Misclassification Rates in Hypertension Diagnosis due to Measurement Errors

Camila Friedman-Gerlicz[1], Claremont McKenna College,
cgerlicz10@cmc.edu
and Isaiah Lilly[1], California State University at Sacramento,
ijlilly@aol.com

Abstract. Using a mixture of two normal distributions, we estimate the false positive and false negative errors in the diagnosis of hypertension. Parameters in the mixture are estimated by the expectation-maximization (EM) algorithm. It is shown that both errors depend on cutoff points. Repeated measurements reduce both errors dramatically. The number of repeated measurements is recommended through a simulation study.

1 Introduction

Hypertension is one of the most important risk factors for coronary heart disease and stroke. Every year, millions of people take screening tests for early detection of hypertension. However, the risk of classifying a truly normal subject as hypertensive (false positive) or missing a truly hypertensive subject (false negative) is quite high in hypertension diagnosis due to random errors of measurements in blood pressure. This paper addresses both misclassification quantitatively.

A false positive (FP, hereafter) error occurs when measurements of patients' blood pressures are above certain cutoff points while their true blood pressures are under those thresholds. A false negative (FN, hereafter) error happens when observations of patients' blood pressures are below cutoff points but their true blood pressure levels are above those thresholds. Subjects falsely classified as hypertensive suffer psychologically, economically by paying and taking unnecessary medication, and experience physically the side effects of various medicines etc. Hypertensive subjects falsely classified as normotensive will not receive treatments and thus run a much higher risk of heart attack, stroke, and death etc.

Many factors have influences on misclassification errors in hypertension diagnosis. It is the purpose of this paper to quantify false positive and false negative errors due to measurement errors. A number of studies have been published in the literature. Rosner (1977), El Lozy (1982), and Moskowitz et al. (1993) proposed statistical frameworks to calculate FP and FN errors. However, all of their calculations are based on the normality of blood pressure data.

In practice, skewed and non-normal distributions or even bimodal distributions from blood pressure data are often observed. Bimodality in large sample is often

[1]Adviser: Xianggui Qu, Oakland University, qu@oakland.edu

(although not always) an indication of two sub-populations. Cicchinelli (1963) first claimed that the skewness in the sample distributions of blood pressure is evidence of the mixture of two sub-distributions. Although normal distribution is still used in hypertension studies due to its simplicity (Marshall, 2008), mixture normal distributions are more powerful in statistical modeling. In fact, Tarpey et al. (2008) showed that a two- or three-normal mixture provides a very good surrogate to some well-known nonnormal distributions. A two-normal mixture distribution is proposed to model blood pressure data and quantify FP and FN errors in this paper.

The paper is arranged as follows. Section 2 lays out the statistical frameworks on normal mixture distributions. Parameter estimation is explained in Section 3. The data from Framingham heart study is used as an example. Section 4 explores the influence of repetition on both FP and FN errors through a simulation study. Concluding remarks are given in Section 5.

2 Statistical Formulation

Let Y be the observed measurement and X is the true level of blood pressure for a subject. The model considered in this paper is

$$Y = X + \epsilon,$$

where $\epsilon \sim N(0, \sigma_\epsilon^2)$ (a normal distribution of mean zero and variance σ_ϵ^2) is the random error of measurement and X is a random variable that is independent of ϵ.

A random variable T follows a mixture two normal distributions, i.e.,

$$T \sim pN(\mu_1, \sigma_1^2) + (1-p)N(\mu_2, \sigma_2^2)$$

if it has a density function

$$\frac{p}{\sqrt{2\pi}\sigma_1} e^{\frac{(t-\mu_1)^2}{2\sigma_1^2}} + \frac{1-p}{\sqrt{2\pi}\sigma_2} e^{\frac{(t-\mu_2)^2}{2\sigma_2^2}},$$

where $0 < p < 1$ is the proportion, μ_1, μ_2, are population means, and $\sigma_1^2 > 0$, $\sigma_2^2 > 0$ are population variances.

Theorem 1. *If $X \sim pN(\mu_1, \sigma_1^2) + (1-p)N(\mu_2, \sigma_2^2)$, $\epsilon \sim N(0, \sigma_\epsilon^2)$, and X and ϵ are independent,*

$$Y = X + \epsilon \sim pN(\mu_1, \sigma_1^2 + \sigma_\epsilon^2) + (1-p)N(\mu_2, \sigma_2^2 + \sigma_\epsilon^2).$$

Conversely, if $Y \sim pN(\mu_1, \sigma_1^2) + (1-p)N(\mu_2, \sigma_2^2)$, $Y = X + \epsilon$, and X and ϵ are independent, then $X \sim pN(\mu_1, \sigma_1^2 - \sigma_\epsilon^2) + (1-p)N(\mu_2, \sigma_2^2 - \sigma_\epsilon^2)$.

Proof. Recall that a normal distribution $N(\mu, \sigma^2)$ has a characteristic function of $e^{it\mu - \frac{1}{2}\sigma^2 t^2}$. If X has a mixture of two normal distributions, its characteristic function is,

$$\begin{aligned} f_X(t) &= E(e^{itX}) = \int_{-\infty}^{+\infty} e^{itx}[\frac{p}{\sigma_1\sqrt{2\pi}} e^{\frac{(x-\mu_1)^2}{2\sigma_1^2}} + \frac{1-p}{\sigma_2\sqrt{2\pi}} e^{\frac{(x-\mu_2)^2}{2\sigma_2^2}}]dx \\ &= \int_{-\infty}^{+\infty} e^{itx} \frac{p}{\sigma_1\sqrt{2\pi}} e^{\frac{(x-\mu_1)^2}{2\sigma_1^2}} dx + \int_{-\infty}^{+\infty} e^{itx} \frac{1-p}{\sigma_2\sqrt{2\pi}} e^{\frac{(x-\mu_2)^2}{2\sigma_2^2}} dx \\ &= pe^{it\mu_1 - \frac{1}{2}\sigma_1^2 t^2} + (1-p)e^{it\mu_2 - \frac{1}{2}\sigma_2^2 t^2}. \end{aligned}$$

Since X and ϵ are independent, the characteristic function of Y is

$$\begin{aligned} f_Y(t) &= E(e^{itY}) = E(e^{it(X+\epsilon)}) = E(e^{itX})E(e^{it\epsilon}) \\ &= pe^{it\mu_1 - \frac{1}{2}(\sigma_1^2+\sigma_\epsilon^2)t^2} + (1-p)e^{it\mu_2 - \frac{1}{2}(\sigma_2^2+\sigma_\epsilon^2)t^2}, \end{aligned}$$

which is a characteristic function of the mixture distribution $pN(\mu_1, \sigma_1^2+\sigma_\epsilon^2)+(1-p)N(\mu_2, \sigma_2^2+\sigma_\epsilon^2)$.

The converse can be proved similarly. □

The assumption that X has a mixture of two normal distributions has both statistical and genetic rationales. First, blood pressure data are quite skewed in practice (Carroll et al. 2006, page 289). Tarpey et al. (2008) showed that a mixture normal distribution usually fits skewed data as well as a single nonnormal distribution statistically.

Second, as is seen in Theorem 2 followed, estimating the density of X is critical to the calculation of FP and FN errors and the mixture normal assumption on X makes the estimation straightforward. Note that estimation of the density function of X based on observed data of Y is a classic deconvolution problem for which no general solution exists. In practice, the skewed data of Y could be modeled directly with an asymmetric distribution with fewer parameters than the normal mixture. However, the deconvolution process becomes much more difficult. Some nonparametric methods have been proposed but the convergent rate of the estimator to the density of X is very slow. More discussion and references on the deconvolution problem can be found in Delaigle (2008).

On the other hand, a mixture model of two components works well from a genetic point of view when there is a major gene dominating the mean quantitative response with additional variability due to environmental and other genetic factors. Blood pressure is one such trait. In fact, Levy et al. (2000) discovered a gene influencing blood pressure on chromosome 17 using data from the Framingham heart study.

Theorem 2 formulates the calculation FP and FN error rates. Considering that the true numbers of hypertensive and normotensive patients are usually unknown

while the total number of patients screened is often recorded, FP and FN errors are calculated as joint probabilities rather than conditional probabilities.

Theorem 2. *Let $\Phi(x)$ be the cumulative distribution function of the standard normal distribution, i.e.,*

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{x^2}{2}} dx,$$

and

$$G(x, y, \rho) = \frac{1}{2\pi\sqrt{1-\rho^2}} \int_{-\infty}^{x} \int_{-\infty}^{y} e^{-\frac{x^2 - 2\rho xy + y^2}{2(1-\rho^2)}} dxdy$$

be the cumulative probability function of the standard bivariate normal distribution with correlation coefficient ρ. If $Y \sim pN(\mu_1, \sigma_1^2) + (1-p)N(\mu_2, \sigma_2^2)$, $Y = X + \epsilon$, $\epsilon \sim N(0, \sigma_\epsilon^2)$, and X and ϵ are independent, then, for a specific cutoff point c,

$$\begin{aligned}
FP &= P(Y > c, X \le c) \\
&= p\Phi\left(\frac{c-\mu_1}{\sqrt{\sigma_1^2 - \sigma_\epsilon^2}}\right) - pG\left(\frac{c-\mu_1}{\sqrt{\sigma_1^2 - \sigma_\epsilon^2}}, \frac{c-\mu_1}{\sigma_1}, \frac{\sqrt{\sigma_1^2 - \sigma_\epsilon^2}}{\sigma_1}\right) \\
&+ (1-p)\Phi\left(\frac{c-\mu_2}{\sqrt{\sigma_2^2 - \sigma_\epsilon^2}}\right) - (1-p)G\left(\frac{c-\mu_2}{\sqrt{\sigma_2^2 - \sigma_\epsilon^2}}, \frac{c-\mu_2}{\sigma_2}, \frac{\sqrt{\sigma_2^2 - \sigma_\epsilon^2}}{\sigma_2}\right)
\end{aligned}$$

$$\begin{aligned}
FN &= P(Y \le c, X > c) \\
&= p\Phi\left(\frac{c-\mu_1}{\sigma_1}\right) - pG\left(\frac{c-\mu_1}{\sqrt{\sigma_1^2 - \sigma_\epsilon^2}}, \frac{c-\mu_1}{\sigma_1}, \frac{\sqrt{\sigma_1^2 - \sigma_\epsilon^2}}{\sigma_1}\right) \\
&+ (1-p)\Phi\left(\frac{c-\mu_2}{\sigma_2}\right) - (1-p)G\left(\frac{c-\mu_2}{\sqrt{\sigma_2^2 - \sigma_\epsilon^2}}, \frac{c-\mu_2}{\sigma_2}, \frac{\sqrt{\sigma_2^2 - \sigma_\epsilon^2}}{\sigma_2}\right)
\end{aligned}$$

provided that $min(\sigma_1^2, \sigma_2^2) > \sigma_\epsilon^2$.

Proof. Let $d(x, y)$ be the joint density of X and Y. Note that $d(x, y) = d_{Y|X=x}(y) d_X(x)$, where $d_{Y|X=x}(y)$ is the conditional density of Y given $X = x$ and $d_X(x)$ is the density of X. Since $Y = X + \epsilon$ and X and ϵ are independent, Theorem 1 shows that

$$d_X(x) = \frac{p}{\sqrt{2\pi(\sigma_1^2 - \sigma_\epsilon^2)}} e^{-\frac{(x-\mu_1)^2}{2(\sigma_1^2 - \sigma_\epsilon^2)}} + \frac{1-p}{\sqrt{2\pi(\sigma_2^2 - \sigma_\epsilon^2)}} e^{-\frac{(x-\mu_2)^2}{2(\sigma_2^2 - \sigma_\epsilon^2)}}.$$

It can be seen that

$$d_{Y|X=x}(y) = \frac{1}{\sqrt{2\pi}\sigma_\epsilon} e^{-\frac{(y-x)^2}{2\sigma_\epsilon^2}}.$$

It follows that

$$\begin{aligned}
FP &= P(Y > c, X \le c) = \int_c^{+\infty} \int_{-\infty}^{c} d(x, y) dxdy \\
&= \int_c^{+\infty} dy \int_{-\infty}^{c} \left[\frac{p}{\sqrt{2\pi(\sigma_1^2 - \sigma_\epsilon^2)}} e^{-\frac{(x-\mu_1)^2}{2(\sigma_1^2 - \sigma_\epsilon^2)}} + \frac{1-p}{\sqrt{2\pi(\sigma_2^2 - \sigma_\epsilon^2)}} e^{-\frac{(x-\mu_2)^2}{2(\sigma_2^2 - \sigma_\epsilon^2)}} \right] \frac{1}{\sqrt{2\pi}\sigma_\epsilon} e^{-\frac{(y-x)^2}{2\sigma_\epsilon^2}} dx,
\end{aligned}$$

Let $g(x, y, \rho)$ be the density of the standard bivariate normal distribution, that is,

$$g(x, y, \rho) = \frac{1}{2\pi\sqrt{1-\rho^2}} e^{-\frac{x^2-2\rho xy+y^2}{2(1-\rho^2)}}.$$

It can be shown that

$$\frac{1}{\sqrt{2\pi(\sigma_1^2-\sigma_\epsilon^2)}} e^{-\frac{(x-\mu_1)^2}{2(\sigma_1^2-\sigma_\epsilon^2)}} \frac{1}{\sqrt{2\pi}\sigma_\epsilon} e^{-\frac{(y-x)^2}{2\sigma_\epsilon^2}} = \frac{1}{\sigma_\epsilon\sqrt{\sigma_1^2-\sigma_\epsilon^2}} g\left(\frac{x-\mu_1}{\sqrt{\sigma_1^2-\sigma_\epsilon^2}}, \frac{y-\mu_1}{\sigma_1}, \frac{\sqrt{\sigma_1^2-\sigma_\epsilon^2}}{\sigma_1}\right),$$

which is the density of a bivariate normal distribution with means μ_1, μ_1, variances $\sigma_1^2 - \sigma_\epsilon^2$, σ_1^2, and correlation coefficient $\frac{\sqrt{\sigma_1^2-\sigma_\epsilon^2}}{\sigma_1}$, respectively.

Similarly,

$$\frac{1}{\sqrt{2\pi(\sigma_2^2-\sigma_\epsilon^2)}} e^{-\frac{(x-\mu_2)^2}{2(\sigma_2^2-\sigma_\epsilon^2)}} \frac{1}{\sqrt{2\pi}\sigma_\epsilon} e^{-\frac{(y-x)^2}{2\sigma_\epsilon^2}} = \frac{1}{\sigma_\epsilon\sqrt{\sigma_2^2-\sigma_\epsilon^2}} g\left(\frac{x-\mu_2}{\sqrt{\sigma_2^2-\sigma_\epsilon^2}}, \frac{y-\mu_2}{\sigma_2}, \frac{\sqrt{\sigma_2^2-\sigma_\epsilon^2}}{\sigma_2}\right).$$

Therefore,

$$\begin{aligned}
FP &= \frac{p}{\sigma_\epsilon\sqrt{\sigma_1^2-\sigma_\epsilon^2}} \int_c^{+\infty} dy \int_{-\infty}^{c} g\left(\frac{x-\mu_1}{\sqrt{\sigma_1^2-\sigma_\epsilon^2}}, \frac{y-\mu_1}{\sigma_1}, \frac{\sqrt{\sigma_1^2-\sigma_\epsilon^2}}{\sigma_1}\right) dx \\
&+ \frac{1-p}{\sigma_\epsilon\sqrt{\sigma_2^2-\sigma_\epsilon^2}} \int_c^{+\infty} dy \int_{-\infty}^{c} g\left(\frac{x-\mu_2}{\sqrt{\sigma_2^2-\sigma_\epsilon^2}}, \frac{y-\mu_2}{\sigma_2}, \frac{\sqrt{\sigma_2^2-\sigma_\epsilon^2}}{\sigma_2}\right) \\
&= p\Phi\left(\frac{c-\mu_1}{\sqrt{\sigma_1^2-\sigma_\epsilon^2}}\right) - pG\left(\frac{c-\mu_1}{\sqrt{\sigma_1^2-\sigma_\epsilon^2}}, \frac{c-\mu_1}{\sigma_1}, \frac{\sqrt{\sigma_1^2-\sigma_\epsilon^2}}{\sigma_1}\right) \\
&+ (1-p)\Phi\left(\frac{c-\mu_2}{\sqrt{\sigma_2^2-\sigma_\epsilon^2}}\right) - (1-p)G\left(\frac{c-\mu_2}{\sqrt{\sigma_2^2-\sigma_\epsilon^2}}, \frac{c-\mu_2}{\sigma_2}, \frac{\sqrt{\sigma_2^2-\sigma_\epsilon^2}}{\sigma_2}\right)
\end{aligned}$$

The formula of calculating FN can be obtained similarly. □

Since there are no closed-form expressions in calculating the cumulative probabilities from both univariate and bivariate normal distribution in Theorem 2, these functions are calculated numerically in Sections 3 and 4 using functions PNORM and PNORM2D in R package (R Development Core Team, 2008).

3 Parameter Estimation

According to Theorem 2, six parameters p, μ_1, μ_2, σ_1^2, σ_2^2, and σ_ϵ^2 in the two-normal mixture have to be estimated from observed data in order to calculate the FP and FN errors. Let $Y_{ij} = X_i + \epsilon_{ij}$ where X_1, X_2, $\cdots$, X_n is an independently and identically distributed (i.i.d.) sample and ϵ_{ij}'s are i.i.d measurement errors from

$N(0, \sigma_\epsilon^2)$ for $i = 1, 2, \cdots, n$, $j = 1, 2, \cdots, m$. Then, $\hat{\sigma}_\epsilon^2 = \frac{1}{n(m-1)} \sum_{i=1}^{n} \sum_{j=1}^{m} (Y_{ij} - \bar{Y}_i)^2$ is an unbiased estimator of σ_ϵ^2 where $\bar{Y}_i = \frac{1}{m} \sum_{j=1}^{m} Y_{ij}$.

Since $\bar{Y}_i = \frac{1}{m} \sum_{j=1}^{m} Y_{ij} = X_i + \frac{1}{m} \sum_{j=1}^{m} \epsilon_{ij}$, $\bar{Y}_1$, $\bar{Y}_2$, $\cdots$, and $\bar{Y}_n$ are i.i.d. By Theorem 1, if X_1, X_2, $\cdots$, X_n is an i.i.d. sample from mixture population $pN(\mu_{X1}, \sigma_{X1}^2) + (1-p)N(\mu_{X2}, \sigma_{X2}^2)$, $\bar{Y}_1$, $\bar{Y}_2$, $\cdots$, and $\bar{Y}_n$ is an i.i.d. sample from

$$pN(\mu_{X1}, \sigma_{X1}^2 + \frac{\sigma_\epsilon^2}{m}) + (1-p)N(\mu_{X2}, \sigma_{X2}^2 + \frac{\sigma_\epsilon^2}{m})$$

and vice versa. Therefore, if $\hat{p}$, $\hat{\mu}_1$, $\hat{\mu}_2$, $\hat{\sigma}_1^2$, $\hat{\sigma}_2^2$ are estimates of p, μ_1, μ_2, σ_1^2, σ_2^2 based on observed $\bar{Y}_i$'s, respectively, FP and FN errors in Theorem 2 can be estimated by substituting p, μ_1, μ_2, σ_1^2, σ_2^2 with their corresponding estimates,and σ_ϵ^2 with $\hat{\sigma}_\epsilon^2/m$, respectively.

In order to estimate all five parameters p, μ_1, μ_2, σ_1^2, and σ_2^2 in the two-normal mixture distribution, it is necessary that all of them are identifiable, i.e., distinct values of five parameters determine different distributions. The lack of identifiability of the five parameters due to the interchanging of two component labels can be easily overcome in practice by the imposition of an appropriate constraint such as $p \geq 0.5$. The expectation-maximization (EM) algorithm (Dempster et al., 1977) is used in this paper because the lack of identifiability is not of concern in its normal course of fitting mixture models (McLachlan and Peel, 2000, page 27).

Given a sample of observations $\bar{Y}_1$, $\bar{Y}_2$, $\cdots$, and $\bar{Y}_n$ from $pN(\mu_1, \sigma_1^2) + (1-p)N(\mu_2, \sigma_2^2)$, the following EM algorithm of fitting a two-normal mixture model is from Everitt and Hand (1981, page 37). The maximum likelihood estimates of the five parameters are calculated by maximizing the likelihood function, $\mathcal{L}$, where

$$\mathcal{L}(p, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2) = \prod_{i=1}^{n} [\frac{p}{\sqrt{2\pi}\sigma_1} e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}} + \frac{1-p}{\sqrt{2\pi}\sigma_2} e^{-\frac{(x-\mu_2)^2}{2\sigma_2^2}}].$$

For convenience, maximum likelihood estimates are usually obtained by maximizing the log-likelihood

$$L(p, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2) = \sum_{i=1}^{n} \ln[\frac{p}{\sqrt{2\pi}\sigma_1} e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}} + \frac{1-p}{\sqrt{2\pi}\sigma_2} e^{-\frac{(x-\mu_2)^2}{2\sigma_2^2}}].$$

By differentiating L with respect to each of the five parameter and equating the corresponding partial derivatives to zero, the maximum likelihood estimates

can be obtained from the following five equations,

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} \hat{P}(c_1|\bar{Y}_i),$$
$$\hat{\mu}_j = \frac{1}{n\hat{p}}\sum_{i=1}^{n} \hat{P}(c_j|\bar{Y}_i)\bar{Y}_i,\ j = 1, 2,$$
$$\hat{\sigma}_j^2 = \frac{1}{n\hat{p}}\sum_{i=1}^{n} \hat{P}(c_j|\bar{Y}_i)(\bar{Y}_i - \hat{\mu}_j)^2,\ j = 1, 2,$$

where

$$\hat{P}(c_1|\bar{Y}_i) = \frac{\hat{p}\hat{\sigma}_2 e^{\frac{(\bar{Y}_i-\hat{\mu}_2)^2}{2\hat{\sigma}_2^2}}}{(1-\hat{p})\hat{\sigma}_1 e^{\frac{(\bar{Y}_i-\hat{\mu}_1)^2}{2\hat{\sigma}_1^2}} + \hat{p}\hat{\sigma}_2 e^{\frac{(\bar{Y}_i-\hat{\mu}_2)^2}{2\hat{\sigma}_2^2}}},$$

and $\hat{P}(c_2|\bar{Y}_i) = 1 - \hat{P}(c_1|\bar{Y}_i)$.

There is no explicit solution to these five equations. The solution could be obtained through an iterative procedure in which initial values of the five parameters are used to estimate $\hat{P}(c_j|\bar{Y}_i)$ first and then uses the five equations to provide new estimates of the five parameters. The process continues until the Euclidean distance between two consecutive estimates of the five parameters is less than a specified number, e.g., 0.0001.

There are many implementations of EM algorithm with different initializations. The EM algorithm in R package MCLUST is used in this paper where initial values are selected from hierarchical clustering and likelihood gain (Fraley and Raftery, 2000 and 2006).

Example 1. The Framingham heart study is a longitudinally prospective study of cardiovascular disease among a population of free living subjects in the town of Framingham, Massachusetts. The study began in 1948 with 5209 men and women of various ages. The Framingham data used in this example are from Carroll et al. (2006, page 112) and are available from `http://www.stat.tamu.edu/ carroll/ eiv.SecondEdition/data.php`. There were $1,615$ men between the ages of 31 to 65 in the data. Two exams of systolic blood pressure were taken two year apart for each subject. In each exam, two repeated measurements were taken by different examiners independently. Since two exams are two year apart and four measurements are recorded by different examiners independently, four measurement errors are independent replicates.

Table 1 lists the maximum likelihood estimates of the five parameters in a mixture of two normal distributions from data of systolic blood pressure with two, three, and four repeated measurements, where $(1, 4)$ stands for data from the first

and the fourth measurements and so on. It can be seen that $\hat{\sigma}_\epsilon$'s are more stable in three- and four-repeat cases than that in the two-repeat case.

Table 1: Maximum likelihood Estimates of Parameters

Repetition	p	μ_1	σ_1^2	μ_2	σ_2^2	σ_ϵ^2
Two $(1,4)$	0.7843	124.8951	132.8374	152.2067	594.1233	120.4467
Three $(1,3,4)$	0.7692	124.7380	127.7964	151.6087	579.6278	97.9631
Four	0.7689	124.7327	121.5739	150.8009	584.8963	96.4834

Figure 1 shows the false positive, false negative, and misclassification errors (sum of false positive and false positive) in systolic blood pressures with two, three, and four repeated measurements. As is observed, misclassification errors depend on cutoff points. Higher cutoff levels in the range of 130 to 180 results lower errors. High errors are observed in cases where either fewer repeated measurements are taken or larger variances exist in measurement errors. Due to estimation variation, the false positive error at 140 with two repeated measurements, 0.0269, is smaller than that of 0.0308 with four repeated measurements.

Table 2 lists misclassification errors at selected cutoff points, 130, 140, 160, and 180. For example, the misclassification rate at systolic points 130 (high-normal) and 140 (stage one hypertension) are 22.71% and 12.86%, respectively if two measurements are recorded while they are 19.54% and 11.68% if four repeated measurements are taken. The number of repeated measurements has larger effects on misclassification errors at lower cutoff points than those at higher cutoff points. The misclassification error at 180 is 1.62% when there are two repeated measurements while it is 1.46% with four repeated measurements. The 9% reduction of misclassification error at 180 is smaller than the 14% reduction at 130. It is also observed that the reduction of misclassification errors in the case of three repeated measurements is not very different from that of taking four repeated measurements.

Table 2: Misclassification Errors in Framingham Data

	Two Repeats			Three Repeats			Four Repeats		
Cutoff	FP	FN	Miss.	FP	FN	Miss.	FP	FN	Miss.
130	0.1431	0.0840	0.2271	0.1185	0.0787	0.1972	0.1170	0.0784	0.1954
140	0.1017	0.0269	0.1286	0.0867	0.0306	0.1173	0.0859	0.0308	0.1168
160	0.0203	0.0132	0.0335	0.0193	0.0127	0.0320	0.0191	0.0123	0.0314
180	0.0107	0.0055	0.0162	0.0099	0.0054	0.0152	0.0095	0.0052	0.0146

4 Simulation Study

The analysis of Framingham data shows that more repeated measurements results in lower misclassification errors. How many repeats do we need and how far are

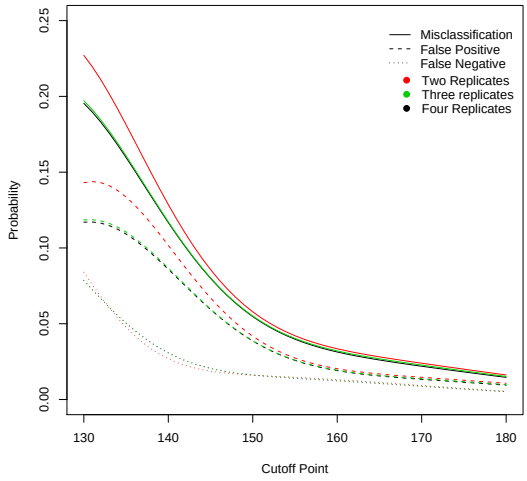

Figure 1: Misclassification Errors in Framingham Data

the estimated errors from the true errors? The simulation study in this section tries to answer these two questions.

Assume that $X \sim 0.75N(120, 12^2) + 0.25N(150, 24^2)$ and $\epsilon \sim N(0, 18^2)$. The variance of measurement error is designed to be larger than that of the Framingham data to see the influence of repeated measurements on estimated errors.

Figure 2 shows misclassification errors when two, four, six, and eight repeated measurements are recorded. It can be seen that the estimated misclassification errors are close to the true values. Repeated measurements improve the error estimates. All errors in the four-repetition case are closer to true errors than those in the two-repetition case. As the number of repeated measurements increases, the gap between true error curves and estimated error curves narrows.

Table 3 provides true and estimate errors at cutoff points 130, 140, 150, 160, 170, and 180. In general, the large estimated errors with two repeated measurements differ from corresponding true errors in the second decimal point (e.g., $0.1684 - 0.1582 = 0.0102$). The differences between estimated errors with four repeated measurements and true errors are in the third decimal points (e.g., $0.1632 - 0.1582 = 0.005$.) Given the variation scale of measurement errors, misclassification errors are estimated well when four measurements are taken. If the measurement error has a large variation, a large number of repeated measurements is needed. Note that, if there are no repeated measurements, σ_ϵ^2 is not estimable and misclassification errors cannot be calculated by the method proposed in this paper.

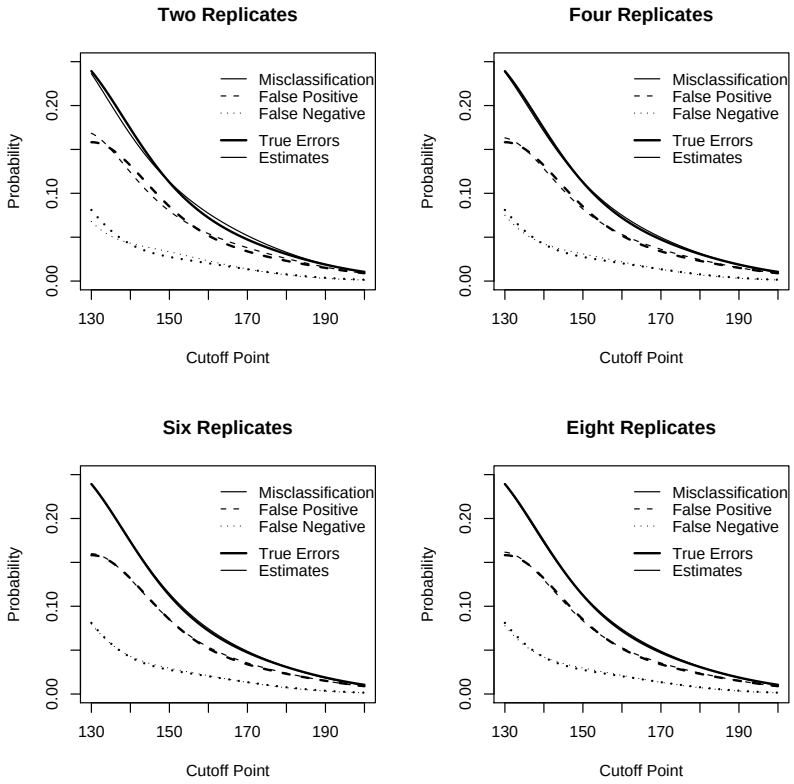

Figure 2: Misclassification Errors and Repeated Measurements

Table 3: Misclassification Errors and Repeats

		Cutoff points					
	Error	130	140	150	160	170	180
True	False Positive	0.1582	0.1317	0.0850	0.0519	0.0338	0.0230
	False Negative	0.0810	0.0420	0.0275	0.0203	0.0135	0.0077
	Misclassification	0.2393	0.1737	0.1125	0.0721	0.0474	0.0308
Two	False Positive	0.1684	0.1239	0.0798	0.0540	0.0380	0.0257
	False Negative	0.0679	0.0431	0.0333	0.0231	0.0138	0.0071
	Misclassification	0.2363	0.1671	0.1131	0.0771	0.0518	0.0328
Four	False Positive	0.1632	0.1280	0.0821	0.0531	0.0362	0.0245
	False Negative	0.0750	0.0425	0.0309	0.0219	0.0135	0.0071
	Misclassification	0.2381	0.1705	0.1130	0.0750	0.0497	0.0316
Six	False Positive	0.1599	0.1309	0.0849	0.0534	0.0355	0.0237
	False Negative	0.0798	0.0431	0.0295	0.0210	0.0132	0.0072
	Misclassification	0.2397	0.1741	0.1144	0.0744	0.0487	0.0309
Eight	False Positive	0.1616	0.1300	0.0834	0.0528	0.0354	0.0239
	False Negative	0.0777	0.0422	0.0296	0.0213	0.0134	0.0073
	Misclassification	0.2393	0.1723	0.1131	0.0740	0.0488	0.0312

5 Conclusion

A mixture of two normal distributions is used to model the distribution of blood pressure data. The mixture model fits the blood pressure data better than a normal distribution. Both FP and FN errors are estimated from the proposed mixture model. It is observed that misclassification errors depend on cutoff points. High cutoff points have low errors. Measurement errors with high variances result in high misclassification errors. Repeated measurements not only provide the variance estimation of measurement errors but also reduce misclassification errors significantly. Though most of our discussion is focusing on systolic blood pressure, the method can also be applied to diastolic blood pressure. The systolic blood pressure was chosen in this paper because numerous studies have found that the determination of systolic blood pressure is more reliable than that of diastolic blood pressure. Moreover, systolic hypertension has been widely accepted as a cause of cardiovascular mortality.

Acknowledgments

The authors thank the associate editor and three anonymous referees for their detailed comments and suggestions that improve the presentation greatly. The study is supported by NSF Research Experience for Undergraduates Program: Computational and Numerical Statistics and Mathematics in Oakland University

References

[1] R.J. CARROLL, D. RUPPERT, L.A. STEFANSKI, and C. CRAINICEANU, Measurement Error in Nonlinear Models: : A Modern Perspective, second edition, Chapman and Hall/CRC, 2006.

[2] A.L. CICCHINRLLI, The composite of two Gaussian distributions as a model for blood pressure in man. Unpublished Ph.D. dissertation, the University of Michigan, 1963.

[3] A.P. DEMPSTER, N.M. LAID, and D.B. RUBIN, Maximum likelihood from incomplete data via the EM algorithm. Journal of Royal Statistical Society, Series B, 39(1977), pp. 1-38.

[4] A. DELAIGLE, An alternative view of the deconvolution problem, Statistica Sinica, 18(2008), pp. 1025-1045.

[5] M. EL LOZY, Simple computation of a bivariate normal integral arising from a problem of misclassification with applications to the diagnosis of hypertension, Comm. Statist.-Theor. Meth., 11(1982), pp. 2195-2205.

[6] C. FRALEY and A.E. RAFTERY, Model-based clustering, discriminant analysis and density estimation, Journal of the American Statistical Association, 97(2002), pp. 611-631.

[7] C. FRALEY and A.E. RAFTERY, MCLUST version 3 for R: Normal mixture modeling and model-based clustering, Technical Report No. 504, Department of

Statistics, University of Washington, September 2006,
http://www.stat.washington.edu/fraley/mclust/tr504.pdf.
[8] D. LEVY, A.L. DeSTEFANO, M.G. LARSON, C.J. O'DONNELL, R.P LIFTON, H. GAVRAS, L.A. CUPPLES, and R.H. MYERS, Evidence for a gene influencing blood pressure phenotypes in subjects from the Framingham study: genome scan linkage results for longitudinal blood pressure phenotypes in subjects from the Framingham heart study, Hypertension, 36(2000), pp. 477483.
[9] T.P. MARSHALL, Blood pressure variability: The challenge of variation, American Journal of Hypertension, 21(2008), pp. 3-4.
[10] G.J. McLACHLAN and D. PEEL, Finite Mixture Models (Wiley Series in Probability and Statistics), Wiley, 2000.
[11] H. MOSKOWITZ, R. PLANTE, and HSIEN-TANG TSAI, A multistage screening model for evaluation and control of misclassification error in the detection of hypertension, Management Science, 39(1993), pp. 307-321.
[12] B. ROSNER and B.F. POLK, The implication of blood pressure variability for clinical and screening purposes, Journal of Chronical Disease, 32(1979), pp. 451-461.
[13] R DEVELOPMENT CORE TEAM, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2008, http://www.R-project.org.
[14] T. TARPEY, D. YUN, and E. PETKOVA (2008), Model misspecification: Finite mixture or homogeneous? Stat. Modelling, 8(2008), pp. 199-218.