UNIVERSITY OF CALIFORNIA
RIVERSIDE

Semi-Parametric Inference for a Semi-Supervised Two-Component Location-Shifted Mixture Model

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Applied Statistics

by

Bradley Mark Lubich

June 2023

Dissertation Committee:
Dr. Daniel R. Jeske, Chairperson
Dr. Weixin Yao
Dr. Jun Li

The Dissertation of Bradley Mark Lubich is approved:

Committee Chairperson

University of California, Riverside

Acknowledgements

I am grateful for my advisor, Dr. Daniel R. Jeske, for his direction and collaboration on this work. I would also like to thank Dr. Weixin Yao for his collaboration on a portion of this work.

The text of this dissertation, in part, is a reprint of the material as it appears in Lubich et al. (2022). The co-authors Daniel R. Jeske and Weixin Yao listed in that publication directed and collaborated on the research which forms the basis for this dissertation.

I would like to thank Dr. Beverly Green and her team with Kaiser Permanente for providing the data set from their blood pressure study funded by the National Heart, Lung, and Blood Institute of the National Institutes of Health (NIH): Grant \#1 R01 HL075263-01; Electronic Communications and Blood Pressure Monitoring.

Lastly, I would like to express my gratitude for the many people who have contributed to my education affording me the opportunity of completing this work including my parents and those from Galt Christian School, Mokelumne River Elementary School, Lodi Christian School, Galt High School, Cal Poly - San Luis Obispo, and University of California - Riverside.

To Marissa for all your love and support.

ABSTRACT OF THE DISSERTATION

Semi-Parametric Inference for a Semi-Supervised Two-Component Location-Shifted Mixture Model

by

Bradley Mark Lubich

Doctor of Philosophy, Graduate Program in Applied Statistics
University of California, Riverside, June 2023
Dr. Daniel R. Jeske, Chairperson

In a randomized clinical trial (RCT) with a control group vs. treatment group design, mixture models (Lindsay, 1995; G. McLachlan and Peel, 2000) can be a good choice for the treatment group response distribution in anticipation that there might be a sub-population of the treated population whose responses have the same distribution as the control group. It is well known that such sub-populations of 'non-responding' treated patients exist in oncology trials (Spear et al., 2001; Manegold et al., 2016). Although it would be ideal to identify a-priori the features that characterize individuals who will respond (a 'responder') to the treatment and those who will not (a 'non-responder'), this dissertation considers inference when such information has yet to be ascertained. Post-hoc sub-group analyses are known to lead to an inflated rate of false discoveries (Lagakos et al., 2006). Assessing the existence of subgroups with mixture model inference before proceeding with identifying subgroups based upon biomarkers can decrease the false discovery rate among sub-group analyses. Jeske and Yao (2020) demonstrated that ignoring the heterogeneity of treatment effects could result in an under-powered experiment and have the risk of missing some useful treatments. When heterogeneity is indeed present and treatment effects are sub-population specific, the average treatment effect obtained by the standard methods can lead to incorrect conclusions. Hence, the use of mixture models to represent the response distribution within the treatment group is compelling and it is desirable to describe the nature of this sub-population specific effect via inference on the corresponding parameters from the mixture distribution. This dissertation explores four methods of point estimation for the parameters. Two of the methods are also used to construct confidence bounds (both intervals and regions) for the parameters. Simulation is used to assess the performances of the various methods and make a recommendation. The recommended methods are illustrated on an example blood pressure data set.

Contents

List of Tables

List of Figures

List of Lists

Chapter 1

Introduction

1.1 Research Direction

In a randomized clinical trial (RCT) with a control group vs. treatment group design, mixture models (Lindsay, 1995; G. McLachlan and Peel, 2000) can be a good choice for the treatment group response distribution in anticipation that there might be a sub-population of the treated population whose responses have the same distribution as the control group. It is well known that such sub-populations of 'non-responding' treated patients exist in oncology trials (Spear et al., 2001; Manegold et al., 2016). Although it would be ideal to identify a-priori the features that characterize individuals who will respond (a 'responder') to the treatment and those who will not (a 'non-responder'), this dissertation considers inference when such information has yet to be confirmed. Post-hoc sub-group analyses are known to lead to an inflated rate of false discoveries (Lagakos et al., 2006). Assessing the existence of subgroups with mixture model inference before proceeding with identifying subgroups based upon biomarkers can decrease the false discovery rate among sub-group analyses. A group fMRI example motivated a recent call for more attention to be given to mixture alternatives for comparing two (alternative) treatments (by a hypothesis test), stating that medical applications, psychiatric-genetics and personalized medicine are important applications where mixtures are plausible alternatives (Rosenblatt and Benjamini, 2018). Jeske and Yao (2020) demonstrated that ignoring the heterogeneity of treatment effects could result in an under-powered experiment and have the risk of missing some useful treatments. When heterogeneity is indeed present and treatment effects are sub-population specific, the average treatment effect obtained by the standard methods can lead to incorrect conclusions. Hence, the use of mixture models to represent the response distribution within the treatment group is compelling and it is desirable to describe the nature of this sub-population specific effect via inference on the corresponding parameters from the mixture distribution.

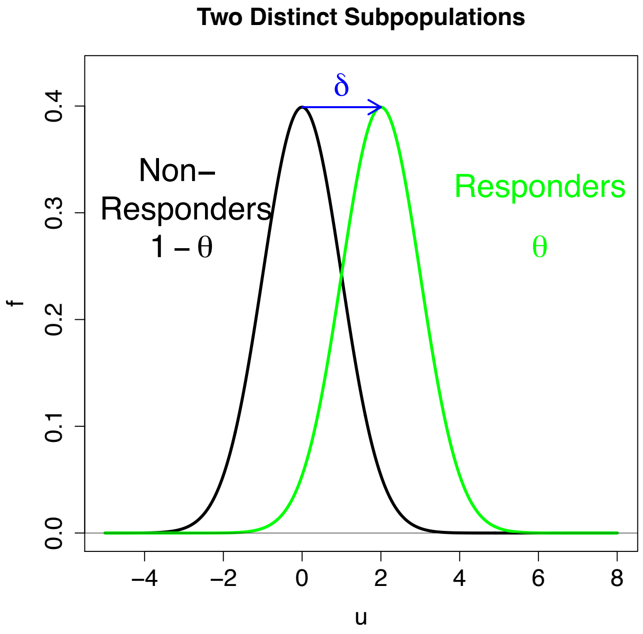


Figure 1.1: Two distinct sub-populations that make up a mixture distribution.

Denote the cumulative distribution functions (CDFs) associated with a response from the control group and the treatment group by F and G, respectively. Mean shift alternatives of the form $G(u) = F(u - \delta)$ are frequently used. This dissertation assumes, without loss of generality, that $\delta > 0$ and uses a mixture model for the responses from the treatment group of the form

$$G(u) = (1-\theta)F(u) + \theta F(u-\delta). \tag{1.1}$$

where $\theta \in (0,1]$ and $F \in \mathcal{F}$ representing the set of all CDFs. Also, $(\theta, \delta) = (0,0)$ is in the parameter space and represents a non-existent treatment effect. In this context, the treatment effect is represented by the pair (θ, δ) and the average treatment effect is $\Delta = \theta\delta$. The parameter θ represents the proportion of responders in the treated population, while δ represents the effect size of the treatment for the responders. (Note that when $\theta = 1$ the model simplifies to a pure mean shift). While it is common to impose that F belongs to a particular parametric family (e.g. Normal), this dissertation aims to limit distributional assumptions with the goal of distribution-robust inference on (θ, δ).

While each individual is either a responder or non-responder, the component membership of a randomly sampled individual is an unobserved random variable. Let $\boldsymbol{Z} = Z_1, ..., Z_n \overset{iid}{\sim} \text{Bernoulli}(\theta)$ represent component membership (0 for non-responder, 1 for responder) for each patient in the treatment group. Let $\boldsymbol{Y} = Y_1, ..., Y_n$ represent the (observed) response from the random sample of treated patients. Thus, the treatment data consists of n pairs - $(\boldsymbol{Y}, \boldsymbol{Z}) = (Y_1, Z_1), ..., (Y_n, Z_n)$ - where Y_i is observed and Z_i is a latent variable. The conditional distributions representing the sub-population responses are $Y_i|(Z_i = 0) \sim F(u)$ and $Y_i|(Z_i = 1) \sim F(u-\delta)$. Therefore Y_i marginally follows G in (1.1). Let $\boldsymbol{X} = X_1, ..., X_m$ represent the patient responses in the control group. Control patients do not respond to the treatment (since they do not receive it) so the distribution for these untreated patients is $F(u)$. By randomization in the RCT $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent random samples with a total sample size denoted $N = m + n$.

The rest of the dissertation is organized as follows. The remainder of Chapter 1 explores mixture models more broadly and further motivates the research direction by showing how it contributes to the existing literature. Chapter 2 explores various estimators for the treatment effect. Chapter 3 discusses confidence bounds (intervals and regions) corresponding to two of the estimates from chapter 2, method of moments and pseudo-likelihood. Chapter 4 compares the performance of the estimators and confidence bounds via extensive simulation studies and provides recommendations. Chapter 5 concludes by demonstrating the utility of the recommended methods on an example blood pressure data set and discusses future work.

1.2 Survey of Mixture Models

Mixture models are also known as ‘latent class models’ or ‘unsupervised learning models’. Sometimes they are used as a means of flexibly modeling data that is difficult to model parametrically. For example, kernel density estimation (KDE) is one very popular non-parametric estimation technique for estimating a density. This technique is actually a special case of mixture modeling. Another application of mixtures is for modeling a population that is thought to be comprised of multiple distinct sub-populations. Inference in these scenarios may focus on both sub-population features as well as the percentage of the population from each sub-population. The number of distinct sub-populations modeled by a mixture distribution may be pre-specified or learned from the observed data.

1.2.1 Basic Definitions and Notation

A mixture distribution G is defined by a distribution function

$$G(u) = \sum_{j=1}^{c} \pi_j F_j(u) \tag{1.2}$$

where $0 \leq \pi_j \leq 1$ for all $j \in \{1, ..., c\}$, $\sum_{j=1}^{c} \pi_j = 1$ and F_j is a distribution function for all $j \in \{1, ..., c\}$. Each F_j is called a component distribution, while each π_j is called a component probability and c is the number of components in the mixture model, which may be known or unknown. Let bolded symbols represent vectors. Thus, $\boldsymbol{F(u)} = [F_1(u), ..., F_c(u)]$ (or just $\boldsymbol{F}$). Since $\sum_{j=1}^{c} \pi_j = 1$, all component probabilities are defined by specification of $c - 1$ of the π_j values. By convention, consider the first $c - 1$ to be the parameter vector $\boldsymbol{\pi} = [\pi_1, ..., \pi_{c-1}]$ (then $\pi_c = 1 - \sum_{j=1}^{c-1} \pi_j$). If c is finite, then G is said to be a finite mixture model. When $G(u)$s have a corresponding probability density function (pdf), $g(u) = \dfrac{d}{du} G(u)$, it may also be written in analogous form to (1.2) as shown below

$$\begin{aligned} g(u) = \frac{\mathrm{d}}{\mathrm{d}u} G(u) &= \frac{\mathrm{d}}{\mathrm{d}u} \sum_{j=1}^{c} \pi_j F_j(u) \\ &= \sum_{j=1}^{c} \pi_j \frac{\mathrm{d}}{\mathrm{d}u} F_j(u) \\ &= \sum_{j=1}^{c} \pi_j f_j(u) \end{aligned} \tag{1.3}$$

where each f_j is the pdf of the corresponding distribution function F_j.

1.2.2 Parametric Specifications

Component distributions $F_j(u)$ are often chosen to be from some parametric family, $F(u; \boldsymbol{\gamma_j})$, which is indexed by a euclidean parameter vector $\boldsymbol{\gamma_j}$. Commonly the same family is chosen for all $j \in \{1, ..., c\}$, though this need not be the case (Grimlund, 1989). Most commonly, this family is chosen to be the Normal family of distributions (Fraley et al., 2012; G. J. McLachlan and Rathnayake, 2014; Maleki et al., 2019). When the random variable is multivariate, the multivariate normal family is commonly used for mixture modeling (NAKAMURA and KONISHI, 1999; Dolan et al., 2004; He et al., 2006; Boldea and Magnus, 2009). Other distributions such as the gamma distribution (Young et al., 2019), t-distribution (Burgess-Hull, 2020), and skewed t-distribution (Lin et al., 2007) have been studied as well.

1.2.3 Infinite Mixture Models

Infinite mixture models also exist where $\boldsymbol{\pi}$ is generalized to be a probability measure H over a parameter vector $\boldsymbol{\gamma}$ such that the infinite mixture distribution is defined by

$$G(u) = \int f(u; \boldsymbol{\gamma}) dH(\boldsymbol{\gamma}), \tag{1.4}$$

where $f(u; \boldsymbol{\gamma})$ is the family of densities indexed by the parameter $\boldsymbol{\gamma}$ and $H(\boldsymbol{\gamma})$ is called the mixing distribution. When $H(\boldsymbol{\gamma})$ is discrete with finite support (1.4) simplifies to (1.2) with finite c.

1.2.4 Framework for Interpretation of Sub-populations

A useful framework for working with mixture distributions is to note that a mixture distribution has the same distribution as the sum of independent variables as follows. Let $\boldsymbol{Z_i} = [1, 0, ..., 0]$ with probability π_1, $\boldsymbol{Z_i} = [0, 1, ..., 0]$ with probability π_2,..., $\boldsymbol{Z_i} = [0, 0, ..., 1]$ with probability π_c, independently for all $i \in \{1, ..., n\}$. That is, let $\boldsymbol{Z_i} \overset{iid}{\sim} Categorical\,(\boldsymbol{\pi})$ for $i \in \{1, ..., n\}$. Let $z_{i,j}$ be the jth element of $\boldsymbol{Z_i}$. Let $X_{i,j} \sim F_j$ independently for all $j \in \{1, ..., c\}$ and for all $i \in \{1, ..., n\}$. Also let $\boldsymbol{X_{i,\cdot}} = [X_{i,1}, ..., X_{i,c}]$ and let $\boldsymbol{X_{i,\cdot}}$ be independent of $\boldsymbol{Z_i}$. Then $Y_i \overset{def}{=} \boldsymbol{Z_i X_{i,\cdot}^T} \overset{iid}{\sim} G(y)$. See the proof below

$$\begin{aligned}
P(Y_i \leq y) &\overset{def}{=} P(\boldsymbol{Z_i X_{i,\cdot}^T} \leq y) \\
&= P\left(\bigcup_{j=1}^{c} \{z_{i,j} = 1 \cap (X_j \leq y)\} \right) \\
&= \sum_{j=1}^{c} P(z_{i,j} = 1) P(X_{i,j} \leq y | z_{i,j} = 1) \\
&= \sum_{j=1}^{c} P(z_{i,j} = 1) P(X_{i,j} \leq y) \\
&= \sum_{j=1}^{c} \pi_j F_j(y) \\
&= G(y)
\end{aligned}$$

where the steps hold by definition, the multiplication and addition rules, independence of $X_{i,\cdot}$ and Z_i, the definitions of Z_i and $X_{i,j}$, and the definition of G in (1.2).

1.2.5 Mixture of Regression Models

Mixture models can also be used to model the distribution of responses (or errors) in the context of models that include covariates, $\boldsymbol{x_i}$. For example, let there be c unobserved groups in the population where the ith observation comes from subgroup j, indicated by $z_{i,j} = 1$. Then

$$Y_i|(\boldsymbol{X}, Z = j) = \boldsymbol{X_i^T}\boldsymbol{\beta} + \epsilon_i, \tag{1.5}$$

where $\epsilon_i \overset{iid}{\sim} \phi(0, \sigma_j^2)$ and ϕ is the normal density. So marginally,

$$Y_i|\boldsymbol{X} \sim \sum_{j=1}^{c} \pi_j \phi(y; X_i^T \beta_j, \sigma_j^2). \tag{1.6}$$

Various extensions exist where each Y_i is multivariate (Soffritti and Galimberti, 2011), ϵ_i is non-normal (Zeller et al., 2016) or even estimated non-parametrically (Hunter and Young, 2012; Hu et al., 2017).

1.2.6 Supervision, Clustering and Classification

The data setup for mixture modeling can be classified according to the availability of $\boldsymbol{Z_i}$ for all observations Y_i for $i \in \{1, ..., N\}$. In the machine learning literature, data settings where no component labels are known are called unsupervised. Data settings where all component labels are known are called supervised. Data settings where a subset of available observations are of known component origin is called semi-supervised. Within the supervised (or semi-supervised) framework it is important to distinguish between types of supervision for accurate modeling (Hosmer Jr, 1973). The data scenarios are listed below.

- Unsupervised
- Supervised
  - Stratified Random Sampling
  - Simple Random Sampling
- Semi-Supervised
  - Stratified Random Sampling
  - Simple Random Sampling

Clustering is the act of grouping unsupervised observations into unique groups, called clusters. If the number of clusters, c, is known, the mixture model approach to clustering corresponds to fitting a mixture distribution with c clusters. However, sometimes the number of clusters is not known and thus the mixture modeling approach then considers c as a parameter rather than a known quantity (NAKAMURA and KONISHI, 1999; G. J. McLachlan and Rathnayake, 2014). The two most common mixture model approaches (G. J. McLachlan et al., 2019) to selecting c are maximizing a penalized log-likelihood and carrying out hypothesis tests using a Likelihood Ratio Test (LRT). For a classic and visual approach to selecting the number of components, silhouette diagrams may be used (Rousseeuw, 1987).

In a supervised setting, the observations are classified into pre-specified groups, indicated by the observed $\boldsymbol{Z}$. If the observations are randomly sampled from the overall population, then $\boldsymbol{Z_1, ..., Z_n} \overset{iid}{\sim} Categorical(\boldsymbol{\pi})$ and inference about $\boldsymbol{\pi}$ can be made directly from $\boldsymbol{Z_1, ..., Z_n}$. However, if stratified sampling is implemented from the sub-populations, then the membership labels $\boldsymbol{Z_1, ..., Z_n} \not\sim Categorical(\boldsymbol{\pi})$ and thus do not provide direct information about $\boldsymbol{\pi}$. In such a case, information is only available about the conditional distributions of $Y_i|(Z = j)$ and thus standard methods like ANOVA or regression may be implemented to conduct statistical inference about the parameters. In a machine learning context, classification of future observations into class membership is often the goal in any form of supervised setting. It should be noted that while an additional sample of observations from (1.2) is necessary for inference on $\boldsymbol{\pi}$ (Ilagan and Falk, 2022), classification techniques using mixture models (or other techniques) may still provide satisfactory classification metrics such as sensitivity $P(C = j|z_{i,j} = 1)$ and specificity $P(C \neq j|z_{i,j} \neq 1)$. Furthermore, posterior predictive probabilities $P(z_{i,j} = 1|C = j)$ may still be useful without this information if components are well-separated.

Lastly, the semi-supervised setting describes when n_1 observations have component labels and n_2 observations do not have component labels ($n_1 + n_2 = n$). The same discussion about stratified versus simple random sampling above applies to the supervised labels $Z_1, ..., Z_{n_1}$. Section 1.2.8 shows that while the sampling method alters how the maximum likelihood estimates are computed, even in the stratified scenario, $Y_1, ..., Y_{n_1}$ may still be utilized for inference on $\boldsymbol{\pi}$ since (Z_i, Y_i) consists of an observation from $Y_i|(z_{i,j} = 1)$ even when $Z_i \not\sim Bern(\pi_j)$.

Besides mixture modeling, many other methods exist in the machine learning literature for clustering (Rokach and Maimon, 2005; Saxena et al., 2017) and classification (Soofi and Awan, 2017; Dogan and Birant, 2021) problems.

1.2.7 Identifiability

For notational purposes, let $\pi_j, \pi_{j'}$ represent two different choices of the value of the same (jth) component probability. Also let π_{k1} and π_{k2} represent two arbitrary but distinct component probabilities. Analogous notation is used for $\boldsymbol{F}$.

One characteristic issue that arises in modeling data with a mixture distribution is identifiability.

Definition 1.2.1 *A distribution $G(u; \boldsymbol{\tau})$ is identifiable if $\boldsymbol{\tau} \neq \boldsymbol{\tau'} \implies G(u; \boldsymbol{\tau}) \neq G(u; \boldsymbol{\tau'})$ where $\boldsymbol{\tau}$ is the parameter vector of G.*

Mixture models carry an inherent identifiability issue called label-switching. For any mixture model (1.2), if $F_{k1} = F'_{k2},\ F'_{k1} = F_{k2}$ and $\pi_{k1} = \pi'_{k2},\ \pi_{k2} = \pi'_{k1}$ then $G(u; \boldsymbol{\tau}) = G(u; \boldsymbol{\tau'})$ where $\boldsymbol{\tau} = (\boldsymbol{\pi}, \boldsymbol{F})$. This means that the model is non-identifiable by the definition of identifiability, but only because of switching the labels of (π_{k1}, F_{k1}) and (π_{k2}, F_{k2}). In settings where the component labels $\{1, 2, ..., c\}$ are arbitrary in their interpretation, as in the unsupervised setting where clustering is the goal, this kind of non-identifiability is not problematic. Thus the definition of identifiability in an unsupervised mixture setting is modified to satisfy **Definition 1.2.1** *up to a permutation in labels.* However, if the various components have distinct interpretations, then this label-switching is an issue. In such scenarios supervised or semi-supervised data is a solution to the label-switching issue.

For an example of label-switching, consider the following normal mixture model

$$G(u; \pi_1, \mu_1, \sigma_1^2, \pi_2, \mu_2, \sigma_2^2) = \pi_1 \Phi(u; \mu_1, \sigma_1^2) + \pi_2 \Phi(u; \mu_2, \sigma_2^2), \tag{1.7}$$

where $\Phi(u)$ is a normal CDF and thus in this context $\boldsymbol{\tau} = (\pi_1, \mu_1, \sigma_1^2, \pi_2, \mu_2, \sigma_2^2)$. Consider that $G(u; .2, 5, 1, .8, 10, 4) = G(u; .8, 10, 4, .2, 5, 1)$, so strictly speaking, the model is not identifiable. However, the only difference is what label (1 or 2) is given to the component that has the smaller mean, mixing proportion, and variance. In scenarios where the aspects of the component distributions ought to match with particular labels, additional restrictions (such as $\pi_1 < ... < \pi_c$ or $\mu_1 < ... < \mu_c$) may be imposed.

Another manner in which identifiability breaks down for (1.2) is by failing to ensure that each component distribution F_j is a 'true component' of G. This occurs when $\pi_j = 0$ for one or more $j \in \{1, ..., c\}$, or when at least two $F_{k1},\ F_{k2}$ are non-distinct.

There are other (less trivial) ways in which non-identifiability of the model can arise. Without any further restrictions on $\boldsymbol{F}$ (other than that each element be distinct), it is not possible to guarantee identifiability of G. Consider the following example where the component distributions, F_j are distinct with positive component probabilities $\pi_j > 0$ but the model is non-identifiable. Let $c = 2$, and $F_1(u) \sim (1 - w_1)N(0, 1) + w_1 N(2, 1)$ and $F_2(u) \sim (1 - w_2)N(0, 1) + w_2 N(2, 1)$ where $w_1 \neq w_2$. If $\pi_2 w_1 + \pi_1 w_2 = p$ for any $p \in [0, 1]$, then $G(u) = (1 - p)N(0, 1) + pN(2, 1)$. For example, if $(w_1 = .4, w_2 = .7, \pi_1 = .4, \pi_2 = .6)$ or $(w_1' = .5, w_2' = .8, \pi_1' = 11/15, \pi_2' = 4/15)$ then $p = .58$. Analogous non-identifiable cases for $c > 2$ abound as well.

These examples highlight the need for further restriction on the mixture model in order to ensure identifiability. One common way to do this is to specify a parametric family for the component distributions. If c is known and finite, then each of $F_1, ..., F_c$ can have a specified family (perhaps the same family, perhaps different). Titterington et al. (1985) showed that with many common families for $\boldsymbol{F}$, G becomes identifiable barring label-switching with a notable exception being the uniform distribution.

To see why (1.3) is not identifiable if all f_j (or more than one) follow a uniform distribution, consider the following example. Let $c = 2$ and choose the following two sets for the parameters ($\pi_1 = 1/4,\ \pi_2 = 3/4,\ f_1 = U_{(-1,1)},\ f_2 = U_{(1,7)}$) and ($\pi_1' = 1/3,\ \pi_2' = 2/3,\ f_1^{'} = U_{(-1,1)},\ f_2^{'} = U_{(-1,7)}$) then $g(u) = g^{'}(u)$ for all u (1.8) thus showing that g is not identifiable.

$$g(u) = g'(u) = \begin{cases} 0 & u < -1 \\ \frac{9}{24} & -1 \leq u < 1 \\ \frac{1}{24} & 1 \leq u < 7 \\ 0 & 7 \leq u \end{cases} \tag{1.8}$$

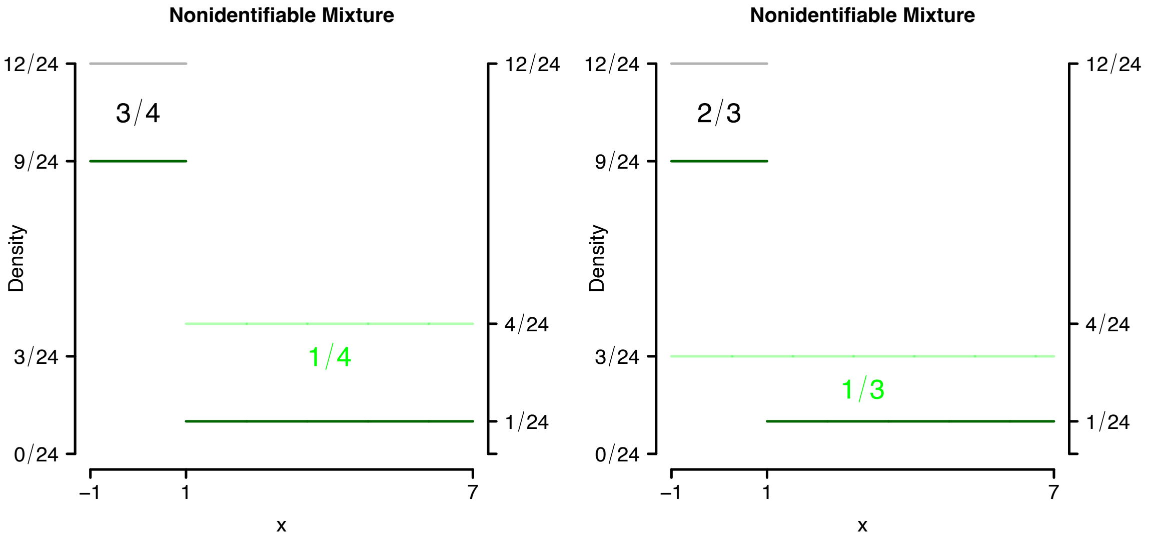


Figure 1.2: Nonidentifiable Mixture of Uniforms. The plot on top represents f_1 in grey, f_2 in light green, and g in dark green. The bottom plot with f_1', f_2', g' is analogous.

1.2.8 The EM Algorithm

Maximum Likelihood Estimation is a stalwart approach for statistical inference dating back to the origin of the field of Statistics (Edgeworth, 1908; Wilks, 1938). Maximum

Likelihood Estimation involves finding the parameter value $\boldsymbol{\tau}$ that maximizes the likelihood function for a given set of data $Y_1, ..., Y_n$ sampled from the model. Since the log is a bijective function that is easier to work with (and the maximizer of the log-likelihood function is the same as that of the likelihood function) attention is given to maximizing the log-likelihood function. An observed random sample $Y_1, ..., Y_n$ from (1.3), where each $F_j(y; \boldsymbol{\gamma_j})$ has a parametric family indexed by $\boldsymbol{\gamma_j}$, provides an (observed) log-likelihood function

$$l_{obs}(\boldsymbol{\tau}; \boldsymbol{Y}) = \sum_{i=1}^{n} log \left\{ \sum_{j=1}^{c} \pi_j f_j(Y_i; \boldsymbol{\gamma_j}) \right\}. \tag{1.9}$$

Analytical maximization of (1.9) is not feasible due to the component summation. However, the latent variable framework from section 1.2.4 makes maximization of the complete log-likelihood possible. The unobserved $\boldsymbol{Z_1, ..., Z_n}$ contain the information missing from (1.9) for the complete data log-likelihood

$$l_c(\boldsymbol{\tau}; \boldsymbol{Y, Z}) = \sum_{i=1}^{n} \sum_{j=1}^{c} z_{i,j} \left\{ log\left(\pi_j\right) + log\left(f_j(Y_i; \boldsymbol{\gamma_j})\right) \right\} \tag{1.10}$$

where $\boldsymbol{\tau} = (\boldsymbol{\pi}, \boldsymbol{\gamma})$ (and $\boldsymbol{\gamma} = [\boldsymbol{\gamma_1}, ..., \boldsymbol{\gamma_c}]$).

The missing data setup of the mixture model lends itself naturally to implementation of the Expectation Maximization (EM) algorithm for maximizing (1.9). Such an algorithm was originally discussed by Day (1969) for the two-component location-shifted (possibly multivariate) normal mixture models. The EM algorithm was then formalized by Dempster et al. (1977), which finds a local maximum of the observed log likelihood (1.9) by iterating back and forth between 'Expectation' and 'Maximization' steps computed from the complete log-likelihood (1.10). The algorithm begins at step $k = 0$ with an initial estimate of the parameter $\boldsymbol{\tau^{(0)}} = (\boldsymbol{\pi^{(0)}}, \boldsymbol{\gamma^{(0)}})$. Then the

'E'-step is performed by computing the conditional expectation of (1.10) given $\boldsymbol{Y}$ and $\boldsymbol{\tau^{(k)}}$, as below

$$Q(\boldsymbol{\tau};\boldsymbol{\tau^{(k)}}) = E_{\boldsymbol{Z}}\left[l_c(\boldsymbol{\tau};\boldsymbol{Y},\boldsymbol{Z})|\boldsymbol{Y},\boldsymbol{\tau}=\boldsymbol{\tau^{(k)}}\right]. \tag{1.11}$$

This reduces to

$$\sum_{i=1}^{n}\sum_{j=1}^{c} E[z_{i,j}|Y_i]\left\{log\left(\pi_j^{(k)}\right)+log\left(f_j(Y_i;\boldsymbol{\gamma_j^{(k)}})\right)\right\} \tag{1.12}$$

by the linearity of expectation. Note that since $z_{i,j}$ is an indicator variable, $E[z_{i,j}|y_i] = P(z_{i,j}=1|y_i)$, and by Bayes' rule

$$P(z_{i,j'}=1|y_i) = \frac{\pi_{j'}f_{j'}(y_i;\boldsymbol{\gamma_{j'}})}{\sum_{j=1}^{c}\pi_j f_j(y_i;\boldsymbol{\gamma_j})}. \tag{1.13}$$

Thus, the 'E' step reduces to computing $p_{i,j}^{(k)} = P\left(z_{i,j}=1|y_i,\boldsymbol{\pi^{(k-1)}},\boldsymbol{\gamma^{(k-1)}}\right)$ for all i,j.

Next, the 'M' step proceeds by maximizing $Q(\boldsymbol{\tau};\boldsymbol{\tau^{(k-1)}})$ over $\boldsymbol{\tau}$ to give an updated value, $\boldsymbol{\tau^{(k)}}$. Note, that maximizing $Q(\boldsymbol{\tau};\boldsymbol{\tau^{(k-1)}})$ can often be done analytically (in particular with f_j all from exponential families) and is more feasible than trying to directly maximize $l_{obs}(\boldsymbol{\tau};\boldsymbol{Y})$. These 'E' and 'M' steps are repeated for increasing k until $l_{obs}(\boldsymbol{\tau^{(k)}};\boldsymbol{Y})$ reaches a (local) maximum. The key property of the EM algorithm for maximizing (1.9) is that

$$l_{obs}(\boldsymbol{\tau^{(k)}};\boldsymbol{Y}) \geq l_{obs}(\boldsymbol{\tau^{(k-1)}};\boldsymbol{Y}) \tag{1.14}$$

for all $k \in \{0,1,2,...\}$. This non-decreasing property ensures that $\boldsymbol{\tau^{(k)}}$ must converge to a local maximum of $l_{obs}(\boldsymbol{\tau};\boldsymbol{Y})$. When there are multiple local maxima (which is common with (1.2) type models), starting the algorithm at various $\boldsymbol{\tau^{(0)}}$ can uncover multiple local maximizers and the value of $l_{obs}(\boldsymbol{\tau}; \boldsymbol{Y})$ can be computed at each to determine which is the largest. While there is no guarantee that this identifies the global maximizer, the chance of obtaining the global maximizer increases in scenarios where a vast array of starting values is computationally feasible or there are computationally inexpensive estimates (like method of moments) that may provide an initial value close to the global maximizer.

1.2.9 Semi-Parametric Modeling

In contrast to modeling (1.2) with parametric assumptions on the component distributions as described in section 1.2.2, much recent work has been done on semi-parametric modeling of (1.2) where the component distributions are estimated non-parametrically. See Xiang et al. (2019) for a recent review. Without parametric specification of the component densities, additional assumptions must be present for identifiability to hold. Chang and Walther (2007a) assume log-concave components for location-shifted distributions, but this is insufficient to ensure identifiability. Bordes, Mottelet, et al. (2006) showed that for two-component location shifted mixture with symmetric component densities, the model is identifiable. Bordes et al. (2007) and Hunter et al. (2007) consider estimation of the component probabilities and location parameters for such a model. Another two-component model was motivated to detect differentially expressed genes under two or more conditions in microarray data where one component is known and the other is symmetric with finite third moment (Bordes, Delmas, et al., 2006). Estimation for this model has since been further explored (Bordes and Vandekerkhove, 2010; Hohmann and Holzmann, 2013; Xiang et al., 2014; Patra and Sen, 2016).

1.3 Model Preliminaries

1.3.1 Identifiability

Consider the special mixture model (1.1) of interest in this dissertation and note that it is a special case of (1.2) where $c = 2$, $\pi_2 = \theta$, $F_1(u) = F(u)$, and $F_2(u) = F(u - \delta)$. The restrictions that $\theta \in (0, 1]$ and $\delta > 0$ or $(\theta, \delta) = (0, 0)$ alone are insufficient to ensure identifiability of (1.1). To see why, consider the following example that produces two distinct triplets (F, θ, δ) and (F', θ', δ') such that $G(u; F, \theta, \delta) = G(u; F', \theta', \delta')$ for all $u \in \mathcal{R}$. Let $(F, \theta, \delta) = (N(0, 1), .7, 1.5)$ and let $(F', \theta', \delta') = (.3N(-2, 1) + .7N(-.5, 1), 1, 2)$. Then $G(u; F, \theta, \delta) = G(u; F', \theta', \delta') = .3N(0, 1) + .7N(1.5, 1)$. Thus in an unsupervised setting - where there is no control data $\boldsymbol{X} \overset{iid}{\sim} F$, but only $\boldsymbol{Y} \overset{iid}{\sim} G$ - the model (1.1) is not identifiable.

As mentioned in sections 1.2.2 and 1.2.9, it is common to restrict F by imposing parametric assumptions or shape constraints to achieve identifiability. However, in the semi-supervised setting, data from $\boldsymbol{X} \overset{iid}{\sim} F$ is available and thus F (a sub-model of G) is identifiable. Consequently in this semi-supervised setting, model (1.1) is identifiable so long as there is no distinct pair of parameters $(\theta, \delta), (\theta', \delta')$ that produce the same equation for G in model (1.1) [for the same F]. Yakowitz and Spragins (1968) showed that finite location-shifted mixture distributions are identifiable for any F. Thus the model is identifiable in this semi-supervised setting without any constraints on F.

1.3.2 Parametric Inference with Treatment Data Only

To motivate the need for semi-parametric inference on (1.1) with control data, consider first the inference problem with only treatment data and parametric assumptions on F. Recall that $(Y_1, Z_1), ..., (Y_n, Z_n)$ represent the n paired observations for the randomly sampled individuals who are given the treatment under consideration. $Y_1, ..., Y_n$ represent the observed response of each individual and $Z_1, ..., Z_n$ represent the unobserved sub-population to which each of the n sampled individuals belongs (each individual is either a non-responder or a responder). Let $Z_i \sim Bernoulli(\theta)$, $Y_i|(Z_i = 0) \sim F(u)$, $Y_i|(Z_i = 1) \sim F(u - \delta)$. Therefore, $Y_1, ..., Y_n \overset{iid}{\sim} G$ in (1.1).

Consider the MLE of $\boldsymbol{\tau} = (\theta, \mu_0, \mu_1, \sigma^2)$ with the correct assumption that $F(u) \sim N(\mu_0, \sigma^2)$ and $F(u - \delta) \sim N(\mu_1, \sigma^2)$ $[\delta = \mu_1 - \mu_0 > 0]$ by implementation of the EM algorithm. The algorithm is defined by

0.) Initialize: $\boldsymbol{\tau^{(0)}}$

1.) E-step: Compute $Q(\boldsymbol{\tau}|\boldsymbol{\tau^{(k-1)}}) = E_{\boldsymbol{Z}}[logL_c(\boldsymbol{\tau; (Y, Z)})|\boldsymbol{Y, \tau^{(k-1)}}]$

2.) M-step: Maximize $Q(\boldsymbol{\tau}|\boldsymbol{\tau^{(k-1)}})$ over $\boldsymbol{\tau}$ to give an updated value $\boldsymbol{\tau^{(k)}}$,

where alternation between 1.) and 2.) repeats until convergence of $\boldsymbol{\tau^{(k)}}$ occurs. Since the only random variables from $logL_c(\boldsymbol{\tau; (Y, Z)})$ in 1.) are $Z_1, ..., Z_n$, it is useful to define $p_i^{(k)} = E[Z_i|\boldsymbol{Y, \tau^{(k-1)}}]$.

Based on the solutions to the expectation and the maximization steps, the initialization and k^{th} steps of the algorithm are

0. Initialize: $\boldsymbol{\tau^{(0)}} = (\theta^{(0)}, \mu_0^{(0)}, \mu_1^{(0)}, \sigma^{2^{(0)}})$ $[\delta^{(0)} = \mu_1^{(0)} - \mu_0^{(0)}]$
1. E-step: $p_i^{(k)} = \dfrac{\theta^{(k-1)} f^{(k-1)}(y_i - \delta^{(k-1)})}{(1-\theta^{(k-1)}) f^{(k-1)}(y_i) \ + \ \theta^{(k-1)} f^{(k-1)}(y_i - \delta^{(k-1)})}$
2. M-step:

$$\theta^{(k)} = \frac{\sum_{i=1}^n p_i^{(k)}}{n}$$

$$\mu_0^{(k)} = \frac{\sum_{i=1}^n (1 - p_i^{(k)}) y_i}{\sum_{i=1}^n (1 - p_i^{(k)})}$$

$$\mu_1^{(k)} = \frac{\sum_{i=1}^n p_i^{(k)} y_i}{\sum_{i=1}^n p_i^{(k)}}$$

$$\sigma^{2^{(k)}} = \frac{\sum_{i=1}^n (1 - p_i^{(k)})(y_i - \mu_0^{(k)})^2 \ + \ p_i^{(k)}(y_i - \mu_1^{(k)})^2}{n}.$$

Notice that since $F(u) \sim N(u; \mu_0, \sigma^2)$, each update of $(\mu_0^{(k)}, \sigma^{2^{(k)}})$ uniquely defines $f^{(k)}$.

Consider the simulation to assess the performance of $\widehat{(\theta, \delta)}$ on 1000 independent data sets when $n = 100$, $F \sim N(0, 1)$, $\theta = .7, \delta = 2$. The simulation results are obtained using the Mclust package in R. Notice from the top panel in **Figure 1.3** that the mixture model has (θ, δ) such that the separation between the sub-population component distributions is not substantial enough to produce a bimodal mixture distribution. However, the effect size is two standard deviations ($K = \delta/\sigma = 2/1 = 2$) and a moderate percentage of the population ($\theta = 70\%$) is a responder. A treatment with 70% responders may be of interest, and for most medical applications it is not feasible to demand an effect size more than two standard deviations. Thus it is desirable to perform inference on a mixture model even if the mixture distribution is not bimodal.

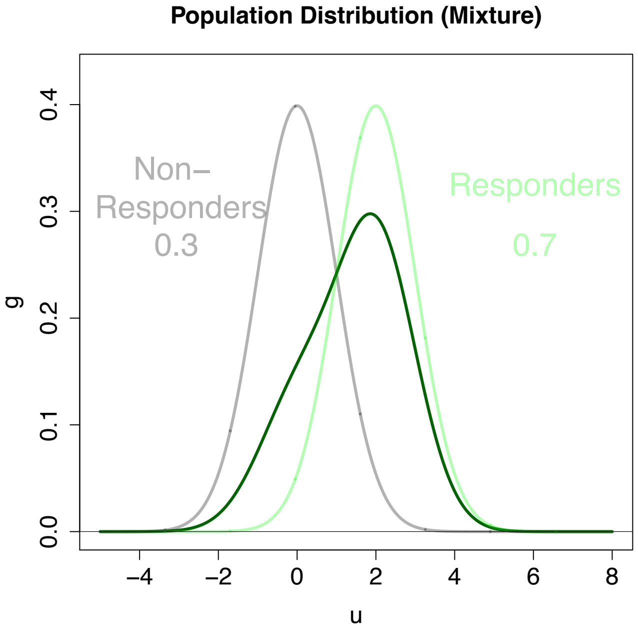


Figure 1.3: Data Collection for Simulation Setting. Observations come from mixture represented by dark green curve. $n = 100, F \sim N(0, 1), \theta = .7, \delta = 2$.

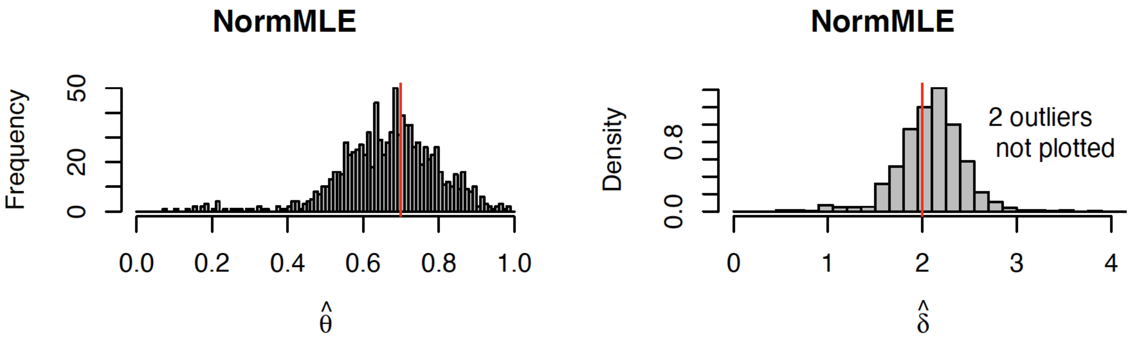


Figure 1.4: Normal Maximum Likelihood Estimate found by EM Algorithm. $n = 100$, $F \sim N(0, 1)$, $\theta = .7$, $\delta = 2$.

Figure 1.4 displays the distributions of $\widehat{\theta}$ and $\widehat{\delta}$ and shows that the estimators are useful. The estimator typically indicates that between 50% to 90% of the population responds to the treatment and the magnitude of the response for the responders is between 1.5 and 2.5 units.

1.3.3 Semi-parametric Inference with Treatment Data Only

As stated in section 1.1, this dissertation aims to conduct inference that is robust to distributional assumptions. An existing method for estimating (θ, δ) is an EM-like algorithm (Bordes, Mottelet, et al., 2006) where F is estimated non-parametrically, which can be implemented using the `mixtools` package in R. First define F_s to be a symmetric distribution function around 0 with density f_s, $f(u) = f_s(u - \mu_0)$ and $f(u - \delta) = f_s(u - \mu_1)$ where $\delta = \mu_1 - \mu_0 > 0$. The symmetry of f_s is necessary for model identifiability. The algorithm is as follows

0. Initialize: $\boldsymbol{\tau^{(0)}} = (\theta^{(0)}, \mu_0^{(0)}, \mu_1^{(0)}, f_s^{(0)})$ $[\delta^{(0)} = \mu_1^{(0)} - \mu_0^{(0)}]$
1. E-step: $p_i^{(k)} = \dfrac{\theta^{(k-1)} f^{(k-1)}(y_i - \delta^{(k-1)})}{(1-\theta^{(k-1)}) f^{(k-1)}(y_i) + \theta^{(k-1)} f^{(k-1)}(y_i - \delta^{(k-1)})}$
2. 'M'-step:

$$\theta^{(k)} = \frac{\sum_{i=1}^{n} p_i^{(k)}}{n}$$

$$\mu_0^{(k)} = \frac{\sum_{i=1}^{n}(1-p_i^{(k)}) y_i}{\sum_{i=1}^{n}(1-p_i^{(k)})}$$

$$\mu_1^{(k)} = \frac{\sum_{i=1}^{n} p_i^{(k)} y_i}{\sum_{i=1}^{n} p_i^{(k)}}$$

$$f_s^{(k)} = ...$$

Now f_s is itself a parameter to be estimated and is no longer uniquely determined by a set of euclidean parameters - as is the case when assuming a particular parametric family for the component distribution. Since finding the particular $f_s^{(k)}$ to maximize $Q(\boldsymbol{\tau}|\boldsymbol{\tau^{(k-1)}})$ is a difficult task, the method of estimation opts rather to estimate f_s by kernel density estimation using a simulation technique to 'complete the data' as follows

(a) Simulate $\tilde{z}_i^{(k)} \sim Bernoulli(p_i^{(k)})$

(b) $\tilde{y}_i^{(k)} = y_i - \mu_{\tilde{z}_i}^{(k)}$

(c) $\widehat{f}_n^{(k)}(u) = \dfrac{1}{nh} \sum_{i=1}^{n} K(\dfrac{u - \tilde{y}_i}{h})$

(d) $f_s^{(k)}(u) = \dfrac{\widehat{f}_n^{(k)}(u) + \widehat{f}_n^{(k)}(-u)}{2}$

In step (a), each observation is randomly assigned to either the 'non-responder' component or the 'responder' component according to its current probability of component membership. In step (b) each observation is 'recentered' by subtracting the current estimate of the center of its respective component assignment. (Note that if all simulation assignments and the center estimates are correct, then $y_1, ..., y_n \overset{iid}{\sim} F_s$.) In step (c), a kernel density estimate is fit on all the 'recentered' data providing a preliminary estimate of f_s. In step (d), an additional symmetrization step is performed to ensure a symmetric estimate of f_s. (A deterministic version of the algorithm (Benaglia et al., 2009) exists as well where the p_js are used directly to assign weighted observations to the group instead of simulating full membership in one group. The deterministic version performs similarly to the stochastic version.) Consider a simulation to assess the performance of $\widehat{(\theta, \delta)}$ on 1000 independent data sets under the same sampling scheme as in **Figure 1.3**

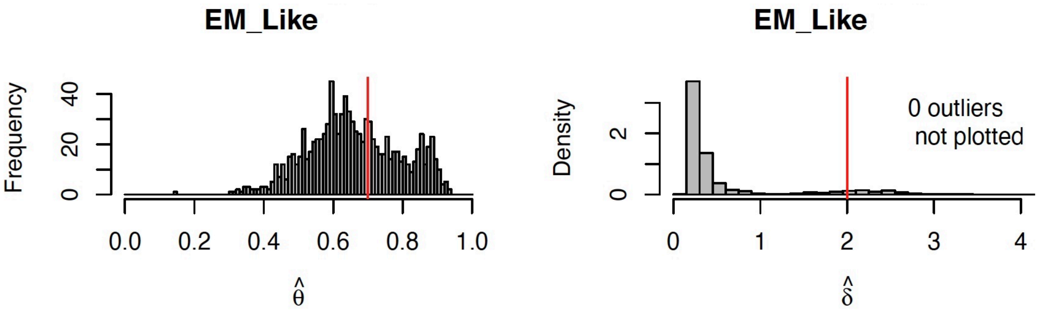


Figure 1.5: EM-like Algorithm with Only Treatment Data. $n = 100$, $F \sim N(0,1)$, $\theta = .7$, $\delta = 2$.

As is very clear from **Figure 1.5**, the simulation study shows that the performance of the EM-like estimator is unsatisfactory in this scenario. Particularly $\widehat{\delta}$ dramatically underestimates δ. By comparing **Figure 1.4** and **Figure 1.5**, it appears that the increased flexibility in allowing F to be any symmetric distribution comes at a steep price. The flexibility in F makes it difficult to identify separate sub-populations when the separation between the components is not pronounced enough to obviously see them in the resulting mixture - even though the model is theoretically identifiable (Bordes, Mottelet, et al., 2006). To verify that the EM-like algorithm can work in some situations, consider a larger effect size, $K = 4$ (see **Figure 1.6**).

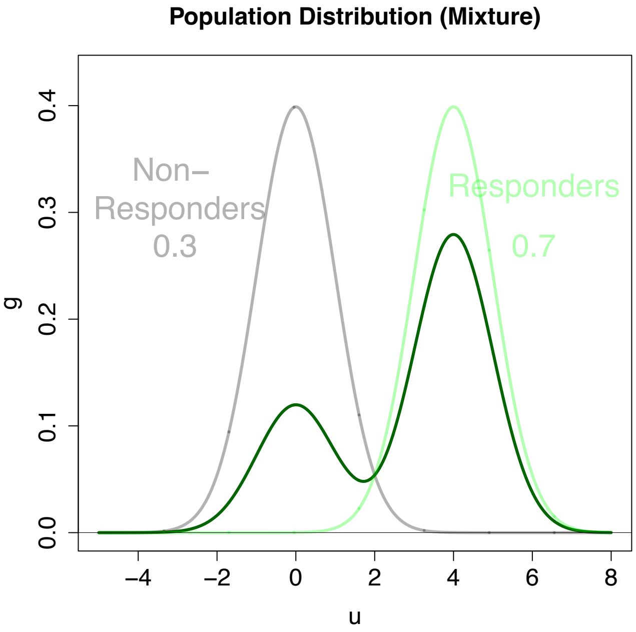


Figure 1.6: Data Collection for Simulation Setting with Large Shift. Observations come from the mixture distribution represented by dark green. $n = 100, F \sim N(0,1), \theta = .7, \delta = 4$

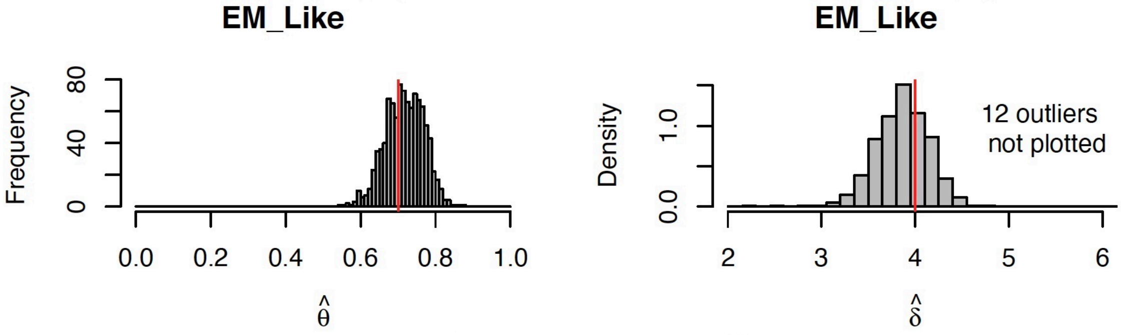


Figure 1.7: EM-like Algorithm with Only Treatment Data. $n = 100$, $F \sim N(0,1)$, $\theta = .7$, $\delta = 4$

As seen from the simulation with a larger effect size, the bias in $\widehat{\delta}$ decreases substantially to something more reasonable, and the variance of $\widehat{\theta}$ decreases as well. However, as previously noted, in many applications of interest it may not be reasonable to assume that the effect size is so large. So for more realistic effect sizes, where the resulting mixture is not so prominently bimodal, it appears that the loss of information about F that results from relaxing the normality assumption makes inference about (θ, δ) problematic. Therefore, consider another way of obtaining information about F - control data.

Recall that the 'non-responder' sub-population is a subset of individuals who - when given the treatment - do not respond. Thus individuals given a control intervention have the same response distribution as individuals who do not respond to the treatment. This dissertation aims to show that when a sample $X_1, ..., X_m \overset{iid}{\sim} F(u)$ from a control group is available, this information proves useful for inference on (θ, δ) without restricting F to be from some known parametric family.

1.3.4 Situating the Research in the Literature

Sections 1.2.2 and 1.2.9 surveyed a host of related literature that exists regarding mixture models where the component distributions are estimated both parametrically and non-parametrically. The specific setup, scope, and focus of the related work can be characterized as in **List 1.1** below. The bolded (and italicized) elements describe how this dissertation research fits into the broader landscape.

- Number of components
  - Unknown
  - **Known**
    - **$c = 2$**
    - $c > 2$ and c finite
    - Infinite
- Data Dimension
  - **Univariate**
  - Multivariate
- Data Type(s)
  - Categorial
  - **Numeric**
  - Mixed
- Assumption on F_js
  - Parametric
  - Symmetric
  - *Log-Concave*
  - *Certain Finite Moments*
  - *Unimodal*
  - *None*
- Assumption between F_js
  - Same Specified Family
  - **Location-shifted**
  - Scaled (same location)
  - Location-Scale
  - Linear Log-ratio
  - None
- Data Collection from
  - G (with Z_i known)
  - **G (with Z_i unknown)**
  - All F_js
  - **Some F_js**
  - No F_js
- Primary Inference Objective
  - G
  - F_js
  - **Euclidean parameters**
  - p_i and z_i prediction

List 1.1: Summary of Mixture Modeling Scenarios

The assumptions on F vary from method to method in this dissertation, but typically have only restrictions that seem broadly applicable to modeling medical outcomes - such as F unimodal with reasonably well-behaved tails. The next chapter presents various approaches to distribution-robust point estimation of (θ, δ).

Chapter 2

Estimation Methods

2.1 Normal Maximum Likelihood Estimator using the EM Algorithm

While model (1.1) does not assume that $F \sim$ Normal, an estimation approach that utilizes a normality assumption may still provide distribution-robust performance. Particularly, as discussed in sections 1.3.2 and 1.3.3, relaxing the normality assumption (by imposing only a symmetric assumption on F) dramatically deteriorates performance of the EM-like estimator in the absence control data. If control data is sparse, then it may be the case that an erroneous distributional assumption (i.e. assume $F \sim$ Normal when $F \not\sim$ Normal) could result in better inference on (θ, δ) than methods that impose no assumptions on F. Thus, adapting the standard EM approach for finding the normal maximum likelihood estimate to the context with additional control data could be beneficial. This estimator provides a benchmark for performance comparisons with other estimators presented in this chapter that have less stringent assumptions on F.

Consider use of the EM algorithm to find the MLE assuming $F \sim N(\mu, \sigma)$ where the available data is $\boldsymbol{X} \overset{iid}{\sim} F$ and $\boldsymbol{Y} \overset{iid}{\sim} G$ from (1.1). The log-likelihood is then

$$\begin{aligned} l_{obs}(\mu, \sigma, \theta, \delta; \boldsymbol{X}, \boldsymbol{Y}) = &\sum_{j=1}^{m} log\left\{f(x_j; \mu, \sigma)\right\} \; + \\ &\sum_{i=1}^{n} log\left\{(1-\theta) f(y_i; \mu, \sigma) + \theta f(y_i - \delta; \mu, \sigma)\right\} \end{aligned} \tag{2.1}$$

and the complete log-likelihood is

$$\begin{aligned} l_{c}(\mu, \sigma, \theta, \delta; \boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}) = &\sum_{j=1}^{m} log\left\{f(x_j; \mu, \sigma)\right\} \; + \\ &\sum_{i=1}^{n} log\left\{(1-z_i) f(y_i; \mu, \sigma) + z_i f(y_i - \delta; \mu, \sigma)\right\}. \end{aligned} \tag{2.2}$$

Let k be the iterating index for the EM algorithm. Based on the solutions to the expectation and the maximization steps, implementation of the algorithm is as follows

0. Initialize: $\boldsymbol{\tau^{(0)}} = (\theta^{(0)}, \mu_0^{(0)}, \mu_1^{(0)}, \sigma^{2^{(0)}})$ $[\delta^{(0)} = \mu_1^{(0)} - \mu_0^{(0)}]$
1. E-step: $p_i^{(k)} = \dfrac{\theta^{(k-1)} f^{(k-1)}(y_i - \delta^{(k-1)})}{(1-\theta^{(k-1)}) f^{(k-1)}(y_i) \; + \; \theta^{(k-1)} f^{(k-1)}(y_i - \delta^{(k-1)})}$
2. M-step:
   (a) $\theta^{(k)} = \dfrac{\sum_{i=1}^{n} p_i^{(k)}}{n}$
   (b) $\mu_0^{(k)} = \dfrac{\sum_{j=1}^{m} x_j \; + \; \sum_{i=1}^{n}(1-p_i^{(k)}) y_i}{m \; + \; \sum_{i=1}^{n}(1-p_i^{(k)})}$
   (c) $\mu_1^{(k)} = \dfrac{\sum_{i=1}^{n} p_i^{(k)} y_i}{\sum_{i=1}^{n} p_i^{(k)}}$
   (d) $\sigma^{2^{(k)}} = \dfrac{1}{m+n} \sum_{j=1}^{m}(x_j - \mu_0^{(k)})^2$
   $+ \dfrac{1}{m+n} \sum_{i=1}^{n} \left\{(1-p_i^{(k)})(y_i - \mu_0^{(k)})^2 \; + \; p_i^{(k)}(y_i - \mu_1^{(k)})^2\right\}.$

Notice that the control data contributes direct information to the updates of $\mu_0^{(k)}$ and $\sigma^{2^{(k)}}$ which uniquely determine $f^{(k)}(u)$. As mentioned in section 1.2.8, best practice is to run the algorithm with multiple starting values and compare the observed log-likelihood at all convergent points, choosing the one with the largest observed log-likelihood as the point estimate. The choice of starting value for implementation using NormEM2loc() (see section A.10 of the Appendix) initializes with the E-step by selecting a percentage of the largest observations in the treatment group to assign as responders $(0\%, 20\%, 40\%, 60\%, 80\%, 100\%)$. The Y_i values assigned $p_i^{(0)} = 1$ are in the group with larger mean, while the Y_i values assigned $p_i^{(0)} = 0$ are in the group with smaller mean. This is equivalent to selecting $\theta^{(0)} \in \{0, .2, .4, .6, .8, 1\}$ with $\mu_0^{(0)}, \mu_1^{(0)}, \sigma^{2^{(0)}}$ computed using equations (b)-(d) (according to the component membership assignment). By definition of the parameter space, if either $\widehat{\theta} = 0$ or $\widehat{\delta} = 0$ then $\widehat{(\theta, \delta)} = (0, 0)$. For a data set simulated from $m = 100,\ n = 100,\ F \sim N(0, 1),\ \theta = .7,\ \delta = 2$ (as shown in **Figure 2.1**), the R output from NormEM2loc() is displayed in **Figure 2.2**. This output displays a matrix with six rows representing the points of convergence from the six different initial values. The convergent point that achieves the highest log likelihood is indicated by a 1 in row 6 of the last column (labeled 'max.LL'). The iterations for the EM algorithm for this run are graphically displayed in **Figure 2.3**.

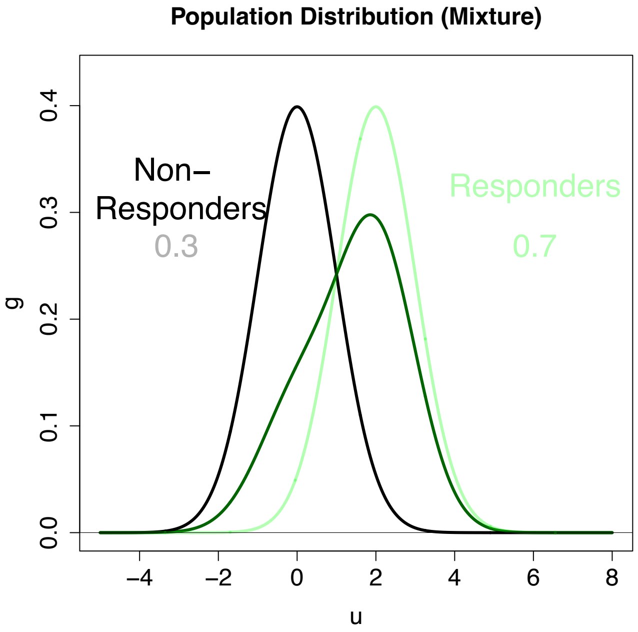
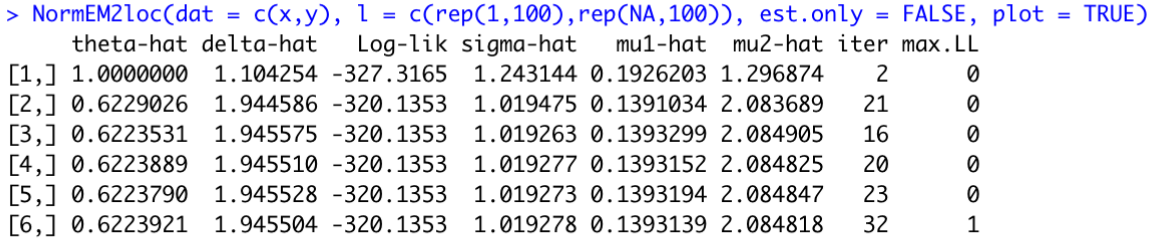


Figure 2.1: Data Collection with Control and Treatment Data. $m = 100$, $n = 100$, $F \sim N(0,1)$, $\theta = .7$, $\delta = 2$.



```
> NormEM2loc(dat = c(x,y), l = c(rep(1,100),rep(NA,100)), est.only = FALSE, plot = TRUE)
     theta-hat delta-hat   Log-lik sigma-hat   mu1-hat  mu2-hat iter max.LL
[1,] 1.0000000  1.104254 -327.3165  1.243144 0.1926203 1.296874    2      0
[2,] 0.6229026  1.944586 -320.1353  1.019475 0.1391034 2.083689   21      0
[3,] 0.6223531  1.945575 -320.1353  1.019263 0.1393299 2.084905   16      0
[4,] 0.6223889  1.945510 -320.1353  1.019277 0.1393152 2.084825   20      0
[5,] 0.6223790  1.945528 -320.1353  1.019273 0.1393194 2.084847   23      0
[6,] 0.6223921  1.945504 -320.1353  1.019278 0.1393139 2.084818   32      1
```

Figure 2.2: Normal EM Output with six initial values.

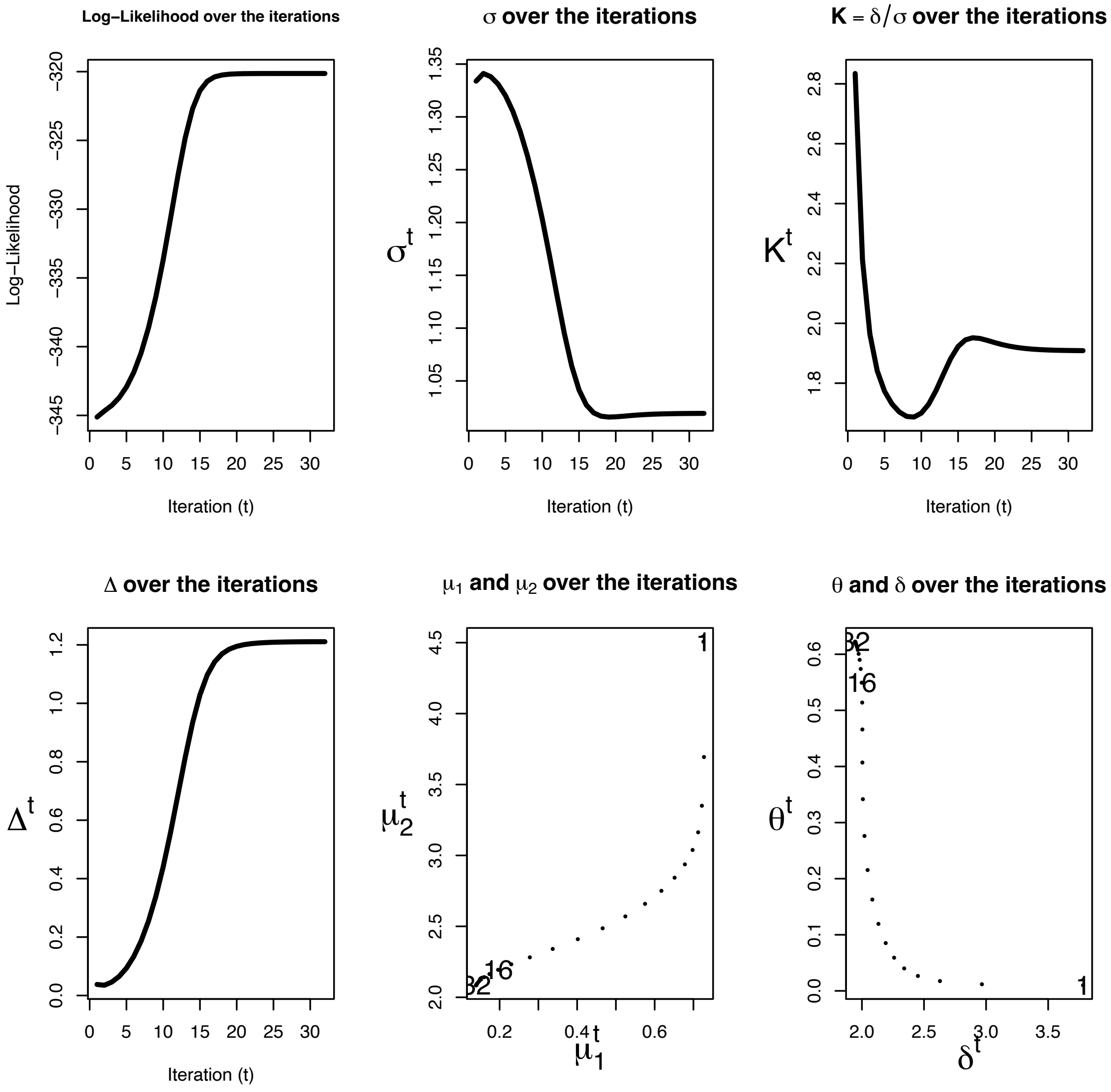


Figure 2.3: Plotting Output of EM Algorithm for Normal MLE.

The point with the highest log-likelihood, indicated by the last column 'max.LL', is the estimate – $\widehat{(\theta, \delta)} = (.62, 1.95)$, $\widehat{\sigma} = 1.02$, $\widehat{\Delta} = \widehat{\theta}\,\widehat{\delta} = 1.21$, $\widehat{K} = \widehat{\delta}/\widehat{\sigma} = 1.91$. **Figure 2.2** shows that multiple starting values for this data set – $\theta^{(0)} \in \{0, .2, .4, .6, .8\}$ – converge to this estimate (with negligible discrepancies). The plots in **Figure 2.3** show the path toward the estimate and verify that the log-likelihood is non-decreasing.

2.2 Method of Moment Estimation

Note that much of section 2.2 below is identical to content from previously published work (Lubich et al., 2022).

Recall that $X_1, ..., X_m \overset{iid}{\sim} F$ and $Y_1, ..., Y_n \overset{iid}{\sim} G$ from (1.1) where $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent for a total sample size of $N = m + n$. Denoting the mean of F and G by μ_X and μ_Y, respectively, then $\mu_Y = \mu_X + \Delta$ (see proposition 2.2.1) and therefore a (modified) method of moment estimator for Δ is $\widehat{\Delta} = (\overline{Y} - \overline{X})_+$ where $t_+ = t$ if $t > 0$ and 0 otherwise. The + operator restricts $\widehat{\Delta}$ to remain in the parameter space. Jeske and Yao (2020) further proposed method of moment estimators for the parameters in model (1.1) of the form

$$\widehat{\Delta} = (\overline{Y} - \overline{X})_+ \tag{2.3}$$

$$\widehat{\theta} = \left\{1 + \frac{(S_Y^2 - S_X^2)_+}{(\overline{Y} - \overline{X})_+^2 + \epsilon_N}\right\}^{-1} \tag{2.4}$$

$$\widehat{\delta} = \frac{\widehat{\Delta}}{\widehat{\theta}} = (\overline{Y} - \overline{X})_+ \left\{1 + \frac{(S_Y^2 - S_X^2)_+}{(\overline{Y} - \overline{X})_+^2 + \epsilon_N}\right\}, \tag{2.5}$$

where $(\overline{X}, S_X^2)$ are the sample mean and variance of the control group observations, $(\overline{Y}, S_Y^2)$ are the same for the treatment group observations, and ϵ_N is a small positive number that bounds the denominators away from zero. If $\epsilon_N = o_p(1)$ as suggested in Lubich et al. (2022), then the method of moment estimators (2.4) and (2.5) are consistent. By definition of the parameter space, if either $\widehat{\theta} = 0$ or $\widehat{\delta} = 0$ then $\widehat{(\theta, \delta)} = (0, 0)$. Another nice property of the estimators is that they do not require any parametric assumptions about the distribution F.

2.2.1 Moments

Here the formulas for some moments of G are stated in terms of (F, θ, δ). These results are utilized to derive the variance of (2.4) and (2.5), which is displayed in section 2.2.3. Let $\mu_X = E[X], \sigma_X^2 = E[(X-\mu_X)^2], \mu_{3c_X} = E[(X-\mu_X)^3]$, and $\mu_{4c_X} = E[(X-\mu_X)^4]$. Similarly, let $\mu_Y = E[Y], \sigma_Y^2 = E[(Y-\mu_Y)^2], \mu_{3c_Y} = E[(Y-\mu_Y)^3]$, and $\mu_{4c_Y} = E[(Y-\mu_Y)^4]$. Let $\mathcal{F}_k$ be the set of all CDFs with finite k^{th} moment.

Proposition 2.2.1 *For $(F, \theta, \delta) \in (\mathcal{F}_4, (0,1], \mathbb{R}^+)$, the moments of $Y \sim G$ can be found in terms of (F, θ, δ) and are as follows*

$$\mu_Y = \mu_X + \theta\delta \tag{2.6}$$

$$\sigma_Y^2 = \sigma_X^2 + \theta(1-\theta)\delta^2 \tag{2.7}$$

$$\mu_{3c_Y} = \mu_{3c_X} + \theta(1-\theta)\delta^3\left[1-2\theta\right] \tag{2.8}$$

$$\mu_{4c_Y} = \mu_{4c_X} + \theta(1-\theta)\delta^4\left[(1-\theta)(1-3\theta) + \theta + 6\sigma_X^2/\delta^2\right]. \tag{2.9}$$

Equations (2.6) – (2.9) are proved in section A.1 of the Appendix. Notice that each moment of $Y \sim G$ can be written in terms of the corresponding moment of $X \sim F$ plus a term that depends on (θ, δ). Equation (2.6) implies that the average effect is given by $\Delta = \theta\delta = E[Y] - E[X]$. The even central moments of Y – (2.7) and (2.9) – can be minimized by letting δ become arbitrarily small or letting θ approach either 0 or 1 (since the additional terms are non-negative). Such cases characterize a scenario where the treatment group's response distribution approaches (a potentially shifted version of) the control group's response distribution. The difference $\mu_{3c_Y} - \mu_{3c_X}$ may be positive or negative, and is 0 when $\theta \in \{.5, 1\}$ or as θ approaches 0.

The bounding parameter ϵ_N is chosen to be of the form $\epsilon_N = S_X^2 a_N$ [where $a_N = o(1)$] so that $\widehat{\theta}$ retains its invariance to location-scale transformations of the data. To see how this $\widehat{\theta}$ maintains the location-scale invariance, consider the estimates on the data $X' = bX + c$ and $Y' = bY + c$. Recall that $S^2_{X'} = b^2 S^2_X$ and $\overline{X'} = b\overline{X} + c$ (and similarly for Y').

$$\begin{aligned}
\widehat{\theta}(X', Y') &= \left\{1 + \frac{(S^2_{Y'} - S^2_{X'})_+}{(\overline{Y'} - \overline{X'})^2_+ + S^2_{X'} a_N}\right\}^{-1} \\
&= \left\{1 + \frac{(b^2 S^2_Y - b^2 S^2_X)_+}{(b\overline{Y} + c - b\overline{X} - c)^2_+ + b^2 S^2_X a_N}\right\}^{-1} \\
&= \left\{1 + \frac{b^2 (S^2_Y - S^2_X)_+}{b^2(\overline{Y} - \overline{X})^2_+ + b^2 S^2_X a_N}\right\}^{-1} \\
&= \left\{1 + \frac{(S^2_Y - S^2_X)_+}{(\overline{Y} - \overline{X})^2_+ + S^2_X a_N}\right\}^{-1} = \widehat{\theta}(X, Y). \qquad (2.10)
\end{aligned}$$

Similarly, $\widehat{K} = \widehat{\delta}/S_X$ − the estimate of $K = \delta/\sigma_X$, which is the magnitude of the effect size for the responders relative to natural variability − is also location-scale invariant. To see this, again consider this estimate on (X', Y'). First,

$$\begin{aligned}
\widehat{\delta}(X', Y') &= (\overline{Y'} - \overline{X'})_+ \left\{1 + \frac{(S^2_{Y'} - S^2_{X'})_+}{(\overline{Y'} - \overline{X'})^2_+ + S^2_{X'} a_N}\right\} \\
&= b(\overline{Y} - \overline{X})_+ \left\{1 + \frac{(S^2_Y - S^2_X)_+}{(\overline{Y} - \overline{X})^2_+ + S^2_X a_N}\right\} \\
&= b\widehat{\delta}(X, Y),
\end{aligned}$$

which implies that

$$\begin{aligned}
\widehat{K}(X', Y') &= \widehat{\delta}(X', Y')/S_{X'} \\
&= b\widehat{\delta}(X, Y)/bS_X \\
&= \widehat{\delta}(X, Y)/S_X = \widehat{K}(X, Y). \qquad (2.11)
\end{aligned}$$

2.2.2 Simulation for Tuning Parameter a_N

This section shows the results of a simulation study for the tuning parameter a_N and recommends a simple function of N that produces near-optimal results for the simulation settings. Recall $\epsilon_N = S_X^2 a_N$ for invariance properties, so a_N uniquely defines the bounding parameter ϵ_N. It is desirable to approximate an optimal a_N with a simple closed-form solution when using the estimators (2.4) and (2.5). As such, the factorial design of the simulation is intended to cover a broadly applicable set of sample sizes and region of the parameter space.

The sample size settings under consideration are $N \in \mathcal{N} = \{60, 120, 240, 480, 960, 1920, 3840, 7680, 15360\}$ such that $m = n = N/2$. For each sample size, 1000 data sets are generated for the $6 \times 3 \times 4 = 72$ combinations with parameter values

- $F \in \{$Normal, Laplace, Skewed Right Normal (SkRNorm), Skewed Right Laplace (SkRLap), Skewed Left Normal (SkLNorm), Skewed Left Laplace (SkLLap)$\}$
  - All distributions are from the 5 parameter skewed generalized T distribution with $\lambda = 0$ for symmetric distributions, $\lambda = .5$ for right skewed distributions and $\lambda = -.5$ for left skewed distributions. Distributions from the generalized Normal family have parameters $p = 2$ and $q = \infty$, while those from the generalized Laplace family have parameters $p = 1$ and $q = \infty$.
  - $\mu_X = 0$ and $\sigma_X = 1$ for all F.
- $\theta \in \{.2, .5, .8\}$
- $\delta \in \{.5, 1, 2, 3\}$

where the choices of F correspond to those described in detail in section A.4 of the Appendix. Henceforth all F in simulations are standardized to have $\mu_X = 0$ and $\sigma_X = 1$ unless otherwise stated.

With these 72 combinations of (F, θ, δ) in hand, consider first for a fixed N, a criterion to decide what choice of a_N produces the best overall performance across the parameter space. Under each of the settings above, the estimates (2.4) and (2.5) are computed with a comprehensive array of 190 choices for $a_N \in \mathcal{A}$ ranging from .0001 to 2.3. $\mathcal{A}$ is chosen to include a range wide enough to sufficiently capture the optimal a_N for each N and to ensure that the grid is dense enough around each optimal a_N to identify it with sufficient precision. See $\mathcal{A}$ below.

.0001	.001	.00125	.0015	.00175	.002	.00225	.0025	.00275	.003
.00325	.0035	.00375	.004	.00425	.0045	.00475	.005	.00525	.0055
.00575	.006	.00625	.0065	.00675	.007	.00725	.0075	.00775	.008
.00825	.0085	.00875	.009	.00925	.0095	.00975	.010	.0105	.011
.0115	.012	.0125	.013	.0135	.014	.0145	.015	.0155	.016
.0165	.017	.0175	.018	.0185	.019	.0195	.02	.021	.022
.023	.024	.025	.026	.027	.028	.029	.03	.031	.032
.033	.034	.035	.036	.037	.038	.039	.04	.041	.042
.043	.044	.045	.046	.047	.048	.049	.05	.0525	.0550
.0575	.0600	.0625	.0650	.0675	.0700	.0725	.0750	.0775	.08
.0825	.085	.0875	.09	.0925	.095	.0975	.10	.1025	.105
.1075	.11	.1125	.115	.1175	.12	.125	.13	.135	.14
.145	.15	.16	.17	.18	.19	.20	.21	.22	.23
.24	.25	.26	.27	.28	.29	.30	.32	.34	.36
.38	.40	.42	.44	.46	.48	.50	.525	.55	.575
.6	.625	.65	.675	.7	.725	.75	.775	.8	.825
.85	.875	.9	.925	.95	.975	1	1.05	1.1	1.15
1.2	1.25	1.3	1.35	1.4	1.45	1.5	1.55	1.6	1.65
1.7	1.75	1.8	1.85	1.9	1.95	2.0	2.1	2.2	2.3

Table 2.1: $\mathcal{A}$

Determine a winning a_N for each N by minimizing the following summary comparative performance metric.

$$\begin{aligned} M(a_N) &= S(\widehat{\theta}_{a_N})S(\widehat{\delta}_{a_N}) \\ &= \prod_{F,\theta,\delta}^{72} \left\{ \sqrt{\frac{MSE(\widehat{\theta}_{a_N})}{\min\limits_{a_N \in \mathcal{A}} MSE(\widehat{\theta}_{a_N})}} \right\}^{1/72} \prod_{F,\theta,\delta}^{72} \left\{ \sqrt{\frac{MSE(\widehat{\delta}_{a_N})}{\min\limits_{a_N \in \mathcal{A}} MSE(\widehat{\delta}_{a_N})}} \right\}^{1/72} . \end{aligned} \tag{2.12}$$

(The performance metric (2.12) does not include $\widehat{\Delta}$ because the method of moment estimate of $\widehat{\Delta}$ is invariant to ϵ_N, and thus a_N.) This process is repeated for each $N \in \mathcal{N}$ with $M(a_N)$ computed for all $a_N \in \mathcal{A}$. For each $N \in \mathcal{N}$, the chosen winner is $a_N^* = \arg\min\limits_{a_N} M(a_N)$. **Figure 2.4** below displays each a_N^* for $N \in \mathcal{N}$ found via simulation as a blue dot along the curve of the $M(a_N)$.

The set of a_N^* for the nine $N \in \mathcal{N}$ indicate that the optimal scaling factor is of the form

$$a_N = \frac{k_1}{N^{k_2}}. \tag{2.13}$$

k_1 and k_2 for the curve fit are chosen to minimize

$$\prod_{N \in \mathcal{N}} \frac{M(k_1/N^{k_2})}{M(a_N^*)} \tag{2.14}$$

over a grid of $k_1 \in \{5, 10, 15, ..., 100\}$ and $k_2 \in \{.05, .10, ..., 2.0\}$. These grid points are chosen to be wide enough to capture the optimal (k_1, k_2). ($M(k_1/N^{k_2})$ for each $N \in \mathcal{N}$ is found by interpolation – see vertical line segments in **Figure 2.4**.) The resulting optimum occurs at $(k_1, k_2) = (20, .95)$. **Figure 2.5** shows the good fit of $a_N = 20/N^{.95}$ to the simulated optima.

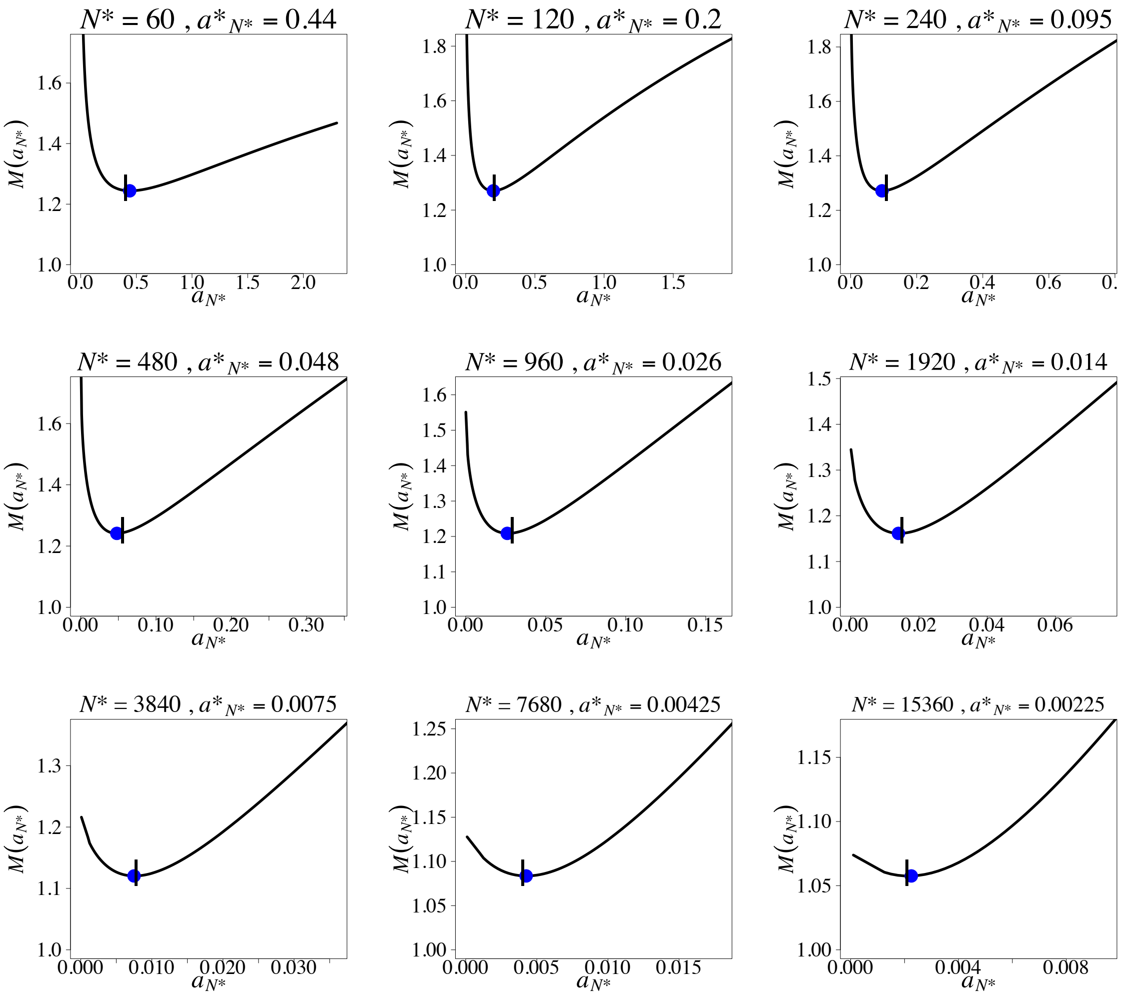


Figure 2.4: Loss for each $N \in \mathcal{N}$ with minimizers, a_N^*, indicated by the blue dots.

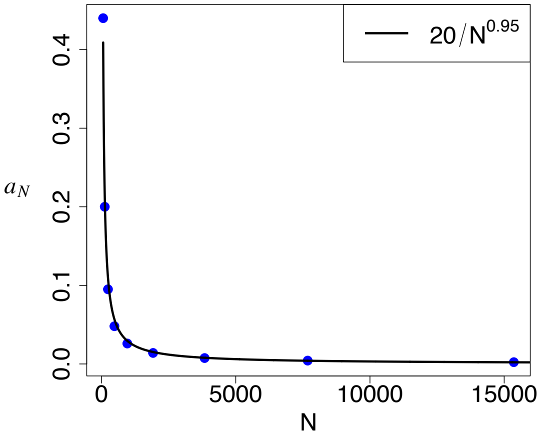


Figure 2.5: Curve fit for a_N

2.2.3 Consistency and Asymptotic Normality

This section states propositions that $\widehat{\theta}$ and $\widehat{\delta}$ in (2.4) and (2.5) are consistent and asymptotically normal estimators of θ and δ, respectively.

Proposition 2.2.2 (Consistency of Moment Estimator) *For any* $(F, \theta, \delta) \in (\mathcal{F}_2, (0,1], \mathbb{R}^+)$, $\widehat{\theta} \xrightarrow{p} \theta$ *and* $\widehat{\delta} \xrightarrow{p} \delta$.

Proposition 2.2.2 is proved in section A.2 of the Appendix. **Proposition 2.2.3** is proved in section A.3 of the Appendix and is stated here for the case of $m = n$.

Proposition 2.2.3 (Normality of Moment Estimator) *For any* $(F, \theta, \delta) \in (\mathcal{F}_4, (0,1), \mathbb{R}^+)$

$$\sqrt{n}(\widehat{\theta} - \theta) \to N(0, \sigma^2_\theta),$$
$$\sqrt{n}(\widehat{\delta} - \delta) \to N(0, \sigma^2_\delta),$$

where

$$\sigma^2_\theta = \left(1 + \frac{\sigma^2_Y - \sigma^2_X}{(\mu_Y - \mu_X)^2}\right)^{-4} \left\{ \frac{4(\sigma^2_Y - \sigma^2_X)^2}{(\mu_Y - \mu_X)^6} \left(\sigma^2_X + \sigma^2_Y\right) - \frac{4(\sigma^2_Y - \sigma^2_X)}{(\mu_Y - \mu_X)^5} \left(\mu_{3c_X} + \mu_{3c_Y}\right) \right.$$
$$\left. + \frac{(\mu_{4c_X} - \sigma^4_X) + (\mu_{4c_Y} - \sigma^4_Y)}{(\mu_Y - \mu_X)^4} \right\}, \qquad (2.15)$$

$$\sigma^2_\delta = \left(1 - \frac{\sigma^2_Y - \sigma^2_X}{(\mu_Y - \mu_X)^2}\right)^2 \left(\sigma^2_X + \sigma^2_Y\right) + 2\left(1 - \frac{\sigma^2_Y - \sigma^2_X}{(\mu_Y - \mu_X)^2}\right)\left(\frac{\mu_{3c_X} + \mu_{3c_Y}}{\mu_Y - \mu_X}\right)$$
$$+ \frac{(\mu_{4c_X} - \sigma^4_X) + (\mu_{4c_Y} - \sigma^4_Y)}{(\mu_Y - \mu_X)^2}. \qquad (2.16)$$

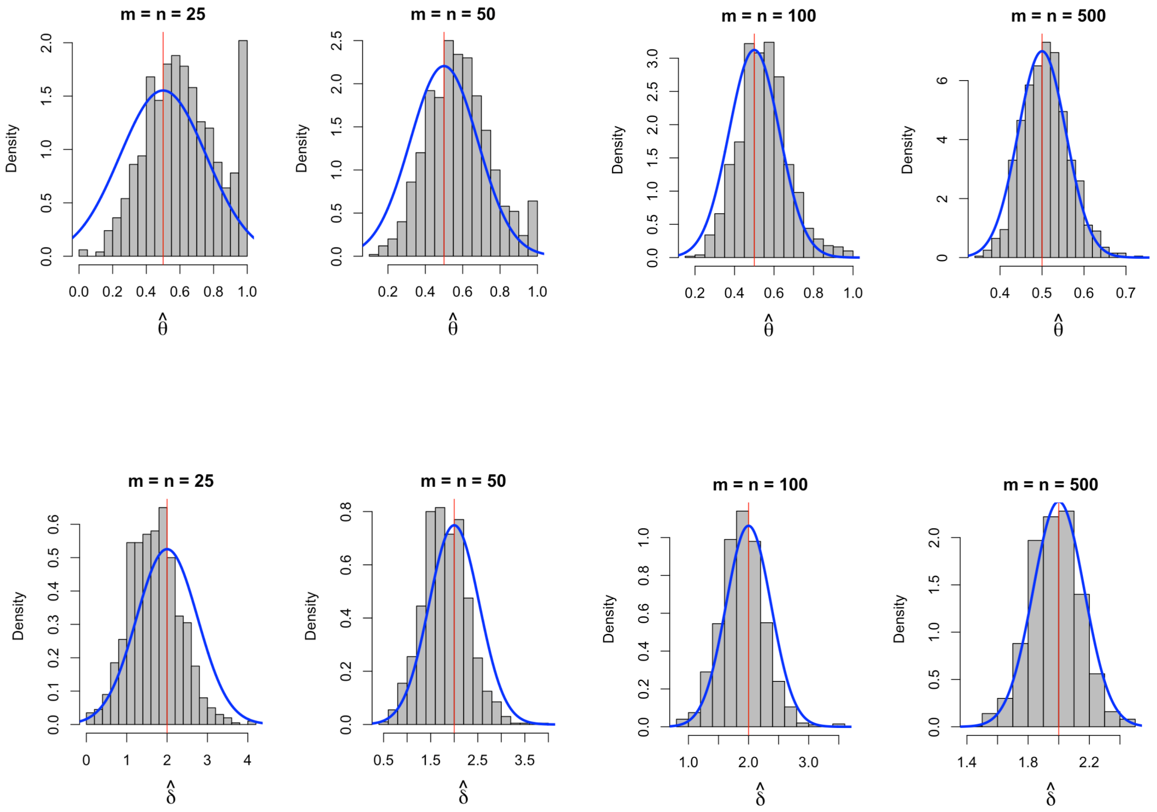


Figure 2.6: Illustration of asymptotic normality for distributions of $\widehat{\theta}$ and $\widehat{\delta}$ for $F \sim$ Laplace, $\sigma_X = 1$, $\theta = .5$, $\delta = 2$ and sample sizes $m = n \in \{25, 50, 100, 500\}$. Blue curve represents approximate distribution based on Proposition 2.2.3.

Figure 2.6 illustrates how fast $\widehat{\theta}$ and $\widehat{\delta}$ converge in distribution to normal. For the selected parameter settings, there is lack of normality due to the bounding of $\widehat{\theta} \leq 1$ in the top left plot of the figure, which subsides as the sample sizes increase. Also note the elimination of the positive bias in $\widehat{\theta}$ and negative bias in $\widehat{\delta}$ as the sample sizes increase.

2.3 Semi-parametric EM-like Algorithm

A generalization of the Normal EM algorithm was proposed by Bordes et al. (2006) for estimating the mixing proportions and mean parameters of location-shifted mixtures that does not assume that the common component distribution is normally distributed, but operates only on treatment data. They showed that this EM-like algorithm produces comparable results to Normal EM when F is Normal and superior results when F is far from normal so long as the mixing proportions are moderate and the components are well separated. This algorithm performs poorly when the components are not well separated as shown in section 1.3.3. Here the algorithm is adapted to incorporate information from the control data that allows for improved performance when the components are not well separated.

This EM-like algorithm has 6 total variations to consider based upon three different inputs variables to the function ssSpEMloc (which can be found in section A.10 of the Appendix) as shown in **Table 2.2**.

EM-like Algorithm Variations			
Version	all.data.f	stochastic	symmetric
1	TRUE	TRUE	TRUE
2	TRUE	TRUE	FALSE
3	TRUE	FALSE	TRUE
4	TRUE	FALSE	FALSE
5	FALSE	FALSE	TRUE
6	FALSE	FALSE	FALSE

Table 2.2: Table of Semi-Parametric EM-like Algorithm Settings

The first option in the algorithm is ‘all.data.f’. When this option is TRUE, $(\mathbf{X}, \mathbf{Y})$ is used to estimate f at each step; whereas when it is FALSE, only $\mathbf{X}$ is used to estimate f. When 'all.data.f = TRUE', there is an option for how to incorporate the weights (component membership probabilities p_i, for treatment observations Y_i) into the estimation of f at each iteration via the 'stochastic' argument. If 'stochastic = TRUE' then each p_i is used to simulate whether the corresponding Y_i came from the non-responder component or the responder component. If 'stochastic = FALSE', then each p_i is used to provide a weighted assignment of each Y_i to the two components when updating $\widehat{f}$. Finally, if 'symmetric = FALSE' a regular kernel density estimate is used to estimate f, while in the 'symmetric = TRUE' case a symmetrization step is added to make this estimate (denoted $\widehat{f_s}$) symmetric about the mean of $\widehat{f}$ (denoted $\widehat{\mu}_0$).

To initialize, begin by using k-means (Hartigan and Wong, 1979) clustering on the treatment data to cluster the data into two groups. The Y_i values assigned $p_i^{(0)} = 1$ are in the group with larger mean, while the Y_i values assigned $p_i^{(0)} = 0$ are in the group with smaller mean. This full group membership assignment is used for the initialization. Let $\delta^{(k)} = \mu_1^{(k)} - \mu_0^{(k)}$. Let $K(\cdot)$ represent the standard normal kernel (used for kernel density estimation) and let h represent the chosen bandwidth. Thereafter a case-by-case description of the iterations between E and 'M' steps is presented below. (The 'M' step does not truly maximize the complete log-likelihood which is why it is surrounded by quotation marks. The updating equations for μ_0 and μ_1 maximize the log-likelihood when $F \sim$ Normal, but not in general. Furthermore, the updating equations for f are based on a kernel density estimate which also does not maximize the log-likelihood function.) By definition of the parameter space, if either $\widehat{\theta} = 0$ or $\widehat{\delta} = 0$ then $\widehat{(\theta, \delta)} = (0, 0)$ for all versions.

1.) all.data.f = TRUE, stochastic = TRUE, symmetric = TRUE

1. E-step: $p_i^{(k)} = \dfrac{\theta^{(k-1)} f_s^{(k-1)}(y_i - \delta^{(k-1)})}{(1-\theta^{(k-1)}) f_s^{(k-1)}(y_i) \ + \ \theta^{(k-1)} f_s^{(k-1)}(y_i - \delta^{(k-1)})}$
2. 'M'-step:
   (a) $\theta^{(k)} = \dfrac{\sum_{i=1}^n p_i^{(k)}}{n}$
   (b) $\mu_0^{(k)} = \dfrac{\sum_{j=1}^m x_j \ + \ \sum_{i=1}^n (1-p_i^{(k)}) y_i}{m + \sum_{i=1}^n \, (1-p_i^{(k)})}$
   (c) $\mu_1^{(k)} = \dfrac{\sum_{i=1}^n p_i^{(k)} y_i}{\sum_{i=1}^n p_i^{(k)}}$
   (d) $f_s^{(k)}(u) =$ Steps i. – iv. below
      i. Simulate $\tilde{z}_i^{(k)} \ \sim$ Bernoulli$(p_i^{(k)})$
      ii. $\tilde{y}_i^{(k)} = y_i - \tilde{z}_i^{(k)} \delta^{(k)}$
      iii. $f^{(k)}(u) = \dfrac{\sum_{j=1}^m K\left((u - x_j)/h\right) \ + \ \sum_{i=1}^n K\left((u - \tilde{y}_i^{(k)})/h\right)}{(m+n)h}$
      iv. $f_s^{(k)}(u) = \dfrac{f^{(k)}(u) \ + \ f^{(k)}(2\mu_0^{(k)} - u)}{2}$

The quantities $p_i^{(k)}$, $\theta^{(k)}$, $\mu_0^{(k)}$, $\mu_1^{(k)}$, $\delta^{(k)}$ are updated in the same manner as the Normal EM algorithm, while f_s is found by using simulated 're-centered' data. If all $\theta^{(k)} = \theta, \mu_0^{(k)} = \mu_0$, and $\mu_1^{(k)} = \mu_1$, then $\tilde{y}_i^{(k)} \overset{iid}{\sim} f(u)$. The estimates for the parameters (θ, μ_0, μ_1) $[\delta = \mu_1 - \mu_0]$ are computed by taking the average of $(\theta^{(k)}, \mu_0^{(k)}, \mu_1^{(k)})$ over all the iterations. The estimate of f is found by using the last iteration, $\widehat{f}(u) = f_s^{\text{'maxiter'}}(u)$ where the number of iterations is pre-specified by the 'maxiter' argument.

2.) all.data.f = TRUE, stochastic = TRUE, symmetric = FALSE

Version 2 of the algorithm is identical to Version 1 with the exception that $f^{(k-1)}$ is used in the E-step (instead of $f_s^{(k-1)}$) and thus the symmetrization step `iv.` is not necessary.

3.) all.data.f = TRUE, stochastic = FALSE, symmetric = TRUE

In Version 3, the p_is are not used to simulate complete data but rather are used directly to provide a weighted assignment of each Y_i into the 'non-responder' and 'responder' components.

1. E-step: $p_i^{(k)} = \dfrac{\theta^{(k-1)} f_s^{(k-1)}(y_i - \delta^{(k-1)})}{(1-\theta^{(k-1)}) f_s^{(k-1)}(y_i) \ + \ \theta^{(k-1)} f_s^{(k-1)}(y_i - \delta^{(k-1)})}$
2. 'M'-step:
   (a) $\theta^{(k)} = \dfrac{\sum_{i=1}^n p_i^{(k)}}{n}$

   (b) $\mu_0^{(k)} = \dfrac{\sum_{j=1}^m x_j \ + \ \sum_{i=1}^n (1-p_i^{(k)}) y_i}{m + \sum_{i=1}^n \ (1-p_i^{(k)})}$

   (c) $\mu_1^{(k)} = \dfrac{\sum_{i=1}^n p_i^{(k)} y_i}{\sum_{i=1}^n p_i^{(k)}}$

   (d) $$f^{(k)}(u) = \frac{1}{(m+n)h} \left\{ \sum_{j=1}^m K\left(\frac{u-x_j}{h}\right) + \sum_{i=1}^n (1-p_i) K\left(\frac{u-y_i}{h}\right) + p_i K\left(\frac{u-(y_i-\delta^{(k)})}{h}\right) \right\}$$

   (e) $f_s^{(k)}(u) = \dfrac{f^{(k)}(u) \ + \ f^{(k)}(2\mu_0^{(k)} - u)}{2}$

The estimates for the parameters $(\theta, \mu_0, \mu_1, f)[\delta = \mu_1 - \mu_0]$ are the final iteration of the corresponding values. The convergence criterion is when $|\theta^{(k)} - \theta^{(k-1)}|$ and $|\mu_0^{(k)} - \mu_0^{(k-1)}|$ and $|\mu_1^{(k)} - \mu_1^{(k-1)}|$ are all less than a pre-specified $\epsilon > 0$.

4.) all.data.f = TRUE, stochastic = FALSE, symmetric = FALSE

Version 4 of the algorithm is identical to Version 3 except that the updating equation for $p_i^{(k)}$ uses $f^{(k-1)}$ instead of $f_s^{(k-1)}$ and thus the symmetrization step in (e) is not necessary.

5.) all.data.f = FALSE, stochastic = FALSE, symmetric = TRUE

Version 5 does not use all of the data to estimate f but only observations of known component origin, $\boldsymbol{X} \sim F$. The motivation for only using control data is that this 'pure' data may provide stability to the algorithm by preventing a dramatic change the estimate of f over the iterations due to inappropriate influence of the Y_is. Therefore, f only needs to be estimated once from the control data.

- $\widehat{\mu}_0 = \dfrac{\sum_{j=1}^{m} x_j}{m}$
- $\widehat{f}(u) = \sum_{j=1}^{m} K\left(\dfrac{u - x_j}{h}\right)$
- $\widehat{f}_s(u) = \dfrac{\widehat{f}(u) + \widehat{f}(2\widehat{\mu}_0 - u)}{2}$

1. E-step: $p_i^{(k)} = \dfrac{\theta^{(k-1)}\widehat{f}_s(y_i - \delta^{(k-1)})}{(1-\theta^{(k-1)})\widehat{f}_s(y_i) \; + \; \theta^{(k-1)}\widehat{f}_s(y_i - \delta^{(k-1)})}$
2. M-step:
   (a) $\theta^{(k)} = \dfrac{\sum_{i=1}^{n} p_i^{(k)}}{n}$
   (b) $\mu_1^{(k)} = \dfrac{\sum_{i=1}^{n} p_i^{(k)} y_i}{\sum_{i=1}^{n} p_i^{(k)}}$

The estimates for the parameters $(\theta, \mu_1, f, \delta)$ are the final iteration of the corresponding values. The convergence criterion is satisfied when $|\theta^{(k)} - \theta^{(k-1)}|$ and $|\mu_0^{(k)} - \mu_0^{(k-1)}|$ and $|\mu_1^{(k)} - \mu_1^{(k-1)}|$ are all less than a pre-specified $\epsilon > 0$.

6.) all.data.f = FALSE, stochastic = FALSE, symmetric = FALSE

Version 6 is the same as Version 5 except $\widehat{f}(u)$ is used in the E-step instead of $\widehat{f}_s(u)$ (and thus $\widehat{f}_s(u)$ need not be computed).

2.3.1 Illustration of EM-like Algorithm Versions

The plots in **Figures 2.7 - 2.12** below describe how the algorithm proceeds for each of the 6 versions on the same data set generated from $m = 100, n = 100, F \sim N(0,1), \theta = .7, \delta = 2$. Each page includes a 2×3 grid of plots of the same variety. The top left plot shows the initial KDE of the mixture (following the k-means initialization). The top center plot shows the final KDE of $\widehat{f}$. The top right plot shows the final KDE of the mixture $\widehat{g}$ using the estimates of $(\widehat{f},\ \widehat{\theta},\ \widehat{\delta})$. The bottom left plot shows the iterations of $\theta^{(k)}$. The bottom center plot shows the iterations of $\mu_0^{(k)}$ (in black) and $\mu_1^{(k)}$ (in red). The bottom right plot shows the iterations of $\delta^{(k)}$.

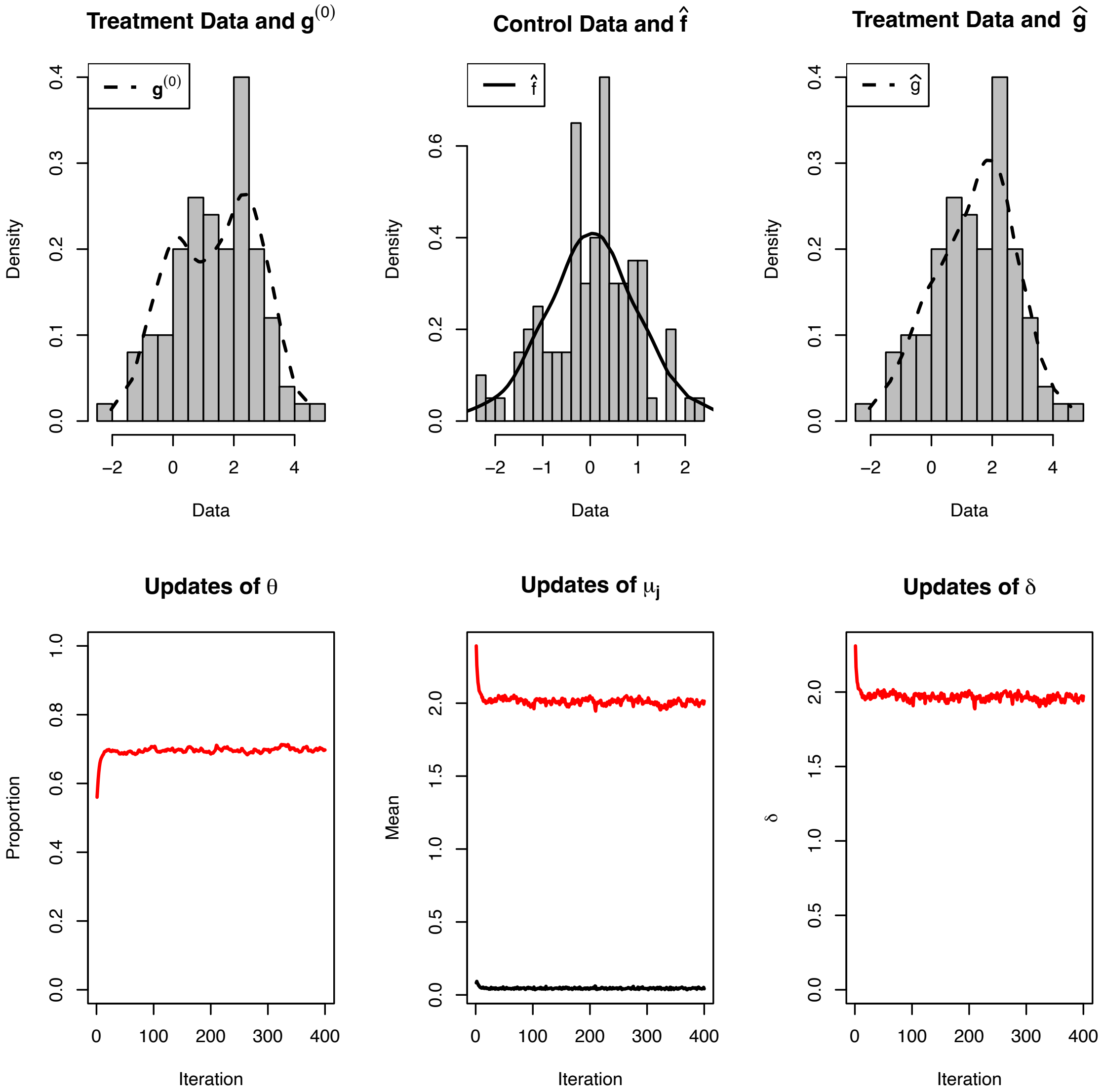


Figure 2.7: EM-like Version 1 on a data set generated from $m = 100, n = 100, F \sim N(0, 1), \theta = .7, \delta = 2$.

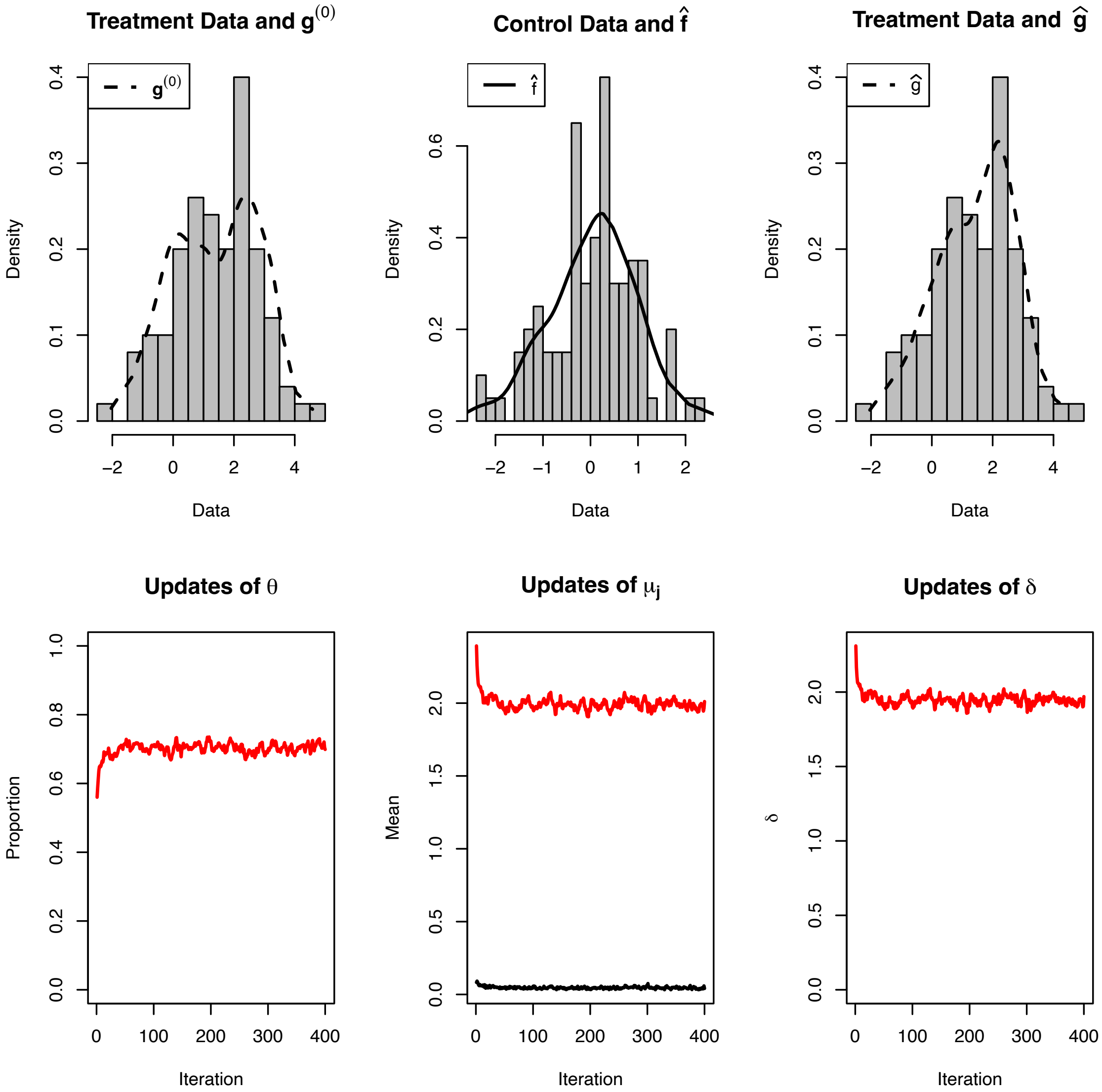


Figure 2.8: EM-like Version 2 on a data set generated from $m = 100, n = 100, F \sim N(0,1), \theta = .7, \delta = 2$.

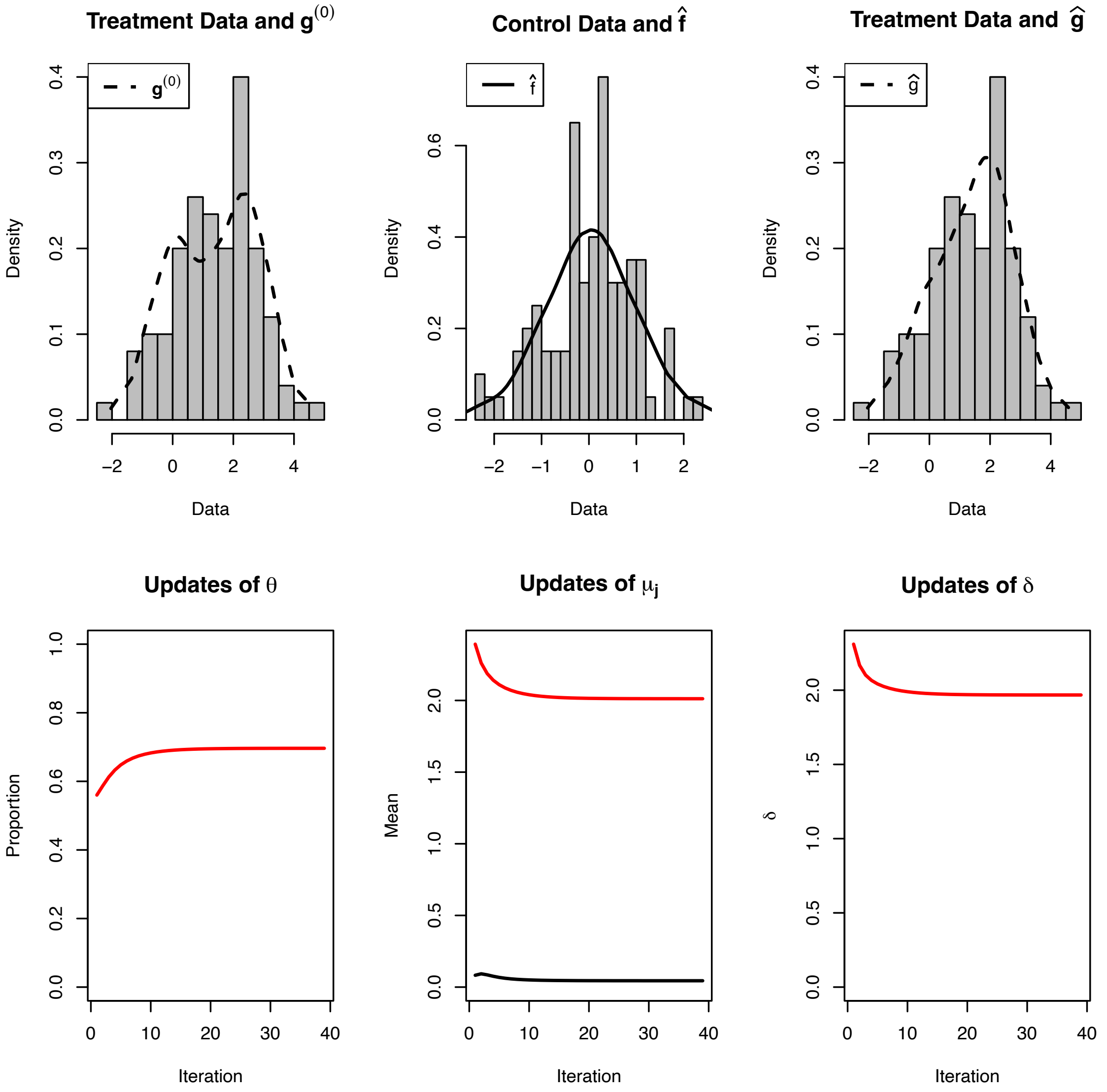


Figure 2.9: EM-like Version 3 on a data set generated from $m = 100, n = 100, F \sim N(0, 1), \theta = .7, \delta = 2$.

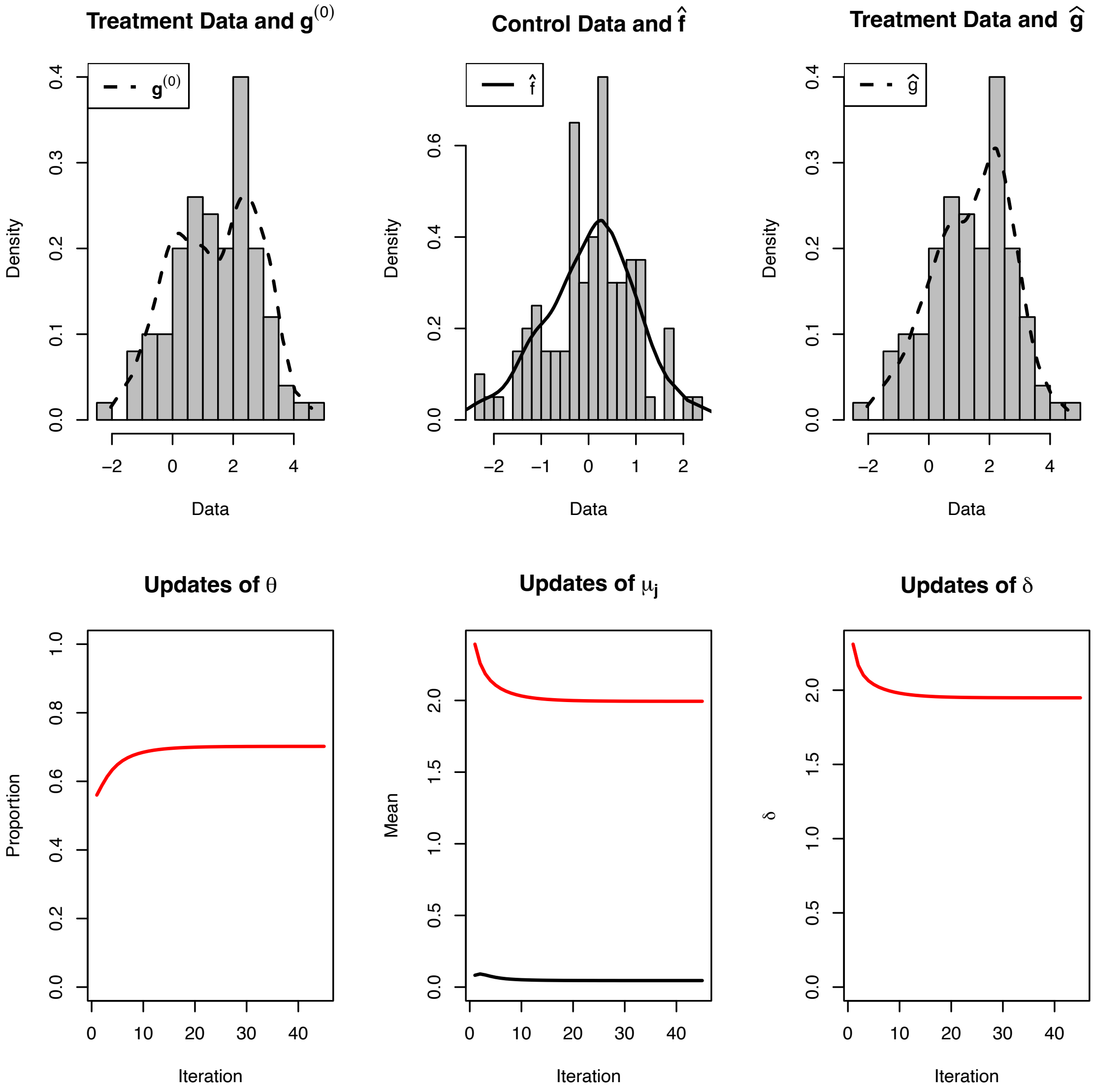


Figure 2.10: EM-like Version 4 on a data set generated from $m = 100, n = 100, F \sim N(0, 1), \theta = .7, \delta = 2$.

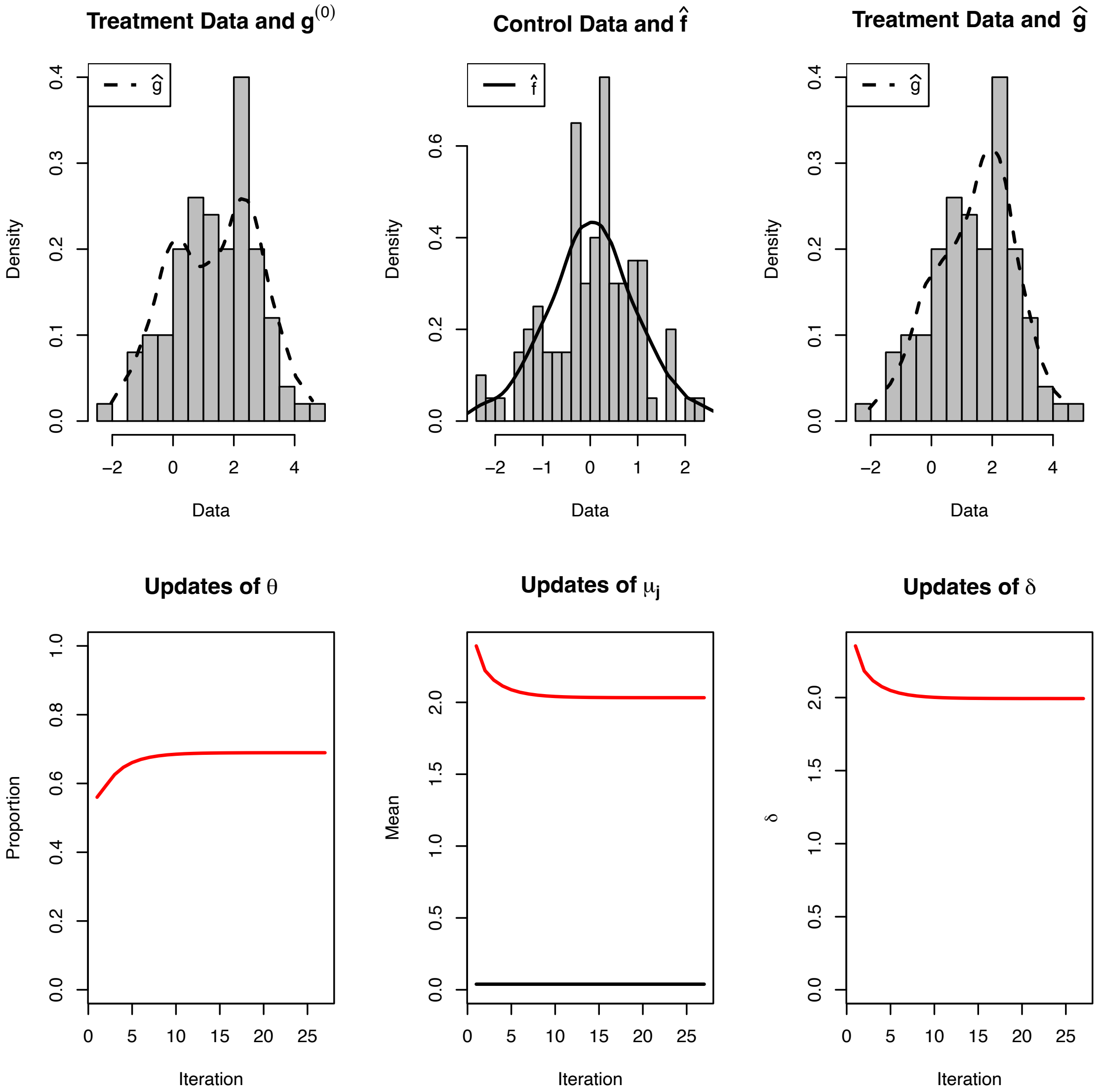


Figure 2.11: EM-like Version 5 on a data set generated from $m = 100, n = 100, F \sim N(0, 1), \theta = .7, \delta = 2$.

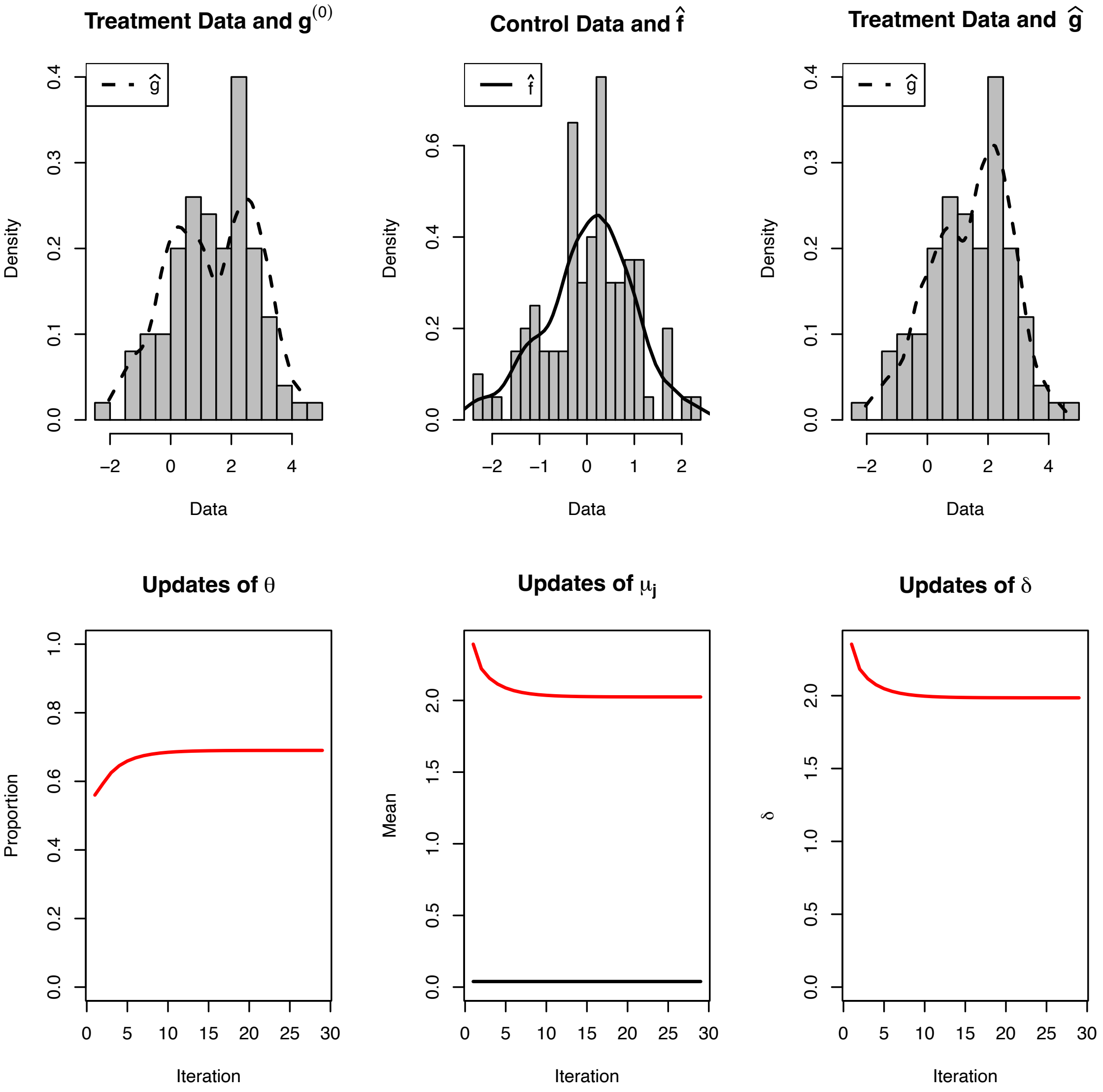


Figure 2.12: EM-like Version 6 on a data set generated from $m = 100, n = 100, F \sim N(0, 1), \theta = .7, \delta = 2$.

Figure 2.13 shows a plot of all the estimates of $\widehat{(\theta,\delta)}_{EMlike}$ for the six versions of the EM-like algorithm on the same data set. The red bulls-eye symbol represents the true (θ,δ). **Figure 2.13** illustrates that the different versions of the algorithm provide distinct but similar estimates of the treatment effect that are also similar to those found by methods that assume (correctly in this case) that $F \sim$ Normal.

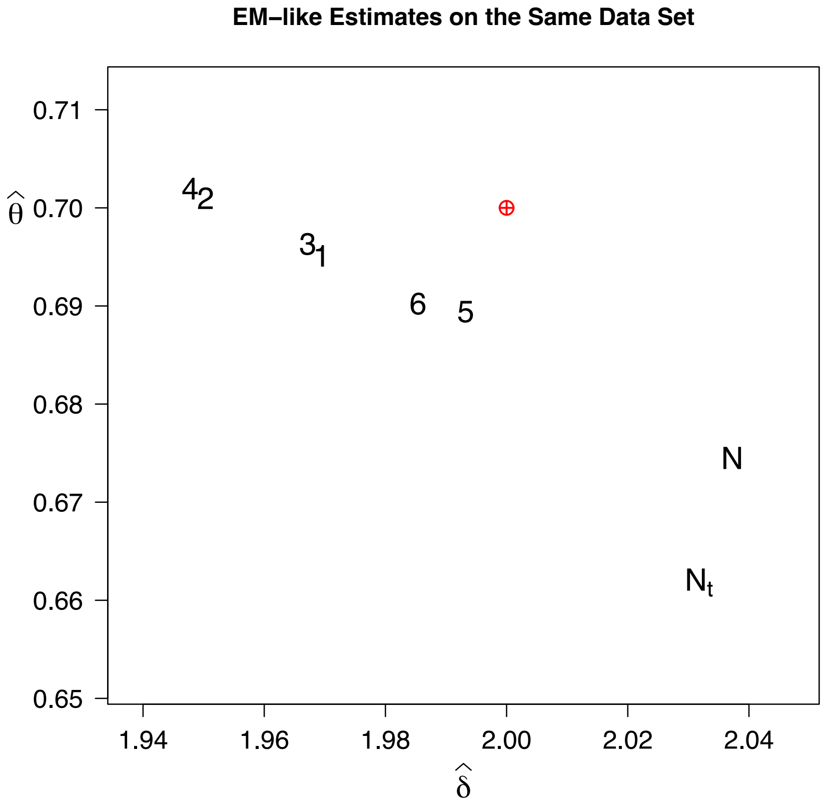


Figure 2.13: Point Estimates for all 6 versions of EM-like Algorithm on a data set generated from $m = 100, n = 100, F \sim N(0,1), \theta = .7, \delta = 2$. Numbers on the plot represent the estimates of the corresponding EM-like versions. N represents the estimate using the EM algorithm with a Normality assumption as described in section 2.1. N_t represents the EM algorithm with a Normality assumption using only treatment data as described in section 1.2.8.

Recall from **Figure 1.5** (copied in the figure below for comparison) that point estimation using this EM-like algorithm under this parameter setting (with $F \sim N(0,1)$ and $(\theta, \delta) = (.7, 2)$) is unsatisfactory without control data. **Figure 2.14** below shows the improvement in distribution-robust inference on (θ, δ) by including $m = 100$ control observations into the algorithm.

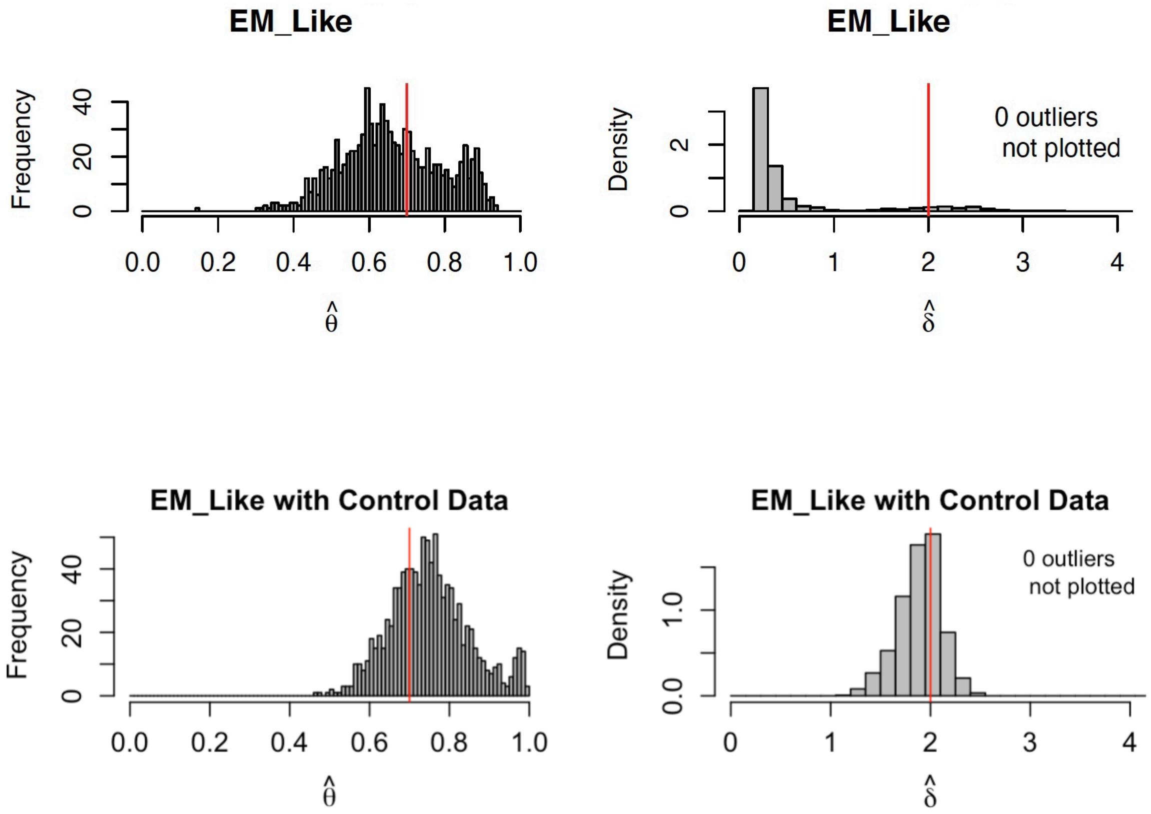


Figure 2.14: EM-like algorithms on 1000 data sets generated from $m = 100, n = 100, F \sim N(0,1), \theta = .7, \delta = 2$. The estimators displayed on the top plots do not use the $m = 100$ control observations, while the estimators on the bottom plots do. Both algorithms use Version 1.

2.3.2 Simulation for Determining Preferable Versions

The concluding pages of this section compare the relative performances of the 6 versions of the EM-like algorithm by comparing results across a wide variety of simulation settings. Performance comparisons should keep in mind that obtaining $\widehat{(\theta,\delta)}_{EMlike}$ is of primary interest and obtaining $\widehat{f}$ is of secondary interest.

The joint distribution of $\widehat{(\theta,\delta)}$ characterizes how well a method estimates (θ,δ). The marginal distributions of $\widehat{\theta}$ and $\widehat{\delta}$ give vital information to the effectiveness of the estimate, but do not fully define the joint distribution. To supplement the marginal distributions, the distribution of $\widehat{\Delta} = \widehat{\theta}\,\widehat{\delta}$ provides additional information about $\widehat{(\theta,\delta)}$. The parameter $\Delta = \theta\delta$ is of particular interest because it represents the average treatment effect in (1.1). Therefore, the distributions of $\widehat{\theta}$, $\widehat{\delta}$, and $\widehat{\Delta}$ together give a comprehensive understanding of the effectiveness of $\widehat{(\theta,\delta)}$ in estimating (θ,δ).

To compare the performance of $\widehat{(\theta,\delta)}_{EMlike}$ under the 6 different variations, consider the simulation study that generates 1000 data sets under each following factorial combinations of settings in **List 2.1**.

- $m = n \in \{25, 50, 100, 500\}$
- $F \in \{$Normal, Laplace, SkRNorm, SkRLap, SkLNorm, SkLLap$\}$
- $\theta \in \{.2, .5, .8\}$
- $\delta \in \{.5, 1, 2, 3\}$

List 2.1: There are 4 sample size settings and 72 unique (F,θ,δ) triples for a total of 288 combinations of $(m = n, F, \theta, \delta)$. The choices of F correspond to those described in section A.4 of the Appendix.

Table 2.3 below displays scores for $\widehat{\theta}$, $\widehat{\delta}$ and $\widehat{\Delta} = \widehat{\theta}\,\widehat{\delta}$ (let $\widehat{\tau}$ represent an estimator for a generic parameter τ). These scores for estimator i are the geometric average of $\sqrt{MSE(\widehat{\tau}_i)/\min_k MSE(\widehat{\tau}_k)}$ across all 72 combinations of (F,θ,δ) (where k indexes the candidate estimators, here $k \in \{1, ..., 6\}$)

$$S(\widehat{\tau}_i) = \prod_{F,\theta,\delta}^{72} \left\{ \sqrt{\frac{MSE(\widehat{\tau}_i)}{\min_k MSE(\widehat{\tau}_k)}} \right\}^{1/72}. \tag{2.17}$$

The score represents the average relative loss in performance of each estimator compared to an 'oracle' estimator (that chooses the optimal algorithm version given the true parameters). For example, estimator i with a score of $S(\widehat{\delta}_i) = 1.10$ has $\sqrt{MSE(\widehat{\delta}_i)}$ that is on average 10% larger than the oracle estimator. A score of $S(\widehat{\delta}_i) = 1$ means that Version i has the smallest $\sqrt{MSE(\widehat{\delta})}$ for each of the 72 simulation settings, so smaller scores (closer to 1) are preferred. **Table 2.3** also presents the geometric average of $S(\widehat{\theta}), S(\widehat{\delta})$, and $S(\widehat{\Delta})$ as a summary score for each estimation method. Each of the 4 sample sizes are shown separately in the table. **Figures 2.15 - 2.16** display the same scores in **Table 2.3** in 4 plots corresponding with the 4 columns of the table. The lowest (and near lowest) scores are highlighted in yellow.

m = n	Version	$S(\widehat{\theta})$	$S(\widehat{\delta})$	$S(\widehat{\Delta})$	$\{S(\widehat{\theta})S(\widehat{\delta})S(\widehat{\Delta})\}^{1/3}$
25	EM-like 1	1.104	1.084	1.056	1.081
	EM-like 2	1.105	1.106	1.051	1.087
	EM-like 3	1.263	1.126	1.080	1.154
	EM-like 4	1.306	1.148	1.098	1.181
	EM-like 5	1.123	1.038	1.033	1.064
	EM-like 6	1.096	1.079	1.018	1.064
50	EM-like 1	1.147	1.143	1.064	1.117
	EM-like 2	1.125	1.149	1.056	1.109
	EM-like 3	1.326	1.192	1.093	1.200
	EM-like 4	1.350	1.192	1.109	1.213
	EM-like 5	1.143	1.074	1.034	1.082
	EM-like 6	1.130	1.112	1.025	1.088
100	EM-like 1	1.224	1.234	1.083	1.178
	EM-like 2	1.132	1.193	1.053	1.125
	EM-like 3	1.396	1.254	1.114	1.249
	EM-like 4	1.346	1.209	1.105	1.216
	EM-like 5	1.161	1.117	1.035	1.103
	EM-like 6	1.157	1.145	1.029	1.109
500	EM-like 1	1.566	1.521	1.186	1.414
	EM-like 2	1.158	1.219	1.046	1.139
	EM-like 3	1.731	1.559	1.215	1.486
	EM-like 4	1.311	1.177	1.085	1.187
	EM-like 5	1.272	1.346	1.080	1.228
	EM-like 6	1.223	1.227	1.052	1.164

Table 2.3: Scores for Estimators of θ, δ, Δ.

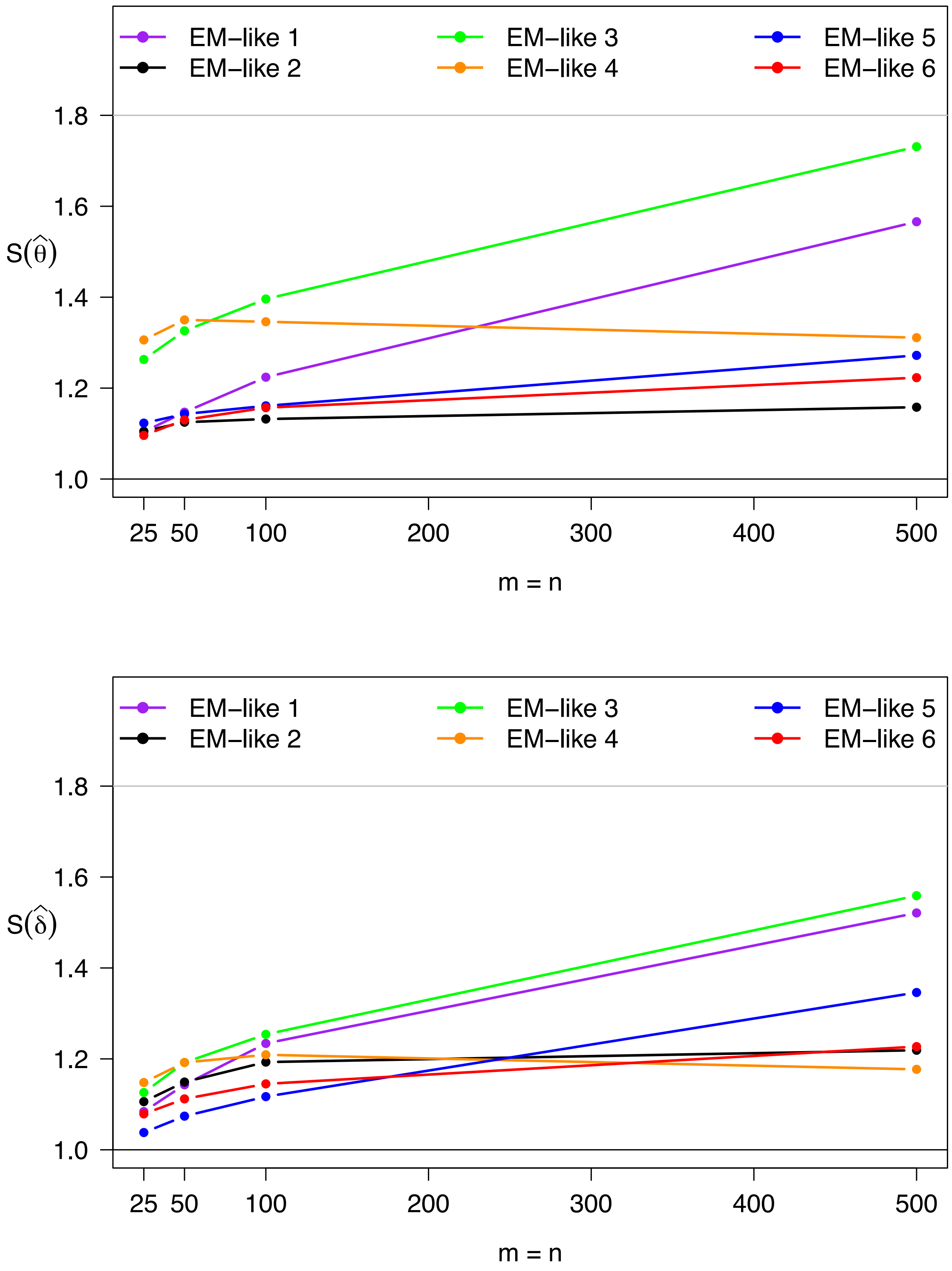


Figure 2.15: EM-like Scores for $\widehat{\theta}$ and $\widehat{\delta}$

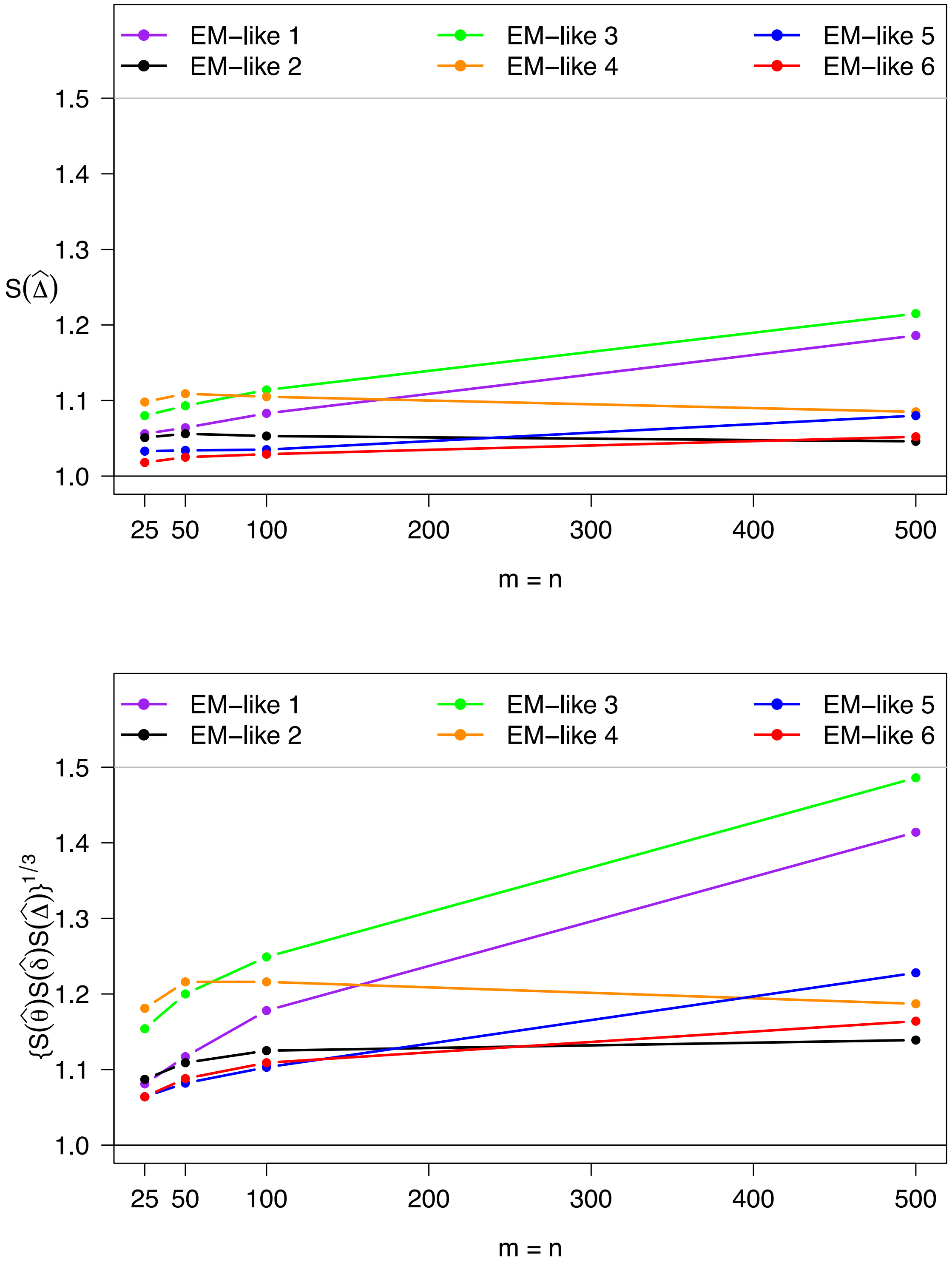


Figure 2.16: EM-like Score for $\widehat{\Delta}$ and an Overall Summary Score

A few trends from **Table 2.3** and **Figures 2.15 - 2.16** emerge. One trend is that the estimators that assume f is symmetric (1, 3, 5) perform comparably to or better than their symmetry-agnostic counterparts (2, 4, 6 respectively) for sufficiently small sample sizes, while for sufficiently large sample sizes the symmetry-agnostic versions dramatically outperform those with a symmetry assumption. The symmetry assumption allows for a decreased variability in $\widehat{f}$ (even if f is not symmetric) which is particularly beneficial for small samples leading to more stable $\widehat{(\theta, \delta)}$. However, for large sample sizes the inconsistency of $\widehat{f}$ for non-symmetric f results in substantially biased $\widehat{f}$ and thus poor estimates. A second trend is that the versions where $\widehat{f}$ is computed only based only on control data (5 and 6) perform the best overall for small to moderate sample sizes while for larger sample sizes the iterative methods are more efficient. Again, the stability of $\widehat{f}(\boldsymbol{X})$ [as opposed to $\widehat{f}(\boldsymbol{X}, \boldsymbol{Y})$] aids in smaller sample size settings. The discrepancies in performance of $\widehat{(\theta, \delta)}_{EMlike5}$ and $\widehat{(\theta, \delta)}_{EMlike6}$ for small sample sizes are minor and version 6 has the advantage of being able to capture skew in f if it exists. For larger sample sizes, the stochastic symmetry-agnostic version (2) performs the best. Note that if (θ, μ_0, μ_1) were known, only the stochastic versions return an f-distributed sample at each iteration from which $\widehat{f}$ is updated (Bordes et al., 2007) - the deterministic versions do not have this property. Since version 2 (and not 1) has a consistent $\widehat{f}$ for all continuous f, it is not surprising that version 2 displays superior performance for sufficiently large samples. While Benaglia et al. (2009) indicate that the deterministic version consistently performs slightly better than the stochastic versions for $\delta \geq 3\sigma_X$ in the case of no control data, these results show that in the presence of control data and smaller δ, the stochastic version indicates superior performance - particularly in estimating the mixing proportion. In light of these observations, a simple robust recommendation for $\widehat{(\theta, \delta)}_{EMlike}$ is

- If $m = n \leq 250$, use version 6
- If $m = n > 250$, use version 2.

Parameter Specific Performance Comparison

To understand how the relative performances of the estimators depend upon (F, θ, δ), the plots on the following pages in **Figures 2.18 - 2.21** compare the 2 recommended versions (2, 6) of the EM-like estimator. Under each of the 288 simulation settings either $\sqrt{MSE(\widehat{\theta}_i)/\min_k MSE(\widehat{\theta})}$ or $\sqrt{MSE(\widehat{\delta}_i)/\min_k MSE(\widehat{\delta}_k)}$ is plotted for each estimator $i \in \{2, 6\}$. Each plot has 72 columns of dots representing the 72 settings of (F, θ, δ). Each column has 2 dots representing the 2 versions of the algorithm. **Figure 2.17** displays the color key for the 2 recommended versions of the EM-like estimator. The best performing estimator under each column's setting has a value of 1 while higher scores represent a relative loss in performance. For example, a dot at 2 represents an estimator that has twice the $\sqrt{MSE}$ as the best estimator under that simulation setting. There are four separate plots for each of the four sample size settings for $\widehat{\theta}$ and another four plots for $\widehat{\delta}$ for a total of eight plots.

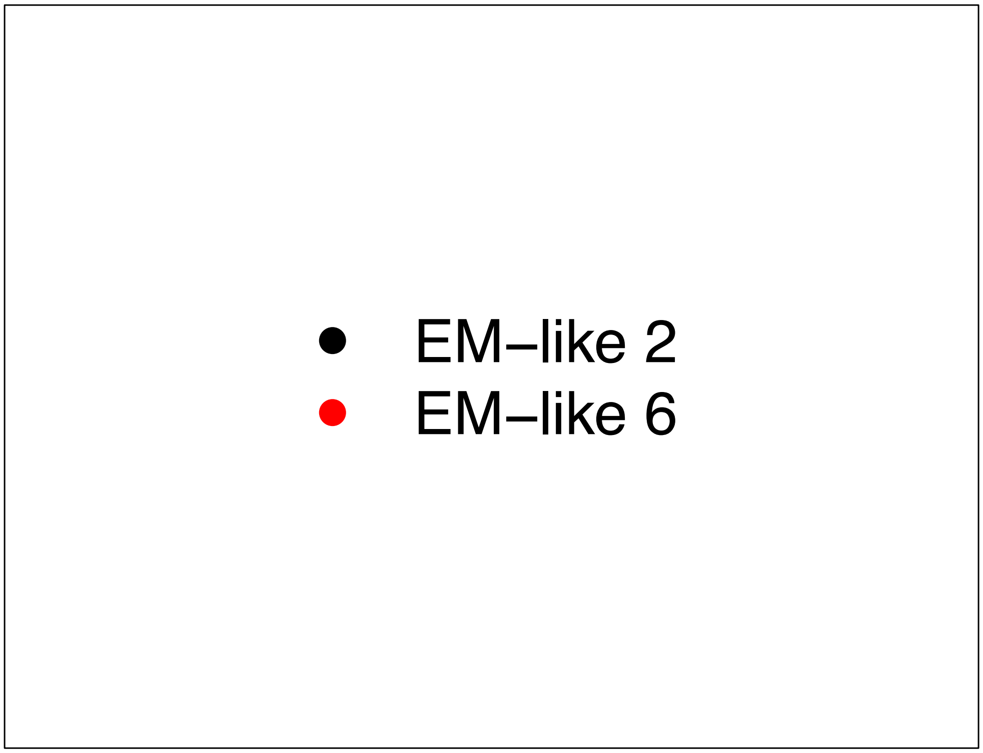


Figure 2.17: Color Key for EM-like Algorithm Versions

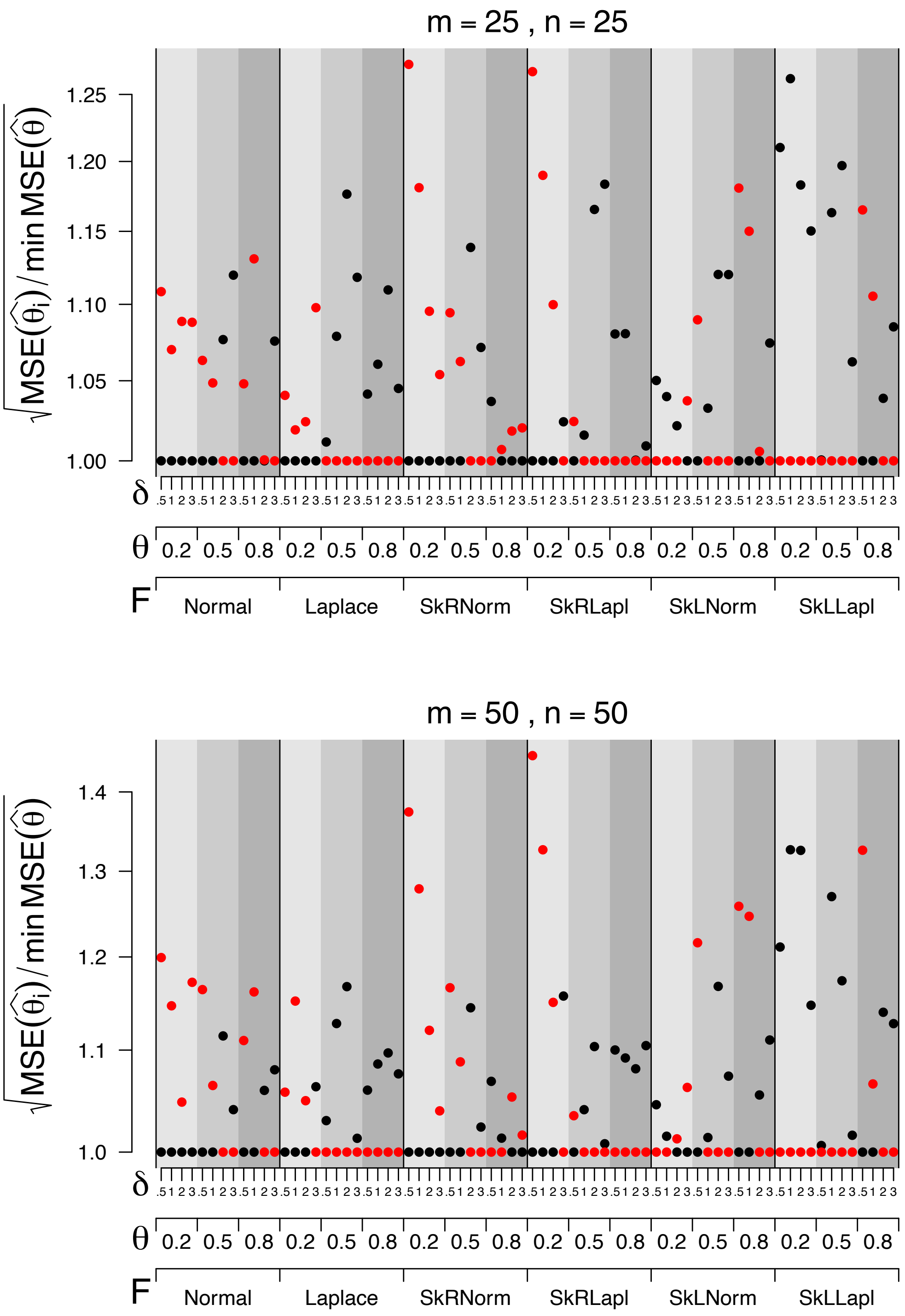


Figure 2.18: Dot Plots comparing the performances of version 2 and version 6 of the EM-like estimator of θ for small sample sizes.

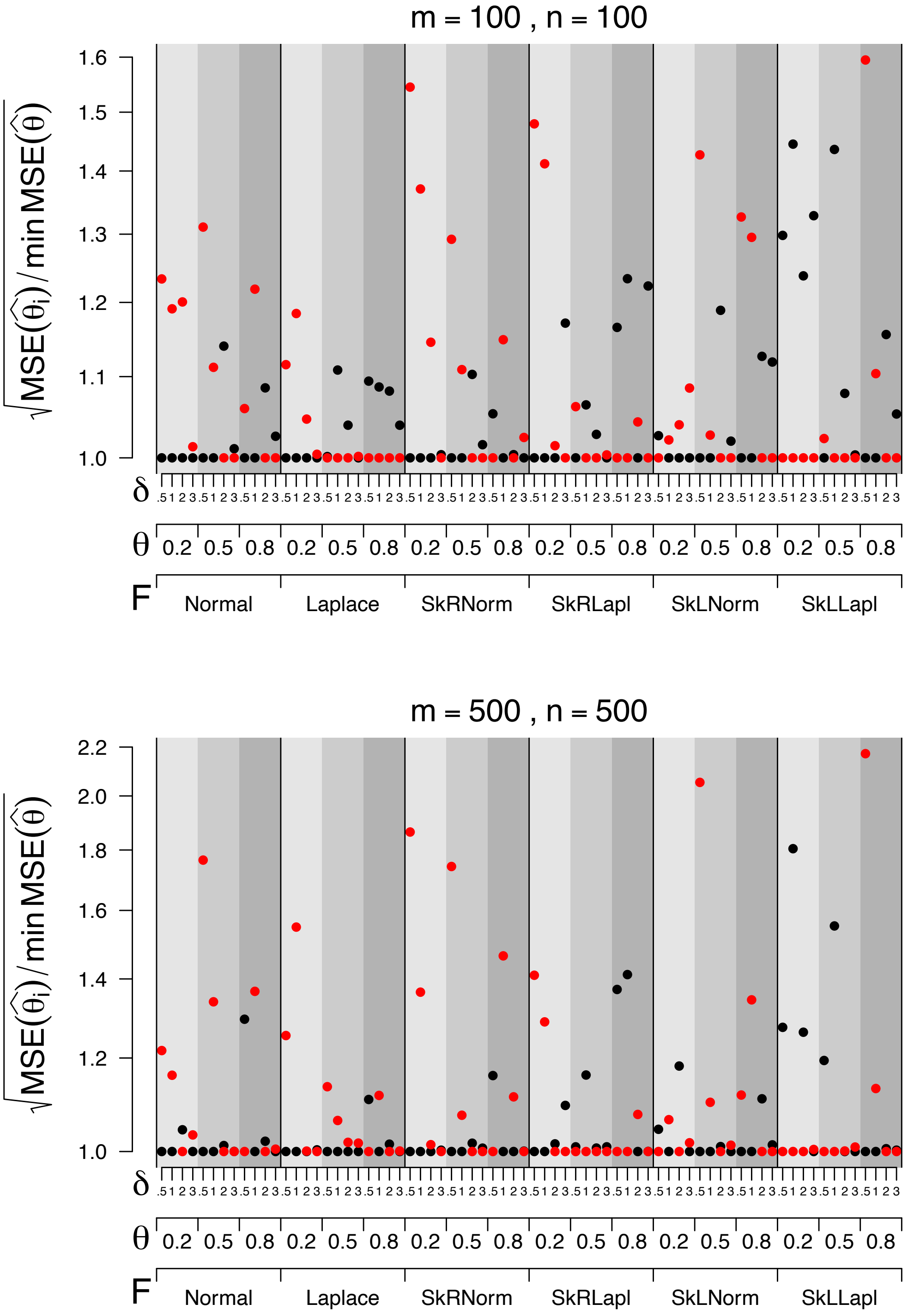


Figure 2.19: Dot Plots comparing the performances of version 2 and version 6 of the EM-like estimator of θ for moderate to large sample sizes.

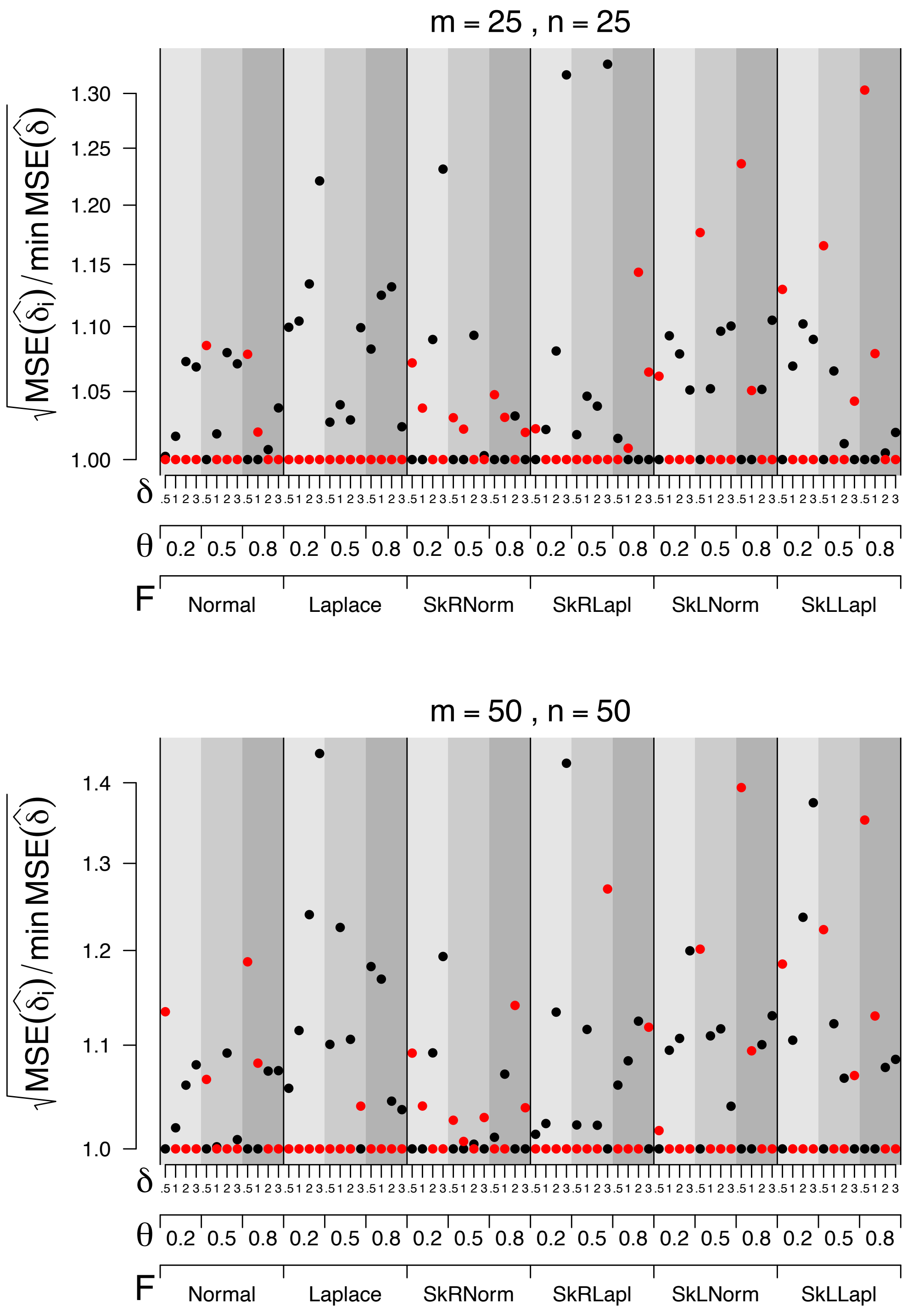


Figure 2.20: Dot Plots comparing the performances of version 2 and version 6 of the EM-like estimator of δ for small sample sizes.

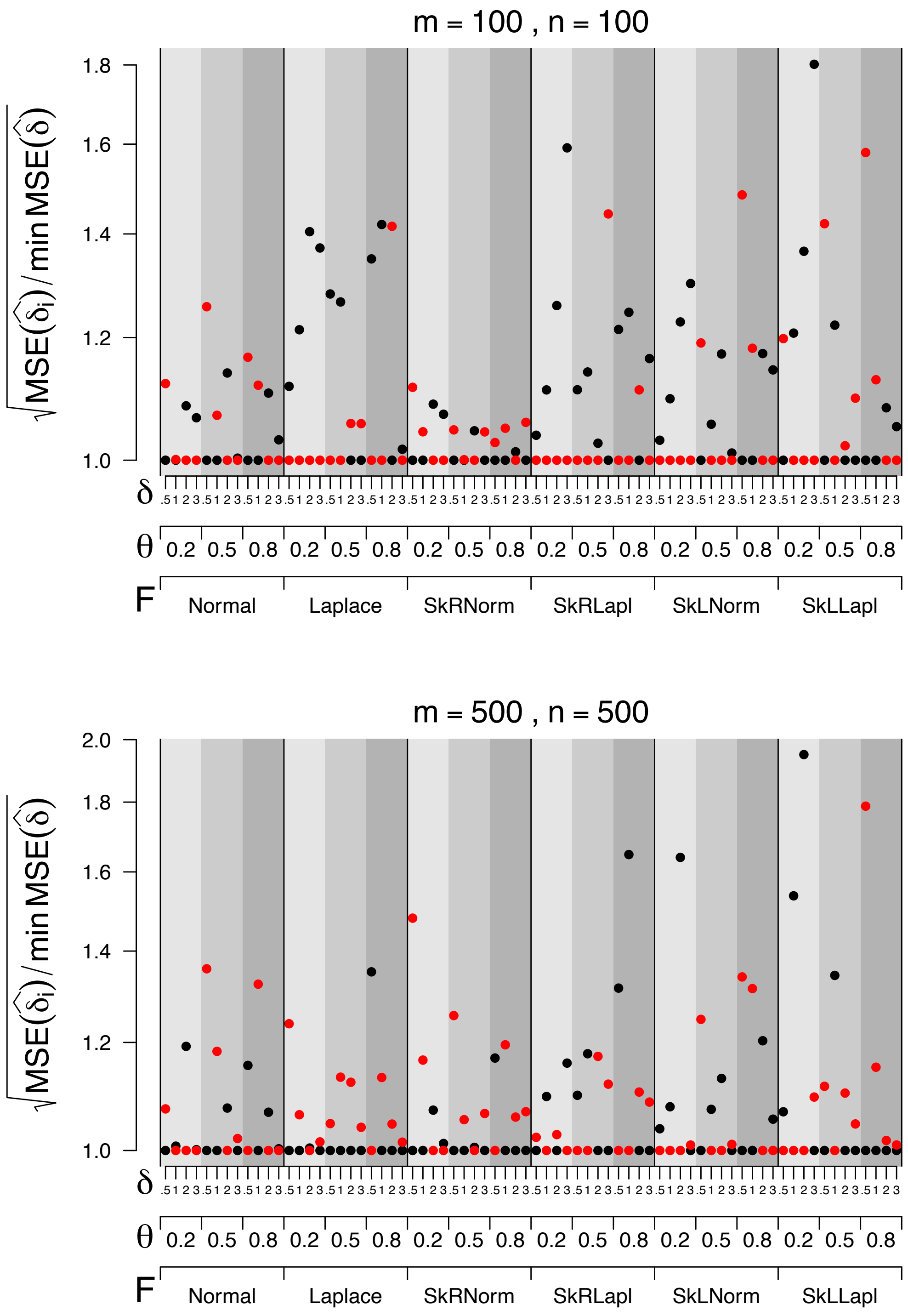


Figure 2.21: Dot Plots comparing the performances of version 2 and version 6 of the EM-like estimator of δ for moderate to large sample sizes.

Figures 2.18 - 2.19 show that for $\widehat{\theta}$ in small sample size settings, version 6 is superior to version 2 $-$ particularly when either $\theta \geq .5$ and F is symmetric or skewed right, or when $\theta = .2$ and F is skewed left. This means that $\widehat{\theta}_{EMlike6}$ prefers θ in the direction of the skew relative to $\widehat{\theta}_{EMlike2}$. Also of note is that version 6 performs better for heavier (Laplace) tailed distributions in small sample sizes. However, for $\widehat{\theta}$ in larger sample size settings version 2 performs better overall. In particular version 2 is preferred for symmetric or light (Normal) tailed distributions while version 6 retains efficiency for skewed Laplace distributions.

Figures 2.18 - 2.19 show that for $\widehat{\delta}$ in small sample size settings, version 6 is superior to version 2 with a notable exception when F is skewed left and δ is small. Version 6 also shows uniformly superior performance for $F \sim$ Laplace and generally superior performance for heavier (Laplace) tailed distributions for small sample sizes. However, for larger sample sizes version 2 performs better overall than version 6, most notably when $F \sim$ Laplace.

Figures 2.18 and **2.20** together show that for $\widehat{(\theta, \delta)}$ in small sample size settings, version 6 prefers large effect sizes for a small subset of the population while version 2 prefers a smaller effect size for a larger subset of the population. **Figures 2.19** and **2.21** show that for $\widehat{(\theta, \delta)}$ in large sample size settings, version 6 is mostly preferable for the skewed left Laplace distribution with $\theta \leq .5$ while version 2 is mostly preferred for symmetric distributions.

2.4 Pseudo-Likelihood Estimator

The likelihood of the model (1.1) when both control and treatment data are present is

$$L(f,\theta,\delta;X,Y)=\prod_{j=1}^{m}[f(x_j)]\prod_{i=1}^{n}[(1-\theta)f(y_i)+\theta f(y_i-\delta)]\,. \tag{2.18}$$

The maximum likelihood estimate of (f,θ,δ) is (f^*,θ^*,δ^*) such that $L(f,\theta,\delta;X=x,Y=y)$ is maximized at $L(f=f^*,\theta=\theta^*,\delta=\delta^*;X=x,Y=y)$. Since the set of possible f is a large space to search over, joint maximization of (f,θ,δ) is difficult. Since the control data provides direct information about f, replacing f with $\widehat{f}$ in (2.18) provides a pseudo-likelihood function, $\widehat{L}(\theta,\delta;X,Y)$, that can be maximized with respect to (θ,δ) alone.

$$\begin{aligned}\widehat{L}(\theta,\delta;X,Y)&=L(\widehat{f},\theta,\delta;X,Y)\\&=\prod_{j=1}^{m}\left[\widehat{f}(x_j)\right]\prod_{i=1}^{n}\left[(1-\theta)\widehat{f}(y_i)+\theta\widehat{f}(y_i-\delta)\right].\end{aligned} \tag{2.19}$$

Thus a dense grid search in a region of plausible (θ,δ) can be used to find the maximum of the pseudo-likelihood. The estimate of the treatment effect for this pseudo-likelihood estimator is

$$\widehat{(\theta,\delta)}_{PsL}=\underset{(\theta,\delta)}{\arg\max}\ \widehat{L}(\theta,\delta;X=x,Y=y). \tag{2.20}$$

For the grid search, the factorial combination of $\theta\in\{.01,.02,...,1\}$ and $\delta\in\{.1S_X,.2S_X,...,6S_X\}$ along with the null case $(0,0)$ is used in the function psl.inf() which can be found in section A.10 of the Appendix. Recalling that $K=\delta/\sigma$ (and $K_j=\delta_j/S_X$) the grid points of δ correspond with $K_j\in\{.1,...,6\}$. An effect size of

$K = 6$ is a utopianly high value as such a case corresponds to virtually no overlap in the responder and non-responder components - thus essentially reducing the estimation problem to the trivial case where the component labels $(Z_1, ..., Z_n)$ are observed. For rare data sets where the grid search suggests that it is plausible that $\widehat{(\theta, \delta)}_{PsL}$ could be beyond $K_j = 6$, the grid can extended to include $K_j \in \{6.1, 6.2, ..., 12\}$ when 'finite.area = FALSE'.

2.4.1 Defining Options for $\widehat{f}$

This section considers how to obtain $\widehat{f}$, an estimate of f. Clearly making use of $X_1, ..., X_m \overset{iid}{\sim} f$ is indispensable. However, $Y_1, ..., Y_n \overset{iid}{\sim} (1-\theta)f(u) + \theta f(u-\delta)$ also contains information about f and could be considered in the estimation of $\widehat{f}$. While there may be some loss of information by excluding treatment data $\boldsymbol{Y}$ in estimation, this provides a computational advantage in estimation such that the estimate of f is not based upon (θ, δ) which relinquishes the need for recursive estimation of the parameters. Furthermore the estimator of f considered here is based only upon $\boldsymbol{X}$ to retain the purity of the estimate of f upon which (θ, δ) are determined from (2.19).

Now consider the manner of obtaining $\widehat{f}(\boldsymbol{X})$, keeping in mind that obtaining $\widehat{(\theta, \delta)}_{PsL}$ is of primary interest, while estimating f is of secondary interest. To decide on a choice of $\widehat{f}$, consider the performance of $\widehat{(\theta, \delta)}_{PsL}$ under four different options for $\widehat{f}$.

The first option for $\widehat{f}$ is kernel density estimation using a standard normal kernel and a default bandwidth formula. Kernel density estimation is a commonly used technique for estimating a density nonparametrically and the two factors that determine the kernel density estimation are the selections of kernel and bandwidth. Density estimation with the normal kernel is common practice and there exist rule of thumb formulas for the bandwidth when using the standard normal kernel (Silverman, 1986). The formula given by Silverman is

$$h = 0.9\text{min}\left(S_X, \frac{IQR(X)}{1.34}\right) m^{-1/5}, \tag{2.21}$$

which has some optimality properties when the true distribution is normally distributed, and yet the IQR portion of the formula provides some robustness for non-normally distributed data.

The second option for $\widehat{f}$ is a kernel density estimate based on a $T(df = 3)$ kernel, while the third option for $\widehat{f}$ is also a kernel density estimate based on a $T(3)$ kernel but standardized to have variance 1. Use of a $T(3)$ kernel is not as common as the normal kernel is, but it is motivated here to possibly induce stability in the estimates $\widehat{(\theta, \delta)}_{PsL}$. Because of the uncommon nature of $T(3)$ kernels there is not a default bandwidth designed for this kernel in the literature, so trying the formula derived for standard normal kernels (2.21) on both a regular $T(3)$ kernel and the standardized $T(3)$ kernel effectively reduces to two different bandwidth selections for the $T(3)$ kernel. These bandwidth selections fully define the second and third options for $\widehat{f}$, respectively.

The fourth and final option for $\widehat{f}$ is a modification of the maximum likelihood estimate for a large class of densities: log-concave densities. First, consider the definition of concave down.

Definition 2.4.1 *A univariate density function f is concave down if $f(\pi x + (1 - \pi)y) \geq \pi f(x) + (1 - \pi) f(y)$ for all $x, y \in \mathbb{R}$ and $0 < \pi < 1$.*

Proposition 2.4.1 *If $f(x)$ is twice differentiable, then $f(x)$ is concave down if and only if $f''(x) \leq 0$ for all $x \in \mathbb{R}$.*

Definition 2.4.2 *A univariate density function f is said to be log-concave if $log(f)$ is concave down.*

Proposition 2.4.2 *If $log(f)$ is twice differentiable, then f is log-concave if and only if $\frac{d^2}{dx^2}log(f(x)) \leq 0$ for all $x \in \mathbb{R}$.*

A list of distributions categorized by their log-concave status is found below in **Table 2.4**

Log-concave	Log-concave (if)	Not Log-concave
Normal	Wishart ($n \geq p+1$)	T
Exponential	Dirichlet (all params > 1)	Cauchy
Uniform	Gamma (shape param > 1)	Pareto
Logistic	χ^2 (df > 2)	Log-normal
Extreme Value	Beta (both params > 1)	F
Laplace	Weibull (shape param > 1)	...

Table 2.4: List of Distributions by their log-concave status.

Note that all of the parameter conditions in the second column of **Table 2.4** correspond to the exponent in the pdf being positive. Furthermore, all log-concave distributions are unimodal (Samworth, 2018).

The log-concave density maximum likelihood estimate - $\widehat{f}_{LCMLE}$ - is nicely summarized in Chang and Walther (2007b):

> Given data $X_1, ..., X_n$ i.i.d. from f, the MLE $\widehat{f}$ of f under the restriction that f be log-concave exists uniquely and has support $[X_{(1)}, X_{(n)}]$. $log(\widehat{f})$ is a piecewise linear function whose knots are a subset of $\{X_1, ..., X_n\}$.

> The MLE can be computed e.g. using the Iterative Convex Minorant Algorithm described in Jongbloed (1998). The resulting algorithms for computing the log-concave MLE $\widehat{f}$ as given in Walther (2002) and Rufibach (2006) provide as output $\widehat{f}(X_i),\ i = 1, ..., n$. This is all that is needed for an EM-type algorithm; of course one can easily compute the entire density $\widehat{f}$ by linearly interpolating between $log\left(\widehat{f}(X_{(i)})\right)$ and $log\left(\widehat{f}(X_{(i+1)})\right)$ and then exponentiating.

This can be easily implemented using the logConDens() function from the logcondens package in R (Dümbgen and Rufibach, 2011). It uses the 'Active Set Algorithm' to perform the computation which is described in Dümbgen and Rufibach (2009) and is faster than the Iterative Convex Minorant Algorithm. To motivate the need for a modification to $\widehat{f}_{LCMLE}(\boldsymbol{X})$, consider that

$$\widehat{f}_{LCMLE}(\boldsymbol{X}) = 0 \qquad \text{for all } x \in \left\{\left(-\infty, X_{(1)}\right) \cup \left(X_{(m)}, \infty\right)\right\}. \tag{2.22}$$

Note that $P(Y_{(1)} < X_{(1)}) > 0$ and that $Y_{(1)} < X_{(1)} \implies Y_{(1)} - \delta < X_{(1)}$ for all $\delta > 0$. This means that if $y_{(1)} < x_{(1)}$ then

$$(1-\theta)\widehat{f}(y_{(1)}) + \theta\widehat{f}(y_{(1)} - \delta) = 0 \qquad \text{for all } \delta > 0, \tag{2.23}$$

which implies that

$$log\widehat{L} = \widehat{l}(\theta, \delta) = \sum_{i=1}^{m} log\left(\widehat{f}(x_i)\right) + \sum_{j=1}^{n} log\left((1-\theta)\widehat{f}(y_j) + \theta\widehat{f}(y_j - \delta)\right) = -\infty \tag{2.24}$$

for all $\delta > 0$ (which contains the whole parameter space). In such a case $\widehat{l}(\theta, \delta)$ cannot be maximized and thus $\widehat{(\theta, \delta)}_{PsL}$ does not exist. Furthermore, the relative frequency of this behavior does not diminish as $m \to \infty$, $n \to \infty$ nor does it diminish for increasing δ. To see this, consider the following.

Recalling the complete data framework of the mixture model described in section 1.2.4, the fraction of observations that are generated from $f(u)$ is $(1-\theta)$. Thus an expected $(1-\theta)n$ of the treatment observations are generated from $f(u)$. There are also m control observations generated from $f(u)$. Each of these observations is equally likely to be the minimum value (there is also a smaller probability that one of the treatment observations from $f(u-\delta)$ is the smallest value). Therefore, if $\widehat{f}_{LCMLE}$ is used

$$\begin{aligned} P(Y_{(1)} < X_{(1)}) &\geq \frac{n(1-\theta)}{m+n(1-\theta)} \\ &= \frac{r(1-\theta)}{1+r(1-\theta)} \quad \text{where } r = n/m \\ \Longrightarrow \lim_{m,n\to\infty} P(\widehat{(\theta,\delta)}_{PsL} \text{ DNE}) &\geq \frac{r_\infty(1-\theta)}{1+r_\infty(1-\theta)}, \end{aligned} \tag{2.25}$$

where $r_\infty = \lim_{m,n\to\infty} n/m$. This lower bound of the limit (2.25) is only 0 if $\theta = 1$ or $r_\infty = 0$.

To avoid this issue caused by $\widehat{f}$ with unbounded support, consider instead the modified log-concave estimate of f below

$$\widehat{f}_{mLC}(x) = \begin{cases} k_1 exp(a_1 x) & x < x_{(1)} \\ \dfrac{m-2}{m}\widehat{f}_{LCMLE}(x) & x \in [x_{(1)}, x_{(m)}] \\ k_2 exp(a_2 x) & x > x_{(m)}. \end{cases} \tag{2.26}$$

The constants (k_1, a_1, k_2, a_2) in (2.26) can be chosen so that each exponential tail has area $1/m$ while ensuring that $\widehat{f}_{mLC}(x)$ is continuous at $x_{(1)}$ and $x_{(m)}$. Note that continuity is achieved by ensuring

$$\lim_{x \to^{-} x_{(1)}} \widehat{f}_{mLC}(x) = \frac{m-2}{m} \widehat{f}_{LCMLE}(x_{(1)}) \tag{2.27}$$

$$\lim_{x \to^{+} x_{(m)}} \widehat{f}_{mLC}(x) = \frac{m-2}{m} \widehat{f}_{LCMLE}(x_{(m)}). \tag{2.28}$$

To solve for the appropriate k_1, a_1 in (2.26), the following two equations must be satisfied

$$k_1 exp(a_1 x_{(1)}) \overset{set}{=} \frac{m-2}{m} \widehat{f}_{LCMLE}(x_{(1)}) \tag{2.29}$$

$$\int_{-\infty}^{x_{(1)}} k_1 exp(a_1 x) dx = \frac{k_1}{a_1} exp(a_1 x_{(1)}) \overset{set}{=} \frac{1}{m}. \tag{2.30}$$

Rearranging (2.29) means that

$$k_1 = \frac{m-2}{m} \widehat{f}_{LCMLE}(x_{(1)}) \ exp(-a_1 x_{(1)}) \tag{2.31}$$

so substituting k_1 into (2.30) gives

$$\frac{1}{a_1} \frac{m-2}{m} \widehat{f}_{LCMLE}(x_{(1)}) \ exp(-a_1 x_{(1)}) exp(a_1 x_{(1)}) = \frac{1}{m} \quad (a_1 > 0) \tag{2.32}$$

and rearranging provides

$$a_1 = (m-2) \widehat{f}_{LCMLE}(x_{(1)}). \tag{2.33}$$

Similarly, to solve for the appropriate k_2, a_2 in (2.26), the following two equations must be satisfied

$$k_2 exp(a_2 x_{(m)}) \overset{set}{=} \frac{m-2}{m} \widehat{f}_{LCMLE}(x_{(m)}) \tag{2.34}$$

$$\int_{x_{(m)}}^{\infty} k_2 exp(a_2 x) dx = -\frac{k_2}{a_2} exp(a_2 x_{(m)}) \overset{set}{=} \frac{1}{m} \quad (a_2 < 0). \tag{2.35}$$

Rearranging (2.34) means that

$$k_2 = \frac{m-2}{m} \widehat{f}_{LCMLE}(x_{(m)})\ exp(-a_2 x_{(m)}) \tag{2.36}$$

so substituting k_2 into (2.35) gives

$$-\frac{(m-2)}{m} \frac{\widehat{f}_{LCMLE}(x_{(m)}) exp(-a_2 x_{(m)})}{a_2}\ exp(a_2 x_{(m)}) = \frac{1}{m} \tag{2.37}$$

and rearanging provides

$$a_2 = -(m-2)\widehat{f}_{LCMLE}(x_{(m)}). \tag{2.38}$$

Note also that because this density estimate has area $1/m$ in each tail, it has the following connection with the emperical cdf of the control data $\widehat{F}_m(x)$

$$\widehat{F}_{mLC}(x_{(1)}) = \widehat{F}_m(x_{(1)}) = \frac{1}{m} \tag{2.39}$$

$$\widehat{F}_{mLC}(x_{(m)}) = \lim_{x \to^- x_{(m)}} \widehat{F}_m(x_{(m)}) = \frac{m-1}{m} \tag{2.40}$$

See **Figure 2.22** below for an example of $\widehat{f}_{LCMLE}(x)$ and $\widehat{f}_{mLC}(x)$ for a random sample of fifty observations from a Skewed Right Laplace distribution.

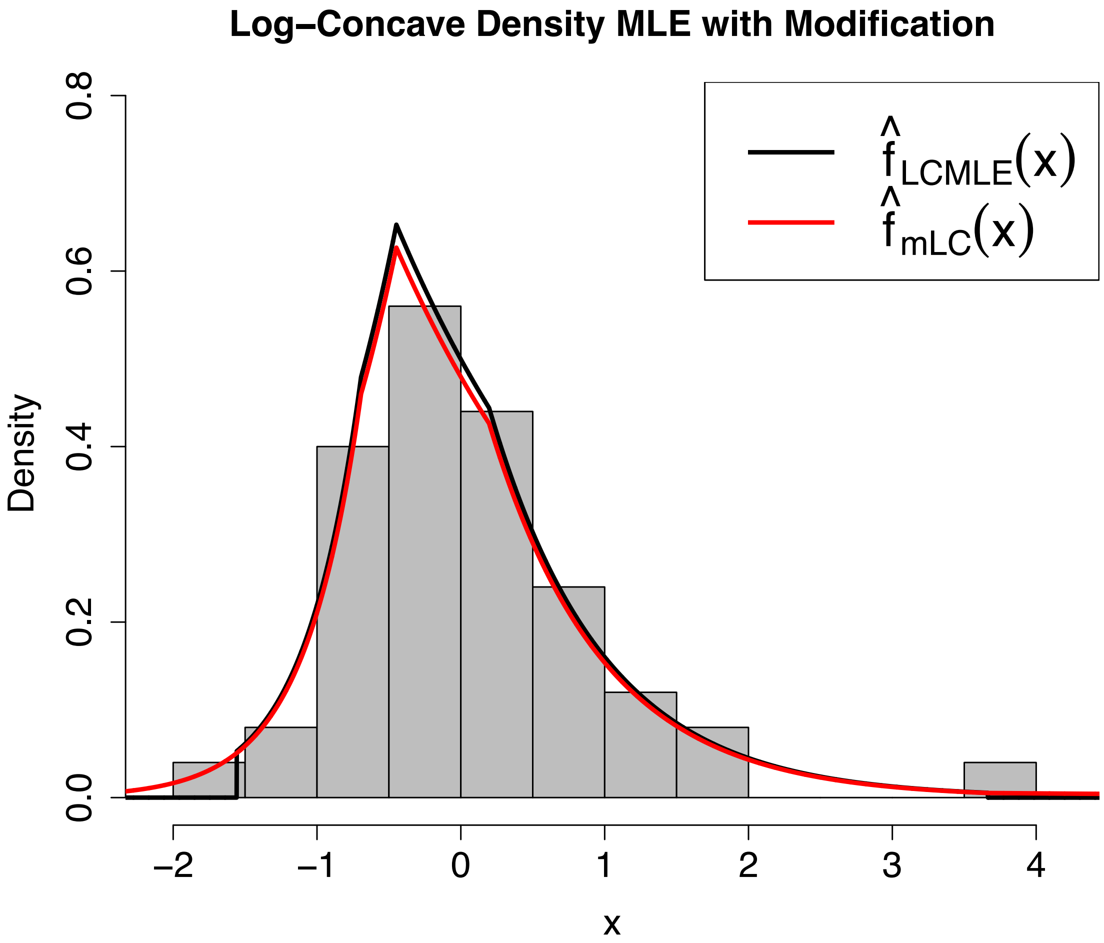


Figure 2.22: Log Concave Density Maximum Likelihood Estimate with Modification. Fifty i.i.d observations are drawn from a skewed generalized T distribution with parameters $(\mu = 0, \sigma = 1, \lambda = .5, p = 1, q = \infty)$.

Even with a fairly small sample size, $m = 50$, the modification to the MLE is minor and achieves tail behavior of a distribution that is not bounded by $X_{(1)}$ and $X_{(m)}$. Thus, to avert the possibility of an unbounded (2.24), the fourth option for $\widehat{f}$ is $\widehat{f}_{mLC}$ as defined above. For a summary of all candidate $\widehat{f}$, see **Table 2.5** below.

$\widehat{f}$ Variations		
$\widehat{f}$	Kernel	Bandwidth
$\widehat{f}_{KDE.Norm}$	Normal	$0.9\text{min}\left(S_X, \dfrac{IQR(X)}{1.34}\right) m^{-1/5}$
$\widehat{f}_{KDE.T3.NA}$	T(3)	$0.9\text{min}\left(S_X, \dfrac{IQR(X)}{1.34}\right) m^{-1/5}$
$\widehat{f}_{KDE.T3.Adj}$	T(3)	$3^{-1/2}0.9\text{min}\left(S_X, \dfrac{IQR(X)}{1.34}\right) m^{-1/5}$
$\widehat{f}_{mLC}$	N/A	N/A

Table 2.5: Candidate $\widehat{f}$s

2.4.2 Simulation for Selecting $\widehat{f}$

To compare the performance of $\widehat{(\theta,\delta)}_{PsL}$ with the four different choices of $\widehat{f}$ presented in **Table 2.5**, consider a simulation study that generates 1000 data sets under each following factorial combinations of settings in **List 2.1**. Recall that (2.17) provides the score to compare robust performance across (F, θ, δ) of estimator i among a set of candidate estimates (indexed by k, here $k \in \{1, ..., 4\}$). **Table 2.6** below displays scores for $\widehat{\theta}$, $\widehat{\delta}$ and $\widehat{\Delta} = \widehat{\theta}\,\widehat{\delta}$. The score represents the average relative loss in performance of each estimator compared to an 'oracle' estimator (that choose the optimal $\widehat{f}$ given the true parameters). For example, an estimator with a score of $S(\widehat{\delta}) = 1.05$ has a $\sqrt{MSE(\widehat{\delta})}$ that is on average 5% larger than the oracle estimator. **Table 2.6** also presents the geometric average of $S(\widehat{\Delta}), S(\widehat{\theta}),$ and $S(\widehat{\delta})$ as a summary score for each estimation method. The smallest scores are highlighted in yellow. **Figures 2.23 - 2.24** display the scores graphically.

m = n	$\widehat{f}$ Variation	$S(\widehat{\theta})$	$S(\widehat{\delta})$	$S(\widehat{\Delta})$	$\{S(\widehat{\Delta})S(\widehat{\theta})S(\widehat{\delta})\}^{1/3}$
25	Normal	1.081	1.133	1.033	1.082
	$T(3)_{NA}$	1.149	1.051	1.024	1.073
	$T(3)_{Adj}$	1.016	1.053	1.020	1.030
	mod LogCon	1.059	1.056	1.026	1.047
50	Normal	1.111	1.232	1.051	1.129
	$T(3)_{NA}$	1.204	1.077	1.040	1.105
	$T(3)_{Adj}$	1.028	1.076	1.024	1.042
	mod LogCon	1.049	1.058	1.017	1.041
100	Normal	1.151	1.363	1.070	1.189
	$T(3)_{NA}$	1.273	1.111	1.048	1.140
	$T(3)_{Adj}$	1.050	1.110	1.022	1.060
	mod LogCon	1.047	1.070	1.009	1.042
500	Normal	1.200	1.626	1.123	1.299
	$T(3)_{NA}$	1.421	1.231	1.074	1.234
	$T(3)_{Adj}$	1.089	1.143	1.025	1.084
	mod LogCon	1.029	1.061	1.004	1.031

Table 2.6: Scores for Estimates of Δ, θ, δ.

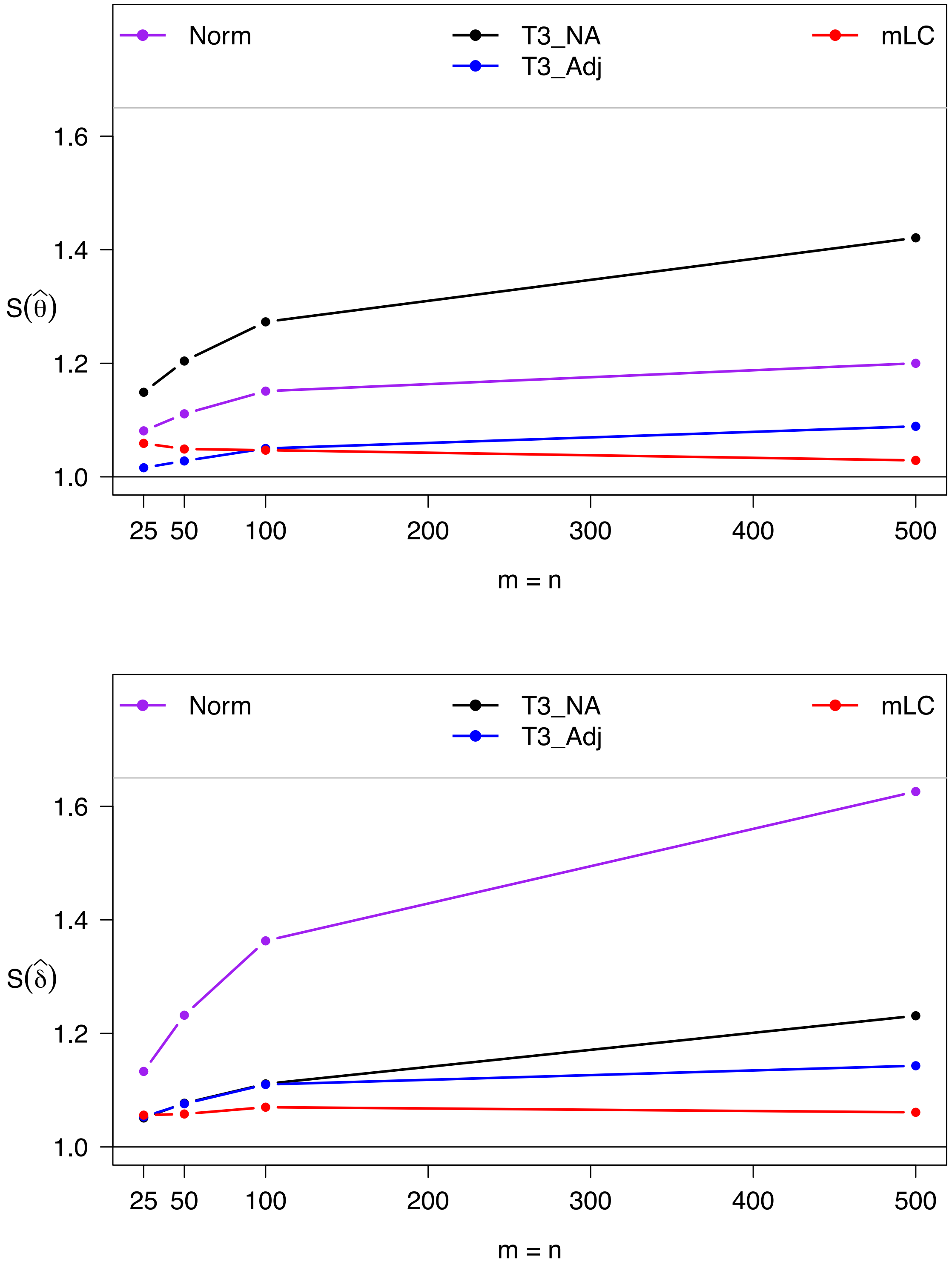


Figure 2.23: Pseudo-Likelihood Scores for $\widehat{\theta}$ and $\widehat{\delta}$

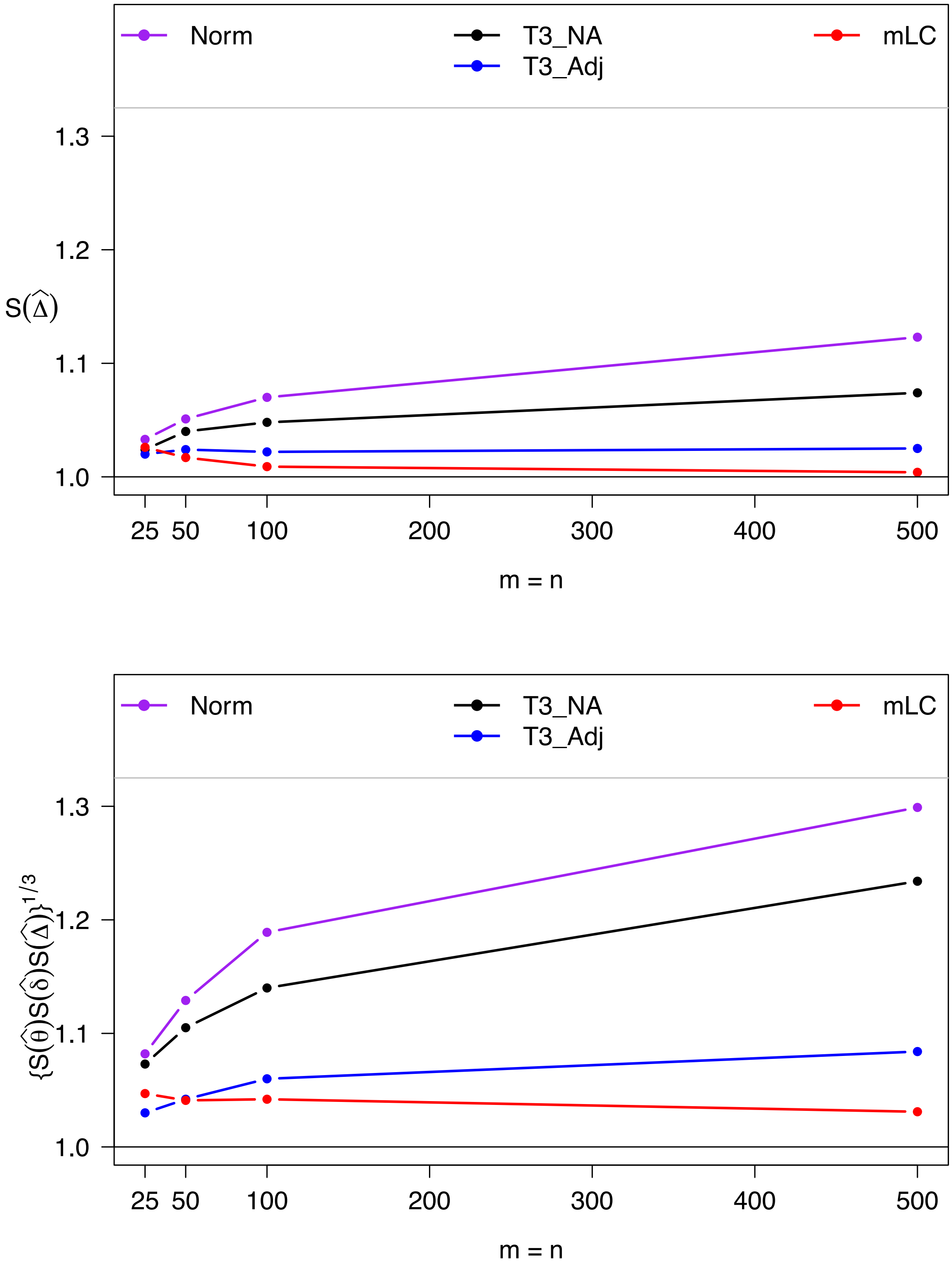


Figure 2.24: Pseudo-Likelihood Score for $\widehat{\Delta}$ and an Overall Summary Score

For some of the smallest sample sizes, one of the kernel density estimates with a $T(3)$ kernel sometimes show the best performance, but for moderate to large sample sizes $\widehat{f}_{mLC}$ produces better estimation. A secondary advantage of $\widehat{f}_{mLC}$ is that it provides a unimodal density estimate while $T(3)_{Adj}$ is almost certain to produce a multimodal estimate of f. Therefore, $\widehat{f}_{mLC}$ is recommended for pseudo-likelihood point estimation.

Parameter Specific Performance Comparison

The plots on the following pages in **Figures 2.26 - 2.29** compare the two best estimator versions under each of the 288 simulation settings by plotting either $\sqrt{MSE(\widehat{\theta}_{T3_{Adj}})/MSE(\widehat{\theta}_{\text{mLC}})}$ or $\sqrt{MSE(\widehat{\delta}_{T3_{Adj}})/MSE(\widehat{\delta}_{\text{mLC}})}$. Each plot has fixed sample sizes and displays one estimator ($\widehat{\theta}$ or $\widehat{\delta}$). Each plot displays the ratio for each of the 72 (F, θ, δ). The axes determine (θ, δ). At each coordinate, the results for all 6 Fs are shown as a 2×3 grid with row representing tail behavior and column representing skew behavior. The ratio is displayed at each grid location with a colored background to facilitate pattern recognition. **Figure 2.25** displays the color key for the $T(3)_{Adj}$ and Log-Concave options for $\widehat{f}$.

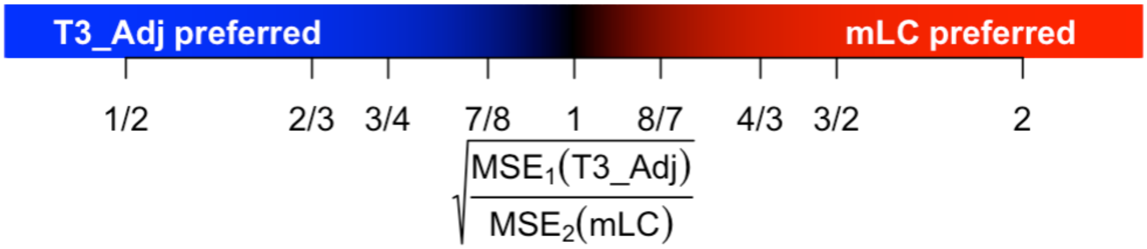
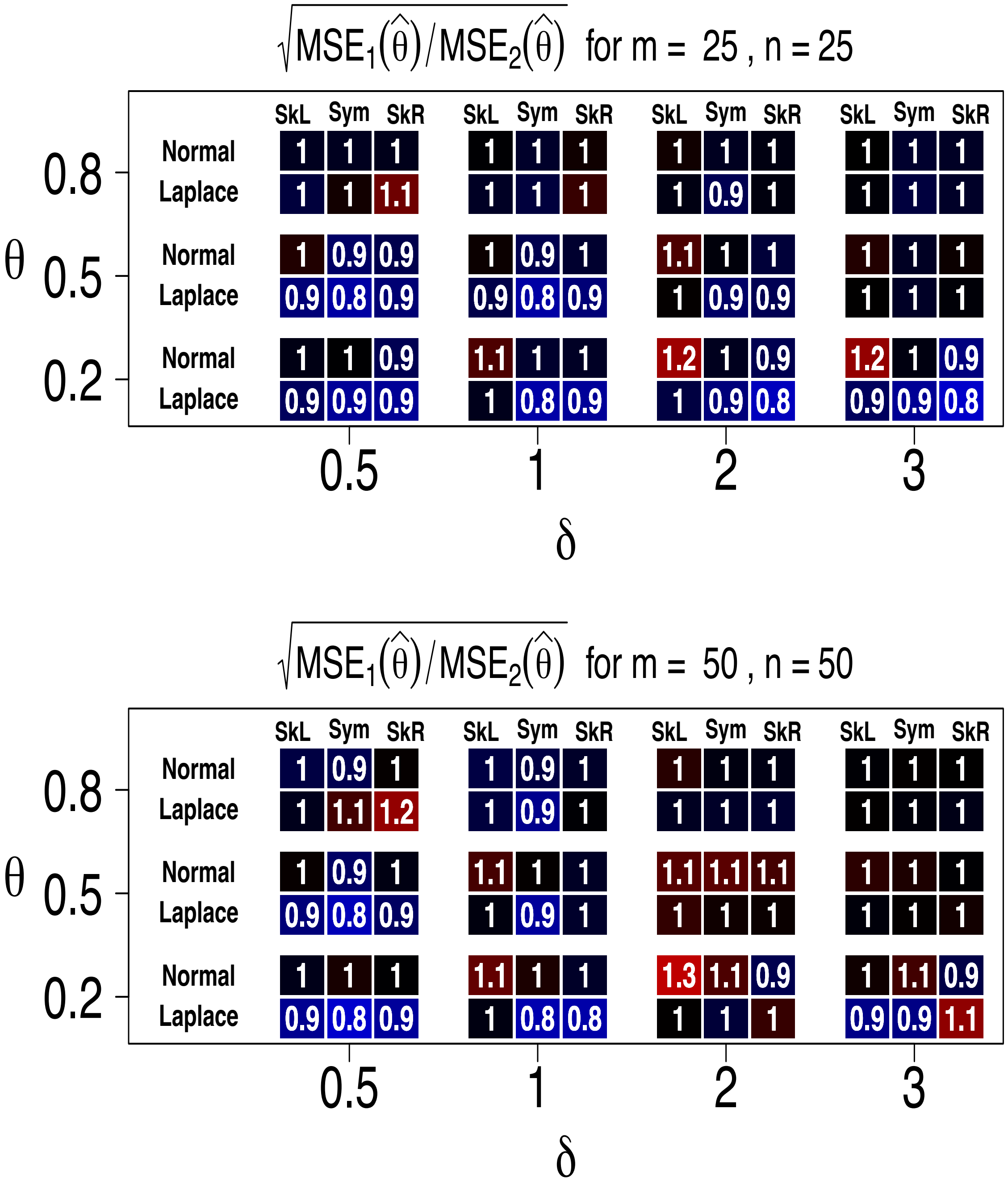


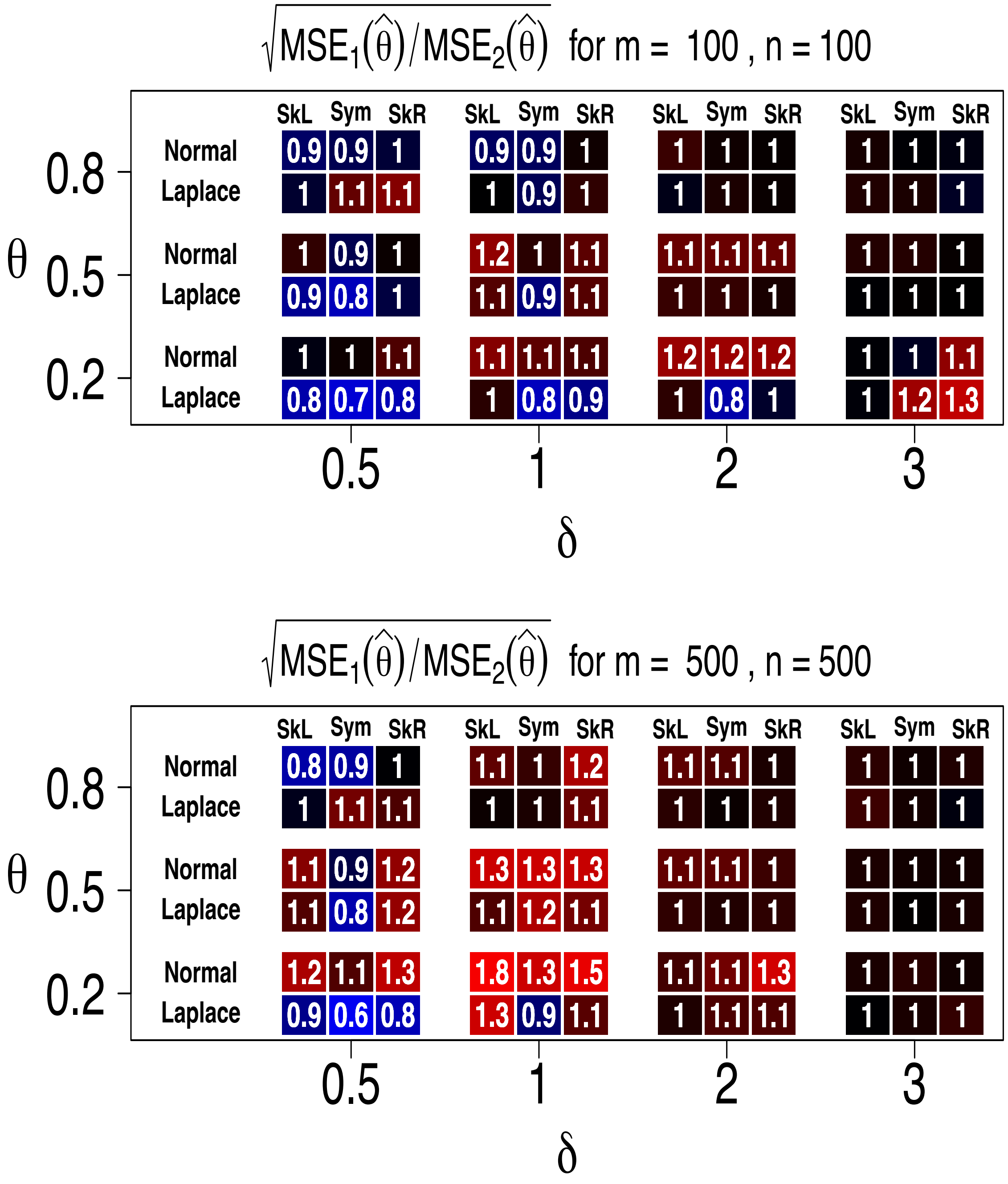
Figure 2.25: Color Key for Candidate $\widehat{f}$



θ		SkL	Sym	SkR	SkL	Sym	SkR	SkL	Sym	SkR	SkL	Sym	SkR
0.8	Normal	1	1	1	1	1	1	1	1	1	1	1	1
	Laplace	1	1	1.1	1	1	1	1	0.9	1	1	1	1
0.5	Normal	1	0.9	0.9	1	0.9	1	1.1	1	1	1	1	1
	Laplace	0.9	0.8	0.9	0.9	0.8	0.9	1	0.9	0.9	1	1	1
0.2	Normal	1	1	0.9	1.1	1	1	1.2	1	0.9	1.2	1	0.9
	Laplace	0.9	0.9	0.9	1	0.8	0.9	1	0.9	0.8	0.9	0.9	0.8
δ		0.5			1			2			3		

θ		SkL	Sym	SkR	SkL	Sym	SkR	SkL	Sym	SkR	SkL	Sym	SkR
0.8	Normal	1	0.9	1	1	0.9	1	1	1	1	1	1	1
	Laplace	1	1.1	1.2	1	0.9	1	1	1	1	1	1	1
0.5	Normal	1	0.9	1	1.1	1	1	1.1	1.1	1.1	1	1	1
	Laplace	0.9	0.8	0.9	1	0.9	1	1	1	1	1	1	1
0.2	Normal	1	1	1	1.1	1	1	1.3	1.1	0.9	1	1.1	0.9
	Laplace	0.9	0.8	0.9	1	0.8	0.8	1	1	1	0.9	0.9	1.1
δ		0.5			1			2			3		

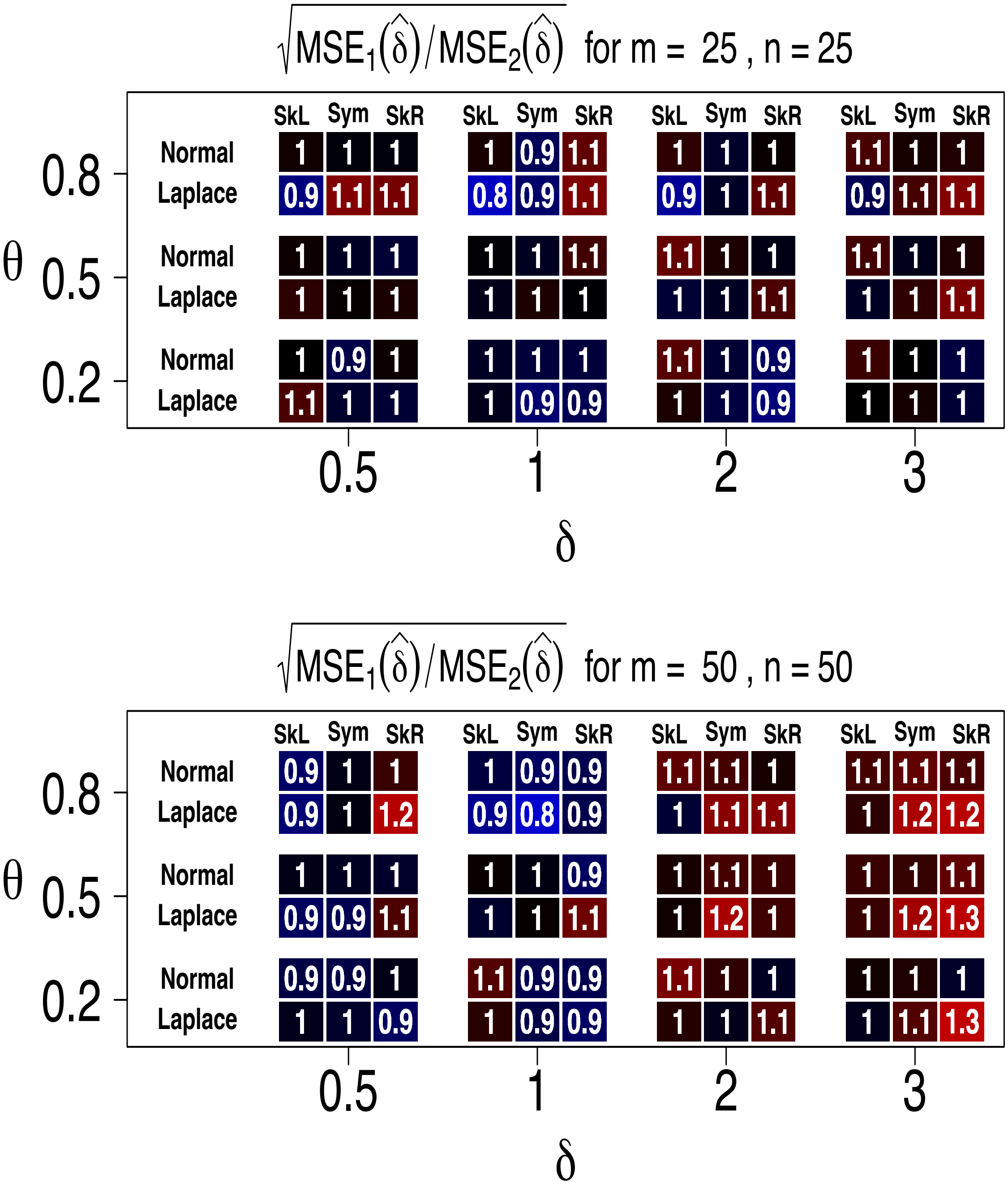
Figure 2.26: Heat Grids comparing the performance of Pseudo-likelihood Estimator Variations of $\widehat{\theta}$ for small sample sizes.



θ		0.5			1			2			3		
		SkL	Sym	SkR	SkL	Sym	SkR	SkL	Sym	SkR	SkL	Sym	SkR
0.8	Normal	0.9	0.9	1	0.9	0.9	1	1	1	1	1	1	1
	Laplace	1	1.1	1.1	1	0.9	1	1	1	1	1	1	1
0.5	Normal	1	0.9	1	1.2	1	1.1	1.1	1.1	1.1	1	1	1
	Laplace	0.9	0.8	1	1.1	0.9	1.1	1	1	1	1	1	1
0.2	Normal	1	1	1.1	1.1	1.1	1.1	1.2	1.2	1.2	1	1	1.1
	Laplace	0.8	0.7	0.8	1	0.8	0.9	1	0.8	1	1	1.2	1.3

θ		0.5			1			2			3		
		SkL	Sym	SkR	SkL	Sym	SkR	SkL	Sym	SkR	SkL	Sym	SkR
0.8	Normal	0.8	0.9	1	1.1	1	1.2	1.1	1.1	1	1	1	1
	Laplace	1	1.1	1.1	1	1	1.1	1	1	1	1	1	1
0.5	Normal	1.1	0.9	1.2	1.3	1.3	1.3	1.1	1.1	1	1	1	1
	Laplace	1.1	0.8	1.2	1.1	1.2	1.1	1	1	1	1	1	1
0.2	Normal	1.2	1.1	1.3	1.8	1.3	1.5	1.1	1.1	1.3	1	1	1
	Laplace	0.9	0.6	0.8	1.3	0.9	1.1	1	1.1	1.1	1	1	1

Figure 2.27: Heat Grids comparing the performance of Pseudo-likelihood Estimator Variations of $\widehat{\theta}$ for moderate to large sample sizes.



$\sqrt{MSE_1(\hat{\delta})/MSE_2(\hat{\delta})}$ for m = 25 , n = 25

θ		δ = 0.5			1			2			3		
		SkL	Sym	SkR	SkL	Sym	SkR	SkL	Sym	SkR	SkL	Sym	SkR
0.8	Normal	1	1	1	1	0.9	1.1	1	1	1	1.1	1	1
	Laplace	0.9	1.1	1.1	0.8	0.9	1.1	0.9	1	1.1	0.9	1.1	1.1
0.5	Normal	1	1	1	1	1	1.1	1.1	1	1	1.1	1	1
	Laplace	1	1	1	1	1	1	1	1	1.1	1	1	1.1
0.2	Normal	1	0.9	1	1	1	1	1.1	1	0.9	1	1	1
	Laplace	1.1	1	1	1	0.9	0.9	1	1	0.9	1	1	1

$\sqrt{MSE_1(\hat{\delta})/MSE_2(\hat{\delta})}$ for m = 50 , n = 50

θ		δ = 0.5			1			2			3		
		SkL	Sym	SkR	SkL	Sym	SkR	SkL	Sym	SkR	SkL	Sym	SkR
0.8	Normal	0.9	1	1	1	0.9	0.9	1.1	1.1	1	1.1	1.1	1.1
	Laplace	0.9	1	1.2	0.9	0.8	0.9	1	1.1	1.1	1	1.2	1.2
0.5	Normal	1	1	1	1	1	0.9	1	1.1	1	1	1	1.1
	Laplace	0.9	0.9	1.1	1	1	1.1	1	1.2	1	1	1.2	1.3
0.2	Normal	0.9	0.9	1	1.1	0.9	0.9	1.1	1	1	1	1	1
	Laplace	1	1	0.9	1	0.9	0.9	1	1	1.1	1	1.1	1.3

Figure 2.28: Heat Grids comparing the performance of Pseudo-likelihood Estimator Variations of $\widehat{\delta}$ for small sample sizes.

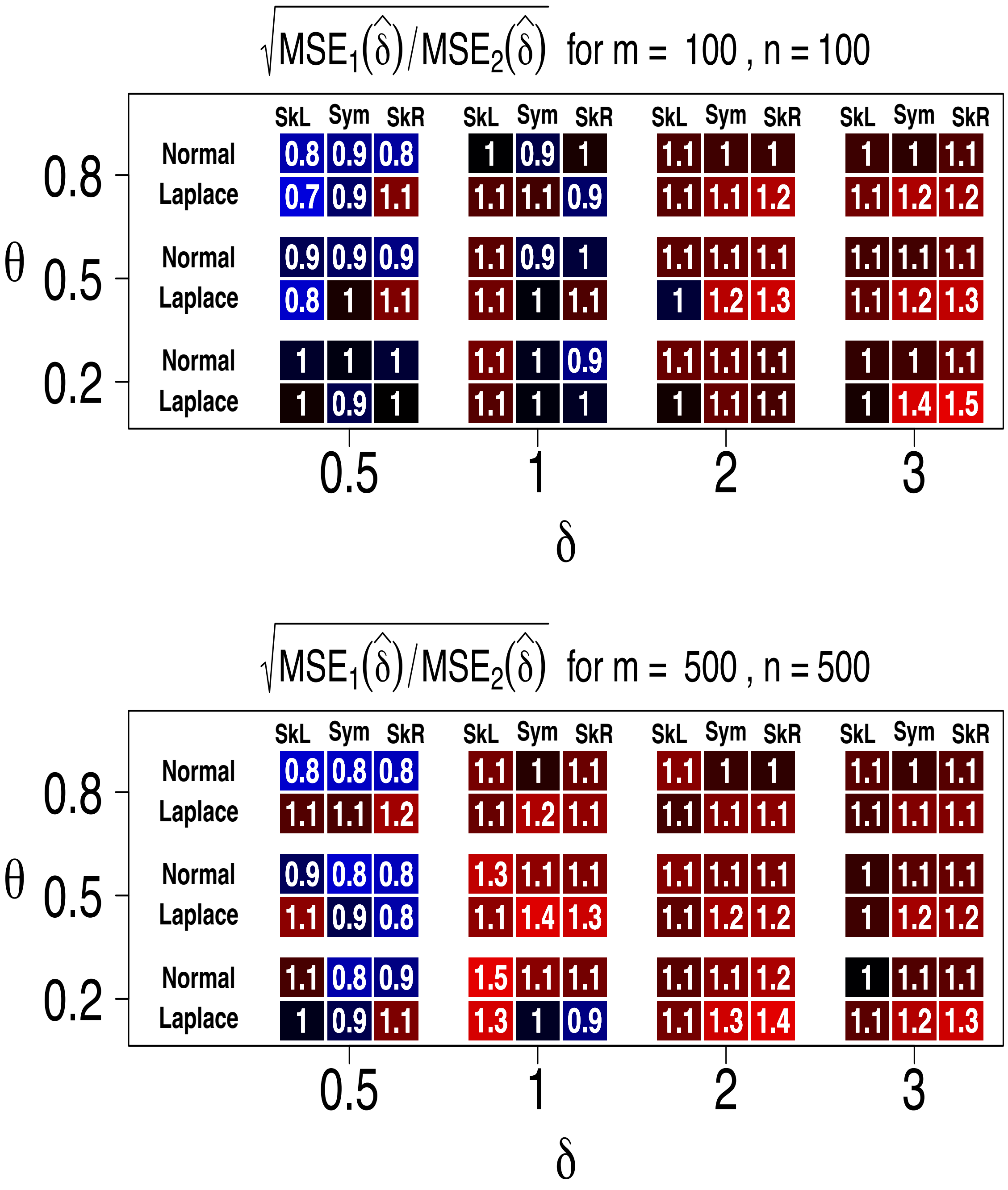


Figure 2.29: Heat Grids comparing the performance of Pseudo-likelihood Estimator Variations of $\widehat{\delta}$ for moderate to large sample sizes.

Figures 2.26 - 2.27 indicates that the $T(3)_{Adj}$ variation of $\widehat{\theta}$ is slightly preferred in small sample sizes, showing improvement with the heavier (Laplace) tailed distributions. As the sample sizes increase, the mLC version shows superior performance for larger δ while $T(3)_{Adj}$ shows superior performance for smaller δ. Across all sample sizes, the $T(3)_{Adj}$ version favors smaller θ for the heavier (Laplace) tailed distributions and small δ. Also across all sample sizes $T(3)_{Adj}$ favors larger θ for lighter (Normal) tailed distributions and small δ.

Figures 2.28 - 2.29 indicate that the discrepancies between $\widehat{\delta}$ for the 2 versions are minor for small sample sizes, though it appears that when $\theta = .8$ and heavy (Laplace) tailed distributions are present $T(3)_{Adj}$ is preferred for skewed left distributions and the log-concave version is preferred for right skewed ones. As the sample sizes increase, the $T(3)_{Adj}$ version becomes increasingly preferable for $\delta = .5$ and lighter (Normal) tails, while the log-concave version becomes preferred for nearly all other scenarios.

Considering the pair $\widehat{(\theta,\delta)}_{PsL}$, the colored grids corroborate the fact that the $T(3)_{Adj}$ variation is slightly preferred with very small sample sizes. The grids also reveal that the only setting for which $\widehat{(\theta,\delta)}_{T(3)_{Adj}}$ is consistently preferred is when F has Gaussian tails, $\delta = .5$ and $\theta \geq .8$ (which is the smallest effect size of the setting closest to the traditional model that assumes $F \sim$ Normal with a pure shift). The primary pattern revealed in the grids is that all other settings eventually give way to comparable or preferential performance of $\widehat{(\theta,\delta)}_{mLC}$. In particular, $\widehat{(\theta,\delta)}_{mLC}$ is uniformly preferred if δ is large enough, where the lower bound for “large enough” decreases as the sample sizes increase.

Chapter 3

Confidence Bounds

A point estimate of (θ, δ) does not quantify the uncertainty surrounding the treatment effect. Confidence bounds can provide information about this uncertainty. This chapter first considers confidence bounds corresponding to the method of moment estimator, with confidence intervals for θ and δ in section 3.1 and confidence regions for (θ, δ) in section 3.2. Then confidence bounds for the pseudo-likelihood method are considered, with confidence regions in section 3.3 and confidence intervals in section 3.4.

3.1 Method of Moment Confidence Intervals

Most of this section is identical to previously published work (Lubich et al., 2022).

Consider confidence intervals that are based on the method of moment estimators in (2.4) and (2.5). Section 3.1.1 considers asymptotic intervals that are based on the asymptotic properties of the moment estimators. Section 3.1.2 considers bootstrap intervals constructed from bootstrap sampling distributions.

3.1.1 Asymptotic Moment Intervals

Consider first asymptotic confidence intervals that rely on the consistency and asymptotic normality of $\widehat{\theta}_{MoM}$ and $\widehat{\delta}_{MoM}$ presented in Propositions 2.2.2 and 2.2.3. These propositions ensure that for sufficiently large (m,n) – here considering $m = n$ – $\widehat{\theta}_{MoM} \mathrel{\dot{\sim}} N(\theta, \sigma_\theta^2/n)$ and $\widehat{\delta}_{MoM} \mathrel{\dot{\sim}} N(\delta, \sigma_\delta^2/n)$, where $\dot{\sim}$ means 'is approximately distributed as'. Therefore the proposed asymptotic $100(1-\alpha)\%$ confidence intervals for θ and δ, respectively, are

$$CI_{MoM}(\theta) = \left(\widehat{\theta} - z_{\alpha/2}\widehat{\sigma}_\theta/\sqrt{n},\ \widehat{\theta} + z_{\alpha/2}\widehat{\sigma}_\theta/\sqrt{n}\right) \tag{3.1}$$

$$CI_{MoM}(\delta) = \left(\widehat{\delta} - z_{\alpha/2}\widehat{\sigma}_\delta/\sqrt{n},\ \widehat{\delta} + z_{\alpha/2}\widehat{\sigma}_\delta/\sqrt{n}\right), \tag{3.2}$$

where $z_{\alpha/2} = \Phi^{-1}(1-\alpha/2)$. The standard errors, $\widehat{\sigma}_\theta$ and $\widehat{\sigma}_\delta$, are found by plugging in the sample moments as estimates for the population moments found in the asymptotic variance formulas and making the same alterations as in the estimators. That is, $(\mu_Y - \mu_X)$ is substituted with $\left(\overline{Y} - \overline{X}\right)_+ + \epsilon_N$ and $(\sigma_Y^2 - \sigma_X^2)$ is substituted with $(S_Y^2 - S_X^2)_+$. Finally, the boundaries of the asymptotic confidence interval are truncated at the edges of the parameter space when necessary.

In addition to the confidence intervals for θ and δ, the natural asymptotic confidence interval for $\Delta = \theta\delta$ can be considered as well. Recalling that $\widehat{\Delta}_{MoM} = (\overline{Y} - \overline{X})_+$, a $100(1-\alpha)\%$ confidence interval for Δ is given by

$$CI_{MoM}(\Delta) = \left(\widehat{\Delta}_{MoM} - z_{\alpha/2}\sqrt{\frac{S_X^2}{m} + \frac{S_Y^2}{n}},\ \widehat{\Delta}_{MoM} + z_{\alpha/2}\sqrt{\frac{S_X^2}{m} + \frac{S_Y^2}{n}}\right), \tag{3.3}$$

where the lower bound is truncated at 0 as needed. The central limit theorem ensures that this has the desired coverage probability. For most distributions, 30 observations per group is ample for satisfactory approximation of the asymptotic distribution (Diez et al., 2012) so long as the overall effect size Δ is large enough so the probability that $\widehat{\Delta}_{MoM} = 0$ is small.

3.1.2 Bootstrap Moment Confidence Intervals

Bootstrapping is a general approach to constructing confidence intervals that does not depend on knowledge of the (asymptotic) distribution of an estimator. Bootstrapping involves repeatedly resampling the observed data with replacement and computing a statistic for each resample to provide a bootstrap sampling distribution of the statistic. This bootstrap sampling distribution can be used for inference. For a more information on bootstrapping, see Efron and Tibshirani (1994). Bootstrap confidence intervals are motivated as an alternative to asymptotic confidence intervals since they may possibly provide better performance, particularly when the sample sizes are small and the asymptotic approximations are imprecise. Consider constructing bootstrap intervals for θ and δ from model (1.1) using $\widehat{\theta}_{MoM}$ and $\widehat{\delta}_{MoM}$ as the statistics respectively. Implementation of bootstrap intervals in this context involves the following general steps

(a) Randomly sample from $X_1, ..., X_m$ and $Y_1, ..., Y_n$ independently with replacement B=1000 times.

(b) For each of these 1000 bootstrap samples, calculate $\widehat{\theta}_b$ and $\widehat{\delta}_b$ to obtain bootstrap sampling distributions.

(c) Determine the bounds of the confidence intervals for θ and δ by using the bootstrap sampling distributions.

Multiple methods for step (c) may be considered. Let $\widehat{\tau}$ represent the method of moment estimate for a generic parameter (either θ or δ). Percentile Bootstrap Intervals select percentiles of the bootstrap distribution, $\widehat{\tau}_b^{(\alpha_1)}$ and $\widehat{\tau}_b^{(\alpha_2)}$, such that $\alpha_1 + (1 - \alpha_2) = \alpha$ and use $[\widehat{\tau}_b^{(\alpha_1)}, \widehat{\tau}_b^{(\alpha_2)}]$ as the $100(1 - \alpha)\%$ confidence interval.

Centered Bootstrap Percentile Intervals use the percentiles in a different manner, $[2\widehat{\tau} - \widehat{\tau}_b^{(\alpha_2)}, 2\widehat{\tau} - \widehat{\tau}_b^{(\alpha_1)}]$, for a $100(1-\alpha)\%$ confidence interval. Another method of bootstrap intervals is called BC_a (Bias-Corrected accelerated) that typically produces better results than the aforementioned approaches (Efron, 1987). Consider implementation of the BC_a confidence intervals

1. Randomly sample from $X_1, ..., X_m$ and $Y_1, ..., Y_n$ independently with replacement B=1000 times.
2. For each of these 1000 bootstrap samples, calculate $\widehat{\theta}_b$ and $\widehat{\delta}_b$ to obtain bootstrap sampling distributions.
3. Calculate the acceleration (a) and bias (z_0) correction terms for both $\widehat{\theta}$ and $\widehat{\delta}$ based on (3.4) and (3.5), respectively.
4. Calculate the percentiles of the bootstrap distributions to use for the confidence interval based on (3.6) and (3.7).

Typically (Efron, 1987) in step 3, z_0 is calculated as $z_0 = \Phi^{-1}\left(\#\left\{\widehat{\tau}_b < \widehat{\tau}\right\}/B\right)$ and

$$a = \frac{\sum_{i=1}^{n}\left(\overline{\tau}_b - \widehat{\tau}_{(-i)}\right)^3}{6\{\sum_{i=1}^{n}\left(\overline{\tau}_b - \widehat{\tau}_{(-i)}\right)^2\}^{3/2}}, \tag{3.4}$$

where $\#$ is the counting operator and $\widehat{\tau}_{(-i)}$ is the estimate with the i^{th} observation removed. However, this formula for z_0 can fail for $\widehat{\theta}$ or $\widehat{\delta}$ because of the bounded nature of the parameter space, and thus the estimators. There is non-zero probability that $\widehat{\delta} = 0$, in which case $z_0 = -\infty$. Consider the following proposed adjustment for the discrete nature of the bootstrap sampling distributions by taking

$$z_0 = \Phi^{-1}\left(\left\{\#\left(\widehat{\tau}_b < \widehat{\tau}\right) + \frac{1}{2}\#\left(\widehat{\tau}_b = \widehat{\tau}\right)\right\}/B\right). \tag{3.5}$$

Step 4 remains unchanged, letting

$$\alpha_l = \Phi\left(z_0 + \frac{z_0 - z_{\alpha/2}}{1 - a(z_0 - z_{\alpha/2})}\right) \tag{3.6}$$

$$\alpha_u = \Phi\left(z_0 + \frac{z_0 + z_{\alpha/2}}{1 - a(z_0 + z_{\alpha/2})}\right), \tag{3.7}$$

giving the BC_a interval $[\widehat{\tau}_b^{(\alpha_l)}, \widehat{\tau}_b^{(\alpha_u)}]$.

3.1.3 Performance Comparison of Moment Intervals

In the following pages, **Tables 3.1** and **3.2** present coverage probabilities of the asymptotic and BC_a method of moment confidence intervals for θ and δ, respectively, for the following combinations of the parameters

- $m = n \in \{25, 50, 100, 500\}$
- $F \in \{\text{Normal}, \text{Logistic}, \text{Laplace}\}$
- $\theta \in \{.5, .8\}$
- $\delta \in \{1, 3\}$

based upon a simulation of 1000 data sets per setting.

For each setting, the distribution of F is standardized to have mean 0 and standard deviation 1. Note that $\delta = 1$ represents a small shift in the component distributions which often results in a unimodal mixture while $\delta = 3$ results in a bimodal mixture. Lastly, **Table 3.3** presents average lengths for the asymptotic method.

Parameters			Interval Method			
			Asymptotic	BC_a	Asymptotic	BC_a
F	θ	δ	m = n = 25		m = n = 50	
Normal	0.5	1	0.88	0.77	0.87	0.81
Normal	0.5	3	0.90	0.94	0.93	0.94
Normal	0.8	1	0.94	0.83	0.95	0.80
Normal	0.8	3	0.90	0.92	0.93	0.94
Logistic	0.5	1	0.95	0.76	0.95	0.80
Logistic	0.5	3	0.91	0.95	0.92	0.94
Logistic	0.8	1	0.94	0.82	0.94	0.84
Logistic	0.8	3	0.88	0.91	0.90	0.94
Laplace	0.5	1	0.88	0.78	0.87	0.86
Laplace	0.5	3	0.92	0.94	0.94	0.95
Laplace	0.8	1	0.92	0.81	0.92	0.82
Laplace	0.8	3	0.92	0.93	0.93	0.92
			m = n = 100		m = n = 500	
Normal	0.5	1	0.91	0.84	0.93	0.94
Normal	0.5	3	0.94	0.95	0.95	0.95
Normal	0.8	1	0.95	0.80	0.96	0.92
Normal	0.8	3	0.94	0.94	0.94	0.95
Logistic	0.5	1	0.95	0.86	0.97	0.95
Logistic	0.5	3	0.93	0.95	0.95	0.96
Logistic	0.8	1	0.94	0.80	0.95	0.92
Logistic	0.8	3	0.93	0.94	0.94	0.96
Laplace	0.5	1	0.89	0.85	0.93	0.95
Laplace	0.5	3	0.94	0.96	0.95	0.95
Laplace	0.8	1	0.94	0.82	0.95	0.93
Laplace	0.8	3	0.93	0.93	0.95	0.94

Table 3.1: Coverage Probabilities of Asymptotic and BC_a 95% Confidence Intervals for θ. Simulated coverage estimates have margin of error ranging from .01 to .03 at 99% confidence depending on coverage. For all F, $\sigma_X = 1$.

Parameters			Interval Method			
			Asymptotic	BC_a	Asymptotic	BC_a
F	θ	δ	m = n = 25		m = n = 50	
Normal	0.5	1	0.99*	0.88	0.99	0.87
Normal	0.5	3	0.96	0.93	0.95	0.94
Normal	0.8	1	0.96	0.91	0.97	0.93
Normal	0.8	3	0.94	0.92	0.96	0.94
Logistic	0.5	1	0.98	0.89	0.99	0.87
Logistic	0.5	3	0.94	0.92	0.95	0.94
Logistic	0.8	1	0.95	0.94	0.96	0.92
Logistic	0.8	3	0.94	0.93	0.94	0.94
Laplace	0.5	1	0.98	0.89	0.99	0.88
Laplace	0.5	3	0.95	0.92	0.96	0.93
Laplace	0.8	1	0.97	0.93	0.97	0.92
Laplace	0.8	3	0.94	0.90	0.96	0.94
			m = n = 100		m = n = 500	
Normal	0.5	1	0.99	0.80	0.96	0.94
Normal	0.5	3	0.95	0.94	0.95	0.95
Normal	0.8	1	0.97	0.91	0.96	0.91
Normal	0.8	3	0.94	0.94	0.95	0.95
Logistic	0.5	1	0.99	0.85	0.97	0.95
Logistic	0.5	3	0.95	0.94	0.94	0.96
Logistic	0.8	1	0.96	0.91	0.96	0.90
Logistic	0.8	3	0.94	0.94	0.96	0.95
Laplace	0.5	1	0.99	0.85	0.98	0.94
Laplace	0.5	3	0.94	0.94	0.96	0.94
Laplace	0.8	1	0.98	0.91	0.97	0.92
Laplace	0.8	3	0.95	0.95	0.94	0.96

Table 3.2: Coverage Probabilities of Asymptotic and BC_a 95% Confidence Intervals for δ. Simulated coverage estimates have margin of error ranging from .01 to .03 at 99% confidence depending on coverage. For all F, $\sigma_X = 1$.

Parameters			$CI(\theta)$			
F	θ	δ	25	50	100	500
Normal	0.5	1	.83 (.02)	.78 (.02)	.72 (.02)	.44 (.01)
Normal	0.5	3	.61 (.01)	.45 (.01)	.32 (.00)	.14 (.00)
Normal	0.8	1	.82 (.02)	.74 (.01)	.61 (.01)	.35 (.00)
Normal	0.8	3	.39 (.01)	.30 (.01)	.22 (.00)	.10 (.00)
Logistic	0.5	1	.87 (.01)	.86 (.01)	.83 (.01)	.55 (.01)
Logistic	0.5	3	.55 (.01)	.42 (.01)	.31 (.00)	.14 (.00)
Logistic	0.8	1	.78 (.02)	.75 (.02)	.66 (.01)	.39 (.01)
Logistic	0.8	3	.37 (.01)	.29 (.01)	.22 (.00)	.10 (.00)
Laplace	0.5	1	.83 (.02)	.81 (.02)	.79 (.02)	.59 (.02)
Laplace	0.5	3	.63 (.01)	.47 (.01)	.33 (.00)	.15 (.00)
Laplace	0.8	1	.84 (.02)	.83 (.01)	.76 (.01)	.45 (.01)
Laplace	0.8	3	.40 (.01)	.32 (.01)	.24 (.00)	.11 (.00)
			$CI(\delta)$			
F	θ	δ	25	50	100	500
Normal	0.5	1	2.28 (.08)	1.98 (.06)	1.56 (.04)	0.78 (.01)
Normal	0.5	3	2.15 (.10)	1.38 (.03)	0.97 (.02)	0.42 (.00)
Normal	0.8	1	1.90 (.05)	1.47 (.03)	1.12 (.02)	0.51 (.00)
Normal	0.8	3	1.34 (.02)	0.94 (.01)	0.66 (.00)	0.29 (.00)
Logistic	0.5	1	3.18 (.26)	2.49 (.20)	1.88 (.08)	0.94 (.01)
Logistic	0.5	3	2.07 (.09)	1.47 (.04)	1.02 (.02)	0.45 (.00)
Logistic	0.8	1	2.07 (.10)	1.63 (.04)	1.27 (.02)	0.60 (.01)
Logistic	0.8	3	1.35 (.03)	0.96 (.02)	0.69 (.01)	0.31 (.00)
Laplace	0.5	1	2.95 (.14)	2.54 (.10)	2.05 (.06)	1.14 (.02)
Laplace	0.5	3	2.33 (.11)	1.62 (.06)	1.13 (.02)	0.51 (.00)
Laplace	0.8	1	2.33 (.10)	1.92 (.05)	1.51 (.03)	0.73 (.01)
Laplace	0.8	3	1.47 (.04)	1.06 (.02)	0.76 (.01)	0.34 (.00)

Table 3.3: Simulated Average Length of Asymptotic 95% CIs for θ and δ when $m = n \in \{25, 50, 100, 500\}$. Average interval length estimates have margin of error at 99% confidence as noted in parentheses. For all F, $\sigma_X = 1$.

Table 3.1 shows that the coverage probability of the 95% asymptotic interval for θ is well calibrated except for the case of very small sample sizes $m = n = 25$. However, the BC_a intervals have far too low coverage probabilities for θ when δ is small, even for moderate sample size (e.g. $m = n = 100$) but well-calibrated coverage probabilities for large δ. As the sample sizes increase, both confidence intervals have coverage probabilities converging toward .95 but the asymptotic interval appears to do so more quickly.

Table 3.2 shows that the coverage probability for the 95% asymptotic interval for δ tends to be conservative when the component distributions are not well separated and are fairly well-calibrated otherwise, even for small sample sizes. Contrarily, the BC_a confidence intervals tend to have coverage probabilities that are too low and this is most notable when the components are not well separated. As the sample sizes increase, both methods have coverage probabilities that converge to .95 rather slowly when δ is small. (There was one data set in the **Table 3.2** setting marked with * for which the asymptotic confidence interval could not be computed. This is possible due to the small sample size and the asymptotic nature of the interval).

Table 3.3 shows average lengths for both parameters of the asymptotic confidence intervals, which were shown to have superior coverage probabilities to the BC_a intervals in **Tables 3.1-3.2**. The intervals for θ are notably smaller when the mixture components are well separated. Also, the intervals for δ are notably smaller when the components are well-separated and also when there are more responders. The tables verify that the average confidence interval length decreases at a rate of $n^{-\frac{1}{2}}$ once the sample size is sufficiently large to ensure that truncation at the edge of the parameter space is rare.

3.2 Method of Moment Confidence Regions

While the method of moment confidence intervals described in section 3.1 provide inference for θ and δ individually, neither produces bounds for the full treatment effect (θ, δ). The foundation for the asymptotic intervals is the limiting distributions found in proposition 2.2.3. An analogous result for the distribution of $\widehat{(\theta, \delta)}_{MoM}$ would provide the basis for asymptotic method of moment confidence regions. However, marginal normality does not imply joint normality and thus the asymptotic results in 2.2.3 do not imply that $\widehat{(\theta, \delta)}_{MoM}$ is asymptotically normally distributed. Since little is known about the distribution of $\widehat{(\theta, \delta)}_{MoM}$, $CR_{MoM}(\theta, \delta)$ can be constructed from $CI_{MoM}(\theta), CI_{MoM}(\delta), CI_{MoM}(\Delta)$ by the following methods

- Using a single confidence interval
- Intersecting two confidence intervals

3.2.1 Confidence Region from Interval

A $100(1-\alpha)\%$ $CI(\Delta)$ is equivalent to a $100(1-\alpha)\%$ $CR(\theta, \delta) = \{(\theta, \delta) : \theta\delta \in CI(\Delta)\}$. Letting Δ_L and Δ_U be the lower and upper bounds of $CI_{MoM}(\Delta)$ respectively, the region can be written in any of the following forms

$$\begin{aligned}
100(1-\alpha)\%\ CR_{MoM\{\Delta\}}(\theta, \delta) &= \{(\theta, \delta) : \Delta_l \leq \theta\delta \leq \Delta_u\} \\
&= \{(\theta, \delta) : \frac{\Delta_l}{\theta} \leq \delta \leq \frac{\Delta_u}{\theta}\} \\
&= \{(\theta, \delta) : \frac{\Delta_l}{\delta} \leq \theta \leq \frac{\Delta_u}{\delta}\}.
\end{aligned}$$

Figure 3.1 below displays this confidence region on a data set with 100 observations per group.

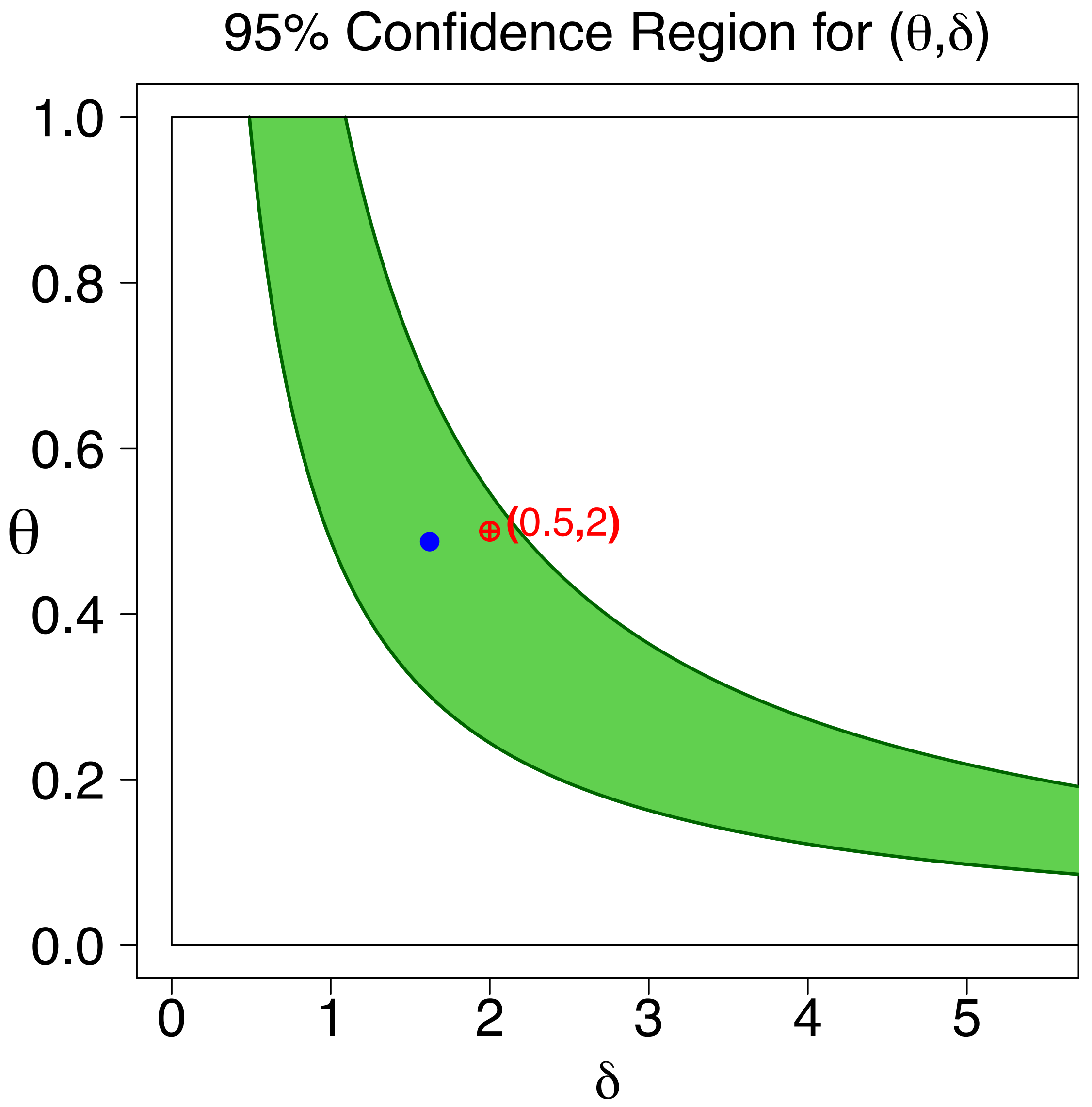


Figure 3.1: Data set of size $m = n = 100$ generated from $F \sim N(0, 1)$ and $(\theta, \delta) = (0.5, 2)$, shown as the red bulls-eye on the plot. The dark-green curves are found from 95% $CI_{MoM}(\Delta) = [0.49, 1.09]$ onto (θ, δ). The blue dot represents the point estimate $\widehat{(\theta, \delta)}_{MoM} = (0.49, 1.62)$ and the light green shaded region is the 95% $CR_{MoM\Delta}(\theta, \delta)$.

3.2.2 Confidence Regions via Intersecting Confidence Intervals

Another way to obtain method of moment confidence regions for (θ, δ) is by intersecting two confidence intervals. For example, a confidence region can be found by intersecting $CI(\theta)$ with $CI(\delta)$ such that

$$CR_{\{\theta,\delta\}}(\theta, \delta) = \{(\theta, \delta) : \theta \in CI(\theta) \cap \delta \in CI(\delta)\}.$$

Similarly, intersecting $CI(\theta)$ with $CI(\Delta)$ produces

$$CR_{\{\theta,\Delta\}}(\theta, \delta) = \{(\theta, \delta) : \theta \in CI(\theta) \ \cap \ \theta\delta \in CI(\Delta)\}$$

and intersecting $CI(\delta)$ with $CI(\Delta)$ produces

$$CR_{\{\delta,\Delta\}}(\theta, \delta) = \{(\theta, \delta) : \delta \in CI(\delta) \ \cap \ \theta\delta \in CI(\Delta)\}.$$

The confidence levels of the confidence intervals can be selected to achieve a conservative confidence region for a corresponding nominal level. The confidence region fails to capture (θ, δ) if at least one interval fails to do so. Let α_{CR} represent the probability that the confidence region fails to capture (θ, δ). Let $\alpha_{CI\delta}$ and $\alpha_{CI\Delta}$ be the probabilities that $CI_{MoM}(\delta)$ and $CI_{MoM}(\Delta)$ fail to capture the true parameter, respectively, and let $\alpha_{CI\delta,CI\Delta}$ be the probability that both fail to capture the true parameter. Then

$$\begin{aligned} \alpha_{CR} &= \alpha_{CI\delta} + \alpha_{CI\Delta} - \alpha_{CI\delta,CI\Delta} \\ &\leq \alpha_{CI\delta} + \alpha_{CI\Delta}. \end{aligned} \tag{3.8}$$

Then a nominal $100(1-\alpha'_{CR})\%$ $CR(\theta,\delta)$ with conservative probability, $(1-\alpha_{CR}) \geq (1-\alpha'_{CR})$, can be obtained by selecting $\alpha_{CI\delta} = \alpha_{CI\Delta} = \alpha'_{CR}/2$. Section 4.2 (in the next chapter) illustrates that $CI_{MoM}(\theta)$ frequently suffers from lower than nominal coverage probability, which motivates selecting $CI_{MoM}(\Delta)$ and $CI_{MoM}(\delta)$ to intersect for a method of moments confidence region for (θ,δ) as described above. **Figure 3.2** below illustrates this confidence region on a data set with 100 observations per group.

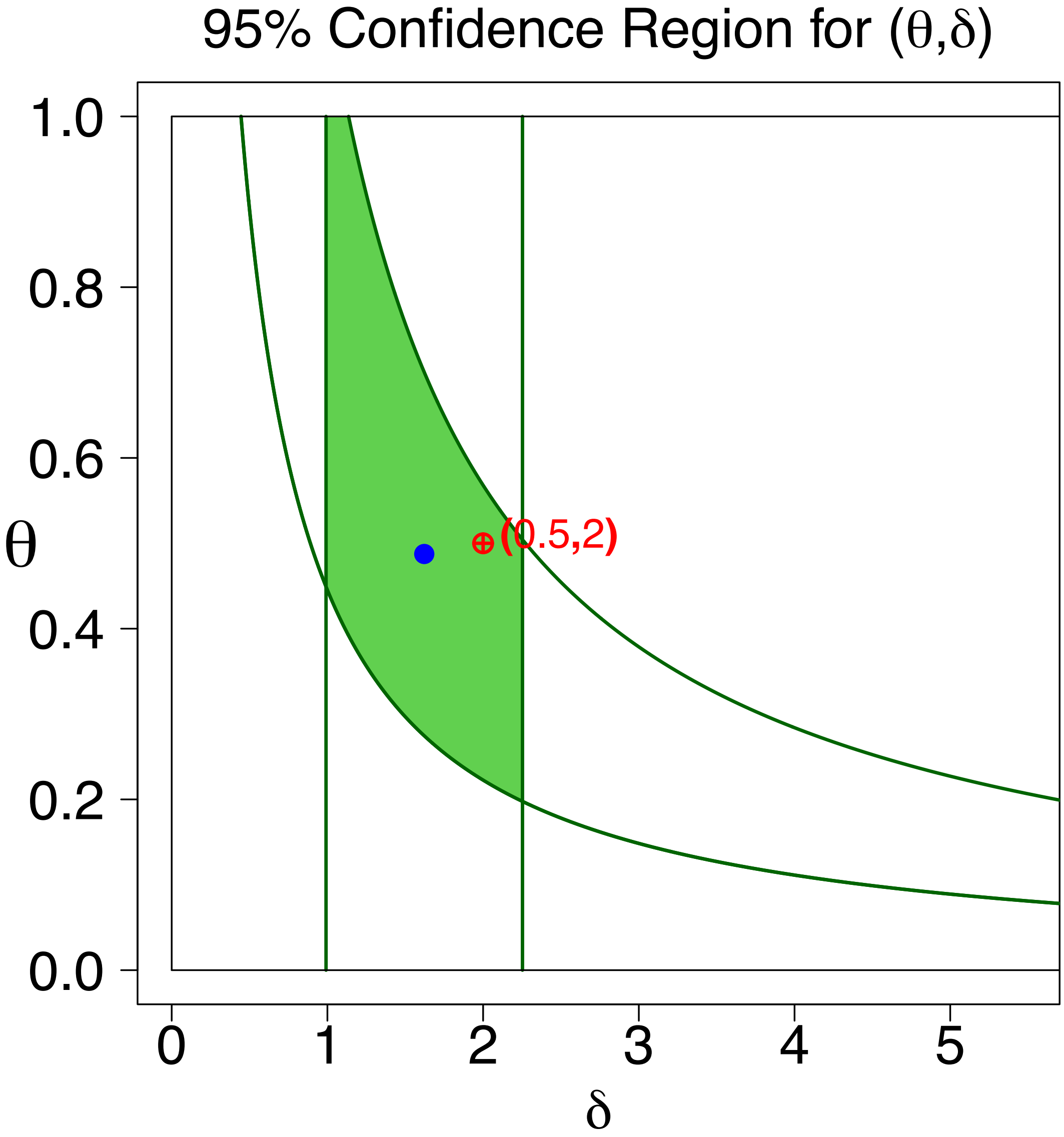


Figure 3.2: Data set of size $m = n = 100$ generated from $F \sim N(0, 1)$ and $(\theta, \delta) = (0.5, 2)$, shown as the red bulls-eye on the plot. The dark-green curves correspond to the bounds of 97.5% $CI_{MoM}(\Delta) = [0.45, 1.14]$ and the vertical dark-green lines correspond to the bounds of 97.5% $CI_{MoM}(\delta) = [0.99, 2.25]$. The blue dot represents the point estimate $\widehat{(\theta, \delta)}_{MoM} = (0.49, 1.62)$ and the light green shaded region is the 95% $CR_{MoM\{\delta,\Delta\}}(\theta, \delta)$.

3.3 Pseudo-Likelihood Confidence Regions

A confidence region for (θ, δ) can be found by inverting a hypothesis test using the likelihood ratio test statistic. Consider first a scenario in which f is known and recall that the likelihood of (1.1) is given by (2.18). Under H_0: $(\theta, \delta) = (\theta_0, \delta_0)$, the likelihood ratio test statistic is given by

$$T_{(\theta_0,\delta_0)} = -2\left[logL\left(\theta_0, \delta_0; X, Y\right) - logL\left(\widehat{\theta}, \widehat{\delta}; X, Y\right)\right]. \tag{3.9}$$

Under the null hypothesis, $T_{(\theta_0,\delta_0)}$ has an asymptotic distribution that is $\chi^2_{df=2}$ (Wilks, 1938) since the hypothesized point null is two-dimensional, so long as (θ_0, δ_0) does not lie on the boundary of the parameter space. The results from Self and Liang (1987) show that if $\theta_0 = 1, \delta_0 > 0$ then $T_{(1,\delta_0)} \sim .5\chi^2_{df=1} + .5\chi^2_{df=2}$ and if $(\theta_0, \delta_0) = (0, 0)$ then the asymptotic distribution of $T_{(0,0)}$ is unknown. Thus, for any true $(\theta_0, \delta_0) \neq (0, 0)$, an α level hypothesis test (asymptotically) fails to reject with probability $(1 - \alpha)$ based on the rejection rule (0 = Fail to Reject, 1 = Reject)

$$\lambda^{\alpha}_{(\theta_0,\delta_0)}(X, Y) = \begin{cases} I\left(T_{(\theta_0,\delta_0)} \geq \chi^2_{2,1-\alpha}\right), & \text{for } \theta_0 \in (0, 1) \\ I\left(T_{(\theta_0,\delta_0)} \geq .5\chi^2_{1,1-\alpha} + .5\chi^2_{2,1-\alpha}\right) & \text{for } \theta_0 = 1. \end{cases} \tag{3.10}$$

Therefore, when a treatment effect exists, a confidence region constructed by

$$100(1-\alpha)\%\ CR_{Lik}(\theta, \delta) = \left\{(\theta, \delta) : \lambda^{\alpha}_{(\theta_0,\delta_0)}(X, Y) = 0\right\} \tag{3.11}$$

has asymptotic coverage probability $(1 - \alpha)$.

However, this confidence region relies upon knowledge of f in (2.18) but in reality, f is unknown. Thus, consider a modification of the confidence region derived in (3.11) by using the pseudo-likelihood (2.19) which plugs in an estimate for f. The analogous likelihood ratio test statistic computed from the pseudo-likelihood for testing H_0: $(\theta, \delta) = (\theta_0, \delta_0)$ is given by

$$\widehat{T}_{(\theta_0,\delta_0)} = -2\left[log\widehat{L}\left(\theta_0, \delta_0; X, Y\right) - log\widehat{L}\left(\widehat{\theta}, \widehat{\delta}; X, Y\right)\right]. \tag{3.12}$$

The results in Liang and Self (1996) and Chen and Liang (2010) suggest that under H_0: $(\theta, \delta) = (\theta_0, \delta_0) \neq (0, 0)$, $\widehat{T}_{(\theta_0,\delta_0)}$ has the same asymptotic distribution as $T_{(\theta_0,\delta_0)}$ so long as the following conditions hold

- $\widehat{f}$ is a consistent estimate of f
- $\lim\limits_{m\to\infty, n\to\infty} n/m = 0$.

Therefore, to improve performance when n and m are both finite with $n/m > 0$, a Satterthwaite approximation is used to model the distribution of $\widehat{T}_{(\theta_0,\delta_0)}$. That is

$$\widehat{T}_{(\theta_0,\delta_0)} \dot{\sim} c_1\chi^2_{d_1}, \tag{3.13}$$

where c_1 and d_1 are functions of m and n that converge to 1 and 2 respectively as $m \to \infty$, $n \to \infty$ and $n/m \to 0$. Thus, the proposed pseudo-likelihood confidence region for capturing the sub-population specific treatment effect (θ, δ) is

$$100(1-\alpha)\%\ CR_{PsL}(\theta, \delta) = \left\{(\theta, \delta) : \widehat{T}_{(\theta,\delta)} < c_1\chi^2_{d_1,1-\alpha}\right\}, \tag{3.14}$$

which has asymptotic coverage probability $(1 - \alpha)$ as $\widehat{f} \to f$, $m \to \infty$, $n \to \infty, n/m \to 0$. For simplicity, this cutoff $(c_1\chi^2_{d_1,1-\alpha})$ is used for all θ (including $\theta = 1$) unless sample sizes are sufficiently large to apply the asymptotic rule from (3.10). In practice, this is done using a dense grid search over a select set of (θ, δ) as described in section 2.4. See section A.10 of the Appendix for details.

The quantities c_1 and d_1 are determined from a large scale simulation by generating 1000 data sets under each following factorial combinations of settings in **List 3.1**.

- $N \in \{60, 120, 180, 300, 600, 1200, 2400, 4800\}$
- $n{:}m \in \{1{:}29, 1{:}19, 1{:}14, 1{:}9, 1{:}5, 1{:}3, 1{:}2, 2{:}3, 1{:}1, 3{:}2, 2{:}1, 3{:}1\}$
- $F \in \{\text{Normal, Laplace, SkRNorm, SkRLap, SkLNorm, SkLLap}\}$
- $\theta \in \{.2, .5, .8\}$
- $\delta \in \{.5, 1, 2, 3\}$

List 3.1: Large Scale Simulation Settings. There are 96 unique sample size pairs and 72 unique (F, θ, δ) triples for a total of 6912 combinations of $(N, n{:}m, F, \theta, \delta)$.

All 72 combinations of (F, θ, δ) are used to determine c_1 and d_1 for each pair of $(N, n/m)$. Specifically, each of the 72 settings produces 1000 realizations of $\widehat{T}_{(\theta_0,\delta_0)}$. Method of moment estimates (3.15) of c_1 and d_1 are calculated from these 72,000 realizations for each $(N, n/m)$.

The derivation of the estimates is provided below.

$$\text{If } X \sim \chi^2_{df=d_1} \implies E[X] = d_1,\ Var(X) = 2d_1.$$

Therefore, since the Satterthwaite approximation indicates $\widehat{T}_{(\theta_0,\delta_0)} \overset{d}{=} c_1 X$

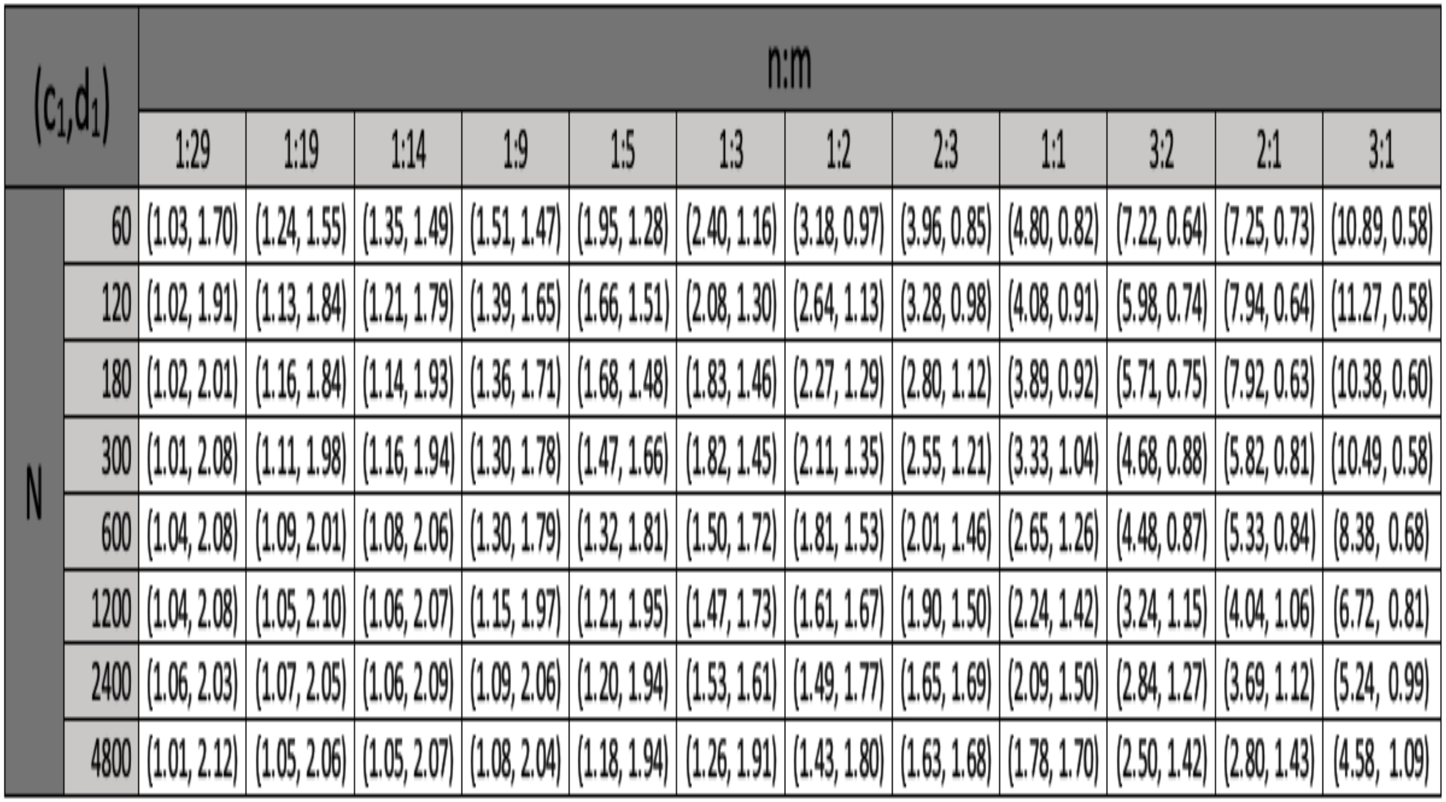
$$E[\widehat{T}_{(\theta_0,\delta_0)}] = c_1 d_1,\ Var(\widehat{T}_{(\theta_0,\delta_0)}) = 2c_1^2 d_1.$$

Let $\overline{T}_{(\theta_0,\delta_0)}$ and $S^2_{\widehat{T}_{(\theta_0,\delta_0)}}$ be the mean and variance of the 72,000 simulated $\widehat{T}_{(\theta_0,\delta_0)}$, respectively. For the method of moment estimates,

$$\begin{aligned}
&\overline{T}_{(\theta_0,\delta_0)} \overset{set}{=} c_1 d_1, \quad S^2_{\widehat{T}_{(\theta_0,\delta_0)}} \overset{set}{=} 2c_1^2 d_1 \\
&\implies c_1 = \frac{\overline{T}_{(\theta_0,\delta_0)}}{d_1} \to S^2_{\widehat{T}_{(\theta_0,\delta_0)}} = 2\frac{\overline{T}^2_{(\theta_0,\delta_0)}}{d_1^2} d_1 \\
&\implies d_1 = \frac{2\overline{T}^2_{(\theta_0,\delta_0)}}{S^2_{\widehat{T}_{(\theta_0,\delta_0)}}} \to c_1 = \overline{T}_{(\theta_0,\delta_0)} \frac{S^2_{\widehat{T}_{(\theta_0,\delta_0)}}}{2\overline{T}^2_{(\theta_0,\delta_0)}}.
\end{aligned}$$

Therefore,

$$c_1 = \frac{S^2_{\widehat{T}_{(\theta_0,\delta_0)}}}{2\overline{T}_{(\theta_0,\delta_0)}}, \qquad d_1 = \frac{2\overline{T}^2_{(\theta_0,\delta_0)}}{S^2_{\widehat{T}_{(\theta_0,\delta_0)}}}. \tag{3.15}$$



(c_1,d_1)		n:m											
		1:29	1:19	1:14	1:9	1:5	1:3	1:2	2:3	1:1	3:2	2:1	3:1
N	60	(1.03, 1.70)	(1.24, 1.55)	(1.35, 1.49)	(1.51, 1.47)	(1.95, 1.28)	(2.40, 1.16)	(3.18, 0.97)	(3.96, 0.85)	(4.80, 0.82)	(7.22, 0.64)	(7.25, 0.73)	(10.89, 0.58)
	120	(1.02, 1.91)	(1.13, 1.84)	(1.21, 1.79)	(1.39, 1.65)	(1.66, 1.51)	(2.08, 1.30)	(2.64, 1.13)	(3.28, 0.98)	(4.08, 0.91)	(5.98, 0.74)	(7.94, 0.64)	(11.27, 0.58)
	180	(1.02, 2.01)	(1.16, 1.84)	(1.14, 1.93)	(1.36, 1.71)	(1.68, 1.48)	(1.83, 1.46)	(2.27, 1.29)	(2.80, 1.12)	(3.89, 0.92)	(5.71, 0.75)	(7.92, 0.63)	(10.38, 0.60)
	300	(1.01, 2.08)	(1.11, 1.98)	(1.16, 1.94)	(1.30, 1.78)	(1.47, 1.66)	(1.82, 1.45)	(2.11, 1.35)	(2.55, 1.21)	(3.33, 1.04)	(4.68, 0.88)	(5.82, 0.81)	(10.49, 0.58)
	600	(1.04, 2.08)	(1.09, 2.01)	(1.08, 2.06)	(1.30, 1.79)	(1.32, 1.81)	(1.50, 1.72)	(1.81, 1.53)	(2.01, 1.46)	(2.65, 1.26)	(4.48, 0.87)	(5.33, 0.84)	(8.38, 0.68)
	1200	(1.04, 2.08)	(1.05, 2.10)	(1.06, 2.07)	(1.15, 1.97)	(1.21, 1.95)	(1.47, 1.73)	(1.61, 1.67)	(1.90, 1.50)	(2.24, 1.42)	(3.24, 1.15)	(4.04, 1.06)	(6.72, 0.81)
	2400	(1.06, 2.03)	(1.07, 2.05)	(1.06, 2.09)	(1.09, 2.06)	(1.20, 1.94)	(1.53, 1.61)	(1.49, 1.77)	(1.65, 1.69)	(2.09, 1.50)	(2.84, 1.27)	(3.69, 1.12)	(5.24, 0.99)
	4800	(1.01, 2.12)	(1.05, 2.06)	(1.05, 2.07)	(1.08, 2.04)	(1.18, 1.94)	(1.26, 1.91)	(1.43, 1.80)	(1.63, 1.68)	(1.78, 1.70)	(2.50, 1.42)	(2.80, 1.43)	(4.58, 1.09)

Table 3.4: Table of all (c_1, d_1) pairs according to the setting of sample sizes $(N, n/m)$.

Table 3.4 illustrates that the Satterthwaite approximation is necessary, as small or treatment-heavy sample size settings indicate (c_1, d_1) values far from the asymptotic result, $(c_1, d_1) \to (1, 2)$. The table also demonstrates the convergence of $\widehat{T}_{(\theta_0,\delta_0)}$ to χ^2_2 as the Satterthwaite constants in the lower-left corner approach $(1, 2)$.

The function psl.inf() (which can be found in section A.10 of the Appendix) implements this Satterthwaite approximation with bi-linear interpolation for sample sizes $(N', n'/m')$ that are not identical to any of the above listed simulation settings. If the sample sizes are such that extrapolation is necessary, a warning is given (e.g. $m = 400, n = 2000$). For appropriately large sample sizes ($m > 4640$, $n > 160$, $n/m < 1/29$) the asymptotic cutoffs from (3.10) are used to determine $CR_{PsL}(\theta, \delta)$.

Figure 3.3 below illustrates this confidence region on a data set with 100 observations per group. The confidence region has an oval-like shape and captures the true parameter from which the data was simulated.

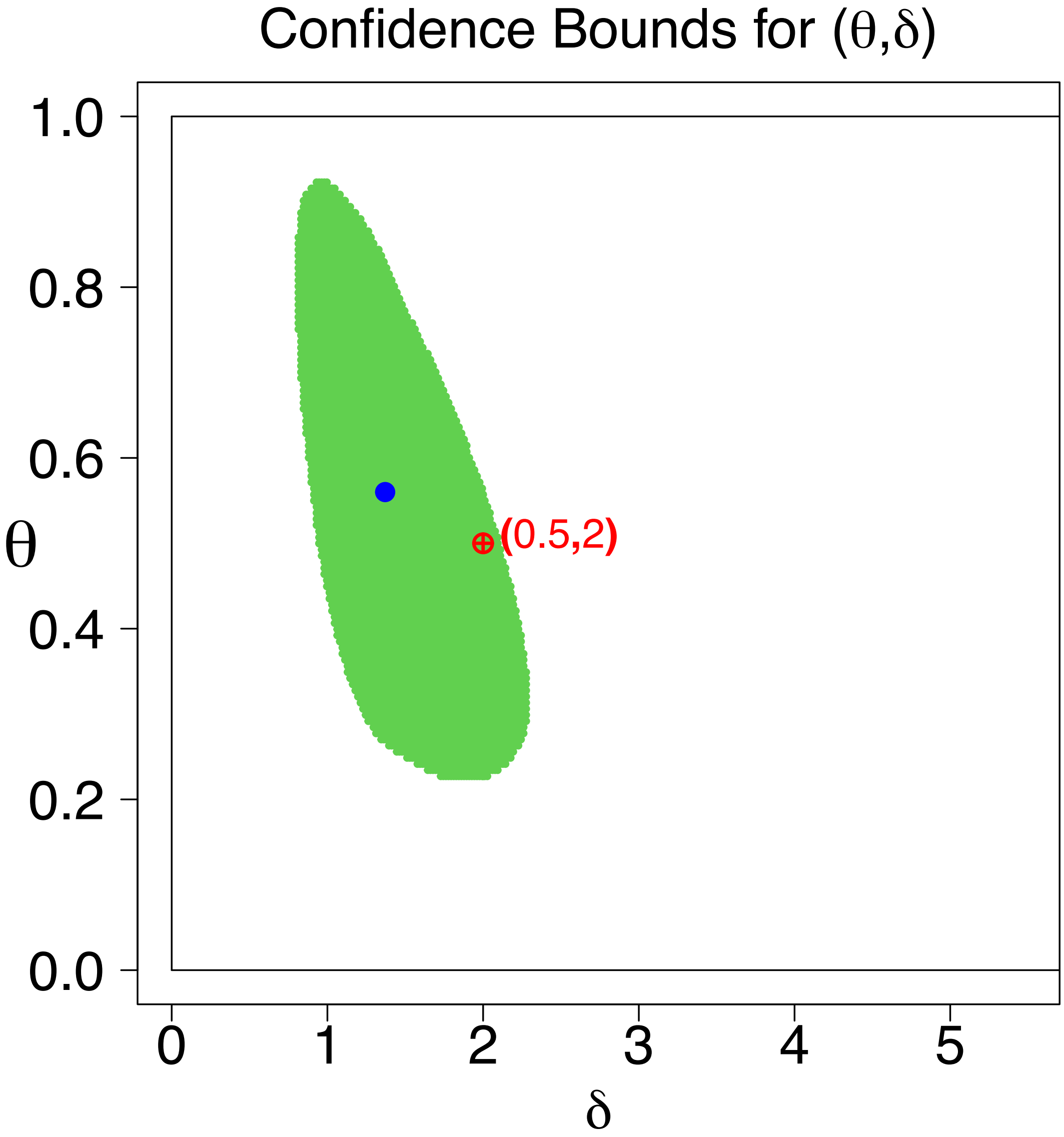


Figure 3.3: Data set of size $m = n = 100$ generated from $F \sim N(0, 1)$ and $(\theta, \delta) = (0.5, 2)$, shown as the red bulls-eye on the plot. The blue dot represents the point estimate $\widehat{(\theta, \delta)}_{PsL} = (0.56, 1.37)$ and the light green shaded region is the 95% $CR_{PsL}(\theta, \delta)$.

3.4 Pseudo-Likelihood Confidence Intervals

Consider constructing confidence intervals that correspond to the pseudo-likelihood estimators in (2.24) by inverting a hypothesis test.

3.4.1 Pseudo-Likelihood Intervals for θ

Because this section considers inference on θ with confidence intervals, it is useful to define profile likelihood for θ. To do so, first consider a scenario in which f is treated as known rather than an unknown parameter.

$$L\left(\theta, \widehat{\delta}(\theta); X, Y\right) = \prod_{j=1}^{m} [f(x_j)] \prod_{i=1}^{n} \left[(1-\theta)f(y_i) + \theta f(y_i - \widehat{\delta}(\theta))\right] \tag{3.16}$$

where $\widehat{\delta}(\theta)$ is the δ that maximizes the likelihood (2.18) for a given θ. For testing H_0: $\theta = \theta_0$, the likelihood ratio test statistic is defined as

$$T_{\theta_0} = -2\left[logL\left(\theta_0, \widehat{\delta}(\theta_0); X, Y\right) - logL\left(\widehat{\theta}, \widehat{\delta}; X, Y\right)\right], \tag{3.17}$$

where T_{θ_0} asymptotically follows a chi-square distribution with 1 degree of freedom so long as $\theta \in (0,1)$ (Wilks, 1938). The results from Self and Liang (1987) show that when $\theta_0 = 1$, $T_{\theta_0} \sim .5\chi^2_0 + .5\chi^2_1$ and when $(\theta_0, \delta_0) = (0,0)$ the distribution is unknown. Therefore, under H_0: $\theta = \theta_0 (\neq 0)$, an α level hypothesis test (asymptotically) fails to reject with probability $(1-\alpha)$ based on the rejection rule (0 = Fail to Reject, 1 = Reject)

$$\lambda^{\alpha}_{\theta_0}(X,Y) = \begin{cases} I\left(T_{\theta_0} \geq \chi^2_{1,1-\alpha}\right), & \text{for } \theta_0 \in (0,1) \\ I\left(T_{\theta_0} \geq .5\chi^2_{0,1-\alpha} + .5\chi^2_{1,1-\alpha}\right) & \text{for } \theta_0 = 1. \end{cases} \tag{3.18}$$

Thus, a $100(1-\alpha)\%$ confidence set for θ defined by the set of all $\theta \in (0,1]$ such that $\lambda_\theta(X,Y) = 0$ has asymptotic coverage probability $1-\alpha$.

However, since f is unknown, the likelihood (2.18) and corresponding profile likelihood (3.16) cannot be used. Consider a similar procedure that instead uses the pseudo-likelihood (2.19), which substitutes an estimate for f. The pseudo-profile likelihood function for θ is defined as

$$\widehat{L}\left(\theta, \widehat{\delta}(\theta); X, Y\right) = \prod_{j=1}^{m}\left[\widehat{f}(x_j)\right]\prod_{i=1}^{n}\left[(1-\theta)\widehat{f}(y_i) + \theta\widehat{f}(y_i - \widehat{\delta}(\theta))\right] \tag{3.19}$$

where $\widehat{\delta}(\theta)$ is the δ that maximizes the pseudo-likelihood (2.19) for a given θ. The pseudo-likelihood ratio test statistic for H_0: $\theta = \theta_0$ is defined as

$$\widehat{T}_{\theta_0} = -2\left[log\widehat{L}\left(\theta_0, \widehat{\delta}(\theta_0); X, Y\right) - log\widehat{L}\left(\widehat{\theta}, \widehat{\delta}; X, Y\right)\right], \tag{3.20}$$

The results from Liang and Self (1996) and Chen and Liang (2010) suggest that $\widehat{T}_{\theta_0}$ has the same asymptotic distribution as T_{θ_0} so long as the following conditions hold

- $\widehat{f}$ is a consistent estimate of f
- $\lim\limits_{m\to\infty, n\to\infty} n/m = 0.$

Therefore, to improve performance when n and m are finite with $n/m > 0$, a Satterthwaite approximation is used to model the distribution of $\widehat{T}_{\theta_0}$. That is,

$$\widehat{T}_{\theta_0} \dot{\sim} c_2\chi^2_{d_2,1-\alpha} \tag{3.21}$$

where c_2 and d_2 are functions of m and n that converge to 1 as $m \to \infty$, $n \to \infty$ and $n/m \to 0$. Thus, a pseudo-likelihood confidence set for θ is given by

$$100(1-\alpha)\%\ CSet_{PsL}(\theta) = \left\{\theta : \widehat{T}_\theta < c_2\chi^2_{d_2,1-\alpha}\right\}. \tag{3.22}$$

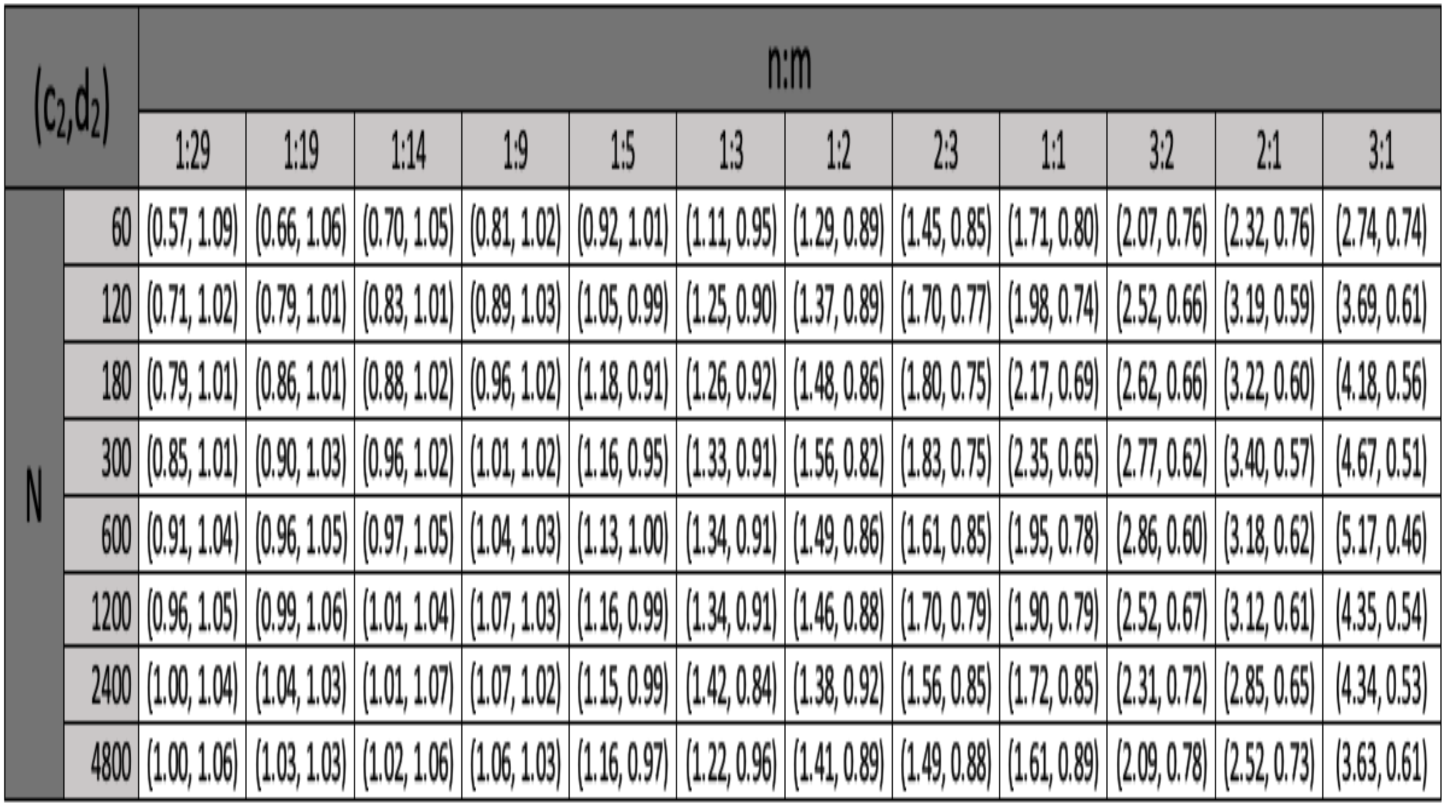
For simplicity, this cutoff ($c_2\chi^2_{d_2,1-\alpha}$) is used for all θ (including $\theta = 1$) unless sample sizes are sufficiently large to apply the asymptotic rule from (3.18). See section A.10 of the Appendix for details.

Since (3.22) is not guaranteed to be an interval, a confidence interval can be defined by

$$100(1-\alpha)\%\ CI_{PsL}(\theta) = [\min CSet_{PsL}(\theta),\ \max CSet_{PsL}(\theta)] \tag{3.23}$$

While the coverage probability of $100(1-\alpha)\% CI_{PsL}(\theta)$ is at least as large as that of $100(1-\alpha)\% CSet_{PsL}(\theta)$, the discrepancy is very minor as a large scale simulation (see chapter 4) indicates that $CSet_{PsL}(\theta)$ is identical to $CI_{PsL}(\theta)$ in 99.68% of data sets - with minor discrepancies when not identical. The quantities c_2 and d_2 are determined from a large scale simulation by generating 1000 data sets under each factorial combinations of settings in **List 3.1**. All 72 combinations of (F, θ, δ) are used to determine c_2 and d_2 for each of the 96 pairs $(N, n/m)$ in the simulation. As illustrated in section 3.3, the 72,000 data sets are used to generate $\widehat{T}_{\theta_0}$ while their corresponding c_2 and d_2 values are computed using the following formulas

$$c_2 = \frac{S^2_{\widehat{T}_{\theta_0}}}{2\overline{T}_{\theta_0}}, \qquad d_2 = \frac{2\overline{T}^2_{\theta_0}}{S^2_{\widehat{T}_{\theta_0}}}. \tag{3.24}$$



(c_2,d_2)		n:m											
		1:29	1:19	1:14	1:9	1:5	1:3	1:2	2:3	1:1	3:2	2:1	3:1
N	60	(0.57, 1.09)	(0.66, 1.06)	(0.70, 1.05)	(0.81, 1.02)	(0.92, 1.01)	(1.11, 0.95)	(1.29, 0.89)	(1.45, 0.85)	(1.71, 0.80)	(2.07, 0.76)	(2.32, 0.76)	(2.74, 0.74)
	120	(0.71, 1.02)	(0.79, 1.01)	(0.83, 1.01)	(0.89, 1.03)	(1.05, 0.99)	(1.25, 0.90)	(1.37, 0.89)	(1.70, 0.77)	(1.98, 0.74)	(2.52, 0.66)	(3.19, 0.59)	(3.69, 0.61)
	180	(0.79, 1.01)	(0.86, 1.01)	(0.88, 1.02)	(0.96, 1.02)	(1.18, 0.91)	(1.26, 0.92)	(1.48, 0.86)	(1.80, 0.75)	(2.17, 0.69)	(2.62, 0.66)	(3.22, 0.60)	(4.18, 0.56)
	300	(0.85, 1.01)	(0.90, 1.03)	(0.96, 1.02)	(1.01, 1.02)	(1.16, 0.95)	(1.33, 0.91)	(1.56, 0.82)	(1.83, 0.75)	(2.35, 0.65)	(2.77, 0.62)	(3.40, 0.57)	(4.67, 0.51)
	600	(0.91, 1.04)	(0.96, 1.05)	(0.97, 1.05)	(1.04, 1.03)	(1.13, 1.00)	(1.34, 0.91)	(1.49, 0.86)	(1.61, 0.85)	(1.95, 0.78)	(2.86, 0.60)	(3.18, 0.62)	(5.17, 0.46)
	1200	(0.96, 1.05)	(0.99, 1.06)	(1.01, 1.04)	(1.07, 1.03)	(1.16, 0.99)	(1.34, 0.91)	(1.46, 0.88)	(1.70, 0.79)	(1.90, 0.79)	(2.52, 0.67)	(3.12, 0.61)	(4.35, 0.54)
	2400	(1.00, 1.04)	(1.04, 1.03)	(1.01, 1.07)	(1.07, 1.02)	(1.15, 0.99)	(1.42, 0.84)	(1.38, 0.92)	(1.56, 0.85)	(1.72, 0.85)	(2.31, 0.72)	(2.85, 0.65)	(4.34, 0.53)
	4800	(1.00, 1.06)	(1.03, 1.03)	(1.02, 1.06)	(1.06, 1.03)	(1.16, 0.97)	(1.22, 0.96)	(1.41, 0.89)	(1.49, 0.88)	(1.61, 0.89)	(2.09, 0.78)	(2.52, 0.73)	(3.63, 0.61)

Table 3.5: Table of all (c_2, d_2) pairs according to the setting of sample sizes $(N, n/m)$.

Table 3.5 illustrates that the Satterthwaite approximation is necessary, as small or treatment-heavy sample size settings indicate (c_2, d_2) values far from the asymptotic result, $(c_2, d_2) \to (1, 1)$. The table also demonstrates the convergence of $\widehat{T}_{\theta_0}$ to χ^2_1 as the Satterthwaite constants in the lower-left corner approach $(1, 1)$.

The function psl.inf() (which can be found in section A.10 of the Appendix) implements this Satterthwaite approximation with bi-linear interpolation for sample sizes $(N', n'/m')$ that are not identical to any of the above listed simulation settings. If the sample sizes are such that extrapolation is necessary, a warning is given (e.g. $m = 400, n = 2000$). For appropriately large sample sizes ($m > 4640$, $n > 160$, $n/m < 1/29$), the asymptotic cutoffs from (3.18) are used to determine $CI(\theta)$.

3.4.2 Pseudo-Likelihood Intervals for δ

Because this section considers inference on δ with a confidence interval, it is useful to define the profile likelihood for δ. To do so, first consider a scenario in which f is treated as known rather than an unknown parameter. The profile likelihood for δ is defined as

$$L\left(\widehat{\theta}(\delta), \delta; X, Y\right) = \prod_{j=1}^{m} [f(x_j)] \prod_{i=1}^{n} \left[(1 - \widehat{\theta}(\delta))f(y_i) + \widehat{\theta}(\delta)f(y_i - \delta)\right], \tag{3.25}$$

where $\widehat{\theta}(\delta)$ is the θ that maximizes the likelihood (2.18) for a given δ.

For testing H_0: $\delta = \delta_0$, the likelihood ratio test statistic is

$$T_{\delta_0} = -2\left[logL\left(\widehat{\theta}(\delta_0), \delta_0; X, Y\right) - logL\left(\widehat{\theta}, \widehat{\delta}; X, Y\right)\right]. \tag{3.26}$$

The profile likelihood ratio test statistic (3.26) asymptotically follows a chi-square distribution with 1 degree of freedom so long as $\delta > 0$ and $\theta < 1$. When $\theta = 1$, T_{δ_0}

has a complicated asymptotic distribution, D_N, described in section 2.3 of Chen and Liang (2010). The distribution of T_{δ_0} is unknown when $(\theta_0, \delta_0) = (0, 0)$. Therefore, under H_0: $\delta = \delta_0 (\neq 0)$, an α level hypothesis test (asymptotically) fails to reject with probability $(1 - \alpha)$ based on the rejection rule

$$\lambda_{\delta_0}^{\alpha}(X, Y) = \begin{cases} I\left(T_{\delta_0} \geq \chi^2_{1,1-\alpha}\right), & \text{for } \theta \in (0, 1) \\ I\left(T_{\delta_0} \geq D_{N,1-\alpha}\right) & \text{for } \theta = 1. \end{cases} \tag{3.27}$$

Thus, a $100(1 - \alpha)\%$ confidence set for δ defined by the set of all $\delta > 0$ such that $\lambda_\delta(X, Y) = 0$ has asymptotic coverage probability $(1 - \alpha)$ for all $\delta > 0$.

However, since f is unknown, the likelihood (2.18) and corresponding profile likelihood (3.25) cannot be used. Consider a similar procedure that instead uses the pseudo-likelihood (2.19), which substitutes an estimate for f. Recall from (2.19) the pseudo-likelihood is given by

$$\begin{aligned} \widehat{L}(\theta, \delta; X, Y) &= L(\widehat{f}, \theta, \delta; X, Y) \\ &= \prod_{j=1}^{m} \left[\widehat{f}(x_j)\right] \prod_{i=1}^{n} \left[(1 - \theta)\widehat{f}(y_i) + \theta \widehat{f}(y_i - \delta)\right]. \end{aligned}$$

In a similar fashion, the pseudo-profile likelihood function for δ is defined as

$$\widehat{L}\left(\widehat{\theta}(\delta), \delta; X, Y\right) = \prod_{j=1}^{m} \left[\widehat{f}(x_j)\right] \prod_{i=1}^{n} \left[(1 - \widehat{\theta}(\delta))\widehat{f}(y_i) + \widehat{\theta}(\delta)\widehat{f}(y_i - \delta)\right], \tag{3.28}$$

where $\widehat{\theta}(\delta)$ is the θ that maximizes (2.19) for a fixed δ. The pseudo-likelihood ratio test statistic for H_0: $\delta = \delta_0$ is defined as

$$\widehat{T}_{\delta_0} = -2\left[log\widehat{L}\left(\widehat{\theta}(\delta_0), \delta_0; X, Y\right) - log\widehat{L}\left(\widehat{\theta}, \widehat{\delta}; X, Y\right)\right]. \tag{3.29}$$

The results from Liang and Self (1996) and Chen and Liang (2010) suggest that $\widehat{T}_{\delta_0}$ has the same asymptotic distribution as T_{δ_0} so long as the following conditions hold

- $\widehat{f}$ is a consistent estimate of f
- $\lim\limits_{m\to\infty, n\to\infty} n/m = 0$.

Therefore, to improve performance when n and m are finite with $n/m > 0$, a Satterthwaite approximation is used to model the distribution of $\widehat{T}_{\delta_0}$. That is,

$$\widehat{T}_{\delta_0} \dot{\sim} c_3\chi^2_{d_3,1-\alpha} \tag{3.30}$$

where c_3 and d_3 are functions of m and n that converge to 1 as $m \to \infty$, $n \to \infty$ and $n/m \to 0$. Thus, a pseudo-likelihood confidence set for δ is given by

$$100(1-\alpha)\%\ CSet_{PsL}(\delta) = \left\{\delta : \widehat{T}_\delta < c_3\chi^2_{d_3,1-\alpha}\right\}. \tag{3.31}$$

Given the complicated nature of D_N, this cutoff ($c_3\chi^2_{d_3,1-\alpha}$) is used for all θ (including $\theta = 1$). See section A.10 of the Appendix for details.

Since (3.31) is not guaranteed to be an interval, a confidence interval can be defined by

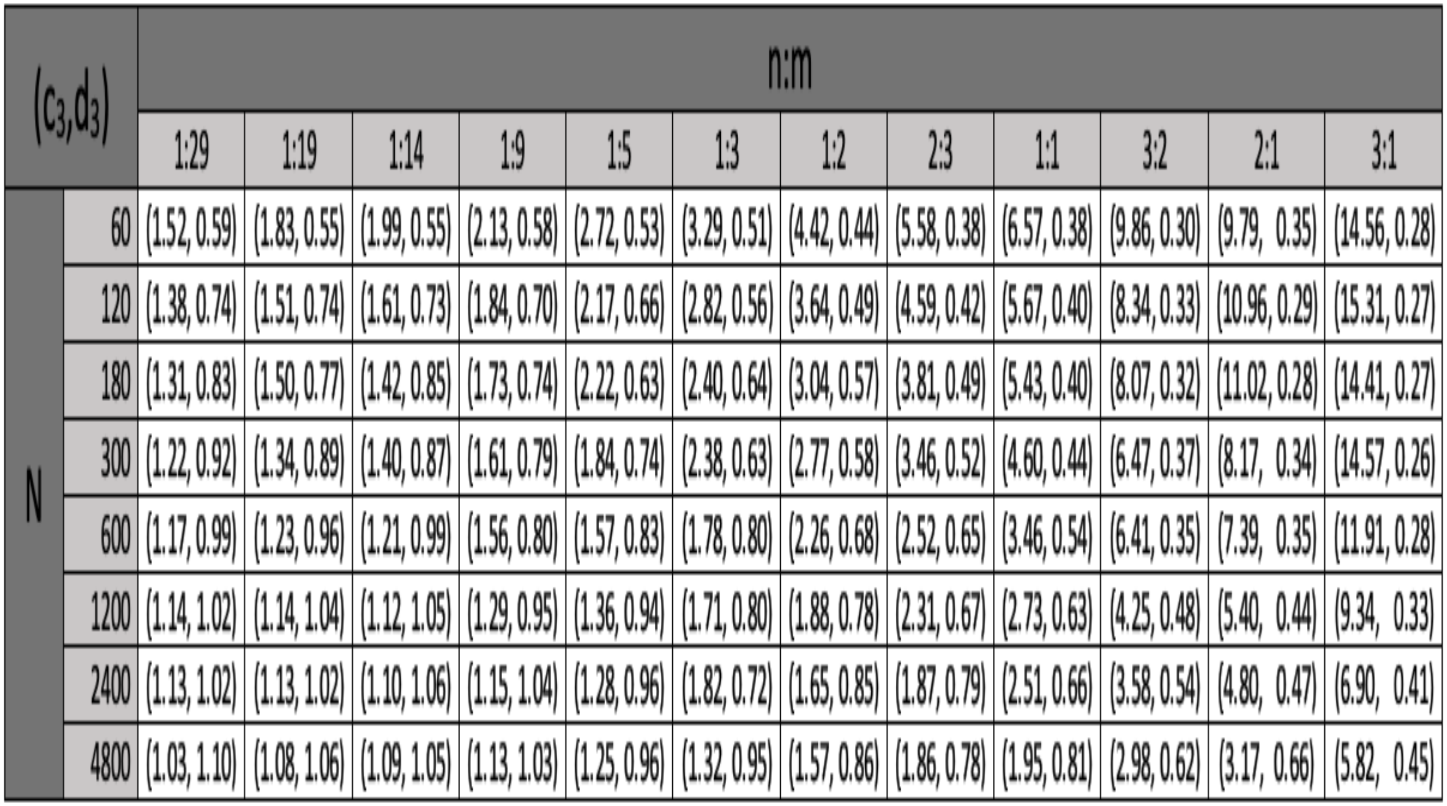
$$100(1-\alpha)\%\ CI_{PsL}(\delta) = \left[\min\, CSet_{PsL}(\delta),\ \max\, CSet_{PsL}(\delta)\right]. \tag{3.32}$$

While the coverage probability of $100(1-\alpha)\%\ CI_{PsL}(\delta)$ is at least as large as that of $100(1-\alpha)\%\ CSet_{PsL}(\delta)$, the discrepancy is very minor as a large scale simulation (see chapter 4) indicates that $CSet_{PsL}(\delta)$ is identical to $CI_{PsL}(\delta)$ in 99.17% of data sets and very similar when not identical. The quantities c_3 and d_3 are determined from a large scale simulation by generating 1000 data sets under each factorial combinations of settings in **List 3.1**. All 72 combinations of (F, θ, δ) are used to determine c_3 and

d_3 for each pair of $(N, n/m)$. As illustrated in section 3.3, the 72,000 data sets are used to generate $\widehat{T}_{\delta_0}$ while their corresponding c_3 and d_3 values are computed using the following formulas

$$c_3 = \frac{S^2_{\widehat{T}_{\delta_0}}}{2\overline{T}_{\delta_0}}, \qquad d_3 = \frac{2\overline{T}^2_{\delta_0}}{S^2_{\widehat{T}_{\delta_0}}}. \tag{3.33}$$

Table 3.6 illustrates that the Satterthwaite approximation is necessary, as small or treatment-heavy sample size settings indicate (c_3, d_3) values far from the asymptotic result, $(c_3, d_3) \rightarrow (1, 1)$. The table also demonstrates the convergence of $\widehat{T}_{\delta_0}$ to χ^2_1 as the Satterthwaite constants in the lower-left corner approach $(1, 1)$.



(c_3,d_3)		n:m											
		1:29	1:19	1:14	1:9	1:5	1:3	1:2	2:3	1:1	3:2	2:1	3:1
N	60	(1.52, 0.59)	(1.83, 0.55)	(1.99, 0.55)	(2.13, 0.58)	(2.72, 0.53)	(3.29, 0.51)	(4.42, 0.44)	(5.58, 0.38)	(6.57, 0.38)	(9.86, 0.30)	(9.79, 0.35)	(14.56, 0.28)
	120	(1.38, 0.74)	(1.51, 0.74)	(1.61, 0.73)	(1.84, 0.70)	(2.17, 0.66)	(2.82, 0.56)	(3.64, 0.49)	(4.59, 0.42)	(5.67, 0.40)	(8.34, 0.33)	(10.96, 0.29)	(15.31, 0.27)
	180	(1.31, 0.83)	(1.50, 0.77)	(1.42, 0.85)	(1.73, 0.74)	(2.22, 0.63)	(2.40, 0.64)	(3.04, 0.57)	(3.81, 0.49)	(5.43, 0.40)	(8.07, 0.32)	(11.02, 0.28)	(14.41, 0.27)
	300	(1.22, 0.92)	(1.34, 0.89)	(1.40, 0.87)	(1.61, 0.79)	(1.84, 0.74)	(2.38, 0.63)	(2.77, 0.58)	(3.46, 0.52)	(4.60, 0.44)	(6.47, 0.37)	(8.17, 0.34)	(14.57, 0.26)
	600	(1.17, 0.99)	(1.23, 0.96)	(1.21, 0.99)	(1.56, 0.80)	(1.57, 0.83)	(1.78, 0.80)	(2.26, 0.68)	(2.52, 0.65)	(3.46, 0.54)	(6.41, 0.35)	(7.39, 0.35)	(11.91, 0.28)
	1200	(1.14, 1.02)	(1.14, 1.04)	(1.12, 1.05)	(1.29, 0.95)	(1.36, 0.94)	(1.71, 0.80)	(1.88, 0.78)	(2.31, 0.67)	(2.73, 0.63)	(4.25, 0.48)	(5.40, 0.44)	(9.34, 0.33)
	2400	(1.13, 1.02)	(1.13, 1.02)	(1.10, 1.06)	(1.15, 1.04)	(1.28, 0.96)	(1.82, 0.72)	(1.65, 0.85)	(1.87, 0.79)	(2.51, 0.66)	(3.58, 0.54)	(4.80, 0.47)	(6.90, 0.41)
	4800	(1.03, 1.10)	(1.08, 1.06)	(1.09, 1.05)	(1.13, 1.03)	(1.25, 0.96)	(1.32, 0.95)	(1.57, 0.86)	(1.86, 0.78)	(1.95, 0.81)	(2.98, 0.62)	(3.17, 0.66)	(5.82, 0.45)

Table 3.6: Table of all (c_3, d_3) pairs according to the setting of sample sizes $(N, n/m)$.

The function psl.inf() (which can be found in section A.10 of the Appendix) implements this Satterthwaite approximation with bi-linear interpolation for sample sizes $(N', n'/m')$ that are not identical to any of the above listed simulation settings. If the sample sizes are such that extrapolation is necessary, a warning is given (e.g. $m = 400, n = 2000$). For appropriately large sample sizes ($m > 4640$, $n > 160$, $n/m < 1/29$), asymptotic cutoffs are used to determine $CI(\delta)$.

Chapter 4

Simulation Studies

4.1 Estimator Performance Comparison

This section compares the 4 different estimators of (θ, δ) presented in Chapter 2: Normal MLE, EM-like Algorithm, Method of Moments, and Pseudo-Likelihood. For Normal MLE, all details of the EM algorithm for finding the maximum of the log-likelihood are described in section 2.1. For the method of moments estimator, the $\epsilon_N = 20S_X^2/N^{.95}$ formula derived in section 2.2.2 is used. For the EM-like algorithm, the sample size dependent recommendation stated in section 2.3.2 is used

- If $m = n \leq 250$, use version 6
- If $m = n > 250$, use version 2

Lastly, for the Pseudo-likelihood estimator, the log-concave maximum likelihood estimate of f described in section 2.4.2 is used.

The estimators are compared by a simulation study that generates 1000 data sets under each following factorial combinations of settings in **List 2.1** (reproduced here for convenience).

- $m = n \in \{25, 50, 100, 500\}$
- $F \in \{$Normal, Laplace, SkRNorm, SkRLap, SkLNorm, SkLLap$\}$
- $\theta \in \{.2, .5, .8\}$
- $\delta \in \{.5, 1, 2, 3\}$

Note that $\sigma_X = 1$ for all F in the simulation so $\delta = \delta/\sigma_X$. The separation between the components (which impacts the performance of the estimators) is determined by δ/σ_X (rather than δ alone). This should be accounted for when interpreting these results for a contest where $\sigma_X \neq 1$, since δ below represents the number of standard deviations of separation between the components. Also, $\delta > 0$ for all simulations which means that skewed right distributions are skewed "in the direction of δ" and skewed left distributions are skewed "in the opposite direction of δ". This directional relationship between the skew of F and the direction of δ is what determines the performance of the estimators, so care should be taken in interpreting the results if $\delta < 0$. The described interpretation of simulation results below assume $\sigma_X = 1,\ \delta > 0$; see section A.5 of the Appendix for an example appropriately translating the results to a context where $\sigma_X \neq 1$ or $\delta < 0$.

For each fixed pair of sample sizes, the score for a specific estimator i defined in (2.17) is used to compare its performance relative to the other estimators (indexed by k, here $k \in \{1, ..., 4\}$). Each estimator's score represents the geometric average loss of that estimator relative to the 'oracle' estimator [that separately chooses the estimator(s) that minimize(s) $\sqrt{MSE(\widehat{\theta})}$, $\sqrt{MSE(\widehat{\delta})}$, and $\sqrt{MSE(\widehat{\Delta})}$ for each $(m = n, F, \theta, \delta)$]. **Table 4.1** displays the scores of the 4 estimators for the simulation. Smaller scores are better and the smallest scores are highlighted in yellow.

m = n	Estimator	$S(\widehat{\theta})$	$S(\widehat{\delta})$	$S(\widehat{\Delta})$	$\{S(\widehat{\Delta})S(\widehat{\theta})S(\widehat{\delta})\}^{1/3}$
25	Normal MLE	1.206	1.454	1.084	1.239
	Moment	1.324	1.243	1.155	1.239
	EM-like	1.099	1.421	1.050	1.179
	Ps-Likelihood	1.123	1.272	1.037	1.140
50	Normal MLE	1.280	1.614	1.120	1.322
	Moment	1.452	1.367	1.180	1.328
	EM-like	1.128	1.490	1.075	1.218
	Ps-Likelihood	1.112	1.277	1.022	1.132
100	Normal MLE	1.372	1.812	1.172	1.429
	Moment	1.612	1.509	1.199	1.429
	EM-like	1.164	1.529	1.095	1.249
	Ps-Likelihood	1.098	1.240	1.015	1.114
500	Normal MLE	1.842	2.372	1.362	1.812
	Moment	2.025	1.871	1.214	1.663
	EM-like	1.237	1.586	1.120	1.300
	Ps-Likelihood	1.138	1.135	1.017	1.095

Table 4.1: Scores for Estimates of θ, δ, Δ.

Table 4.1 indicates that the Pseudo-Likelihood estimator has the most robust performance for all $\widehat{\theta}$, $\widehat{\delta}$, $\widehat{\Delta}$ and all sample sizes with the exception of when $m = n = 25$, $S(\widehat{\theta}_{PsL})$ and $S(\widehat{\delta}_{PsL})$ are close seconds to $S(\widehat{\theta}_{EMlike})$ and $S(\widehat{\delta}_{MoM})$ respectively. **Figures 4.1 - 4.2** plot the scores over the sample sizes. The Normal MLE and Moment estimators have the least desirable performance - with the Normal MLE unsurprisingly performing relatively worse as the sample size increases (as the Normality assumption only holds for one sixth of the simulation settings). The next best estimator is the EM-like algorithm which is the second-best for almost all settings. The Pseudo-Likelihood estimator becomes increasingly more efficient relative to the other estimators as the sample sizes increase, particularly for $\widehat{\delta}$ and $\widehat{\Delta}$.

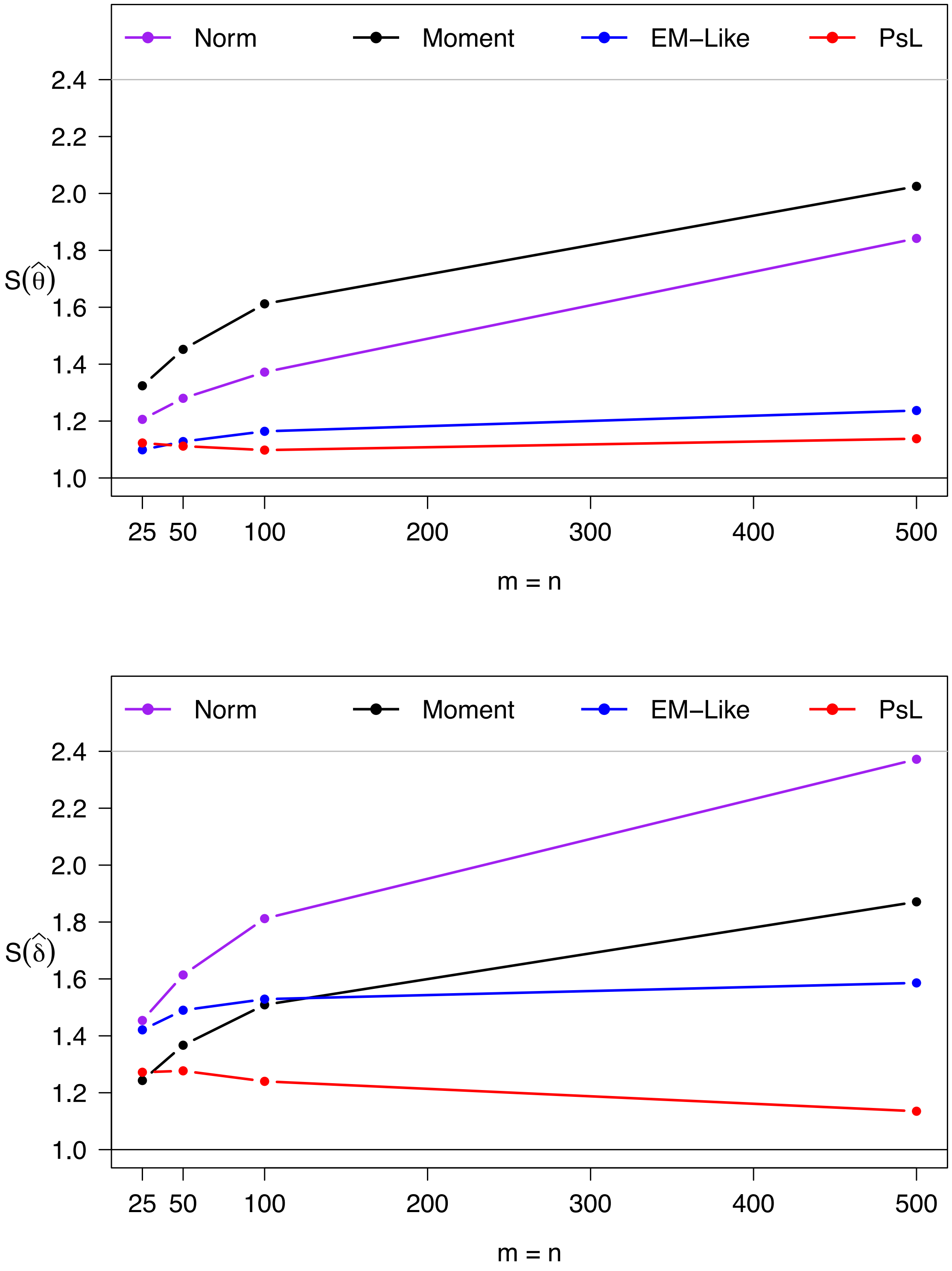


Figure 4.1: Estimator Scores for $\widehat{\theta}$ and $\widehat{\delta}$

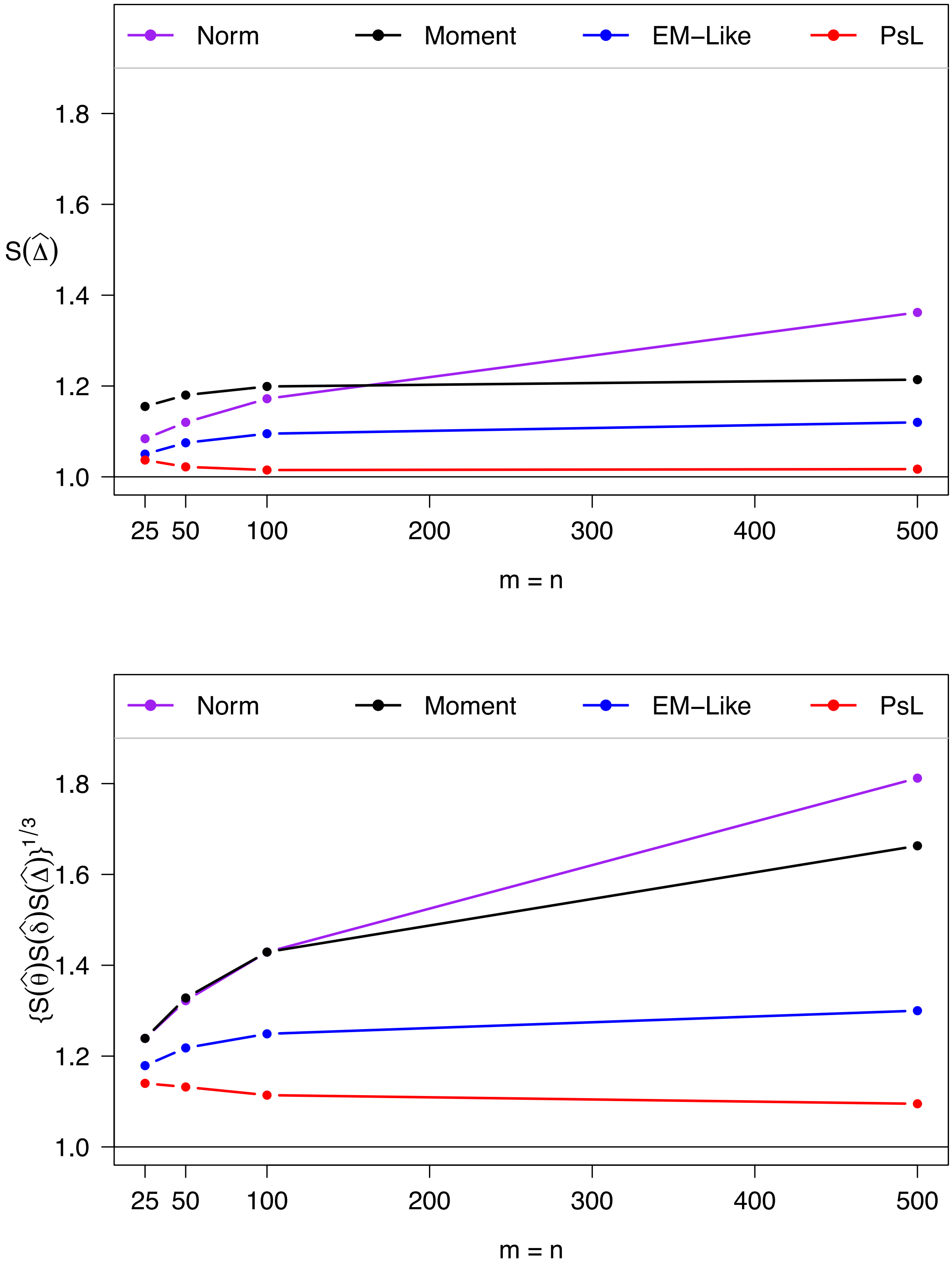


Figure 4.2: Estimator Scores for $\widehat{\Delta}$ and an Overall Summary Score

To understand the performances of the best two estimators (EM-like and Pseudo-Likelihood) for every simulation setting, consider the scatterplots of $\sqrt{MSE}$ in **Figure 4.3** below. When θ is more difficult to estimate, the EM-like estimate appears to have slightly lower $\sqrt{MSE(\widehat{\theta})}$, while scenarios when estimation is easier lend themselves to smaller $\sqrt{MSE(\widehat{\theta})}$ for the pseudo-likelihood method. For estimation of δ, the pseudo-likelihood method tends to have smaller $\sqrt{MSE(\widehat{\delta})}$ especially in the cases where estimation is easier.

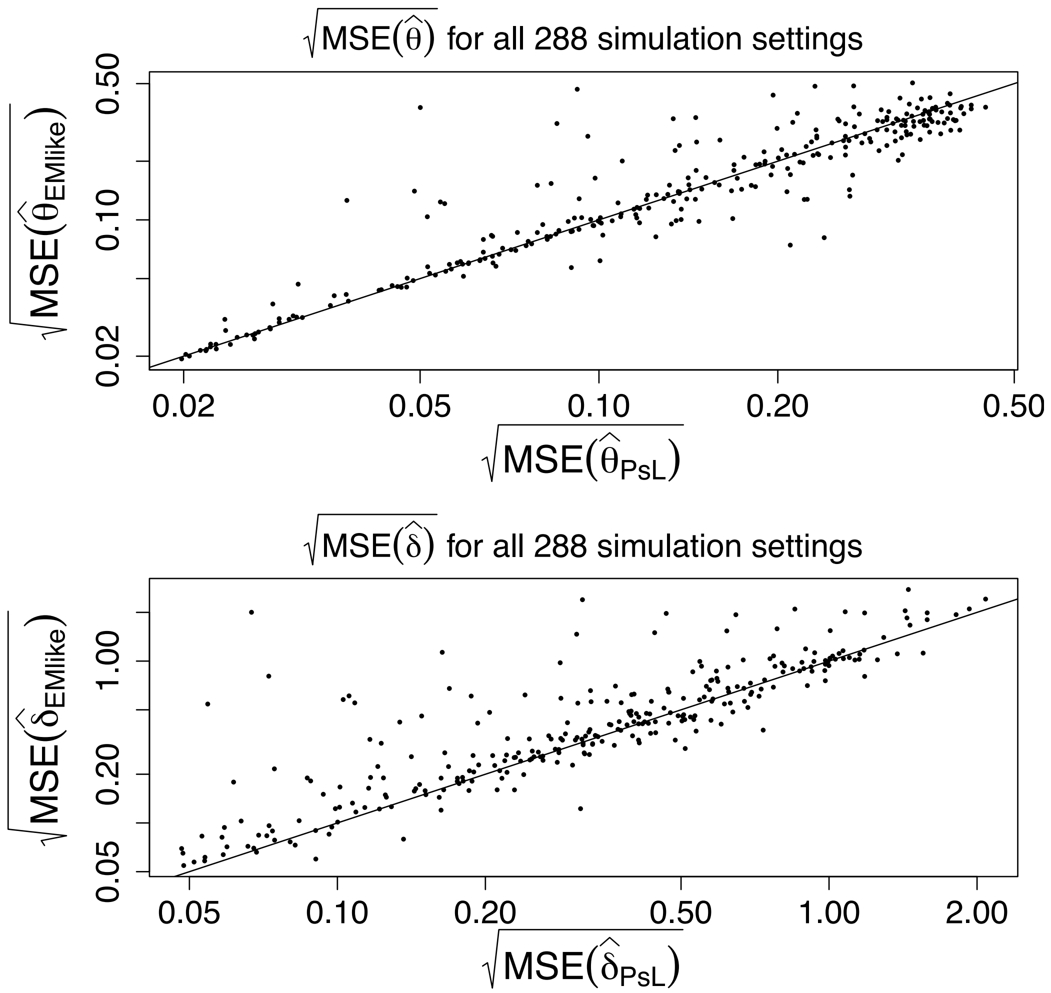


Figure 4.3: Scatterplots of $\sqrt{MSE}$ comparing EM-like and Pseudo-likelihood estimators for both $\widehat{\theta}$ and $\widehat{\delta}$.

To better understand the relative performances of these two estimators for each the parameter setting, consider the dot plots in **Figures 4.4 - 4.7** of $\sqrt{MSE/\min MSE}$ under each setting. Blue dots represent the EM-like estimator and red dots represent the Pseudo-likelihood estimator. Simulation setting is denoted on the x axes. Under each setting, the estimator at 1 is the best estimator.

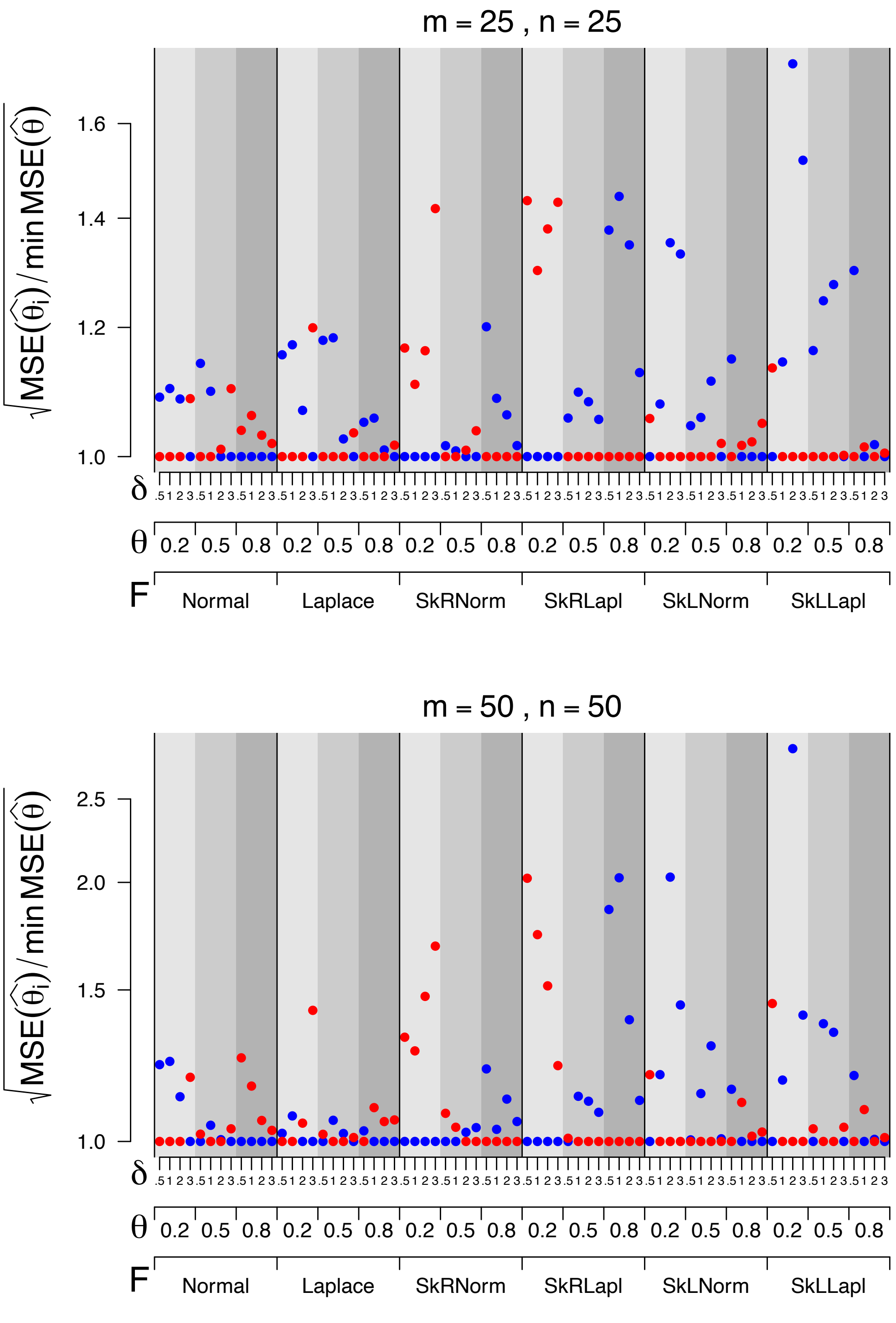


Figure 4.4: Dot Plots comparing the performance of estimators of θ for small sample sizes. Blue dots represent the EM-like estimator and red dots represent the Pseudo-likelihood estimator.

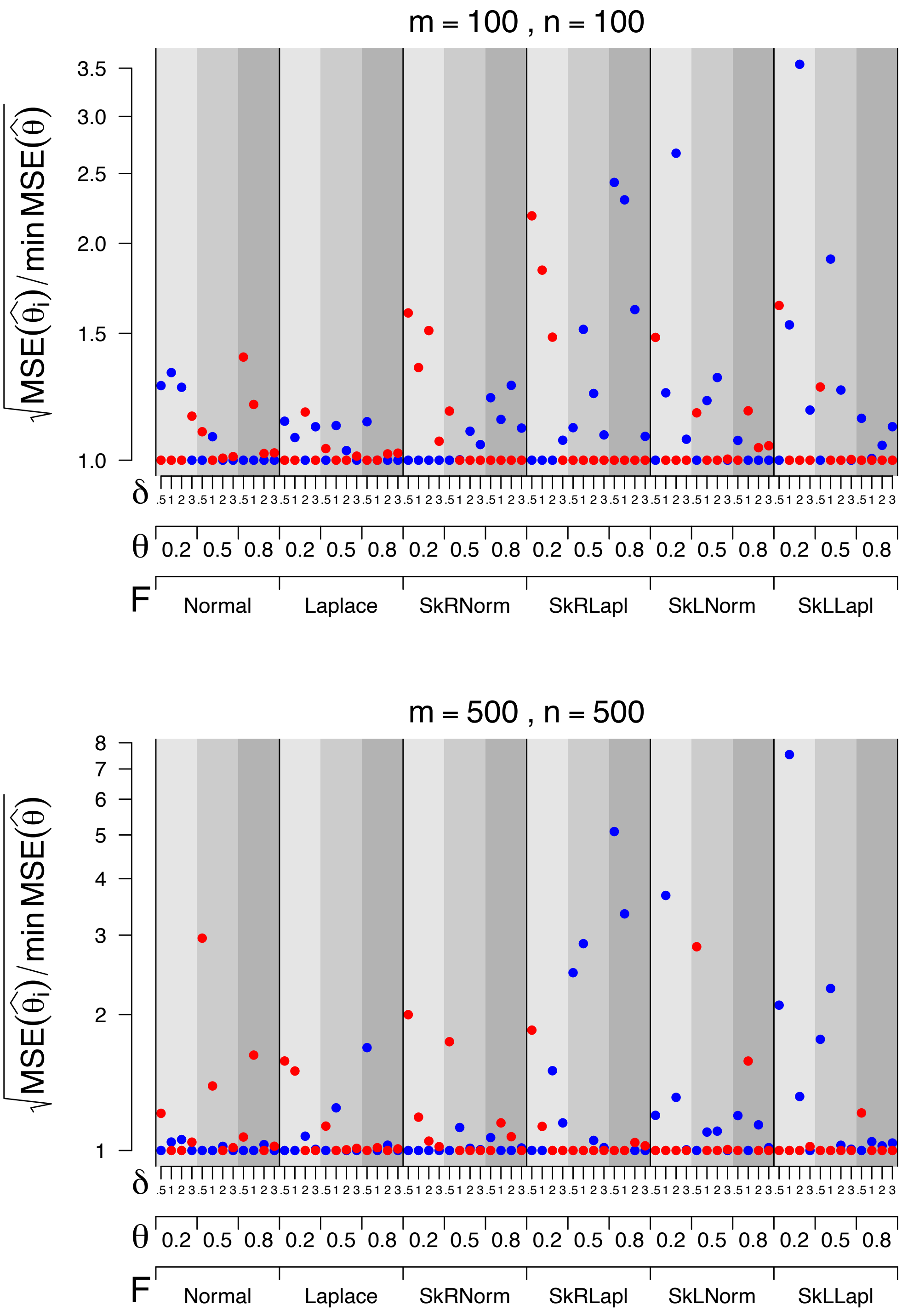
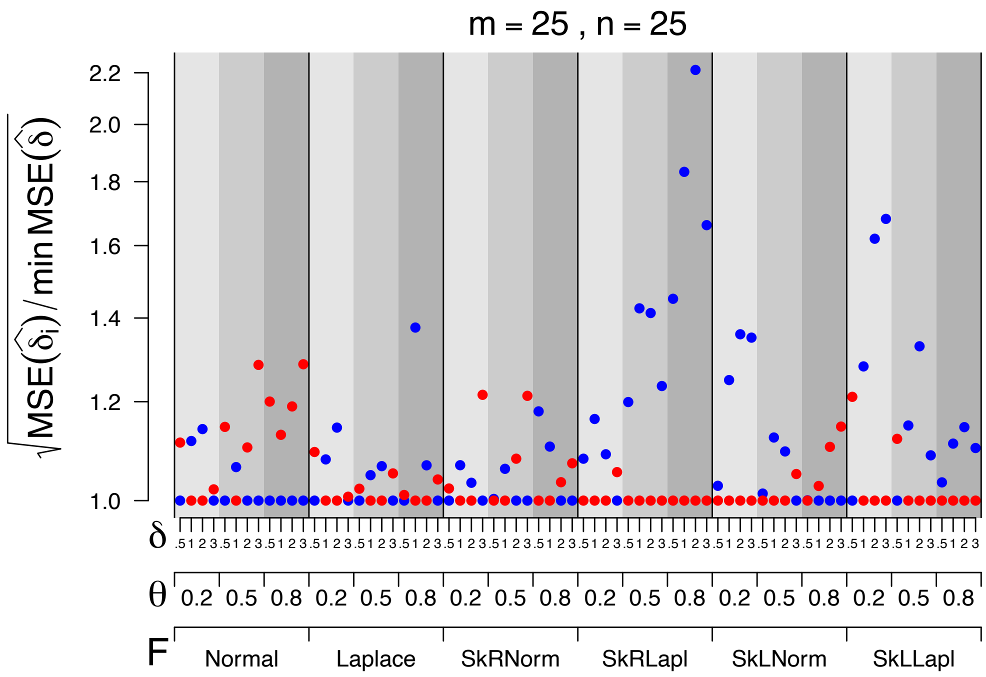


Figure 4.5: Dot Plots comparing the performance of estimators of θ for moderate to large sample sizes. Blue dots represent the EM-like estimator and red dots represent the Pseudo-likelihood estimator.



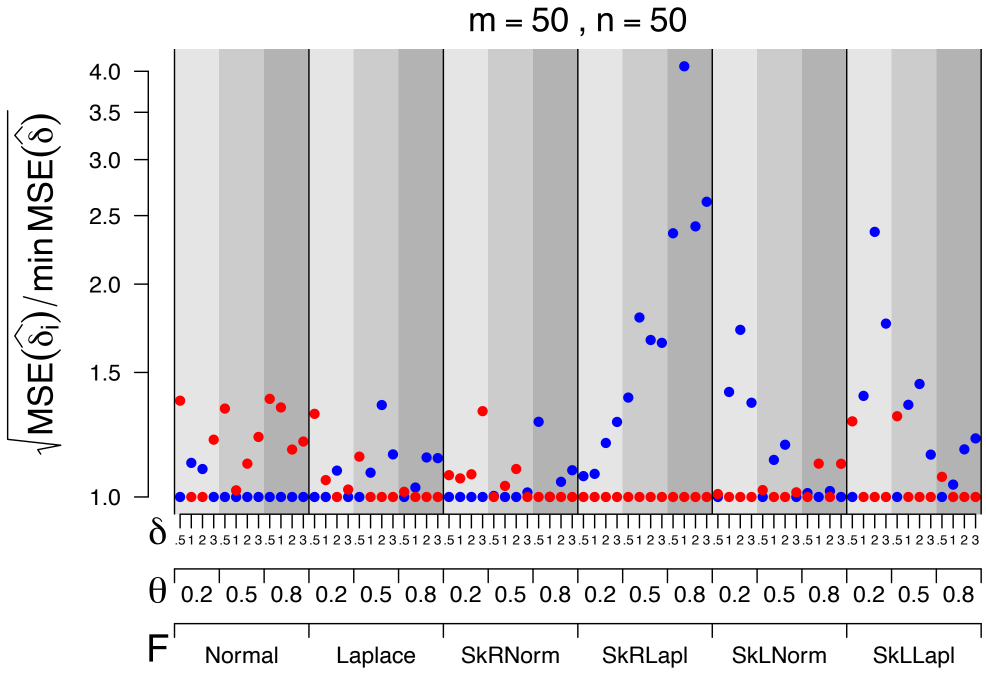


Figure 4.6: Dot Plots comparing the performance of estimators of δ for small sample sizes. Blue dots represent the EM-like estimator and red dots represent the Pseudo-likelihood estimator.

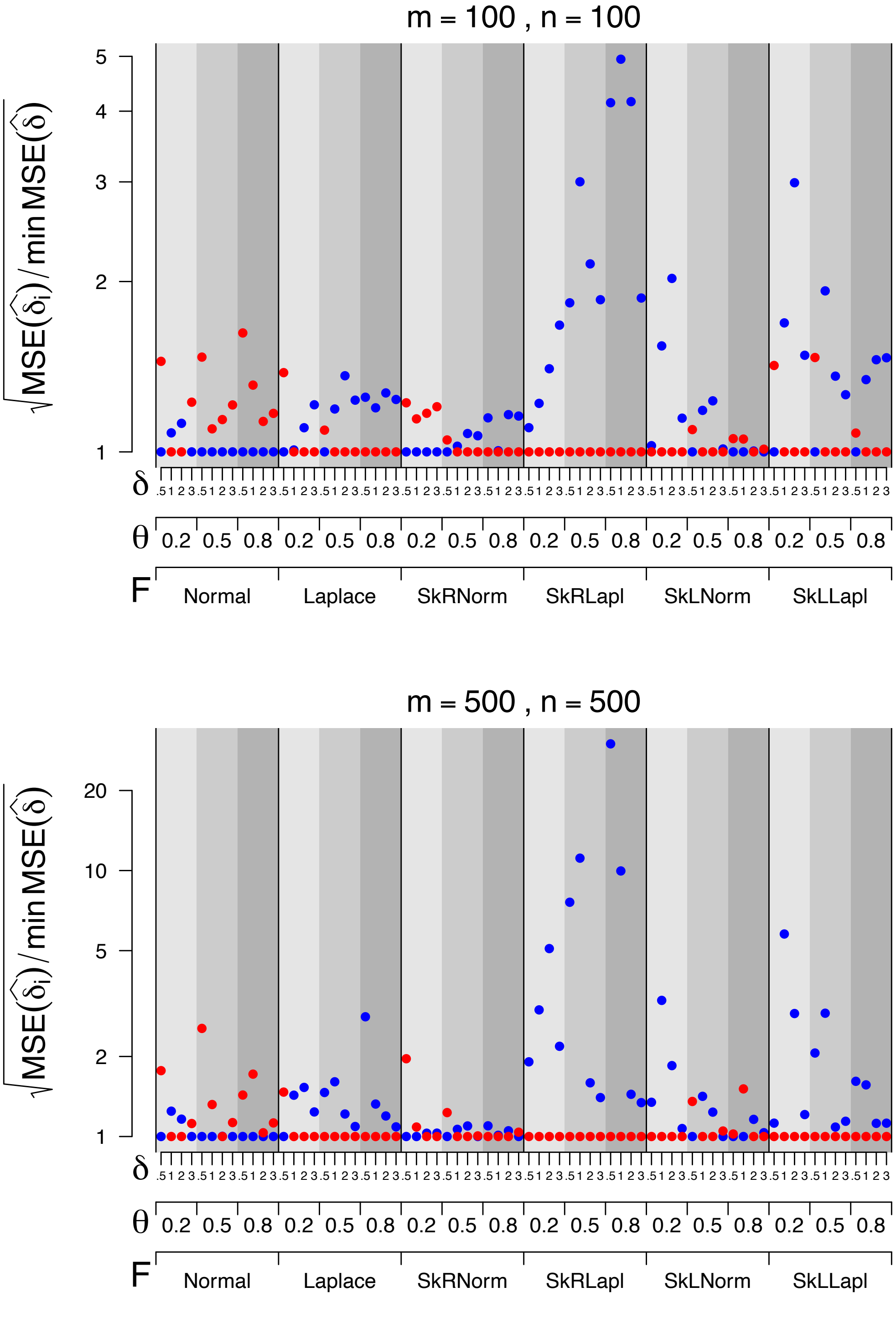


Figure 4.7: Dot Plots comparing the performance of estimators of δ for moderate to large sample sizes. Blue dots represent the EM-like estimator and red dots represent the Pseudo-likelihood estimator.

The dot plots in **Figures 4.4 - 4.7** reveal a few trends. For small sample sizes in particular, $\widehat{\theta}_{PsL}$ is preferable for larger θ with skewed right distributions and for smaller θ with skewed left distributions. Conversely, $\widehat{\theta}_{EMlike}$ is preferred for smaller θ with skewed right distributions and for larger θ with skewed left distributions. The figures also reveal that $\widehat{\theta}_{PsL}$ shows superior performance compared to $\widehat{\theta}_{EMlike}$ for heavy-tailed and left-skewed distributions while the EM-like algorithm is better for Normal and Skewed-right Normal distributions. Similarly, $\widehat{\delta}_{EMlike}$ shows better results for $F \sim$ Normal and mixed results for Skewed-right Normal while $\widehat{\delta}_{PsL}$ demonstrates dramatically superior performance for all other F. Since the heuristic EM-like algorithm mimics a normal EM-algorithm and uses normal kernels for kernel density estimation, it is not surprising to see this preference for $F \sim$ Normal.

4.1.1 Optimal Sample Size Allocation

Given that $\widehat{(\theta, \delta)}_{PsL}$ achieves the most efficient estimation, it is natural to consider what sample size allocation produces optimal results. Optimal sample size allocation is determined from a simulation study that generates 1000 data sets under each factorial combination of the following settings.

- $N \in \{60, 120, 180, 300, 600, 1200, 2400, 4800\}$
- $n{:}m \in \{1{:}3, 1{:}2, 2{:}3, 1{:}1, 3{:}2, 2{:}1, 3{:}1\}$
- $F \in \{\text{Normal, Laplace, SkRNorm, SkRLap, SkLNorm, SkLLap}\}$
- $\theta \in \{.2, .5, .8\}$
- $\delta \in \{.5, 1, 2, 3\}$

For each pair of $(N, n{:}m)$, the performance across all 72 (F, θ, δ) is summarized by the average $\sqrt{MSE(\widehat{\theta})}$, $\sqrt{MSE(\widehat{\delta})}$, and $\sqrt{MSE(\widehat{\Delta})}$ in **Tables 4.2, 4.3,** and **4.4** respectively. Each row represents a total sample size N and each column represents a randomization ratio (ratios on the left of the table assign more patients to the control group while those on the right assign more patients to the treatment group). The optimal randomization ratio for each row is highlighted in yellow. For scenarios where the experimental cost of assigning a patient to each group is equal, **Tables 4.2 - 4.4** show that a sample allocation of three patients assigned to the treatment group for every two patients assigned to the control group is optimal for precise estimation of the treatment effect regardless of the total sample size. For settings where the cost associated with assigning a patient to one group is more costly than another, the settings of sample sizes that are within budget can be identified and the table entries can be used to identify the within-budget setting with the best performance.

		n:m						
		1:3	1:2	2:3	1:1	3:2	2:1	3:1
	60	0.285	0.268	0.257	0.247	0.243	0.247	0.251
	120	0.228	0.213	0.204	0.196	0.194	0.196	0.202
	180	0.201	0.186	0.177	0.171	0.168	0.169	0.176
N	300	0.167	0.153	0.148	0.141	0.139	0.140	0.144
	600	0.129	0.118	0.113	0.107	0.105	0.107	0.111
	1200	0.097	0.089	0.084	0.081	0.079	0.079	0.083
	2400	0.072	0.066	0.062	0.059	0.057	0.058	0.060
	4800	0.052	0.047	0.045	0.042	0.041	0.041	0.043

Table 4.2: Average $\sqrt{MSE(\widehat{\theta})}$ across 72 (F, θ, δ).

		n:m						
		1:3	1:2	2:3	1:1	3:2	2:1	3:1
N	60	0.763	0.715	0.683	0.662	0.644	0.656	0.659
	120	0.622	0.577	0.542	0.520	0.512	0.516	0.531
	180	0.543	0.498	0.474	0.458	0.439	0.450	0.447
	300	0.444	0.400	0.382	0.364	0.349	0.344	0.360
	600	0.322	0.292	0.270	0.248	0.244	0.250	0.259
	1200	0.216	0.199	0.180	0.168	0.163	0.170	0.176
	2400	0.145	0.122	0.116	0.108	0.108	0.111	0.120
	4800	0.091	0.078	0.074	0.069	0.069	0.071	0.078

Table 4.3: Average $\sqrt{MSE(\widehat{\delta})}$ across 72 (F, θ, δ).

		n:m						
		1:3	1:2	2:3	1:1	3:2	2:1	3:1
N	60	0.309	0.276	0.263	0.251	0.248	0.252	0.266
	120	0.217	0.196	0.185	0.176	0.173	0.175	0.186
	180	0.178	0.160	0.151	0.143	0.141	0.143	0.150
	300	0.139	0.125	0.118	0.111	0.109	0.110	0.116
	600	0.098	0.088	0.083	0.078	0.077	0.078	0.081
	1200	0.070	0.062	0.059	0.055	0.054	0.055	0.058
	2400	0.049	0.044	0.042	0.039	0.038	0.039	0.041
	4800	0.035	0.031	0.030	0.028	0.027	0.028	0.029

Table 4.4: Average $\sqrt{MSE(\widehat{\Delta})}$ across 72 (F, θ, δ).

4.1.2 Parameter Specific Performance

With the optimal ratio $n{:}m = 3{:}2$ in hand, consider the pseudo-likelihood estimator performance across (F, θ, δ) for each $N \in \{60, 120, 180, 300, 600, 1200, 2400, 4800\}$. The heat grids in **Figures 4.9 - 4.12** present grids of $\sqrt{MSE(\widehat{\theta})}$ for each sample size, N. Similarly, **Figures 4.13 - 4.16** present grids of $\sqrt{MSE(\widehat{\delta})}$ for each sample size, N. To aid in pattern recognition across (F, θ, δ) for each fixed N, every $\sqrt{MSE}$ entry contains a colored background indicating the size of $\sqrt{MSE}$ relative to the median $\sqrt{MSE}$, M. The color key in **Figure 4.8** below indicates that a black background represents the median while bright red represents $\sqrt{MSE}$ much smaller than the median and bright blue represents $\sqrt{MSE}$ much larger than the median. For each grid, red represents easier cases of estimation while blue indicates more difficult ones.

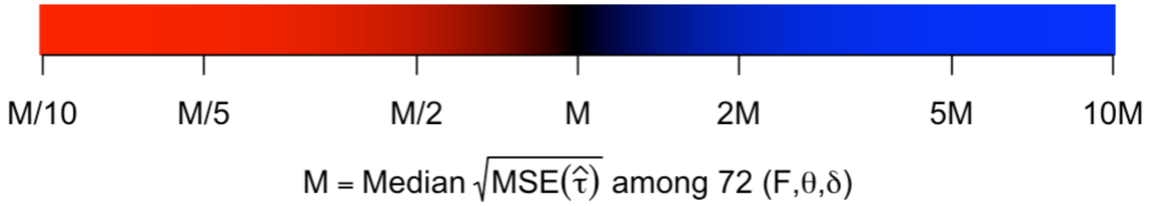
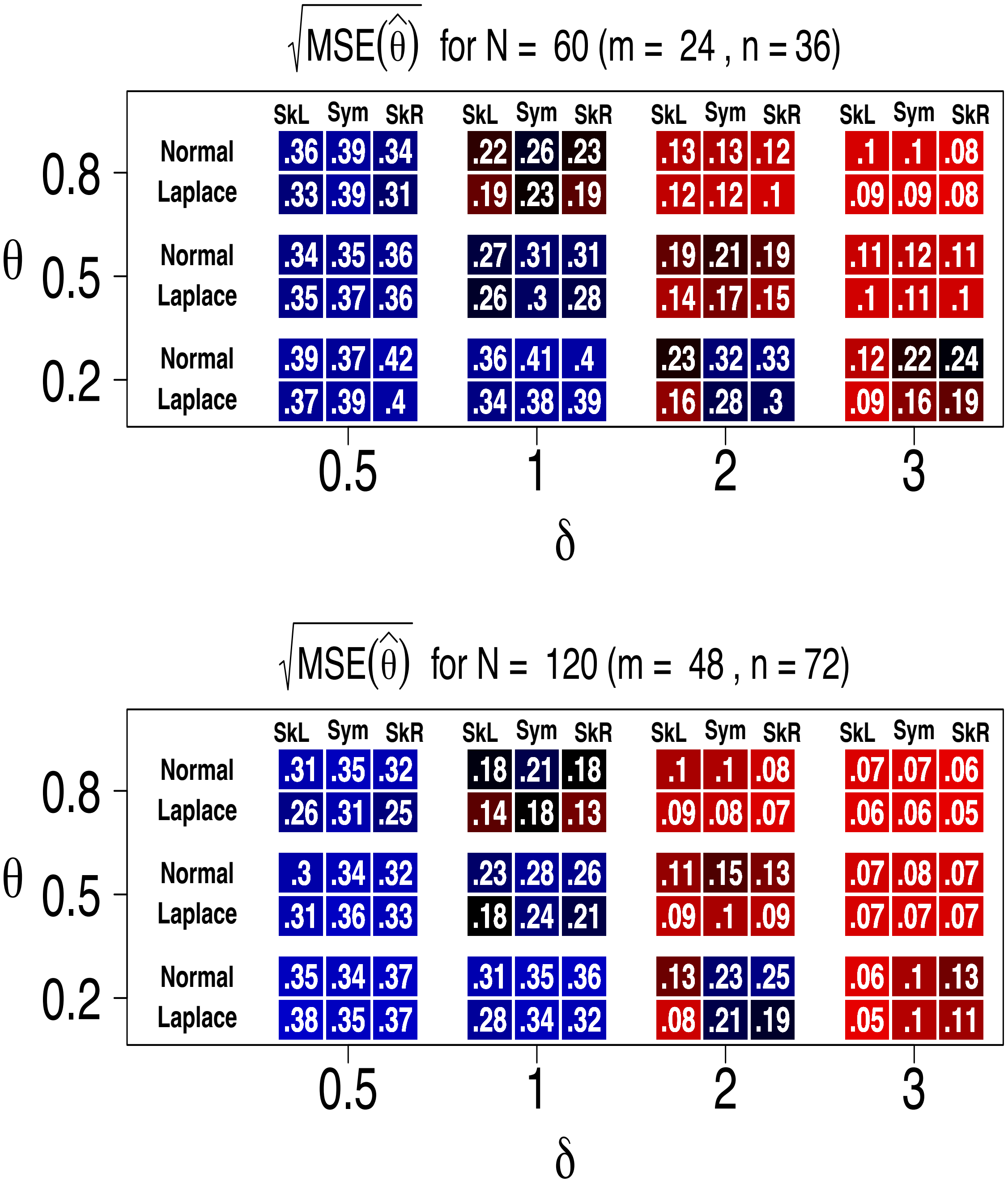


Figure 4.8: Color Key for $\sqrt{MSE(\widehat{\tau}_{PsL})}$



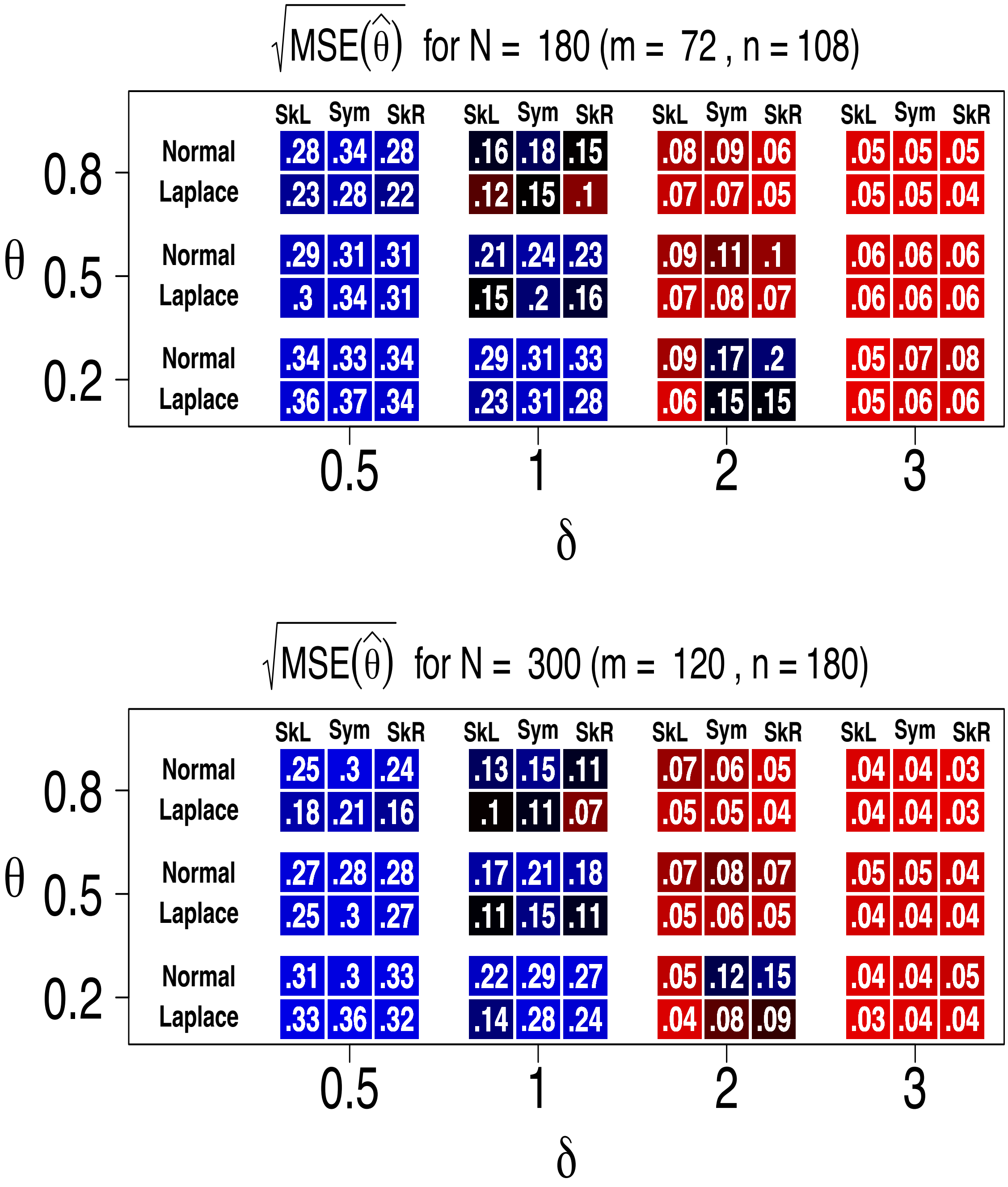
$\sqrt{\mathrm{MSE}(\widehat{\theta})}$ for N = 60 (m = 24 , n = 36)

θ		0.5			1			2			3		
		SkL	Sym	SkR	SkL	Sym	SkR	SkL	Sym	SkR	SkL	Sym	SkR
0.8	Normal	.36	.39	.34	.22	.26	.23	.13	.13	.12	.1	.1	.08
	Laplace	.33	.39	.31	.19	.23	.19	.12	.12	.1	.09	.09	.08
0.5	Normal	.34	.35	.36	.27	.31	.31	.19	.21	.19	.11	.12	.11
	Laplace	.35	.37	.36	.26	.3	.28	.14	.17	.15	.1	.11	.1
0.2	Normal	.39	.37	.42	.36	.41	.4	.23	.32	.33	.12	.22	.24
	Laplace	.37	.39	.4	.34	.38	.39	.16	.28	.3	.09	.16	.19

$\sqrt{\mathrm{MSE}(\widehat{\theta})}$ for N = 120 (m = 48 , n = 72)

θ		0.5			1			2			3		
		SkL	Sym	SkR	SkL	Sym	SkR	SkL	Sym	SkR	SkL	Sym	SkR
0.8	Normal	.31	.35	.32	.18	.21	.18	.1	.1	.08	.07	.07	.06
	Laplace	.26	.31	.25	.14	.18	.13	.09	.08	.07	.06	.06	.05
0.5	Normal	.3	.34	.32	.23	.28	.26	.11	.15	.13	.07	.08	.07
	Laplace	.31	.36	.33	.18	.24	.21	.09	.1	.09	.07	.07	.07
0.2	Normal	.35	.34	.37	.31	.35	.36	.13	.23	.25	.06	.1	.13
	Laplace	.38	.35	.37	.28	.34	.32	.08	.21	.19	.05	.1	.11

Figure 4.9: Heat Grids for Pseudo-likelihood $\widehat{\theta}$ for $N \in \{60, 120\}$



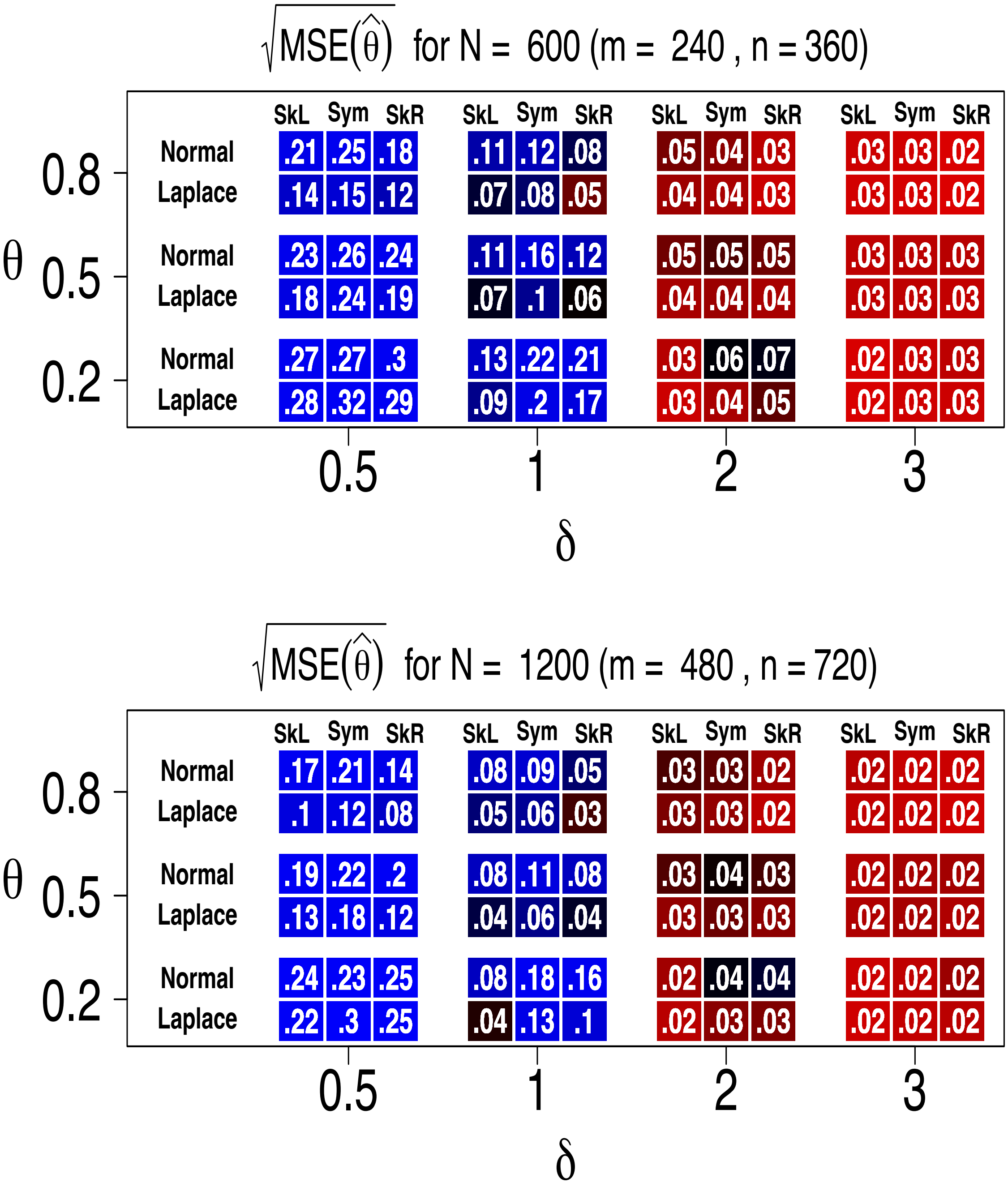
$\sqrt{\mathrm{MSE}(\hat{\theta})}$ for N = 180 (m = 72 , n = 108)

θ		0.5			1			2			3		
		SkL	Sym	SkR	SkL	Sym	SkR	SkL	Sym	SkR	SkL	Sym	SkR
0.8	Normal	.28	.34	.28	.16	.18	.15	.08	.09	.06	.05	.05	.05
	Laplace	.23	.28	.22	.12	.15	.1	.07	.07	.05	.05	.05	.04
0.5	Normal	.29	.31	.31	.21	.24	.23	.09	.11	.1	.06	.06	.06
	Laplace	.3	.34	.31	.15	.2	.16	.07	.08	.07	.06	.06	.06
0.2	Normal	.34	.33	.34	.29	.31	.33	.09	.17	.2	.05	.07	.08
	Laplace	.36	.37	.34	.23	.31	.28	.06	.15	.15	.05	.06	.06

$\sqrt{\mathrm{MSE}(\hat{\theta})}$ for N = 300 (m = 120 , n = 180)

θ		0.5			1			2			3		
		SkL	Sym	SkR	SkL	Sym	SkR	SkL	Sym	SkR	SkL	Sym	SkR
0.8	Normal	.25	.3	.24	.13	.15	.11	.07	.06	.05	.04	.04	.03
	Laplace	.18	.21	.16	.1	.11	.07	.05	.05	.04	.04	.04	.03
0.5	Normal	.27	.28	.28	.17	.21	.18	.07	.08	.07	.05	.05	.04
	Laplace	.25	.3	.27	.11	.15	.11	.05	.06	.05	.04	.04	.04
0.2	Normal	.31	.3	.33	.22	.29	.27	.05	.12	.15	.04	.04	.05
	Laplace	.33	.36	.32	.14	.28	.24	.04	.08	.09	.03	.04	.04

Figure 4.10: Heat Grids for Pseudo-likelihood $\widehat{\theta}$ for $N \in \{180, 300\}$



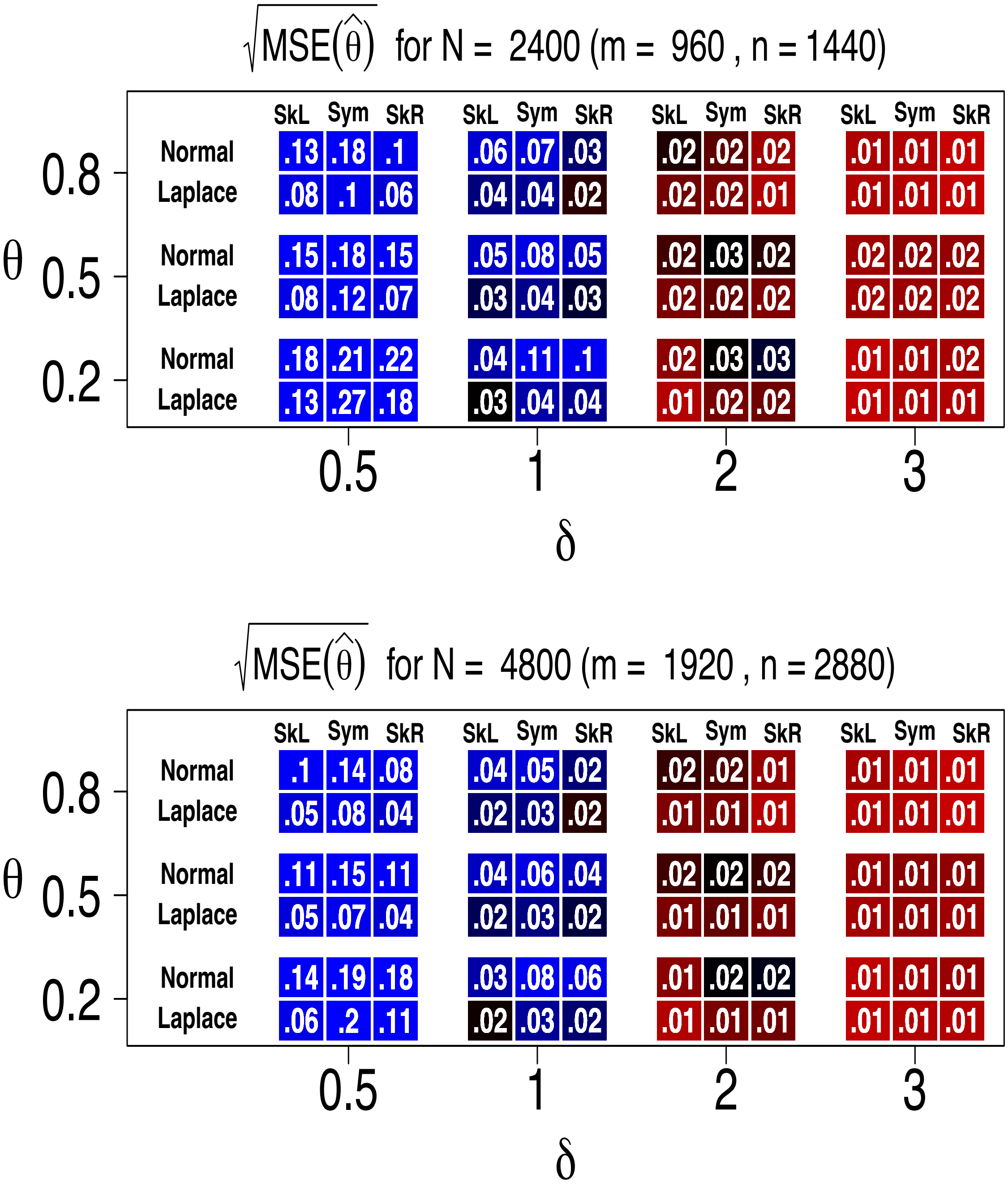
$\sqrt{\mathrm{MSE}(\widehat{\theta})}$ for N = 600 (m = 240 , n = 360)

θ		δ = 0.5			1			2			3		
		SkL	Sym	SkR	SkL	Sym	SkR	SkL	Sym	SkR	SkL	Sym	SkR
0.8	Normal	.21	.25	.18	.11	.12	.08	.05	.04	.03	.03	.03	.02
	Laplace	.14	.15	.12	.07	.08	.05	.04	.04	.03	.03	.03	.02
0.5	Normal	.23	.26	.24	.11	.16	.12	.05	.05	.05	.03	.03	.03
	Laplace	.18	.24	.19	.07	.1	.06	.04	.04	.04	.03	.03	.03
0.2	Normal	.27	.27	.3	.13	.22	.21	.03	.06	.07	.02	.03	.03
	Laplace	.28	.32	.29	.09	.2	.17	.03	.04	.05	.02	.03	.03

$\sqrt{\mathrm{MSE}(\widehat{\theta})}$ for N = 1200 (m = 480 , n = 720)

θ		δ = 0.5			1			2			3		
		SkL	Sym	SkR	SkL	Sym	SkR	SkL	Sym	SkR	SkL	Sym	SkR
0.8	Normal	.17	.21	.14	.08	.09	.05	.03	.03	.02	.02	.02	.02
	Laplace	.1	.12	.08	.05	.06	.03	.03	.03	.02	.02	.02	.02
0.5	Normal	.19	.22	.2	.08	.11	.08	.03	.04	.03	.02	.02	.02
	Laplace	.13	.18	.12	.04	.06	.04	.03	.03	.03	.02	.02	.02
0.2	Normal	.24	.23	.25	.08	.18	.16	.02	.04	.04	.02	.02	.02
	Laplace	.22	.3	.25	.04	.13	.1	.02	.03	.03	.02	.02	.02

Figure 4.11: Heat Grids for Pseudo-likelihood $\widehat{\theta}$ for $N \in \{600, 1200\}$



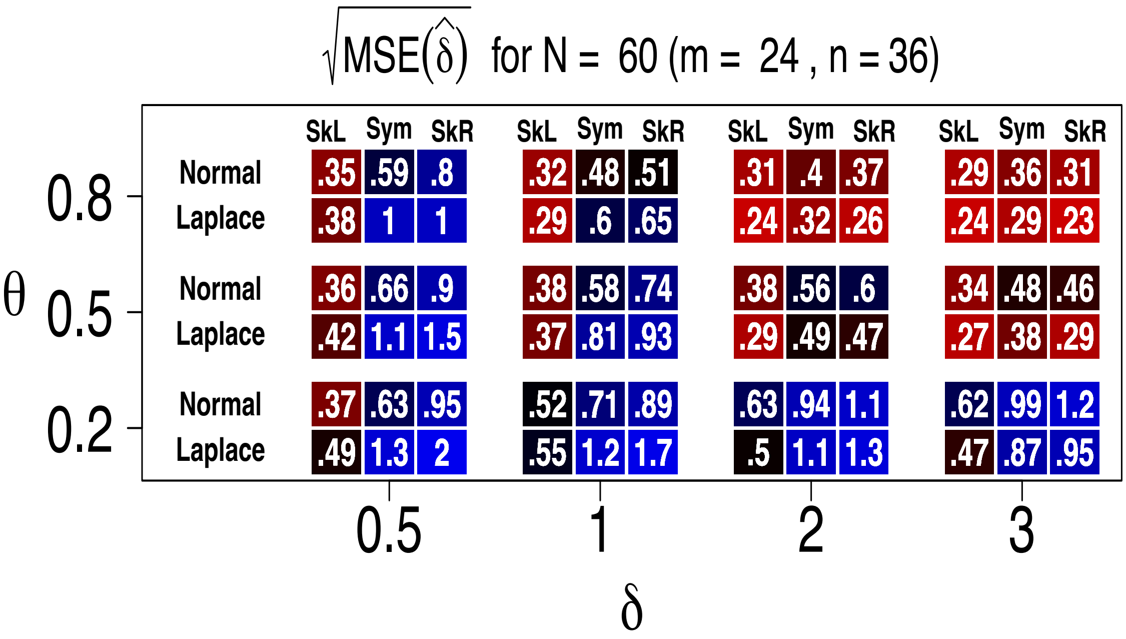
$\sqrt{\text{MSE}(\widehat{\theta})}$ for N = 2400 (m = 960 , n = 1440)

θ		δ = 0.5			δ = 1			δ = 2			δ = 3		
		SkL	Sym	SkR	SkL	Sym	SkR	SkL	Sym	SkR	SkL	Sym	SkR
0.8	Normal	.13	.18	.1	.06	.07	.03	.02	.02	.02	.01	.01	.01
	Laplace	.08	.1	.06	.04	.04	.02	.02	.02	.01	.01	.01	.01
0.5	Normal	.15	.18	.15	.05	.08	.05	.02	.03	.02	.02	.02	.02
	Laplace	.08	.12	.07	.03	.04	.03	.02	.02	.02	.02	.02	.02
0.2	Normal	.18	.21	.22	.04	.11	.1	.02	.03	.03	.01	.01	.02
	Laplace	.13	.27	.18	.03	.04	.04	.01	.02	.02	.01	.01	.01

$\sqrt{\text{MSE}(\widehat{\theta})}$ for N = 4800 (m = 1920 , n = 2880)

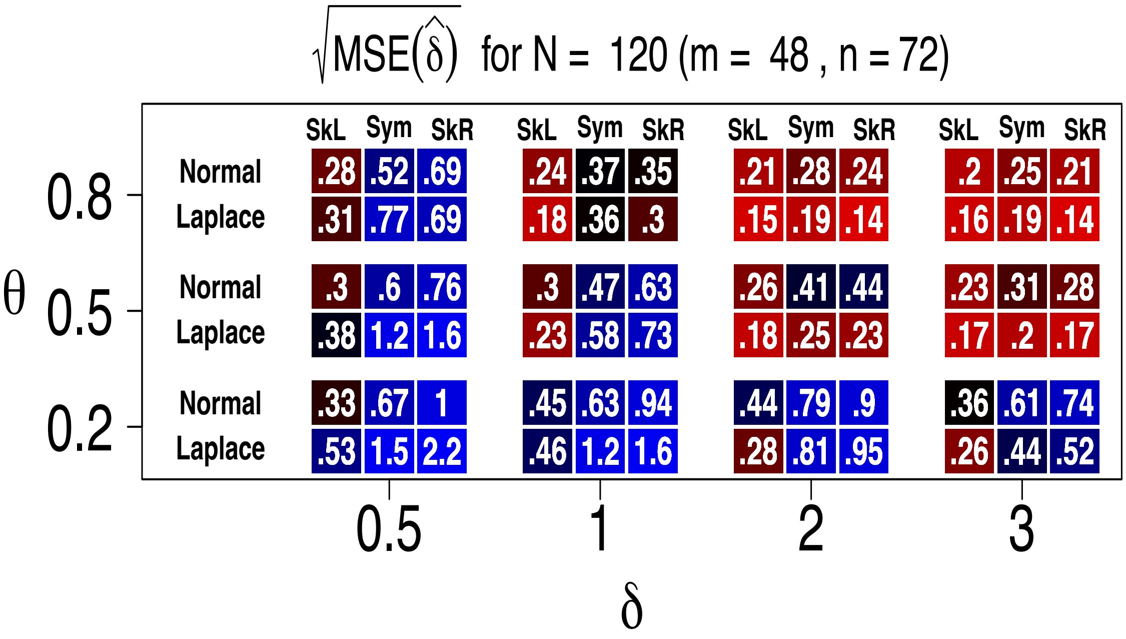
θ		δ = 0.5			δ = 1			δ = 2			δ = 3		
		SkL	Sym	SkR	SkL	Sym	SkR	SkL	Sym	SkR	SkL	Sym	SkR
0.8	Normal	.1	.14	.08	.04	.05	.02	.02	.02	.01	.01	.01	.01
	Laplace	.05	.08	.04	.02	.03	.02	.01	.01	.01	.01	.01	.01
0.5	Normal	.11	.15	.11	.04	.06	.04	.02	.02	.02	.01	.01	.01
	Laplace	.05	.07	.04	.02	.03	.02	.01	.01	.01	.01	.01	.01
0.2	Normal	.14	.19	.18	.03	.08	.06	.01	.02	.02	.01	.01	.01
	Laplace	.06	.2	.11	.02	.03	.02	.01	.01	.01	.01	.01	.01

Figure 4.12: Heat Grids for Pseudo-likelihood $\widehat{\theta}$ for $N \in \{2400, 4800\}$



$\sqrt{\text{MSE}(\hat{\delta})}$ for N = 60 (m = 24 , n = 36)

		δ = 0.5			δ = 1			δ = 2			δ = 3		
θ		SkL	Sym	SkR	SkL	Sym	SkR	SkL	Sym	SkR	SkL	Sym	SkR
0.8	Normal	.35	.59	.8	.32	.48	.51	.31	.4	.37	.29	.36	.31
	Laplace	.38	1	1	.29	.6	.65	.24	.32	.26	.24	.29	.23
0.5	Normal	.36	.66	.9	.38	.58	.74	.38	.56	.6	.34	.48	.46
	Laplace	.42	1.1	1.5	.37	.81	.93	.29	.49	.47	.27	.38	.29
0.2	Normal	.37	.63	.95	.52	.71	.89	.63	.94	1.1	.62	.99	1.2
	Laplace	.49	1.3	2	.55	1.2	1.7	.5	1.1	1.3	.47	.87	.95



$\sqrt{\text{MSE}(\hat{\delta})}$ for N = 120 (m = 48 , n = 72)

		δ = 0.5			δ = 1			δ = 2			δ = 3		
θ		SkL	Sym	SkR	SkL	Sym	SkR	SkL	Sym	SkR	SkL	Sym	SkR
0.8	Normal	.28	.52	.69	.24	.37	.35	.21	.28	.24	.2	.25	.21
	Laplace	.31	.77	.69	.18	.36	.3	.15	.19	.14	.16	.19	.14
0.5	Normal	.3	.6	.76	.3	.47	.63	.26	.41	.44	.23	.31	.28
	Laplace	.38	1.2	1.6	.23	.58	.73	.18	.25	.23	.17	.2	.17
0.2	Normal	.33	.67	1	.45	.63	.94	.44	.79	.9	.36	.61	.74
	Laplace	.53	1.5	2.2	.46	1.2	1.6	.28	.81	.95	.26	.44	.52

Figure 4.13: Heat Grids for Pseudo-likelihood $\widehat{\delta}$ for $N \in \{60, 120\}$

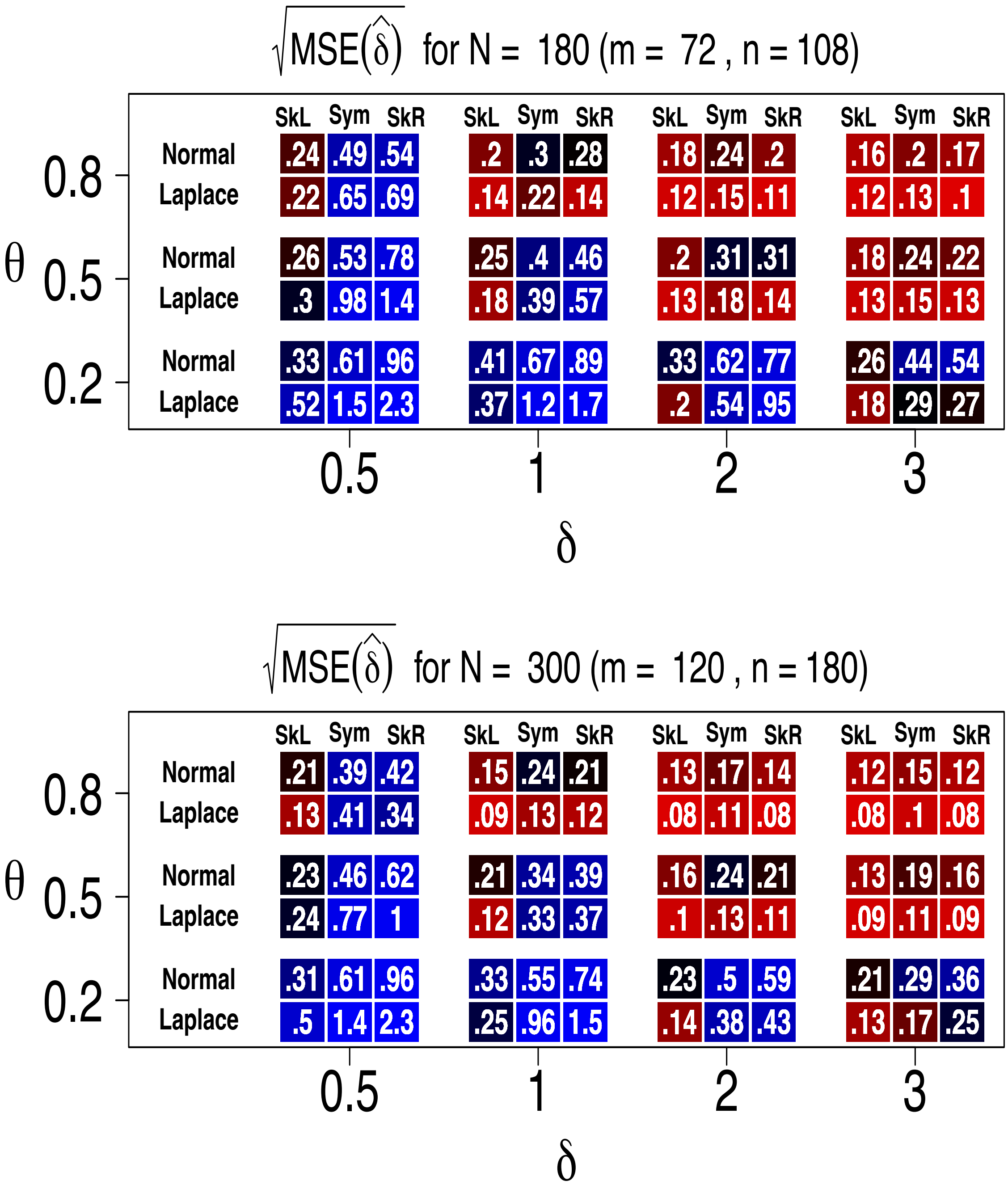
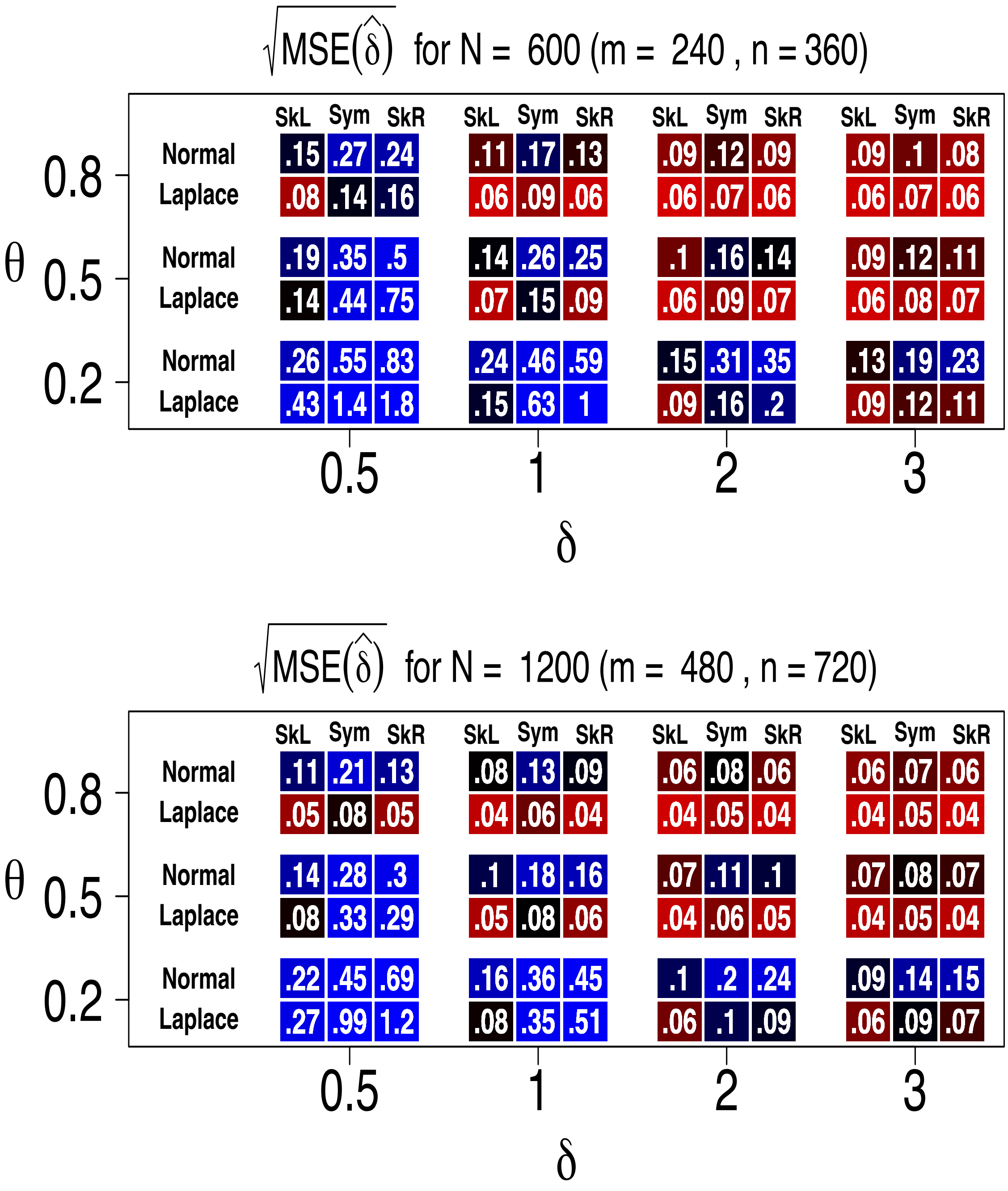


Figure 4.14: Heat Grids for Pseudo-likelihood $\widehat{\delta}$ for $N \in \{180, 300\}$



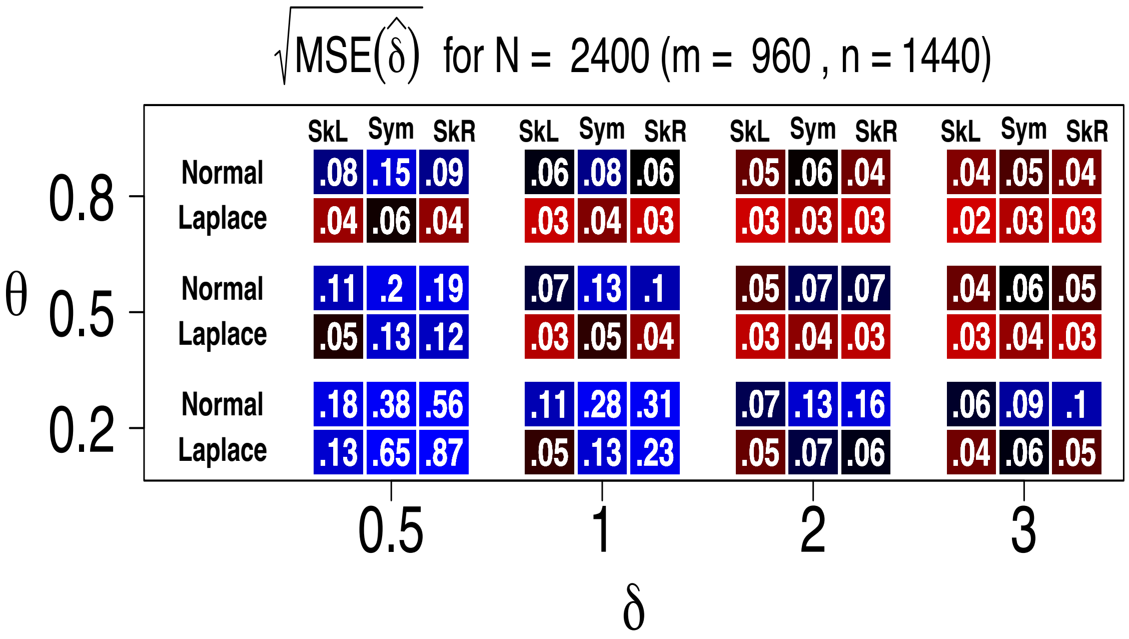
$\sqrt{\mathrm{MSE}(\hat{\delta})}$ for N = 600 (m = 240 , n = 360)

θ		0.5			1			2			3		
		SkL	Sym	SkR	SkL	Sym	SkR	SkL	Sym	SkR	SkL	Sym	SkR
0.8	Normal	.15	.27	.24	.11	.17	.13	.09	.12	.09	.09	.1	.08
	Laplace	.08	.14	.16	.06	.09	.06	.06	.07	.06	.06	.07	.06
0.5	Normal	.19	.35	.5	.14	.26	.25	.1	.16	.14	.09	.12	.11
	Laplace	.14	.44	.75	.07	.15	.09	.06	.09	.07	.06	.08	.07
0.2	Normal	.26	.55	.83	.24	.46	.59	.15	.31	.35	.13	.19	.23
	Laplace	.43	1.4	1.8	.15	.63	1	.09	.16	.2	.09	.12	.11

$\sqrt{\mathrm{MSE}(\hat{\delta})}$ for N = 1200 (m = 480 , n = 720)

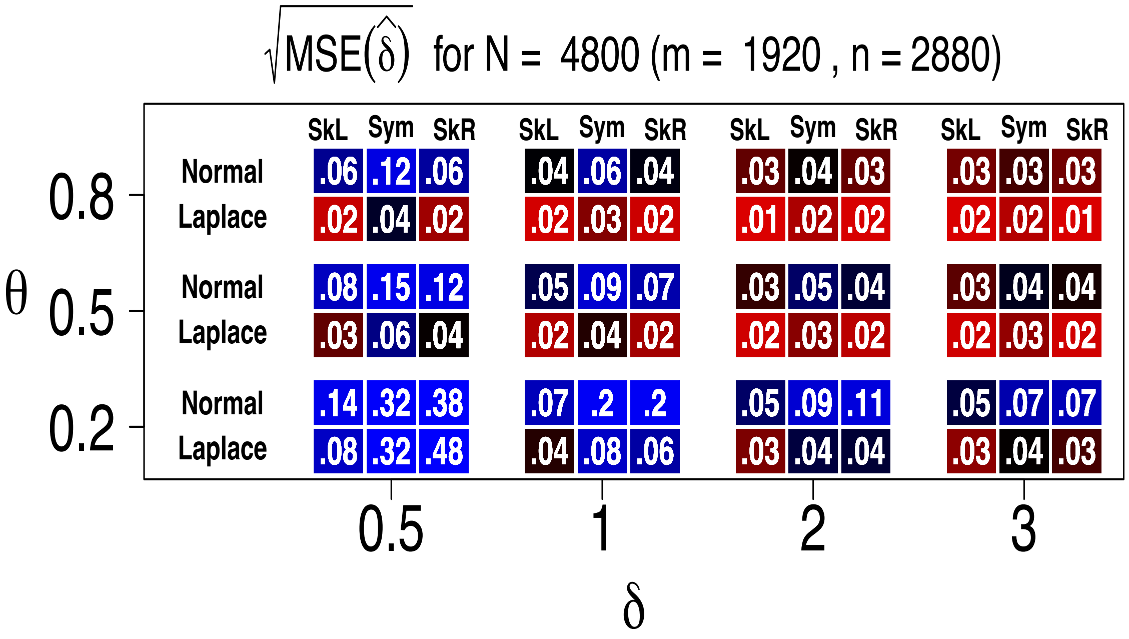
θ		0.5			1			2			3		
		SkL	Sym	SkR	SkL	Sym	SkR	SkL	Sym	SkR	SkL	Sym	SkR
0.8	Normal	.11	.21	.13	.08	.13	.09	.06	.08	.06	.06	.07	.06
	Laplace	.05	.08	.05	.04	.06	.04	.04	.05	.04	.04	.05	.04
0.5	Normal	.14	.28	.3	.1	.18	.16	.07	.11	.1	.07	.08	.07
	Laplace	.08	.33	.29	.05	.08	.06	.04	.06	.05	.04	.05	.04
0.2	Normal	.22	.45	.69	.16	.36	.45	.1	.2	.24	.09	.14	.15
	Laplace	.27	.99	1.2	.08	.35	.51	.06	.1	.09	.06	.09	.07

Figure 4.15: Heat Grids for Pseudo-likelihood $\widehat{\delta}$ for $N \in \{600, 1200\}$



$\sqrt{\text{MSE}(\hat{\delta})}$ for N = 2400 (m = 960 , n = 1440)

θ		δ = 0.5			1			2			3		
		SkL	Sym	SkR	SkL	Sym	SkR	SkL	Sym	SkR	SkL	Sym	SkR
0.8	Normal	.08	.15	.09	.06	.08	.06	.05	.06	.04	.04	.05	.04
	Laplace	.04	.06	.04	.03	.04	.03	.03	.03	.03	.02	.03	.03
0.5	Normal	.11	.2	.19	.07	.13	.1	.05	.07	.07	.04	.06	.05
	Laplace	.05	.13	.12	.03	.05	.04	.03	.04	.03	.03	.04	.03
0.2	Normal	.18	.38	.56	.11	.28	.31	.07	.13	.16	.06	.09	.1
	Laplace	.13	.65	.87	.05	.13	.23	.05	.07	.06	.04	.06	.05



$\sqrt{\text{MSE}(\hat{\delta})}$ for N = 4800 (m = 1920 , n = 2880)

θ		δ = 0.5			1			2			3		
		SkL	Sym	SkR	SkL	Sym	SkR	SkL	Sym	SkR	SkL	Sym	SkR
0.8	Normal	.06	.12	.06	.04	.06	.04	.03	.04	.03	.03	.03	.03
	Laplace	.02	.04	.02	.02	.03	.02	.01	.02	.02	.02	.02	.01
0.5	Normal	.08	.15	.12	.05	.09	.07	.03	.05	.04	.03	.04	.04
	Laplace	.03	.06	.04	.02	.04	.02	.02	.03	.02	.02	.03	.02
0.2	Normal	.14	.32	.38	.07	.2	.2	.05	.09	.11	.05	.07	.07
	Laplace	.08	.32	.48	.04	.08	.06	.03	.04	.04	.03	.04	.03

Figure 4.16: Heat Grids for Pseudo-likelihood $\hat{\delta}$ for $N \in \{2400, 4800\}$

Figures 4.9 - 4.12 show that for the smallest sample sizes, $\widehat{\theta}$ has better performance for larger $\theta\delta$. As the sample size increases, θ has less of an impact on $\sqrt{MSE(\widehat{\theta})}$ than δ does. Notable discrepancies in performance persist even with very large sample sizes (e.g. $N = 4800$) depending upon how well separated the component distributions are. For a fixed (θ, δ), $\sqrt{MSE(\widehat{\theta})}$ is lower for the Laplace-tailed distributions than for the Normal-tailed distributions. The only exception to this is $(\theta, \delta) = (.2, .5)$ which always prefers $F \sim$ Normal instead of $F \sim$ Laplace. Also, right skewed distributions are favorable for larger θ while left skewed distributions are favorable for small θ.

Figures 4.13 - 4.16 show that the performance of $\widehat{\delta}$ is also superior for larger $\theta\delta$. Both θ and δ have a notable impact on $\sqrt{MSE(\widehat{\delta})}$ for all sample sizes, with smaller values of each resulting in higher $\sqrt{MSE(\widehat{\delta})}$. The component distribution, F, has a more prominent impact on the performance of $\widehat{\delta}$ than $\widehat{\theta}$. Skewed left distributions have much smaller $\sqrt{MSE(\widehat{\delta})}$ than symmetric or skewed right distributions, particularly for the more difficult cases (small $\theta\delta$, small sample size). In easy cases ($\theta\delta$ large, F skewed left, larger sample size) the Laplace-tailed distributions are preferred, whereas for more difficult cases the Normal-tailed distributions are preferred. As the sample sizes increase, more of the 72 (F, θ, δ) simulation settings prefer the Laplace-tailed distributions. Only when $(\theta, \delta) = (.2, .5)$ are the symmetric and skewed right Normal distributions preferred to the corresponding symmetric and skewed right Laplace distributions for all sample sizes considered in the simulation.

4.2 Interval Performance Comparison

This section compares confidence intervals for θ and δ. Specifically, the asymptotic moment intervals described in section 3.1.1 and the pseudo-likelihood intervals described in section 3.4 are compared. Note that the asymptotic intervals use $a_N = log(N^2)/N$ as in Lubich et al. (2022) since it provides better performance for the confidence intervals than the $a_N = 20/N^{.95}$ optimized for point estimation in section 2.2.2. The pseudo-likelihood intervals use $\widehat{f}_{mLC}(x)$ from (2.26) to be congruent with the pseudo-likelihood point estimate.

The most ideal confidence interval procedure is one that always captures the true parameter with an arbitrarily small interval. With finite data sets this is not possible, as intervals with such certainty would necessarily contain the entire parameter space. Therefore researchers specify a sufficiently large success rate, called the confidence level (commonly 90% or 95%), for which intervals should capture the parameter. If multiple methods of constructing a confidence interval achieve coverage probabilities $(1 - \alpha)$ at least as large as the researcher's confidence requirement, then the method that produces narrower intervals is preferred.

Therefore the primary criterion to assess the performance of the confidence interval methods is whether coverage probability $(1 - \alpha)$ is sufficiently high, while the secondary criterion is interval length. Since the coverage probability of a confidence interval method may vary depending upon the parameters (F, θ, δ) and these parameters are unknown, it is also important to assess the prevalence of sufficiently high coverage probability across a variety of (F, θ, δ). Thus, coverage probability is assessed via simulation by determining if the method produces a sufficiently high coverage probability for a sufficient number of parameter settings. If this coverage probability assessment is satisfactory for multiple methods, then the average lengths are used to determine the preferable method.

To carry out the performance comparison, 1000 data sets are generated under each of the factorial combinations of the settings from **List 3.1** (and displayed below)

- $N \in \{60, 120, 180, 300, 600, 1200, 2400, 4800\}$
- $n{:}m \in \{1{:}29, 1{:}19, 1{:}14, 1{:}9, 1{:}5, 1{:}3, 1{:}2, 2{:}3, 1{:}1, 3{:}2, 2{:}1, 3{:}1\}$
- $F \in$ {Normal, Laplace, SkRNorm, SkRLap, SkLNorm, SkLLap}
- $\theta \in \{.2, .5, .8\}$
- $\delta \in \{.5, 1, 2, 3\}$

where the choices of F correspond to those described in detail in section A.4 of the Appendix.

In the sections that follow, the performance is summarized across all (F, θ, δ) to provide a recommendation on which method to use. While prior information about (F, θ, δ) may not be readily available, researchers may be able to anticipate whether their proposed treatment has a small or large overall effect size. Thus, the 36 simulation settings for which $\theta\delta \leq .5\sigma_X$ and the 36 settings for which $\theta\delta > .5\sigma_X$ are assessed separately. Assessment is done for each pair of sample sizes $(N, n{:}m)$ so that a sample-size dependent recommendation can be given.

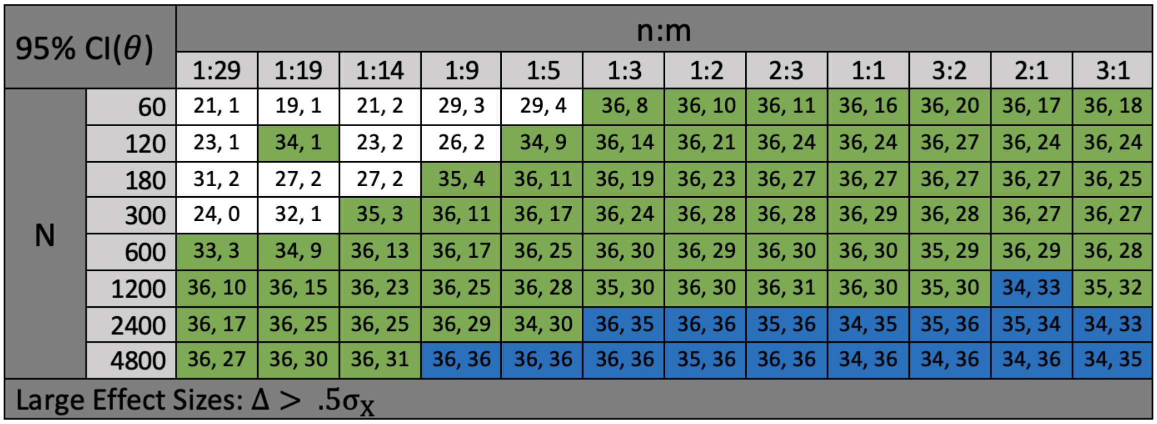
For 95% confidence intervals, coverage probability is said to be sufficiently high if the simulated coverage probability is at least .925 and is considered to apply to a sufficiently wide variety of settings if at least 33/36 simulation settings achieve this. The two methods that are compared according to this criterion are the pseudo-likelihood and method of moment intervals. For a particular sample size setting, when both methods satisfy the coverage probability criterion the method with narrower average intervals in more of the 36 settings is recommended. (Section A.9 of the Appendix verifies that these recommendations may be applied to 90% confidence intervals as well).

Note that very rarely confidence intervals cannot be computed, so performance measures (coverage probability, average length) are computed among data sets where the confidence intervals can be produced. Method of moment confidence intervals for θ fail to compute in 0.0000868% of data sets. Method of moment confidence intervals for δ fail to compute in 0.4588% of data sets. Pseudo-likelihood intervals for θ and for δ each fail to compute in 0.00033% of data sets.

4.2.1 Confidence Intervals for θ

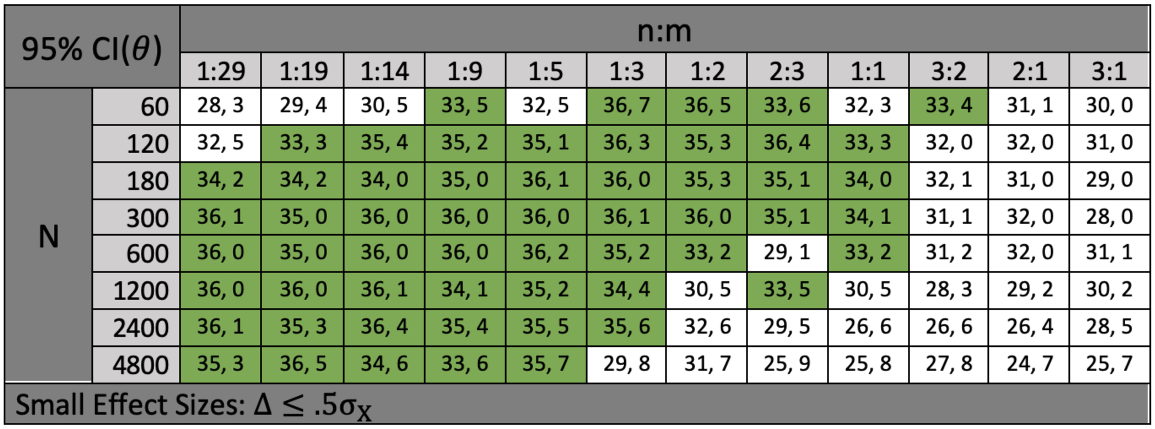
The tables below are indexed by N in the rows and $n{:}m$ in the columns, indicating the pair of sample sizes that the cell represents. In each cell, the pair of numbers represent how many of the 36 settings have sufficient coverage probability for the two methods $-$ $95\% CI_{PsL}(\theta)$ and $95\% CI_{MoM}(\theta)$ respectively. (For example, a cell with entry $-$ $36, 33$ $-$ indicates that 36/36 settings produce coverage probability at least .925 for $95\% CI_{PsL}(\theta)$ and 33/36 settings produce simulated coverage probability at least .925 for $95\% CI_{MoM}(\theta)$.) To aid in pattern recognition, each cell has a background color corresponding to whether or not it meets the coverage probability criterion. Sample size settings where neither method meets the coverage probability criterion have no background color. Sample size settings where both methods meet the coverage probability criterion have blue background color. Sample size settings where only $95\% CI_{MoM}(\theta)$ meets the coverage probability criterion have gold background color. Sample size settings where only $95\% CI_{PsL}(\theta)$ meets the coverage probability criterion have green background color.

Sufficient Coverage Probability Tables



95% CI(θ)		n:m											
		1:29	1:19	1:14	1:9	1:5	1:3	1:2	2:3	1:1	3:2	2:1	3:1
N	60	21, 1	19, 1	21, 2	29, 3	29, 4	36, 8	36, 10	36, 11	36, 16	36, 20	36, 17	36, 18
	120	23, 1	34, 1	23, 2	26, 2	34, 9	36, 14	36, 21	36, 24	36, 24	36, 27	36, 24	36, 24
	180	31, 2	27, 2	27, 2	35, 4	36, 11	36, 19	36, 23	36, 27	36, 27	36, 27	36, 27	36, 25
	300	24, 0	32, 1	35, 3	36, 11	36, 17	36, 24	36, 28	36, 28	36, 29	36, 28	36, 27	36, 27
	600	33, 3	34, 9	36, 13	36, 17	36, 25	36, 30	36, 29	36, 30	36, 30	35, 29	36, 29	36, 28
	1200	36, 10	36, 15	36, 23	36, 25	36, 28	35, 30	36, 30	36, 31	36, 30	35, 30	34, 33	35, 32
	2400	36, 17	36, 25	36, 25	36, 29	34, 30	36, 35	36, 36	35, 36	34, 35	35, 36	35, 34	34, 33
	4800	36, 27	36, 30	36, 31	36, 36	36, 36	36, 36	35, 36	36, 36	34, 36	34, 36	34, 36	34, 35
Large Effect Sizes: $\Delta >$ .5σ_X													

Table 4.5: Each cell entry represents how many of the 36 (F, θ, δ) with $\Delta > .5\sigma_X$ correspond to $95\%CI_{PsL}(\theta), 95\%CI_{MoM}(\theta)$ that have simulated coverage probability at least .925. Color coded backgrounds emphasize when this number is at least 33 - neither method: white, both methods: blue, pseudo-likelihood only: green, method of moments only: gold.



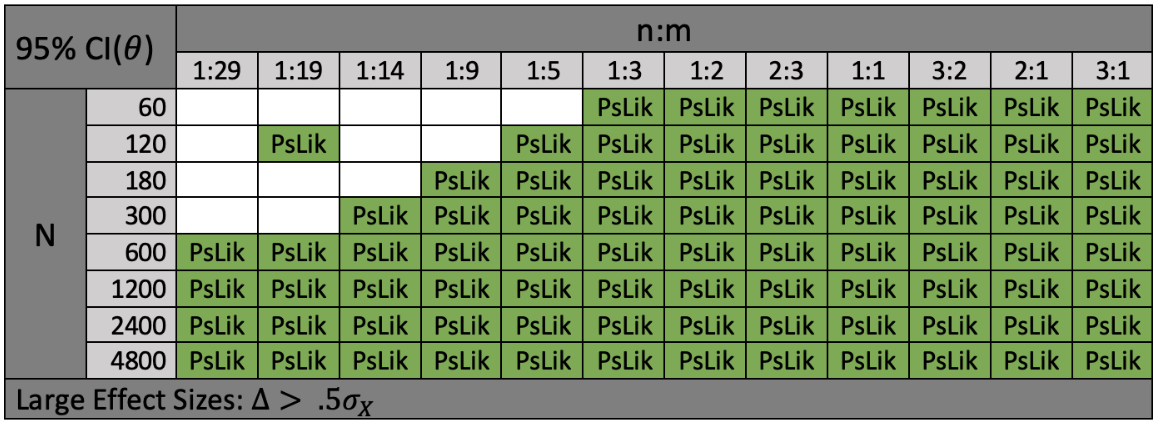
95% CI(θ)		n:m											
		1:29	1:19	1:14	1:9	1:5	1:3	1:2	2:3	1:1	3:2	2:1	3:1
N	60	28, 3	29, 4	30, 5	33, 5	32, 5	36, 7	36, 5	33, 6	32, 3	33, 4	31, 1	30, 0
	120	32, 5	33, 3	35, 4	35, 2	35, 1	36, 3	35, 3	36, 4	33, 3	32, 0	32, 0	31, 0
	180	34, 2	34, 2	34, 0	35, 0	36, 1	36, 0	35, 3	35, 1	34, 0	32, 1	31, 0	29, 0
	300	36, 1	35, 0	36, 0	36, 0	36, 0	36, 1	36, 0	35, 1	34, 1	31, 1	32, 0	28, 0
	600	36, 0	35, 0	36, 0	36, 0	36, 2	35, 2	33, 2	29, 1	33, 2	31, 2	32, 0	31, 1
	1200	36, 0	36, 0	36, 1	34, 1	35, 2	34, 4	30, 5	33, 5	30, 5	28, 3	29, 2	30, 2
	2400	36, 1	35, 3	36, 4	35, 4	35, 5	35, 6	32, 6	29, 5	26, 6	26, 6	26, 4	28, 5
	4800	35, 3	36, 5	34, 6	33, 6	35, 7	29, 8	31, 7	25, 9	25, 8	27, 8	24, 7	25, 7
Small Effect Sizes: $\Delta \leq$ .5σ_X													

Table 4.6: Each cell entry represents how many of the 36 (F, θ, δ) with $\Delta \leq .5\sigma_X$ correspond to $95\%CI_{PsL}(\theta), 95\%CI_{MoM}(\theta)$ that have simulated coverage probability at least .925. Color coded backgrounds emphasize when this number is at least 33 - neither method: white, both methods: blue, pseudo-likelihood only: green, method of moments only: gold.

Tables 4.5 - 4.6 show that at least one method of constructing $95\% CI(\theta)$ provides satisfactory coverage probability for some, but not all, sample size settings. In particular, for large effect sizes, if treatment data is sparse (e.g. $n \leq 15$) then neither method is apt to achieve the coverage probability criterion. Also, for small effect sizes, if the total sample size is either very small (e.g. $N \leq 60$), very large, or control data is proportionally insufficient then neither method produces sufficient coverage probability. For all other sample size scenarios, the pseudo-likelihood intervals have satisfactory coverage probability. For large effect sizes and large enough group sizes, the method of moment confidence intervals are also sufficient.

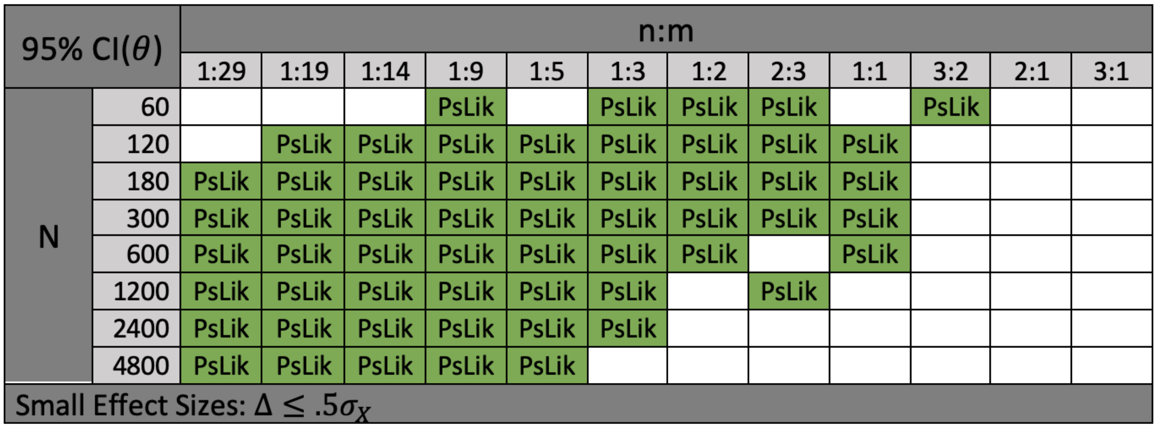
Tables 4.7 - 4.8 below give a sample-size dependent recommendation for which interval method to use by assessing average lengths when both methods satisfy the coverage probability criterion.

Recommendation Tables



95% CI(θ)		n:m											
		1:29	1:19	1:14	1:9	1:5	1:3	1:2	2:3	1:1	3:2	2:1	3:1
N	60						PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik
	120		PsLik			PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik
	180				PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik
	300			PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik
	600	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik
	1200	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik
	2400	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik
	4800	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik
Large Effect Sizes: $\Delta >$ .5σ_X													

Table 4.7: Each cell entry represents the recommended method − $95\%CI_{PsL}(\theta)$ or $95\%CI_{MoM}(\theta)$ (blank white cell means neither method is recommended). The recommended method achieves simulated coverage probability at least .925 for at least 33 of the 36 simulation settings where $\Delta > .5\sigma_X$. If the non-recommended method also meets the coverage probability criterion, the recommended method has smaller average interval length in more settings than the alternate method.



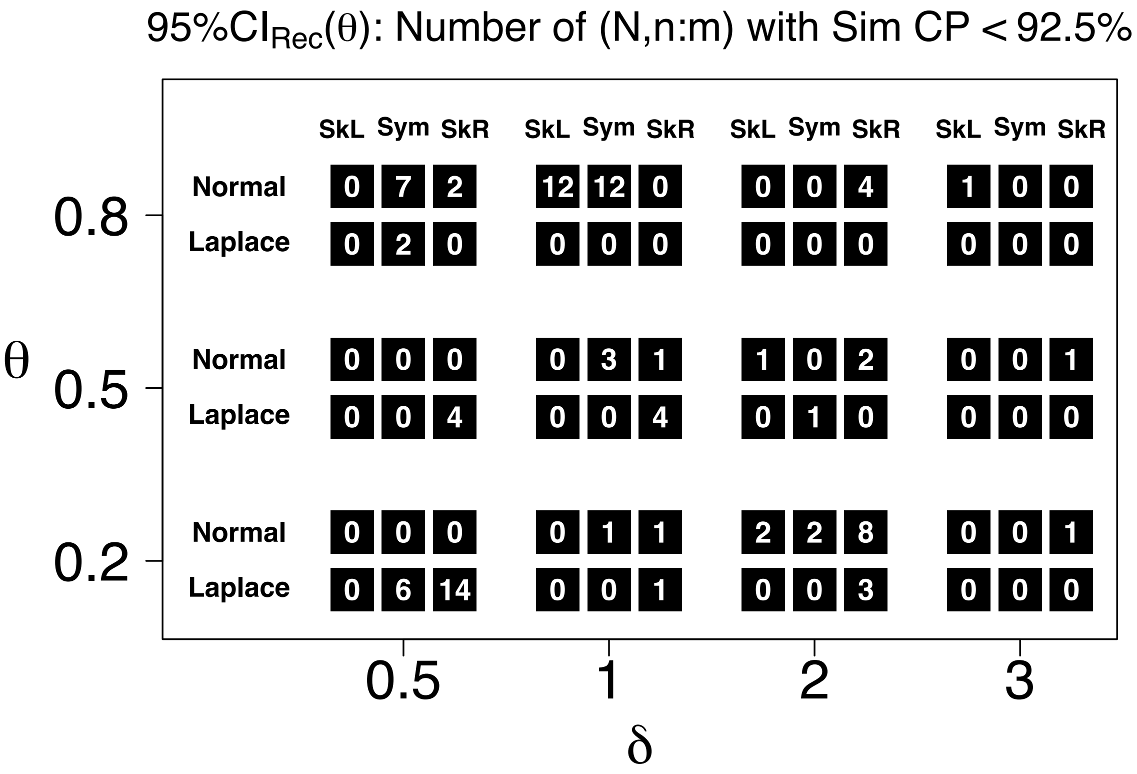
95% CI(θ)		n:m											
		1:29	1:19	1:14	1:9	1:5	1:3	1:2	2:3	1:1	3:2	2:1	3:1
N	60				PsLik		PsLik	PsLik	PsLik		PsLik		
	120		PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik			
	180	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik			
	300	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik			
	600	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik		PsLik			
	1200	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik		PsLik				
	2400	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik						
	4800	PsLik	PsLik	PsLik	PsLik	PsLik							
Small Effect Sizes: $\Delta \leq .5\sigma_X$													

Table 4.8: Each cell entry represents the recommended method − $95\%CI_{PsL}(\theta)$ or $95\%CI_{MoM}(\theta)$ (blank white cell means neither method is recommended). The recommended method achieves simulated coverage probability at least .925 for at least 33 of the 36 simulation settings where $\Delta \leq .5\sigma_X$. If the non-recommended method also meets the coverage probability criterion, the recommended method has smaller average interval length in more settings than the alternate method.

Tables 4.7 - 4.8 indicate that pseudo-likelihood $95\% CI(\theta)$ is always preferable to the method of moment $95\% CI(\theta)$. For the scenarios (e.g. $n \leq 15$, $\Delta > .5\sigma_X$ or $N \geq 2400$, $n/N \geq .40$, $\Delta \leq .5\sigma_X$) where neither method is recommended, the projection of confidence regions for (θ, δ) onto θ can be used (see section 4.3). More specifically, let the projected confidence interval include $\theta^{'}$ if and only if there exists a $\delta^{'}$ such that $(\theta^{'}, \delta^{'}) \in 95\% CR(\theta, \delta)$. This approach produces a $95\% CI_{Proj}(\theta)$ with conservative probability [since $P\left(\theta \in CI_{Proj}(\theta)\right) \geq P\left(\theta \in CR(\theta, \delta)\right)$].

Parameter Specific Performance

While the sample sizes $(N, m{:}n)$ alone can be observed and pre-determined, it is of interest to identify any patterns that may exist regarding the kinds of parameter values (F, θ, δ) that result in insufficient coverage probability. **Figure 4.17** below identifies which (F, θ, δ) results in simulated coverage probability too low (less than .925 for 95% confidence intervals) for the recommended method in **Tables 4.7 - 4.8** (aggregated across all possible sample sizes). When using the recommended method, 97.72% of simulated settings result in satisfactory coverage probability while only 2.28% do not. The parameter settings of those 2.28% are displayed below.



θ		SkL	Sym	SkR	SkL	Sym	SkR	SkL	Sym	SkR	SkL	Sym	SkR
0.8	Normal	0	7	2	12	12	0	0	0	4	1	0	0
	Laplace	0	2	0	0	0	0	0	0	0	0	0	0
0.5	Normal	0	0	0	0	3	1	1	0	2	0	0	1
	Laplace	0	0	4	0	0	4	0	1	0	0	0	0
0.2	Normal	0	0	0	0	1	1	2	2	8	0	0	1
	Laplace	0	6	14	0	0	1	0	0	3	0	0	0
δ			0.5			1			2			3	

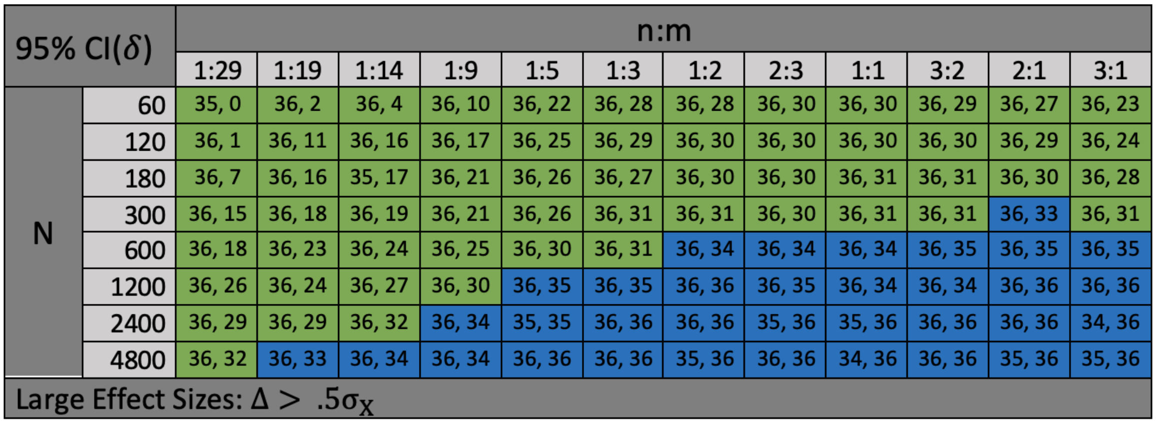
Figure 4.17: Each square corresponds to one (F, θ, δ), as labeled by the axes. The number in each square represents how many of the sample size settings $(N, n{:}m)$ produce insufficient coverage probability in the simulation by using the recommendations in **Tables** 4.7 **-** 4.8.

Figure 4.17 indicates that achieving satisfactory coverage probability is most difficult when δ is small and θ is close to a boundary (either 0 or 1). In particular, if θ is near 1 then the lighter-tailed (Normal-tailed) distributions are more likely to produce low coverage probability. However, if θ is near 0, then the distributions with a heavy and long upper tail are more likely to produce lower than nominal coverage probability. For the other scenarios, the confidence intervals frequently have satisfactory coverage probability.

4.2.2 Confidence Intervals for δ

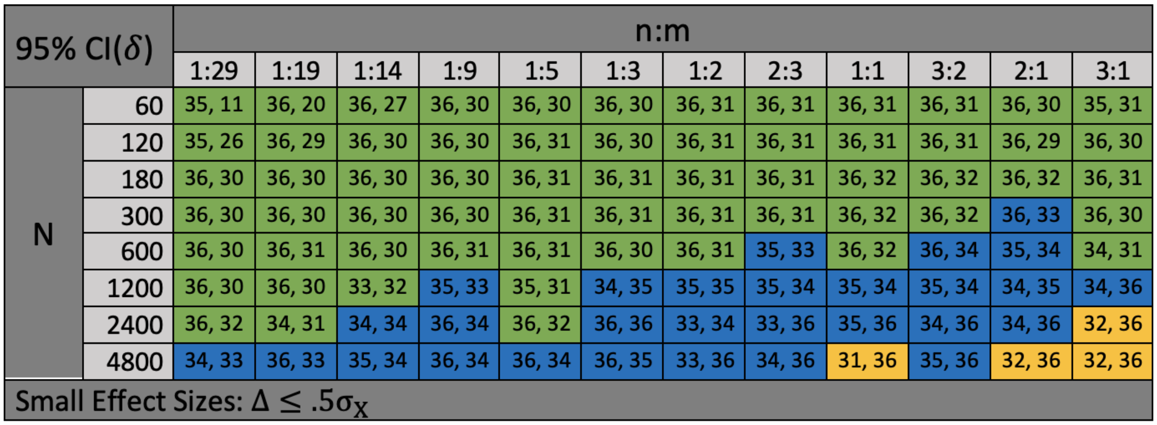
The tables below are indexed by N in the rows and n:m in the columns, indicating the pair of sample sizes that the cell represents. In each cell, the pair of numbers represent how many of the 36 settings have sufficient coverage probability for the two methods $-$ $95\%CI_{PsL}(\delta)$ and $95\%CI_{MoM}(\delta)$ respectively. (For example, a cell with entry $-$ $36, 33$ $-$ indicates that 36/36 settings produce coverage probability at least .925 for $95\%CI_{PsL}(\delta)$ and 33/36 settings produce coverage probability at least .925 for $95\%CI_{MoM}(\delta)$.) To aid in pattern recognition, each cell has a background color corresponding to whether or not it meets the coverage probability criterion. Sample size settings where neither method meets the coverage probability criterion have no background color. Sample size settings where both methods meet the coverage probability criterion have blue background color. Sample size settings where only $95\%CI_{MoM}(\delta)$ meets the coverage probability criterion have gold background color. Sample size settings where only $95\%CI_{PsL}(\delta)$ meets the coverage probability criterion have green background color.

Sufficient Coverage Probability Tables



95% CI(δ)		n:m											
		1:29	1:19	1:14	1:9	1:5	1:3	1:2	2:3	1:1	3:2	2:1	3:1
N	60	35, 0	36, 2	36, 4	36, 10	36, 22	36, 28	36, 28	36, 30	36, 30	36, 29	36, 27	36, 23
	120	36, 1	36, 11	36, 16	36, 17	36, 25	36, 29	36, 30	36, 30	36, 30	36, 30	36, 29	36, 24
	180	36, 7	36, 16	35, 17	36, 21	36, 26	36, 27	36, 30	36, 30	36, 31	36, 31	36, 30	36, 28
	300	36, 15	36, 18	36, 19	36, 21	36, 26	36, 31	36, 31	36, 30	36, 31	36, 31	36, 33	36, 31
	600	36, 18	36, 23	36, 24	36, 25	36, 30	36, 31	36, 34	36, 34	36, 34	36, 35	36, 35	36, 35
	1200	36, 26	36, 24	36, 27	36, 30	36, 35	36, 35	36, 36	36, 35	36, 34	36, 34	36, 36	36, 36
	2400	36, 29	36, 29	36, 32	36, 34	35, 35	36, 36	36, 36	35, 36	35, 36	36, 36	36, 36	34, 36
	4800	36, 32	36, 33	36, 34	36, 34	36, 36	36, 36	35, 36	36, 36	34, 36	36, 36	35, 36	35, 36
Large Effect Sizes: $\Delta >$ $.5\sigma_X$													

Table 4.9: Each cell entry represents how many of the 36 (F, θ, δ) with $\Delta > .5\sigma_X$ correspond to $95\%CI_{PsL}(\delta), 95\%CI_{MoM}(\delta)$ that have simulated coverage probability at least .925. Color coded backgrounds emphasize when this number is at least 33 - neither method: white, both methods: blue, pseudo-likelihood only: green, method of moments only: gold.



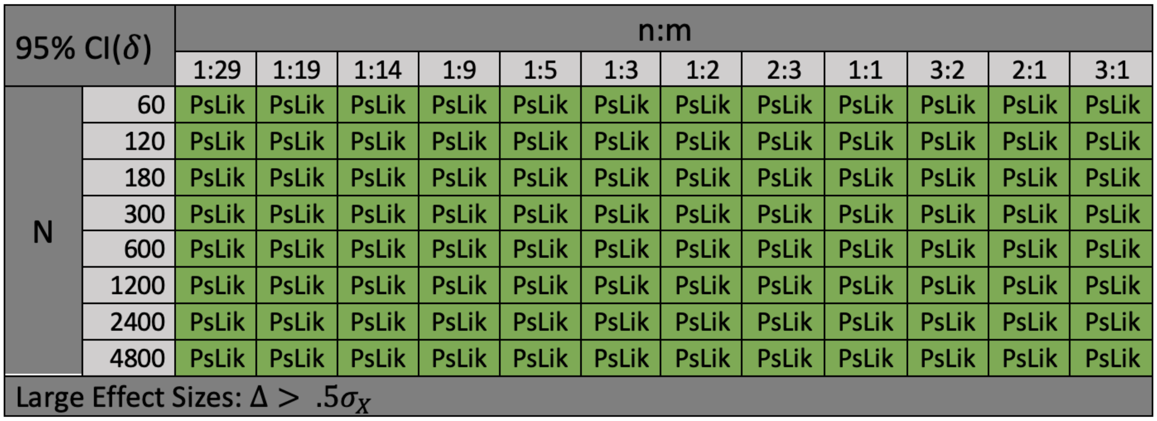
95% CI(δ)		n:m											
		1:29	1:19	1:14	1:9	1:5	1:3	1:2	2:3	1:1	3:2	2:1	3:1
N	60	35, 11	36, 20	36, 27	36, 30	36, 30	36, 30	36, 31	36, 31	36, 31	36, 31	36, 30	35, 31
	120	35, 26	36, 29	36, 30	36, 30	36, 31	36, 30	36, 31	36, 31	36, 31	36, 31	36, 29	36, 30
	180	36, 30	36, 30	36, 30	36, 30	36, 31	36, 31	36, 31	36, 31	36, 32	36, 32	36, 32	36, 31
	300	36, 30	36, 30	36, 30	36, 30	36, 31	36, 31	36, 31	36, 31	36, 32	36, 32	36, 33	36, 30
	600	36, 30	36, 31	36, 30	36, 31	36, 31	36, 30	36, 31	35, 33	36, 32	36, 34	35, 34	34, 31
	1200	36, 30	36, 30	33, 32	35, 33	35, 31	34, 35	35, 35	35, 34	35, 34	35, 34	34, 35	34, 36
	2400	36, 32	34, 31	34, 34	36, 34	36, 32	36, 36	33, 34	33, 36	35, 36	34, 36	34, 36	32, 36
	4800	34, 33	36, 33	35, 34	36, 34	36, 34	36, 35	33, 36	34, 36	31, 36	35, 36	32, 36	32, 36
Small Effect Sizes: $\Delta \leq .5\sigma_X$													

Table 4.10: Each cell entry represents how many of the 36 (F, θ, δ) with $\Delta \leq .5\sigma_X$ correspond to $95\%CI_{PsL}(\delta), 95\%CI_{MoM}(\delta)$ that have simulated coverage probability at least .925. Color coded backgrounds emphasize when this number is at least 33 - neither method: white, both methods: blue, pseudo-likelihood only: green, method of moments only: gold.

Tables 4.9 - 4.10 show that at least one method of constructing $95\% CI(\delta)$ provides satisfactory coverage probability for all sample size settings. For one third of the settings, both the pseudo-likelihood and method of moments meet the coverage probability criterion. Scenarios where method of moments does not achieve sufficient coverage performance include when the treatment group size is not large (e.g. $n \leq 160$). If the effect size is small, total sample size is large, and group allocation favors the treatment group, then the pseudo-likelihood intervals for δ may fail to achieve satisfactory coverage probabilities.

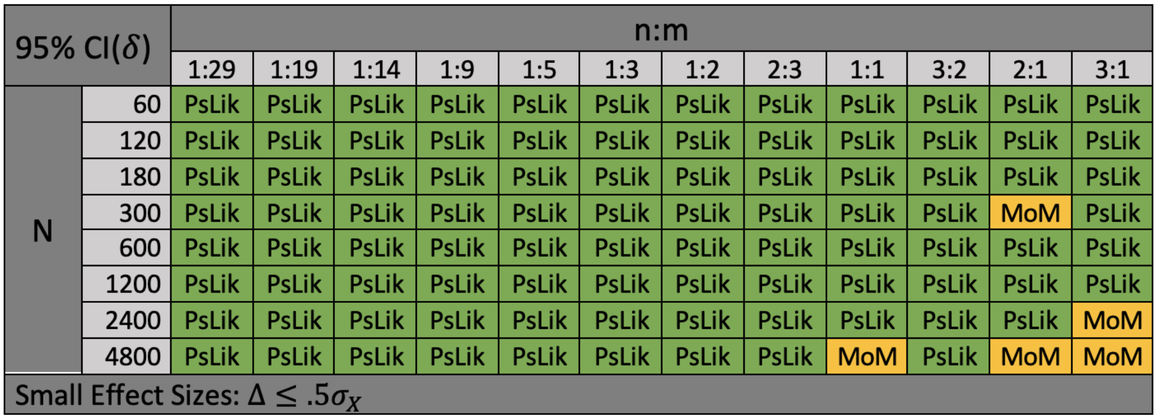
Tables 4.11 - 4.12 below give a sample-size dependent recommendation for which interval method to use by breaking the ties using average lengths when both methods satisfy the coverage probability criterion. Since intervals for δ can occasionally be very large for both methods (and can be infinite for pseudo-likelihood, see section A.7 of the Appendix), all intervals for δ are truncated above at $6S_X$ when comparing lengths. (Recall that an effect size of $\delta = 6\sigma_X$ is a utopianly high effect size.)

Recommendation Tables



95% CI(δ)		n:m											
		1:29	1:19	1:14	1:9	1:5	1:3	1:2	2:3	1:1	3:2	2:1	3:1
N	60	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik
	120	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik
	180	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik
	300	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik
	600	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik
	1200	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik
	2400	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik
	4800	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik
Large Effect Sizes: $\Delta >\ .5\sigma_X$													

Table 4.11: Each cell entry represents the recommended method – $95\%CI_{PsL}(\delta)$ or $95\%CI_{MoM}(\delta)$. The recommended method achieves simulated coverage probability at least .925 for at least 33 of the 36 simulation settings where $\Delta > .5\sigma_X$. If the non-recommended method also meets the coverage probability criterion, the recommended method has smaller average interval length in more settings than the alternate method.



95% CI(δ)		n:m											
		1:29	1:19	1:14	1:9	1:5	1:3	1:2	2:3	1:1	3:2	2:1	3:1
N	60	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik
	120	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik
	180	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik
	300	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	MoM	PsLik
	600	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik
	1200	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik
	2400	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	MoM
	4800	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	MoM	PsLik	MoM	MoM
Small Effect Sizes: $\Delta \leq .5\sigma_X$													

Table 4.12: Each cell entry represents the recommended method – $95\%CI_{PsL}(\delta)$ or $95\%CI_{MoM}(\delta)$. The recommended method achieves simulated coverage probability at least .925 for at least 33 of the 36 simulation settings where $\Delta \leq .5\sigma_X$. If the non-recommended method also meets the coverage probability criterion, the recommended method has smaller average interval length in more settings than the alternate method.

Tables 4.11 **-** 4.12 indicate that the pseudo-likelihood method is recommended rather than method of moments for $95\% CI(\delta)$ in nearly any sample size setting except when the total sample size is very large, group allocation favors the treatment group.

Parameter Specific Performance

While the sample sizes $(N, m{:}n)$ alone can be observed and pre-determined, it is of interest to identify any patterns that may exist regarding the kinds of parameter values (F, θ, δ) that result in insufficient coverage probability. **Figure** 4.18 below identifies which (F, θ, δ) results in simulated coverage probability too low (less than .925 for 95% confidence intervals) for the recommended method in **Tables** 4.11 **-** 4.12 (aggregated across all possible sample sizes). When using the recommended method, 99.12% of simulated settings result in satisfactory coverage probability while only 0.88% do not. The parameter settings of those 0.88% are displayed below.

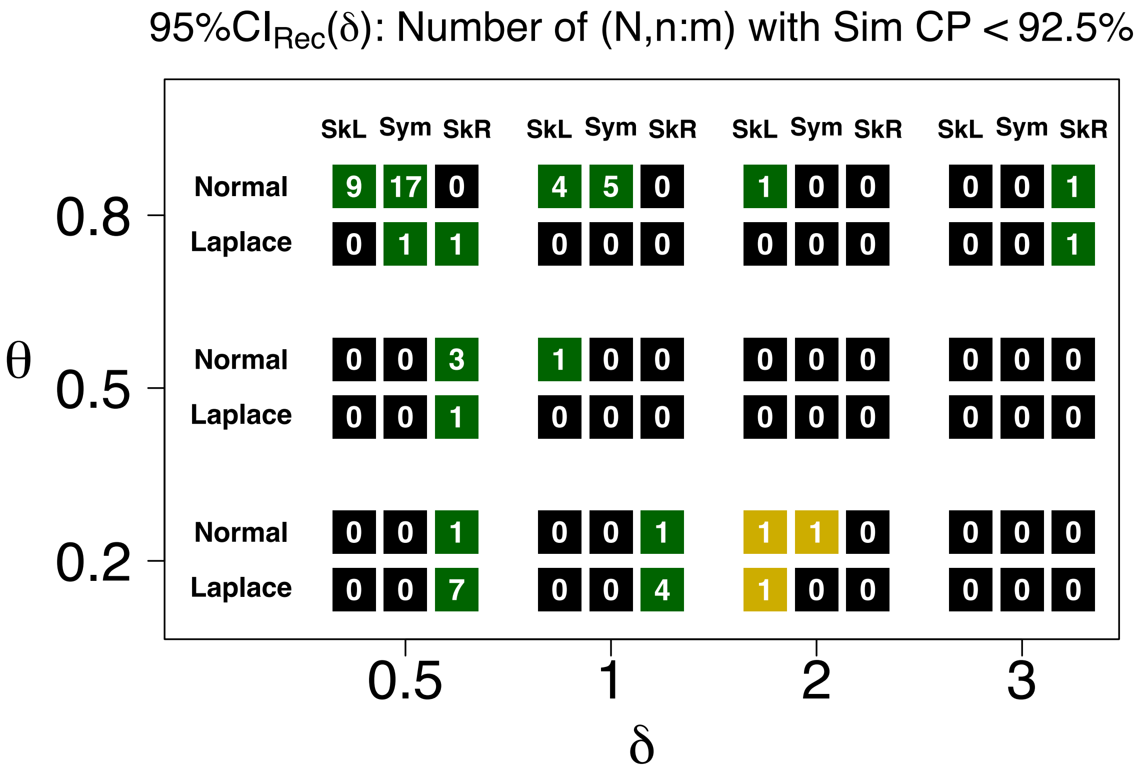


Figure 4.18: Each square corresponds to one (F, θ, δ), as labeled by the axes. The number in each square represents how many of the sample size settings $(N, n{:}m)$ produce insufficient coverage probability in the simulation by using the recommendations in **Tables** 4.11 **-** 4.12. The squares with a green background indicate that more than 3/4 of the time, the insufficient coverage probability occurs when using the pseudo-likelihood interval. The squares with a gold background indicate that more than 3/4 of the time, the insufficient coverage probability occurs when using the method of moment interval.

Figure 4.18 indicates that when θ is near 1 and δ is small, then the recommendation is also more likely to produce lower than nominal coverage probability. For all other scenarios, the confidence intervals almost always have satisfactory coverage probability.

4.3 Confidence Region Performance Comparison

Analogous to the ideal confidence interval procedure, the ideal procedure for constructing a confidence region for (θ, δ) is one that always captures the true parameter pair with an arbitrarily small region. Since this is not possible with finite data sets, a researcher may specify a sufficiently large success rate, called the confidence level (commonly 90% or 95%), for which regions should capture the parameter. If multiple methods of constructing a confidence region achieve coverage probabilities $(1 - \alpha)$ at least as large as the researcher's confidence requirement, then the method that produces smaller areas is preferred.

Therefore the primary criterion to assess the performance of the confidence region methods is whether coverage probability $(1 - \alpha)$ is sufficiently high across a sufficient number of parameter space settings (F, θ, δ). (For 95% confidence regions, simulated coverage probability at least .925 for at least 33/36 settings is defined as satisfactory.) When both methods have satisfactory coverage probability, the method with the smaller average area is preferred. (Areas are based on confidence regions truncated by $\delta \leq 6S_X$ since both confidence region methods compared below can be unbounded in δ [the method of moment region always is].)

To carry out the performance comparison, 1000 data sets are generated under each of the factorial combinations of the the settings in **List 3.1** (and displayed below)

- $N \in \{60, 120, 180, 300, 600, 1200, 2400, 4800\}$
- $n{:}m \in \{1{:}29, 1{:}19, 1{:}14, 1{:}9, 1{:}5, 1{:}3, 1{:}2, 2{:}3, 1{:}1, 3{:}2, 2{:}1, 3{:}1\}$
- $F \in \{$Normal, Laplace, SkRNorm, SkRLap, SkLNorm, SkLLap$\}$
- $\theta \in \{.2, .5, .8\}$
- $\delta \in \{.5, 1, 2, 3\}$

In the sections that follow, the performance is summarized across all (F, θ, δ) to provide a recommendation on which method to use. The 36 simulation settings for which $\theta\delta \leq .5\sigma_X$ and the 36 settings for which $\theta\delta > .5\sigma_X$ are assessed separately. An assessment is done for each pair of sample sizes $(N, n{:}m)$ so that a sample-size dependent recommendation can be given. The two methods of constructing confidence regions compared are the method of moment region corresponding to $95\%CI_{MoM}(\Delta)$ and the pseudo-likelihood $95\%CR_{PsL}(\theta, \delta)$. (This method of moment region is chosen because it is more competitive with $95\%CR_{PsL}(\theta, \delta)$ than the region found by intersecting two confidence intervals. See section A.8 of the Appendix for a comparison of the two types of method moment regions described in section 3.2). Note that very rarely confidence regions cannot be computed, so performance measures (coverage probability, average area) are computed among data sets where the confidence regions can be produced. The method of moment region never fails to compute. The pseudo-likelihood region fails to compute in 0.00033% of data sets.

Tables 4.13 - 4.14 below are indexed by N in the rows and $n{:}m$ in the columns, indicating the pair of sample sizes that the cell represents. In each cell, the pair of numbers represent how many of the 36 settings have sufficient coverage probability for the two methods $-$ $95\%CR_{PsL}(\theta, \delta)$ and $95\%CR_{MoM\Delta}(\theta, \delta)$ respectively. (For example, a cell with entry $-$ $36, 33$ $-$ indicates that 36/36 settings produce coverage probability at least .925 for $95\%CR_{PsL}(\theta, \delta)$ and 33/36 settings produce coverage probability at least .925 for $95\%CR_{MoM\Delta}(\theta, \delta)$.) To aid in pattern recognition, cells have colored background according to whether or not they meet the coverage probability criterion. Sample size settings where neither method meets the coverage probability criterion have no background color. Sample size settings where both methods meet the coverage probability criterion have blue background color. Sample size settings where only $95\%CR_{MoM\Delta}(\theta, \delta)$ meets the coverage probability criterion have gold background color. Sample size settings where only $95\%CR_{PsL}(\theta, \delta)$ meets the coverage probability criterion have green background color. (Section A.9 of the Appendix verifies that these recommendations may be applied to 90% confidence regions as well).

Sufficient Probability Tables

95% CR(θ, δ)		n:m											
		1:29	1:19	1:14	1:9	1:5	1:3	1:2	2:3	1:1	3:2	2:1	3:1
N	60	29, 0	36, 0	36, 2	36, 7	36, 15	36, 33	36, 36	36, 36	36, 35	36, 36	36, 36	36, 36
	120	33, 1	31, 3	36, 6	36, 16	36, 34	36, 36	36, 36	36, 36	36, 36	36, 36	36, 36	36, 36
	180	27, 3	35, 12	36, 19	36, 27	36, 36	36, 36	36, 36	36, 36	36, 36	36, 36	36, 36	36, 36
	300	31, 11	36, 28	36, 30	36, 35	36, 36	36, 36	36, 36	36, 36	36, 36	36, 36	36, 36	35, 36
	600	34, 29	36, 32	36, 36	36, 35	36, 36	36, 36	36, 36	36, 36	36, 36	36, 36	36, 36	35, 36
	1200	36, 36	36, 35	36, 36	36, 36	36, 36	35, 36	36, 36	36, 36	36, 36	36, 36	35, 36	34, 36
	2400	36, 36	36, 36	36, 36	36, 36	35, 36	36, 36	35, 36	34, 36	35, 36	34, 36	36, 36	34, 36
	4800	36, 36	36, 36	36, 36	36, 36	36, 36	36, 36	36, 36	36, 36	33, 36	35, 36	34, 36	32, 36
Large Effect Sizes: $\Delta > .5\sigma_X$													

Table 4.13: Each cell entry represents how many of the 36 (F, θ, δ) with $\Delta > .5\sigma_X$ correspond to $95\%CR_{PsL}(\theta, \delta), 95\%CR_{MoM\Delta}(\theta, \delta)$ that have simulated coverage probability at least .925. Color coded backgrounds emphasize when this number is at least 33 - neither method: white, both methods: blue, pseudo-likelihood only: green, method of moments only: gold.

95% CR(θ, δ)		n:m											
		1:29	1:19	1:14	1:9	1:5	1:3	1:2	2:3	1:1	3:2	2:1	3:1
N	60	36, 1	36, 9	36, 24	36, 28	36, 36	36, 36	36, 36	36, 36	36, 36	36, 36	36, 36	33, 36
	120	36, 19	36, 29	36, 31	36, 35	36, 35	36, 36	36, 36	36, 36	36, 36	36, 36	36, 36	33, 36
	180	36, 25	36, 29	36, 33	36, 34	36, 36	36, 36	36, 36	36, 36	36, 36	36, 36	36, 36	34, 36
	300	36, 29	36, 33	36, 36	36, 36	36, 36	36, 36	36, 36	36, 36	36, 36	35, 36	36, 36	30, 36
	600	36, 34	36, 35	36, 36	36, 36	36, 36	35, 36	36, 36	34, 36	35, 36	35, 36	34, 36	31, 36
	1200	35, 36	36, 36	35, 36	34, 36	35, 36	34, 36	33, 36	31, 36	30, 36	33, 36	32, 36	25, 36
	2400	35, 36	34, 36	35, 36	35, 36	35, 36	36, 36	33, 36	30, 36	31, 36	29, 36	30, 36	25, 36
	4800	34, 36	36, 36	33, 36	36, 36	34, 36	32, 36	30, 36	26, 36	27, 36	27, 36	25, 36	21, 36
Small Effect Sizes: $\Delta \leq .5\sigma_X$													

Table 4.14: Each cell entry represents how many of the 36 (F, θ, δ) with $\Delta \leq .5\sigma_X$ correspond to $95\%CR_{PsL}(\theta, \delta), 95\%CR_{MoM\Delta}(\theta, \delta)$ that have simulated coverage probability at least .925. Color coded backgrounds emphasize when this number is at least 33 - neither method: white, both methods: blue, pseudo-likelihood only: green, method of moments only: gold.

Tables 4.13 - 4.14 indicate that at least one confidence region provides satisfactory coverage probability for nearly any sample size setting. For most settings, both the pseudo-likelihood and method of moment confidence regions produce satisfactory coverage probability. If treatment data is sparse (e.g. $n \leq 30$) then the method of moment region may be unsatisfactory and if it is very sparse (e.g. $n \leq 10$) then the pseudo-likelihood may be as well. If the total sample size is very large and the group allocation is nearly balanced or favors the treatment group, then the pseudo-likelihood coverage probability may be unsatisfactory $-$ particularly if the overall effect size is small.

Recommendation Tables

95% CR(θ, δ)		n:m											
		1:29	1:19	1:14	1:9	1:5	1:3	1:2	2:3	1:1	3:2	2:1	3:1
N	60		PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	MoM	MoM	MoM
	120	PsLik		PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik
	180		PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik
	300		PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik
	600	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik
	1200	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik
	2400	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik
	4800	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	MoM
Large Effect Sizes: $\Delta >$ $.5\sigma_X$													

Table 4.15: Each cell entry represents the recommended method – $95\%CR_{PsL}(\theta, \delta)$ or $95\%CR_{MoM}(\theta, \delta)$ (blank white cell means neither method is recommended). The recommended method achieves simulated coverage probability at least .925 for at least 33 of the 36 simulation settings where $\Delta > .5\sigma_X$. If the non-recommended method also meets the coverage probability criterion, the recommended method has smaller average area in more settings than the alternate method.

95% CR(θ, δ)		n:m											
		1:29	1:19	1:14	1:9	1:5	1:3	1:2	2:3	1:1	3:2	2:1	3:1
N	60	PsLik	PsLik	PsLik	PsLik	MoM	MoM	MoM	MoM	MoM	MoM	MoM	MoM
	120	PsLik	PsLik	PsLik	MoM	PsLik	PsLik	PsLik	PsLik	PsLik	Tie	MoM	PsLik
	180	PsLik	PsLik	Tie	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik
	300	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	MoM
	600	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	MoM
	1200	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	MoM	MoM	PsLik	MoM	MoM
	2400	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	PsLik	MoM	MoM	MoM	MoM	MoM
	4800	PsLik	PsLik	PsLik	PsLik	PsLik	MoM	MoM	MoM	MoM	MoM	MoM	MoM
Small Effect Sizes: $\Delta \leq$ $.5\sigma_X$													

Table 4.16: Each cell entry represents the recommended method – $95\%CR_{PsL}(\theta, \delta)$ or $95\%CR_{MoM}(\theta, \delta)$. The recommended method achieves simulated coverage probability at least .925 for at least 33 of the 36 simulation settings where $\Delta \leq .5\sigma_X$. If the non-recommended method also meets the coverage probability criterion, the recommended method has smaller average area in more settings than the alternate method.

Tables 4.15 - 4.16 show which confidence region method is recommended for each sample size setting. For large effect sizes, the pseudo-likelihood regions are preferred for all sample size settings except sometimes when $m \leq 40$ or when $N \geq 4800$, $n/N \geq .75$. For small overall effect sizes, the pseudo-likelihood method is almost always recommended except sometimes when N is very small or N is very large with group allocation that is nearly even or treatment heavy.

Parameter Specific Performance

While the sample sizes $(N, m{:}n)$ alone can be observed and pre-determined, it is of interest to identify any patterns that may exist regarding the kinds of parameter values (F, θ, δ) that result in insufficient coverage probability. **Figure 4.19** below identifies which (F, θ, δ) results in simulated coverage probability too low (less than .925 for 95% confidence regions) for the recommended method in **Tables 4.15 - 4.16** (aggregated across all possible sample sizes). When using the recommended method, 98.97% of simulated settings result in satisfactory coverage probability while only 1.03% do not. The parameter settings of those 1.03% are displayed below.

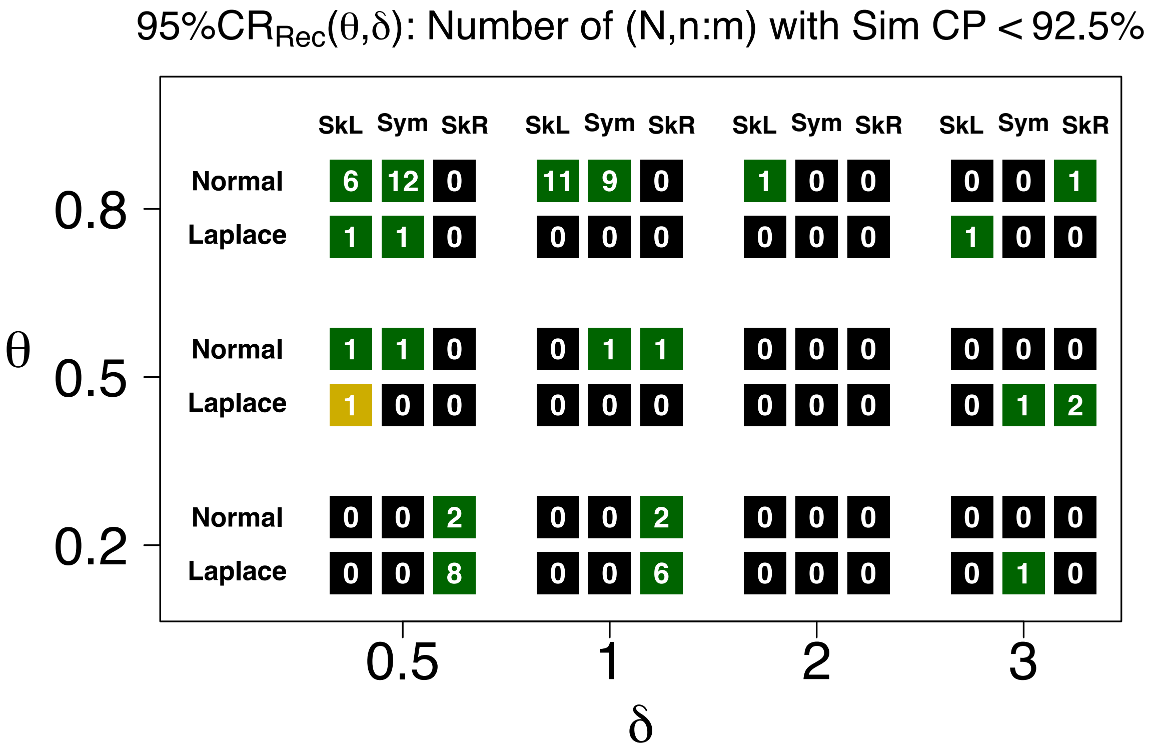


Figure 4.19: Each square corresponds to one (F, θ, δ), as labeled by the axes. The number in each square represents how many of the sample size settings $(N, n{:}m)$ produce insufficient coverage probability in the simulation by using the recommendations in **Tables** 4.15 **-** 4.16. The squares with a green background indicate that more than 3/4 of the time, the insufficient coverage probability occurs when using the pseudo-likelihood region. The squares with a gold background indicate that more than 3/4 of the time, the insufficient coverage probability occurs when using the method of moment region.

Figure 4.19 indicates coverage probability may drop below nominal when δ is small. Particularly some difficulty arises if θ is close to 1 and F has lighter (normal) tails or if θ is small and F is skewed right. For the other scenarios, the confidence regions frequently have satisfactory coverage probability.

Chapter 5

Conclusion

5.1 Example

Note that a small portion of section 5.1 below is identical to content from previously published work (Lubich et al., 2022).

To conclude exploration of inference approaches for (1.1), consider the following demonstration of the recommended analysis of an example blood pressure data set provided by Kaiser Permamente's Electronic Blood Pressure Study (Green et al., 2008). In this study $m = 246$ patients did not receive collaborative care management support provided by clinical pharmacists via the Web, while $n = 237$ patients did. Patients were randomly assigned to the two groups. Summary statistics of the reduction in DBP for the two groups are $\overline{X} = 3.793, S_X^2 = 71.78, \overline{Y} = 6.354, S_Y^2 = 89.73, N = 246 + 237 = 483$. **Figure 5.1** below displays histograms of the responses for the two groups.

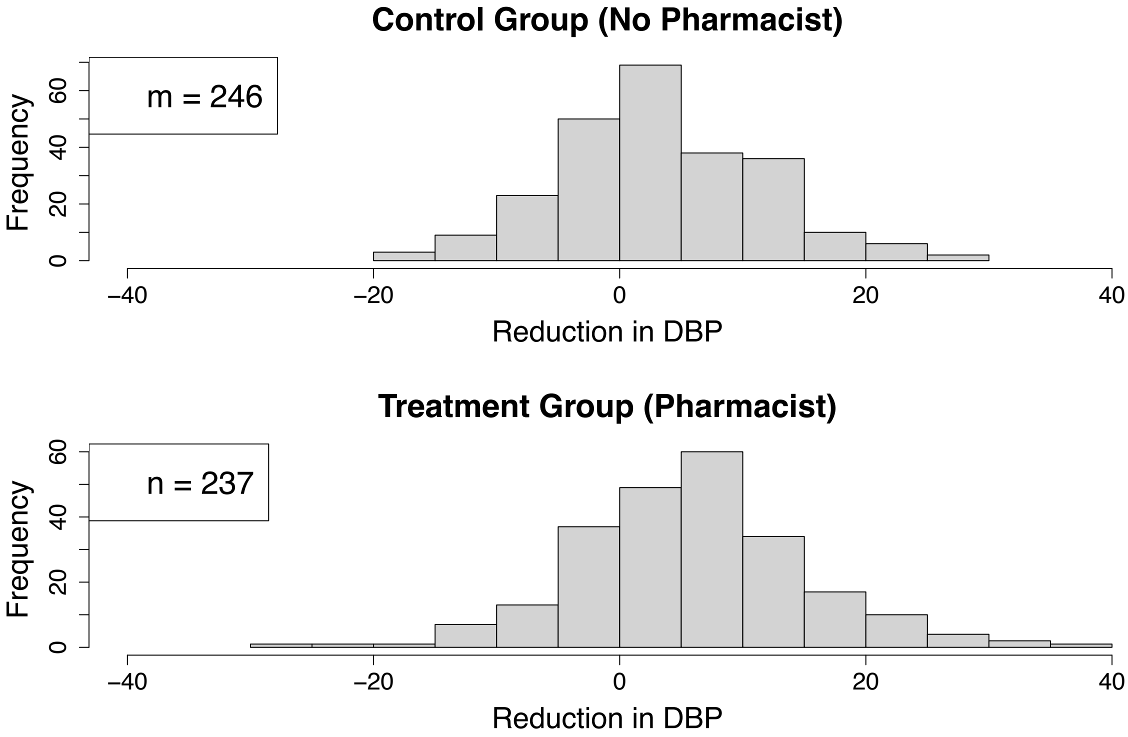


Figure 5.1: Reduction in Diastolic Blood Pressure by Group.

Consider an approach to modeling the data with (1.1). In this context, θ represents the proportion of hypertensive patients that respond to additional pharmacist intervention and δ represents the magnitude of the reduction in DBP for those responding patients. The observed difference in average reduction in DBP is $\overline{Y} - \overline{X} = 2.56$, which under the naive assumption that $\theta = 1$ estimates the effect of the pharmacist intervention for the entire treated population. Section 4.1 indicates that the pseudo-likelihood point estimate is always the preferred method. Sections 4.2 and 4.3 provide scenario-dependent recommendations for the preferred method. Since $\overline{Y} - \overline{X} = 2.56$ is an estimate of Δ and $2.56 = .3S_X \leq .5S_X$, the data suggests that $\Delta < .5\sigma_X$. Therefore **Tables 4.8, 4.12** and **4.16** provide the recommended method for constructing confidence intervals and a confidence region, respectively. The sample size scenario here is nearest to the ($N = 600, n{:}m = 1{:}1$) table entries and thus each table recommends use of the pseudo-likelihood method for confidence intervals and a confidence region, respectively. (While the data strongly suggests that $\Delta \leq .5\sigma_X$, the corresponding tables with $\Delta > .5\sigma_X$ also recommend the pseudo-likelihood method for all inference procedures).

The point estimate for is $\widehat{(\theta, \delta)}_{PsL} = (0.48, 5.93)$, with an estimated average effect size of $\widehat{\Delta}_{PsL} = 2.85$. The point estimate indicates that just under half of patients benefit from the pharmacist intervention and the magnitude for those who benefit is about a 6 mmHg reduction in DBP. For patients who do benefit, the estimated treatment effect $\widehat{\delta}_{PsL} = 5.93$ is more than double the naive estimate $\overline{Y} - \overline{X} = 2.56$ that assumes an effect on the entire treated population. To test the model's goodness of fit, the estimated probability integral transform is applied to the treatment data using the estimate $\widehat{(\theta, \delta)}_{PsL}$ and the emperical CDF of the control data $\widehat{F}_m$ forming the set of $U_i = (1 - \widehat{\theta})\widehat{F}_m(y_i) + \widehat{\theta}\widehat{F}_m(y_i - \widehat{\delta})$ for $i \in \{1, ..., n\}$. This follows approximately a uniform distribution as shown in **Figure 5.2**, indicating that the model is a good fit for the data.

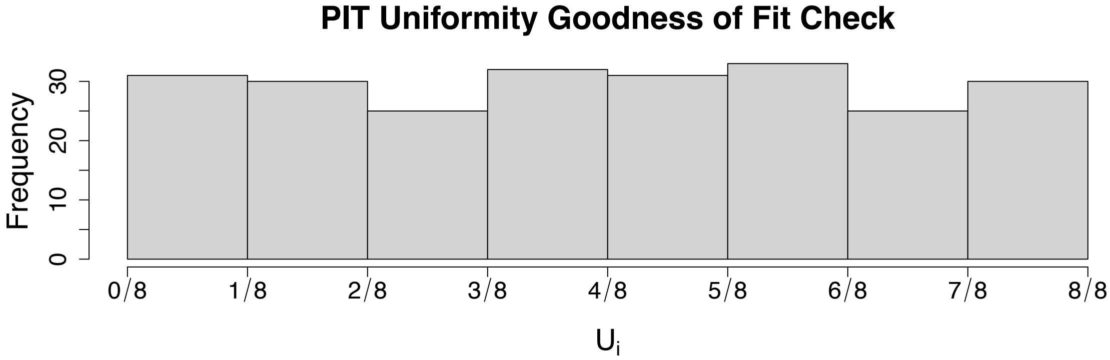


Figure 5.2: Approximately Uniform PIT Transformation of Treatment Data.

Table 5.1 below displays 90% and 95% confidence intervals for the parameters of interest.

Level	$CI(\theta)$	$CI(\delta)$
90%	[.23,.72]	[3.6,10.0]
95%	[.19,.84]	[2.6,11.7]

Table 5.1: Pseudo-likelihood confidence intervals.

While the treatment effect is only fully characterized by the pair (θ, δ), it is possible that the primary interest may be inference on a single parameter. Consider a scenario where θ alone may be of interest. Suppose that Kaiser Permanente has already implemented the additional pharmacist intervention as a component of their standard care for the population of their members. If a large proportion of patients experience some benefit from the treatment (say, at least 85%), then it may not be worthwhile to look for alternate treatment options for the small subset (e.g. less than 15%) of non-responders for whom the availability of additional pharmacist care has no effect on reducing DBP. However, if only half of members benefit (and half do not) then it may be worth trying to identify features of members who would benefit and those that would not. Since the $90\% CI(\theta) = [.23, .72]$ indicates that between 23% and 72% of patients respond (meaning that between 28% and 77% do not benefit), there is evidence of a substantial proportion of treated patients do not benefit. Therefore, it may prove useful to characterize the kinds of patients who will not benefit from the intervention so that they may be cost-effectively referred to a treatment option that is more likely to provide a reduction in DBP.

Similarly, consider a scenario where δ alone may be of interest. Suppose Kaiser already has confidence in their ability to later identify which sub-population of members will benefit from a treatment and is only interested in determining if the treatment has a clinically meaningful effect for the correct sub-population to be referred to this service. Then 90% $CI_{PsL}(\delta) = [3.6, 10.0]$ provides the desired information, indicating that the effect of the clinical pharmacist support is between a 3.6 and 10.0 mmHg reduction in DBP. For example, if a 3.0 mmHg reduction in DBP is considered clinically meaningful then the confidence interval indicates that this treatment has a clinically meaningful effect on responding patients.

To quantify the uncertainty surrounding the full treatment effect (θ, δ) consider **Figure 5.3** below that displays the 90% Pseudo-likelihood confidence region for (θ, δ).

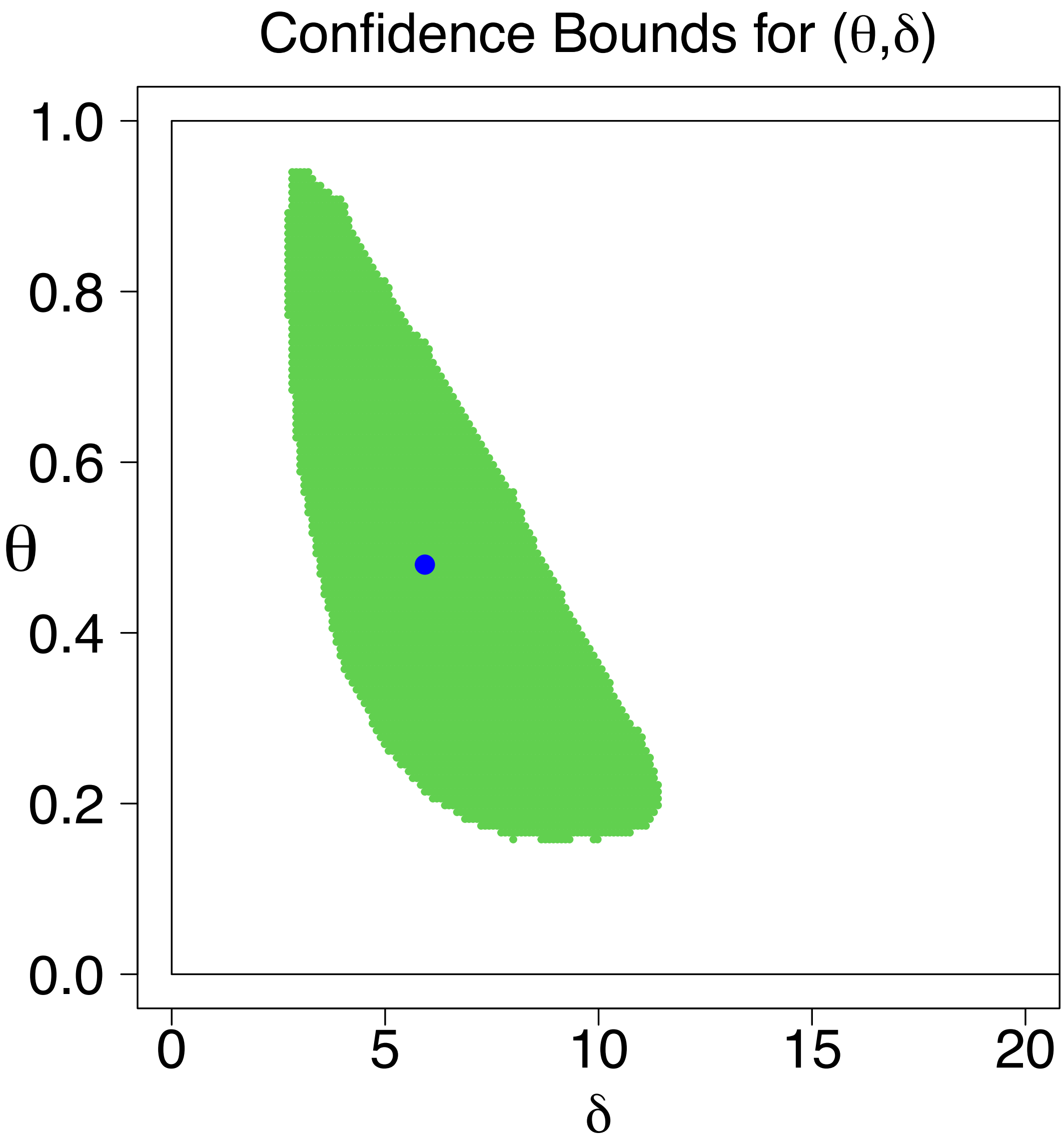


Figure 5.3: 90% Pseudo-likelihood confidence region for (θ, δ). The blue dot near the center of the region is $\widehat{(\theta, \delta)}_{PsL} = (0.48, 5.93)$.

Table 5.2 below displays points that encompass the edge of the 90% Pseudo-likelihood confidence region for (θ, δ).

δ_l	θ	δ_u	θ_l	δ	θ_u
8.0	.15	10.0	.77	2.7	.89
6.0	.2	11.4	.57	3	.95
4.6	.3	10.7	.35	4	.91
3.8	.4	9.6	.26	5	.82
3.3	.5	8.5	.20	6	.74
3.0	.6	7.5	.17	7	.64
2.8	.7	6.3	.15	8	.57
2.7	.8	5.1	.15	9	.46
2.8	.9	4.1	.15	10	.37
2.8	.94	3.2	.17	11	.28
			.20	11.4	.21

Table 5.2: The left side of the table provides the range of δ that lie in the 90% pseudo-likelihood confidence region for an array of $\theta \in \{.15, .2, .3, ..., .9, .94\}$. The right side of the table provides the range of θ that lie in the confidence region for an array of $\delta \in \{2.7, 3, 4, ..., 11, 11.4\}$.

It is important to consider the practical implications of a treatment effect in the confidence region. Consider a few select descriptions of plausible treatment effects from the confidence region. The treatment may only benefit 15% of patients by a magnitude of 8 mmHg. The treatment may benefit 94% of patients with only a 3 mmHg magnitude reduction. The treatment may benefit 60% of patients with a reduction of 7.5 mmHg. The treatment may benefit only 20% of patients but have an effect of reducing DBP by 11.4 mmHg. An important observation about the region is that it does not contain $(\theta, \delta) = (0, 0)$ nor does it contain any points for which $\theta = 1$. Together these observations indicate that there exists a subset of the treated population that does not benefit from the treatment and a subset that does.

5.2 Future Work

More focused exploration on hypothesis testing could prove useful. In particular, it would be interesting to see if a randomization test based on this pseudo-likelihood estimator provides a more powerful test for a treatment effect $-$ H_0: $\theta\delta = 0 \iff F(u) = G(u)$ for all u $-$ than standard non-parametric tests such as an asymptotic Z-test or Wilcoxon Rank Sum test. Also, it would be of interest to investigate a formal test of H_0: $\theta = 1$ by using the pseudo-likelihood ratio test statistic (also with a Satterthwaite approximation for finite-samples). Such a hypothesis test could function as a model-checking test to verify that (1.1) should be used rather than a pure shift alternative. To see if the pseudo-likelihood's efficiency gains over method of moment translate to a multi-stage clinical trial setting, a group sequential clinical trial setting as described in Friel (2022) could utilize the pseudo-likelihood approach. Since all inference procedures in this dissertation fall under the frequentist umbrella, bayesian point estimation and credible regions may be a worthwhile research direction.

An extension (5.1) of model (1.1) allows for the responder distribution to be a location-scale change from the control group rather than assuming only location-shift.

$$G(u) = (1-\theta)F(u) + \theta F\left(\frac{u-\delta}{\gamma}\right) \tag{5.1}$$

Another extension (5.2) of model (1.1) considers the possibility that, in addition to the subset of individuals who do not respond to the treatment, there may also be a subset of individuals for which the treatment is harmful (when contraindications have yet to be established). Such a model can be written as

$$G(u) = \pi_{-1}F(u-\delta_{-1}) + \pi_0 F(u) + \pi_1 F(u-\delta_1), \tag{5.2}$$

where $\pi_i > 0$ for all $i \in \{-1, 0, 1\}$ and $\pi_{-1} + \pi_0 + \pi_1 = 1$ while $\delta_{-1} < 0$ and $\delta_1 > 0$. Without loss of generality, harmful effects are represented by δ_{-1} and beneficial effects are represented by δ_1.

It might be interesting to see how the inference procedures in this dissertation perform for distribution (F) other than the 6 considered in the simulations. Note that (5.1) simplifies to (1.1) if $\gamma = 1$ and (5.2) simplifies to (1.1) when $(\pi_{-1}, \delta_{-1}) = (0, 0)$. It would be interesting to see how useful the inference procedures on (1.1) are if (5.1) is true and γ is slightly different from 1, or if (5.2) is true and (π_{-1}, δ_{-1}) are relatively small.

Bibliography

Edgeworth, F. Y. (1908). On the probable errors of frequency-constants. *Journal of the Royal Statistical Society*, *71*(2), 381–397.

Wilks, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The annals of mathematical statistics*, *9*(1), 60–62.

Yakowitz, S. J., & Spragins, J. D. (1968). On the identifiability of finite mixtures. *The Annals of Mathematical Statistics*, *39*(1), 209–214.

Day, N. E. (1969). Estimating the components of a mixture of normal distributions. *Biometrika*, *56*(3), 463–474.

Hosmer Jr, D. W. (1973). A comparison of iterative maximum likelihood estimates of the parameters of a mixture of two normal distributions under three different types of sample. *Biometrics*, 761–770.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, *39*(1), 1–22.

Hartigan, J. A., & Wong, M. A. (1979). Algorithm as 136: A k-means clustering algorithm. *Journal of the royal statistical society. series c (applied statistics)*, *28*(1), 100–108.

Titterington, D. M., Afm, S., Smith, A. F., Makov, U., et al. (1985). *Statistical analysis of finite mixture distributions* (Vol. 198). John Wiley & Sons Incorporated.

Silverman, B. (1986). Density estimation for statistics and data analysis, chapman and hall, london, 1986. *Crossref, á*.

Efron, B. (1987). Better bootstrap confidence intervals. *Journal of the American statistical Association*, *82*(397), 171–185.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, *20*, 53–65.

Self, S. G., & Liang, K.-Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, *82*(398), 605–610.

Grimlund, R. A. (1989). Panel on nonstandard distributions," statistical models and analysis in auditing"(book review). *The Accounting Review*, *64*(2), 372.

Efron, B., & Tibshirani, R. J. (1994). *An introduction to the bootstrap*. CRC press.

Lindsay, B. G. (1995). Mixture models: Theory, geometry and applications. *NSF-CBMS regional conference series in probability and statistics*, i–163.

Liang, K.-Y., & Self, S. G. (1996). On the asymptotic behaviour of the pseudo-likelihood ratio test statistic. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(4), 785–796.

Theodossiou, P. (1998). Financial data and the skewed generalized t distribution. *Management Science*, *44*(12-part-1), 1650–1661.

Dodge, Y., & Rousson, V. (1999). The complications of the fourth central moment. *The American Statistician*, *53*(3), 267–269.

NAKAMURA, N., & KONISHI, S. (1999). Estimation of number of components for multivariate normal mixture models based on information criteria. *Ouyou toukeigaku*, *27*(3), 165–180.

McLachlan, G., & Peel, D. (2000). *D.(2000), finite mixture models*. Wiley, New York.

Spear, B. B., Heath-Chiozzi, M., & Huff, J. (2001). Clinical application of pharmacogenetics. *Trends in molecular medicine*, *7*(5), 201–204.

Dolan, C. V., Jansen, B. R., & Van der Maas, H. L. (2004). Constrained and unconstrained multivariate normal finite mixture modeling of piagetian data. *Multivariate Behavioral Research*, *39*(1), 69–98.

Cho, E., Cho, M. J., & Eltinge, J. (2005). The variance of sample variance from a finite population. *International Journal of Pure and Applied Mathematics*, *21*(3), 389.

Rokach, L., & Maimon, O. (2005). Clustering methods. *Data mining and knowledge discovery handbook* (pp. 321–352). Springer.

Bordes, L., Delmas, C., & Vandekerkhove, P. (2006). Semiparametric estimation of a two-component mixture model where one component is known. *Scandinavian journal of statistics*, *33*(4), 733–752.

Bordes, L., Mottelet, S., & Vandekerkhove, P. (2006). Semiparametric estimation of a two-component mixture model. *The Annals of Statistics*, *34*(3), 1204–1232. https://doi.org/10.1214/009053606000000353

He, Y., Pan, W., & Lin, J. (2006). Cluster analysis using multivariate normal mixture models to detect differential gene expression with microarray data. *Computational statistics & data analysis*, *51*(2), 641–658.

Lagakos, S. W. et al. (2006). The challenge of subgroup analyses-reporting without distorting. *New England Journal of Medicine*, *354*(16), 1667.

Bordes, L., Chauveau, D., & Vandekerkhove, P. (2007). A stochastic em algorithm for a semiparametric mixture model. *Computational Statistics & Data Analysis*, *51*(11), 5429–5443.

Chang, G. T., & Walther, G. (2007a). Clustering with mixtures of log-concave distributions. *Computational Statistics & Data Analysis*, *51*(12), 6242–6251.

Chang, G. T., & Walther, G. (2007b). Clustering with mixtures of log-concave distributions. *Comput. Stat. Data Anal.*, *51*, 6242–6251.

Hunter, D. R., Wang, S., & Hettmansperger, T. P. (2007). Inference for mixtures of symmetric distributions. *The Annals of Statistics*, 224–251.

Lin, T. I., Lee, J. C., & Hsieh, W. J. (2007). Robust mixture modeling using the skew t distribution. *Statistics and computing*, *17*(2), 81–92.

Zhang, L. (2007). Sample mean and sample variance: Their covariance and their (in) dependence. *The American Statistician*, *61*(2), 159–160.

Green, B. B., Ralston, J. D., Fishman, P. A., Catz, S. L., Cook, A., Carlson, J., Tyll, L., Carrell, D., & Thompson, R. S. (2008). Electronic communications and home blood pressure monitoring (e-bp) study: Design, delivery, and evaluation framework. *Contemporary clinical trials*, *29*(3), 376–395.

Benaglia, T., Chauveau, D., & Hunter, D. R. (2009). An em-like algorithm for semi-and nonparametric estimation in multivariate mixtures. *Journal of Computational and Graphical Statistics*, *18*(2), 505–526.

Boldea, O., & Magnus, J. R. (2009). Maximum likelihood estimation of the multivariate normal mixture model. *Journal of the American Statistical Association*, *104*(488), 1539–1549.

Dümbgen, L., & Rufibach, K. (2009). Maximum likelihood estimation of a log-concave density and its distribution function: Basic properties and uniform consistency. *Bernoulli*, *15*(1), 40–68.

Bordes, L., & Vandekerkhove, P. (2010). Semiparametric two-component mixture model with a known component: An asymptotically normal estimator. *Mathematical Methods of Statistics*, *19*(1), 22–41.

Chen, Y., & Liang, K.-Y. (2010). On the asymptotic behaviour of the pseudo-likelihood ratio test statistic with boundary problems. *Biometrika*, *97*(3), 603–620.

Dümbgen, L., & Rufibach, K. (2011). Logcondens: Computations related to univariate log-concave density estimation. *Journal of Statistical Software*, *39*, 1–28.

Soffritti, G., & Galimberti, G. (2011). Multivariate linear regression with non-normal errors: A solution based on mixture models. *Statistics and Computing*, *21*(4), 523–536.

Diez, D. M., Barr, C. D., & Cetinkaya-Rundel, M. (2012). *Openintro statistics*. OpenIntro Boston, MA, USA:

Fraley, C., Raftery, A. E., Murphy, T. B., & Scrucca, L. (2012). *Mclust version 4 for r: Normal mixture modeling for model-based clustering, classification, and density estimation* (tech. rep.). Technical report.

Hunter, D. R., & Young, D. S. (2012). Semiparametric mixtures of regressions. *Journal of Nonparametric Statistics*, *24*(1), 19–38.

Hohmann, D., & Holzmann, H. (2013). Semiparametric location mixtures with distinct components. *Statistics*, *47*(2), 348–362.

McLachlan, G. J., & Rathnayake, S. (2014). On the number of components in a gaussian mixture model. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *4*(5), 341–355.

Xiang, S., Yao, W., & Wu, J. (2014). Minimum profile hellinger distance estimation for a semiparametric mixture model. *Canadian Journal of Statistics*, *42*(2), 246–267.

Davis, C. (2015). The skewed generalized t distribution tree package vignette.

Manegold, C., Adjei, A., Bussolino, F., Cappuzzo, F., Crino, L., Dziadziuszko, R., Ettinger, D., Fennell, D., Kerr, K., Le Chevalier, T., et al. (2016). Novel active agents in patients with advanced nsclc without driver mutations who have progressed after first-line chemotherapy. *ESMO open*, *1*(6), e000118.

Patra, R. K., & Sen, B. (2016). Estimation of a two-component mixture model with applications to multiple testing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *78*(4), 869–893.

Zeller, C. B., Cabral, C. R., & Lachos, V. H. (2016). Robust mixture regression modeling based on scale mixtures of skew-normal distributions. *Test*, *25*(2), 375–396.

Hu, H., Yao, W., & Wu, Y. (2017). The robust em-type algorithms for log-concave mixtures of regression models. *Computational statistics & data analysis*, *111*, 14–26.

Saxena, A., Prasad, M., Gupta, A., Bharill, N., Patel, O. P., Tiwari, A., Er, M. J., Ding, W., & Lin, C.-T. (2017). A review of clustering techniques and developments. *Neurocomputing*, *267*, 664–681.

Soofi, A. A., & Awan, A. (2017). Classification techniques in machine learning: Applications and issues. *Journal of Basic & Applied Sciences*, *13*, 459–465.

Rosenblatt, J. D., & Benjamini, Y. (2018). On mixture alternatives and wilcoxon's signed-rank test. *The American Statistician*, *72*(4), 344–347.

Samworth, R. J. (2018). Recent progress in log-concave density estimation. *Statistical Science*, *33*(4), 493–509.

Maleki, M., Contreras-Reyes, J. E., & Mahmoudi, M. R. (2019). Robust mixture modeling based on two-piece scale mixtures of normal family. *Axioms*, *8*(2), 38.

McLachlan, G. J., Lee, S. X., & Rathnayake, S. I. (2019). Finite mixture models. *Annual review of statistics and its application*, *6*, 355–378.

Xiang, S., Yao, W., & Yang, G. (2019). An overview of semiparametric extensions of finite mixture models. *Statistical science*, *34*(3), 391–404.

Young, D. S., Chen, X., Hewage, D. C., & Nilo-Poyanco, R. (2019). Finite mixture-of-gamma distributions: Estimation, inference, and model-based clustering. *Advances in Data Analysis and Classification*, *13*(4), 1053–1082.

Burgess-Hull, A. J. (2020). Finite mixture models with student t distributions: An applied example. *Prevention Science*, *21*(6), 872–883.

Jeske, D. R., & Yao, W. (2020). Sample size calculations for mixture alternatives in a control group vs. treatment group design. *Statistics*, *54*(1), 97–113.

Dogan, A., & Birant, D. (2021). Machine learning and data mining in manufacturing. *Expert Systems with Applications*, *166*, 114060.

Friel, D. C. (2022). *Wilcoxon rank sum tests to detect one-sided mixture alternatives in group sequential clinical trials*. University of California, Riverside.

Ilagan, M. J., & Falk, C. F. (2022). Supervised classes, unsupervised mixing proportions: Detection of bots in a likert-type questionnaire.

Lubich, B., Jeske, D., & Yao, W. (2022). Statistical inference for method of moments estimators of a semi-supervised two-component mixture model. *The American Statistician*, 1–8.

Appendix

A.1 Proof of Moments of G (2.6 - 2.9)

To derive equations (2.6) - (2.9), consider the following relationship from model (1.1)

$$Y \overset{d}{=} (1-Z)X + Z(X+\delta)$$

where $Z \sim \text{Bernoulli}(\theta)$ independent of $X \sim F$. This relationship holds because $Y \sim G$ from (1.1). Thus,

$$\begin{aligned} \mu_Y \equiv E[Y] &= E\left[(1-Z)X + Z(X+\delta)\right] \\ &= \mu_X + \theta\delta. \end{aligned} \tag{2.6}$$

To calculate (2.7), we first attain $E[Y^2]$ in terms of (F, θ, δ). Letting $a = (1-Z)X$ and $b = Z(X+\delta)$

$$\begin{aligned} E[Y^2] = E\left[(a+b)^2\right] &= E\left[\left(a^2 + ab + b^2\right)\right] \\ &= E\left[a^2 + b^2\right] \\ &= (1-\theta)\, E[X^2] + \theta E\left[E[X^2] + 2\mu_X\delta + \delta^2\right] \\ &= E[X^2] + 2\mu_X\theta\delta + \theta\delta^2. \end{aligned}$$

Notice that the terms for which both a and b have a non-zero exponents - k_a and k_b - are 0 because $(1-Z)^{k_a}Z^{k_b} = 0$ with probability 1 whenever $k_a > 0$ and $k_b > 0$. Then, we have that

$$\begin{aligned}
\sigma_Y^2 \equiv E[(Y-\mu_Y)^2] &= E[Y^2] - E[Y]^2 \\
&= E[X^2] + 2\theta\delta\mu_X + \theta\delta^2 - (\mu_X^2 + 2\theta\delta\mu_X + \theta^2\delta^2) \\
&= (E[X^2] - \mu_X^2) + (\theta - \theta^2)\delta^2 \\
&= \sigma_X^2 + \theta(1-\theta)\delta^2. \qquad (2.7)
\end{aligned}$$

To calculate (2.8), first attain $E[Y^3]$ in terms of (F, θ, δ).

$$\begin{aligned}
E[Y^3] &= E[(a+b)^3] \\
&= E[a^3 + 3a^2b + 3ab^2 + b^3] \\
&= E[a^3 + b^3] \\
&= E[X^3] + 3\theta\delta E[X^2] + 3\theta\delta^2\mu_X + \theta\delta^3,
\end{aligned}$$

again noting that $a^{k_a}b^{k_b} = 0$ if $k_a > 0$ and $k_b > 0$. Then, we have that

$$\begin{aligned}
\mu_{3cy} \equiv E[(Y-\mu_Y)^3] &= E[Y^3] - 3E[Y^2]\mu_Y + 2\mu_Y^3 \\
&= (E[X^3] - 3E[X^2]\mu_X + 2\mu_X^3) + \theta\delta^3 - 3\theta^2\delta^3 + 2\theta^3\delta^3 \\
&= \mu_{3cx} + \theta(1-\theta)\delta^3[1-2\theta]. \qquad (2.8)
\end{aligned}$$

To calculate (2.9), we first attain $E[Y^4]$ in terms of (F, θ, δ).

$$\begin{aligned}
E[Y^4] = E[(a+b)^4] = E[a^4 + b^4] \\
&= E[X^4] + 4\theta\delta E[X^3] + 6\theta\delta^2 E[X^2] + 4\theta\delta^3 \mu_X + \theta\delta^4,
\end{aligned}$$

again noting that $a^{k_a} b^{k_b} = 0$ if $k_a > 0$ and $k_b > 0$. Then, we have that

$$\begin{aligned}
\mu_{4cy} \equiv E[(Y - \mu_Y)^4] &= E[Y^4] - 4\mu_Y E[Y^3] + 6\mu_Y^2 E[Y^2] - 3\mu_Y^4 \\
&= \big(E[X^4] - 4\mu_X E[X^3] + 6\mu_X^2 E[X^2] - 3\mu_X^4\big) \\
&+ 6\theta\delta^2 E[X^2] + \theta\delta^4 - 6\theta\delta^2\mu_X^2 - 6\theta^2\delta^2 E[X^2] \\
&- 4\theta^2\delta^4 + 6\theta^2\delta^2\mu_X^2 + 6\theta^3\delta^4 - 3\theta^4\delta^4 \\
&= \mu_{4cx} + \theta\delta^4 \left[(1 - 4\theta + 6\theta^2 - 3\theta^3) + 6(1-\theta)(\sigma_X^2/\delta^2)\right] \\
&= \mu_{4cx} + \theta\delta^4 \left[\big((1-3\theta)(1-\theta)^2 + \theta(1-\theta)\big) + 6(1-\theta)(\sigma_X^2/\delta^2)\right] \\
&= \mu_{4cx} + \theta(1-\theta)\delta^4 \left[(1-3\theta)(1-\theta) + \theta + 6\sigma_X^2/\delta^2\right]. \qquad (2.9)
\end{aligned}$$

A.2 Consistency

Consider first a proof of proposition 2.2.2, consistency of both $\widehat{\theta}$ and $\widehat{\delta}$ in estimating θ and δ respectively. The proof shows that the + operator and ϵ_N modifications do not negate the natural consistency of the moment estimator so long as $\epsilon_N \to 0$ as $m, n \to \infty$. First consider $f\left(\overline{X}, \overline{Y}, S_Y^2, S_X^2\right)$, an approximation of $\widehat{\theta}$

$$\widehat{\theta} = \left\{1 + \frac{(S_Y^2 - S_X^2)_+}{(\overline{Y} - \overline{X})_+^2 + \epsilon_N}\right\}^{-1}$$
$$f\left(\overline{X}, \overline{Y}, S_Y^2, S_X^2\right) = \left\{1 + \frac{(S_Y^2 - S_X^2)}{(\overline{Y} - \overline{X})^2}\right\}^{-1}$$

If $m \to \infty$ and $n \to \infty$, then clearly $f\left(\overline{X}, \overline{Y}, S_Y^2, S_X^2\right) \xrightarrow{p} \theta$ since $\left(\overline{Y} - \overline{X}\right)^2 \xrightarrow{p} (\mu_Y - \mu_X)^2$, $(S_Y^2 - S_X^2) \xrightarrow{p} (\sigma_Y^2 - \sigma_X^2)$, and $\theta = \left\{1 + \frac{(\sigma_Y^2 - \sigma_X^2)}{(\mu_Y - \mu_X)^2}\right\}^{-1}$. Thus it suffices to show that $(S_Y^2 - S_X^2)_+ \xrightarrow{p} (\sigma_Y^2 - \sigma_X^2)$ and $\left(\left(\overline{Y} - \overline{X}\right)_+^2 + \epsilon_N\right) \xrightarrow{p} (\mu_Y - \mu_X)$.

Since $(S_Y^2 - S_X^2) \xrightarrow{p} (\sigma_Y^2 - \sigma_X^2)$, this means that $\forall\ \epsilon > 0$ and $\forall\ \omega > 0$, $\exists\ \{m_0, n_0\}$ such that $\forall\ m > m_0$ and $\forall\ n > n_0$

$$\begin{aligned} & P\left(\left|\left(S_Y^2 - S_X^2\right) - \left(\sigma_Y^2 - \sigma_X^2\right)\right| > \epsilon\right) < \omega \\ \Leftrightarrow\ & P\left(\left(S_Y^2 - S_X^2\right) - \left(\sigma_Y^2 - \sigma_X^2\right) < -\epsilon\right) + P\left(\left(S_Y^2 - S_X^2\right) - \left(\sigma_Y^2 - \sigma_X^2\right) > \epsilon\right) < \omega. \end{aligned}$$

Also,

$$\begin{aligned} & P\left(\left|\left(S_Y^2 - S_X^2\right)_+ - \left(\sigma_Y^2 - \sigma_X^2\right)\right| > \epsilon\right) < \omega \\ \Leftrightarrow\ & P\left(\left(S_Y^2 - S_X^2\right)_+ - \left(\sigma_Y^2 - \sigma_X^2\right) < -\epsilon\right) + P\left(\left(S_Y^2 - S_X^2\right)_+ - \left(\sigma_Y^2 - \sigma_X^2\right) > \epsilon\right) < \omega. \end{aligned}$$

so since

$$P\left(\left(S_Y^2 - S_X^2\right)_+ - \left(\sigma_Y^2 - \sigma_X^2\right) < -\epsilon\right) < P\left(\left(S_Y^2 - S_X^2\right) - \left(\sigma_Y^2 - \sigma_X^2\right) < -\epsilon\right),$$

this means that

$$(S_Y^2 - S_X^2) \xrightarrow{p} (\sigma_Y^2 - \sigma_X^2) \implies (S_Y^2 - S_X^2)_+ \xrightarrow{p} (\sigma_Y^2 - \sigma_X^2).$$

An analogous argument shows that $\left(\overline{Y} - \overline{X}\right)_+^2 \xrightarrow{p} (\mu_Y - \mu_X)$. Therefore $\left(\left(\overline{Y} - \overline{X}\right)_+^2 + \epsilon_N\right) \xrightarrow{p} (\mu_Y - \mu_X)$ so $\widehat{\theta} \xrightarrow{p} \theta$. The consistency of $\widehat{\delta}$ immediately follows because $\widehat{\delta} = \left(\overline{Y} - \overline{X}\right)_+ / \widehat{\theta}$ and $\delta = (\mu_Y - \mu_X)/\theta$.

A.3 Asymptotic Normality

Here is a proof for the asymptotic normality and derivation of the asymptotic variance of $\widehat{\theta}$ and $\widehat{\delta}$. First consider proposition 2.2.3 for $\widehat{\theta}$,

$$\begin{aligned} \widehat{\theta} &= \left\{1 + \frac{(S_Y^2 - S_X^2)_+}{(\overline{Y} - \overline{X})_+^2 + \epsilon_N}\right\}^{-1} \\ f\left(\overline{X}, \overline{Y}, S_Y^2, S_X^2\right) &= \left\{1 + \frac{(S_Y^2 - S_X^2)}{(\overline{Y} - \overline{X})^2}\right\}^{-1} \end{aligned}$$

Using a first order taylor series expansion gives

$$\begin{aligned} & f\left(\overline{X}, \overline{Y}, S_X^2, S_Y^2\right) = \\ & f\left(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2\right) \; + \; \frac{\partial f}{\partial \overline{X}}\bigg|_{\overline{X}=\mu_X} \left(\overline{X} - \mu_X\right) \; + \; \frac{\partial f}{\partial \overline{Y}}\bigg|_{\overline{Y}=\mu_Y} \left(\overline{Y} - \mu_Y\right) \\ & \qquad\qquad + \; \frac{\partial f}{\partial S_X^2}\bigg|_{S_X^2=\sigma_X^2} \left(S_X^2 - \sigma_X^2\right) \; + \; \frac{\partial f}{\partial S_Y^2}\bigg|_{S_Y^2=\sigma_Y^2} \left(S_Y^2 - \sigma_Y^2\right) + o(1) \end{aligned}$$

$$\begin{aligned} = f\left(\mu_X, \mu_Y, \sigma_Y^2, \sigma_X^2\right) \; + \; \left\{1 + \frac{(\sigma_Y^2 - \sigma_X^2)}{(\mu_Y - \mu_X)^2}\right\}^{-2} \Bigg\{ & \frac{2(\sigma_Y^2 - \sigma_X^2)(\mu_X - \overline{X})}{(\mu_Y - \mu_X)^3} \\ & + \; \frac{2(\sigma_Y^2 - \sigma_X^2)(\overline{Y} - \mu_Y)}{(\mu_Y - \mu_X)^3} \\ & + \; \frac{S_X^2 - \sigma_X^2}{(\mu_Y - \mu_X)^2} \\ & + \; \frac{\sigma_Y^2 - S_Y^2}{(\mu_Y - \mu_X)} \Bigg\} + o(1) \end{aligned} \tag{3}$$

Now since $\overline{X}, \overline{Y}, S_X^2, S_Y^2$ are all unbiased estimators of $\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2$, respectively, $E[f\left(\overline{X}, \overline{Y}, S_X^2, S_Y^2\right)] = 0$ with accuracy to the first order expansion. Furthermore, since that converge in distribution to a normal distribution by the central limit theorem, $f\left(\overline{X}, \overline{Y}, S_X^2, S_Y^2\right)$ also converges in distribution to a normal with asymptotic variance equal to $Var\left(f\left(\overline{X}, \overline{Y}, S_X^2, S_Y^2\right)\right)$

Now to derive the variance of $\widehat{\theta}$ by taking the variance of (3), begin by noting that any covariance terms between X and Y are 0 because X and Y are independent. Also, utilizing the following variance (Cho et al., 2005) and covariance (Dodge and Rousson, 1999; Zhang, 2007) results

$$Var\left(S^2\right) = \frac{1}{n}\left(\mu_{4c} - \frac{n-3}{n-1}\sigma^4\right) \tag{4}$$

$$Cov\left(\overline{X}, S^2\right) = \frac{\mu_{3c}}{n}. \tag{5}$$

provides a first order taylor series approximate variance of $\widehat{\theta}$

$$\begin{aligned} &Var\left(\widehat{\theta}\right) = \\ &(1+o(1))\left\{1 + \frac{(\sigma_Y^2 - \sigma_X^2)}{(\mu_Y - \mu_X)^2}\right\}^{-4} \left\{\frac{4\left(\sigma_Y^2 - \sigma_X^2\right)^2}{(\mu_Y - \mu_X)^6}\left(\frac{\sigma_X^2}{m} + \frac{\sigma_Y^2}{n}\right)\right. \\ &\qquad - \frac{4(\sigma_Y^2 - \sigma_X^2)}{(\mu_Y - \mu_X)^5}\left(\frac{\mu_{3cx}}{m} + \frac{\mu_{3cy}}{n}\right) \\ &\qquad \left. + \frac{1}{(\mu_Y - \mu_X)^4}\left(\frac{(\mu_{4cx} - \frac{m-3}{m-1}\sigma_X^4)}{m} + \frac{(\mu_{4cy} - \frac{n-3}{n-1}\sigma_Y^4)}{n}\right)\right\}. \end{aligned} \tag{6}$$

The special case of $m = n$ gives the asymptotic variance formula in (2.15) as desired.

Now to prove proposition 2.2.3 for the case of $\widehat{\delta}$

$$\begin{aligned} \widehat{\delta} = \frac{\widehat{\Delta}}{\widehat{\theta}} &= (\overline{Y} - \overline{X})_+ \left\{1 + \frac{(S_Y^2 - S_X^2)_+}{(\overline{Y} - \overline{X})_+^2 + \epsilon_N}\right\} \\ g\left(\overline{X}, \overline{Y}, S_Y^2, S_X^2\right) &= (\overline{Y} - \overline{X})\left\{1 + \frac{(S_Y^2 - S_X^2)}{(\overline{Y} - \overline{X})^2}\right\} \end{aligned}$$

Using a first order taylor series expansion

$$
\begin{aligned}
& g\left(\overline{X}, \overline{Y}, S_Y^2, S_X^2\right) = \\
& g(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2) \ + \ \frac{\partial g}{\partial \overline{X}}\bigg|_{\overline{X}=\mu_X} (\overline{X} - \mu_X) \ + \ \frac{\partial g}{\partial \overline{Y}}\bigg|_{\overline{Y}=\mu_Y} (\overline{Y} - \mu_Y) \\
& \qquad + \frac{\partial g}{\partial S_X^2}\bigg|_{S_X^2=\sigma_X^2} (S_X^2 - \sigma_X^2) + \frac{\partial g}{\partial S_Y^2}\bigg|_{S_Y^2=\sigma_Y^2} (S_Y^2 - \sigma_Y^2) \\
& \qquad + o(1)
\end{aligned}
$$

$$
\begin{aligned}
= g(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2) &+ \Big(\frac{(\sigma_Y^2 - \sigma_X^2)}{(\mu_Y - \mu_X)^2} - 1\Big)(\overline{X} - \mu_X) \\
&+ \Big(1 - \frac{(\sigma_Y^2 - \sigma_X^2)}{(\mu_Y - \mu_X)^2}\Big)(\overline{Y} - \mu_Y) \\
&+ \ \frac{-(S_X^2 - \sigma_X^2)}{(\mu_Y - \mu_X)} \ + \ \frac{(S_Y^2 - \sigma_Y^2)}{(\mu_Y - \mu_X)} + o(1). \qquad (7)
\end{aligned}
$$

Now since $\overline{X}, \overline{Y}, S_X^2, S_Y^2$ are all unbiased estimators of $\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2$ respectively, $E[g\left(\overline{X}, \overline{Y}, S_Y^2, S_X^2\right)] = \delta$ with accuracy to the first order expansion. Furthermore, since $\overline{X}, \overline{Y}, S_X^2, S_Y^2$ each converge in distribution to a normal distribution by the central limit theorem, $f\left(\overline{X}, \overline{Y}, S_X^2, S_Y^2\right)$ also converges in distribution to a normal with asymptotic variance equal to $Var\left(f\left(\overline{X}, \overline{Y}, S_X^2, S_Y^2\right)\right)$. Now to derive the variance of

$\widehat{\delta}$ by taking the variance of (7)

$$\begin{aligned}
&Var\left(\widehat{\delta}\right) = \\
&(1+o(1))\Bigg\{ \left(\frac{(\sigma_Y^2-\sigma_X^2)}{(\mu_Y-\mu_X)^2}-1\right)^2 \left(\frac{\sigma_X^2}{m}+\frac{\sigma_Y^2}{n}\right) \\
&\qquad + \; 2\left(1-\frac{\sigma_Y^2-\sigma_X^2}{(\mu_Y-\mu_X)^2}\right)\frac{(\mu_{3cx}/m+\mu_{3cy}/n)}{(\mu_Y-\mu_X)} \\
&\qquad + \; \frac{1}{(\mu_Y-\mu_X)^2}\left(\frac{(\mu_{4cx}-\frac{m-3}{m-1}\sigma_X^4)}{m}+\frac{(\mu_{4cy}-\frac{n-3}{n-1}\sigma_Y^4)}{n}\right)\Bigg\}. \qquad (8)
\end{aligned}$$

The special case of $m = n$ gives the asymptotic variance formula in (2.16) as desired.

A.4 Simulation Settings

The performance of the methods in the coming chapters are compared across an array of the simulation settings (F, θ, δ). Robust methodology is desired so that performance is satisfactory across the parameter space, which motivates the following primary simulation settings on which to measure performance.

- $F \in \{$Normal, Laplace, Skewed Right Normal (SKRN), Right Skewed Laplace (SKLL), Skewed Left Normal (SKLN), Skewed Left Laplace (SKLL)$\}$
  - All distributions are from the 5 parameter skewed generalized T distribution with $\lambda = 0$ for symmetric distributions, $\lambda = .5$ for right skewed distributions and $\lambda = -.5$ for left skewed distributions. Distributions from the generalized Normal family have parameters $p = 2$, and $q = \infty$ while those from the generalized Laplace family have parameters $p = 1$ and $q = \infty$.
  - $\mu_X = 0$ and $\sigma_X = 1$ for all F.
- $\theta \in \{.2, .5, .8\}$
- $\delta \in \{.5, 1, 2, 3\}$

The six choices of F are chosen to allow for a variety of distributional shapes − in particular to vary tail heaviness and skewness. This distribution has multiple parameterizations. A vignette by Davis (2015), which can be accessed in R programming by the command vignette(“sgt”), displays the parameterization as well as much of the content summarizing this family of distributions described here. Consider how the 5 parameters of the Skewed Generalized T Distribution $(\mu, \sigma, \lambda, p, q)$, introduced by Theodossiou (1998), allow for this kind of flexibility. The skew of the distribution is controlled by $-1 < \lambda < 1$ where $\lambda < 0$ for skewed left distributions, $\lambda = 0$ for symmetric distributions, and $\lambda > 0$ for skewed right distributions. The parameters $p > 0$

and $q > 0$ jointly control the tail behavior (with smaller values of each corresponding to heavier tailed distributions). The density is given by

$$f_{SGT}(x;\mu,\sigma,\lambda,p,q) = p\left\{2v\sigma q^{1/p}B(1/p,q)\left(\frac{|x-\mu+m|^p}{q(v\sigma)^p(\lambda\text{sign}(x-\mu+m)+1)^p}+1\right)^{\frac{1}{p}+q}\right\}^{-1} \tag{9}$$

where $B(\cdot,\cdot)$ represents the Euler Beta function

$$B(x,y) = \int_0^1 t^{x-1}(1-t)^{y-1}dt = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}$$

and

$$m = (B(1/p,q))^{-1}\left(2v\sigma\lambda q^{1/p}B(2/p,q-1/p)\right),$$

$$v = q^{-1/p}\left[(3\lambda^2+1)\left(\frac{B(3/q,q-2/p)}{B(1/p,q)}\right)-4\lambda^2\left(\frac{B(2/q,q-1/p)}{B(1/p,q)}\right)^2\right]^{-1/2}$$

so that μ represent the mean and σ is the standard deviation so long as they exist (the h^{th} moment exists if $pq > h$). See Figure A.1 below for families of distributions that are special cases.

Specifically, in the case of $q \to \infty$, we have the six distributions used for the simulation shown in Table A.1 below.

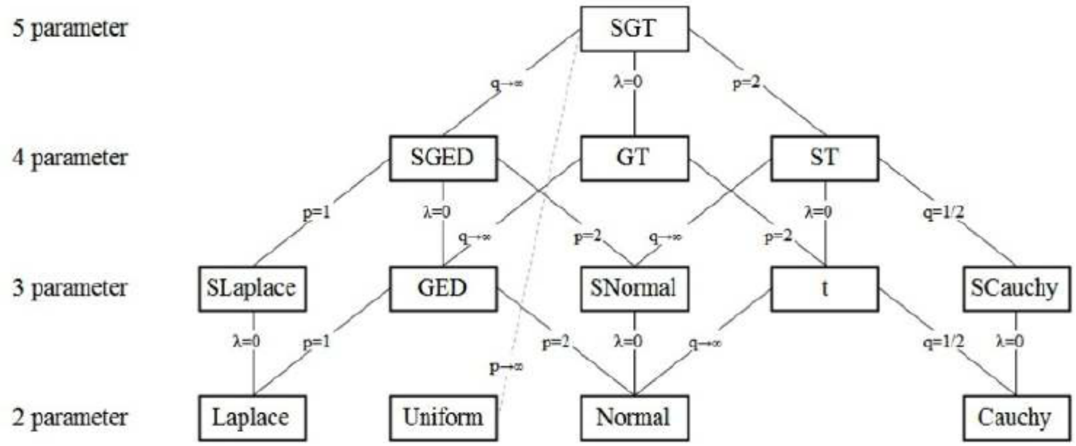


Figure A.1: Diagram of Skewed Generalized T Family Tree

		λ		
		-.5	0	.5
p	2	Skewed Left Normal	Normal	Skewed Right Normal
	1	Skewed Left Laplace	Laplace	Skewed Right Laplace

Table A.1: Distributions used for Simulations

Figure A.2 displays the pdfs for the six distributions in the simulation.

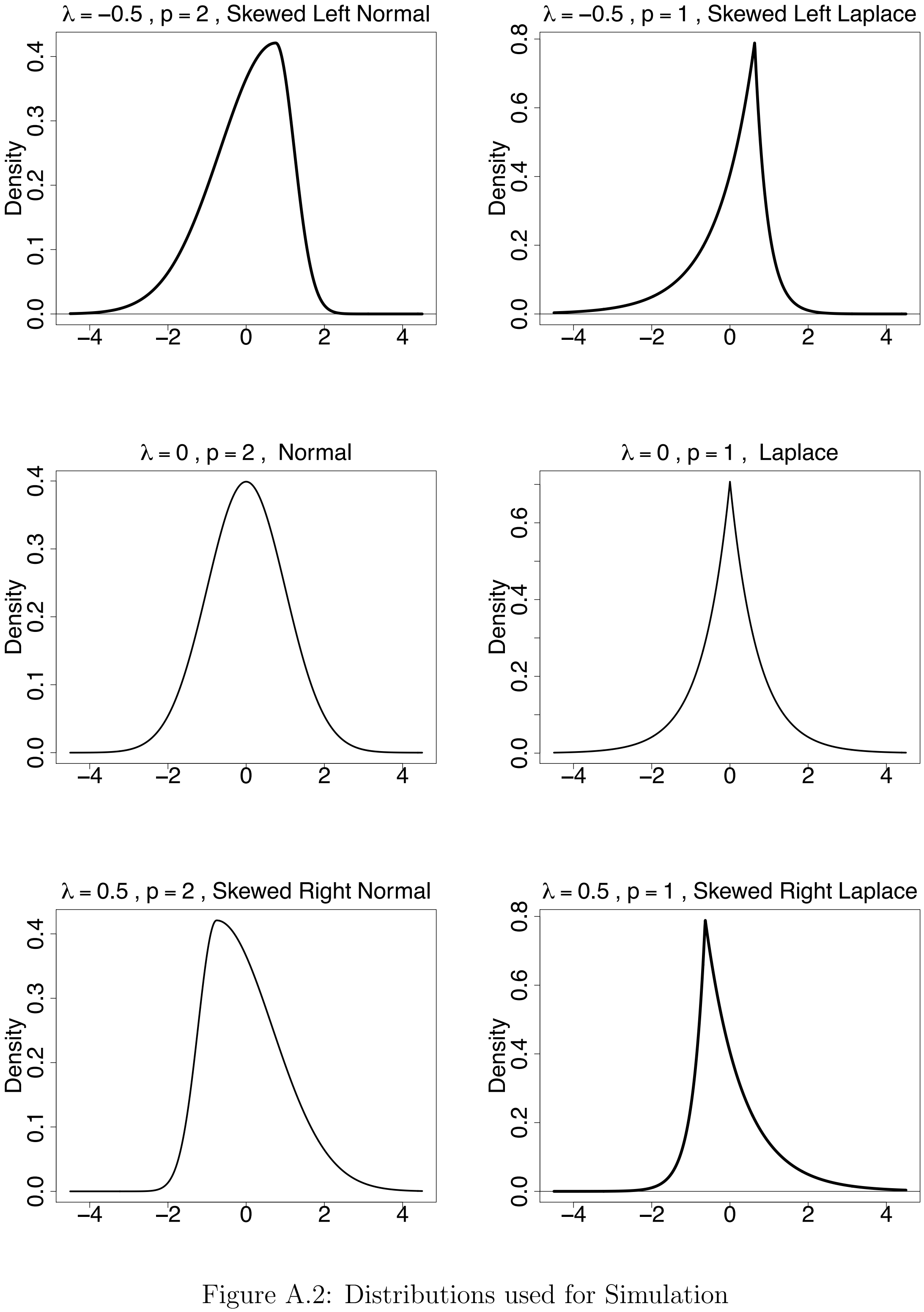


Figure A.2: Distributions used for Simulation

Consider the intuitive measure of tail heaviness in that if $\lim_{x \to +\infty} f_1(x)/f_2(x) = 0$ then f_2 is said to have a heavier upper tail than f_1, while if $\lim_{x \to -\infty} f_1(x)/f_2(x) = 0$ then f_2 is said to have a heavier lower tail than f_1. If both conditions are satisfied, then f_2 is said to be heavier tailed than f_1. Note the difference in the distributions for the Normal tails (Skewed Normal Family $p = 2$) and the Laplace tails (Skewed Laplace Family $p = 1$) by considering the special cases of (9) for the Skewed Normal and Skewed Laplace families presented below.

$$f_{SNorm}(x; \mu, \sigma, \lambda) = (v_n \sigma \sqrt{\pi})^{-1} exp \left(-\frac{|x - \mu + m_n|}{v_n \sigma (1 + \lambda sign(x - \mu + m_n))} \right)^2 \tag{10}$$

where

$$m_n = \frac{2 v_n \sigma \lambda}{\sqrt{\pi}} \tag{11}$$

and

$$v_n = \sqrt{2\pi} \left[(\pi - 8\lambda^2 + 3\pi\lambda^2) \right]^{-1/2}. \tag{12}$$

Similarly,

$$f_{SLap}(x; \mu, \sigma, \lambda) = (v_l \sigma 2)^{-1} exp \left(-\frac{|x - \mu + m_l|}{v_l \sigma (1 + \lambda sign(x - \mu + m_l))} \right) \tag{13}$$

where

$$m_l = 2 v_l \sigma \lambda \tag{14}$$

and

$$v_l = \left[2(1 + \lambda^2) \right]^{-1/2}. \tag{15}$$

Now to compare the tails of the distributions,

$$\begin{aligned}\frac{f_{SNorm}(x;\mu_n,\sigma_n,\lambda_n)}{f_{SLap}(x;\mu_l,\sigma_l,\lambda_l)} &= \frac{(\sqrt{\pi}v_n\sigma_n)^{-1}exp\left\{-\frac{|x-\mu_n+m_n|}{v_n\sigma_n(1+\lambda_n sign(x-\mu_n+m_n))}\right\}^2}{(2v_l\sigma_l)^{-1}exp\left\{-\frac{|x-\mu_l+m_l|}{v_l\sigma_l(1+\lambda_l sign(x-\mu_l+m_l))}\right\}} \\ &= \frac{2v_l\sigma_l}{\sqrt{\pi}v_n\sigma_n}exp\left\{-\left(\frac{|x-(\mu_n-m_n)|^2}{c_n}\right)+\frac{|x-(\mu_l-m_l)|}{c_l}\right\}\end{aligned} \tag{16}$$

where $c_n = v_n^2\sigma_n^2(1+\lambda_n sign(x-\mu_n+m_n))^2$ and $c_l = v_l\sigma_l(1+\lambda_l sign(x-\mu_l+m_l))$ for brevity.

$$= \frac{2v_l\sigma_l}{\sqrt{\pi}v_n\sigma_n}exp\left\{\frac{-c_l(x^2-2(\mu_n-m_n)x+(\mu_n-m_n)^2)+c_n[x-(\mu_l-m_l)]}{c_nc_l}\right\}. \tag{17}$$

Note that there exists (L_n, U_n, L_l, U_l) such that $0 < L_n < c_n < U_n < \infty$ and $0 < L_l < c_l < U_l < \infty$ for all x (because $-1 < \lambda < 1$). Thus,

$$lim_{x\to\pm\infty}\frac{f_{SNorm}(x;\mu_n,\sigma_n,\lambda_n)}{f_{SLap}(x;\mu_l,\sigma_l,\lambda_l)} = lim_{x\to\pm\infty}exp(-x^2) = 0. \tag{18}$$

Therefore, the Skewed Laplace Distribution has heavier tails than Skewed Normal Distribution.

The settings of θ for the simulation study cover situations where the treatment is only effective on a small proportion of the treated population ($\theta = .2$) to the case where a large majority do ($\theta = .8$). The settings for δ are chosen to cover a range of treatment effect sizes for the responders that is large enough to be of practical importance ($.5\sigma_X \leq \delta$) but also realistic ($\delta \leq 3\sigma_X$).

A.5 Interpreting Simulation Results

The table below provides some examples for appropriately identifying the simulation results that communicate the performance statistics relevant to a novel scenario (where possibly $\delta < 0$ or $\sigma_X \neq 1$). The column on the left of **Table A.2** indicates scenarios that may reflect a real-life data set, while the column on the right indicates the simulation results corresponding to the scenario on the left.

m	n	F_r	θ_r	δ_r	σ_r	m	n	F_s	θ_s	δ_s	σ_s
90	90	SkR Norm	0.8	3.0	6.0	90	90	SkR Norm	0.8	0.5	1.0
480	720	SkL Norm	0.2	-2.0	1.0	480	720	SkR Norm	0.2	2.0	1.0
1920	2880	SkR Lap	0.5	-1.0	$0.\overline{3}$	1920	2880	SkL Lap	0.5	3.0	1.0

Table A.2: Corresponding settings for interpretation. Let r subscript a real scenario and s subscript the corresponding simulated scenario in **List 3.1**.

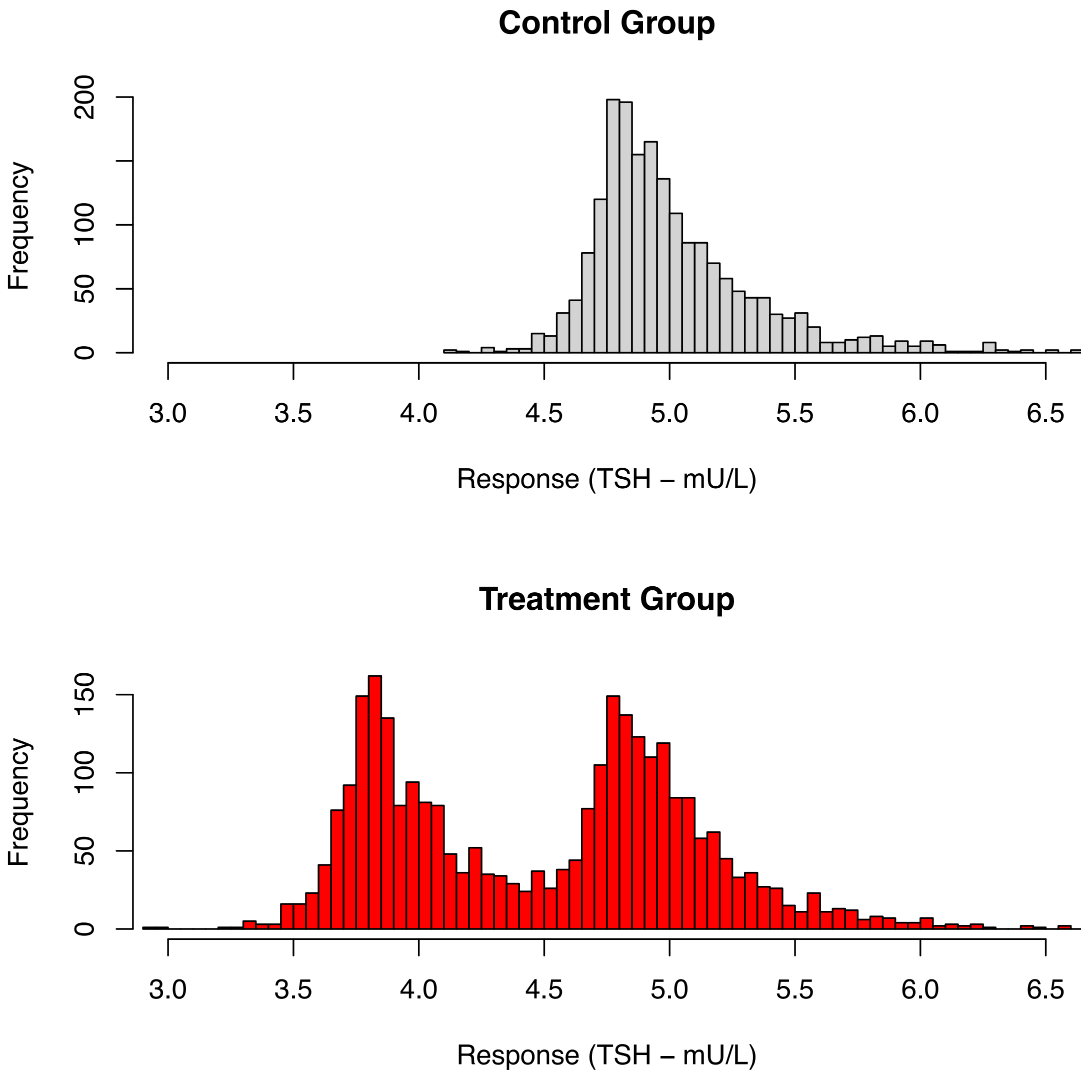


Figure A.3: The "real" data in the histograms represents a scenario for treating hypothyroid that corresponding to the left side of **Table A.2** in the last row.

Note that in the "real" scenario that results in data as displayed in **Figure A.3** $\sqrt{MSE(\widehat{\theta}_r)}$ (for example) is the same as the simulation setting on the last row of the right column ($m = 1920$, $n = 2880$, $F_s \sim SkLLap$, $\theta_s = 0.5$, $\delta_s = 3.0$, $\sigma_{X_s} = 1.0$) and has $\sqrt{MSE(\widehat{K_r})}$ the same as on the from the simulated setting, where $\widehat{K_r} = \widehat{\delta}_r / S_{X_r}$ (and $\widehat{K_s} = \widehat{\delta}_s / S_{X_s}$). Note that, even though the data is clearly skewed right in this scenario, the simulation results that communicate the relevant performance metrics (e.g. $\sqrt{MSE}$, coverage probability, average interval lengths, etc.) are from the skewed left Laplace distribution. Similarly, even though the magnitude of δ_r may seem small $\delta_r = -1$, this is a pronounced effect relative to the natural variability in the data (as observed in **Figure A.3**), so the appropriate simulation to reference for performance metrics has $\delta_s = 3$. The scenarios correspond to each other because both of the following are true.

1. Both the real data set and the corresponding simulation settings have a treatment effect in the opposite direction of the skew in F (F_r skewed right and $\delta_r < 0 \iff F_s$ skewed left and $\delta_s > 0$).
2. The magnitude of the treatment effect relative to the natural variability in F is the same ($\delta_r/\sigma_{X_r} = 1/0.\overline{3} = 3 = 3/1 = \delta_s/\sigma_{X_s}$).

A.6 Area Calculations for Confidence Regions

The area of pseudo-likelihood confidence regions is calculated from a dense grid of points that encompass the region. To identify such a dense grid of points that encompasses the region, a lighter grid search is first done across the set of $\theta \in \{.01, .02, ..., .99, 1.0\} \times \delta \in \{.1S_X, .2S_X, ..., 5.9S_X, 6.0S_X\}$ to compute the pseudo-likelihood test statistic (3.12) at each grid point. The four boundaries of the dense (100×100) grid are selected to encapsulate the confidence region as follows.

- The upper boundary for θ is selected as the smallest θ_u such that all (θ, δ) with $\theta \geq \theta_u$ have $\widehat{T}_{(\theta,\delta)} \geq c_1\chi^2_{d_1,1-\alpha}$.
- The upper boundary for δ is selected as the smallest δ_u such that all (θ, δ) with $\delta \geq \delta_u$ have $\widehat{T}_{(\theta,\delta)} \geq c_1\chi^2_{d_1,1-\alpha}$.
- The lower boundary for θ is selected as the largest θ_l such that all (θ, δ) with $\theta \leq \theta_l$ have $\widehat{T}_{(\theta,\delta)} \geq c_1\chi^2_{d_1,1-\alpha}$.
- The lower boundary for δ is selected as the largest δ_l such that all (θ, δ) with $\delta \leq \delta_l$ have $\widehat{T}_{(\theta,\delta)} \geq c_1\chi^2_{d_1,1-\alpha}$.

Therefore, by construction, all edge points of the dense region are not contained in the confidence region. The area is calculated as the area of the rectangular region formed by the dense grid search, $A = (\theta_u - \theta_l)(\delta_u - \delta_l)$, times the proportion of the dense grid points contained in the confidence region with a half-weight adjustment given to the edge points for improved numerical accuracy (see **Figure A.4** below). Let X be the number of points in the dense grid search that are in the confidence region. Let K be the number of grid points, $K = 100 \times 100 = 10,000$.

$$\text{Area } CR(\theta,\delta)_{PsL} = \frac{1}{2}\left(\frac{X}{K} + \frac{X}{K-400}\right)(\theta_u - \theta_l)(\delta_u - \delta_l) \tag{19}$$

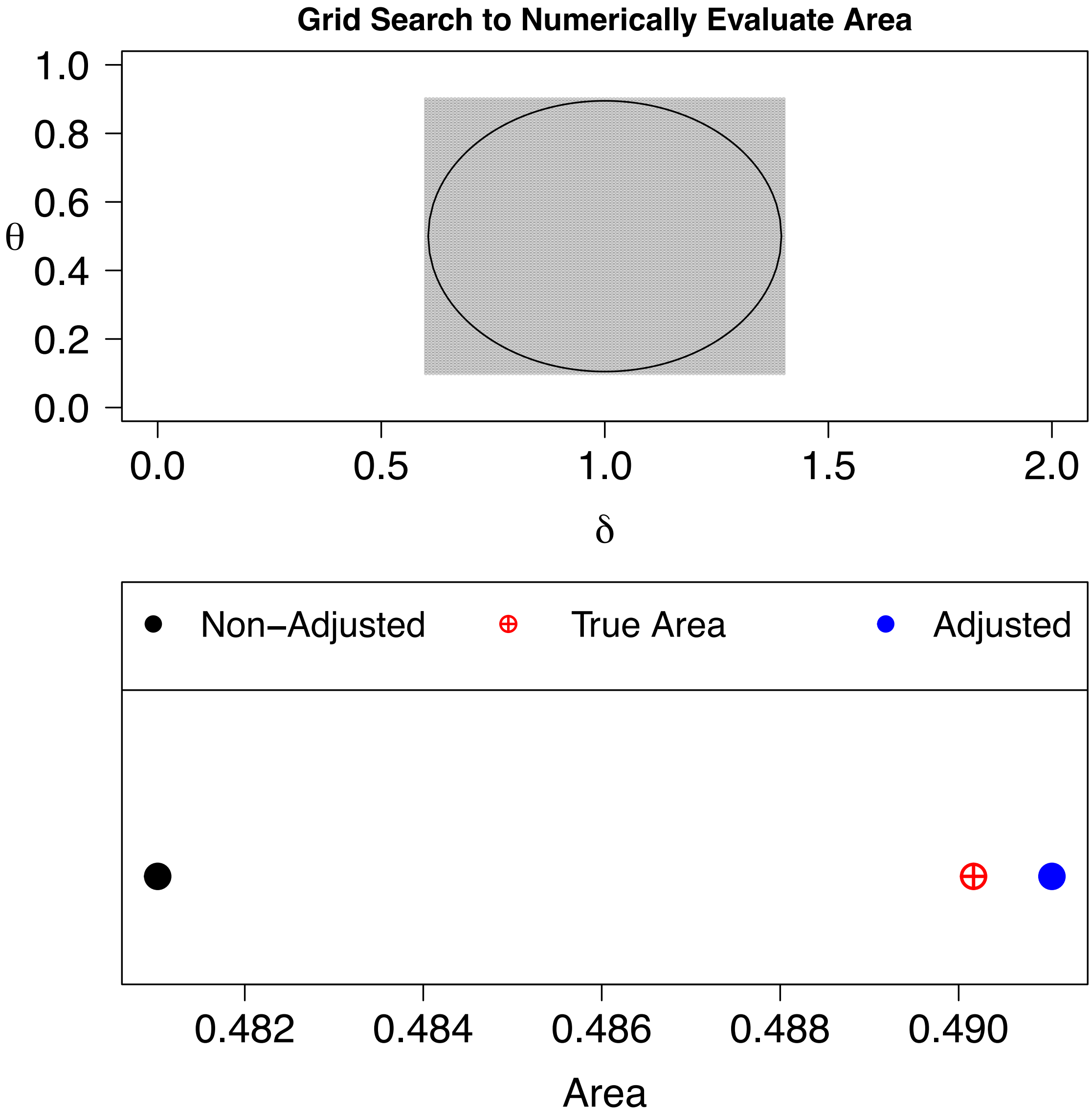


Figure A.4: Circle with radius $r = .395$ and center at $(\theta, \delta) = (.5, 1.0)$. Grid rectangular region is defined by $\theta_l = 0.1$, $\theta_u = 0.9$, $\delta_l = 0.6$, $\delta_u = 1.4$. True area is $\pi r^2 = 0.4902$, denoted as the red bulls-eye in the lower graph. The blue "Adjusted" number represents the numerical approximation to the area (0.4910) from equation (19), while the "Non-Adjusted" number is the proportion of grid points that are inside the circle times the area formed by the rectangle (0.4810).

For the method of moment confidence regions, the area is an analytical expression of the boundary equations. The exact calculation depends upon how the boundaries intersect each other. Let Δ_l and Δ_u represent the lower and upper bounds of Δ used to construct the region. Let δ_l and δ_u represent the lower and upper bounds of δ used to construct the region. Since some confidence regions can have infinite area, the simulations in section 4.3 compute areas of the regions truncated at a maximum feasible value of $\delta = 6S_X$. Let δ_u be the minimum of the computed upper bound of the confidence interval for δ and $6S_X$. For the confidence regions that only use $CI_{MoM}(\Delta)$, naturally $\delta_l = 0$ and $\delta_u = 6S_X$. The 9 scenarios for the possible forms of confidence regions listed below and shown in the **Figure A.5**.

1. $\Delta_l = 0,\ \delta_l < \Delta_u,\ \delta_u \leq \Delta_u$
2. $\Delta_l = 0,\ \delta_l < \Delta_u,\ \delta_u > \Delta_u$
3. $\Delta_l = 0,\ \delta_l \geq \Delta_u,\ \delta_u > \Delta_u$
4. $\Delta_l > 0,\ \delta_l < \Delta_l,\ \delta_u \leq \Delta_l$
5. $\Delta_l > 0,\ \delta_l < \Delta_l,\ \Delta_l < \delta_u \leq \Delta_u$
6. $\Delta_l > 0,\ \delta_l < \Delta_l,\ \delta_u > \Delta_u$
7. $\Delta_l > 0,\ \delta_l \geq \Delta_l,\ \delta_u \leq \Delta_u$
8. $\Delta_l > 0,\ \Delta_l \leq \delta_l < \Delta_u,\ \delta_u > \Delta_u$
9. $\Delta_l > 0,\ \delta_l \geq \Delta_u,\ \delta_u > \Delta_u$

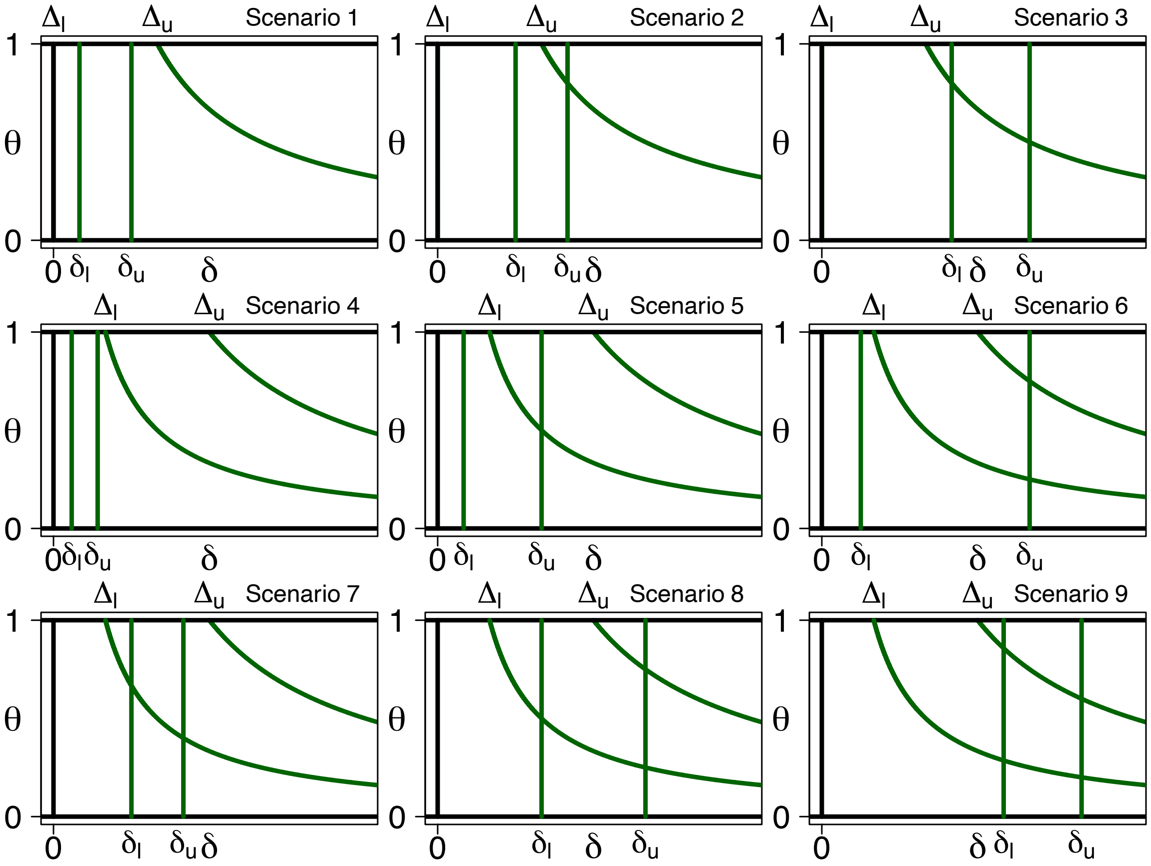


Figure A.5: Nine different scenarios for Method of Moment confidence regions.

The corresponding formulas for each of the confidence region areas are as follows

1. $(\delta_u - \delta_l)$
2. $(\Delta_u - \delta_l) + \Delta_u log\left(\frac{\delta_u}{\Delta_u}\right)$
3. $\Delta_u log\left(\frac{\delta_u}{\delta_l}\right)$
4. 0
5. $(\delta_u - \Delta_l) - \Delta_l log\left(\frac{\delta_u}{\Delta_l}\right)$
6. $(\Delta_u - \Delta_l) - \Delta_l log\left(\frac{\Delta_u}{\Delta_l}\right) + \Delta_u log\left(\frac{\delta_u}{\Delta_u}\right) - \Delta_l log\left(\frac{\delta_u}{\Delta_u}\right)$
7. $(\delta_u - \delta_l) - \Delta_l log\left(\frac{\delta_u}{\delta_l}\right)$
8. $(\Delta_u - \delta_l) - \Delta_l log\left(\frac{\Delta_u}{\delta_l}\right) + \Delta_u log\left(\frac{\delta_u}{\Delta_u}\right) - \Delta_l log\left(\frac{\delta_u}{\Delta_u}\right)$
9. $\Delta_u log\left(\frac{\delta_u}{\delta_l}\right) - \Delta_l log\left(\frac{\delta_u}{\delta_l}\right)$

Below is the derivation for area computations (recall that $\Delta = \theta\delta$).

$$\begin{aligned}
A1 &= \int_{\delta_l}^{\delta_u} \int_0^1 1 \frac{d}{d\theta} \frac{d}{d\delta} \\
&= \int_{\delta_l}^{\delta_u} \theta|_{\theta=0}^{\theta=1} \frac{d}{d\delta} \\
&= \int_{\delta_l}^{\delta_u} 1 \frac{d}{d\delta} \\
&= \delta|_{\delta=\delta_l}^{\delta=\delta_u} \\
&= (\delta_u - \delta_l). \qquad (20)
\end{aligned}$$

$$\begin{aligned}
A2 &= \int_{\delta_l}^{\Delta_u} \int_0^1 1 \frac{d}{d\theta} \frac{d}{d\delta} + \int_{\Delta_u}^{\delta_u} \int_0^{\Delta_u/\delta} 1 \frac{d}{d\theta} \frac{d}{d\delta} \\
&= (\Delta_u - \delta_l) + \int_{\Delta_u}^{\delta_u} \int_0^{\Delta_u/\delta} 1 \frac{d}{d\theta} \frac{d}{d\delta} \\
&= (\Delta_u - \delta_l) + \int_{\Delta_u}^{\delta_u} \frac{\Delta_u}{\delta} \frac{d}{d\delta} \\
&= (\Delta_u - \delta_l) + \Delta_u log(\delta)|_{\delta=\Delta_u}^{\delta=\delta_u} \\
&= (\Delta_u - \delta_l) + \Delta_u log(\delta_u) - \Delta_u log(\Delta_u) \\
&= (\Delta_u - \delta_l) + \Delta_u log\left(\frac{\delta_u}{\Delta_u}\right). \qquad (21)
\end{aligned}$$

$$\begin{aligned}
A3 &= \int_{\delta_l}^{\delta_u} \int_{0}^{\Delta_u/\delta} 1 \frac{d}{d\theta} \frac{d}{d\delta} \\
&= \int_{\delta_l}^{\delta_u} \frac{\Delta_u}{\delta} \frac{d}{d\delta} \\
&= \Delta_u log(\delta)|_{\delta=\delta_l}^{\delta=\delta_u} \\
&= \Delta_u log(\delta_u) - \Delta_u log(\delta_l) \\
&= \Delta_u log \left(\frac{\delta_u}{\delta_l} \right). \qquad (22)
\end{aligned}$$

$$A4 = 0 \quad \text{(because } [\delta_l, \delta_u] \text{ and } [\Delta_l, \Delta_u] \text{ are mutually exclusive).} \qquad (23)$$

$$\begin{aligned}
A5 &= \int_{\Delta_l}^{\delta_u} \int_{\Delta_l/\delta}^{1} 1 \frac{d}{d\theta} \frac{d}{d\delta} \\
&= \int_{\Delta_l}^{\delta_u} 1 - \frac{\Delta_l}{\delta} \frac{d}{d\delta} \\
&= [\delta - \Delta_l log(\delta)] \,|_{\delta=\Delta_l}^{\delta=\delta_u} \\
&= \{\delta_u - \Delta_l log(\delta_u)\} - \{\Delta_l - \Delta_l log(\Delta_l)\} \\
&= (\delta_u - \Delta_l) - \Delta_l log \left(\frac{\delta_u}{\Delta_l} \right). \qquad (24)
\end{aligned}$$

$$
\begin{aligned}
A6 &= \int_{\Delta_l}^{\Delta_u} \int_{\Delta_l/\delta}^{1} 1 \frac{d}{d\theta}\frac{d}{d\delta} + \int_{\Delta_u}^{\delta_u} \int_{\Delta_l/\delta}^{\Delta_u/\delta} 1 \frac{d}{d\theta}\frac{d}{d\delta} \\
&= \left\{ (\Delta_u - \Delta_l) - \Delta_l log \left(\frac{\Delta_u}{\Delta_l} \right) \right\} + \int_{\Delta_u}^{\delta_u} \int_{\Delta_l/\delta}^{\Delta_u/\delta} 1 \frac{d}{d\theta}\frac{d}{d\delta} \\
&= \left\{ (\Delta_u - \Delta_l) - \Delta_l log \left(\frac{\Delta_u}{\Delta_l} \right) \right\} + \int_{\Delta_u}^{\delta_u} \frac{\Delta_u}{\delta} - \frac{\Delta_l}{\delta} \frac{d}{d\delta} \\
&= \left\{ (\Delta_u - \Delta_l) - \Delta_l log \left(\frac{\Delta_u}{\Delta_l} \right) \right\} + \\
&\quad [\Delta_u log\,(\delta_u) - \Delta_l log\,(\delta_u)] - [\Delta_u log(\Delta_u) - \Delta_l log(\Delta_u)] \\
&= (\Delta_u - \Delta_l) - \Delta_l log \left(\frac{\Delta_u}{\Delta_l} \right) + \Delta_u log \left(\frac{\delta_u}{\Delta_u} \right) - \Delta_l log \left(\frac{\delta_u}{\Delta_u} \right). \qquad (25)
\end{aligned}
$$

$$
\begin{aligned}
A7 &= \int_{\delta_l}^{\delta_u} \int_{\Delta_l/\delta}^{1} 1 \frac{d}{d\theta}\frac{d}{d\delta} \\
&= [\delta - \Delta_l log(\delta)] \,|_{\delta=\delta_l}^{\delta=\delta_u} \\
&= \{\delta_u - \Delta_l log(\delta_u)\} - \{\delta_l - \Delta_l log(\delta_l)\} \\
&= (\delta_u - \delta_l) - \Delta_l log \left(\frac{\delta_u}{\delta_l} \right). \qquad (26)
\end{aligned}
$$

$$
\begin{aligned}
A8 &= \int_{\delta_l}^{\Delta_u} \int_{\Delta_l/\delta}^{1} 1 \frac{d}{d\theta}\frac{d}{d\delta} + \int_{\Delta_u}^{\delta_u} \int_{\Delta_l/\delta}^{\Delta_u/\delta} 1 \frac{d}{d\theta}\frac{d}{d\delta} \\
&= \left\{ (\Delta_u - \delta_l) - \Delta_l log \left(\frac{\Delta_u}{\delta_l} \right) \right\} + \int_{\Delta_u}^{\delta_u} \int_{\Delta_l/\delta}^{\Delta_u/\delta} 1 \frac{d}{d\theta}\frac{d}{d\delta} \\
&= \left\{ (\Delta_u - \delta_l) - \Delta_l log \left(\frac{\Delta_u}{\delta_l} \right) \right\} + \int_{\Delta_u}^{\delta_u} \frac{\Delta_u}{\delta} - \frac{\Delta_l}{\delta} \frac{d}{d\delta} \\
&= (\Delta_u - \delta_l) - \Delta_l log \left(\frac{\Delta_u}{\delta_l} \right) + \Delta_u log \left(\frac{\delta_u}{\Delta_u} \right) - \Delta_l log \left(\frac{\delta_u}{\Delta_u} \right). \qquad (27)
\end{aligned}
$$

$$
\begin{aligned}
A9 &= \int_{\delta_l}^{\delta_u} \int_{\Delta_l/\delta}^{\Delta_u/\delta} 1 \frac{d}{d\theta}\frac{d}{d\delta} \\
&= \int_{\delta_l}^{\delta_u} \frac{\Delta_u}{\delta} - \frac{\Delta_l}{\delta}\frac{d}{d\delta} \\
&= \Delta_u \log\left(\frac{\delta_u}{\delta_l}\right) - \Delta_l log\left(\frac{\delta_u}{\delta_l}\right). \qquad (28)
\end{aligned}
$$

A.7 Proof Relating Infinte and "Null-Containing" Intervals

Let the parameter space of (θ, δ) be $\{(\theta, \delta) : \theta \in (0,1]$ and $\delta \in (0, \infty)$, or $(\theta, \delta) = (0,0)\}$.

Claim:

$$0 \in CSet_{PsL}(\delta) \iff CSet_{PsL}(\delta) = [0, \infty)$$
$$0 \in CSet_{PsL}(\theta) \iff CSet_{PsL}(\theta) = [0, 1]$$

Observation 1: $\widehat{L}(0, \delta) = \widehat{L}(0,0) = \prod_{j=1}^{m} \widehat{f}(x_j) \prod_{i=1}^{n} \widehat{f}(y_i)$ for all $\delta \in \mathcal{R}$. (Note, this does not say anything about the parameter space, it is simply an observation about the function $\widehat{L}(\cdot, \cdot)$).

Observation 2: $\lim\limits_{\theta \to 0^+} L(\theta, \delta') = L(0, \delta')$ for a fixed δ' (since the pseudo-likelihood is a continuous function of θ for any given δ') . This means that for any $\epsilon > 0$, there exists a $\theta' > 0$ such that $|L(0, \delta') - L(\theta', \delta')| < \epsilon$.

Proof:

$(\impliedby)$ $CSet_{PsL}(\delta) = [0, \infty) \implies 0 \in CSet_{PsL}(\delta)$ trivially.

$(\implies)$ Now to see that $0 \in CSet_{PsL}(\delta) \implies CSet_{PsL}(\delta) = [0, \infty)$, recall that

$$100(1-\alpha)\%\ CSet_{PsL}(\delta) = \left\{\delta : \widehat{T}_\delta < c_3 \chi^2_{d_3, 1-\alpha}\right\}, \qquad (3.31)$$

where

$$\widehat{T}_\delta = -2\left[log\widehat{L}\left(\widehat{\theta}(\delta), \delta; X, Y\right) - log\widehat{L}\left(\widehat{\theta}, \widehat{\delta}; X, Y\right)\right]. \qquad (3.29)$$

Rearranging to put this in terms of the profile (pseudo)-likelihood gives

$$\begin{gathered} 0 \in CSet_{PsL}(\delta) \\ \Longleftrightarrow \\ \widehat{L}(\widehat{\theta}(0), 0; X, Y) > \widehat{L}(\widehat{\theta}, \widehat{\delta}; X, Y) exp\left\{-\frac{c_3\chi^2_{d_3,1-\alpha}}{2}\right\} \overset{def}{=} k. \end{gathered} \tag{29}$$

Now, since $\widehat{L}(\widehat{\theta}(\delta), \delta; X, Y) = \max_{\theta} \widehat{L}(\theta, \delta; X, Y)$, then for a given $\delta^{'}$, $\delta^{'} \in CSet_{PsL}(\delta)$ $\Longleftrightarrow$ $\exists\, \theta^{'}$ such that $\widehat{L}(\theta^{'}, \delta^{'}) > k$. Because $0 \in CSet_{PsL}(\delta)$, then $\widehat{L}(0, 0) > k$. By **observation 1**, this implies that for any fixed $\delta^{'}$, $L(0, \delta^{'}) > k$. Let $\epsilon = L(0, \delta^{'}) - k$. Then, by **observation 2** there exists a $\theta^{'} > 0$ such that $|L(0, \delta^{'}) - L(\theta^{'}, \delta^{'})| < L(0, \delta^{'}) - k$.

<u>Case 1:</u> $L(0, \delta^{'}) < L(\theta^{'}, \delta^{'})$

$$\begin{aligned} k < L(0, \delta^{'}) &< L(\theta^{'}, \delta^{'}) \\ \Longrightarrow\ L(\theta^{'}, \delta^{'}) &> k. \end{aligned}$$

<u>Case 2:</u> $L(\theta^{'}, \delta^{'}) < L(0, \delta^{'})$

$$\begin{aligned} L(0, \delta^{'}) - L(\theta^{'}, \delta^{'}) &< L(0, \delta^{'}) - k \\ \Longrightarrow\ L(\theta^{'}, \delta^{'}) &> k. \end{aligned}$$

Since $\delta^{'}$ is arbitrary, then for any $\delta^{'}$ there exists a $\theta^{'} > 0$ such that $L(\theta^{'}, \delta^{'}) > k$. Therefore, $\delta^{'} \in CSet_{PsL}(\delta)$ for all $\delta^{'} \in [0, \infty)$. A similar argument shows that $0 \in CSet_{PsL}(\theta) \implies CSet(\theta) = [0, 1]$, that has an analogous **observation 2** based on the continuity of $\widehat{f}$ in the pseudo-likelihood (2.19).

Discussion of Interval Agreement and Interpretation

Note that the cutoff k is different for $CSet(\theta)$ and $CSet(\delta)$ only due to the Satterthwaite approximation. In the limit (when $c_2 = c_3 = 1$ and $d_2 = d_3 = 1$), the intervals always agree on when to include or exclude 0 from the intervals.

The non-compact nature of the parameter space reflects the difficulty interpreting a single parameter. If one parameter is 0, then the other parameter does not have a real interpretation. In general, we say that θ represents the proportion of responders and δ is the magnitude of the shift for the responder subpopulation. However, if $\theta = 0$ then there are no responders, so δ can no longer be meaningfully interpreted as the magnitude of the shift for the responders (since there are none). Similarly, if $\delta = 0$, then there is no shift for the "responders". In such a case, θ can no longer be meaningfully interpreted as the proportion of the responders (since these "responders" are really no different than non-responders). So while in some circumstances a practitioner might care to focus primarily on one variable, say θ, inference about it cannot be totally divorced from inference for δ since θ is meaningless if $\delta = 0$. One nice asymptotic property of the pseudo-likelihood intervals is that $0 \in CSet(\delta) \iff 0 \in CSet(\theta)$.

However, this is not true for the method of moment intervals nor for sample sizes where the Satterthwaite approximation is used for the pseudo-likelihood intervals. For example, when investigating the Kaiser data set on reduction in diastolic blood pressure, the moment intervals came out to be

$$95\%\ CI_{MoM}(\theta) = [0.13, 0.50]$$

$$95\%\ CI_{MoM}(\delta) = [0.00, 16.1].$$

While the interval for θ is interesting (in that it does not contain either 0 or 1), it's not meaningful if $\delta = 0$, which is in $CI_{MoM}(\delta)$. This highlights the danger of using a single method of moment interval (for either θ or δ) to conclude that a treatment effect exists.

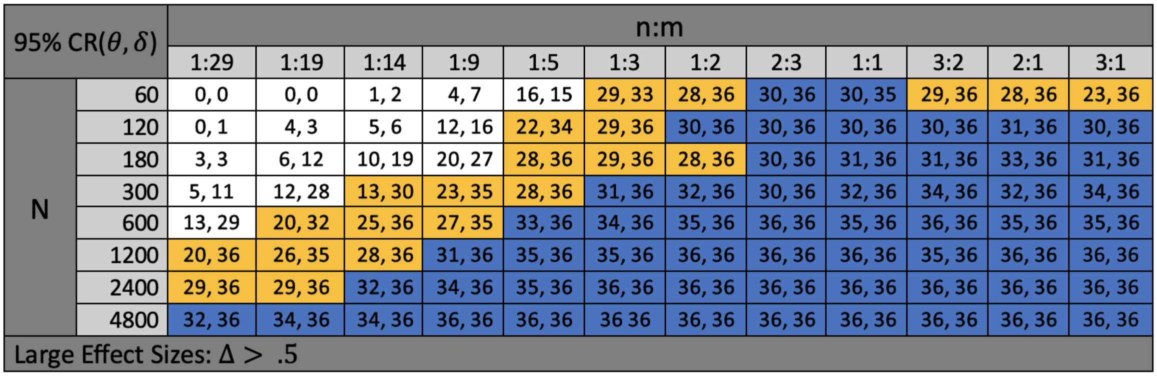
Another frustrating aspect of the moment intervals is that the marginal intervals $CI_{MoM}(\theta)$ or $CI_{MoM}(\delta)$ can contain 0 even when there's clearly enough evidence from the moments that there truly is a treatment effect - as seen by the corresponding $CI(\Delta)$. For example, for the same data set from which the two above moment intervals are calculated, the confidence interval for the overall treatment effect is

$$95\% \; CI_{MoM}(\Delta) = [1.0, 4.2].$$

Surely if the average treatment effect is somewhere between a 1.0 and 4.2 point reduction in DBP, then the effect on the responders (which is subset of all treated patients) should also be at least 1.0. However, the corresponding lower bound for 95% $CI_{MoM}(\delta)$ is 0. So since the treatment effect is characterized by (θ, δ) and interpretation of one interval is incomplete without consideration of the other, this suggests that a better way to characterize the uncertainty surrounding the treatment effect is with a confidence region.

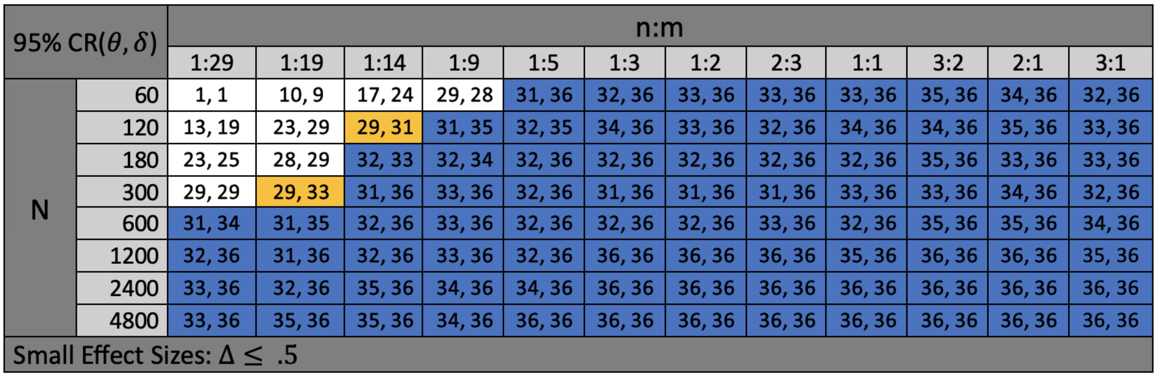
A.8 Comparing Method of Moment Regions

A.8.1 Sufficient Probability Tables (95% Regions)



95% CR(θ, δ)		n:m											
		1:29	1:19	1:14	1:9	1:5	1:3	1:2	2:3	1:1	3:2	2:1	3:1
N	60	0, 0	0, 0	1, 2	4, 7	16, 15	29, 33	28, 36	30, 36	30, 35	29, 36	28, 36	23, 36
	120	0, 1	4, 3	5, 6	12, 16	22, 34	29, 36	30, 36	30, 36	30, 36	30, 36	31, 36	30, 36
	180	3, 3	6, 12	10, 19	20, 27	28, 36	29, 36	28, 36	30, 36	31, 36	31, 36	33, 36	31, 36
	300	5, 11	12, 28	13, 30	23, 35	28, 36	31, 36	32, 36	30, 36	32, 36	34, 36	32, 36	34, 36
	600	13, 29	20, 32	25, 36	27, 35	33, 36	34, 36	35, 36	36, 36	35, 36	36, 36	35, 36	35, 36
	1200	20, 36	26, 35	28, 36	31, 36	35, 36	35, 36	36, 36	36, 36	36, 36	35, 36	36, 36	36, 36
	2400	29, 36	29, 36	32, 36	34, 36	35, 36	36, 36	36, 36	36, 36	36, 36	36, 36	36, 36	36, 36
	4800	32, 36	34, 36	34, 36	36, 36	36, 36	36 36	36, 36	36, 36	36, 36	36, 36	36, 36	36, 36
Large Effect Sizes: Δ > .5													

Table A.3: Each cell entry represents how many of the 36 (F, θ, δ) with $\Delta > .5\sigma_X$ correspond to $CR_{MoM\Delta}(\theta, \delta), CR_{MoM\{\delta,\Delta\}}(\theta, \delta)$ that have simulated coverage probability at least .925. Color coded backgrounds emphasize when this number is at least 30 - neither method: white, both methods: blue, Δ region only: dark orange, $\Delta \cap \delta$ region only: gold.

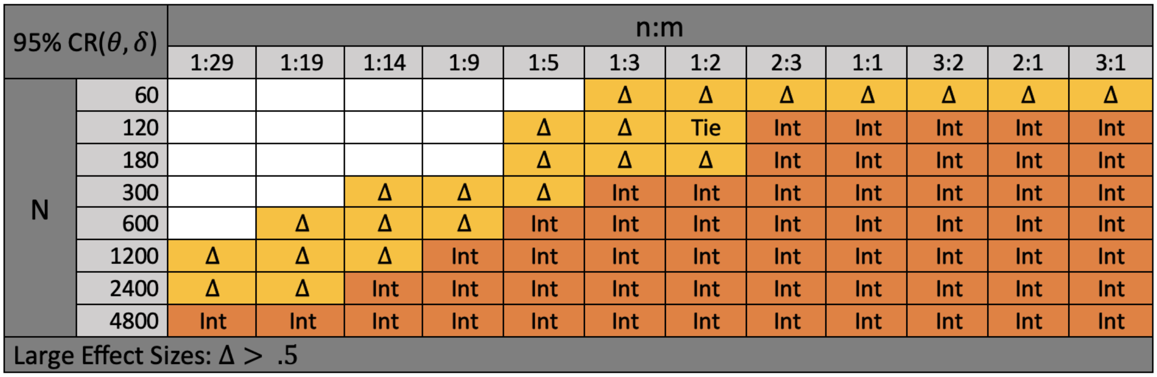


95% CR(θ, δ)		n:m											
		1:29	1:19	1:14	1:9	1:5	1:3	1:2	2:3	1:1	3:2	2:1	3:1
N	60	1, 1	10, 9	17, 24	29, 28	31, 36	32, 36	33, 36	33, 36	33, 36	35, 36	34, 36	32, 36
	120	13, 19	23, 29	29, 31	31, 35	32, 35	34, 36	33, 36	32, 36	34, 36	34, 36	35, 36	33, 36
	180	23, 25	28, 29	32, 33	32, 34	32, 36	32, 36	32, 36	32, 36	32, 36	35, 36	33, 36	33, 36
	300	29, 29	29, 33	31, 36	33, 36	32, 36	31, 36	31, 36	31, 36	33, 36	33, 36	34, 36	32, 36
	600	31, 34	31, 35	32, 36	33, 36	32, 36	32, 36	32, 36	33, 36	32, 36	35, 36	35, 36	34, 36
	1200	32, 36	31, 36	32, 36	33, 36	32, 36	36, 36	36, 36	36, 36	35, 36	36, 36	36, 36	35, 36
	2400	33, 36	32, 36	35, 36	34, 36	34, 36	36, 36	36, 36	36, 36	36, 36	36, 36	36, 36	36, 36
	4800	33, 36	35, 36	35, 36	34, 36	36, 36	36, 36	36, 36	36, 36	36, 36	36, 36	36, 36	36, 36
Small Effect Sizes: Δ ≤ .5													

Table A.4: Each cell entry represents how many of the 36 (F, θ, δ) with $\Delta \leq .5\sigma_X$ correspond to $CR_{MoM\Delta}(\theta, \delta), CR_{MoM\{\delta,\Delta\}}(\theta, \delta)$ that have simulated coverage probability at least .925. Color coded backgrounds emphasize when this number is at least 30 - neither method: white, both methods: blue, Δ region only: dark orange, $\Delta \cap \delta$ region only: gold.

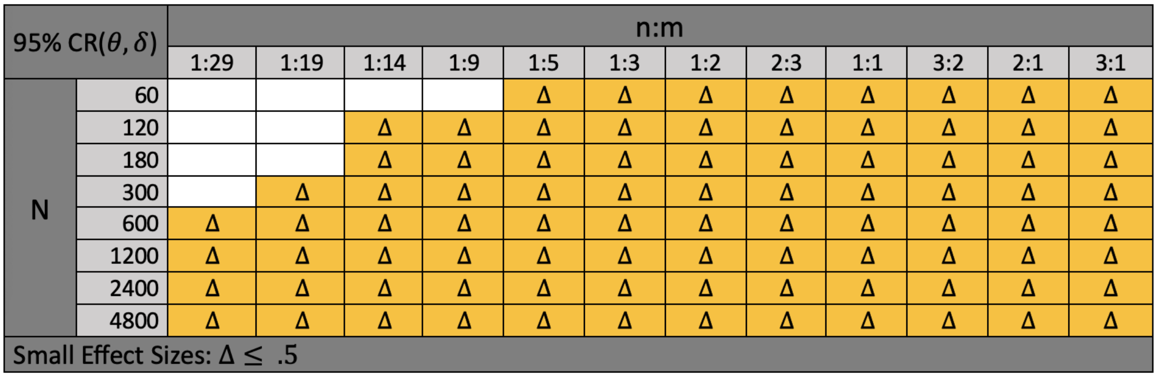
Tables A.3 - A.4 indicate that at least one confidence region provides satisfactory coverage probability for nearly any sample size setting. For most settings, both the Δ region and method of moment confidence regions produce satisfactory coverage probability. If treatment data is sparse (e.g. $n \leq 20$) both the method of moment regions may be unsatisfactory.

A.8.2 Recommendation Tables



95% CR(θ, δ)		n:m											
		1:29	1:19	1:14	1:9	1:5	1:3	1:2	2:3	1:1	3:2	2:1	3:1
N	60						Δ	Δ	Δ	Δ	Δ	Δ	Δ
	120					Δ	Δ	Tie	Int	Int	Int	Int	Int
	180					Δ	Δ	Δ	Int	Int	Int	Int	Int
	300			Δ	Δ	Δ	Int	Int	Int	Int	Int	Int	Int
	600		Δ	Δ	Δ	Int	Int	Int	Int	Int	Int	Int	Int
	1200	Δ	Δ	Δ	Int	Int	Int	Int	Int	Int	Int	Int	Int
	2400	Δ	Δ	Int	Int	Int	Int	Int	Int	Int	Int	Int	Int
	4800	Int	Int	Int	Int	Int	Int	Int	Int	Int	Int	Int	Int
Large Effect Sizes: Δ > .5													

Table A.5: Each cell entry represents the recommended method – $CR_{MoM\Delta}(\theta, \delta)$ or $CR_{MoM\{\delta,\Delta\}}(\theta, \delta)$ (blank white cell means neither method is recommended). The recommended method achieves simulated coverage probability at least .925 for at least 30 of the 36 simulation settings where $\Delta > .5\sigma_X$. If the non-recommended method also meets the coverage probability criterion, the recommended method has smaller average interval length in more settings than the alternate method.



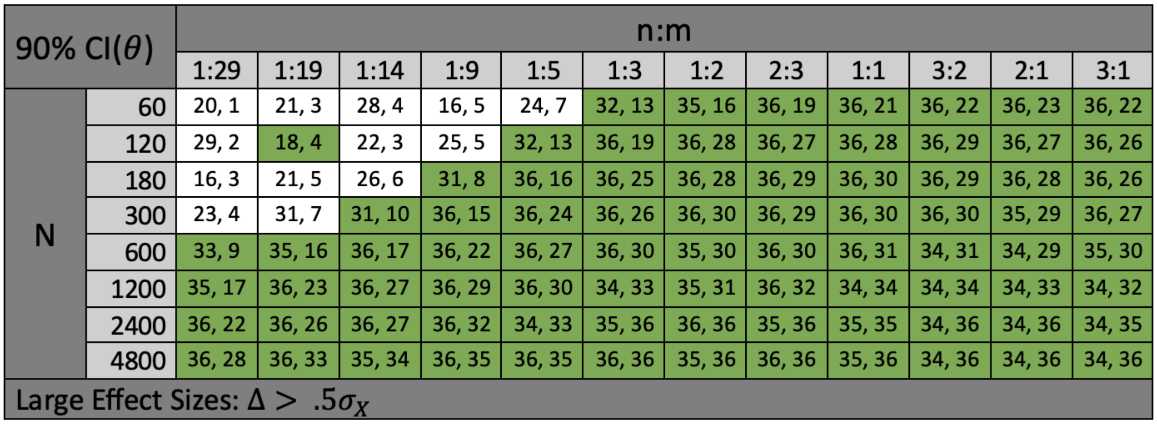
95% CR(θ, δ)		n:m											
		1:29	1:19	1:14	1:9	1:5	1:3	1:2	2:3	1:1	3:2	2:1	3:1
N	60					Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ
	120			Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ
	180			Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ
	300		Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ
	600	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ
	1200	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ
	2400	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ
	4800	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ
Small Effect Sizes: Δ ≤ .5													

Table A.6: Each cell entry represents the recommended method – $CR_{MoM\Delta}(\theta, \delta)$ or $CR_{MoM\{\delta,\Delta\}}(\theta, \delta)$. The recommended method achieves simulated coverage probability at least .925 for at least 30 of the 36 simulation settings where $\Delta \leq .5\sigma_X$. If the non-recommended method also meets the coverage probability criterion, the recommended method has smaller average interval length in more settings than the alternate method.

Tables A.5 - A.6 show which confidence region method is recommended for each sample size setting. For large effect sizes, the $\Delta \cap \delta$ regions win almost all tie-breakers based on area and are preferred when the total and treatment group sample sizes are sufficiently large (e.g. $N \geq 120$ or $N \geq 300 and n \geq 120$). For small overall effect sizes, the Δ regions are always preferred and are recommended so long as there are more than 10 treatment observations.

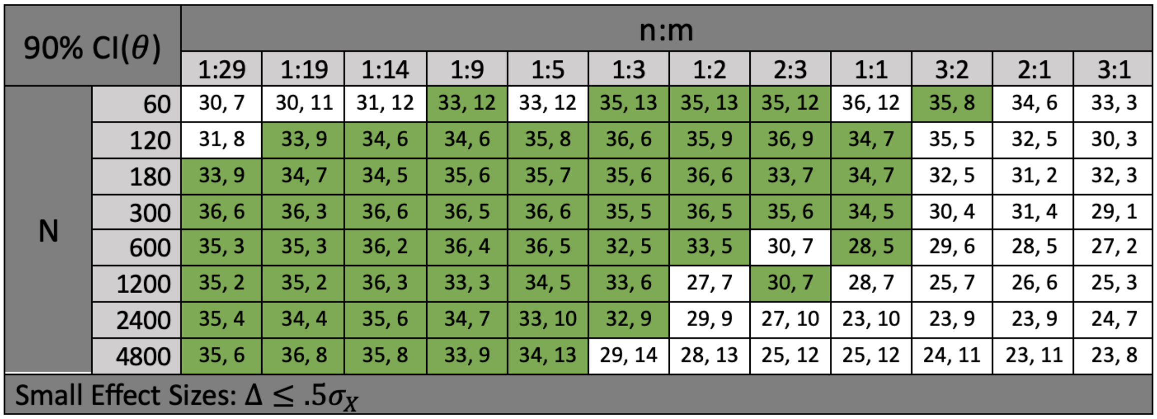
A.9 90% Confidence Tables

Note that the 90% confidence interval simulation coverage probability cutoff is chosen to be .865 because the simulation error for a 90% confidence interval is larger than that of a 95% confidence interval by $\left(\frac{\sqrt{.90(1-.90)/1000}}{\sqrt{.95(1-.95)/1000}} =\right)$ 138%, so .865 is the same number of standard errors (due to simulation error) below .90 as .925 is below .95. Each table below has colored background according to the recommendations (green is pseudo-likelihood, gold is method of moment, white is neither), but information about coverage probabilities is for the 90% intervals. The tables below show that the recommended method consistently has sufficient coverage probability across a large number of (F, θ, δ) while the cases where neither method is recommended consistently has both methods with fewer (F, θ, δ) scenarios where coverage probability is sufficient.



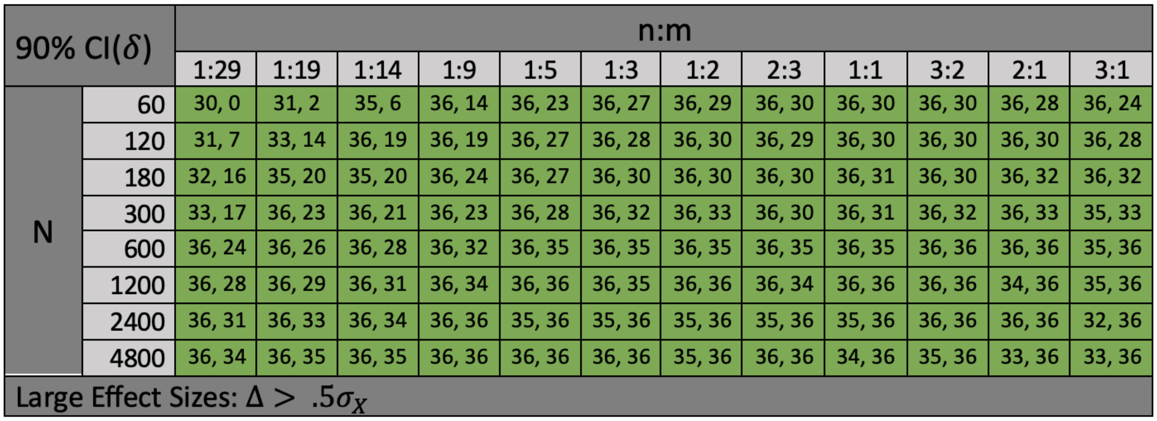
90% CI(θ)		n:m											
		1:29	1:19	1:14	1:9	1:5	1:3	1:2	2:3	1:1	3:2	2:1	3:1
N	60	20, 1	21, 3	28, 4	16, 5	24, 7	32, 13	35, 16	36, 19	36, 21	36, 22	36, 23	36, 22
	120	29, 2	18, 4	22, 3	25, 5	32, 13	36, 19	36, 28	36, 27	36, 28	36, 29	36, 27	36, 26
	180	16, 3	21, 5	26, 6	31, 8	36, 16	36, 25	36, 28	36, 29	36, 30	36, 29	36, 28	36, 26
	300	23, 4	31, 7	31, 10	36, 15	36, 24	36, 26	36, 30	36, 29	36, 30	36, 30	35, 29	36, 27
	600	33, 9	35, 16	36, 17	36, 22	36, 27	36, 30	35, 30	36, 30	36, 31	34, 31	34, 29	35, 30
	1200	35, 17	36, 23	36, 27	36, 29	36, 30	34, 33	35, 31	36, 32	34, 34	34, 34	34, 33	34, 32
	2400	36, 22	36, 26	36, 27	36, 32	34, 33	35, 36	36, 36	35, 36	35, 35	34, 36	34, 36	34, 35
	4800	36, 28	36, 33	35, 34	36, 35	36, 35	36, 36	35, 36	36, 36	35, 36	34, 36	34, 36	34, 36
Large Effect Sizes: $\Delta >$ .5σ_X													

Table A.7: Each cell entry represents how many of the 36 (F, θ, δ) with $\Delta > .5\sigma_X$ correspond to $90\% CI_{PsL}(\theta), 90\% CI_{MoM}(\theta)$ that have simulated coverage probability at least .865. Color coded backgrounds correspond to the recommendations from **Figure** 4.7.



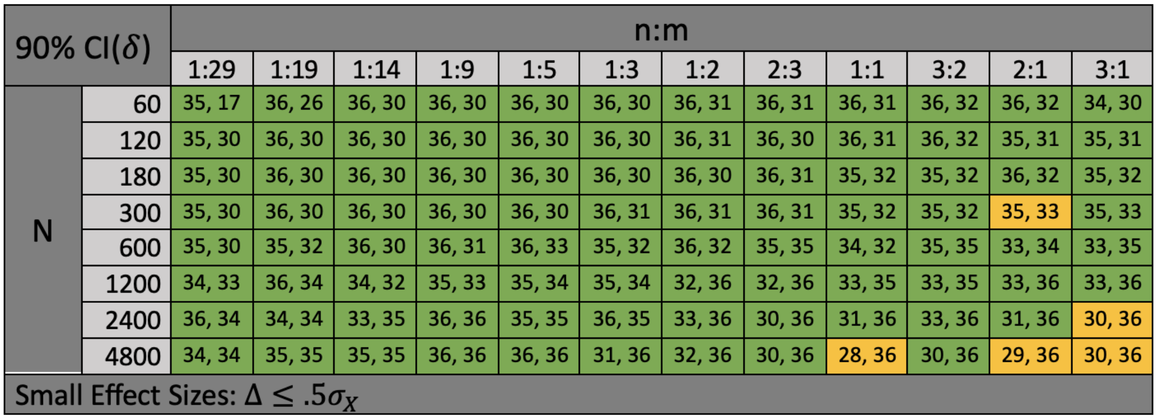
90% CI(θ)		n:m											
		1:29	1:19	1:14	1:9	1:5	1:3	1:2	2:3	1:1	3:2	2:1	3:1
N	60	30, 7	30, 11	31, 12	33, 12	33, 12	35, 13	35, 13	35, 12	36, 12	35, 8	34, 6	33, 3
	120	31, 8	33, 9	34, 6	34, 6	35, 8	36, 6	35, 9	36, 9	34, 7	35, 5	32, 5	30, 3
	180	33, 9	34, 7	34, 5	35, 6	35, 7	35, 6	36, 6	33, 7	34, 7	32, 5	31, 2	32, 3
	300	36, 6	36, 3	36, 6	36, 5	36, 6	35, 5	36, 5	35, 6	34, 5	30, 4	31, 4	29, 1
	600	35, 3	35, 3	36, 2	36, 4	36, 5	32, 5	33, 5	30, 7	28, 5	29, 6	28, 5	27, 2
	1200	35, 2	35, 2	36, 3	33, 3	34, 5	33, 6	27, 7	30, 7	28, 7	25, 7	26, 6	25, 3
	2400	35, 4	34, 4	35, 6	34, 7	33, 10	32, 9	29, 9	27, 10	23, 10	23, 9	23, 9	24, 7
	4800	35, 6	36, 8	35, 8	33, 9	34, 13	29, 14	28, 13	25, 12	25, 12	24, 11	23, 11	23, 8
Small Effect Sizes: $\Delta \leq$ .5σ_X													

Table A.8: Each cell entry represents how many of the 36 (F, θ, δ) with $\Delta \leq .5\sigma_X$ correspond to $90\% CI_{PsL}(\theta), 90\% CI_{MoM}(\theta)$ that have simulated coverage probability at least .865. Color coded backgrounds correspond to the recommendations from **Figure** 4.8.



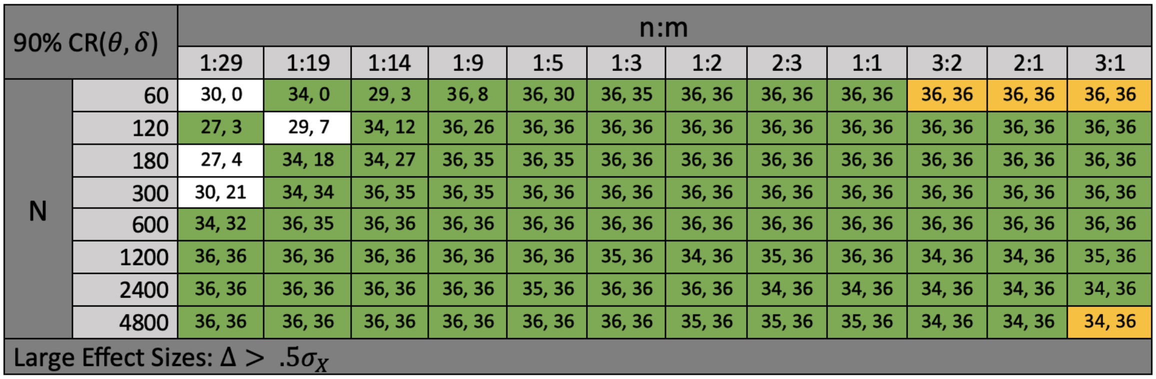
90% CI(δ)		n:m											
		1:29	1:19	1:14	1:9	1:5	1:3	1:2	2:3	1:1	3:2	2:1	3:1
N	60	30, 0	31, 2	35, 6	36, 14	36, 23	36, 27	36, 29	36, 30	36, 30	36, 30	36, 28	36, 24
	120	31, 7	33, 14	36, 19	36, 19	36, 27	36, 28	36, 30	36, 29	36, 30	36, 30	36, 30	36, 28
	180	32, 16	35, 20	35, 20	36, 24	36, 27	36, 30	36, 30	36, 30	36, 31	36, 30	36, 32	36, 32
	300	33, 17	36, 23	36, 21	36, 23	36, 28	36, 32	36, 33	36, 30	36, 31	36, 32	36, 33	35, 33
	600	36, 24	36, 26	36, 28	36, 32	36, 35	36, 35	36, 35	36, 35	36, 35	36, 36	36, 36	35, 36
	1200	36, 28	36, 29	36, 31	36, 34	36, 36	36, 35	36, 36	36, 34	36, 36	36, 36	34, 36	35, 36
	2400	36, 31	36, 33	36, 34	36, 36	35, 36	35, 36	35, 36	35, 36	35, 36	36, 36	36, 36	32, 36
	4800	36, 34	36, 35	36, 35	36, 36	36, 36	36, 36	35, 36	36, 36	34, 36	35, 36	33, 36	33, 36
Large Effect Sizes: $\Delta > .5\sigma_X$													

Table A.9: Each cell entry represents how many of the 36 (F, θ, δ) with $\Delta > .5\sigma_X$ correspond to $90\% CI_{PsL}(\delta), 90\% CI_{MoM}(\delta)$ that have simulated coverage probability at least .865. Color coded backgrounds correspond to the recommendations from **Figure 4.11**.



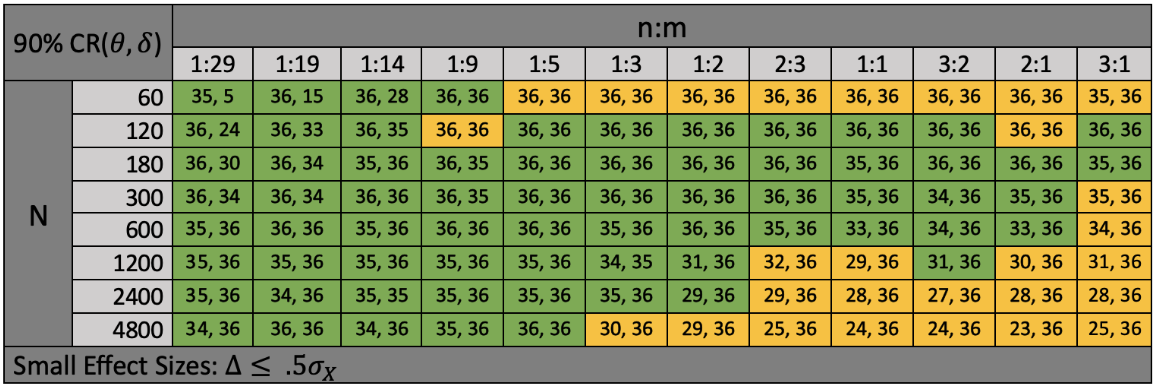
90% CI(δ)		n:m											
		1:29	1:19	1:14	1:9	1:5	1:3	1:2	2:3	1:1	3:2	2:1	3:1
N	60	35, 17	36, 26	36, 30	36, 30	36, 30	36, 30	36, 31	36, 31	36, 31	36, 32	36, 32	34, 30
	120	35, 30	36, 30	36, 30	36, 30	36, 30	36, 30	36, 31	36, 30	36, 31	36, 32	35, 31	35, 31
	180	35, 30	36, 30	36, 30	36, 30	36, 30	36, 30	36, 30	36, 31	35, 32	35, 32	36, 32	35, 32
	300	35, 30	36, 30	36, 30	36, 30	36, 30	36, 31	36, 31	36, 31	35, 32	35, 32	35, 33	35, 33
	600	35, 30	35, 32	36, 30	36, 31	36, 33	35, 32	36, 32	35, 35	34, 32	35, 35	33, 34	33, 35
	1200	34, 33	36, 34	34, 32	35, 33	35, 34	35, 34	32, 36	32, 36	33, 35	33, 35	33, 36	33, 36
	2400	36, 34	34, 34	33, 35	36, 36	35, 35	36, 35	33, 36	30, 36	31, 36	33, 36	31, 36	30, 36
	4800	34, 34	35, 35	35, 35	36, 36	36, 36	31, 36	32, 36	30, 36	28, 36	30, 36	29, 36	30, 36
Small Effect Sizes: $\Delta \leq .5\sigma_X$													

Table A.10: Each cell entry represents how many of the 36 (F, θ, δ) with $\Delta \leq .5\sigma_X$ correspond to $90\% CI_{PsL}(\delta), 90\% CI_{MoM}(\delta)$ that have simulated coverage probability at least .865. Color coded backgrounds correspond to the recommendations from **Figure 4.12**.



90% CR(θ, δ)		n:m											
		1:29	1:19	1:14	1:9	1:5	1:3	1:2	2:3	1:1	3:2	2:1	3:1
N	60	30, 0	34, 0	29, 3	36, 8	36, 30	36, 35	36, 36	36, 36	36, 36	36, 36	36, 36	36, 36
	120	27, 3	29, 7	34, 12	36, 26	36, 36	36, 36	36, 36	36, 36	36, 36	36, 36	36, 36	36, 36
	180	27, 4	34, 18	34, 27	36, 35	36, 35	36, 36	36, 36	36, 36	36, 36	36, 36	36, 36	36, 36
	300	30, 21	34, 34	36, 35	36, 35	36, 36	36, 36	36, 36	36, 36	36, 36	36, 36	36, 36	36, 36
	600	34, 32	36, 35	36, 36	36, 36	36, 36	36, 36	36, 36	36, 36	36, 36	36, 36	36, 36	36, 36
	1200	36, 36	36, 36	36, 36	36, 36	36, 36	35, 36	34, 36	35, 36	36, 36	34, 36	34, 36	35, 36
	2400	36, 36	36, 36	36, 36	36, 36	35, 36	36, 36	36, 36	34, 36	34, 36	34, 36	34, 36	34, 36
	4800	36, 36	36, 36	36, 36	36, 36	36, 36	36, 36	35, 36	35, 36	35, 36	34, 36	34, 36	34, 36
Large Effect Sizes: $\Delta >$ $.5\sigma_X$													

Table A.11: Each cell entry represents how many of the 36 (F, θ, δ) with $\Delta > .5\sigma_X$ correspond to $90\%CR_{PsL}(\theta, \delta), 90\%CR_{MoM\Delta}(\theta, \delta)$ that have simulated coverage probability at least .865. Color coded backgrounds correspond to the recommendations from **Figure 4.15**.



90% CR(θ, δ)		n:m											
		1:29	1:19	1:14	1:9	1:5	1:3	1:2	2:3	1:1	3:2	2:1	3:1
N	60	35, 5	36, 15	36, 28	36, 36	36, 36	36, 36	36, 36	36, 36	36, 36	36, 36	36, 36	35, 36
	120	36, 24	36, 33	36, 35	36, 36	36, 36	36, 36	36, 36	36, 36	36, 36	36, 36	36, 36	36, 36
	180	36, 30	36, 34	35, 36	36, 35	36, 36	36, 36	36, 36	36, 36	35, 36	36, 36	36, 36	35, 36
	300	36, 34	36, 34	36, 36	36, 35	36, 36	36, 36	36, 36	36, 36	35, 36	34, 36	35, 36	35, 36
	600	35, 36	36, 36	35, 36	36, 36	36, 36	35, 36	36, 36	35, 36	33, 36	34, 36	33, 36	34, 36
	1200	35, 36	35, 36	35, 36	35, 36	35, 36	34, 35	31, 36	32, 36	29, 36	31, 36	30, 36	31, 36
	2400	35, 36	34, 36	35, 35	35, 36	35, 36	35, 36	29, 36	29, 36	28, 36	27, 36	28, 36	28, 36
	4800	34, 36	36, 36	34, 36	35, 36	36, 36	30, 36	29, 36	25, 36	24, 36	24, 36	23, 36	25, 36
Small Effect Sizes: $\Delta \leq$ $.5\sigma_X$													

Table A.12: Each cell entry represents how many of the 36 (F, θ, δ) with $\Delta \leq .5\sigma_X$ correspond to $90\%CR_{PsL}(\theta, \delta), 90\%CR_{MoM\Delta}(\theta, \delta)$ that have simulated coverage probability at least .865. Color coded backgrounds correspond to the recommendations from **Figure 4.16**.

A.10 R Code

Normal MLE (with Control Data) Code

```
NormEM2loc = function(dat, l, eps = 1e-5, maxiter = 1000,
                      plot = FALSE, verbose = FALSE, est.only = TRUE)
{
  if(length(dat) != length(l)){stop("data vector/matrix (dat) does not match label vector (l) in length")}
  # Input
  mu0 = 2  # Assumes two components
  # Assumes delta > 0
  # Input

  p.mean.dnorm = function(X,p){return(sum(colMeans(p*dnorm(X)))/bw)}

  N <- length(dat)
  if(length(mu0) > 1){g <- length(mu0)} else{g <- mu0}
  unlabeled = which(is.na(l))
  labeled = which(!is.na(l))
  labs = unique(l[labeled])

  ### Just initialize by different cut-points with absolute membership
  if(length(unlabeled)>0)
  {
    ord.unlab = order(dat[unlabeled])
    starts = min(c(length(unlabeled),6))
    ind = round(seq(from = 1, to = length(unlabeled), length.out = starts))
    M = matrix(NA,nrow=length(unlabeled),ncol=starts)
    count = 0
    for(i in ind)
    {
      count = count+1
      M[,count] = rep(1,length(unlabeled)) +
        (1:length(unlabeled)) %in% ord.unlab[c(rep(F,i-1),rep(T,length(unlabeled)-i+1))]
    }
  }



  list.lambda = list()
  list.mu = list()
  list.sigma = list()
  list.Log.lik = list()
  max.iters = NULL
  ans = matrix(NA,nrow=0,ncol=8)
  colnames(ans) = c("theta-hat","delta-hat","Log-lik","sigma-hat","mu1-hat","mu2-hat","iter","max.LL")

  if(length(labeled)>0){starts.seq = 1:starts}else{starts.seq = 2:starts} # Assumes labeled data comes from component 1
  for(s in starts.seq)
  {

    # Initialize z.hat
    l[unlabeled] = M[,s] # Fill in unlabeled
    z.hat <- matrix(0, nrow = N, ncol = g)
    for (j in 1:g)
    {
      z.hat[l == j, j] <- 1
    }
    z.hat[l==.5,] <- c(.5,.5) # assumes g = 2

    iter <- 0
    finished <- FALSE
    lambda <- mu <- matrix(0, maxiter, g)
    sigma <- Log.lik <- NULL

    while (!finished) {
      iter <- iter + 1
      t0 = proc.time()
      lambda[iter, ] <- colMeans(z.hat[unlabeled,])
      mu[iter, ] <- colMeans(sweep(z.hat, 1, dat, "*"))/colMeans(z.hat) # changed to colMeans on top
      ei = matrix(dat - rep(mu[iter,],each = N), ncol = g)
      sigma[iter] <- sqrt(sum(z.hat*(ei^2))/N)
      fkernel <- matrix(dnorm(dat, mean = rep(mu[iter,],each = N), sd = sigma[iter]), ncol = g)
      Log.lik[iter] = sum(log(lambda[iter,1]*dnorm(dat[unlabeled],
                                                   mean=mu[iter,1],
                                                   sd=sigma[iter]) +
                          lambda[iter,2]*dnorm(dat[unlabeled],
                                               mean=mu[iter,2],
                                               sd=sigma[iter]))) +
              sum(dnorm(dat[labeled],mean=mu[iter,1],sd=sigma[iter],log=TRUE)) #!# to check log-likelihood
      lambda.f <- sweep(fkernel, 2, lambda[iter, ], "*")
      z.hat[unlabeled,] <- lambda.f[unlabeled,]/rowSums(lambda.f[unlabeled,])
      finished <- iter >= maxiter
      if (iter > 1)
      {
        change <- Log.lik[iter] - Log.lik[iter-1]
        finished <- (finished | (change < eps))
      }


      if(plot & verbose & iter==1)
      {
        if(length(labeled)>1)
        {
          hist(dat[labeled],col="grey",
               breaks = (10 + length(labeled)/20)/(max(log10(max(length(labeled)-1000,1))/1.5,1)),
               freq = F, main = "Histogram of Labeled Data and Initial KDE", xlab = "Data")
          legend("topleft", lty = 1, lwd = 2, col = g, legend = "f Density Estimate")
          for(j in 1:g)
          {
            lines(x=sort(dat[unlabeled]) - mu[iter, j] + mu[iter,1],
                  y=fkernel[unlabeled[order(dat[unlabeled])],j],col=j, lwd = 2) # the + mu[iter,1] is project specific
          }
        }

        hist(dat[unlabeled],col="grey",
             breaks = (10 + length(unlabeled)/20)/(max(log10(max(length(unlabeled)-1000,1))/1.5,1)),
             freq = F, main = "Histogram of Unlabeled Data and Initial KDE", xlab = "Data")
        legend("topleft",lty = 2, lwd = 2, col = 1, legend = "Mixture Estimate")
        lines(x=sort(dat[unlabeled]),
              y=rowSums(sweep(fkernel[unlabeled[order(dat[unlabeled])],],2,lambda[iter,],"*")),
              lty = 2, lwd = 2)
      }
      if (verbose) {
        t1 <- proc.time()
        cat("iteration ", iter, "  lambda ", round(lambda[iter,
        ], 4), "  mu ", round(mu[iter, ], 4))
        cat(" time", (t1 - t0)[3], "\n")
      }
      if(diff(mu[iter,])<0){
        mu[iter,] = rep(mean(dat),2)
        lambda[iter,] <- c(1,0)
        sigma[iter] = sd(dat)*(length(dat)-1)/length(dat)
        Log.lik[iter] = sum(log(lambda[iter,1]*dnorm(dat[unlabeled],mean=mu[iter,1],sd=sigma[iter]) +
                                  lambda[iter,2]*dnorm(dat[unlabeled],mean=mu[iter,2],sd=sigma[iter]))) +
          sum(dnorm(dat[labeled],mean=mu[iter,1],sd=sigma[iter],log=TRUE))
        break}
    } #Ends While loop
    if(plot & verbose){
      plot(Log.lik, type = "l", lwd = 3,
           main = "Log-Likelihood over the iterations",
           xlab = "Iteration (t)", ylab = "Log-Likelihood"
      )

      plot(sigma, type = "l", lwd = 3,
           main = "Sigma over the iterations",
           xlab = "Iteration (t)", ylab = expression(sigma^t)
      )

      plot(apply(mu[1:iter,],1,diff)/sigma, type = "l", lwd = 3,
           main = "K = delta/sigma over the iterations",
           xlab = "Iteration (t)", ylab = expression(K^t)
      )

      plot(lambda[1:iter,2]*apply(mu[1:iter,],1,diff), type = "l", lwd = 3,
           main = "Delta over the iterations",
           xlab = "Iteration (t)", ylab = expression(Delta^t)
      )

      plot(x = mu[1:iter,1], y = mu[1:iter,2],
           type = "p", pch = 16, cex = .5,
           main = "Mu1 and Mu2 over the iterations",
           xlab = expression(mu[1]^t), ylab = expression(mu[2]^t)
      )
      text(1,x = mu[1,1], y = mu[1,2], cex = 1.5)
      text(floor(iter/2), x = mu[floor(iter/2),1], y = mu[floor(iter/2),2], cex = 1.5)
      text(iter, x = mu[iter,1], y = mu[iter,2], cex = 1.5)

      plot(x = apply(mu[1:iter,],1,diff), lambda[1:iter,2],
           type = "p", pch = 16, cex = .5,
           main = "theta and delta over the iterations",
           xlab = expression(delta^t), ylab = expression(theta^t))
      text(1,x = diff(mu[1,]), y = lambda[1,2], cex = 1.5)
      text(floor(iter/2),x = diff(mu[floor(iter/2),]), y = lambda[floor(iter/2),2], cex = 1.5)
      text(iter,x = diff(mu[iter,]), y = lambda[iter,2], cex = 1.5)

    }
```

```
    list.lambda = c(list.lambda,list(matrix(lambda,ncol=g)))
    list.mu = c(list.mu,list(matrix(mu,ncol=g)))
    list.sigma = c(list.sigma,list(sigma))
    list.Log.lik = c(list.Log.lik,list(Log.lik))
    max.iters = c(max.iters,iter)

    ans = rbind(ans,c(lambda[iter,2], mu[iter,2] - mu[iter,1],
                      Log.lik[iter],sigma[iter],mu[iter,1],mu[iter,2],iter,NA))

  } # Ends for loop
  if(plot)
  {
    est.ind = which.max(ans[,3])
    est.iter = max.iters[which.max(ans[,3])]
    plot(list.Log.lik[[est.ind]], type = "l", lwd = 3,
         main = "Log-Likelihood over the iterations", cex.main = .9,
         xlab = "Iteration (t)", ylab = "Log-Likelihood"
    )

    plot(list.sigma[[est.ind]], type = "l", lwd = 3,
         main = bquote(bold(sigma ~ "over the iterations")),
         xlab = "Iteration (t)", ylab = "", cex.main = 1.25
    )
    mtext(expression(sigma^t),side = 2, line = 2.5, las = 1, cex = 1.25) # Add y label manually

    plot(apply(list.mu[[est.ind]][1:est.iter,],1,diff)/list.sigma[[est.ind]], type = "l", lwd = 3,
         main = bquote(bold(K == delta/sigma ~ "over the iterations")),
         xlab = "Iteration (t)", ylab = "", cex.main = 1.25
    )
    mtext(expression(K^t),side = 2, line = 2.5, las = 1, cex = 1.25) # Add y label manually


    plot(list.lambda[[est.ind]][1:est.iter,2]*apply(list.mu[[est.ind]][1:est.iter,],1,diff), type = "l", lwd = 3,
         main = bquote(bold(Delta ~ "over the iterations")),
         xlab = "Iteration (t)", ylab = "", cex.main = 1.25
    )
    mtext(expression(Delta^t), side = 2, line = 2.5, las = 1, cex = 1.25) # Add y label manually

    plot(x = list.mu[[est.ind]][1:est.iter,1], y = list.mu[[est.ind]][1:est.iter,2],
         type = "p", pch = 16, cex = .5,
         main = bquote(bold(mu[1] ~ "and" ~ mu[2] ~ "over the iterations")),
         xlab = "", ylab = "", cex.main = 1.25
    )
    mtext(expression(mu[1]^t), side = 1, line = 2.5, las = 1, cex = 1.25) # Add x label manually
    mtext(expression(mu[2]^t), side = 2, line = 2.5, las = 1, cex = 1.25) # Add y label manually
    text(1,x = list.mu[[est.ind]][1,1], y = list.mu[[est.ind]][1,2], cex = 1.5)
    text(floor(est.iter/2), x = list.mu[[est.ind]][floor(est.iter/2),1],
         y = list.mu[[est.ind]][floor(est.iter/2),2], cex = 1.5)
    text(est.iter, x = list.mu[[est.ind]][est.iter,1], y = list.mu[[est.ind]][est.iter,2], cex = 1.5)

    plot(x = apply(list.mu[[est.ind]][1:est.iter,],1,diff), list.lambda[[est.ind]][1:est.iter,2],
         type = "p", pch = 16, cex = .5,
         main = bquote(bold(theta ~ "and" ~ delta ~ "over the iterations")),
         xlab = "", ylab = "")
    mtext(expression(delta^t), side = 1, line = 2.5, las = 1, cex = 1.25) # Add x label manually
    mtext(expression(theta^t), side = 2, line = 2.5, las = 1, cex = 1.25) # Add y label manually
    text(1,x = diff(list.mu[[est.ind]][1,]), y = list.lambda[[est.ind]][1,2], cex = 1.5)
    text(floor(est.iter/2),x = diff(list.mu[[est.ind]][floor(est.iter/2),]),
         y = list.lambda[[est.ind]][floor(est.iter/2),2], cex = 1.5)
    text(est.iter,x = diff(list.mu[[est.ind]][est.iter,]), y = list.lambda[[est.ind]][est.iter,2], cex = 1.5)
  }

  ans[,8] = (ans[,"Log-lik"]==max(ans[,"Log-lik"]))
  if(est.only){if(ans[which.max(ans[,"Log-lik"]),1]<.0001){
    return(c(0,0))}else{return(ans[which.max(ans[,"Log-lik"]),1:2])}
    }else{return(ans)}

  #return(c(lambda[iter,2], mu[iter,2] - mu[iter,1],Log.lik[iter]))
}

# Generate m = 100 observations from N(0,1) for the control group and n = 100 observations from .3N(0,1) + .7N(2,1) for the trt group.
m = 100
n = 100
true.theta = .7
true.delta = 2
x = rnorm(m)
z = sample(c(0,1), size = n, replace = TRUE, prob = c(1-true.theta,true.theta))
y = rnorm(n) + true.delta*z


# Find the Normal Maximum Likelihood
NormEM2loc(dat = c(x,y), l = c(rep(1,m),rep(NA,n)), plot = TRUE)
```

Semi-Supervised Semi-Parametric EM-like Algorithm

```
## Semi-Parametric EM Algorithm(s) - EM# - 6 versions
ssSpEMloc = function(dat, l, bw = bw.nrd0(dat[!is.na(l) & l==1]), eps = min(c(1e-5*sd(dat[!is.na(l) & l==1]),1e-3)), maxiter = 100,
                     all.data.f = FALSE, stochastic = FALSE, symmetric = FALSE,
                     plot = FALSE, verbose = FALSE, est.only = TRUE,
                     delta.pos = TRUE)
{

  ### Warnings and Errors
  if(all.data.f == FALSE & stochastic == TRUE){
    warning("stochastic = TRUE only works with all.data.f = TRUE. Output uses all.data.f = TRUE")
    }
  if(length(dat) != length(l)){
    stop("data vector/matrix (dat) does not match label vector (l) in length")
    }


  # Internally Define Kernel Density Estimation Function
  KDE = function(f.data,y,bw=bw.nrd0(f.data),df=3,var.adj=TRUE)
  {
    dat = c(f.data,y)
    std.dat = (dat - mean(dat))/sd(dat)
    std.datx = (f.data - mean(f.data))/sd(f.data)
    std.daty = (y - mean(y))/sd(y)
    if(df=="adj1"){df <- 3 + 1/( max(0,max(abs(std.dat))-3) )}
    if(df=="adj2"){df <- 3 + 1/( max(0,max(abs(std.datx)-3,max(abs(std.daty)))-3) )}
    if(df < 3){df <- 3}
    sig = sqrt(1/(1-2/df))
    f.hat = NULL
    if(var.adj)
    {
      #!# Vectorize KDE operations
      #dist = sweep(matrix(f.data),2,y)
      #dens = dt(sig*dist/bw, df=df)/bw
      for(i in 1:length(y))
      {
        f.hat[i] = sig*mean(dt( sig*(f.data-y[i])/bw , df=df))/bw
      }
    }else
    {
      for(i in 1:length(y))
      {
        f.hat[i] = mean(dt( (f.data-y[i])/bw , df=df))/bw
      }
    }
    return(f.hat)
  }



  ## Input ##
  mu0 = 2  # two components hard-coded in
  # l = 1 indicates component 1 (i.e. 'non-responder' or control data)
  # This function assumes delta = mu2 - mu1 > 0
  ## Input ##

  # total sample size
  n <- length(dat)

  # Number of components (hard-coded to be 2)
  if(length(mu0) > 1){m <- length(mu0)}else{m <- mu0}

  # which observations do not have labels (and come from the mixture)
  unlabeled = which(is.na(l))

  # which observations are labeled (and come directly from the labeled component)
  labeled = which(!is.na(l))

  # which components provide at least one labeled observation
  labs = unique(l[labeled])

  # Store these initial labels that actually come from the data (not predictions to be updated in the algorithm)
  init.class = l

  ## Fill in initialization of class membership for observations without component label ##
  if(length(unlabeled)>0)
  {
    obj = kmeans(dat[unlabeled], centers = mu0)
    shifted = obj$cluster == which.max(obj$centers) # This cluster labeling assumes two components and delta > 0. component 2
    unshifted = obj$cluster == which.min(obj$centers) # This cluster labeling assumes two components and delta > 0. component 1
    if( !all.equal(shifted + unshifted, rep(1,length(unlabeled))) ){stop("uhoh")} # Sanity check
    init.class[unlabeled][shifted] = 2
    init.class[unlabeled][unshifted] = 1
  }
```

```
## Initialize Necessary Elements ##
z.hat <- matrix(0, nrow = n, ncol = m)
fkernel <- matrix(0, nrow = n, ncol = m)
p.mean.dnorm = function(X,p){return(sum(colMeans(p*dnorm(X)))/bw)}
tt0 <- proc.time()
#lambda <- rep(1/m, m)
#kmeans <- kmeans(dat, mu0)
for (j in 1:m) {
  z.hat[init.class == j, j] <- 1
}
iter <- 0
if (stochastic) {
  sumpost <- matrix(0, n, m)
}
finished <- FALSE
lambda <- mu <- matrix(0, maxiter, m)


if(all.data.f)
{
  while (!finished) {
    iter <- iter + 1
    t0 <- proc.time()

    # theta-hat = the average weight in component 2 of the z.hats among unlabeled data.
    lambda[iter, ] <- colMeans(z.hat[unlabeled,])

    # mu1-hat average of control and (1-z.hat) weighted treatment obs. mu2-hat is z.hat weighted average of treatment data.
    mu[iter, ] <- apply(sweep(z.hat, 1, dat, "*"), 2, mean)/colMeans(z.hat)

    ## Compute f-hat at ui - muj for all i,j.
    if(stochastic)
    {
      ### Generate simulated component membership (for labeled data, it’s automatically the known label...
        # for unlabeled data, according to the current weight for each component).
      z = matrix(0, nrow = n, ncol = m)
      z[labeled,] = z.hat[labeled,]
      z[unlabeled,] <- t(apply(z.hat[unlabeled,], 1, function(prob) rmultinom(1, 1, prob)))

      # Recenter each observation so that combined re-centered data has mean 0.
      dat.t <- dat-apply(sweep(z,2,mu[iter, ],"*"),1,sum)
      if(symmetric)
      {
        for(j in 1:m)
        {
          for(i in unlabeled)
          {
            ### KDE with normal kernel, one version on re-centered data, one version on mirror image of re-centered data
              # - those two version averaged ensures symmetric f-hat.
            fkernel[i,j] = mean(c( mean((1/bw)*dnorm(((dat[i]-mu[iter,j])-dat.t)/bw)),
                                   mean((1/bw)*dnorm((-(dat[i]-mu[iter,j])-dat.t)/bw)) ))
          }
        }
      }else{
        for(j in 1:m)
        {
          for(i in unlabeled)
          {
            ### KDE with normal kernel on re-centered data.
            fkernel[i,j] = mean((1/bw)*dnorm(((dat[i]-mu[iter,j])-dat.t)/bw))
          }
        }
      }
    }else
    {

      ## Begin Deterministic KDE ##
      if(symmetric)
      {
        for(j in 1:m)
        {
          M = matrix(((dat[1] - mu[iter,j]) - (rep(dat,m)-rep(mu[iter,],each=n)))/bw,nrow=n,ncol=m)
          for(i in 1:n)
          {
            # M.prime is the ’reflected’ data used for the symmetrization step.
            M.prime = M - 2*(dat[i]-mu[iter,j])/bw

            #Symmetric f-hat
            fkernel[i,j] = mean(c(p.mean.dnorm(M,z.hat),p.mean.dnorm(M.prime,z.hat)))
            M = M - M[min(i+1,n),j]
          }
        }
      }else{
        for(j in 1:m)
        {
          M = matrix(((dat[1] - mu[iter,j]) - (rep(dat,m)-rep(mu[iter,],each=n)))/bw,nrow=n,ncol=m)
          for(i in 1:n)
          {
            # f-hat
            fkernel[i,j] = p.mean.dnorm(M,z.hat)
            M = M - M[min(i+1,n),j]
          }
        }
      }
      ## End deterministic KDE ##

    }

    # updated lambda*f (i.e. pi*f, (1-theta,theta)*f).
    lambda.f <- sweep(fkernel, 2, lambda[iter, ], "*")

    # update weights
    z.hat[unlabeled,] <- lambda.f[unlabeled,]/rowSums(lambda.f[unlabeled,])

    ## Determine if time to stop ##
    finished <- iter >= maxiter
    if (stochastic) {
        # keep track of cumulative sum of weights for stochastic estimate of mixing proportions
        sumpost <- sumpost + z.hat
    }else if (iter > 1) {
      change <- c(lambda[iter, ] - lambda[iter - 1, ],
                  mu[iter, ] - mu[iter - 1, ])
      finished <- finished | (max(abs(change)) < eps)
    }

    ## Possible Output ##
    if(plot & iter==1 & length(labeled)>1 & length(unlabeled)>1)
    {
      hist(dat[labeled],col="grey",breaks = (10 + n/20)/(max(log10(max(n-1000,1))/1.5,1)),
           freq = F, main = "Histogram of Labeled Data and Initial KDE", xlab = "Data")
      for(j in 1:m)
      {
        # the + mu[iter,1] is project specific, plot F(u) instead of F with mean 0
        lines(x=sort(dat[unlabeled]) - mu[iter, j] + mu[iter,1],
              y=fkernel[unlabeled[order(dat[unlabeled])],j],col=j, lwd = 2)
        if(!all.data.f){break}
      }
      legend("topleft", lty = 1, lwd = 2, col = j, legend = "f Density Estimate")

      hist(dat[unlabeled],col="grey",breaks = (10 + n/20)/(max(log10(max(n-1000,1))/1.5,1)),
           freq = F, main = "Histogram of Unlabeled Data and Initial KDE", xlab = "Data")
      legend("topleft",lty = 2, lwd = 2, col = 1, legend = "Mixture Estimate")
      lines(x=sort(dat[unlabeled]),
            y=rowSums(sweep(fkernel[unlabeled[order(dat[unlabeled])],],2,lambda[iter,],"*")),
            lty = 2, lwd = 2)
    }
    if (verbose) {
      t1 <- proc.time()
      cat("iteration ", iter, "  lambda ", round(lambda[iter,
      ], 4), "  mu ", round(mu[iter, ], 4))
      cat(" time", (t1 - t0)[3], "\n")
    }

  } #Ends While loop
}# Ends if all.data.f

if(!all.data.f)
{
  # Compute the mean for each component with labeled data (just mu_1 for dissertation)
  mu[, labs] <- rep( apply(sweep(matrix(z.hat[labeled,labs],nrow=length(labeled)), 1, dat[labeled], "*"),
                           2, mean)/colMeans(matrix(z.hat[labeled,labs],nrow=length(labeled))), each = nrow(mu) )

  ## Centered KDE on control data ##

  # Recenter labeled observation so that combined re-centered data has mean 0.
  dat.t <- dat[labeled]-apply(sweep(matrix(z.hat[labeled,labs],nrow=length(labeled)),2,mu[1, labs],"*"),1,sum)

  while(!finished){
    iter <- iter+1
    t0 <- proc.time()

    # mixing proportions calculated as average component weight among unlabeled observations.
    lambda[iter, ] <- colMeans(z.hat[unlabeled,])

    # Computes averages for components without labeled data by using weighted average of (unlabeled) observations.
    # (mu2-hat is z.hat weighted average of treatment data).
    mu[iter, -labs] <- apply(sweep(matrix(z.hat[unlabeled,-labs],nrow=length(unlabeled)), 1, dat[unlabeled], "*"),
                             2, mean)/colMeans(matrix(z.hat[unlabeled,-labs],nrow=length(unlabeled)))

    if(symmetric)
    {
      for(j in 1:m)
      {
        fkernel[,j] <- apply(cbind(KDE(f.data = dat.t, y = dat - mu[iter,j], df=Inf, bw=bw, var.adj=TRUE),
                                    KDE(f.data = dat.t, y = -(dat - mu[iter,j]), df=Inf, bw=bw, var.adj=TRUE)
        ),1,mean) # Average of kernel density estimates on re-centered and mirror image of re-centered data.
      }
    }else{
      for(j in 1:m)
      {
        # KDE for re-centered data
        fkernel[,j] <- KDE(f.data = dat.t, y = dat - mu[iter,j], df=Inf, bw=bw, var.adj=TRUE)
      }
    }

    # mixing proportions times f
    lambda.f <- sweep(fkernel, 2, lambda[iter, ], "*")

    # computes updated component weighhts for unlabeled data
    z.hat[unlabeled,] <- lambda.f[unlabeled,]/rowSums(lambda.f[unlabeled,])

    ## Determine if time to stop ##
    finished <- iter >= maxiter
    if (iter > 1) {
      change <- c(lambda[iter, ] - lambda[iter - 1, ],
                  mu[iter, ] - mu[iter - 1, ])
      finished <- finished | (max(abs(change)) < eps)
    }

    ## Possible Output ##
    if(plot & iter==1 & length(labeled) > 1 & length(unlabeled) > 1)
    {
      hist(dat[labeled],col="grey",breaks = (10 + n/20)/(max(log10(max(n-1000,1))/1.5,1)),
           freq = F, main = "Histogram of Labeled Data and Initial KDE", xlab = "Data")
      comp.ind = 1; if(!all.data.f){comp.ind = 2}
      for(j in comp.ind:m)
      {
        lines(x=sort(dat[unlabeled]) - mu[iter, j] + mu[iter,1],
              y=fkernel[unlabeled[order(dat[unlabeled])],j],col=j, lwd = 2)
      }
      legend("topleft", lty = 1, lwd = 2, col = j, legend = "f Density Estimate")

      hist(dat[unlabeled],col="grey",breaks = (10 + n/20)/(max(log10(max(n-1000,1))/1.5,1)),
           freq = F, main = "Histogram of Unlabeled Data and Initial KDE", xlab = "Data")
      legend("topleft",lty = 2, lwd = 2, col = 1, legend = "Mixture Estimate")
      lines(x=sort(dat[unlabeled]),
            y=rowSums(sweep(fkernel[unlabeled[order(dat[unlabeled])],],2,lambda[iter,],"*")),
            lty = 2, lwd = 2)
    }
    if (verbose) {
      t1 <- proc.time()
      cat("iteration ", iter, "  lambda ", round(lambda[iter,
      ], 4), "  mu ", round(mu[iter, ], 4))
      cat(" time", (t1 - t0)[3], "\n")
    }

  } # End While Loop

} # End if !all.data.f


### Finishing Touches ###
if (verbose) {
  tt1 <- proc.time()
  cat("lambda ", round(lambda[iter, ], 4))
  cat(", total time", (tt1 - tt0)[3], "s\n")
}

if(plot)
{
  hist(dat[labeled],col="grey",breaks = (10 + n/20)/(max(log10(max(n-1000,1))/1.5,1)),
       freq = F, main = "Histogram of Labeled Data and Final KDE", xlab = "Data")
  comp.ind = 1; if(!all.data.f){comp.ind = 2}
  for(j in comp.ind:m)
  {
    lines(x=sort(dat[unlabeled]) - mu[iter, j] + mu[iter,1],
          y=fkernel[unlabeled[order(dat[unlabeled])],j],col=j, lwd = 2)
  }
  legend("topleft", lty = 1, lwd = 2, col = j, legend = "f Density Estimate")

  hist(dat[unlabeled],col="grey",breaks = (10 + n/20)/(max(log10(max(n-1000,1))/1.5,1)),
       freq = F, main = "Histogram of Unlabeled Data and Final KDE", xlab = "Data")
  legend("topleft",lty = 2, lwd = 2, col = 1, legend = "Mixture Estimate")
  lines(x=sort(dat[unlabeled]),
        y=rowSums(sweep(fkernel[unlabeled[order(dat[unlabeled])],],2,lambda[iter,],"*")),
        lty = 2, lwd = 2)

  plot(x = 1:iter, y = lambda[1:iter,1], ylim = c(0,1), type = "l", lwd = 2,
       main = "Estimates of Mixing Proportions throughout Algorithm", ylab = "Proportion")
  for(j in 2:m)
```

```
    {
      lines(x = 1:iter, y = lambda[1:iter,j], col = j, lwd = 2)
    }


    plot(x = 1:iter, y = mu[1:iter,1], ylim = range(mu), type = "l", lwd = 2,
         main = "Estimates of Component Means throughout Algorithm", ylab = "Mean")
    for(j in 2:m)
    {
      lines(x = 1:iter, y = mu[1:iter,j], col = j, lwd = 2)
    }

    if(m==2){
      plot(x = 1:iter, y = apply(mu[1:iter,],1,diff),
           ylim = range(c(0,apply(mu[1:iter,],1,diff))),
           type = "l", col = 2, lwd = 2, main = bquote("Estimates of" ~ delta ~ "throughout Algorithm"),
           ylab = bquote(delta))
    }
  }



  if (stochastic) {
    if(est.only)
    {
      if(m==2 & delta.pos & (diff(mu[iter,])<=0 | lambda[iter,2]<.0001)){return(c(0,0))}
      return(c(colMeans(lambda)[2],diff(colMeans(mu))))
    }else{
      return(structure(list(data = dat, posteriors = sumpost/iter,
                            bandwidth = bw, lambdahat = colMeans(lambda),
                            muhat = colMeans(mu), symmetric = symmetric),
                       class = "Adapted from npEM"))
    }
  }
  else {
    if(est.only)
    {
      if(m==2 & delta.pos & (diff(mu[iter,])<=0 | lambda[iter,2]<.0001)){return(c(0,0))}
      return(c(lambda[iter,2],diff(mu[iter,])))
    }else{
      return(structure(list(data = dat, posteriors = z.hat, bandwidth = bw,
                            lambdahat = lambda[iter, ], muhat = mu[iter, ],
                            symmetric = symmetric), class = "Adapted from npEM"))
    }
  }
}

# Generate 50 observations from N(0,1) for the control group and 50 observations from .3N(0,1) + .7N(2,1) for the treatment group.
x = rnorm(50)
z = sample(c(0,1), size = 50, replace = TRUE, prob = c(.3,.7))
y = rnorm(50) + 2*z

# Find the Normal Maximum Likelihood
ssSpEMloc(dat = c(x,y), l = c(rep(1,50),rep(NA,50)), plot = TRUE)
```

Pseudo-Likelihood Inference Code

Dependencies

```
    install.packages("sgt", repos = "http://cran.us.r-project.org")
    library("sgt")
    install.packages("logcondens", repos = "http://cran.us.r-project.org")
    library("logcondens")
```

Function Code

```
psl.inf = function(f.data,y,f.est="mLCD",bw = bw.nrd0(f.data),df=3, var.adj = TRUE, level, finite.area = FALSE,
                  plot=FALSE,true.theta=NA,true.delta=NA,
                  mu = NA, sigma = NA, lambda = NA, p = NA, q = NA)
{

  ### Initialize Important Quantities ###
  # sequence of possible theta for grid search
  th = seq(from = .01, to = 1, length.out = 100)
    # Add true.theta to grid search
    if(!is.na(true.theta)){
      less.th = sum(th < true.theta)
      if(less.th < length(th)){
        last.th.ind <- (less.th+1):length(th)
      }else{last.th.ind <- 0}
      th <- c(th[0:less.th],true.theta,th[last.th.ind])
    }
  # sequence of possible delta for grid search
  del = seq(from = .1*sd(f.data), to = 6*sd(f.data), length.out = 60)
    # Add true.delta to grid search if known
    if(!is.na(true.delta)){
      less.del = sum(del < true.delta)
      if(less.del < length(del)){
        last.del.ind <- (less.del+1):length(del)
      }else{last.del.ind <- 0}
      del <- c(del[0:less.del],true.delta,del[last.del.ind])
    }

  # Internally define helper function for kernel density estimation
  KDE = function(f.data,y,bw=bw.nrd0(f.data),df=3,var.adj=TRUE)
  {
    dat = c(f.data,y)
    std.dat = (dat - mean(dat))/sd(dat)
    std.datx = (f.data - mean(f.data))/sd(f.data)
    std.daty = (y - mean(y))/sd(y)
    if(df=="adj1"){df <- 3 + 1/( max(0,max(abs(std.dat))-3) )}
    if(df=="adj2"){df <- 3 + 1/( max(0,max(abs(std.datx)-3,max(abs(std.daty)))-3) )}
    if(df < 3){df <- 3}
    sig = sqrt(1/(1-2/df))
    f.hat = NULL
    if(var.adj)
    {
      #!# Vectorize KDE operations
      #dist = sweep(matrix(f.data),2,y)
      #dens = dt(sig*dist/bw, df=df)/bw
      for(i in 1:length(y))
      {
        f.hat[i] = sig*mean(dt( sig*(f.data-y[i])/bw , df=df))/bw
      }
    }else
    {
      for(i in 1:length(y))
      {
        f.hat[i] = mean(dt( (f.data-y[i])/bw , df=df))/bw
      }
    }
    return(f.hat)
  }

  # Internally define helper function for modifying Log-Condave Maximum Likelihood Density Estimate
  mod.fhat = function(res,eval)
  {

    ends = range(res$knots)
    n = res$n
    w = c(1/n,(n-2)/n,1/n)

    ### Create indices for eval vector saying which segment it’s in.
    lower = which(eval < ends[1])
    middle = which(eval >= ends[1] & eval <= ends[2])
    upper = which(eval > ends[2])
    if(!all(sort(c(lower,middle,upper)) == 1:length(eval))){"missing indices?"}

    h1 = exp(res$phi[1])
    a1 = (n-2)*h1
    k1 = w[1]*(n-2)*h1*exp(-(n-2)*h1*ends[1])

    h2 = exp(res$phi[res$m])
    a2 = -(n-2)*h2
    k2 = w[3]*(n-2)*h2*exp((n-2)*h2*ends[2])


    mfhat = NULL
    mfhat[lower] = k1*exp(a1*eval[lower])
    mfhat[middle] = w[2]*fhat(res=res,eval=eval[middle])
    mfhat[upper] = k2*exp(a2*eval[upper])

    if(any(is.na(mfhat))){"Why still NAs?"}

    return(mfhat)
  }


  if(toupper(f.est) %in% c("KDE","KERNEL","KERN")){
    obj = NA

    # f-hat, density estimate (based on control data, using kernel density estimate) evaluated at y values
    fhat.y = KDE(f.data = f.data, y = y, bw = bw, df = df, var.adj = var.adj)
  }
  if(toupper(f.est) %in% c("LCD","LOG-CONCAVE","LOG-CON","LOG CONCAVE",
                           "LOG CON","LCON","LOG","LC","LCDENS","LC-DENS","MLCD"))
  {
    # Log-Concave MLE
    obj = activeSetLogCon(f.data)

    # f-hat, density estimate (based on control data, using modified log-concave MLE) evaluated at y values
    fhat.y = mod.fhat(res=obj,eval=y) #!# mod.fhat() helper function defined outside
  }
  if(toupper(f.est) %in% c("TRUTH","F"))
  {
    obj = NA

    # f-"hat", True/(user specified) density of f evaluated at y values
    fhat.y = dsgt(y,mu = mu, sigma = sigma, lambda = lambda, p = p, q = q)
  }

  # Function to evaluate fhat(y - deli)
  eval.fhat.yd = function(f.data,y,f.est,
                          bw,df,var.adj,
                          deli,obj,
                          mu,sigma,lambda,p,q)
  {
    if(toupper(f.est) %in% c("KDE","KERNEL","KERN"))
    {
      fhat.yd = KDE(f.data = f.data, y = y - deli, bw = bw, df = df, var.adj = var.adj)
    }
    if(toupper(f.est) %in% c("LCD","LOG-CONCAVE","LOG-CON","LOG CONCAVE",
                             "LOG CON","LCON","LOG","LC","LCDENS","LC-DENS","MLCD"))
    {
      fhat.yd = mod.fhat(res=obj,eval=y - deli)
    }
    if(toupper(f.est) %in% c("TRUTH","F"))
    {
      fhat.yd = dsgt(y - deli,mu = mu, sigma = sigma, lambda = lambda, p = p, q = q)
    }
    return(fhat.yd)
  }

  # Define function for computing bound (used for confidence intervals and regions)
  any.in = function(x,cl,df,c=1) any(x < c*qchisq(p=cl,df=df))

  # Define function for bi-linearly interpolating mean and var (used for Satterthwaite approximation)
  bilinear = function(M,N,p.trt,N.in,p.in)
  {
    if(nrow(M) != length(N)){stop("Length of N does not equal number of rows of M")}
    if(ncol(M) != length(p.trt)){stop("Length of p.trt does not equal number of columns of M")}
    rownames(M) <- N
    colnames(M) <- p.trt

    p.trt.ind = min(max(sum(p.in >= p.trt),1),length(p.trt)-1)
    N.ind = min(max(sum(N.in >= N),1),length(N)-1)
    x = p.in; y = N.in

    x0 = p.trt[p.trt.ind]
    y0 = N[N.ind]
    x1 = p.trt[p.trt.ind+1]
    y1 = N[N.ind+1]
    z00 = M[N.ind,p.trt.ind]
    z01 = M[N.ind,p.trt.ind+1]
    z10 = M[N.ind+1,p.trt.ind]
    z11 = M[N.ind+1,p.trt.ind+1]

    z.star = ( (x1-x)*(y1-y)*z00 + (x1-x)*(y-y0)*z10 + (x-x0)*(y1-y)*z01 + (x-x0)*(y-y0)*z11 )/( (x1-x0)*(y1-y0) )
    return(z.star)
  }


  # Define Satterthwaite constants
  C.CR <-
  matrix(c(1.026035,1.242881,1.353304,1.506900,1.947223,2.400741,3.182174,3.958607,4.799132,7.224214,7.252960,10.890202,
           1.022448,1.133521,1.209424,1.391784,1.656656,2.080179,2.636975,3.279664,4.080320,5.977822,7.938542,11.269434,
           1.019073,1.164717,1.142916,1.355192,1.677559,1.830471,2.272151,2.801275,3.889390,5.713839,7.915159,10.382594,
           1.012556,1.107921,1.156220,1.302759,1.473298,1.824668,2.109494,2.547117,3.325433,4.676027,5.816139,10.491207,
           1.039824,1.093934,1.076780,1.295238,1.324490,1.498298,1.806582,2.011168,2.645757,4.479356,5.330594,8.378515,
           1.043142,1.049957,1.064548,1.152818,1.214041,1.465731,1.609443,1.897742,2.238591,3.239241,4.035796,6.722422,
           1.064287,1.069622,1.055648,1.091301,1.199641,1.526891,1.485453,1.649860,2.092417,2.840655,3.687886,5.243302,
           1.012544,1.050787,1.053520,1.083536,1.181289,1.256195,1.427919,1.631067,1.783403,2.495140,2.798502,4.578211),
         nrow = 8, byrow = TRUE)

  DF.CR <-
  matrix(c(1.701194,1.548824,1.494056,1.474888,1.283580,1.161816,0.9736668, 0.8522719,0.8160677, 0.6404665,0.7258481, 0.5837087,
          1.910312,1.839065,1.788471,1.654943,1.510889,1.304269,1.1304029, 0.9825586,0.9068518, 0.7360351,0.6413862, 0.5770284,
          2.010585,1.844756,1.928034,1.709017,1.480560,1.459013,1.2921745, 1.1205915,0.9219614, 0.7481736,0.6264377, 0.6026179,
          2.084614,1.976540,1.937847,1.778442,1.659618,1.451513,1.3455644, 1.2102273,1.0422029, 0.8757224,0.8080819, 0.5836541,
          2.081533,2.014705,2.060340,1.788118,1.810456,1.717652,1.5327706, 1.4615703,1.2591096, 0.8720247,0.8443209, 0.6811000,
          2.081224,2.097099,2.072250,1.967987,1.951054,1.730244,1.6712588, 1.5045396,1.4223475, 1.1495261,1.0588689, 0.8118687,
          2.034170,2.046209,2.093438,2.057180,1.938396,1.611027,1.7693809, 1.6873963,1.4960799, 1.2741177,1.1195940, 0.9910001,
          2.123243,2.058405,2.067950,2.042236,1.939096,1.905736,1.8001788, 1.6824635,1.7003569, 1.4152951,1.4306898, 1.0945199),
        nrow = 8, byrow = TRUE)


  C.CIt <-
  matrix(c(0.5739984, 0.6560228,0.6979546, 0.8081725,0.9242172, 1.107447, 1.294294, 1.454438, 1.712100, 2.073506, 2.324835, 2.739284,
          0.7120568, 0.7921105,0.8323754, 0.8928814,1.0502815, 1.249342, 1.372993, 1.702619, 1.982571, 2.524152, 3.189236, 3.693035,
          0.7893309, 0.8562170,0.8820219, 0.9573503,1.1805273, 1.264927, 1.478778, 1.804763, 2.174147, 2.622138, 3.216853, 4.178737,
          0.8533257, 0.8967342,0.9574057, 1.0060719,1.1617061, 1.330631, 1.556792, 1.825316, 2.345892, 2.767181, 3.402797, 4.671675,
          0.9149942, 0.9554293,0.9694660, 1.0420609,1.1301589, 1.335296, 1.491548, 1.609600, 1.950919, 2.861848, 3.181003, 5.168521,
          0.9627990, 0.9883437,1.0061804, 1.0660749,1.1612572, 1.336950, 1.457690, 1.702376, 1.896527, 2.524949, 3.115720, 4.354668,
          0.9997521, 1.0430513,1.0099895, 1.0653691,1.1538371, 1.423905, 1.380139, 1.564071, 1.722495, 2.313429, 2.847196, 4.338659,
          0.9956396, 1.0291604,1.0221148, 1.0642393,1.1621291, 1.220362, 1.408159, 1.491989, 1.605369, 2.089509, 2.522134, 3.625007),
        nrow = 8, byrow = TRUE)

  DF.CIt <-
  matrix(c(1.087390, 1.055937, 1.051993, 1.018971, 1.0120624,0.9475902, 0.8858178,0.8477451, 0.8039635,0.7573021, 0.7590984,0.7394837,
           1.019960, 1.005323, 1.008099, 1.029073, 0.9879237,0.8981221, 0.8893386,0.7697848, 0.7412096,0.6623974, 0.5936911,0.6129038,
           1.006291, 1.005329, 1.021087, 1.019935, 0.9125467,0.9192865, 0.8553957,0.7472854, 0.6915622,0.6579413, 0.5959874,0.5554533,
           1.014999, 1.032704, 1.016386, 1.024034, 0.9535650,0.9061006, 0.8205744,0.7526594, 0.6499711,0.6242470, 0.5703290,0.5123080,
           1.041496, 1.047536, 1.049122, 1.032562, 0.9983934,0.9101461, 0.8611576,0.8454547, 0.7750930,0.6045156, 0.6169245,0.4640618,
           1.051141, 1.059369, 1.042806, 1.027734, 0.9851017,0.9100837, 0.8812822,0.7934464, 0.7850725,0.6725143, 0.6111315,0.5408882,
           1.040141, 1.025528, 1.065488, 1.024694, 0.9875698,0.8377217, 0.9209586,0.8492124, 0.8507647,0.7155059, 0.6533312,0.5299317,
           1.059026, 1.034749, 1.055034, 1.029376, 0.9704596,0.9566411, 0.8882094,0.8838344, 0.8880207,0.7771873, 0.7253455,0.6134246),
         nrow = 8, byrow = TRUE)

  C.CId <-
  matrix(c(1.516477,1.826422,1.990346,2.132463,2.722614,3.291681,4.420750,5.584964,6.571250,9.860396,9.792822,14.559449,
          1.377650,1.506431,1.609982,1.837078,2.173961,2.817640,3.644735,4.589144,5.666289,8.335783,10.960068, 15.306485,
          1.308230,1.497846,1.415831,1.726081,2.224273,2.404352,3.035308,3.809549,5.431025,8.074415,11.021878, 14.406966,
          1.217408,1.337588,1.396507,1.611503,1.843107,2.384813,2.765119,3.458886,4.595699,6.474964,8.166269,14.566650,
          1.171666,1.231406,1.206510,1.560753,1.570748,1.775119,2.259675,2.516555,3.456203,6.409431,7.387063,11.905145,
          1.135017,1.135115,1.123657,1.286124,1.357646,1.709172,1.883729,2.310775,2.734692,4.249965,5.397095,9.338454,
          1.125844,1.134289,1.104515,1.147843,1.282499,1.822858,1.648062,1.866824,2.509362,3.580416,4.802758,6.899954,
          1.028051,1.075499,1.091011,1.132110,1.247353,1.320929,1.565746,1.857433,1.954477,2.984488,3.174651,5.823208),
        nrow = 8, byrow = TRUE)

  DF.CId <-
  matrix(c(0.5887816,0.5528131, 0.5458533,0.5776927, 0.5335293,0.5080249, 0.4358518,0.3775613, 0.3827662,0.3043313, 0.3496716,0.2829981,
           0.7390238,0.7413282, 0.7315318,0.6964397, 0.6610254,0.5637184, 0.4856496,0.4244074, 0.4016887,0.3316179, 0.2946872,0.2738511,
           0.8272066,0.7725987, 0.8490924,0.7427438, 0.6286789,0.6403420, 0.5661177,0.4868548, 0.3967950,0.3233705, 0.2792186,0.2730339,
           0.9212071,0.8852327, 0.8687061,0.7912667, 0.7415143,0.6260229, 0.5828540,0.5152537, 0.4403457,0.3736606, 0.3428074,0.2586485,
           0.9922391,0.9613626, 0.9943895,0.8033920, 0.8332030,0.7987648, 0.6829315,0.6511951, 0.5449091,0.3481682, 0.3495558,0.2809752,
           1.0187917,1.0426874, 1.0517253,0.9507281, 0.9429417,0.7998040, 0.7780296,0.6731247, 0.6335426,0.4812147, 0.4376909,0.3290643,
           1.0196227,1.0248727, 1.0589298,1.0402121, 0.9553533,0.7180629, 0.8481006,0.7894588, 0.6633031,0.5413594, 0.4674560,0.4081975,
           1.0976480,1.0588452, 1.0475996,1.0279192, 0.9635527,0.9514298, 0.8620989,0.7766074, 0.8088941,0.6197419, 0.6585104,0.4544030),
         nrow = 8, byrow = TRUE)


  rownames(C.CR) <- rownames(DF.CR) <- rownames(C.CIt) <- rownames(DF.CIt) <- rownames(C.CId) <- rownames(DF.CId) <-
    c("60", "120", "180", "300", "600", "1200", "2400","4800")
  colnames(C.CR) <- colnames(DF.CR) <- colnames(C.CIt) <- colnames(DF.CIt) <- colnames(C.CId) <- colnames(DF.CId) <-
    c("1/30", "1/20", "1/15", "1/10", "1/6", "1/4", "1/3", "2/5", "1/2", "3/5", "2/3", "3/4")

  # Calculate "c" and "df" estimates based on N.in and p.in
  N = c(60,120,180,300,600,1200,2400,4800)
```

```
p.trt = c(1/29,1/19,1/14,1/9,1/5,1/3,1/2,2/3,1,3/2,2,3)/(1 + c(1/29,1/19,1/14,1/9,1/5,1/3,1/2,2/3,1,3/2,2,3))

# use interpolation or extrapolation or asymptotic results to specify Satterthwaite constants
# (with corresponding message when appropriate)
N.in = length(f.data) + length(y); p.in = length(y)/N.in
if(!(N.in < 60 | N.in > 4800 | p.in < 1/30 | p.in > 3/4 | toupper(f.est) %in% c("TRUTH","F")))
{
  c.CR = bilinear(M=C.CR,N=N,p.trt=p.trt,N.in = N.in, p.in = p.in)
  df.CR = bilinear(M=DF.CR,N=N,p.trt=p.trt,N.in = N.in, p.in = p.in)

  c.CIt = bilinear(M=C.CIt,N=N,p.trt=p.trt,N.in = N.in, p.in = p.in)
  df.CIt = bilinear(M=DF.CIt,N=N,p.trt=p.trt,N.in = N.in, p.in = p.in)

  c.CId = bilinear(M=C.CId,N=N,p.trt=p.trt,N.in = N.in, p.in = p.in)
  df.CId = bilinear(M=DF.CId,N=N,p.trt=p.trt,N.in = N.in, p.in = p.in)
}else if(!(N.in > 4800 & p.in < 1/30 & length(y) > 160) & !(toupper(f.est) %in% c("TRUTH","F")))
{
  warning("Extrapolation of Sattertwaite approximation")
  c.CR = bilinear(M=C.CR,N=N,p.trt=p.trt,N.in = N.in, p.in = p.in)
  df.CR = bilinear(M=DF.CR,N=N,p.trt=p.trt,N.in = N.in, p.in = p.in)

  c.CIt = bilinear(M=C.CIt,N=N,p.trt=p.trt,N.in = N.in, p.in = p.in)
  df.CIt = bilinear(M=DF.CIt,N=N,p.trt=p.trt,N.in = N.in, p.in = p.in)

  c.CId = bilinear(M=C.CId,N=N,p.trt=p.trt,N.in = N.in, p.in = p.in)
  df.CId = bilinear(M=DF.CId,N=N,p.trt=p.trt,N.in = N.in, p.in = p.in)
}else{
  message("Asymptotic Results Used")
  c.CR <- c.CIt <- c.CId <- 1
  df.CR <- 2
  df.CIt <- df.CId <- 1
}

# Define an object to contain Likelihood information for the grid search
Star = matrix(NA, nrow = length(th), ncol = length(del))

# Evaluate the log-likelihood at the null (theta,delta) <- (0,0)
fn.null = psL(theta = 0, fhat.y = fhat.y, fhat.yd = fhat.y)

# Initialize the maximum value of log-likelihood
fn.max <- fn.null; est <- c(0,0)

# Initialize quantities for plotting profile likelihood
fni = NULL; fni.plot = matrix(nrow=length(th),ncol=0); theta.max = NULL; fni.max = NULL # ???

### Light Grid Search ###
i = 0 # counter
for(deli in del) # for each possible delta in the grid
{
  i = i+1 # counter

  # Evaluate the (pseudo) log-likelihood for across all theta (the specified deltai in this loop)
  fhat.yd <- eval.fhat.yd(f.data=f.data,y=y,f.est=f.est,
                          bw=bw,df=df,var.adj=var.adj,
                          deli=deli,obj=obj,
                          mu=mu,sigma=sigma,lambda=lambda,p=p,q=q)

  # Store the (pseudo) log-likelihood
  Star[,i] <- fni <- psL(theta = th, fhat.y = fhat.y, fhat.yd = fhat.yd)

  # The largest theta for this value of delta (conditional theta that maximizes pseudo-LL)
  theta.max <- c(theta.max,th[which.max(fni)])

  # The maximum value of pseudo-likelihood given deltai
  fni.max <- c(fni.max,fni[which.max(fni)])

  # The value of pseudo-log likelihood given deltai
  fni.plot = cbind(fni.plot,fni)

  # Keep track of max ps-log likelihood, and corresponding (theta,delta)
  if(max(fni) > fn.max){
    fn.max <- fni[which.max(fni)]
    est <- c(th[which.max(fni)],deli)
    row.max = which.max(fni); col.max = i
  }
}

# If the confidence region contains a point on the boundary (deltai = 6*Sx)
# and if ‘finite.area‘ is FALSE, then extend the light grid search to 12*Sx
more = FALSE
if( ( -2*(fni.max[length(fni.max)] - fn.max) < c.CR*qchisq(p=level,df=df.CR) ) & (finite.area == FALSE) )
{
  more = TRUE
  del <- c(del,del[length(del)] + seq(from = .1*sd(f.data), to = 6*sd(f.data), length.out = 60))
  Star = cbind(Star,matrix(NA, nrow = length(th), ncol = 60))

  # for each possible delta in the grid
  for(deli in del[length(del)-60] + seq(from = .1*sd(f.data), to = 6*sd(f.data), length.out = 60))
  {
    i = i+1 # counter

    fhat.yd <- eval.fhat.yd(f.data=f.data,y=y,f.est=f.est,
                            bw=bw,df=df,var.adj=var.adj,
                            deli=deli,obj=obj,
                            mu=mu,sigma=sigma,lambda=lambda,p=p,q=q)

    # The (pseudo) log-likelihood
    Star[,i] <- fni <- psL(theta = th, fhat.y = fhat.y, fhat.yd = fhat.yd)

    # The largest theta for this value of delta (conditional theta that maximizes pseudo-LL)
    theta.max <- c(theta.max,th[which.max(fni)])

    # The maximum value of pseudo-log likelihood given deltai
    fni.max <- c(fni.max,fni[which.max(fni)])

    # The value of pseudo-log likelihood given deltai
    fni.plot = cbind(fni.plot,fni)

    # Keep track of max ps-likelihood, and corresponding (theta,delta)
    if(max(fni) > fn.max){
      fn.max <- fni[which.max(fni)]
      est <- c(th[which.max(fni)],deli)
      row.max = which.max(fni)
      col.max = i
    }
  }
}

# Calculate the pseudo-likelihood test statistic
Star <- -2*(Star - fn.max)

### Dense Grid Search that encapsulates the confidence region found from the light grid search ###
ind.thrange = range(which(apply(Star,1,any.in,cl=level,c = c.CR, df=df.CR))) + c(-1,1)
ind.thrange[1] = max(c(ind.thrange[1],1))
ind.thrange[2] = min(c(ind.thrange[2],length(th)))

ind.delrange = range(which(apply(Star,2,any.in,cl=level,c = c.CR, df= df.CR))) + c(-1,1)
ind.delrange[1] = max(c(ind.delrange[1],1))
ind.delrange[2] = min(c(ind.delrange[2],length(del)))

theta.dense.grid = seq(th[ind.thrange[1]],th[ind.thrange[2]],length.out=100 + 100*more)
delta.dense.grid = seq(del[ind.delrange[1]],del[ind.delrange[2]],length.out=100 + 100*more)

### Add the true theta to the grid search if it is known/specified
if(!is.na(true.theta)){
  if(true.theta > min(theta.dense.grid) & true.theta < max(theta.dense.grid))
  {
    less.th.grid = sum(theta.dense.grid < true.theta)
    last.th.grid.ind <- (less.th.grid+1):length(theta.dense.grid)
    theta.dense.grid <- c(theta.dense.grid[0:less.th.grid],true.theta,theta.dense.grid[last.th.grid.ind])
  }else
  {
    true.th.range.del <- range(del[order(Star[less.th+1,])[1:(length(del)/10)]])
  }
}

### Add the true delta to the grid search if it is known/specified
if(!is.na(true.delta)){
  if(true.delta > min(delta.dense.grid) & true.delta < max(delta.dense.grid))
  {
    less.del.grid = sum(delta.dense.grid < true.delta)
    last.del.grid.ind <- (less.del.grid+1):length(delta.dense.grid)
    delta.dense.grid <- c(delta.dense.grid[0:less.del.grid],true.delta,delta.dense.grid[last.del.grid.ind])
  }
}


lt = length(theta.dense.grid); ld = length(delta.dense.grid)

# Keep track of theta*delta for all dense grid points
Delta.dense = rep(theta.dense.grid,ld)*rep(delta.dense.grid,each=lt)

# Define an object to contain likelihood information for the dense grid search
Star.Dense = matrix(NA, nrow = lt, ncol = ld)

# Implement dense grid search
j = 0 #counter
for(deli in delta.dense.grid)
{
  j = j+1 # counter
  fhat.yd <- eval.fhat.yd(f.data=f.data,y=y,f.est=f.est,
                          bw=bw,df=df,var.adj=var.adj,
                          deli=deli,obj=obj,
                          mu=mu,sigma=sigma,lambda=lambda,p=p,q=q)
  Star.Dense[,j] = psL(theta = theta.dense.grid, fhat.y = fhat.y, fhat.yd = fhat.yd)
}

# update maximum ps-log likelihood based on dense grid serach
old.fn.max <- fn.max
fn.max <- max(fn.max,max(Star.Dense))

# Calculate the pseudo-likelihood test statistic for the dense grid
Star.Dense <- -2*(Star.Dense - fn.max)

## Check Null Estimate
# (is it contained the confidence region/intervals
# [based on the distribution for the interior of the parameter space])
TS.null = -2*(fn.null - fn.max)
null.CR = any.in(TS.null,cl=level,df=df.CR,c=c.CR)
null.CIt = any.in(TS.null,cl=level,df=df.CIt,c=c.CIt)
null.CId = any.in(TS.null,cl=level,df=df.CId,c=c.CId)
null.CID = any.in(TS.null,cl=level,df=1,c=1)
null = c(TS.null,null.CR,null.CIt,null.CId,null.CID)

# If specified by user, include the true theta in the grid search
if(!is.na(true.theta))
{
  if(true.theta > min(theta.dense.grid) & true.theta < max(theta.dense.grid))
  {
    prof.TS.th.truth = min(Star.Dense[less.th.grid+1,])
  }else
  {
    prof.TS.th.truth = Inf
    j = 0 #counter
    for(deli in seq(true.th.range.del[1],true.th.range.del[2],length.out = 100 + 100*more))
    {
      j = j+1 # counter
      fhat.yd <- eval.fhat.yd(f.data=f.data,y=y,f.est=f.est,
                              bw=bw,df=df,var.adj=var.adj,
                              deli=deli,obj=obj,
                              mu=mu,sigma=sigma,lambda=lambda,p=p,q=q)

      prof.TS.th.truth = min(c(prof.TS.th.truth,-2*(psL(theta = true.theta, fhat.y = fhat.y, fhat.yd = fhat.yd) - fn.max)))
    }
  }
}

# If specified by user, include the true delta in the grid search
if(!is.na(true.delta))
{
  if(true.delta > min(delta.dense.grid) & true.delta < max(delta.dense.grid))
  {
    prof.TS.del.truth = min(Star.Dense[,less.del.grid+1])
  }else
  {
    fhat.yd <- eval.fhat.yd(f.data=f.data,y=y,f.est=f.est,
                            bw=bw,df=df,var.adj=var.adj,
                            deli=true.delta,obj=obj,
                            mu=mu,sigma=sigma,lambda=lambda,p=p,q=q)

    prof.TS.del.truth = min(-2*( psL(theta = seq(.01,1,length=1000), fhat.y = fhat.y, fhat.yd = fhat.yd) - fn.max ))
  }
}


# For computing a 50-50 mixture of chi-squares with df = 1 and df = 2
qmix.chisq = function(x,p,df=c(1,2)) abs(mean(pchisq(x,df=df)) - p)
boundary.cutoff = optim(par = qchisq(p=level,df=1),
                        fn = qmix.chisq, p = level, df=c(1,2),
                        method = "Brent", lower = qchisq(p=level,df=1),
                        upper = qchisq(p=level,df=2))$par


## Compute CR Area
N = length(Star.Dense)
X = sum(Star.Dense < c.CR*qchisq(p=level,df=df.CR))
if((N.in > 4800 & p.in < 1/30 & length(y) > 160) | (f.est %in% c("TRUTH","F")))
{
  th.bound.ind <- which(theta.dense.grid == 1)
  X = X + sum(Star.Dense[th.bound.ind,] < boundary.cutoff) -
    sum(Star.Dense[th.bound.ind,] < c.CR*qchisq(p=level,df=df.CR))
}
if(null.CR & !finite.area){
  Area=Inf}else{
    Area=.5*(X/N + X/(N-2*lt-2*ld))*diff(range(theta.dense.grid))*diff(range(delta.dense.grid))
  }

## Compute Intervals and note whether or not CSet is naturally an interval
TF.th = apply(Star.Dense,1,any.in,cl=level,df=df.CIt,c=c.CIt)
TF.del = apply(Star.Dense,2,any.in,cl=level,df=df.CId,c=c.CId)
```

```
if(sum(TF.th) > 0){ind.thmarg = range(which(TF.th))}else{ind.thmarg = NA}
if(sum(TF.del) > 0){ind.delmarg = range(which(TF.del))}else{ind.delmarg = NA}

# If 0 in CSet(theta) => CSet(theta) = (0,1)
if(null.CIt){CI.th = c(0,1)}else{
  if(!any(is.na(ind.thmarg))){CI.th = theta.dense.grid[ind.thmarg]}else{CI.th = c(NA,NA)}
}

# If 0 in CSet(delta) => CSet(delta) = (0,Inf)
# If finite.area == FALSE then adjust interval for delta according to the above result
if(null.CId & !finite.area){CI.del <- c(0,Inf)}else{
  if(!any(is.na(ind.delmarg))){CI.del = delta.dense.grid[ind.delmarg]}else{CI.del = c(NA,NA)}
}

# If 0 in CSet(Delta) => CSet(Delta) = (0,Inf)
# If finite.area == FALSE then adjust interval for Delta according to the above result
if(null.CID & !finite.area){CI.Del <- c(0,Inf)}else{
  CI.Del = range(Delta.dense[which(Star.Dense < qchisq(p=level,df=1))]) #!# what do I do here? leave as asymptotic, I guess.
}

# Store the endpoints for the intervals
if(!any(is.na(ind.thmarg))){
  th.int = all(TF.th[seq(ind.thmarg[1],ind.thmarg[2])])
}else{
    th.int = NA
    }
if(!any(is.na(ind.delmarg))){
  del.int = all(TF.del[seq(ind.delmarg[1],ind.delmarg[2])])
}else{
    del.int = NA
    }
ints = c(th.int = th.int, del.int = del.int)


### Determine whether CR captured truth
if(!any(is.na(c(true.theta,true.delta)))){
  fhat.yd <- eval.fhat.yd(f.data=f.data,y=y,f.est=f.est,
                          bw=bw,df=df,var.adj=var.adj,
                          deli=true.delta,obj=obj,
                          mu=mu,sigma=sigma,lambda=lambda,p=p,q=q)

  TS.th.del.truth = -2*(psL(theta=true.theta,fhat.y=fhat.y,fhat.yd=fhat.yd)-fn.max)
  if(true.theta == 1 & ( (N.in > 4800 & p.in < 1/30 & length(y) > 160) | f.est %in% c("TRUTH","F") ))
  {
    CR.cap = TS.th.del.truth < boundary.cutoff
  } else{
    CR.cap = TS.th.del.truth < c.CR*qchisq(p=level,df=df.CR)
  }
}else{TS.th.del.truth = NA; CR.cap = NA}



## Plot the results
if(plot)
{
  par(mar = c(4,6,3,1), cex.lab = 2, cex.main = 2, cex.axis = 2)

  CI.thresh = fn.max + (-c.CId*qchisq(p=level,df=df.CId)/2)
  CR.thresh = fn.max + (-c.CR*qchisq(p=level,df=df.CR)/2)
  plot(x = del, y = fni.max, type = "l", lwd = 3, pch = 16,
       ylim = range(c(CR.thresh,CI.thresh,fni.max)),
       xlab = expression(delta), ylab = "Max Log-Lik",
       main = bquote("Maximum Log-Lik given" ~ delta)
  )
  abline(h = CI.thresh, lwd = 2, lty = 2)
  segments(x0 = CI.del[1], y0 = -10^200, y1 = CI.thresh, lwd = 2, lty = 2, col = "darkgreen")
  segments(x0 = CI.del[2], y0 = -10^200, y1 = CI.thresh, lwd = 2, lty = 2, col = "darkgreen")
  abline(h = CR.thresh)
  points(x = 0, y = fn.null, pch = 19)

  plot(x = del, y = theta.max, type = "l", lwd = 3, pch = 16,
       ylim = c(0,1),
       xlab = expression(delta), ylab = bquote(hat(theta)(delta)[MLE]),
       main = bquote(theta ~ "that maximizes PsL given" ~ delta))
  points(x = 0, y = 0, pch = 19)

  plot(x = c(0,max(del)), y = c(0,1.05), col = "white",
       xlab = bquote(delta), ylab = "",
       main = bquote("Confidence Bounds for("*theta*","*delta*")")
  )
  mtext(text = bquote(theta), side = 2, cex = 2.5, las = 2, line = 3)

  for(i in 1:length(theta.dense.grid))
  {
    if(i < length(theta.dense.grid)){
      for(j in 1:length(delta.dense.grid))
      {
          points(x = delta.dense.grid[j], y = theta.dense.grid[i],
                 col = 1 + 2*(Star.Dense[i,j]< c.CR*qchisq(level,df=df.CR)), pch = 19, cex = .5)
        }
      }else{
        for(j in 1:length(delta.dense.grid))
        {
          if(N.in > 4800 & p.in < 1/30 & length(y) > 160)
          {
            points(x = delta.dense.grid[j], y = theta.dense.grid[i],
                   col = 1 + 2*(Star.Dense[i,j]< boundary.cutoff), pch = 19, cex = .5)
          }else{
            points(x = delta.dense.grid[j], y = theta.dense.grid[i],
                   col = 1 + 2*(Star.Dense[i,j]< c.CR*qchisq(level,df=df.CR)), pch = 19, cex = .5)
          }
        }
      }
    }

    lines(x = CI.Del[1]/seq(.001,1,.001), y = seq(.001,1,.001), lwd = 2, col = "darkgreen")
    lines(x = CI.Del[2]/seq(.001,1,.001), y = seq(.001,1,.001), lwd = 2, col = "darkgreen")
    segments(x0 = -.15, y0 = CI.th[1], y1 = CI.th[2], lwd = 5, col = "darkgreen")
    segments(y0 = CI.th[1], x0 = -.15, x1 = CI.del[2], lty = 2, lwd = 3, col = "darkgreen")
    if(!null.CIt){text(x = 0, y = CI.th[1], labels = round(CI.th[1],3), pos = 1, font = 2, col = "darkgreen")}
    segments(y0 = CI.th[2], x0 = -.15, x1 = CI.del[2], lty = 2, lwd = 3, col = "darkgreen")
    text(x = 0, y = CI.th[2], labels = round(CI.th[2],3), pos = 3, font = 2, col = "darkgreen")
    segments(x0 = CI.del[1], x1 = CI.del[2], y0 = -.03, lwd = 5, col = "darkgreen")
    segments(x0  = CI.del[1], y0 = -.03, y1 = CI.th[2], lty = 2, lwd = 3, col = "darkgreen")
    if(!null.CId){text(x = CI.del[1], y = -.02, labels = round(CI.del[1],2), pos = 2, font = 2, col = "darkgreen")}
    segments(x0 = CI.del[2], y0 = -.03, y1 = CI.th[2], lty = 2, lwd = 3, col = "darkgreen")
    text(x = CI.del[2], y = -.02, labels = round(CI.del[2],2), pos = 4, font = 2, col = "darkgreen")
    points(x = true.delta, y = true.theta, col = "red", lwd = 2, cex = 1.5, pch = 10)
    segments(x0 = true.delta, y0 = -.15, y1 = true.theta, col = "red", lwd = 2, lty = 2)
    segments(x0 = -.15, x1 = true.delta, y0 = true.theta, col = "red", lwd = 2, lty = 2)
    text(x = true.delta, y = true.theta, pos = 4,
         labels = bquote(bold("(" * .(round(true.theta,2)) * "," * .(round(true.delta,2)) * ")")), col = "red")

    if(null.CR){points(x = 0, y = 0, col = 3, pch = 15, cex = 1.5)}
    points(x = est[2], y = est[1], col = "blue", pch = 19)
  }
  if(!any(is.na(c(true.theta,true.delta)))){
    TS.truth = c(TS.CR.Truth = TS.th.del.truth,
                 TS.prof.th.truth = prof.TS.th.truth,
                 TS.prof.del.truth = prof.TS.del.truth)}
  else{
    TS.truth = rep(NA,3)
    }
  ans = list(c(theta.hat = est[1], delta.hat = est[2], Delta.hat = prod(est)),
             CI.th,CI.del,CI.Del,
             ints,null.CR,Area,CR.cap,
             TS.truth,TS.null)
  names(ans) = c("Est",
                 paste0(100*level,c("CI_theta", "CI_delta","CI_Delta",
                                    "Ints","Null_CR","Area","CR_Cap")),
                 "TS_Truth","TS_Null")
  return(ans)
}

# Generate m = 100 observations from N(0,1) for the control group and n = 100 observations from .3N(0,1) + .7N(2,1) for the trt group.
m = 100
n = 100
true.theta = .7
true.delta = 2
x = rnorm(m)
  z = sample(c(0,1), size = n, replace = TRUE, prob = c(1-true.theta,true.theta))
y = rnorm(n) + true.delta*z

# Find the Normal Maximum Likelihood
psl.inf(f.data = x, y = y, plot = TRUE, level = .95, true.theta = .7, true.delta = 2)
```