Multiple Comparisons with the Best ROC Curve

Jason C. Hsu
Department of Statistics
The Ohio State University
Columbus, OH 43210-1247
hsu.1@osu.edu

Peihua Qiu
School of Statistics
University of Minnesota
Minneapolis, MN 55455
qiu@stat.umn.edu

Lin Yee Hin
Department of Obstetrics & Gynaecology
Prince of Wales Hospital
Hong Kong S.A.R.

Donald O. Mutti
College of Optometry
The Ohio State University
Columbus, OH 43210-1240
mutti.2@osu.edu

Karla Zadnik
College of Optometry
The Ohio State University
Columbus, OH 43210-1240
zadnik.4@osu.edu

January 22, 2004

Abstract

The accuracy of a medical diagnostic tool depends on its *specificity*, the probability that it classifies a normal person as normal, and its *sensitivity*, the probability that it classifies a diseased person as diseased. The receiver operating characteristic (ROC) curve of such a tool is its sensitivity plotted against (1−specificity) as the threshold defining "normal" versus "diseased" ranges over all possible values. A common, global measure of the accuracy of a diagnostic tool is the area under the curve (AUC), the curve being the ROC curve. Thus, one way to compare the accuracies of medical diagnostic tools is to compare their AUCs. By comparing each diagnostic tool with the truly most accurate diagnostic tool, one can eliminate diagnostic tools that are not the most accurate, and discover diagnostic tools which are either the most accurate or practically the most accurate. This article shows how the method of multiple comparison with the best (MCB) for normal error general linear models can be adapted to compare diagnostic tools in terms of AUCs of their ROC curves. MCB of AUCs of ROC curves is illustrated by comparing diagnostic variables for predicting the need for emergency Cesarean section, and for predicting the onset of juvenile myopia.

1 Receiver Operating Characteristic (ROC) Curves

For two well-defined groups, diseased patients and non-diseased ("normal") subjects, let T denote a diagnostic tool measurement. Define a decision rule by t_0, a threshold value of T, such that if $T > t_0$ the person is classified as positive (diseased) and if $T \leq t_0$, the person is classified as negative ("normal"). For a given threshold, define *specificity* as the probability that a normal person is classified as normal (true negative) and *sensitivity* as the probability that a diseased person is classified as diseased (true positive), with some choice of t_0. The theoretical receiver operating characteristic (ROC) curve is the function of sensitivity versus (1−specificity) as the threshold t_0 ranges over all possible values. On the y-axis is sensitivity, or the true-positive fraction. On the x-axis is (1 − specificity), or the false positive fraction.

One convenient global measure of the diagnostic accuracy of a laboratory tool is the area under its ROC curve. The area under the ROC curve measures the probability, denoted by θ, that in a randomly selected pair of normal and diseased individuals the diagnostic tool allows them to be correctly identified. Let X denote the diagnostic tool measurement T for the "normal" subject and Y the measurement for a diseased patient. Then $\theta = P\{X < Y\}$. An area of $\theta = 0.8$, for example, means that a randomly selected individual from the diseased group has a diagnostic tool measurement Y larger than the measurement X for a randomly selected individual from the non-diseased group 80% of the time. Suppose measurements from a diagnostic tool applied to m diseased patients and n non-diseased patients are available. An unbiased estimate of $P\{X < Y\}$ is the area under the curve (AUC) of the empirical ROC plot, which is also the Mann-Whitney version of the two-sample rank-sum statistic of Wilcoxon (cf. Bamber 1975).

2 Multiple Comparisons of ROC Curves

If k (≥ 2) diagnostic tools are to be compared, then a global approach is to compare their θ's, denoted by $\theta_1, \ldots, \theta_k$, considering diagnostic tool i as better than diagnostic tool j if $\theta_i > \theta_j$.

Suppose measurements from k diagnostic tools applied to the same m diseased and n non-diseased patients are available. For $i = 1, \ldots, k$, let $\hat{\theta}_i$ be the Mann-Whitney statistic of the $m+n$ measurements provided by the ith diagnostic tool. Then $\hat{\theta} = (\hat{\theta}_1, \ldots, \hat{\theta}_k)'$ is an unbiased estimate of $\theta = (\theta_1, \ldots, \theta_k)'$ and is asymptotically normally distributed. Its asymptotic variance-covariance matrix can be consistently estimated, by $\mathbf{V} = \{v_{ij}\}$ say. (Exact expressions for $\hat{\theta}_i$ and $\mathbf{V}$ are given in DeLong, DeLong, and Clarke-Pearson, 1988.) One reasonable strategy to compare the $\theta_1, \ldots, \theta_k$ is to derive analogues of multiple comparison methods for comparing normal populations by basing them on $\hat{\theta}_i$ and $\mathbf{V}$. These methods will then be asymptotically valid. Toward this end, let v^i_j denote the estimated variance of $\hat{\theta}_i - \hat{\theta}_j$, i.e., $v^i_j = v_{ii} + v_{jj} - 2v_{ij}$.

DeLong, DeLong, and Clarke-Pearson (1988) derived an analogue of Scheffé's method. Their method provides asymptotically correct simultaneous confidence intervals

$$\sum_{i=1}^{k} c_i \theta_i \in \sum_{i=1}^{k} c_i \hat{\theta}_i \pm \sqrt{(k-1)\chi^2_{\alpha,k-1}}(c'\mathbf{V}c)^{1/2}$$

for all contrast vectors $c = (c_1, \ldots, c_k)'$ such that $c_1 + \cdots + c_k = 0$, where $\chi^2_{\alpha,k-1}$ is the upper α quantile of the χ^2 distribution with $k-1$ degrees of freedom. We will compare Scheffé's confidence intervals for pairwise differences $\theta_i - \theta_j$ with those given by the multiple comparison with the best (MCB) method we propose using two examples.

Tukey's method of pairwise comparisons of normal means is derived by pivoting the Studentized range statistic. McClish (1998) considered an analogue of the Studentized range statistic in the setting of comparing AUC under ROC curves, but only used it to test $H_0 : \theta_1 = \cdots = \theta_k$.

In many situations, not all pairwise comparisons are of interest. If one diagnostic tool is a standard or control diagnostic tool, the kth diagnostic tool say, and of primary interest is which new diagnostic tools are better than this control, then an analogue of one-sided Dunnett's normal means method for multiple comparison with a control (MCC) is

$$\theta_i - \theta_k > \hat{\theta}_i - \hat{\theta}_k - d\sqrt{v_k^i} \text{ for } i = 1, \ldots, k-1,$$

where d is the upper α quantile of the maximum of $k-1$ random variables from a multivariate normal distribution with means zero and correlation matrix

$$\mathbf{R}_{-k} = [diag(\mathbf{V}_{-k})]^{-1}\mathbf{V}_{-k}[diag(\mathbf{V}_{-k})]^{-1} = \{r_{ij}^k\}.$$

Here $\mathbf{C}_{-k}$ is the matrix such that

$$\theta_{-k} = (\theta_j - \theta_k,\ j \neq k)' = \mathbf{C}_{-k}\theta$$

so $\mathbf{V}_{-k} = \mathbf{C}_{-k}\mathbf{V}\mathbf{C}'_{-k}$, and v_k^i denotes the estimated variance of $\hat{\theta}_i - \hat{\theta}_k$. The critical value d can be computed exactly when $\mathbf{R}_{-k}$ has a *one-factor* structure, that is, there exist constants $\lambda_1, \ldots, \lambda_k$ with all $|\lambda_i| < 1$ such that $r_{ij}^k = \lambda_i\lambda_j$ for all $i \neq j$ (using function PROBMC in SAS, for example). Having a one-factor structure means $\hat{\theta}_i - \hat{\theta}_k,\ i = 1, \ldots, k-1$, are conditionally independent, and this conditional independence facilitates critical value computation. The factor-analytic approximation of Hsu (1992) can be used to deterministically approximate d when $\mathbf{R}_{-k}$ does not have a one-factor structure. The idea of the factor-analytic approximation is to use factor analysis algorithms in multivariate analysis to find the correlation matrix $\mathbf{R}_{-k}^{fa}$ with a one-factor structure that most closely approximate the correlation matrix $\mathbf{R}_{-k}$ of $\hat{\theta}_i - \hat{\theta}_k,\ i = 1, \ldots, k-1$, and use the approximate correlation matrix $\mathbf{R}_{-k}^{fa}$ to compute the critical value d. The variance reduction technique of Hsu and Nelson (1998) can be used to efficiently approximate d by simulation. This technique is a *control variate* technique. Random vectors of $\hat{\theta}_i - \hat{\theta}_k,\ i = 1, \ldots, k-1$, are generated with correlation matrix $\mathbf{R}_{-k}$ and correlation matrix $\mathbf{R}_{-k}^{fa}$ using the same random variates (i.e., same seeds). Instead of estimating the quantile of $\max_{i=1,\ldots,k-1}[(\hat{\theta}_i - \hat{\theta}_k)/\sqrt{v_k^i}]$ generated under $\mathbf{R}_{-k}$ directly, one estimates the *difference* between this unknown quantile and the known quantile of $\max_{i=1,\ldots,k-1}[(\hat{\theta}_i - \hat{\theta}_k)/\sqrt{v_k^i}]$ generated under $\mathbf{R}_{-k}^{fa}$, with the statistic $\max_{i=1,\ldots,k-1}[(\hat{\theta}_i - \hat{\theta}_k)/\sqrt{v_k^i}]$ generated under $\mathbf{R}_{-k}^{fa}$ serving as the control variate.

Another situation in which not all pairwise comparisons are of interest is when the comparisons of primary interest are comparisons with the unknown best diagnostic

tool. For example, suppose among five diagnostic tools two are much inferior than the other three. Then it is not of primary interest which of those two diagnostic tools is worse; the inference that neither is best suffices. Suppose the second best diagnostic tool is almost as good as the true best diagnostic tool. Then identifying both as practically the best is useful, for there may be other considerations (e.g. cost and efficiency) impacting on the choice of the diagnostic tool. Multiple comparisons with the best (MCB) of ROC curves compares each diagnostic tool with the best of the other diagnostic tools. MCB has been developed for and applied in linear model settings (e.g., Hsu 1984, Edwards and Hsu 1983, Horrace and Schmidt 1999). In this article, we describe the ideas behind the development and use them to derive an asymptotically valid MCB method for comparing ROC curves. (Thus, it should be understood that the probabilistic statements given in this section are valid only asymptotically, as m and n approach infinity.) For ease of presentation, in describing the ideas behind MCB, it is assumed that there is only one best medical diagnostic tool. But the MCB result stated in Theorem 1 is valid without this assumption.

The idea behind MCB is to ask, with $i = 1, \ldots, k$ in turn, the question

"Is there sufficient evidence that the ith diagnostic tool is *not* the best?" (1)

If the probability of incorrectly answering "yes" to the ith question is controlled at the level α for each i, then the collection of diagnostic tools for which the answer is "no" constitutes a $100(1-\alpha)\%$ confidence set for the best diagnostic tool (because exactly one diagnostic tool is best). Let (k) denote the unknown index of the best diagnostic tool. Each question can be answered by a $100(1-\alpha)\%$ confidence MCC analysis with the ith diagnostic tool as the "control," and this analysis can be *1-sided* because it is impossible for a diagnostic tool to be better than the best. Adjusting for the multiplicity of executing k MCC analyses simultaneously is not necessary, because it is impossible to make more than one mistake in answering the k questions (1). Collating the k MCC analyses, MCB provides simultaneous confidence intervals for

$$\theta_i - \max_{j \neq i} \theta_j = \min_{j \neq i} (\theta_i - \theta_j),\ i = 1, \ldots, k.$$

If

$$\theta_i - \max_{j \neq i} \theta_j > 0,$$

then diagnostic tool i is the best diagnostic tool. On the other hand, if

$$\theta_i - \max_{j \neq i} \theta_j < 0,$$

then diagnostic tool i is not the best diagnostic tool. Further, even if the ith diagnostic tool is not the best, but nevertheless

$$\theta_i - \max_{j \neq i} \theta_j > -\delta$$

where δ is a small positive number, then the ith diagnostic tool is at least close to the best.

For each fixed i, let $\mathbf{C}_{-i}$ be the matrix such that

$$\theta_{-i} = (\theta_j - \theta_i,\ j \neq i)' = \mathbf{C}_{-i}\theta.$$

If $\hat{\theta}_i$ is, for the ith diagnostic tool, the proportion of pairs of measurement (X,Y) so that the measurement X from a normal patient is less than the measurement Y from a diseased patient (among all possible pairs of measurements), then

$$\hat{\theta}_{-i} = (\hat{\theta}_j - \hat{\theta}_i,\ j \neq i)' = \mathbf{C}_{-i}\hat{\theta}$$

is asymptotically normal with mean $\mathbf{C}_{-i}\theta$ and a variance-covariance matrix which is consistently estimated by $\mathbf{V}_{-i} = \mathbf{C}_{-i}\mathbf{V}\mathbf{C}'_{-i}$. Let v^i_j denote the estimated variance of $\hat{\theta}_i - \hat{\theta}_j$, i.e., $v^i_j = v_{ii} + v_{jj} - 2v_{ij}$.

For each $i,\ i = 1,\ldots,k$, suppose the constant d^i is such that

$$P\{\hat{\theta}_i - \hat{\theta}_j - (\theta_i - \theta_j) > -d^i\sqrt{v^i_j} \text{ for all } j, j \neq i\} = 1 - \alpha, \tag{2}$$

i.e., it is the one-sided MCC asymptotic critical value with the ith diagnostic tool as the control. Then for that i,

$$\hat{\theta}_i - \hat{\theta}_j + d^i\sqrt{v^i_j} \text{ for all } j,\ j \neq i$$

form $100(1-\alpha)\%$ simultaneous upper confidence bounds for $\theta_i - \theta_j$ for all $j,\ j \neq i$. As they are simultaneous upper confidence bounds on $\min_{j\neq i}\{\theta_i - \theta_j\}$ as well, $\min_{j\neq i}\{\hat{\theta}_i - \hat{\theta}_j + d^i\sqrt{v^i_j}\}$ is a $100(1-\alpha)\%$ upper confidence bound for $\min_{j\neq i}\{\theta_i - \theta_j\}$. In particular,

$$\min_{j\neq(k)}\{\hat{\theta}_{(k)} - \hat{\theta}_j + d^{(k)}\sqrt{v^{(k)}_j}\}$$

is a $100(1-\alpha)\%$ upper confidence bound for $\min_{j\neq(k)}\{\theta_{(k)} - \theta_j\}$.

The parameter $\min_{j\neq i}\{\theta_i - \theta_j\},\ i = 1,\ldots,k$, is positive when $i = (k)$, negative otherwise. If we use the notation $D^+_i = (\min_{j\neq i}\{\hat{\theta}_i - \hat{\theta}_j + d^i\sqrt{v^i_j}\})^+$, where $x^+ = \max\{x,0\}$, then $D^+_{(k)}$ is a $100(1-\alpha)\%$ upper confidence bound for $\min_{j\neq i}\{\theta_i - \theta_j\}$ with $i = (k)$, while for each $i \neq (k)$, D^+_i is trivially a 100% upper confidence bound for $\min_{j\neq i}\{\theta_i - \theta_j\}$. Therefore, $D^+_i,\ i = 1,\ldots,k$, are simultaneous $100(1-\alpha)\%$ upper confidence bounds for $\min_{j\neq i}\{\theta_i - \theta_j\},\ i = 1,\ldots,k$.

Further, for each i, a size-α test for

$$H_{0i} : \min_{j\neq i}\{\theta_i - \theta_j\} > 0$$

which answers the question (1) is to accept (answer "no") when

$$\min_{j\neq i}\{\hat{\theta}_i - \hat{\theta}_j + d^i\sqrt{v^i_j}\} > 0,$$

or, equivalently, to accept when $D^+_i > 0$. Therefore, $G = \{i : D^+_i > 0\}$ is a $100(1-\alpha)\%$ confidence set for the unknown index (k) of the best medical diagnostic tool.

Finally, $\hat{\theta}_i - \hat{\theta}_{(k)} - d^{(k)}\sqrt{v_i^{(k)}}$ for all i, $i \neq (k)$, form $100(1-\alpha)\%$ simultaneous lower confidence bounds for $\theta_i - \theta_{(k)} = \theta_i - \max_{j \neq i} \theta_j$ for $i \neq (k)$. On the other hand, for $i = (k)$, trivially $\theta_i - \max_{j \neq i} \theta_j > 0$. Therefore, since G is a confidence set for the unknown index (k) of the best diagnostic tool, if we define

$$D_i^- = \begin{cases} 0 \text{ if } G = \{i\} \\ \min_{j \in G, j \neq i} \{\hat{\theta}_i - \hat{\theta}_j - d^j \sqrt{v_i^j}\} \text{ otherwise;} \end{cases}$$

then D_i^-, $i = 1, \ldots, k$, are simultaneous $100(1-\alpha)\%$ lower confidence bounds for $\min_{j \neq i}\{\theta_i - \theta_j\}$, $i = 1, \ldots, k$.

Note that it is the same $100(1-\alpha)\%$ probability event in (2) from which the confidence limits D_i^+, D_i^-, $i = 1, \ldots, k$, are derived. We thus have the following result, which can be proven rigorously along the lines of Theorem 7.3.1 of Hsu (1996).

Theorem 1 For all θ, as $m, n \to \infty$,

$$P_\theta\{\theta_i - \max_{j \neq i} \theta_j \in [D_i^-, D_i^+] \text{ for } i = 1, \ldots, k\} \geq 1 - \alpha.$$

Note that the techniques for computing MCC critical values discussed previously apply to the computation of d^i.

In the examples which follow, we use the factor-analytic approximation of Hsu (1992) to deterministically approximate d^i. In these examples, the diagnostic tools are candidate variables for predicting whether an outcome will occur. Cost of measurement and simplicity considerations make it desirable to use a single diagnostic variable, so the problems are cast as finding the best or almost the best single-variable predictor. In situations where measurement on all variables are readily available, Su and Liu (1993) and Reiser and Faraggi (1997) discuss how to take combinations of the variables to increase accuracy. But how to provide a probabilistic guarantee similar to Theorem 1 that a particular combination is the best or almost the best combination predictor remains a future research problem.

3 Prediction of emergency Cesarean section example

To compare the ability of ultrasound and clinical assessment to predict the need for Cesarean section, 105 Hong Kong Chinese with singleton pregnancies in cephalic presentation were recruited at random from women admitted to the labor ward of the Prince of Wales Hospital during January 1993 (Stock *et al*, 1994). Assessment of all the individuals was performed by the same obstetrician who was not involved in their clinical management. In addition, the attending obstetricians were blinded from the results of the study obstetricians.

The fundal height (FH), was measured, and the clinical estimation of the fetal weight (CLINICAL) was made during clinical examination. The biparietal diameter (BPD), abdominal circumference (AC), and femur length (FL) were measured on fetal ultrasonography. These five variables are the candidate diagnostic variables for predicting the need for emergency Cesarean section.

The point estimates of areas under the ROC curves are

$$\begin{aligned}
\hat{\theta}_{BPD} &= 0.637 \\
\hat{\theta}_{FL} &= 0.854 \\
\hat{\theta}_{AC} &= 0.748 \\
\hat{\theta}_{FH} &= 0.638 \\
\hat{\theta}_{CLINICAL} &= 0.680
\end{aligned}$$

with estimated variance-covariance matrix

$$\mathbf{V} = 10^{-3} \begin{bmatrix} 8.71 & 2.49 & 4.07 & 6.01 & 5.78 \\ 2.49 & 2.79 & 1.67 & 2.75 & 3.84 \\ 4.07 & 1.67 & 5.31 & 4.16 & 3.50 \\ 6.01 & 2.75 & 4.16 & 7.72 & 6.36 \\ 5.78 & 3.84 & 3.50 & 6.36 & 8.57 \end{bmatrix}$$

To execute MCB, one first calculates the upper confidence bounds $D_1^+, \ldots, D_5^+$ using the critical values $(d^1, \ldots, d^5)$, which at the 95% level are $(2.133, 2.160, 2.134, 2.189, 2.164)$ based on the factor-analytic approximation to the correlation matrices derived from $\mathbf{V}_{-1}, \ldots, \mathbf{V}_{-5}$. The upper confidence bounds turn out to be $0, 0.255, 0.041, 0, 0$. Therefore, the confidence set G for the index of the unknown best diagnostic variable is $\{2,3\}$. In fact, if one tests for $i = 1, \ldots, 5$ the null hypotheses

$$H_{0i} : \theta_i > \theta_j \text{ for all } j \neq i,$$

then the p-values are $0.012, 0.997, 0.163, 0.004, 0.007$. One then calculates the lower confidence bounds $D_i^-, i = 1, \ldots, 5$, which for each i is

$$D_i^- = \min_{j \in \{2,3\}, j \neq i} \{\hat{\theta}_i - \hat{\theta}_j - d^j \sqrt{v_j^i}\}$$

in this case. They turn out to be $-0.391, -0.041, -0.255, -0.369, -0.305$. Therefore, at the 95% confidence level, the MCB confidence intervals for each of FH, CLINICAL, BPD, AC, FL minus the best of the other diagnostic variables are:

$$\begin{aligned}
\theta_{BPD} - \max\{\theta_{FH}, \theta_{CLINICAL}, \theta_{AC}, \theta_{FL}\} &\in [-0.391, 0] \\
\theta_{FL} - \max\{\theta_{FH}, \theta_{CLINICAL}, \theta_{BPD}, \theta_{AC}\} &\in [-0.041, 0.255] \\
\theta_{AC} - \max\{\theta_{FH}, \theta_{CLINICAL}, \theta_{BPD}, \theta_{FL}\} &\in [-0.255, 0.041] \\
\theta_{FH} - \max\{\theta_{CLINICAL}, \theta_{BPD}, , \theta_{AC}, \theta_{FL}\} &\in [-0.369, 0] \\
\theta_{CLINICAL} - \max\{\theta_{FH}, \theta_{BPD}, \theta_{AC}, \theta_{FL}\} &\in [-0.305, 0]
\end{aligned}$$

So, at the 95% confidence level, one can say BPD, FH, and CLINICAL are not the best diagnostic variables. AC is within 0.255 of the best, while FL is within 0.041 of the best. Note that for AC and FL, their confidence intervals for $\theta_i - \max_{j \neq i} \theta_j$ are reflections of each other with respect to zero. That is because AC and FL are the only two diagnostic variables that can be best, so that the difference between each and the best of the others is the difference between each other.

For this emergency Cesarean section data, Scheffé's method is not able to declare any two diagnostic variables to be different at the 95% confidence level, in contrast to the MCB analysis. Simultaneous 95% confidence intervals for each of FH, CLINICAL, BPD, AC, FL minus the best of the other diagnostic variables that can be deduced from Scheffé's method are:

$$\begin{aligned}
\theta_{BPD} - \max\{\theta_{FH}, \theta_{CLINICAL}, \theta_{AC}, \theta_{FL}\} &\in (-0.465, 0.033) \\
\theta_{FL} - \max\{\theta_{FH}, \theta_{CLINICAL}, \theta_{BPD}, \theta_{AC}\} &\in (-0.107, 0.319) \\
\theta_{AC} - \max\{\theta_{FH}, \theta_{CLINICAL}, \theta_{BPD}, \theta_{FL}\} &\in (-0.319, 0.107) \\
\theta_{FH} - \max\{\theta_{CLINICAL}, \theta_{BPD}, , \theta_{AC}, \theta_{FL}\} &\in (-0.435, 0.003) \\
\theta_{CLINICAL} - \max\{\theta_{FH}, \theta_{BPD}, \theta_{AC}, \theta_{FL}\} &\in (-0.361, 0.013)
\end{aligned}$$

Further, if a diagnostic variable with an AUC within 0.05 (say) of the best can be considered practically the best, then MCB infers FL is practically the best diagnostic variable but Scheffé's method fails to do so.

4 Prediction of myopia example

In optometry, accurate prediction of myopia onset and identification of children at high risk for myopia onset is important in eventually preventing and controlling abnormal myopic eye growth. During the period from 1989 to 1993, measurements on four candidate predictor variables were taken from 554 children in the Orinda Longitudinal Study of Myopia (Zadnik *et al*, 1993) who were not myopic by the occasion of their third grade visit. Through 1994, 45 of them had developed myopia. The candidate predictor variables are: mean cycloplegic sphere power of the refractive error (WSMEAN), corneal power in the vertical meridian from the third ring of the photokeratoscope photograph (CS3V), Gullstrand crystalline lens power (GLP), and axial length (AL.MN).

The point estimates for θ are:

$$\begin{aligned}
\hat{\theta}_{WSMEAN} &= 0.875 \\
\hat{\theta}_{GLP} &= 0.605 \\
\hat{\theta}_{AL.MN} &= 0.614 \\
\hat{\theta}_{CS3V} &= 0.608
\end{aligned}$$

with estimated variance-covariance matrix

$$\mathbf{V} = 10^{-3} \begin{bmatrix} 0.785 & -0.167 & 0.320 & -0.084 \\ -0.167 & 1.780 & 0.836 & -0.384 \\ 0.320 & 0.836 & 2.043 & -1.324 \\ -0.084 & -0.384 & -1.324 & 1.818 \end{bmatrix}$$

The question is whether the superiority of WSMEAN reflected in it having the largest $\hat{\theta}$ and the dominant ROC curve in Figure 1 can be ruled out as being due to chance.

To execute MCB, one first calculates the upper confidence bounds $D_1^+, \ldots, D_4^+$ using the critical values $(d^1, \ldots, d^4)$, which at the 99% level are $(2.699, 2.572, 2.566, 2.542)$

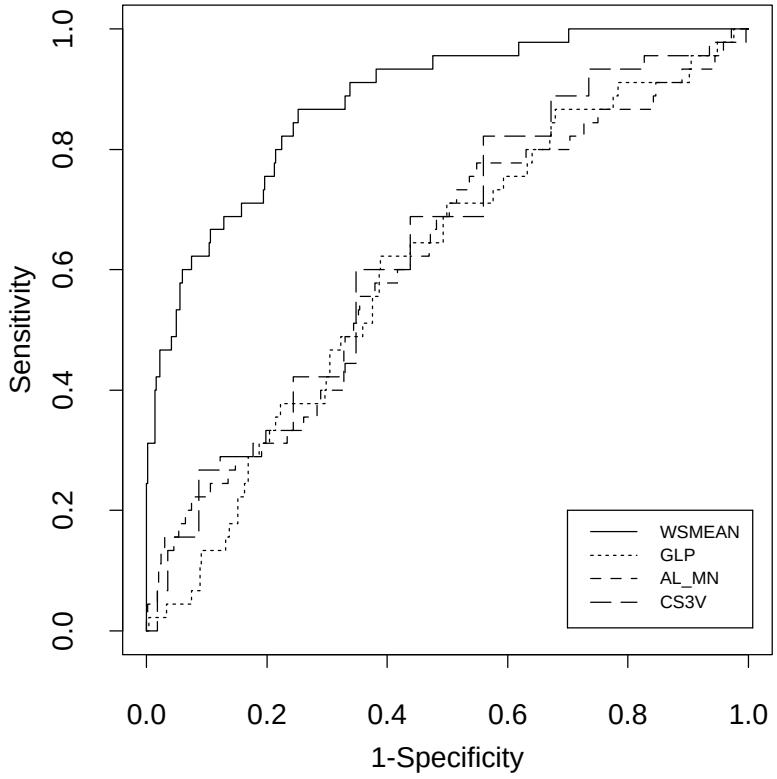

Figure 1: ROC curves of predictors of myopia

based on the factor-analytic approximation to the correlation matrices derived from $\mathbf{V}_{-1}, \ldots, \mathbf{V}_{-4}$. The upper confidence bounds turn out to be $0.387, 0, 0, 0$. Therefore, at the 99% confidence level, we can infer WSMEAN is the best diagnostic variable. In fact, if one tests for $i = 1, \ldots, 4$ the null hypotheses

$$H_{0i} : \theta_i > \theta_j \text{ for all } j \neq i,$$

then the p-values are $1, 0, 0, 0$ to at least the fourth decimal place. One then calculates the lower confidence bounds $D_i^-, i = 1, \ldots, 4$, which for each i is

$$D_i^- = \begin{cases} 0 \text{ for } i = 1 \\ \hat{\theta}_i - \hat{\theta}_1 - d^1 \sqrt{v_1^i} \text{ for } i = 2, 3, 4 \end{cases}$$

in this case. They turn out to be $0, -0.415, -0.387, -0.409$. Therefore, at the 99% confidence level, the MCB confidence intervals for each of WSMEAN, GLP, AL.MN, CS3V minus the best of the other diagnostic variables are:

$$\begin{aligned} \theta_{WSMEAN} - \max\{\theta_{GLP}, \theta_{AL.MN}, \theta_{CS3V}\} &\in [0, 0.387] \\ \theta_{GLP} - \theta_{WSMEAN} &\in [-0.415, 0] \\ \theta_{AL.MN} - \theta_{WSMEAN} &\in [-0.387, 0] \\ \theta_{CS3V} - \theta_{WSMEAN} &\in [-0.409, 0] \end{aligned}$$

For this optometry data set, Scheffé's 99% confidence intervals for pairwise differences are as follow:

$$\begin{aligned} \theta_{WSMEAN} - \max\{\theta_{GLP}, \theta_{AL.MN}, \theta_{CS3V}\} &\in (0.118, 0.391) \\ \theta_{GLP} - \theta_{WSMEAN} &\in (-0.420, -0.118) \\ \theta_{AL.MN} - \theta_{WSMEAN} &\in (-0.391, -0.129) \\ \theta_{CS3V} - \theta_{WSMEAN} &\in (-0.414, -0.119) \end{aligned}$$

Comparing the inferences given by MCB and Scheffé's method, both infer WSMEAN to be the best diagnostic variable. But for this data set Scheffé's method has the advantage that it gives, in addition, lower bounds on *how much* WSMEAN is better than the other diagnostic variables.

5 Adequacy of the Normal Approximation

A small simulation study was conducted to assess the adequacy of the normal approximation. Applying the MCB method of comparing ROC curves to simulated data with known θ's repeatedly, we observe the true simultaneous coverage probability of MCB confidence intervals based on normal approximation to be slightly lower than the nominal confidence level $1 - \alpha$, as described below.

For $i = 1, \ldots, k$, let X_i denote the test score for a "normal" patient and Y_i the test score for a diseased patient. Independent random samples of $X = (X_1, \ldots, X_k)$ and

$Y = (Y_1, \ldots, Y_k)$ with $k = 4$ of sizes $m = n = 50$ were generated from the multivariate normal distribution with $E(X) = (0,0,0,0)$, $E(Y) = \mu$ and all variance equal to one and all covariance equal 0.5 for various μ. With $\mu = (0.5, 1.0, 1.5, 2.0)$ so that $\theta = (0.638, 0.760, 0.856, 0.921)$ for example, the estimated true simultaneous coverage probabilities of 95% and 90% MCB confi dence intervals based on normal approximation are 92.6% and 86.9% respectively (each based on 10,000 simulations runs). The reason for the under-coverage appears to be as follows.

To compare the *i*th and *j*th diagnostic techniques based on the asymptotic normality of $\hat{\theta}_i$, $\hat{\theta}_j$, one refers

$$Z = \frac{\hat{\theta}_i - \hat{\theta}_j - (\theta_i - \theta_j)}{\hat{\sigma}_{\hat{\theta}_i - \hat{\theta}_j}}$$

to the standard normal distribution or a t distribution for all θ. Figure 2 shows the densities of two such Z's (estimated using kernel smoothing), one for $\theta = (0.76, 0.92)$ (labeled MCC parameter > 0 in the fi gure) and the other for $\theta = (0.92, 0.76)$ (labeled MCC parameter < 0 in the fi gure). Note the longer tail of the latter density. The reason for this is there is a correlation between $\hat{\theta}_i - \hat{\theta}_j$ and $\hat{\sigma}_{\hat{\theta}_i - \hat{\theta}_j}$, which depends on θ. So Z is not as pivotal as one would hope for. How to improve the coverage probability of the MCB method remains a research problem.

References

Bamber, D. (1975). The area above the ordinary dominance graph and the area below the receiver operating characteristic curve. *Journal of Mathematical Psychology*, 12:387–415.

Campbell, G. (1994). General methodology I: Advances in statistical methodology for the evaluation of diagnostic and laboratory tests. *Statistics in Medicine*, 13:499–508.

DeLong, E. R., DeLong, D. M., and Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics*, 44:837–845.

Edwards, D. G. and Hsu, J. C. (1983). Multiple comparisons with the best treatment. *Journal of the American Statistical Association*, 78:965–971.

Hanley, J. A. and McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143:29–36.

Hanley, J. A. and McNeil, B. J. (1983). A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology*, 148:839–843.

Horrace, W. C. and Schmidt, P. (1999). Multiple comparisons with the best, with economic applications. *Journal of Applied Economics*, to appear.

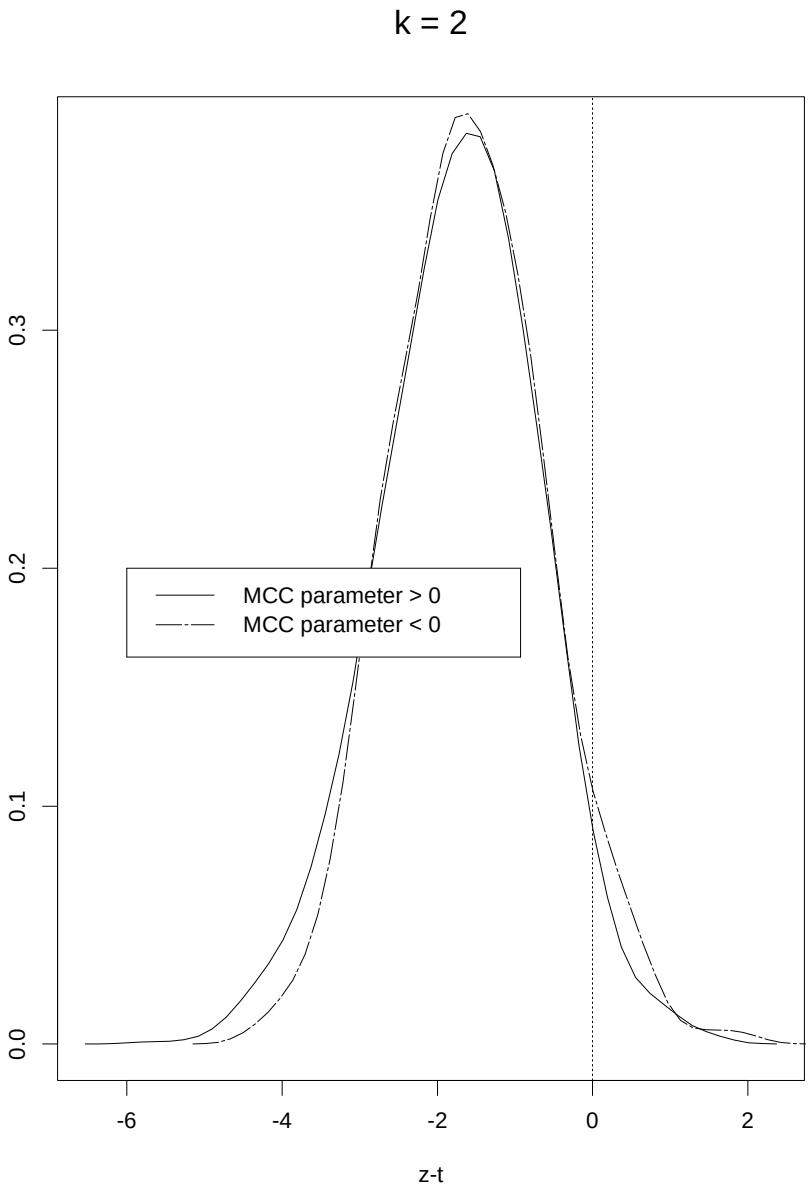

Figure 2: Densities of $Z - t_{.05,\nu}$

Hsu, J. C. (1984). Constrained two-sided simultaneous confi dence intervals for multiple comparisons with the best. *Annals of Statistics*, 12:1136–1144.

Hsu, J. C. (1992). The factor analytic approach to simultaneous inference in the general linear model. *Journal of Graphical and Computational Statistics*, 1:151–168.

Hsu, J. C. (1996). *Multiple Comparisons: Theory and Methods*. Chapman & Hall, London.

Hsu, J. C. and Nelson, B. L. (1998). Multiple comparisons in the general linear model. *Journal of Computational and Graphical Statistics*, 7:23–41.

McClish, D. K. (1987). Comparing the areas under more than two independent roc curves. *Medical Decision Making*, 7:149–155.

Reiser, B. and Faraggi, D. (1997). Confi dence intervals for the generalized Roc criterion. *Biometrics*, 53:644–652.

Stock, A., Wong, W., Rogers, M., and Chang, A. (1994). Prediction of caesarean section from ultrasound and clinical assessment of fetal size. *Australian and New Zealand Journal of Obstetrics and Gynaecology*, 34:393–398.

Su, J. Q. and Liu, J. S. (1993). Linear combinations of multiple diagnostic markers. *Journal of the American Statistical Association*, 88:1350–1355.

Zadnik, K., Mutti, D. O., Friedman, N. E., and Adams, A. J. (1993). Initial cross-sectional results from the Orinda Longitudinal Study of Myopia. *Optometry and Vision Science*, 70:750–758.

Zweig, M. H. and Campbell, G. (1993). Receiver-operating characteristic (ROC) plots: A fundamental evaluation too in clinical medicine. *Clinical Chemistry*, 39:561–577.