Spring 5-16-2011

Statistical Analysis in Empirical Bayes and in Causal inference

Hui Nie
University of Pennsylvania, niehui@wharton.upenn.edu

Statistical Analysis in Empirical Bayes and in Causal inference

Abstract
In Part I titled Empirical Bayes Estimation, we discuss the estimation of a heteroscedastic multivariate normal mean in terms of the ensemble risk. We first derive the ensemble minimax properties of various estimators that shrink towards zero through the empirical Bayes method. We then generalize our results to the case where the variances are given as a common unknown but estimable chi-squared random variable scaled by different known factors. We further provide a class of ensemble minimax estimators that shrink towards the common mean. We also make comparison and show differences between results from the heteroscedastic case and those from the homoscedastic model.

In Part II titled Causal Inference Analysis, we study the estimation of the causal effect of treatment on survival probability up to a given time point among those subjects who would comply with the assignment to both treatment and control when both administrative censoring and noncompliance occur. In many clinical studies with a survival outcome, administrative censoring occurs when follow-up ends at a pre-specified date and many subjects are still alive. An additional complication in some trials is that there is noncompliance with the assigned treatment. We first discuss the standard instrumental variable method for survival outcomes and parametric maximum likelihood methods, and then develop an efficient plug-in nonparametric empirical maximum likelihood estimation (PNEMLE) approach. The PNEMLE method does not make any assumptions on outcome distributions, and makes use of the mixture structure in the data to gain efficiency over the standard instrumental variable method. Theoretical results of the PNEMLE are derived and the method is illustrated by an analysis of data from a breast cancer screening trial. From our limited mortality analysis with administrative censoring times 10 years into the follow-up, we find a significant benefit of screening is present after 4 years (at the 5% level) and this persists at 10 years follow-up.

Degree Type
Dissertation

Degree Name
Doctor of Philosophy (PhD)

Graduate Group
Statistics

First Advisor
Lawrence Brown

Subject Categories
Biometry | Statistical Theory | Statistics and Probability

STATISTICAL ANALYSIS IN EMPIRICAL BAYES AND IN CAUSAL INFERENCE

Hui Nie

A DISSERTATION

in

Statistics

For the Graduate Group in Managerial Science and Applied Economics

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy

2011

Supervisor of Dissertation
Lawrence Brown, Professor of Statistics

Graduate Group Chairperson
Eric Bradlow, Professor of Marketing, Statistics and Education

Dissertation Committee
Lawrence Brown, Professor of Statistics
Edward George, Professor of Statistics
Dylan Small, Associate Professor of Statistics
Linda Zhao, Professor of Statistics

STATISTICAL ANALYSIS IN EMPIRICAL BAYES AND IN CAUSAL INFERENCE

Dedicated to

my parents Daosheng Nie and Zengmin Liang

my wife Wenting Zhu

Acknowledgements

First and foremost, I would like to offer my deepest gratitude to my advisor, Professor Lawrence Brown. It is my fortune to have Professor Brown as my academic father. As a student of Professor Brown, I receive careful guidance and warm mentoring from him. He leads me to the castle of Statistics and helps me explore the treasure inside it. From his encouragement, I feel that he is always on my back and supports me whenever needed. Besides research, he also cares very much about my personal life at the graduate school. Without him, I will not have such a wonderful experience at Penn.

I would also like to show my deep appreciation for all the faculty members, staffs and colleagues in the Statistics Department at University of Pennsylvania. They make my Ph.D. life enjoyable. Especially, I would like to thank my committee members – Professor Edward George, Professor Dylan Small and Professor Linda Zhao – for their invaluable helps in my graduate study. I want to show my special thank to Professor Edward George, who encourages me with my research all the time. I am also extremely grateful for Professor Dylan Small, who introduces me the

word of causal inference and is always very supportive in my graduate study. I also show my appreciation for Professor Linda Zhao, who recruits me to the department and gives me lots of helps in the past five years.

Lastly and most importantly, I would like to thank my parents Daosheng Nie and Zengmin Liang as well as my wife Wenting Zhu for their constant support and love. I dedicate this thesis to them.

ABSTRACT

STATISTICAL ANALYSIS IN EMPIRICAL BAYES AND IN CAUSAL INFERENCE

Hui Nie

Lawrence Brown (Advisor)

In Part I titled Empirical Bayes Estimation, we discuss the estimation of a heteroscedastic multivariate normal mean in terms of the ensemble risk. We first derive the ensemble minimax properties of various estimators that shrink towards zero through the empirical Bayes method. We then generalize our results to the case where the variances are given as a common unknown but estimable chi-squared random variable scaled by different known factors. We further provide a class of ensemble minimax estimators that shrink towards the common mean. We also make comparison and show differences between results from the heteroscedastic case and those from the homoscedastic model.

In Part II titled Causal Inference Analysis, we study the estimation of the causal effect of treatment on survival probability up to a given time point among those subjects who would comply with the assignment to both treatment and control when both administrative censoring and noncompliance occur. In many clinical

studies with a survival outcome, administrative censoring occurs when follow-up ends at a pre-specified date and many subjects are still alive. An additional complication in some trials is that there is noncompliance with the assigned treatment. We first discuss the standard instrumental variable method for survival outcomes and parametric maximum likelihood methods, and then develop an efficient plug-in nonparametric empirical maximum likelihood estimation (PNEMLE) approach. The PNEMLE method does not make any assumptions on outcome distributions, and makes use of the mixture structure in the data to gain efficiency over the standard instrumental variable method. Theoretical results of the PNEMLE are derived and the method is illustrated by an analysis of data from a breast cancer screening trial. From our limited mortality analysis with administrative censoring times 10 years into the follow-up, we find a significant benefit of screening is present after 4 years (at the 5% level) and this persists at 10 years follow-up.

Contents

List of Tables

List of Figures

Part I: Empirical Bayes Estimation

Ensemble Minimax Estimation for Multivariate Normal Means

Chapter 1

Introduction and Background Knowledge

1.1 Estimation of Multivariate Normal Mean

We consider the problem of simultaneously estimating the mean parameter $\theta = (\theta_1, \cdots, \theta_p)$ from independent normal observations $X \sim N(\theta, \Sigma)$, where Σ is a diagonal matrix with the elements $\{\sigma_1^2, \cdots, \sigma_p^2\}$. For any estimator $\hat{\theta}$, our loss function is the ordinary squared error loss

$$L(\hat{\theta}, \theta) = \sum_{i=1}^{p} (\hat{\theta}_i - \theta_i)^2.$$

The conventional risk function is the expected value of the loss function with respect to θ. That is,

$$R(\theta, \hat{\theta}) = E_\theta(L(\hat{\theta}, \theta)) = \sum_{i=1}^{p} E_\theta(\hat{\theta}_i - \theta_i)^2.$$

James and Stein (1961) study the homoscedastic case in which $\sigma^2 = \sigma_1^2 = \cdots = \sigma_p^2$. In that case they prove the astonishing result that the James-Stein shrinkage estimator

$$\delta_{J-S}(X) = \left(1 - \frac{C\sigma^2}{\|X\|^2}\right) X \tag{1.1.1}$$

and its positive part

$$\delta_{J-S}^{+}(X) = \left(1 - \frac{C\sigma^2}{\|X\|^2}\right)_{+} X \tag{1.1.2}$$

dominate the usual MLE $\delta_0(X) = X$ for $0 \leq C \leq 2(p-2)$ and $p \geq 3$. The discovery by James and Stein has led to a wide application of shrinkage techniques in many important problems. References include Efron and Morris (1975), Fay and Herriot (1979), Rubin (1981), Morris (1983a), Green and Strawderman (1985), Jones (1991) and Brown (2008). The theoretical properties of shrinkage estimators have also been extensively studied in the literature under the homoscedastic Gaussian model. Since Stein's discovery, "shrinkage" has been developed as a broad statistical framework in many aspects (Stein, 1962; Strawderman, 1971; Efron and Morris, 1971,1972a,1972b and 1973; Casella, 1980; Hastie et al., 2003).

There is also some literature discussing the properties of the James-Stein shrinkage estimators under heteroscedasticity. James and Stein (1961) discuss the estimation problem under heteroscedasticity where the loss function is weighted by the inverse of the variances. This problem can be transformed to the homoscedastic case under ordinary squared error loss. Brown (1975) shows that the James-Stein estimator is not always minimax and hence does not necessarily dominate the usual

MLE under ordinary squared error loss when the variances are not equal. Specifically, the James-Stein shrinkage estimator does not dominate the usual MLE when the largest variance is larger than the sum of the rest. Moreover, Casella (1980) argues that the James-Stein shrinkage estimator may not be a desirable shrinkage estimator under heteroscedasticity even if it is minimax. Minimax estimators in general shrink most on the coordinates with smaller variances, while Bayes estimators shrink most on large variance coordinates.

In many applications, θ_i are thought to follow some exchangeable prior distribution π. It is then natural to consider the compound risk function which is then the Bayes risk with respect to the prior π

$$\overline{R}(\pi, \hat{\theta}) = E_\pi(R(\theta, \hat{\theta})) = \int R(\theta, \hat{\theta})\pi(d\theta).$$

Efron and Morris (1971, 1972a, 1972b and 1973) address this problem from both the Bayes and empirical Bayes perspective. They extensively develop this framework. Especially, they consider a prior distribution of the form $\theta \sim N_p(0, \tau^2 I)$ with $\tau^2 \in [0, \infty)$, and they use the term "ensemble risk" for the compound risk. Morris and Lysy (2009) discuss the motivation and importance of shrinkage estimation in this multi-level normal model. The ensemble risk is described as the level-II risk in Morris and Lysy (2009).

By introducing a set of ensemble risks $\overline{R}(\pi, \hat{\theta})$ ($\pi \in \mathcal{P}$), we can then discuss ensemble minimaxity and other properties with respect to a set of prior distributions $\mathcal{P}$. We elaborate the definitions of ensemble minimaxity and other properties in

Chapter 2. The previously cited papers (and others) discuss the desirability of the ensemble risks with respect to the normal priors $\theta \sim N_p(0, \tau^2 I)$ with $\tau^2 \in [0, \infty)$. In this paper, we will concentrate on the ensemble minimaxity of various estimators in this respect.

Brown (2008) discusses the connection between the parametric empirical Bayes estimator and the random effects model. In fact, the estimation problem of group means in a one way random effects model with infinite degrees of freedom for errors (and hence known error variance) is equivalent to the above problem. Our ensemble risk then corresponds to the ordinary risk function in the random effects model.

The more familiar unbalanced one-way random effects model is exactly equivalent to the generalization considered in Section 4.2. Again, ensemble risk in the empirical Bayes sense corresponds to ordinary prediction risk for the random effects model. We close Section 4.2 with a summary statement describing estimators proven to dominate the ordinary least squares group means in the random effect model.

1.2 Empirical Bayes Method

Empirical Bayes method is a powerful tool in statistical decision theory. It has a very long history. Robbins (1951) propose a subminimax decision rule for the classification problem through empirical Bayes perspective. The key idea of empirical Bayes method is that the underlying relationships among the coordinates of the

parameters allow use of the observed data to estimate some features of the prior distribution. The most obvious empirical Bayes problems are when the parameters arise from some common population so that we can imagine creating a probabilistic model for the population and also interpret this model as the prior distribution. The simplest version of this situation is when the parameters are i.i.d. from some prior distribution. Random effects models are one of this types of problems. Although it uses the Bayesian idea, from the modeling perspective, empirical Bayes problems can be considered to be problems in classical statistics.

Empirical Bayes methods can be categorized in two different ways. One division is between parametric empirical Bayes and nonparametric empirical Bayes. In the former, one assumes that the prior distribution of the parameters is in some parametric class with unknown hyperparameters. In the latter, one assumes only that the parameters are i.i.d. A different categorization of empirical Bayes analysis is given by its operational focus. The most natural focus is to use the data to estimate the prior information or the posterior distribution. The subsequent analysis turns to a typical Bayesian fashion once this is done. The other possible operational focus is to represent the Bayes rule in terms of the unknown prior, and then use the data to estimate the Bayes rule directly.

There is lots of previous literature discussing the application of empirical Bayes methods in the theoretical estimation problems of multivariate normal means. Efron and Morris (1971, 1972a, 1972b and 1973) give an interpretation of Stein's (1962) as-

tonishing results through Bayes and parametric empirical Bayes perspective. They point out that the James-Stein estimator is an empirical Bayes estimator where we use the observed data to estimate the Bayes rule, and they propose a class of minimax empirical Bayes estimators that shrink towards common means instead of zero. Strawderman (1971) proposes a class of minimax Bayes estimators with respect to the harmonic prior. Casella (1980) derives conditions that are necessary and sufficient for minimaxity of a large class of ridge regression estimators.

The discovery of the theoretical knowledge has also led to a wide application of empirical Bayes methods in many important problems. Carter and Rolph (1974) use empirical Bayes methods to estimate fire alarm possibilities. Efron and Morris (1975) and Brown (2008) predict the batting averages of players in Major League Baseball via empirical Bayes methods. Fay and Herriot (1979) estimate the income for small places. Rubin (1981) apply empirical Bayes techniques in the law school validity studies. Jones (1985) studies the house-price variation in Southampton. Green and Strawderman (1985) use empirical Bayes estimation in individual tree volume equation development.

In our paper, we use the idea of empirical Bayes methods to construct various shrinkage estimators that have better properties than the usual maximum likelihood estimator. Our judgement criteria is the "ensemble risk", which will be defined in the following chapter.

Chapter 2

Definition of Ensemble Minimaxity

As discussed above, we study in this paper the behavior of shrinkage estimators based on the ensemble risk

$$\overline{R}(\pi, \hat{\theta}) = E_\pi(R(\theta, \hat{\theta})) = \int R(\theta, \hat{\theta})\pi(d\theta) \ .$$

If the prior $\pi(\theta)$ is known, the resulting posterior mean $E_\pi(\theta|x)$ is then the optimal estimate under the sum of the squared error loss. However, it is often infeasible to exactly specify the prior. To avoid excessive dependence on the choice of prior, it is natural to consider a set of priors $\mathcal{P}$ on θ and study the properties of various estimators based on the corresponding set of ensemble risks. As in the classic decision theory, there rarely exists an estimator that achieves the minimum ensemble risk uniformly for all $\pi \in \mathcal{P}$. A more realistic goal as pursued in this paper is to

study the ensemble minimaxity (defined shortly) of familiar shrinkage estimators.

Recall that with ordinary risk $R(\theta, \delta)$, an estimator δ is said to dominate another estimator δ' if

$$R(\theta, \delta) \leq R(\theta, \delta^{'})$$

holds for each $\theta \in \Theta$ with strict inequality for at least one θ. The estimator δ is inadmissible if there exists another procedure which dominates δ; otherwise δ is admissible. δ is said to be minimax if

$$\sup_{\theta \in \Theta} R(\theta, \delta) = \inf_{\delta^{'}} \sup_{\theta \in \Theta} R(\theta, \delta^{'}) \ ,$$

that is, the estimator attains the minimum worst-case risk. Similarly for the case of ensemble risk we have the following definitions.

Ensemble admissibility and minimaxity. Given a set of priors $\mathcal{P}$, an estimator δ is said to dominate another estimator δ' with respect to $\mathcal{P}$ if

$$\overline{R}(\pi, \delta) \leq \overline{R}(\pi, \delta^{'})$$

holds for each $\pi \in \mathcal{P}$ with strict inequality for at least one π. The estimator δ is ensemble inadmissible with respect to $\mathcal{P}$ if there exists another procedure which dominates δ, otherwise δ is ensemble admissible. The estimator δ is ensemble minimax with respect to $\mathcal{P}$ if

$$\sup_{\pi \in \mathcal{P}} \overline{R}(\pi, \delta) = \inf_{\delta^{'}} \sup_{\pi \in \mathcal{P}} \overline{R}(\pi, \delta^{'}) \ .$$

The motivation for the above definitions comes from the use of the empirical Bayes methods in simultaneous inference. Efron and Morris (1972a), building from

Stein (1962), derive the James-Stein estimator through the parametric empirical Bayes model with $\theta_i \sim N(0, \tau^2)$. Note that in such an empirical Bayes model, τ^2 is the unknown parameter. (Parameter here refers to an unknown non-random quantity.) Ensemble admissibility and minimaxity with respect to $\mathcal{P} = \{\theta_i \sim N(0, \tau^2) : 0 < \tau^2 < \infty\}$ is then exactly the counterpart of ordinary admissibility and minimaxity in the empirical Bayes model. Consistent with this, we also confine $\mathcal{P}$ to be the one given above. Another reason for preferring such a set $\mathcal{P}$ is because it enjoys the conjugate minimaxity property (Morris, 1983a). From now on, mention of this underlying set $\mathcal{P}$ will be omitted whenever confusion is unlikely. As an explicit notation in this setting, we define $\overline{R}_{\tau^2}(\delta) = \overline{R}(\pi, \delta)$ for $\pi = N(0, \tau^2)$.

Note that ensemble minimaxity can also be interpreted as a particular case of Gamma minimaxity studied in the context of robust Bayes analysis (Good, 1952; Berger, 1979). However, in such studies, a "large" set consisting of many diffuse priors are usually included in the analysis. Since this is quite different from our formulation of the problem, we use the term ensemble minimaxity throughout our paper, following the Efron and Morris papers cited above.

Chapter 3

Main Results on Ensemble Minimaxity

In this chapter, we discuss the ensemble minimaxity of various shrinkage estimators. We first present a general theorem characterizing a class of shrinkage estimators that are ensemble minimax. We then study the ensemble minimaxity of James-Stein-type shrinkage estimators, along with several supplementary theorems highlighting the difference and similarity between our results and those obtained in the homoscedastic case. Finally, we investigate the ensemble minimaxity of the parametric empirical Bayes estimator via method of moment estimation, a case with several open problems unresolved during our study. Throughout the current discussion, the variances σ_i^2 are assumed to be known; the case of unknown σ_i^2 is addressed in the next chapter.

3.1 General Theory

As discussed in Section 1.1, when $p \geq 3$ and $0 \leq C \leq 2(p-2)$, both δ_{J-S} in (1.1.1) and δ^{+}_{J-S} in (1.1.2) are known to be minimax under the homoscedastic model. However, this is not always the case under the heteroscedastic model. Brown (1975) shows that for any $C > 0$, if $\sum \sigma_i^2 \leq 2\max\{\sigma_i^2\}$, both δ_{J-S} in (3.1.1) and δ^{+}_{J-S} in (3.1.2) are no longer minimax in the ordinary sense. This is one motivation for instead studying the ensemble minimaxity for these shrinkage estimators. The following theorem shows that $\delta_0(X) = X$ is ensemble minimax with respect to $\mathcal{P}$.

Theorem 1. *δ_0 is ensemble minimax with respect to $\mathcal{P}$.*

Proof. See Chapter 6. □

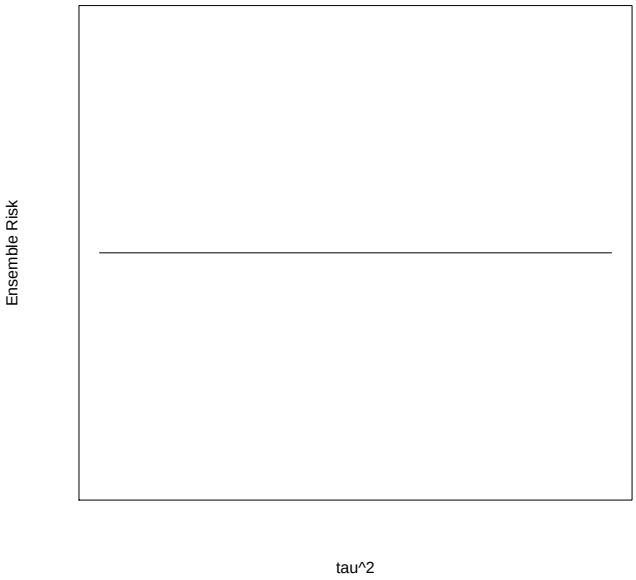
As we can see in Figure 3.1, δ_0 has a constant ensemble risk for any prior in $\mathcal{P}$. Although δ_0 is ensemble minimax, it is possible to construct a class of ensemble minimax estimators that dominates δ_0 with respect to $\mathcal{P}$. Before presenting our main result, we first give a lemma concerning the evaluation of the ensemble risk $\overline{R}_{\tau^2}(\delta)$ that will be repeatedly used in subsequent discussion.

Lemma 1. *The ensemble risk of any estimator δ with the form $\delta_i(X) = (1 - h_i(X))X_i$ can be written as*

$$\overline{R}_{\tau^2}(\delta) = \sum_{i=1}^{p} \left[E_x \left(\frac{\sigma_i^2}{\tau^2 + \sigma_i^2} X_i - h_i(X) X_i \right)^2 + \frac{\tau^2 \sigma_i^2}{\tau^2 + \sigma_i^2} \right] ,$$

where the expectation is taken with respect to the joint distribution of X such that $X_i \sim N(0, \tau^2 + \sigma_i^2)$ and all the coordinates are jointly independent.

Figure 3.1: Ensemble Risk of $\delta_0(X) = X$



Proof. See Chapter 6. □

Under the heteroscedastic model, we define the James-Stein-type estimator δ_{J-S} as

$$(\delta_{J-S}(X))_i = \left(1 - \frac{C\sigma_i^2}{\|X\|^2}\right) X_i \ . \tag{3.1.1}$$

Its positive part δ^+_{J-S} is then

$$(\delta^+_{J-S}(X))_i = \left(1 - \frac{C\sigma_i^2}{\|X\|^2}\right)_+ X_i \ . \tag{3.1.2}$$

Estimators of this general form appear as a generalization of the original James-Stein proposal in Brown (1966). See also Efron and Morris (1971). To study the ensemble minimaxity of these two estimators, we first present a general result that characterizes a class of shrinkage estimators that are ensemble minimax.

The general result refers to estimators with the form $\delta_i(X) = (1 - h_i(X))X_i$, as in Lemma 1, and in which h_i is symmetric in the sense that

$$h_i(X) = \mathfrak{h}_i(X_1^2, \cdots, X_p^2). \tag{3.1.3}$$

In addition, we define

$$W = \sum_{j=1}^{p} \frac{X_j^2}{\tau^2 + \sigma_j^2} \tag{3.1.4}$$

$$T_i = \frac{X_i^2}{W(\tau^2 + \sigma_i^2)},\ i = 1, \cdots, p. \tag{3.1.5}$$

In this way $X_i^2 = (\tau^2 + \sigma_i^2)WT_i$, and $\mathfrak{h}$ can be rewritten as a function of $\underline{T} = (T_1, \cdots, T_p)$ and W. With a minor extension of the notation, write $\mathfrak{h}(\underline{T}, W) = \mathfrak{h}((\tau^2 + \sigma_1^2)WT_1, \cdots, (\tau^2 + \sigma_p^2)WT_p)$.

Theorem 2. *An estimator δ with the form $\delta_i(X) = (1 - h_i(X))X_i$ is ensemble minimax if each shrinkage factor $h_i(X)$ satisfies the following conditions:*

(1) $h_i(X) \geq 0$, $\forall X$.

(2) $h_i(X)$ *can be written in the form (3.1.3).*

(3) $\mathfrak{h}_i(\underline{T}, W)$ *is decreasing in* W *for fixed* $\underline{T}$.

(4) $\mathfrak{h}_i(\underline{T}, W)W$ *is increasing in* W *for fixed* $\underline{T}$.

(5)

$$E\left[\sup_{\underline{T}} \mathfrak{h}_i(\underline{T}, W)\right] \leq \frac{2\sigma_i^2}{\sigma_i^2 + \tau^2}\ .$$

Proof. See Chapter 6. □

Note that most of the conditions in Theorem 2 are rather intuitive to understand. Condition (1) simply means that the estimator is indeed a genuine shrinkage estimator, and never an expander. Condition (2) implies the shrinkage estimator has a certain natural symmetry property. Condition (3) requires the amount of shrinkage to decrease when the distance of the data vector is further away from the origin. Condition (5) controls the expected overall amount of shrinkage according to the ratio of the variability of the observation and that of the prior, but this condition is less intuitive than the others.

Let $\mu \in \mathcal{R}^p$. Consider estimation of the linear combination $\mu^t\theta$ under squared error loss $L_{lc}(d, \theta) = (d - \mu^t\theta)^2$. Ordinary minimaxity and ensemble minimaxity can be defined for this loss. As a Corollary to Theorem 2, we have

Corollary 1. *Assume conditions (1)-(5) of Theorem 2. Then the estimator* $\hat{\eta} = \mu^t\delta$ *is an ensemble minimax estimator of* $\eta = \mu^t\theta$.

Proof. From the proof of Theorem 2, we see that ensemble minimaxity is actually achieved for each coordinate, that is, for any $i = 1, \cdots, p$,

$$\overline{R}_{lc}(\delta_i, \theta_i) \leq \sigma_i^2 \ .$$

This proves validity of the corollary. □

3.2 James-Stein-type Shrinkage Estimators

With Theorem 2, we then proceed to study the ensemble minimaxity of certain shrinkage estimators. These estimators include the original James-Stein estimators. As we will show, the original James-Stein estimator is often but not always ensemble minimax. Consider the estimator δ_{GS} with the form

$$(\delta_{GS}(X))_i = \left(1 - \frac{\lambda_i \sigma_i^2}{\nu_i \sigma_i^2 + \|X\|^2}\right) X_i \ , \tag{3.2.1}$$

where λ_i and ν_i are properly chosen constants. Consider also its positive part version δ_{GS}^+ given by

$$(\delta_{GS}^+(X))_i = \left(1 - \frac{\lambda_i \sigma_i^2}{\nu_i \sigma_i^2 + \|X\|^2}\right)_+ X_i \ . \tag{3.2.2}$$

Note that these forms are generalizations of the original James-Stein forms, as can be seen by setting $\nu_i = 0$, $\lambda_i = C$.

The following two corollaries state conditions under which δ_{GS} in (3.2.1) and δ_{GS}^+ in (3.2.2) are ensemble minimax.

Corollary 2. *δ_{GS} in (3.2.1) is ensemble minimax if $p \geq 3$ and for any $i = 1, \cdots, p$, $0 \leq \lambda_i \leq 2(p-2)$ and $\nu_i \geq (\lambda_i/2 - (p-2) \cdot \sigma_{min}^2/\sigma_i^2)_+$ with $\sigma_{min}^2 = \min_i\{\sigma_i^2\}$.*

Proof. See Chapter 6. □

Corollary 3. *δ_{GS}^+ in (3.2.2) is ensemble minimax if $p \geq 3$ and for any $i = 1, \cdots, p$, $0 \leq \lambda_i \leq 2(p-2)$ and $\nu_i \geq [\lambda_i - (p-2)(1 + \sigma_{min}^2/\sigma_i^2)]_+$ with $\sigma_{min}^2 = \min_i\{\sigma_i^2\}$.*

Proof. See Chapter 6. □

Remarks: When $\nu_i = 0$ and $\lambda_i = C$, δ_{GS} in (3.2.1) and δ^+_{GS} in (3.2.2) reduce to the James-Stein estimators in (3.1.1) and (3.1.2). In the case where σ_i^2 are all equal, Corollary 2 and 3 show that δ_{J-S} in (3.1.1) and δ^+_{J-S} in (3.1.2) are each ensemble minimax when $0 \leq C \leq 2(p-2)$. This reaches the same conclusion as James and Stein (1961).

When the values of σ_i^2 are not all equal, the results in Corollary 2 and 3 do not always establish ensemble minimaxity of δ_{J-S} in (3.1.1) and δ^+_{J-S} in (3.1.2) for the entire range $0 \leq C \leq 2(p-2)$. Specializing the conditions of Corollary 2 to the case where $\lambda_i = C$ and $\nu_i = 0$ yields that δ_{J-S} in (3.1.1) is ensemble minimax if

$$C \leq 2(p-2)\frac{\sigma^2_{min}}{\sigma^2_{max}}. \tag{3.2.3}$$

Thus, for any $C > 0$, there are configurations of $\sigma_1^2, \cdots, \sigma_p^2$ for which the conditions in Corollary 2 fail to prove δ_{J-S} in (3.1.1) is ensemble minimax.

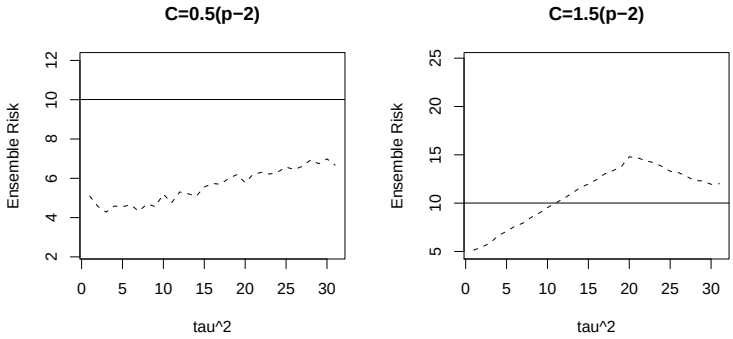
For δ^+_{J-S} in (3.1.2), the situation is a little different. Specializing the conditions of Corollary 3 to the case $\lambda_i = C$ and $\nu_i = 0$ yields that δ^+_{J-S} in (3.1.2) is ensemble minimax if

$$C \leq (p-2)(1 + \frac{\sigma^2_{min}}{\sigma^2_{max}}). \tag{3.2.4}$$

Hence, when $C \leq p-2$, the conditions in Corollary 3 are always satisfied by δ^+_{J-S} in (3.1.2). However, for any $C > p-2$, there are configurations of $\sigma_1^2, \cdots, \sigma_p^2$ for which Corollary 3 fails to prove δ^+_{J-S} in (3.1.2) is minimax.

Theorem 3 and 4 below address ensemble minimaxity of δ_{J-S} in (3.1.1) when $C > 0$ and δ^+_{J-S} in (3.1.2) when $C > p-2$. They state conditions under which

Figure 3.2: Ensemble Risk of δ^+_{J-S} (Dash Line) and $\delta_0(X) = X$ (Solid Line)



δ_{J-S} in (3.1.1) and δ^+_{J-S} in (3.1.2) can fail to be ensemble minimax when $C > 0$ or $C > p-2$, respectively. There is a gap between the conditions in Corollaries 2 and 3. We do not as yet have a formulation of a sharp necessary and sufficient condition for ensemble minimaxity of δ_{J-S} in (3.1.1) and δ^+_{J-S} in (3.1.2) in the case of general $\sigma_1^2, \cdots, \sigma_p^2$.

Theorem 3. *For any $C = c(p-2)$ with $c > 1$, there exists some sufficiently large $p \geq 3$ and some $\sigma_1^2, \cdots, \sigma_p^2$ such that δ^+_{J-S} in (3.1.2) is not ensemble minimax.*

Proof. See Chapter 6. □

Figure 3.2 shows a specific example of the relationship between the ensemble risk of δ^+_{J-S} and that of $\delta_0(X) = X$. Let $p = 1000$, $\sigma_1^2 = 10$ and $\sigma_2^2 = \cdots = \sigma_p^2 = 0.000001$. On the left panel of Figure 3.2, we can find that when $C = 0.5(p-2)$, the ensemble risk of δ^+_{J-S} is smaller than that of δ_0 for any τ^2. On the right panel of Figure 3.2, however, when $C = 1.5(p-2)$, δ^+_{J-S} does not always dominate δ_0.

The above results show that for δ^+_{J-S} in (3.1.2) to be ensemble minimax, the

constant C must have a much smaller upper bound under the general heteroscedastic model than under the homoscedastic one. Furthermore, it is proved below that δ_{J-S} in (3.1.1) is not even always ensemble minimax regardless of the choice of C.

Theorem 4. *For any $C > 0$, there exists some $\sigma_1^2,\cdots,\sigma_p^2$ such that δ_{J-S} in (3.1.1) is not ensemble minimax.*

Proof. See Chapter 6. □

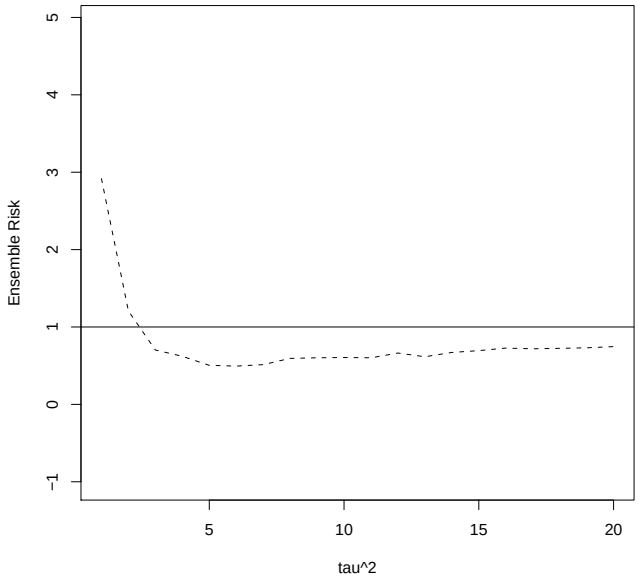
The fact that δ_{J-S} in (3.1.1) is not in general ensemble minimax may be quite surprising at first glance. However, this is to be expected given the form of δ_{J-S}. Under the heteroscedastic model, δ_{J-S} may have a non-negligible probability of dramatically over-shrinking, which causes the performance of the shrinkage estimator to deteriorate; while such an issue is always well controlled in the homoscedastic case.

Figure 3.3 shows a specific example of the relationship between the ensemble risk of δ_{J-S} and that of $\delta_0(X) = X$. Here we set $p = 10$, $\sigma_1^2 = 1$, $\sigma_2^2 = \cdots = \sigma_p^2 = 0.000001$ and $C = 5$. We find that δ_{J-S} does not always dominate δ_0.

Similar to the homoscedastic case, δ_{J-S}^+ in (3.1.2) is always better than δ_{J-S} in (3.1.1) in terms of ensemble risk under the heteroscedastic model. A general result is given in the following theorem.

Theorem 5. *Let δ be any estimator with the form $\delta_i(X) = (1 - h_i(X))X_i$ and δ^+ be its positive part estimator such that $\delta_i^+(X) = (1 - h_i(X))_+X_i$. If $P(\delta \neq \delta^+) > 0$, then δ^+ dominates δ with respect to $\mathcal{P}$.*

Figure 3.3: Ensemble Risk of δ_{J-S} (Dash Line) and $\delta_0(X) = X$ (Solid Line)



Proof. See Chapter 6. □

Corollary 4. *For any constant $C \geq 0$, δ^+_{J-S} in (3.1.2) dominates δ_{J-S} in (3.1.1) with respect to $\mathcal{P}$.*

Proof. Directly from Theorem 5. □

3.3 Parametric Empirical Bayes Estimators

Carter and Rolph (1974), Brown (2008) and Efron and Morris (1973, 1975) each derive parametric empirical Bayes estimators for the heteroscedastic problem. The first two papers use method of moments to estimate the hyperparameter τ^2. (Morris and Lysy (2009) also discuss such estimators.) We will discuss here ensemble

minimaxity of such empirical Bayes estimators.

In contrast, Efron and Morris (1973, 1975) use a maximum likelihood method for this step. The resulting estimation does not have an explicit closed form, although it is easily calculated numerically. For this reason we have (so far) been less successful in settling the ensemble minimaxity of this empirical Bayes version, and we do not address this issue here.

In this subsection, we treat the special case of shrinkage to 0. The previously cited references (and others) involve shrinkage to a common mean. This generalization is treated in Section 4.2. While our results in the present subsection shed some light on the ensemble minimaxity of these estimators, they are unfortunately not as nearly complete as our preceding results about generalized James-Stein estimators.

As mentioned above, if τ^2 is known, the optimal estimator of θ_i $(i = 1, \cdots, p)$ would be

$$(\delta_B(X))_i = \left(1 - \frac{\sigma_i^2}{\tau^2 + \sigma_i^2}\right) X_i \ , \tag{3.3.1}$$

which is the Bayes estimator. However in the empirical Bayes setting, τ^2 is an unknown hyperparameter to be estimated. The idea of the parametric empirical Bayes method is to use $\{X_i\}$ to obtain an estimate of τ^2 and then substitute the estimate of τ^2 into (3.3.1) to yield a final estimator of $\{\theta_i\}$. Below we use the method of moments estimator

$$\widetilde{\tau}^2 = \frac{1}{p}\sum_{i=1}^{p}(X_i^2 - \sigma_i^2),$$

and its positive part

$$\widetilde{\tau}_+^2 = \frac{1}{p}\left[\sum_{i=1}^{p}(X_i^2 - \sigma_i^2)\right]_+ .$$

In practice, some other constant "$1/C$" is oftens used in lieu of "$1/p$" above.

The corresponding parametric empirical Bayes estimator is then given by

$$(\delta_{PEB}(X))_i = \left(1 - \frac{\sigma_i^2}{\sigma_i^2 + \frac{1}{C}\sum_{j=1}^{p}(X_j^2 - \sigma_j^2)}\right) X_i \ , \tag{3.3.2}$$

along with its positive part estimator

$$(\delta_{PEB}^+(X))_i = \left(1 - \frac{\sigma_i^2}{\sigma_i^2 + \frac{1}{C}(\sum_{j=1}^{p}(X_j^2 - \sigma_j^2))_+}\right) X_i \ . \tag{3.3.3}$$

Note that the form of the parametric empirical Bayes estimator δ_{PEB} in (3.3.2) differs from the James-Stein-type estimator δ_{J-S} in (3.1.1) in the use of the term $C\sigma_i^2 + \sum_{j=1}^{p}(X_j^2 - \sigma_j^2)$ instead of $\sum_{j=1}^{p} X_j^2$ in the denominator. Therefore the former denominator can be much smaller than the latter and hence lead to over-shrinkage. Not surprisingly, the conditions needed for ensemble minimaxity appear somewhat more restrictive than in the James-Stein case.

The following corollary contains conditions that guarantee ensemble minimaxity for the parametric empirical Bayes estimators. Simulation results (not reported here) lead us to conjecture that ensemble minimaxity holds under somewhat less restrictive conditions.

Corollary 5. *Assume*

$$p \le \frac{\sum_{j=1}^{p}\sigma_j^2}{\sigma_{min}^2} \le C \le 2(p-2) \ . \tag{3.3.4}$$

Then both δ_{PEB} in (3.3.2) and δ_{PEB}^+ in (3.3.3) are ensemble minimax.

Remark: In the homoscedastic case, Condition (3.3.4) requires $p \leq 2(p-2)$. This is satisfied if and only if $p \geq 4$. In that case, δ_{PEB} in (3.3.2) and δ^+_{PEB} in (3.3.3) are ensemble minimax if $4 \leq p \leq C \leq 2(p-2)$.

Proof. See Chapter 6. □

Theorem 6. *Let $p \geq 1$ and $C > 0$. Then there exists some $\sigma_1^2, \cdots, \sigma_p^2$ such that δ_{PEB} in (3.3.2) is not ensemble minimax.*

Proof. See Chapter 6. □

Unfortunately, we have been unable to obtain a complete answer on the ensemble minimaxity of the positive part estimators δ^+_{PEB} in (3.3.3). Nevertheless, the following theorem indicates that, unlike in the case of James-Stein-type estimators, C can not be too small.

Theorem 7. *For any p, there exists some sufficiently small C and some $\sigma_1^2, \cdots, \sigma_p^2$ such that δ^+_{PEB} in (3.3.3) is not ensemble minimax.*

Proof. See Chapter 6. □

One interesting observation here is that as $C \to 0$, δ^+_{PEB} reduces to the hard-threshold estimator $X1_{\{\|X\|^2 > p\sigma^2\}}$ under the homoscedastic model. The above theorem simply indicates that the hard-threshold estimator is worse than the ordinary MLE in terms of ensemble risk when $\tau^2 > \sigma^2$.

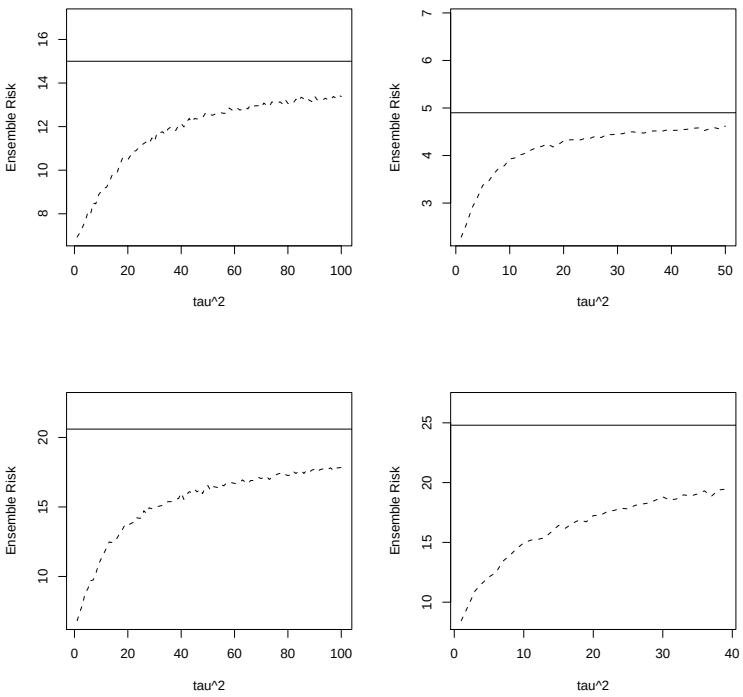
Similar to the James-Stein estimator, δ^+_{PEB} in (3.3.3) is better than δ_{PEB} in (3.3.2) in terms of ensemble risk.

Corollary 6. *For any constant $C \geq 0$, δ^+_{PEB} in (3.3.3) dominates δ_{PEB} in (3.3.2) with respect to $\mathcal{P}$.*

Proof. Directly from Theorem 5. □

Figure 3.4 shows the relationship between the ensemble risk of δ^+_{PEB} and that of $\delta_0(X) = X$ in various simulations. On the upper left graph, we have $p = 5$, $\Sigma = \text{diag}\{1, 2, 3, 4, 5\}$ and $C = 2$. On the upper right graph, we have $p = 7$, $\Sigma = \text{diag}\{2.1, 0.5, 1.2, 0.8, 0.3\}$ and $C = 1$. On the lower left graph, we have $p = 10$, $\Sigma = \text{diag}\{2.9, 1.3, 1.1, 1.6, 2.9, 1.4, 2.5, 2.0, 3.6, 1.3\}$ and $C = 9$. On the lower right graph, we have $p = 14$, $\Sigma = \text{diag}\{0.9, 1.4, 1.8, 1.0, 1.8, 2.5, 0.3, 2.3, 3.4, 2.9, 0.8, 2.8, 2.2, 0.7\}$ and $C = 20$. Although simulation results lend support to the conjecture that δ^+_{PEB} in (3.3.3) is ensemble minimax when $1 \leq C \leq 2(p-2)$ (the lower bound could be much smaller) and $p \geq 3$, a rigorous proof is still yet to be found.

Figure 3.4: Ensemble Risk of δ^{+}_{PEB} (Dash Line) and $\delta_0(X) = X$ (Solid Line)



Chapter 4

Generalization to Other Cases

4.1 Unknown Variances Case

The discussion so far has been focused on the ensemble minimaxity of shrinkage estimators assuming the variances σ_i^2 to be known. It is common in many circumstances that variances are unknown and have to be estimated from data. Here we consider the case where $X_i \sim N(\theta_i, \sigma^2\gamma_i)$ for $i = 1, \cdots, p$ with unknown σ^2 but known γ_i. Denote $\Gamma = diag\{\gamma_1, \cdots, \gamma_p\}$. We also assume that σ^2 is estimated by $M \sim \sigma^2\chi_m^2/m$ where M is independent of X, an assumption which is satisfied in applications in which a pooled estimate of σ^2 is used. In particular, this setting corresponds to the one-way random effects setting of Section 4.2 with $\gamma_i = 1/J_i$ where J_i is the number of observations in group i. We will discuss the ensemble minimaxity of some shrinkage estimators. First of all, we give two lemmas that will

be used in our later proof. The first one is the generalization of Lemma 1 to the unknown variances case.

Lemma 2. *The ensemble risk of any estimator δ with the form $\delta_i(X, M) = (1 - h_i(X, M))X_i$ has the following representation*

$$\overline{R}_{\tau^2}(\delta) = \sum_{i=1}^{p} E\left[\left(\frac{\sigma^2\gamma_i}{\tau^2 + \sigma^2\gamma_i}X_i - h_i(X, M)X_i\right)^2 + \frac{\tau^2\sigma^2\gamma_i}{\tau^2 + \sigma^2\gamma_i}\right] , \tag{4.1.1}$$

where the expectation is taken with respect to the joint distribution of (X, M) where each $X_i \sim N(0, \tau^2 + \sigma^2\gamma_i)$ and $M \sim \sigma^2\chi^2_m/m$, and they are jointly independent.

Proof. See Chapter 6. □

The second lemma is an inequality concerning expectations of non-negative random variables.

Lemma 3. *For a non-negative random variable M and two non-negative functions $\mu(M)$ and $\mu'(M)$, if the ratio $r(M) = \mu(M)/\mu'(M)$ is non-decreasing in M, we then have*

$$\frac{E(M\mu(M))}{E(\mu(M))} \geq \frac{E(M\mu'(M))}{E(\mu'(M))}$$

assuming all expectations are finite and non-zero.

Proof. See Chapter 6. □

Define δ_{GSV} with the form

$$(\delta_{GSV}(X))_i = \left(1 - \frac{\lambda_i M\gamma_i}{\nu_i M\gamma_i + \|X\|^2}\right) X_i \ . \tag{4.1.2}$$

We have the following theorem characterizing the ensemble minimaxity of $\delta_{GSV}(X)$ in (4.1.2). The upper bound for λ_i is slightly smaller since we are now estimating σ_i^2, a phenomenon observed in similar studies under the homoscedastic model.

Theorem 8. *δ_{GSV} in (4.1.2) is ensemble minimax if $p \geq 3$, $m \geq 3$ and for any $i = 1, \cdots, p$, $0 \leq \lambda_i \leq \frac{2m(p-2)}{m+2}$ and $\nu_i \geq (\frac{m+2}{2(m-2)}\lambda_i - \frac{m\gamma_{min}(p-2)}{\gamma_i(m-2)})_+$.*

Proof. See Chapter 6. □

For the corresponding positive part estimator δ_{GSV}^+ given by

$$(\delta_{GSV}^+(X))_i = \left(1 - \frac{\lambda_i M\gamma_i}{\nu_i M\gamma_i + \|X\|^2}\right)_+ X_i \ , \tag{4.1.3}$$

as in the case of known variance, a slightly stronger result holds.

Theorem 9. *δ_{GSV}^+ in (4.1.3) is ensemble minimax if $p \geq 3$, $m \geq 3$ and for any $i = 1, \cdots, p$, $0 \leq \lambda_i \leq \frac{2m(p-2)}{m+2}$ and $\nu_i \geq (\frac{m+2}{m-2}\lambda_i - \frac{m(p-2)}{m-2}(1 + \frac{\gamma_{min}}{\gamma_i}))_+$.*

Proof. See Chapter 6. □

When $\lambda_i = C$ and $\nu_i = 0$, δ_{GSV} in (4.1.2) and δ_{GSV}^+ in (4.1.3) reduce to the James-Stein estimator and its positive part for the unknown variance case. Similar to the known variances case, the choice of C in the above theorems is different from that in the homoscedastic case. For the homoscedastic case and ordinary minimaxity, the upper bound of the constant C can be chosen to be as large as $\frac{2m(p-2)}{m+2}$ for the original James-Stein estimators. While for our case, the upper

bound becomes $\frac{m(p-2)}{m+2}$. Like in Theorem 3, it can be shown that the bound can not be easily improved. However, we omit the result here for simplicity.

We can also extend the parametric Bayes estimator δ_{PEB} in (3.3.2) and δ^+_{PEB} in (3.3.3) to the unknown variance case. Consider δ_{PEBV} with the form

$$(\delta_{PEBV}(X))_i = \left(1 - \frac{CM\gamma_i}{CM\gamma_i + (\sum_{j=1}^p X_j^2 - \sum_{j=1}^p M\gamma_j)}\right) X_i \tag{4.1.4}$$

and δ^+_{PEBV} with the form

$$(\delta^+_{PEBV}(X))_i = \left(1 - \frac{CM\gamma_i}{CM\gamma_i + (\sum_{j=1}^p X_j^2 - \sum_{j=1}^p M\gamma_j)_+}\right) X_i. \tag{4.1.5}$$

The following corollary gives the conditions that guarantee the ensemble minimaxity of δ_{PEBV} in (4.1.4) and δ^+_{PEBV} in (4.1.5).

Corollary 7. *Assume $m \geq 6$, $p \geq 3$ and*

$$p \leq \frac{\sum_{j=1}^p \gamma_i}{\gamma_{min}} \leq C \leq \frac{2m(p-2)}{m+2}. \tag{4.1.6}$$

Then both δ_{PEBV} in (4.1.4) and δ^+_{PEBV} in (4.1.5) are ensemble minimax.

Proof. See Chapter 6. □

4.2 Shrinkage towards the Common Mean

In the sections above, we discuss the ensemble minimaxity properties of the estimators that shrink towards zero under the heteroscedastic model. We will generalize our method to provide a class of ensemble minimax estimators that shrink towards

the common mean in this section. Assume that $X \sim N(\theta, \Sigma)$ and $\theta \sim N(\mu 1, \tau^2 I)$, where Σ is the covariance matrix. We first present a lemma whose proof is sufficiently simple to be omitted.

4.2.1 General Theory

Lemma 4. *There exists an orthogonal matrix Q with the form*

$$Q = \begin{pmatrix} \frac{1}{\sqrt{p}} 1^T \\ Q_2 \end{pmatrix} ,$$

such that $T = Q\Sigma Q^T$ can be written in the block matrix form

$$T = \begin{pmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{pmatrix}$$

where T_{11} is 1×1, and $T_{22} = diag\{t_{22}, \cdots, t_{pp}\}$ is a $(p-1) \times (p-1)$ diagonal matrix.

From the fact that Q is orthogonal, we have

$$\begin{aligned} Q_2 1 &= 0 \\ Q_2 Q_2^T &= I_{p-1} \\ Q_2^T Q_2 &= I_p - \frac{1}{p} 1 1^T . \end{aligned}$$

Moreover, we also have $T_{22} = Q_2 \Sigma Q_2^T$. Since Σ is positive definite, we can easily verify that T_{22} is also positive definite. Therefore, $t_{ii} > 0$ for all $i = 2, \cdots, p$. Assume $p \geq 4$. Let $Y = QX$,$\eta = Q\theta$ and $Y_{(2)} = (Y_2, \cdots, Y_p)^T$. Then we have

$$Y = \begin{pmatrix} \frac{1}{\sqrt{p}} 1^T \\ Q_2 \end{pmatrix} (\overline{X} 1 + (X - \overline{X} 1)) = \begin{pmatrix} \sqrt{p} \overline{X} \\ Q_2 (X - \overline{X} 1) \end{pmatrix} ,$$

which implies $Y_1 = \sqrt{p}\overline{X}$ and $Y_{(2)} = Q_2(X - \overline{X}1)$. Note that $Q\mu 1 = (\sqrt{p}\mu, 0, \cdots, 0)^T$ and $Q\text{diag}(\tau^2 I_p)Q^T = \tau^2 I_p$. Consider the estimator δ_{cm} with the form

$$\delta_{cm}(X) = Q^T \left(Y_1, \xi_2(Y_{(2)}), \cdots, \xi_p(Y_{(2)})\right)^T \ , \tag{4.2.1}$$

where $\xi_i(Y_{(2)})$ is any ensemble minimax estimator for $\eta_{(2)}$, $\forall i = 2, \cdots, p$. We then have the following result.

Theorem 10. *For $p \geq 4$, δ_{cm} in (4.2.1) is ensemble minimax.*

Proof. See Chapter 6. □

Note that δ_{cm} in (4.2.1) can be interpreted as "shrinking" towards the overall mean since it can be written as $\delta_{cm}(X) = \overline{X}1 + Q_2^T \xi(Q_2(X - \overline{X}1))$, which is a generalized shrinkage estimator.

Furthermore, if we assume that $\xi_i(Y_{(2)}) = (1 - h_i(\|Y_{(2)}\|^2))Y_i$ for $i = 2, \cdots, p$, we have

$$\begin{aligned} \delta_{cm}(X) &= \overline{X}1 + Q_2^T \text{diag}\{1 - h_2(\|Y_{(2)}\|^2), \cdots, 1 - h_p(\|Y_{(2)}\|^2)\}Y_{(2)} \\ &= \overline{X}1 + Q_2^T \text{diag}\{1 - h_2(\|Y_{(2)}\|^2), \cdots, 1 - h_p(\|Y_{(2)}\|^2)\}Q_2(X - \overline{X}1) \ , \end{aligned}$$

which, along with the fact that $\|Y_{(2)}\|^2 = \|Q_2(X - \overline{X}1)\|^2 = \|X - \overline{X}1\|^2$, implies

$$\delta_{cm}(X) = \overline{X}1 + D \cdot (X - \overline{X}1)$$

with $D = Q_2^T \text{diag}\{1 - h_2(\|X - \overline{X}1\|^2), \cdots, 1 - h_p(\|X - \overline{X}1\|^2)\}Q_2$.

4.2.2 Random Effects Models

The standard one-way random effects model involves observations of independent variables Y_{ij}, $i = 1, \cdots, p$, $j = 1, \cdots, J_i$ under the distributional assumptions

$$\begin{aligned} Y_{ij}|\theta_i &\sim N(\theta_i, \sigma^2) \text{ (independent)} \\ \theta_i|\mu, \tau^2 &\sim N(\mu, \tau^2) \text{ (independent)} \ . \end{aligned}$$

Here, the unknown parameters are σ^2, μ, τ^2. To fit with previous notation, let

$$\begin{aligned} X_i &= Y_{i\cdot} = \frac{1}{J_i} \sum_{j=1}^{J_i} Y_{ij} \\ M &= \frac{1}{n-p} \sum_{i}^{p} \sum_{j=1}^{J_i} (Y_{ij} - Y_{i\cdot})^2, \ m = \sum_{i=1}^{p} (J_i - 1) = n - p \ . \end{aligned}$$

Thus, M denotes the usual mean squared errors and m denotes the degrees of freedom for error.

The goal in random effects estimation (sometimes referred to as "prediction") is to estimate (predict) the values of θ_i under ordinary squared error loss

$$L(\delta, \theta) = \sum_{i=1}^{p} (\delta_i - \theta_i)^2 \ .$$

The usual estimator (predictor) is of course $\delta_0(X) = X$. This problem is clearly mathematically equivalent to the ensemble minimaxity formulation. Hence, ensemble minimaxity in the hierarchical formulation is identical to ordinary minimaxity for the estimation of $\{\theta_i\}$ in the random effects model.

We construct a class of ensemble minimax generalized shrinkage estimators following the approach in Section 4.2.1. Let $\Gamma = diag\{1/J_1, \cdots, 1/J_p\}$. From Lemma

4, there exists an orthogonal matrix Q with the form

$$Q = \begin{pmatrix} \frac{1}{\sqrt{p}} 1^T \\ Q_2 \end{pmatrix} ,$$

such that $T = Q\Gamma Q^T$ can be written in the block matrix form

$$T = \begin{pmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{pmatrix}$$

where T_{11} is 1×1, and $T_{22} = \text{diag}\{t_{22}, \cdots, t_{pp}\}$ is a $(p-1) \times (p-1)$ diagonal matrix.

Assume $p \geq 4$. Let $U = QX$,$\eta = Q\theta$ and $U_{(2)} = (U_2, \cdots, U_p)^T$. Consider the estimator δ_{cmv} with the form

$$\delta_{cmv}(X) = Q^T \left(Y_1, \xi_2(U_{(2)}, M), \cdots, \xi_p(U_{(2)}, M)\right)^T , \tag{4.2.2}$$

where $\xi_i(U_{(2)}, M)$ is any ensemble minimax estimator for $\eta_{(2)}$, $\forall i = 2, \cdots, p$. Note that δ_{cmv} in (4.2.2) can be interpreted as "shrinking" towards the overall mean since it can be written as $\delta_{cmv}(X) = \overline{X}1 + Q_2^T \xi(Q_2(X - \overline{X}1), M)$, which is a generalized shrinkage estimator. Following the similar approach in Theorem 10, it is easy to verify that δ_{cmv} in (4.2.2) is ensemble minimax.

Especially, if we choose δ^+_{GSV} as ξ here, we get the estimator $\delta^+_{GSV;re}$ with the form

$$\delta^+_{GSV;re}(X) = \overline{X}1 + Q_2^T A Q_2(X - \overline{X}1) , \tag{4.2.3}$$

where $A = diag\{(1 - \frac{\lambda_2 M t_{22}}{\nu_2 M t_{22} + ||X - \overline{X}1||^2})_+, \cdots, (1 - \frac{\lambda_p M t_{pp}}{\nu_p M t_{pp} + ||X - \overline{X}1||^2})_+\}$. We have the following corollary that shows its ensemble mimimax property.

Corollary 8. *$\delta^{+}_{GSV;re}$ in (4.2.3) is ensemble minimax if $p \geq 4$, $m \geq 3$ and for any $i = 2, \cdots, p$, $0 \leq \lambda_i \leq \frac{2m(p-3)}{m+2}$ and $\nu_i \geq (\frac{m+2}{m-2}\lambda_i - \frac{m(p-3)}{m-2}(1 + \frac{t_{min}}{t_{ii}}))_+$. Hence, $\delta^{+}_{GSV;re}$ is minimax for the random effects model under these conditions and dominates the usual estimator $\delta_0(X) = X$.*

In the interest of space we omit the formal proof. If a single version of the above estimators is to be used for all $J_1, \cdots, J_p$, the preferred and simple choice would be $\delta^{+}_{GSV;re}$ in (4.2.3) with $\lambda_i = \frac{m(p-3)}{m+2}$ and $\nu_i = 0$.

Chapter 5

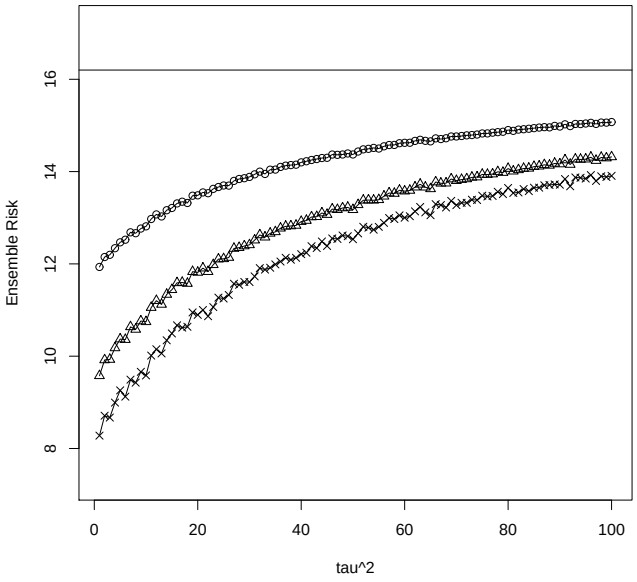
Further Discussions

In this paper, we propose a class of ensemble minimax estimators that dominate the usual estimator $\delta_0(X) = X$. But there are still many interesting open questions in this area. Three of them are listed as below.

Q1. How shall we choose the shrinkage factor?

We provide a class of ensemble minimax estimators in our paper, however, we do not specify how to choose the shrinkage factors among all possibilities. For example, δ^+_{J-S} is ensemble minimax if $p \geq 3$ and $0 \leq C \leq p-2$. But which C works best among all possible values? We conjecture that the estimators which have the largest shrinkage factors are favorable, but we do not have a rigorous proof for our conjecture. Figure 5.1 provides an example from a simulation study. In this example, we have $p = 6$, $\Sigma = diag\{0.9, 1.8, 2.3, 2.5, 4.1, 4.6\}$. The solid line is the ensemble risk of $\delta_0(X) = X$. The circle line, triangle line and cross line are the

Figure 5.1: Ensemble Risk of δ^+_{J-S} with different C and $\delta_0(X) = X$



ensemble risks of δ^+_{J-S} with $C = 1$, $C = 2$ and $C = 3$, respectively. As we can see from Figure 5.1, δ^+_{J-S} with $C = 3$ has the smallest ensemble risk. Theoretical research of the choice of the shrinkage factor will be an interesting topic for future work.

Q2. Are these ensemble minimax estimators ensemble admissible?

Another interesting question is to investigate the ensemble admissibility of various ensemble minimax estimators we propose here. We conjecture that all the ensemble admissible estimators are either Bayes estimators or limit of Bayes estimators. Therefore, we suspect that the estimators we proposed here are not ensemble admissible, and some Bayes estimators or limit of Bayes estimators will be a better candidate. Further research is needed in this area.

Q3. Can we apply empirical Bayes methods to solve other problem?

In this paper, we use empirical Bayes methods to construct various ensemble minimax estimators. The key idea is to use the data to extract information for the underlying structure of the unknown parameters. We can apply this powerful tool to solve other problems. One possible theoretical application is to solve the Robbins classification problems under heteroscedastic and unknown variances assumptions using empirical Bayes methods. It would also be great if we can apply this method to deal with data from real world.

Chapter 6

Proofs and Supplemental Materials

Proof of Lemma 1

Proof. By definition, we have

$$\overline{R}_{\tau^2}(\delta) = \int\int L(\theta, \delta(x)) P_{x|\theta}(dx)\pi_\tau(d\theta) = \sum_{i=1}^{p} E(X_i - \theta_i)^2 \ .$$

Note that

$$\begin{pmatrix} \theta_i \\ X_i \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau^2 & \tau^2 \\ \tau^2 & \tau^2 + \sigma_i^2 \end{pmatrix}\right)$$

and (θ_i, X_i) are jointly independent for different i, we have from the property of

conditional expectation

$$\begin{aligned}
\overline{R}_{\tau^2}(\delta) &= \sum_{i=1}^{p}[E_x(\delta_i(X) - E(\theta_i|X))^2 + E_x(E((\theta_i - E(\theta_i|X_i))^2|X))] \\
&= \sum_{i=1}^{p}\left[E_x\left((1-h_i(X))X_i - \frac{\tau^2}{\tau^2+\sigma_i^2}X_i\right)^2 + \frac{\tau^2\sigma_i^2}{\tau^2+\sigma_i^2}\right] \\
&= \sum_{i=1}^{p}\left[E_x\left(\frac{\sigma_i^2}{\tau^2+\sigma_i^2}X_i - h_i(X)X_i\right)^2 + \frac{\tau^2\sigma_i^2}{\tau^2+\sigma_i^2}\right]
\end{aligned}$$

where E_x is used to emphasize that the expectation is taken with respect to the marginal distribution of X, i.e., each coordinate X_i has the normal distribution $N(0, \tau^2 + \sigma_i^2)$ and they are jointly independent. □

Proof of Lemma 2

Proof. The proof follows the same approach as in Lemma 1. Once we condition on M. First note that

$$\begin{aligned}
\overline{R}_{\tau^2}(\delta) &= E\left[\sum_{i=1}^{p}(\delta_i(X,M) - \theta_i)^2\right] \\
&= \sum_{i=1}^{p}E[E[(\delta_i(X,M) - \theta_i)^2|M]]
\end{aligned}$$

Since given M, (θ_i, X_i) is an independent array whose distribution is

$$\begin{pmatrix} \theta_i \\ X_i \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau^2 & \tau^2 \\ \tau^2 & \tau^2+\sigma^2\gamma_i \end{pmatrix}\right).$$

Conditioning on M, we have, as in Lemma 1,

$$\begin{aligned}
& E[(\delta_i(X,M) - \theta_i)^2|M] \\
= \; & E\left[\left(\frac{\sigma^2\gamma_i}{\tau^2+\sigma^2\gamma_i}X_i - h_i(X,M)X_i\right)^2|M\right] + \frac{\tau^2\sigma^2\gamma_i}{\tau^2+\sigma^2\gamma_i},
\end{aligned}$$

which then implies

$$\begin{aligned}\overline{R}_{\tau^2}(\delta) &= \sum_{i=1}^{p} E[E[(\delta_i(X,M)-\theta_i)^2]|M] \\ &= E\left[\sum_{i=1}^{p}\left(\frac{\sigma^2\gamma_i}{\tau^2+\sigma^2\gamma_i}X_i - h_i(X,M)X_i\right)^2 + \frac{\tau^2\sigma^2\gamma_i}{\tau^2+\sigma^2\gamma_i}\right] .\end{aligned}$$

□

Proof of Lemma 3

Proof. First we use $\mu'(M)$ to induce a new probability distribution

$$P_{\mu'}(M \in A) = \int_A \frac{\mu'(m)}{E(\mu'(M))}dm .$$

Using this change of measure, we have

$$\frac{E[M\mu(M)]}{E[\mu(M)]} = E_{\mu'}[M \cdot r(M)] \times \frac{E[\mu'(M)]}{E[\mu(M)]}$$

and

$$\frac{E(M\mu'(M))}{E(\mu'(M))} = E_{\mu'}(M) ,$$

the original inequality then becomes a direct application of covariance inequality under the new probability $P_{\mu'}$. □

Proof of Theorem 1

Proof. For any $\pi = N(0,\tau^2) \in \mathcal{P}$, the ensemble risk of δ_0 with respect to π is given as

$$\overline{R}_{\tau^2}(\delta_0) = E_{\tau^2}[E_\theta \sum_{i=1}^{p}(X_i-\theta_i)^2] = \sum_{i=1}^{p}\sigma_i^2 .$$

Moreover, the Bayes risk is $\sum_{i=1}^{p} \frac{\sigma_i^2\tau^2}{\sigma_i^2+\tau^2}$. Therefore, for any estimator $\delta^{'}$, we have

$$\overline{R}_{\tau^2}(\delta^{'}) \geq \sum_{i=1}^{p} \frac{\sigma_i^2\tau^2}{\sigma_i^2+\tau^2},\ \forall \tau^2 \in (0,\infty)$$

Thus,

$$\inf_{\tau^2} \sup_{\delta'} \overline{R}_{\tau^2}(\delta^{'}) \geq \sum_{i=1}^{p} \sigma_i^2\ .$$

Hence, δ_0 is ensemble minimax with respect to $\mathcal{P}$. □

Proof of Theorem 2

Proof. From Lemma 1. and the fact that

$$\overline{R}_{\tau^2}(\delta_0) = \sum_{i=1}^{p} \sigma_i^2,$$

it suffices to show that for each i,

$$E\left(\frac{\sigma_i^2}{\sigma_i^2+\tau^2}X_i - h_i(X)X_i\right)^2 \leq \frac{\sigma_i^4}{\tau^2+\sigma_i^2}\ , \tag{6.0.1}$$

which is equivalent to

$$E\left[h_i(X)^2X_i^2\right] \leq \frac{2\sigma_i^2}{\tau^2+\sigma_i^2}E\left[h_i(X)X_i^2\right]\ .$$

To prove the above inequality, first note that condition (2) indicates

$$\begin{aligned} E\left[h_i(X)^2X_i^2\right] &= E\left[\mathfrak{h}_i(\underline{T},W)^2(\tau^2+\sigma_i^2)T_iW\right] \\ &= E\left[E(\mathfrak{h}_i(\underline{T},W)(\tau^2+\sigma_i^2)T_iW \times \mathfrak{h}_i(\underline{T},W)|\underline{T})\right]\ . \end{aligned}$$

From condition (3), (4) and the covariance inequality, we then have

$$E\left[h_i(X)^2X_i^2\right] \leq E[E(\mathfrak{h}_i(\underline{T},W)(\tau^2+\sigma_i^2)T_iW|\underline{T}) \times E(\mathfrak{h}_i(\underline{T},W)|\underline{T})]\ ,$$

which implies

$$E\left[h_i(X)^2 X_i^2\right] \leq E\left[E(\mathfrak{h}_i(\underline{T},W)(\tau^2+\sigma_i^2)T_iW|\underline{T}) \times E\left(\sup_{\underline{T}} \mathfrak{h}_i(\underline{T},W)|\underline{T}\right)\right] .$$

Based on the independence of $\underline{T}$ and W, we have

$$\begin{aligned} E\left[h_i(X)^2 X_i^2\right] &\leq E\left[E(\mathfrak{h}_i(\underline{T},W)(\tau^2+\sigma_i^2)T_iW|\underline{T}) \times E\left(\sup_{\underline{T}} \mathfrak{h}_i(\underline{T},W)\right)\right] \\ &= E\left(\sup_{\underline{T}} \mathfrak{h}_i(\underline{T},W)\right) \times E[\mathfrak{h}_i(\underline{T},W)T_iW] , \end{aligned}$$

which along from condition (5) shows

$$E\left[h_i(X)^2 X_i^2\right] \leq \frac{2\sigma_i^2}{\sigma_i^2+\tau^2} E(\mathfrak{h}_i(\underline{T},W)(\tau^2+\sigma_i^2)T_iW) = \frac{2\sigma_i^2}{\sigma_i^2+\tau^2} E(h_i(X)X_i^2) .$$

This proves the ensemble minimaxity of δ. □

Proof of Theorem 3

Proof. Fix $\sigma_1^2 > 0$ and set $\sigma_2^2 = \cdots = \sigma_p^2 = \sigma^2$. Let $\sigma^2 \to 0$. In order to show that δ_{J-S}^+ dominates δ_0 with respect to $\mathcal{P}$, we have to prove

$$E\left[h_1(X)^2 X_1^2\right] \leq \frac{2\sigma_1^2}{\tau^2+\sigma_1^2} E\left[h_1(X)X_1^2\right] \tag{6.0.2}$$

with

$$h_1(X) = \frac{c(p-2)\sigma_1^2}{\|X\|^2} \wedge 1 .$$

However, the law of large numbers implies that

$$\frac{c(p-2)\sigma_1^2}{\|X\|^2} = \frac{c(p-2)\sigma_1^2}{p} \frac{p}{\|X\|^2} \to \frac{c\sigma_1^2}{\tau^2}$$

as $p \to \infty$. Let $\sigma_1^2 < \tau^2 < c\sigma_1^2$, we then have

$$h_1(X) \to 1$$

as $p \to \infty$. Since $0 < h_1(X) \leq 1$, the dominated convergence theorem implies

$$E(h_1(X)^2 X_1^2) \to \tau^2 + \sigma_1^2$$

and

$$E(h_1(X) X_1^2) \to \tau^2 + \sigma_1^2 \ .$$

However, our choice of σ_1^2 and τ^2 indicates

$$1 > \frac{2\sigma_1^2}{\tau^2 + \sigma_1^2},$$

which means that the inequality (6.0.2) does not hold for large p. Hence, δ^+_{J-S} in (3.1.2) is not ensemble minimax for some p and $\sigma_1^2, \cdots, \sigma_p^2$. □

Proof of Theorem 4

Proof. Fix $\sigma_1^2 = 1$, and let $\sigma_2^2 = \cdots = \sigma_p^2 = \sigma^2$. It suffices to show

$$\lim_{\substack{\tau^2 \to 0 \\ \sigma^2 \to 0}} \overline{R}_{\tau^2}(\delta_{J-S}) = \infty \ .$$

From Lemma 1, we have

$$\begin{aligned}
\lim_{\substack{\tau^2 \to 0 \\ \sigma^2 \to 0}} \overline{R}_{\tau^2}(\delta_{J-S}) &\geq \lim_{\substack{\tau^2 \to 0 \\ \sigma^2 \to 0}} E\left(\frac{1}{\tau^2+1}X_1 - \frac{C}{\|X\|^2}X_1\right)^2 \\
&\geq \lim_{\substack{\tau^2 \to 0 \\ \sigma^2 \to 0}} \frac{1}{2}E\left(\frac{C}{\|X\|^2}X_1\right)^2 - \lim_{\substack{\tau^2 \to 0 \\ \sigma^2 \to 0}} E\left(\frac{1}{\tau^2+1}X_1\right)^2 \\
&\geq \lim_{\substack{\tau^2 \to 0 \\ \sigma^2 \to 0}} \frac{1}{2}E\left(\frac{C^2X_1^2}{\|X\|^4}\right) - 1 \ .
\end{aligned}$$

Since the last term is finite, it is sufficient to prove

$$\lim_{\substack{\tau^2 \to 0 \\ \sigma^2 \to 0}} E\left(\frac{X_1^2}{\|X\|^4}\right) = \infty \ .$$

Let $X_1 = \sqrt{1+\tau^2}Z_1$, $Z_1 \sim N(0,1)$, and $X_i = \sqrt{\tau^2+\sigma^2}Z_i$, $Z_i \sim N(0,1)$, $\forall i = 2, \cdots, p$. Therefore,

$$E\left(\frac{X_1^2}{\|X\|^4}\right) = (1+\tau^2)E\left(\frac{Z_1^2}{((1+\tau^2)Z_1^2 + (\tau^2+\sigma^2)\sum_{i=2}^p Z_i^2)^2}\right) \ .$$

Note that $\frac{Z_1^2}{((1+\tau^2)Z_1^2+(\tau^2+\sigma^2)\sum_{i=2}^p Z_i^2)^2}$ is increasing when both σ^2 and τ^2 decrease to zero and

$$\lim_{\substack{\tau^2 \to 0 \\ \sigma^2 \to 0}} \frac{X_1^2}{\|X\|^4} = \frac{1}{Z_1^2}.$$

From the monotone convergence theorem, we have

$$\lim_{\substack{\tau^2 \to 0 \\ \sigma^2 \to 0}} E\left(\frac{X_1^2}{\|X\|^4}\right) = E\left(\frac{1}{Z_1^2}\right) = \infty$$

which completes the proof. □

Proof of Theorem 5

Proof. First note that δ_i^+ is equivalently written as $\delta_i^+(X) = (1 - h_i^+(X))X$, where $h_i^+(X) = h_i(x) \wedge 1$. From Lemma 1, we have

$$\begin{aligned} & \overline{R}_{\tau^2}(\delta^+) - \overline{R}_{\tau^2}(\delta) \\ = & \sum_{i=1}^{p} E\left[\left(h_i^+(X) + h_i(X) - \frac{2\sigma_i^2}{\tau^2+\sigma_i^2}\right)(h_i(X) - h_i^+(X))X_i^2\right] . \end{aligned}$$

Since for any $i = 1, \cdots, p$,

$$h_i(X) - h_i^+(X) = \begin{cases} h_i(X) - 1 & \text{, if } h_i(X) > 1 \\ 0 & \text{, if } h_i(X) \leq 1 \end{cases} ,$$

we then have

$$\begin{aligned} & \overline{R}_{\tau^2}(\delta^+) - \overline{R}_{\tau^2}(\delta) \\ = & \sum_{i=1}^{p} E\left[\left(h_i^+(X) + h_i(X) - \frac{2\sigma_i^2}{\tau^2+\sigma_i^2}\right)(h_i(X) - h_i^+(X))X_i^2\right] \\ = & \sum_{i=1}^{p} E\left[\left(1 + h_i(X) - \frac{2\sigma_i^2}{\tau^2+\sigma_i^2}\right)(h_i(X) - 1)X_i^2 I_{\{h_i(X)>1\}}\right] \\ \geq & \sum_{i=1}^{p} E\left[\left(2 - \frac{2\sigma_i^2}{\tau^2+\sigma_i^2}\right)(h_i(X) - 1)X_i^2 I_{\{h_i(X)>1\}}\right] > 0 \end{aligned}$$

which completes the proof. □

Proof of Theorem 6

Proof. When $p = 1$, set $\tau^2 = 1$. From Lemma 1, we only need to show

$$\lim_{\sigma_1^2 \to 0} E\left(\frac{1}{1+\sigma_1^2}X_1 - \frac{C}{X_1^2 + (C-1)\sigma_1^2}X_1\right)^2 > \frac{1}{1+\sigma_1^2} .$$

Since

$$\begin{aligned} & E\left(\frac{1}{1+\sigma_1^2}X_1 - \frac{C}{X_1^2+(C-1)\sigma_1^2}X_1\right)^2 \\ \geq\ & \frac{1}{2}E\left(\frac{C}{X_1^2+(C-1)\sigma_1^2}X_1\right)^2 - E\left(\frac{1}{1+\sigma_1^2}X_1\right)^2 \\ =\ & \frac{1}{2}E\left(\frac{C^2X_1^2}{(X_1^2+(C-1)\sigma_1^2)^2}\right) - \frac{1}{1+\sigma_1^2} \end{aligned}$$

where the last term is finite, it is then sufficient to show

$$\lim_{\sigma_1^2\to 0} E\left(\frac{X_1^2}{(X_1^2+(C-1)\sigma_1^2)^2}\right) = \infty\ . \tag{6.0.3}$$

When $C < 1$, this is trivial. In fact, it holds for any σ_1^2. When $C \geq 1$, let $X_1 = \sqrt{1+\sigma_1^2}Z_1$, $Z_1 \sim N(0,1)$, we have

$$E\left(\frac{X_1^2}{(X_1^2+(C-1)\sigma_1^2)^2}\right) = E\left(\frac{(1+\sigma_1^2)Z_1^2}{((1+\sigma_1^2)Z_1^2+(C-1)\sigma_1^2)^2}\right)\ .$$

Since $\frac{Z_1^2}{((1+\sigma_1^2)Z_1^2+(C-1)\sigma_1^2)^2}$ is increasing as $\sigma_1^2 \to 0$, and

$$\lim_{\sigma_1^2\to 0} \frac{Z_1^2}{((1+\sigma_1^2)Z_1^2+(C-1)\sigma_1^2)^2} = \frac{1}{Z_1^2}\ ,$$

from monotone convergence theorem, we have

$$\lim_{\sigma_1^2\to 0} E\left(\frac{Z_1^2}{((1+\sigma_1^2)Z_1^2+(C-1)\sigma_1^2)^2}\right) = E\left(\frac{1}{Z_1^2}\right) = \infty\ .$$

Note that $1+\sigma_1^2 \to 1$ as $\sigma_1^2 \to 0$, (6.0.3) is then verified. Hence, when $p = 1$, δ_{PEB} in (3.3.2) is not ensemble minimax.

For the case where $p \geq 2$, let $\sigma_1^2 = 1$ and $\sigma_2^2 = \cdots = \sigma_p^2 = C$. Again from

Lemma 1, we have

$$\begin{aligned}
\overline{R}_{\tau^2}(\delta_{PEB}) &\geq E\left(\frac{1}{\tau^2+1}x_1 - \frac{C}{C+\|X\|^2-1-(p-1)C}X_1\right)^2 \\
&\geq \frac{1}{2}E\left(\frac{C^2X_1^2}{(\|X\|^2-1-(p-2)C)^2}\right) - E\left(\frac{1}{(\tau^2+1)^2}X_1^2\right) \\
&= \frac{1}{2}E\left(\frac{C^2X_1^2}{(\|X\|^2-1-(p-2)C)^2}\right) - \frac{1}{\tau^2+1} \\
&= \frac{1}{2}E\left[E\left(\frac{C^2X_1^2}{(\|X\|^2-1-(p-2)C)^2} \mid X_1\right)\right] - \frac{1}{\tau^2+1} \ .
\end{aligned}$$

For any $X_1^2 < 1+(p-2)C$, it is not difficult to see that

$$E\left(\frac{C^2X_1^2}{(\|X\|^2-1-(p-2)C)^2} \mid X_1\right) = \infty$$

and

$$P(X_1^2 < 1+(p-2)C) > 0 \ ,$$

we then have

$$E\left(\frac{C^2X_1^2}{(\|X\|^2-1-(p-2)C)^2}\right) = \infty \ ,$$

which implies

$$\overline{R}_{\tau^2}(\delta_{PEB}) = \infty \ . \tag{6.0.4}$$

Therefore, δ_{PEB} in (3.3.2) is not ensemble minimax. To sum up, there exists some $\sigma_1^2, \cdots, \sigma_p^2$ such that δ_{PEB} in (3.3.2) is not ensemble minimax. □

Proof of Theorem 7

Proof. Let $\sigma_1^2 = \cdots = \sigma_p^2 = 1$ and $\tau^2 = 2$. Similarly as above, to show that δ_{PEB}^+ in (3.3.3) is ensemble minimax, we would need to have

$$E\left(\sum_{i=1}^p h_i^2(X)X_i^2\right) \leq \frac{2}{3}E\left(\sum_{i=1}^p h_i(X)X_i^2\right) \tag{6.0.5}$$

with

$$h_i(X) = \frac{1}{1 + \frac{1}{C}(\|X\|^2 - p)_+} \ .$$

Notice that $h_i(X) \to I_{\{\|X\|^2 \le p\}}$ as $C \to 0$ and $h_i(X) \le 1$, from dominant convergence theorem, we have

$$E\left(\sum_{i=1}^{p} h_i^2(X) X_i^2\right) \to E\left[\|X\|^2 I_{\{\|X\|^2 \le p\}}\right]$$

and

$$E\left(\sum_{i=1}^{p} h_i(X) X_i^2\right) \to E\left[\|X\|^2 I_{\{\|X\|^2 \le p\}}\right]$$

as $C \to 0$. Hence, as $C \to 0$, (6.0.5) would no longer always hold. Thus, there exists some sufficiently small C and some $\sigma_1^2, \cdots, \sigma_p^2$ such that δ_{PEB}^+ in (3.3.3) is not ensemble minimax. □

Proof of Theorem 8

Proof. As in the proof for the known variance case, based on Lemma 2, it suffices to show

$$E\left[\left(\frac{\lambda_i M\gamma_i}{\nu_i M\gamma_i + \|X\|^2}\right)^2 X_i^2\right] \le \frac{2\sigma^2\gamma_i}{\sigma^2\gamma_i + \tau^2} E\left(\frac{\lambda_i M\gamma_i}{\nu_i M\gamma_i + \|X\|^2} X_i^2\right) \ .$$

Conditioning on M and following the proof in Theorem 2 and Corollary 2, we know

$$\begin{aligned} & E\left[\left(\frac{\lambda_i M\gamma_i}{\nu_i M\gamma_i + \|X\|^2}\right)^2 X_i^2\right] \\ \le \ & E\left[E\left(\frac{\lambda_i M\gamma_i}{\nu_i M\gamma_i + \|X\|^2} X_i^2 | M\right) \times E\left(\frac{\lambda_i M\gamma_i}{\nu_i M\gamma_i + (\sigma^2\gamma_{min} + \tau^2)W} | M\right)\right] \ . \end{aligned}$$

The difficulty here is that a direct application of the covariance inequality on the two conditional expectation is no longer helpful since they are both increasing in M.

However, by moving the M in the numerator of the second conditional expectation to the first one, the covariance inequality can then be applied, i.e.,

$$\begin{aligned}
& E\left[\left(\frac{\lambda_i M\gamma_i}{\nu_i M\gamma_i + \|X\|^2}\right)^2 X_i^2\right] \\
\leq\ & E\left[E\left(\frac{\lambda_i M\gamma_i}{\nu_i M\gamma_i + \|X\|^2}X_i^2|M\right) \times E\left(\frac{\lambda_i M\gamma_i}{\nu_i M\gamma_i + (\sigma^2\gamma_{min} + \tau^2)W}|M\right)\right] \\
=\ & E\left[E\left(\frac{\lambda_i M^2\gamma_i}{\nu_i M\gamma_i + \|X\|^2}X_i^2|M\right) \times E\left(\frac{\lambda_i \gamma_i}{\nu_i M\gamma_i + (\sigma^2\gamma_{min} + \tau^2)W}|M\right)\right] \\
=\ & E\left[E\left(\frac{\lambda_i M^2\gamma_i}{\nu_i M\gamma_i + \|X\|^2}X_i^2|M\right)\right] \times E\left[E\left(\frac{\lambda_i \gamma_i}{\nu_i M\gamma_i + (\sigma^2\gamma_{min} + \tau^2)W}|M\right)\right] \\
=\ & E\left(\frac{\lambda_i M^2\gamma_i}{\nu_i M\gamma_i + \|X\|^2}X_i^2\right) \times E\left(\frac{\lambda_i \gamma_i}{\nu_i M\gamma_i + (\sigma^2\gamma_{min} + \tau^2)W}\right) .
\end{aligned}$$

Now let

$$\mu_s(M) = \frac{\lambda_i M\gamma_i}{\nu_i M\gamma_i + s} ,$$

notice that the ratio $r(M) = \mu_s(M)/\mu_{s'}(M)$ is non-decreasing in M for $s > s'$, from Lemma 3, we then have

$$\frac{E\left(\frac{\lambda_i M^2\gamma_i}{\nu_i M\gamma_i + \|X\|^2}|X\right)}{E\left(\frac{\lambda_i M\gamma_i}{\nu_i M\gamma_i + \|X\|^2}|X\right)} \leq \lim_{\|X\|^2\to\infty} \frac{E\left(\frac{\lambda_i M^2\gamma_i}{\nu_i M\gamma_i + \|X\|^2}|X\right)}{E\left(\frac{\lambda_i M\gamma_i}{\nu_i M\gamma_i + \|X\|^2}|X\right)} = \lim_{\|X\|^2\to\infty} \frac{E\left(\frac{\lambda_i M^2\gamma_i \cdot \|X\|^2}{\nu_i M\gamma_i + \|X\|^2}|X\right)}{E\left(\frac{\lambda_i M\gamma_i \cdot \|X\|^2}{\nu_i M\gamma_i + \|X\|^2}|X\right)} .$$

Applying monotone convergence theorem gives us

$$\lim_{\|X\|^2\to\infty} \frac{E\left(\frac{\lambda_i M^2\gamma_i \cdot \|X\|^2}{\nu_i M\gamma_i + \|X\|^2}|X\right)}{E\left(\frac{\lambda_i M\gamma_i \cdot \|X\|^2}{\nu_i M\gamma_i + \|X\|^2}|X\right)} = \frac{E(M^2)}{E(M)} = \frac{(m+2)\sigma^2}{m} ,$$

which along with the previous inequality implies

$$E\left(\frac{\lambda_i M^2\gamma_i}{\nu_i M\gamma_i + \|X\|^2}|X\right) \leq \frac{(m+2)\sigma^2}{m} E\left(\frac{\lambda_i M\gamma_i}{\nu_i M\gamma_i + \|X\|^2}|X\right) .$$

Multiplying both sides by X_i^2 and taking expectation leads to

$$E\left(\frac{\lambda_i M^2\gamma_i}{\nu_i M\gamma_i + \|X\|^2}X_i^2\right) \leq \frac{(m+2)\sigma^2}{m} E\left(\frac{\lambda_i M\gamma_i}{\nu_i M\gamma_i + \|X\|^2}X_i^2\right) .$$

Since we have already shown that

$$E\left[\left(\frac{\lambda_i M\gamma_i}{\nu_i M\gamma_i+\tau^2 W}\right)^2 X_i^2\right] \le E\left(\frac{\lambda_i M^2\gamma_i}{\nu_i M\gamma_i+\|X\|^2}X_i^2\right)\times E\left(\frac{\lambda_i\gamma_i}{\nu_i M\gamma_i+(\sigma^2\gamma_{min}+\tau^2)W}\right) ,$$

in order to prove

$$E\left[\left(\frac{\lambda_i M\gamma_i}{\nu_i M\gamma_i+\|X\|^2}\right)^2 X_i^2\right] \le \frac{2\sigma^2\gamma_i}{\sigma^2\gamma_i+\tau^2}E\left(\frac{\lambda_i M\gamma_i}{\nu_i M\gamma_i+\|X\|^2}X_i^2\right) ,$$

it is then sufficient to show

$$\frac{(m+2)\sigma^2}{m}E\left(\frac{\lambda_i\gamma_i}{\nu_i M\gamma_i+(\sigma^2\gamma_{min}+\tau^2)W}\right) \le \frac{2\sigma^2\gamma_i}{\sigma^2\gamma_i+\tau^2} .$$

As in the proof of Corollary 2, using the covariance inequality twice, we have

$$\begin{aligned}
& \frac{(m+2)\sigma^2}{m}E\left(\frac{\lambda_i\gamma_i}{\nu_i M\gamma_i+(\sigma^2\gamma_{min}+\tau^2)W}\right)\\
= & \frac{(m+2)\sigma^2}{m}E\left[E\left(\frac{\lambda_i\gamma_i/W}{\nu_i M\gamma_i/W+(\sigma^2\gamma_{min}+\tau^2)}\right)|M\right]\\
= & \frac{(m+2)\sigma^2}{m}E\left[\frac{E(\lambda_i\gamma_i/W|M)}{E[(\nu_i M\gamma_i/W+(\sigma^2\gamma_{min}+\tau^2))|M]}\right]\\
= & \frac{(m+2)\sigma^2}{m}E\left[\frac{\lambda_i\gamma_i}{\nu_i M\gamma_i+(p-2)(\sigma^2\gamma_{min}+\tau^2)}\right]\\
= & \frac{(m+2)\sigma^2}{m}E\left[\frac{\lambda_i\gamma_i/M}{\nu_i\gamma_i+(p-2)(\sigma^2\gamma_{min}+\tau^2)/M}\right]\\
\le & \frac{(m+2)\sigma^2}{m}\frac{E(\lambda_i\gamma_i/M)}{E(\nu_i\gamma_i+(p-2)(\sigma^2\gamma_{min}+\tau^2)/M)}\\
= & \frac{(m+2)\sigma^2}{m}\frac{\lambda_i\gamma_i\cdot m/(m-2)}{\nu_i\sigma^2\gamma_i+(p-2)(\sigma^2\gamma_{min}+\tau^2)\cdot m/(m-2)}\\
= & \frac{(m+2)\lambda_i\gamma_i\cdot\sigma^2}{(m-2)\nu_i\sigma^2\gamma_i+m(p-2)(\sigma^2\gamma_{min}+\tau^2)} .
\end{aligned}$$

Now applying the condition $0\le\lambda_i\le\frac{2m(p-2)}{m+2}$ and $\nu_i\ge(\frac{m+2}{2(m-2)}\lambda_i-\frac{m\gamma_{min}(p-2)}{\gamma_i(m-2)})_+$, we finally have

$$\frac{(m+2)\lambda_i\gamma_i\cdot\sigma^2}{(m-2)\nu_i\sigma^2\gamma_i+m(p-2)(\sigma^2\gamma_{min}+\tau^2)} \le \frac{2\sigma^2\gamma_i}{\sigma^2\gamma_i+\tau^2}$$

which completes the proof. □

Proof of Theorem 9

Proof. The proof follows similar steps in the proofs of Corollary 3 and Theorem 8, therefore, we will skip most of the details and only highlight the parts that are substantially different. First let the shrinkage factor

$$h_i^+(X, M) = \min(1, h_i(X, M)) = \min\left(1, \frac{\lambda_i M \gamma_i}{\nu_i M \gamma_i + \|X\|^2}\right) .$$

As before, we have to prove

$$E\left[h_i^+(X, M)^2 X_i^2\right] \leq \frac{2\sigma^2\gamma_i}{\sigma^2\gamma_i + \tau^2} E\left[h_i^+(X, M) X_i^2\right] .$$

When $\sigma^2\gamma_i \geq \tau^2$, the above inequality is trivial. From now on, assume $\sigma^2\gamma_i < \tau^2$. As in the proof of Theorem 8, we have

$$E\left[h_i^+(X, M)^2 X_i^2\right] \leq E\left(h_i^+(X, M) M X_i^2\right) \times E\left(\frac{\lambda_i \gamma_i}{\nu_i M \gamma_i + (\sigma^2\gamma_{min} + \tau^2) W}\right) .$$

Define

$$\mu_s(M) = \min\left(1, \frac{\lambda_i M \gamma_i}{\nu_i M \gamma_i + s}\right) .$$

Note that the ratio $r(M) = \mu_s(M)/\mu'_s(M)$ is still non-decreasing in M for $s > s'$. As in Theorem 8, applying Lemma 3 and monotone convergence theorem leads to

$$E\left(h_i^+(X, M) M X_i^2\right) \leq \frac{(m+2)\sigma^2}{m} E\left(h_i^+(X, M) X_i^2\right) .$$

It is then sufficient to show

$$\frac{(m+2)\sigma^2}{m} E\left(\frac{\lambda_i \gamma_i}{\nu_i M \gamma_i + (\sigma^2\gamma_{min} + \tau^2) W}\right) \leq \frac{2\sigma^2\gamma_i}{\sigma^2\gamma_i + \tau^2}$$

whose proof follows exactly the same argument used in the last part of the proof of Corollary 3. □

Proof of Theorem 10

Proof. Since $\xi(Y_{(2)})$ is an ensemble minimax estimator for $\eta_{(2)}$, we have that

$$E\left[\sum_{i=2}^{p}\left(\xi_i(Y_{(2)})-\eta_i\right)^2\right] \leq \text{trace}(T_{22}) \ ,$$

which along with

$$E[(Y_1-\eta_1)^2] = T_{11}$$

and $\text{trace}(T) = \text{trace}(\Sigma)$ implies

$$E\left[(Y_1-\eta_1)^2+\sum_{i=2}^{p}\left(\xi_i(Y_{(2)})-\eta_i\right)^2\right] \leq \text{trace}(\Sigma) \ .$$

Therefore, we have

$$\begin{aligned} & E\left[\sum_{i=1}^{p}\left((\delta_c(X))_i-\theta_i\right)^2\right] \\ = \ & E\left[(Q^T(Y_1-\eta_1,\xi_2(Y_{(2)})-\eta_2,\cdots,\xi_p(Y_{(2)})-\eta_p)^T)^T\right. \\ & \left.\cdot\,(Q^T(Y_1-\eta_1,\xi_2(Y_{(2)})-\eta_2,\cdots,\xi_p(Y_{(2)})-\eta_p)^T)\right] \\ = \ & E\left[(Y_1-\eta_1)^2+\sum_{i=2}^{p}\left(\xi_i(Y_{(2)})-\eta_i\right)^2\right] \leq \text{trace}(\Sigma) \end{aligned}$$

which completes the proof. □

Proof of Corollary 2

Proof. It is sufficient for us to verify that the conditions in Theorem 2 are satisfied by

$$\begin{aligned} h_i(X) &= \frac{\lambda_i \sigma_i^2}{\nu_i \sigma_i^2 + \|X\|^2} \\ &= \frac{\lambda_i \sigma_i^2}{\nu_i \sigma_i^2 + \sum_{j=1}^{p} (\sigma_i^2 + \tau^2) T_i W} \\ &= \mathfrak{h}_i(\underline{T}, W) \ . \end{aligned}$$

Clearly, the shrinkage factor $h_i(X)$ satisfies conditions (1)-(4). For (5), define $g_i(W)$ as

$$g_i(W) = \frac{\lambda_i \sigma_i^2}{\nu_i \sigma_i^2 + (\sigma_{min}^2 + \tau^2) W} \ .$$

Then, $\sup_{\underline{T}} h_i(\underline{T}, W) \leq g_i(W)$. Using the covariance inequality, we have

$$\begin{aligned} E[g_i(W)] &= E\left[\frac{\lambda_i \sigma_i^2 / W}{\nu_i \sigma_i^2 / W + \sigma_{min}^2 + \tau^2}\right] \leq \frac{E[\lambda_i \sigma_i^2 / W]}{E[\nu_i \sigma_i^2 / W + \sigma_{min}^2 + \tau^2]} \\ &= \frac{\lambda_i \sigma_i^2}{\nu_i \sigma_i^2 + (p-2)(\sigma_{min}^2 + \tau^2)} \ . \end{aligned}$$

From the condition $0 \leq \lambda_i \leq 2(p-2)$ and $\nu_i \geq (\lambda_i/2 - (p-2) \cdot \sigma_{min}^2 / \sigma_i^2)_+$, it is then easy to verify

$$\frac{\lambda_i \sigma_i^2}{\nu_i \sigma_i^2 + (p-2)(\sigma_{min}^2 + \tau^2)} \leq \frac{2\sigma_i^2}{\sigma_i^2 + \tau^2} \ ,$$

which completes the proof. □

Proof of Corollary 3

Proof. As in the proof of Corollary 2, it is sufficient for us to verify that conditions

in Theorem 2 are satisfied by $h_i^+(X) = h_i(X) \wedge 1$, where

$$\begin{aligned} h_i(X) &= \frac{\lambda_i \sigma_i^2}{\nu_i \sigma_i^2 + \|X\|^2} \\ &= \frac{\lambda_i \sigma_i^2}{\nu_i \sigma_i^2 + \sum_{j=1}^p (\sigma_i^2 + \tau^2) T_i W} \\ &= \mathfrak{h}_i(\underline{T}, W) \ . \end{aligned}$$

Conditions (1)-(4) are straightforward. If $\tau^2 \leq \sigma_i^2$, (5) is also automatically satisfied.

Assuming $\tau^2 > \sigma_i^2$, define $g_i(W)$ as

$$g_i(W) = \frac{\lambda_i \sigma_i^2}{\nu_i \sigma_i^2 + (\sigma_{min}^2 + \tau^2) W} \ .$$

Note that $\sup_{\underline{T}} \mathfrak{h}_i(\underline{T}, W) \leq g_i(W)$. Using the covariance inequality we have

$$\begin{aligned} E[g_i(W)] &= E\left[\frac{\lambda_i \sigma_i^2 / W}{\nu_i \sigma_i^2 / W + \sigma_{min}^2 + \tau^2}\right] \leq \frac{E[\lambda_i \sigma_i^2 / W]}{E[\nu_i \sigma_i^2 / W + \sigma_{min}^2 + \tau^2]} \\ &= \frac{\lambda_i \sigma_i^2}{\nu_i \sigma_i^2 + (p-2)(\sigma_{min}^2 + \tau^2)} \ . \end{aligned}$$

Now we only need to show

$$\frac{\lambda_i \sigma_i^2}{\nu_i \sigma_i^2 + (p-2)(\sigma_{min}^2 + \tau^2)} \leq \frac{2\sigma_i^2}{\sigma_i^2 + \tau^2}$$

for $\tau^2 > \sigma_i^2$, which is equivalent to

$$2\sigma_i^2((p-2) - \nu_i) \leq (\sigma_i^2 + \tau^2)(2(p-2) - \lambda_i) + 2(p-2)\sigma_{min}^2 \ .$$

Since $0 \leq \lambda_i \leq 2(p-2)$ and $\nu_i \geq [\lambda_i - (p-2)(1 + \sigma_{min}^2/\sigma^2)]_+$, we have

$$\begin{aligned} 2\sigma_i^2((p-2) - \nu_i) &\leq 2\sigma_i^2(2(p-2) - \lambda_i) + 2(p-2)\sigma_{min}^2 \\ &\leq (\sigma_i^2 + \tau^2)(2(p-2) - \lambda_i) + 2(p-2)\sigma_{min}^2 \ , \end{aligned}$$

which completes the proof. □

Proof of Corollary 5

Proof. Set $\lambda_i = C$ and $\nu_i = C - \frac{\sum_{j=1}^p \sigma_j^2}{\sigma_i^2}$. Condition (3.3.4) guarantees that $\nu_i \geq 0$ and $\lambda_i \leq 2(p-2)$. It is evident that $\delta_{PEB} = \delta_{GS}$. A little care with the positive part conditions shows that also $\delta^+_{PEB} = \delta^+_{GS}$.

It then follows from Corollary 2 that $\delta_{PEB} = \delta_{GS}$ is ensemble minimax if

$$\text{diff} = \nu_i - [\frac{C}{2} - (p-2)\frac{\sigma^2_{min}}{\sigma_i^2}] \geq 0. \tag{6.0.6}$$

Substituting and simplifying yields

$$\begin{aligned} \text{diff} &= \frac{C}{2} - [\frac{\sum_{j=1}^p \sigma_j^2}{\sigma_i^2} - (p-2)\frac{\sigma^2_{min}}{\sigma_i^2}] \\ &\geq \frac{C}{2} - \frac{\sum_{j=1}^p \sigma_j^2 - (p-2)\sigma^2_{min}}{\sigma^2_{min}}, \end{aligned}$$

since $\sum_{j=1}^p \sigma_j^2 \geq p\sigma^2_{min} \geq (p-2)\sigma^2_{min}$. Hence, from (3.3.4),

$$\begin{aligned} \text{diff} &\geq \frac{C}{2} + p - 2 - \frac{\sum_{j=1}^p \sigma_j^2}{\sigma^2_{min}} \\ &\geq \frac{1}{2}[C - \frac{\sum_{j=1}^p \sigma_j^2}{\sigma^2_{min}}] \\ &\geq 0\ . \end{aligned}$$

This verifies (6.0.6) and proves δ_{PEB} is ensemble minimax.

The proof for δ^+_{PEB} is similar, but easier. λ_i and ν_i are defined as before. Condition (3.3.4) is still required in order that $0 \leq \lambda_i \leq 2(p-2)$ and $\nu_i \geq 0$. Truth of (6.0.6) validates the remaining condition in Corollary 3, and hence proves δ^+_{PEB} is ensemble minimax. □

Proof of Corollary 7

Proof. Set $\lambda_i = C$ and $\nu_i = C - \frac{\sum_{j=1}^{p}\gamma_i}{\gamma_{min}}$. Condition (4.1.6) guarantees that $\nu_i \geq 0$ and $\lambda_i \leq \frac{2m(p-2)}{m+2}$. It is evident that $\delta_{PEBV} = \delta_{GSV}$. A little care with the positive part conditions shows that also $\delta^+_{PEBV} = \delta^+_{GSV}$.

It then follows from Theorem 8 that $\delta_{PEBV} = \delta_{GSV}$ is ensemble minimax if

$$\text{diff} = \nu_i - [\frac{m+2}{2(m-2)}C - \frac{m(p-2)\gamma_{min}}{(m-2)\gamma_i}] \geq 0. \tag{6.0.7}$$

Substituting and simplifying yields

$$\begin{aligned}
\text{diff} &= \frac{m-6}{2(m-2)}C + \frac{m(p-2)\gamma_{min}}{(m-2)\gamma_i} - \frac{\sum_{j=1}^{p}\gamma_j}{\gamma_i} \\
&\geq \frac{m-6}{2(m-2)}\frac{\sum_{j=1}^{p}\gamma_j}{\gamma_i} + \frac{m(p-2)\gamma_{min}}{(m-2)\gamma_i} - \frac{\sum_{j=1}^{p}\gamma_j}{\gamma_i} \\
&= \frac{\gamma_{min}}{\gamma_i}[\frac{m(p-2)}{m-2} - \frac{m+2}{2(m-2)}\frac{\sum_{j=1}^{p}\gamma_j}{\gamma_{min}}] \\
&\geq \frac{\gamma_{min}}{\gamma_i}[\frac{m(p-2)}{m-2} - \frac{m+2}{2(m-2)}\frac{2m(p-2)}{m+2}] \\
&= 0\ .
\end{aligned}$$

This verifies (6.0.7) and proves δ_{PEBV} is ensemble minimax.

The proof for δ^+_{PEBV} is similar, but easier. λ_i and ν_i are defined as before. Condition (4.1.6) is still required in order that $0 \leq \lambda_i \leq \frac{2m(p-2)}{m+2}$ and $\nu_i \geq 0$. Truth of (6.0.7) validates the remaining condition in Theorem 9, and hence proves δ^+_{PEBV} is ensemble minimax. □

Bibliography

[1] Basu, D. (1955). On Statistics Independent of a Complete Sufficient Statistic. *Sankhya, Series A*, **15**, 377–380.

[2] Berger, J.O. (1985). *Statistical Decision Theory and Bayesian Analysis*, Springer, New York.

[3] Berger, R.L. (1979). Gamma Minimax Robustness of Bayes Rules. *Communications in Statistics: Theory and Methods*, **8**, 543–560.

[4] Brown, L.D. (1966). On the Admissibility of Invariant Estimators of One or More Location Parameters. *Ananals of Mathematical Statistics*, **37**, 1087–1136.

[5] Brown, L.D. (1975). Estimation with Incomplete Specified Loss Functions (the Case with Several Location Parameters). *Journal of the American Statistical Association*, **70**, 417–427.

[6] Brown, L.D. (2008). In-season Prediction of Batting Averages: a Field Test of Empirical Bayes and Bayes Methodologies. *Annals of Applied Statistics*, **2**, 113–152.

[7] Carter, G.M. and Rolph, J.E. (1974). Empirical Bayes Methods Applied to Estimating Fire Alarm Probabilities. *Journal of the American Statistical Association*, **69**, 880–885.

[8] Casella, G. (1980). Minimax Ridge Regression Estimation. *Ananals of Statistics*, **8**, 1036–1056.

[9] Copas, J.B. (1983). Regression, Prediction and Shrinkage. *Journal of the Royal Statistical Society, Series B (Methodological)*, **45**, 311–354.

[10] Efron, B. and Morris, C. (1971). Limiting the Risk of Bayes and Empirical Bayes Estimators - Part I: Bayes Case. *Journal of the American Statistical Association*, **66**, 807–815.

[11] Efron, B. and Morris, C. (1972a). Limiting the Risk of Bayes and Empirical Bayes Estimators - Part II: The Empirical Bayes Case. *Journal of the American Statistical Association*, **67**, 130–139.

[12] Efron, B. and Morris, C. (1972b). Empirical Bayes on Vector Observations: An Extension of Stein's Method. *Biometrika*, **59**, 335–347.

[13] Efron, B. and Morris, C. (1973). Stein's Estimation Rule and Its Competitors – An Empirical Bayes Approach. *Journal of the American Statistical Association*, **68**, 117–130.

[14] Efron, B. and Morris, C. (1975). Data Analysis Using Stein's Estimator and Its Generalizations. *Journal of the American Statistical Association*, **70**, 311–319.

[15] Fay, R.E.III and Herriot, R.A. (1979). Estimates of Income for Small Places: An Application of James-Stein Procedures to Census Data. *Journal of the American Statistical Association*, **74**, 269–277.

[16] Good, I.J. (1952). Rational Decisions. *Journal of the Royal Statistical Society, Series B (Methodological)*, **14**, 107–114.

[17] Green, J. and Strawderman, W.E. (1985). The Use of Bayes/Empirical Bayes Estimation in Individual Tree Volume Equation Development. *Forest Science*, **31**, 975–990.

[18] Hastie, T., Tibshirani, R., C. and Friedman, J. (2001). *The Elements of Statsitical Learning: Data Mining, Inference and Prediction*, Springer, New York.

[19] James, W. and Stein, C. (1961). Estimation with Quadratic Loss. *Proceedings of 4th Berkeley Symposium on Probability and Statistics*, **I**, 367–379.

[20] Jones, K. (1991). Specifying and Estimating Multi-level Models for Geographical Research *Transactions of the Institute of British Geographers*, **16**, 148–159.

[21] Lindley, D.V. and Smith, A.F.M. (1972). Bayes Estimates for the Linear Model. *Journal of the Royal Statistical Society, Series B (Methodological)*, **34**, 1–41.

[22] Morris, C. (1983a). Parametric Empirical Bayes Inference: Theory and Applications. *Journal of the American Statistical Association*, **78**, 47–55.

[23] Morris, C. (1983b). Parametric Empirical Bayes Confidence Intervals. in *Scientific Inference, Data Analysis, and Robustness*, eds. Box, G.E.P., Leonard, T. and Wu, C.F.J., New York: Academic Press, 25–50.

[24] Morris, C. and Lysy, M. (2009). Shrinkage Estimation in Multi-level Normal Models. *Preprint*

[25] Robbins, H. (1951). Asymptotically Subminimax Solutions of Compound Statistical Decision Problems. *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, **1**, University of California Press, Berkeley.

[26] Robbins, H. (1964). The Empirical Bayes Approach to Statistical Decision Problems. *The Annals of Mathematical Statistics*, **35**, 1–20.

[27] Rubin, D. (1981). Using Empirical Bayes Techniques in the Law School Validity Studies. *Journal of the American Statistical Association*, **75**, 801–827.

[28] Strawderman, W. (1971). Proper Bayes Estimators of the Multivariate Normal Mean. *Annals of Mathematical Statistics*, **42**, 385–388.

Part II: Application In Causal Inference

Inference for the Effect of Treatment on Survival Probability in Randomized Trials with Noncompliance and Administrative Censoring

Chapter 7

Introduction and Background Knowledge

In randomized trials with a survival outcome, two common problems are administrative censoring and noncompliance. Administrative censoring means follow-up ends at a pre-specified date when many subjects have not failed yet. Noncompliance means a subject does not take his or her assigned treatment. A trial that had both administrative censoring and noncompliance is the HIP study, a randomized trial of breast cancer screening (Joffe, 2001). Other examples of trials involving both administrative censoring and noncompliance are Kubik (1990), Follmann (2000) and Oken (2005). When there is noncompliance, in addition to the intent-to-treat effect, it is often of interest to estimate the causal effect of actually receiving the active treatment compared to receiving the control. Knowledge of this effect is useful for

predicting the impact of the treatment in a setting for which compliance patterns might differ from the randomized trial and for scientific understanding of the treatment (Sommer and Zeger, 1991; Sheiner and Rubin, 1995; Cheng and Small, 2006; Small et al., 2006).

There is a lot of literature studying the causal effect of a treatment on a continuous, binary or multinomial outcome when there is noncompliance. A few papers have considered trials with a survival time as the outcome in the presence of noncompliance. Robins and Tsiatis (1991) consider a structural accelerated failure time model in which treatment multiplies the failure time by a constant factor for each subject, and developed semiparametric estimators for this model. Joffe (2001) provides a good discussion of their approach and comparisons with other survival analysis methods. Loeys and Goetghebeur (2003) and Cuzick et al. (2007) consider a structural proportional hazards model in which the hazard of the potential failure time under treatment for a certain group of subjects is proportional to the hazard of the potential failure time under control for these same subjects. Both the structural accelerated failure time model and the structural proportional hazards model are semiparametric models, the parametric part being the effect of the treatment on the distribution of failure times. In this paper we use empirical likelihood (a nonparametric approach) to estimate the effect of treatment on survival at specific times in the presence of non-compliance and administrative censoring. Our work builds on Baker (1998). Baker extends the model and assumptions for settings with

noncompliance of Baker and Lindeman (1994) and Angrist et al. (1996) to discrete-time survival data. He derives closed form expressions for the maximum likelihood estimates of the hazards of compliers (subjects who would only receive the treatment if assigned to the treatment) in the treatment and control groups when these estimates lie in the interior of the parameter space. For the effect of the treatment at a specific time, Baker's estimator is analogous to the standard instrumental variable (IV) estimator in the setting with a survival outcome. However, this estimator can provide negative estimates of hazards (Baker, 1998) and be inefficient in some situations. The reason for the inefficiency is the same as the reason that the standard IV estimator is inefficient in the non-survival setting: standard IV methods do not fully use the mixture structure implied by the latent compliance model (Imbens and Rubin, 1997; Cheng et al., 2009a; Cheng et al., 2009b). The nonparametric approach developed in this paper makes full use of this mixture structure and thus has the potential to be more efficient than the standard IV method.

Chapter 8

Model Framework

8.1 Notation

We assume that the treatment has two levels. Let $\underline{\mathrm{R}}$ be the indicator vector of randomization assignments for all subjects. Its individual element $R_i = r_i \in \{0, 1\}$ indicates the randomization assignment for subject i: $R_i = 1$ if subject i is assigned active treatment (hereafter, it is referred to as 'treatment'), $R_i = 0$ for control. We also let $\underline{A}^{\underline{r}}$ denote the vector of potential treatments received under randomization assignments $\underline{r}$. Its individual element $A_i^{\underline{r}} = a_i \in \{0, 1\}$, is equal to 1 if subject i takes the treatment and 0 if subject i takes the control under the randomization assignments $\underline{r}$.

Let $\underline{T}^{\underline{r},\underline{a}}$ be the vector of the potential failure times under randomization assignments $\underline{r}$ and treatments received $\underline{a}$. Its individual element $T_i^{\underline{r},\underline{a}}$ is the potential

failure time for subject i with the vector of randomization assignments $\underline{r}$ and the vector of treatment received $\underline{a}$. Let $\underline{C}$ denote the vector of administrative censoring times for all subjects with individual element C_i as the administrative censoring time for subject i, i.e., C_i is the time between the date of enrollment for subject i and the prespecified date at which follow-up finishes. Subject i would get censored under randomization assignments $\underline{r}$ and treatments received $\underline{a}$ if $T_i^{\underline{r},\underline{a}} > C_i$. Let $Y_i^{\underline{r},\underline{a}} = \min\{T_i^{\underline{r},\underline{a}}, C_i\}$ denote the length of subject i's follow-up time and let $\Delta_i^{\underline{r},\underline{a}} = I\{T_i^{\underline{r},\underline{a}} \leq C_i\}$ be an indicator of failure for subject i under $\underline{r}$, $\underline{a}$; $\Delta_i^{\underline{r},\underline{a}} = 1$ if failure occurs before censoring and $\Delta_i^{\underline{r},\underline{a}} = 0$ otherwise.

8.2 Assumptions

We make the same five assumptions as Angrist, Imbens and Rubin (1996) made for the non-survival setting and then an additional assumption for the survival setting.

Assumption 1: Stable Unit Treatment Value Assumption (SUTVA) (Rubin, 1978).

a. If $r_i = r_i^{'}$, then $A_i^{\underline{r}} = A_i^{\underline{r}^{'}}$ for all i.

b. If $r_i = r_i^{'}$ and $a_i = a_i^{'}$, then $T_i^{\underline{r},\underline{a}} = T_i^{\underline{r}^{'},\underline{a}^{'}}$ for all i.

The SUVTA assumption allows us to write $T_i^{\underline{r},\underline{a}}$, $Y_i^{\underline{r},\underline{a}}$, $\Delta_i^{\underline{r},\underline{a}}$ and $A_i^{\underline{r}}$ as $T_i^{r_i,a_i}$, $Y_i^{r_i,a_i}$, $\Delta_i^{r_i,a_i}$ and $A_i^{r_i,a_i}$ respectively for subject i.

Assumption 2: Random Assignment

The treatment assignment R_i is random: $Pr(\underline{R} = \underline{u}) = Pr(\underline{R} = \underline{u}^{'})$ for all $\underline{u}$

and $\underline{u}'$ such that $l^T\underline{u} = l^T\underline{u}'$, where l is the vector with all elements equal to one.

Assumption 3: Exclusion Restriction

$T_i^{r_i,a_i} = T_i^{r_i',a_i}$ for all r_i, r_i', a_i and all subjects i.

According to this assumption, the randomization assignment does not affect the potential failure time except through its effect on the treatment received. Thus, we write $T_i^{r_i,a_i}$, $Y_i^{r_i,a_i}$ and $\Delta_i^{r_i,a_i}$ as $T_i^{a_i}$, $Y_i^{a_i}$ and $\Delta_i^{a_i}$ respectively for subject i.

Assumption 4: Nonzero Average Causal Effect of R on A

$E(A_i^1 - A_i^0) \neq 0$.

This assumption requires randomization assignment R to have an effect on the average probability of receiving treatment.

Assumption 5: Monotonicity (Imbens and Angrist, 1994).

$A_i^1 \geq A_i^0$ for all i, which rules out treatment defiers.

Besides Assumptions 1 - 5 from Angrist et al. (1996), we also make the following assumption as in Kaplan and Meier (1958).

Assumption 6: Independence of Failure Times and Censoring Times

The distributions of potential failure times T and administrative censoring times C are independent of each other. Type I censoring (i.e., censoring times are the same for all subjects) and random censoring are two special cases.

8.3 Compliance Classes

A subject in a two-arm trial can be classified into one of four compliance classes: always-takers ($A^1 = 1, A^0 = 1$), who will always take the treatment no matter which group they are assigned to; compliers ($A^1 = 1, A^0 = 0$), who will comply with their assignments; never-takers ($A^1 = 0, A^0 = 0$), who will never take the treatment no matter which group they are assigned to; and defiers ($A^1 = 0, A^0 = 1$), who will do the opposite of their assigned treatment. Note the monotonicity assumption rules out the existence of defiers. First assume

Assumption 7: Only Compliers and Never-takers in Our Model

This occurs in a trial in which only subjects assigned to the treatment have the opportunity to receive the treatment, e.g., a single consent design (Zelen, 1979). Let $I = 1$ if the subject is a complier and $I = 0$ if he or she is a never-taker. Let $\pi_c = P(I = 1)$. We will later extend our results to more general trials with always-takers.

8.4 Model Structure

Under Assumption 7, the compliance status is observed for the treatment group but not for the control group. If $R = 1$ and $A = 1$, we know that the subject is a complier; if $R = 1$ and $A = 0$, we know that the subject is a never-taker. However, in the control group, we cannot tell which group the subject belongs to, and hence

have a mixture of compliers and never-takers in the control arm. We can organize the data as follows, where there are $m_1 + m_2$ subjects in the treatment group and N subjects in the control group:

1. Compliers in the Treatment Group

   For $i = 1, \cdots, m_1$, assume that $R_i = 1$ and $A_i = 1$, which means that they are compliers in the treatment group ($I_i = 1$). Let Y_i denote the observed follow-up time for subject i and Δ_i denote the censoring indicator. Let the n_1 ordered unique failure times corresponding to $\{Y_i\}_{i=1}^{m_1}$ be $0 < T_1^{(1)} < \cdots < T_{n_1}^{(1)} < \infty$, where $n_1 \leq m_1$. For $j = 1, \cdots, n_1$, let $d_j^{(1)}$ be the number of failures at $T_j^{(1)}$ and $r_j^{(1)}$ be the number of subjects at risk of failure just prior to $T_j^{(1)}$.

2. Never-takers in the Treatment Group

   For $i = m_1 + 1, \cdots, m_1 + m_2$, assume that $R_i = 1$ and $A_i = 0$, which means that they are never-takers in the treatment group ($I_i = 0$). Again, let Y_i denote the observed follow-up time for subject i and Δ_i denote the censoring indicator. Let the n_2 ordered unique failure times corresponding to $\{Y_i\}_{i=m_1+1}^{m_1+m_2}$ be $0 < T_1^{(2)} < \cdots < T_{n_2}^{(2)} < \infty$, where $n_2 \leq m_2$. For $j = 1, \cdots, n_2$, let $d_j^{(2)}$ be the number of failures at $T_j^{(2)}$ and $r_j^{(2)}$ be the number of subjects at risk of failure just prior to $T_j^{(2)}$.

3. Mixture in the Control Group

   For $i = m_1 + m_2 + 1, \cdots, m_1 + m_2 + N$, assume that $R_i = 0$, which shows that they are subjects in the control group. Similarly, let the n_3 ordered unique

failure times corresponding to $\{Y_i\}_{i=m_1+m_2+1}^{m_1+m_2+N}$ be $0 < T_1^{(3)} < \cdots < T_{n_3}^{(3)} < \infty$, where $n_3 \leq N$.

We assume that the survival functions of compliers in the treatment group and compliers in the control group at time V are given by $S_{c1}(V)$ and $S_{c0}(V)$ respectively. Never-takers in both the treatment group and control group have the same survival function $S_{nt}(V)$ at time V because of the exclusion restriction (Assumption 3). Furthermore, we let $S_{T|R=1}(V)$ and $S_{T|R=0}(V)$ denote the survival probabilities at time point V of the mixture distribution in the treatment group and that in the control group, respectively. Notice that

$$S_{T|R=1}(V) = \pi_c S_{c1}(V) + (1 - \pi_c) S_{nt}(V)$$

$$S_{T|R=0}(V) = \pi_c S_{c0}(V) + (1 - \pi_c) S_{nt}(V)$$

Chapter 9

Main Results

9.1 Standard Instrumental Variable Estimation

Under Assumptions 1 - 6, the compliers are the only subgroup for which a randomized trial provides information on the causal effect of receiving treatment (Angrist et al., 1996). By Proposition 1 in Angrist et al. (1996), the difference between the survival probability at a time point V of compliers in the treatment group and that in the control group is given by

$$W(V) = \frac{S_{T|R=1}(V) - S_{T|R=0}(V)}{E[A|R=1] - E[A|R=0]}$$

which is the difference of the survival probability at time V in the two arms divided by the proportion of compliers. Under Assumption 7, note that $E[A|R=0] = 0$ since there are no subjects with treatment in the control group. The standard IV estimator is given by substituting estimators of the quantities in the above formula,

i.e.,

$$\hat{W}(V) = \frac{\hat{S}_{T|R=1}(V) - \hat{S}_{T|R=0}(V)}{\hat{E}[A|R=1] - \hat{E}[A|R=0]} \tag{9.1.1}$$

where $\hat{E}$ denotes the sample mean, and $\hat{S}_{T|R=1}(V)$, $\hat{S}_{T|R=0}(V)$ represent the Kaplan-Meier estimators in the treatment arm and the control arm, respectively. In the survival setting, Assumption 6 is needed to ensure that these Kaplan-Meier estimators are consistent. (9.1.1) is also the standard IV estimator for general trials under Assumptions 1 - 6, where always-takers exist in both arms. The estimators in Baker (1998) are equivalent to the standard IV estimators. For the standard IV method in the non-survival setting, Angrist et al. (1996) provides the foundation and Cheng et al. (2009b) discusses various properties.

Although the standard IV estimator is very useful, it does not take full advantages of the mixture structure of the outcomes in two arms. The likelihood approach which uses the mixture information provides considerable efficiency gains over the standard IV estimation.

9.2 Parametric Maximum Likelihood Estimation

The maximum likelihood approach is a powerful tool under a parametric model. We use the EM-algorithm to find the maximum likelihood estimators of the parameters in our model, and then get the maximum likelihood estimator of the difference between the survival probability of the compliers in the treatment group and the survival probability of the compliers in the control group at a specific time V.

We illustrate this method through an example with assumptions of Weibull distributions. However, our method can be easily applied under assumptions of other distributions.

Example

Assume that the compliers in the treatment group, the compliers in the control group and the never-takers have Weibull distributions with parameters ρ_{c1}, κ_{c1}, ρ_{c0}, κ_{c0} and ρ_{nt}, κ_{nt}, respectively. The likelihood function can be written as:

$$L_{obs} = \prod_{i=1}^{m_1} \pi_c L_i^{c1} \prod_{i=m_1+1}^{m_1+m_2} (1-\pi_c) L_i^{nt} \prod_{i=m_1+m_2+1}^{m_1+m_2+N} (\pi_c L_i^{c0} + (1-\pi_c) L_i^{nt})$$

where

$$\begin{aligned} L_i^{c1} &= (\kappa_{c1}\rho_{c1}(\rho_{c1}Y_i)^{\kappa_{c1}-1})^{\Delta_i} \exp\left(-(\rho_{c1}Y_i)^{\kappa_{c1}}\right),\ i = 1, \cdots, m_1 \\ L_i^{c0} &= (\kappa_{c0}\rho_{c0}(\rho_{c0}Y_i)^{\kappa_{c0}-1})^{\Delta_i} \exp\left(-(\rho_{c0}Y_i)^{\kappa_{c0}}\right),\ i = m_1+m_2+1, \cdots, m_1+m_2+N \\ L_i^{nt} &= (\kappa_{nt}\rho_{nt}(\rho_{nt}Y_i)^{\kappa_{nt}-1})^{\Delta_i} \exp\left(-(\rho_{nt}Y_i)^{\kappa_{nt}}\right),\ i = m_1+1, \cdots, m_1+m_2+N \end{aligned}$$

Viewing the compliance class as missing data, the complete data likelihood function is

$$L_c = \prod_{i=1}^{m_1} \pi_c L_i^{c1} \prod_{i=m_1+1}^{m_1+m_2} (1-\pi_c) L_i^{nt} \prod_{i=m_1+m_2+1}^{m_1+m_2+N} (\pi_c L_i^{c0})^{I_i} ((1-\pi_c) L_i^{nt})^{1-I_i} \qquad (9.2.1)$$

We directly find the MLEs for ρ_{c1} and κ_{c1} by solving the two equations below

$$\frac{\sum_{i=1}^{m_1}\Delta_i}{\kappa_{c1}}+\sum_{i=1}^{m_1}\Delta_i\log Y_i = \frac{\sum_{i=1}^{m_1}\Delta_i\sum_{i=1}^{m_1}Y_i^{\kappa_{c1}}\log Y_i}{\sum_{i=1}^{m_1}Y_i^{\kappa_{c1}}}$$

$$\rho_{c1} = (\frac{\sum_{i=1}^{m_1}\Delta_i}{\sum_{i=1}^{m_1}Y_i^{\kappa_{c1}}})^{\frac{1}{\kappa_{c1}}}$$

For the other parameters, we use the EM-algorithm to find their MLEs.

E-step Let $\hat{I}_i$ denote $E(I_i|\hat{\theta}^{(t)})$, where $\hat{\theta}^{(t)}$ is the vector of estimates of π_c, ρ_{c0}, κ_{c0}, ρ_{nt} and κ_{nt} at the tth iteration of the EM-algorithm. For $i = m_1+m_2+1,\cdots,m_1+m_2+N$, we have

$$\hat{I}_i = \frac{\pi_c L_i^{c0}}{\pi_c L_i^{c0}+(1-\pi_c)L_i^{nt}} \tag{9.2.2}$$

where L_i^{c0}, L_i^{nt} and π_c are evaluated at $\hat{\theta}^{(t)}$. For $i = m_1+1,\cdots,m_1+m_2$, $\hat{I}_i = 0$. Because the complete data log-likelihood function is linear in I_i, the expected value of the complete data log-likelihood given $\hat{\theta}^{(t)}$ is obtained by substituting $\hat{I}_i$ into (9.2.1).

M-Step After substituting (9.2.2) into (9.2.1), we get the maximizers of π_c, ρ_{c0}, κ_{c0}, ρ_{nt} and κ_{nt} by setting the first derivatives as zero and solving five equations below.

$$\pi_c = \frac{m_1 + \sum_{i=m_1+m_2+1}^{m_1+m_2+N} \hat{I}_i}{m_1 + m_2 + N}$$

$$\rho_{c0} = \left(\frac{\sum_{i=m_1+m_2+1}^{m_1+m_2+N} \hat{I}_i \Delta_i}{\sum_{i=m_1+m_2+1}^{m_1+m_2+N} \hat{I}_i Y_i^{\kappa_{c0}}}\right)^{\frac{1}{\kappa_{c0}}}$$

$$\rho_{nt} = \left(\frac{\sum_{i=m_1+1}^{m_1+m_2+N} (1-\hat{I}_i) \Delta_i}{\sum_{i=m_1+1}^{m_1+m_2+N} (1-\hat{I}_i) Y_i^{\kappa_{nt}}}\right)^{\frac{1}{\kappa_{nt}}}$$

$$\begin{aligned} & \frac{\sum_{i=m_1+m_2+1}^{m_1+m_2+N} \Delta_i \hat{I}_i}{\kappa_{c0}} + \sum_{i=m_1+m_2+1}^{m_1+m_2+N} \Delta_i \hat{I}_i \log Y_i \\ = & \frac{\sum_{i=m_1+m_2+1}^{m_1+m_2+N} \Delta_i \hat{I}_i \sum_{i=m_1+m_2+1}^{m_1+m_2+N} \hat{I}_i Y_i^{\kappa_{c0}} \log Y_i}{\sum_{i=m_1+m_2+1}^{m_1+m_2+N} \hat{I}_i Y_i^{\kappa_{c0}}} \end{aligned}$$

$$\begin{aligned} & \frac{\sum_{i=m_1+1}^{m_1+m_2+N} \Delta_i (1-\hat{I}_i)}{\kappa_{nt}} + \sum_{i=m_1+1}^{m_1+m_2+N} \Delta_i (1-\hat{I}_i) \log Y_i \\ = & \frac{\sum_{i=m_1+1}^{m_1+m_2+N} \Delta_i (1-\hat{I}_i) \sum_{i=m_1+1}^{m_1+m_2+N} (1-\hat{I}_i) Y_i^{\kappa_{nt}} \log Y_i}{\sum_{i=m_1+1}^{m_1+m_2+N} (1-\hat{I}_i) Y_i^{\kappa_{nt}}} \end{aligned}$$

We run this EM-algorithm until the parameters converge. We choose the starting values as $\tilde{\pi}_c = m_1/(m_1+m_2)$, $\tilde{\rho}_{nt}$ and $\tilde{\kappa}_{nt}$ which maximize $\prod_{i=m_1+1}^{m_1+m_2} L_i^{nt} \times \prod_{i=m_1+m_2+1}^{m_1+m_2+N} (L_i^{nt})^{1-\tilde{\pi}_c}$ as well as $\tilde{\rho}_{c0}$ and $\tilde{\kappa}_{c0}$ which maximize $\prod_{i=m_1+m_2+1}^{m_1+m_2+N} (L_i^{c0})^{\tilde{\pi}_c}$. Then the estimated difference between the survival probability of the compliers in the treatment group and the survival probability of the compliers in the control group at a specific time V is given by

$$\hat{W}(V) = \hat{S}_{c1}(V) - \hat{S}_{c0}(V) = \exp\left(-(\hat{\rho}_{c1}V)^{\hat{\kappa}_{c1}}\right) - \exp\left(-(\hat{\rho}_{c0}V)^{\hat{\kappa}_{c0}}\right)$$

9.3 Nonparametric Empirical Likelihood Estimation

If the parametric assumptions are not correct, then the MLE could be biased. Therefore, in this section, we propose a nonparametric approach based on empirical likelihood (Owen, 2001). Under the exclusion restriction (ER) assumption, the never-takers have the same distribution in both treatment arm and control arm, so it is natural to incorporate this constraint in the empirical likelihood. However, this corresponds to an infinite number of estimating equations and leads to poor performance of empirical likelihood as in the examples in Chapter 10 of Owen (2001). Instead, we impose a finite subset of the infinite set of constraints as in Owen (2001). In particular, we impose the constraint that the survival probabilities of the never-takers at the time point we are focusing on are the same in both arms. Our approach, which we call plug-in nonparametric empirical maximum likelihood estimation (PNEMLE), uses three steps to find an approximation to the empirical maximum likelihood estimator of the difference between the survival probability of the compliers in the treatment group and that in the control group at a specific time V subject to the constraints that (a) the survival probabilities of the never-takers in two arms at the time point we are focusing on are the same and (b) the proportions of compliers are the same in both arms.

- Step I: Estimate $S_{c1}(V)$, $S_{nt}(V)$ and π_c in the treatment group. We estimate

$S_{c1}(V)$, $S_{nt}(V)$ by Kaplan-Meier estimators $\hat{S}_{c1}(V)$, $\hat{S}_{nt}(V)$, and we use the observed fraction of compliers in the treatment group $\hat{\pi}_c = \hat{E}(I = 1|R = 1)$ to estimate π_c.

- Step II: Estimate $S_{c0}(V)$ in the control group. We get our estimator $\hat{S}_{c0}(V)$ by applying the nonparametric empirical likelihood approach to model the distribution of the control group with constraints $S_{nt}(V) = \hat{S}_{nt}(V)$ and $\pi_c = \hat{\pi}_c$.

- Step III: Estimate $W(V) = S_{c1}(V) - S_{c0}(V)$ by $\hat{W}(V) = \hat{S}_{c1}(V) - \hat{S}_{c0}(V)$, where $\hat{S}_{c1}(V)$ and $\hat{S}_{c0}(V)$ are obtained in Step I and Step II.

Since Step I and Step III are straightforward, we only focus on Step II. For $j = 1, \cdots, n_3$, let $p_j^{c0} = P(T = T_j^{(3)} \mid R = 0, I = 1)$ and $p_j^{nt} = P(T = T_j^{(3)} \mid R = 0, I = 0)$. Then the empirical likelihood for the observed data in the control group is given by

$$L_{obs} = \prod_{i=m_1+m_2+1}^{m_1+m_2+N} (\hat{\pi}_c L_i^1 + (1 - \hat{\pi}_c) L_i^2) \tag{9.3.1}$$

where

$$L_i^1 = \{\sum_{j=1}^{n_3} p_j^{c0} I(T_j^{(3)} = Y_i)\}^{\Delta_i} \{\sum_{j=1}^{n_3} p_j^{c0} I(T_j^{(3)} > Y_i)\}^{1-\Delta_i}$$
$$L_i^2 = \{\sum_{j=1}^{n_3} p_j^{nt} I(T_j^{(3)} = Y_i)\}^{\Delta_i} \{\sum_{j=1}^{n_3} p_j^{nt} I(T_j^{(3)} > Y_i)\}^{1-\Delta_i}$$

Our constraints are given by

$$\sum_{j=1}^{n_3} p_j^{nt} I(T_j^{(3)} \geq V) = \hat{S}_{nt}(V) \tag{9.3.2}$$

$$\sum_{j=1}^{n_3} p_j^{c0} \leq 1 \tag{9.3.3}$$

$$\sum_{j=1}^{n_3} p_j^{nt} \leq 1 \tag{9.3.4}$$

$$p_j^{c0} \geq 0, j = 1, \cdots, n_3 \tag{9.3.5}$$

$$p_j^{nt} \geq 0, j = 1, \cdots, n_3 \tag{9.3.6}$$

We want to maximize (9.3.1) under the constraints (9.3.2) - (9.3.6). Since failure times and censoring times are independent, we can use the hazard function to get the equivalent form of our optimization problem. For $j = 1, \cdots, n_3$, let $\lambda_j = P(T = T_j^{(3)} \mid T \geq T_j^{(3)}, R = 0, I = 1)$ and $\xi_j = P(T = T_j^{(3)} \mid T \geq T_j^{(3)}, R = 0, I = 0)$. Then the optimization function (9.3.1) with constraints (9.3.2) - (9.3.6) has the equivalent form

$$L_{obs} = \prod_{i=m_1+m_2+1}^{m_1+m_2+N} (\hat{\pi}_c L_i^3 + (1 - \hat{\pi}_c) L_i^4) \tag{9.3.7}$$

where

$$L_i^3 = \{ \prod_{j:T_j^{(3)} < Y_i} (1 - \lambda_j) - \prod_{j:T_j^{(3)} \leq Y_i} (1 - \lambda_j) \}^{\Delta_i} \{ \prod_{j:T_j^{(3)} \leq Y_i} (1 - \lambda_j) \}^{1-\Delta_i}$$

$$L_i^4 = \{ \prod_{j:T_j^{(3)} < Y_i} (1 - \xi_j) - \prod_{j:T_j^{(3)} \leq Y_i} (1 - \xi_j) \}^{\Delta_i} \{ \prod_{j:T_j^{(3)} \leq Y_i} (1 - \xi_j) \}^{1-\Delta_i}$$

subject to

$$\prod_{j:T_j^{(3)}<V} (1-\xi_j) = \hat{S}_{nt}(V) \tag{9.3.8}$$

$$0 \leq \lambda_j \leq 1, j = 1, \cdots, n_3 \tag{9.3.9}$$

$$0 \leq \xi_j \leq 1, j = 1, \cdots, n_3 \tag{9.3.10}$$

We want to use the EM-algorithm to solve this mixture problem. Here, our complete data likelihood function is given by

$$L_c = \prod_{i=m_1+m_2+1}^{m_1+m_2+N} (\hat{\pi}_c L_i^3)^{I_i} \{(1-\hat{\pi}_c) L_i^4\}^{1-I_i} \tag{9.3.11}$$

E-Step Since the complete data log-likelihood is linear in I_i, the E-step just involves substituting $\hat{I}_i = \hat{E}(I_i|\lambda_j^{(t)}, \xi_j^{(t)})$ into (9.3.11), where for all $i = m_1+m_2+1, \cdots, m_1+m_2+N$, $\hat{I}_i$ is given by

$$\hat{I}_i = \frac{\hat{\pi}_c L_i^3}{\hat{\pi}_c L_i^3 + (1-\hat{\pi}_c) L_i^4} \tag{9.3.12}$$

M-Step After plugging (9.3.12) into (9.3.11), we can get the maximizers of λ and ξ as below.

(1) Maximizers of λ_j, $\forall j = 1, \cdots, n_3$

For each Y_i, we consider it as $\hat{I}_i$ "subjects" instead of one subject. Let d_j^{c0} be the number of failures at $T_j^{(3)}$, and r_j^{c0} be the number of subjects at risk of failure just prior to $T_j^{(3)}$. Then the maximization problem is equivalent to maximizing

$$\prod_{j=1}^{n_3} \lambda_j^{d_j^{c0}} (1-\lambda_j)^{r_j^{c0}-d_j^{c0}} \tag{9.3.13}$$

Therefore, the maximizer of λ_j is given by

$$\hat{\lambda}_j = d_j^{c0}/r_j^{c0} \tag{9.3.14}$$

(2) Maximizers of ξ_i, $\forall i = 1, \cdots, n_3$

For each Y_i, we consider it as $1 - \hat{I}_i$ "subjects" instead of one subject. Let d_j^{nt} be the number of failures at $T_j^{(3)}$, and r_j^{nt} be the number of subjects at risk of failure just prior to $T_j^{(3)}$. Then our maximization problem is equivalent to maximizing

$$\prod_{j=1}^{n_3} \xi_j^{d_j^{nt}} (1 - \xi_j)^{r_j^{nt} - d_j^{nt}} \tag{9.3.15}$$

subject to

$$\prod_{j:T_j^{(3)} < V} (1 - \xi_j) = \hat{S}_{nt}(V) \tag{9.3.16}$$

$$0 \leq \xi_j \leq 1, j = 1, \cdots, n_3 \tag{9.3.17}$$

We easily use the Lagrange Multiplier method to solve this optimization problem. The maximizer of ξ_j is given by

$$\hat{\xi}_j = \begin{cases} \frac{d_j^{nt}}{r_j^{nt}}, & if \quad T_j^{(3)} \geq V \\ \frac{d_j^{nt}}{r_j^{nt} - \alpha}, & if \quad T_j^{(3)} < V \end{cases} \tag{9.3.18}$$

where α is uniquely obtained by solving

$$\log \hat{S}_{nt}(V) - \sum_{j:T_j^{(3)} < V} \log \left(1 - \frac{d_j^{nt}}{r_j^{nt} - \alpha}\right) = 0$$

Theoretical properties in Section 9.5 show that the EM sequence converges to the unique global maximum no matter where we start our algorithm. One possible

way to get the initial values is to run the maximization step by assuming $\hat{I}_i = m_1/(m_1+m_2)$. From Theorem 12, the EM sequence converges to the unique global maximum. We repeat this EM-algorithm until $\{\lambda_j^{(t)}\}_{j=1}^{n_3}$ and $\{\xi_j^{(t)}\}_{j=1}^{n_3}$ converge. Assume that $\{\lambda_j^{(t)}\}_{j=1}^{n_3}$ converge to $\{\lambda_j^{MLE}\}_{j=1}^{n_3}$, and $\{\xi_j^{(t)}\}_{j=1}^{n_3}$ converge to $\{\xi_j^{MLE}\}_{j=1}^{n_3}$. Then the estimator of $S_{c0}(V)$ is given by

$$\hat{S}_{c0}(V) = \prod_{j:T_j^{(3)}<V} (1-\lambda_j^{MLE}) \tag{9.3.19}$$

In summary, our EM-algorithm in Step II is described as below:

E-Step Estimate I_i through (9.3.12).

M-Step Estimate λ_i and ξ_i through (9.3.14) and (9.3.18).

9.4 Extension to Trials under Assumptions 1 - 6

In this chapter, we extend our PNEMLE approach to more general trials under Assumptions 1 - 6 in which the control group has access to the treatment. For such trials, we have one more compliance class, the always-takers, in addition to the compliers and the never-takers. If $R = 1$ and $A = 1$, we know that the subject is either a complier or an always-taker; if $R = 1$ and $A = 0$, the subject is a never-taker; if $R = 0$ and $A = 1$, the subject is an always-taker; and if $R = 0$ and $A = 0$, the subject is either a complier or a never-taker. The always-takers are identifiable in the control group and the never-takers are identifiable in the treatment group. Let $S_{at}(V)$, $S_{nt}(V)$, $S_{c1}(V)$ and $S_{c0}(V)$ be the survival probability of the always-

takers, the never-takers, the compliers in the treatment group and the compliers in the control group at time point V, respectively. We also use π_{at}, π_{nt} and π_c to denote the proportion of the always-takers, the never-takers and the compliers.

Similar to the approach in Section 9.3 of our paper, we follow five steps below to estimate the difference $W(V) = S_{c1}(V) - S_{c0}(V)$ for trials under Assumptions 1 - 6.

- Step I: Estimate $S_{at}(V)$ and π_{at} from the always-takers (R=0, A=1) in the control group (R=0). We estimate $S_{at}(V)$ by the Kaplan-Meier estimator $\hat{S}_{at}(V)$, and we use the observed fraction of compliers in the treatment group $\hat{\pi}_{at} = \hat{E}(I = 1|R = 0)$ to estimate π_{at}.

- Step II: Estimate $S_{nt}(V)$ and π_{nt} from the never-takers (R=1, A=0) in the treatment group (R=1). We estimate $S_{nt}(V)$ by the Kaplan-Meier estimator $\hat{S}_{nt}(V)$, and we use the observed fraction of compliers in the treatment group $\hat{\pi}_{nt} = \hat{E}(I = 0|R = 1)$ to estimate π_{nt}.

- Step III: Estimate $S_{c1}(V)$ from the mixture of the compliers and the always-takers (R=1, A=1) in the treatment group (R=1). We get our estimator $\hat{S}_{c1}(V)$ by applying the nonparametric empirical likelihood approach to model the distribution of the mixture of the compliers and the always-takers in the control group with constraints $S_{at}(V) = \hat{S}_{at}(V)$, $\pi_{at} = \hat{\pi}_{at}$ and $\pi_c = 1 - \hat{\pi}_{at} - \hat{\pi}_{nt}$.

- Step IV: Estimate $S_{c0}(V)$ from the mixture of the compliers and the never-takers (R=0, A=0) in the control group (R=0). We get our estimator $\hat{S}_{c0}(V)$ by applying the nonparametric empirical likelihood approach to model the distribution of the mixture of the compliers and the never-takers in the control group with constraints $S_{nt}(V) = \hat{S}_{nt}(V)$, $\pi_{nt} = \hat{\pi}_{nt}$ and $\pi_c = 1 - \hat{\pi}_{at} - \hat{\pi}_{nt}$.

- Step V: Estimate $W(V) = S_{c1}(V) - S_{c0}(V)$ by $\hat{W}(V) = \hat{S}_{c1}(V) - \hat{S}_{c0}(V)$, where $\hat{S}_{c1}(V)$ and $\hat{S}_{c0}(V)$ are obtained in Step III and Step IV.

9.5 Theoretical Properties of PNEMLE

In this section, we establish theoretical properties of our PNEMLE estimator, which are proved in Chapter 12.

9.5.1 Existence and Uniqueness

Theorem 11. *For any specific time V and plug-in $\hat{S}_{nt}(V)$, $\hat{\pi}_c$, the maximization problem (9.3.1) under constraints (9.3.2) - (9.3.6) has a unique global maximum.*

Proof. See Chapter 12. □

9.5.2 Convergence of EM-algorithm

Theorem 12. *The EM sequence converges to the unique global maximum of the maximization problem in Theorem 11.*

Proof. See Chapter 12. □

9.5.3 Asymptotic Consistency

We discuss the asymptotic consistency of PNEMLE. For a specific time V, let $r_v^{(1)}$, $r_v^{(2)}$ and r_v^G be the number of subjects at risk of failure just prior to V in compliers in the treatment group, never-takers in the treatment group and the mixture in the control group respectively. Let G be the distribution of the mixture in the control group, that is, $G = \pi_c F_{c0} + (1 - \pi_c)F_{nt}$. We also assume that the distributions of compliers and never-takers in the control group overlap at least minimally, which means that the never-takers in the control group can be neither all in the lower $1 - \pi_c$ quantile nor all in the upper $1 - \pi_c$ quantile. This implies that

$$\frac{S_G(V) - \pi_c}{1 - \pi_c} I_{(V \le G^{-1}(1-\pi_c))} < S_{nt} < 1 + \frac{S_G(V) - (1 - \pi_c)}{1 - \pi_c} I_{(V \ge G^{-1}(\pi_c))} \tag{9.5.1}$$

Asymptotic consistency is shown in the following theorem.

Theorem 13. *If (9.5.1) and the conditions below are satisfied*

(i) $r_v^{(1)} \to \infty, r_v^{(2)} \to \infty, r_v^G \to \infty,$ *as* $m_1 + m_2 + N \to \infty$

(ii) $\frac{m_1 + m_2}{N} \to c,$ *as* $m_1 + m_2 + N \to \infty,$ *where* c *is a finite constant*

(iii) $0 < \pi_c < 1$

then we have that

$$\hat{S}_{c1}(V) - \hat{S}_{c0}(V) \xrightarrow{P} S_{c1}(V) - S_{c0}(V), \text{ as } m_1 + m_2 + N \to \infty$$

Proof. See Chapter 12. □

9.6 Estimation of Confidence Intervals via Bootstrap Method

PNEMLE provides us a powerful tool to obtain the point estimate of the difference between the survival probability of compliers in the treatment group and that of compliers in the control group. However, we are interested in not only the point estimate but also the confidence interval (CI). Efron and Tibshirani (1994) suggest using the Bootstrap Method to obtain the confidence interval for censored data sets $\{(Y_i, \Delta_i)\}_{i=1}^{m_1+m_2+N}$. We can construct the CI based on bootstrap percentiles following the steps below.

- Step I: Draw a Bootstrap sample $\{(Y_i^*, \Delta_i^*)\}_{i=1}^{m_1+m_2+N}$. For compliers in the treatment group $\{(Y_i, \Delta_i)\}_{i=1}^{m_1}$, we sample with replacement by putting mass $\frac{1}{m_1}$ at each point (Y_i, Δ_i) in order to get Bootstrap sample $\{(Y_i^*, \Delta_i^*)\}_{i=1}^{m_1}$. For never-takers in the treatment group $\{(Y_i, \Delta_i)\}_{i=m_1+1}^{m_1+m_2}$, we sample with replacement by putting mass $\frac{1}{m_2}$ at each point (Y_i, Δ_i) in order to get Bootstrap sample $\{(Y_i^*, \Delta_i^*)\}_{i=m_1+1}^{m_1+m_2}$. For mixtures in the control group $\{(Y_i, \Delta_i)\}_{i=m_1+m_2+1}^{m_1+m_2+N}$, we sample with replacement by putting mass $\frac{1}{N}$ at each point (Y_i, Δ_i) in order to get Bootstrap sample $\{(Y_i^*, \Delta_i^*)\}_{i=m_1+m_2+1}^{m_1+m_2+N}$. We join the three Bootstrap samples together to get a Bootstrap sample $\{(Y_i^*, \Delta_i^*)\}_{i=1}^{m_1+m_2+N}$.

- Step II: Estimate PNEMLE $\hat{W}^*(V)$ for this Bootstrap sample following the procedures in Section 9.3.

- Step III: Independently repeat steps I and II B times and obtain $\{\hat{W}_b^*(V)\}_{b=1}^B$. Find the lower $\frac{\alpha}{2}$ percentile $\hat{W}_{LOW}^*(V)$ and the upper $\frac{\alpha}{2}$ percentile $\hat{W}_{UP}^*(V)$. The $(1-\alpha)$ confidence interval is given by $(\hat{W}_{LOW}^*(V), \hat{W}_{UP}^*(V))$.

We can also use the BC_a method (a bias corrected version of the bootstrap percentile method). Please refer to Chapter 12 for details. As we see in Section 9.7, both types of Bootstrap CIs have reasonably good coverage probability.

9.7 Simulation Studies

In this chapter, we conduct simulation studies to compare our PNEMLE method to the standard IV estimation and the Weibull parametric estimation under various outcome distributions (see Table 9.1) and $\pi_c = 0.5$ (see Table 9.2) or $\pi_c = 0.2$ (see Table 9.3). In Table 9.1, the density function of Weibull distribution with ρ and κ is $\rho\kappa(\rho x)^{\kappa-1}\exp\left(-(\rho x)^{\kappa}\right)$, the density function of lognormal distribution with μ and σ is given by $\frac{1}{x\sigma\sqrt{2\pi}}e^{-(log(x)-\mu)^2/2\sigma^2}$, the density function of loglogistic distribution with a and s is given by $a(x/s)^a/(x(1+(x/s)^a)^2)$, and the density function of gamma distribution with κ and θ is given by $x^{\kappa-1}\frac{\exp(-x/\theta)}{\Gamma(\kappa)\theta^{\kappa}}$. In all settings, we set the probability of being assigned to treatment as π_c, and results are obtained from 1000 simulated data sets with a sample size of $2K(K = 100$ or $200)$. The administrative censoring time C is uniformly distributed on the interval $[C_0, C_0 + \Delta C]$. For each setting, we consider three values of V: close to zero, in the middle, and close to $C_0 + \Delta C$. We only present results from single consent trials, in which we only have

compliers and never-takers in both arms. However, the method can be directly extended to more general trials.

Before discussing our simulation results, we consider the factors likely to affect the size of the efficiency gain of our PNEMLE estimator over the standard IV estimator. Let C_{max} denote the maximum censoring time in the control group. The PNEMLE estimator is $\hat{S}_{c1}(V) - \hat{S}_{c0}(V)$. Similarly, the standard IV estimator is written as $\hat{S}_{c1}(V) - \tilde{S}_{c0}(V)$, where

$$\tilde{S}_{c0}(V) = \frac{\hat{S}_{T|R=0}(V) - \hat{S}_{T|R=1}(V) + \hat{\pi}_c \hat{S}_{c1}(V)}{\hat{\pi}_c} \tag{9.7.1}$$

(see Imbens and Rubin, 1997). The estimate of $S_{c0}(V)$ that PNEMLE uses, $\hat{S}_{c0}(V)$, is always between 0 and 1 because of the constraints applied but the estimate of $S_{c0}(V)$ that standard IV uses, $\tilde{S}_{c0}(V)$, might not be between 0 and 1. When $\tilde{S}_{c0}(V)$ is not between 0 and 1, we expect PNEMLE to be a better estimate than standard IV because PNEMLE incorporates the knowledge that $0 \leq S_{c0}(V) \leq 1$ whereas standard IV is implicitly based on an estimate of $S_{c0}(V)$ that is not between 0 and 1. For similar reasoning in the non-survival setting, see Imbens and Rubin (1997) and Cheng et al. (2009b). Three factors which affect the probability that $0 \leq \tilde{S}_{c0}(V) \leq 1$ are (a) sample size; (b) the time point V and (c) the proportion of compliers π_c. Asymptotically $\tilde{S}_{c0}(V) \xrightarrow{P} S_{c0}(V)$. Consequently, it is more likely that $\tilde{S}_{c0}(V)$ escapes from the interval 0 to 1 for small samples. Similarly, $\tilde{S}_{c0}(V)$ is more likely to escape from [0,1] when $S_{c0}(V)$ is near to 0 or near to 1, which will tend to happen as V gets closer to zero or possibly C_{max}. The variance of $\tilde{S}_{c0}(V)$ is

Table 9.1: Outcome Distributions of the Simulation Studies.

Group	Compliers with Treatment	Compliers with Control	Never-Takers
E	Exponential with hazard 0.6	Exponential with hazard 1.5	Exponential with hazard 0.3
W	Weibull with $\rho = 0.67$, $\kappa = 1.2$	Weibull with $\rho = 2$, $\kappa = 0.8$	Weibull with $\rho = 1$, $\kappa = 0.8$
LN	Lognormal with $\mu = 2$, $\sigma = 1$	Lognormal with $\mu = 3$, $\sigma = 1$	Lognormal with $\mu = 1$, $\sigma = 1$
LL	Loglogistic with $a = 2$, $s = 1.5$	Loglogistic with $a = 1$, $s = 0.5$	Loglogistic with $a = 1.5$, $s = 2$
G	Gamma with $\kappa = 2$, $\theta = 0.5$	Gamma with $\kappa = 3$, $\theta = 0.5$	Gamma with $\kappa = 1$, $\theta = 1$

approximately proportional to $\frac{1}{\pi_c^2}$ for fixed sample size, $S_{T|R=1}(V)$, $S_{T|R=0}(V)$ and $S_{c1}(V)$, so that $\tilde{S}_{c0}(V)$ is more likely to escape from [0,1] when π_c is small. Thus, we expect PNEMLE to gain more over standard IV for small sample sizes, when V is closer to the boundaries of [0, C_{max}] and when π_c is small, because in these settings, $\tilde{S}_{c0}(V)$ is more likely to escape from [0,1].

Table 9.2 shows the relative biases, i.e., ((estimated-true)/true)*100%, and the root mean squared errors (RMSE) under various outcome distributions with $\pi_c = 0.5$. Note that "*" sign in Table 9.2 means that the difference of the RMSEs between PNEMLE and IV method in that row is significant at 5% level. The Monte-Carlo estimate of the SE of differences in the RMSE is estimated using the delta method (see Chapter 12). We see from Table 9.2 that both PNEMLE and the standard IV method provide approximately unbiased estimates for each V — the relative bias is at most 11.1% and is much less in most simulation results. PNEMLE is always at least as efficient as the standard IV method, and PNEMLE is sometimes much more efficient than the standard IV method, gaining as much as 28% in RMSE. The gain in efficiency for PNEMLE is bigger when V is close to zero or $C_0 + \Delta C$, and is bigger for the smaller sample size $K = 100$ than $K = 200$.

Weibull parametric estimation provides an approximately unbiased estimator with smaller RMSE compared to the other two methods when the outcome distribution is actually Weibull (note that the exponential is a special case of Weibull). However, it could have large bias and large RMSE when the underlying distribution is not Weibull. In summary, PNEMLE gives us a more efficient nonparametric estimation method than the standard IV method and avoids the potential for large bias and RMSE of a mis-specified parametric model.

To examine the effect of π_c on efficiency and bias, we conduct a similar simulation but change π_c from 0.5 to 0.2. Again, "*" sign in Table 9.3 means that the difference of the RMSEs between PNEMLE and IV method in that row is significant at 5% level. The Monte-Carlo estimate of the SE of differences in the RMSE is estimated using the delta method (see Chapter 12). Simulation results in Table 9.3 shows that PNEMLE has a little more efficiency gain but larger relative bias compared to results with $\pi_c = 0.5$ in Table 9.2. Considerable literature on IV estimators for uncensored outcomes shows that the estimators have a finite sample bias which asymptotically goes to zero (Nagar, 1959; Bound et al., 1995; Stock et al., 2002). The literature shows that the finite sample bias is approximately inversely proportional to the concentration parameter (see discussion below equation (12) in Bound et al., 1995), which in our setting is $\frac{2K\pi_c}{1-\pi_c}$. This is consistent with the results in Tables 9.2 and 9.3, where the bias is larger when $\pi_c = .2, K = 200$ and the concentration parameter is 100 than the bias when $\pi_c = .5, K = 100$ and the concentration parameter is 200.

Table 9.2: Estimates of the Difference between $S_{c1}(V)$ and $S_{c0}(V)$ when $\pi_c = 0.5$

						Relative bias with			RMSE with		
Group	V	C_0	ΔC	K	True	PNEMLE	IV	Para	PNEMLE	IV	Para
E	0.1	2	0.2	100	0.081	-0.446	1.45	-6.57	0.0736*	0.0757	0.0546
	1	2	0.2	100	0.326	-2.13	3.24	1.63	0.143	0.150	0.125
	2.1	2	0.2	100	0.241	-8.34	4.03	-3.80	0.107*	0.141	0.0854
	0.1	2	0.2	200	0.081	-0.909	-0.767	-3.41	0.0508	0.0510	0.0400
	1	2	0.2	200	0.326	2.43	2.59	3.17	0.102	0.103	0.0844
	2.1	2	0.2	200	0.241	-5.88	-0.563	-3.79	0.0825	0.104	0.0612
W	0.15	2	0.2	100	0.256	-0.313	-3.84	-2.66	0.110*	0.124	0.102
	1	2	0.2	100	0.363	-0.0141	1.00	1.29	0.133	0.139	0.116
	2.05	2	0.2	100	0.186	-6.42	2.10	-5.01	0.0890*	0.105	0.0807
	0.15	2	0.2	200	0.256	0.608	0.910	-1.25	0.0764	0.0779	0.0676
	1	2	0.2	200	0.363	-0.334	-0.262	1.26	0.0928	0.0933	0.0823
	2.05	2	0.2	200	0.186	-3.12	0.886	-2.87	0.0629*	0.0711	0.0543
LN	4	30	2	100	-0.216	-11.1	6.39	-21.7	0.112*	0.155	0.0821
	16	30	2	100	-0.370	-0.118	-0.118	0.959	0.124	0.124	0.108
	31	30	2	100	-0.256	2.02	2.04	12.4	0.0891	0.0891	0.0898
	4	30	2	200	-0.216	-7.04	0.597	-21.9	0.0803	0.0982	0.0659
	16	30	2	200	-0.370	1.07	1.07	0.741	0.0906	0.0906	0.0774
	31	30	2	200	-0.256	0.243	0.265	12.4	0.0645	0.0645	0.0684
LL	0.04	2.5	0.2	100	0.0734	2.57	6.15	2.44	0.0402	0.0403	0.0310
	1	2.5	0.2	100	0.359	1.66	1.85	16.7	0.139	0.141	0.169
	2.6	2.5	0.2	100	-0.270	-3.34	1.49	58.0	0.126*	0.133	0.128
	0.04	2.5	0.2	200	0.0734	1.69	5.06	5.08	0.0275	0.0275	0.0213
	1	2.5	0.2	200	0.359	0.366	0.366	14.7	0.0978	0.0978	0.134
	2.6	2.5	0.2	200	-0.270	-4.63	-3.51	68.1	0.0888	0.0907	0.104
G	0.4	2	0.2	100	-0.144	-7.53	4.17	1.46	0.110*	0.128	0.0934
	1	2	0.2	100	-0.271	-1.10	-0.793	3.39	0.143	0.145	0.123
	2.1	2	0.2	100	-0.132	1.87	2.06	0.132	0.0974	0.0982	0.0914
	0.4	2	0.2	200	-0.144	-6.86	2.50	1.89	0.0687*	0.0827	0.0498
	1	2	0.2	200	-0.271	1.41	1.41	2.15	0.102	0.102	0.0847
	2.1	2	0.2	200	-0.132	2.26	2.45	0.004	0.0717	0.0717	0.0660

Table 9.3: Estimates of the Difference between $S_{c1}(V)$ and $S_{c0}(V)$ when $\pi_c = 0.2$

						Relative bias with			RMSE with		
Group	V	C_0	ΔC	K	True	PNEMLE	IV	Para	PNEMLE	IV	Para
W	0.15	2	0.2	100	0.256	11.9	45.2	-5.72	0.260*	0.604	0.246
	1	2	0.2	100	0.363	-14.1	5.56	-10.1	0.270*	0.370	0.263
	2.05	2	0.2	100	0.186	-33.2	5.71	-25.9	0.207*	0.279	0.197
	0.15	2	0.2	200	0.256	2.22	17.8	-3.79	0.193*	0.339	0.194
	1	2	0.2	200	0.363	-7.87	0.703	-6.54	0.203	0.250	0.188
	2.05	2	0.2	200	0.186	-19.0	3.54	-13.9	0.142	0.193	0.130
LN	4	30	2	100	-0.216	-59.2	-11.5	-19.4	0.269*	0.478	0.115
	16	30	2	100	-0.370	3.74	4.12	24.8	0.236	0.257	0.207
	31	30	2	100	-0.256	2.69	2.81	25.5	0.170	0.172	0.168
	4	30	2	200	-0.216	-30.6	-10.1	-10.2	0.185*	0.272	0.0792
	16	30	2	200	-0.370	1.64	4.41	21.7	0.173	0.182	0.171
	31	30	2	200	-0.256	0.518	0.518	20.7	0.123	0.123	0.130

We conducted a simulation study to check the coverage probability of 95% bootstrap CI. We set $\pi_c = 0.5$, $K = 100$ and simulated 1000 data sets with $2K$ bootstrap samples for each data set. Table 9.4 shows that both types of Bootstrap CIs have reasonably good coverage probability.

Table 9.4: Coverage Probability of the 95% Bootstrap Confidence Intervals for PNEMLE

Group	V	C_0	ΔC	Percentile	BC_α
W	0.15	2	0.2	92.5%	93.0%
	1	2	0.2	93.5%	92.3%
	2.05	2	0.2	91.9%	91.7%
LN	4	30	2	93.7%	94.1%
	16	30	2	92.4%	93.1%
	31	30	2	95.8%	94.9%

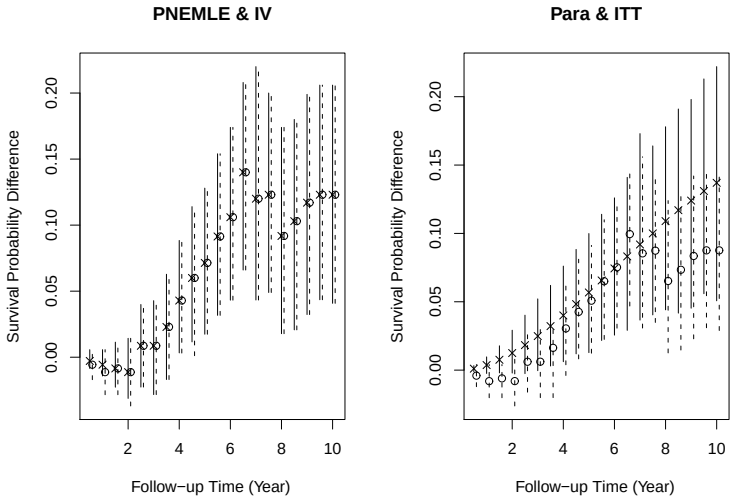
Chapter 10

Application to HIP Study

The HIP study was a randomized trial that began in 1963 and was aimed at examining the effects of periodic screening on breast cancer mortality. More than 60,000 women were randomized into two groups at the beginning of the study. Women in the treatment arm received an initial screening examination and three annual follow-up visits. Women in the control group received usual care. There is a lot of literature studying this data set. Joffe (2001) uses G-estimation of the accelerated failure time model (AFTM) with artificial censoring to analyze the data and concluded that screening increased the mean time to death from breast cancer by 22% with 95% confidence interval $(5.1\%, 63.2\%)$. However, this approach depends on a parametric model for how the treatment affects failure time for its validity. Baker (1998) extends the noncompliance setting to survival outcomes and estimated that screening saved $\$16,000$ cost-effectiveness per life year with 95% confidence inter-

val ($10,000, $51,000). Baker's estimate is equivalent to the standard IV estimate. Similar to Joffe (2001), we consider the first 10 years of each woman's follow-up, because this reduces the attenuation of the effects of the screening in the initial three years by later periods in which both groups receive the same treatments. Therefore, the administrative censoring times for all subjects are 10 years in our study. Furthermore, we follow Baker (1998) in conducting a limited mortality analysis by only considering data from subjects whose breast cancer was diagnosed within the first 7 years of study. Note that our sample is different from the samples of Joffe and Baker because we combine their sample selection rules. We provide estimates of the difference of survival probability of compliers in the treatment group and that in the control group for every half year as well as the 95% confidence intervals. In the left panel of Figure 10.1, we show the results of PNEMLE (cross, with solid line confidence interval) and the standard IV (circle, with dashed line confidence interval). For the right panel of Figure 10.1, we show the results of Parametric Weibull (cross, with solid line confidence interval) and ITT (circle, with dashed line confidence interval). We contrast PNEMLE with several other approaches. From the left Panel of Figure 10.1, we find that both PNEMLE and standard IV provide similar estimates and CIs at each time point. The 95% CIs are strictly above zero after 4 years, which means that there is strong evidence that the treatment has a beneficial effect for compliers after 4 years. Compliers who received treatment have 12.3% (4.08%, 20.6%) higher probability to survive over 10 years than those who

Figure 10.1: Results from HIP study.



received control. The right Panel of Figure 10.1 shows the parametric estimates under Weibull assumption and the intent-to-treat (ITT) estimates as well as 95% CIs. The parametric estimator would be more efficient if we knew that the underlying distribution of failure times is Weibull. However, from the discussion in Chapter 12, there is evidence to cast doubt on the validity of the Weibull model for never-takers in the treatment arm. The ITT estimates the effect of assignment to treatment on survival in contrast to the PNEMLE which estimates the effect of actually receiving treatment on survival. The ITT estimates are substantially smaller than the PNEMLE estimates, meaning that there is a substantial amount of noncompliance.

Chapter 11

Conclusion

In this paper, we develop a more efficient nonparametric method than the standard IV to estimate the difference between the survival probability of the compliers in the treatment group and that in the control group at some specific time. PNEMLE does not rely on parametric assumptions, which is an advantage over the parametric method and accelerated failure time model. An interesting problem for future work is to estimate the whole distribution of potential failure times through a nonparametric approach. In addition, under the current setting, we assume that distributions of censoring times and failure times are independent. Estimation of the causal effect under dependence is a potential research topic for further studies. The first step is to extend this method to the cases when censoring is at random given baseline covariates. It will also be important to investigate the cases where the probability of the compliance depends on a baseline covariate such as co-morbidity.

Chapter 12

Proofs and Supplementary Materials

Proof of Theorem 11

Proof. We prove this theorem by showing that it is a convex optimization problem. The maximization problem (9.3.1) under the constraints (9.3.2) - (9.3.6) is equivalent to minimizing

$$-\log L_{obs} = \sum_{i=m_1+m_2+1}^{m_1+m_2+N} -\log\left(\hat{\pi}_c L_i^1 + (1-\hat{\pi}_c)L_i^2\right) \tag{12.0.1}$$

subject to

$$\sum_{j=1}^{n_3} p_j^{nt} I(T_j^{(3)} \geq V) = \hat{S}_{nt}(V) \tag{12.0.2}$$

$$-1 + \sum_{j=1}^{n_{(3)}} p_j^{c0} \leq 0 \tag{12.0.3}$$

$$-1 + \sum_{j=1}^{n_{(3)}} p_j^{nt} \leq 0 \tag{12.0.4}$$

$$-p_j^{c0} \leq 0, j = 1, \cdots, n_3 \tag{12.0.5}$$

$$-p_j^{nt} \leq 0, j = 1, \cdots, n_3 \tag{12.0.6}$$

Here, L_i^1 and L_i^2 are the same as before.

Let $x = (p_1^{c0}, \cdots, p_{n_3}^{c0}, p_1^{nt}, \cdots, p_{n_3}^{nt})$. Then note that

- $\hat{\pi}_c L_i^1 + (1 - \hat{\pi}_c) L_i^2$ is a linear function of x and hence is a concave function
- $-log(z)$ is strictly convex and decreasing
- Summation preserves convexity

Then according to Boyd and Vandenberghe (2004, Chapter 3), we have that our objective function (12.0.1) is strictly convex.

Moreover, for our constraints (12.0.2) - (12.0.6), we have that

- The inequality constraint functions in (12.0.3) - (12.0.6) are all linear functions of x and hence are convex functions
- The equality constraint function (12.0.2) is a linear combination of x and thus is affine

Therefore, our maximization problem is a strictly convex optimization problem. Hence, according to Boyd and Vandenberghe (2004, Chapter 4), it has a unique global maximum. □

Proof of Theorem 12

Proof. The proof for this theorem is similar to that of Lemma 1 in Cheng, Small, Tan and Ten Have (2009b). There are two major differences. One is that our optimization problem is given by (12.0.1) - (12.0.6), however, like Cheng et al. (2009b), our problem is also a convex optimization problem. The other is that our constraint parameter space is given by $\Theta = \{p_i^{c0}, p_i^{nt}, i = 1, \cdots, n_3 : q_i = \hat{\pi}_c p_i^{c0} + (1 - \hat{\pi}_c)p_i^{nt}, \sum_i p_i^{nt} I(T_i^{(3)} \geq V) = \hat{S_{nt}}(V), \sum_i p_i^{nt} \leq 1, \sum_i p_i^{c0} \leq 1, p_i^{nt} \geq 0, p_i^{c0} \geq 0\}$. The set Θ, as in Cheng et al. (2009b), is a convex set because the feasible set of a convex optimization problem is also convex. Because our optimization problem is convex and the constraint parameter space is a convex set, we follow the same steps as in Cheng et al. (2009b) to prove the theorem. □

Proof of Theorem 13

Proof. **Step 1**

From Kaplan and Meier (1958), under the conditions $r_v^{(1)} \to \infty$ and $r_v^{(2)} \to \infty$, we know that, as $m_1 + m_2 + N \to \infty$,

$$\hat{S}_{c1}(V) \xrightarrow{P} S_{c1}(V) \tag{12.0.7}$$

$$\hat{S}_{nt}(V) \xrightarrow{P} S_{nt}(V) \tag{12.0.8}$$

By the Law of Large Numbers, as $m_1 + m_2 + N \to \infty$,

$$\hat{\pi}_c \overset{a.s.}{\to} \pi_c \tag{12.0.9}$$

Step 2

Let $q_i^{KP} = \frac{d_i^G}{r_i^G} \prod_{j=1}^{i-1}(1 - \frac{d_j^G}{r_j^G})$, which is the Kaplan-Meier estimator of the mixture in the control group. Consider the conditions (9.3.2) - (9.3.6) along with

$$\hat{\pi}_c p_i^{c0} + (1 - \hat{\pi}_c) p_i^{nt} = q_i^{KP} \tag{12.0.10}$$

Here, notice that with (9.3.3) - (9.3.6) and (12.0.10), sometimes (9.3.2) cannot be satisfied. Given (9.3.3) - (9.3.6) and (12.0.10), we want to find the lower bound $\hat{S_{nt}^L}(V)$ and upper bound $\hat{S_{nt}^M}(V)$ of $\sum_{j=1}^{n_3} p_j^{nt} I(T_j^{(3)} \geq V)$.

Define $l = \sup\{i : \sum_{j=1}^{i-1} \frac{q_j^{KP}}{1-\hat{\pi}_c} \leq 1\}$, then

$$p_i^{nt,L} = \begin{cases} \frac{q_i^{KP}}{1-\hat{\pi}_c}, i = 1, \cdots, l-1 \\ 1 - \sum_{j=1}^{l-1} \frac{q_j^{KP}}{1-\hat{\pi}_c}, i = l \\ 0, i = l+1, \cdots, n_3 \end{cases}$$

and

$$\hat{S_{nt}^L}(V) = \sum_{j=1}^{n_3} p_j^{nt,L} I(T_j^{(3)} \geq V)$$

Define $m = \inf\{i : \sum_{j=i+1}^{n_3} \frac{q_j^{KP}}{1-\hat{\pi}_c} \leq 1\}$, then

$$p_i^{nt,M} = \begin{cases} 0, i = 1, \cdots, m-1 \\ 1 - \sum_{j=m+1}^{n_3} \frac{q_j^{KP}}{1-\hat{\pi}_c}, i = m \\ \frac{q_i^{KP}}{1-\hat{\pi}_c}, i = m+1, \cdots, n_3 \end{cases}$$

and

$$\hat{S_{nt}^{M}}(V) = \sum_{j=1}^{n_3} p_j^{nt,M} I(T_j^{(3)} \geq V)$$

Therefore, $\sum_{j=1}^{n_3} p_j^{nt} I(T_j^{(3)} \geq V) \in [\hat{S_{nt}^{L}}(V), \hat{S_{nt}^{M}}(V)]$

Similarly, define $\tilde{l} = \sup\{i : \sum_{j=1}^{i-1} \frac{q_j^{KP}}{1-\pi_c} \leq 1\}$, then

$$p_i^{nt,\tilde{L}} = \begin{cases} \frac{q_i^{KP}}{1-\pi_c}, i = 1, \cdots, \tilde{l}-1 \\ 1 - \sum_{j=1}^{\tilde{l}-1} \frac{q_j^{KP}}{1-\pi_c}, i = \tilde{l} \\ 0, i = \tilde{l}+1, \cdots, n_3 \end{cases}$$

and

$$\tilde{S_{nt}^{L}}(V) = \sum_{j=1}^{n_3} p_j^{nt,\tilde{L}} I(T_j^{(3)} \geq V)$$

Define $\tilde{m} = \inf\{i : \sum_{j=i+1}^{n_3} \frac{q_j^{KP}}{1-\pi_c} \leq 1\}$, then

$$p_i^{nt,\tilde{M}} = \begin{cases} 0, i = 1, \cdots, \tilde{m}-1 \\ 1 - \sum_{j=\tilde{m}+1}^{n_3} \frac{q_j^{KP}}{1-\pi_c}, i = \tilde{m} \\ \frac{q_i^{KP}}{1-\pi_c}, i = \tilde{m}+1, \cdots, n_3 \end{cases}$$

and

$$\tilde{S_{nt}^{M}}(V) = \sum_{j=1}^{n_3} p_j^{nt,\tilde{M}} I(T_j^{(3)} \geq V)$$

We want to show that, as $m_1 + m_2 + N \to \infty$,

$$\hat{S_{nt}^{L}}(V) - \tilde{S_{nt}^{L}}(V) \overset{a.s.}{\to} 0$$

$$\hat{S_{nt}^{M}}(V) - \tilde{S_{nt}^{M}}(V) \overset{a.s.}{\to} 0$$

Since

$$
\begin{aligned}
& |\hat{S_{nt}^L}(V) - \tilde{S_{nt}^L}(V)| \\
= \; & |\sum_{j=1}^{n_3} p_j^{nt,L} I(T_j^{(3)} \geq V) - \sum_{j=1}^{n_3} p_j^{nt,\tilde{L}} I(T_j^{(3)} \geq V)| \\
\leq \; & |\sum_{j=1}^{\min\{l,\tilde{l}\}} p_j^{nt,L} I(T_j^{(3)} \geq V) - \sum_{j=1}^{\min\{l,\tilde{l}\}} p_j^{nt,\tilde{L}} I(T_j^{(3)} \geq V)| \\
& + |\sum_{j=\min\{l,\tilde{l}\}+1}^{\max\{l,\tilde{l}\}} p_j^{nt,L} I(T_j^{(3)} \geq V) - \sum_{j=\min\{l,\tilde{l}\}+1}^{\max\{l,\tilde{l}\}} p_j^{nt,\tilde{L}} I(T_j^{(3)} \geq V)|
\end{aligned}
$$

where $\sum_{j=b}^{c} a_j = 0$ if $b > c$.

Note that if $\hat{\pi_c} \geq \pi_c$, then $l \leq \tilde{l}$. We get that

$$
\begin{aligned}
& \sum_{j=1}^{\min\{l,\tilde{l}\}} p_j^{nt,L} I(T_j^{(3)} \geq V) - \sum_{j=1}^{\min\{l,\tilde{l}\}} p_j^{nt,\tilde{L}} I(T_j^{(3)} \geq V) \\
= \; & (\frac{1}{1-\hat{\pi_c}} - \frac{1}{1-\pi_c}) \sum_{j=1}^{l} q_j^{KP} I(T_j^{(3)} \geq V) \\
& + (1 - \frac{1}{1-\hat{\pi_c}} \sum_{j=1}^{l} q_j^{KP}) I(T_l^{(3)} \geq V)
\end{aligned}
$$

Similarly, if $\hat{\pi_c} < \pi_c$, then $l \geq \tilde{l}$. We also get that

$$
\begin{aligned}
& \sum_{j=1}^{\min\{l,\tilde{l}\}} p_j^{nt,L} I(T_j^{(3)} \geq V) - \sum_{j=1}^{\min\{l,\tilde{l}\}} p_j^{nt,\tilde{L}} I(T_j^{(3)} \geq V) \\
= \; & (\frac{1}{1-\hat{\pi_c}} - \frac{1}{1-\pi_c}) \sum_{j=1}^{\tilde{l}} q_j^{KP} I(T_j^{(3)} \geq V) \\
& - (1 - \frac{1}{1-\pi_c} \sum_{j=1}^{\tilde{l}} q_j^{KP}) I(T_{\tilde{l}}^{(3)} \geq V)
\end{aligned}
$$

Therefore, combining these two situation, we get that

$$
\begin{aligned}
& |\sum_{j=1}^{\min\{l,\tilde{l}\}} p_j^{nt,L} I(T_j^{(3)} \geq V) - \sum_{j=1}^{\min\{l,\tilde{l}\}} p_j^{nt,\tilde{L}} I(T_j^{(3)} \geq V)| \\
= \; & |\max\{\frac{1}{1-\hat{\pi_c}} - \frac{1}{1-\pi_c}, \frac{1}{1-\pi_c} - \frac{1}{1-\hat{\pi_c}}\} \sum_{j=1}^{\min\{l,\tilde{l}\}} q_j^{KP} I(T_j^{(3)} \geq V) \\
& +(1 - \max\{\frac{1}{1-\hat{\pi_c}}, \frac{1}{1-\pi_c}\} \sum_{j=1}^{\min\{l,\tilde{l}\}} q_j^{KP}) I(T_{\min\{l,\tilde{l}\}}^{(3)} \geq V)| \\
\leq \; & |(\frac{1}{1-\hat{\pi_c}} - \frac{1}{1-\pi_c}) \sum_{j=1}^{\min\{l,\tilde{l}\}} q_j^{KP}| + |1 - \sum_{j=1}^{\min\{l,\tilde{l}\}} q_j^{KP} \max\{\frac{1}{1-\hat{\pi_c}}, \frac{1}{1-\pi_c}\}|
\end{aligned}
$$

From (12.0.9), we know that $|\frac{1}{1-\hat{\pi_c}} - \frac{1}{1-\pi_c}| \overset{a.s.}{\to} 0$. And from the fact that $|\sum_{j=1}^{\min\{l,\tilde{l}\}} q_j^{KP}| \leq 1$,

$$
|(\frac{1}{1-\hat{\pi_c}} - \frac{1}{1-\pi_c}) \sum_{j=1}^{\min\{l,\tilde{l}\}} q_j^{KP}| \overset{a.s.}{\to} 0
$$

For the second part, from definitions, it is easy to verify that

$$
\min\{1-\hat{\pi_c}, 1-\pi_c\} \leq \sum_{j=1}^{\min\{l,\tilde{l}\}} q_j^{KP} \leq \max\{1-\hat{\pi_c}, 1-\pi_c\}
$$

Thus,

$$
1 - \max\{\frac{1-\hat{\pi_c}}{1-\pi_c}, \frac{1-\pi_c}{1-\hat{\pi_c}}\} \leq 1 - \sum_{j=1}^{\min\{l,\tilde{l}\}} q_j^{KP} \max\{\frac{1}{1-\hat{\pi_c}}, \frac{1}{1-\pi_c}\} \leq 0
$$

Therefore,

$$\begin{aligned} & |1 - \sum_{j=1}^{\min\{l,\tilde{l}\}} q_j^{KP} \max\{\frac{1}{1-\hat{\pi_c}}, \frac{1}{1-\pi_c}\}| \\ \leq & |1 - \max\{\frac{1-\hat{\pi}_c}{1-\pi_c}, \frac{1-\pi_c}{1-\hat{\pi}_c}\}| \\ = & |1 - \frac{1-\min\{\hat{\pi}_c, \pi_c\}}{1-\max\{\hat{\pi}_c, \pi_c\}}| \end{aligned}$$

Again, from (9) and the condition that $0 < \pi_c < 1$, we know that

$$|1 - \sum_{j=1}^{\min\{l,\tilde{l}\}} q_j^{KP} \max\{\frac{1}{1-\hat{\pi_c}}, \frac{1}{1-\pi_c}\}| \overset{a.s.}{\to} 0$$

Hence,

$$|\sum_{j=1}^{\min\{l,\tilde{l}\}} p_j^{nt,L} I(T_j^{(3)} \geq V) - \sum_{j=1}^{\min\{l,\tilde{l}\}} p_j^{nt,\tilde{L}} I(T_j^{(3)} \geq V)| \overset{a.s.}{\to} 0$$

Next, if $\hat{\pi}_c \geq \pi_c$, then $l \leq \tilde{l}$. We also get that

$$\begin{aligned} & |\sum_{j=\min\{l,\tilde{l}\}+1}^{\max\{l,\tilde{l}\}} p_j^{nt,L} I(T_j^{(3)} \geq V) - \sum_{j=\min\{l,\tilde{l}\}+1}^{\max\{l,\tilde{l}\}} p_j^{nt,\tilde{L}} I(T_j^{(3)} \geq V)| \\ = & \sum_{j=l+1}^{\tilde{l}} p_j^{nt,\tilde{L}} I(T_j^{(3)} \geq V) \\ \leq & \sum_{j=l+1}^{\tilde{l}} p_j^{nt,\tilde{L}} \\ = & 1 - \frac{\sum_{j=1}^{l} q_j^{KP}}{1-\pi_c} \end{aligned}$$

From the definitions of $l, \tilde{l}$ and the fact that $\hat{\pi}_c \geq \pi_c$, we get that

$$1 - \hat{\pi}_c \leq \sum_{j=1}^{l} q_j^{KP} \leq 1 - \pi_c$$

Therefore,

$$\begin{aligned} & |\sum_{j=\min\{l,\tilde{l}\}+1}^{\max\{l,\tilde{l}\}} p_j^{nt,L} I(T_j^{(3)} \geq V) - \sum_{j=\min\{l,\tilde{l}\}+1}^{\max\{l,\tilde{l}\}} p_j^{nt,\tilde{L}} I(T_j^{(3)} \geq V)| \\ \leq & \ 1 - \frac{\sum_{j=1}^{l} q_j^{KP}}{1-\pi_c} \\ \leq & \ 1 - \frac{1-\hat{\pi}_c}{1-\pi_c} \end{aligned}$$

If $\hat{\pi}_c < \pi_c$, then $\tilde{l} \leq l$. We get that

$$\begin{aligned} & |\sum_{j=\min\{l,\tilde{l}\}+1}^{\max\{l,\tilde{l}\}} p_j^{nt,L} I(T_j^{(3)} \geq V) - \sum_{j=\min\{l,\tilde{l}\}+1}^{\max\{l,\tilde{l}\}} p_j^{nt,\tilde{L}} I(T_j^{(3)} \geq V)| \\ = & \ \sum_{j=\tilde{l}+1}^{l} p_j^{nt,L} I(T_j^{(3)} \geq V) \\ \leq & \ \sum_{j=\tilde{l}+1}^{l} p_j^{nt,L} \\ = & \ 1 - \frac{\sum_{j=1}^{\tilde{l}} q_j^{KP}}{1-\hat{\pi}_c} \end{aligned}$$

Again, from the definitions of $l, \tilde{l}$ and the fact that $\hat{\pi}_c \geq \pi_c$, we get that

$$1 - \pi_c \leq \sum_{j=1}^{\tilde{l}} q_j^{KP} \leq 1 - \hat{\pi}_c$$

Therefore,

$$\begin{aligned} & |\sum_{j=\min\{l,\tilde{l}\}+1}^{\max\{l,\tilde{l}\}} p_j^{nt,L} I(T_j^{(3)} \geq V) - \sum_{j=\min\{l,\tilde{l}\}+1}^{\max\{l,\tilde{l}\}} p_j^{nt,\tilde{L}} I(T_j^{(3)} \geq V)| \\ \leq & \ 1 - \frac{\sum_{j=1}^{\tilde{l}} q_j^{KP}}{1-\hat{\pi}_c} \\ \leq & \ 1 - \frac{1-\pi_c}{1-\hat{\pi}_c} \end{aligned}$$

Combining these two case, we get that

$$\begin{aligned} & |\sum_{j=\min\{l,\tilde{l}\}+1}^{\max\{l,\tilde{l}\}} p_j^{nt,L} I(T_j^{(3)} \geq V) - \sum_{j=\min\{l,\tilde{l}\}+1}^{\max\{l,\tilde{l}\}} p_j^{nt,\tilde{L}} I(T_j^{(3)} \geq V)| \\ \leq & \ 1 - \frac{\min\{1-\pi_c, 1-\hat{\pi}_c\}}{\max\{1-\pi_c, 1-\hat{\pi}_c\}} \\ = & \ 1 - \frac{1-\max\{\pi_c, \hat{\pi}_c\}}{1-\min\{\pi_c, \hat{\pi}_c\}} \end{aligned}$$

From (12.0.9) and the condition that $0 < \pi_c < 1$, we know that

$$|\sum_{j=\min\{l,\tilde{l}\}+1}^{\max\{l,\tilde{l}\}} p_j^{nt,L} I(T_j^{(3)} \geq V) - \sum_{j=\min\{l,\tilde{l}\}+1}^{\max\{l,\tilde{l}\}} p_j^{nt,\tilde{L}} I(T_j^{(3)} \geq V)| \xrightarrow{a.s.} 0$$

To sum up, we verify that

$$\hat{S_{nt}^L}(V) - \tilde{S_{nt}^L}(V) \xrightarrow{a.s.} 0$$

Similarly, we can also prove that

$$\hat{S_{nt}^M}(V) - \tilde{S_{nt}^M}(V) \xrightarrow{a.s.} 0$$

Step 3

In this step, we would like to show that, as $m_1 + m_2 + N \to \infty$,

$$\tilde{S_{nt}^L}(V) \xrightarrow{P} \frac{S_G(V) - \pi_c}{1-\pi_c} I_{V \leq G^{-1}(1-\pi_c)} \tag{12.0.11}$$

$$\tilde{S_{nt}^M}(V) \xrightarrow{P} 1 + \frac{S_G(V) - (1-\pi_c)}{1-\pi_c} I_{V \geq G^{-1}(\pi_c)} \tag{12.0.12}$$

As $m_1 + m_2 + N \to \infty$,

$$\begin{aligned} \tilde{S_{nt}^{L}}(V) &= \sum_{j=1}^{n_3} p_j^{nt,\tilde{L}} I(T_j^{(3)} \geq V) \\ &= \sum_{j=1}^{\tilde{l}-1} \frac{q_j^{KP}}{1-\pi_c} I(T_j^{(3)} \geq V) + p_{\tilde{l}} I(T_{\tilde{l}}^{(3)} \geq V) \\ &\xrightarrow{P} \frac{(1-\pi_c) - (1 - S_G(V))}{1-\pi_c} I_{(V \leq G^{-1}(1-\pi_c))} \\ &= \frac{S_G(V) - \pi_c}{1-\pi_c} I_{(V \leq G^{-1}(1-\pi_c))} \end{aligned}$$

Similarly,

$$\tilde{S_{nt}^{M}}(V) \xrightarrow{P} 1 + \frac{S_G(V) - (1-\pi_c)}{1-\pi_c} I_{(V \geq G^{-1}(\pi_c))}$$

Step 4

From (9.5.1) which is given by

$$\frac{S_G(V) - \pi_c}{1-\pi_c} I_{(V \leq G^{-1}(1-\pi_c))} < S_{nt} < 1 + \frac{S_G(V) - (1-\pi_c)}{1-\pi_c} I_{(V \geq G^{-1}(\pi_c))}$$

, along with (12.0.8), (12.0.11) and (12.0.12), we verify that, as $m_1 + m_2 + N \to \infty$,

$$\frac{S_G(V) - \pi_c}{1-\pi_c} I_{(V \leq G^{-1}(1-\pi_c))} \leq \sum_{j=1}^{n_3} p_j^{nt} I(T_j^{(3)} \geq V) \geq 1 + \frac{S_G(V) - (1-\pi_c)}{1-\pi_c} I_{(V \geq G^{-1}(\pi_c))}$$

is asymptotically valid in probability. Hence, (9.3.2) is asymptotically satisfied in probability under maximization constraints (9.3.3) - (9.3.6) and (12.0.10). Therefore, the maximization problem (9.3.1) under constraints (9.3.2) - (9.3.6) in the paper is asymptotically equivalent to the maximization problem (9.3.1) under constraints (9.3.3) - (9.3.6) and (12.0.10) in probability. Note that q_i^{KP} in (12.0.10) is

actually the Kaplan-Meier estimator of the mixture in the control group. Therefore, from Kaplan and Meier (1958), since $r_v^G \to \infty$, as $m_1 + m_2 + N \to \infty$, we have that

$$\hat{S}_G(V) \xrightarrow{P} S_G(V) \tag{12.0.13}$$

Step 5

From (12.0.8), (12.0.9), (12.0.13) and the fact that $\hat{S}_G(V)$,$\hat{S}_{nt}(V)$,$\hat{\pi}_c$ are all bounded, we get that

$$\begin{aligned} \hat{S}_{c0}(V) &= \frac{\hat{S}_G(V) - \hat{\pi}_c \hat{S}_{nt}(V)}{1 - \hat{\pi}_c} \\ &\xrightarrow{P} \frac{S_G(V) - \pi_c S_{nt}(V)}{1 - \pi_c} \\ &= S_{c0}(V) \end{aligned}$$

Hence, along with (12.0.7), we verify that

$$\hat{S}_{c1}(V) - \hat{S}_{c0}(V) \xrightarrow{P} S_{c1}(V) - S_{c0}(V)$$

□

Standard Errors Calculation through Delta Method

The difference of the RMSEs of two estimators δ_1 and δ_2 are defined as:

$$T = \sqrt{E((\delta_1 - \theta)^2)} - \sqrt{E((\delta_2 - \theta)^2)}$$

Therefore, we can get the variance of T as below:

$$
\begin{aligned}
Var(T) &= \sigma^2_{\delta_1}(\frac{\partial T}{\partial \delta_1})^2 + \sigma^2_{\delta_2}(\frac{\partial T}{\partial \delta_2})^2 + 2\sigma_{\delta_1\delta_2}(\frac{\partial T}{\partial \delta_1})(\frac{\partial T}{\partial \delta_2}) \\
&= \sigma^2_{\delta_1}(\frac{E(\delta_1-\theta)}{\sqrt{E((\delta_1-\theta)^2)}})^2 + \sigma^2_{\delta_2}(\frac{E(\delta_2-\theta)}{\sqrt{E((\delta_2-\theta)^2)}})^2 \\
&\quad +2\sigma_{\delta_1\delta_2}(\frac{E(\delta_1-\theta)}{\sqrt{E((\delta_1-\theta)^2)}})(\frac{E(\delta_2-\theta)}{\sqrt{E((\delta_2-\theta)^2)}})
\end{aligned}
$$

From Delta method, we can estimate the standard error of T as

$$
\begin{aligned}
\hat{Var}(T) &= \hat{Var}(\delta_1)(\frac{\hat{Bias}(\delta_1)}{\hat{RMSE}(\delta_1)})^2 + \hat{Var}(\delta_2)(\frac{\hat{Bias}(\delta_2)}{\hat{RMSE}(\delta_2)})^2 \\
&+ 2\hat{Cov}(\delta_1,\delta_2)(\frac{\hat{Bias}(\delta_1)}{\hat{RMSE}(\delta_1)})(\frac{\hat{Bias}(\delta_2)}{\hat{RMSE}(\delta_2)})
\end{aligned}
$$

where $\hat{Var}(\delta_1)$ and $\hat{Var}(\delta_2)$ are sample variances of δ_1 and δ_2, $\hat{Cov}(\delta_1,\delta_2)$ is the sample covariance, $\hat{Bias}(\delta_1)$ and $\hat{Bias}(\delta_2)$ are sample biases, as well as $\hat{RMSE}(\delta_1)$ and $\hat{RMSE}(\delta_2)$ are sample RMSEs. All of the above quantities are available through the simulation studies.

Test for Parametric Weibull Assumption

We will give the detailed discussion on whether the potential failure times of compliers and never-takers in our limited mortality analysis of the HIP data set follow the Weibull distribution. From Cox and Oakes (1984), if T follows the Weibull distribution with density function $\rho\kappa(\rho x)^{\kappa-1}\exp(-(\rho x)^{\kappa})$, it is easy to verify that

$$
log(H(t)) = \kappa\log(\rho) + \kappa\log(t),
$$

where $H(t)$ is the cumulative hazard function of T. Therefore, $\log(H(t))$ and $\log(t)$ should have a linear relationship when T follows a Weibull distribution. We estimate $\log(H(t))$ by $\log(\hat{H}(t))$ through the Kaplan-Meier estimator. Figure 12.1 shows us the result for compliers in the treatment arm. The mean squared error (MSE) for the regression between $\log(\hat{H}(t))$ and $\log(t)$ is 0.0306. We use this MSE as a test statistic for testing that T follows a Weibull distribution and use the parametric Bootstrap method to perform the test. If we assume that the failure times of compliers in the treatment arm follow a Weibull distribution with parameters κ_{c1} and ρ_{c1}, the MLEs for κ_{c1} and ρ_{c1} are given by $\hat{\kappa}_{c1} = 1.939$ and $\hat{\rho}_{c1} = 0.579$. We generate random samples for Weibull distribution with parameters $\hat{\kappa}_{c1} = 1.939$ and $\hat{\rho}_{c1} = 0.579$, calculate $\log(\hat{H}(t))$ through the Kaplan-Meier method, fit the linear regression between $\log(\hat{H}(t))$ and $\log(t)$, and get the MSE for this regression. We repeat this procedure 10,000 times. The estimated p-value is given by 0.1858. Therefore, there is not strong evidence for us to reject the null hypothesis that the failure times of compliers in the treatment group follow the Weibull distribution. However, from Figure 12.2 which shows us the result for never-takers in the treatment arm, it does not reveal a strong linear pattern between $\log(\hat{H}(t))$ and $\log(t)$. The MSE for the regression between $\log(\hat{H}(t))$ and $\log(t)$ is 0.0924. If we assume that the failure times of compliers in the treatment arm follow a Weibull distribution with parameters κ_{nt} and ρ_{nt}, the MLEs for κ_{nt} and ρ_{nt} are given by $\hat{\kappa}_{nt} = 1.202$ and $\hat{\rho}_{nt} = 0.444$. We follow the same approach as above to conduct

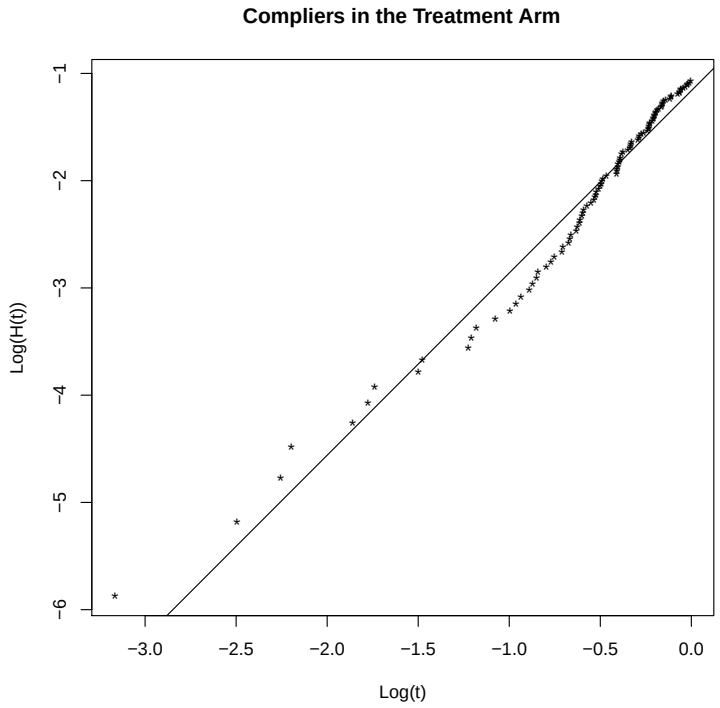


Figure 12.1: Compliers in the Treatment Arm

the hypothesis testing, and the estimated p-value is given by 0.0418. Thus, there is evidence for us to cast doubts on the validity of the Weibull model for never-takers in the treatment arm.

Details of BC_a method

Besides Bootstrap percentile method, Efron and Tibshirani (1994) suggest using BC_a method to obtain the confidence interval for censored data sets $\{(Y_i, \Delta_i)\}_{i=1}^{m_1+m_2+N}$. We can construct the BC_a confidence interval following the steps below.

- Step I: Draw a Bootstrap sample $\{(Y_i^*, \Delta_i^*)\}_{i=1}^{m_1+m_2+N}$. For compliers in the treatment group $\{(Y_i, \Delta_i)\}_{i=1}^{m_1}$, we sample with replacement by putting mass

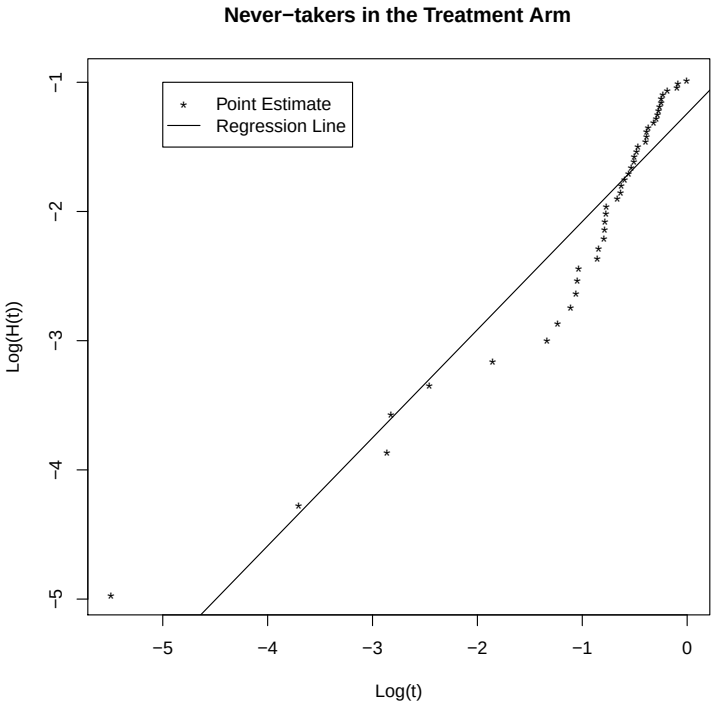


Figure 12.2: Never-takers in the Treatment Arm

$\frac{1}{m_1}$ at each point (Y_i, Δ_i) in order to get Bootstrap sample $\{(Y_i^*, \Delta_i^*)\}_{i=1}^{m_1}$. For never-takers in the treatment group $\{(Y_i, \Delta_i)\}_{i=m_1+1}^{m_1+m_2}$, we sample with replacement by putting mass $\frac{1}{m_2}$ at each point (Y_i, Δ_i) in order to get Bootstrap sample $\{(Y_i^*, \Delta_i^*)\}_{i=m_1+1}^{m_1+m_2}$. For mixtures in the control group $\{(Y_i, \Delta_i)\}_{i=m_1+m_2+1}^{m_1+m_2+N}$, we sample with replacement by putting mass $\frac{1}{N}$ at each point (Y_i, Δ_i) in order to get Bootstrap sample $\{(Y_i^*, \Delta_i^*)\}_{i=m_1+m_2+1}^{m_1+m_2+N}$. We join three Bootstrap samples together to get a Bootstrap sample $\{(Y_i^*, \Delta_i^*)\}_{i=1}^{m_1+m_2+N}$.

- Step II: Estimate PNEMLE $\hat{W}^*(V)$ for this Bootstrap sample following the procedures in Section 9.3.

- Step III: Independently repeat steps I and II B times and obtain $\{\hat{W}_b^*(V)\}_{b=1}^{B}$. Find the lower α_1 percentile $\hat{W}_{LOW}^*(V)$ and the upper α_2 percentile $\hat{W}_{UP}^*(V)$, where α_1 and α_2 are given as

$$\begin{aligned} \alpha_1 &= \Phi(\hat{z}_0 + \frac{\hat{z}_0 + z^{(\alpha)}}{1 - \hat{\alpha}(\hat{z}_0 + z^{(\alpha)})}) \\ \alpha_2 &= \Phi(\hat{z}_0 + \frac{\hat{z}_0 + z^{(1-\alpha)}}{1 - \hat{\alpha}(\hat{z}_0 + z^{(1-\alpha)})}) \end{aligned}$$

The value of $\hat{z}_0$ is obtained directly from the proportion of Bootstrap replications less than the original estimate $\hat{W}(V)$, that is,

$$\hat{z}_0 = \Phi^{-1}(\frac{\text{No. of } \{\hat{W}_b^*(V) < \hat{W}(V)\}}{B})$$

Let $\hat{W}_{(.)(V)} = \frac{\sum_{b=1}^{B} \hat{W}_b^*(V)}{B}$. The value of $\hat{\alpha}$ is obtained from

$$\hat{\alpha} = \frac{\sum_{b=1}^{B} (\hat{W}_{(.)}(V) - \hat{W}_b(V))^3}{6(\sum_{b=1}^{B} (\hat{W}_{(.)}(V) - \hat{W}_b(V))^2)^{\frac{3}{2}}}$$

- Step IV: The $(1 - 2\alpha)$ confidence interval is given by $(\hat{W}^*_{LOW}(V), \hat{W}^*_{UP}(V))$.

Bibliography

[1] Angrist, J.D., Imbens, G.W. and Rubin, D.B. (1996). Identification of Causal Effects Using Instrumental Variables. *Journal of the American Statistical Association*, **91**, 444–455.

[2] Baker, S.G. (1998). Analysis of Survival Data from a Randomized Trial with All-or-None Compliance: Estimating the Cost-Effectiveness of a Cancer Screening Program. *Journal of the American Statistical Association*, **93**, 929–934.

[3] Baker, S.G. and Lindeman, K.S. (1994). The Paired Availability Design: a Proposal for Evaluating Epidural Analgesia during Labor. *Statistics in Medicine*, **13**, 2269–2278.

[4] Boyd, S. and Vandenberghe, L. (2004). Convex Optimization. Cambridge University Press.

[5] Bound, J., Jaeger, D.A. and Baker, R.M. (1995). Problems with Instrumental Variables Estimation When the Correlation Between the Instruments

and the Endogeneous Explanatory Variable is Weak. *Journal of the American Statistical Association*, **90**, 443–450.

[6] Cheng, J. and Small, D.S. (2006). Bounds on Causal Effects in Three-arm Trials with Noncompliance. *Journal of the Royal Statistical Society, Series B (Methodological)*, **68**, 815–836.

[7] Cheng, J., Qin, J. and Zhang, B (2009a). Semiparametric Estimation and Inference for Distributional and General Treatment Effects. *Journal of the Royal Statistical Society, Series B (Methodological)*, **71**, 881–904.

[8] Cheng, J., Small, D.S., Tan, Z. and Ten Have, T.R. (2009b). Efficient Nonparametric Estimation of Causal Effects in Randomized Trials with Noncompliance. *Biometrika*, **96**, 19–36.

[9] Cox, D.R. and Oakes, D. (1984). Analysis of Survival Data. Chapman & Hall/CRC.

[10] Cuzick, J., Sasieni, P., Myles, J. and Tyler, J. (2007). Estimating the Effect of Treatment in a Proportional Hazards Model in the Presence of Noncompliance and Contamination. *Journal of the Royal Statistical Society, Series B (Methodological)*, **69**, 565–588.

[11] Efron, B. and Tibshirani, R.J. (1994). An Introduction to the Bootstrap. Chapman & Hall/CRC.

[12] Follmann, D.A. (2000). On the Effect of Treatment among Would-Be Treatment Compliers: An Analysis of the Multiple Risk Factor Intervention Trial. *Journal of the American Statistical Association*, **95**, 1101–1109.

[13] Greenland, S., Lanes, S. and Jara, M. (2008). Estimating Effects from Randomized Trials with Discontinuations: the Need for Intent-to-treat Design and G-estimation. *Clinical Trials*, **5**, 5–13.

[14] Imbens, G.W. and Angrist, J.D. (1994). Identification and Estimation of Local Average Treatment Effects. *Econometrica*, **62**, 467–476.

[15] Imbens, G.W. and Rubin, D.B. (1997). Bayesian Inference for Causal Effects in Randomized Exper- iments with Noncompliance. *The Annals of Statistics*, **25**, 305–327.

[16] Joffe, M.M. (2001). Administrative and Artificial Censoring in Censored Regression Models. *Statistics in Medicine*, **20**, 2287–2304.

[17] Joffe, M.M., Small, D.S., Ten Have, T., Brunelli, S. and Feldman, H.I. (2008). Extended Instrumental Variables Estimation for Overall Effects. *The International Journal of Biostatistics*, **4**, Artical 4.

[18] Kaplan, E.L. and Meier, P. (1958). Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association*, **53**, 457–481.

[19] Kubik, A., Parkin, D.M., Khlat, M. et al. (1990). Lack of benefit from semi-annual screening for cancer of the lung: follow-up report of a randomized controlled trial on a population of high-risk males in Czechoslovakia. *International Journal of Cancer*, **45**, 26–33.

[20] Loeys, T. and Goetghebeur, E. (2003). A Causal Proportional Hazards Estimator for the Effect of Treatment Actually Received in a Randomized Trial with All-or-nothing Compliance. *Biometrics*, **59**, 100–105.

[21] Nagar, A.L. (1959). The Bias and Moment Matrix of the General k-Class Estimators of the Parameters in Simultaneous Equations. *Econometrica*, **27**, 575–595.

[22] Oken, M.M., Marcus, P.M., Hu, P. et al. (2005). Baseline chest radiograph for lung cancer detection in the randomized prostate, lung, colorectal and ovarian cancer screening trial. *Journal of National Cancer Institute*, **97**, 1832–1839.

[23] Owen, A.B. (2001). Empirical Likelihood. Chapman & Hall/CRC.

[24] Robins, J.M. and Finkelstein, D.M. (2000). Correcting for Noncompliance and Dependent Censoring in an AIDS Clinical Trial with Inverse Probability of Censoring Weighted (IPCW) Log-rank Tests. *Biometrics*, **56**, 779-788.

[25] Robins, J.M. and Tsiatis, A.A. (1991). Correcting for Non-compliance in Randomized Trials Using Rank Preserving Structural Failure Time Models. *Communications in Statistics, Theory and Methods*, **20**, 2609-2631.

[26] Rubin, D.B. (1978). Bayesian Inference for Causal Effects. *The Annals of Statistics*, **6**, 34–58.

[27] Sheiner, L.B. and Rubin, D.B. (1995). Intention-to-treat Analysis and the Goals of Clinical Trials. *Clinical Pharmacology and Therapeutics*, **57**, 6–15.

[28] Small, D.S., Ten Have, T.R., Joffe, M.M. and Cheng, J. (2006). Random Effects Logistic Models for Analyzing Efficacy of a Longitudinal Randomized Treatment with Non-adherence. *Statistics in Medicine*, **25**, 1981–2007.

[29] Sommer, A. and Zeger, S.L. (1991). On Estimating Efficacy from Clinical Trials. *Statistics in Medicine*, **10**, 45–52.

[30] Stock, J.H., Wright, J.H. and Yogo, M. (2002). A Survey of Weak Instruments and Weak Identification in Generalized Method of Moments. *Journal of Business and Economic Statistics*, **20**, 518-529.

[31] Zelen, M. (1979). A New Design for Randomized Clinical Trials. *New England Journal of Medicine*, **300**, 1242–1245.