$ is constructed using the Laplace approximation (where the family f satisfies the standard smoothness assumptions for the Laplace approximation to work) to expand $l_j(\beta,\alpha)$ around the MLE, $(\hat\beta_j,\hat\alpha_j)$ from study j and ignoring the higher order terms $o(||\beta_j-\hat\beta_j||^2)$. Then, (β_j,α_j) is integrated out in the Gaussian integral to obtain

$$l_j(\beta,\alpha)=l_j(\hat\beta_j,\hat\alpha_j)-\frac{1}{2}\left[\begin{pmatrix}\hat\beta_j\\ \hat\alpha_j\end{pmatrix}-\begin{pmatrix}\hat\beta_j\\ \hat\alpha_j\end{pmatrix}\right]^T\Delta_j^{-1}\left[\begin{pmatrix}\hat\beta_j\\ \hat\alpha_j\end{pmatrix}-\begin{pmatrix}\hat\beta_j\\ \hat\alpha_j\end{pmatrix}\right]-\frac{1}{2}\log|\Delta_j| \quad (4.7)$$

where $\Delta_j=\left[\begin{pmatrix}\Sigma_{\beta\beta}&\Sigma_{\beta\alpha}\\ \Sigma_{\alpha\beta}&\Sigma_{\alpha\alpha}\end{pmatrix}+\left[\begin{pmatrix}I_{\beta_j\beta_j}&I_{\beta_j\alpha_j}\\ I_{\alpha_j\beta_j}&I_{\alpha_j\alpha_j}\end{pmatrix}\Big|_{\hat\beta_j,\hat\alpha_j}\right]^{-1}\right]^{-1}$.

The AD part of likelihood for each $k\in S_2$ is given by

$$l_j(\beta)=-\frac{1}{2}\left(\hat\beta_j-\beta\right)^T\left(\widehat{V(\hat\beta_j)})+\Sigma_{\beta\beta}\right)^{-1}\left(\hat\beta_j-\beta\right)-\frac{1}{2}\log|\widehat{V(\hat\beta_j)})+\Sigma_{\beta\beta}|. \quad (4.8)$$

Putting (4.7) and (4.8) together, we have the full log likelihood for (β,α) as

$$\begin{aligned}l(\beta,\alpha) \;=\; &\sum_{j\in S_1}l_j(\hat\beta_j,\hat\alpha_j)-\frac{1}{2}\sum_{j\in S_1}\left[\begin{pmatrix}\hat\beta_j\\ \hat\alpha_j\end{pmatrix}-\begin{pmatrix}\hat\beta_j\\ \hat\alpha_j\end{pmatrix}\right]^T\Delta_j^{-1}\left[\begin{pmatrix}\hat\beta_j\\ \hat\alpha_j\end{pmatrix}-\begin{pmatrix}\hat\beta_j\\ \hat\alpha_j\end{pmatrix}\right]-\frac{1}{2}\sum_{j\in S_1}\log|\Delta_j|\\ &-\frac{1}{2}\sum_{j\in S_2}(\hat\beta_j-\beta)^T\left(\widehat{V(\hat\beta_j)})+\Sigma_{\beta\beta}\right)^{-1}(\hat\beta_j-\beta)-\frac{1}{2}\sum_{j\in S_2}\log|\widehat{V(\hat\beta_j)}+\Sigma_{\beta\beta}|. \quad (4.9)\end{aligned}$$

which then can be maximized to obtain the estimates of mean parameters and the variance components. To reduce the computational burden, we lower the dimension

by expressing the mean parameter as a function of the variance components in (4.10).

$$\begin{pmatrix} \hat{\beta}(\Sigma) \\ \hat{\alpha}(\Sigma) \end{pmatrix} = \left[\sum_{j \in S_1} \Delta_j + \sum_{j \in S_2} \Omega_j \right]^{-1} \left[\sum_{j \in S_1} \Delta_j \begin{pmatrix} \hat{\beta}_j \\ \hat{\alpha}_j \end{pmatrix} + \sum_{j \in S_2} \Omega_j \begin{pmatrix} \hat{\beta}_j \\ \hat{\alpha}_j \end{pmatrix} \right] \tag{4.10}$$

where

$$\Delta_j = \Delta_j(\Sigma) = \left[\begin{pmatrix} \Sigma_{\beta\beta} & \Sigma_{\beta\alpha} \\ \Sigma_{\alpha\beta} & \Sigma_{\alpha\alpha} \end{pmatrix} + \left[\begin{pmatrix} I_{\beta_j\beta_j} & I_{\beta_j\alpha_j} \\ I_{\alpha_j\beta_j} & I_{\alpha_j\alpha_j} \end{pmatrix} \Big|_{\hat{\beta}_j, \hat{\alpha}_j} \right]^{-1} \right]^{-1}$$

$$\Omega_j = \Omega_j(\Sigma_{\beta\beta}) = \begin{bmatrix} (\Sigma_{\beta\beta} + \widehat{V(\hat{\beta}_j)})^{-1} & 0 \\ 0 & 0 \end{bmatrix}.$$

We then maximize (4.9) with respect of Σ to obtain $\hat{\Sigma}$ and use the expression from (4.10) to obtain the combined estimator, $\hat{\beta}(\hat{\Sigma})$ denoted by $\hat{\beta}_{IPD-AD-MA}$. The estimated variance of β is then given by

$$\widehat{V(\hat{\beta}_{IPD-AD-MA})} = \left[\sum_{j \in S_1} \Delta_j(\hat{\Sigma}) + \sum_{j \in S_2} \Omega_j(\hat{\Sigma}_{\beta\beta}) \right]^{-1}_{[1,1]}. \tag{4.11}$$

In the GLMM framework, for the IPD-MA estimator, we essentially combine the MLE estimates and the estimated variances for all the parameters in the model, including the parameter of interest and the nuisance parameter whereas for the AD-MA estimator, we integrate MLE estimates and the estimated variances for only the parameter of interest. Thus, there may be a loss of information for the AD-MA estimator when compared to the IPD-MA estimator depending on the correlation of the parameter of interest with the nuisance parameters.

In GLMM, we use the data to get the MLE estimates and integrate them to derive the IPD-MA estimator. In LMM, however, one can directly use the data for

the IPD studies instead of using the MLE estimates. While the general methodology of aggregation of GLMM studies can be applied to linear mixed models with the identity link, explicit computation of the likelihood and the estimators in the LMM allows more efficient strategies for data aggregation.

4.3.1 Selection in the Logistic Model

For a special case of generalized linear mixed model, we provide the variance expression for IPD selection. We consider a logistic model with similar setup as (1), where the treatment effect, β is fixed and the trial effect is random with with $\alpha_j \sim N(\alpha, \sigma_\alpha^2)$. Suppose further that, n_{jT} and n_{jC} be the number of persons allocated for treatment and control, respectively, for jth study where $n_{jT} + n_{jC} = n_j$. Let x_{ji} denote the treatment allocation for individual i in study j, with proportion of treatment π_j for $j = 1, \ldots, k$ and $i = 1, \ldots, n_j$. The response y_{ji} is binary with probability of success,

$$\begin{aligned} P(y_{ji} = 1 | x_{ji} = 1) = \quad p_{j1} &= \frac{\exp(\alpha_j + \beta)}{1 + \exp(\alpha_j + \beta)}, \\ P(y_{ji} = 1 | x_{ji} = 0) = \quad p_{j0} &= \frac{\exp(\alpha_j)}{1 + \exp(\alpha_j)}. \end{aligned} \tag{4.12}$$

Using the variance expression in (4.11), the variance of the combined estimator can be simplified to

$$V(\hat{\beta}_{IPD-AD-MA}) = \left[\sum_{j \in S_1} \frac{h_j}{c_j} (g_j - \tilde{g})^2 + \sum_{j=1}^{k} h_j^{-1} \right]^{-1} \tag{4.13}$$

where

$$h_j = a_j^{-1} + b_j^{-1} = V(\hat{\beta}_j),$$

$$g_j = \frac{1}{(a_j/b_j)^{-1} + 1},$$

$$c_j = \sigma_\alpha^2 h_j + a_j^{-1} b_j^{-1},$$

$$a_j = n_{jT}\ p_{j1}(1 - p_{j1}),$$

$$b_j = n_{jC}\ p_{j0}(1 - p_{j0}).$$

The problem of using (4.13) to select the "best" combination of IPD and AD is identical to the selection problem stated in (4.6). In order to use the sequential selection algorithm proposed in section 4.2.2, the unknowns g_j and h_j/c_j need to be estimated using AD. There are two possible solutions.

First, with the assumption of rare disease, $a_j/b_j = (n_{jT}/n_{jC}) * (\text{odds ratio}_j)$ where the odds ratio$_j$ can be estimated from $\hat{\beta}_j$. Therefore, with this assumption h_j/c_j and g_j can be computed as

$$\begin{aligned} g_j &= \frac{1}{(n_{jT}/n_{jC}) * \exp(\hat{\beta}_j) + 1} \\ \frac{h_j}{c_j} &= \frac{1/\widehat{V(\hat{\beta}_j)}}{\sigma_\alpha^2/\widehat{V(\hat{\beta}_j)}) + g_j(1 - g_j)} \end{aligned} \tag{4.14}$$

provided we have AD for all studies along with n_{jT} and n_{jC}. We can then compute the variance expression (4.13) to choose the IPD studies that 'maximize' the efficiency of the combined estimator using the sequential algorithm (1) .

The second solution requires knowledge of the number of cases and controls for each study in addition to treatment and control group totals. In fact, with this information we can reconstruct the individual 2×2 contingency tables and hence

estimate g_j and h_j/c_j even without the rare disease assumption. However, efficient selection of IPD studies is of no additional advantage for binary data in RCTs in case the 2×2 contingency tables are available since one can construct the IPD from the table.

4.3.2 Simulation results

We illustrate the advantages of combining estimates in GLMM through a limited simulation experiment using binary data and the logistic link. Suppose X_{ji} denotes the treatment allocation for individual i in study j, with proportion of treatment, π_j where $j = 1, \ldots, k$ and $i = 1, \ldots, n_j$. The response y_{ji} is binary with probability of success as in model (4.12) where both β_j and α_j are assumed to be random with $\beta_j|\beta \sim N(\beta, \Sigma_\beta)$ and $\alpha_j|\alpha \sim N(\alpha, \Sigma_\alpha)$, independently, where the true parameters are taken to be $\beta = 0.5, \alpha = 0.5, \sigma_\beta^2 = 0.5$ and $\sigma_\alpha^2 = 0.5$. We simulated the true π_j's from a uniform distribution over the interval $[0, 1]$. The results are shown in Table 4.6. We present the estimates, estimated bias and standard error for different choices of $n = n_j$ and k along with the relative efficiency with respect to the IPD-MA estimator. Similar to the linear model, we could have different combinations of IPD and AD for the IPD-AD-MA estimator, even when the percentage of IPD is fixed. However, since there can be many such combinations for $k = 50$, we only report the results for one randomly chosen combination for each possibility of no IPD, 20% IPD, 40% IPD, 60% IPD, 80% IPD and all IPD.

The mean square error (MSE) for the all AD estimator is 0.021 for $k = 50, n =$

100, 0.013 for $k = 50, n = 500$, 0.012 for $k = 100, n = 100$ and 0.006 for $k = 100, n = 500$. The MSE for the all IPD estimator is 0.019 when $k = 50, n = 100$ and 0.010 for $k = 100, n = 100$ and stays the same for other choices of k and n. While the bias doesn't change much when n is fixed and k is varying, the MSE decreases with increase in either sample size or number of studies or both. On the other hand, the relative efficiency approaches 1 as percentage of IPD studies is increased for each choice of k and n.

4.3.3 Real data analysis

We consider the dataset in *Yusuf et al. (1985)* which includes results from the long-term trials of oral beta blockers on its effectiveness for reducing mortality.[108] The data consists of 2×2 tables from 22 clinical trials. We illustrate the aggregation of (log) odds ratio under the logistic model (4.12) with independent random effects for the study effect and treatment effect. Although the full data is available, for the purpose of illustration we assume access to IPD for some trials and access to only meta-analysis results for the remaining trials. The cohort size for the smallest trial is 77 whereas the largest trial has a cohort of 3837 individuals. To show the significance of selection for the logistic model, we randomly sampled $n_j = 50$ individuals for each study j.

We assume we have access to the AD for each study,i.e. the MLE estimates of the log-odds ratio and its standard error along with the treatment and control group totals. We apply the sequential algorithm using the estimates in (4.14) where

Table 4.6: Model parameter estimates for different choices of n and k

Scenario	**(% IPD, % AD)**	**Estimate**	**Bias**	**Std. Error**	**Rel. eff.**
k=50, n=100	(0, 100)	0.43	-0.07	0.127	0.926
	(20, 80)	0.43	-0.07	0.127	0.925
	(40, 60)	0.431	-0.069	0.126	0.94
	(60, 40)	0.433	-0.067	0.125	0.961
	(80, 20)	0.436	-0.064	0.125	0.982
	(100, 0)	0.44	-0.06	0.126	1.000
k=50, n=500	(0, 100)	0.482	-0.018	0.113	0.975
	(20, 80)	0.482	-0.018	0.113	0.976
	(40, 60)	0.483	-0.017	0.113	0.979
	(60, 40)	0.483	-0.017	0.113	0.983
	(80, 20)	0.485	-0.015	0.113	0.986
	(100, 0)	0.488	-0.012	0.112	1.000
k=100, n=100	(0, 100)	0.438	-0.062	0.09	0.875
	(20, 80)	0.438	-0.062	0.09	0.873
	(40, 60)	0.439	-0.061	0.09	0.89
	(60, 40)	0.441	-0.059	0.089	0.911
	(80, 20)	0.444	-0.056	0.089	0.942
	(100, 0)	0.448	-0.052	0.088	1.000
k=100, n=500	(0, 100)	0.482	-0.018	0.074	0.973
	(20, 80)	0.482	-0.018	0.074	0.975
	(40, 60)	0.483	-0.017	0.074	0.979
	(60, 40)	0.484	-0.016	0.074	0.986
	(80, 20)	0.485	-0.015	0.074	0.994
	(100, 0)	0.488	-0.012	0.074	1.000

the σ_α^2 is estimated using a randomly selected pilot IPD studies. In practice, it can be estimated using IPD studies that are conveniently accessible since the σ_α^2 seems to have little or no effect in choosing the final set of selected IPD studies. The estimated variance for the IPD-MA estimator and AD-MA estimator are 0.033 and 0.044, respectively, which implies a 26% loss in efficiency. If we include the best 5 IPD studies chosen using the sequential algorithm, the estimated variance for the IPD-AD-MA estimator is 0.035, resulting in a mere 6% loss in efficiency relative to the IPD-MA estimator, which is a gain of 26% efficiency relative to the AD-MA estimator. However, the IPD-AD-MA estimator with the worst 5 IPD studies results in a estimated variance of 0.044, which is no any efficiency gain from the AD-MA estimator.

The pooled estimate (log) odds ratio with its standard error for the IPD-MA estimator, the best combination of IPD-AD studies with $k_1 = 2, 5$ and 8, respectively and the AD-MA estimator are reported in Table 4.7. The IPD trials in the table are ordered with respect to the g_j, analogous to the proportion of treatment, π_j in the linear mixed effects model. The standard error for the best combined estimator decreases as number of IPD studies is increased. In conclusion, a combined estimator with just 8 selected IPD studies out of the total 22 studies can achieve efficiency very close to the IPD-MA estimator.

Table 4.7: Pooled estimate and standard error of the (log) odds ratio for the IPD-MA estimator, the best IPD-AD-MA estimator with 2 IPDs, 5 IPDs and 8 IPDs, respectively and the AD-MA estimator

IPD Trials	$\hat{\boldsymbol{\beta}}$	**se**$(\hat{\boldsymbol{\beta}})$
None	0.001	0.211
1,22	0.006	0.206
1:3,21:22	0.017	0.204
1:5,20:22	0.001	0.200
1:22	0.057	0.199

4.4 Discussion

In this paper, we provide a method of combining information across independent studies for a generalized mixed effects model in a multivariate setup. We show that the combined estimator is efficient compared to the all IPD estimator asymptotically through simulation. For a much simpler linear mixed effects model with heterogeneous studies, we investigate the performance of the combined estimator for various distribution of treatment proportion across studies. We advocate a method for selection of AD studies over IPD studies to ensure a fully efficient combined estimator. When the number of studies is large, we propose an approximate sequential algorithm to select the best combination of IPD and AD studies.

We assume the same set of covariates for each study. In the future, we plan to expand our method to include disparate covariate information for multivariate

mixed effects models. Another possible direction would be to quantify the degree of unbalancedness of proportion of treatments among studies. That would help in investigating the impact of selecting AD studies for multiple treatment problems as well. It would be interesting to extend the sequential algorithm to a more general GLM framework, which requires a result on optimality of the AD estimators in terms of the design parameters. Once such a result has been established then an analogous algorithm could be devised by optimizing a measure of departure from the optimality condition.

We also look at an interesting application of our novel selection algorithm in the context of the recently popular 'Split and Conquer' approach to analysis of 'Big Data'. With the increasing need of resources to store and analyze large data, meta-analysis methods are becoming very popular. Due to the availability and accessibility of massive amount of data in several fields, a new research paradigm is focused towards divide and recombine (D&R) approach for big data analysis.[109–111] The main goal of our method is to aid in the analysis of Big Data like Genome Wide Association Studies, Electronic Health Records, large socio-economic survey data, using the split, analyze and aggregate approach in order to reduce computational and storage limitations but still provide estimates with efficiency close to that from the IPD analysis.

Chapter 5: Future Work

Applications of constrained inference lie beyond curve estimation and include areas such as financial economics and imaging data restoration. While we have looked into some aspects of the constrained statistical inference, there are still many ingredients that we need to explore. In this dissertation, we have focused on a particular scale mixture distribution and a two component model for the prior. Other scale mixture distributions worth exploring are Bayesian lasso with positive constraints, with hard thresholding properties for non-negative mean vectors. Discrete mixture models where the mixing kernel for the positive means could be chosen in a more flexible manner, belonging to flexible families on the non-negative orthant, e.g. product of gamma densities where heavy tailed priors are used for the hyperparamters are another interesting area of perusal. Another possible direction of research is to look at the connection between the definition of sparsity advocated in this dissertation to penalized and constrained regression. An important application of our work is testing for parameter for which the null set could be the boundary of the polyhedral cone versus an alternative where the parameter belong to the interior of the cone using Bayesian model comparison. In chapter 2, we raise the problem of choosing a linear transformation of the data so that the intrinsic volume or statis-

tical dimension of the intersection of cone and the transformed cone is maximized. Though computationally challenging, it is definitely worth investigating.

For chapter 3, a future plan is to develop similar design algorithms for a more general setup with multiple treatments and expand our method to include disparate covariate information for multivariate mixed effects models. Besides, a potential application of the meta-analysis selection algorithm worth exploring is in the context of the divide and recombine (DR) approach for Big Data analyses. Another possible direction, that would help us investigate the impact of selecting aggregate studies for multiple treatment problems, is be to quantify the degree of unbalancedness of proportion of treatments among studies. Since most meta-analysis is performed on log-linear and other generalized linear models, it would be interesting to extend the sequential algorithm to a more general framework, which requires a result on optimality of the aggregate estimators in terms of the design parameters. Once such a result has been established then an analogous extension is to devise an algorithm by optimizing a measure of departure from the optimality condition.

Bibliography

[1] James W, Stein C. Estimation with Quadratic Loss. In: The Regents of the University of California; 1961. ISSN: 0097-0433.

[2] Donoho DL, Johnstone IM. Ideal spatial adaptation by wavelet shrinkage. *Biometrika* 1994; 81(3): 425–455. Publisher: Oxford Academic.

[3] Tibshirani R. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* 1996; 58(1): 267–288.

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

[5] Mitchell TJ, Beauchamp JJ. Bayesian Variable Selection in Linear Regression. *Journal of the American Statistical Association* 1988; 83(404): 1023–1032.

[6] Berger JO, Strawderman WE. Choice of hierarchical priors: admissibility in estimation of normal means. *Ann. Statist.* 1996; 24(3): 931–951.

[7] Johnson VE, Rossell D. Bayesian Model Selection in High-Dimensional Settings. *J Am Stat Assoc* 2012; 107(498).

[8] Johnson VE, Rossell D. On the use of non-local prior densities in Bayesian hypothesis tests. *Journal of the Royal Statistical Society Series B* 2010; 72(2): 143–170. Publisher: Royal Statistical Society.

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

[10] Carvalho CM, Polson NG, Scott JG. The horseshoe estimator for sparse signals. *Biometrika* 2010; 97(2): 465–480.

[11] Johnstone IM, Silverman BW. Needles and straw in haystacks: Empirical Bayes estimates of possibly sparse sequences. *Ann. Statist.* 2004; 32(4): 1594–1649.

[12] Brown LD, Greenshtein E. Nonparametric empirical Bayes and compound decision approaches to estimation of a high-dimensional vector of normal means. *Ann. Statist.* 2009; 37(4): 1685–1704.

[13] Dharmadhikari S, Sudhakar D, Joag-Dev K. *Unimodality, Convexity, and Applications*. Elsevier Science . 1988.

[14] https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE1572. Accessed: 2022-01-25.

[15] Liao X, Meyer MC. coneproj: An R Package for the Primal or Dual Cone Projections with Routines for Constrained Regression. *Journal of Statistical Software* 2014; 61: 1–22.

[16] Meyer MC. A Simple New Algorithm for Quadratic Programming with Applications in Statistics. *Communications in Statistics - Simulation and Computation* 2013; 42(5): 1126–1139.

[17] Dykstra RL. An Algorithm for Restricted Least Squares Regression. *Journal of the American Statistical Association* 1983; 78(384): 837-842.

[18] Fraser DAS, Massam H. A Mixed Primal-Dual Bases Algorithm for Regression under Inequality Constraints. Application to Concave regression. *Scandinavian Journal of Statistics* 2016; 16: 65-74.

[19] Hildreth C. Point Estimates of Ordinates of Concave Functions. *Journal of the American Statistical Association* 1954; 49(267): 598-619.

[20] Karmarkar N. A new polynomial-time algorithm for linear programming. *Combinatorica* 1984; 4(4): 373–395.

[21] Hestenes MR. Multiplier and gradient methods. *J Optim Theory Appl* 1969; 4(5): 303–320.

[22] Fathi Y, Murty KG. A critical index algorithm for nearest point problems on simplicial cones. 1982. Accepted: 2006-09-11T19:32:29Z Publisher: Springer-Verlag; The Mathematical Programming Society, Inc.

[23] Powell MJD. Algorithms for nonlinear constraints that use lagrangian functions. *Mathematical Programming* 1978; 14(1): 224–248.

[24] Duguid p. Studies in linear and non-linear programming, by K. J. Arrow, L. Hurwicz and H. Uzawa. Stanford University Press, 1958. 229 pages.. *Canadian Mathematical Bulletin*; 3(3). doi: 10.1017/S0008439500025522

[25] Dimiccoli M. Fundamentals of cone regression. *Statistics Surveys* 2016; 10(none): 53 – 99.

[26] Silvapulle MJ, Sen PK. *Constrained Statistical Inference: Order, Inequality, and Shape Constraints*. John Wiley & Sons . 2011.

[27] Danaher MR, Roy A, Chen Z, Mumford SL, Schisterman EF. Minkowski–Weyl Priors for Models With Parameter Constraints: An Analysis of the BioCycle Study. *J Am Stat Assoc* 2012; 107(500): 1395–1409.

[28] Motzkin TS, Raiffa H, Thompson GL, Thrall RM. 3. The Double Description Method. In: Princeton University Press. 2016 (pp. 51–74)

[29] Meyer MC. An extension of the mixed primal–dual bases algorithm to the case of more constraints than dimensions. *Journal of Statistical Planning and Inference* 1999; 81(1): 13–31.

[30] Zolotykh NY. New modification of the double description method for constructing the skeleton of a polyhedral cone. *Computational Mathematics and Mathematical Physics* 2012; 52: 146–156.

[31] Fukuda K, Prodon A. Double description method revisited. In: Deza M, Euler R, Manoussakis I. , eds. *Combinatorics and Computer Science*Lecture Notes in Computer Science. Springer; 1996: 91–111.

[32] Chernikova NV. Algorithm for finding a general formula for the non-negative solutions of a system of linear inequalities. *USSR Computational Mathematics and Mathematical Physics* 1965; 5(2): 228–233.

[33] Fukuda K. cdd, cddplus and cddlib Homepage. https://people.inf.ethz.ch/fukudak/cdd_home/index.html; . Accessed: 2022-01-25.

[34] Boor dC. *A Practical Guide to Splines*. New York, Springer . 2001.

[35] Sasabuchi S. A test of a multivariate normal mean with composite hypotheses determined by linear inequalities. *Biometrika* 1980; 67(2): 429–439.

[36] Berger RL. Uniformly More Powerful Tests for Hypotheses concerning Linear Inequalities and Normal Means. *Journal of the American Statistical Association* 1989; 84: 192-199.

[37] Liu H, Berger RL. Uniformly More Powerful, One-sided Tests for Hypotheses about linear Inequalities. *Annals of Statistics* 1995; 23: 55-72.

[38] Berger RL, Hsu JC. Bioequivalence trials, intersection-union tests and equivalence confidence sets. *Statistical Science* 1996; 11: 283-319.

[39] Lee S, Lim J, Vannucci M, Petkova E, Preter M, Klein D. Order-Preserving Dimension Reduction Procedure for the Dominance of Two Mean Curves with Application to Tidal Volume Curves. *Biometrics* 2008; 64: 931–9.

[40] Bartholomew DJ. A Test Of Homogeneity for Ordered Alternatives. *Biometrika* 1959; 46(1/2): 36–48.

[41] Bartholomew DJ. A Test of Homogeneity of Means Under Restricted Alternatives. *Journal of the royal statistical society series b-methodological* 1961; 23: 239-272.

[42] Kudo A. A multivariate analogue of the one-sided test. *Biometrika* 1963; 50(3-4): 403–418.

[43] Perlman MD. One-Sided Testing Problems in Multivariate Analysis. *Ann. Math. Statist.* 1969; 40(2): 549–567.

[44] Shi NZ, Kudô A. The Most Stringent Somewhere Most Powerful One Sided Test of the Multivariate Normal Mean. *Memoirs of the Faculty of Science, Kyushu University. Series A, Mathematics* 1987; 41(1): 37–44.

[45] Tsai, M.-T. and Sen, P. K (1993). On the local optimality of optimal linear tests for restricted alternatives. Vol.3, No.1. .

[46] Tang DI, Gnecco C, Geller NL. An approximate likelihood ratio test for a normal mean vector with nonnegative components with application to clinical trials. *Biometrika* 1989; 76(3): 577–583.

[47] Follmann D. A Simple Multivariate Test for One-Sided Alternatives. *Journal of the American Statistical Association* 1996; 91(434): 854-861.

[48] Javanmard A, Lee JD. A flexible framework for hypothesis testing in high-dimensions. *arXiv preprint arXiv:1704.07971* 2017.

[49] Yu M, Gupta V, Kolar M. Constrained High Dimensional Statistical Inference. *arXiv:1911.07319* 2019.

[50] Wei Y, Wainwright MJ, Guntuboyina A. The geometry of hypothesis testing over convex cones: Generalized likelihood tests and minimax radii. *arXiv:1703.06810 [cs, math, stat]* 2018.

[51] Hsieh HK. On Asymptotic Optimality of Likelihood Ratio Tests for Multivariate Normal Distributions. *The Annals of Statistics* 1979; 7(3): 592 – 598.

[52] Moreau JJ. Décomposition orthogonale d'un espace hilbertien selon deux cônes mutuellement polaires. *Comptes rendus hebdomadaires des séances de l'Académie des sciences* 1962; 255: 238-240.

[53] Tiago De Oliveira J. The asymptotical independence of the sample mean and the extremes.. *Rev. Fac. Cienc. Univ. Lisboa A* 1961; 8: 299–310.

[54] Lo GS. A note on the asymptotic normality of sums of extreme values. *Journal of Statistical Planning and Inference* 2000; 22: 127–136.

[55] David L. Donoho JCH, Stern AS. Maximum Entropy and the Nearly Black Object. *Journal of the Royal Statistical Society. Series B* 1992; 54: 41–81.

[56] Johnstone IM. On Minimax Estimation of a Sparse Normal Mean Vector. *The Annals of Statistics* 1994; 22: 271–289.

[57] Agarwala PJA. Sparse Priors on Closed Convex Polyhedral Cones. *arxiv??* 2022.

[58] Cai TT, Liu W, Xia Y. Two-sample test of high dimensional means under dependence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2014; 76(2): 349–372.

[59] Hall P, Jin J. Innovated higher criticism for detecting sparse signals in correlated noise. *Ann. Statist* 2010; 38: 1686—1732.

[60] Friedman J, Hastie T, Tibshirani R. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* 2007; 9(3): 432–441.

[61] Witten DM, Friedman JH, Simon N. New Insights and Faster Computations for the Graphical Lasso. *Journal of Computational and Graphical Statistics* 2011; 20(4): 892-900.

[62] Meinshausen N, Bühlmann P. High-dimensional graphs and variable selection with the Lasso. *The Annals of Statistics* 2006; 34(3): 1436 – 1462.

[63] Amelunxen D, Lotz M. Intrinsic Volumes of Polyhedral Cones: A Combinatorial Perspective. *Discrete & Computational Geometry* 2017; 58(2): 371–409.

[64] McCoy MB, Tropp JA. From Steiner Formulas for Cones to Concentration of Intrinsic Volumes. *Discrete & Computational Geometry* 2014; 51: 926-963.

[65] Neyman J, Pearson ES. On the Problem of the Most Efficient Tests of Statistical Hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* 1933; 231: 289–337.

[66] Gumbel EJ. Les valeurs extrêmes des distributions statistiques. *Annales de l'institut Henri Poincaré* 1935; 5(2): 115–158.

[67] Gumbel EJ. The Return Period of Flood Flows. *Ann. Math. Statist.* 1941; 12(2): 163–190.

[68] Csorgo S, Haeusler E, Mason DM. The Asymptotic Distribution of Extreme Sums. *Ann. Probab.* 1991; 19(2): 783–811.

[69] Chalmers I. The Cochrane collaboration: preparing, maintaining, and disseminating systematic reviews of the effects of health care. *Ann N Y Acad Sci* 1993; 703: 156–163; discussion 163–165.

[70] Ahmed I, Sutton AJ, Riley RD. Assessment of publication bias, selection bias, and unavailable data in meta-analyses using individual participant data: a database survey. *BMJ* 2012; 344.

[71] Jones AP, Riley RD, Williamson PR, Whitehead A. Meta-analysis of individual patient data versus aggregate data from longitudinal clinical trials. *Clinical Trials* 2009; 6(1): 16–27.

[72] Cooper H, Patall EA. The relative benefits of meta-analysis conducted with individual participant data versus aggregated data.. *Psychological Methods* 2009; 14(2): 165.

[73] Riley RD. Commentary: like it and lump it? Meta-analysis using individual participant data. *International Journal of Epidemiology* 2010; 39(5): 1359–1361.

[74] DerSimonian R, Laird N. Meta-analysis in Clinical Trials. *Controlled clinical trials* 1986; 7(3): 177–188.

[75] DerSimonian R, Laird N. Meta-analysis in clinical trials revisited. *Contemporary Clinical Trials* 2015; 45: 139–145.

[76] Viechtbauer W. Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software* 2010; 36(3): 1–48.

[77] Veroniki AA, Jackson D, Viechtbauer W, et al. Methods to estimate the between-study variance and its uncertainty in meta-analysis. *Research Synthesis Methods* 2016; 7(1): 55–79.

[78] Panityakul T, Bumrungsup C, Knapp G. On estimating residual heterogeneity in random-effects meta-regression: a comparative study. *Journal of Statistical Theory and Applications* 2013; 12(3): 253–265.

[79] Sidik K, Jonkman JN. A comparison of heterogeneity variance estimators in combining results of studies. *Statistics in Medicine* 2007; 26(9): 1964–1981.

[80] Viechtbauer W. Bias and efficiency of meta-analytic variance estimators in the random-effects model. *Journal of Educational and Behavioral Statistics* 2005; 30(3): 261–293.

[81] Hedges LV, Olkin I. *Statistical methods for meta-analysis*. Academic Press . 2014.

[82] Hartung J, Knapp G, Sinha BK. *Statistical meta-analysis with applications*. 738. John Wiley & Sons . 2011.

[83] McCulloch CE, Neuhaus JM. Generalized linear mixed models. *Encyclopedia of Biostatistics* 2005; 4.

[84] Van Houwelingen HC, Arends LR, Stijnen T. Advanced methods in meta-analysis: multivariate approach and meta-regression. *Statistics in Medicine* 2002; 21(4): 589–624.

[85] Ritz J, Demidenko E, Spiegelman D. Multivariate meta-analysis for data consortia, individual patient meta-analysis, and pooling projects. *Journal of Statistical Planning and Inference* 2008; 138(7): 1919–1933.

[86] Olkin I, Sampson A. Comparison of meta-analysis versus analysis of variance of individual patient data. *Biometrics* 1998: 317–322.

[87] Mathew T, Nordström K. On the equivalence of meta-analysis using literature and using individual patient data. *Biometrics* 1999; 55(4): 1221–1223.

[88] Whitehead A. *Meta-analysis of controlled clinical trials*. 7. John Wiley & Sons . 2002.

[89] Mathew T, Nordström K. Comparison of one-step and two-step meta-analysis models using individual patient data. *Biometrical Journal: Journal of Mathematical Methods in Biosciences* 2010; 52(2): 271–287.

[90] Lin DY, Zeng D. On the relative efficiency of using summary statistics versus individual-level data in meta-analysis. *Biometrika* 2010; 97(2): 321–332.

[91] Liu D, Liu RY, Xie M. Multivariate meta-analysis of heterogeneous studies using only summary statistics: efficiency and robustness. *Journal of the American Statistical Association* 2015; 110(509): 326–340.

[92] Thomas D, Platt R, Benedetti A. A comparison of analytic approaches for individual patient data meta-analyses with binary outcomes. *BMC Medical Research Methodology* 2017; 17(1): 1–12.

[93] Collaboration FS. Systematically missing confounders in individual participant data meta-analysis of observational cohort studies. *Statistics in Medicine* 2009; 28(8): 1218–1237.

[94] Kundu P, Tang R, Chatterjee N. Generalized meta-analysis for multiple regression models across studies with disparate covariate information. *Biometrika* 2019; 106(3): 567–585.

[95] Riley RD, Dodd SR, Craig JV, Thompson JR, Williamson PR. Meta-analysis of diagnostic test studies using individual patient data and aggregate data. *Statistics in Medicine* 2008; 27(29): 6111–6136.

[96] Riley RD, Lambert PC, Staessen JA, et al. Meta-analysis of continuous outcomes combining individual patient data and aggregate data. *Statistics in Medicine* 2008; 27(11): 1870–1893.

[97] Riley RD, Simmonds MC, Look MP. Evidence synthesis combining individual patient data and aggregate data: a systematic review identified current practice and possible methods. *Journal of Clinical Epidemiology* 2007; 60(5): 431–e1.

[98] Idris NRN, Abdullah MH. A study on the effects of different levels of data on the overall meta-analysis estimates. *Far East Journal of Mathematical Sciences* 2015; 96(1): 73.

[99] Yamaguchi Y, Sakamoto W, Goto M, et al. Meta-analysis of a continuous outcome combining individual patient data and aggregate data: a method based on simulated individual patient data. *Research Synthesis Methods* 2014; 5(4): 322–351.

[100] Guyot P, Ades A, Ouwens MJ, Welton NJ. Enhanced secondary analysis of survival data: reconstructing the data from published Kaplan-Meier survival curves. *BMC Medical Research Methodology* 2012; 12(1): 1–13.

[101] Riley RD, Steyerberg EW. Meta-analysis of a binary outcome using individual participant data and aggregate data. *Research Synthesis Methods* 2010; 1(1): 2–19.

[102] Donegan S, Williamson P, D'Alessandro U, Garner P, Smith CT. Combining individual patient data and aggregate data in mixed treatment comparison meta-analysis: individual patient data may be beneficial if only for a subset of trials. *Statistics in Medicine* 2013; 32(6): 914–930.

[103] Saramago P, Sutton AJ, Cooper NJ, Manca A. Mixed treatment comparisons using aggregate and individual participant level data. *Statistics in Medicine* 2012; 31(28): 3516–3536.

[104] Ravva P, Karlsson MO, French JL. A linearization approach for the model-based analysis of combined aggregate and individual patient data. *Statistics in Medicine* 2014; 33(9): 1460–1476.

[105] Sutton AJ, Kendrick D, Coupland CA. Meta-analysis of individual-and aggregate-level data. *Statistics in Medicine* 2008; 27(5): 651–669.

[106] Jackson C, Best N, Richardson S. Improving ecological inference using individual-level data. *Statistics in Medicine* 2006; 25(12): 2136–2159.

[107] Varin C, Reid N, Firth D. An overview of composite likelihood methods. *Statistica Sinica* 2011: 5–42.

[108] Yusuf S, Peto R, Lewis J, Collins R, Sleight P. Beta blockade during and after myocardial infarction: an overview of the randomized trials. *Progress in Cardiovascular Diseases* 1985; 27(5): 335–371.

[109] Lee JY, Brown JJ, Ryan LM. Sufficiency revisited: Rethinking statistical algorithms in the big data era. *The American Statistician* 2017; 71(3): 202–208.

[110] Chen X, Xie Mg. A split-and-conquer approach for analysis of extraordinarily large data. *Statistica Sinica* 2014: 1655–1684.

[111] Cheung MWL, Jak S. Analyzing big data in psychology: A split/analyze/meta-analyze approach. *Frontiers in Psychology* 2016; 7: 738.