Published in final edited form as:

J Stat Plan Inference. 2011 January 1; 141(1): 549–558. doi:10.1016/j.jspi.2010.07.004.

Exact confidence interval estimation for the difference in diagnostic accuracy with three ordinal diagnostic groups

Lili Tian^{a,*}, Chengjie Xiong^b, Chin-Ying Lai^a, and Albert Vexler^a

^aDepartment of Biostatistics, University at Buffalo, 249 Farber Hall, 3435 Main St. Bldg. 26 Buffalo, NY 14214-3000, USA

^bDivision of Biostatistics, Washington University at St. Louis, Buffalo, St. Louis, MO 63110, USA

Abstract

In the cases with three ordinal diagnostic groups, the important measures of diagnostic accuracy are the volume under surface (*VUS*) and the partial volume under surface (*PVUS*) which are the extended forms of the area under curve (*AUC*) and the partial area under curve (*PAUC*). This article addresses confidence interval estimation of the difference in paired *VUS*s and the difference in paired *PVUS*s. To focus especially on studies with small to moderate sample sizes, we propose an approach based on the concepts of generalized inference. A Monte Carlo study demonstrates that the proposed approach generally can provide confidence intervals with reasonable coverage probabilities even at small sample sizes. The proposed approach is compared to a parametric bootstrap approach and a large sample approach through simulation. Finally, the proposed approach is illustrated via an application to a data set of blood test results of anemia patients.

Keywords

Diagnostic accuracy; Receiver operating characteristic (ROC); curve; Generalized pivot; Generalized test variable

1. Introduction

Receiver-operating characteristic (ROC) curves, which can be constructed by plotting the false-positive rate (i.e. 1-specificity) against the true-positive rate (i.e. sensitivity), have been common tools for evaluating the performance of diagnostic tests. The area under the ROC curve (*AUC*) has been widely used as a quantitative index of discriminating ability of a biomarker, measured on a continuous scale, between two states of a disease; e.g., Shapiro (1999), Zhou et al. (2002) and Pepe (2003). In practice, investigators often need to compare the diagnostic accuracies between two biomarkers or diagnostic tests. The comparison of the overall diagnostic accuracy between two biomarkers measured simultaneously on an individual is frequently addressed by comparing the resulting paired *AUC*s. For example, Delong et al. (1988) presented a non-parametric approach to the analysis of areas under correlated ROC curves; Obuchowski (1997) and Zhou et al. (2002) applied two one-sided tests to evaluate the two-sided equivalence of two diagnostic procedures; Wieand et al. (1989) proposed non-parametric and parametric tests for the same problem; Molodianovitch et al. (2006) extended Wieand et al.'s approach to non-normal data by using Box–Cox

transformation; Vexler et al. (2008) applied the maximum likelihood technique to compare *AUC*s for data with limits of detection; Liu et al. (2006) proposed to use the standardized difference for assessing equivalence of paired *AUC*s; and Li et al. (2008) show that the generalized variable approach is very appropriate to make inference about paired *AUC*s.

Under many circumstances, the research interest only lies in lower part of range of false-positive rates, that is, a minimum acceptable specificity is imposed. Therefore, the partial area under ROC curve (*PAUC*) over a practicably relevant range of false positive rate can be considered as a reasonable summary measure of diagnostic accuracy. There exist a few methods for estimating and comparing *PAUC*s. For example, McClish (1989, 1990) proposed a method for comparing *PAUC*s under the assumption of binormal model; Thompson and Zucchini (1989) proposed an ANOVA model to compare *PAUC*s; Jiang et al. (1996) extended Mcclish's work to highly sensitive diagnostic tests; Zhang et al. (2002) presented a non-parametric method for comparing *PAUC*s; Dodd and Pepe (2003) discussed a parsimonious regression model of *PAUC*; recently, Li et al. (2007) proposed a generalized variable approach for comparing paired *PAUC*s for normally distributed data.

In practice, there exist many disease processes with three ordinal disease classes. For example, mild cognitive impairment (MCI) and/or early stage Alzheimer's disease is a transitional stage between the cognitive changes of normal aging and the more serious problems caused by Alzheimer's disease (AD) stated in Xiong et al. (2006). As another example, in a study of iron deficiency related anemia by Wians et al. (2001), non-pregnant women with anemia and a ferritin concentration less than 20 $\mu\text{g/l}$ were considered to have iron deficiency anemia (IDA) and the ones with anemia and a ferritin concentration greater than 240 $\mu\text{g/l}$ were considered to have anemia of chronic disease (ACD), while the ones with a ferritin concentration between 20 and 240 $\mu\text{g/l}$ were considered to belong to the intermediate group. Since patients at different disease states require different treatments, it is important to have good diagnostic tests which can discriminate among these three ordinal diagnostic groups. Thereafter, we refer the state between "diseased" and "healthy" as "intermediate", in other words, transitional or early/mild diseased. To be specific, denote Y_1 , Y_2 and Y_3 the scores of a biomarker or results of a diagnostic test and let F_1 , F_2 and F_3 be the corresponding cumulative distribution functions for non-diseased, intermediate and diseased groups, respectively. Assume the results of a diagnostic test are measured on continuous scale and higher values indicate greater severity of the disease. Let $p_1 = F_1(c_1)$, $p_3 = 1 - F_3(c_3)$, where c_1 and c_3 are threshold values ($c_1 < c_3$) for non-diseased group and diseased group, respectively, be true classification rates for non-diseased and diseased groups, respectively. Then the probability that a randomly selected subject from intermediate group has a score between c_1 and c_3 is

$$p_2 = F_2(c_3) - F_2(c_1) = F_2 \left[F_3^{-1}(1 - p_3) \right] - F_2 \left[F_1^{-1}(p_1) \right]. \quad (1)$$

As a function of (p_1, p_3) , $p_2 = p_2(p_1, p_3)$ defines a surface in the three-dimensional space (p_1, p_3, p_2) , called ROC surface. The point $(p_1, p_3, p_2) = (1, 1, 1)$ indicate perfect discrimination ability of the marker between three ordinal disease groups. The volume under surface (*VUS*) can be used as a summary measure of the diagnostic accuracy. One can prove that *VUS*, analogous measure to the *AUC* for binary classification, equals to the probability that a random selected triple with one individual from each diagnostic group have the correct ordering (i.e. $Y_1 < Y_2 < Y_3$). More details were given by Nakas and Yiannoutsos (2004) and Xiong et al. (2006). Similar to *PAUC* for the cases with two diagnostic groups, the partial volume under surface (*PVUS*) has been proposed to denote the diagnostic accuracy with pre-specified minimum classification rates for the cases with three ordinal diagnostic groups. Recently, the traditional ROC analysis based on a binary gold-standard for the true disease status has been extended to three diagnostic groups; e.g., Mossman (1999), Dreiseitl

et al. (2000) and Heckerling (2001). Furthermore, Nakas and Yiannoutsos (2004) proposed distribution-free approaches for hypothesis testing for a single *VUS* and paired *VUS*s; Xiong et al. (2006) developed an asymptotic approach for confidence interval estimation of *VUS* and *PVUS* for normally distributed data; Nakas and Alonzo (2007) and Alonzo and Nakas (2007) proposed non-parametric inference procedures for diagnostic accuracy with three disease classes under umbrella ordering; and Xiong et al. (2007) developed a large sample approach for comparing several *VUS*s for normally distributed data. Most recently, Li and Fine (2008) proposed a method for ROC analysis with multiple classes.

The aim of this paper is to develop an approach for confidence interval estimation of the difference in paired *VUS*s and paired *PVUS*s based on the concepts of generalized inference. The generalized variables (GV) and generalized pivots were introduced by Tsui and Weerahandi (1989) and Weerahandi (1993); see the book by Weerahandi (2003) or a detailed discussion. A brief summary of the concepts is included in Appendix. The concepts of generalized confidence interval and generalized *P*-value have been successfully applied to a variety of practical settings where standard exact solutions do not exist for confidence intervals and hypothesis testing. It has been shown that generalized variable approaches typically have good performance at small sample sizes; e.g. Weerahandi (1995), Weerahandi and Berger (1999), Krishnamoorthy and Lu (2003), Tian and Cappelleri (2004), Iyer et al. (2004), and Krishnamoorthy et al. (2009). Especially, as mentioned aforehand, GV approaches were proposed to construct an exact test for equivalence of diagnostic accuracy based on paired *PAUC*s by Li et al. (2007) and to estimate confidence interval of the difference in paired *AUC*s by Li et al. (2008) for normally distributed data.

This paper is organized as follows. Section 2 presents the preliminary knowledge about *VUS* and *PVUS*. In Section 3, the GV approaches for confidence interval estimation of the difference in paired *VUS*s and paired *PVUS*s are proposed. In Section 4, simulation results are presented to evaluate the performance of the proposed approach. In Section 5, the proposed approach is applied to a data set from a study of anemia subjects. Section 6 presents summary and discussion. Appendix contains a brief review of the basic concepts of generalized inference.

2. Preliminaries

In the following, we will briefly review the definitions of *VUS* and *PVUS* for which more details can be found in Xiong et al. (2006).

Denote Y_1 , Y_2 and Y_3 to be the scores of a biomarker or the results by a diagnostic test and let F_1 , F_2 and F_3 be the corresponding cumulative distribution functions for non-diseased, intermediate and diseased groups, respectively. Assume the responses are measured on a continuous scale and higher values indicate greater severity of the disease. Let c_1 and c_3 be threshold values ($c_1 < c_3$) for the non-diseased group and the diseased group, respectively. Following Eq. (1), the volume under the ROC surface can be proved to be equal to the probability that Y_1 , Y_2 and Y_3 are in correct order, that is

$$VUS = P(Y_1 < Y_2 < Y_3) = \int_0^1 \int_0^{1-F_3[F_1^{-1}(p_1)]} \left\{ F_2[F_3^{-1}(1-p_3)] - F_2[F_1^{-1}(p_1)] \right\} dp_3 dp_1. \quad (2)$$

The partial volume under surface is defined by

$$PVUS = \iint_{D_{p_{10}p_{30}}} \left\{ F_2[F_3^{-1}(1-p_3)] - F_2[F_1^{-1}(p_1)] \right\} dp_3 dp_1, \quad (3)$$

where $D_{p_{10}p_{30}} = \{(p_1, p_3) | p_{10} \leq p_1 \leq 1, p_{30} \leq p_3 \leq 1 - F_3[F_1^{-1}(p_1)]\}$ with p_{10} and p_{30} as the minimum desired rates for non-diseased and diseased groups, respectively. In other words, p_{10} and p_{30} represent the minimum desired specificity and the sensitivity between non-diseased and diseased groups. When non-diseased, intermediate and diseased groups can be discriminated perfectly, $PVUS$ reach its maximum value as $PVUS_{max} = (1 - p_{10})(1 - p_{30})$. A value of $PVUS$ closer to $PVUS_{max}$ indicate a better discrimination ability of this biomarker among three ordinal diagnostic groups. As $p_{10} = p_{30} = 0$, $PVUS = VUS$.

Assume $Y_i \sim N(\mu_i, \sigma_i^2)$ where $i = 1, 2, 3$. Define ratios $a = \sigma_2/\sigma_1$, $b = (\mu_1 - \mu_2)/\sigma_1$, $c = \sigma_2/\sigma_3$ and $d = (\mu_3 - \mu_2)/\sigma_3$. The volume under the ROC surface and the partial volume under surface as stated in Xiong et al. (2006) are

$$VUS = \int_{-\infty}^{\infty} \Phi(as - b) \Phi(-cs + d) \phi(s) ds, \quad (4)$$

$$PVUS = \int_{[\Phi^{-1}(p_{10})+b]/a}^{[d-\Phi^{-1}(p_{30})]/c} [\Phi(as - b) \Phi(-cs + d) - p\Phi(-cs + d) - q\Phi(as - b) + pq] \phi(s) ds. \quad (5)$$

3. The generalized variable approach

3.1. Differences in paired VUS s and PVUS s

In this paper, we focus on the inference about the differences in paired VUS s and $PVUS$ s for the cases with paired data. Let $\mathbf{Y}_1$, $\mathbf{Y}_2$ and $\mathbf{Y}_3$ be two-dimensional vectors denoting the scores for markers A and B measured simultaneously for the non-diseased, intermediate, and diseased groups, respectively. Specifically, assume

$$\mathbf{Y}_i = \begin{pmatrix} Y_{iA} \\ Y_{iB} \end{pmatrix} \sim N_2(\mu_i, \Sigma_i) \quad \text{for } i=1, 2, 3, \quad (6)$$

where

$$\mu_i = \begin{pmatrix} \mu_{iA} \\ \mu_{iB} \end{pmatrix} \quad \text{and} \quad \Sigma_i = \begin{pmatrix} \sigma_{iA}^2 & \sigma_{iAB}^2 \\ \sigma_{iAB}^2 & \sigma_{iB}^2 \end{pmatrix}.$$

For the marker A , we denote the ratios $a_A = \sigma_{2A}/\sigma_{1A}$, $b_A = (\mu_{1A} - \mu_{2A})/\sigma_{1A}$, $c_A = \sigma_{2A}/\sigma_{3A}$ and $d_A = (\mu_{3A} - \mu_{2A})/\sigma_{3A}$. The volume under the surface and the partial volume under the surface for marker A are

$$VUS_A = P(Y_{1A} < Y_{2A} < Y_{3A}) = \int_{-\infty}^{\infty} \Phi(a_A s - b_A) \Phi(-c_A s + d_A) \phi(s) ds, \quad (7)$$

$$PVUS_A = \int_{[\Phi^{-1}(p_{10})+b_A]/a_A}^{[d_A-\Phi^{-1}(p_{30})]/c_A} [\Phi(a_A s - b_A) \Phi(-c_A s + d_A) - p\Phi(-c_A s + d_A) - q\Phi(a_A s - b_A) + pq] \phi(s) ds, \quad (8)$$

respectively. It is clear that, for the marker B , the volume under surface VUS_B and partial volume under surface $PVUS_B$ can be obtained by replacing A with B in Eqs. (7) and (8).

In this article, we propose methods to estimate the confidence intervals of the difference between the paired VUS and $PVUS$, i.e.

$$\Delta VUS = VUS_A - VUS_B, \quad (9)$$

$$\Delta PVUS = PVUS_A - PVUS_B. \quad (10)$$

3.2. Generalized pivots for ΔVUS and $\Delta PVUS$

Denote $N_p(\mu_i, \Sigma_i)$ to be a p -variate normal distribution functions with mean vector μ_i and covariate matrix Σ_i . Assume that $Y_{1,1}, Y_{2,1}, \dots, Y_{n_1,1}$ consist a sample from $N_2(\mu_1, \Sigma_1)$ for non-diseased group; $Y_{1,2}, Y_{2,2}, \dots, Y_{n_2,2}$ consist a sample from $N_2(\mu_2, \Sigma_2)$ for intermediate group; and $Y_{1,3}, Y_{2,3}, \dots, Y_{n_3,3}$ consist a sample from $N_2(\mu_3, \Sigma_3)$ for diseased group. For the i th population, let $\bar{Y}_i$ and S_i be the sample mean vector and sample covariance matrix, respectively. It is well-known that $\bar{Y}_i$ and S_i are mutually independent as well as

$$\bar{Y}_i \sim N_2(\mu_i, \Sigma_i/n_i) \quad \text{and} \quad U_i = \frac{n_i - 1}{n_i} S_i \sim W_2(n_i - 1, \Sigma_i/n_i), \quad i=1, 2, 3, \quad (11)$$

where $W_p(m, \Sigma)$ denotes a p -dimensional Wishart distribution with degrees of freedom m and scale matrix Σ .

The generalized pivots for μ_i can be given as (Lin et al., 2007)

$$R_{\mu_i} = \bar{y}_i - (\mathbf{u}_i^{1/2} \mathbf{W}_i^{-1} \mathbf{u}_i^{1/2})^{1/2} \mathbf{Z}_i \quad \text{for } i=1, 2, 3, \quad (12)$$

where $\mathbf{Z}_i \sim N_2(0, I_2)$ with I_2 as a 2 by 2 identity matrix and the generalized pivot for Σ_i can be given as

$$R_{\Sigma_i} = n_i \mathbf{u}_i^{1/2} \mathbf{W}_i^{-1} \mathbf{u}_i^{1/2} \quad \text{for } i=1, 2, 3, \quad (13)$$

where $\mathbf{W}_i \sim W_2(n_i - 1, I_2)$. Note that

$$R_{\mu_i} = \begin{pmatrix} R_{\mu_{iA}} \\ R_{\mu_{iB}} \end{pmatrix} \quad \text{and} \quad R_{\Sigma_i} = \begin{pmatrix} R_{\sigma_{iA}^2} & R_{\sigma_{iAB}^2} \\ R_{\sigma_{iAB}^2} & R_{\sigma_{iB}^2} \end{pmatrix}$$

for $i = 1, 2, 3$. Hence we can obtain generalized pivots $R_{a_A}, R_{b_A}, R_{c_A}, R_{d_A}$ for a_A, b_A, c_A, d_A in the following forms:

$$R_{a_A} = \frac{R_{\sigma_{2A}}}{R_{\sigma_{1A}}}, \quad (14)$$

$$R_{b_A} = \frac{R_{\mu_{1A}} - R_{\mu_{2A}}}{R_{\sigma_{1A}}}, \quad (15)$$

$$R_{c_A} = \frac{R_{\sigma_{2A}}}{R_{\sigma_{3A}}}, \quad (16)$$

$$R_{d_A} = \frac{R_{\mu_{3A}} - R_{\mu_{2A}}}{R_{\sigma_{3A}}}. \quad (17)$$

The generalized pivots R_{VUS_A} and R_{PVUS_A} for VUS and $PVUS$ for marker A can be derived by substituting a_A, b_A, c_A, d_A with their corresponding generalized pivots $R_{a_A}, R_{b_A}, R_{c_A}, R_{d_A}$ as follows:

$$R_{VUS_A} = \int_{-\infty}^{+\infty} \Phi(R_{a_A}s - R_{b_A}) \Phi(-R_{c_A}s + R_{d_A}) \phi(s) ds, \quad (18)$$

$$R_{PVUS_A} = \int_{\left(\Phi^{-1}(p_{10}) + R_{b_A}\right)/R_{a_A}}^{\left(R_{d_A} - \Phi^{-1}(p_{30})\right)/R_{c_A}} \left[\Phi(R_{a_A}s - R_{b_A}) \Phi(-R_{c_A}s + R_{d_A}) - p\Phi(-R_{c_A}s + R_{d_A}) - q\Phi(R_{a_A}s - R_{b_A}) + pq \right] \phi(s) ds. \quad (19)$$

For marker B , we can obtain generalized pivots $R_{a_B}, R_{b_B}, R_{c_B}, R_{d_B}$ for a_B, b_B, c_B, d_B similarly to (14)–(17). Using $R_{a_B}, R_{b_B}, R_{c_B}$ and R_{d_B} one can calculate R_{PVUS_B} and R_{VUS_B} in the similar way as R_{VUS_A} and R_{PVUS_A} .

One can easily check that $R_{VUS_A}, R_{PVUS_A}, R_{VUS_B}$ and R_{PVUS_B} are bona fide generalized pivots as follows. For given $\bar{y}_i$ and $\mathbf{u}_i$ ($i = 1, 2, 3$), the following holds: (1) the distributions of $R_{VUS_A}, R_{PVUS_A}, R_{VUS_B}$ and R_{PVUS_B} are independent of any unknown parameters, and (2) the value of $R_{VUS_A}, R_{PVUS_A}, R_{VUS_B}$ and R_{PVUS_B} are $VUS_A, PVUS_A, VUS_B$ and $PVUS_B$ as $\bar{Y}_i = \bar{y}_i, \mathbf{U}_i = \mathbf{u}_i$ for $i = 1, 2, 3$.

Remark 3.1—Note that R_{μ_i} and R_{Σ_i} are defined using bivariate mean vector $\bar{Y}_i$ and the corresponding scaled sample variance matrix $\mathbf{U}_i$ which incorporate markers A and B simultaneously. Therefore, the facts that markers A and B are paired and hence VUS_A and VUS_B (or $PVUS_A$ and $PVUS_B$) are not independent are taken care of automatically in this proposed approach.

Furthermore, the generalized pivots for $\Delta PVUS$ and ΔVUS can be defined as

$$R_{\Delta VUS} = R_{VUS_A} - R_{VUS_B}, \quad (20)$$

$$R_{\Delta PVUS} = R_{PVUS_A} - R_{PVUS_B}. \quad (21)$$

For testing the hypothesis $H_0 : \Delta VUS = \Delta VUS_0$ vs. $H_1 : \Delta VUS > \Delta VUS_0$, the generalized test variable is defined as

$$T_{\Delta VUS} = R_{\Delta VUS} - \Delta VUS_0. \quad (22)$$

It is clear that $T_{\Delta VUS}$ is a bona fide generalized test variable. In a similar manner, the generalized variable

$$T_{\Delta PVUS} = R_{\Delta PVUS} - \Delta PVUS_0. \quad (23)$$

for testing the hypothesis $H_0 : \Delta PVUS = \Delta PVUS_0$ vs. $H_1 : \Delta PVUS > \Delta PVUS_0$.

3.3. Computing algorithm

For a given data set containing outcomes of markers A and B which are measured simultaneously on non-diseased, intermediate and diseased groups, respectively, the

generalized confidence intervals for ΔVUS and $\Delta PVUS$ can be obtained following the simulation steps:

1. Compute the sample mean vector $\bar{\mathbf{y}}_i$ and sample covariance matrix $\mathbf{s}_i$ for $i = 1, 2, 3$.
2. Generate $Z_i \sim N_2(0, I_2)$ and $\mathbf{W}_i \sim W_2(n_i - 1, I_2)$. Calculate R_{μ_i} and R_{Σ_i} for $i = 1, 2, 3$, following (12) and (13).
3. Compute $R_{aA}, R_{bA}, R_{cA}, R_{dA}$ and $R_{aB}, R_{bB}, R_{cB}, R_{dB}$ following (14)–(17).
4. Compute $R_{\Delta VUS}$ and $R_{\Delta PVUS}$ following (20) and (21).
5. Repeat Steps 2–5 a total m times (in general, m is set as 2000) and obtain an array of $R_{\Delta VUS}$'s values and an array of $R_{\Delta PVUS}$'s values.
6. Rank the array of $R_{\Delta VUS}$'s and the array of $R_{\Delta PVUS}$'s from small to large.

Denote $R_{\Delta VUS}(\alpha)$ as the 100 α th percentile of $R_{\Delta VUS}$'s. Then $(R_{\Delta VUS}(\alpha/2), R_{\Delta VUS}(1-\alpha/2))$ is a two-sided 100(1- α)% confidence interval of ΔVUS . The percentage that $R_{\Delta VUS}$'s are less than or equal to ΔVUS_0 is a Monte Carlo estimate of the generalized P -value for testing $\Delta VUS = \Delta VUS_0$ vs. $\Delta VUS > \Delta VUS_0$. Similarly, the generalized P -values for testing $\Delta VUS = \Delta VUS_0$ vs. $\Delta VUS < \Delta VUS_0$ can be obtained. The confidence interval estimation and hypothesis testing about $\Delta PVUS$ can be done similarly.

4. A simulation study

Simulation studies were performed to assess the coverage probabilities of the proposed confidence interval estimations for ΔVUS and $\Delta PVUS$. We generated data from the bivariate normal distributions $N_2(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1), N_2(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2), N_2(\boldsymbol{\mu}_3, \boldsymbol{\Sigma}_3)$ for non-diseased, intermediate and diseased groups, respectively, with $\boldsymbol{\mu}_1 = (0, 0)'$, $\boldsymbol{\mu}_2 = (2, 3)'$, $\boldsymbol{\mu}_3 = (5, 4)'$, or $\boldsymbol{\mu}_1 = (0, 0)'$, $\boldsymbol{\mu}_2 = (0.5, 0.3)'$, $\boldsymbol{\mu}_3 = (1.0, 0.6)'$ and with three different configurations of $(\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2, \boldsymbol{\Sigma}_3)$ as follow:

$$\text{config.1: } \left(\begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix} \right),$$

$$\text{config.2: } \left(\begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}, \begin{pmatrix} 4 & 2 \\ 2 & 4 \end{pmatrix}, \begin{pmatrix} 16 & 8 \\ 8 & 16 \end{pmatrix} \right),$$

$$\text{config.2: } \left(\begin{pmatrix} 1 & 1 \\ 1 & 4 \end{pmatrix}, \begin{pmatrix} 1 & 1 \\ 1 & 4 \end{pmatrix}, \begin{pmatrix} 1 & 1 \\ 1 & 4 \end{pmatrix} \right).$$

The different configurations of $(\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2, \boldsymbol{\Sigma}_3)$ were chosen to represent three possible scenarios of covariance structures of the three disease groups. Config. 1 stands for the scenario with equal variances for markers A and B across three disease groups; Config. 2 stands for the scenario with equal variances for markers A and B but with increasing variances across three disease groups from non-diseased group to diseased group; Config. 3 stands for the scenario with unequal variances for markers A and B .

Table 1 presents the coverage probabilities of proposed confidence intervals for ΔVUS at nominal level 0.95 based on 2000 random samples, in comparison with those of the large sample approach (Xiong et al., 2007) and a parametric bootstrap approach. To estimate the

confidence intervals by the proposed generalized variable approach, within each of the 2000 random samples, 2000 $R_{\Delta VUS}$'s were calculated using the algorithms presented in Section 3.3. The large sample approach proposed by Xiong et al. (2007) for parametric comparison of VUS s can provide confidence intervals for the differences in paired VUS s. Additionally, we also consider percentile intervals by parametric bootstrap. Overall speaking, the proposed exact confidence intervals provide reasonable coverage whether or not the sample sizes are imbalanced across diagnostic groups except that it tends to be slightly conservative as sample sizes are small. When the sample sizes are the same across diagnostic groups, the large sample approach works reasonably well even as the sample sizes are quite small; however, the imbalance in sample sizes also seems to be associated with poor coverage probabilities in comparing paired VUS s, regardless of whether the covariance matrices are the same across diagnostic groups. This similar phenomenon was also observed in simulation studies regarding the coverage probability of Student t -confidence intervals when two means are compared from two normal distributions with unequal variances (Milliken and Johnson, 1992). Note that for the setting $\mu_1 = (0,0)'$, $\mu_2 = (2,3)'$, $\mu_3 = (5,4)'$, some of the true VUS is large, and the Fisher z -transformation was then used in calculating the CI for the difference on VUS , as recommended by Xiong et al. (2007). The parametric bootstrap approach generally performs well; however, for scenarios with small sample sizes, it also can underestimate the coverage probabilities regardless of settings for the covariance matrices. For example, as the sample sizes equal (10, 10, 5), the coverage probabilities by the bootstrap approach can be as low as 0.90.

Table 2 presents simulation results for coverage probabilities of confidence intervals for $\Delta PVUS$ by the proposed approach and the parametric bootstrap approach. Note that the large sample approach for comparing paired $PVUS$ s has not been developed, and therefore it is not included in the comparison. The desired minimum classification rates for non-diseased and diseased groups were set as $p_{10} = p_{30} = 0.5$, i.e. $\Delta PAUC$ is obtained for the region with both the minimum desired specificity and sensitivity for diseased and non-diseased groups as 0.5, similar to the setting in the asymptotic approach by Xiong et al. (2006). In general, the proposed approach works well except that for a few scenarios it tends to be slightly liberal. The parametric bootstrap approach works well for balance cases, but can slightly underestimate the coverage probabilities, e.g., as the sample sizes equal (50, 20, 10), the coverage probabilities can be as low as 0.92.

To investigate the robustness of the proposed test, a simulation study was conducted for the mixture of normal data. The results are shown in Table 3. For each parameter setting presented, the mixture of normal data was generated as a mixture of two normal distributions; i.e.,

$$Y_i = \begin{pmatrix} Y_{iA} \\ Y_{iB} \end{pmatrix} \sim (1 - \alpha) N_2 \left(\mu_i, \frac{1}{1.1} \Sigma_i \right) + \alpha N_2 \left(\mu_i, \frac{2}{1.1} \Sigma_i \right),$$

where $\alpha = 0.1$ for $i = 1, 2, 3$. Note that $\mathbf{Y}_i$ from such a normal mixture distribution has mean μ_i and variance Σ_i . The results presented in Table 3 show that the proposed approach generally has satisfactory coverage probabilities for normal mixture data.

Remark 4.1

As stated in Section 3, the proposed approach can easily provide P -values for hypothesis testing. An additional Simulation study shows that the proposed test has satisfying type-I error control. These simulation results are not presented in this article, but they are available upon requests.

5. An example

In this section, the proposed approach for confidence interval estimations of ΔVUS and $\Delta PVUS$ is illustrated via a data set of blood test results for patients with anemia (Wians et al., 2001). These data were also used by Obuchowski (2006). A total of 134 patients with anemia underwent a series of blood tests. To eliminate the bias which might cause by gender, we will limit our analysis to the 55 female study patients. Ferritin concentration provides a useful screening test for iron deficiency anemia (IDA). Non-pregnant women with anemia and a ferritin concentration less than 20 $\mu\text{g/l}$ were assigned to the IDA group, while those with anemia and a ferritin concentration greater than 240 $\mu\text{g/l}$ were assigned to be the anemia of chronic disease (ACD) group. The intermediate group consists of the women with ferritin concentration between 20 and 240 $\mu\text{g/l}$. There were 12, 14, 29 female study subjects in ACD, intermediate and IDA groups, respectively. We are interested in comparison of the diagnostic accuracy between two rapid blood tests, i.e. total iron binding capacity (TIBC) and per cent transferrin saturation (%TS), for discriminating between the ACD, intermediate and IDA groups. The multivariate normality for each group was tested by Henze–Zirkler (1990) test and not rejected with P -values as 0.14, 0.26 and 0.46 for ACD, intermediate and IDA groups, respectively.

The sample mean for ACD group is $\bar{\mathbf{x}}_1 = (TIBC, \%TS)' = (214:00, 6:42)'$ and the sample covariance is

$$\mathbf{s}_1 = \begin{pmatrix} 2803.82 & -54.45 \\ -54.45 & 9.17 \end{pmatrix},$$

the sample mean for intermediate group is $\bar{\mathbf{x}}_2 = (282:64, 5:07)'$ and the sample covariance is

$$\mathbf{s}_2 = \begin{pmatrix} 3881.79 & -16.90 \\ -16.90 & 6.69 \end{pmatrix},$$

and the sample mean for IDA group is $\bar{\mathbf{x}}_3 = (430:14, 3:53)'$ and the sample covariance is

$$\mathbf{s}_3 = \begin{pmatrix} 8107.34 & -56.57 \\ -56.57 & 3.54 \end{pmatrix},$$

The point estimates of VUS and $PVUS$ are 0.7036 and 0.1218 for TIBC, and 0.3607 and 0.0114 for %TS, respectively. Note that TIBC increases from the ACD group to the IDA group while %TS decreases from ACD group to IDA group. The 95% confidence intervals by the proposed approach are (0.1103, 0.5139) for ΔVUS and (0.038, 0.1515) for $\Delta PVUS$ respectively. Both results show that TIBC has better diagnostic ability than %TS in discriminating subjects with anemia among ACD group, intermediate group and IDA group.

6. Summary and discussion

This article focuses on the confidence interval estimation of the differences in paired volumes under surfaces (VUS) and paired partial volumes ($PVUS$) under surfaces based on the generalized inference theory. In addition to confidence interval estimation, the proposed approach can easily provide P -values for hypothesis testing. The proposed approach is a numeric method which involves generating multivariate normal data and can be performed

using a few straightforward simulation steps presented in Section 3.3. Considering the facts that (1) the large sample approach for comparing two *VUS*s can be very liberal; (2) the large sample approach for comparing two *PVUS*s has not been developed and it is expected to be not so straightforward; and (3) the parametric bootstrap approach also can underestimate the coverage probabilities, the proposed approach based on generalized inference can serve as a good candidate for confidence interval estimate of the difference of paired *VUS*s and paired *PVUS*s, especially in cases with small to medium sample sizes.

All the inference procedures based on the generalized variable theory requires parametric assumptions. For example, when there are two diagnostic groups, the generalized variable approaches proposed by Li et al. (2007, 2008) for inferences about ΔAUC and $\Delta PAUC$ are based on normality assumptions of the data distributions. In parallel, when there are three diagnostic groups, the proposed generalized variable approach for ΔVUS and $\Delta PVUS$ also utilize normality assumptions. It is well-known that *AUC* and *PAUC* are invariant measures of diagnostic accuracy under any monotonic transformation. Similarly, *VUS* and *PVUS* are also invariant measures under any monotonic transformations. Therefore, the proposed approach is expected to have wide practical applications due to the fact that it not only accommodates multivariate normal data, also any data from distributions which can be transformed into normal by monotonic transformation. In their paper (Xiong et al., 2006) on asymptotic approach based on normality for confidence interval estimation for the *VUS* and *PVUS*, Xiong et al.'s stated that it is important to check the model assumption of the original and transformed data before applying their asymptotic approach. Similarly, we recommend to check the model assumption of the original and transformed data before applying the proposed generalized variable approach. To investigate the robustness of the proposed approach, a simulation study based on mixture of multivariate normal distribution is performed and simulation results show that the proposed approach give satisfactory results.

Acknowledgments

The work by Dr. Xiong was partly supported by Grants NIH/NIA R01 AG029672, AG003991, AG005681, and AG026276 from the National Institute on Aging and Grant NIRG-08-91082 from the Alzheimer's Association.

Appendix A. Generalized pivots and generalized test variables

In the following, the basic concepts for generalized inference developed by Tsui and Weerahandi (1989) and Weerahandi (1993) are described.

Suppose that $\mathbf{Y} = (Y_1, Y_2, \dots, Y_n)'$ form a random sample from a distribution which depends on the parameters $\theta = (\psi, \nu)$ where ψ is the parameter of interest and ν' is a vector of nuisance parameters. A generalized pivot $R(\mathbf{Y}; \mathbf{y}, \psi, \nu)$, where $\mathbf{y}$ is an observed value of $\mathbf{Y}$, for interval estimation defined in Weerahandi (1993), has the following two properties:

1. $R(\mathbf{Y}; \mathbf{y}, \psi, \nu)$ has a distribution free of unknown parameters.
2. The value of $R(\mathbf{y}; \mathbf{y}, \psi, \nu)$ is ψ .

Let that R_α be the 100α th percentile of R . Then R_α becomes the $100(1 - \alpha)\%$ lower bound for ψ and $(R_{\alpha/2}, R_{1-\alpha/2})$ becomes a $100(1 - \alpha)\%$ two-sided generalized confidence interval for ψ .

Now consider testing $H_0 : \psi = \psi_0$ vs. $H_1 : \psi > \psi_0$ where ψ_0 is a specified quantity. A generalized test variable of the form $T(\mathbf{Y}; \mathbf{y}, \psi, \nu)$ where $\mathbf{y}$ is an observed value of $\mathbf{Y}$, is chosen to satisfy the following three conditions (Tsui and Weerahandi, 1989):

1. For fixed y , the distribution of $T(Y; y, \psi, \nu)$ is free of the vector of nuisance parameters ν .
2. The value of $T(Y; y, \psi, \nu)$ at $Y = y$ is free of any unknown parameters.
3. For fixed y and ν , and for all t , $\Pr[T(Y; y, \psi, \nu) > t]$ is a monotonic function of ψ .

A generalized extreme region is defined as $C = \{Y : T(Y; y, \psi, \nu) \geq T(y; y, \psi, \nu) \text{ if } T(Y; y, \psi, \nu) \text{ is stochastically increasing in } \psi \text{ (i.e. } \Pr[T(Y; y, \psi, \nu) > t] \text{ is a non-decreasing function of } \psi)\}$. If $T(Y; y, \psi, \nu)$ is stochastically decreasing in ψ (i.e. $\Pr[T(Y; y, \psi, \nu) > t]$ is a non-increasing function of ψ), a generalized extreme region is defined as $C = \{Y : T(Y; y, \psi, \nu) \leq T(y; y, \psi, \nu)\}$. Then the generalized P -value is defined as $P(C|\psi)$.

References

- Alonzo T, Nakas C. Comparison of ROC umbrella volumes with an application to the assessment of lung cancer diagnostic markers. *Biometrical Journal*. 2007; 49:654–664. [PubMed: 17726716]
- Delong ER, Delong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrika*. 1988; 44:837–845.
- Dodd LE, Pepe MS. Partial AUC estimation and regression. *Biometrics*. 2003; 59:614–623. [PubMed: 14601762]
- Dreiseitl S, Ohno-Machado L, Binder M. Comparing three-class diagnostic tests by three-way ROC analysis. *Medical Decision Making*. 2000; 20:323–331. [PubMed: 10929855]
- Heckerling PS. Parametric three-way receiver operating characteristic surface analysis using mathematica. *Medical Decision Making*. 2001; 21:409–417. [PubMed: 11575490]
- Henze N, Zirkler B. A class of invariant consistent tests for multivariate normality. *Communications in Statistics: Theory and Methods*. 1990; 19:3595–3617.
- Iyer HK, Wang CM, Mathew T. Models and confidence intervals for true values in interlaboratory trials. *Journal of the Acoustical Society of America*. 2004; 99:1060–1071. (12).
- Jiang Y, Metz CE, Nishikawa RM. A receiver operating characteristic partial area index for highly sensitive diagnostic tests. *Radiology*. 1996; 201:621–625. [PubMed: 8939207]
- Krishnamoorthy K, Lu Y. Inference on the common means of several normal populations based on the generalized variable method. *Biometrics*. 2003; 59:237–247. [PubMed: 12926708]
- Krishnamoorthy K, Lin Y, Xia Y. Confidence limits and prediction limits for a Weibull distribution based on the generalized variable approach. *Journal of Statistical Planning and Inference*. 2009; 139:2675–2684.
- Li C, Liao C, Liu J. A non-inferiority test for diagnostic accuracy based on the paired partial areas under ROC curves. *Statistics in Medicine*. 2007; 27:1762–1776. [PubMed: 17968858]
- Li C, Liao C, Liu J. On the exact interval estimation for the difference in paired areas under the ROC curves. *Statistics in Medicine*. 2008; 27:224–242. [PubMed: 17139702]
- Li J, Fine J. ROC analysis with multiple classes and multiple tests: methodology and its application in microarray studies. *Biostatistics*. 2008; 9:566–576. [PubMed: 18304996]
- Lin SH, Lee JC, Wang RS. Generalized inferences on the common mean vector of several multivariate normal populations. *Journal of Statistical Planning and Inference*. 2007; 137:2240–2249.
- Liu J, Ma M, Wu C, Tan J. Tests of equivalence and non-inferiority for diagnostic accuracy based on the paired areas under ROC curves. *Statistics in Medicine*. 2006; 25:1219–1238. [PubMed: 16158400]
- McClish DK. Analyzing a portion of the ROC Curve. *Medical Decision Making*. 1989; 9:190–195. [PubMed: 2668680]
- McClish DK. Determining a range of false-positive rates for which ROC curves differ. *Medical Decision Making*. 1990; 10:283–297. [PubMed: 2233158]
- Milliken, GA.; Johnson, DE. *Analysis of Messy Data, Volume 1: Designed Experiments*. Chapman & Hall/CRC; New York: 1992.

- Molodianovitch K, Faraggi D, Reiser B. Comparing the areas under two correlated ROC curves: parametric and non-parametric approaches. *Biometrical Journal*. 2006; 48:745–757. [PubMed: 17094340]
- Mossman D. Three-way ROCs. *Medical Decision Making*. 1999; 19:78–89. [PubMed: 9917023]
- Nakas C, Yiannoutsos C. Ordered multiple-class ROC analysis with continuous measurement. *Statistics in Medicine*. 2004; 23:3437–3449. [PubMed: 15505886]
- Nakas C, Alonzo T. ROC graphs for assessing the ability of a diagnostic marker to detect three disease classes with an umbrella ordering. *Biometrics*. 2007; 63:603–609. [PubMed: 17688513]
- Obuchowski N. Testing for equivalence of diagnostic tests. *American Journal of Radiology*. 1997; 168:13–17.
- Obuchowski N. An ROC-type measure of diagnostic accuracy when the gold standard is continuous-scale. *Statistics in Medicine*. 2006; 25:481–493. [PubMed: 16287217]
- Pepe MS. *The Statistical Evaluation of Medical Tests for Classification and Prediction* (2002). Oxford Statistical Science Series. 2003; vol. 28
- Shapiro D. The interpretation of diagnostic tests. *Statistical Methods in Medical Research*. 1999; 8:113–134. [PubMed: 10501649]
- Thompson ML, Zucchini W. On the statistical analysis of ROC curves. *Statistics in Medicine*. 1989; 8:1277–1290. [PubMed: 2814075]
- Tian L, Cappelleri JC. A new approach for interval estimation and hypothesis testing of a certain intraclass correlation coefficient: the generalized variable method. *Statistics in Medicine*. 2004; 23:2125–2135. [PubMed: 15211607]
- Tsui KW, Weerahandi S. Generalized P -values in significance testing of hypotheses in the presence of nuisance parameters. *Journal of American Statistical Association*. 1989; 84:602–607.
- Vexler A, Liu A, Eliseeva E, Schisterman EF. Maximum likelihood ratio tests for comparing the discriminatory ability of biomarkers subject to limit of detection. *Biometrics*. 2008; 64:895–903. [PubMed: 18047527]
- Weerahandi S. Generalized confidence intervals. *Journal of American Statistical Association*. 1993; 88:899–905.
- Weerahandi S. ANOVA under unequal error variances. *Biometrics*. 1995; 51:589–599.
- Weerahandi, S. *Exact Statistical Methods for Data Analysis*. Springer; New York: 2003.
- Weerahandi S, Berger VW. Exact inference for growth curves with intraclass correlation structure. *Biometrics*. 1999; 55:921–924. [PubMed: 11315029]
- Wians FH Jr, Urban JE, Keffer JH, Kroft SH. Discriminating between iron deficiency anemia and anemia of chronic disease using traditional indices of iron status vs transferrin receptor concentration. *American Journal of Clinical Hematopathology*. 2001; 115:112–118.
- Wieand S, Gail MH, James BR, James KL. A family of nonparametric statistics for comparing diagnostic markers with paired or unpaired data. *Biometrika*. 1989; 76:585–592.
- Xiong C, van Belle G, Miller J, Morris J. Measuring and estimating diagnostic accuracy when there are three ordinal diagnostic groups. *Statistics in Medicine*. 2006; 25:1251–1273. [PubMed: 16345029]
- Xiong C, van Belle G, Miller JP, Yan Y, Gao F, Feng S, Yu K, Morris JC. A parametric comparison of diagnostic accuracy with three ordinal diagnostic groups. *Biometrical Journal*. 2007; 49(5):682–693. [PubMed: 17763377]
- Zhang DD, Zhou X-H, Freeman DH, Freeman JL. A non-parametric method for the comparison of partial areas under ROC curves and its application to large health care data sets. *Statistics in Medicine*. 2002; 21:701–705. [PubMed: 11870811]
- Zhou, X-H.; Obuchowski, N.; McClish, D. *Statistical Methods in Diagnostic Medicine*. Wiley; New York: 2002.

Table 1

Empirical coverage probabilities (94–96% considered satisfactory) of approximate 95% two-sided confidence bounds for ΔVUS (based on 2000 simulations).

Sample sizes	Config. of $\Sigma_1, \Sigma_2, \Sigma_3^a$								
	1			2			3		
	Proposed ^b	Large ^c	Boot ^d	Proposed	Large	Boot	Proposed	Large	Boot
$\mu_1 = (0,0)', \mu_2 = (2,3)', \mu_3 = (5,4)'$									
(5, 5, 5)	97	96	92	95	96	91	96	95	92
(10, 10, 10)	96	95	93	96	96	93	96	95	94
(10, 10, 5)	96	92	91	94	94	90	96	93	92
(20, 20, 20)	96	95	94	96	95	95	96	94	94
(20, 10, 10)	97	92	93	96	91	92	96	93	93
(30, 20, 10)	96	94	94	94	94	93	95	94	95
(50, 50, 50)	96	95	95	95	96	96	97	95	96
(50, 20, 10)	96	92	94	95	92	93	95	93	92
$\mu_1 = (0,0)', \mu_2 = (0.5,0.3)', \mu_3 = (1,0.6)'$									
(5,5, 5)	95	93	93	95	95	91	95	91	93
(10, 10, 10)	96	95	94	95	96	93	96	95	95
(10, 10, 5)	96	95	91	94	96	90	95	94	93
(20, 20, 20)	96	96	95	94	95	95	96	93	95
(20, 10, 10)	96	87	94	96	90	93	95	89	93
(30, 20, 10)	95	91	93	95	93	92	95	92	93
(50, 50, 50)	96	94	95	95	94	95	96	95	95
(50, 20, 10)	95	87	93	95	87	93	95	88	94

^aConfig. of $\Sigma_1, \Sigma_2, \Sigma_3$: 1: $\begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$; 2: $\begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}, \begin{pmatrix} 4 & 2 \\ 2 & 4 \end{pmatrix}, \begin{pmatrix} 16 & 8 \\ 8 & 16 \end{pmatrix}$; 3: $\begin{pmatrix} 1 & 1 \\ 1 & 4 \end{pmatrix}, \begin{pmatrix} 1 & 1 \\ 1 & 4 \end{pmatrix}, \begin{pmatrix} 1 & 1 \\ 1 & 4 \end{pmatrix}$

^bProposed: The proposed generalized approach.

^cLarge: The large sample approach by Xiong et al. (2007).

^dBoot: The parametric bootstrap approach.

Table 2

Empirical coverage probabilities (94–96% considered satisfactory) of approximate 95% two-sided confidence bounds for DPVUS (based on 2000 simulations).

Sample sizes	Config. of $\Sigma_1, \Sigma_2, \Sigma_3^a$					
	<u>1</u>		<u>2</u>		<u>3</u>	
	Proposed ^b	Boot ^d	Proposed	Boot	Proposed	Boot
$\mu_1 = (0,0)', \mu_2 = (2,3)', \mu_3 = (5,4)'$						
(10, 10, 10)	96	93	94	93	94	93
(20, 20, 20)	95	94	95	94	95	94
(20, 10, 10)	96	93	94	92	94	92
(30, 20, 10)	96	93	94	93	94	93
(50, 50, 50)	95	94	95	94	95	94
(50, 20, 10)	96	93	94	93	94	92
$\mu_1 = (0,0)', \mu_2 = (0.5,0.3)', \mu_3 = (1,0.6)'$						
(10, 10, 10)	96	94	95	93	96	95
(20, 20, 20)	96	95	96	95	96	95
(20, 10, 10)	95	94	95	93	95	94
(30, 20, 10)	94	93	95	93	95	94
(50, 50, 50)	95	95	95	95	95	95
(50, 20, 10)	95	94	95	94	95	94

Note: See bottom of Table 1 for footnotes a, b, and d.

Table 3

Empirical coverage probabilities (94–96% considered satisfactory) of approximate 95% two-sided confidence bounds for ΔAUS and $\Delta PAUS$ with minimum classification rate 0.5 for both non-diseased and diseased groups for mixture of normal data: a robust study (based on 2000 simulations).

Config. ^a of $\Sigma_1, \Sigma_2, \Sigma_3$		Sample sizes				
		(10, 10, 10)	(20, 20, 20)	(50, 50, 50)	(50, 20, 20)	(50, 20, 10)
$\mu_1 = (0,0)', \mu_2 = (2,3)', \mu_3 = (5,4)'$						
1	ΔVUS	96	96	96	96	97
	$\Delta PVUS$	96	95	94	95	96
2	ΔVUS	95	95	96	96	95
	$\Delta PVUS$	93	94	93	95	94
3	ΔVUS	96	97	97	96	95
	$\Delta PVUS$	92	95	94	94	94
$\mu_1 = (0,0)', \mu_2 = (0.5,0.3)', \mu_3 = (1.0,0.6)'$						
1	ΔVUS	96	96	95	96	95
	$\Delta PVUS$	96	95	95	95	95
2	ΔVUS	95	95	96	96	95
	$\Delta PVUS$	95	95	96	96	94
3	ΔVUS	95	96	96	95	95
	$\Delta PVUS$	96	96	96	95	95

Note: See bottom of Table 1 for Configurations of $\Sigma_1, \Sigma_2, \Sigma_3$.

The mixture of bivariate normal is defined as $Y_i = \begin{pmatrix} Y_{iA} \\ Y_{iB} \end{pmatrix} \sim (1 - \alpha) N_2(\mu_i, (1/1.1) \Sigma_i) + \alpha N_2(\mu_i, (2/1.1) \Sigma_i)$ where $\alpha = 0:1$ for $i = 1, 2, 3$. Note that Y_i from such a normal mixture distribution has mean μ_i and variance Σ_i .