Bias in Estimation of a Mixture of Normal Distributions

Spencer Lourens, Ying Zhang
Department of Biostatistics
University of Iowa, Iowa City, IA, 52442,
Jeffrey D. Long
Departments of Psychiatry and Biostatistics
University of Iowa, Iowa City, IA, 52442
and
Jane S. Paulsen
Departments of Psychiatry, Neurology and Psychology
University of Iowa, Iowa City, IA, 52442

October 14, 2013

Abstract

Estimating parameters in a mixture of normal distributions dates back to the 19th century when Pearson originally considered data consisting of ratios of forehead to body length of crabs from the Bay of Naples. Since then, many real world applications of mixtures have led to various proposed methods for studying similar problems. Among them, maximum likelihood estimation (MLE) and the continuous empirical characteristic function (CECF) methods have drawn the most attention. However, the performance of these competing estimation methods has not been thoroughly studied in the literature and conclusions have not been consistent in published research. In this article, we review this classical problem with a focus on estimation bias. An extensive simulation study is conducted to compare the estimation bias between the MLE and CECF methods over a wide range of disparity values. We use the overlapping coefficient (OC) to measure the amount of disparity, and provide a practical guideline for estimation quality in mixtures of normal distributions. Application to an ongoing multi-site Huntingtons disease study is illustrated for ascertaining cognitive biomarkers of disease progression.

keywords: Biomarkers; Disparity Index; EM algorithm; Mixtures;

1 Introduction

Mixture (or mixing) distributions refer to composite distributions constructed by mixing a number (K) of component distributions. Estimation of mixture distributions is a classical statistical problem which has been studied for over 100 years. The first account of mixture data being analyzed was documented by Pearson [12] in 1894, in which a series of equations were derived in order to estimate parameters denoting crab characteristics in the case where the number of components is two (K = 2). The word mixture is used because the density function of a random observation is a mixing of several (unique) component density functions of the form $f(x) = \sum_{i=1}^{K} \eta_i f_i(x)$, with $\sum_{i=1}^{K} \eta_i = 1$ and $f_i(x), \;\; i = 1, \ldots, K$ being unique densities with known form. Each observation comes from one of the K (unique) component distributions with unknown membership status. The aim of mixture modeling is to estimate the parameters of each component density, f_i, as well as the mixing

parameters, η_i. This problem can be also regarded as a missing data problem, as the group membership or the component distribution from which an observation is generated is not known. We'd not only like to be able to estimate the parameters from a mixture, but also to gain an understanding of estimation performance over a wide range of settings. Mixtures have been used to analyze data arising in Hydrology [9], Economics [19], Ecology [12], as well as many other fields [3]. Partial mixtures, for which group membership is known for a certain subjects in the data set, have also been considered [5].

Several estimation methods for normal mixtures have been proposed in the literature. Among them, the ordinary maximum likelihood estimation method (MLE) appears to be a straightforward choice, as the likelihood for the mixture is easy to establish. However, directly calculating the MLE via optimizing the likelihood for a mixture of normal distributions is difficult and numerical algorithms can lead to computational issues such as non-convergence, as noted in [19]. Hosmer [5] developed an estimation method for the case K = 2, which can be viewed as a special case of the well-known Expectation-Maximization (EM) algorithm for computing the MLE in missing data problems [2]. He found that the MLE may not perform well with regards to bias in the small sample case, especially when the two distributions are poorly separated. Leytham [9] corroborated Hosmer's work in regards to the estimation of the means and variances in normal mixtures through simulation, but claimed that estimation of quantiles for normal mixtures may be approximately unbiased. Moreover, results regarding the MLE for normal mixtures are inconsistent in the literature. Some researchers report unbiased estimation via the MLE [10], but others conclude otherwise [9, 19]. Mixtures of normal distributions may have a model identifiability issue and the likelihood can be unbounded for some special cases [1, 20]. This means the EM algorithm may be converging to a local maximum of the likelihood and may then yield biased estimation for model parameters of interest. In response to the unbounded likelihood of normal mixtures, alternative methods such as the Moment Generating Functions method (MGF) [13], the Discrete Empirical Characteristic Function method (DECF) [17], and the Continuous Empirical Characteristic Function method (CECF) [19] have been proposed. They can be viewed as special cases of Generalized Methods of Moments (GMM) [20]. Limited simulation studies have provided some numerical evidence for the merit of these methods compared to MLE [19]. The question is whether this set of GMM methods performs better than the MLE in general.

Our study of mixture distributions is motivated by empirical issues with progression marker data in Huntington Disease (HD). HD is an autosomal dominant neurodegenerative disease caused by the trinucleotide cytosine-adenine-guanine (CAG) expansion in the gene of the protein huntingtin. Clinical symptoms of HD include progressive motor dysfunction, cognitive decline, and psychiatric disturbance [6]. Individuals who have a CAG repeat of length 36 or greater are referred to as at-risk of HD, and individuals who are at-risk of HD but have not yet received the HD motor diagnosis are described as being prodromal-HD (prHD). As the disease progresses, prHD individuals exhibit impairments noted above, often with daily functioning deterioration as

a result, and patients with larger CAG repeats often deteriorate at much faster rates than those with smaller CAG repeats. Although the CAG expansion is vital for determining the at-risk status of an individual, it is estimated that fewer than 5% of those at-risk (i.e. having at least one parent diagnosed with HD in their lifetime) of having an expansion length of 36 or greater are willing to undergo the genetic testing to ascertain their at-risk status. This is because knowing their at-risk status may largely affect their family lives, employment opportunities, and insurance coverage, in addition to adding psychological disturbance [18]. As a result, at-risk status information is not always known, and in the past, researchers have used proxies for at-risk status. For example, [8] used the information that at least one parent was diagnosed with HD as a proxy for at-risk status. However, using such information as a proxy can result in biased estimation and invalid inference for understanding disease progression in HD. We believe that discovering critical HD progression biomarkers to serve as a proxy for the at-risk status of HD for individuals who are unwilling to undergo gene testing is profound, because these individuals can receive a timely treatment, if available, and avoid the aforementioned negative impacts associated with a positive result from the gene test for HD. Essentially, the task is to study the distributions of potential HD progression biomarkers for both HD at-risk (prHD) and healthy control cohorts when the information of CAG is unknown.

In this article, we review the existing methods for estimating normal mixtures and conduct an overarching numerical experiment to examine their estimation performance with focus on comparing the bias between the MLE via EM algorithm and CECF method. In addition to the numerical experiment, we apply the methods to HD data from the PREDICT-HD study. Intuitively, the estimation of a mixture distribution should be largely influenced by the disparity between the component distributions. We use the OC [7] as a disparity index for quantifying the difference between the two component distributions, and we study the estimation performance over a wider range of this disparity index than considered by previous authors. We aim to provide a practical guide for the validity of the methods in terms of estimation bias in relation to the disparity index.

The rest of the paper is organized as follows. Section 2 provides an overview of the competing methods proposed in the literature and provides a disparity index to quantify the difference between two distributions. Section 3 presents an extensive simulation study comparing the performance of the MLE via the EM algorithm and the CECF method under various settings, along with the index values. Section 4 applies this index to PREDICT-HD data to ascertain HD cognitive biomarkers as potential proxy variables for HD at-risk status. Section 5 gives our concluding remarks and some guidance regarding analyzing normal mixtures.

2 Overview of the Methods for Estimation of a Normal Mixture

In this section, we provide an overview of the estimation methods mentioned in Section 1, specifically in the case that the data come from a mixture of two component normal distributions. Suppose we observe a random sample of continuous outcomes $Y_1, Y_2, \cdots, Y_n$ that are distributed according to a normal mixture of $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$. Let $\mathbf{R_1}$ and $\mathbf{R_2}$ denote the two latent groups, $D_i = 1[Y_i \in \mathbf{R_2}]$ the indicator for outcome Y_i coming from Group 2 for $i = 1, 2, \cdots, n$, and $\eta = P(Y_i \in \mathbf{R_2})$, the probability that Y_i comes from Group 2. In the mixture problem, the information of group membership $D_1, D_2, \cdots, D_n$ is unknown and η, the mixing parameter, must also be regarded as an unknown parameter in the analysis. For the two component normal mixture, the probability density function (PDF) for Y_i, $i = 1, 2, \cdots, n$ is:

$$f_{y_i}(y) = \eta f_2(y) + (1 - \eta) f_1(y) \quad i = 1, 2, \cdots, n \tag{1}$$

where

$$f_j(y) = \frac{1}{\sqrt{2\pi}\sigma_j} \exp\left(-\frac{(y - \mu_j)^2}{2\sigma_j^2}\right) \quad j = 1, 2.$$

2.1 MLE via the EM Algorithm

Although the likelihood for the unknown parameter $\theta = (\eta, \mu_2, \mu_1, \sigma_2^2, \sigma_1^2)$ can be easily established for the observed data with the PDF given in (1), the numerical algorithm for computing the MLE is not stable, as demonstrated in [19]. In a mixture setting, we do not observe the complete data, (Y_i, D_i), for each subject, as component membership (D_i) is missing for all individuals under study. This means that mixture problems can be considered missing data problems. Since this is a missing data problem, the EM algorithm is a natural alternative for computing the MLE. To apply the EM algorithm, the "complete" data likelihood is formed as if D_i are observed. It turns out the log complete likelihood is a linear function of unobserved data D_i for $i = 1, 2, \cdots, n$. Hence, the conditional expectation of each latent observation D_i, given the observed data and current estimate of unknown parameters, needs to be evaluated and then be substituted into the (log) complete likelihood for D_i. The EM algorithm is particularly effective for this situation, because both the E-step and the M-step have explicit solutions as given by [9, 10]. We briefly present their solutions here. Given a current estimate of θ, $\hat{\theta}_{(k)} = (\hat{\eta}_{(k)}, \hat{\mu}_{2,(k)}, \hat{\mu}_{1,(k)}, \hat{\sigma}^2_{2,(k)}, \hat{\sigma}^2_{1,(k)})$, the estimate $\hat{\theta}_{(k+1)}$ can be explicitly updated by

$$\hat{\eta}_{(k+1)} = \frac{\sum_{i=1}^{n} q_{ik}}{n}$$
$$\hat{\mu}_{2,(k+1)} = \frac{\sum_{i=1}^{n} q_{ik} y_i}{\sum_{i=1}^{n} q_{ik}}$$
$$\hat{\mu}_{1,(k+1)} = \frac{\sum_{i=1}^{n} (1-q_{ik}) y_i}{\sum_{i=1}^{n} (1-q_{ik})}$$
$$\hat{\sigma}^2_{2,(k+1)} = \frac{\sum_{i=1}^{n} q_{ik} (y_i - \hat{\mu}_{2,(k)})^2}{\sum_{i=1}^{n} q_{ik}}$$
$$\hat{\sigma}^2_{1,(k+1)} = \frac{\sum_{i=1}^{n} (1-q_{ik}) (y_i - \hat{\mu}_{1,(k)})^2}{\sum_{i=1}^{n} (1-q_{ik})}$$

where

$$q_{ik} = E(D_i | y_i; \hat{\theta}_{(k)}) = \frac{\hat{\eta}_{(k)} \hat{f}_{2,(k)}(y_i)}{\hat{\eta}_{(k)} \hat{f}_{2,(k)}(y_i) + (1-\hat{\eta}_{(k)}) \hat{f}_{1,(k)}(y_i)}$$

with $\hat{f}_{j,(k)}(y) = \frac{1}{\sqrt{2\pi}\hat{\sigma}_{j,(k)}} \exp\left(-\frac{(y-\hat{\mu}_{j,(k)})^2}{2\hat{\sigma}^2_{j,(k)}}\right)$ for $j = 1, 2$. By choosing an arbitrary initial value, $\hat{\theta}_{(0)}$, this iterative procedure can be easily implemented and forced to stop when the difference of the estimates in adjacent iterations is sufficiently small, say less than $10e^{-6}$.

2.2 The CECF method

Motivated by the estimation method involving minimizing a distance between the empirical characteristic function and the population-based characteristic function originally proposed by Heathcote [4], Xu and Knight [19] developed the CECF method. For observed continuous outcomes $\mathbf{y} = (y_1, y_2, \cdots, y_n)$, they consider minimizing

$$c(\theta, \mathbf{y}) = \int_{-\infty}^{+\infty} \|C_n(\mathbf{y}, r) - C(\theta, r)\|^2 G(r) dr \tag{2}$$

where $C_n(\mathbf{y}, r) = \sum_{i=1}^{n} \exp(iry_i)/n$ denotes the empirical characteristic function, $C(\theta, r) = E(\exp(itY))$ the characteristic function, and $G(r)$ a weight function. Specifically, for a normal mixture,

$$C(\theta, r) = \eta \exp(i\mu_2 r - 1/2\sigma_2^2 r^2) + (1-\eta) \exp(i\mu_1 r - 1/2\sigma_1^2 r^2).$$

Therefore, the choice of $G(r) = \exp(-br^2)$ makes (2) integrable and results in an explicit function of unknown parameter θ and the tuning parameter b. Heathcote [4] did not consider the optimal choice of b, and instead set it to 1, as was commonly done in the past for this type of problem. For a given value b, the minimization problem (2) is straightforward. Xu and Knight [19] chose the optimal b by iteratively solving for the θ value that minimizes (2) at a given b and updating the value b at the value which minimizes the trace (or determinant) of the resulting variance matrix for the current estimate of θ. This procedure continues until the change in the optimal θ values is sufficiently small. They demonstrated, in a limited simulation study, that the CECF

method is comparable to the standard MLE in terms of estimation efficiency. They also showed that the CECF may still be a valid estimation method in cases when the MLE is problematic.

2.3 The DECF Method

Similar to the CECF method, the DECF method considers minimizing the distance between sample quantities and population analogs over a fixed set of grid points, $\mathbf{r} = (r_1, r_2 \ldots, r_m)$. That is, the unknown parameter θ is estimated by minimizing

$$e(\theta, \mathbf{y}, \mathbf{r}) = \sum_{i=1}^{m} \|C_n(\mathbf{y}, r_i) - C(\theta, r_i)\|^2 \tag{3}$$

where $C_n(\mathbf{y}, r_i)$ and $C(\theta, r_i)$ are the same as defined for the CECF method. The performance of the DECF methods depends on the choice of grid points $\mathbf{r}$, both the number and location of the nodes, $r_i, i = 1, \cdots, m$. Work has been done to show that as the grid becomes finer and more extended, the DECF becomes more efficient [20]. As also noted in [17, 20], the DECF method is actually a special case of Generalized Methods of Moments (GMM). Multiple authors [17, 20] have noted that the estimation efficiency of the DECF could be increased by rescaling the weight matrix used by GMM. (the weight matrix is the identity matrix in the presentation above) The CECF method is generally preferred over the DECF, as the distance is defined over the whole continuum of r values in $(-\infty, \infty)$ and hence, it does not require specification of the grid points $\mathbf{r}$.

2.4 The MGF method

The MGF method developed in [13] is very similar in nature to the DECF method. The only difference is that the moment generating function is used to replace the characteristic function of the DECF method. That is, the unknown parameter θ is estimated by minimizing

$$m(\theta, \mathbf{y}, \mathbf{r}) = \sum_{i=1}^{m} (M_n(\mathbf{y}, r_i) - M(\theta, r_i))^2 \tag{4}$$

where $M_n(\mathbf{y}, r) = \sum_{i=1}^{n} \exp(r y_i)/n$ and $M(\theta, r) = \eta \exp(\mu_2 r - 1/2\sigma_2^2 r^2) + (1 - \eta) \exp(\mu_1 r - 1/2\sigma_1^2 r^2)$. Again, choice of the number of points and their location must be made to facilitate the use of this method. Moreover, the possible non-existence of the moment generating function for fat-tailed distributions (i.e. Cauchy) makes the MGF method less desirable than the DECF method in practice [17].

2.5 A Disparity Index

For a mixture distribution, estimation quality largely depends on the difference between the component distributions. If the distributions have a large overlap it will be difficult to identify the group membership of observations and to estimate each component's parameters. Therefore, it is highly desirable to define an index which measures the difference between the two latent distributions in order to develop a guideline regarding estimation quality for a mixture distribution.

For normal mixtures, Hosmer [5] defined an index,

$$\mathrm{H} = \frac{\|\mu_2 - \mu_1\|}{\min(\sigma_2, \sigma_1)}$$

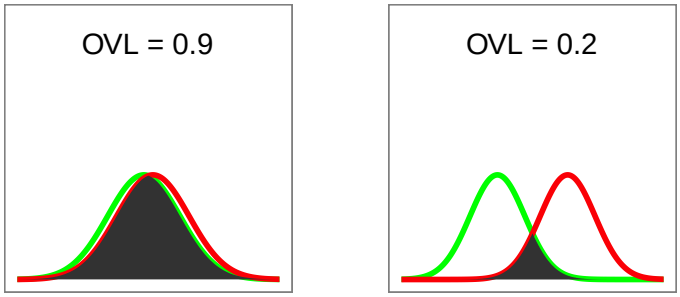
to measure the separation between the two normal distributions. This measure, however, cannot capture divergence of the two latent normal distributions due to a difference in variance alone. The simulation study conducted in [5] only considered the performance of MLE for the case of $\mu_1 \neq \mu_2$. Whenever $\mu_1 = \mu_2$, $\mathrm{H} \equiv 0$ regardless of what the variances equal. For the case of $\sigma_2 > \sigma_1$, with σ_2 increasing, it will be shown via simulation that estimation quality improves, eventually resulting in negligible bias. However, the value of H will not change in this situation, thus, H does not properly index the observed improvement in estimation performance. As a result of these observations, the proper term for describing the difference between two normal distributions that make up the mixture distribution is "disparity". The disparity between two distributions not only accounts for mean separation, but also for differences in variability. One measure that considers both the means and the variances is Nityasuddhi's D [10], which is defined as ,

$$\mathrm{D} = \frac{1}{2}\sum_{i=1}^{2}(\mu_i - \bar{\mu})^2 + \sum_{i=1}^{2}(\sigma_i^2 - \bar{\sigma}^2)^2$$

where $\bar{\mu} = (\mu_1 + \mu_2)/2$ and $\bar{\sigma}^2 = (\sigma_1^2 + \sigma_2^2)/2$. However, this index can yield similar values for two opposing cases in which estimation quality will be very different. For instance, similar D values may result due to a difference in means, while the variances are the same, or due to a difference in variances, while the means are the same. That is to say, the same D value may be observed when only the variances differ, or when only the means differ. Our simulation shows that much smaller differences in means are necessary for estimation to have negligible bias, while differences in variances must be larger for estimation to have negligible bias. Thus, two different underlying parameter values may yield the same Nityasuddhi's D value, even if estimation performance varies substantially in both cases.

Ideally, a good disparity index should always have a large value when estimation quality is good, and a small value when estimation is bad. Intuitively, the shared (or overlapping) area under the two normal distributions is key to determining the estimation quality, as the observations from this area obscure their group membership.

Figure 1: OC Example Plots



Distributions with little overlap tend to be easily separated and result in parameter estimation with small bias. However, for mixtures where the component distributions have large overlap, severe bias might result. Inman and Edwin [7] have studied the OC for the case of normal distributions and derived an explicit formula to calculate its value,

$$\text{OC} = \begin{cases} 2\Phi\left(-\frac{|\mu_1-\mu_2|}{\sigma}\right) & \text{if } \sigma_1 = \sigma_2 = \sigma \\ 1 + \Phi\left(\frac{X_1-\mu_1}{\sigma_1}\right) + \Phi\left(\frac{X_2-\mu_2}{\sigma_2}\right) - \Phi\left(\frac{X_1-\mu_2}{\sigma_2}\right) - \Phi\left(\frac{X_2-\mu_1}{\sigma_1}\right) & \text{if } \sigma_1 \neq \sigma_2. \end{cases} \tag{5}$$

In (5), Φ denotes the cumulative distribution function of the standard normal distribution, X_1 and X_2 are given by

$$X_1 = \frac{\mu_1\sigma_2^2 - \mu_2\sigma_1^2 - \sigma_1\sigma_2\left[(\mu_1-\mu_2)^2 + (\sigma_2^2-\sigma_1^2)\log\left(\frac{\sigma_2^2}{\sigma_1^2}\right)\right]^{1/2}}{\sigma_2^2-\sigma_1^2}$$

and

$$X_2 = \frac{\mu_1\sigma_2^2 - \mu_2\sigma_1^2 + \sigma_1\sigma_2\left[(\mu_1-\mu_2)^2 + (\sigma_2^2-\sigma_1^2)\log\left(\frac{\sigma_2^2}{\sigma_1^2}\right)\right]^{1/2}}{\sigma_2^2-\sigma_1^2}.$$

We propose the use of DI = 1 - OC as a disparity index. Note that DI satisfies our requirements specified in the above paragraph. Namely, large values of DI (large disparity) reflect cases where estimation quality is good, and small values of DI (small disparity) reflect cases where estimation quality is poor. This index does not suffer from the sub-optimal properties of the indices mentioned above, as it allows for variances alone to contribute to the disparity between the two normal distributions. Two examples of normal mixture distributions, one with large disparity (DI = 0.8) and one with small disparity (DI = 0.1), are shown in Figure 1 for illustration purposes. The shaded portion depicts the overlap. Notice that small values of OC reflect cases where there is large disparity between the component distributions, and large values reflect cases where there is small disparity.

Table 1: Simulation Settings

Case	DI	η	μ_2	μ_1	σ_2^2	σ_1^2
A1	0.1	0.5	1.25	1	1	1
A2	0.3	0.5	1.77	1	1	1
A3	0.55	0.5	2.51	1	1	1
A4	0.8	0.5	3.56	1	1	1
B1	0.1	0.5	1	1	1.5	1
B2	0.3	0.5	1	1	3.6	1
B3	0.55	0.5	1	1	13.2	1
B4	0.8	0.5	1	1	101	1
C1	0.1	0.5	1.1	1	1.48	1
C2	0.3	0.5	1.5	1	3.22	1
C3	0.55	0.5	2.2	1	11.5	1
C4	0.8	0.5	6.55	1	18.0	1

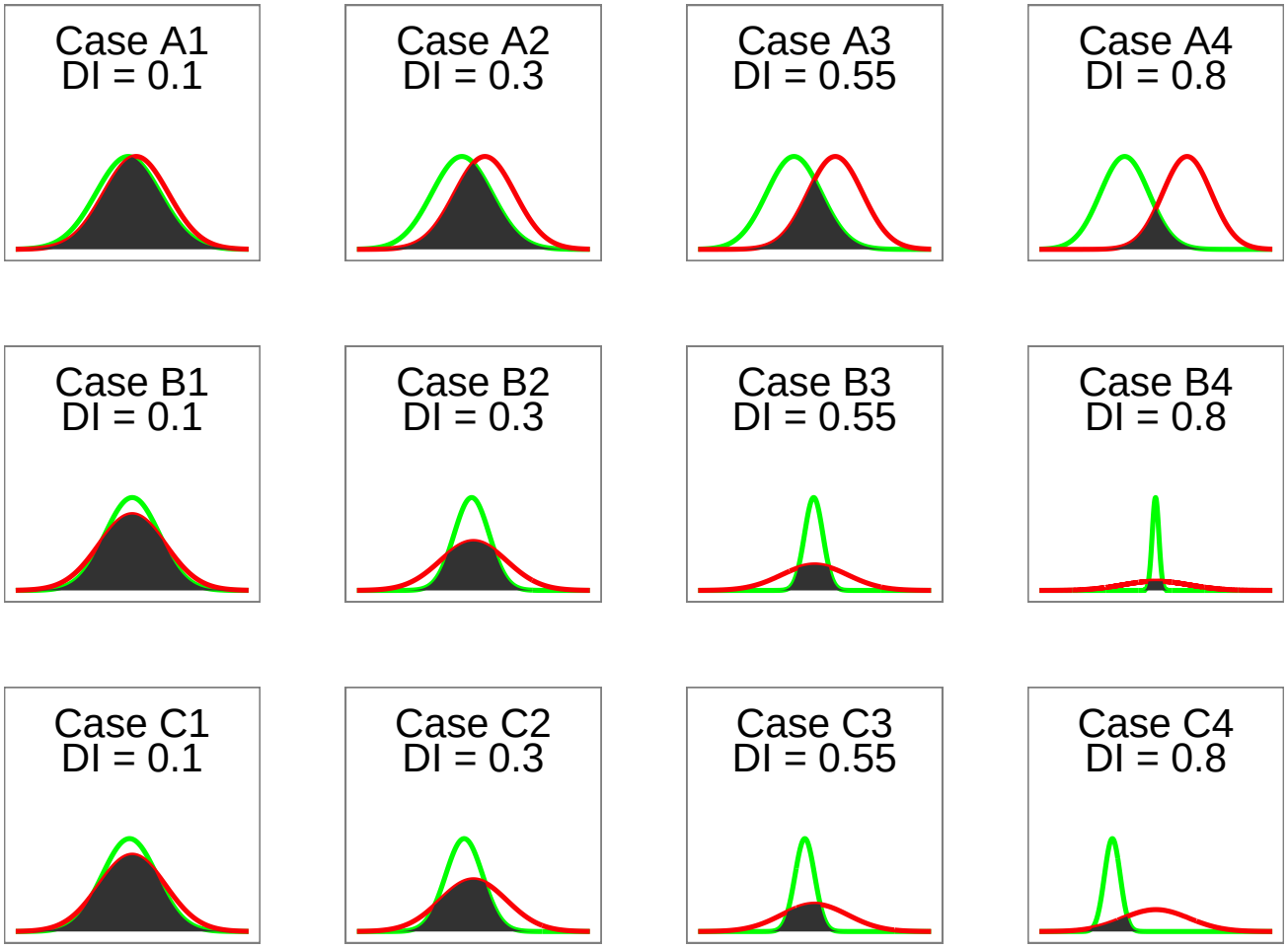
3 Simulation Study

In this section, we describe a simulation study to examine the estimation performance of the methods discussed above with focus on the estimation bias. As the CECF, DECF, and MGF methods are very similar in nature and the CECF has the merit of not requiring the identification of the optimal number and locations of the grid points, we only include the CECF in the study and compare it to the MLE (via the EM algorithm). The DI is used to quantify the amount of disparity between the two component distributions. Our simulation conditions are summarized in Table 1. They describe the amount of disparity due to mean and variance differences in various scenarios. Figure 2 provides a visual display of the two component distributions for the cases given in Table 1 with the shading depicting distribution overlap. It is worth noting that our study covers a much broader range of disparity values than any other studies conducted in the literature [5, 10, 19].

For each case listed in Table 1, we conducted a Monte Carlo simulation study with 1000 trials. Hence, the estimation bias and Monte Carlo Standard Deviation (MCSD) was calculated based on the results from the 1000 trials. Relative bias and relative MCSD are reported in the table. Relative bias is defined as the bias divided by the value of the parameter being estimated. Relative MCSD is defined in an analogous manner. For example, if the bias and MCSD are 0.05 and 1.50, respectively, and the parameter value is 2, then the relative bias is 0.025 and relative MCSD is 0.75. It is worth noting that for approximately 5% of the trials for cases with the least disparity, the EM algorithm for the MLE did not lead to numerical convergence, while the CECF method did not have any numerical problems. When this occurred, data were regenerated and the simulation continued until 1000 trials were completed. This convergence issue was not observed in cases with at least moderate disparity between the two component distributions.

Table 2 summarizes the simulation results for Case A with sample sizes 100, 200 and 500. It appears that there is substantial bias in estimating both mean and variance parameters using the MLE when the DI is less than or equal to 0.3. When DI=0.3 the CECF has smaller bias in estimating the means but larger larger bias in

Figure 2: Graphical Representation of Simulation Settings



Note: The above graphs are not on the same scale.

Table 2: Simulation Results - Bias/MCSD : Case A

		Bias (MCSD) of Estimation MLE/CECF Case A									
		MLE method					**CECF method**				
Case	n	η	μ_2	μ_1	σ_2^2	σ_1^2	η	μ_2	μ_1	σ_2^2	σ_1^2
A1	100	0.026	0.417	-0.519	-0.374	-0.383	-0.061	0.369	-0.363	-0.398	-0.352
		(0.597)	(0.574)	(0.664)	(0.424)	(0.443)	(0.654)	(0.509)	(0.570)	(0.586)	(0.562)
DI = 0.1	200	0.031	0.315	-0.436	-0.275	-0.301	-0.016	0.347	-0.312	-0.338	-0.357
		(0.566)	(0.515)	(0.614)	(0.414)	(0.412)	(0.682)	(0.446)	(0.510)	(0.549)	(0.510)
	500	0.119	0.203	-0.441	-0.143	-0.266	-0.014	0.260	-0.313	-0.321	-0.309
		(0.535)	(0.406)	(0.590)	(0.396)	(0.393)	(0.714)	(0.430)	(0.518)	(0.493)	(0.583)
A2	100	0.033	0.181	-0.343	-0.321	-0.327	-0.019	0.105	-0.183	-0.336	-0.301
		(0.717)	(0.429)	(0.699)	(0.476)	(0.493)	(0.671)	(0.373)	(0.668)	(0.628)	(0.988)
DI = 0.3	200	0.063	0.129	-0.258	-0.212	-0.221	0.003	0.066	-0.134	-0.285	-0.261
		(0.601)	(0.411)	(0.644)	(0.443)	(0.524)	(0.698)	(0.351)	(0.594)	(0.587)	(0.698)
	500	0.110	0.074	-0.236	-0.106	-0.173	0.002	0.044	-0.091	-0.257	-0.255
		(0.560)	(0.396)	(0.608)	(0.413)	(0.448)	(0.734)	(0.305)	(0.569)	(0.519)	(0.510)
A3	100	-0.014	0.078	-0.117	-0.202	-0.156	-0.018	0.025	-0.040	-0.228	-0.169
		(0.585)	(0.309)	(0.689)	(0.508)	(0.538)	(0.631)	(0.278)	(0.679)	(0.635)	(0.709)
DI = 0.55	200	-0.024	0.068	-0.079	-0.127	-0.077	0.010	0.005	-0.062	-0.146	-0.136
		(0.580)	(0.295)	(0.638)	(0.479)	(0.490)	(0.649)	(0.257)	(0.646)	(0.629)	(0.599)
	500	-0.049	0.052	-0.026	-0.088	-0.023	0.016	0.001	-0.035	-0.098	-0.100
		(0.556)	(0.241)	(0.557)	(0.407)	(0.414)	(0.602)	(0.220)	(0.560)	(0.493)	(0.474)
A4	100	-0.009	0.006	-0.007	-0.040	-0.024	-0.014	-0.004	0.038	-0.005	0.041
		(0.354)	(0.133)	(0.485)	(0.492)	(0.501)	(0.420)	(0.143)	(0.518)	(0.680)	(0.708)
DI = 0.8	200	-0.010	0.006	0.005	-0.013	0.009	-0.021	0.004	0.028	-0.013	0.046
		(0.260)	(0.096)	(0.345)	(0.361)	(0.373)	(0.302)	(0.100)	(0.376)	(0.459)	(0.502)
	500	-0.001	0.001	0.003	-0.002	0.011	-0.013	0.003	0.019	-0.010	0.040
		(0.163)	(0.058)	(0.211)	(0.219)	(0.230)	(0.189)	(0.063)	(0.240)	(0.267)	(0.301)

Table 3: Simulation Results - Bias/MCSD : Case B

		Bias (MCSD) of Estimation MLE/CECF Case B									
		MLE method					**CECF method**				
Case	n	η	μ_2	μ_1	σ_2^2	σ_1^2	η	μ_2	μ_1	σ_2^2	σ_1^2
B1	100	-0.075	0.005	-0.038	-0.287	-0.326	0.043	0.005	-0.471	0.055	-0.493
		(0.571)	(1.283)	(0.590)	(0.463)	(0.403)	(0.681)	(1.073)	(0.590)	(0.678)	(0.424)
DI = 0.1	200	-0.021	-0.045	-0.035	-0.211	-0.288	0.122	0.030	-0.497	0.076	-0.488
		(0.580)	(1.199)	(0.567)	(0.459)	(0.379)	(0.712)	(0.973)	(0.569)	(0.527)	(0.433)
	500	-0.055	-0.070	-0.082	-0.132	-0.164	0.142	0.007	-0.490	0.097	-0.431
		(0.558)	(1.143)	(0.444)	(0.513)	(0.338)	(0.745)	(0.899)	(0.523)	(0.497)	(0.452)
B2	100	-0.069	-0.072	-0.0002	-0.135	-0.081	0.063	-0.009	0.017	-0.098	-0.283
		(0.552)	(1.595)	(0.414)	(0.514)	(0.589)	(0.630)	(1.287)	(0.467)	(0.686)	(0.634)
DI = 0.3	200	-0.060	0.043	-0.004	-0.010	-0.024	0.123	0.019	-0.002	-0.016	-0.251
		(0.488)	(1.218)	(0.267)	(0.436)	(0.521)	(0.608)	(0.973)	(0.348)	(0.652)	(0.618)
	500	-0.011	-0.012	0.0009	0.033	-0.022	0.123	-0.010	0.001	0.011	-0.176
		(0.370)	(0.414)	(0.147)	(0.266)	(0.370)	(0.510)	(0.515)	(0.228)	(0.436)	(0.481)
B3	100	-0.013	-0.002	0.013	-0.0008	0.024	0.020	-0.017	0.019	-0.010	-0.066
		(0.274)	(0.865)	(0.227)	(0.313)	(0.619)	(0.291)	(0.922)	(0.229)	(0.382)	(0.560)
DI = 0.55	200	-0.0010	0.013	0.004	0.003	0.006	0.025	0.010	0.005	-0.012	-0.050
		(0.178)	(0.385)	(0.155)	(0.203)	(0.359)	(0.194)	(0.424)	(0.165)	(0.243)	(0.390)
	500	0.0012	0.009	0.004	-0.007	0.004	0.011	0.003	0.004	-0.014	-0.020
		(0.109)	(0.243)	(0.095)	(0.123)	(0.195)	(0.119)	(0.269)	(0.099)	(0.141)	(0.225)
B4	100	-0.005	0.015	0.008	-0.024	-0.006	0.006	0.053	0.008	-0.050	-0.021
		(0.134)	(1.458)	(0.172)	(0.215)	(0.323)	(0.150)	(1.856)	(0.182)	(0.305)	(0.349)
DI = 0.8	200	0.0002	0.042	0.0004	-0.011	-0.006	0.004	0.048	0.0004	-0.024	-0.008
		(0.094)	(0.997)	(0.123)	(0.160)	(0.227)	(0.098)	(1.238)	(0.131)	(0.215)	(0.249)
	500	0.0001	0.028	0.002	-0.011	0.003	0.001	0.019	0.003	-0.016	0.003
		(0.061)	(0.623)	(0.080)	(0.099)	(0.136)	(0.063)	(0.775)	(0.086)	(0.134)	(0.152)

Table 4: Simulation Results - Bias/MCSD : Case C

		Bias (MCSD) of Estimation MLE/CECF Case C									
		MLE					**CECF method**				
Case	n	η	μ_2	μ_1	σ_2^2	σ_1^2	η	μ_2	μ_1	σ_2^2	σ_1^2
C1	100	-0.076	0.124	-0.109	-0.309	-0.341	0.038	0.030	0.043	-0.280	-0.497
		(0.586)	(1.191)	(0.610)	(0.470)	(0.396)	(0.678)	(0.990)	(0.604)	(0.671)	(0.415)
DI = 0.1	200	-0.028	0.084	-0.092	-0.207	-0.280	0.140	0.025	-0.0005	-0.247	-0.492
		(0.582)	(1.063)	(0.530)	(0.470)	(0.382)	(0.694)	(0.858)	(0.576)	(0.517)	(0.420)
	500	-0.032	0.091	-0.124	-0.121	-0.171	0.127	0.003	0.017	-0.242	-0.415
		(0.551)	(1.006)	(0.448)	(0.500)	(0.338)	(0.743)	(0.839)	(0.513)	(0.468)	(0.486)
C2	100	-0.121	0.302	-0.049	-0.197	-0.076	0.030	0.191	0.007	-0.149	-0.262
		(0.559)	(1.006)	(0.412)	(0.464)	(0.565)	(0.644)	(0.874)	(0.457)	(0.681)	(0.626)
DI = 0.3	200	-0.089	0.275	-0.022	-0.068	-0.033	0.079	0.204	-0.033	-0.117	-0.230
		(0.529)	(0.820)	(0.277)	(0.438)	(0.522)	(0.609)	(0.654)	(0.349)	(0.523)	(0.585)
	500	-0.063	0.136	-0.008	0.004	0.003	0.064	0.124	-0.028	-0.040	-0.141
		(0.418)	(0.449)	(0.176)	(0.291)	(0.391)	(0.523)	(0.474)	(0.229)	(0.401)	(0.469)
C3	100	-0.019	0.072	0.003	-0.030	0.027	-0.0001	0.114	-0.006	-0.057	-0.013
		(0.296)	(0.459)	(0.237)	(0.292)	(0.625)	(0.334)	(0.507)	(0.233)	(0.409)	(0.606)
DI = 0.55	200	-0.003	0.029	-0.011	-0.004	0.004	0.006	0.040	-0.010	-0.014	-0.005
		(0.190)	(0.208)	(0.157)	(0.198)	(0.373)	(0.231)	(0.276)	(0.167)	(0.278)	(0.429)
	500	-0.0004	0.010	-0.0008	-0.008	0.004	-0.005	0.015	0.001	0.004	0.018
		(0.115)	(0.115)	(0.099)	(0.122)	(0.210)	(0.134)	(0.140)	(0.103)	(0.163)	(0.253)
C4	100	-0.046	0.047	0.027	-0.100	0.055	-0.044	0.031	0.025	-0.158	0.040
		(0.158)	(0.145)	(0.183)	(0.249)	(0.400)	(0.193)	(0.180)	(0.194)	(0.343)	(0.433)
DI = 0.8	200	-0.022	0.025	0.007	-0.044	0.028	-0.024	0.019	0.010	-0.078	0.020
		(0.111)	(0.099)	(0.124)	(0.179)	(0.257)	(0.134)	(0.128)	(0.135)	(0.272)	(0.294)
	500	-0.011	0.012	0.005	-0.026	0.022	-0.016	0.014	0.006	-0.054	0.026
		(0.067)	(0.059)	(0.080)	(0.107)	(0.148)	(0.086)	(0.085)	(0.089)	(0.179)	(0.180)

estimating the variances compared to the MLE. Both methods lead to negligible bias for estimating the means in the case of DI=0.55, especially when sample size is greater than 200. But the estimation of the variances for the CECF is still seemingly biased even when n = 500, with 9.8% and 10% relative bias for σ_2^2 and σ_1^2, respectively. When there is a large amount of disparity between the two component distributions, for instance when DI=0.8 in our study, both MLE and CECF work very well. However, the MLE method outperforms the CECF method with regards to both estimation bias and standard error. This is not surprising, as the MLE is the efficient estimation method when it works.

Table 3 summarizes the simulation results for Case B with sample sizes 100, 200 and 500, respectively. In this scenario, the bias in estimation of the means is small and for the most part negligible for all cases for the MLE, but it is not the case for the CECF when DI = 0.1. Clearly when DI = 0.1, the MLE outperforms the CECF, but both methods are too biased in estimating the variances to be useful in practice. When DI = 0.3, the estimation bias appears to be acceptable for the MLE if sample size is greater than or equal to 200. But the estimation bias for the smaller variance is still relatively large for the CECF. When DI$\geq$0.55, the estimation bias is virtually negligible for all the model parameters under both methods, but the MLE is apparently preferred over the CECF as it has smaller MCSD. The estimation bias of the mixing parameter η is relatively small for the MLE, even in the case of small disparity. As the disparity increases, the MLE works better than the CECF, in terms of having smaller MCSD. However, estimation of the mixing parameter is highly variable (under both methods) in cases with small disparity (DI$\leq$0.3).

Table 4 summarizes the simulation results for Case C with sample sizes 100, 200 and 500, respectively. When both means and variances are allowed to vary between the two component distributions, the simulation results are similar to Case A. That is, in the small sample or small disparity cases (n = 100/200 or DI = 0.1/0.3), the CECF tends to have smaller estimation bias for the means but larger bias for variances. When the amount of disparity is large (DI$\geq$0.45), the MLE is clearly the winner between the two competing methods.

As a concluding remark for the simulation study, the MLE may be generally preferred over the CECF when a variance difference is the source of the disparity between the two distributions or when the DI is large, say greater than or equal to 0.55. However, use of the MLE in cases where the DI is less than 0.55 should be with caution, particularly when the disparity between the distributions is purely due to a separation of the means. Though the CECF is an alternative method for estimating the means with less bias than the MLE when separation of the means is the source of small disparity between the distributions, it still results in biased estimation for the means when the DI is small. In general, the MLE is a better method for estimating the variances than the CECF, as it results in less estimation bias as well as smaller standard error.

4 Application to the PREDICT-HD Data

The PREDICT-HD study is an ongoing observational study of prHD participants at 32 sites in the United States, Canada, Australia, Germany, Spain and the United Kingdom [11]. Comprehensive longitudinal data have been collected, including more than 80 variables from over 1300 research participants who underwent genetic testing for the HD mutation. As mentioned in the introduction, if an individual's CAG repeat length is greater than or equal to 36, they are considered at-risk for HD, or prHD.

In this section, we apply the DI to the PREDICT-HD data with the aim of identifying possible sensitive cognitive biomarkers that may distinguish between prHD (at-risk) individuals and healthy (non-at-risk) controls, particularly when CAG repeat length is masked or unknown. If CAG repeat length is not observed for individuals under study, then the observed data are a mixture from the control group and the prHD group. The size of the study sample and the longitudinal nature of the study may facilitate opportunities to discover disease biomarkers whose progress may indicate an individual's at-risk status without knowledge of their CAG repeat length. We focus on the following five cognitive measures: Symbol Digit Modalities Test (SDMT), Stroop Color Test (STROOP-C), Stroop Word Test (STROOP-W), Trail Making Test A (TRAILS-A), and Trail Making Test B (TRAILS-B).

We now provide some background regarding the potential biomarkers from Predict-HD that we will analyze. SDMT [15] involves a simple substitution task to pair specific numbers with given geometric figures within a fixed amount of time. Individuals with cerebral dysfunction usually perform poorly on the SDMT, which is indicated by a smaller value of this measure. The task of the Stroop tests [16] is to look at pages of colored words, reading words or naming colors as quickly as possible within a fixed amount of time. A smaller value of these measures may determine the individual's cognitive inflexibility. The Trails A test [14], a measure of speeded attention, requires individuals to draw the lines connecting the numbers 1,2,3,4 etc in order until reaching the end. The Trails B test [14] asks individuals to draw the lines connecting the numbers 1,2,3,4 etc and the letters A,B,C,D etc in alternating order. The total times (in seconds) needed to complete each of these tasks are recorded. A larger value of these measures is indicative of cognitive impairment.

The PREDICT-HD study has collected these cognitive measures longitudinally for both prHD individuals and healthy controls. Having at-risk information allows us to estimate the group characteristics for these outcomes and their corresponding DI values. For this analysis, let us denote $\mathbf{R_1}$ and $\mathbf{R_2}$ as the two latent groups with $\mathbf{R_1}$ representing the control group and $\mathbf{R_2}$ the prHD group. The parameters of interest are: $\theta = (\eta, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2)$, and parameters with the subscript 2 correspond to the prHD group, while those with the subscript 1 correspond to the control group. Since PREDICT-HD records CAG length, we can estimate the means and variances using their sample estimates and substitute them into (5) to obtain an estimate of

Table 5: Characteristics of the Five Cognitive Measures in the PREDICT-HD Study

Cognitive Variables	Sample Sizes (n_1, n_2)	$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\sigma}_1$	$\hat{\sigma}_2$	DI*
At 40-42 Age Window:						
SDMT	(31,231)	56.48	52.03	9.59	12.45	0.192
STROOP-C	(31,229)	85.77	78.54	9.19	14.84	0.308
STROOP-W	(31,230)	104.29	98.88	15.42	19.05	0.153
TRAILS-A	(20,153)	20.45	26.29	6.39	10.92	0.343
TRAILS-B	(20,151)	48.05	66.81	19.33	36.06	0.370
At 50-52 Age Window:						
SDMT	(65,183)	54.54	48.11	7.76	11.88	0.308
STROOP-C	(65,183)	82.11	71.90	10.96	13.02	0.336
STROOP-W	(65,183)	105.63	90.88	16.16	16.00	0.354
TRAILS-A	(42,103)	25.10	29.96	6.20	11.56	0.344
TRAILS-B	(41,104)	54.34	78.70	18.39	43.38	0.465

OC, OC* and calculate the DI by DI$^* = 1 - $ OC*. The results are summarized in Table 5.

Table 5 presents the characteristics of the two cohorts and their corresponding DI* for the five cognitive measures mentioned above at two age windows: 40-42 and 50-52. We considered the 40-42 age window and 50-52 age window so that we could determine whether the ability to estimate parameters for prHD and control individuals changed over time. We anticipated that there would be more disparity between the prHD and control groups by age 50-52, as prHD individuals will have had more time to progress, which would be reflected by higher DI values and thus, better estimation performance. Based on the simulation results presented above, it appears that only the Trails-B measure in the age window 50-52 may have a chance of providing a reasonable estimate of the model parameters when the genetic information of CAG repeat length is unknown or not considered. This is because the DI for Trails-B in the 50-52 window is the largest, at DI = 0.465. To estimate all parameters, including means/variances/the mixing proportion, we then apply both the MLE (via the EM algorithm) and the CECF methods to the Trails-B data, as if CAG repeat length were not observed, and compare their performance in light of our simulation results presented above. These estimates are summarized in Table 6. This real data example resembles the simulation scenario C3 with sample size around 100 where the estimation is less biased for the CECF method. In this setting, we can only compare our incomplete data estimation results with the complete data estimation results from Table 5, since we do not know the true parameter values. The complete data estimation serves as reference, because estimation based on the complete data should result in smaller variances, less biased estimation, and more efficiency in general. Indeed, the CECF method yields closer estimates than the MLE method when inspecting the values given at the bottom row of Table 5. Nevertheless, the resulting estimates for the mixing parameter are far from the complete data mixing proportion for both methods, since estimation under the complete information yields $\hat{\eta} = \frac{104}{145} = 0.717$. This is mainly due to the fact that the estimation standard error is quite large for $\hat{\eta}$ and implies that when there is quite bit overlap between the two distributions, the methods may yield reasonable estimates for the mean and variance parameters but could have difficulty correctly estimating the mixing

Table 6: Estimates of Model Parameters in TRAILS-B at Age Window 50-52 for the PREDICT-HD data

Methods	$\hat{\eta}$	$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\sigma}_1$	$\hat{\sigma}_2$
MLE	0.26	55.87	118.05	13.83	51.14
CECF	0.31	53.96	96.95	14.15	43.37

parameter. Both methods largely overestimate μ_2, and slightly underestimate σ_1. This is not a surprise, as the DI is only 0.465 and our sample size is 145.

5 Final Remarks

Estimation of normal mixtures is a classical problem that has been widely researched. While the MLE and the CECF appear to be the most popular methods among many others, their estimation properties have not been extensively studied. This is probably due to the well-known fact that the normal mixture can be an ill-posed model when the disparity between the component distributions is small [1,20]. In this article, we utilize the OC to quantify the disparity between the two distributions and then empirically examined when the methods can lead to reasonable estimates of the model parameters. The results provide an instructive guideline regarding the use of the existing methods. Generally speaking, when there is enough disparity, the MLE is still a more favorable method in practice, particularly when a difference in variances is the major source of the disparity between the two component distributions. When a difference in means is the major source of the disparity, the MLE may not lead to estimation with negligible bias when the DI is small, and in this case, the CECF may be a reasonable alternative.

Our simulation study implies that neither the MLE nor the CECF method will yield a satisfactory outcome with regards to accurately estimating parameters for prHD individuals and healthy controls based on cross-sectional cognitive measures in PREDICT-HD data, as the DI values for these measures are too small at the times considered. The amount of overlap present led to the CECF and MLE largely overestimating the mean for prHD individuals and underestimating the amount of variability in the control group, relative to estimation when complete information is known. Since HD is a progressive disease, investigating the disparity between longitudinal trajectories of these cognitive measures between the prHD and healthy control may provide better estimation of parameters for these cohorts. A future research direction is to develop an index which measures the disparity between the two groups based on longitudinal data. Latent class modeling of generalized linear mixed-effects models could be used for the groups' longitudinal trajectories, in order to identify sensitive cognitive markers for indirectly ascertaining at-risk status in similar cohorts.

Acknowledgement

This research was supported by grants NS040068 and 5R01NS054893 from the National Institutes for Health, National Institute of Neurological Disorders and Stroke, and a grant A3917 from CHDI Foundation, Inc.

References

[1] Jiahua Chen and Pengfei Li. Hypothesis test for normal mixture models: The em approach. *The Annals of Statistics*, 37:2523–2542, 2009.

[2] A.P. Dempster and N.M. Laird. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Issue 1*, 39B:1–38, 1977.

[3] Brian Everitt and D.J. Hand. *Finite Mixture Distributions*. Chapman and Hall, 1981.

[4] C.R. Heathcote. The integrated squared error estimation of parameters. *Biometrika, Issue 2*, 64:255–264, 1977.

[5] David W. Hosmer Jr. A comparison of iterative maximum likelihood estimates of the parameters of a mixture of two normal distributions under three different types of sample. *Biometrics*, 29:761–770, 1973.

[6] Group HsDCR. A novel gene containing a trinucleotide repeat that is expanded and unstable on huntington's disease chromosones. *Cell*, 72:971–983, 1993.

[7] Henry F. Inman and Edwin L. Jr. Bradley. The overlapping coefficient as a measure of agreement between probability distributions and point estimation of the overlap of two normal densities. *Communications in Statistics - Theory and Methods*, 18:3851–3874, 1989.

[8] Douglas R. Langbehn and Jane S. Paulsen. Predictors of diagnosis in huntington disease. *Neurology*, 68, 2007.

[9] K.M. Leytham. Maximum likelihood estimates for the parameters of mixture distributions. *Water Resources Research, Issue 7*, 20:896–902, 1984.

[10] Dechavudh Nityasuddhi and Dankmar Bohning. Asymptotic properties of the em algorithm estimate for normal mixture models with component specific variances. *Computation Statistics and Data Analysis*, 41:591–601, 2003.

[11] J.S. Paulsen. Early detection of huntington's disease. *Future Neurology*, 5:85–104, 2010.

[12] Karl Pearson. Contributions to the mathematical theory of evolution. *Philos. Trans. Royal Society of London*, 185:71–110, 1894.

[13] Richard E. Quandt and James B. Ramsey. Estimating mixtures of normal distributions and switching regressions. *Journal of the American Statistical Association, Issue 364*, 73:730–738, 1978.

[14] RM Reitan. Validity of the trail making test as an indicator of organic brain damage. *Perceptual and Motor Skills*, 8:271–276, 1958.

[15] A. Smith. Symbol digit modalities test (sdmt) manual (revised). *Los Angeles: Western Psychological Services*, 1982.

[16] JR Stroop. Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, 18:643–662, 1935.

[17] K. Tran. Estimating mixtures of normal distributions via empirical characteristic function. *Econometric Reviews, Issue 2*, 17:167–183, 1998.

[18] Francis O. Walker. Huntington's disease. *The Lancet, Issue 9557*, 369:218–228, 2007.

[19] Dinghai Xu and John Knight. Continuous empirical characteristic function estimation of mixtures of normal parameters. *Econometric Reviews, Issue 1*, 30:25–50, 2011.

[20] Jun Yu. Empirical characteristic function estimation and its applications. *Econometric Reviews, Issue 2*, 23:93–123, 2004.