Hypothesis Testing for High Dimensional Constrained Normal Means

by

Neha Agarwala

A dissertation submitted to University of Maryland, Baltimore County
in conformity with the requirements for the degree of
Doctor of Philosophy

Baltimore, Maryland
January, 2020

Committee Members:

- Dr. Anindya Roy (Advisor), Department of Mathematics and Statistics
- Dr. Junyong Park (Co-advisor), Department of Mathematics and Statistics
- Dr. Thomas Mathew, Department of Mathematics and Statistics
- Dr. Yaakov Malinovsky, Department of Mathematics and Statistics
- Dr. Jinglai Shen, Department of Mathematics and Statistics
- Dr. Emanuel Ben-David, U.S. Census Bureau

Dear Professors:

I would like to extend my sincere thanks for serving on my prelims oral committee and all the discussions we had in our meetings. It is my pleasure to discuss my research progress till date, and have your valuable comments on my research proposal. I am hereby enclosing my research proposal titled "Hypothesis Testing for High Dimensional Constrained Normal Means" for your review. As a gentle reminder, the exam is scheduled on January 24, 2020 at Maths and Psycology Building in the University of Maryland, Baltimore County commencing from 11 AM.

Warm Regards
Neha Agarwala

Contents

Chapter 1

Introduction

High dimensional testing problem is generating considerable interest due to the availability and accessibility of massive amount of data in several fields. Modern statistical problems, however, involve natural constraints on model parameters. For such tests, it is not fitting to apply standard tests designed for unrestricted alternatives. Traditional statistical theory have mostly focused on methods developed for large samples and small number of features. Modern data sets have large sample size but even larger number of features. Given a n dimensional normal distribution with covariance I, 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. In the first chapter, our goal is to develop a computationally efficient test for the entire spectrum of alternatives.

For the next chapter, we propose a method for combining treatment effects across trials where the response can be binary or continuous. For some studies with meta-analysis (MA) data, the associated Individual Patient Data (IPD) maybe available, albeit at some extra effort or cost to the analyst. We consider the case when treatment effects are fixed and common across studies and evaluate the wisdom of choosing MA when IPD is available by studying the relative efficiency of analyzing all IPD studies versus combining various percentages of MA and IPD studies. For many different models design constraints under which the MA estimators are the IPD estimators, and hence fully efficient, are known. For such models we advocate a selection procedure that chooses MA studies over IPD studies in a manner that force least departure from design constraints and hence ensures a fully efficient combined MA and IPD estimator.

Chapter 2

Hypothesis testing

2.1 Introduction

High dimensional testing problem is generating considerable interest due to the availability and accessibility of massive amount of data in several fields. Modern statistical problems, however, involve natural constraints on model parameters. For such tests, it is not fitting to apply standard tests designed for unrestricted alternatives. Such tests do not take into account the specified direction of deviation from the null and hence may have unreasonably low power.

In general, the prior information on the parameters may provide very different 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 alternative, it seems rational to reject H_0 when $\bar{x}$ is relatively large.

Constrained hypothesis testing has gained popularity for its wide range of scientific applications in emerging fields like genetic studies, social network analysis, drug testing, etc. Some common examples in this setting are testing non-negativity or monotonicity in treatment effects, cone-constrained testing in linear regression, etc. Let $\mathbf{X} \sim \mathbf{N}_n(\boldsymbol{\mu}, \Sigma)$ where Σ is known. The general problem of testing in convex cones can be formulated as

$$\mathcal{H}_0 : \boldsymbol{\mu} \in \mathcal{M} \quad vs. \quad \mathcal{H}_1 : \boldsymbol{\mu} \in \mathcal{C} \backslash \mathcal{M}. \tag{2.1}$$

where $\mathcal{M}$ is a linear space $\subseteq \mathcal{C}$, a closed convex cone. In this paper, we focus on the special case when $\mathcal{M} = \{\mathbf{0}\}$.

The literature on multivariate one-sided tests dates back to Kudo (1963) who studied the likelihood ratio test for positive orthant under the assumption of normality [11]. Perlman (1969) extended his work to the case of unknown covariance matrix, giving upper and lower bounds on the null distribution of the test statistic [19]. 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.

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 [20]. 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. However, LRT may not be most powerful even for a specific part of the positive orthant. Also, the LRT may not be a Bayes test for such constrained alternatives and hence, it is not generally a most stringent test (even asymptotically). Tsai and Sen (1993) showed that OLT is uniformly locally more powerful than the LRT [1].

Traditional statistical theory have mostly focused on methods developed for large samples and small number of features. On the other hand, modern scientific world is fast moving towards the regime of high dimensional data. 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 [22]. 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 [3]. Javanmard and Lee (2019) have developed a framework for testing very general hypotheses regarding the model parameters assuming sparsity and approximate sparsity structure [10]. 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 [24]. Wei *et al.* (2018) considered hypothesis testing on whether the parameters lie on some closed convex cone [23]. They provided a sharp characterization of the 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.

We consider the classical one-sided normal mean testing problem. Suppose $\mathbf{X} \sim \mathbf{N}_n(\boldsymbol{\mu}, \mathbf{I}_n)$. The testing problem in the canonical form is

$$\mathcal{H}_0 : \boldsymbol{\mu} = \mathbf{0} \quad vs. \quad \mathcal{H}_1 : \boldsymbol{\mu} \geq \mathbf{0}. \tag{2.2}$$

We adopt the following convention: "$\boldsymbol{\mu} \geq \mathbf{0}$" is to be interpreted as "all coordinates non-negative with at least one strictly positive". In other words, we want to develop an omnibus test to detect if $\boldsymbol{\mu} \in \mathcal{K} \backslash \{\mathbf{0}\}$ where $\mathcal{K}$ is non-negative orthant. We mention a few application areas where situations like these might arise:

– Consider an experiment to establish that a new treatment for controlling fever is better than a placebo. For each patient, let X_1 and X_2 denote the reduction in body temperature and body ache due to treatment. Suppose it is known that the new treatment is expected to be at least as good as the placebo. Thus, it is assumed *a prioiri* that $\boldsymbol{\mu} \geq \mathbf{0}$. The inference problem for concluding that the new treatment is better than the placebo can be phrased as:

$$\mathcal{H}_0 : \mu_1 = 0, \quad \mu_2 = 0 \quad vs. \quad \mathcal{H}_1 : \mu_1 \geq 0,\ \mu_2 \geq 0,\ ||\boldsymbol{\mu}||^2 > 0. \tag{2.3}$$

– The second example is motivated from Lee *et al.* (2008) [12]. 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.

2.1.1 Generalized Likelihood Ratio Test (GLRT)

The likelihood ratio test (LRT) or sometimes the literature call it GLRT is the ratio of the likelihood at the null value to the likelihood at the MLE. One could think of it

as the Neyman Pearson statistic where the alternative value is being estimated by the MLE. The GLRT is expected to have better overall performance.

The GLRT statistic for the above testing problem (2.2) when $\mathcal{K}$ is a polyhedral cone is $T = ||P_{\mathcal{K}}\mathbf{X}||^2$ where $P_{\mathcal{K}}\mathbf{X}$ is the least-squares projection of $\mathbf{X}$ onto $\mathcal{K}$ and the MLE of $\boldsymbol{\mu}$ under the alternative. 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 is not exactly a chi-squared distribution, but instead has the form

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

where χ^2_m is a standard chi-squared variable with m degrees of freedom. The quantity $P(m,k)$ is the probability that the maximum-likelihood estimate of $\boldsymbol{\mu}$ under the alternative hypothesis belongs to one of the m-dimensional faces of $\mathcal{K}$ under the assumption that the null hypothesis is true. This distribution is called a *chi-bar-squared* distribution. We now explain the GLRT when $n = 2$.

Suppose $\mathbf{X} \sim \mathbf{N}_2(\boldsymbol{\mu}, \mathbf{I}_2)$ where $\boldsymbol{\mu} = (\mu_1, \mu_2)^T$. We want to find the MLE based on one observation on $\mathbf{X}$ and subject to $\boldsymbol{\mu} \in \mathcal{K}$ where $\mathcal{K}$ is $\{(\mu_1, \mu_2)^T : \mu_1 \geq 0, \mu_2 \geq 0\}$.

Hence, the likelihood is

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

$-2l(\boldsymbol{\mu})$ is equal to the squared distance between $\mathbf{X}$ and $\boldsymbol{\mu}$. Thus, the MLE is the projection of $\mathbf{X}$ onto $\mathcal{K}$ (figure 2.1) 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 $P_{\mathcal{K}}\mathbf{X}$ the least-squares projection of X onto $\mathcal{K}$.

Thus the LRT for testing $\mathcal{H}_0 : \boldsymbol{\mu} = \mathbf{0}$ *vs.* $\mathcal{H}_1 : \boldsymbol{\mu} \geq \mathbf{0}$ is simply

$$LRT = \mathbf{X}^2 - (\mathbf{X} - \hat{\boldsymbol{\mu}})^2 = \hat{\boldsymbol{\mu}}^2. \tag{2.4}$$

This implies

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

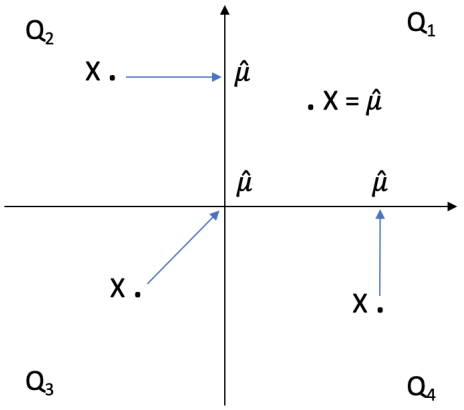

Figure 2.1: Projection of $\mathbf{X}$ onto $\mathcal{K}$

The $LRT = ||P_{\mathcal{K}}\mathbf{X}||^2$ 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 $\mathcal{H}_0$,

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

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

$$\begin{aligned} P(LRT \leq c) &= \sum P(LRT \leq c \;\&\; \mathbf{X} \in \mathbf{Q}_i) \\ &= \sum P(LRT \leq c \mid \mathbf{X} \in \mathbf{Q}_i) P(\mathbf{X} \in \mathbf{Q}_i) \\ &= \frac{1}{4} + \frac{1}{2} P(\chi_1^2 \leq c) + \frac{1}{4} P(\chi_2^2 \leq c). \end{aligned} \tag{2.5}$$

where $\mathbf{Q}_1, \mathbf{Q}_2, \mathbf{Q}_3, \mathbf{Q}_4$ denote the four quadrants in the two-dimensional plane. Un-

der $\mathcal{H}_0$, LRT is in fact a mixture of χ^2 distributions with weights being the probability that $\mathbf{X}$ falls in the corresponding cone $\mathbf{Q}_i$.

Power of the Chi-bar test statistic is

$$\begin{aligned} P_{\mathcal{H}_1}(LRT > c) &= \sum P_{\mathcal{H}_1}(LRT > c \;\&\; \mathbf{X} \in \mathbf{Q}_i) \\ &= \sum P_{\mathcal{H}_1}(LRT > c \mid \mathbf{X} \in \mathbf{Q}_i) P(\mathbf{X} \in \mathbf{Q}_i). \end{aligned} \tag{2.6}$$

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.

We say a test is uniformly most powerful (UMP) when its power function is uniformly greater than that of any other α level test for all $\boldsymbol{\mu} \in \mathcal{H}_1$. Karlin and Rubin gives the necessary conditions for a LRT to be UMP. These conditions are fullfiled for many one-sided (univariate) tests. UMP test does not exist for the above test of hypothesis, i.e., the chi-bar test statistic is not optimal in the most powerful sense for the alternative $\mathcal{H}_1$. However, UMP is a very stringent requirement. We seek to find alternative tests that are optimal in some sense and also easy to compute.

2.2 Theory and Methods

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. We propose investigating a slew of tests for their performance in the one sided testing problem. The first set of tests in subsection 2.2.1 are all hybrid tests comprising linear tests with tests consisting of top order statistics. In subsection 2.2.4 we look at modification of the LRT by replacing the MLE with estimators that have better risk properties compared to MLE, particularly in the sparse cases.

2.2.1 Test Statistic 1

Let us define

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

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

We know M_n and S_n are asymptotically independent and their distributions are known in one normal mean problem. Therefore for $\alpha \in [0,1]$, we consider the test statistic

$$T(\alpha) = \alpha \frac{M_n - \beta}{v} + (1-\alpha)S_n = \alpha G_n + (1-\alpha)S_n. \tag{2.7}$$

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 $\frac{\pi^2}{6}$ [6] [7]. The idea is that M_n takes care of the sparse signals whereas S_n focuses on the dense signals. Since α is unknown, we define a test statistic

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

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

Selection of α through grid points

In practice, we may take some grid points of α such that $0 = \alpha_0 < \alpha_1 < \cdots < \alpha_m = 1$ and compute use the test statistic $\mathcal{P} = \min_{1 \le k \le m} p_{\alpha_k}$. Under H_0, S_n has the standard normal limiting distribution and $P\big[M_n \le z\big] \to G(z) : e^{-e^{-z}}$ the so-called 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 $P_{H_o}(\mathcal{P} < p) \le 0.05$. Therefore,

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

where a_k and b_k are intercept and slope for each line $y = a_k + b_k x$, where $a_k = \beta + \nu q_k(1-p)/\alpha_k$ and $b_k = -\nu(1-\alpha_k)/\alpha_k$. We use A_k & B_k to denote the intersection point for each line to be the minimum line and $B_m = q_1(1-p)$.

To illustrate how equation (7) looks for $m = 6$, consider a choice of α_k's, $\alpha_1 = 0 < \alpha_2 = 0.2 < \alpha_3 = 0.4 < \alpha_4 = 0.6 < \alpha_5 = 0.8 < \alpha_6 = 1$. Let us call the 4 intersection points from 5 lines as $x_i, i = 1, 2, 3, 4$. Refer to the plot of the lines below.

Thus equation (7) can be re-written as:

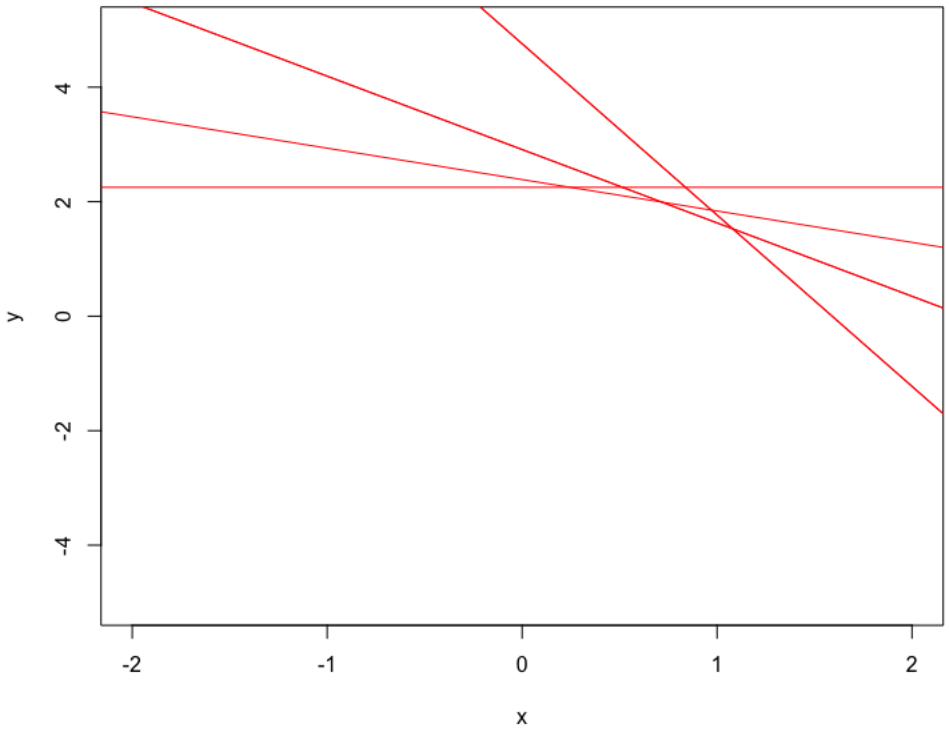

Figure 2.2: Lines from the example

$$P_{H_o}(\mathcal{P} < p) = 1 - \left(\int_{-\infty}^{x_1} \phi(y)\ e^{-e^{-a_6}}\ dy + \int_{x_1}^{x_2} \phi(y)\ e^{-e^{-(a_5+b_5 y)}}\ dy + \right.$$
$$\int_{x_2}^{x_3} \phi(y)\ e^{-e^{-(a_4+b_4 y)}}\ dy + \int_{x_3}^{x_4} \phi(y)\ e^{-e^{-(a_3+b_3 y)}}\ dy + \quad (2.8)$$
$$\left. \int_{x_4}^{q_1(1-p)} \phi(y)\ e^{-e^{-(a_2+b_2 y)}}\ dy \right).$$

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,

$$\begin{aligned} 1-p &= P_{H_o}(T_{\alpha_k} \leq t) \\ &= \begin{cases} F_{S_n}(t), & \text{for } \alpha_k = 0, \\ F_{M_n}(t), & \text{for } \alpha_k = 1, \\ \int F_{M_n}\big(\frac{t-(1-\alpha_k)s}{\alpha_k}\big)\ f_{S_n}(s)\ ds & \text{for } 0 < \alpha_k < 1. \end{cases} \end{aligned} \tag{2.9}$$

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

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

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

Selection of α corresponding to the strongest signal

Another way of determining α adaptively is finding the $\alpha^* \in [0, 1]$ that maximizes the test statistic

$$T^*(\alpha) = \frac{\alpha G_n\ +\ (1-\alpha)S_n}{\sqrt{\alpha^2 + (1-\alpha)^2}}.$$

In other words, we want to use the test statistic

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

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 $t \in R$ such that $P_{H_o}(T > t) \leq 0.05$.

$$\begin{aligned} 0.05 &= P_{H_o}(T > t) \\ &= P_{H_o}\Big(\sqrt{S_n^2 + G_n^2} > t, G_n > 0, S_n > 0\Big) + \\ &\quad P_{H_o}\Big(-\sqrt{S_n^2 + G_n^2} > t, G_n < 0, S_n < 0\Big) + \\ &\quad P_{H_o}\Big(G_n > t, G_n > 0, S_n < 0\Big) + \\ &\quad P_{H_o}\Big(S_n > t, G_n < 0, S_n > 0\Big) \\ &= p_1 + p_2 + p_3 + p_4. \end{aligned} \tag{2.10}$$

where

$$\begin{aligned} p_1 &= \int_0^{|t|} P_{H_0}\big[M_n > \beta + v\sqrt{t^2 - s^2}\big]\phi_{S_n}(s)ds, \\ p_2 &= \int_{|t|}^0 P_{H_0}\big[\beta - v\sqrt{t^2 - s^2} \leq M_n \leq \beta\big]\phi_{S_n}(s)ds, \\ p_3 &= P_{H_0}[S_n < 0]\ P_{H_0}[M_n > \max(\beta, \beta + tv)], \\ p_4 &= P_{H_0}[S_n > \max(t, 0)]\ P_{H_0}[M_n < \beta]. \end{aligned} \tag{2.11}$$

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

2.2.2 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

$$\begin{aligned} T(\gamma) &= \sum_{i=1}^{n} x_i I(x_i > \text{sgn}(\gamma)\sqrt{2|\gamma| \log n}) \\ L(\gamma) &= \frac{T(\gamma) - E_{H_0}(T(\gamma))}{\sqrt{V_{H_0}(T(\gamma))}}. \end{aligned} \tag{2.12}$$

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

$$\begin{aligned} E_{H_0}(T(\gamma)) &= n\ \phi(\gamma^*) \\ V_{H_0}(T(\gamma)) &= n\Big(\gamma^*\ \phi(\gamma^*) + \Phi(-\gamma^*) - \phi^2(\gamma^*)\Big). \end{aligned} \tag{2.13}$$

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

$$\mathcal{P} = \min_{\gamma \in \Gamma} p_\gamma. \tag{2.14}$$

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

Under H_0,

$$Pr\left[\frac{X_{(n)} - a_n}{b_n} \leq z\right] \to G(z) : e^{-e^{-z}}. \tag{2.15}$$

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,

$$
\begin{aligned}
0.05 &= P_{H_o}(\mathcal{P} < p) \\
&= P_{H_o}\Big(\min_{\gamma\in\Gamma} p_\gamma < p\Big) \\
&= 1 - P_{H_o}\Big(p_\gamma \geq p \;\forall\; \gamma \in \Gamma\Big) \\
&= 1 - P_{H_o}\Big(L(\gamma) \leq q_\gamma(1-p) \;\forall\; \gamma \in \Gamma\Big) \\
&= 1 - P_{H_o}\Big(L(\gamma_k) \leq q_{\gamma_k}(1-p) \;\forall\; \gamma = 1(1)m\Big)\; P_{H_o}\Big(L(\infty) \leq q_\infty(1-p)\Big).
\end{aligned} \tag{2.16}
$$

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

$$
Power = P_{H_1}\Big(\min_{\gamma\in\Gamma} p_\gamma < p^*\Big). \tag{2.17}
$$

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 $r \in \mathcal{N} = \{1, ..., n\}$

$$\begin{aligned} M_n(r) &= \sum_{i=1}^{r} X_{(n-i+1)} \\ G_n(r) &= \frac{M_n(r) - E_{H_0}(M_n(r))}{\sqrt{\widehat{Var_{H_0}(M_n(r))}}}. \end{aligned} \tag{2.18}$$

We define our test statistic as

$$T = \max_{\gamma \in \Gamma r \in \mathcal{N}} \{L(\gamma), G_n(r)\}. \tag{2.19}$$

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 [4]. 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.

2.2.3 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 wanted to investigate statistics with powers in the range from linear to quadratic.

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

$$\begin{aligned} T(\zeta) &= \sum_{i=1}^{n} x_i^{\zeta} I(x_i > 0) \\ L(\zeta) &= \frac{T(\zeta) - E_{H_0}(T(\zeta))}{\sqrt{\widehat{Var_{H_0}(T(\zeta))}}}. \end{aligned}$$

where

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

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

It is easy to see that $L = (L(\zeta_1), L(\zeta_2), \ldots, L(\zeta_m))$ 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,

$$\begin{aligned} 0.05 &= P_{H_o}(\mathcal{P} < p) \\ &= 1 - P_{H_o}\Big(L(\zeta_k) \leq q_{\zeta_k}(1-p)\ \forall\ \zeta = 1(1)m\Big)\ P_{H_o}\Big(L(\infty) \leq q_\infty(1-p)\Big). \end{aligned} \tag{2.20}$$

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 (2.15).

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 2.2.2.

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 $r \in \mathcal{N} = \{1, ..., n\}$

$$M_n(r) = \sum_{i=1}^{r} X_{(n-i+1)}$$
$$G_n(r) = \frac{M_n(r) - E_{H_0}(M_n(r))}{\sqrt{\widehat{Var_{H_0}(M_n(r))}}}. \tag{2.21}$$

We define our test statistic as

$$T = \max_{\zeta \in \zeta r \in \mathcal{N}} \{L(\zeta), G_n(r)\}. \tag{2.22}$$

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.

2.2.4 Modification of the GLRT

In section 2.1.1, we noticed the likelihood ratio test statistic for testing point null against non-negative orthant is $\|\mathbf{X}\|^2 - (\|\mathbf{X} - \hat{\boldsymbol{\mu}}\|)^2$ where $\hat{\boldsymbol{\mu}}$ is the non-negative orthant constrained MLE. Ideally we would like to provide an improved estimator under the constraint to have better performance than the Chibar.

It is known that the MLE suffers from several problems in the high dimensional case. Particularly, it becomes inadmissible in higher dimensions and since Stein's famous result [5] researchers have advocated different types of shrinkage estimation for improving the risk of the MLE in normal mean estimation problem. The loss of efficiency for MLE is even more stark in higher dimensional cases where very sparse parameter configurations are possible. For multivariate case, in normal one-sided mean testing problem, in higher dimension, the LRT is no longer optimum. The finite sample power can be uniformly improved [14].

Therefore, a natural question would be whether the performance of likelihood ratio omnibus test can be improved if one replaces MLE with a more desirable estimator in the high dimensional case.

Thus we propose investigating power characteristics of tests of the form $T_{\tilde{\boldsymbol{\mu}}} = \|\mathbf{X}\|^2 - (\|\mathbf{X} - \tilde{\boldsymbol{\mu}}\|)^2$, where $\tilde{\boldsymbol{\mu}}$ are suitably chosen estimators. Given that the expected gain in power is mostly in the sparse configuration, one would like to have an estimator that directly account for sparsity and shrinks the estimator accordingly. In this context we think of using a Bayes estimator or related estimator that takes a priori knowledge of sparsity into account.

Specifically, suppose the coordinates μ_i are coming from a mixture $\pi\delta_o + (1-\pi)g_1(\mu_i)$. Then we could use an estimator adapted to that prior configuration.

We therefore investigate Empirical Bayes (EB) estimation to provide one such estimator for $\boldsymbol{\mu}$.

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

$$p(y|\mu) = e^{\mu y - \Phi(\mu)} h_o(y), \quad \Phi(\mu) = \frac{\mu^2}{2}, h_o(y) = e^{-\frac{y^2}{2}}$$

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

Therefore, the marginal of y is

$$\begin{aligned} p_g(y) &= \int p(y|\mu)g(\mu)d\mu \\ &= \int \phi(y-\mu)g(\mu)d\mu \\ &= \pi\phi(y) + (1-\pi)\int \phi(y-\mu)g_1(\mu)d\mu \\ p_g'(y) &= -\pi y\phi(y) - (1-\pi)\int (y-\mu)\phi(y-\mu)g_1(\mu)d\mu \end{aligned} \tag{2.23}$$

An estimator of μ is then given by

$$\begin{aligned} \tilde{\mu} = E(\mu|y) &= \frac{\int \mu\phi(y-\mu)g(\mu)}{\int \phi(y-\mu)g(\mu)} \\ &= \frac{(1-\pi)\int \mu\phi(y-\mu)g_1(\mu)d\mu}{\pi\phi(y) + (1-\pi)\int \phi(y-\mu)g_1(\mu)d\mu} \end{aligned} \tag{2.24}$$

If $g_1(\mu) = \frac{e^{-\beta\mu}\mu^{\alpha-1}\beta^\alpha}{\Gamma(\alpha)}$, then

$$\begin{aligned} \int \mu\phi(y-\mu)g_1(\mu)d\mu &= \frac{\beta^\alpha}{\Gamma(\alpha)\sqrt{2\pi}} e^{\frac{-y^2}{2}+\frac{\beta^2}{2}-y\beta}\int_{\mu>0} \mu^{\alpha+1-1} e^{\frac{-(\mu-(y-\beta))^2}{2}}d\mu \\ &= \frac{e^{\frac{-y^2}{2}}}{\sqrt{2\pi}} H(\alpha+1, y, \beta) \end{aligned} \tag{2.25}$$

so that $E(\mu|y)$ reduces to

$$E(\mu|y) = \frac{H(\alpha+1, y, \beta)}{\frac{\pi}{1-\pi} + H(\alpha, y, \beta)}$$

where

$$\begin{aligned} H(\gamma, y, \beta) &= \frac{\beta^{(\gamma-1)}}{\Gamma(\gamma-1)} e^{\frac{\beta^2}{2} - y\beta} \int_{x>0} x^{\gamma} e^{\frac{-(x-(y-\beta))^2}{2}} dx \\ &= \frac{\beta^{(\gamma-1)}}{\Gamma(\gamma-1)} e^{\frac{\beta^2}{2} - y\beta} \Gamma(\gamma+1) e^{\frac{-(y-\beta)^2}{4}} D_{(-\gamma-1)}(y-\beta) \end{aligned}$$

and $\Phi(\alpha, \gamma, z) = \sum_{n=0}^{\infty} \frac{\alpha^{\bar{n}} z^n}{\gamma^{\bar{n}} n!}$ is the Kummer's Confluent hypergeometric function, $z^{\bar{n}} = \frac{\Gamma(z+n)}{\Gamma(z)}$ and

$$D_{(-\gamma-1)}(\mu/\sigma) = 2^{-(\gamma+1)/2} e^{-\mu^2/4\sigma^2} \left[\frac{\sqrt{\pi}}{\Gamma(\frac{\gamma+2}{2})} \Phi\left(\Gamma\left(\frac{\gamma+1}{2}\right), \frac{1}{2}, \frac{\mu^2}{2\sigma^2} \right) - \frac{\sqrt{2\pi}(\frac{\mu}{\sigma})}{\Gamma(\frac{\gamma+1}{2})} \Phi\left(\Gamma\left(\frac{\gamma+2}{2}\right), \frac{3}{2}, \frac{\mu^2}{2\sigma^2} \right) \right]$$

Provided we can estimate π and parameters of g_1, we could then substitute them into the expression for $\tilde{\boldsymbol{\mu}}$ and use the EB estimator to compute a likelihood ratio statistic. While the EB may be of independent interest in this problem, we will investigate performance of the modified LRT.

Another prior distribution that we propose to investigate is a Horseshoe like prior [3] restricted to the positive orthant. Thus, a priori we have

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

where $N(\mu, \sigma)_+$ is the $N(\mu, \sigma)$ restricted to the positive half, and $C(0,1)$ is the standard Cauchy restricted to positive half and $p(\tau)$ is some suitable prior on τ.

2.3 Numerical Studies

We compare the performance of our test statistics with LRT and Neyman-Pearson (NP) test statistic [17] at the center, i.e., the densest case with all μ_i equal, using numerical studies.

The NP test statistic for any fixed $\boldsymbol{\mu_1} \in \mathcal{H}_1$ is $\frac{\mathbf{X}^\top \boldsymbol{\mu_1}}{\boldsymbol{\mu_1}^\top \boldsymbol{\mu_1}}$. The NP test statistic (T_{NP}) at the center, i.e., for any $\boldsymbol{\mu_1} = \mu \mathbf{1}$ is $\frac{\sum X_i}{\sqrt{n}}$. The power of a size-$\alpha$ test using T_{NP} is $1 - \Phi\big(z_\alpha - \frac{\sum \mu_i}{\sqrt{n}}\big)$ where z_α is the upper-α point of a standard normal distribution. By NP lemma, the NP test statistic is most powerful for testing a simple null hypothesis against a simple alternative hypothesis. Hence T_{NP} is most powerful for testing $\mathcal{H}_0 : \boldsymbol{\mu} = \mathbf{0}$ *vs.* $\mathcal{H}_1 : \boldsymbol{\mu} = \boldsymbol{\mu_1}$ where $\boldsymbol{\mu_1} \in \mathcal{K}$.

For our simulation purposes we consider $n = 500$ and $\boldsymbol{\mu} \in \mathcal{K}$, equidistant from $\mathbf{0}$ with $||\boldsymbol{\mu}||_2 = \frac{\sqrt{n}}{4}$. We choose 50000 simulations to study the behaviour of the power curve using Monte Carlo for the test statistics at 5% level of significance. We investigate the power of our test statistics as we move from sparse to dense $\boldsymbol{\mu} \in \mathcal{K}$ keeping the $\mathcal{L}^2$ norm fixed. To illustrate the concept of sparsity level, suppose $||\boldsymbol{\mu}||_2 = \frac{\sqrt{n}}{4}$ and dimension is 500, the most sparse case would be $\boldsymbol{\mu} = (\frac{\sqrt{n}}{4}, 0, ..., 0)$ and the most dense case would be $\boldsymbol{\mu} = (\frac{1}{4}, \frac{1}{4}, ..., \frac{1}{4})$ with non-negative values in all 500 coordinates such that $\mathcal{L}^2$ norm is $\frac{\sqrt{n}}{4}$.

We simulate $\boldsymbol{\mu}$ for different sparsity level starting from the most sparse case $\boldsymbol{\mu} = (\frac{\sqrt{n}}{4}, 0, ..., 0)$ to the most dense case for $n = 500$ as follows:

- First, we simulate $\mu_{[1:i]}$ from truncated normal with mean 0 and variance 1 bounded below at 0 and take the rest of the components of $\boldsymbol{\mu}$ as 0.
- Then we normalize the $\boldsymbol{\mu}$ vector so that $||\boldsymbol{\mu}||_2 = \sqrt{n}/4$.
- Repeat steps 1 and 2 for each $i = 1, \ldots, n$ to generate 500 different $\boldsymbol{\mu}$ vectors.

We also report average power and minimum power for strong sparsity as well as for extremely dense configuration of $\boldsymbol{\mu}$. For j=1,2

$$\text{Average power}_j = \int P(\boldsymbol{\mu}) g_j(\boldsymbol{\mu}) d\boldsymbol{\mu}$$

$$\text{Minimum power}_j = \min_{\boldsymbol{\mu} \sim g_j(\boldsymbol{\mu})} P(\boldsymbol{\mu})$$

where $g_1(\boldsymbol{\mu})$ is the empirical limit of the sampling scheme for strong sparsity and $g_2(\boldsymbol{\mu})$ for extremely dense configuration of $\boldsymbol{\mu}$. For strong sparsity, we simulate $\boldsymbol{\mu}$ for $R = 10000$ times as follows:

- Take $\mu_1 = 1$. Simulate μ_i from $U(0, \mu_{i-1})$ for $i = 2, ..., n$.
- Normalize $\boldsymbol{\mu}$ to have a norm $\sqrt{n}/4$.

For extremely dense configuration, we simulate $\boldsymbol{\mu}$ for $R = 10000$ times as follows:

- Take $\mu_1 = 1$. Simulate μ_i from $U(0, 1)$ for $i = 2, ..., n$.
- Sort and normalize $\boldsymbol{\mu}$ to have a norm $\sqrt{n}/4$.

2.3.1 Test Statistic 1

Selection of α through grid points

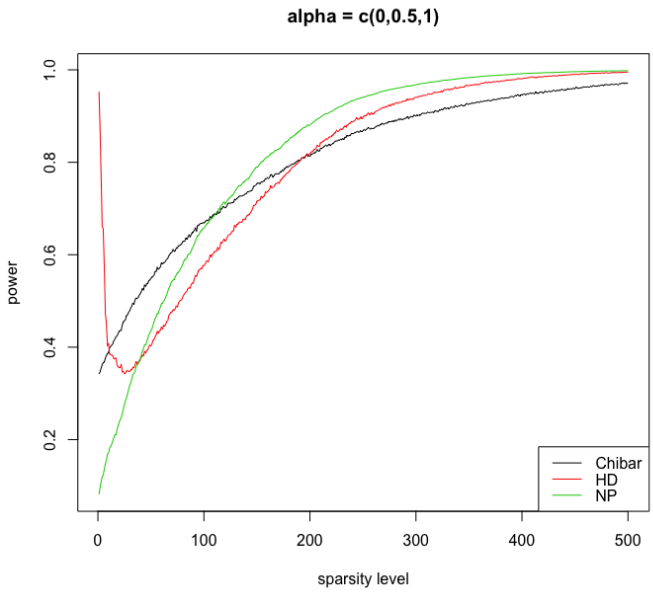

Figure 2.3: Power curve for different sparsity level

Power	Test	Strong Sparsity	Dense
Min Power	Chibar	0.34	0.97
	HD	0.04	1.00
	NP	0.08	1.00
Avg Power	Chibar	0.36	0.98
	HD	0.18	1.00
	NP	0.11	1.00

Table 2.1: Type-I error, minimum power and average power for $T(\alpha)$ corresponding to the strongest signal

Selection of α corresponding to the strongest signal

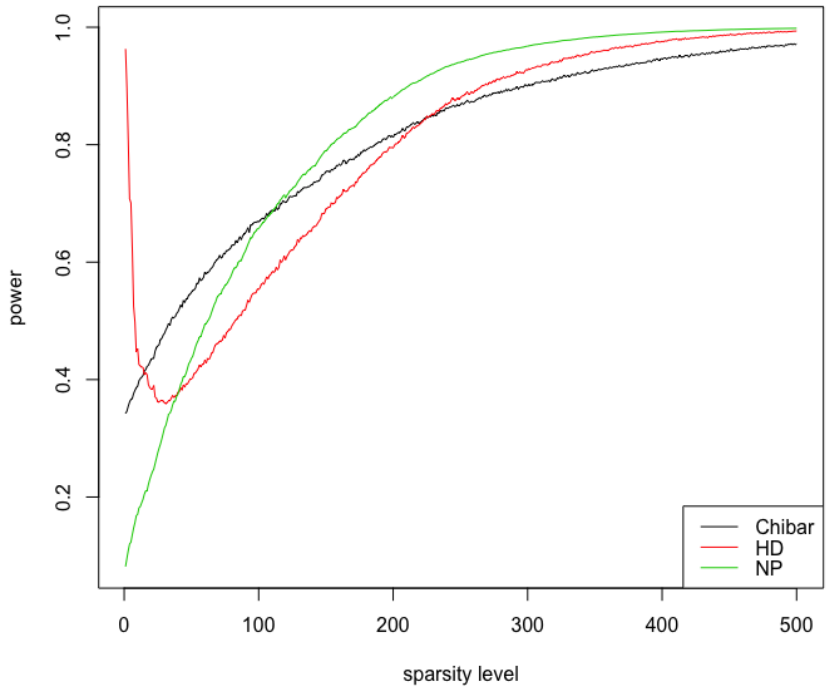

Figure 2.4: Power curve for different sparsity level

2.3.2 Test Statistic 2

$L(\gamma)$ and maximum

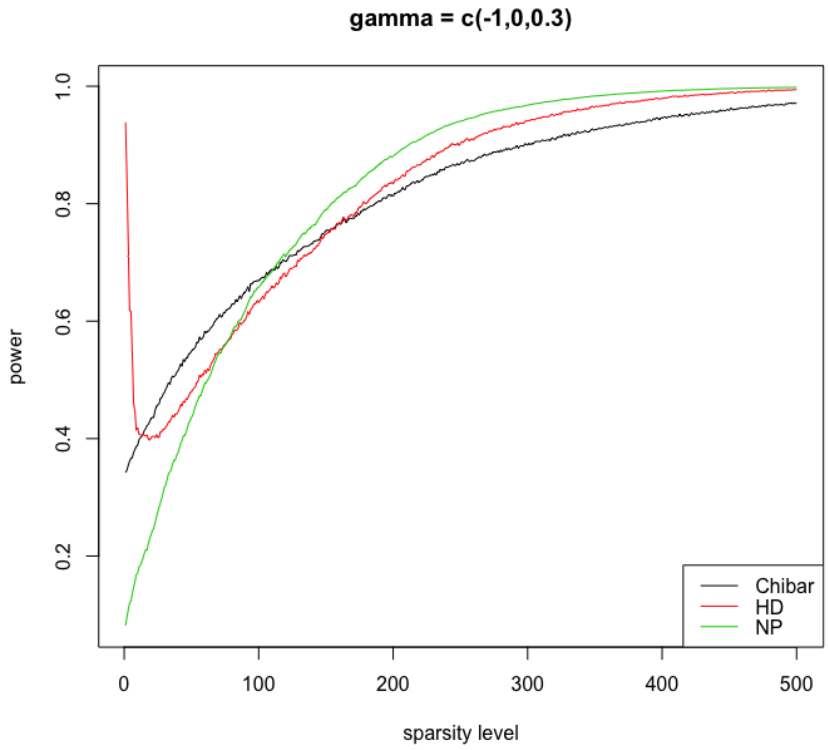

Figure 2.5: Power curve for different sparsity level

$L(\gamma)$ and sum of r extreme order statistics

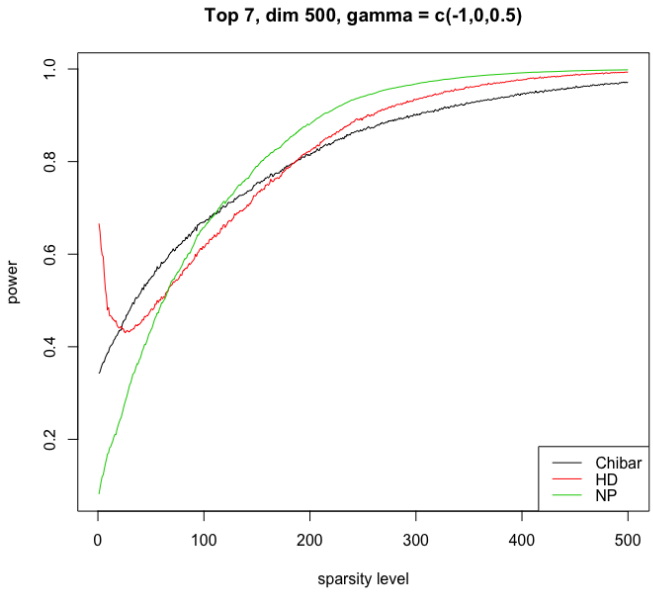

Figure 2.6: Power curve for different sparsity level

Power	Test	Strong Sparsity	Dense
Min Power	Chibar	0.34	0.98
	HD	0.47	0.99
	NP	0.08	1.00
Avg Power	Chibar	0.36	0.98
	HD	0.63	1.00
	NP	0.11	1.00

Table 2.2: Minimum power and average power for $L(\gamma)$ and sum of r extreme order statistics

2.3.3 Test Statistic 3

$L(\zeta)$ and maximum

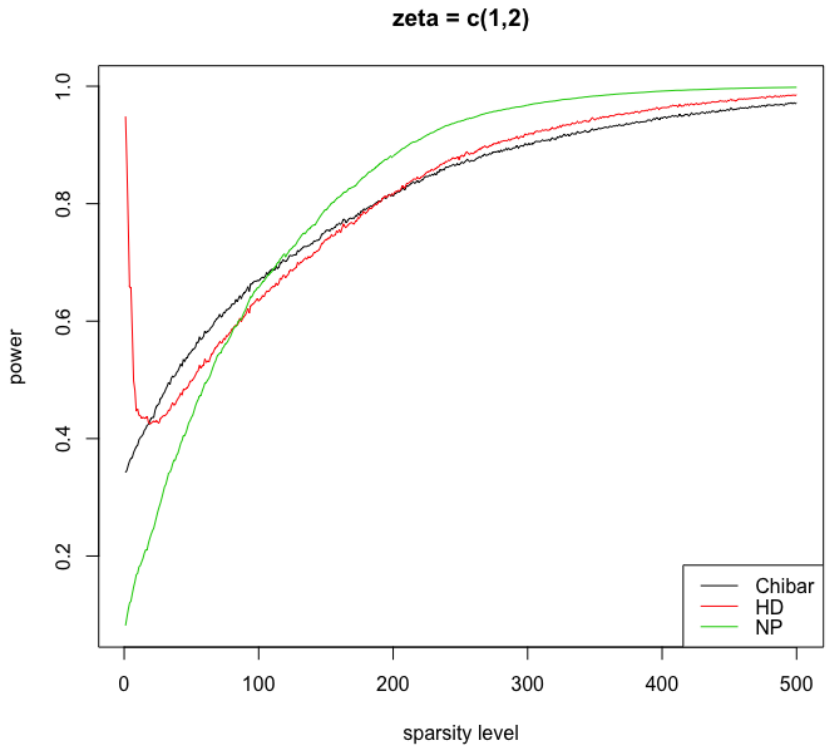

Figure 2.7: Power curve for different sparsity level

$L(\zeta)$ and sum of r extreme order statistics

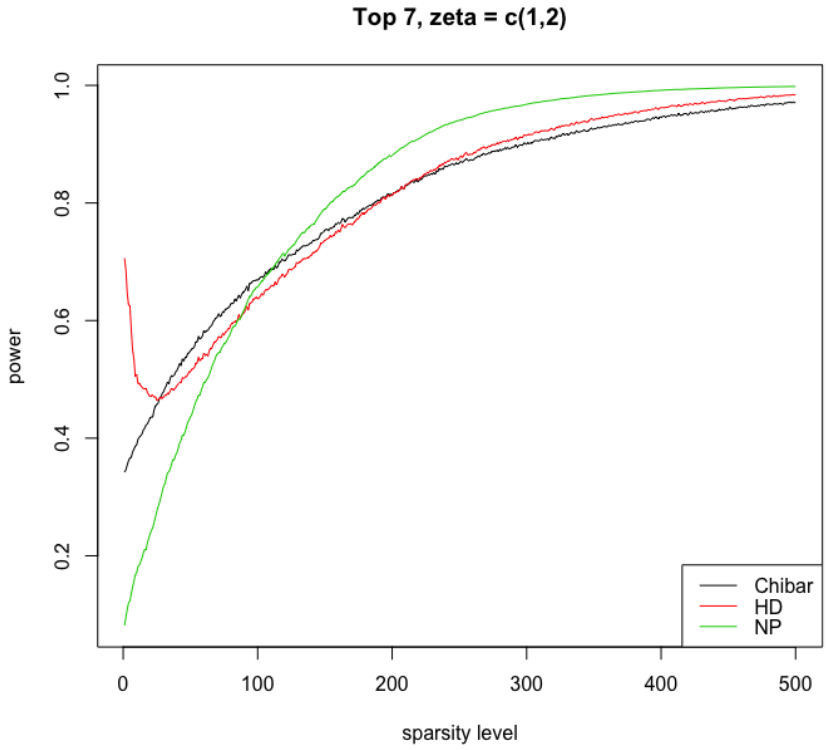

Figure 2.8: Power curve for different sparsity level

2.4 Some remarks and Future Work

For tests where we use the asymptotic independence, Ho and Hsing (1996, Journal of Applied Probability) [8], Hsing (1995, Annals of Prob.)) [9] showed that when x_i's are strong mixing process or stationary normal, then $\max_i x_i$ and $\sum_{i=1}^{n} z_i$ are asymptotically independent. Therefore, under H_0, if $x \sim N(0, \Sigma)$ where Σ satisfies stationary condition, the above test is still asymptotically valid. For more general Σ, it is in general a challenging problem. However, we may think about some approximate solution.

When $x \sim N(\mu, \Sigma)$ for known Σ, Tang et al. (Biometrika, 1989) [21] proposed a transformation, $z = Ax \sim N(A\mu, I)$ from $A^T A = \Sigma^{-1}$. A is not unique, but Tang et al. suggested A such that the center of $\mathcal{O}^+ = \{\mu | \mu_i \geq 0\}$ is equal to that of $A\mathcal{O}^+ = \{A\mu | \mu \in \mathcal{O}^+\}$. The cone $A\mathcal{O}^+$ is approximated by $\mathcal{O}^+$, but it is not clear

Power	Test	Strong Sparsity	Dense
Min Power	Chibar	0.33	0.97
	HD	0.47	0.99
	NP	0.08	1.00
Avg Power	Chibar	0.36	0.98
	HD	0.66	0.99
	NP	0.11	1.00

Table 2.3: Minimum power and average power for $L(\zeta)$ and sum of r extreme order statistics

how good this approximation is in high dimension. With this, Tang et al. proposed an approximate LR which is

$$P(g(z) \geq c) = \sum_{i=0}^{n} \binom{n}{i} / 2^n P(\chi_i^2 \geq c)$$

where $g(z) = \sum_{i=1}^{n} (\max(z_i, 0))^2$. When $n \to \infty$, the computation of probability is not that obvious. There may be some approximate algorithm (maybe based on (importance) sampling). Instead, when we obtain z from the transformation A, we apply $(\max_i z_i, \sum_i z_i)$ as we did for $\Sigma = I$.

A case when the linear transformation will leave the cone unchanged is when the matrix square root of Σ is a positive matrix. Thus, if we could choose a positive square root of Σ then the testing problem remains invariant. The following result shows that a positive square root exists if Σ itself is a positive matrix.

Proposition 1. Let Ω be a positive definite matrix. Then there exists a square root, A of Ω, i.e., $AA' = \Omega$ with all entries non-negative iff all entries of Ω are non-negative.

However, when not all pairwise correlations are non-negative, we will have to choose a transformation which will preserve some of the basic features of the testing problem. Exact invariance of the proposed tests may not be achieved in such a case. We can target specific alternatives of $\boldsymbol{\mu}$ to get an approximate test. For example, if we want to leave some pre-specified alternative, $\boldsymbol{\mu_1}, \ldots, \boldsymbol{\mu_k}$, $k < n$ unchanged then we would seek a square root of Ω of the form $A = \Omega^{1/2} Q$ where $\Omega^{1/2}$ is the symmetric square root of Ω and Q is an orthogonal matrix such that $\boldsymbol{\mu_1}, \ldots, \boldsymbol{\mu_k}$ are eigenvectors of A. This can be achieved by searching for the right Q over $O(n)$, the group of orthogonal matrices with determinant one.

When exact solution is not possible, we would want to pose the optimization problem of choosing Q by minimizing a suitable objective function that describes the distance between the specified alternative vectors $\boldsymbol{\mu_1}, \ldots, \boldsymbol{\mu_k}$ from the eigenspace of A, i.e., $\min_{Q \in O(n)} d\big((\boldsymbol{\mu_1}, \ldots, \boldsymbol{\mu_k}),\ \Omega^{1/2} Q(\boldsymbol{\mu_1}, \ldots, \boldsymbol{\mu_k})\big)$ where d is some distance criteria.

When Σ is unknown, we have m observational vectors $\mathbf{x}_i \sim N(\mu, \Sigma)$ for $1 \leq i \leq m$ where n is the dimensionality of vector and m is sample size. We may consider some estimator $\hat{\Sigma}$ or $\hat{\Sigma}^{-1}$ assuming possibly some sparse structure on precision matrix as in Cai et al. (JRSS-B, 2014) [2] in high dimensional mean test.

One drawback is that we have no idea how $A\mathcal{O}^+$ approximate $\mathcal{O}^+$ in high dimension. More specifically, it is not clear whether $\max_i z_i$ support the large value of $\mu_i > 0$ under the alternative.

Another approach is to consider the transformation $\Sigma^{-1} \equiv \Omega$. Cai et al. (2014) [2] also considered this transformation, however, Cai et al. (2014) used the maximum of the transformed data. In fact, after transformation based on Ω, if we take sum of them, $1^T \Omega X$, then this becomes the O'brien test. More specifically, after transformation $Z = \Omega X = (z_1, \ldots, z_n)^T$, the standardized form $\frac{1^T \Omega X}{\sqrt{1^T \Omega 1}} = \frac{\sum_{i=1}^n z_i}{\sqrt{1^T \Omega 1}}$ is the Obrien's test. On the other hand, Cai et al. used

$$\max_{1 \leq i \leq n} \frac{z_i^2}{w_{ii}} \tag{2.26}$$

where w_{ii} is the ith diagonal term in Ω.

Cai et.al. is not directly applicable to the conic alternative case since the cone defined by ΩC could be any closed cone in $\mathbb{R}^d$. As in method1 we need to investigate the change in alternative cone due to the linear transformation.

Chapter 3

Integrating Meta-analysis and Individual Patient Data

3.1 Introduction

Individual Patient Data (IPD) is the gold standard in statistical methods. However getting IPD usually involves extra cost and hence we may have limited resources. With the increasing need of resources to store large data, meta-analysis methods are becoming very popular. When both the MA and IPD studies are available, combining these two levels of data could improve the overall meta-analysis estimates, compared to utilizing MA studies alone. The general idea of this paper is to aggregate information from IPD studies and MA studies (summary statistics) which includes results from published journals,etc to obtain a combined estimate for the parameter of interest.

One common application of combining studies is estimating the effect of several treatments in a multicenter trial. We start with the case of continuous response and provide the combined treatment effects across trials when the treatment effect is fixed and common across trials. Assuming the observations within and between the studies are independent, we investigate the loss of efficiency from using combined estimator with various percentages of MA and IPD studies. When treatments are fixed and trial effects are random, Mathew and Nordstorm (2010) [16] derived the necessary and sufficient condition for the IPD estimator to coincide with meta-analysis estimator for a general within trial covariance matrix. The condition for equality requires that the fraction of observations corresponding to any given treatment is same across trials for the linear model. In practice, it is more likely to have studies with differential allocation to treatments. For such models, we studied the relative efficiency of analyzing IPD versus combining IPD & MA studies under systematic departures from the same allocation proportion condition.

For treatment vs. control comparison with continuous outcome, Olkin and Sampson (1998) showed MA estimator is equivalent to IPD estimator if no study-by-treatment interactions and variances are constant across trials [18]. Mathew and Nordstrom (1999) further showed that this equivalence holds even if the error variances are different across trials [15]. However, Mathew and Nordstorm, 2010 provided conditions on the loss in efficiency for random effects linear model for fixed k and fixed n_k. Hence we provide the loss of efficiency theoretically for fixed and finite number of studies and sample size in each study for the MA and IPD combined estimator and show the impact of distribution of sample sizes in different studies on the performance of the combined estimator.

For all commonly used parametric and semi-parametric models, Lin and Zeng (2010) showed that IPD estimator has no gain in efficiency over MA estimator asymptotically in fixed effects model [13]. In the context of random effects model, we provide the MA and IPD combined estimator and study its performance empirically for a generalized model. This is more general setup where the covariate or response may be categorical or continuous and the common parameter of interest can be multidimensional. For such models, we do not provide selection criteria based on treatment allocation across trials to minimize the loss of efficiency. The condition for equality of IPD estimator and MA estimator for a general model is known asymptotically(Lin and Zeng 2010). Nevertheless, we explore the performance of the combined estimator through relative efficiency.

3.2 Methods

We start with the problem of combining a one-dimensional parameter of interest, the treatment-control effect for a linear model. We finally extend it to estimating multi-dimensional parameter for a generalized linear model for discrete/continuous covariate. For linear models with treatment vs control comparison, we propose a method to select the IPD studies among the available studies so as to get the maximum efficiency in terms of the combined estimator.

3.2.1 Linear Mixed Effects Model

Consider that there is one continuous outcome of interest and assume that there are two groups, namely treatment(T) and control(C) group for all k independent studies. Suppose, for k_1 studies in S_1, IPD studies are available and for the remaining $k-k_1 = k_2$ studies in S_2 we have access to only MA results where $S_1 \cup S_2 = S$, the total set of studies. Suppose further that, n_{i1} and n_{i2} be the number of persons in each group for ith study with $n_{i1} + n_{i2} = n_i$. The model for response y_{ij} for the jth patient in study i is

$$\begin{aligned} y_{ij} &= \alpha_i + \beta x_{ij} + \epsilon_{ij}, \quad i = 1, ..., k_1 \\ \epsilon_{ij} &\sim N(0, \sigma_i^2), \quad \alpha_i \sim N(\alpha, \sigma_\alpha^2) \end{aligned} \tag{3.1}$$

where α_i and ϵ_{ij} are assumed independent and x_{ij} is 0/1 denoting the treatment and control groups respectively.

Aggregation

For the above model, α acts as a common nuisance parameter across studies. This means that even if we had access to all IPD studies, the MA estimator and the IPD estimator doesn't necessarily coincide (Mathew and Nordstorm 2010) [16]. The two estimators coincide if and only if the vectors $(n_{i1}/n_i, n_{i2}/n_i)$ are all equal for $i = 1, ..., k$. Let n_{ipd} be the total number of patients for the IPD studies. For any finite and fixed k and n_k, IPD estimator is more efficient than MA estimator. However IPD estimator has no efficiency gain over MA estimator asymptotically. In meta-analysis literature, we generally assume σ_i^2 and σ_α^2 are known.
For study i, $\mathbf{y}$ is normal with

$$\begin{aligned} E(\boldsymbol{y_i}) &= \alpha \mathbf{1}_{\boldsymbol{n_i}} + \beta X_i \\ Cov(\boldsymbol{y_i}) &= H_i = \sigma_\alpha^2 \mathbf{1}_{\boldsymbol{n_i}} \mathbf{1}_{\boldsymbol{n_i}}^{\boldsymbol{T}} + \sigma_i^2 I_{n_i} \end{aligned} \tag{3.2}$$

For the k_2 MA studies, assuming the same model as (1), we have the restricted maximum likelihood estimates (REML), $\hat{\beta}_i$ and their estimated variances, $\hat{\sigma}_i^2$. So the model for meta-analysis study is

$$\begin{aligned} \hat{\beta}_i &\sim N(\beta, \hat{\sigma}_i^2), \quad i = 1, \ldots, k_2 \\ \hat{\sigma}_i^2 &= \frac{n_{i1} n_{i2}}{n_i \sigma_i^2} = \frac{n_i \pi_i (1 - \pi_i)}{\sigma_i^2} \end{aligned} \tag{3.3}$$

where $\pi_i = n_{i1}/n_i$ is the proportion of treatments in study i.

With the above model, the combined estimator of (α, β) and the variance of the combined estimator are

$$\begin{pmatrix} \hat{\alpha}_{comb} \\ \hat{\beta}_{comb} \end{pmatrix} = (U^T \Sigma^{-1} U)^{-1} U^T \Sigma^{-1} Y^* \tag{3.4}$$

$$Cov \begin{pmatrix} \hat{\alpha}_{comb} \\ \hat{\beta}_{comb} \end{pmatrix} = (U^T \Sigma^{-1} U)^{-1} \tag{3.5}$$

where

$$Y^* = \begin{pmatrix} \mathbf{y} \\ \hat{\boldsymbol{\beta}} \end{pmatrix} \quad U = \begin{pmatrix} \mathbf{1_{n_{ipd}}} & X \\ 0 & \mathbf{1_{k_2}} \end{pmatrix} \quad \Sigma = \begin{pmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{pmatrix}$$

$$\Sigma_1 = \begin{pmatrix} H_1 & & \\ & \ddots & \\ & & H_{k_1} \end{pmatrix} \quad \Sigma_2 = \begin{pmatrix} \hat{\sigma}_1^2 & & \\ & \ddots & \\ & & \hat{\sigma}_{k_2}^2 \end{pmatrix}$$

Our parameter of interest is β. It is easy to see that the estimator in (3.4) is unbiased. Hence we want to explore the variance for the combined estimate for the parameter of interest which is given by

$$v(\hat{\beta}_{comb}) = \Big[\sum_{i \in S_1} \frac{n_i \pi_i (1 + n_i(1-\pi_i) b_i)}{\sigma_i^2 a_i} + \sum_{i \in S_2} \frac{n_i \pi_i (1-\pi_i)}{\sigma_i^2} - \frac{(\sum_{i \in S_1} \frac{n_i \pi_i}{\sigma_i^2 a_i})^2}{\sum_{i \in S_1} \frac{n_i}{\sigma_i^2 a_i}} \Big]^{-1} \tag{3.6}$$

where $a_i = 1 + n_i b_i,\ b_i = \frac{\sigma_\alpha^2}{\sigma_i^2}$ & $\pi_i = \frac{n_{i1}}{n_i}$. Assuming all $n_i = n$ and $\sigma_i = \sigma$, the variance expression simplifies to

$$v(\hat{\beta}_{comb}) = \frac{\sigma^2}{n} \Big[\frac{nb}{a} \sum_{i \in S_1} \pi_i (1-\pi_i) + \sum_{i \in S_2} \pi_i (1-\pi_i) + \frac{1}{ak_1} \Big(\sum_{i \in S_1} \pi_i \Big) \Big(\sum_{i \in S_1} (1-\pi_i) \Big) \Big]^{-1} \tag{3.7}$$

where $a = 1 + nb,\ b = \frac{\sigma_\alpha^2}{\sigma^2}$.

The ideal situation is having access to all IPD studies. So we want to compare the efficiency of $\hat{\beta}_{comb}$ with the "best" estimator when all IPD studies are available. The variance of the unbiased minimum variance unbiased estimator of β when all k studies are IPD is

$$v(\hat{\beta}_{allIPD}) = \Big[\sum_{i=1}^{k} \frac{n_i \pi_i (1 + n_i(1-\pi_i) b_i)}{\sigma_i^2 a_i} - \frac{\big(\sum_{i=1}^{k} \frac{n_i \pi_i}{\sigma_i^2 a_i} \big)^2}{\sum_{i=1}^{k} \frac{n_i}{\sigma_i^2 a_i}} \Big]^{-1} \tag{3.8}$$

which simplies to the following when $n_i = n$ and $\sigma_i = \sigma$

$$v\left(\hat{\beta}_{allIPD}\right) = \sigma^2 \Big[\frac{n^2 b}{a} \sum_{i=1}^{k} \pi_i (1-\pi_i) + \frac{n}{ak} \Big(\sum_{i=1}^{k} \pi_i \Big) \Big(\sum_{i=1}^{k} (1-\pi_i) \Big) \Big]^{-1} \tag{3.9}$$

Note that the above expression is different in case of a fixed effects model is

$$v\left(\hat{\beta}_{allIPD(FE)}\right) = \sigma^2\left[\frac{n}{k}\Big(\sum_{i=1}^{k}\pi_i\Big)\Big(\sum_{i=1}^{k}(1-\pi_i)\Big)\right]^{-1} \tag{3.10}$$

whereas variance for estimator of β assuming only MA results are available for all k studies is

$$v(\hat{\beta}_{allMA}) = \left[\sum_{i=1}^{k}\frac{n_i\pi_i(1-\pi_i)}{\sigma_i^2}\right]^{-1} \tag{3.11}$$

For an all meta-analysis aggregation, the expression of variance for a random effects model isn't different from a fixed effects model since we get the estimates first for each study and then combine the information across studies whereas for IPD studies we combine the data first and then get the variance.

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

$$\frac{\left[\frac{n^2b}{a}\sum_{i\in S_1}\pi_i(1-\pi_i) + n\sum_{i\in S_2}\pi_i(1-\pi_i) + \frac{n}{ak_1}\Big(\sum_{i\in S_1}\pi_i\Big)\Big(\sum_{i\in S_1}(1-\pi_i)\Big)\right]}{\left[\frac{n^2b}{a}\sum_{i=1}^{k}\pi_i(1-\pi_i) + \frac{n}{ak}\Big(\sum_{i=1}^{k}\pi_i\Big)\Big(\sum_{i=1}^{k}(1-\pi_i)\Big)\right]} \tag{3.12}$$

We will use the above expression for relative efficiency for simulation purposes in section 3.3.

Selection

The bigger question here is whether it matters which IPD studies are selected. If yes, how can we minimize the loss in efficiency?

The variance of the combined estimator in equation (3.8) can be simplified to

$$v(\hat{\beta}_{comb}) = \left[\sum_{i\in S_1}\frac{n_i}{\sigma_i^2 a_i}(\pi_i - \tilde{\pi}_i)^2 + \sum_{i\in S}\frac{n_i\pi_i(1-\pi_i)}{\sigma_i^2}\right]^{-1} \tag{3.13}$$

where $\tilde{\pi_i} = \frac{\sum \frac{\pi_i n_i}{\sigma_i^2 a_i}}{\sum \frac{n_i}{\sigma_i^2 a_i}}$.

We see that the expression for variance in equation (3.13) do not involve y and depends only on π_i's, n and σ's. 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 "maximization" of $\sum_{i \in S_1} \frac{n_i}{\sigma_i^2 a_i}(\pi_i - \tilde{\pi_i})^2$. Since the π_i's are fixed, finding the optimum combination of IPD studies and MA studies is not exactly a maximization problem. Essentially this is a combinatorial optimization problem which can be formulated as follows:

Given $(x_1, y_1), \ldots, (x_k, y_k)$ with $x_i > 0$, $y_i > 0$, $\max_{A \subset \{1,\ldots,k\}, |A|=k_1} \sum_{i \in A} y_i (x_i - \bar{x}_A)^2$
where $\bar{x}_A = k^{-1} \sum_{i \in A} x_i$.

For an interesting special case $n_i = n$ and $\sigma_i^2 = \sigma^2$, the variance further reduces to

$$v(\hat{\beta}_{comb}) = \frac{\sigma^2}{n} \left[\sum_{i \in S_1} \frac{1}{a} (\pi_i - \tilde{\pi_i})^2 + \sum_{i \in S} \pi_i (1 - \pi_i) \right]^{-1} \tag{3.14}$$

for this special case, we propose an algorithm to find the optimum combination of IPD and AD studies. Firstly we arrange the studies in increasing order of their π_i values. We include studies with extreme π_i values in the IPD subset, 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 estimator coincides with 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 MA studies, we propose allocating studies with similar proportion of treatments to MA study subset and studies with widely varying proportion of treatments to the IPD set.

3.2.2 Generalized Linear Model

The generalized linear model is a generic model where the response variable may be continuous, categorical, survival, longitudinal data, etc. Consider a similar setup as linear model where we have access to full data, (Y_{ki}, X_{ki}), $i = 1, ..., n_k$ for k_1 studies in S_1 whereas for k_2 studies in S_2, only the MA results for the parameter of interest, β are available where $S_1 \cup S_2 = S$. Here β can be a multi-dimensional parameter.

The IPD data $(Y_{ki}, X_{ki}|\beta_k, \eta_k) \sim f_k(Y_{ki}, X_{ki}; \beta_k, \eta_k)$ where f may have distribution models other than a normal distribution. We assume a random effects model for β_k and η_k where $\beta_k|\beta \sim N(\beta, \Sigma_\beta)$ and $\eta_k|\eta \sim N(\eta, \Sigma_\eta)$ and η is common nuisance parameter. This is a more general model and can be reduced to a mixed effects model or fixed effects model with the corresponding variances are taken to be zero. This setup allows having additional covariate, Z which is independent of X and the parameter corresponding to Z can be included in η. This work is inspired by Dr Dinchen's recent work in meta-analysis.

For the k_2 MA studies, we have the estimates with their variances, $(\hat{\beta}_k, \hat{var}(\hat{\beta}_k))$ and $\hat{\beta}_k|\beta_k \sim N(\beta_k, \hat{var}(\hat{\beta}_k)), \beta_k|\beta \sim N(\beta, \Sigma_\beta)$ where we later denote $\hat{var}(\hat{\beta}_k) = M_{k_{11}}$.

We assume that all studies are independent. We use the likelihood from IPD and MA studies to combine information and estimate the common parameter of interest.

Aggregation

The full likelihood combining IPD studies and MA studies is

$$\begin{aligned} L(\beta, \eta) &= \prod_{k \in S_1} L_k(\beta, \eta) \prod_{k \in S_2} L_k(\beta) \\ \log L(\beta, \eta) = l(\beta, \eta) &= \sum_{k \in S_1} l_k(\beta, \eta) + \sum_{k \in S_2} l_k(\beta) \end{aligned} \tag{3.15}$$

The IPD part of log-likelihood for study $k \in S_1$ is

$$\begin{aligned} l_k(\beta, \eta) &= \log L_k(\beta, \eta) \\ &= \log \int \frac{L_k(\beta_k, \eta_k)}{C_1} e^{-\frac{1}{2} \begin{pmatrix} \beta_k - \beta \\ \eta_k - \eta \end{pmatrix}^T \begin{pmatrix} \Sigma_\beta & 0 \\ 0 & \Sigma_\eta \end{pmatrix}^{-1} \begin{pmatrix} \beta_k - \beta \\ \eta_k - \eta \end{pmatrix}} d\beta_k d\eta_k \end{aligned} \tag{3.16}$$

where $C_1 = \left|A\right|^{-\frac{1}{2}}$ and $A = \begin{pmatrix} \Sigma_\beta & 0 \\ 0 & \Sigma_\eta \end{pmatrix}^{-1}$

Expanding $L_k(\beta_k, \eta_k)$ around the MLE and ignoring the higher order terms,

$$L_k(\beta_k, \eta_k) = \frac{L_k(\hat{\beta}_k, \hat{\eta}_k)}{C_2} e^{-\frac{1}{2} \begin{pmatrix} \beta_k - \hat{\beta}_k \\ \eta_k - \hat{\eta}_k \end{pmatrix}^T \begin{pmatrix} I_{\beta_k \beta_k} & I_{\beta_k \eta_k} \\ I_{\eta_k \beta_k} & I_{\eta_k \eta_k} \end{pmatrix} \Big|_{\hat{\beta}_k, \hat{\eta}_k} \begin{pmatrix} \beta_k - \hat{\beta}_k \\ \eta_k - \hat{\eta}_k \end{pmatrix}} \tag{3.17}$$

where $C_2 = \left|B_k\right|^{-\frac{1}{2}}$ and $B_k = \begin{pmatrix} I_{\beta_k\beta_k} & I_{\beta_k\eta_k} \\ I_{\eta_k\beta_k} & I_{\eta_k\eta_k} \end{pmatrix}\Big|_{\hat{\beta}_k,\hat{\eta}_k}$

Combining (3.16) and (3.17), the IPD part of likelihood for study k is

$$l_k(\beta,\eta) = l_k(\hat{\beta}_k,\hat{\eta}_k) - \frac{1}{2}\begin{pmatrix}\hat{\beta}_k-\beta \\ \hat{\eta}_k-\eta\end{pmatrix}^T\left(\begin{pmatrix}\Sigma_\beta & 0 \\ 0 & \Sigma_\eta\end{pmatrix} + \left(\begin{pmatrix} I_{\beta_k\beta_k} & I_{\beta_k\eta_k} \\ I_{\eta_k\beta_k} & I_{\eta_k\eta_k}\end{pmatrix}\Big|_{\hat{\beta}_k,\hat{\eta}_k}\right)^{-1}\right)^{-1}\begin{pmatrix}\hat{\beta}_k-\beta \\ \hat{\eta}_k-\eta\end{pmatrix} - \log C_3 \quad (3.18)$$

where $C_3 = |\Delta_k|^{-\frac{1}{2}}$ and $\Delta_k = (A^{-1} + B_k^{-1})^{-1}$

The MA part of likelihood for study k,

$$\begin{aligned} l_k(\beta) &= \log L_k(\beta) \\ &= \log \int \frac{1}{|M_{k_{11}}|^{1/2}|\Sigma_\beta|^{1/2}} e^{-\frac{1}{2}\left(\hat{\beta}_k-\beta_k\right)^T M_{k_{11}}^{-1}\left(\hat{\beta}_k-\beta_k\right)} e^{-\frac{1}{2}\left(\beta_k-\beta\right)^T\Sigma_\beta^{-1}\left(\beta_k-\beta\right)} d\beta_k \end{aligned}$$

where $M_k = B_k^{-1}$ and $M_{k_{11}} = (I_{\hat{\beta}_k,\hat{\beta}_k} - I_{\hat{\beta}_k,\hat{\eta}_k}(I_{\hat{\eta}_k,\hat{\eta}_k})^{-1}I_{\hat{\eta}_k,\hat{\beta}_k})^{-1}$

Integraing out β_k, we have

$$l_k(\beta) = -\frac{1}{2}\left(\hat{\beta}_k-\beta\right)^T\left(M_{k_{11}}+\Sigma_\beta\right)^{-1}\left(\hat{\beta}_k-\beta\right) - \log\left|M_{k_{11}}+\Sigma_\beta\right|^{\frac{1}{2}} \quad (3.19)$$

Putting (3.18) and (3.19) in (3.15) we have the full likelihood for (β,η),

$$\begin{aligned} l(\beta,\eta) = \sum_{k\in S_1} l_k(\hat{\beta}_k,\hat{\eta}_k) - \sum_{k\in S_1}\frac{1}{2}\begin{pmatrix}\hat{\beta}_k-\beta \\ \hat{\eta}_k-\eta\end{pmatrix}^T\begin{pmatrix}\Delta_{k_{11}} & \Delta_{k_{12}} \\ \Delta_{k_{21}} & \Delta_{k_{22}}\end{pmatrix}\begin{pmatrix}\hat{\beta}_k-\beta \\ \hat{\eta}_k-\eta\end{pmatrix} - \sum_{k\in S_1}\log|\Delta_k|^{-\frac{1}{2}} \\ - \sum_{k\in S_2}\frac{1}{2}\left(\hat{\beta}_k-\beta\right)^T\left(M_{k_{11}}+\Sigma_\beta\right)^{-1}\left(\hat{\beta}_k-\beta\right) - \sum_{k\in S_2}\log\left|M_{k_{11}}+\Sigma_\beta\right|^{\frac{1}{2}} \end{aligned}$$

Note that $\Delta_k = (A^{-1} + B_k^{-1})^{-1} = (A^{-1} + M_k)^{-1} = \begin{pmatrix} \Sigma_\beta + M_{k_{11}} & M_{k_{12}} \\ M_{k_{21}} & \Sigma_\eta + M_{k_{22}} \end{pmatrix}^{-1}$

To find the the combined estimator $\hat{\beta}$, we set $\frac{\partial l(\beta,\eta)}{\partial \beta} = 0$ and $\frac{\partial l(\beta,\eta)}{\partial \eta} = 0$ and solve the following equations simultaneously.

$$\hat{\beta} = \Big(\sum_{k \in S_1} \Delta_{k_{11}} + \sum_{k \in S_2} (M_{k_{11}} + \Sigma_\beta)^{-1} \Big)^{-1} \Big(\sum_{k \in S_1} (\Delta_{k_{11}} \hat{\beta}_k + \Delta_{k_{12}} (\hat{\eta}_k - \hat{\eta})) + \sum_{k \in S_2} (M_{k_{11}} + \Sigma_\beta)^{-1} \hat{\beta}_k \Big)$$

$$\hat{\eta} = \Big(\sum_{k \in S_1} \Delta_{k_{22}} \Big)^{-1} \Big(\sum_{k \in S_1} (\Delta_{k_{22}} \hat{\eta}_k + \Delta_{k_{21}} (\hat{\beta}_k - \hat{\beta})) \Big)$$

|||||| my edits |||||

$$l(\beta, \eta) = \sum_{k \in S_1} l_k(\hat{\beta}_k, \hat{\eta}_k) - \sum_{k \in S_1} \frac{1}{2} \begin{pmatrix} \hat{\beta}_k - \beta \\ \hat{\eta}_k - \eta \end{pmatrix}^T \begin{pmatrix} \Delta_{k_{11}} & \Delta_{k_{12}} \\ \Delta_{k_{21}} & \Delta_{k_{22}} \end{pmatrix} \begin{pmatrix} \hat{\beta}_k - \beta \\ \hat{\eta}_k - \eta \end{pmatrix} - \sum_{k \in S_1} \log |\Delta_k|^{-\frac{1}{2}}$$

$$- \sum_{k \in S_2} \frac{1}{2} \left(\hat{\beta}_k - \beta \right)^T (M_{k_{11}} + \Sigma_\beta)^{-1} \left(\hat{\beta}_k - \beta \right) - \sum_{k \in S_2} \log |M_{k_{11}} + \Sigma_\beta|^{\frac{1}{2}}$$

|||||| my edits |||||

3.3 Numerical Results

3.3.1 Linear Mixed Effects Model

For simulation studies, we present two main scenarios to show the importance of selection of studies for the combined estimator. For simplicity, we considered both k and n to be small and equal to 10 and $\beta = 1.5, \alpha = 0.5, \sigma^2 = 2.5, \sigma_\alpha^2 = 0.025$.

For $\pi_i = 0.3$ for all $i = 1(1)10$ and $\beta = 1.5$, the combined estimator performs quite well as is expected shown in table 3.1. This is true as the proportion of treatment is equal for all studies. However when the π_i are not equal, it we need to choose the IPD studies so as the reach a specified efficiency.

Table 3.1: **Relative efficiency (RE) of combined estimator to IPD estimator for different percentage of MA studies in the estimator**

% of study	0	20	40	60	80	100
Estimate	1.503	1.502	1.503	1.503	1.503	1.503
RE	0.970	0.983	0.986	0.991	0.997	1

Uniform distribution of proportion of treatment

For this scenario, we used *runif* to generate $k = 10$ random proportion of treatment over the interval $[0, 1]$ in table 3.2:

Table 3.2: **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

If the desired relative efficiency is 0.95, we could achieve that with lots of possible combinations of 60% IPD and 40% MA studies. However 60% IPD may not be cost effective. But if we choose the right combination of 40% IPD and 60% MA, 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 and 3:8 for MA.

The combinations 2:9 in 20% IPD, for example, represents that the study no 2:9 accounts for 80% MA studies and study no 1 & 10 account for 20% IPD and this distribution of studies between MA and IPD has the maximum relative efficiency among all combinations in 20% IPD group.

Bathtub distribution of proportion of treatment

This is an example of unbalanced proportion of treatment in table 3.3.

Table 3.3: **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 importance of selection of right IPD and MA study for the combined estimate is more clearly seen in this case. The all-meta estimator has a relative efficiency of 0.44 wrt to the all IPD estimator.

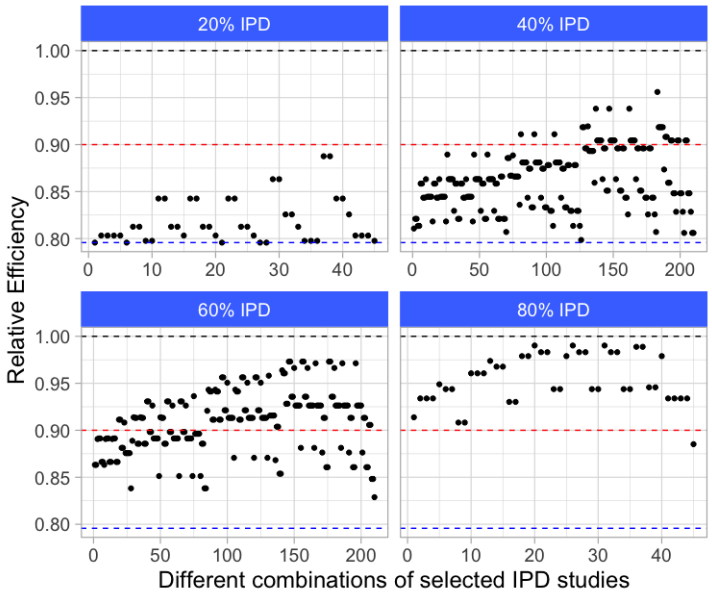

Figure 3.1: The relative efficiency of all 45 possible combinations for each of 20% IPD and 80% IPD, 210 possible combinations for each of 40% IPD and 60% IPD are plotted. Black dashed line represents the RE for all IPD studies which is 1, blue dashed line is RE for all MA studies which is 0.79 and the red dashed line is the desired RE, say 0.9, for example.

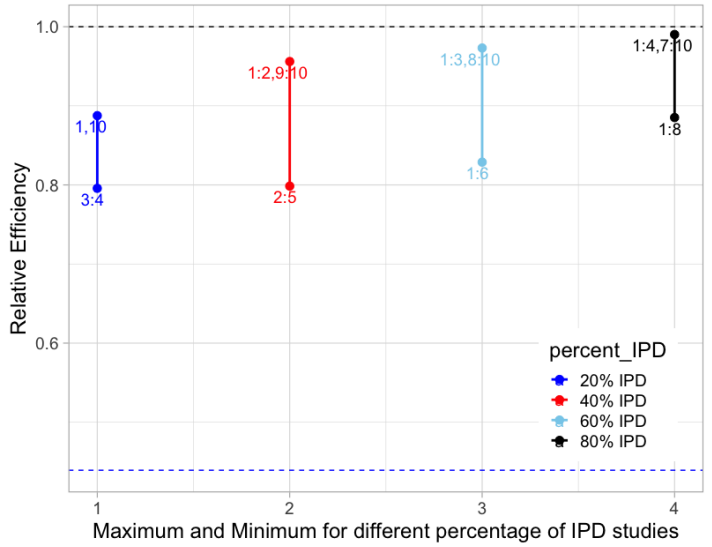

Figure 3.2: Plot showing the maximum and minimum relative efficiency and the MA combination within each percentage of IPD and MA studies.

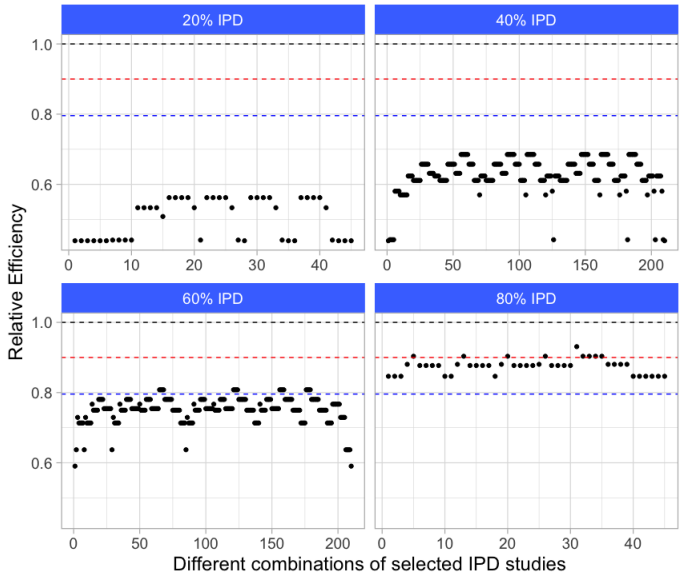

Figure 3.3: The relative efficiency of all 45 possible combinations for each of 20% IPD and 80% IPD, 210 possible combinations for each of 40% IPD and 60% IPD are plotted. Black dashed line represents the RE for all IPD studies which is 1, blue dashed line is RE for all MA studies which is 0.79 and the red dashed line is the desired RE, say 0.9, for example.

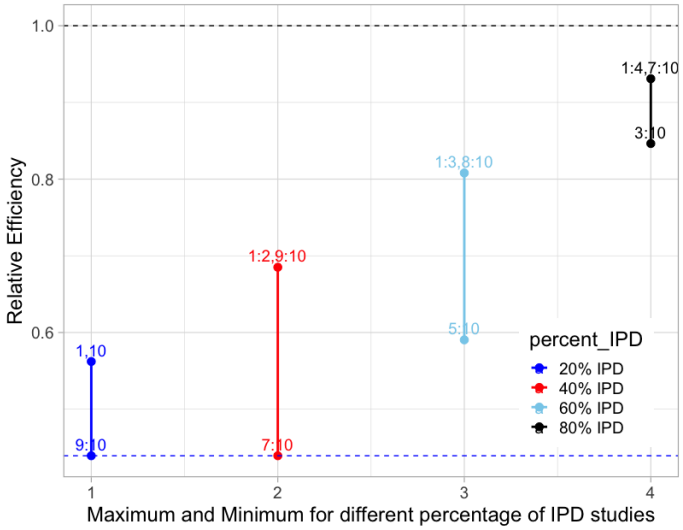

Figure 3.4: Plot showing the maximum and minimum relative efficiency and the MA combination within each percentage of IPD and MA studies.

The situation can worsen for severely unbalanced distribution of proportion of treatment. Note that the loss in efficiency is the issue with fixed small sample size studies. Asymptotically all-MA combined estimator are quite efficient compared to all-IPD combined estimator. Going forward, we need to quantify the unbalanced distribution of proportion of treatment among studies. We plan to extend the impact of selecting MA studies for multiple treatment problems as well.

3.3.2 Logistic Model

We illustrate an application of combining estimates in GLM through logistic model. Simulation results for different proportion of treatment across studies is presented. Here X_{ki} is 0/1 variable with proportion of treatment π_i and the response y_{ki} is binary with probability of success, $p_{ki} = \frac{exp(\alpha_k+\beta_k X_{ki})}{1+exp(\alpha_k+\beta_k X_{ki})}$.

Both β_k and α_k are assumed random with $\beta_k|\beta \sim N(\beta, \Sigma_\beta)$ and $\eta_k|\eta \sim N(\eta, \Sigma_\eta)$ where the true parameters are assumed: $\beta = 0.5, \alpha = 0.8, \sigma^2_\beta = 0.5, \sigma^2_\alpha = 0.5$. We present the estimates along with the estimated bias and estimated standard error for different choices of $n = n_i$ and k. We also present the relative efficiency of the combined estimator with respect to the aggregated estimator obtained when IPD data is available for all k studies.

We assume a differential proportion of treatment across studies with each of the proportions $(0.1, 0.4, 0.5, 0.6, 0.9)$ appearing $k/5$ times. The analytic form of the variance of the combined estimator is not yet explored. We therefore calculated the standard deviation and bias of the $\hat{\beta}$ empirically. The results are shown in table 3.4.

3.4 Discussion

In this paper, we assume that the covariate set is the same in each study. We plan to cover the case of disparate covariate information for multiple regression model across AD and IPD studies.

Table 3.4: **Model parameter estimates for different scenarios**

Scenario	(% IPD, % MA)	Estimate	Bias	Standard Deviation	Relative efficiency
k=100, n=100	(0, 100)	0.43	-0.07	0.05	0.66
	(20, 80)	0.44	-0.06	0.05	0.76
	(40, 60)	0.44	-0.06	0.05	0.81
	(60, 40)	0.44	-0.06	0.05	0.89
	(80, 20)	0.45	-0.05	0.05	0.94
	(100, 0)	0.45	-0.05	0.05	1.00
k=100, n=500	(0, 100)	0.49	-0.01	0.02	0.93
	(20, 80)	0.50	0.00	0.02	0.97
	(40, 60)	0.50	0.00	0.02	0.99
	(60, 40)	0.50	0.00	0.02	0.99
	(80, 20)	0.50	0.00	0.02	1.00
	(100, 0)	0.50	0.00	0.02	1.00

Bibliography

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

[2] T. Tony Cai, Weidong Liu, and Yin Xia. Two-sample test of high dimensional means under dependence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(2):349–372, 2014.

[3] Carvalho and Carvalho. The horseshoe estimator for sparse signals. *Biometrika*, 97(2):465–480, 2010.

[4] Sandor Csorgo, Erich Haeusler, and David M. Mason. The Asymptotic Distribution of Extreme Sums. *Ann. Probab.*, 19(2):783–811, April 1991.

[5] Bradley Efron and Carl Morris. Stein's Paradox in Statistics. *Scientific American*, 236(5):119–127, 1977.

[6] E. J. Gumbel. Les valeurs extrÃªmes des distributions statistiques. *Annales de l'institut Henri PoincarÃ©*, 5(2):115–158, 1935.

[7] E. J. Gumbel. The Return Period of Flood Flows. *Ann. Math. Statist.*, 12(2):163–190, June 1941.

[8] Hwai-Chung Ho and Tailen Hsing. On the asymptotic joint distribution of the sum and maximum of stationary normal random variables. *Journal of Applied Probability*, 33(1):138–145, March 1996.

[9] Tailen Hsing. A Note on the Asymptotic Independence of the Sum and Maximum of Strongly Mixing Stationary Random Variables. *Ann. Probab.*, 23(2):938–947, April 1995.

[10] Adel Javanmard and Jason D. Lee. A flexible framework for hypothesis testing in high-dimensions. *arXiv preprint arXiv:1704.07971*, 2017.

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

[12] Sang Lee, Johan Lim, Marina Vannucci, Eva Petkova, Maurice Preter, and Donald Klein. Order-Preserving Dimension Reduction Procedure for the Dominance of Two Mean Curves with Application to Tidal Volume Curves. *Biometrics*, 64:931–9, February 2008.

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

[14] Huimei Liu and Roger L. Berger. Uniformly More Powerful, One-Sided Tests for Hypotheses About Linear Inequalities. *Ann. Statist.*, 23(1):55–72, February 1995.

[15] Thomas Mathew and Kenneth Nordstrom. On the Equivalence of Meta-Analysis Using Literature and Using Individual Patient Data. *Biometrics*, 55(4):1221–1223, 1999.

[16] Thomas Mathew and Kenneth NordstrÃ¶m. Comparison of one-step and two-step meta-analysis models using individual patient data. *Biom J*, 52(2):271–287, April 2010.

[17] J. Neyman and E. S. Pearson. 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*, 231:289–337, 1933.

[18] I. Olkin and A. Sampson. Comparison of meta-analysis versus analysis of variance of individual patient data. *Biometrics*, 54(1):317–322, March 1998.

[19] Michael D. Perlman. One-Sided Testing Problems in Multivariate Analysis. *Ann. Math. Statist.*, 40(2):549–567, April 1969.

[20] Ning-Zhong Shi and Akio KudÃ´. 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*, 41(1):37–44, 1987.

[21] Dei-In Tang, Clare Gnecco, and Nancy L. Geller. An approximate likelihood ratio test for a normal mean vector with nonnegative components with application to clinical trials. *Biometrika*, 76(3):577–583, September 1989.

[22] Robert Tibshirani. Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[23] Yuting Wei, Martin J. Wainwright, and Adityanand Guntuboyina. The geometry of hypothesis testing over convex cones: Generalized likelihood tests and minimax radii. *arXiv:1703.06810 [cs, math, stat]*, March 2018. arXiv: 1703.06810.

[24] Ming Yu, Varun Gupta, and Mladen Kolar. Constrained High Dimensional Statistical Inference. *arXiv preprint arXiv:1911.07319*, 2019.