IMPROVED SAMPLE SIZE RE-ESTIMATION IN ADAPTIVE CLINICAL TRIALS WITHOUT UNBLINDING

by

Chen Teel

B.S. Business Administration, Nanjing University of Science and Technology, Nanjing 2003

M.A. Statistics, University of Pittsburgh, Pittsburgh, PA 2007

Submitted to the Graduate Faculty of

the Arts & Sciences in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2011

UNIVERSITY OF PITTSBURGH

ARTS & SCIENCES

This dissertation was presented

by

Chen Teel

It was defended on

December 14, 2010

and approved by

Dr. Allan Sampson

Dr. Yu Cheng

Dr. Leon Gleser

Dr. Abdus Wahed

Dissertation Director: Dr. Allan Sampson

Allan R. Sampson, PhD

IMPROVED SAMPLE SIZE RE-ESTIMATION IN ADAPTIVE CLINICAL TRIALS WITHOUT UNBLINDING

Chen Teel, PhD

University of Pittsburgh, 2011

Sample size calculations in clinical trials depend on good estimates of the standard deviation. Due to the uncertainty in the planning phase, adaptive sample size designs have been used to re-estimate the standard deviation based on interim data and adjust the sample size as necessary. Our research concentrates on carrying out the sample size re-estimation without obtaining the treatment identities.

Gould and Shih[15] treated the interim data as coming from a mixture of two normal distributions with a common standard deviation. To adjust the sample size, they used the EM algorithm to obtain the MLE of the standard deviation while preserving the blind. However, their approach has been criticized in the literature and our simulation studies show that Gould and Shih's[15] EM algorithm sometimes obtains incorrect boundary modes as estimates of the standard deviation. We establish a new procedure to re-estimate the sample size without breaking the blind but using additional information concerning the randomization structure at the interim. We enhance their EM procedure by utilizing the conditional Bernoulli model to incorporate the available information that equal numbers of subjects are observed at the interim stage. Properties of the enhanced EM estimator are investigated in detail.

Furthermore, we use the full information of the blocked randomization schedule in the enhanced EM algorithm that the numbers of subjects are equal across treatment groups within each randomization block. With increased information that occurs with an increasing

number of blocks, the accuracy of the standard deviation estimation improves and there is small bias when the block size is small. Moreover, for the case of two treatment groups, the preservation of the actual type I error rate when using the standard t-test at the end of the trial is verified through a simulation study. The actual power and the expected sample size are analytically computed and simulated. The enhanced procedure with large numbers of blocks is shown to adaptively maintain the power at a minimal sample size cost. Results are extended to handle multi-center trials.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

I am indebted to all the faculty and graduate students for all of their support and guidance throughout my years studying at Pitt. However, there are several individuals who I want to personally acknowledge for their contributions to my growth. I want to thank my closest advisors, who without their contributions, this dissertation would have been immensely more difficult. First, I want to thank Dr. Sampson for his research support, personal guidance, and showing an enduring confidence in my abilities. He has inspired me to improve my education and my command of the English language since my first year at Pitt and has been a wonderful mentor for my graduate student career. I would like to thank my co-advisor Dr. Park for his patience, research support, and his prompt late night responses to my emails.

I want to thank all of my committee members. Thank you to Dr. Gleser for his accessibility and for being an outstanding educator. Dr. Cheng for her willingness to support me on short notice after Dr. Park departed. Thanks Dr. Wahed for his constructive suggestions regarding my dissertation. With the help of my committee, my dissertation has improved.

I also want to recognize other faculty and staff members instrumental in my personal development. I appreciate Dr. Iyengar for proving guidance and his commitment to answering my questions. Dr. Krafty has been willing to share his knowledge of career advice. Mary and Kim for taking care of me for the last six years.

Lastly, I want to thank my friends and family. I want to thank Jo for being such a trustworthy and loyal friend. I want to thank my husband, Jon, for persistently cheering me up during the critical stages of my graduate program. Also, I would like to thank my parents and my parents-in-law for their love and support.

1.0 INTRODUCTION

1.1 BACKGROUND

1.1.1 Adaptive designs in clinical trials

Adaptive clinical trial designs allow modifying the design specifications or statistical procedures of an on-going trial based on the analysis of the interim data. Possible adaptations used in clinical trials include[9][6]: sample size re-estimation, dropping or adding treatment arms, adaptive dose finding, and adaptive hypothesis design (e.g., switching from a superiority hypothesis to a non-inferiority hypothesis, changing primary endpoints). Compared to traditional clinical trials, in which data are not analyzed until the end of the study, adaptive designs based on interim data analysis can be more flexible and efficient for identifying clinical benefits[9]. Adaptive designs can also increase the probability of success, and potentially reduce costs and resources for drug development. These advantages come at a price. Adaptive trials are challenging to implement, and methods used for adaptive trials must be carefully chosen to protect type I error and also to maintain the trial validity and integrity.

In recent years, sample size recalculations based on interim data have become increasingly popular. In designing a clinical trial, determination of the sample size is a key step. It is important to have a sufficient number of subjects in order to achieve the desired power for detecting a clinically meaningful difference if such a difference truly exists. On the other hand, if fewer subjects than planned can detect this difference, it is desirable to reduce the number of subjects in the trial, particularly for trials in which subjects might be exposed to an inferior or possibly toxic treatment. To deal with these concerns, sample size adaptive designs can be used to adjust the sample size at the interim stage to avoid an underpowered

or overpowered trial.

There are basically two types of adaptive designs for sample size recalculations, unblinded and blinded sample size recalculations. Unblinded sample size recalculation methods break the blind of the treatment identities at the interim stage and use this information to adjust the sample size of the ongoing trial. Blinded sample size recalculation methods adjust the sample size without breaking the blind at the interim stage. In this dissertation, we focus on sample size re-estimation in adaptive designs when the blinding is maintained.

1.1.2 Sample size re-estimation for normal data

Our specific focus is on sample size re-estimation when the primary trial outcome measure can be viewed as following a normal distribution, or at least well approximated by a normal distribution. However, our proposed method in this dissertation can be used under any exponential family assumption. Suppose we plan a clinical trial aimed to compare an experimental treatment with a control treatment where the primary endpoints are normally distributed. For normal data, sample size is determined by type I error, the power at the treatment effect to be detected, and the standard deviation of the primary outcome variable[11].

The value of the population standard deviation of the primary endpoint is generally unknown in the planning stage. In the planning stage, an estimate of the standard deviation is typically based on previous or similar trials. This estimate can be unreliable for a variety of reasons. Study populations can differ, study conduct can vary, and primary endpoints can be measured differently. Moreover, even in identical settings, studies can differ in their variability for unknown reasons. Underestimating the standard deviation in the design phase causes the trial to be underpowered. Overestimating the standard deviation in the design phase is wasteful of time and money on the trial, as well as possibly being ethically problematic. Therefore, It is desirable to get a more accurate estimate of standard deviation using interim data.

Treatment effect is defined to be the difference in the mean of a primary endpoint between the control and treatment groups. Typically in the planning stage the treatment effect is chosen to be a clinically meaningful difference of interest. It can also represent a difference

that would make further development of the experimental treatment cost feasible. Re-estimation of the treatment effect based on interim data can be used to adjust the sample size, but requires unblinding the data and thus is sometimes controversal[27, 25, 26]. Unblinding can introduce potential bias and add more complexity to the studies. In our research, we focus on sample size re-estimation based on the nuisance parameter, σ, rather than on the treatment effect, because as we indicated we want to maintain the blind at the interim stage.

Gould[14] proposed a procedure of sample size re-estimation for binomial trials which does not break the blind. An initial sample size is calculated based on type I error and the power for the assumed treatment effect and the anticipated overall event rate for the binary primary endpoint. Gould's method adjusts sample size based on the estimated overall event rate at the interim stage, which is available without breaking the blind. The standard chi-square test is used to test the null hypothesis of equal proportions at the end of the trial. His simulation studies showed that there is not substantial type I error rate inflation by using the chi-square test as if no adjustment to the sample size occurred.

1.2 UNBLINDED SAMPLE SIZE RE-ESTIMATION

In this subsection, we briefly review several methods that have been used to re-estimate σ using unblinded data at the interim stage.

1.2.1 Stein's method

Stein proposed a two-stage procedure in 1945[33] that can be used in two sample clinical trials as follows. First, calculate the planned sample size based on an initial guess of the standard deviation and use a sub-sample of the assumed number of subjects as first stage sample. After subjects in the first stage finish the trial, we calculate the within-group standard deviation and use it determine the new final sample size. More subjects are recruited until the new sample size is reached. At the end of the trial, compute a standard t-statistic using the first stage's within group standard deviation in the denominator.

Since the estimate of standard deviation in the Stein's t-statistic is only based on the first stage's data, it can be shown that this t-statistic follows a t-distribution. Also it is shown that the desired power is guaranteed[24]. However, Stein's procedure has not been frequently used in clinical trials. Because Stein's procedure only uses the standard deviation's estimate from the first stage in the final test statistic, it may be a bad estimator when first stage's sample size is small or the standard deviation of the primary endpoint changes over the course of the trial[28].

1.2.2 The naive t-test

Wittes and Brittain modified the Stein procedure and presented the idea of an internal pilot study[36]. In this approach, they treat the first fraction of the planned sample as an 'internal pilot' and recalculate sample size using an estimate of σ from the internal pilot data. The study then continues with the recalculated sample size as the target for the overall sample size. The data are analyzed at the end of the trial as if they had been collected in a fixed sample study.

Wittes and Brittain's procedure is similar to Stein's, in that the sample size adjustment is based on the within-group standard deviation of the first stage's sample. The difference is that they use the within-group standard deviation of the entire sample data in the denominator of the standard t-statistic at the end of the trial. This test statistic uses the t-distribution as its reference distribution. In such a setting with no adjustment for sample size re-estimation, this approach is called the naive method. Since the total sample size is adaptive, the t-statistic under the null hypothesis does not actually follow a t-distribution. Its advantage, however, is that it uses all the data to estimate the standard deviation.

Simulation studies showed that this naive method assures the desired power is reached, but it may inflate the type I error rate, especially when the first stage sample size is small[36][2][37]. Kieser and Friede[16] analytically computed the upper bound for the actual type I error rate and proposed an adjustment for the critical value for the naive t-test. Miller[22] adjusted the variance estimator in the test statistic with an additive correction. He showed through simulation that the actual type I error is very close to the nominal level

of α.

1.3 BLINDED SAMPLE SIZE RE-ESTIMATION

The aforementioned methods calculate the within-group standard deviation from the separate standard deviations in the treatment and control groups which requires breaking the blind. For the blinded design, the blind is not broken at the interim stage and only broken at the end of the trial. Maintaining the blind in the sample size recalculation has clear operational advantages. Unblinding the trial for the interim analysis usually requires an independent external group, such as Independent Data Monitoring Committee, to conduct the sample size re-estimation. This may introduce unnecessary complexity and prolong the study of the trial. For a blinded design, it can be conducted by in-house personnel. Blinding the treatment identity also helps preserve the integrity of the trial[12]. By unblinding the treatment assignment, an investigator who infers the apparent treatment effect from the interim data might have a tendency to treat remaining subjects with some bias. In facts, an investigator would potentially be able to estimate the interim treatment effect if they can infer the first stage's within-group variance and obtain the pooled variance[24].

1.3.1 Pooled one sample standard deviation procedure

Gould and Shih[15] used a simple adjustment procedure to re-estimate sample size without unblinding the data. First, the pooled variance is calculated from the blinded internal pilot data, treating both treatments' data as coming from a single population. Then the adjusted variance is based on the one sample pooled variance and the hypothesized treatment effect. The adjusted standard deviation is an unbiased estimator of σ if the hypothesized treatment effect holds. The potential problem is that the adjusted one-sample standard deviation depends on the observed treatment effect. If the true treatment effect is bigger than the assumed one, the calculated sample size could be unnecessarily large[22]. An alternative is to use the one sample pooled standard deviation without adjustment. It has been argued

that the overestimation of the one sample pooled standard deviation is not large in typical clinical trials[24]. Kieser and Friede[17] showed through analytical computations that the type I error rate in the usual t-test is not inflated if sample size is recalculated with the adjusted or unadjusted one-sample standard deviation of the pooled data, and also that the desired power is achieved.

1.3.2 EM algorithm

Gould and Shih[15] proposed another method, by using the EM algorithm, to estimate the standard deviation assuming the data follow a mixture of normal distributions. This does not require breaking the treatment blind. They showed that the EM algorithm reasonably estimates the standard deviation no matter the assumed value of treatment difference. Details concerning their EM procedure are provided in Chapter 2.

Gould and Shih[15]'s implementation of the EM algorithm for the mixture of normal distributions has raised some issues in subsequent literature. Friede and Kieser[12] indicated that Gound and Shih[15]'s procedure has critical deficiencies. First, they showed through simulation study that the estimate of within-group standard deviation depends on the initial value of standardized treatment effects when implementing the EM algorithm. They also showed that the EM algorithm Gould and Shih[15] used converges very slowly. Thus, an inadequate stopping criteria makes the algorithm stop at incorrect values before the estimator stabilizes. Waksman[35] examined Gould and Shih[15]'s published computer program and argued that Gould and Shih[15]'s EM algorithm is independent of the initial values as long as the stopping criteria are strict enough. Waksman explained that Gould and Shih[15] altered the estimate of standard deviation in the M-step by subtracting 1 from the total sample size in the denominator. When the alteration is removed from the program, the results are changed significantly. He used simulation to show that the EM estimate of the standard deviation is independent of the initial values of the EM algorithm. Waksman also suggested that using a sufficiently strict stopping criteria leads the EM algorithm to obtain the MLE.

1.4 SETTINGS AND NOTATION

Consider a clinical trial where we want to compare two treatments, a control group and a treatment group. For simplicity, we assume that equal numbers, $N/2$, of subjects are assigned to each group. The primary endpoint is assumed normally distributed with mean μ_c in the control group, mean μ_t in the treatment group and with a common standard deviation σ. Define the true treatment difference as $\delta = \mu_t - \mu_c$. The goal of the trial is to compare these two groups, i.e., $H_0 : \mu_c = \mu_t$ versus $H_1 : \mu_c \neq \mu_t$ at the end of the trial.

At the beginning of the trial, the planned sample size N can be obtained as

$$\frac{4\tilde{\sigma}^2(z_{\alpha/2} + z_\beta)^2}{\Delta^2}, \tag{1.1}$$

where $\tilde{\sigma}$ is an initial guess of the standard deviation, α is the type I error rate, $1 - \beta$ is the desired power, $z_{\alpha/2}$ and z_β are the upper $\alpha/2$ and β quantiles of a standard normal distribution, and Δ is the assumed treatment effect for which the power is desired. We typically obtain $\tilde{\sigma}$ based on experience or from a previous study with the same endpoint. When the data of the first N_1 subjects out of N are available, we re-estimate σ from these N_1 observations without knowing the treatment identities, that is, without breaking the blind. For example N_1 can be half the initially planned sample size, i.e., $N_1 = N/2$. The new estimator $\hat{\sigma}$ is obtained from the blinded data and used to determine a new sample size N' that is given by

$$\frac{4\hat{\sigma}^2(z_{\alpha/2} + z_\beta)^2}{\Delta^2}. \tag{1.2}$$

Here, N is the originally planned sample size and N' is the recalculated sample size based on (1.2) using the first N_1 observations. Taking into account that we already have N_1 subjects at the interim stage, the new recalculated sample size N' can be adjusted following different sample size capping rules. For example, Birkett and Day [2] proposed the unrestricted rule, where N' is at least N_1, that is, the adjusted final sample size for the entire study is $N_{\text{adj}} = \max(N_1, N')$. Thus in the second stage, a further $N_2 = \max(N_1, N') - N_1$ subjects are recruited. Further rules for the adjusted final sample size are given in Chapter 4.

1.5 OVERVIEW

In our research, we aim to re-estimate the sample size by utilizing the blinded data at the interim of the adaptive clinical trials. We estimate the standard deviation by extending the ideas of Gould and Shih[15] and using further details motivated by the practical setting of clinical trials.

We enhance Gould and Shih's[15] EM procedure by utilizing the information of the blocked randomization schedule observed in clinical trials. The computational details used to modify the EM algorithm when having this additional information are discussed in Chapter 2.

In Chapter 3, we further explore the effects of initial values and the convergence properties for our enhanced EM algorithm in comparison to Gould and Shih's[15]'s EM algorithm. Simulation studies are conducted to compare the estimates from both EM algorithms. In addition, we compare the estimates from the enhanced EM algorithm with different block sizes.

In Chapter 4, the actual type I error rates are simulated under different scenarios by using the standard t-test at the end of the adaptive studies and compared across EM procedures. The actual power and the expected sample size are simulated in a similar way and also computed using an analytical method. We show the benefits of using our adaptive sample size procedure with large numbers of blocks.

EM procedures for single center trials are extended to multi-center trials in Chapter 5. A preliminary simulation study is conducted for a two-center trial setting.

Finally, Chapter 6 presents our conclusions and lays down the foundation for future work.

2.0 ENHANCED EM ALGORITHM ESTIMATION

2.1 MOTIVATION

It is known that for the two-sample t-test, the standardized treatment difference in the two means (i.e., effect size) affects power calculation. For a given sample size, the larger the effect size, the larger the power. For fixed total sample size, the standard deviation of the difference in sample means is minimized when the two treatments have equal sample sizes. Thus most clinical trials generally allocate patients to equal-sized groups to get the best power when comparing two treatments.

The simplest randomization for two treatment groups is complete randomization with $p = 0.5$. Simple randomization, however, is not typically used in clinical trials because it can lead a substantial imbalance in the number of subjects assigned to each treatment group. The imbalance would reduce the test's ability to detect the true difference between two treatments. In clinical trials, we minimally want to keep equal numbers of subjects in the two treatment groups at the end of the trial, where each subject has the same probability to be assigned to either the control group or the treatment group.

To improve complete randomization, block randomization is often applied in clinical trials[30]. Within each block, equal numbers of subjects are randomly allocated to the control group and the treatment group. The block size must be an even number, and usually is not given in the clinical trial protocol. In addition to keeping a group balance at the end of the trial, block randomization also periodically keeps the balance of patients between two treatments. This is very important because time confounding can be guarded against especially for a clinical trial which takes a long time to complete. During the trial, medical equipment, concomitant medications and staff can change. It is also possible that

the disease severity of patients entering the trial earlier is significantly different from that of patients entering the trial towards the end. The balancing of numbers of patients makes the two treatments intermittently more comparable over time. In the case of multiple-center trials, not only are centers blocked but within centers blocks are also used to avoid an imbalance that would happen within a center. For instance, if the trial is ended before one center completes enrollment, we can still guarantee that there are equal numbers of patients assigned to each treatment within the center. In general to protect against possibly guessing the next patient's allocation when small block sizes are used, block sizes are usually chosen randomly, that is, we may use a combination of different block sizes, e.g., $2, 4, 6$ and 8 during the randomization.

As noted, Gould and Shih[15] used the EM algorithm to estimate the within-group standard deviation without unblinding the data at the interim stage. They planned N patients in total with $N/2$ patients assigned to the control group and $N/2$ patients assigned to the treatment group. For the N_1 patients at the interim stage, they keep the treatment identities blinded, so that the treatment indicators z_i's follow independent Bernoulli distributions with probability 0.5 for $i = 1, ..., N_1$. Clearly this does not guarantee an equal number of patients in each group. Gould and Shih[15] chose not to use any block information concerning the randomization at the interim stage. At the end of the trial, they use the standard t-test as if the total sample size is fixed. Under this basis, their simulation showed that the actual type I error rate is not inflated.

We propose a new procedure using the EM algorithm which untilizes the additional information that equal numbers of subjects are assigned to each treatment at the interim stage. The new proposed procedure is called the *enhanced* EM algorithm. To avoid confusion, we call the EM algorithm used by Gould and Shih the *conventional* EM algorithm in our dissertation. In this new procedure, given the condition that the sum of z_i's for the N_1 subjects at the interim stage equals to $N_1/2$, z_i does not follow an independent Bernoulli distribution any longer. The distribution of $z_1, ..., z_{N_1}$ given the sum, $\sum_{i=1}^{N_1} z_i$, follows the so-called conditional Bernoulli distribution, which we discuss in detail in Section 2.3. We show that we can enter the additional information of balanced treatment allocation into the EM algorithm. The critical remaining issue which we discuss in our research is whether the

type I error of the standard t-test used at the end of the trial is inflated or not.

Suppose that we additionally knew we had two equal sized blocks for the interim data each with size $N_1/2$. Thus, in addition to knowing equal numbers of subjects from the two treatments at the interim stage, we can gain a little more information by knowing that within each half of the interim data, the numbers of subjects from two treatments are also equal. In this case, the treatment identities given their sum within each half of the interim data is also conditional Bernoulli distributed. We can use the information about block sizes further in the enhanced EM algorithm. Conceptually, as we continue obtaining more information about blocks, we know more balancing points at the interim stage. In this case, we show that we get better estimates of the within-group standard deviation by using more available information about the block sizes.

Operationally, there may be a concern that as we reveal more information about block sizes, the interim data may not be considered fully blinded. We do not address this potential operational issue, other than to show type I error is preserved. Our ultimate goal is to assess when using the full randomization block information whether or not the type I error rate is inflated when we use the standard two-sample t-test at the end of the trial.

2.2 THE CONVENTIONAL EM ALGORITHM

2.2.1 The EM algorithm

In the usual approach to maximum likelihood estimation, we set the first derivatives of a log-likelihood function equal to zero, and find the maximum likelihood estimates by solving for the unknown parameters in the equation. In the case where the underlying density is a mixture of two distributions, it is difficult to find such analytical solutions for maximum likelihood estimates. The EM algorithm [10] is an iterative algorithm developed to find maximum likelihood estimates (MLEs) from the perspective of incomplete data and can be used to obtain MLEs for mixture distributions.

Incomplete data arise from data missing by error or data involving some latent variables

that are conceptually missing. The notation Y_{obs} denotes the incomplete data, i.e., the observed data; Y_{mis} denotes the missing data; and $Y_{\text{com}} = (Y_{\text{obs}}, Y_{\text{mis}})$ denotes the complete data. The complete data are assumed to have a joint density function $f(Y_{\text{com}}|\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is a vector of parameters. The EM algorithm greatly reduces the complexity of the maximum likelihood estimation by taking advantage of the complete data[20].

The EM algorithm starts with an initial guess of the parameters, $\boldsymbol{\theta}^{(0)}$, and then it iterates between two steps, the expectation step (E-step) and the maximization step (M-step). The E-step calculates the conditional expectation of the complete-data log-likelihood function given the observed data Y_{obs} and the current parameter estimates. Specifically, the E-step computes

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = E\left[\ell(\boldsymbol{\theta}|Y_{\text{com}})|Y_{\text{obs}}, \boldsymbol{\theta}^{(t)}\right],$$

where $\boldsymbol{\theta}^{(t)}$ denotes the estimate of $\boldsymbol{\theta}$ at the t^{th} iteration.

In the M-step, we maximize the expectation computed in the E-step with respect to $\boldsymbol{\theta}$, and update the estimate of $\boldsymbol{\theta}$, i.e.,

$$\boldsymbol{\theta}^{(t+1)} = \arg\max_{\theta} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}).$$

The E-step and M-step are alternated repeatedly until certain convergence criteria are met. The purpose of the EM algorithm is to maximize the observed-data log-likelihood function $\log \ell(\boldsymbol{\theta}|Y_{\text{obs}})$. It can be shown from Jensen's inequality that the observed-data log-likelihood function evaluated at $\boldsymbol{\theta}^{(t)}$ is monotonically increasing on every iteration of the EM algorithm[10]. The monotone convergence property of the EM algorithm guarantees finding a local maximum of the observed-data log-likelihood.

2.2.2 Gould-Shih's EM procedure without unblinding

In the blinded design, treatment identities are unknown when the interim analysis is to be done after the first N_1 of the planned sample size are available. Each observed primary endpoint y_i, for $i = 1, ..., N_1$, is either from one treatment group or the other, so that its treatment identity is missing. Our goal is to recalculate the sample size based only on

the blinded data estimators. We obtain the maximum likelihood estimate of the nuisance parameter σ at the interim stage, and use this estimate to adjust the second stage sample size in a study comparing treatment and control. The observed data $\mathbf{y} = (y_1, y_2, ..., y_{N_1})$ are treated as a mixture of two normal distributions, with parameters μ_1, μ_2, and σ, denoted collectively by $\boldsymbol{\theta}$. The density function of y_i is given by

$$f(y_i|\boldsymbol{\theta}) = \frac{1}{2}f(y_i|\mu_1, \sigma) + \frac{1}{2}f(y_i|\mu_2, \sigma)\,. \tag{2.1}$$

The observed-data likelihood function is given by the product of the sums of two normal distributions, i.e.,

$$L(\boldsymbol{\theta}|\mathbf{y}) = \prod_{i=1}^{N_1}\left\{\frac{1}{2}f(y_i|\mu_1, \sigma) + \frac{1}{2}f(y_i|\mu_2, \sigma)\right\}, \tag{2.2}$$

where we use the fact that subjects are equally randomized to two treatments. Gould and Shih[15] treated maximizing (2.2) as an incomplete-data problem and used the EM algorithm for maximum likelihood estimation in a mixture of two normal distributions. The observed data are the primary endpoints, i.e., $Y_{\text{obs}} = \{y_i\}_{i=1}^{N_1}$. The complete data refer to primary endpoints and the missing group identities, i.e., $Y_{\text{com}} = (\{y_i\}_{i=1}^{N_1}, \{z_i\}_{i=1}^{N_1})$, where z_i denotes the group identity indicator for subject i with $i = 1, ..., N_1$ and $z_i = 1$ or $z_i = 0$ indicates a subject i is drawn from $N(\mu_1, \sigma)$ or $N(\mu_2, \sigma)$, respectively. Because we assume that subjects are randomly assigned equally to the two treatments, $z_1, ..., z_{N_1}$ are modeled as independent Bernoulli distributions with probability 0.5, i.e.,

$$z_i = \begin{cases} 1 & \text{with probability } 0.5 \\ 0 & \text{with probability } 0.5. \end{cases}$$

Thus, y_i is assumed to follow $N(\mu_1, \sigma)$ when $z_i = 1$ and $N(\mu_2, \sigma)$ when $z_i = 0$. The conditional density function of y_i given z_i is

$$\begin{aligned} f(y_i|z_i, \boldsymbol{\theta}) &= f(y_i|\mu_1, \sigma)^{z_i} f(y_i|\mu_2, \sigma)^{1-z_i} \\ &= (2\pi)^{-1/2}\left(\frac{1}{\sigma}\right)^{z_i} \exp\left\{-\frac{(y_i-\mu_1)^2}{2\sigma^2}z_i\right\}\left(\frac{1}{\sigma}\right)^{1-z_i} \exp\left\{-\frac{(y_i-\mu_2)^2}{2\sigma^2}(1-z_i)\right\} \\ &= (2\pi)^{-1/2}\left(\frac{1}{\sigma}\right) \exp\left[-\frac{1}{2\sigma^2}\left\{z_i(y_i-\mu_1)^2 + (1-z_i)(y_i-\mu_2)^2\right\}\right]. \end{aligned} \tag{2.3}$$

The joint density function of y_i and z_i is the product of the marginal distribution of group identity and the conditional distribution of the primary endpoint y_i given the group identity,

$$\begin{aligned} f(y_i, z_i|\boldsymbol{\theta}) &= f(y_i|z_i, \mu_1, \mu_2, \sigma) \times p(z_i|\mu_1, \mu_2, \sigma) \\ &= f(y_i|\mu_1, \sigma)^{z_i} f(y_i|\mu_2, \sigma)^{1-z_i} \frac{1}{2}^{z_i} \left(1 - \frac{1}{2}\right)^{1-z_i} \\ &= \frac{1}{2}(2\pi)^{-1/2} \left(\frac{1}{\sigma}\right) \exp\left[-\frac{1}{2\sigma^2} \left\{ z_i(y_i - \mu_1)^2 + (1 - z_i)(y_i - \mu_2)^2 \right\} \right], \end{aligned} \tag{2.4}$$

which is the joint density function for a pair of complete data (y_i, z_i). Hence, the complete-data log-likelihood function is given by

$$\ell(\boldsymbol{\theta}|\mathbf{y}, \mathbf{z}) = -N_1 \log 2 - \frac{N_1}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{N_1} \left\{ z_i(y_i - \mu_1)^2 + (1 - z_i)(y_i - \mu_2)^2 \right\} - \frac{N_1}{2} \log 2\pi. \tag{2.5}$$

The E-step computes the conditional expectation of the complete-data log-likelihood given the observed data and the current parameter estimates,

$$\begin{aligned} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) &= E\left[\ell(\boldsymbol{\theta}|\mathbf{y}, \mathbf{z}) | \mathbf{y}, \boldsymbol{\theta}^{(t)} \right] \\ &= -\frac{N_1}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{N_1} \left[(y_i - \mu_1)^2 E(z_i|\mathbf{y}, \boldsymbol{\theta}^{(t)}) + (y_i - \mu_2)^2 \{1 - E(z_i|\mathbf{y}, \boldsymbol{\theta}^{(t)})\} \right] \\ &\quad + \text{constant}. \end{aligned} \tag{2.6}$$

Because the conditional expectation of the complete data log likelihood is linear in z_i, it amounts to computing the conditional expectations of the missing treatment identity z_i in (2.6). Specifically, the conditional probability of z_i is written as

$$p(z_i|y_i, \boldsymbol{\theta}) = \frac{f(y_i, z_i|\mu_1, \mu_2, \sigma)}{f(y_i|\mu_1, \mu_2, \sigma)}, \tag{2.7}$$

where $f(y_i, z_i|\mu_1, \mu_2, \sigma)$ is given in (2.4) and $f(y_i|\mu_1, \mu_2, \sigma)$ from (2.1). Thus, we have

$$\begin{aligned} p(z_i|y_i, \boldsymbol{\theta}) &= \frac{f(y_i|\mu_1, \sigma)^{z_i} \times f(y_i|\mu_2, \sigma)^{1-z_i}}{f(y_i|\mu_1, \sigma) + f(y_i|\mu_2, \sigma)} \\ &= \left\{ \frac{f(y_i|\mu_1, \sigma)}{f(y_i|\mu_1, \sigma) + f(y_i|\mu_2, \sigma)} \right\}^{z_i} \left\{ \frac{f(y_i|\mu_2, \sigma)}{f(y_i|\mu_1, \sigma) + f(y_i|\mu_2, \sigma)} \right\}^{1-z_i}, \end{aligned} \tag{2.8}$$

that is, conditional on the observed data and the parameter estimates, the missing treatment identity follows a Bernoulli distribution with probability $f(y_i|\mu_1^{(t)}, \sigma^{(t)})/\{f(y_i|\mu_1^{(t)}, \sigma^{(t)}) + f(y_i|\mu_2^{(t)}, \sigma^{(t)})\}$.

Then the $E(z_i|\mathbf{y}, \boldsymbol{\theta}^{(t)})$ of (2.6) in the E-step is written as

$$E(z_i|\mathbf{y}, \boldsymbol{\theta}^{(t)}) = p(z_i = 1|\mathbf{y}, \boldsymbol{\theta}^{(t)}) = \frac{f(y_i|\mu_1^{(t)}, \sigma^{(t)})}{f(y_i|\mu_1^{(t)}, \sigma^{(t)}) + f(y_i|\mu_2^{(t)}, \sigma^{(t)})}. \tag{2.9}$$

The M-step maximizes the conditional expectation of the complete-data log-likelihood computed in the E-step. Thus we update $\boldsymbol{\theta}^{(t+1)}$ with

$$\begin{aligned}
\mu_1^{(t+1)} &= \frac{\sum_{i=1}^{N_1} y_i E(z_i|\mathbf{y}, \boldsymbol{\theta}^{(t)})}{\sum_{i=1}^{N_1} E(z_i|\mathbf{y}, \boldsymbol{\theta}^{(t)})} \\
\mu_2^{(t+1)} &= \frac{\sum_{i=1}^{N_1} y_i \{1 - E(z_i|\mathbf{y}, \boldsymbol{\theta}^{(t)})\}}{\sum_{i=1}^{N_1} \{1 - E(z_i|\mathbf{y}, \boldsymbol{\theta}^{(t)})\}} \\
\sigma^{2(t+1)} &= \frac{1}{N_1} \sum_{i=1}^{N_1} \left[E(z_i|\mathbf{y}, \boldsymbol{\theta}^{(t)})(y_i - \mu_1^{(t)})^2 + \{1 - E(z_i|\mathbf{y}, \boldsymbol{\theta}^{(t)})\}(y_i - \mu_2^{(t)})^2 \right].
\end{aligned} \tag{2.10}$$

We iterate between the E-step and the M-step until certain convergence criteria are satisfied. When the EM algorithm converges, we obtain the local maximum for the observed data likelihood.

2.3 CONDITIONAL BERNOULLI DISTRIBUTION

2.3.1 Conditional Bernoulli model

Suppose z_i's are independent Bernoulli random variables with probability p_i's respectively, for $i = 1, ..., N_1$. The conditional Bernoulli model is developed by Chen, Dempster and Liu[7] as the conditional distribution of $\mathbf{z} = (z_1, z_2, ..., z_{N_1})$ given that $\sum_{i=1}^{N_1} z_i = n$, where n is the number of $z_i = 1$ out of N_1 observations. To motivate the derivation of the conditional Bernoulli distribution, we first introduce the Poisson-Binomial distribution, which is the distribution of $\sum_{i=1}^{N_1} z_i$ when not all the p_i's are equal. If all the p_i's are equal, it would become the binomial distribution. Under the Poisson-Binomial distribution, the probability

that $\sum_{i=1}^{N_1} z_i = n$ is the sum of the probabilities of $(z_1, ..., z_{N_1})$, where n of them are equal to 1 and $(N_1 - n)$ of them are equal to 0, that is,

$$
\begin{aligned}
p\Big(\sum_{i=1}^{N_1} z_i = n\Big) &= \sum_{\forall Z}\left\{\prod_{i=1}^{N_1} p_i^{z_i}(1-p_i)^{1-z_i} \times 1\Big(\sum_{i=1}^{N_1} z_i = n\Big)\right\} \\
&= p_1 p_2 \cdots p_{n-1} p_n (1-p_{n+1})(1-p_{n+2}) \cdots (1-p_{N_1}) \\
&\quad + p_1 p_2 \cdots p_{n-1} p_{n+1} (1-p_n)(1-p_{n+2}) \cdots (1-p_{N_1}) \\
&\quad + \cdots + (1-p_1)(1-p_2) \cdots (1-p_n) p_{n+1} p_{n+2} \cdots p_{N_1} \qquad (2.11) \\
&= \{(1-p_1)(1-p_2) \cdots (1-p_{N_1})\} \\
&\quad \times \left\{\left(\frac{p_1}{1-p_1} \times \frac{p_2}{1-p_2} \times, \cdots, \times \frac{p_n}{1-p_n}\right) + \cdots \right. \\
&\quad \left. + \left(\frac{p_{n+1}}{1-p_{n+1}} \times \frac{p_{n+2}}{1-p_{n+2}} \times \cdots \times \frac{p_{N_1}}{1-p_{N_1}}\right)\right\},
\end{aligned}
$$

where $1(\cdot)$ denotes the indicator function. We let w_i denote the odds, $p_i/(1-p_i)$, so that the second term in (2.11) becomes the sum of the product of all possible $\binom{N_1}{n}$ combinations of w_i's, and thus (2.11) can be rewritten (Chen and Liu[8]) as

$$
p\Big(\sum_{i=1}^{N_1} z_i = n\Big) = \left\{\prod_{i=1}^{N_1}(1-p_i)\right\} \sum_{1 \le i_1 < \cdots < i_n \le N_1} (w_{i_1} \cdots w_{i_n}), \qquad (2.12)
$$

where $i_1 < \cdots < i_n$ denotes an ordered set of n indices with values between 1 and N_1. There are $\binom{N_1}{n}$ possible combinations of distinct $i_1 < \cdots < i_n$ from $\{1, ..., N_1\}$. The joint

distribution of $\mathbf{z} = (z_1, ..., z_{N_1})$ and the sum of z_i is given by

$$\begin{aligned} p\Big(\mathbf{z}, \sum_{i=1}^{N_1} z_i = n\Big) &= p(\mathbf{z}) \times p\Big(\sum_{i=1}^{N_1} z_i = n | \mathbf{z}\Big) \\ &= \Big\{ \prod_{i=1}^{N_1} p_i^{z_i} (1-p_i)^{1-z_i} \Big\} \times 1\Big(\sum_{i=1}^{N_1} z_i = n\Big) \\ &= \Big\{ (1-p_1)(1-p_2)\cdots(1-p_{N_1}) \Big\} \\ &\quad \times \Big\{ \Big(\frac{p_1}{1-p_1}\Big)^{z_1} \Big(\frac{p_2}{1-p_2}\Big)^{z_2} \cdots \Big(\frac{p_{N_1}}{1-p_{N_1}}\Big)^{z_{N_1}} \Big\} \times 1\Big(\sum_{i=1}^{N_1} z_i = n\Big) \\ &= \Big\{ \prod_{i=1}^{N_1} (1-p_i) \Big\} \prod_{i=1}^{N_1} w_i^{z_i} \times 1\Big(\sum_{i=1}^{N_1} z_i = n\Big). \end{aligned} \tag{2.13}$$

By using (2.12) and (2.13), we obtain that the conditional Bernoulli distribution has the form

$$\begin{aligned} p\Big(\mathbf{z} | \sum_{i=1}^{N_1} z_i = n\Big) &= \frac{p(\mathbf{z}, \sum_{i=1}^{N_1} z_i = n)}{p(\sum_{i=1}^{N_1} z_i = n)} \\ &= \frac{\Big\{ \prod_{i=1}^{N_1} (1-p_i) \Big\} \prod_{i=1}^{N_1} w_i^{z_i} \times 1\big(\sum_{i=1}^{N_1} z_i = n \big)}{\Big\{ \prod_{i=1}^{N_1} (1-p_i) \Big\} \sum_{1 \le i_1 < \cdots < i_n \le N_1} \big(w_{i_1} \cdots w_{i_n} \big)} \\ &= \frac{\prod_{i=1}^{N_1} w_i^{z_i} \times 1\big(\sum_{i=1}^{N_1} z_i = n \big)}{\sum_{1 \le i_1 < \cdots < i_n \le N_1} \big(w_{i_1} \cdots w_{i_n} \big)}, \end{aligned} \tag{2.14}$$

which is Chen and Liu[8]'s equation (3). If $p_i = \frac{1}{2}$ for all i's, then (2.14) can be simplified as $1/\binom{N_1}{n}$.

2.3.2 Recursive generation of R function

The computation of the conditional Bernoulli distribution in (2.14) requires the summation over the product of all $\binom{N_1}{n}$ combinations of w_i's in the denominator. Even with moderate N_1 and n, the computation would not be practical. This is because the summation of $\binom{N_1}{n}$ terms is computationally prohibitive when n and N_1 are large. In the context of retrospective case control studies, Gail, Lubin and Rubinstein[13] earlier developed an efficient recursive method to calculate the summation in (2.14).

Let C denote any set contained in S and let $|C|$ denote the cardinality of a set. The recursive method is based on the function $R(k, C)$,

$$R(k, C) = \sum_{B \subset C, |B|=k} \left(\prod_{i \in B} w_i \right), \tag{2.15}$$

for any non-empty set, $C \subset S$ and $1 \leq k \leq |C|$. We define $R(0, C) = 1$, and $R(k, C) = 0$ for any $k > |C|$.

In the R function, when $k = n$ and the set C includes all N_1 units in $\{1, ..., N_1\}$, the denominator of the conditional Bernoulli distribution in (2.14) is denoted by $R(n, S)$. Then Chen and Liu observed that (2.12) can be rewritten as

$$p\left(\sum_{i=1}^{N_1} z_i = n \right) = \left\{ \prod_{i=1}^{N_1} (1 - p_i) \right\} R(n, S),$$

and also (2.14) can be rewritten as

$$p\left(\mathbf{z} | \sum_{i=1}^{N_1} z_i = n \right) = \frac{\prod_{i=1}^{N_1} w_i^{z_i} \times 1\left(\sum_{i=1}^{N_1} z_i = n \right)}{R(n, S)} \quad i = 0, 1, ..., N_1.$$

The recursive relationship for computing $R(k, C)$ proposed by Gail, Lubin and Rubinstein is as follows. For any $C \subset S$, $1 \leq k \leq |C|$ and $C \setminus \{k\}$ denoting the complement of k in C, we have

$$R(k, C) = R(k, C \setminus \{k\}) + w_k R(k - 1, C \setminus \{k\}), \tag{2.16}$$

which implies that for $S = \{1, 2, ..., N_1\}$ and $\sum_{i=1}^{N_1} z_i = n$ for $i = 1, ..., N_1$, $R(n, S)$ is written as follows

$$R(n, S) = R(n, S \setminus \{i\}) + w_i R(n - 1, S \setminus \{i\}). \tag{2.17}$$

We illustrate the recursive computation of $R(n, S)$ when $n = 2$ and $N_1 = 4$, i.e., $R(n, S) = R(2, \{1, 2, 3, 4\})$. There exist $\binom{4}{2} = 6$ combinations of a pair of w_i's from four distinct w_i's. Thus, $R(2, \{1, 2, 3, 4\})$ is equal to $w_1w_2 + w_1w_3 + w_1w_4 + w_2w_3 + w_2w_4 + w_3w_4$. To compute $R(2, \{1, 2, 3, 4\})$, the recursive formula in (2.17) can be used by first removing the largest index in $\{1, 2, 3, 4\}$. That is, we have

$$\begin{aligned} R(2, \{1, 2, 3, 4\}) &= R(2, S \setminus \{4\}) + w_4 R(1, S \setminus \{4\}) \\ &= R(2, \{1, 2, 3\}) + w_4 R(1, \{1, 2, 3\}) \,. \end{aligned} \tag{2.18}$$

Then we use the recursive formula again to get $R(2, \{1, 2, 3\})$ and $R(1, \{1, 2, 3\})$, i.e.,

$$\begin{aligned} R(2, \{1, 2, 3\}) &= R(2, \{1, 2\}) + w_3 R(1, \{1, 2\}) \\ R(1, \{1, 2, 3\}) &= R(1, \{1, 2\}) + w_3 R(0, \{1, 2\}), \end{aligned} \tag{2.19}$$

where

$$\begin{aligned} R(2, \{1, 2\}) &= R(2, \{1\}) + w_2 R(1, \{1\}) = w_2 R(1, \{1\}) = w_2 w_1 \\ R(1, \{1, 2\}) &= R(1, \{1\}) + w_2 R(0, \{1\}) = w_1 + w_2 \\ R(0, \{1, 2\}) &= 1, \end{aligned} \tag{2.20}$$

since $R(1, \{0\}) = 0$, $R(0, \{0\}) = 1$, $R(0, \{1\}) = 1$ and $R(1, \{1\}) = R(1, \{0\}) + w_1 R(0, \{0\}) = w_1$. With (2.19) and (2.20), $R(2, \{1, 2, 3, 4\})$ in (2.18) can be re-written as the product of every two w's, i.e, $w_1w_2 + w_1w_3 + w_1w_4 + w_2w_3 + w_2w_4 + w_3w_4$.

Figure 1 illustrates how the recursive procedure is used to calculate $R(2, \{1, 2, 3, 4\})$ graphically, referring to Table 1 in Chen and Liu[8]. The entry in the cell corresponding to row 2 and column 4, denoted by $cell(2, 4)$, corresponds to $R(2, \{1, 2, 3, 4\})$. It can be also seen from Figure 1 that the recursive procedure requires $nn_1 - n^2$ additions and $nn_1 - n^2 + n$ multiplications to get $cell(n, N_1)$ because $N_1 - n$ additions and $N_1 - n + 1$ multiplications are required for each row. This is $O(nn_1)$ operations in total[8]. As compared to $\binom{N_1}{n}$ operations required without using the recursive formula in (2.17), it significantly reduces the cost of computation.

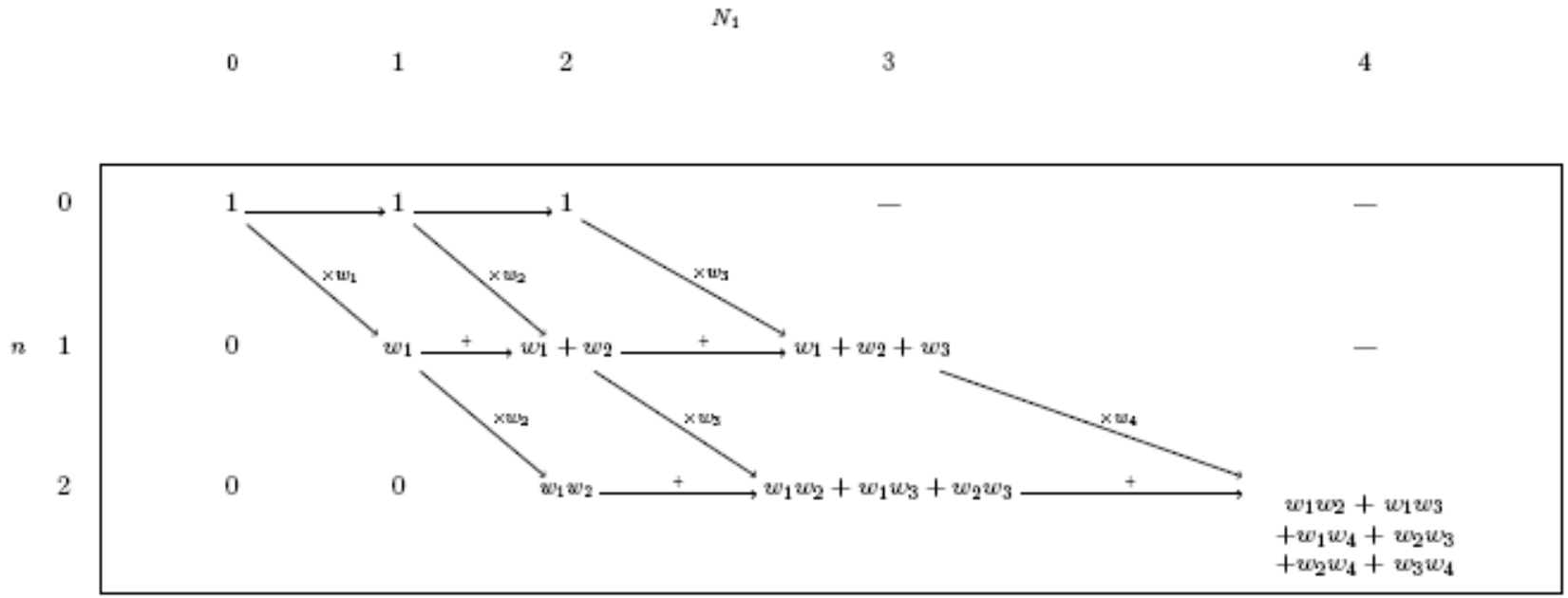


Figure 1: (Chen and Liu, 1997) Recursive generation of $R(n, S)$ for $n = 2, N_1 = 4$; $R(2, \{1, 2, 3, 4\})$ is given in $cell(2, 4)$.

2.4 ENHANCED EM ALGORITHM

2.4.1 Applying conditional Bernoulli model into EM algorithm

In this section, we propose the enhanced EM algorithm for a mixture of normal distributions, which is constructed to take advantage of the observed information of a known number of observations from each group. We use this new algorithm to improve blinded adaptive designs in comparison to Gould and Shih's conventional algorithm. The enhanced EM algorithm is also used to estimate $\boldsymbol{\theta} = (\mu_1, \mu_2, \sigma)$ based on the first N_1 available observations from the mixture of normal distributions without knowing subjects' treatment assignments. Additionally, the enhanced EM algorithm takes into account the fact that there are exactly $N_1/2$ subjects in each treatment group at the interim (i.e., $n = N_1/2$). Gould and Shih[15] ignored this information by treating the group identities z_i's as independent Bernoulli variables. Given the condition $\sum_{i=1}^{N_1} z_i = N_1/2$, however, z_i's are no longer independently distributed and, in fact, $\mathbf{z}$ follow a conditional Bernoulli distribution. When we construct the enhanced EM algorithm, we incorporate this observed information.

We also treat the unobserved treatment identities as missing data. However, our observed data include the fact that $\sum_{i=1}^{N_1} z_i = N_1/2$ in addition to primary endpoints $y_1, ..., y_{N_1}$, i.e., $Y_{\text{obs}} = (\{y_i\}_{i=1}^{n_1}, \sum_{i=1}^{N_1} z_i)$. The complete data likelihood function is given by

$$L\left(\boldsymbol{\theta}|\mathbf{y}, \mathbf{z}, \sum_{i=1}^{N_1} z_i\right) = f\left(\mathbf{y}|\mathbf{z}, \sum_{i=1}^{N_1} z_i, \boldsymbol{\theta}\right) \times p\left(\mathbf{z}|\sum_{i=1}^{N_1} z_i, \boldsymbol{\theta}\right), \tag{2.21}$$

where we know $\sum_{i=1}^{N_1} z_i = N_1/2$ and $p(z_i = 1) = 0.5$. Therefore, the joint probability of the z_i is $1/\binom{N_1}{N_1/2}$, which is a uniform distribution on the subsets of $\overset{N_1}{\underset{1}{\times}}\{0, 1\}$ where there are $N_1/2$ values of 1 in the set. Once we know the z_i's, the summation of z_i's is immediately known. We have

$$f\left(\mathbf{y}|\mathbf{z}, \sum_{i=1}^{N_1} z_i, \boldsymbol{\theta}\right) = f(\mathbf{y}|\mathbf{z}, \boldsymbol{\theta}). \tag{2.22}$$

We assume y_i is normally distributed given z_i. When $z_i = 1$, y_i is distributed with $N(\mu_1, \sigma)$ and when $z_i = 0$, y_i is distributed with $N(\mu_2, \sigma)$. Thus y_i has the conditional density function given by

$$f(y_i|z_i, \boldsymbol{\theta}) = f(y_i|\mu_1, \sigma)^{z_i} \times f(y_i|\mu_2, \sigma)^{1-z_i}. \tag{2.23}$$

Therefore, the complete data likelihood function can be calculated as

$$L\left(\boldsymbol{\theta}|\mathbf{y},\mathbf{z},\sum_{i=1}^{N_1} z_i\right) = \prod_{i=1}^{N_1}\left\{f(y_i|\mu_1,\sigma)^{z_i} f(y_i|\mu_2,\sigma)^{1-z_i}\right\}\frac{1}{\binom{N_1}{N_1/2}}\,. \tag{2.24}$$

Then the complete data log-likelihood function has the following form

$$\begin{aligned}\ell\left(\boldsymbol{\theta}|\mathbf{y},\mathbf{z},\sum_{1}^{N_1} z_i\right) = -\frac{N_1}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{N_1}\left\{z_i(y_i-\mu_1)^2 + (1-z_i)(y_i-\mu_2)^2\right\} \\ - \log\left\{\binom{N_1}{N_1/2}\right\} - \frac{N_1}{2}\log 2\pi,\end{aligned} \tag{2.25}$$

which is linear in z_i with respect to $\boldsymbol{\theta}$.

In the E-step, the conditional expectation of the complete data log-likelihood in (2.25), $\ell(\boldsymbol{\theta}|\mathbf{y},\mathbf{z},\sum_{i=1}^{N_1} z_i)$ given the observed data, $\mathbf{y}$ and $\sum_{i=1}^{N_1} z_i$, is defined by

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = E\left[\ell\left(\boldsymbol{\theta}|\mathbf{y},\mathbf{z},\sum_{i=1}^{N_1} z_i\right)\middle|\mathbf{y},\sum_{i=1}^{N_1} z_i,\boldsymbol{\theta}^{(t)}\right]\,. \tag{2.26}$$

Since the complete data log-likelihood function is linear in z_i with respect to $\boldsymbol{\theta}$, $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ in (2.26) is reduced to a function of the conditional expectation of z_i. Hence, (2.26) is rewritten as

$$\begin{aligned}Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = -\frac{N_1}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{N_1}\Bigg[(y_i-\mu_1)^2 E\big(z_i\big|\mathbf{y},\sum_{i=1}^{N_1} z_i = \frac{N_1}{2},\boldsymbol{\theta}^{(t)}\big) \\ + (y_i-\mu_2)^2\left\{1 - E\big(z_i\big|\mathbf{y},\sum_{i=1}^{N_1} z_i = \frac{N_1}{2},\boldsymbol{\theta}^{(t)}\big)\right\}\Bigg] + \text{constant}.\end{aligned} \tag{2.27}$$

To calculate the conditional expectation of the missing data z_i's, we first find the density function of $\mathbf{z}$ given the observed data and the summation of z_i's, which is written as

$$
\begin{aligned}
p\left(\mathbf{z}|\mathbf{y}, \sum_{i=1}^{N_1} z_i = \frac{N_1}{2}, \boldsymbol{\theta}\right) &= \frac{f(\mathbf{y}|\mathbf{z}, \sum_{i=1}^{N_1} z_i = \frac{N_1}{2}, \boldsymbol{\theta}) \times p(\mathbf{z}|\sum_{i=1}^{N_1} z_i = \frac{N_1}{2})}{\sum_{\forall \mathbf{z}} f(\mathbf{y}|\mathbf{z}, \sum_{i=1}^{N_1} z_i = \frac{N_1}{2}, \boldsymbol{\theta}) \times p(\mathbf{z}|\sum_{i=1}^{N_1} z_i = \frac{N_1}{2})} \\
&= \frac{\prod_{i=1}^{N_1} \left(f(y_i|\mu_1, \sigma)^{z_i} f(y_i|\mu_2, \sigma)^{1-z_i}\right) \frac{1}{\binom{N_1}{N_1/2}}}{\sum_{\forall \mathbf{z}} \left\{ \prod_{i=1}^{N_1} \left(f(y_i|\mu_1, \sigma)^{z_i} f(y_i|\mu_2, \sigma)^{1-z_i}\right) \frac{1}{\binom{N_1}{N_1/2}} \right\}} \\
&= \frac{\prod_{i=1}^{N_1} \left(\frac{f(y_i|\mu_1,\sigma)}{f(y_i|\mu_1,\sigma)+f(y_i|\mu_2,\sigma)}\right)^{z_i} \left(1 - \frac{f(y_i|\mu_1,\sigma)}{f(y_i|\mu_1,\sigma)+f(y_i|\mu_2,\sigma)}\right)^{1-z_i}}{\sum_{\forall \mathbf{z}} \left\{ \prod_{i=1}^{N_1} \left(\frac{f(y_i|\mu_1,\sigma)}{f(y_i|\mu_1,\sigma)+f(y_i|\mu_2,\sigma)}\right)^{z_i} \left(1 - \frac{f(y_i|\mu_1,\sigma)}{f(y_i|\mu_1,\sigma)+f(y_i|\mu_2,\sigma)}\right)^{1-z_i} \right\}} \\
&= \frac{\prod_{i=1}^{N_1} p_i^{z_i}(1-p_i)^{1-z_i}}{\sum_{\forall \mathbf{z}} \left\{ \prod_{i=1}^{N_1} p_i^{z_i}(1-p_i)^{1-z_i} \right\}} \\
&= \frac{\prod_{i=1}^{N_1} (1-p_i) \times \prod_{i=1}^{N_1} {w_i}^{z_i}}{\sum_{\forall \mathbf{z}} \left\{ \prod_{i=1}^{N_1} (1-p_i) \times \prod_{i=1}^{N_1} {w_i}^{z_i} \right\}}.
\end{aligned}
\tag{2.28}
$$

where $p_i = f(y_i|\mu_1, \sigma)/(f(y_i|\mu_1, \sigma)+f(y_i|\mu_2, \sigma))$ and $w_i = p_i/(1-p_i)$. The product of $(1-p_i)$ can be canceled out in the numerator and denominator. Corresponding to the definition in (2.14), $\mathbf{z}$ given $\mathbf{y}$, $\sum_{i=1}^{N_1} z_i = N_1/2$ and $\boldsymbol{\theta}$ in (2.28) is conditional Bernoulli distribution with $\mathbf{p} = (p_1, .., p_{N_1})$.

Therefore, the E-step is computed by using the conditional Bernoulli distribution

$$
\begin{aligned}
&E\left(z_i|\mathbf{y}, \sum_{i=1}^{N_1} z_i = \frac{N_1}{2}, \boldsymbol{\theta}^{(t)}\right) = p\left(z_i = 1|\mathbf{y}, \sum_{i=1}^{N_1} z_i = \frac{N_1}{2}, \boldsymbol{\theta}^{(t)}\right) \\
&= \frac{p(z_i = 1, \sum_{i=1}^{N_1} z_i = N_1/2|\mathbf{y}, \boldsymbol{\theta}^{(t)})}{p(\sum_{i=1}^{N_1} z_i = N_1/2|\mathbf{y}, \boldsymbol{\theta}^{(t)})} \\
&= \frac{p(z_i = 1|\mathbf{y}, \boldsymbol{\theta}^{(t)}) p(\sum_{j\neq i} z_j = N_1/2 - 1|\mathbf{y}, \boldsymbol{\theta}^{(t)})}{p(\sum_{i=1}^{N_1} z_i = N_1/2|\mathbf{y}, \boldsymbol{\theta}^{(t)})} \\
&= \frac{w_i R(N_1/2 - 1, S \setminus \{i\})}{R(N_1/2, S)},
\end{aligned}
\tag{2.29}
$$

where $p(z_i = 1|\mathbf{y}, \boldsymbol{\theta}^{(t)}) = p_i$,
$p(\sum_{j\neq i} z_j = \frac{N_1}{2} - 1|\mathbf{y}, \boldsymbol{\theta}^{(t)}) = \left\{\prod_{j\neq i}(1-p_j)\right\} R(N_1/2 - 1, S \setminus \{i\})$, and

$p(\sum_{i=1}^{N_1} z_i = \frac{N_1}{2}|\mathbf{y}, \boldsymbol{\theta}^{(t)}) = \left\{\prod_{i=1}^{N_1}(1-p_i)\right\} R(N_1/2, S)$.

In the M-step, we maximize $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ with respect to $\boldsymbol{\theta}$. We update the parameter estimates for the $(t+1)^{th}$ iteration as follows:

$$\mu_1^{(t+1)} = \frac{\sum_{i=1}^{N_1} y_i \times E(z_i|\mathbf{y}, \sum_{i=1}^{N_1} z_i, \boldsymbol{\theta}^{(t)})}{\sum_{i=1}^{N_1} E(z_i|\mathbf{y}, \sum_{i=1}^{N_1} z_i, \boldsymbol{\theta}^{(t)})}$$

$$\mu_2^{(t+1)} = \frac{\sum_{i=1}^{N_1} y_i \times \{1 - E(z_i|\mathbf{y}, \sum_{i=1}^{N_1} z_i, \boldsymbol{\theta}^{(t)})\}}{\sum_{i=1}^{N_1} \{1 - E(z_i|\mathbf{y}, \sum_{i=1}^{N_1} z_i, \boldsymbol{\theta}^{(t)})\}}$$

$$\sigma^{2(t+1)} = \frac{1}{N_1}\sum_{i=1}^{N_1}\left[E\left(z_i|\mathbf{y}, \sum_{i=1}^{N_1} z_i, \boldsymbol{\theta}^{(t)}\right)(y_i - \mu_1^{(t)})^2 + \left\{1 - E\left(z_i|\mathbf{y}, \sum_{i=1}^{N_1} z_i, \boldsymbol{\theta}^{(t)}\right)\right\}(y_i - \mu_2^{(t)})^2\right]. \tag{2.30}$$

The conventional EM algorithm uses the primary endpoints y_i's as the only observed data, i.e., $Y_{\text{obs}} = \{y_i\}_{i=1}^{N_1}$. By contrast, our enhanced EM algorithm uses the summation of z_i as additional observed data. That is, we additionally know the number of subjects in the treatment and control groups are both $N_1/2$, i.e., $Y_{\text{obs}} = (\{y_i\}_{i=1}^{N_1}, \sum_{i=1}^{N_1} z_i)$. For both EM algorithms, the complete data log-likelihood function is linear in z_i. Thus the $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ becomes a function of $E(z_i|Y_{\text{obs}}, \theta^{(t)})$. The difference in observed information results in the two EM algorithms maximizing slightly different observed data likelihood functions.

2.4.2 Enhancement of R function

Although (2.29) shows that the conditional expectation of the missing data can be calculated from the recursive relationship of the R function defined in (2.17), this computation can be numerically unstable even for a moderate sample size of N_1. Numerical errors can occur when p_i in (2.29) is close to 1 and thus the corresponding w_i becomes large, where $p_i = f(y_i|\mu_1, \sigma)/(f(y_i|\mu_1, \sigma) + f(y_i|\mu_2, \sigma))$ and $w_i = p_i/(1-p_i)$.

In the E-step of the EM algorithm used to fit the mixture of two normal distributions with $\mu_1 < \mu_2$, observations from the first treatment group tend to have big p_i's close to 1, and the observations from the second treatment group tend to have small p_i's close to 0. As shown in (2.15), the R function consists of a sum of a product of w_i's. Thus, the computation of the R function becomes numerically unstable because of divergence when some p_i's tend

to one and the corresponding w_i's tend to infinity. This numerical problem is illustrated in the following cases.

First, when there is a big treatment effect and thus two treatment groups are well separated, the probability that observation i belongs to the first group, i.e., p_i will tend to be one. For example, suppose that two treatment groups follow $N(0,1)$ and $N(3,1)$, respectively. In this case, we may observe -2 from the first group because it is likely under $N(0,1)$. The resulting p_i is then 0.9999725 and the corresponding w_i equals 36315. Because the R function is a sum of a product of w_i's, such a large w_i causes inflation of the R function and its computation can be numerically unstable. Second, in the large sample case, we may obtain some extreme observations with p_i close to one. For example, suppose two treatment groups follow $N(0,1)$ and $N(\delta,1)$, respectively, for $\delta > 0$. When the sample size is large, it is likely that we observe some extreme observations from the first group, say, -3. If there is at least an appreciable treatment effect, say, $\delta = 2$, such an observation has p_i close to one and the corresponding w_i is large, i.e., $w_i = 2981$. Third, as the sample size becomes larger, it is also likely that the R function grows quickly. Even when there are no extreme observations, a product of relatively large w_i's can still cause inflation of the R function, thereby making its computation numerically unstable.

This numerical problem motivates us to modify the R function and the E-step in (2.29) accordingly. We note that the E-step is computed as the ratio of two R functions, so that canceling out a big common factor between the numerator and denominator of the E-step can make its computation numerically stable. Specifically, we consider factoring out a product of some largest w_i's and model the remaining expression of the R function, denoted by R^*. We thus develop a new recursive relationship for the R^* function and express the E-step in (2.29) in terms of the R^* function. Because of canceling out a product of some largest w_i's between the numerator and denominator, the computation of the E-step in (2.29) becomes numerically stable. The modified R function, i.e., $R^*(k,C)$ is defined as

$$R^*(k,C) = \frac{R(k,C)}{w_{[|C|-k+1]} w_{[|C|-k+2]} ... w_{[|C|]}}, \tag{2.31}$$

where we denote $w_{[1]}, w_{[2]}, ..., w_{[|C|]}$ as the ordered w_i's from the smallest to the largest. That is, $R^*(k,C)$ is the original R function divided by a product of the k largest w_i's.

Figure 2 displays the arithmetic operations of $R^*(n, S)$ with $S = \{1, ..., N_1\}$ and $\sum_{i=1}^{N_1} z_i = n$ for the simple case when $n = 2$ and $S = \{1, 2, 3, 4\}$, where $w = (w_1, w_2, w_3, w_4)$ and $w_4 < w_3 < w_2 < w_1$, i.e., $w_{[1]} = w_4, w_{[2]} = w_3, w_{[3]} = w_2$, and $w_{[4]} = w_1$. Starting from the upper-left corner of the table, i.e., $cell(0, 0)$, $R^*(n, S)$ is generated at the lower-right corner $cell(n, N_1)$. For $i = 1, ..., n$ and $j = 1, ..., N_1$,

$$cell(i, j) = cell(i, j-1) + cell(i-1, j-1) \times \frac{w_j}{w_{[N_1-i+1]}}.$$

In the example, $R^*(2, \{1, 2, 3, 4\})$ is given in $cell(2, 4)$. It is calculated by $cell(2, 3) + cell(1, 3) \times \frac{w_4}{w_{[3]}}$, where $w_{[3]}$ is the second largest w, which is w_2 here. We can see from Figure 2 that the new recursive requires the same number of operations, i.e., $O(nn_1)$, as using the original recursive procedure as shown in Figure 1. Thus, the cost of computation remains the same.

Using our new R^* function, the conditional expectation $E(z_i|\mathbf{y}, \boldsymbol{\theta}^{(t)}, \sum_{i=1}^{N_1} z_i = n)$ is modified as follows.

$$\begin{aligned} E\left(z_i|\mathbf{y}, \boldsymbol{\theta}^{(t)}, \sum_{i=1}^{N_1} z_i = n\right) &= \frac{w_i R(n-1, S \setminus \{i\})}{R(n, S)} \\ &= \frac{R(n-1, S \setminus \{i\}) / \prod_{i=1}^{n-1} w_{[N_1-i+1]}}{R(n, S) / \prod_{i=1}^{n} w_{[N_1-i+1]}} \times \frac{\prod_{i=1}^{n-1} w_{[N_1-i+1]}}{\prod_{i=1}^{n} w_{[N_1-i+1]}} \times w_i \\ &= \frac{R^*(n-1, S \setminus \{i\})}{R^*(n, S)} \times \frac{\prod_{i=1}^{n-1} w_{[N_1-i+1]}}{\prod_{i=1}^{n} w_{[N_1-i+1]}} \times w_i \,. \end{aligned} \tag{2.32}$$

There are two cases for computing the ratio of $\prod_{i=1}^{n-1} w_{[N_1-i+1]}$ and $\prod_{i=1}^{n} w_{[N_1-i+1]}$, depending on the relative size of w_i.

$$\frac{\prod_{i=1}^{n-1} w_{[N_1-i+1]}}{\prod_{i=1}^{n} w_{[N_1-i+1]}} = \begin{cases} \frac{1}{w_{[N_1-n+1]}}, & \text{if } w_i < w_{[N_1-n+1]} \\ \frac{1}{w_i}, & \text{if } w_i \geq w_{[N_1-n+1]} \end{cases}. \tag{2.33}$$

Then (2.32) can be rewritten as

$$E\left(z_i|\mathbf{y}, \boldsymbol{\theta}^{(t)}, \sum_{i=1}^{N_1} z_i = n\right) = \begin{cases} \frac{R^*(n-1, S\setminus\{i\})}{R^*(n,S)} \times \frac{w_i}{w_{[N_1-n+1]}}, & \text{if } w_i < w_{[N_1-n+1]} \\ \frac{R^*(n-1, S\setminus\{i\})}{R^*(n,S)}, & \text{if } w_i \geq w_{[N_1-n+1]} \end{cases}, \tag{2.34}$$

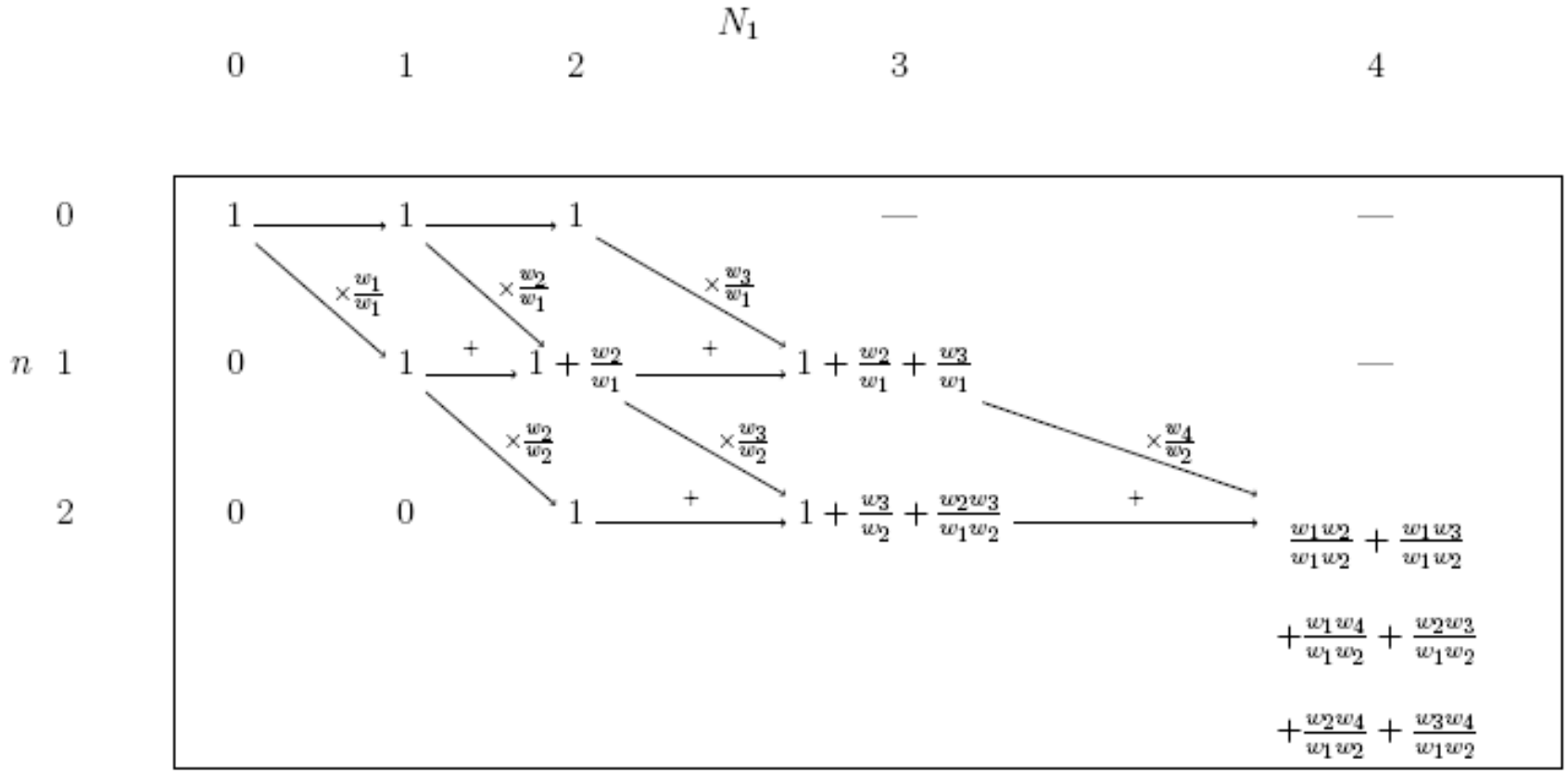


Figure 2: New Recursive generation of $R^*(n,S)$ for $n=2, N_1=4$; w_1 is the largest w and w_2 is the second largest w.

where

$$R^*(n,S) = \frac{R(n,S)}{w_{[N_1-n+1]}w_{[N_1-n+2]}...w_{[N_1]}} \tag{2.35}$$

and

$$R^*(n-1, S\setminus\{i\}) = \frac{R(n-1, S\setminus\{i\})}{\text{product of the } n-1 \text{ largest } w\text{'s after excluding } w_i}\,. \tag{2.36}$$

The conditional expectation in (2.29) is calculated by using (2.34) after setting $n = N_1/2$. Our enhanced R function computed in this way is numerically stable for any given vector $\mathbf{p} = (p_1, p_2, ..., p_{N_1})$.

2.4.3 Idea of using randomized block design

Now suppose block randomization is used in a clinical trial and that we have the information of block sizes at the interim stage. We can use this additional information in the enhanced EM algorithm. We denote by $m_1, m_2, ..., m_B$ the different block sizes for a total of B blocks among the N_1 patients at the interim stage, i.e., $m_1 + m_2 + ... + m_B = N_1$. Within each block, an equal number of subjects is randomly allocated to either the control group or the treatment group.

For notation simplicity, we use equal block sizes to illustrate the procedure of parameter estimation using the enhanced EM algorithm. This procedure can be easily modified for varying block sizes. When the block sizes are fixed, the observations at the interim stage are divided into N_1/m blocks with each block of size m. The extreme case in a clinical trial would be $m = 2$, that is, for every two patients we assign one subject to the control group and the other to the treatment group.

We begin with the simplest case to demonstrate the enhanced EM algorithm. Assume we know that there were two blocks used for the N_1 subjects at the interim stage, that is, the first half of subjects and the second half of subjects are both balanced blocks of size $N_1/2$. Within each of these two blocks, there are $N_1/4$ subjects in the control group and $N_1/4$ in the experimental group. Thus for the enhanced EM algorithm, we are observing

the summation of z_i for each block, which equals to $N_1/4$. The observed data are now $Y_{\text{obs}} = (\{y_i\}_{i=1}^{N_1}, \sum_{i=1}^{N_1/2} z_i, \sum_{i=(N_1/2)+1}^{N_1} z_i)$, and the complete data likelihood function is

$$L\left(\boldsymbol{\theta}|\mathbf{y}, \sum_{i=1}^{N_1/2} z_i, \sum_{i=(N_1/2)+1}^{N_1} z_i\right) = \sum_{\forall \mathbf{z}} \left[\prod_{i=1}^{N_1/2} \left\{ f(y_i|\mu_1,\sigma)^{z_i} f(y_i|\mu_2,\sigma)^{1-z_i} \right\} \cdot \frac{1}{\binom{N_1/2}{N_1/4}} \right. \\ \left. \times \prod_{i=(N_1/2)+1}^{N_1} \left\{ f(y_i|\mu_1,\sigma)^{z_i} f(y_i|\mu_2,\sigma)^{1-z_i} \right\} \cdot \frac{1}{\binom{N_1/2}{N_1/4}} \right]. \tag{2.37}$$

The complete data log-likelihood over the entire trial is just the summation of the complete data log-likelihood in each block, i.e.,

$$\ell\left(\boldsymbol{\theta}|\mathbf{y}, \mathbf{z}, \sum_{i=1}^{N_1/2} z_i, \sum_{i=(N_1/2)+1}^{N_1} z_i\right) = \\ -\frac{N_1}{4}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{N_1/2}\left\{ z_i(y_i-\mu_1)^2 + (1-z_i)(y_i-\mu_2)^2 \right\} - \log\binom{N_1/2}{N_1/4} - \frac{1}{4}\log 2\pi \\ -\frac{N_1}{4}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=(N_1/2)+1}^{N_1}\left\{ z_i(y_i-\mu_1)^2 + (1-z_i)(y_i-\mu_2)^2 \right\} - \log\binom{N_1/2}{N_1/4} - \frac{1}{4}\log 2\pi\,. \tag{2.38}$$

In the E-step, the conditional expectation of the complete data log-likelihood function given the observed data and the current iterate of parameters is

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = E\left[\ell\left(\boldsymbol{\theta}|\mathbf{y}, \mathbf{z}, \sum_{i=1}^{N_1/2} z_i, \sum_{i=(N_1/2)+1}^{N_1} z_i\right) \middle| \mathbf{y}, \sum_{i=1}^{N_1/2} z_i = \frac{N_1}{4}, \sum_{i=(N_1/2)+1}^{N_1} z_i = \frac{N_1}{4}, \boldsymbol{\theta}^{(t)} \right] \\ = -\frac{N_1}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{N_1}\left[(y_i-\mu_1)^2 E\left(z_i|\mathbf{y}, \sum_{i=1}^{N_1/2} z_i = \frac{N_1}{4}, \sum_{i=(N_1/2)+1}^{N_1} z_i = \frac{N_1}{4}, \boldsymbol{\theta}^{(t)} \right) \right. \\ \left. + (y_i-\mu_2)^2 \left\{ 1 - E\left(z_i|\mathbf{y}, \sum_{i=1}^{N_1/2} z_i = \frac{N_1}{4}, \sum_{i=(N_1/2)+1}^{N_1} z_i = \frac{N_1}{4}, \boldsymbol{\theta}^{(t)} \right) \right\} \right] \\ + \text{constant}\,. \tag{2.39}$$

Because the complete data likelihood function is linear in z_i with respect to $\boldsymbol{\theta}$, the E-step is equivalent to computing

$$E\left(z_i|\mathbf{y}, \sum_{i=1}^{N_1/2} z_i = \frac{N_1}{4}, \sum_{i=(N_1/2)+1}^{N_1} z_i = \frac{N_1}{4}, \boldsymbol{\theta}^{(t)}\right) = E\left(z_i|\mathbf{y}, \sum_{i=1}^{N_1/2} z_i = \frac{N_1}{4}, \boldsymbol{\theta}^{(t)}\right), \tag{2.40}$$

when $i = 1, ..., N_1/2$; and

$$E\left(z_i|\mathbf{y}, \sum_{i=1}^{N_1/2} z_i = \frac{N_1}{4}, \sum_{i=(N_1/2)+1}^{N_1} z_i = \frac{N_1}{4}, \boldsymbol{\theta}^{(t)}\right) = E\left(z_i|\mathbf{y}, \sum_{i=(N_1/2)+1}^{N_1} z_i = \frac{N_1}{4}, \boldsymbol{\theta}^{(t)}\right), \tag{2.41}$$

when $i = (N_1/2 + 1), ..., N_1$. We can obtain $E\big(z_i|\mathbf{y}, \sum_{i=1}^{N_1/2} z_i = N_1/4, \boldsymbol{\theta}^{(t)}\big)$ from (2.34) for $n = N_1/4$ and $S = \{1, 2, ..., N_1/2\}$. Similarly, we can get $E(z_i|\mathbf{y}, \sum_{(N_1/2)+1}^{N_1} z_i = N_1/4, \boldsymbol{\theta}^{(t)})$ for $n = N_1/4$ and $S = \{(N_1/2 + 1), ..., N_1\}$. The M-step does not change, where we use (2.30) to update parameter estimates.

In the more general cases, we have block size equal to m for the N_1 patients at the interim. We let E_b denote the conditional expectation of z_i given the current iterate of parameters and the observed data in each block. That is,

$$E\left(z_i|\mathbf{y}, \sum_{i=1}^{m} z_i = \frac{N_1}{m}, \sum_{i=m+1}^{2m} z_i = \frac{N_1}{m}, ..., \sum_{i=N_1-m+1}^{N_1} z_i = \frac{N_1}{m}, \boldsymbol{\theta}^{(t)}\right).$$

So

$$E_b = \begin{cases} E(z_i|\mathbf{y}, \sum_{i=1}^{m} z_i, \boldsymbol{\theta}^{(t)}), & \text{if } 1 \le i \le m \\ E(z_i|\mathbf{y}, \sum_{m+1}^{2m} z_i, \boldsymbol{\theta}^{(t)}), & \text{if } m+1 < i \le 2m \\ ... & \\ E(z_i|\mathbf{y}, \sum_{N_1-m+1}^{N_1} z_i, \boldsymbol{\theta}^{(t)}), & \text{if } N_1 - m + 1 < i \le N_1 \ , \end{cases} \tag{2.42}$$

where $1 + m \times (b - 1) \le i \le m \times b$ for $b = 1, 2, ..., N_1/(m)$. The conditional expectation $E(z_i|\mathbf{y}, \sum_{i=1}^{N_1} z_i, \boldsymbol{\theta}^{(t)})$ in the Q function is computed based on (2.42).

In the case of the conventional EM algorithm and knowing the block sizes, it is clear that there is no gain in the observed information. Specifically we are still assuming that within

each block, the probability of a subject assigned to the control or the experimental group is 0.5. The complete data likelihood function in each block is given by

$$L(\boldsymbol{\theta}|\mathbf{y},\mathbf{z}) = \prod_{i=1+m(b-1)}^{mb} \left\{ f(y_i|\mu_1,\sigma)^{z_i} f(y_i|\mu_2,\sigma)^{1-z_i} \right\} \left(\frac{1}{2}\right)^m, \tag{2.43}$$

and hence by independence, the complete data likelihood function of the entire data at the interim stage is the product of the complete data likelihood function of each block. Clearly this product is the same as the complete data likelihood function used for the conventional EM algorithm, assuming no information of block sizes, that is the complete data likelihood function is still given by

$$\begin{aligned} L(\boldsymbol{\theta}|\mathbf{y},\mathbf{z}) &= \prod_{b=1}^{N_1/m} \left[\prod_{i=1+m(b-1)}^{mb} \left\{ f(y_i|\mu_1,\sigma)^{z_i} f(y_i|\mu_2,\sigma)^{1-z_i} \right\} \left(\frac{1}{2}\right)^m \right] \\ &= \prod_{i=1}^{N_1} \left\{ f(y_i|\mu_1,\sigma)^{z_i} f(y_i|\mu_2,\sigma)^{1-z_i} \right\} \left(\frac{1}{2}\right)^{N_1}. \end{aligned} \tag{2.44}$$

The observed data used by the conventional EM algorithm are the primary endpoints y_i's regardless of whether we have information of block sizes or not. Thus breaking the data into blocks does not increase the observed information in the conventional EM algorithm.

2.5 IDENTIFIABILITY AND LABEL SWITCHING

A family of distributions is identifiable with respect to a parameter if distinct values of this parameter correspond to distinct cumulative distribution functions[5]. In our mixture model where the control group is from $N(\mu_c, \sigma)$, and the experimental group is from $N(\mu_t, \sigma)$, without any restrictions on the means, the means of two treatments are not identifiable. However the mixture distribution is identifiable in σ for fixed μ_c and μ_t.

Note that problems with identifiability can be resolved by redefining the model[5]. We use μ_1 to denote the treatment with a smaller mean, i.e., $\mu_1 = \min(\mu_c, \mu_t)$ and use μ_2 to denote the treatment with a bigger mean, i.e., $\mu_2 = \max(\mu_c, \mu_t)$. With this parametrization, our mixture distribution is identifiable with respect to μ_1 and μ_2. To use the estimation of μ_1

and μ_2, in practice, it requires us making an assumption of the real relationship between μ_{t} and μ_{c}. On the other hand, σ is identifiable since two treatments share the common standard deviation. Also, the absolute difference between two treatments is identifiable because it has the same value even if we do not know the order of the estimates of μ_{t} and μ_{c}.

Because $\mu_1 < \mu_2$, we use as initial values $\mu_1^{(0)} < \mu_2^{(0)}$ in the first iteration of EM estimates. We show theoretically with this initial value that $\mu_1^{(t)} < \mu_2^{(t)}$ is guaranteed at every iteration of the EM algorithm, that is, once we begin with $\mu_1^{(0)} < \mu_2^{(0)}$ the inequality is preserved as an iteration goes on. Thus there are no label switching problems in using the EM algorithm when we parameterize with μ_1 and μ_2. Details are described in the Appendix A.

3.0 ASSESSMENT OF THE ENHANCED EM ALGORITHM ESTIMATES

3.1 A SIMPLE EXAMPLE

Consider a motivating example of a hypothetical clinical trial, where four patients' primary endpoints are observed at the interim stage of the trial. We assume the first and third patients are in a control group, and the second and forth patients in a treatment group. Let (y_1, y_2, y_3, y_4) denote the observed primary endpoints and (z_1, z_2, z_3, z_4) denote their treatment identities. Suppose treatment identities are blinded, i.e., we do not know which of the four observations are from the treatment group or are from the control group. By having the four observations from two populations without knowing their identities, we can use both the conventional EM algorithm and the enhanced EM algorithm to estimate two populations' parameters.

We assume z_i are Bernoulli distributed with $P(z_i = 1) = 0.5$ for $i = 1, 2, 3, 4$. The enhanced EM algorithm uses the additional observed information $\sum_{i=1}^{4} z_i = 2$ which is not used in the conventional EM algorithm. The difference between the two EM algorithms lies in the conditional probability of z_i given the observed data Y_{obs} and the parameters. In the conventional EM algorithm the observed information is $Y_{\text{obs}} = \{y_i\}_{i=1}^{4}$, the conditional probability z_i given $\mathbf{y}$ and $\boldsymbol{\theta}$ is independently Bernoulli distributed. In the enhanced EM algorithm the observed information is $Y_{\text{obs}} = (\{y_i\}_{i=1}^{4}, \sum_{i=1}^{4} z_i = 2)$, the conditional probability $\mathbf{z}$ given the observed data and $\boldsymbol{\theta}$ is conditional Bernoulli distributed.

If we just know each patient has an equal probability 0.5 of being in either a control group or a treatment group, there are 2^4 combinations of assigning (z_1, z_2, z_3, z_4) to (y_1, y_2, y_3, y_4). When we know $\sum_{i=1}^{4} z_i = 2$, however, there are $\binom{4}{2} = 6$ combinations of assigning 2 patients in the control group and 2 patients in the treatment group. The reduction in the number of

Table 1: All possible combinations of the group indicators z_1, z_2, z_3 and z_4

	Conventional EM				Enhanced EM				Enhanced EM with 2 blocks			
	z_1	z_2	z_3	z_4	z_1	z_2	z_3	z_4	z_1	z_2	z_3	z_4
all in control	0	0	0	0								
all in treatment	1	1	1	1								
	1	0	0	0								
3 in control	0	1	0	0								
1 in treatment	0	0	1	0								
	0	0	0	1								
	0	1	1	1								
1 in control	1	0	1	1								
3 in treatment	1	1	0	1								
	1	1	1	0								
	1	1	0	0	1	1	0	0				
	1	0	1	0	1	0	1	0	1	0	1	0
2 in control	1	0	0	1	1	0	0	1	1	0	0	1
2 in treatment	0	1	1	0	0	1	1	0	0	1	1	0
	0	1	0	1	0	1	0	1	0	1	0	1
	0	0	1	1	0	0	1	1				

possible combinations is illustrated in Table 1. If we were further to know that the first two and the last two patients are blocked and balanced to have one patient in the control group and the other in the treatment group, we have $\binom{2}{1} \times \binom{2}{1} = 4$ combinations. For example, one of these 4 possibilities is the case that the first and third patients are in a control group, and the second and forth patients in a treatment group. Thus, we have higher probabilities in the EM algorithm of statistically guessing the true treatment assignment when we know more information on blocks.

It is clear that the enhanced EM algorithm has a relative advantage with a blocked design as compared to the conventional EM algorithm. With blocking we get more information because we narrow down the possible number of treatment identifications. Therefore, we expect to obtain increasingly better estimates when we use more observed information.

3.2 INITIAL VALUES FOR THE ENHANCED EM ALGORITHM VERSUS THE CONVENTIONAL EM ALGORITHM

3.2.1 Review on choosing the initial values for the EM algorithm

In literature, a number of people address the issue of initial values. When there are multiple modes in the likelihood function, different initial values of the EM algorithm may converge to different modes. In the case of the five parameter setting $(\mu_1, \mu_2, \sigma_1, \sigma_2, p)$ for a two component mixture normal model with unequal variances and unknown mixing proportion, it is known that the surface for the likelihood function tends to be multimodal[21][29]. Bohning, Schlattmann and Lindsay[3] illustrated the multimodal likelihood with a particular example, where the mixture probability is fixed at $p = 0.5$ and the mean of one population is fixed at 0. They showed that the EM algorithm converged to multiple local maxima for the MLE when the two initial values for the means are not well separated. Lindsay[19] suggested using a different number of starting values, let the algorithm run a long time, and select as the maximum likelihood estimator that local maximum in the interior of the parameter space with the largest likelihood. As for distributions other than the normal distribution,

Seidel, Mosler and Alker[31] showed that the EM algorithm for the mixture of exponential distributions produces different local modes, depending on the initial values of parameters.

In our three parameter setting (μ_1, μ_2, σ) space, where the common standard deviation is unknown and p is known, the MLEs exist and are consistent[1][21]. But the EM estimates may converge to the boundary of parameter space instead of the meaningful interior of the parameter space[23]. We use a single set of simulated data as an illustration to better understand why and whether the conventional or the enhanced EM algorithm gets stuck at the boundary modes.

3.2.2 Illustrative examples concerning initial values for the two EM algorithms

Suppose we have a mixture of two normal samples with total sample size of 20, where ten observations are sampled from $N(0, 1)$ and the other ten are from $N(1, 1)$. We investigate whether the EM estimates depend on the initial values of the EM algorithm. We use a dotplot to display this 20 observations. As seen in Figure 3, our simulated dataset is representative and the two sample means are well separately.

Figure 3: Dot plots of the 10 observations from each of normal distribution N(0,1) and N(1,1)

For this simulated dataset, we calculate the EM estimates according to the conventional EM algorithm and the enhanced EM algorithm by using different initial values by varying the initial standardized treatment effect $d^{(0)} = (\mu_2^{(0)} - \mu_1^{(0)})/\sigma^{(0)}$. The separate choices of $\mu_1^{(0)}$, $\mu_2^{(0)}$ and $\sigma^{(0)}$ do not affect the estimation of EM algorithm, as long as they provide the same value of $d^{(0)}$. Both EM procedures were initialized using values of $d^{(0)}$ running from 0.00625 to 2 by increments of 0.0625. The stopping criterion used in the EM algorithms is

whether the estimates for all three parameters from successive iterations, say iterations $t-1$ and t, satisfy

$$\sqrt{(\mu_1^{(t)} - \mu_1^{(t-1)})^2 + (\mu_2^{(t)} - \mu_2^{(t-1)})^2 + (\sigma^{(t)} - \sigma^{(t-1)})^2} < 10^{-5}.$$

Figures 4 and 5 show respectively the dependence of the conventional and enhanced EM algorithms on the initializing values of the standardized treatment effect, $d^{(0)}$. In Figure 4, we find that the estimates of the conventional EM algorithm result in identical estimates when $d^{(0)}$ is large enough. For small values of $d^{(0)}$, the conventional EM estimates $\hat{\mu}_1$ and $\hat{\mu}_2$ are fairly close. As shown in this particular example, $\hat{\mu}_1$ and $\hat{\mu}_2$ are both roughly equal to 0.63 when $d^{(0)}$ is less than 0.375. That is, the conventional EM estimates occur near $\mu_1 = \mu_2$, which is on the boundary of the parameter space. We call such an estimate the "boundary mode" of the likelihood surface. The boundary mode implies that there exists only one component which is incorrect since there exist two groups with different means. Figure 5 shows that the estimates of μ_1, μ_2 and σ from the enhanced EM algorithm do not vary no matter what initial values are used.

In general, the convergence of the conventional EM algorithm to the meaningful interior modes depends on the initial value of d, the true parameters, data sample size, and even the specific dataset. For the specific simulated data we used, we get stable interior modes of μ_1 and μ_2 when values of $d^{(0)}$ are big enough. However, for some datasets which we examined in detail in our setting, we have not been able to obtain interior estimates no matter how we adjust the initial values. In those cases, $\hat{\mu}_1$ and $\hat{\mu}_2$ are always stuck at the boundary modes, and estimate of σ is not the value that maximizes the observed log-likelihood function.

The actual likelihood surface for the three-parameter setting in the two component mixture normal model is complicated to illustrate because it involves a three dimensional plot. For both EM algorithms, the means (μ_1 and μ_2) and the standard deviation (σ) are conditionally marginally maximized, i.e., we iterate between the maximization of μ_1 and μ_2 given σ and the maximization of σ given μ_1 and μ_2. To illustrate the reason why the conventional EM algorithm gets stuck at the boundary mode but the enhanced EM algorithm does not, we use the profile likelihood function between iterations. We continue to use the same dataset

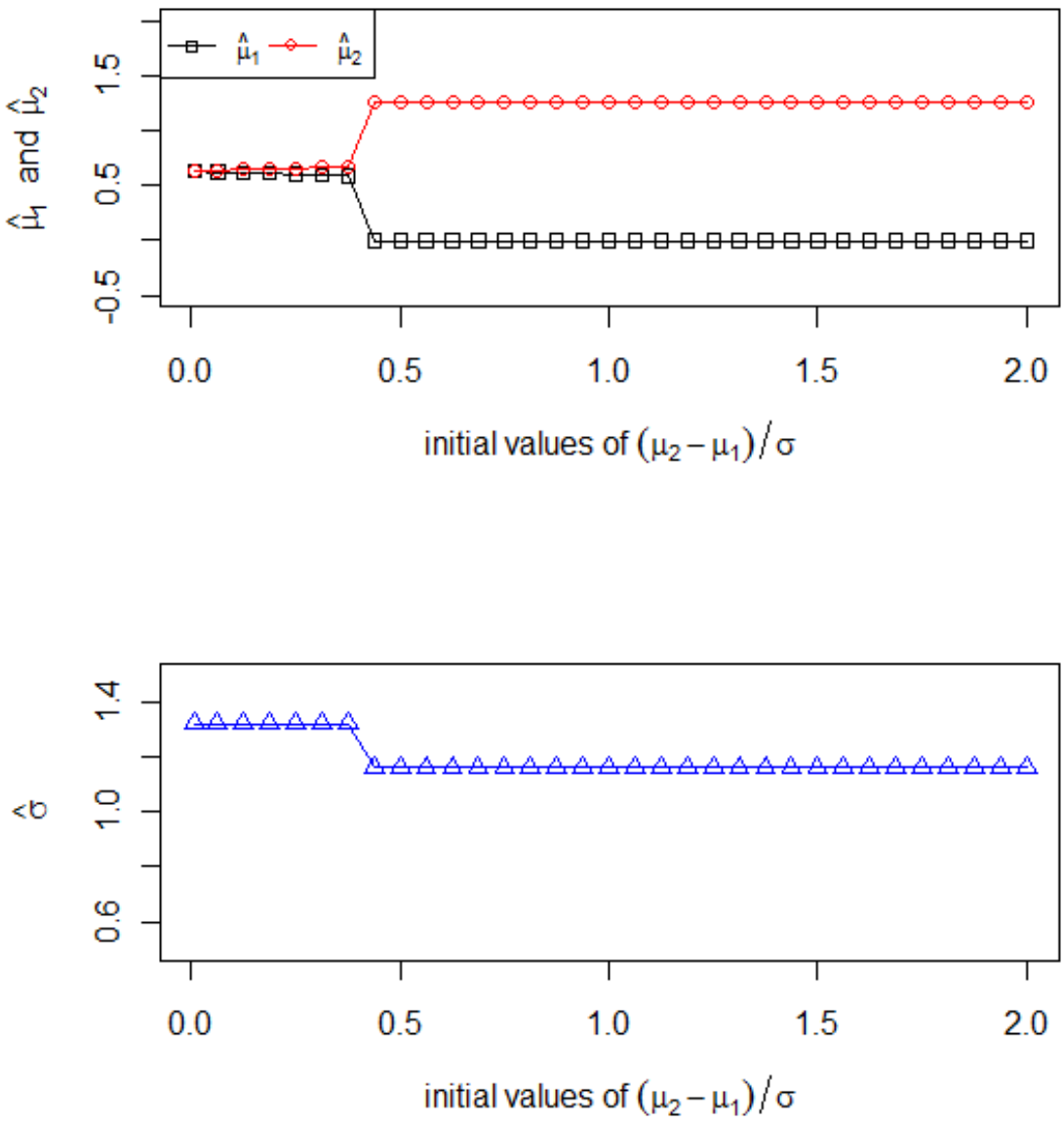


Figure 4: Conventional EM algorithm estimates for a representative dataset, with varying initial values of standardized treatment effect. Initial values of $(\mu_2 - \mu_1)/\sigma$ are set as 0.006255 to 2 by 0.0625. (Simulated sample has sample size 20 with ten from $N(0, 1)$ and the other ten from $N(1, 1)$).

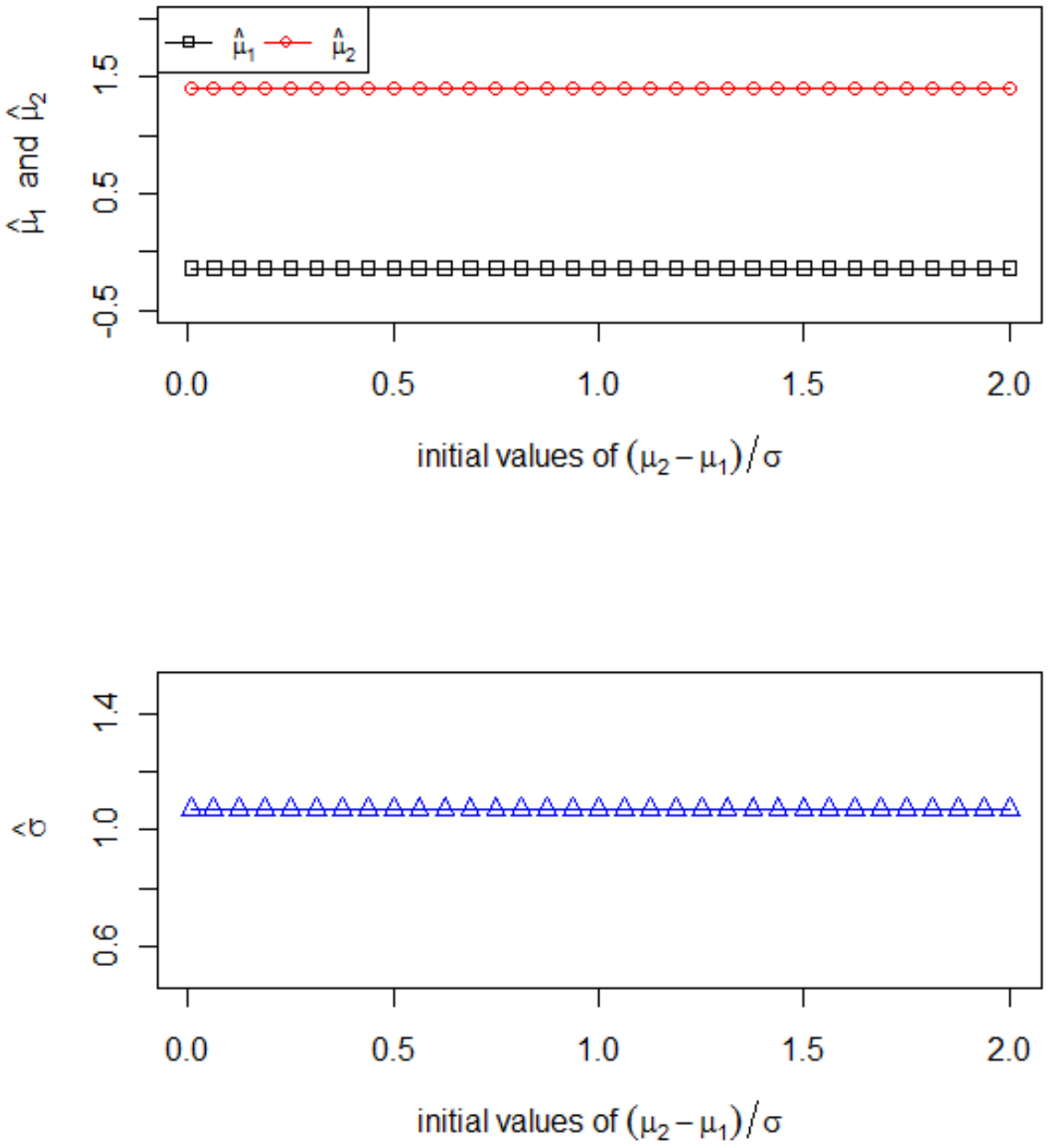


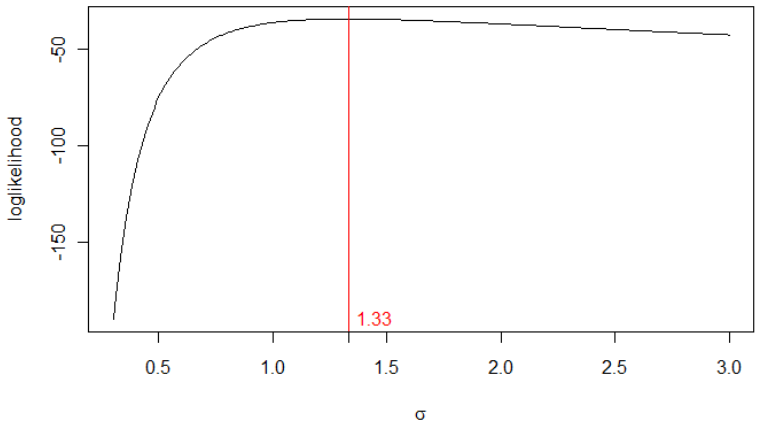
Figure 5: Enhanced EM algorithm estimates for a representative dataset, with varying initial values of standardized treatment effect. Initial values of $(\mu_2 - \mu_1)/\sigma$ are set as 0.006255 to 2 by 0.0625. (Simulated sample has sample size 20 with ten from $N(0, 1)$ and the other ten from $N(1, 1)$).

we simulated as shown in Figure 4 and 5, i.e., a mixture of 20 observations, ten from $N(0,1)$ and the other ten from $N(1,1)$, to illustrate the empirical evidence.

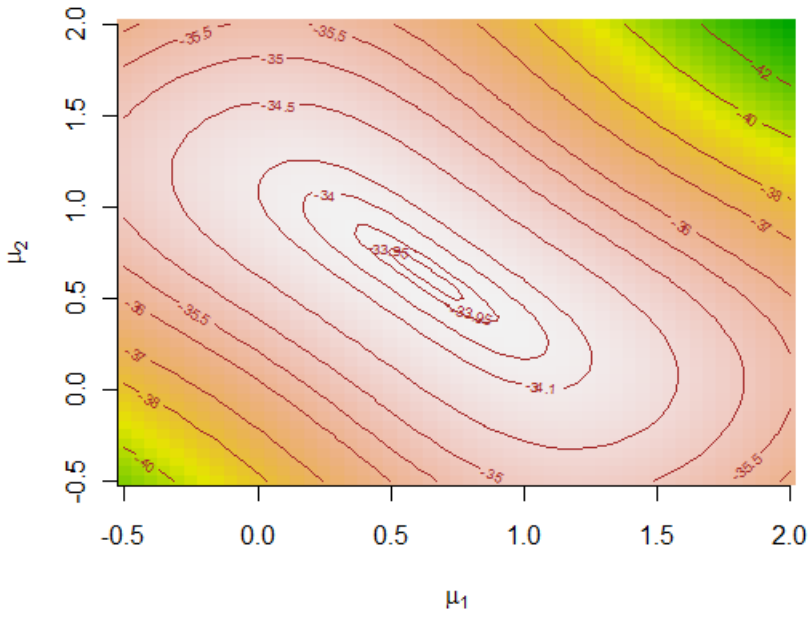
First, we plot the profile log-likelihood function of the conventional EM algorithm for σ when means are started at $\mu_1^{(0)} = \mu_2^{(0)}$, i.e, $d^{(0)} = 0$. In the illustrated example, $\mu_1^{(0)} = \mu_2^{(0)} = 0.5$. Figure 6 (a) shows that the profile log-likelihood function is maximized at $\sigma^{(0)} = 1.33$ given $\mu_1^{(0)} = \mu_2^{(0)}$. In Figure 6 (b), we draw a contour (heatmap) graph of $\mu_1^{(1)}$ and $\mu_2^{(1)}$ given $\sigma^{(0)}$ at 1.33 which is found from the previous step. When $\sigma^{(0)}$ is fixed at the value which maximizes the profile log-likelihood function given $\mu_1^{(0)} = \mu_2^{(0)}$, the profile log-likelihood function of ($\mu_1^{(1)}$, $\mu_2^{(1)}$) becomes unimodal with mode at $\mu_1 = \mu_2$. So the values of $\mu_1^{(1)}$ and $\mu_2^{(1)}$ that maximize the profile likelihood would be necessarily $\mu_1^{(1)} = \mu_2^{(1)}$. Then, the values of $\sigma^{(1)}$ that maximizes the resulting profile log-likelihood function remains at $\sigma^{(0)}$, and thus the profile log-likelihood function for (μ_1, μ_2) is still unimodal. Because of being trapped by the boundary mode, the conventional EM algorithm does not find the interior mode when it begins with $\mu_1^{(0)} = \mu_2^{(0)}$.

For comparison, we use the enhanced EM algorithm for the same dataset. Figure 7 (a) shows the profile log-likelihood functions of the enhanced EM algorithm when means are started at $\mu_1^{(0)} = \mu_2^{(0)}$. The profile log-likelihood function for σ is maximized at $\sigma^{(0)} = 1.33$ given $\mu_1^{(0)} = \mu_2^{(0)} = 0.5$. Similarly, we a draw contour plot of $\mu_1^{(1)}$ and $\mu_2^{(1)}$ given $\sigma^{(0)}$=1.33 in Figure 7 (b). The profile log-likelihood surface for (μ_1, μ_2) becomes slightly bimodal. The values of μ_1 and μ_2 that maximize the resulting profile likelihood are $\mu_1^{(1)} = 0.35$ and $\mu_2^{(1)} = 0.9$ as shown in Figure 7 (b). Next, we fix the means at $\mu_1^{(1)} = 0.35$, $\mu_2^{(1)} = 0.9$ and plot the profile log-likelihood function for σ as in Figure 7 (c), where we can find that $\sigma^{(1)} = 1.29$ maximizes the log-likelihood. Given $\sigma^{(1)} = 1.29$, the profile log-likelihood surface for μ_1 and μ_2 has two modes which are further apart as shown in Figure 7 (d). By iterating the conditional maximization, we were away from the boundary modes near $\mu_1 = \mu_2$. So the enhanced EM algorithm has the nice property that it always obtains the interior modes. This occurs because the σ that maximizes the profile log-likelihood function of the enhanced EM algorithm makes μ_1 and μ_2 separate further on the next iteration in comparison to the previous iteration.

We also simulated other datasets and tried different parameter settings which reflect

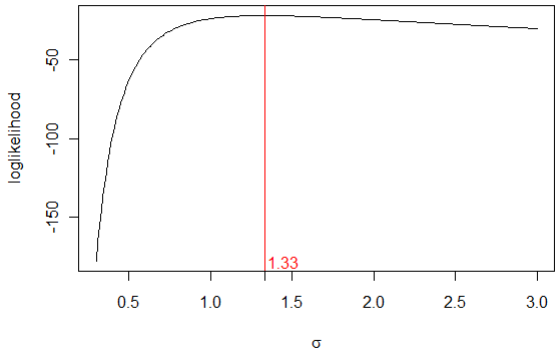


(a) σ versus profile log-likelihood function when $\mu_1 = \mu_2 = 0.5$

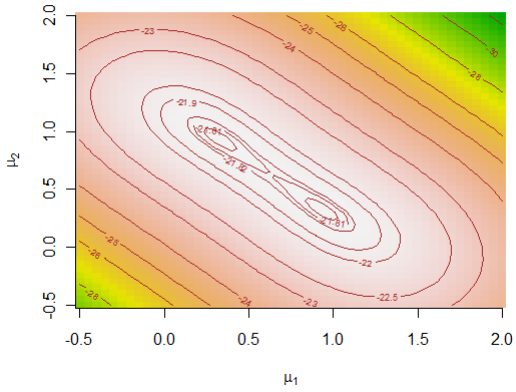


(b) contour plot of μ_1 and $\mu 2$ given $\sigma = 1.33$

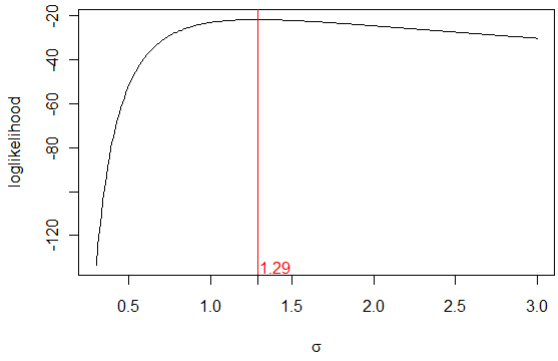
Figure 6: Profile log-likelihood function of the conventional EM algorithm



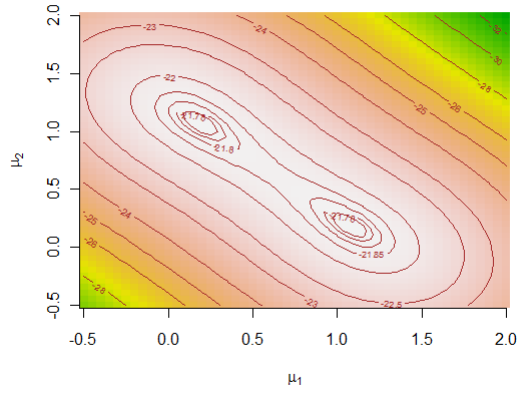
(a) σ versus profile log-likelihood function when $\mu_1 = \mu_2 = 0.5$



(b) contour plot of μ_1 and μ_2 given $\sigma = 1.33$



(c) σ versus profile log-likelihood function when $\mu_1 = 0.35$ and $\mu_2 = 0.9$



(d) contour plot of μ_1 and μ_2 given $\sigma = 1.29$

Figure 7: Profile log-likelihood function of the enhanced EM algorithm

common situations in clinical trials. The graphs show the same features as in our illustrated example. In the simulated data we used the identical initial values of the two means, which is the most extreme case. When the initial values are not identical, however, the conventional EM estimates still can be stuck around the boundary mode at a certain iteration t if the estimates from the pervious iteration $\mu_1^{(t-1)}$ and $\mu_2^{(t-1)}$ are very close to each other. Generally speaking, when the initial standardized treatment effect $d^{(0)}$ is large, it is more likely for the conventional EM algorithm to obtain meaningful interior modes than when $d^{(0)}$ is small. But sometimes, even when $d^{(0)}$ is quite large, after a certain iteration, the conventional EM estimates of μ_1 and μ_2 become very close and they remain stuck around the boundary mode for the rest of the iterations.

3.3 SIMULATION STUDY FOR COMPARING TWO EM ALGORITHMS

For the conventional EM algorithm, though we only make use of the fact that each subject has 0.5 probability to be assigned to each treatment, we design the stimulation study with equal subjects in each treatment at the interim stage. (Gould and Shih[15] did the same in their simulation study[15].) While Gould and Shih[15]'s EM algorithm does not use the assumption of equal number of subjects in each group, the enhanced EM algorithm uses this additional information.

We conduct a simulation study to compare the performance of the two EM algorithms for a reasonable range of parameters values. For simplicity, the true value of σ is set to 1 and μ_1 is set to 0. Let $\delta = \mu_2 - \mu_1$ denote the true treatment effect, so that μ_2 has the same value as δ. The values of δ are set to $0.1, 0.2, 0.35, 0.5, 0.75, 1, 1.5$ and 2. We consider the sample sizes $N_1 = 20$ and $N_1 = 80$ at which we would obtain our interim data. In our simulation study, we generate an equal number of observations from two normal distributions with means μ_1 (control group) and μ_2 (experimental group) and with common standard deviation σ. The two EM algorithms are used to estimate μ_1, μ_2, σ and $(\mu_2-\mu_1)/\sigma$ for each set of observations. The value of $(\mu_2-\mu_1)/\sigma$ was chosen for the inherent interest in this parameter in clinical trial. Gould and Shih[15] noted that the conventional EM algorithm "does not estimate reliably

the true difference between the treatment means". Thus we also compare the estimates of $(\mu_2 - \mu_1)/\sigma$ between two EM algorithms to assess if the enhanced EM algorithm improve this estimate.

For each combination of true parameters, we generate 1000 datasets for each of which we run both EM algorithms. The identical initial values are chosen for both the conventional and enhanced EM procedures on the same dataset as starting values for μ_1, μ_2 and σ. Specifically we use the overall sample mean minus and plus 1.5 as the initial values of μ_1 and μ_2, and the overall sample standard deviation as the initial value of σ. We also apply the same stopping rule for both EM algorithms, that is, the EM algorithm stops when $\sqrt{(\hat{\mu}_1^{(t)} - \hat{\mu}_1^{(t-1)})^2 + (\hat{\mu}_2^{(t)} - \hat{\mu}_2^{(t-1)})^2 + (\hat{\sigma}^{(t)} - \hat{\sigma}^{(t-1)})^2} < 10^{-5}$ is satisfied, where $\hat{\mu}_1^{(t)} - \hat{\mu}_1^{(t-1)}$ denotes the difference between the estimates of μ_1 at the t^{th} iteration and at the $(t-1)^{th}$ iteration, etc, or stop if we hit 20,000 EM iterations. We use 20,000 for the purpose of making simulations manageable. For a given dataset, if the stopping rule is not satisfied at the 20,000th iteration, we could continuously run more iterations and use a trace plot to check the convergence behavior of the estimation.

Results of the simulation study are presented in Tables 2 and 3, which show the bias, variance and mean square errors of estimators based on the 1000 simulated data for each set of parameters.

In Table 2 where the interim sample size is $N_1 = 20$, when δ is 0.1, the conventional EM algorithm gets smaller MSEs for the estimators of μ_1 and μ_2 than the enhanced EM algorithm, but the enhanced EM algorithm gets smaller MSEs for estimating σ and $(\mu_2 - \mu_1)/\sigma$. As δ increases, the differences between the MSEs from the two algorithms decrease. When δ reaches 0.5, the enhanced EM algorithm has smaller MSEs than the conventional EM algorithm for all estimators. We see that both EM algorithms obtain better estimates as δ increases, while the enhanced EM algorithm obtains better estimates more quickly as δ increases. For the largest δ that we used ($\delta = 2$), the MSEs for the enhanced EM estimates are much smaller than the conventional EM estimates. For all values of δ's when $N_1 = 20$, the enhanced EM estimate of σ has smaller MSE than the conventional EM estimate. Importantly, even when δ is small, the enhanced EM algorithm obtains a better estimate of σ, which is an important feature for our ultimate goal of adjusting the sample

size.

Table 3 shows that when $N_1 = 80$, as we would expect, we obtain less biased estimates and smaller MSEs than $N_1 = 20$. As with $N_1 = 20$, the two EM procedures produce better estimates as δ increases. When δ is small, the conventional EM algorithm still obtains better estimates. As we shall see later, this comparison can be misleading because the conventional EM estimates include quite a few boundary modes which are favored in the case of small δ. When δ reaches 0.75, the enhanced EM algorithm has smaller MSEs for μ_1, μ_2, $(\mu_2 - \mu_1)/\sigma$ and very close MSEs for σ. After δ reaching 1, all enhanced EM estimators have smaller MSEs then the conventional EM estimators. The improvement in the enhanced EM algorithm is slower for $N_1 = 80$ than for $N_1 = 20$. The one possible interpretation is that the enhanced EM algorithm takes more advantage of the information of equal numbers of subjects than the conventional EM algorithm does when the sample size is small. When the sample size is large, the impact of using the equal allocation of treatment identities decreases. When we have a large sample size, there is not lot of information gained if we exactly assign half of all subjects into one treatment group or assign subjects with the probability 50% to that treatment group. But if we have a small sample, for example, the sample size is 6, knowing 3 subjects in each treatment provides significantly more information than just knowing there is 0.5 probability of a subject assigning to each treatment.

On the other hand, Tables 2 and 3 both show when decomposing the MSE that the enhanced EM estimates always have a smaller variance but a bigger bias. When we examined the histograms of the 1000 simulated estimates, we find that the distribution of the estimators from the enhanced EM algorithm is bell-shaped while the histogram of the conventional EM estimators is more skewed and outliers exist.

Figures 8 and 9 show the relationship between the 1000 simulated EM estimates of μ_1 and μ_2 when $N_1 = 20$ and $N_1 = 80$, $\mu_1 = 0, \mu_2 = 1$ and $\sigma = 1$. From Figures 8 and 9, we notice that the conventional EM estimates of (μ_1, μ_2) compose two apparent clusters, whereas the enhanced EM algorithm has a single cluster. Comparing the enhanced EM estimates with the interior conventional EM estimates, two clouds are roughly centering around the true value of μ_1 and μ_2 while the conventional EM estimators have bigger variations. For the conventional EM algorithm, most of the estimates are in a cloud around the true value $(0, 1)$,

Table 2: Comparisons of two EM estimates when $N_1 = 20$. True parameters used to generate sample are set as $\sigma = 1$, $\mu_1 = 0$ and $\delta = 0.1, 0.2, 0.35, 0.5, 0.75, 1, 1.5, 2$. 1000 sample are generated from each parameters configuration.

		Enhanced EM estimates				Conventional EM estimates			
δ		μ_1	μ_2	σ	$\frac{\mu_2-\mu_1}{\sigma}$	μ_1	μ_2	σ	$\frac{\mu_2-\mu_1}{\sigma}$
0.1	Bias	-0.5724	0.5807	-0.2912	1.8087	-0.4677	0.4778	-0.2629	1.6188
	Variance	0.0877	0.0957	0.0296	0.6403	0.1697	0.1759	0.0467	1.5990
	MSE	0.4152	0.4328	0.1144	3.9109	0.3882	0.4040	0.1157	4.2177
0.2	Bias	-0.5032	0.5264	-0.2884	1.6918	-0.3982	0.4179	-0.2630	1.5092
	Variance	0.0846	0.0943	0.0330	0.7466	0.1691	0.1810	0.0505	1.7408
	MSE	0.3377	0.3712	0.1161	3.6079	0.3275	0.3554	0.1196	4.0166
0.35	Bias	-0.4516	0.4416	-0.2780	1.5197	-0.3399	0.3298	-0.2473	1.3038
	Variance	0.0940	0.0864	0.0318	0.6500	0.1751	0.1702	0.0489	1.5391
	MSE	0.2979	0.2813	0.1091	2.9587	0.2905	0.2788	0.1100	3.2375
0.5	Bias	-0.3859	0.3989	-0.2772	1.4197	-0.2794	0.2939	-0.2496	1.2264
	Variance	0.0827	0.0905	0.0308	0.6474	0.1695	0.1794	0.0476	1.5778
	MSE	0.2315	0.2495	0.1076	2.6623	0.2474	0.2656	0.1099	3.0803
0.75	Bias	-0.2868	0.2775	-0.2491	1.1632	-0.1583	0.1642	-0.2146	0.9429
	Variance	0.0885	0.0884	0.0372	0.6935	0.1704	0.1839	0.0577	1.6645
	MSE	0.1707	0.1653	0.0992	2.0458	0.1953	0.2107	0.1037	2.5518
1	Bias	-0.2066	0.1966	-0.2096	0.9239	-0.1050	0.0890	-0.1869	0.7898
	Variance	0.0950	0.0987	0.0367	0.6681	0.2081	0.2064	0.0579	1.7101
	MSE	0.1376	0.1373	0.0806	1.5210	0.2190	0.2141	0.0927	2.3322
1.5	Bias	-0.0857	0.0984	-0.1535	0.6391	0.0080	0.0112	-0.1324	0.5386
	Variance	0.1134	0.1069	0.0377	0.7317	0.2267	0.2285	0.0607	1.6308
	MSE	0.1206	0.1165	0.0613	1.1395	0.2265	0.2284	0.0781	1.9193
2	Bias	-0.0032	0.0144	-0.1052	0.4005	0.0674	-0.0624	-0.0838	0.3251
	Variance	0.1099	0.1193	0.0410	0.7165	0.2218	0.2343	0.0652	1.4569
	MSE	0.1098	0.1194	0.0520	0.8762	0.2261	0.2380	0.0721	1.5612

Table 3: Comparisons of two EM estimates when $N_1 = 80$. True parameters used to generate sample are set as $\sigma = 1$, $\mu_1 = 0$ and $\delta = 0.1, 0.2, 0.35, 0.5, 0.75, 1, 1.5, 2$. 1000 sample are generated from each parameters configuration.

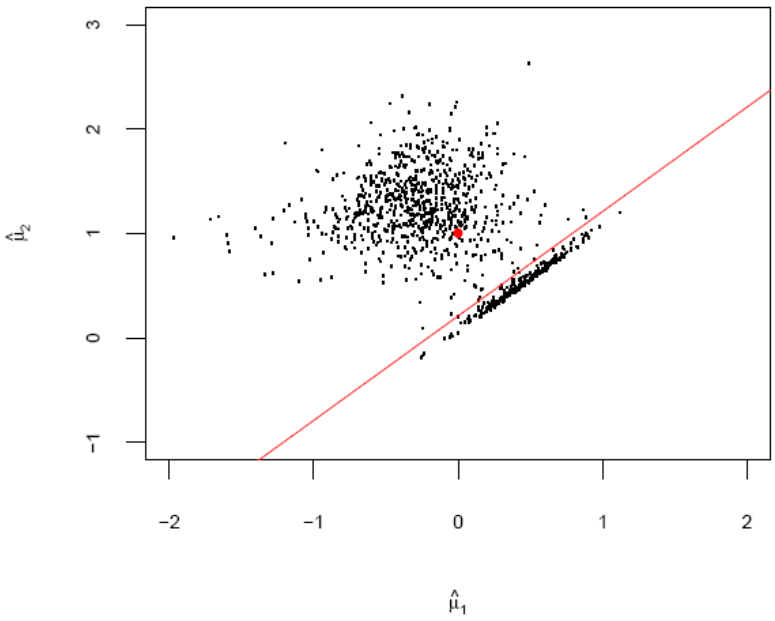
		Enhanced EM estimates				Conventional EM estimates			
δ		μ_1	μ_2	σ	$\frac{\mu_2-\mu_1}{\sigma}$	μ_1	μ_2	σ	$\frac{\mu_2-\mu_1}{\sigma}$
0.1	Bias	-0.4591	0.4619	-0.1802	1.2246	-0.3382	0.3402	-0.1470	0.9376
	Variance	0.0484	0.0461	0.0152	0.3827	0.0984	0.0944	0.0206	0.7185
	MSE	0.2592	0.2594	0.0476	1.8820	0.2127	0.2100	0.0421	1.5969
0.2	Bias	-0.4231	0.4240	-0.1794	1.1605	-0.2944	0.2933	-0.1432	0.8558
	Variance	0.0430	0.0447	0.0168	0.3773	0.0962	0.0978	0.0231	0.7377
	MSE	0.2219	0.2244	0.0490	1.7236	0.1828	0.1837	0.0436	1.4693
0.35	Bias	-0.3623	0.3468	-0.1706	1.0121	-0.2327	0.2200	-0.1347	0.7126
	Variance	0.0445	0.0457	0.0172	0.3931	0.0981	0.0991	0.0233	0.7504
	MSE	0.1757	0.1660	0.0463	1.4172	0.1522	0.1474	0.0414	1.2575
0.5	Bias	-0.2973	0.2936	-0.1553	0.8704	-0.1740	0.1714	-0.1217	0.5918
	Variance	0.0485	0.0491	0.0165	0.3744	0.1068	0.1068	0.0227	0.7315
	MSE	0.1368	0.1353	0.0406	1.1317	0.1334	0.1361	0.0374	1.0810
0.75	Bias	-0.1933	0.1886	-0.1310	0.6324	-0.0662	0.0621	-0.0948	0.3471
	Variance	0.0504	0.0486	0.0175	0.3765	0.1098	0.1051	0.0237	0.7185
	MSE	0.0877	0.0842	0.0346	0.7760	0.1141	0.1089	0.0327	0.8382
1	Bias	-0.1113	0.0981	-0.1020	0.4273	0.0081	-0.0239	-0.0671	0.1674
	Variance	0.0539	0.0557	0.0187	0.3862	0.1151	0.1183	0.0259	0.7333
	MSE	0.0662	0.0653	0.0291	0.5684	0.1150	0.1188	0.0304	0.7606
1.5	Bias	0.0013	0.0007	-0.0428	0.1442	0.1015	-0.0962	-0.0109	-0.0538
	Variance	0.0567	0.0575	0.0203	0.3684	0.1237	0.1276	0.0293	0.6911
	MSE	0.0567	0.0575	0.0221	0.3889	0.1339	0.1368	0.0294	0.6933
2	Bias	0.0158	-0.0147	-0.0230	0.0889	0.0627	-0.0605	-0.0055	0.0047
	Variance	0.0450	0.0441	0.0199	0.3242	0.0954	0.0965	0.0290	0.5557
	MSE	0.0452	0.0443	0.0204	0.3318	0.0993	0.1001	0.0290	0.5558

but quite a few estimates fall on a diagonal line as shown in Figure 8 (a) and 9 (a). The estimates that fall on the diagonal line are the boundary modes we mentioned in Section 3.2.2. We can roughly separate the two clusters of estimates by drawing a straight line on the scatter plot of the conventional EM estimates. For the enhanced EM algorithm all estimators are nicely spread around the true value (0,1) as shown in Figure 8 (b) and 9 (b).

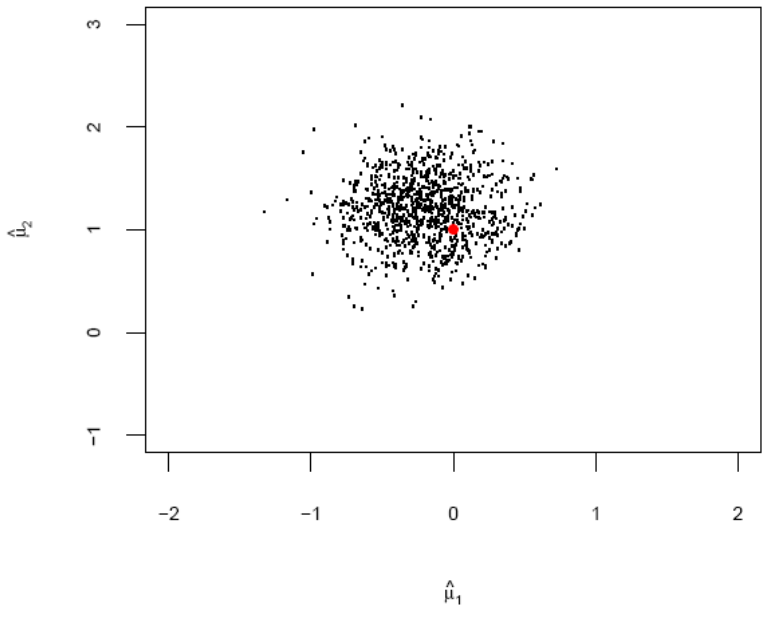
The conventional EM estimates get trapped in the boundary of the parameter space. These boundary estimates would be favorable when the true difference between μ_1 and μ_2 is small since the true difference is close to the boundary at $\mu_1 = \mu_2$. This results in the conventional EM algorithm that produces estimates with smaller bias than the enhanced EM when the true values of μ_1 and μ_2 are near the boundary. However, we emphasize that the small bias results from the fact that the conventional EM algorithm fails to find interior estimates. Even with the case where the true difference between μ_1 and μ_2 is large, the conventional EM estimates get trapped in the boundary of the parameter space, which can greatly mislead inference about parameters. In Appendix B, we illustrate more empirical evidence about this behavior when $\delta = 0.1$ and $\delta = 2$.

We randomly picked out two datasets whose conventional EM estimates fall under the straight line in Figure 9(a), and then created box plots of these two datasets in Figure 10. As seen from the box plots, the sample means of two treatment groups are well separated. Therefore, it is obvious the conventional EM estimates for these two datasets are the non-meaningful boundary modes.

The arbitrary straight line only roughly separates the correct conventional EM estimates around the true values and the incorrect estimates stuck at the boundary modes. It is not necessary that all the EM estimates below the straight line are boundary estimates. Particularly when δ is small, it is hard to tell the estimates below the straight line are boundary estimates or meaningful estimates falling in the boundary area. But through this illustrative example as shown in Figure 8(b) and 9(b), we see that the conventional EM algorithm produces inferior estimates around the diagonal line $\mu_1 = \mu_2$ while the enhanced EM algorithm does not. We just roughly separate the incorrect conventional EM estimates with the intention to find out how the conventional EM algorithm estimate the true parameters if we only consider the meaningful estimates.

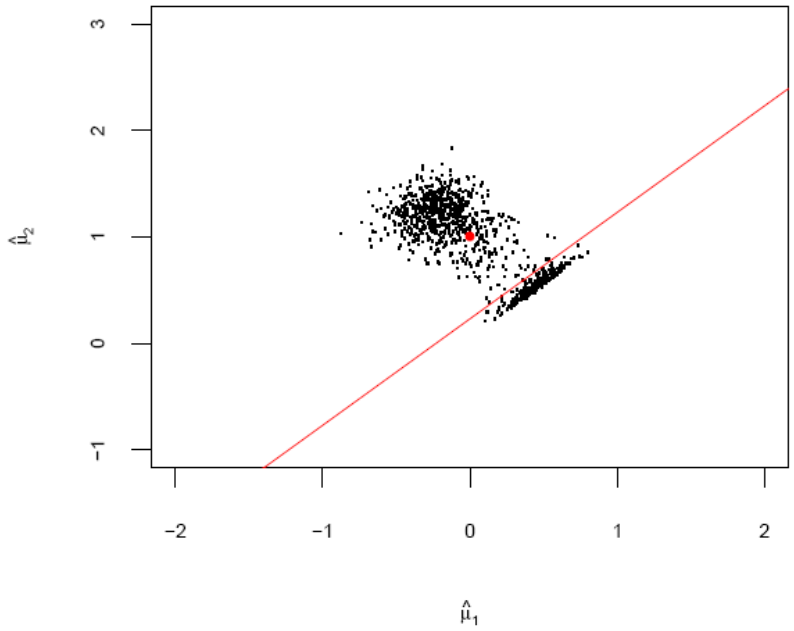


(a) Scatterplot of the Conventional EM estimate of μ_1 versus μ_2

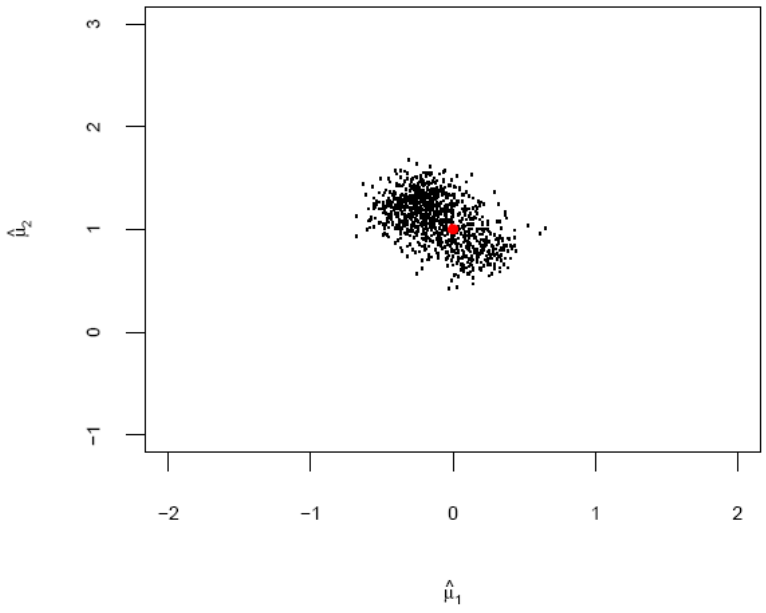


(b) Scatterplot of the Enhanced EM estimate of μ_1 versus μ_2

Figure 8: Comparison of two EM estimates of μ_1 and μ_2 when $N_1 = 20$ ($\mu_1 = 0$, $\mu_2 = 1$ and $\sigma = 1$, stimulater 1000). The red point denotes the true value of (μ_1, μ_2) on the scatterplot.



(a) Scatterplot of the Conventional EM estimate of μ_1 versus μ_2



(b) Scatterplot of the Enhanced EM estimate of μ_1 versus μ_2

Figure 9: Comparison of two EM estimates of μ_1 and μ_2 when $N_1 = 80$ ($\mu_1 = 0$, $\mu_2 = 1$ and $\sigma = 1$, stimulater 1000). The red point denotes the true value of (μ_1, μ_2) on the scatterplot.

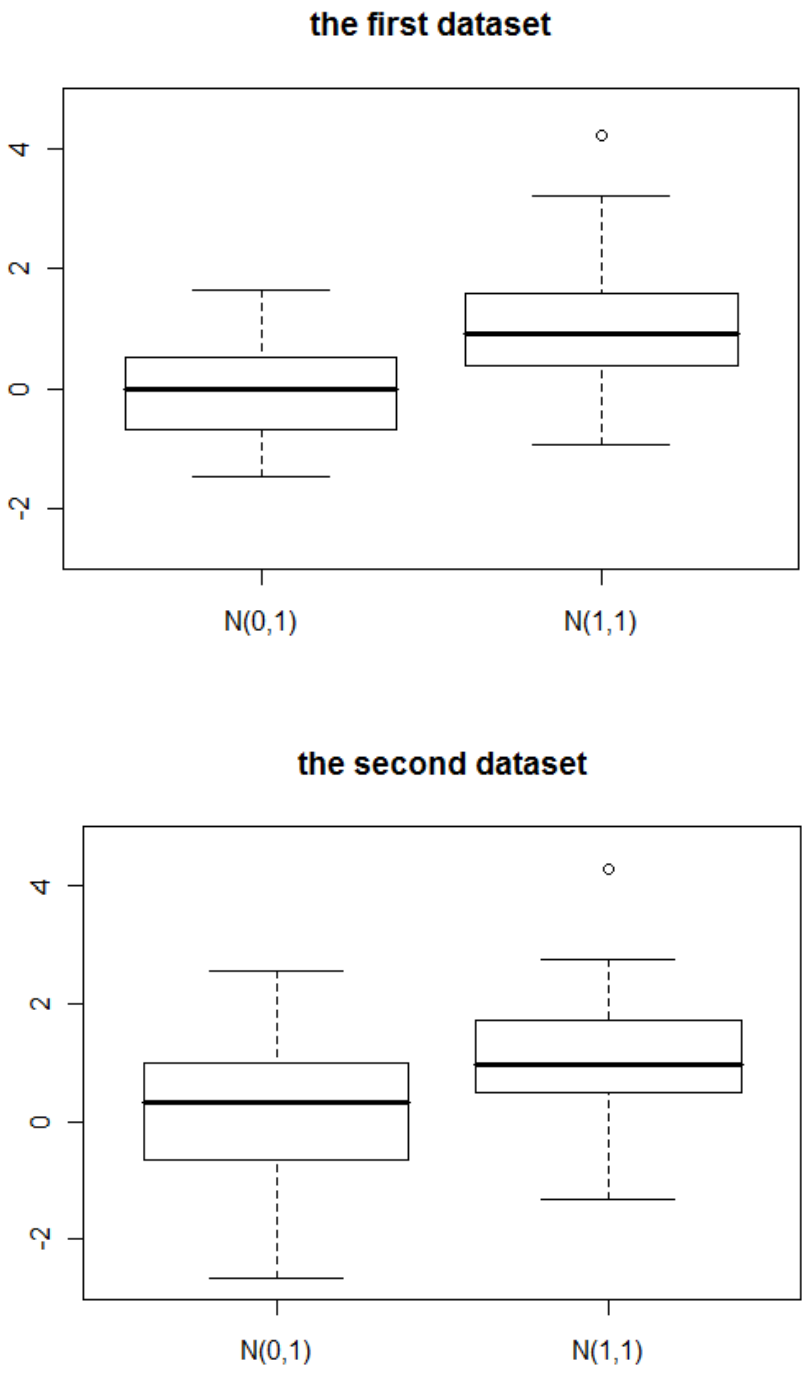


Figure 10: Side-by-side box plots for two randomly chosen datasets which have boundary conventional EM estimates.

For the same dataset, the proportion of the conventional EM estimates that get stuck at the boundary modes depend on the initial values, as we mentioned in Section 3.2.2. Consider the case $\delta = 1$ as shown in Figure 8(a). As an example, we use the straight line $\hat{\mu}_2 = \hat{\mu}_1 + 0.21$ to separate the boundary modes and meaningful estimates and calculate the proportion below the line. With the starting values we used ($\mu_2^{(0)} - \mu_1^{(0)} = 3$ and $\sigma^{(0)}$ is the overall sample standard deviation), there are roughly 240 out of 1000 (24.0%) estimates below the straight line on the scatter plot. For the same 1000 datasets using different initial values such as $\mu_2^{(0)} - \mu_1^{(0)} = 0.1$, $\mu_2^{(0)} - \mu_1^{(0)} = 1$ and $\mu_2^{(0)} - \mu_1^{(0)} = 10$ and the same straight line to separate the conventional EM estimates, we find the proportions of the conventional EM estimates stuck around the boundary of the parameter space are 36.3%, 24.8% and 23.5%, respectively. Note that the proportion decreases as the distance between the starting values of μ_1 and μ_2 gets further apart.

To fully illustrate the effectiveness of the enhanced EM algorithm, we calculate the mean of the Euclidean distance between the true value (μ_1, μ_2) and their EM estimates. We obtain the enhanced EM estimates, the conventional EM estimates, and the conventional EM estimates around the true values for different parameter configurations when $N_1 = 20$ and $N_1 = 80$, and compare the mean of Euclidean distance in Table 4 and Table 5. We find that the mean Euclidean distance of enhanced EM estimates is always smaller than that of the interior conventional EM estimates for different values of δ.

Figure 11 and 12 show the probability density function of σ by using the Gaussian kernel smoother for the simulated 1000 EM estimates for $\delta = 1$ when $N_1 = 20$ and $N_1 = 80$, respectively. The conventional EM estimates of σ have a bimodal distribution. This is because some estimates of (μ_1, μ_2) are stuck at the boundary modes and thus the corresponding conventional EM estimates are incorrect. It is obvious that the enhanced EM algorithm can obtain estimates of σ with smaller bias through comparing the enhanced EM estimates of σ with the meaningful interior conventional EM estimates of σ.

Therefore, if we do not consider the incorrect estimates from the conventional EM algorithm, i.e., the estimates below the straight line as shown in Figure 8 and 9, then the enhanced EM algorithm obtains estimates with smaller bias than the conventional EM algorithm. This is shown in the comparison of these two EM algorithms in Appendix B.

Table 4: Comparisons of the mean of Euclidean distance between (μ_1, μ_2) and their EM estimates when $N_1 = 20$. True parameters used to generate sample are set as $\sigma = 1$, $\mu_1 = 0$ and $\delta = 0.1, 0.2, 0.35, 0.5, 0.75, 1, 1.5, 2$. 1000 samples are generated from each parameter's configuration.

δ	Enhanced EM	Conventional EM	Conventional EM around (μ_1, μ_2)
0.1	0.8809	0.7909	0.9597
0.2	0.7961	0.7309	0.9087
0.35	0.7154	0.6800	0.8042
0.5	0.6416	0.6507	0.7393
0.75	0.5270	0.5920	0.6065
1	0.4721	0.6013	0.5622
1.5	0.4334	0.5943	0.4955
2	0.4210	0.5660	0.4637

Table 5: Comparisons of the mean of Euclidean distance between (μ_1, μ_2) and their EM estimates when $N_1 = 80$. True parameters used to generate sample are set as $\sigma = 1$, $\mu_1 = 0$ and $\delta = 0.1, 0.2, 0.35, 0.5, 0.75, 1, 1.5, 2$. 1000 samples are generated from each parameter's configuration.

δ	Enhanced EM	Conventional EM	Conventional EM around (μ_1, μ_2)
0.1	0.6749	0.5440	0.7718
0.2	0.6230	0.5121	0.7320
0.35	0.5329	0.4831	0.6240
0.5	0.4689	0.4791	0.5601
0.75	0.3702	0.4486	0.4303
1	0.3326	0.4439	0.3524
1.5	0.2959	0.4173	0.2832
2	0.2515	0.3277	0.2618

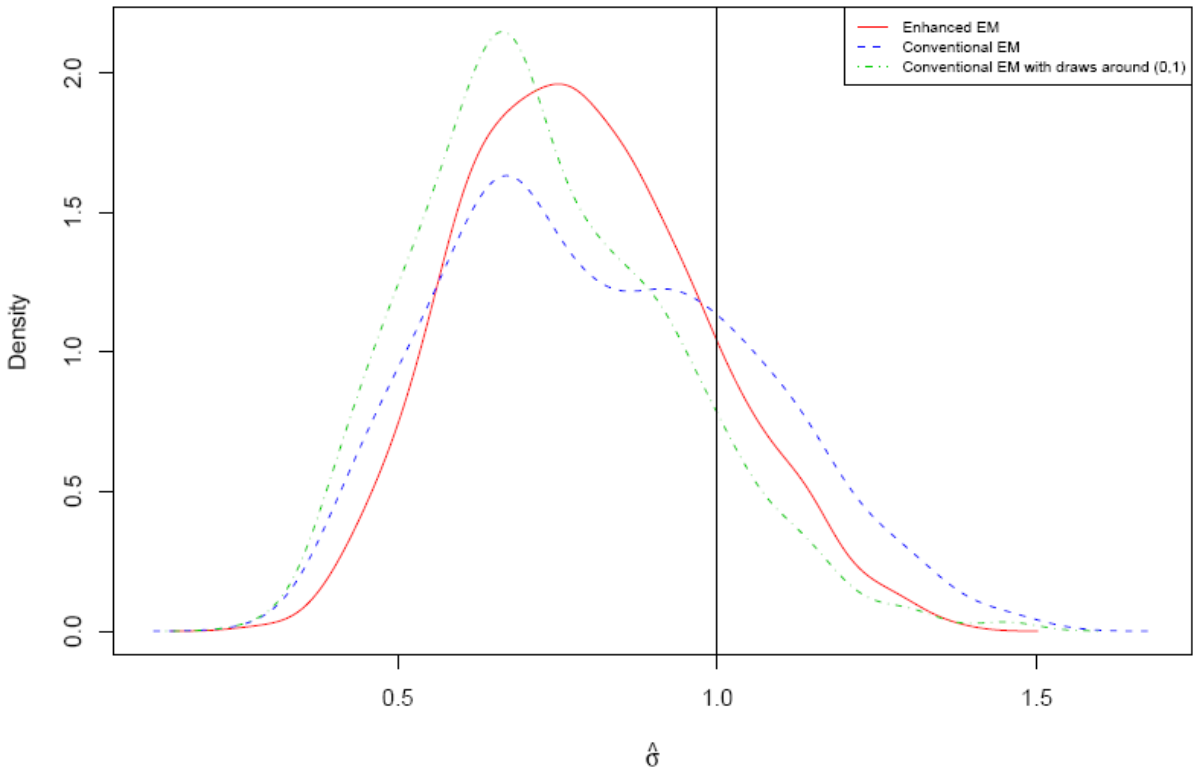


Figure 11: Gaussian kernel smoother of EM estimates of σ when $\mu_1 = 0$, $\mu_2 = 1$ and $\sigma = 1$ for $N_1 = 20$

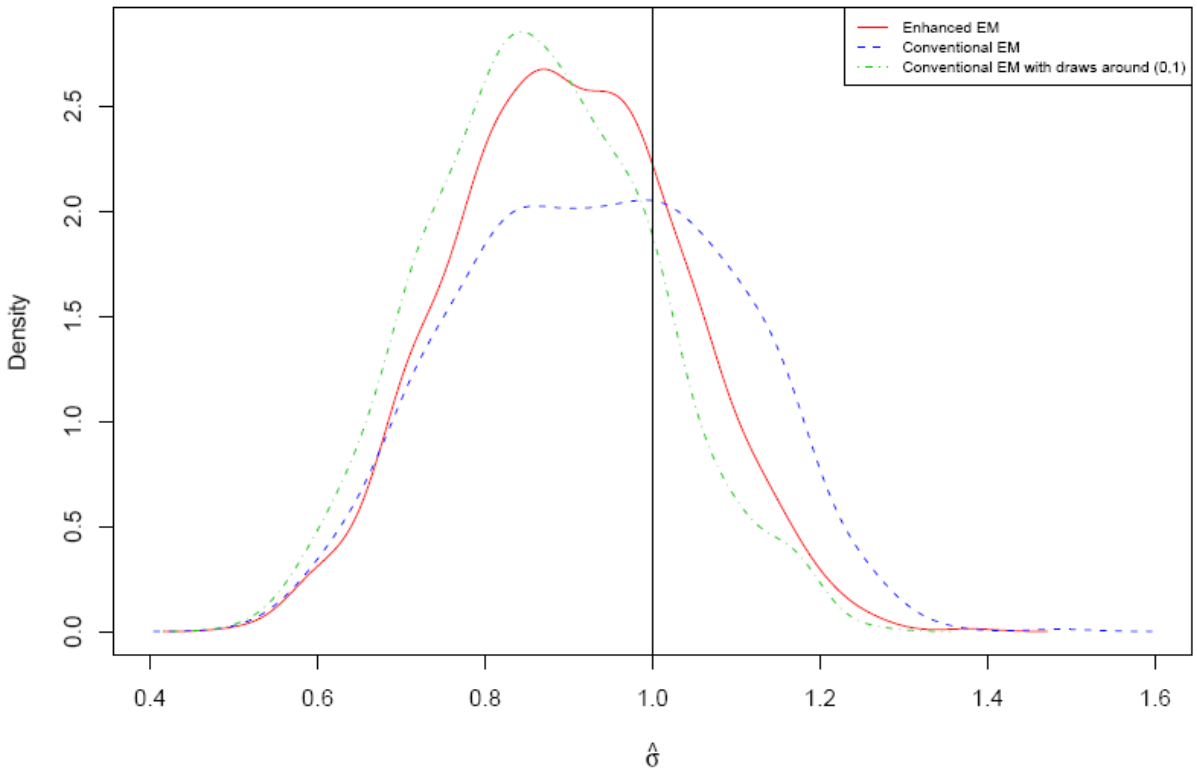


Figure 12: Gaussian kernel smoother of EM estimates of σ when $\mu_1 = 0$, $\mu_2 = 1$ and $\sigma = 1$ for $N_1 = 80$

3.4 SIMULATION STUDY OF THE ENHANCED EM ALGORITHM WITH BLOCK DESIGN

This simulation study is carried out to evaluate the two EM procedures when we have information concerning the block sizes of the randomization at the interim stage. We still maintain the treatment identities blinded at the interim stage, but now assume that the block size used for the block randomization is known. In practice, various block sizes are usually used. For simplicity, we use fixed block sizes to illustrate the performance of the EM procedures as the block size changes. In an actual trial, if we knew the various random block sizes, we could easily apply the EM procedures as described in Section 2.4.3.

We assume the parameter configuration $\delta = 0.5$ ($\mu_1 = 0, \mu_2 = 0.5$, and $\sigma = 1$) and interim stage sample size $N_1 = 80$ for our simulation study. We use all the possible block sizes $2, 4, 8, 10, 16, 20,$ and 40 in the simulation study. For example, when we have the information that the block size is 10, there are 8 blocks and within each block 5 patients are randomly assigned to the control group and the other 5 are assigned to the treatment group. When the block size is 80, the situation corresponds to the study design without blocking as we discussed in Section 3.3. Throughout this simulation study we use 1000 as our simulation size.

The simulation results are given in Table 6. We compare the simulated biases, variances and the mean squared errors of the enhanced EM estimates by taking into account the various block sizes. As noted in Section 2.4.3, there is no difference in the results for the conventional EM estimation no matter how many block sizes we have. The conventional EM algorithm estimates are given in the last row of Table 6. In addition, the MSE's for both EM estimates are plotted in Figure 13.

As was apparent from comparing Tables 2 and 3, Table 6 confirms that the bias and the variance of enhanced EM estimates both become smaller as the block sizes decreases. We can see from Figure 13 that the enhanced EM algorithm begins to have smaller MSEs than the conventional EM algorithm when the block size decreases to 40. As the block size reduces to 2, which is the minimum block size that could be used in a clinical trial, the enhanced estimates are much better than the conventional EM estimates.

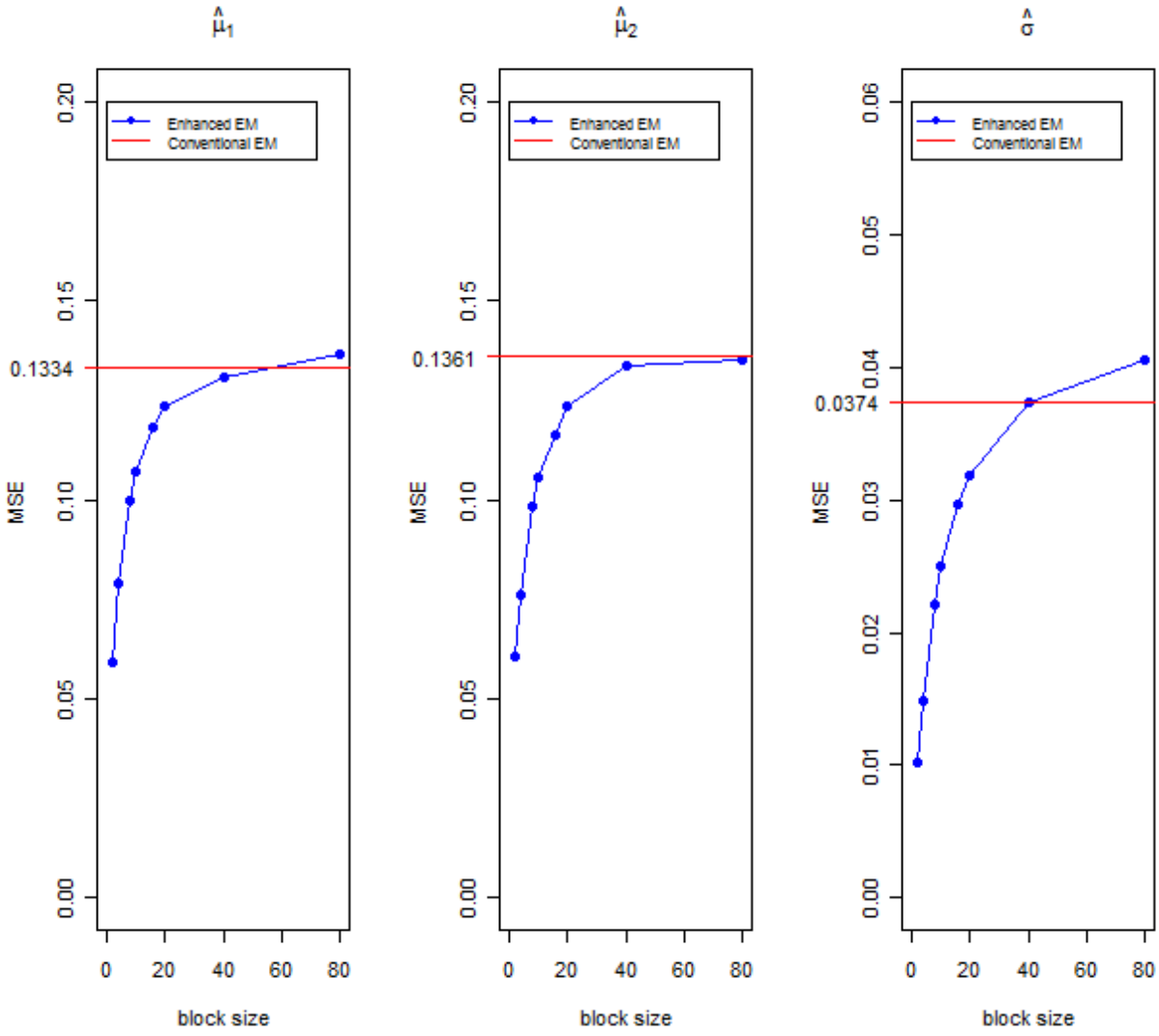


Figure 13: MSEs for the EM estimates $\hat{\mu_1}$, $\hat{\mu_2}$ and $\hat{\sigma}$ with block size $2, 4, 8, 10, 16, 20, 40$ and 80 with $N_1 = 80$ and $\delta = 0.5$.

Table 6: Estimates of the conventional EM algorithm and the enhanced EM algorithm with block design when $N_1 = 80$ and $\delta = 0.5$. Block sizes are $2, 4, 8, 10, 16, 20, 40$ and 80.

Enhanced EM algorithm						
block size	number of blocks	statistics	μ_1	μ_2	σ	$\frac{\mu_1-\mu_2}{\sigma}$
80	1	Bias	-0.2973	0.2936	-0.1553	0.8704
		Variance	0.0485	0.0491	0.0165	0.3744
		MSE	0.1368	0.1353	0.0406	1.1317
40	2	Bias	-0.2472	0.2436	-0.1393	0.7446
		Variance	0.0700	0.0744	0.0180	0.4937
		MSE	0.1310	0.1337	0.0374	1.0476
20	4	Bias	-0.1862	0.1825	-0.1170	0.5809
		Variance	0.0892	0.0905	0.0182	0.5707
		MSE	0.1237	0.1237	0.0319	0.9075
16	5	Bias	-0.1694	0.1658	-0.1084	0.5301
		Variance	0.0895	0.0890	0.0179	0.5509
		MSE	0.1181	0.1164	0.0297	0.8314
10	8	Bias	-0.1276	0.1240	-0.0902	0.4113
		Variance	0.0907	0.0906	0.0169	0.5271
		MSE	0.1069	0.1058	0.0250	0.6958
8	10	Bias	-0.1042	0.1006	-0.0793	0.3423
		Variance	0.0892	0.0882	0.0159	0.4892
		MSE	0.1000	0.0983	0.0221	0.6058
4	20	Bias	-0.0474	0.0438	-0.0506	0.1725
		Variance	0.0768	0.0740	0.0123	0.3599
		MSE	0.0789	0.0759	0.0148	0.3893
2	40	Bias	-0.0032	-0.0004	-0.0284	0.0467
		Variance	0.0590	0.0607	0.0094	0.2386
		MSE	0.0590	0.0607	0.0102	0.2405
Conventional EM algorithm		Bias	-0.1740	0.1714	-0.1217	0.5918
		Variance	0.1068	0.1068	0.0227	0.7315
		MSE	0.1334	0.1361	0.0374	1.0810

We choose $\delta = 0.5$, which is a moderate size of δ, to compare the conventional EM estimates with the enhanced EM estimates for various block sizes. The results shown in Table 6 are representative. We also examined the enhanced EM algorithm for other values of δ. The mean bias and variance always get smaller as the block size decreases for different δ. As we saw in Table 3, when δ is bigger than 0.75 for $N_1 = 80$, the enhanced EM algorithm without blocking has better MSEs than that of the conventional EM algorithm. So as the block size decreases for the cases with δ bigger than 0.75, the enhanced EM algorithm shows more advantages than the conventional EM algorithm. When δ is small, e.g., $\delta = 0.1$, the enhanced EM algorithm starts to beat the conventional EM algorithm when the block size decreases to 20. When $\delta = 0.1$ and the block size is equal to 20, the MSEs of μ_1, μ_2, σ and $(\mu_1 - \mu_2)/\sigma$ are 0.2052, 0.2070, 0.0364 and 1.4030 respectively, which are all smaller than the corresponding values of the conventional EM algorithm as shown in Table 3. In clinical trials, the block sizes used are usually 2, 4, 6, 8 and 10[4]. For two-treatment trials block sizes of 2 and 4 are commonly used[38]. Thus, if we use this additional information with the enhanced EM algorithm, we can always obtain better estimates than using the conventional EM algorithm.

4.0 TYPE I ERROR AND POWER RESULTS: SINGLE-CENTER TRIAL

4.1 EVALUATING THE EFFECT ON TYPE I ERROR RATE

4.1.1 Illustrating actual type I errors in adaptive sample size design

One concern of sample size re-estimation procedures is that using the standard t-test at the end of the trial may inflate the type I error. The type I error rate could be inflated because the final adjusted sample size is a random variable containing information from the interim study but the adaption is not taken into account for the test statistic and the critical value. The t-statistic is not precisely t distributed any more since the components of the t-statistic are both from the first and second stage of the adaptive design. Nonetheless, in our blinded adaptive setting the standard t-test has traditionally been viewed as a good approximation to the actual test statistic.

It is known the type I error rate may be inflated when the adjusted sample size is based on the unblinded pooled variance estimate[2] [36]. We aim to evaluate if the actual type I error is controlled at the nominal level for our blinded design. One would intuitively expect that under the blinded sample size re-estimation case, the type I error rate should not be affected since the adjusted sample size provides no information about the true treatment effects.

Kieser and Friede[17] used analytical methods to compute the actual type I error rate when the standard t-test statistic is applied in evaluating their blinded sample size re-estimation procedure. Their sample size adjustment is based on the pooled one sample variance from the internal pilot. Due to the simple form of their variance estimator, they were able to split the test statistics into components which are independent random variables

and obtained the joint density of these components. Therefore, the density function of the test statistic can be derived as a product of the densities of its components which they show separately follow chi-square and normal distributions. The actual type I error probability for their procedure can be obtained by integrating the density function of the test-statistic over the rejection region of the t-test. Kieser and Friede showed through numerical integration that actual type I error is controlled at the nominal level. However, their procedure relies on the simple form of the estimates they use and the special features of the t-test. Unfortunately a general method for other forms of blinded estimates of the variance are not available to obtain analytical computation of the actual type I error rate for any given test. Clearly we cannot obtain an explicit form for the re-estimated sample size that is calculated from the EM estimator of variance. Hence, simulation is necessary to evaluate the type I error rate of the adaptive procedure.

In Gould and Shih[15]'s paper, they showed through simulation that the conventional EM procedure preserves the type I error rate of their blinded adaptive design. Under the null hypothesis, observations for the two treatment groups are from one population. Therefore, intuitively knowledge that subjects have equal probabilities to be assigned to each treatment should not be different from knowledge that equal numbers of subjects are assigned to each treatment. Hence, we argue that under the null hypothesis, the enhanced EM procedure intuitively does not use any additional information comparing to the conventional EM procedure. Since the conventional EM procedure has been shown to control the type I error, we would expect the enhanced EM procedure should not inflate the type I error.

4.1.2 Simulation study for actual type I error

4.1.2.1 Purpose of the Simulation Study As we have noted, Gould and Shihm used simulation to show that using the conventional EM algorithm to estimate the variance from the blinded data at the interim stage and to re-calculate the sample size based on this EM estimator does not affect the type I error rate of the standard t-statistic. We introduce the enhanced EM algorithm in this dissertation, which requires more information at the interim stage of the trial than Gould and Shih[15]'s approach required. In our settings, by revealing

the information that there is an equal number of subjects from each treatment at the interim, the individual treatment identities remain blinded, but collectively the blind could be viewed as compromised to some extent. Our goal then is to evaluate the effects of revealing more information on the type I error. In our simulation study, we explore and compare the actual type I error rates by using different EM algorithms under various sample size capping rules over a range of true parameter values of σ.

Keeping the initial estimate of standard deviation fixed, we consider using different true σ because we want to look at the effect on type I error for different adjusted sample sizes. When σ is bigger than the initial estimate, the re-calculated sample size tends to be bigger than the planned sample size, and the chance of rejecting the null hypothesis may increase as the adjusted sample size increases. On the other hand, when σ is smaller than the initial estimate, the re-estimated sample size decreases which also can increase the chance of rejecting the null hypothesis.

Furthermore, we are interested in the effects on type I error rate when using the enhanced EM algorithm with block design. We explained the details of block designs in Section 2.4.3, where block size means the minimum known balance point in numbers of patients on each treatment throughout the trial. For example, block size is 4 means within every four patients, there are two in experimental group and the other two in the control group. In most actual studies, random block sizes are used, that is instead of choosing a constant number as block size, we commonly use varying block sizes. For example, the random block size can result in sizes like 4, 2, 6, 8, 4, 2, 6... throughout the trial. The particular block sizes we choose are only illustrative for the enhanced EM algorithm. In a specific study we can utilize the information of any possible block sizes in the enhanced EM algorithm. We also know the enhanced EM estimators substantially improve as block size decreases; however, this full block information may also makes the design 'less unblinded'. Under the block design, we simulate two representative block sizes 2 and 4. If the actual type I error rate is preserved under the nominal level when block size is 2, one expects that the enhanced EM procedure with larger block sizes will continue to control the type I error rate since block size 2 used the most available information to estimate σ.

4.1.2.2 Description of the Simulation Study Without the loss of generality, we assume the initial estimate of the common standard deviation is 1. For simplicity, we assume the clinical meaningful treatment difference, Δ, is 0.443 to make the initial sample size 160, for a nominal type I error of 0.05 and 80% power. The initial sample size is obtained by using (1.1), i.e.,

$$N = 4 \cdot 1^2 \cdot (z_{0.025} + z_{0.2})^2 / 0.443^2 \approx 160\,.$$

where N is the total initial planned sample size for two treatment groups. We use N_1 to denote the sample size for the first stage study and N' as the total recalculated sample size. As noted in Chapter 1.4, N_{adj} is the adjusted final sample size based on applying different sample size rules to N'.

Since our sample size adjustment procedure is not based on the observed treatment difference, the assumed treatment difference, Δ, remains the same throughout the simulation. We choose to use 25% ($N_1 = 40$) and 50% ($N_1 = 80$) of the initial sample size to conduct the interim analysis. Wittes et al[37] show the choice of internal pilot between 25% to 75% of the expected sample size is practical in clinical trials to keep a balance between the requirement for adjusting the sample size reasonably early in the study and the requirement for including sufficient first stage data to achieve a stable estimate of the variance.

To evaluate the effect on type I error rate, obviously the true treatment difference, δ, is set to 0. A range of true σ values (0.5, $1/\sqrt{2}$, 1, $\sqrt{2}$ and 2) are selected. Because the actual value of $\mu_1 = \mu_2$ is not relevant to the t-test under H_0, we set $\mu_1 = \mu_2 = 0$. We generate 3000 samples from $N(0, \sigma)$ where each sample has sample size N_1.

We apply four EM procedures (conventional EM, enhanced EM, enhanced EM with block size 4 and enhanced EM with block size 2) for obtaining the estimate of σ. Three sample size capping rules are used to obtain the final adjusted sample size: unrestricted design rule[2], restricted design rule[36] and the rule Gould and Shih[15] used in their paper. For the unrestricted design, we increase the sample size when the recalculated sample size N' is bigger than the first stage sample size N_1. If N' is smaller, then N_1 is the final adjusted sample size and the trial is stopped at the interim. For the restricted design, we increase the sample size when the recalculated sample size N' is bigger than the initially planned sample size N, otherwise N is used as the final sample size. In Gould and Shih[15]'s paper,

they increased the sample size to the recalculated sample size N' when $N'/N > 1.33$ and $N'/N < 2$. If N' is smaller than $1.33N$, there will be no sample size adjustment and initial sample size N is used for the study. Also Gould and Shih[15] capped the maximum sample size as $2N$ when $N' > 2N$ as a practical limitation. Specifically, we obtain the recalculated sample size N' by using the estimate of σ from different EM procedures using (1.2). The adjusted sample sizes, denoted as N_{adj}, are obtained by applying the three capping rules to each recalculated sample size.

In our simulation study, for each true value of the standard deviation, we generate 3000 random samples from $N(0, \sigma^2)$, where each random sample has size $(N_1 + 1500)$, 1500 being an arbitrary large number. N_1 is the sample size for the first stage of the study. We conduct the interim analysis at N_1 and calculate the adjusted total sample size N_{adj}. Then we take $(N_{\text{adj}} - N_1)$ observations out of the remaining 1500 simulated values and run the t-test. This is repeated for each of the 3000 samples. By doing so, we guarantee there is a high proportion of the data that are common for each scenario, so that different EM algorithms and different capping rules for the same σ are more comparable.

After all N_{adj} subjects are generated, we compute the standard t-statistic as if the sample size were fixed:

$$t = \frac{\bar{y}_1 - \bar{y}_2}{S_{pool}\sqrt{4/N_{adj}}}, \tag{4.1}$$

where $\bar{y}_1$ and $\bar{y}_2$ are the sample means of two treatments, and S_{pool} is the pooled sample standard deviation for the entire dataset. We use $N_{\text{adj}} - 2$ as degrees of freedom for the t-test, where the rejection region is two-sided ($\alpha = 0.05$). After computing the test statistic, we count the number of rejections under the null hypothesis in the 3000 tests for each scenario.

4.1.2.3 Simulation Results For the cases $N_1 = 40$ and $N_1 = 80$, Tables 7 and 8 display the proportions and numbers of rejections under the null hypothesis among the 3000 samples for each scenario of σ. A two-sided exact binomial confidence interval for the rejection proportion is also calculated in each cell of these two tables. In both cases, $N_1 = 40$ and $N_1 = 80$, the 95% confidence interval for the proportion of rejections always includes the nominal type I error of 0.05. It is clear the blinded sample size adjustment through the EM algorithm if it has any effect on the significance level, it is negligible. In addition, we note

Table 7: Simulated type I error rate and confidence interval when $N_1 = 80$. True parameters used to generate the sample are set as $\delta = 0$ and $\sigma = 0.5$, $1/\sqrt{2}$, 1, $\sqrt{2}$ and 2. 3000 sample are generated from each value of σ.

		Conventional EM	Enhanced EM	Enhanced EM with block size 4	Enhanced EM with block size 2
True σ	capping rule				
$\frac{1}{2}$	Unrestricted	0.0527 (158) (0.0449, 0.0613)	0.0527 (158) (0.0449, 0.0613)	0.0527 (158) (0.0449, 0.0613)	0.0527 (158) (0.0449, 0.0613)
	Restricted	0.0507 (152) (0.0431, 0.0591)	0.0507 (152) (0.0431, 0.0591)	0.0507 (152) (0.0431, 0.0591)	0.0507 (152) (0.0431, 0.0591)
	Gould-Shih's	0.0507 (152) (0.0431, 0.0591)	0.0507 (152) (0.0431, 0.0591)	0.0507 (152) (0.0431, 0.0591)	0.0507 (152) (0.0431, 0.0591)
$\frac{1}{\sqrt{2}}$	Unrestricted	0.0523 (157) (0.0446, 0.0609)	0.0520 (156) (0.0443, 0.0606)	0.0530 (159) (0.0453, 0.0616)	0.0530 (159) (0.0453, 0.0616)
	Restricted	0.0493 (148) (0.0417, 0.0577)	0.0493 (148) (0.0417, 0.0577)	0.0493 (148) (0.0417, 0.0577)	0.0493 (148) (0.0417, 0.0577)
	Gould-Shih's	0.0493 (148) (0.0417, 0.0577)	0.0493(148) (0.0417, 0.0577)	0.0493 (148) (0.0417, 0.0577)	0.0493 (148) (0.0417, 0.0577)
1	Unrestricted	0.0467 (140) (0.0394, 0.0548)	0.0453 (136) (0.0382, 0.0534)	0.0490 (147) (0.0416, 0.0573)	0.0530 (159) (0.0453, 0.0616)
	Restricted	0.0510 (153) (0.0434, 0.0595)	0.0497 (149) (0.0442, 0.0581)	0.0507 (152) (0.0431, 0.0591)	0.0513 (154) (0.0437, 0.0598)
	Gould-Shih's	0.0507 (152) (0.0431, 0.0591)	0.0503 (151) (0.0428, 0.0588)	0.0507 (152) (0.0431, 0.0591)	0.0510 (153) (0.0434, 0.0595)
$\sqrt{2}$	Unrestricted	0.0490 (147) (0.0416, 0.0573)	0.0493 (148) (0.0419, 0.0577)	0.0527 (158) (0.0449, 0.0613)	0.0530 (159) (0.0453, 0.0616)
	Restricted	0.0490 (147) (0.0416, 0.0573)	0.0507 (152) (0.0431, 0.0591)	0.0527 (158) (0.0449, 0.0613)	0.0530 (159) (0.0453, 0.0616)
	Gould-Shih's	0.0493 (148) (0.0419, 0.0577)	0.0510 (153) (0.0434, 0.0595)	0.0553 (166) (0.0474, 0.0641)	0.0520 (156) (0.0443, 0.0606)
2	Unrestricted	0.0490 (147) (0.0416, 0.0573)	0.0503 (151) (0.0428, 0.0588)	0.0477 (143) (0.0403, 0.0559)	0.0470 (141) (0.0397, 0.0552)
	Restricted	0.0490 (147) (0.0416, 0.0573)	0.0503 (151) (0.0428, 0.0588)	0.0477 (143) (0.0403, 0.0559)	0.0470 (141) (0.0397, 0.0552)
	Gould-Shih's	0.0493 (148) (0.0419, 0.0577)	0.0473 (142) (0.0400, 0.0556)	0.0470 (141) (0.0397, 0.0552)	0.0477 (143) (0.0403, 0.0559)

Table 8: Simulated type I error rate and confidence interval when $N_1 = 40$. True parameters used to generate the sample are set as $\delta = 0$ and $\sigma = 0.5$, $1/\sqrt{2}$, 1, $\sqrt{2}$ and 2. 3000 sample are generated from each value of σ.

		Conventional EM	Enhanced EM	Enhanced EM with block size 4	Enhanced EM with block size 2
True σ	capping rule				
$\frac{1}{2}$	Unrestricted	0.0530 (159) (0.0453, 0.0616)	0.0527 (158) (0.0449, 0.0613)	0.0540 (162) (0.0462, 0.0627)	0.0553 (166) (0.0474, 0.0641)
	Restricted	0.0487 (146) (0.0412, 0.0570)	0.0487 (146) (0.0412, 0.0570)	0.0487 (146) (0.0412, 0.0570)	0.0487 (146) (0.0412, 0.0570)
	Gould-Shih's	0.0487(146) (0.0412, 0.0570)	0.0487(146) (0.0412, 0.0570)	0.0487(146) (0.0412, 0.0570)	0.0487(146) (0.0412, 0.0570)
$\frac{1}{\sqrt{2}}$	Unrestricted	0.0470 (141) (0.0397, 0.0552)	0.0437 (131) (0.0366, 0.0516)	0.0507 (152) (0.0431, 0.0591)	0.0493 (148) (0.0419, 0.0577)
	Restricted	0.0507 (152) (0.0431, 0.0591)	0.0507 (152) (0.0431, 0.0591)	0.0507 (152) (0.0431, 0.0591)	0.0507 (152) (0.0431, 0.0591)
	Gould-Shih's	0.0507 (152) (0.0431, 0.0591)	0.0507 (152) (0.0431, 0.0591)	0.0507 (152) (0.0431, 0.0591)	0.0507 (152) (0.0431, 0.0591)
1	Unrestricted	0.0460 (138) (0.0388, 0.0541)	0.0457 (137) (0.0385, 0.0538)	0.0457 (137) (0.0385, 0.0538)	0.0497 (149) (0.0422, 0.0581)
	Restricted	0.0423 (127) (0.0354, 0.0502)	0.0430 (129) (0.0360, 0.0509)	0.0437 (131) (0.0366, 0.0516)	0.0430 (129) (0.0360, 0.0509)
	Gould-Shih's	0.0433 (130) (0.0363, 0.0512)	0.0430 (129) (0.0360, 0.0509)	0.0447 (134) (0.0376, 0.0527)	0.0440 (132) (0.369, 0.0520)
$\sqrt{2}$	Unrestricted	0.0467 (140) (0.0394, 0.0548)	0.0507 (152) (0.0431, 0.0591)	0.0450 (135) (0.0379, 0.0530)	0.0477 (143) (0.0403, 0.0559)
	Restricted	0.0497 (149) (0.0422, 0.0581)	0.0533 (160) (0.0456, 0.0620)	0.0447 (134) (0.0376, 0.0527)	0.0477 (143) (0.0403, 0.0559)
	Gould-Shih's	0.0497 (149) (0.0422, 0.0581)	0.0510 (153) (0.0434, 0.0595)	0.0440 (132) (0.0369, 0.0520)	0.0483 (145) (0.0409, 0.0566)
2	Unrestricted	0.0500 (150) (0.0425, 0.0584)	0.0533 (160) (0.0456, 0.0620)	0.0490 (147) (0.0416, 0.0573)	0.0523 (157) (0.0446, 0.0609)
	Restricted	0.0513 (154) (0.0437, 0.0598)	0.0530 (159) (0.0453, 0.0616)	0.0490 (147) (0.0416, 0.0573)	0.0523 (157) (0.0446, 0.0609)
	Gould-Shih's	0.0533 (160) (0.0456, 0.0620)	0.0530 (159) (0.0453, 0.0616)	0.0520 (156) (0.0443, 0.0606)	0.0520 (156) (0.0443, 0.0606)

that the actual type I error from different EM procedures are all quite similar to each other, which will allow us to later compare power among procedures directly without adjusting the critical value of the test.

When the null hypothesis of $H_0 : \mu_1 - \mu_2 = 0$ holds true ($\delta = 0$), the number of simulations rejecting H_0 should be binomially distributed according to $B(3000, 0.05)$[15]. In Figure 14, we produce a figure analogous to Gould and Shih[15]'s Figure 2 and plot the observed cumulative density functions (CDFs) of the numbers of rejections of H_0 in 3000 runs by using different EM procedures, and compare them with the theoretical CDF of the binomial distribution with probability 0.05. We use the number of rejections shown in Tables 7 and 8 (30 cases for each procedure) as the empirical distributions of the rejection frequencies for three EM algorithms.

Figure 14 shows the distributions of the rejection frequencies for the different EM algorithms fall closely together. We see in our figure that the four blinded sample size adjustment procedures have a very similar type I error rates. We also notice from Figure 14 that none of the EM procedures inflate the type I error materially. However, Figure 14 does display the probabilities of obtaining large numbers or small numbers of rejections are both smaller than expected. From the observed CDF, it seems the distribution of the number of rejections in 3000 samples is under dispersed, i.e., the variance of the actual number of rejections is less than the variance of Binomial (3000, 0.05). To gain some understanding of why this happens, we consider some simple calculations. Let θ denote the true type I error rate over different scenarios in our simulation study. Further, consider θ to be a random variable and following a distribution with $E(\theta) = p$. Let $\mathbf{T}$ denote the number of rejections among 3000 samples for a scenario with type I error θ, that is, $\mathbf{T}|\theta \sim Binomial(3000, \theta)$. It follows that

$$\begin{aligned} Var(\mathbf{T}) &= E\{var(\mathbf{T}|\theta)\} + VarE(\mathbf{T}|\theta) \\ &= E(3000\theta - 3000\theta^2) + Var(3000\theta) \\ &= 3000p - 3000E(\theta^2) + 3000^2 Var(\theta) \\ &= 3000p(1 - 3000p) + 3000E(\theta^2)(3000 - 1)\,. \end{aligned} \tag{4.2}$$

We are interested if p, which is $E(\theta)$, can possibly be equal to the nominal type I error rate of 0.05 as desired. We observed in Figure 14 that under dispersed true distribution

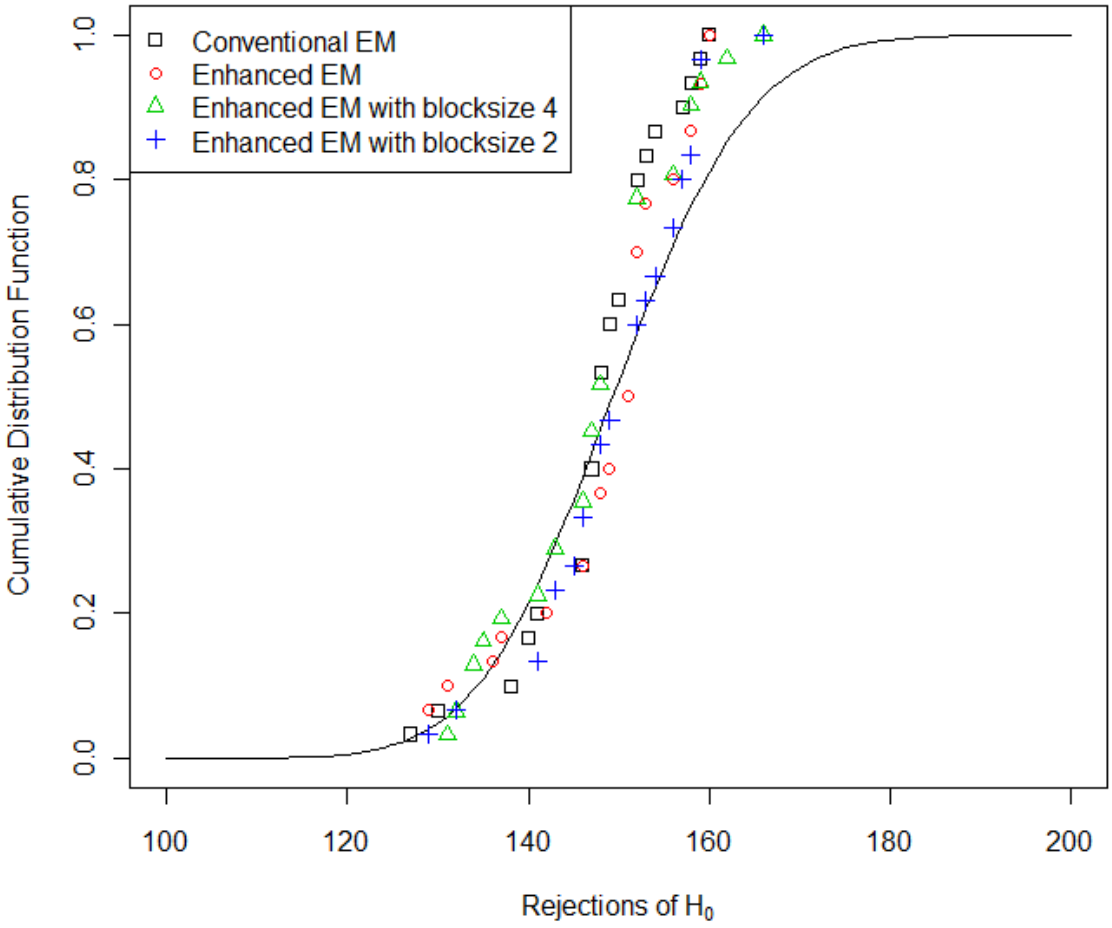


Figure 14: Observed and expected CDF of rejections of H_0 in 3000 simulations for each true value of σ under three sample size capping rules and four difference EM procedures.

of the number of rejections, suggesting that $Var(\mathbf{T}) < 3000 \cdot 0.05(1-0.05)$. Suppose that $E(\theta) = p = 0.05$, so that substituting (4.2) into the previous inequality with $p = 0.05$, we obtain

$$-3000^2(0.05)^2 + 3000(0.05)^2 + 3000E(\theta^2)(3000-1) < 0\,, \tag{4.3}$$

which reduces to

$$E(\theta^2) < (0.05)^2 = \{E(\theta)\}^2. \tag{4.4}$$

The later inequality cannot hold due to Jenson's inequality. Therefore, $E(\theta) < 0.05$. The question remains as to whether $var(\theta) > 0$ or $var(\theta) = 0$. The latter corresponds to $T \sim B(n,p)$ where $p < 0.05$, i.e., the true type I error rate is consistent across various scenarios with a value less than 0.05. To explore this possibility, we fit various binomial c.d.f's with $p < 0.05$ and none of them provided an adequate fit to the observed c.d.f. Our conclusion is that, in fact, $var(\theta) > 0$, and the different scenarios have different true type I error rates which on average are < 0.05.

The plot Gould and Shih[15] used in their paper (Figure 2) shows apparent close agreement in their simulations to the expected CDF curve. There could be multiple reasons why we did not get a graph similar to Gould and Shih[15]'s. First, we conduct a different simulation study from theirs. Gould and Shih used a selection of 6 different values of σ and 3 true mean difference values δ to generate the samples, as well 6 initially assumed values of σ to calculate the planning sample size, 2 sample size capping rules and 2 values for the interim analysis timings. Hence, their simulations were conducted under 432 cases. In our simulation study, under the condition $\delta = 0$, $\Delta = 0.443$ and $\tilde{\sigma} = 1$, we fixed the first stage sample size at 80 or 40, and we use 5 different values of σ to generate the samples. In order to make different EM procedures and different sample size capping rules comparable, we simulate a large proportion of the same data across the different EM procedures. Hence, for each value of σ, the number of rejections for different capping rules are correlated. Therefore, our simulation in essence was conducted under 10 independent cases for each EM algorithm. Second, as Waksman[35] pointed out, in each iteration of estimation Gould and Shih[15] altered their estimate of σ^2 in the M-step by subtracting 1 from the total interim sample size in the denominator. We, however, removed this alteration in our simulation study. Third, besides the sample size capping rules Gould and Shih[15] used, we also applied the restricted

and the unrestricted capping rule in the simulation study. All these reasons may explain why Gould and Shih[15]'s figure looks somewhat different than our Figure 14.

We used block sizes 2 and 4 as representative to illustrate the enhanced EM procedure with block design. The simulation results show the actual type I error rate is preserved at the nominal level 0.05 when the block sizes are 2 and 4. The enhanced EM procedure with bigger block sizes should improve the control of the type I error rate since it uses less available information to estimate σ. As a result we need not run more detailed simulations for larger block sizes.

4.2 EVALUATING THE EFFECT ON POWER

The EM procedure we use does not use any information of the treatment difference at the interim and does not estimate the absolute true treatment difference reliably. In other words, the pre-specified treatment difference used in the sample size calculation reflects the clinical benefits and does not necessarily need to be a good estimate of the true value. On the other hand, we try to estimate the nuisance parameter σ accurately from the interim data and use it to determine the appropriate adjusted sample size. The main purpose of our procedure is to compensate for the effects of σ's misidentification on the actual power and sample size. In this section, we want to evaluate the effect on power when σ is misspecified. We also look at how our procedure handles the power when the true treatment difference is misspecified, even this is not what our procedure designed for.

We briefly review the approaches to power and sample size evaluation that other researchers performed for blinded sample size re-estimation based on the nuisance parameter. Gould and Shih[15] used simulation studies to explore the effects of a range of parameter values on the likelihood of rejecting H_0 under the alternative hypothesis. Kieser and Friede[17] obtained the power by integrating the joint density of the test statistic components over the rejection region of the t-test under the alternative hypothesis.

4.2.1 Analytical calculation of the actual power and the expected sample size

Let us first consider an analytical approach to attempt to derive the power function. For large samples when we can assume σ is known, the unconditional power can be written as follows,

$$\begin{aligned}
\text{power} &= P\left(|\bar{y}_1 - \bar{y}_2| > z_{\alpha/2} \cdot \frac{2\sigma}{\sqrt{N_{\text{adj}}}}\right) \\
&= P\left(\bar{y}_1 - \bar{y}_2 > z_{\alpha/2} \cdot \frac{2\sigma}{\sqrt{N_{\text{adj}}}}\right) + P\left(\bar{y}_1 - \bar{y}_2 < -z_{\alpha/2} \cdot \frac{2\sigma}{\sqrt{N_{\text{adj}}}}\right) \\
&= \int\int_{\{(\bar{y}_1-\bar{y}_2, N_{adj}):\bar{y}_1-\bar{y}_2 > z_{\alpha/2}\cdot\frac{2\sigma}{\sqrt{N_{\text{adj}}}}\}} f(\bar{y}_1 - \bar{y}_2, N_{\text{adj}})\, d(\bar{y}_1 - \bar{y}_2)\, d(N_{\text{adj}}) \\
&\quad + \int\int_{\{(\bar{y}_1-\bar{y}_2, N_{adj}):\bar{y}_1-\bar{y}_2 < -z_{\alpha/2}\cdot\frac{2\sigma}{\sqrt{N_{\text{adj}}}}\}} f(\bar{y}_1 - \bar{y}_2, N_{\text{adj}})\, d(\bar{y}_1 - \bar{y}_2)\, d(N_{\text{adj}}) \\
&= \int_{\Re\{N_{\text{adj}}\}} P_{\bar{y}_1-\bar{y}_2|N_{adj}}(\bar{y}_1 - \bar{y}_2 > z_{\alpha/2} \cdot \frac{2\sigma}{\sqrt{N_{\text{adj}}}}) f(N_{\text{adj}})\, d(N_{\text{adj}}) \\
&\quad + \int_{\Re\{N_{\text{adj}}\}} P_{\bar{y}_1-\bar{y}_2|N_{\text{adj}}}(\bar{y}_1 - \bar{y}_2 < -z_{\alpha/2} \cdot \frac{2\sigma}{\sqrt{N_{\text{adj}}}}) f(N_{\text{adj}})\, d(N_{\text{adj}})\,,
\end{aligned} \tag{4.5}$$

where $f(\bar{y}_1 - \bar{y}_2, N_{\text{adj}})$ is the joint density function of random variables $\bar{y}_1 - \bar{y}_2$ and N_{adj}, and $f(N_{\text{adj}})$ is the marginal distribution of N_{adj}, where N_{adj} is the total final adjusted sample size for two treatment groups which is used in the test statistic. Note that, N_{adj} is a function of the estimate of the standard deviation at the interim, $\hat{\sigma}$, and it also depends on the sample size rules applied. Hence, the integration region for N_{adj}, $\Re\{N_{\text{adj}}\}$, changes when using different sample size rules. Thus, the sample size rules provide the corresponding integration region for $\hat{\sigma}$. To calculate the power, we need the conditional density function of $\bar{y}_1 - \bar{y}_2$ given N_{adj}. Since N_{adj} is just a function of $\hat{\sigma}$, it is easier to calculate (4.5) using $f(\bar{y}_1 - \bar{y}_2|\hat{\sigma})$. Using this conditional distribution, we have that (4.5)

$$\begin{aligned}
&\int_0^\infty P_{\bar{y}_1-\bar{y}_2|\hat{\sigma}}(\bar{y}_1 - \bar{y}_2 > z_{\alpha/2} \cdot \frac{2\sigma}{\sqrt{N_{\text{adj}}(\hat{\sigma})}}) f(\hat{\sigma})\, d(\hat{\sigma}) \\
&\quad + \int_0^\infty P_{\bar{y}_1-\bar{y}_2|\hat{\sigma}}(\bar{y}_1 - \bar{y}_2 < -z_{\alpha/2} \cdot \frac{2\sigma}{\sqrt{N_{\text{adj}}(\hat{\sigma})}}) f(\hat{\sigma})\, d(\hat{\sigma})\,,
\end{aligned} \tag{4.6}$$

where $f(\hat{\sigma})$ is the marginal density function of $\hat{\sigma}$

We know the sample mean difference can be broken into components as

$$\bar{y}_1 - \bar{y}_2 = \frac{N_1(\bar{y}_{11} - \bar{y}_{12}) + N_2(\hat{\sigma})(\bar{y}_{21} - \bar{y}_{22})}{N_1 + N_2(\hat{\sigma})}, \tag{4.7}$$

where the second stage sample size N_2 is a function of $\hat{\sigma}$. It is obvious that the observations from the second stage are independent of $\hat{\sigma}$, because $\hat{\sigma}$ is estimated from the first stage. Hence, given $\hat{\sigma}$, $\bar{y}_{21} - \bar{y}_{22}$ follows a normal distribution with mean $\mu_1 - \mu_2$ and variance $2\sigma^2/N_2(\hat{(\sigma)})$.

Clearly, if we want to focus on the first stage data and the treatment assignments were known, then we would get the estimator of the standard deviation as the pooled sample standard deviation, which is known to be independent of $\bar{y}_{11} - \bar{y}_{12}$. However, little theory appears to be known for the independence between the sample mean difference and the standard deviation estimate from an EM procedure. Our approach is to explore this dependence by simulation and compute sample correlations between the sample mean difference from the first stage study and the EM estimate of σ at the interim based on our simulations. We simulated 3000 samples with each having sample size of 80. Forty observations are from a population following $N(0, 1)$ and the other forty are from $N(1, 1)$. The scatterplots between the EM estimate of σ and the treatment mean difference for these 80 observations are plotted in Figure 15. As seen in Figure 15, the treatment difference appears uncorrelated with the enhanced EM estimator with block size 2 and 4. The two variables appear not quite independent when using the conventional EM procedure or enhanced EM procedure without block design. But ever then, the correlations between these two variables is quite small. Hence, when using the enhanced EM algorithm with block size 2 or 4 for the estimate of σ in a blinded adaptive design, we feel comfortable for the analytic calculations in assuming that $\hat{\sigma}$ and $\bar{y}_{11} - \bar{y}_{12}$ are independent random variables. Therefore, given $\hat{\sigma}$, $\bar{y}_{11} - \bar{y}_{12}$ can be assumed to be normally distributed with mean $\mu_1 - \mu_2$ and variance $2\sigma^2/N_1$.

Based on the proceeding, given $\hat{\sigma}$, the summation of the two components of the entire sample mean difference given in (4.7), $\bar{y}_1 - \bar{y}_2 = N_1 \cdot (\bar{y}_{11} - \bar{y}_{12})/N_{\text{adj}}(\hat{\sigma}) + N_2(\hat{\sigma}) \cdot (\bar{y}_{21} - \bar{y}_{22})/N_{\text{adj}}(\hat{\sigma})$, follows a normal distribution with mean $\mu_1 - \mu_2$ and variance $2\sigma^2/N_{\text{adj}}(\hat{\sigma})$. If we

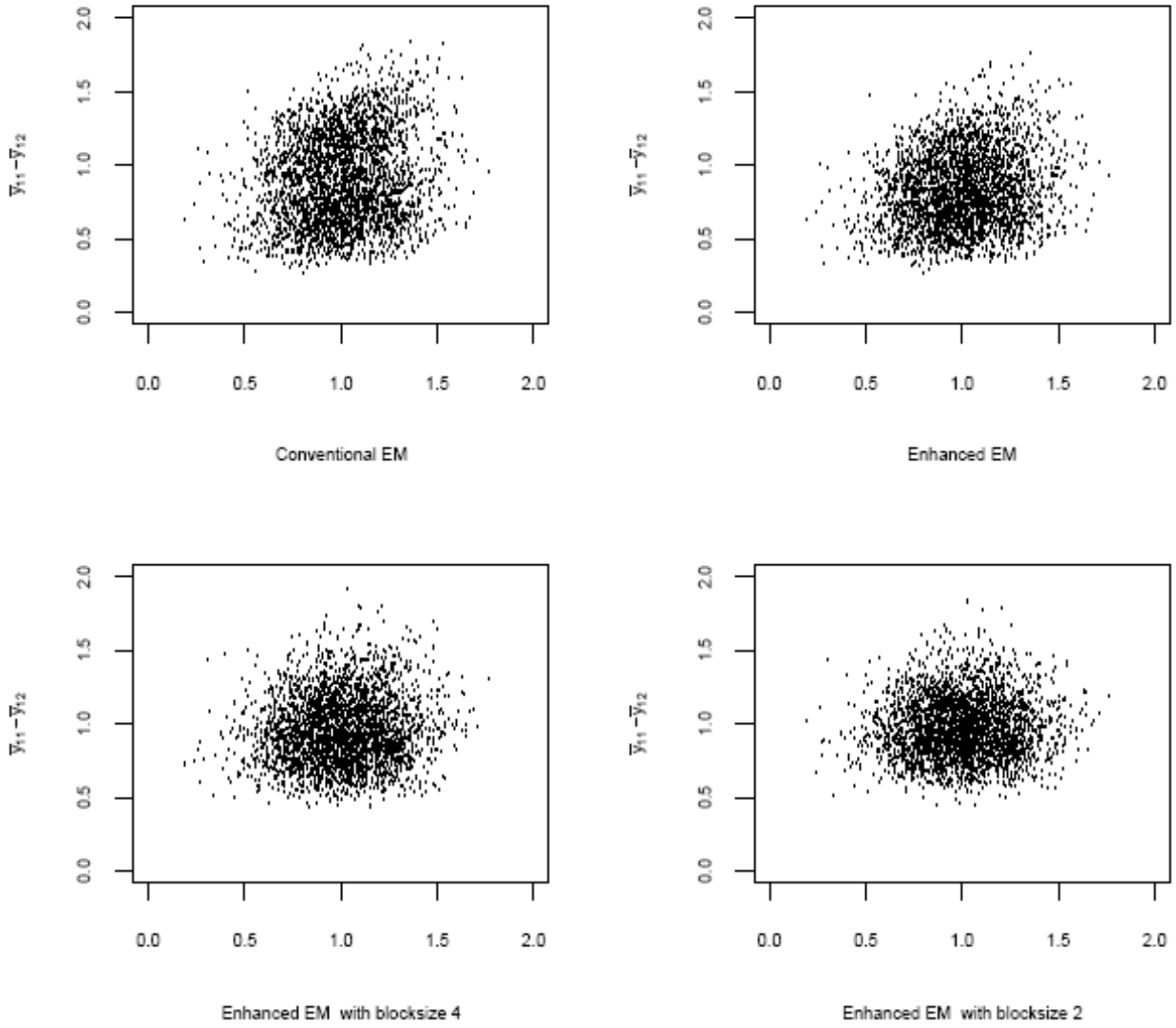


Figure 15: Scatterplots of the observed treatment difference at the interim versus the EM estimates at the interim ($N_1 = 80$). EM estimators are calculated based on 3000 simulated samples which are generated from N(0,1) and N(1,1).

standardize the distribution of $\bar{y}_1 - \bar{y}_2$, we have $P\left(\frac{\bar{y}_1-\bar{y}_2-\delta}{\sqrt{4\sigma^2/N_{\text{adj}}(\hat{\sigma})}} < \frac{-z_{\alpha/2}(2\sigma)/\sqrt{N_{\text{adj}}(\hat{\sigma})}-\delta}{\sqrt{4\sigma^2/N_{\text{adj}}(\hat{\sigma})}} \,\middle|\, \hat{\sigma}\right) \sim N(0,1)$. Hence, we can rewrite (4.6) as follows:

$$\text{power} = 1 - \int_0^\infty \Phi\left(z_{\alpha/2} - \frac{\delta}{\sqrt{4\sigma^2/N_{\text{adj}}(\hat{\sigma})}}\right) f(\hat{\sigma})\, d(\hat{\sigma}) + \int_0^\infty \Phi\left(-z_{\alpha/2} - \frac{\delta}{\sqrt{4\sigma^2/N_{\text{adj}}(\hat{\sigma})}}\right) f(\hat{\sigma})\, d(\hat{\sigma}), \tag{4.8}$$

where Φ is the CDF of the standard normal distribution. Since $\Phi\big(-z_{\alpha/2} - \frac{\delta}{\sqrt{4\sigma^2/N_{\text{adj}}(\hat{\sigma})}}\big)$ will be a very small number, we ignore it in the calculation of the power, so that

$$\text{power} \approx 1 - \int_0^\infty \Phi\left(z_{\alpha/2} - \frac{\delta}{\sqrt{4\sigma^2/N_{\text{adj}}(\hat{\sigma})}}\right) f(\hat{\sigma})\, d(\hat{\sigma}). \tag{4.9}$$

First, we evaluate the power for the case $N_1 = 80$ under the unrestricted sample size rule, i.e. $N_{\text{adj}} = \max(N_1, N')$. We still assume the desired treatment mean difference to achieve 80% power is set to 0.443 as stated in Section 4.1 and the initial sample size N is 160. By implementing the EM estimate $\hat{\sigma}$ in the sample size calculation formula (1.2), we obtain $N'(\hat{\sigma}) = 4\hat{\sigma}^2 \cdot (z_{0.025} + z_{0.2})^2/0.443^2 = 160 \cdot \hat{\sigma}^2$. The second part of (4.9) can be written as

$$\int_0^\infty \Phi\left(z_{\alpha/2} - \frac{\delta}{\sqrt{4\sigma^2}} \cdot \sqrt{\max(N_1, N'(\hat{\sigma}))}\right) f(\hat{\sigma}) d(\hat{\sigma}). \tag{4.10}$$

The final sample size $N_{\text{adj}}(\hat{\sigma})$ is a different function depends on two regions of $\hat{\sigma}$. So that if $\hat{\sigma}^2 \leq 1/2$ then $\max(N_1, N'(\hat{\sigma})) = N_1$, i.e., $N_{\text{adj}}(\hat{\sigma}) = 80$. If $\hat{\sigma}^2 > 1/2$, then $\max(N_1, N'(\hat{\sigma})) = N'(\hat{\sigma})$, i.e., $N_{\text{adj}}(\hat{\sigma}) = 160 \cdot \hat{\sigma}^2$.

Since there does not exist a simple close form for the EM estimator, $\hat{\sigma}$, one cannot explicitly obtain the distribution of the EM estimator. However, as seen in Figure 16, the distributions of the $\hat{\sigma}$'s from different enhanced EM procedures are all approximately normal distributed, especially for the case of small block size 2 or 4. On the other hand, the distribution of the conventional EM estimator seems to be a mixture of two normal distributions because it has boundary modes that we mentioned in Chapter 3. Nonetheless, since the two components of the conventional EM estimates are close to each other and it is difficult to approximate the mixture distribution, we approximate the distribution of the conventional EM estimates by a normal distribution. Thus, for different EM procedures, we simply approximate the distribution of $\hat{\sigma}$ by a normal distribution, denoted by $\hat{\sigma} \sim N(\mu^*, \sigma^{*2})$

where μ^* and σ^{*2} are the mean and variance calculated from 3000 simulated samples. To analytically do the numerical integration, it's reasonable to use $\mu^* + 4 \cdot \sigma^*$ as the upper limit of $\hat{\sigma}$ under the assumed normal distribution. When doing the numerical integration for the standard normal we use -4 as a reasonable lower limit for the standard normal variable. Therefore, we re-write (4.10) as follows:

$$\begin{aligned}
&\int_0^{\sqrt{1/2}} \Phi\left(z_{\alpha/2} - \frac{\delta \cdot \sqrt{80}}{\sqrt{4\sigma^2}}\right) f(\hat{\sigma}) d(\hat{\sigma}) + \int_{\sqrt{1/2}}^{\mu^*+4\cdot\sigma^*} \Phi\left(z_{\alpha/2} - \frac{\delta \cdot \sqrt{160}}{\sqrt{4\sigma^2}} \cdot \hat{\sigma}\right) f(\hat{\sigma}) d(\hat{\sigma}) \\
&= \int_0^{\sqrt{1/2}} \left(\int_{-4}^{z_{\alpha/2} - \frac{\delta\cdot\sqrt{80}}{\sqrt{4\sigma^2}}} \phi(z) d(z) \right) f(\hat{\sigma}) d(\hat{\sigma}) + \int_{\sqrt{1/2}}^{\mu^*+4\cdot\sigma^*} \left(\int_{-4}^{z_{\alpha/2} - \frac{\delta\cdot\sqrt{160}}{\sqrt{4\sigma^2}}\cdot\hat{\sigma}} \phi(z) d(z) \right) f(\hat{\sigma}) d(\hat{\sigma}) \,,
\end{aligned} \tag{4.11}$$

where $\phi(z)$ is the pdf of the standard normal distribution.

As an example to illustrate the numerical calculation steps, considering the setting when the true treatment difference is $\delta = 0.443$, the true standard deviation is $\sigma = 1$ and we use the enhanced EM procedure. Specifically, we apply the enhanced EM algorithm to estimate the common standard deviation, σ, for each of 3000 simulated random samples. We found that the mean of these 3000 enhanced estimators is 0.8443 and the sample variance is 0.0161 (We do not use the simulation results from Chapter 3 because for these calculations we want more accurate simulation results.), so that we can assume $\hat{\sigma} \sim N(0.8443, 0.0161)$. Then the reasonable upper limit of $\hat{\sigma}$ in the integration is $0.8443 + 4 \cdot \sqrt{0.0161} \approx 1.35$. Thus, (4.11) becomes to:

$$\int_0^{\sqrt{1/2}} \left(\int_{-4}^{1.96 - \frac{0.443}{2}\sqrt{80}} \phi(z) d(z) \right) f(\hat{\sigma}) d(\hat{\sigma}) + \int_{\sqrt{1/2}}^{1.35} \left(\int_{-4}^{1.96 - \frac{0.443}{2}\sqrt{160}\cdot\hat{\sigma}} \phi(z) d(z) \right) f(\hat{\sigma}) d(\hat{\sigma}) \,, \tag{4.12}$$

where $f(\hat{\sigma}) = \frac{1}{\sqrt{2\pi\cdot(0.0161)}} \exp\left(-\frac{(\hat{\sigma}-0.8443)^2}{2\cdot(0.0161)}\right)$. MATLAB is used to compute the numerical integration.

The analytical computation varies based on the different sample size capping rules used. The restricted sample size rule, $N_{\text{adj}} = \max(N, N')$, has two different forms on two different regions of $\hat{\sigma^2}$. Specifically, if $\hat{\sigma}^2 \leq 1$ then $\max(N_1, N'(\hat{\sigma})) = N$, i.e., $N_{\text{adj}}(\hat{\sigma}) = 160$. If $\hat{\sigma}^2 > 1$,

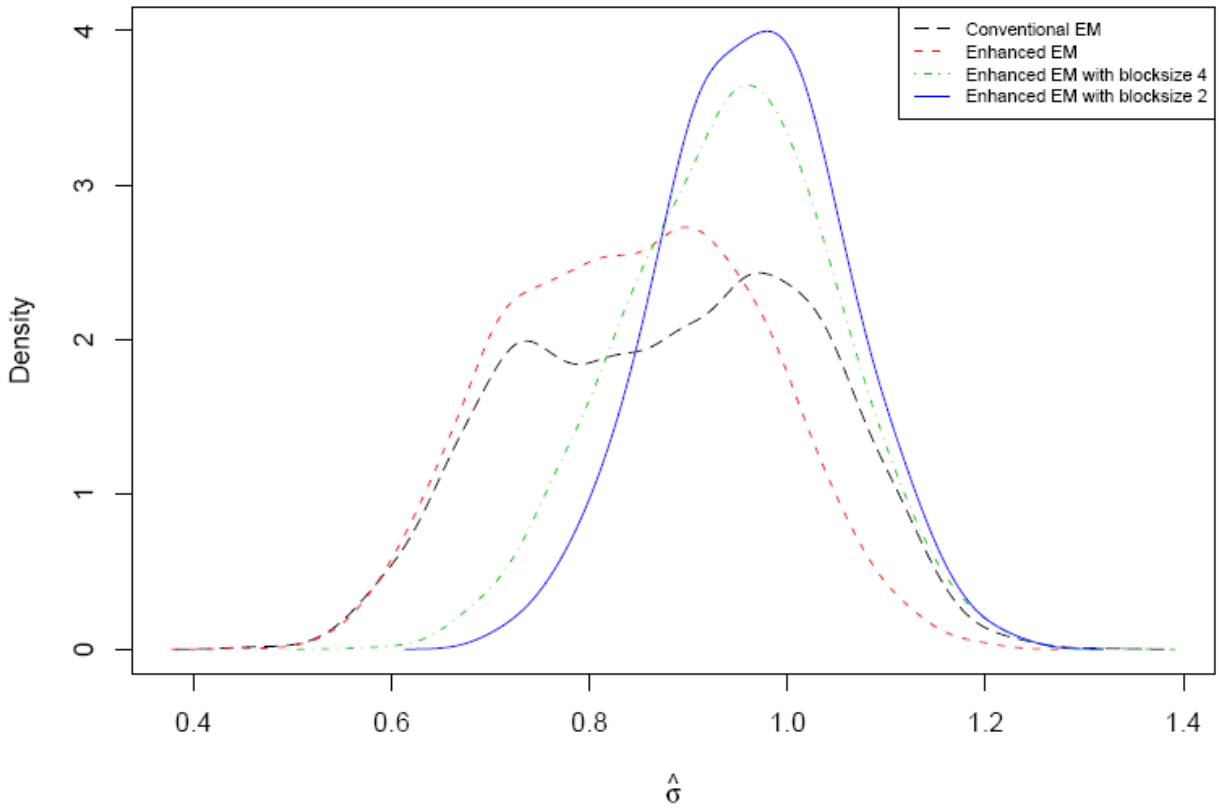


Figure 16: Gaussian kernel smoother of various types of EM estimates of σ when $\mu_1 = 0$, $\mu_2 = 0.443$ and $\sigma = 1$ for $N_1 = 80$ in 3000 simulation runs.

then $\max(N_1, N'(\hat{\sigma})) = N'(\hat{\sigma})$, i.e., $N_{\text{adj}}(\hat{\sigma}) = 160 \cdot \hat{\sigma}^2$. So that on two different regions of $\hat{\sigma}$ we obtain

$$\begin{aligned}
power &\approx 1 - \int_0^{+\infty} \Phi\left(z_{\alpha/2} - \frac{\delta}{\sqrt{4\sigma^2}} \cdot \sqrt{\max(N, N'(\hat{\sigma}))}\right) f(\hat{\sigma}) d(\hat{\sigma}) \\
&= 1 - \int_0^1 \Phi\left(z_{\alpha/2} - \frac{\delta \cdot \sqrt{160}}{\sqrt{4\sigma^2}}\right) f(\hat{\sigma}) d(\hat{\sigma}) - \int_1^{\mu^*+4\cdot\sigma^*} \Phi\left(z_{\alpha/2} - \frac{\delta \cdot \sqrt{160}}{\sqrt{4\sigma^2}} \cdot \hat{\sigma}\right) f(\hat{\sigma}) d(\hat{\sigma}) \\
&= 1 - \int_0^1 \left(\int_{-4}^{z_{\alpha/2} - \frac{\delta}{\sqrt{4\sigma^2}} \cdot \sqrt{160}} \phi(z) d(z) \right) f(\hat{\sigma}) d(\hat{\sigma}) - \int_1^{\mu^*+4\cdot\sigma^*} \left(\int_{-4}^{z_{\alpha/2} - \frac{\delta}{\sqrt{4\sigma^2}} \sqrt{160} \cdot \hat{\sigma}} \phi(z) d(z) \right) f(\hat{\sigma}) d(\hat{\sigma}) .
\end{aligned} \tag{4.13}$$

For the Gould and Shih[15]'s sample rule, N_{adj} is a function on three different regions. Specifically, if $\hat{\sigma}^2 \leq 1.33$ then $N_{\text{adj}}(\hat{\sigma}) = 160$, that is, no sample size adjustment and the initially planned sample size, N, is used for the study. If $1 < \hat{\sigma}^2 \leq 2$ then $N_{\text{adj}}(\hat{\sigma}) = N'(\hat{\sigma}) = 160 \cdot \hat{\sigma}^2$. If $\hat{\sigma}^2 > 2$ then $N_{\text{adj}}(\hat{\sigma}) = 320$, that is, twice of the initially planned sample size, $2N$, is used for the study. Hence,

$$\begin{aligned}
power &\approx 1 - \int_0^{\sqrt{1.33}} \Phi\left(z_{\alpha/2} - \frac{\delta \cdot \sqrt{160}}{\sqrt{4\sigma^2}}\right) f(\hat{\sigma}) d(\hat{\sigma}) - \int_{\sqrt{1.33}}^{\sqrt{2}} \Phi\left(z_{\alpha/2} - \frac{\delta \cdot \sqrt{160}}{\sqrt{4\sigma^2}} \cdot \hat{\sigma}\right) f(\hat{\sigma}) d(\hat{\sigma}) \\
&\quad - \int_{\sqrt{2}}^{\mu^*+4\cdot\sigma^*} \Phi\left(z_{\alpha/2} - \frac{\delta \cdot \sqrt{320}}{\sqrt{4\sigma^2}}\right) f(\hat{\sigma}) d(\hat{\sigma}) \\
&= 1 - \int_0^{\sqrt{1.33}} \left(\int_{-4}^{z_{\alpha/2} - \frac{\delta}{\sqrt{4\sigma^2}} \cdot \sqrt{160}} \phi(z) d(z) \right) f(\hat{\sigma}) d(\hat{\sigma}) - \int_{\sqrt{1.33}}^{\sqrt{2}} \left(\int_{-4}^{z_{\alpha/2} - \frac{\delta}{\sqrt{4\sigma^2}} \sqrt{160} \cdot \hat{\sigma}} \phi(z) d(z) \right) f(\hat{\sigma}) d(\hat{\sigma}) \\
&\quad - \int_{\sqrt{2}}^{\mu^*+4\cdot\sigma^*} \left(\int_{-4}^{z_{\alpha/2} - \frac{\delta}{\sqrt{4\sigma^2}} \sqrt{320}} \phi(z) d(z) \right) f(\hat{\sigma}) d(\hat{\sigma}) .
\end{aligned} \tag{4.14}$$

Therefore, for a specific design with various σ and δ, we can compute the unconditional power approximately by integration. When $N_1 = 40$, we use the same procedures to evaluate the actual unconditional power with a slight adjustment on the integration regions for the unrestricted design since the interim sample size changes from 80 to 40. Specifically, under the unrestricted rule, if $\hat{\sigma}^2 \leq 1/4$ then $N_{\text{adj}}(\hat{\sigma}) = \max(N_1, N'(\hat{\sigma})) = 40$ and if $\hat{\sigma}^2 > 1/4$, then $N_{\text{adj}}(\hat{\sigma}) = \max(N_1, N'(\hat{\sigma})) = 160 \cdot \hat{\sigma}^2$. For the restricted and Gould and Shih[15]'s sample size rule, the integration on $\hat{\sigma}$ keep the same format when $N_1 = 40$ as when $N_1 = 80$.

We can also obtain the expectation of the adjusted sample size in a similar way. Still taking $N_1 = 80$ as an example, in the unrestricted design, we have

$$\begin{aligned} E(N_{\text{adj}}) &= E\{\max(N_1, N'(\hat{\sigma}))\} \\ &= \int_0^{+\infty} \max(N_1, 160\hat{\sigma}^2) f(\hat{\sigma}) d(\hat{\sigma}) \\ &= \int_0^{\sqrt{1/2}} 80 \cdot f(\hat{\sigma}) d(\hat{\sigma}) + \int_{\sqrt{1/2}}^{\mu^*+4\cdot\sigma^*} 160\hat{\sigma}^2 \cdot f(\hat{\sigma}) d(\hat{\sigma}) \,. \end{aligned} \tag{4.15}$$

For the restricted sample size rule, the expected sample size for the same example can be expressed as:

$$\begin{aligned} E(N_{\text{adj}}) &= \int_0^{+\infty} \max(N, N'(\hat{\sigma})) f(\hat{\sigma}) d(\hat{\sigma}) \\ &= \int_0^{1} 160 \cdot f(\hat{\sigma}) d(\hat{\sigma}) + \int_{1}^{\mu^*+4\cdot\sigma^*} 160\hat{\sigma}^2 \cdot f(\hat{\sigma}) d(\hat{\sigma}) \,. \end{aligned} \tag{4.16}$$

For Gould and Shih[15]'s sample size rule, the expected sample size can be calculated as follows:

$$E(N_{\text{adj}}) = \int_0^{\sqrt{1.33}} 160 \cdot f(\hat{\sigma}) d(\hat{\sigma}) + \int_{\sqrt{1.33}}^{\sqrt{2}} 160\hat{\sigma}^2 \cdot f(\hat{\sigma}) d(\hat{\sigma}) + \int_{\sqrt{2}}^{\mu^*+4\cdot\sigma^*} 320 \cdot f(\hat{\sigma}) d(\hat{\sigma}) \,. \tag{4.17}$$

The analytical computation of the actual power and the expected sample size are shown in Table 9 and 10. We note again that these are only approximate results. First, we do not have independence between the sample treatment mean difference and the final adjusted sample size, especially when using the conventional and enhanced EM estimator without the block design. Also we assume that $\hat{\sigma}$ estimated from the interim follows a normal distribution. Moreover, as shown in Figure 16, the distributions of EM estimates of σ are slightly skewed, but we would expect to have a better approximation when the sample sizes are larger.

Table 9: Numerical integration results for the actual power and the expected sample size when $N_1 = 80$ true treatment difference, δ, are set to 0.35, 0.443, and 0.5; standard deviation σ are set to $1/\sqrt{2}$, 1, and $\sqrt{2}$ for each δ.

			Conventional EM	Enhanced EM	Enhanced EM with block size 4	Enhanced EM with block size 2
δ	σ	SS Capping rule				
0.35	$\frac{1}{\sqrt{2}}$	Unrestricted	0.6137 (84)	0.6056 (82)	0.6166 (84)	0.6183 (84)
		Restricted	0.8791 (160)	0.8791 (160)	0.8791 (160)	0.8791 (160)
		Gould-Shih's	0.8791 (160)	0.8791 (160)	0.8791 (160)	0.8791 (160)
	1	Unrestricted	0.4946 (128)	0.4648 (118)	0.5421 (142)	0.5630 (150)
		Restricted	0.6126 (166)	0.6049 (162)	0.6149 (168)	0.6166 (168)
		Gould-Shih's	0.6046 (164)	0.6011 (162)	0.6034 (162)	0.6028 (162)
	$\sqrt{2}$	Unrestricted	0.4787 (244)	0.4490 (224)	0.5388 (282)	0.5614 (296)
		Restricted	0.4876 (250)	0.4585 (230)	0.5393 (282)	0.5615 (296)
		Gould-Shih's	0.4650 (234)	0.4401 (218)	0.5188 (266)	0.5426 (282)
0.443	$\frac{1}{\sqrt{2}}$	**Unrestricted**	**0.8141 (84)**	**0.8063 (82)**	**0.8159 (84)**	**0.8177 (86)**
		Restricted	**0.9774 (160)**	**0.9774 (160)**	**0.9774 (160)**	**0.9774 (160)**
		Gould-Shih's	**0.9774 (160)**	**0.9774 (160)**	**0.9774 (160)**	**0.9774 (160)**
	1	**Unrestricted**	**0.6898 (130)**	**0.6588 (120)**	**0.7415 (144)**	**0.7638 (152)**
		Restricted	**0.8115 (166)**	**0.8047 (164)**	**0.8134 (168)**	**0.8153 (168)**
		Gould-Shih's	**0.8041 (164)**	**0.8009 (162)**	**0.8029 (162)**	**0.8025 (162)**
	$\sqrt{2}$	**Unrestricted**	**0.6658 (248)**	**0.6343 (226)**	**0.7340 (284)**	**0.7585 (298)**
		Restricted	**0.6774 (252)**	**0.6463 (230)**	**0.7349 (284)**	**0.7586 (298)**
		Gould-Shih's	**0.6532 (236)**	**0.6241 (220)**	**0.7137 (268)**	**0.7398 (282)**
0.5	$\frac{1}{\sqrt{2}}$	Unrestricted	0.8968 (86)	0.8907 (82)	0.8979 (86)	0.8988 (86)
		Restricted	0.9940 (160)	0.9940 (160)	0.9940 (160)	0.9940 (160)
		Gould-Shih's	0.9940 (160)	0.9940 (160)	0.9940 (160)	0.9940 (160)
	1	Unrestricted	0.7869 (132)	0.7603 (122)	0.8331 (146)	0.8514 (152)
		Restricted	0.8947 (168)	0.8895 (164)	0.8960 (168)	0.8970 (168)
		Gould-Shih's	0.8888 (164)	0.8863 (162)	0.8878 (164)	0.8874 (162)
	$\sqrt{2}$	Unrestricted	0.7623 (250)	0.7344 (228)	0.8262 (284)	0.8494 (298)
		Restricted	0.7751 (254)	0.7479 (234)	0.8270 (284)	0.8495 (300)
		Gould-Shih's	0.7527 (238)	0.7258 (222)	0.8090 (268)	0.8343 (284)

Table 10: Numerical integration results for the actual power and the expected sample size when $N_1 = 40$ true treatment difference, δ, are set to 0.35, 0.443, and 0.5; standard deviation σ are set to $1/\sqrt{2}$, 1, and $\sqrt{2}$ for each δ.

			Conventional EM	Enhanced EM	Enhanced EM with block size 4	Enhanced EM with block size 2
δ	σ	SS Capping rule				
0.35	$\frac{1}{\sqrt{2}}$	Unrestricted	0.4775 (62)	0.4445 (56)	0.5205 (68)	0.5469 (74)
		Restricted	0.8791 (160)	0.8791 (160)	0.8791 (160)	0.8791 (160)
		Gould-Shih's	0.8791 (160)	0.8791 (160)	0.8791 (160)	0.8791 (160)
	1	Unrestricted	0.4540 (116)	0.4196 (104)	0.5136 (134)	0.5421 (144)
		Restricted	0.6142 (168)	0.6048 (162)	0.6169 (168)	0.6205 (170)
		Gould-Shih's	0.6070 (164)	0.6015 (162)	0.6065 (164)	0.6073 (164)
	$\sqrt{2}$	Unrestricted	0.4369 (220)	0.4058 (198)	0.5079 (264)	0.5401 (284)
		Restricted	0.4604 (232)	0.4297 (212)	0.5128 (266)	0.5417 (286)
		Gould-Shih's	0.4392 (218)	0.4134 (202)	0.4873 (248)	0.5140 (264)
0.443	$\frac{\mathbf{1}}{\sqrt{\mathbf{2}}}$	**Unrestricted**	**0.6689 (64)**	**0.6361 (58)**	**0.7156 (70)**	**0.7394 (74)**
		Restricted	**0.9774 (160)**	**0.9774 (160)**	**0.9774 (160)**	**0.9774 (160)**
		Gould-Shih's	**0.9774 (160)**	**0.9774 (160)**	**0.9774 (160)**	**0.9774 (160)**
	1	**Unrestricted**	**0.6285 (116)**	**0.5951 (106)**	**0.6997 (134)**	**0.7358 (144)**
		Restricted	**0.8108 (166)**	**0.8039 (162)**	**0.8141 (168)**	**0.8179 (170)**
		Gould-Shih's	**0.8048 (164)**	**0.8011 (162)**	**0.8052 (164)**	**0.8062 (164)**
	$\sqrt{\mathbf{2}}$	**Unrestricted**	**0.6128 (224)**	**0.5805 (202)**	**0.6966 (266)**	**0.7309 (286)**
		Restricted	**0.6464 (236)**	**0.6148 (216)**	**0.7039 (268)**	**0.7333 (286)**
		Gould-Shih's	**0.6232 (220)**	**0.5947 (204)**	**0.6775 (248)**	**0.7066 (264)**
0.5	$\frac{1}{\sqrt{2}}$	Unrestricted	0.7670 (64)	0.7374 (58)	0.8084 (70)	0.8324 (74)
		Restricted	0.9940 (160)	0.9940 (160)	0.9940 (160)	0.9940 (160)
		Gould-Shih's	0.9940 (160)	0.9940 (160)	0.9940 (160)	0.9940 (160)
	1	Unrestricted	0.7224 (118)	0.6920 (106)	0.7950 (136)	0.8234 (144)
		Restricted	0.8935 (168)	0.8885 (164)	0.8962 (170)	0.8985 (170)
		Gould-Shih's	0.8889 (164)	0.8862 (162)	0.8893 (164)	0.8899 (166)
	$\sqrt{2}$	Unrestricted	0.7064 (226)	0.6763 (204)	0.7842 (262)	0.8174 (282)
		Restricted	0.7437 (238)	0.7145 (216)	0.7926 (266)	0.8202 (284)
		Gould-Shih's	0.7224 (222)	0.6943 (204)	0.7698 (246)	0.7983 (262)

4.2.2 Simulation study for actual power and expected sample size

4.2.2.1 Purpose of the Simulation Study While the approach of Section 4.1.1 provides an interesting approximation to power and expected sample sizes, simulation appears to be the only approach to accurately assess these quantities. In our simulation study, we aim to compare the actual power and the expected sample size among different EM procedures under various sample size capping rules over a range of σ's and treatment difference δ's. We would like to show through simulation, that our proposed blinded sample size adjustment procedure can maintain the desired power when σ is misspecified in the planning phase in a range of scenarios.

Another interest is the effects of the block sizes on the enhanced EM procedure. Recall that block size decreases, the enhanced EM estimates of σ improve and the value of $\hat{\sigma}$ increases since in general the EM procedures tend to underestimate σ. We show in our simulations that the enhanced EM algorithm with block design appears to better preserve the power.

4.2.2.2 Description of Simulation Study Our simulation results in Section 4.1.2.3 indicate that the actual type I errors are controlled at 0.05 when using the t-statistic in (4.1). Therefore, it is meaningful to compare the actual powers of our procedures with the planned power of 0.8. For the 0.05 level test, given a clinical meaningful treatment difference for Δ of 0.443, the initial sample size is calculated as 160. Two interim points are chosen at $N_1 = 40$ and $N_1 = 80$ to examine the effects of timing on the adaptive design. The true values of the common standard deviations are examined at $1/\sqrt{2}, 1$ and $\sqrt{2}$.

Under the alternative hypothesis, 3000 samples with sample size N_1 are generated from $N(0, \sigma)$ and $N(\delta, \sigma)$. Since we need to handle the randomization by blocks for the enhanced EM procedure, all data are generated in pairs, that is, for every two patients, one is from the experimental group and the other is from the control group. Equal number of patients in each treatment group are kept at both the interim and the end of the trial. For the enhanced EM procedure, we can use the full block information that block size is 2, or we can assume a larger block size in the design for the algorithm. For the conventional EM procedure, we

estimate σ from the same blocked data but without using this block information. Actually in Gould and Shih[15]'s simulation study, they keep the number of patients balanced at the interim, but obviously do not use this information in their algorithm.

Furthermore, the true treatment differences, δ, examined in our simulation study are set to $0.35, 0.443$ (which is equal to the assumed value for calculating initial sample size N), and 0.5. By generating data from distributions with smaller or bigger than assumed treatment difference, we can obtain the values of the actual power and the expected sample size under the underpowered or overpowered situations through simulation studies. Note that when $\delta = 0.35$, the design assumptions are incorrect and the study necessarily is under powered, and when $\delta = 0.5$, overpowered. The results for $\delta = 0.443$ are the ones which provide the most insight about the value of blinded sample size re-estimation.

For each scenario, we use four EM algorithms (conventional EM, enhanced EM, enhanced EM with block size 2, and enhanced with block size 4) to re-estimate σ and three sample size capping rules for the final adjusted sample size. Similar to the previous simulation studies for the actual type I error, we add more observations to each sample as necessary and conduct the t-test and count the number of rejections. Power is estimated by the proportion of samples which reject. And the mean number of final adjusted sample size N_{adj} estimates the expected sample size.

Also viewing each scenario as a fixed sample size design with planned sample size 160, for each combination of true δ and true σ, we calculate the actual power achieved as a reference guide. As designed when $\delta = 0.443$ and $\sigma = 1$, the power is the designed value of 0.8. Also, we calculate the sample size under fixed design when the treatment difference and the common standard deviation are both correctly presumed as a reference against which do compare to the expected sample size. Assign when $\delta = 0.443$ and $\sigma = 1$, the fixed sample size is 160.

4.2.2.3 Comparing analytical calculation with simulation results

Tables 11 and 13 show the simulation results of actual power from the 3000 simulated samples for the conventional EM procedure, the enhanced EM procedure and the enhanced EM procedures with block sizes 2 and 4. The last column in the tables gives the power for the fixed sample

Table 11: Simulation results for the actual power when $N_1 = 80$. True parameters are used to generate the sample are set as $\delta = 0.35$, 0.443, and 0.5; and $\sigma = 1/\sqrt{2}$, 1 and $\sqrt{2}$. 3000 samples are generated from each parameter configuration.

			Conventional EM	Enhanced EM	Enhanced EM with block size 4	Enhanced EM with block size 2	Fixed sample size $N = 160$
δ	σ	SS Capping rule					
0.35	$\frac{1}{\sqrt{2}}$	Unrestricted	0.5980	0.5893	0.5997	0.6000	0.88
		Restricted	0.8653	0.8653	0.8653	0.8653	
		Gould-Shih's	0.8653	0.8653	0.8653	0.8653	
	1	Unrestricted	0.4960	0.4647	0.5403	0.5637	0.60
		Restricted	0.6173	0.6147	0.6127	0.6207	
		Gould-Shih's	0.6093	0.6083	0.6087	0.6090	
	$\sqrt{2}$	Unrestricted	0.4950	0.4593	0.5453	0.5640	0.35
		Restricted	0.5003	0.4663	0.5453	0.5637	
		Gould-Shih's	0.4793	0.4500	0.5253	0.5480	
0.443	$\frac{\mathbf{1}}{\sqrt{\mathbf{2}}}$	**Unrestricted**	**0.8003**	**0.7917**	**0.8030**	**0.8080**	**0.98**
		Restricted	**0.9783**	**0.9783**	**0.9783**	**0.9783**	
		Gould-Shih's	**0.9783**	**0.9783**	**0.9783**	**0.9783**	
	1	**Unrestricted**	**0.6827**	**0.6587**	**0.7467**	**0.7713**	**0.80**
		Restricted	**0.7993**	**0.7947**	**0.8077**	**0.8087**	
		Gould-Shih's	**0.7993**	**0.7903**	**0.7930**	**0.7930**	
	$\sqrt{\mathbf{2}}$	**Unrestricted**	**0.6557**	**0.6307**	**0.7303**	**0.7577**	**0.51**
		Restricted	**0.6707**	**0.6410**	**0.7313**	**0.7577**	
		Gould-Shih's	**0.6507**	**0.6193**	**0.7110**	**0.7393**	
0.5	$\frac{1}{\sqrt{2}}$	Unrestricted	0.8863	0.8793	0.8903	0.8953	0.99
		Restricted	0.9933	0.9933	0.9933	0.9933	
		Gould-Shih's	0.9933	0.9933	0.9933	0.9933	
	1	Unrestricted	0.7803	0.7570	0.8270	0.8490	0.89
		Restricted	0.8900	0.8833	0.8937	0.8953	
		Gould-Shih's	0.8817	0.8810	0.8817	0.8813	
	$\sqrt{2}$	Unrestricted	0.7497	0.7277	0.8140	0.8453	0.61
		Restricted	0.7617	0.7393	0.8143	0.8453	
		Gould-Shih's	0.7420	0.7193	0.7983	0.8330	

Table 12: Simulation results for the means of the adjusted sample size when $N_1 = 80$. True parameters used to generate samples are set at $\delta = 0.35$, 0.443, and 0.5; and $\sigma = 1/\sqrt{2}$, 1 and $\sqrt{2}$. 3000 samples are generated from each parameter configuration.

			Conventional EM	Enhanced EM	Enhanced EM with block size 4	Enhanced EM with block size 2	Fixed sample size to achieve 80% power
δ	σ	SS Capping rule					
0.35	$\frac{1}{\sqrt{2}}$	Unrestricted	84	82	84	86	130
		Restricted	160	160	160	160	
		Gould-Shih's	160	160	160	160	
	1	Unrestricted	128	118	144	150	258
		Restricted	166	162	168	168	
		Gould-Shih's	162	160	162	162	
	$\sqrt{2}$	Unrestricted	246	226	282	298	514
		Restricted	250	230	282	298	
		Gould-Shih's	236	220	268	282	
0.443	$\frac{\mathbf{1}}{\sqrt{\mathbf{2}}}$	**Unrestricted**	**84**	**82**	**86**	**86**	**80**
		Restricted	**160**	**160**	**160**	**160**	
		Gould-Shih's	**160**	**160**	**160**	**160**	
	1	**Unrestricted**	**132**	**120**	**146**	**152**	**160**
		Restricted	**166**	**164**	**168**	**168**	
		Gould-Shih's	**166**	**162**	**162**	**162**	
	$\sqrt{\mathbf{2}}$	**Unrestricted**	**248**	**228**	**284**	**298**	**320**
		Restricted	**252**	**232**	**284**	**298**	
		Gould-Shih's	**238**	**222**	**268**	**282**	
0.5	$\frac{1}{\sqrt{2}}$	Unrestricted	86	84	86	86	64
		Restricted	160	160	160	160	
		Gould-Shih's	160	160	160	160	
	1	Unrestricted	134	122	146	152	126
		Restricted	168	164	168	170	
		Gould-Shih's	162	162	162	162	
	$\sqrt{2}$	Unrestricted	250	230	284	300	252
		Restricted	256	234	284	300	
		Gould-Shih's	240	224	268	284	

Table 13: Simulation results for the actual power when $N_1 = 40$. True parameters are used to generate the sample are set at $\delta = 0.35$, 0.443, and 0.5; and $\sigma = 1/\sqrt{2}$, 1, $\sqrt{2}$. 3000 samples are generated from each parameter configuration.

			Conventional EM	Enhanced EM	Enhanced EM with block size 4	Enhanced EM with block size 2	Fixed sample size $N = 160$
δ	σ	SS Capping rule					
0.35	$\frac{1}{\sqrt{2}}$	Unrestricted	0.4663	0.4360	0.5203	0.5483	0.88
		Restricted	0.8793	0.8793	0.8793	0.8793	
		Gould-Shih's	0.8793	0.8793	0.8793	0.8793	
	1	Unrestricted	0.4563	0.4233	0.5260	0.5560	0.60
		Restricted	0.6087	0.6000	0.6203	0.6217	
		Gould-Shih's	0.6007	0.5963	0.6053	0.6043	
	$\sqrt{2}$	Unrestricted	0.4360	0.4027	0.5200	0.5540	0.35
		Restricted	0.4620	0.4263	0.5217	0.5547	
		Gould-Shih's	0.4417	0.4097	0.4933	0.5170	
0.443	$\frac{\mathbf{1}}{\sqrt{\mathbf{2}}}$	**Unrestricted**	**0.6590**	**0.6213**	**0.7117**	**0.7433**	**0.98**
		Restricted	**0.9750**	**0.9750**	**0.9750**	**0.9750**	
		Gould-Shih's	**0.9750**	**0.9750**	**0.9750**	**0.9750**	
	1	**Unrestricted**	**0.6060**	**0.5843**	**0.6897**	**0.7277**	**0.80**
		Restricted	**0.7963**	**0.7933**	**0.8050**	**0.8050**	
		Gould-Shih's	**0.7853**	**0.7943**	**0.7990**	**0.7997**	
	$\sqrt{\mathbf{2}}$	**Unrestricted**	**0.6240**	**0.5900**	**0.7043**	**0.7440**	**0.51**
		Restricted	**0.6507**	**0.6153**	**0.7117**	**0.7460**	
		Gould-Shih's	**0.6337**	**0.5973**	**0.6897**	**0.7250**	
0.5	$\frac{1}{\sqrt{2}}$	Unrestricted	0.7520	0.7190	0.8060	0.8400	0.99
		Restricted	0.9937	0.9937	0.9937	0.9937	
		Gould-Shih's	0.9937	0.9937	0.9937	0.9937	
	1	Unrestricted	0.7097	0.6800	0.7910	0.8240	0.89
		Restricted	0.8867	0.8830	0.8927	0.8947	
		Gould-Shih's	0.8823	0.8793	0.8830	0.8837	
	$\sqrt{2}$	Unrestricted	0.7017	0.6767	0.7913	0.8240	0.61
		Restricted	0.7423	0.7157	0.8007	0.8223	
		Gould-Shih's	0.7240	0.6937	0.7747	0.8000	

Table 14: Simulation results for the means of the adjusted sample size when $N_1 = 40$. True parameters are used to generate the sample are set at $\delta = 0.35$, 0.443, and 0.5; and $\sigma = 1/\sqrt{2}$, 1, and $\sqrt{2}$. 3000 samples are generated from each parameter configuration.

			Conventional EM	Enhanced EM	Enhanced EM with block size 4	Enhanced EM with block size 2	Fixed sample size to achieve 80% power
δ	σ	SS Capping rule					
0.35	$\frac{1}{\sqrt{2}}$	Unrestricted	62	56	70	74	130
		Restricted	160	160	160	160	
		Gould-Shih's	160	160	160	160	
	1	Unrestricted	118	106	136	144	258
		Restricted	168	164	168	170	
		Gould-Shih's	164	162	164	164	
	$\sqrt{2}$	Unrestricted	222	200	264	286	514
		Restricted	234	212	268	286	
		Gould-Shih's	218	202	248	264	
0.443	$\frac{\mathbf{1}}{\sqrt{\mathbf{2}}}$	**Unrestricted**	**64**	**58**	**70**	**74**	**80**
		Restricted	**160**	**160**	**160**	**160**	
		Gould-Shih's	**160**	**160**	**160**	**160**	
	1	**Unrestricted**	**118**	**106**	**136**	**146**	**160**
		Restricted	**166**	**164**	**168**	**170**	
		Gould-Shih's	**164**	**162**	**164**	**166**	
	$\sqrt{\mathbf{2}}$	**Unrestricted**	**224**	**204**	**266**	**286**	**320**
		Restricted	**238**	**216**	**270**	**288**	
		Gould-Shih's	**222**	**206**	**250**	**264**	
0.5	$\frac{1}{\sqrt{2}}$	Unrestricted	66	60	72	76	64
		Restricted	160	160	160	160	
		Gould-Shih's	160	160	160	160	
	1	Unrestricted	118	106	138	146	126
		Restricted	168	164	170	172	
		Gould-Shih's	164	162	164	166	
	$\sqrt{2}$	Unrestricted	226	204	264	284	252
		Restricted	240	216	266	284	
		Gould-Shih's	224	206	248	262	

size design ($N = 160$) under different scenarios of δ and σ's. Similarly, Tables 12 and 14 show the mean adjusted sample sizes among 3000 simulations for each value of δ and σ.

When we compare the simulation results of this section with analytical calculations of Section 4.2.1, we find the actual powers and the expected sample sizes from both methods are very similar. This means our analytical method did a good job of approximating the power. In the simulation study, we calculated the adjusted sample size based on $\hat{\sigma}$ for each of the 3000 simulated samples, and we randomly generated additional $N_{\text{adj}} - N_1$ observations for each sample. Thus, even through there are two samples which have the same $\hat{\sigma}$ and we added the same number of patients to each of the two samples, it is not necessary that the actual power are the same for these two studies, i.e., both reject or accept the null hypothesis. In effect, given $\hat{\sigma}$, the simulation study estimates the conditional power instead of using the formula of the analytic study. In the analytical calculations, the distribution of $\hat{\sigma}$ is still based on the 3000 estimators, but given the estimate of σ from the interim, the conditional power is analytically calculated. Also, we believe the sample mean and standard deviation obtained from estimating σ for 3000 times will not change too much from estimating σ 10000 times. Therefore, the integration approach to power should always show similar results even with large simulation studies to estimate the distribution of $\hat{\sigma}$. Since the simulated actual power from 3000 samples are close to the analytical calculation of the actual power, we did not think it necessary to use a larger number of samples than 3000 to conduct the simulation study.

In the analytical calculation, we integrated the conditional power given $\hat{\sigma}$ while in the simulation studies we simulated the probability of rejecting the alternative hypothesis under each specific value of $\hat{\sigma}$. We note that for the case of using the enhanced EM algorithm with small block sizes, we can provide very good approximations to the simulated power and the estimated expected sample size. One might speculate that instead of using a normal approximation to the distribution of $\hat{\sigma}$, we might obtain more precise results by integrating over the kernel smoother (as shown in Figure 16).

4.2.2.4 Interpretation of the simulation results The chief purpose of the blinded sample size re-estimation is to mitigate the effect of false assumptions about σ on the power

of a trial. As we can see from Table 11 and 13 that, for both interim points $N_1 = 40$ and $N_1 = 80$, the enhanced EM procedures with appropriate small block sizes ensure that the study has better power properties than using the conventional EM procedure.

When the true mean difference is correctly assumed ($\delta = 0.443$) is the situation that we view as the most interesting. For both the interim points $N_1 = 40$ and $N_1 = 80$, when the true standard deviation $\sigma = 1$ and $\sigma = \sqrt{2}$, the enhanced EM procedures with small blocks sizes approach the planned power most closely among different EM procedures. Specifically, even the variance is underestimated as half of the true value in the planning phase of the study, the power can reach around 75% for both interim analysis sample sizes by using the enhanced EM procedure with block size 2. When $\sigma = 1/\sqrt{2}$, the advantage on power for the enhanced EM procedure with small block size is still obvious in the unrestricted capping rule. Specifically, even through when the variance is overestimated as twice the true value in the planning phase of the study, after the sample size adjustment using the enhanced EM procedure with block size 2, the power is adequate when $N_1 = 80$ and is 74.3% when $N_1 = 40$. In the restricted and Gould and Shih[15]'s capping rules for $\sigma = 1/\sqrt{2}$, all EM procedures overpower the study because these two capping rules require the adjusted sample size be bigger than the already abundant planned initial sample size.

We can also compare the expected sample sizes from different EM procedures as shown in Tables 12 and 14. Our enhanced EM procedure with small block sizes did what it is designed to do: it increases the sample size when the true standard deviation was greater than anticipated, and decreases the sample size when the opposite was true (This is most reflected in the restricted design). In Table 15, we use $N_1 = 80$ and $\delta = 0.443$ as an example (from Table 12) to compare the number of patients needed in the fixed design to achieve the same power as in the adaptive design. The expected sample size needed for the EM procedures is only slightly larger than that of the fixed design. For example, when $\sigma = 1$ and using the unrestricted capping rule for the enhanced EM procedure with block size 2, the mean adjusted sample size from 3000 simulations is 152. The corresponding actual power is 0.7713. In the fixed design without sample size adaption at the interim, assuming the true parameters $\delta = 0.443$ and $\sigma = 1$ are used in the fixed sample size calculation, then 150 patients are needed to achieve the same power of 0.7712. The two more patients, difference

Table 15: Adjusted versus fixed sample size for achieving the same power when $N_1 = 80$. $\delta = 0.443$ and $\sigma = 1/\sqrt{2}, 1$ and $\sqrt{2}$.

			Conventional EM	Enhanced EM	Enhanced EM with block size 4	Enhanced EM with block size 2
δ	σ	SS Capping rule	Adjusted Sample Size (Fixed)			
0.443	$\frac{1}{\sqrt{2}}$	Unrestricted	84 (82)	82 (80)	86 (82)	86 (82)
		Restricted	160 (162)	160 (162)	160 (162)	160 (162)
		Gould-Shih's	160 (162)	160 (162)	160 (162)	160 (162)
	1	Unrestricted	132 (122)	120 (116)	146 (142)	152 (150)
		Restricted	166 (160)	164 (158)	168 (164)	168 (164)
		Gould-Shih's	166 (160)	162 (158)	162(158)	162(158)
	$\sqrt{2}$	Unrestricted	248 (230)	228 (216)	284 (272)	298 (290)
		Restricted	252 (236)	232 (220)	284 (272)	298 (290)
		Gould-Shih's	238 (226)	222 (210)	268 (260)	282 (276)

between 152 and 150, that the enhanced EM procedure needed are the cost of using our enhanced adaptive design. Therefore, the adaptive design does not have too much of an expected penalty cost in comparison to using the fixed design, and has the obvious benefits. If we look at simulation results from Gould and Shih[15]'s conventional EM procedure, they need 132 patients on average to achieve power at 0.6827. With correct assumptions on σ and δ, 122 patients would be needed to achieve the same power in the fixed design. Therefore, not only do we get better power than Gould and Shih[15]'s, when our assumptions are wrong about σ, the cost of our design appropriately compared to the fixed design is less than Gould and Shih's.

Figure 17 shows the histograms of the adjusted sample size for different EM procedures when the treatment difference is correctly assumed, $N_1 = 80$ and the initial standard deviation is underestimated ($\sigma = \sqrt{2}$) and using the restricted adjusted sample size rule. Both adjusted sample sizes from the conventional EM and the enhanced EM procedures are skewed to the right, and there is a high frequency of sample sizes adjusted at 160 due to the capping rule. Therefore, without the restricted capping rule to force the adjusted sample size be at least 160, the conventional and the enhanced EM procedure would be even worse, i.e., more

underpowered, than the enhanced EM procedure with small block sizes. The distribution of the adjusted sample size under the enhanced EM produce with block size 2 is close to the true sample size 320 and very little skewed. This is because the adjusted sample size is a function of $\hat{\sigma}$, and $\hat{\sigma}$ from enhanced EM with 2 block procedure fits tightly around the true σ. Similarly, Figure 18 shows the histograms of the adjusted sample size by different EM procedures when $N_1 = 80$ and the treatment difference and the initial standard deviation are both assumed correctly and using the unrestricted sample size rule. For this scenario, we get the similar conclusion that a proportion of the estimates from the conventional and enhanced EM algorithm without block design make the study largely underpowered. The enhanced EM procedure with small block size can obtain a good estimate of the standard deviation which leads to a more accurate adjusted sample size. For other EM procedures, a big proportion of the estimates are underestimates. Even with the aid of the capping rules, the power of the study is still much lower than using the enhanced EM procedure with small blocks.

In this subsection's simulation study, we inspected the scenarios when there is a difference δ, i.e., there exists a meaningful difference between the experimental group and the control group. On the other hand, when $\delta = 0$ (or there is a neglectable small difference between treatments), the expected sample size, that is the cost of conducting the clinical trial when null hypothesis is true, from Gould and Shih's[15] procedure will be smaller than that of the enhanced EM procedure with small block sizes. The reason for this is because the Gould and Shih's[15] EM algorithm tends to underestimate the standard deviation. Hence, their procedure is less accurate for estimating σ which leads the adjusted sample size to be smaller. But our enhanced EM procedure still treats the treatment difference as the initially specified value and tries to improve the estimate of σ as close as possible to the true value, thereby increases the sample size.

We also inspect the case when the treatment difference is misspecified in our simulation study. The simulation results in Table 11-14 show that the quality of our enhanced EM procedure does not change based on the misspecification of δ. The enhanced EM procedure with small block sizes still tends to give relatively better power than other EM procedures especially for the case when δ is over estimated in the planning phase of the trial. The

enhanced EM procedure with block size 2 can attain the largest power among four EM procedures even though all EM procedures lead to an underpowered study due to the over-estimation of the treatment difference. Only when δ is underestimated (such as $\delta = 0.5$) and σ is underestimated (such as $\sigma = 1/\sqrt{2}$) in the planning phase, and the use of the restricted or Gould and Shih's[15]'s rule will inflate the actual power. This is because of the impact from the capping rules which do not down adjust the overpowered initial sample size. In this scenario, different EM procedures overpower to a comparable extent.

4.3 DISCUSSION

Due to the unknown distribution and the complicated form of the EM estimates of σ, we used simulation studies to investigate the properties of the actual type I error rate and the power and compared them among different EM estimates. From the simulation studies, we can conclude that after adjusting the sample size for the ongoing trial based on blinded sample size re-estimation, we can still use the standard t-test and that the type I error rate is preserved. Even if we used the enhanced EM algorithm with block size 2 where there is more information about the randomization schedule, the type I error rate will still be controlled at the nominal level. The type I error rates when using different EM procedures and different sample size capping rules are all quite similar.

When using the enhanced EM procedures, we need to pay particular attention to a couple of issues. One consideration is the information revealed on block size, i.e., the minimum unit for the treatment balance. Pharmaceutical companies should have well defined operational strategies to conduct these designs. To avoid revealing the randomized block sizes, the implementation of the enhanced EM procedure could be pre-programmed taking the results of the randomization code directly. Hence, the sample size adjustment procedure could be implemented while the block size is not revealed. Our simulation in this chapter showed this level of information about blocks does not comprise the type I error rate even with the block size equal to 2. Furthermore, we recommend re-estimating sample size only once and the implementation plan should be stated clearly in the protocol before the trial started. In

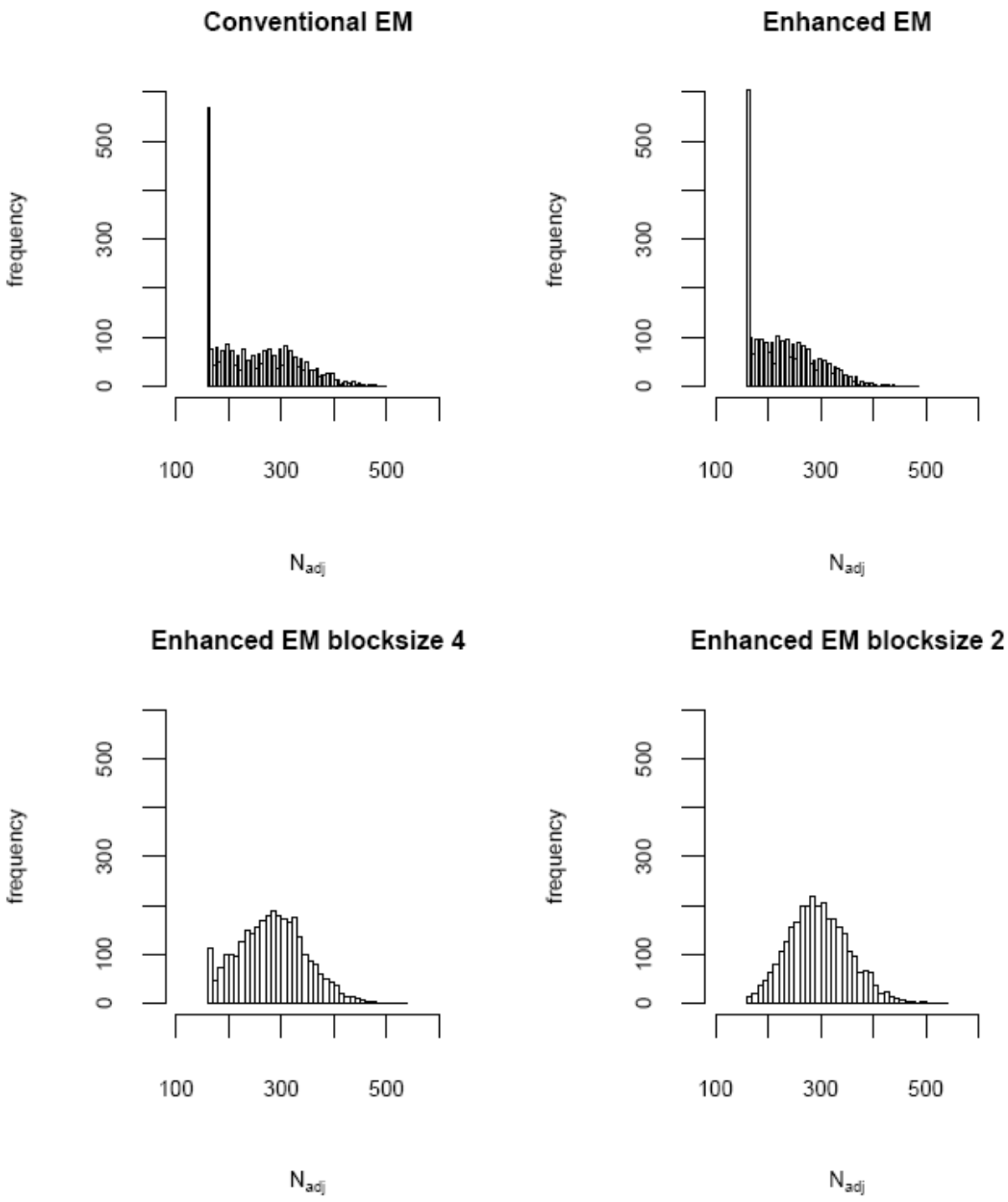


Figure 17: Histogram of adjusted sample sizes among 3000 simulations of different EM procedures in the restricted rule with $\delta = 0.443$ and $\sigma = \sqrt{2}$ when $N_1 = 80$.

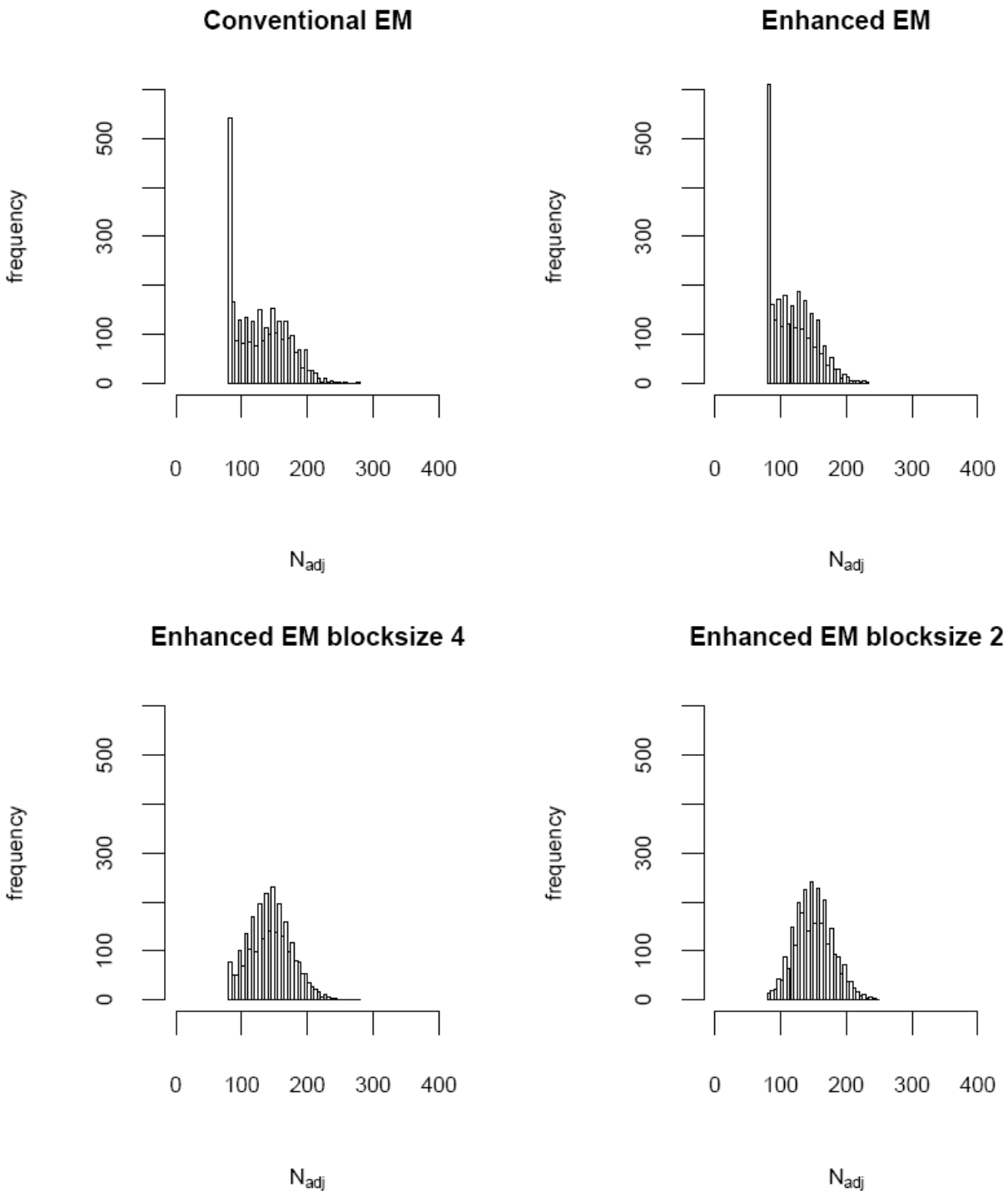


Figure 18: Histogram of adjusted sample sizes among 3000 simulations of different EM procedures in the unrestricted rule with $\delta = 0.443$ and $\sigma = 1$ when $N_1 = 80$.

our simulation study, we consider both a quarter and a half of the planned sample size as the interim point. The effect on the type I error rate is negligible but since the estimates from the halfway interim point are more precise than using the one quarter interim point, the former interim point produces a slightly improved power for the study.

5.0 EXTENSIONS TO MULTI-CENTER TRIALS

5.1 BACKGROUND

One concern in the single-center trial is the supply of patients. A single-center trial sometimes cannot recruit a sufficient number of patients within a required time period. In a multi-center research trial, a study is conducted simultaneously at more than one medical center or clinic following an agreed protocol. In other words, patients within each of the many centers are randomly assigned to one of the two treatments and the recruitment is accelerated. Clearly it is easier to recruit a large sample in a short-period of time for a multi-center trial than for a single-center trial. Another benefit for a multi-center trial is that patients from a variety of institutions can be studied so the study results can be generalized to a more broad population.

Like the single-center trial, the assumptions made about the standard deviation in the planning stage of the multi-center trial are usually uncertain, so that again there is doubt about the planned power. We want to estimate the standard deviation at the interim taking into account the sample size at the centers. Based on this estimate we want to adjust the sample size accordingly, so that the study design will be more efficient to detect the treatment differences. In this chapter, we will go through the details on re-estimating the sample size in multi-center trials using the EM procedures we developed in earlier sections. We again only consider blinded sample size re-estimation in multi-center trials.

5.1.1 Statistical model

For simplicity, our notation is for randomized two-arm multi-center trials with 2 centers. We discuss how to extend our technique to more centers in a later subsection. Here, we consider treating the participating centers as a fixed effect in the linear model. The following is the fixed effects model containing treatment group, center, and treatment-by-center interaction:

$$y_{ijk} = \tau_j + c_k + (\tau c)_{jk} + \varepsilon_{ijk}, \quad \varepsilon_{ijk} \sim N(0, \sigma) i.i.d, \tag{5.1}$$

where y_{ijk} denotes the primary endpoint from the ith patient, receiving the jth treatment in the kth center ($j, k = 1, 2$, and $i = 1, ..., n_{jk}$). We assume balanced randomization between two treatment arms within each center, $n_{1k} = n_{2k}$. The treatment effects τ_j and the center effects c_k are both fixed. The measurement errors for the ith patient in treatment j and center k, ε_{ijk}, are assumed to be independent, normally distributed with mean 0 and common standard deviation σ. Note that we have not included a 'ground mean' effect in our model.

5.1.2 Sample size re-estimation procedure

The power of the F-test for treatment effects involves the calculation of the non-central parameter of the F distribution. With a specified type I error, power, number of centers, and clinical assumptions of treatment means, the total sample size can be determined at the planning phase of the trial[18]. In the fixed sample size design, there is an alternative simple way to calculate the required total sample size for comparing two treatment groups. By assuming an additive model and equal numbers of patients at each center and within a center equal numbers of each treatment, we can use the large-sample approximation formula,

$$N = 4\sigma^2(z_{\alpha/2} + z_\beta)/\Delta^2\,, \tag{5.2}$$

where $\Delta = \tau_2 - \tau_1$ is the treatment difference meant to detect and N is the total sample size for the whole multi-center trial. We lose an extra 2 degrees of freedom when comparing to the two sample t-test in the single-center trial. But when the total sample size is large

relative to the number of centers, the effects of the loss of degrees of freedom is negligible. When treatment-by-center interaction exists, this sample size formula does not hold since the treatment sums of square divided by σ^2 and the error sums of square divided by σ^2 are no longer independent χ^2 distributions. But it has been shown that the adverse effect on the power of the clinical trials by incorrectly assuming treatment-by-center effect does not exist is very small [32]. So in the planning phase of the trial, it is plausible to calculate the initial sample size using (5.2) by assuming no interaction before the trial starts.

Suppose we use the normal approximation formula (5.2) to calculate the total sample size when $n_{1k} = n_{2k}, \forall\, k$. The approximate initial sample size N is calculated based on the the magnitude of the standard deviation assumed in advance as $\tilde{\sigma}$. We recruit a proportion of the initial sample size, e.g., $N/2$ for the internal pilot study, then estimate the standard deviation, $\hat{\sigma}$, based on the data we have from the patients who have already finished the trial. The treatment identities are kept blinded at the interim. The re-calculated sample size N' is based on $\hat{\sigma}$ by using the sample size formula (5.2) again.

5.1.3 Analytical method

Like in single-center trial, we apply different sample size capping rules within each center and recruit additional patients as N_{adj} suggested to complete the trial. At the end of the study, we analyze the trial as for randomized block designs. To test the equality of the two treatment effects, the statistical hypotheses are:

$$\mathbf{H_0} : \tau_1 = \tau_2 = 0 \;\; \textit{versus} \;\; \mathbf{H_1} : \tau_1 \neq \tau_2 \,. \tag{5.3}$$

The test statistic to be used is:

$$F = \frac{SST/1}{SSE/(N_{\text{adj}} - q - 1)} \,, \tag{5.4}$$

where SST and SSE are the sums of squares associated with the treatment effect and the residual error, respectively, and q is the appropriate degrees of freedom for either an additive model or one with interaction. When H_0 holds, F in (5.4) is assumed to be distributed as $F[1-\alpha; 1, (N_{\text{adj}} - q - 1)]$, where we again ignore the adaption.

5.2 BLINDED SAMPLE SIZE RE-ESTIMATION PROCEDURES IN MULTI-CENTER TRIALS

5.2.1 Blinded variance estimation methods in two center designs when treatment-by-center interaction does not exist

We assume there is no interaction at the interim, so an additive linear model is used when re-estimating σ. We start with the situation when information of block sizes is not considered in the sample size re-estimation procedures. In Gould-Shih[15]'s paper introducing the conventional EM procedure, they did not extend their work to multi-center trials. We show here how to implement both conventional and enhanced EM algorithm in estimating σ in a two center clinical trial.

Suppose at the end of the first stage study, there are total N_1 patients at the interim. For simplicity, we assume there are $N_1/2$ patients randomized to each center. Since the treatment identity j is blinded when we estimate the standard deviation at the interim, we use the notation $y_{i.k}$ to denote a primary endpoint from patient i in center k. That is, in center 1, primary endpoints are $y_{1.1}, y_{2.1}, ..., y_{\frac{N_1}{2}.1}$; and in center 2, primary endpoints are $y_{1.2}, y_{2.2}, ..., y_{\frac{N_1}{2}.2}$. If center 1 and center 2 have different sample sizes, this simple notation still applies. Let $z_{i.k}$ denote the treatment identities for ith patient in treatment j and center k. In center 1, when a patient is randomized to the control group, then $z_{i.1} = 1$ and $y_{i.1} \sim N(\tau_1 + c_1, \sigma)$; when a patient is randomized to the experimental group, then $z_{i.1} = 0$ and $y_{i.1} \sim N(\tau_2 + c_1, \sigma)$, where $i = 1, ..., N_1/2$. Similarly in center 2, when a patient is randomized to the control group, then $z_{i.2} = 1$ and $y_{i.2} \sim N(\tau_1 + c_2, \sigma)$; when a patient is randomized to the experimental group, then $z_{i.2} = 0$ and $y_{i.2} \sim N(\tau_2 + c_2, \sigma)$.

For the conventional EM algorithm, it is assumed the probability of each patients being assigned to each treatment group is 0.5 within both centers, i.e., $P(z_{i.k} = 1) = P(z_{i.k} = 0) =$

0.5 for $k = 1, 2$. Here the complete data likelihood function is given by

$$\begin{aligned} L(\boldsymbol{\theta}; \mathbf{y}..., \mathbf{z}_{..\mathbf{1}}, \mathbf{z}_{..\mathbf{2}}) &= \prod_{i=1}^{\frac{N_1}{2}} \left\{ f(y_{i.1}|z_{i.1}, \tau_1 + c_1, \tau_2 + c_1, \sigma) \times p(z_{i.1}|\tau_1 + c_1, \tau_2 + c_1, \sigma) \right\} \\ &\quad \cdot \prod_{i=1}^{\frac{N_1}{2}} \left\{ f(y_{i.2}|z_{i.2}, \tau_1 + c_2, \tau_2 + c_2, \sigma) \times p(z_{i.2}|\tau_1 + c_2, \tau_2 + c_2, \sigma) \right\} \\ &= \prod_{i=1}^{\frac{N_1}{2}} \Bigg\{ f(y_{i.1}|\tau_1 + c_1, \sigma)^{z_{i.1}} \cdot f(y_{i.1}|\tau_2 + c_1, \sigma)^{1-z_{i.1}} \cdot \frac{1}{2}^{z_{i.1}} (1 - \frac{1}{2})^{1-z_{i.1}} \\ &\quad \cdot f(y_{i.2}|\tau_1 + c_2, \sigma)^{z_{i.2}} \cdot f(y_{i.2}|\tau_2 + c_2, \sigma)^{1-z_{i.2}} \cdot \frac{1}{2}^{z_{i.2}} (1 - \frac{1}{2})^{1-z_{i.2}} \Bigg\}. \end{aligned} \tag{5.5}$$

To make the model identifiable, we assume without loss of generality that $c_1 = 0$. Then (5.5) is proportional to

$$\begin{aligned} \propto \prod_{i=1}^{\frac{N}{2}} \Bigg[\left(\frac{1}{\sigma}\right)^{z_{i.1}} \exp\left\{ -\frac{(y_{i.1} - \tau_1)^2}{2\sigma^2} z_{i.1} \right\} \cdot \left(\frac{1}{\sigma}\right)^{z_{i.1}} \exp\left\{ -\frac{(y_{i.1} - \tau_2)^2}{2\sigma^2} (1 - z_{i.1}) \right\} \\ \cdot \left(\frac{1}{\sigma}\right)^{z_{i.2}} \exp\left\{ -\frac{(y_{i.2} - \tau_1 - c_2)^2}{2\sigma^2} z_{i.2} \right\} \cdot \left(\frac{1}{\sigma}\right)^{z_{i.2}} \exp\left\{ -\frac{(y_{i.2} - \tau_2 - c_2)^2}{2\sigma^2} (1 - z_{i.2}) \right\} \Bigg]. \end{aligned} \tag{5.6}$$

The complete data log likelihood function can then be written as

$$\begin{aligned} \ell(\boldsymbol{\theta}|\mathbf{y}..., \mathbf{z}...) &= -\frac{N}{2} \log \sigma - \frac{1}{2\sigma^2} \sum_{i=1}^{\frac{N}{2}} z_{i.1} (y_{i.1} - \tau_1)^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{\frac{N}{2}} (1 - z_{i.1}) (y_{i.1} - \tau_2)^2 \\ &\quad - \frac{N}{2} \log \sigma - \frac{1}{2\sigma^2} \sum_{i=1}^{\frac{N}{2}} z_{i.2} (y_{i.2} - \tau_1 - c_2)^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{\frac{N}{2}} (1 - z_{i.2}) (y_{i.2} - \tau_2 - c_2)^2. \end{aligned} \tag{5.7}$$

The E-step computes the conditional expectation of the complete-data log-likelihood given the observed data, and the current parameter estimates, $\boldsymbol{\theta} = (\tau_1, \tau_2, c_2, \sigma)$, that is,

$$\begin{aligned} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) &= E\left[\ell(\boldsymbol{\theta}|\mathbf{y}..., \mathbf{z}...) | \mathbf{y}..., \boldsymbol{\theta}^{(t)} \right] \\ &= -N \log \sigma - \frac{1}{2\sigma^2} \sum_{i=1}^{\frac{N}{2}} \Big[(y_{i.1} - \tau_1)^2 E(z_{i.1}|\mathbf{y}_{..\mathbf{1}}, \boldsymbol{\theta}_{\mathbf{1}}^{(\mathbf{t})}) + (y_{i.1} - \tau_2)^2 \{1 - E(z_{i.1}|\mathbf{y}_{..\mathbf{1}}, \boldsymbol{\theta}_{\mathbf{1}}^{(\mathbf{t})})\} \\ &\quad + (y_{i.2} - \tau_1 - c_2)^2 E(z_{i.2}|\mathbf{y}_{..\mathbf{2}}, \boldsymbol{\theta}_{\mathbf{2}}^{(\mathbf{t})}) + (y_{i.2} - \tau_2 - c_2)^2 \{1 - E(z_{i.2}|\mathbf{y}_{..\mathbf{2}}, \boldsymbol{\theta}_{\mathbf{2}}^{(\mathbf{t})})\} \Big]. \end{aligned} \tag{5.8}$$

Because the conditional expectation of the complete data log likelihood is linear in z_{ij}, the E-step in (5.8) is reduced to computing the conditional expectations of the missing treatment identity $z_{i.k}$, i.e., $E(z_{i.k}|\mathbf{y}_{..\mathbf{k}}, \theta^{(t)})$. Specifically, the conditional expectation of $z_{i.k}$ can be written as:

$$
\begin{aligned}
E(z_{i.1}|\mathbf{y}_{..\mathbf{1}}, \theta_1^{(t)}) &= P(z_{i.1} = 1|\mathbf{y}_{..\mathbf{1}}, \theta_1^{(t)}) = \frac{f(y_{i.1}|\tau_1^{(t)}, \sigma^{(t)})}{f(y_{i.1}|\tau_1^{(t)}, \sigma^{(t)}) + f(y_{i.1}|\tau_2^{(t)}, \sigma^{(t)})} \\
E(z_{i.2}|\mathbf{y}_{..\mathbf{2}}, \theta_2^{(t)}) &= P(z_{i.2} = 1|\mathbf{y}_{..\mathbf{2}}, \theta_2^{(t)}) = \frac{f(y_{i.2}|\tau_1^{(t)} + c_2^{(t)}, \sigma^{(t)})}{f(y_{i.2}|\tau_1^{(t)} + c_2^{(t)}, \sigma^{(t)}) + f(y_{i.2}|\tau_2^{(t)} + c_2^{(t)}, \sigma^{(t)})} .
\end{aligned}
\tag{5.9}
$$

The M-step maximizes the conditional expectation of the complete-data log-likelihood computed in the E-step. Thus, we update the parameters with

$$
\begin{aligned}
\tau_1^{(t)} &= \frac{\frac{N_1}{2}\sum_{i=1}^{N_1/2} E(z_{i.1})y_{i.1} + \frac{N_1}{2}\sum_{i=1}^{N_1/2} E(z_{i.2})y_{i.2} - \sum_{i=1}^{N_1/2} E(z_{i.2})\sum_{i=1}^{N_1/2} y_{i.2} + \sum_{i=1}^{N_1/2} E(z_{i.2})\sum_{i=1}^{N_1/2} y_{i.1}}{\frac{N_1}{2}\sum_{i=1}^{N_1/2} E(z_{i.1}) + \frac{N_1}{2}\sum_{i=1}^{N_1/2} E(z_{i.2}) - \left\{\sum_{i=1}^{N_1/2} E(z_{i.2})\right\}^2 + \sum_{i=1}^{N_1/2} E(z_{i.1})\sum_{i=1}^{N_1/2} E(z_{i.2})} \\
\tau_2^{(t)} &= \frac{\sum_{i=1}^{N_1/2} y_{i.1} - \tau_1 \sum_{i=1}^{N_1/2} E(z_{i.1}|\mathbf{y}_{..\mathbf{1}}, \boldsymbol{\theta}_{\mathbf{1}}^{(\mathbf{t})})}{\sum_{i=1}^{N_1/2}\left\{1 - E(z_{i.1}|\mathbf{y}_{..\mathbf{1}}, \boldsymbol{\theta}_{\mathbf{1}}^{(\mathbf{t})})\right\}} \\
c_2^{(t)} &= \frac{\sum_{i=1}^{N_1/2} y_{i.2} - \tau_1 \sum_{i=1}^{N_1/2} E(z_{i.2}|\mathbf{y}_{..\mathbf{2}}, \boldsymbol{\theta}_{\mathbf{2}}^{(\mathbf{t})}) - \tau_2 \sum_{i=1}^{N/2}\left\{1 - E(z_{i.2}|\mathbf{y}_{..\mathbf{2}}, \boldsymbol{\theta}_{\mathbf{2}}^{(\mathbf{t})})\right\}}{N_1/2} \\
\sigma^{2(t+1)} &= \frac{1}{N_1}\sum_{i=1}^{N_1/2}\Big[E(z_{i.1}|\mathbf{y}_{..\mathbf{1}}, \boldsymbol{\theta}_{\mathbf{1}}^{(\mathbf{t})})(y_{i.1} - \tau_1^{(t)})^2 + \left\{1 - E(z_{i.1}|\mathbf{y}_{..\mathbf{1}}, \boldsymbol{\theta}_{\mathbf{1}}^{(\mathbf{t})})\right\}(y_{i.1} - \tau_2^{(t)})^2 \\
&\quad + E(z_{i.2}|\mathbf{y}_{..\mathbf{2}}, \boldsymbol{\theta}_{\mathbf{2}}^{(\mathbf{t})})(y_{i.2} - \tau_1^{(t)} - c_2^{(t)})^2 + \left\{1 - E(z_{i.2}|\mathbf{y}_{..\mathbf{2}}, \boldsymbol{\theta}_{\mathbf{2}}^{(\mathbf{t})})\right\}(y_{i.2} - \tau_2^{(t)} - c_2^{(t)})^2\Big] .
\end{aligned}
\tag{5.10}
$$

For the enhanced EM algorithm, we also treat the unobserved treatment identities as missing data. However, we assume that we conduct interim analysis at a balance point of numbers of patients between the two treatment groups. We utilize the additional observed information at the interim that equal numbers of patients, $N_1/4$, are randomized to either experimental or control treatment group within each center. Our complete data now include

the fact that $\sum_{i=1}^{N_1/2} z_{i.1} = N_1/4$ and $\sum_{i=1}^{N_1/2} z_{i.2} = N_1/4$, in addition to the primary endpoints $y_{1.1}, ..., y_{\frac{N_1}{2}.1}, y_{1.2}, ..., y_{\frac{N_1}{2}.2}$. The complete data likelihood function becomes

$$L(\boldsymbol{\theta}; \mathbf{y}_{...}, \mathbf{z}_{..\mathbf{1}}, \mathbf{z}_{..\mathbf{2}}, \sum_{i=1}^{N_1/2} z_{i.1}, \sum_{i=1}^{N_1/2} z_{i.2}) = \prod_{i=1}^{\frac{N}{2}} \Big\{ f(y_{i.1}|\tau_1 + c_1, \sigma)^{z_{i.1}} \cdot f(y_{i.1}|\tau_2 + c_1, \sigma)^{1-z_{i.1}} \cdot f(y_{i.2}|\tau_1 + c_2, \sigma)^{z_{i.2}} \cdot f(y_{i.2}|\tau_2 + c_2, \sigma)^{1-z_{i.2}} \cdot \frac{1}{\binom{N_1/2}{N_1/4}} \cdot \frac{1}{\binom{N_1/2}{N_1/4}} \Big\}. \tag{5.11}$$

We also set the restriction that $c_1 = 0$ in the enhanced EM algorithm. The M-step stays the same as (5.10), but the E-step is computed differently since we condition the treatment identity $z_{i.k}$ on more observed information. The E-step is computed by using the conditional Bernoulli distribution.

$$\begin{aligned} &\text{in center 1 :} \qquad E(z_{i.1}|\mathbf{y}_{..\mathbf{1}}, \sum_{i=1}^{N_1/2} z_{i.1}, {\boldsymbol{\theta}_{\mathbf{1}}}^{(t)}) = \frac{w_{i.1} R(\frac{N_1}{4} - 1, S \setminus \{i\})}{R(N_1/4, S)}, \\ &\text{and,} \\ &\text{in center 2 :} \qquad E(z_{i.2}|\mathbf{y}_{..\mathbf{2}}, \sum_{i=1}^{N_1/2} z_{i.2}, {\boldsymbol{\theta}_{\mathbf{2}}}^{(t)}) = \frac{w_{i.2} R(\frac{N_1}{4} - 1, S \setminus \{i\})}{R(N_1/4, S)}, \end{aligned} \tag{5.12}$$

where $w_{i.1} = f(y_{i.1}|\tau_1, \sigma)/f(y_{i.1}|\tau_2, \sigma)$, $w_{i.2} = f(y_{i.2}|\tau_1 + c_2, \sigma)/f(y_{i.2}|\tau_2 + c_2, \sigma)$ and $S = \{1, 2, ..., N_1/2\}$ for $i = 1, ..., N_1/2$. To guarantee the numerical stability of the R function, we use the R^* introduced in Section 2.4.2 in the computation.

We also note that the condition $\sum_{i=1}^{N_1/2} z_{i.1} = \sum_{i=1}^{N_1/2} z_{i.2} = N_1/4$ holds true in the enhanced EM algorithm. So taking the treatment identities in center 1 as an example, $z_{1.1}, z_{2.1}, ..., z_{\frac{N_1}{2}.1}$ given $\sum_{i=1}^{N_1/2} z_{i.1}$ follows conditional Bernoulli distribution. And we have:

$$\begin{aligned} \sum_{i=1}^{N_1/2} E(z_{i.1}|\sum_{i=1}^{N_1/2} z_{i.1}) &= E(z_{1.1}|\sum_{i=1}^{N_1/2} z_{i.1}) + ... + E(z_{\frac{N_1}{2}.1}|\sum_{i=1}^{N_1/2} z_{i.1}) \\ &= E(z_{1.1} + ... + z_{\frac{N_1}{2}.1}|\sum_{i=1}^{N_1/2} z_{i.1}) = \frac{N_1}{4}. \end{aligned} \tag{5.13}$$

Therefore, the enhanced EM estimates in the M-step can be simplified as:

$$\tau_1^{(t)} = \frac{2}{N_1}\Big(\sum_{i=1}^{N_1/2} E(z_{i.1}|\mathbf{y}_{..\mathbf{1}}, \boldsymbol{\theta}_{\mathbf{1}}^{(\mathbf{t})}, \sum_{i=1}^{N_1/2} z_{i.1}) y_{i.1} + \sum_{i=1}^{N_1/2} E(z_{i.2}|\mathbf{y}_{..\mathbf{2}}, \boldsymbol{\theta}_{\mathbf{2}}^{(\mathbf{t})}, \sum_{i=1}^{N_1/2} z_{i2}) y_{i.2} - \frac{1}{2}\sum_{i=1}^{N_1/2} y_{i.2} + \frac{1}{2}\sum_{i=1}^{N_1/2} y_{i.1}\Big)$$

$$\tau_2^{(t)} = \frac{2}{N_1}\Big(-\sum_{i=1}^{N_1/2} E(z_{i.1}|\mathbf{y}_{..\mathbf{1}}, \boldsymbol{\theta}_{\mathbf{1}}^{(\mathbf{t})}, \sum_{i=1}^{N_1/2} z_{i.1}) y_{i.1} - \sum_{i=1}^{N_1/2} E(z_{i.2}|\mathbf{y}_{..\mathbf{2}}, \boldsymbol{\theta}_{\mathbf{2}}^{(\mathbf{t})}, \sum_{i=1}^{N_1/2} z_{i.2}) y_{i.2} + \frac{1}{2}\sum_{i=1}^{N_1/2} y_{i.2} + \frac{3}{2}\sum_{i=1}^{N_1/2} y_{i.1}\Big)$$

$$c_2^{(t)} = \frac{2}{N_1}\Big(\sum_{i=1}^{N_1/2} y_{i.2} - \sum_{i=1}^{N_1/2} y_{i.1}\Big)$$

$$\begin{aligned}\sigma^{2(t+1)} = \frac{1}{N_1}\sum_{i=1}^{N_1/2} \Big[& E(z_{i.1}|\mathbf{y_1}, \boldsymbol{\theta}_{\mathbf{1}}^{(\mathbf{t})}, \sum_{i=1}^{N_1/2} z_{i.1})(y_{i.1} - \tau_1^{(t)})^2 + \big\{1 - E(z_{i.1}|\mathbf{y_1}, \boldsymbol{\theta}_{\mathbf{1}}^{(\mathbf{t})}, \sum_{i=1}^{N_1/2} z_{i.1})\big\}(y_{i.1} - \tau_2^{(t)})^2 \\ & + E(z_{i.2}|\mathbf{y}_{..\mathbf{2}}, \boldsymbol{\theta}_{\mathbf{2}}^{(\mathbf{t})}, \sum_{i=1}^{N_1/2} z_{i.2})(y_{i.2} - \tau_1^{(t)} - c_2^{(t)})^2 + \big\{1 - E(z_{i.2}|\mathbf{y}_{..\mathbf{2}}, \boldsymbol{\theta}_{\mathbf{2}}^{(\mathbf{t})}, \sum_{i=1}^{N_1/2} z_{i.2})\big\}(y_{i.2} - \tau_2^{(t)} - c_2^{(t)})^2 \Big].\end{aligned} \tag{5.14}$$

Observe that $c_2^{(t)}$ remain the same for iteration to iteration, i.e., $\hat{c}_2 = 2/N_1 \cdot (\sum_{i=1}^{N_1/2} y_{i.2} - \sum_{i=1}^{N_1/2} y_{i.1})$. If we have more block information at the interim, we can conduct the enhanced EM algorithm with block design as well as in the single center trial. We assume for simplest case that the block size is $N_1/4$ within each center, i.e., in both center 1 and 2, the numbers of patients are balanced in the first block and also balanced in the second block. Therefore, our observed information includes the summation of treatment identities for both the first and second $N_1/4$ observations in each center. The M-step keeps the same as in the enhanced EM algorithm. We use center 1 as an example to illustrate the changes in the E-step. We have

$$E\bigg(z_{i.1}|\mathbf{y}_{..\mathbf{1}}, \sum_{i=1}^{N_1/4} z_{i.1} = \frac{N_1}{8}, \sum_{i=(N_1/4)+1}^{N_1/2} z_{i.1} = \frac{N_1}{8}, \boldsymbol{\theta}_{\mathbf{1}}^{(\mathbf{t})}\bigg) = E\bigg(z_{i.1}|\mathbf{y}_{..\mathbf{1}}, \sum_{i=1}^{N_1/4} z_{i.1} = \frac{N_1}{8}\bigg), \tag{5.15}$$

when $i = 1, ..., N_1/4$; and

$$E\bigg(z_{i.1}|\mathbf{y}_{..\mathbf{1}}, \sum_{i=1}^{N_1/4} z_{i.1} = \frac{N_1}{8}, \sum_{i=(N_1/4)+1}^{N_1/2} z_{i.1} = \frac{N_1}{8}, \boldsymbol{\theta}_{\mathbf{1}}^{(\mathbf{t})}\bigg) = E\bigg(z_{i.1}|\mathbf{y}_{..\mathbf{1}}, \sum_{i=(N_1/4)+1}^{N_1/2} z_{i.1} = \frac{N_1}{8}\bigg), \tag{5.16}$$

when $i = N_1/4 + 1, ..., N_1/2$.

5.2.2 When assuming treatment-by-center interaction exists

When a clinical trial is conducted at more than one center, it is possible there exists a difference in treatment effects among different centers. If we assume the treatment-by-center interaction exists in the complete block randomization model, this interaction ideally would need to be considered when re-estimating σ at the interim.

Because we are working with blinded data, all that we can estimate at each center is the absolute values of the difference in treatment means. Therefore, it is impossible to separate a quantitative interaction from a qualitative interaction without further strong assumptions. In the case of a single center, the identifiability of the absolute values of the difference does not impact the estimate of σ^2. However, this is not true for center-by-treatment introduction.

In Appendix C, we provide an algorithm which under certain assumptions does estimate the parameters assuming a treatment-by-interaction, but further research is required to examine the effects of starting values on the EM algorithm. Our initial simulations suggest that, for example, if we suspect a quantitative interaction and choose starting values to reflect this, the estimation will be appropriate.

5.2.3 Enhanced EM procedure trials with more than two center: treatment-by-center interaction does not exist

When there are just two centers, we noted when assuming no interaction in the randomized block design model that the enhanced EM estimator of the center effect $\hat{c}_2$ given in (5.14) is a constant over iterations. Thus we can always estimate the center effect from the interim data without using the algorithm. This leads to our being able to estimate the other parameters using a simpler way. We illustrate this for a moment when there are just two centers. Specifically, we subtract $\hat{c}_2$ from all the observations in center 2, then pool these observations with the observations in center 1. Then the observed data can be treated as arising from one center. Hence, it is now clear that we can use the same enhanced EM algorithm to estimate the common standard deviation as we used for a single trial study. The only difference is that we need to use the enhanced EM algorithm with block size $N_1/2$ since each center is considered as a block and patients are balanced within each center.

We show in detail the calculation steps have no difference between the two-center estimates and the single-center "shortcut" estimates. Suppose the observations in center 1 are $y_{1.1}, ..., y_{\frac{N_1}{2}.1}$ and the observations in center 2 are $y_{1.2}, ..., y_{\frac{N_1}{2}.2}$. We take out the estimate of the center effect $\hat{c}_2$, which is a constant, from the observations in center 2. That is, the observations now in center 2 are $\mathbf{y'_{i.2}} = (y_{(1.2} - \hat{c}_2, ..., y_{\frac{N_1}{2}.2} - \hat{c}_2)$. Since center effect is taken out from center 2, we treat all the observations coming from one single center, and the first and second half of subjects are both balanced blocks of size $N_1/2$. We can use the enhanced EM algorithm with block size $N_1/2$ as introduced in Section 2.4.3.

Taking the estimation of σ as an example, at iteration t, we get the estimate of σ for the next iteration as

$$\sigma^{2(t+1)} = \frac{1}{N_1}\Bigg(\sum_{i=1}^{N_1/2} \Big[E(z_{i.1}|\mathbf{y}_{..\mathbf{1}}, \boldsymbol{\theta}^{(\mathbf{t})}, \sum_{i=1}^{N_1/2} z_{i.1})(y_{i.1} - \tau_1^{(t)})^2 + \big\{1 - E(z_{i.1}|\mathbf{y}_{..\mathbf{1}}, \boldsymbol{\theta}^{(\mathbf{t})}, \sum_{i=1}^{N_1/2} z_{i.1})\big\}(y_{i.1} - \tau_2^{(t)})^2 \Big]$$
$$+ \sum_{i=1}^{N_1/2} \Big[E(z_{i.2}|\mathbf{y'}_{..\mathbf{2}}, \boldsymbol{\theta}^{(\mathbf{t})}, \sum_{i=\frac{N_1}{2}+1}^{N_1} z_{i.2})(y'_{i.2})^2 + \big\{1 - E(z_{i.2}|\mathbf{y'}_{..\mathbf{2}}, \boldsymbol{\theta}^{(\mathbf{t})}, \sum_{i=\frac{N_1}{2}+1}^{N_1} z_{i.2})\big\}(y'_{i.2})^2 \Big] \Bigg). \tag{5.17}$$

In the second block, where the observations are from center 2, the conditional expectation of the missing identity, $E(z_{i.2}|\mathbf{y'}_{...}, \boldsymbol{\theta}^{(\mathbf{t})}, \sum_{i=1}^{N_1/2} z_{i.2})$, is a function of $w'_{i.2}$'s, and

$$w_{i.2}^{'(t)} = \frac{p_{i.2}^{'(t)}}{1 - p_{i.2}^{'(t)}} = \frac{f(y'_{i.2}|\tau_1^{(t)}, \sigma^{(t)})}{f(y'_{i.2}|\tau_2^{(t)}, \sigma^{(t)})}. \tag{5.18}$$

We compare (5.17) with the enhanced EM estimates of σ in two-center trial as shown in (5.14), i.e.,

$$\sigma^{2(t+1)} = \frac{1}{N_1} \sum_{i=1}^{N_1/2} \Big[E(z_{i.1}|\mathbf{y}_{..\mathbf{1}}, \boldsymbol{\theta}_{\mathbf{1}}^{(\mathbf{t})}, \sum_{i=1}^{N_1/2} z_{i.1})(y_{i.1} - \tau_1^{(t)})^2 + \big\{1 - E(z_{i.1}|\mathbf{y}_{..\mathbf{1}}, \boldsymbol{\theta}_{\mathbf{1}}^{(\mathbf{t})}, \sum_{i=1}^{N_1/2} z_{i.1})\big\}(y_{i.1} - \tau_2^{(t)})^2$$
$$+ E(z_{i.2}|\mathbf{y}_{..\mathbf{2}}, \boldsymbol{\theta}_{\mathbf{2}}^{(\mathbf{t})}, \sum_{i=\frac{N_1}{2}+1}^{N_1} z_{i.2})(y_{i.2} - \tau_1^{(t)} - c_2^{(t)})^2 + \big\{1 - E(z_{i.2}|\mathbf{y}_{..\mathbf{2}}, \boldsymbol{\theta}_{\mathbf{2}}^{(\mathbf{t})}, \sum_{i=\frac{N_1}{2}+1}^{N_1} z_{i.2})\big\}(y_{i.2} - \tau_2^{(t)} - c_2^{(t)})^2 \Big]. \tag{5.19}$$

In center 2, the conditional expectation is also a function of $w_{i.2}$'s, and it is denoted as:

$$w_{i.2} = \frac{p_{i.2}}{1 - p_{i.2}} = \frac{f(y_{i.2}|\tau_1^{(t)} + c_2^{(t)}, \sigma^{(t)})}{f(y_{i.2}|\tau_2^{(t)} + c_2^{(t)}, \sigma^{(t)})}, \tag{5.20}$$

which is equivalent to (5.18) since $c_2^{(t)}$ is a constant through iterations. Let $c_2^{(t)} = \hat{c}_2$, and $y'_{i.2} = y_{i.2} - \hat{c}_2$. Hence, $w_{i.2}$ and $w'_{i.2}$ are equivalent. Therefore, in (5.19) the estimate of σ in tth iteration in the same as in (5.17). We can use similar steps show that the estimation of τ_1 and τ_2 remain the same too. In this two center study, the estimates of treatment means and the standard deviation by assuming no interaction is equivalent to the estimates from using the single center enhanced EM estimation with block size $N_1/2$ for adjusted observations. If we want to use the full block size information in the two-center trial, i.e., more balance points within each center, we can adjust the block size of the single center's enhanced EM algorithm correspondingly.

Since we can simplify the estimation of σ in the enhanced EM algorithm, we can easily extend our estimation results if we have multiple centers (center size > 2). For example, when there are three centers involved in the study, we can estimate the center effects in center 2 and 3 as $\hat{c}_2 = 2/N_1(\sum_{i=1}^{N_1/2} y_{i.2} - \sum_{i=1}^{N_1/2} y_{i.1})$ and $\hat{c}_3 = 2/N_1(\sum_{i=1}^{N_1/2} y_{i.3} - \sum_{i=1}^{N_1/2} y_{i.1})$ separately. Then, obtain the new observations in center 2 as $y'_{i.2} = (y_{i.2} - \hat{c}_2$ and in center 3 as $y'_{i.2} = y_{i.3} - \hat{c}_3)$. We can still use the enhanced EM procedure for the single center study to solve the estimates. The only difference is this time we will use enhanced EM algorithm with three blocks.

5.3 SIMULATION STUDIES FOR A TWO CENTER TRIAL ASSUMING NO CENTER-TREATMENT INTERACTION

We conducted a very limited simulation study to investigate the estimates from our EM procedures in multi-center trials. In the simulation study, we compared two treatment groups in a two-center trial. For simplicity, we only consider the situation that the center sizes are equal in the two centers. Suppose we planned to test 160 patients before the trial starts and conduct the interim analysis when there are 40 patients in each center that have already completed the study.

In Table 16, we demonstrate the comparison between different EM estimators for two chosen center effects, $c_2 = 0.1$ and $c_2 = 0.5$. We are interested to know if smaller or bigger

center effects will have an impact on the re-estimation of the standard deviation. Meantime, the true treatment differences in the two centers are both set to 0.5 since we want to assume no center-treatment interaction. In center 1, 1000 samples with each having sample size 40 are generated from $N(0,1)$ and $N(0.5,1)$; in center 2 another 1000 samples with each having sample size 40 are generated from $N(c_2,1)$ and $N(0.5+c_2,1)$. Thus there are totally 1000 datasets consisting of observations from both centers.

We observe from the simulation results in Table 16 that the estimators from the enhanced EM algorithm have a larger bias and smaller variance than the estimates from the conventional EM algorithm. However, the enhanced EM estimators with block sizes are greatly improved with a much smaller bias when the block size is small (block size is 4 in our simulation). These are similar conclusions as in the single center trial case. Hence, everything we learned in earlier chapters from single trials can apply to multi-center trial as well.

As we can see from Table 16, the EM estimates do not seem to vary much depending on whether or not the center effect is equal to 0.1 or 0.5. Therefore, we believe that the properties of our estimates from EM procedures do not depend on the value of center effects.

We do note that the results for the enhanced EM estimates of Table 16 do not coincide exactly with those from Table 6 when block size is 40. The reason for this is because we need to estimate the center effect c_2. The shortcut single-center enhanced EM procedure described in Section 5.2.3 uses observations in center 1 as one block and observations in center 2 after subtracting the estimate of c_2 as the other block. The estimate of c_2 is estimated as the the sample mean difference of the observations between two centers, so that it varies from sample to sample. Hence, after subtracting the estimate of c_2 from observations in center 2, the observations in center 2 are not exactly distributed as a mixture normal of $N(\tau_1,\sigma)$ and $N(\tau_2,\sigma)$. So the estimation results are different from when we use enhanced EM procedure in single center trial with data generated from $N(\tau_1,\sigma)$ and $N(\tau_2,\sigma)$. To be clear, if there is no center effects in the data, i.e., $c_2 = 0$, and we set $c_2 = 0$ at every iteration of the multi-center enhanced EM algorithm, the estimation results from the two-center trial would be identical to those from the enhanced EM algorithm in the single-center trial with two blocks.

Table 16: Comparisons of EM estimates when $N_1 = 80$ in a two-center trial. True parameters used to generate samples are set as $\sigma = 1$, $\tau_1 = 0$, $\delta = 0.5$, and $c_2 = 0.1$ and 0.5. 1000 sample are generated from each parameters configuration.

		Enhanced EM estimates				Conventional EM estimates			
c_2		τ_1	τ_2	c_2	σ	τ_1	τ_2	c_2	σ
0.1	Bias	-0.3158	0.3168	-0.0073	-0.1804	-0.1445	0.1643	-0.0239	-0.1279
	Variance	0.0500	0.0499	0.0504	0.0154	0.1108	0.1379	0.0577	0.0249
	MSE	0.1497	0.1502	0.0504	0.0479	0.1316	0.1648	0.0583	0.0413
		Enhanced EM with block size 4							
	Bias	-0.0839	0.0850	-0.0073	-0.0770				
	Variance	0.0827	0.0806	0.0504	0.0124				
	MSE	0.0896	0.0877	0.0504	0.0183				
0.5	Bias	-0.3391	0.3249	0.0019	-0.1876	-0.1804	0.1883	-0.0182	-0.1384
	Variance	0.0507	0.0478	0.0497	0.0147	0.1149	0.1370	0.0601	0.0246
	MSE	0.1656	0.1533	0.0496	0.0499	0.1474	0.1723	0.0604	0.0437
		Enhanced EM with block size 4							
	Bias	-0.0881	0.0740	0.0019	-0.0760				
	Variance	0.0847	0.0868	0.0497	0.0137				
	MSE	0.0924	0.0922	0.0496	0.0195				

6.0 CONCLUSIONS AND FUTURE WORK

6.1 CONCLUSIONS

In this dissertation, our research concentrates on sample size re-estimation without breaking the blind in adaptive clinical trials. With normally distributed primary endpoints, we adjust the sample size for the ongoing trial based on the re-estimation of the standard deviation.

Gould and Shih[15] used the information that the probability of each subject assigned to treatment or control group is 0.5, so that based on a mixture distribution for the N_1 subjects, the EM algorithm can be used to obtain the MLE of the standard deviation. With this assumption, Gould and Shih[15] obviously treated the treatment identities as independent Bernoulli random variables, so there is no assumption that the numbers of subjects within each treatment group are equal at the interim stage. In practice, however, clinicians often use block randomization designs in clinical trial and as a result the numbers of subjects within each treatment group are equal at certain interim points in the study. We use this additional information to obtain more accurate MLE's of the standard deviation. This use of additional information requires us to change the EM algorithm used by Gould and Shih[15].

For similar adaptive designs, the typical approach at study end is to use the standard t-statistic to compare the two treatments ignoring the sample size re-estimation. Hence, this ignores the fact that the final t-statistic does not truly follow the t-distribution under the null hypothesis. However, we are able to show that with our new adaptive design which makes use of the block-randomization details, there is no inflation in the type I error using the usual t approach.

In Chapter 2.0, we give details of Gould and Shih[15]'s EM algorithm. Then we propose how to modify this EM algorithm when the information of equal numbers of subjects at

the interim is available. Since this means we also observe the summation of the missing treatment identities at the interim, the joint density function of missing treatment identities are not independent, and this joint density follows a conditional Bernoulli distribution. We obtain the conditional marginal density function of the treatment identities in the E-step of the EM algorithm. One of the challenges of the computation in this enhanced EM algorithm is the numerical instability in the mixture distributions setting. We develop a new recursive function in order to solve this problem. From a clinical trials perspective, the enhanced EM algorithm with block design is a practical application since small blocks are frequently used. Therefore, we further modify the E-step in our enhanced EM algorithm when we have the information of block sizes and show how this additional block information enters the enhanced EM algorithm.

In Chapter 3.0, a simple example is presented to illustrate the properties of the two EM algorithms. Then we refine Waksman[35]'s result and show that for certain settings the conventional EM estimates depend on the starting values for the conventional EM algorithm. On the other hand our enhanced EM algorithm shows little impact due to the starting values chosen and also shows a nice property of converging to interior estimates. We also investigate, using simulation, the reason why the conventional EM estimates depend on the starting values, and why this is not the case for the enhanced EM estimates. Through more general simulation studies with different parameter combinations, we compare the estimates of the two EM algorithms. We also simulate and compare the enhanced EM estimates when using different block sizes. As the block size decreases, the accuracy of enhanced EM estimation improves, while the conventional EM algorithm cannot utilize the block information. Especially when the block size is small, which is the common case in clinical trials, the bias and variance of the enhanced EM estimator is much smaller than that of the conventional EM estimator.

In Chapter 4.0, we first evaluate the actual type I error rate when using the standard t-test at the end of the trial through a simulation study. Different scenarios are considered including data generated from various values of the true standard deviation and different sample size capping rules. The simulation results show that the type I error rates from the different EM procedures are all controlled at the nominal level. Then, we analytically

compute and simulate the actual power and the expected sample size. The analytical results for power and expected sample size are quite similar to the simulation results and both show that the enhanced EM procedure with block design has a nice power property and adjusts the final sample size to a more appropriate size with a smaller penalty cost.

In Chapter 5.0, we extend the EM procedures to the setting of multi-center trials. In addition to the treatment effect, we also consider the study center as a blocking effect in the sample size re-estimation procedure. We develop the detailed steps for estimating the standard deviation at the interim when assuming for the primary endpoints the treatment-by-center interaction does not exist. We also perform a simulation study and show similar comparative performances of the various EM estimators to the single-center trial case.

6.2 FUTURE WORK

6.2.1 Kieser and Friede's simple procedure for blinded sample size re-estimation

As we mentioned in Chapter 1.0, Kieser and Friede[17] proposed using simple blinded variance estimators for normally distributed data's sample size recalculation. They presented two methods, one using an adjusted and the other using an unadjusted one sample variance based on the pooled interim data and ignoring the fact that observations at the interim are from two treatment groups.

The unadjusted one sample variance S^2_{unadj} is defined as follows:

$$S^2_{unadj} = \frac{1}{N_1 - 1} \sum_{i=1}^{N_1} (y_i - \bar{y})^2 \,, \tag{6.1}$$

where $\bar{y}$ is the grand mean of the interim data. We know S^2_{unadj} is a biased estimator of σ^2 when $\mu_1 \neq \mu_2$. Decomposition of the sum of squares in (6.1) becomes:

$$\begin{aligned} \sum_{i=1}^{N_1} (y_i - \bar{y})^2 &= N_1(\bar{y}_1 - \bar{y}_2)^2/4 + \sum_{i \in \text{group1}} (y_i - \bar{y}_1)^2 + \sum_{i \in \text{group2}} (y_i - \bar{y}_2)^2 \\ &= N_1 \hat{\delta}/4 + (N_1 - 2)s^2 \end{aligned} \tag{6.2}$$

where $\hat{\delta}$ is the unobserved interim treatment effect estimate and s^2 is the unobserved two sample variance. Based on the blinded data, the one sample variance estimator can be adjusted by the bias under the alternative hypothesis that the assumed treatment difference is Δ:

$$S^2_{adj} = \frac{(N_1 - 1)S^2_{unadj} - N_1\Delta^2/4}{N_1 - 2}\,. \tag{6.3}$$

Kieser and Friede[17] applied both the restricted and unrestricted sample size rules to calculate the final sample size.

Waksman[35] compared Gould and Shih[15]'s EM estimator of the standard deviation with the unadjusted one-sample pooled standard deviation on the same simulated data for different configurations. He showed through the comparisons that the unadjusted one-sample estimator generally has a smaller mean square error than the conventional EM estimator when the true treatment difference is less than 0.5 but a larger mean square error when the true treatment difference is bigger than 1. In the future research, we plan to compare our enhanced EM estimator when utilizing full block information with Kieser and Friede[17]'s simple estimator.

Kieser and Friede[17] also showed through numerical integration that the nominal type I error rate of the t-test is controlled for multiple parameter combinations they selected and the desired power is ensured by using the simple procedure. Since we showed that our enhanced EM procedure for sample size adjustment also preserve the type I error and obtain the desired power, it will be meaningful to compare the actual power and expected sample size between our procedures and Kieser and Friede[17]'s procedure.

6.2.2 Dealing with dropouts

It is common in clinical trials that missing data occurs when subjects do not complete the study and drop out of the trial without the primary endpoints being measured. When using the enhanced EM algorithm to estimate the standard deviation, we assume, at the interim, there are equal numbers of subjects in each of the two treatment groups. However, we recognize that in reality dropouts could happen. Intention-to-treat analysis is typically used to cover the issue of missing data and many imputation methods have been suggested to

forecast what the missing measurement might have been. But we would still be interested in the effects of dropouts on the enhanced EM algorithm, if no ITT data is available from the dropouts.

We assume that dropout rates are not treatment related. For the conventional EM algorithm, the probability of a subject assigned to each treatment is still 0.5 even if some subjects drop out of the trial. Because in the conventional EM algorithm, each observation is independently Bernoulli distributed with probability 0.5, even if dropouts occur, the distribution of the mixture likelihood does not change. When using the enhanced EM algorithm, our assumption is based on the exact numbers of subjects in each treatment. Knowing the total number of dropouts at the interim, we could deal with this by modifying the assumptions of the enhanced EM algorithm and consider all possible scenarios of the distribution of the number of dropouts between the two treatment groups.

If the number of subjects at the interim is less than the planned number, since the blind is maintained, we cannot figure out how many missing observations there are for each treatment group. The enhanced EM algorithm requires knowing how many subjects there are in each treatment group at the interim. One approach to handle dropouts at the interim is to compute the enhanced EM estimates assuming the true numbers of subjects remaining are equal. This will help us investigate the robustness of the enhanced EM procedure to the mistakes.

For example, suppose we plan to do the interim analysis after 80 subjects' observations are available and there are 2 subjects who drop out of the trial at the interim. Further, suppose these two subjects are actually both from the first treatment group. When we look at the blinded interim data, we do not know exactly the number of missing data in each treatment group besides knowing there are totally 78 observations. In practice, we suggest using the enhanced EM algorithm to estimate the common standard deviation by assuming balanced dropouts, that is, there are 39 subjects who finish the trial for both treatment groups.

In the future, in a simulation study, we plan to generate 3000 datasets for each parameter combination with δ is equal to 0.1, 0.2, 0.5, 0.75, 1 and 2. For each dataset, 38 observations are generated from $N(0,1)$ and 40 observations are generated from $N(\delta,1)$. We plan

to compare the enhanced EM estimates of σ by both assuming the sums of the treatment indicators in two treatments are 38 versus 40 (the correct assumption), and 39 versus 39 (the approximating assumption). If there is a noticeable difference between correctly and incorrectly assuming the dropout distribution, we plan to compare the enhanced EM estimates while assuming 39 subjects within each treatment group with the conventional EM estimates while assuming each subject having 0.5 probability in each treatment group when the sample size is 78.

If the simulation results generally show the lack of robustness by using the enhanced EM algorithm to estimate the common standard deviation, we may further explore the dropout problem by using the weighted average of the enhanced EM estimates. For the above example, there are three possibilities when there are 2 missing data in 80 observations: 2 dropouts in the first treatment group with probability 0.25; 2 dropouts in the second treatment group with probability 0.25; and 1 dropout from each treatment group with probability 0.5. We can obtain the enhanced EM estimates by assuming the different scenarios and then take the probability weighted average of the EM estimators.

Although it is beyond the scope of this dissertation, there is perhaps an even more mathematically elegant approach to handling dropouts that unfortunately is algorithmically complex. The idea is to treat the dropout problem as an exact estimation problem involving mixture distributions. We use $L_{38,40}(\boldsymbol{\theta}|\mathbf{y})$ to denote the likelihood function for the interim data assuming there are 38 observations in the first treatment group and 40 observations in the second treatment group; $L_{40,38}(\boldsymbol{\theta}|\mathbf{y})$ as the likelihood function assuming there are 40 observations in the first treatment group and 38 observations in the second treatment group; and $L_{39,39}(\boldsymbol{\theta}|\mathbf{y})$ assuming that there are 39 observations in both treatment group. Each likelihood function above is based on the mixture model in which we know the sum of the treatment identities, i.e, the likelihood function assumed for the enhanced EM algorithm. Overall, the likelihood function for the interim data with dropouts can be explained as a mixture of likelihood functions as follows,

$$L(\boldsymbol{\theta}|\mathbf{y}) = 0.25 \times L_{38,40}(\boldsymbol{\theta}|\mathbf{y}) + 0.25 \times L_{40,38}(\boldsymbol{\theta}|\mathbf{y}) + 0.5 \times L_{39,39}(\boldsymbol{\theta}|\mathbf{y})\,. \tag{6.4}$$

In theory, one can possibly use the standard EM algorithm in concert with the enhanced EM algorithm to obtain estimators for this mixture distribution.

Therefore, when there are dropouts at the interim in clinical trials, we can consider the strategy 'Intention to treat', where we assume the patients are analyzed according to the groups as they were originally randomly assigned. So that there are no missing data. Alternatively, we propose how the enhanced EM algorithm can possibly handle the missing data in this subsection. One assumption we can make for the interim data is the equal allocation of subjects. Or we can assume all possible scenarios of missingness and weight each enhanced EM estimate by its probability. Further simulation studies are needed to verify the robustness of our suggested methods.

6.3 SUMMARY

From a regulatory point of view, blinded re-estimation is preferred for adaptive clinical trials. The current Gould-Shih[15]'s EM procedure does not take into account the commonly used block randomization schemes. In our research, we enhanced the EM procedure through using the available additional information about the randomization block sizes and show this improves the estimates of the standard deviation significantly and leads to a more appropriate power for the study without inflating the type I error rate. Furthermore, our enhanced EM procedure can be applied in multi-center trials with the same properties for estimates. Our enhanced EM procedure is highly attractive due to its pragmatism in making sample size adjustment for on-going clinical trials.

APPENDIX A

INVESTIGATION ON LABEL SWITCHING IN THE EM ALGORITHM

A.1 LABEL SWITCHING OF THE CONVENTIONAL EM ALGORITHM

Suppose we fit a mixture of two normal components with a mixing proportion equal to 0.5 and a common standard deviation σ. When the mixing proportion is 0.5, the mixture distribution is symmetric in the components and the likelihood is invariant under the permutation of the component labels. Thus it is hard to identify the estimates of two component means when the labels switch during iterations of the EM algorithm. In this appendix we examine whether the means of two components can be pushed apart[21] by imposing the identifiability constraint on the model parameters[34]; i.e. $\mu_1 < \mu_2$. We show that $\mu_1^{(t)} < \mu_2^{(t)}$ at any iteration t in computation of EM estimates with the condition that the starting values satisfies $\mu_1^{(0)} < \mu_2^{(0)}$.

Without loss of generality, we assume σ is set to 1. The conditional expectation in (2.9) given the initial values $\mu_1^{(0)}$ and $\mu_2^{(0)}$ becomes

$$E(z_i|\mathbf{y}, \mu_1^{(0)}, \mu_2^{(0)}) = \frac{1}{1 + \exp\left\{\frac{1}{2}(\mu_1^{(0)})^2 - \frac{1}{2}(\mu_2^{(0)})^2 + y_i(\mu_2^{(0)} - \mu_1^{(0)})\right\}}. \tag{A.1}$$

For notation convenience, in the following proof we denote

$$c(y_i) = \exp\left\{\frac{1}{2}(\mu_1^{(0)})^2 - \frac{1}{2}(\mu_2^{(0)})^2 + y_i(\mu_2^{(0)} - \mu_1^{(0)})\right\},$$

where y_i is the ith subject at the interim for $i = 1, ..., N_1$. Since it is known that $\mu_1^{(0)} < \mu_2^{(0)}$, $c(y_i)$ is a monotonically increasing function of y_i.

At the first iteration of EM algorithm, we update $\mu_1^{(1)}$ by substituting $c(y_i)$ for the conditional expectation in (A.1), as shown in (2.20)

$$\begin{aligned}\mu_1^{(1)} &= \frac{y_1 E(z_1|\mathbf{y}, \mu_1^{(0)}, \mu_2^{(0)}) + y_2 E(z_2|\mathbf{y}, \mu_1^{(0)}, \mu_2^{(0)}) + \cdots + y_{N_1} E(z_{N_1}|\mathbf{y}, \mu_1^{(0)}, \mu_2^{(0)})}{E(z_1|\mathbf{y}, \mu_1^{(0)}, \mu_2^{(0)}) + E(z_2|\mathbf{y}, \mu_1^{(0)}, \mu_2^{(0)}) + \cdots + E(z_{N_1}|\mathbf{y}, \mu_1^{(0)}, \mu_2^{(0)})} \\ &= \frac{y_1 \frac{1}{1+c(y_1)} + y_2 \frac{1}{1+c(y_2)} + \cdots + y_{N_1} \frac{1}{1+c(y_{N_1})}}{\frac{1}{1+c(y_1)} + \frac{1}{1+c(y_2)} + \cdots + \frac{1}{1+c(y_{N_1})}} .\end{aligned} \tag{A.2}$$

We then compare it with the sample mean of observations at the interim stage $\bar{y}$,

$$\mu_1^{(1)} - \frac{y_1 + y_2 + \cdots + y_{N_1}}{N_1} .$$

The denominator of the difference is bigger than 0 and the numerator of $\mu_1^{(1)} - \bar{y}$ is as follows

$$N_1 \left\{ \frac{y_1}{1+c(y_1)} + \cdots + \frac{y_{N_1}}{1+c(y_{N_1})} \right\} - (y_1 + \cdots + y_{N_1}) \left\{ \frac{1}{1+c(y_1)} + \cdots + \frac{1}{1+c(y_{N_1})} \right\} \tag{A.3}$$

We expand (A.3) in the following form

$$\begin{aligned}&(y_1 - y_2) \left\{ \frac{1}{1+c(y_1)} - \frac{1}{1+c(y_2)} \right\} + (y_1 - y_3) \left\{ \frac{1}{1+c(y_1)} - \frac{1}{1+c(y_3)} \right\} + \cdots \\ &\quad + (y_i - y_j) \left\{ \frac{1}{1+c(y_i)} - \frac{1}{1+c(y_j)} \right\} + \cdots + (y_{N_1 - 1} - y_{N_1}) \left\{ \frac{1}{1+c(y_{N_1-1})} - \frac{1}{1+c(y_{N_1})} \right\},\end{aligned} \tag{A.4}$$

where $i = 1, ..., N_1$ and $i \neq j$. Thus, (A.3) is the sum of the product of every pairwise difference between two observations and the difference of the two corresponding functions of $c(y_i)$, i.e, $1/(1 + c(y_i))$. Because $c(y_i)$ is positive and monotonically increasing on y_i, when $y_i < y_j$ we have $1/\{1 + c(y_i)\} > 1/\{1 + c(y_j)\}$ and when $y_i > y_j$ we have $1/\{1 + c(y_i)\} < 1/\{1 + c(y_j)\}$. So each component of the summation is negative. That is $\mu_1^{(1)} < \bar{y}$.

Similarly, we update $\mu_2^{(1)}$ in the first iteration by substituting $c(y_i)$ for the conditional expectation in (A.1)

$$\begin{aligned}\mu_2^{(1)} &= \frac{y_1\{1 - E(z_1|\mathbf{y}, \mu_1^{(0)}, \mu_2^{(0)})\} + y_2\{1 - E(z_2|\mathbf{y}, \mu_1^{(0)}, \mu_2^{(0)})\} + \cdots + y_{N_1}\{1 - E(z_{N_1}|\mathbf{y}, \mu_1^{(0)}, \mu_2^{(0)})\}}{\{1 - E(z_1|\mathbf{y}, \mu_1^{(0)}, \mu_2^{(0)})\} + \{1 - E(z_2|\mathbf{y}, \mu_1^{(0)}, \mu_2^{(0)})\} + \cdots + \{1 - E(z_{N_1}|\mathbf{y}, \mu_1^{(0)}, \mu_2^{(0)})\}} \\ &= \frac{y_1 \frac{c(y_1)}{1+c(y_1)} + y_2 \frac{c(y_2)}{1+c(y_2)} + \cdots + y_{N_1} \frac{c(y_{N_1})}{1+c(y_{N_1})}}{\frac{c(y_1)}{1+c(y_1)} + \frac{c(y_2)}{1+c(y_2)} + \cdots + \frac{c(y_{N_1})}{1+c(y_{N_1})}} .\end{aligned} \tag{A.5}$$

We also compare it with the sample mean of observations at the interim stage $\bar{y}$,

$$\mu_2^{(1)} - \frac{y_1 + y_2 + \cdots + y_{N_1}}{N_1}\,.$$

The denominator of the difference is positive and the numerator is as follows

$$N_1\left\{\frac{y_1 c(y_1)}{1+c(y_1)} + \cdots + \frac{y_{N_1} c(y_{N_1})}{1+c(y_{N_1})}\right\} - (y_1 + \cdots + y_{N_1})\left\{\frac{c(y_1)}{1+c(y_1)} + \cdots + \frac{c(y_{N_1})}{1+c(y_{N_1})}\right\}. \quad \text{(A.6)}$$

We expand (A.6) in the following form

$$\begin{aligned}(y_1 - y_2)\left\{\frac{c(y_1)}{1+c(y_1)} - \frac{c(y_2)}{1+c(y_2)}\right\} + (y_1 - y_3)\left\{\frac{c(y_1)}{1+c(y_1)} - \frac{c(y_3)}{1+c(y_3)}\right\} + \cdots \\ + (y_i - y_j)\left\{\frac{c(y_i)}{1+c(y_i)} - \frac{c(y_j)}{1+c(y_j)}\right\} + \cdots + (y_{N_1-1} - y_{N_1})\left\{\frac{c(y_{N_1-1})}{1+c(y_{N_1-1})} - \frac{c(y_{N_1})}{1+c(y_{N_1})}\right\}.\end{aligned} \quad \text{(A.7)}$$

Similar as in our explanation of (A.4), when $y_i < y_j$ we have $c(y_i)/\{1+c(y_i)\} < c(y_j)/\{1+c(y_j)\}$ and when $y_i > y_j$ we have $c(y_i)/\{1+c(y_i)\} > c(y_j)/\{1+c(y_j)\}$. Since each part of the summation is positive, we get (A.7) $-\,\bar{\mathrm{y}} > 0$, i.e., $\mu_2^{(1)} > \bar{y}$.

Therefore, by using $\bar{y}$ as a mediator we show that $\mu_1^{(1)} < \mu_2^{(1)}$ under the constraint $\mu_1^{(0)} < \mu_2^{(0)}$. If we simply replace the iteration number to t and repeat the same proof steps we can show $\mu_1^{(t)} < \mu_2^{(t)}$ by knowing $\mu_1^{(t-1)} < \mu_2^{(t-1)}$ at any iteration t. So we conclude the label switching problem is solved while imposing the constraint $\mu_1 < \mu_2$ in the conventional EM algorithm.

A.2 LABEL SWITCHING OF THE ENHANCED EM ALGORITHM

In the enhanced EM algorithm, we assume equal subjects from each treatment are observed at the interim stage. The observed data likelihood is invariant under the relabeling of two mixture components. After putting the constraint $\mu_1 < \mu_2$ on the parameter space, we can also show the estimates of the means retain their order at each iteration.

Without the loss of generality, the common standard deviation σ is set to 1. Start with $\mu_1^{(0)} < \mu_2^{(0)}$, the conditional expectation of z_i (for $i = 1, ..., N_1$) given the observed data at the interim stage and the sum of z_i's in (2.29) is written as

$$E\left(z_i|\mathbf{y}, \sum_{i=1}^{N_1} z_i = \frac{N_1}{2}, \mu_1^{(0)}, \mu_2^{(0)}\right) = \frac{w_i R(n-1, S\setminus\{i\})}{R(n,S)}\,, \tag{A.8}$$

where $n = N_1/2$ and $S = \{1, 2, ..., N_1\}$. w_i is a monotonically decreasing function of y_i since $\mu_1^{(0)} < \mu_2^{(0)}$ and $w_i^{(0)}$ can be re-written as follows

$$\begin{aligned} w_i^{(0)} =& \frac{p_i^{(0)}}{1-p_i^{(0)}} = \frac{f_{1i}/(f_{1i}+f_{2i})}{f_{2i}/(f_{1i}+f_{2i})} \\ &= \exp\left\{ -\frac{1}{2}(\mu_1^{(0)})^2 + \frac{1}{2}(\mu_2^{(0)})^2 + y_i\mu_1^{(0)} - y_i\mu_2^{(0)} \right\}. \end{aligned}$$

In the first iteration of the enhanced EM algorithm, we update the estimate of μ_1 in (2.30) as follows

$$\begin{aligned} \mu_1^{(1)} &= \frac{y_1\frac{w_1R(n-1,S\setminus\{1\})}{R(n,S)} + y_2\frac{w_2R(n-1,S\setminus\{2\})}{R(n,S)} + \cdots + y_{N_1}\frac{w_{N_1}R(n-1,S\setminus\{N_1\})}{R(n,S)}}{\frac{w_1R(n-1,S\setminus\{1\})}{R(n,S)} + \frac{w_2R(n-1,S\setminus\{2\})}{R(n,S)} + \cdots + \frac{w_{N_1}R(n-1,S\setminus\{N_1\})}{R(n,S)}} \\ &= \frac{y_1w_1R(n-1,S\setminus\{1\}) + y_2w_2R(n-1,S\setminus\{2\}) + \cdots + y_{N_1}w_{N_1}R(n-1,S\setminus\{N_1\})}{w_1R(n-1,S\setminus\{1\}) + w_2R(n-1,S\setminus\{2\}) + \cdots + w_{N_1}R(n-1,S\setminus\{N_1\})}\,. \end{aligned} \tag{A.9}$$

We then compare $\mu_1^{(1)}$ with the sample mean of observations at the interim stage $\bar{y}$,

$$\mu_1^{(1)} - \frac{y_1 + y_2 + \cdots + y_{N_1}}{N_1}\,.$$

The denominator of the difference is bigger than 0 because w_i and $R(n-1, S \setminus \{i\})$ are both positive numbers for any i. We calculate the numerator of $\mu_1^{(1)} - \bar{y}$ in the following form

$$
\begin{aligned}
& N_1\Big\{y_1 w_1 R(n-1, S \setminus \{1\}) + y_2 w_2 R(n-1, S \setminus \{2\}) + \cdots + y_{N_1} w_{N_1} R(n-1, S \setminus \{N_1\})\Big\} \\
& - (y_1 + \cdots + y_{N_1})\Big\{w_1 R(n-1, S \setminus \{1\}) + w_2 R(n-1, S \setminus \{2\}) + \cdots + w_{N_1} R(n-1, S \setminus \{N_1\})\Big\} \\
& = (y_1 - y_2)\Big\{w_1 R(n-1, S \setminus \{1\}) - w_2 R(n-1, S \setminus \{2\})\Big\} \\
& \quad + (y_1 - y_3)\Big\{w_1 R(n-1, S \setminus \{1\}) - w_3 R(n-1, S \setminus \{3\})\Big\} + \cdots \\
& \quad + (y_i - y_j)\Big\{w_i R(n-1, S \setminus \{i\}) - w_j R(n-1, S \setminus \{j\})\Big\} + \cdots \\
& \quad + (y_{N_1-1} - y_{N_1})\Big\{w_{N_1-1} R(n-1, S \setminus \{N_1 - 1\}) - w_{N_1} R(n-1, S \setminus \{N_1\})\Big\}.
\end{aligned}
\tag{A.10}
$$

Let us look at the R function, it has the following relation as shown in (2.17)

$$
w_i R(n-1, S \setminus \{i\}) = R(n, S) - R(n, S \setminus \{i\}), \tag{A.11}
$$

where $R(n, S \setminus \{i\})$ means the summation over all possible n combinations of w's excluding w_i. So if $w_i > w_j$ then $R(n, S \setminus \{i\}) < R(n, S \setminus \{j\})$ for $i, j = 1, ..., N_1$. The proof is straightforward. When $w_i > w_j$, the product of n distinct w's only excluding w_i is smaller than the product of n distinct w's only excluding w_j and the product of n distinct w's excluding w_i and w_j are the same. It is also obvious that when $w_i > w_j$, $w_i R(n-1, S \setminus \{i\}) > w_j R(n-1, S \setminus \{j\})$. Because w_i is decreasing on y_i, when $y_i < y_j$ we have $w_i R(n-1, S \setminus \{i\}) > w_j R(n-1, S \setminus \{j\})$ and when $y_i > y_j$ we have $w_i R(n-1, S \setminus \{i\}) < w_j R(n-1, S \setminus \{j\})$. Therefore each component of the summation in (A.10) is smaller than 0. That is, $\mu_1^{(1)} - \bar{y} < 0$.

Now we show $\mu_2^{(1)}$ bigger than $\bar{y}$. In the first iteration of the enhanced EM algorithm, $\mu_2^{(1)}$ can be expanded as follows by using the relation in (A.11)

$$
\mu_2^{(1)} = \frac{y_1 R(n, S \setminus \{1\}) + y_2 R(n, S \setminus \{2\}) + \cdots + y_{N_1} R(n, S \setminus \{N_1\})}{R(n, S \setminus \{i\}) + R(n, S \setminus \{2\}) + \cdots + R(n, S \setminus \{N_1\})}. \tag{A.12}
$$

Similarly, the denominator of $\mu_2^{(1)} - \bar{y}$ is positive since R function is always positive. The numerator of the difference can be written as follows

$$(y_1 - y_2)\Big\{R(n, S \setminus \{1\}) - R(n, S \setminus \{2\})\Big\} + (y_1 - y_3)\Big\{R(n, S \setminus \{1\}) - R(n, S \setminus \{3\})\Big\} + \cdots$$
$$+ (y_i - y_j)\Big\{R(n, S \setminus \{i\}) - R(n, S \setminus \{j\})\Big\} + \cdots + (y_{N_1 - 1} - y_{N_1})\Big\{R(n, S \setminus \{N_1 - 1\}) - R(n, S \setminus \{N_1\})\Big\}. \tag{A.13}$$

We know when $y_i > y_j$, it makes $w_i < w_j$, so $R(n, S \setminus \{i\}) > R(n, S \setminus \{j\})$ and $y_i < y_j$, $R(n, S \setminus \{i\}) < R(n, S \setminus \{j\})$. Therefore each component of (A.13) is positive, which make the sum positive. That is, $\mu_2^{(1)} - \bar{y} > 0$.. By repeating the same proof steps, we can show at any iteration t, $\mu_1^{(t)} < \mu_2^{(t)}$. Our conclusion is that with knowing $\mu_1 < \mu_2$, label switching does not happen at any iteration of the enhanced EM algorithm.

APPENDIX B

COMPARISON OF THE ENHANCED EM ESTIMATES WITH THE INTERIOR CONVENTIONAL EM ESTIMATES

The empirical evidence showed in Section 3.3 indicates that the conventional EM algorithm obtains non-meaningful boundary estimates. So we make an arbitrary straight line which is parallel to the line $\hat{\mu}_1 = \hat{\mu}_2$ to separate the EM estimates around the true parameter values and the boundary modes. Notice that the arbitrary line we make only can roughly but not accurately pick up all the correct conventional EM estimates. It is difficult to decide if estimates around the line $\hat{\mu}_1 = \hat{\mu}_2$ are incorrect boundary estimates or they are meaningful estimates due to the bias especially for the case when δ is small. Fortunately, our illustrated method used to separate the conventional EM estimates is good enough to show how the conventional EM algorithm performs when the estimates are not stuck at the boundary modes compared to the enhanced EM algorithm.

In Tables 17 and 18, we used the same 1000 datasets for each parameter configuration as in Tables 2 and 3, so the enhanced EM estimates are the same as in Tables 2 and 3 for each δ. For the conventional EM algorithm, we plot a scatter plot of $\hat{\mu}_1$ versus $\hat{\mu}_2$ for each δ then use an arbitrary line to separate the two clusters of estimates as we did for Figures 8 and 9. We consider estimates above the straight line as meaningful conventional EM estimates and compare the mean bias, variance and the MSE of the 1000 estimates with those of the enhanced EM estimates. The arbitrary lines we used for each scenario slightly vary for different δ and datasets. When $N_1 = 20$, it is obvious the enhanced EM algorithm outperforms the conventional EM algorithm. The enhanced EM estimates have

smaller mean bias and variance even when δ is as small as 0.1. When $N_1 = 80$, the enhanced EM algorithm obtains smaller mean bias and similar or slightly bigger variance than the conventional EM algorithm. Overall for different values of δ, the enhanced EM estimates have smaller MSEs than that of the conventional EM estimates.

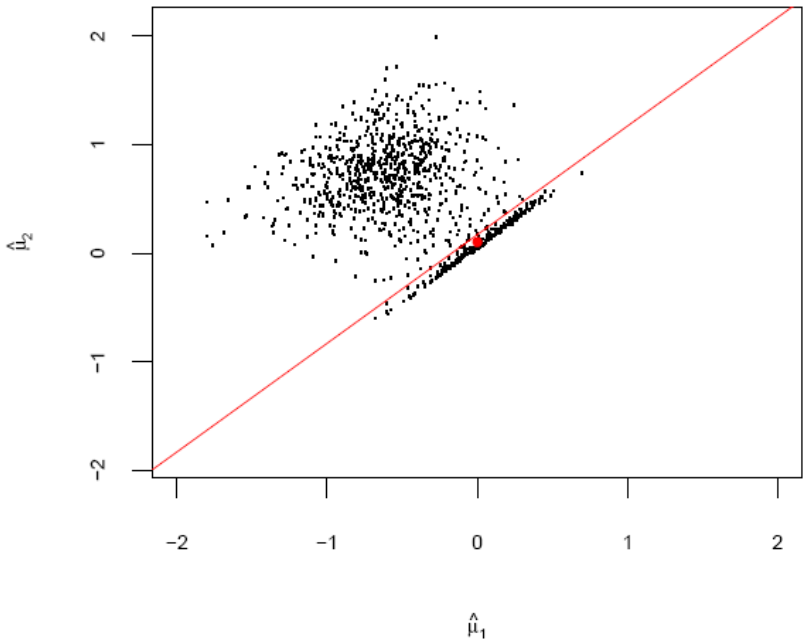
Figures 19 to 22 visually display how we separate the conventional EM estimates and the performance of the enhanced EM estimates on μ_1 and μ_2 for the case when δ is small (0.1) and when δ is big (2).

Table 17: Comparisons of then enhanced EM estimates with the meaningful conventional EM estimates when $N_1 = 20$. True parameters used to generate sample are set as $\sigma = 1$, $\mu_1 = 0$ and $\delta = 0, 0.1, 0.2, 0.35, 0.5, 0.75, 1, 1.5, 2$. 1000 sample are generated from each parameters configuration.

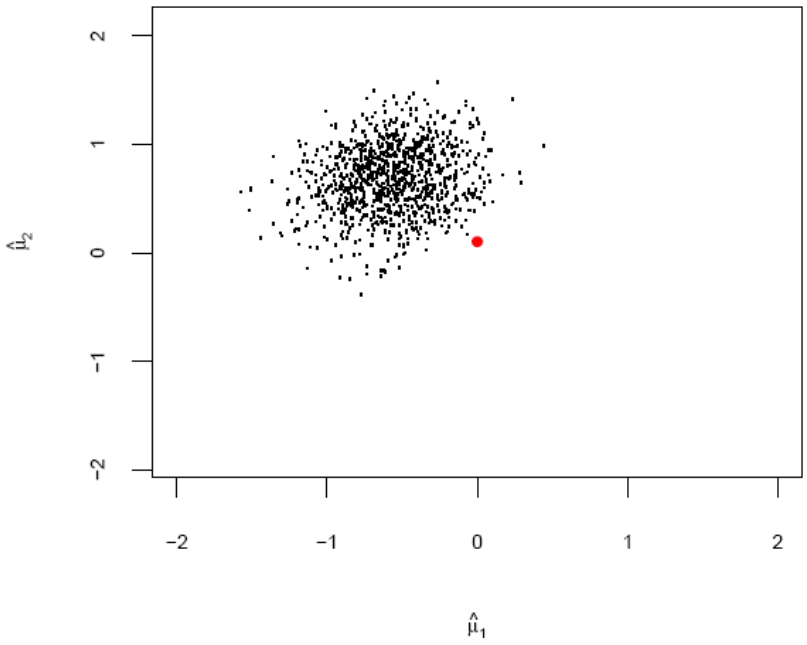
		Enhanced EM estimates				Interior Conventional EM estimates			
δ		μ_1	μ_2	σ	$\frac{\mu_2-\mu_1}{\sigma}$	μ_1	μ_2	σ	$\frac{\mu_2-\mu_1}{\sigma}$
0.1	Bias	-0.5724	0.5807	-0.2912	1.8087	-0.6096	0.6233	-0.3270	2.1038
	Variance	0.0877	0.0957	0.0296	0.6403	0.1147	0.1179	0.0357	1.0384
	MSE	0.4152	0.4328	0.1144	3.9109	0.4861	0.5062	0.1426	5.4630
0.2	Bias	-0.5032	0.5264	-0.2884	1.6918	-0.5699	0.5990	-0.3461	2.1195
	Variance	0.0846	0.0943	0.0330	0.7466	0.1048	0.1085	0.0345	1.0202
	MSE	0.3377	0.3712	0.1161	3.6079	0.4294	0.4672	0.1542	5.511
0.35	Bias	-0.4516	0.4416	-0.2780	1.5197	-0.5037	0.4911	-0.3241	1.8462
	Variance	0.0940	0.0864	0.0318	0.6500	0.1104	0.1074	0.0344	0.9185
	MSE	0.2979	0.2813	0.1091	2.9587	0.3640	0.3485	0.1394	4.3256
0.5	Bias	-0.3859	0.3989	-0.2772	1.4197	-0.4445	0.4564	-0.3257	1.7686
	Variance	0.0827	0.0905	0.0308	0.6474	0.1021	0.1181	0.0309	0.9115
	MSE	0.2315	0.2495	0.1076	2.6623	0.2995	0.3263	0.1369	4.0382
0.75	Bias	-0.2868	0.2775	-0.2491	1.1632	-0.3183	0.3352	-0.2944	1.4807
	Variance	0.0885	0.0884	0.0372	0.6935	0.1094	0.1130	0.0426	1.0611
	MSE	0.1707	0.1653	0.0992	2.0458	0.2106	0.2252	0.1293	3.2521
1	Bias	-0.2066	0.1966	-0.2096	0.9239	-0.2773	0.2680	-0.2679	1.3306
	Variance	0.0950	0.0987	0.0367	0.6681	0.1328	0.1208	0.0405	1.0302
	MSE	0.1376	0.1373	0.0806	1.5210	0.2095	0.1925	0.1122	2.7994
1.5	Bias	-0.0857	0.0984	-0.1535	0.6391	-0.1364	0.1566	-0.1961	0.9408
	Variance	0.1134	0.1069	0.0377	0.7317	0.1391	0.1393	0.0435	1.0125
	MSE	0.1206	0.1165	0.0613	1.1395	0.1576	0.1637	0.0819	1.8963
2	Bias	-0.0032	0.0144	-0.1052	0.4005	-0.0354	0.0470	-0.1331	0.5948
	Variance	0.1099	0.1193	0.0410	0.7165	0.1431	0.1441	0.0453	0.9510
	MSE	0.1098	0.1194	0.0520	0.8762	0.1442	0.1462	0.0629	1.3038

Table 18: Comparisons of then enhanced EM estimates with the meaningful conventional EM estimates when $N_1 = 80$. True parameters used to generate sample are set as $\sigma = 1$, $\mu_1 = 0$ and $\delta = 0, 0.1, 0.2, 0.35, 0.5, 0.75, 1, 1.5, 2$. 1000 sample are generated from each parameters configuration.

		Enhanced EM estimates				Interior Conventional EM estimates			
δ		μ_1	μ_2	σ	$\frac{\mu_2-\mu_1}{\sigma}$	μ_1	μ_2	σ	$\frac{\mu_2-\mu_1}{\sigma}$
0.1	Bias	-0.4591	0.4619	-0.1802	1.2246	-0.5257	0.5226	-0.2150	1.4447
	Variance	0.0484	0.0461	0.0152	0.3827	0.0426	0.0418	0.0154	0.3621
	MSE	0.2592	0.2594	0.0476	1.8820	0.3189	0.3149	0.0616	2.4486
0.2	Bias	-0.4231	0.4240	-0.1794	1.1605	-0.4769	0.4808	-0.2203	1.3681
	Variance	0.0430	0.0447	0.0168	0.3773	0.0447	0.0414	0.0151	0.3721
	MSE	0.2219	0.2244	0.0490	1.7236	0.2720	0.2726	0.0636	2.2432
0.35	Bias	-0.3623	0.3468	-0.1706	1.0121	-0.4220	0.4094	-0.2090	1.2295
	Variance	0.0445	0.0457	0.0172	0.3931	0.0408	0.04418	0.0162	0.3781
	MSE	0.1757	0.1660	0.0463	1.4172	0.2188	0.2093	0.0598	1.8893
0.5	Bias	-0.2973	0.2936	-0.1553	0.8704	-0.3650	0.3642	-0.1974	1.1019
	Variance	0.0485	0.0491	0.0165	0.3744	0.0403	0.0440	0.0136	0.3300
	MSE	0.1368	0.1353	0.0406	1.1317	0.1735	0.1766	0.0526	1.5437
0.75	Bias	-0.1933	0.1886	-0.1310	0.6324	-0.2576	0.2485	-0.1666	0.8324
	Variance	0.0504	0.0486	0.0175	0.3765	0.0439	0.04256	0.0158	0.3386
	MSE	0.0877	0.0842	0.0346	0.7760	0.1102	0.1043	0.0435	1.0310
1	Bias	-0.1113	0.0981	-0.1020	0.4273	-0.172	0.1592	-0.1349	0.6168
	Variance	0.0539	0.0557	0.0187	0.3862	0.0471	0.0477	0.0178	0.3508
	MSE	0.0662	0.0653	0.0291	0.5684	0.0766	0.0730	0.0359	0.7307
1.5	Bias	0.0013	0.0007	-0.0428	0.1442	-0.0307	0.0376	-0.0639	0.2487
	Variance	0.0567	0.0575	0.0203	0.3684	0.0523	0.0547	0.0185	0.3414
	MSE	0.0567	0.0575	0.0211	0.3889	0.0532	0.0561	0.0226	0.4029
2	Bias	0.0158	-0.0147	-0.0230	0.0889	0.083	-0.0038	-0.0301	0.1264
	Variance	0.0450	0.0441	0.0199	0.3242	0.0514	0.0482	0.0203	0.3440
	MSE	0.0452	0.0443	0.0204	0.3318	0.0514	0.0482	0.0212	0.3597

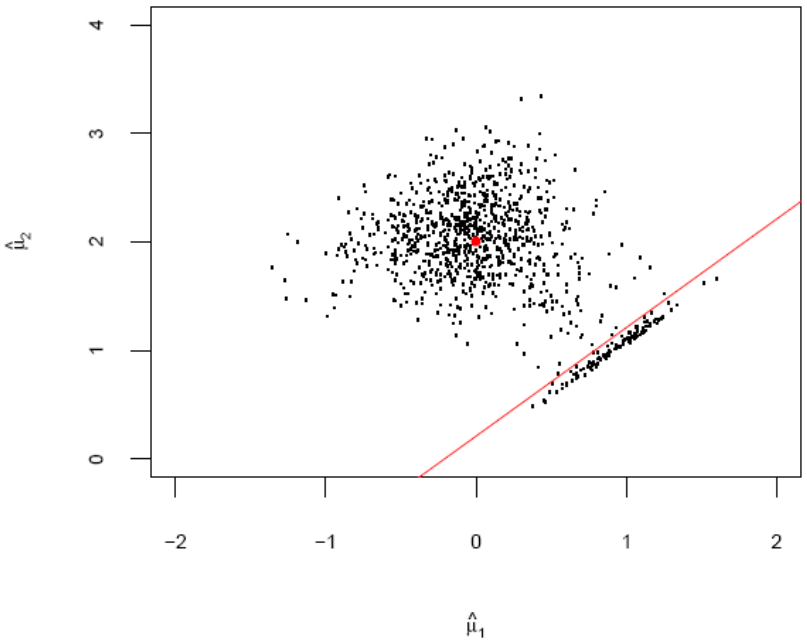


(a) Scatterplot of the Conventional EM estimate of μ_1 versus μ_2

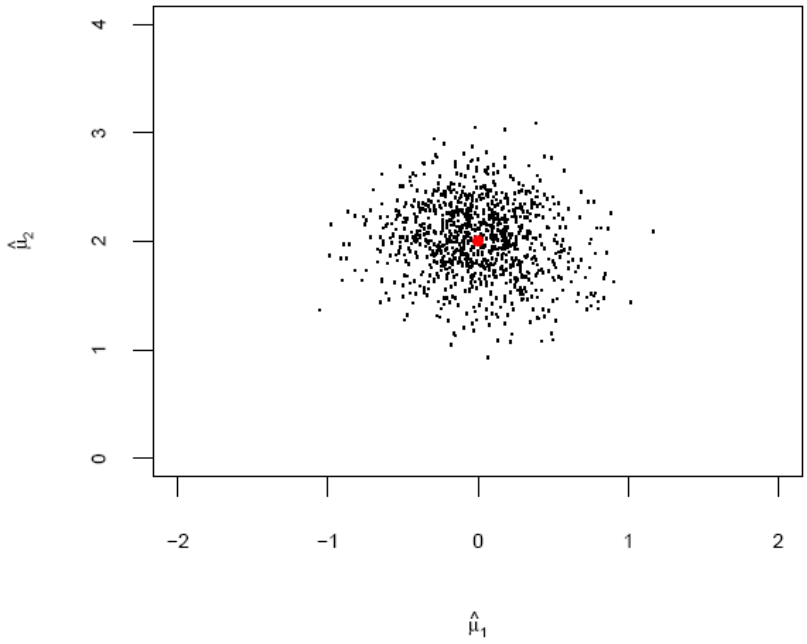


(b) Scatterplot of the Enhanced EM estimate of μ_1 versus μ_2

Figure 19: Comparison of two EM estimates of μ_1 and μ_2 when $N_1 = 20$ ($\mu_1 = 0$, $\mu_2 = 0.1$ and $\sigma = 1$, stimulater 1000). The red point denotes the true value of (μ_1, μ_2) on the scatterplot.

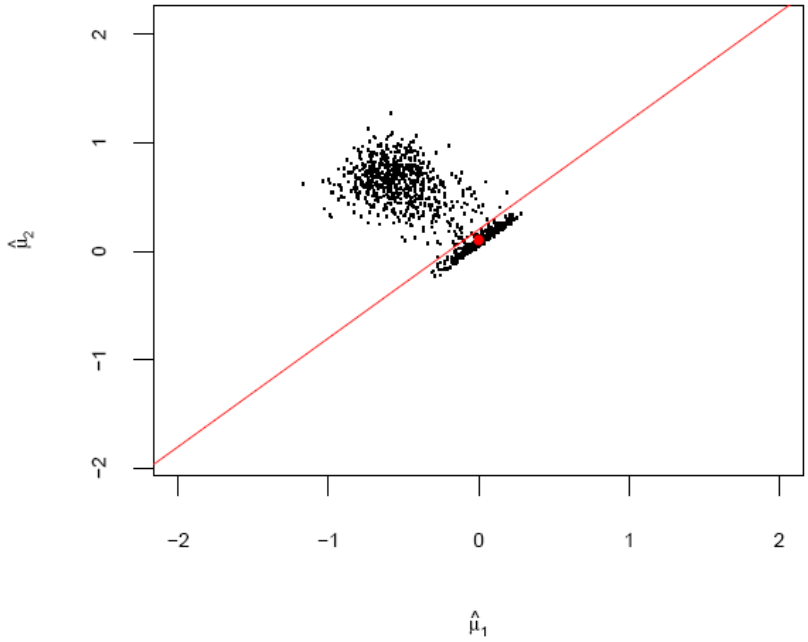


(a) Scatterplot of the Conventional EM estimate of μ_1 versus μ_2

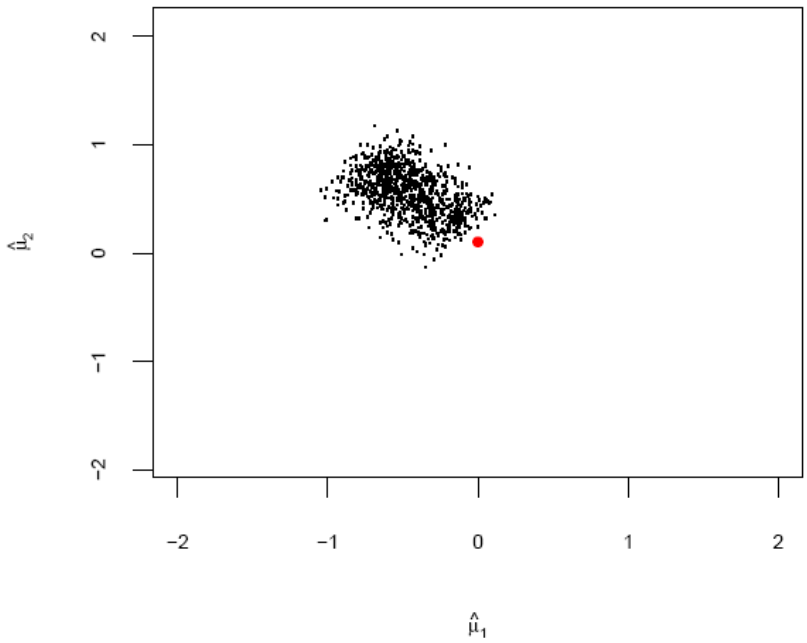


(b) Scatterplot of the Enhanced EM estimate of μ_1 versus μ_2

Figure 20: Comparison of two EM estimates of μ_1 and μ_2 when $N_1 = 20$ ($\mu_1 = 0$, $\mu_2 = 2$ and $\sigma = 1$, stimulater 1000). The red point denotes the true value of (μ_1, μ_2) on the scatterplot.

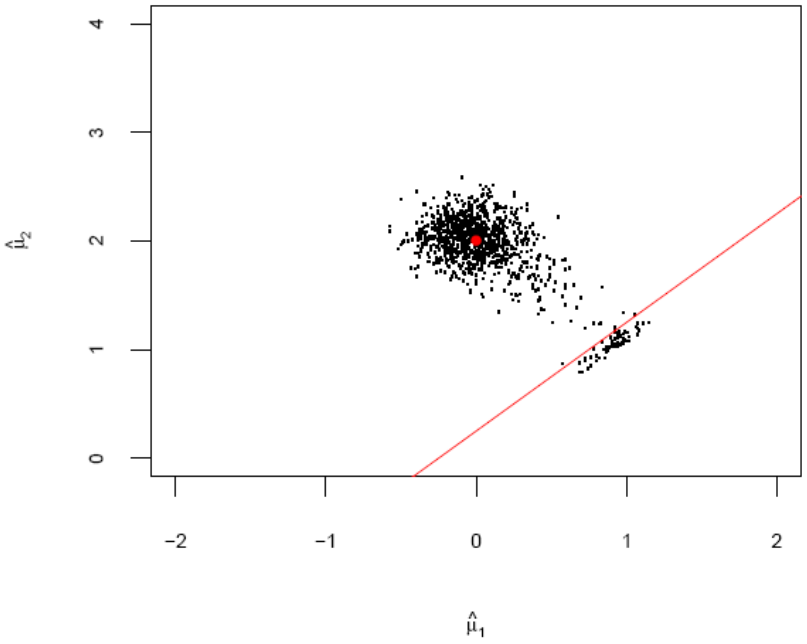


(a) Scatterplot of the Conventional EM estimate of μ_1 versus μ_2

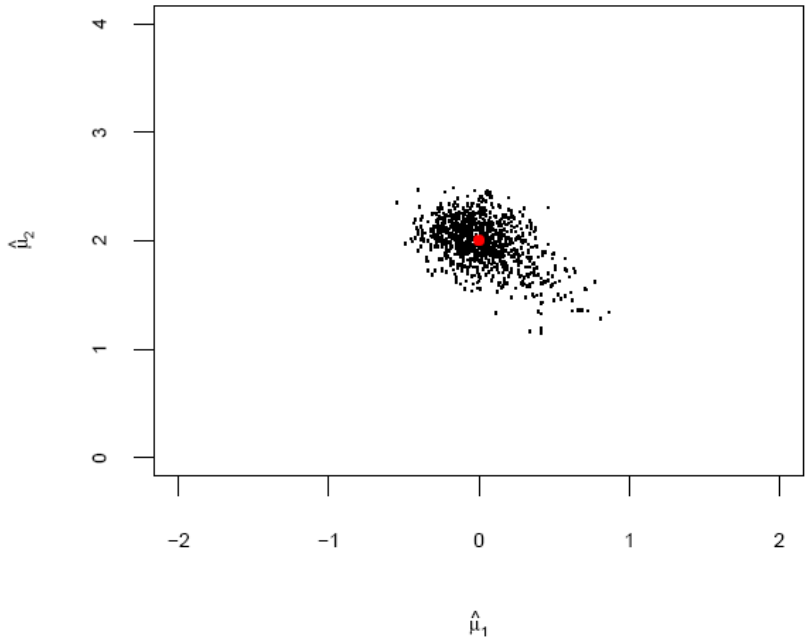


(b) Scatterplot of the Enhanced EM estimate of μ_1 versus μ_2

Figure 21: Comparison of two EM estimates of μ_1 and μ_2 when $N_1 = 80$ ($\mu_1 = 0$, $\mu_2 = 0.1$ and $\sigma = 1$, stimulater 1000). The red point denotes the true value of (μ_1, μ_2) on the scatterplot.



(a) Scatterplot of the Conventional EM estimate of μ_1 versus μ_2



(b) Scatterplot of the Enhanced EM estimate of μ_1 versus μ_2

Figure 22: Comparison of two EM estimates of μ_1 and μ_2 when $N_1 = 80$ ($\mu_1 = 0$, $\mu_2 = 2$ and $\sigma = 1$, stimulater 1000). The red point denotes the true value of (μ_1, μ_2) on the scatterplot.

APPENDIX C

BLINDED RE-ESTIMATION OF STANDARD DEVIATION WHEN TREATMENT-BY-CENTER INTERACTION EXISTS IN TWO CENTER DESIGNS

C.1 WHEN TREATMENT-BY-CENTER INTERACTION EXISTS

As noted in Section 5.2.2, the difficulties is that the distribution is not identifiable, and only with assumptions about the nature of the interactions will the following lead to coherent answers. Since the center effect is not consistent between the two treatment groups, we assume different parameters for treatment means within each center. In center 1, if a patient is in the control group, i.e., $z_{i.1} = 1$, then $y_{i.1} \sim N(\tau_{11}, \sigma)$; if he/she is in the experimental group, i.e., $z_{i.1} = 0$, then $y_{i.1} \sim N(\tau_{21}, \sigma)$ where $i = 1, ..., N_1/2$. Similarly in center 2, if a patient is in the control group, i.e., $z_{i.2} = 1$, then $y_{i.2} \sim N(\tau_{12}, \sigma)$; if he/she is in the experimental group, i.e., $z_{i.2} = 0$, then $y_{i.2} \sim N(\tau_{22}, \sigma)$. This setup is very similar to having two independent blocked studies with sparate parameters for each, except in our case there is σ^2 in common.

For the conventional EM algorithm, it is assumed that the probability of each patient assigning to each treatment group is 0.5 within both centers. Therefore, the complete data

likelihood function is given by

$$L(\boldsymbol{\theta}; \mathbf{y}_{\cdots}, \mathbf{z}_{\cdot\cdot\mathbf{1}}, \mathbf{z}_{\cdot\cdot\mathbf{2}}) = \prod_{i=1}^{\frac{N_1}{2}} \Big\{ f(y_{i.1}|\tau_{11}, \sigma)^{z_{i.1}} \cdot f(y_{i.1}|\tau_{21}, \sigma)^{1-z_{i.1}} \cdot \frac{1}{2}^{z_{i.1}} (1-\frac{1}{2})^{1-z_{i.1}} \\ \cdot f(y_{i.2}|\tau_{12}, \sigma)^{z_{i.2}} \cdot f(y_{i.2}|\tau_{22}, \sigma)^{1-z_{i.2}} \cdot \frac{1}{2}^{z_{i.2}} (1-\frac{1}{2})^{1-z_{i.2}} \Big\}. \tag{C.1}$$

Similarly as in the case when interaction does not exist, the $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ function is linear in $z_{i.k}$. In the E-step, the conditional expectations of $z_{i.k}$ is calculated as:

$$E(z_{i.1}|\mathbf{y}_{\cdot\cdot\mathbf{1}}, \theta_1^{(t)}) = P(z_{i.1} = 1|\mathbf{y}_{\cdot\cdot\mathbf{1}}, \boldsymbol{\theta}_{\mathbf{1}}^{(\mathbf{t})}) = \frac{f(y_{i.1}|\tau_{11}^{(t)}, \sigma^{(t)})}{f(y_{i.1}|\tau_{11}^{(t)}, \sigma^{(t)}) + f(y_{i.1}|\tau_{21}^{(t)}, \sigma^{(t)})} \\ E(z_{i.2}|\mathbf{y}_{\cdot\cdot\mathbf{2}}, \theta_2^{(t)}) = P(z_{i.2} = 1|\mathbf{y}_{\cdot\cdot\mathbf{2}}, \boldsymbol{\theta}_{\mathbf{1}}^{(\mathbf{t})}) = \frac{f(y_{i.2}|\tau_{12}^{(t)}, \sigma^{(t)})}{f(y_{i.2}|\tau_{12}^{(t)}, \sigma^{(t)}) + f(y_{i.2}|\tau_{22}^{(t)}, \sigma^{(t)})}. \tag{C.2}$$

In the M-step, we treat the missing information as known and substitute the conditional expectation computed in the E-step in the Q-function. The M-step maximizes the conditional expectation of the complete-data log-likelihood. Thus we update $\boldsymbol{\theta_1}^{(t+1)}$ with

$$\tau_{11}^{(t)} = \frac{\sum_{i=1}^{N_1/2} y_{i.1} E(z_{i.1}|\mathbf{y}_{\cdot\cdot\mathbf{1}}, \boldsymbol{\theta}_{\mathbf{1}}^{(\mathbf{t})})}{\sum_{i=1}^{N_1/2} E(z_{i.1}|\mathbf{y}_{\cdot\cdot\mathbf{1}}, \boldsymbol{\theta}_{\mathbf{1}}^{(\mathbf{t})})} \\ \tau_{21}^{(t)} = \frac{\sum_{i=1}^{N_1/2} y_{i.1} \{1 - E(z_{i.1}|\mathbf{y}_{\cdot\cdot\mathbf{1}}, \boldsymbol{\theta}_{\mathbf{1}}^{(\mathbf{t})})\}}{\sum_{i=1}^{N_1/2} \{1 - E(z_{i.1}|\mathbf{y}_{\cdot\cdot\mathbf{1}}, \boldsymbol{\theta}_{\mathbf{1}}^{(\mathbf{t})})\}} \\ \tau_{12}^{(t)} = \frac{\sum_{i=1}^{N_1/2} y_{i.2} E(z_{i.2}|\mathbf{y}_{\cdot\cdot\mathbf{2}}, \boldsymbol{\theta}_{\mathbf{2}}^{(\mathbf{t})})}{\sum_{i=1}^{N_1/2} E(z_{i.2}|\mathbf{y}_{\cdot\cdot\mathbf{2}}, \boldsymbol{\theta}_{\mathbf{2}}^{(\mathbf{t})})} \\ \tau_{22}^{(t)} = \frac{\sum_{i=1}^{N_1/2} y_{i.2} \{1 - E(z_{i.2}|\mathbf{y}_{\cdot\cdot\mathbf{2}}, \boldsymbol{\theta}_{\mathbf{2}}^{(\mathbf{t})})\}}{\sum_{i=1}^{N_1/2} \{1 - E(z_{i.2}|\mathbf{y}_{\cdot\cdot\mathbf{1}}, \boldsymbol{\theta}_{\mathbf{2}}^{(\mathbf{t})})\}} \\ \sigma^{2(t+1)} = \frac{1}{N_1} \sum_{i=1}^{N_1/2} \Big[E(z_{i.1}|\mathbf{y}_{\cdot\cdot\mathbf{1}}, \boldsymbol{\theta}_{\mathbf{1}}^{(\mathbf{t})})(y_{i.1} - \tau_{11}^{(t)})^2 + \{1 - E(z_{i.1}|\mathbf{y}_{\cdot\cdot\mathbf{1}}, \boldsymbol{\theta}_{\mathbf{1}}^{(\mathbf{t})})\}(y_{i.2} - \tau_{21}^{(t)})^2 \\ + E(z_{i.2}|\mathbf{y}_{\cdot\cdot\mathbf{2}}, \boldsymbol{\theta}_{\mathbf{2}}^{(\mathbf{t})})(y_{i.2} - \tau_{12}^{(t)})^2 + \{1 - E(z_{i.2}|\mathbf{y}_{\cdot\cdot\mathbf{2}}, \boldsymbol{\theta}_{\mathbf{2}}^{(\mathbf{t})})\}(y_{i.2} - \tau_{22}^{(t)})^2 \Big]. \tag{C.3}$$

For the enhanced EM algorithm, the complete data likelihood function becomes

$$L(\boldsymbol{\theta}; \mathbf{y}_{\cdots}, \mathbf{z}_{\cdot\cdot\mathbf{1}}, \mathbf{z}_{\cdot\cdot\mathbf{2}}, \sum_{i=1}^{N_1/2} z_{i.1}, \sum_{i=1}^{N_1/2} z_{i.2}) = \prod_{i=1}^{\frac{N}{2}} \Big\{ f(y_{i.1}|\tau_{11}, \sigma)^{z_{i.1}} \cdot f(y_{i.1}|\tau_{21}, \sigma)^{1-z_{i.1}} \\ \cdot f(y_{i.2}|\tau_{12}, \sigma)^{z_{i.2}} \cdot f(y_{i.2}|\tau_{22}, \sigma)^{1-z_{i.2}} \cdot \frac{1}{\binom{N_1/2}{N_1/4}} \cdot \frac{1}{\binom{N_1/2}{N_1/4}} \Big\}. \tag{C.4}$$

The E-step stays the same as in (5.12), with different $w_{i..}$, where $w_{i.1} = f(y_{i.1}|\tau_{11}, \sigma)/f(y_{i.1}|\tau_{21}, \sigma)$ for center 1 and $w_{i.2} = f(y_{i.2}|\tau_{21}, \sigma)/f(y_{i.2}|\tau_{22}, \sigma)$ for center 2. For the enhanced EM algorithm with block design, the E-step also does not change as in (5.15) and (5.16). The M-step stays the same as in (C.3).

C.2 COMPARISON AMONG EM PROCEDURES UNDER THE ASSUMPTION OF INTERACTION OR NO INTERACTION

In Appendix C1, we suggest an EM procedures in detail when treatment-by-center interaction exists in the two-center trial. Due to the lack of information in the design stage, possible treatment-by-center interactions are usually neglected when planning the sample size. We also may not want to consider the interaction in the sample size re-calculation procedure.

In a simulation study of our suggested algorithm, we considered a couple of scenarios where two centers have the same or different treatment effects, then we conducted the EM procedures by both assuming whether interactions exist or not. For each scenario, we generated 1000 samples with a sample size of 80 across two centers at the interim. We are interested in the influence the mis-specification on interactions in the EM algorithm has on the estimation of the standard deviation. The specific scenarios we are considering as follows:

Scenario 1: in Center 1, $y_{i.1} \sim N(0, 1)$ and $N(0.5, 1)$; in Center 2, $y_{i.2} \sim N(0.1, 1)$ and $N(0.6, 1)$ (no interaction; small difference between centers).

Scenario 2: in Center 1, $y_{i.1} \sim N(0, 1)$ and $N(0.5, 1)$; in Center 2, $y_{i.2} \sim N(0.5, 1)$ and $N(1, 1)$ (no interaction; moderate difference between centers).

Scenario 3: in Center 1, $y_{i.1} \sim N(0, 1)$ and $N(0.5, 1)$; in Center 2, $y_{i.2} \sim N(0.1, 1)$ and $N(0.3, 1)$ (small quantitative interaction; small difference between centers).

Scenario 4: in Center 1, $y_{i.1} \sim N(0, 1)$ and $N(0.5, 1)$; in Center 2, $y_{i.2} \sim N(0.5, 1)$ and $N(0, 1)$ (moderate qualitative interaction; no difference between centers).

In each of the scenarios, when using the EM algorithm assuming non-interaction, we make the starting values of τ_2 bigger than the starting values of τ_1, specifically we set τ_1 and τ_2 as

the mean of the primary endpoints plus and minus a constant respectively; when using the EM algorithm assuming interaction, we make the starting values of difference between $\tau_1 1$ and $\tau_2 1$ and the difference between $\tau_1 2$ and $\tau_2 2$ in the same direction, that is, within each center, the mean of the treatment group is always set to be bigger than the mean of the control group.

Table 19 shows that estimates of σ vary little when using the EM algorithm with or without assuming interactions regardless of whether interaction truly exists or not. In Scenario 1 and 2, two centers have the same treatment effect for both center effect small or moderate. The EM procedure assuming no interaction as introduced in Section 5.2.1 has smaller bias and MSE on estimating σ than assuming interaction exists as introduced in Appendix C1. Also take Scenario 3 and 4 as examples, either quantitative or qualitative interaction exists in two center's observations. If we use the EM procedure without assuming interaction, it shows in Table 19 that the bias and the MSE of σ is slightly smaller than we use the EM procedure assuming interactions. It is surprising that even interaction does exist in the data the estimate of σ by assuming no interaction in the EM algorithm is close to the estimate assuming interaction. If the data are unblinded, using additive model when interaction exists supposes to inflate the estimate. The inflation does not happen here could be caused by the non-identifiability of the parameters as we stated in Section 5.2.2.

Besides the simulation results we have shown in Table 19, we also consider the scenarios with more exaggerating interactions between the center and the treatment in two centers. In those cases, we still can use the EM algorithm assuming no interaction, the simulation shows similar results as in Table 19 that the estimation does not differ much from the EM algorithm assuming interaction. At this point, the results of this simulation need further explanation in the context. We include Appendix C as a indication of how one might proceed to handle interaction using blinded data.

Table 19: Comparing EM estimates of σ with and without assuming interaction in a two-center trial for four different scenarios. 1000 samples with sample size $N_1 = 80$ are generated for each scenario.

EM algorithm	Scenario		without interaction	with interaction
Conventional	S1: no interaction $c_2 = 0.1$	Bias	-0.1279	-0.1595
		Variance	0.0249	0.0190
		MSE	0.0413	0.0445
	S2: no interaction $c_2 = 0.5$	Bias	-0.1384	-0.1645
		Variance	0.0246	0.0194
		MSE	0.0437	0.0465
	S3: quantitative interaction	Bias	-0.1496	-0.1783
		Variance	0.0234	0.0184
		MSE	0.0458	0.0502
	S4: qualitative interaction	Bias	-0.1374	-0.1666
		Variance	0.0228	0.0180
		MSE	0.0417	0.0457
Enhanced	S1: no interaction $c_2 = 0.1$	Bias	-0.1804	-0.1891
		Variance	0.0154	0.0147
		MSE	0.0479	0.0505
	S2: no interaction $c_2 = 0.5$	Bias	-0.1876	-0.1955
		Variance	0.0147	0.0143
		MSE	0.0499	0.0525
	S3: quantitative interaction	Bias	-0.1981	-0.2061
		Variance	0.0145	0.0139
		MSE	0.0538	0.0563
	S4: qualitative interaction	Bias	-0.1890	-0.1984
		Variance	0.0133	0.0128
		MSE	0.0490	0.0521
Enhanced block size 4	S1: no interaction $c_2 = 0.1$	Bias	-0.0770	-0.0925
		Variance	0.0124	0.0111
		MSE	0.0183	0.0197
	S2: no interaction $c_2 = 0.5$	Bias	-0.0760	-0.0906
		Variance	0.0137	0.0123
		MSE	0.0195	0.0205
	S3: quantitative interaction	Bias	-0.0795	-0.0943
		Variance	0.0120	0.0108
		MSE	0.0183	0.0197
	S4: qualitative interaction	Bias	-0.0763	-0.0903
		Variance	0.0113	0.0103
		MSE	0.0171	0.0185

BIBLIOGRAPHY

[1] Basford, K. and McLachlan, G. (1985). Likelihood estimation with normal mixture models. *Journal of Applied Statistics*, 34:282–289.

[2] Birkett, M. and Day, S. (1994). Internal pilot studies for estimating sample size. *Statistics in Medicine*, 22:2455–2463.

[3] Bohning, D., Schlattmann, P., and Lindsay, B. (1992). Computer-assisted analysis of mixtures: Statistical algorithms. *Biometrics*, 48:283–303.

[4] Campbell, M. and Swinscow, T. (2009). *Statistics Square One*. Wiley-Blackwell.

[5] Casella, G. and Berger, R. (2002). *Statistical Inference*. Duxbury.

[6] Chang, M. (2007). *Design and Analysis of Experiments with SAS*. Chapman and Hall.

[7] Chen, X., Dempster, A. P., and Liu, J. (1994). Weighted finite population sampling to maximize entropy. *Biometrika*, 81.

[8] Chen, X. and Liu, J. (1997). Statistical application of the poisson-binomial and conditional Bernoulli distributions. *Statistica Sinica*, 7.

[9] Chow, S.-C. and Chang, M. (2006). *Adaptive Design Methods in Clinical Trials*. Chapman and Hall.

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

[11] Friede, T. and Kieser, M. (2001). A comparison of methods for adaptive sample size adjustment. *Statistics in Medicine*, 20:3861–3873.

[12] Friede, T. and Kieser, M. (2002). On the inappropriateness of an EM algorithm based procedure for blinded sample size re-estimation. *Statistics in Medicine*, 21:165–176.

[13] Gail, M. H., Lubin, J. H., and Rubinstein, L. V. (1981). Likelihood calculation for matched case-control studies and survival studies with tied death times. *Biometrika*, 68.

[14] Gould, L. (1992). Interim analyses for monitoring clinical trials that do not materially affect the type I error rate. *Statistics in Medicine*, 11:55–66.

[15] Gould, L. A. and Shih, W. J. (1992). Sample size re-estimation without unblinding for normally distributed outcomes with unknown variance. *Communications in Statistics*, 21(10):2833–2853.

[16] Kieser, M. and Friede, T. (2000). Re-calculating the sample size in internal pilot study designs with control of the type I error rate. *Statistics in Medicine*, 19:901–911.

[17] Kieser, M. and Friede, T. (2003). Simple procedures for blinded sample size adjustment that do not affect the type I error rate. *Statistics in Medicine*, 22:3571–3581.

[18] Lawson, J. (2010). *Adaptive Design Theory and Implementation Using SAS and R.* Chapman and Hall.

[19] Lindsay, B. (1989). Review of mixture models: Inference and applications to clustering by Mclachlan and Basford. *Journal of the American Statistical Association*, 84:337–338.

[20] McLachlan, G. and Krishnan, T. (2008). *The EM algorithm and Extensions.* Wiley.

[21] McLachlan, G. and Peel, D. (2000). *Finite Mixture Models.* Wiley.

[22] Miller, F. (2005). Variance estimation in clinical studies with interim sample size reestimation. *Biometrics*, 61:355–361.

[23] Nettleton, D. (1999). Convergence properties of the EM algorithm in constrained parameter spaces. *The Canadian Journal of Statistics*, 27.

[24] Proschan, M. (2005). Two-stage sample size re-estimation based on a nuisance parameter: A review. *Journal of Biopharmaceutical Statistics*, 15:559–574.

[25] Proschan, M. (2009). Sample size re-estimation in clinical trials. *Biometrical Journal*, 51:348–357.

[26] Proschan, M., Lan, K., and Wittes, J. (2007). *Statistical Monitoring of Clinical Trials: A Unified Approach.* Springer.

[27] Proschan, M., Liu, Q., and Hunsberger, S. (2003). Practical midcourse sample size modification in clinical trials. *Controlled Clinical Trials*, 24.

[28] Proschan, M. and Wittes, J. (2000). An improved double sampling procedure based on the variance. *Biometrics*, 56:1183–1187.

[29] Redner, R. and Walker, H. (1984). Mixtures densities, maximum likelihood and the EM algorithm. *SIAM review*, 26.

[30] Rosenberger, W. F. and Lachin, J. M. (2002). *Randomization in clinical trials: theory and practice.* Wiley.

[31] Seidel, W., Mosler, K., and Alker, M. (2000). A cautionary note on likelihood ratio tests in mixture models. *Annals of the Institute of Statistical Mathematics*, 52:481–487.

[32] Senn, S. (1998). Some controversies in planning and analysing multi-center trials. *Statistics in medicine*, 17.

[33] Stein, C. (1945). A two-sample test for a linear hypothesis whose power is independent of the variance. *Annals of Mathematical Statistics*, 16:243–258.

[34] Stephens, M. (2000). Dealing with label switching in mixture models. *Journal of Royal Statistical Society Ser. B.*, 62.

[35] Waksman, J. (2007). Assessment of the Gould-Shih procedure for sample size re-estimation. *Pharmaceutical Statistics*, 6:53–65.

[36] Wittes, J. and Brittain, E. (1990). The role of internal pilot studies in increasing the efficiency of clinical trials. *Statistics in Medicine*, 9:65–72.

[37] Wittes, J., Schabenberger, O., Zucker, D., Brittain, E., and Proschan, M. (1999). Internal pilot studies I: Type I error rate of the naive t-test. *Statistics in Medicine*, 18:3481–3491.

[38] Xing, B. and Ganju, J. (2005). A method to estimate the variance of an endpoint from an on-going blinded trial. *Statistics in Medicine*, 24:1807–1814.