Statistics and Its Interface Volume 10 (2017) 229–238

Variable selection in ROC curve analysis with focused information criteria

Baoying Yang*,†, Xin Huang, and Gengsheng Qin

In *Receiver Operating Characteristic* (ROC) curve analysis, many factors such as the study subject's characteristics or operating conditions of a medical test may affect the diagnostic accuracy of the test. ROC regression models are introduced to accommodate effects of the covariates. If many covariates are available, variable selection problem arises. The area under the ROC curve (AUC) is a popular one-number summary index of the discriminatory accuracy of a medical test. In this paper, we propose a variable selection method based on the Focused Information Criteria (FIC) with focus on the AUC index. In particular, the FIC is developed in a placement-value model for ROC regression. The proposed method is illustrated through simulation studies and a real data example.

Keywords and phrases: AUC, Diagnostic test, Placement value model, Variable selection, ROC.

1. INTRODUCTION

Diagnostic tests have been developed in medical studies, and it is important to assess the diagnostic accuracy of the tests (Swets and Pickett, 1982). Suppose that a diagnostic test produces a continuous measurement Y for an individual. For a given threshold value c, the individual will be classified into the diseased group if $Y > c$, otherwise he/she will be classified into the non-diseased group. The *Receiver Operating Characteristic* (ROC) curve of the diagnostic test is the plot of sensitivity versus one minus specificity for all possible threshold values. Zhou *et al.* (2002) and Pepe (2003) demonstrated that the ROC curve is a fundamental tool for the evaluation of the diagnostic accuracy of a medical test.

When covariates such as characteristics (age, gender, general health status, etc.) of study subjects or operating conditions of a diagnostic test are available, various ROC regression models have been introduced into ROC analysis for the evaluation or control of the possible effects of covariates. For example, Faraggi and Reiser (2002) modeled the test results with diseased status and covariates by linear regression models. Pepe and Cai (2004) directly modeled the ROC curve on covariates based on a placement-value model. When many covariates are available, variable selection becomes an important issue in ROC analysis.

In statistical analysis, a variety of variable selection methods have been proposed, which aim at selecting a best-subset of variables associated with some criteria, and focus on explaining phenomena under investigation via the average prediction quality, regardless of purposes of the selection. The AIC (Akaike, 1973), BIC (Schwarz, 1978), and DIC (Spiegelhalter *et al.*, 2002), to name just a few, are examples of such methods, with various contexts and applications. However, in ROC analysis, we are interested in the discriminatory accuracy of the test to distinguish diseased subjects from non-diseased subjects instead of predicting the test results of a diagnostic test conducted on new subjects. Since the area under the ROC curve (AUC), expressed as $AUC = P(Y^D > Y^{\bar{D}})$ with $Y^{\bar{D}}$ and Y^D being the respective test results in the non-diseased group and the diseased group, is the most popular summary index of the discriminatory accuracy of a test, it is necessary to develop new covariates selection procedures with focus on the AUC in ROC analysis.

Claeskens and Hjort (2003) proposed the Focused Information Criteria (FIC) with a totally different point of view that the model selector should instead focus on the accuracy of estimation for interest parameters. By this criterion, one needs to estimate the mean squared error (MSE) of the focus estimator and select the candidate model under which the MSE is minimized. Hjort and Claeskens (2006) developed the FIC for the Cox hazard regression model and applied it to a study of skin cancer. More recently, Wang *et al.* (2011) reported that in many clinical settings, a commonly encountered problem is to assess accuracy of a screening test for early detection of a disease. Following their idea, we believe that the FIC can also be used for variable selection in designing a medical test. An example is a research study conducted to design a new screening test by selecting variables from an existing screener with a hierarchical structure among variables.

Another motivation example is a medical study to be to discussed in detail in Section 5 on an experimental hearing device developed to diagnose hearing impairment. In the latter study, the test result, called distortion product otoacoustic emission (DPOAE), was used to measure the strength of

*Corresponding author.
†The research of Baoying Yang was supported by National Natural Science Foundation of China (NNSFC, No. 11501472) and the Soft Science Research Program in Sichuan Province of China (No. 2015ZR0211).

the cochlear response from two sounds emitted into a single ear at different combinations of frequencies and intensities (Stover, Gorga, and Neely 1996; Pepe 2003; Dodd and Pepe 2003), and the audiometric threshold was used to assess the severity of hearing impairment subjects. The potential influence factors/covariates for the DPOAE were: frequency level, intensity level, the hearing threshold level and their interaction terms. Evaluating the effects of these covariates and selecting the most important covariates with focus on the discrimination capacity could improve the diagnostic accuracy of the hearing impairment. Motivated by this real example and the need of the development of new variable selection methods in medical and biological applications, we will propose variable selection methods with focus on AUC based on the FIC criteria in this paper.

The rest of this paper is organized as follows. In Section 2, we introduce the FIC with focus on the AUC. In Section 3, we develop the FIC with focus on the AUC based on a placement value model. We examine the performance of the FIC through simulation studies in Section 4 and through an application to the audiology data in Section 5. We conclude the paper with some discussion in Section 6 and relegate the technical proofs to the Appendix.

2. FIC WITH FOCUS ON AUC

In this section, we will propose the FIC with focus on AUC. We first introduce general model assumptions for test results. Assume that there are n_1 diseased subjects with test results Y_i^D and covariate vectors $\mathbf{Z}_i^D = (Z_{i1}^D, \cdots, Z_{iq_1}^D)^t$ $(i = 1, 2, \cdots, n_1)$ from the diseased population with density function $f_1(y^D|\boldsymbol{\xi}_1, \mathbf{Z}^D)$, and n_2 non-diseased subjects with test results $Y_j^{\bar{D}}$ and covariate vectors $\mathbf{Z}_j^{\bar{D}} = (Z_{j1}^{\bar{D}}, \cdots, Z_{jq_2}^{\bar{D}})^t$ $(j = 1, 2, \cdots, n_2)$ from the non-diseased population with density function $f_2(y^{\bar{D}}|\boldsymbol{\xi}_2, \mathbf{Z}^{\bar{D}})$, where $\boldsymbol{\xi}_k = (\boldsymbol{\theta}_k, \boldsymbol{\eta}_k), k = 1, 2$, are the parameters in the diseased and non-diseased models respectively, $\boldsymbol{\eta}_k = (\eta_{k,1}, \cdots, \eta_{k,q_k})^t$ are parameters associated with the covariates, and $\boldsymbol{\theta}_k$ includes the parameters appearing in all the candidate models. For example, Faraggi and Reiser (2002) considered the following linear regression models for test results:

$$\begin{aligned} Y_i^D &= \eta_{1,0} X_i^D + \boldsymbol{\eta}_1^t \mathbf{Z}_i^D + \varepsilon_{1i}, \quad i = 1, \cdots, n_1, \\ (2.1)\ Y_j^{\bar{D}} &= \eta_{2,0} X_j^{\bar{D}} + \boldsymbol{\eta}_2^t \mathbf{Z}_j^{\bar{D}} + \varepsilon_{2j}, \quad j = 1, \cdots, n_2, \end{aligned}$$

where ε_{ki}'s follow $N(0, \sigma_k^2)$. X^D and $X^{\bar{D}}$ can be the intercept term or the variable which always included in the model. The parameters $\eta_{k,0}$ and σ_k are always included in the models and can be denoted as $\boldsymbol{\theta}_k = (\eta_{k,0}, \sigma_k), k = 1, 2$, and the parameters $\boldsymbol{\eta}_k = (\eta_{k,1}, \cdots, \eta_{k,q_k})$ for $k = 1, 2$ are the regression coefficients associated with the covariates.

In practice, more information on covariates is usually obtained for diseased subjects (i.e., $q_1 \geq q_2$). When $q_1 > q_2$, we can set $\mathbf{Z}_j^{\bar{D}} = (Z_{j1}^{\bar{D}}, \cdots, Z_{jq_2}^{\bar{D}}, 0, \cdots, 0)^t$ where the number of 0's equals $q_1 - q_2$. So, without loss of generality, we assume that $q_1 = q_2 = q$ in the rest of the paper. Our goal is to select most important covariates using the FIC with focus on AUC in the assumed models for the test results. Similar to Claeskens and Hjort (2003), we assume that $\boldsymbol{\xi}_k^0 = (\boldsymbol{\theta}_k^0, \boldsymbol{\eta}_k^0)$ are fixed parametric vectors of a null model for the test results, and $\boldsymbol{\xi}_k = (\boldsymbol{\theta}_k, \boldsymbol{\eta}_k) = (\boldsymbol{\theta}_k^0, \boldsymbol{\eta}_k^0 + \boldsymbol{\delta}_k/\sqrt{n_k})$ with $\boldsymbol{\delta}_k = O(1)$ are unknown parametric vectors of the true models which are in a local neighborhood of the null model. Obviously, the null model is the true model with $\boldsymbol{\delta}_k = 0$, and the true model and the null model are close to each other when sample size is big because $\boldsymbol{\eta}_k$ is not far from $\boldsymbol{\eta}_k^0$ with the departure $\boldsymbol{\delta}_k/\sqrt{n_k} = O(1/\sqrt{n_k})$. The local model used here is a natural extension of the null model.

There are many candidate models in the model selection for test results. One is called the full model which includes all the available covariates. Another is called the narrow model which is a special case of the full model in which $\boldsymbol{\eta}_k = (\eta_{k,1}, \cdots, \eta_{k,q})$ with some known $\boldsymbol{\eta}_{k,i}$'s. Let S be an index set indicating which covariate variables are selected in the model. The full model is indexed by $S = \{1, 2, \cdots, q\}$. If S is a subset of $\{1, 2, \cdots, q\}$, then the model indexed by S is a sub-model representing a candidate model between the full model and the narrow model excluding all the covariates. Let π_S be the projection matrix mapping a vector $(a_1, \cdots, a_q)^t$ to vector $(a_l, l \in S)^t$. Then $\boldsymbol{\eta}_{k,S} \equiv \pi_S \boldsymbol{\eta}_k = \pi_S \boldsymbol{\eta}_k^0 + \pi_S \boldsymbol{\delta}_k/\sqrt{n_k}$, and $(\boldsymbol{\theta}_k, \boldsymbol{\eta}_k)$, $(\boldsymbol{\theta}_k, \boldsymbol{\eta}_{k,S})$ are the parameter vectors of the full model and sub-model respectively.

As mentioned in Section 1, we consider the FIC with focus on the AUC. Note that the covariate-specific AUC at a covariate vector $\mathbf{Z}_0 = (Z_{0,1}, \cdots, Z_{0,q})$ can be expressed as

$$\begin{aligned} AUC(\mathbf{Z}_0) &= P(Y^D > Y^{\bar{D}}|\mathbf{Z}_0) \\ &= g(\boldsymbol{\theta}_1^0, \boldsymbol{\eta}_1^0 + \boldsymbol{\delta}_1/\sqrt{n_1}, \boldsymbol{\theta}_2^0, \boldsymbol{\eta}_2^0 + \boldsymbol{\delta}_2/\sqrt{n_2}|\mathbf{Z}_0) \\ &\equiv g(\mathbf{Z}_0), \end{aligned}$$

where $g(\cdot)$ is a non-negative function with values in (0,1). Since the parameters in the models for the test results are unknown, this covariate-specific AUC is still unknown but can be estimated. Let $(\widehat{\boldsymbol{\theta}}_k, \widehat{\boldsymbol{\eta}}_{k,S})$ be estimators of $(\boldsymbol{\theta}_k, \boldsymbol{\eta}_{k,S})$ under a candidate (sub-model) model S for the test results. Then the covariate-specific AUC under this model can be estimated by $\widehat{g}_S(\mathbf{Z}_0) = g(\widehat{\boldsymbol{\theta}}_1, \widehat{\boldsymbol{\eta}}_{1,S}, \widehat{\boldsymbol{\theta}}_2, \widehat{\boldsymbol{\eta}}_{2,S}|\mathbf{Z}_0)$. Targeting on covariates selection with focus on AUC, we define the value of the FIC under model S at a covariate vector $\mathbf{Z}_0$ as an estimate of the asymptotic mean square error of $\sqrt{n_1 + n_2}(\widehat{g}_S(\mathbf{Z}_0) - g(\mathbf{Z}_0))$. We will select the model with the smallest FIC value among all the possible candidate models.

To state our main theorem, we first introduce some notation. Assume that the Fisher information matrix of the full model, evaluated at the null point $(\boldsymbol{\theta}_k^0, \boldsymbol{\eta}_k^0)$, is non-singular. The log-likelihood function based on test results from the diseased population can be written as

$$l_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} \log f_1(Y_i^D|\boldsymbol{\theta}_1, \boldsymbol{\eta}_1, \mathbf{Z}^D).$$

Then the information matrix evaluated at the null point is

$$J^D_{\text{full}} = \lim_{n_1\to\infty} Var\begin{pmatrix} \sqrt{n_1}\frac{\partial l_1}{\partial \boldsymbol{\theta}_1}(\boldsymbol{\theta}_1^0,\boldsymbol{\eta}_1^0) \\ \sqrt{n_1}\frac{\partial l_1}{\partial \boldsymbol{\eta}_1}(\boldsymbol{\theta}_1^0,\boldsymbol{\eta}_1^0)\end{pmatrix} \equiv \begin{pmatrix} J^D_{00} & J^D_{01} \\ J^D_{10} & J^D_{11}\end{pmatrix},$$

and the inverse matrix of J^D_{full} can be written as

$$(J^D_{\text{full}})^{-1} = \begin{pmatrix} J^{00,D} & J^{01,D} \\ J^{10,D} & J^{11,D}\end{pmatrix},$$

where the sub-matrix $J^{11,D}$ of $(J^D_{\text{full}})^{-1}$ can be expressed as $J^{11,D} = J^D_{11} - (J^D_{00})^{-1}J^D_{01} \equiv K^D$.

Similarly, we can find the Fisher information matrix evaluated at the null point for sub-model S:

$$J^D_S = \begin{pmatrix} J^D_{00} & J^D_{01}\pi_S^t \\ \pi_S J^D_{10} & \pi_S J^D_{11}\pi_S^t\end{pmatrix},$$

and its inverse matrix:

$$(J^D_S)^{-1} = \begin{pmatrix} J^{00,D}_S & J^{01,D}_S \\ J^{10,D}_S & J^{11,D}_S\end{pmatrix},$$

where $J^{11,D}_S = (\pi_S(K^D)^{-1}\pi_S^t)^{-1} \equiv K^D_S$.

The following lemma, adopted from Claeskens and Hjort (2003), provides the asymptotic distribution of the MLEs for the parameters in sub-model S.

Lemma. *Under the ordinary regularity conditions given in Claeskens and Hjort (2003), we have that*

$$\sqrt{n_1}\begin{pmatrix} \frac{\partial l_1}{\partial \boldsymbol{\theta}_1}(\boldsymbol{\theta}_1^0,\boldsymbol{\eta}_1^0) \\ \frac{\partial l_1}{\partial \boldsymbol{\eta}_1}(\boldsymbol{\theta}_1^0,\boldsymbol{\eta}_1^0)\end{pmatrix} \xrightarrow{d} \begin{pmatrix} J^D_{01} \\ J^D_{11}\end{pmatrix}\boldsymbol{\delta}_1 + \begin{pmatrix} M^D \\ N^D\end{pmatrix},$$

and

$$\begin{aligned}
&\sqrt{n_1}\begin{pmatrix} \widehat{\boldsymbol{\theta}}_{1,S} - \boldsymbol{\theta}^0_{1,S} \\ \widehat{\boldsymbol{\eta}}_{1,S} - \boldsymbol{\eta}^0_{1,S}\end{pmatrix} \xrightarrow{d} \begin{pmatrix} A^D_S \\ B^D_S\end{pmatrix} \\
\equiv\ & (J^D_S)^{-1}\begin{pmatrix} J^D_{01}\boldsymbol{\delta}_1 + M^D \\ \pi_S J^D_{11}\boldsymbol{\delta}_1 + N^D_S\end{pmatrix} \\
\sim\ & N_{p+q}\left((J^D_S)^{-1}\begin{pmatrix} J^D_{01} \\ \pi_S J^D_{11}\end{pmatrix}\boldsymbol{\delta}_1, (J^D_S)^{-1}\right),
\end{aligned}$$

where $\begin{pmatrix} M^D \\ N^D\end{pmatrix} \sim N_{p+q}(0, J^D_{full})$, $N^D_S = \pi_S N^D$.

Let $H^D_S = (K^D)^{-\frac{1}{2}}\pi_S^t K^D_S \pi_S (K^D)^{-\frac{1}{2}}$, and $\boldsymbol{\omega}^D = J^D_{10}(J^D_{00})^{-1}\frac{\partial g}{\partial \boldsymbol{\theta}_1} - \frac{\partial g}{\partial \boldsymbol{\eta}_1}$, where the partial derivatives are evaluated at null point $(\boldsymbol{\theta}^0_1, \boldsymbol{\eta}^0_1)$. Here and hereafter, we can define similar notations for the non-diseased sample with the subscript D replaced by $\bar{D}$. From the above lemma, we can derive the approximate mean squared error of the covariate-specific AUC estimator $\widehat{g}_S$.

Theorem 1. *If sub-model S is fitted, and $n_2/n_1 \to \rho$ $(0 < \rho < \infty)$, then $\sqrt{n_1+n_2}(\widehat{g}_S(\mathbf{Z}_0) - g(\mathbf{Z}_0))$ converges in distribution to*

$$\begin{aligned}
\boldsymbol{\Lambda}_S \ =\ & c_1(\frac{\partial g}{\partial \boldsymbol{\theta}_1})^t (J^D_{00})^{-1}M^D + c_2(\frac{\partial g}{\partial \boldsymbol{\theta}_2})^t (J^{\bar{D}}_{00})^{-1}M^{\bar{D}} \\
& + c_1(\boldsymbol{\omega}^D)^t\left[\boldsymbol{\delta}_1 - (K^D)^{\frac{1}{2}}H^D_S(K^D)^{-\frac{1}{2}}(\boldsymbol{\delta}_1 + W^D)\right] \\
& + c_2(\boldsymbol{\omega}^{\bar{D}})^t\left[\boldsymbol{\delta}_2 - (K^{\bar{D}})^{\frac{1}{2}}H^{\bar{D}}_S(K^{\bar{D}})^{-\frac{1}{2}}(\boldsymbol{\delta}_2 + W^{\bar{D}})\right],
\end{aligned}$$

where $c_1 = \sqrt{1+\rho}$, $c_2 = \sqrt{1+\rho^{-1}}$, $M^D \sim N_p(0, J^D_{00})$, $W^D \sim N_q(0, K^D)$ with M^D and W^D being independent, $M^{\bar{D}} \sim N_p(0, J^{\bar{D}}_{00})$, $W^{\bar{D}} \sim N_q(0, K^{\bar{D}})$ with $M^{\bar{D}}$ and $W^{\bar{D}}$ being independent, and the partial derivatives are evaluated at null point $(\boldsymbol{\theta}^0_1, \boldsymbol{\eta}^0_1)$.

The proof of Theorem 1 is regelated to the Appendix. From Theorem 1, it can be verified that

$$\begin{aligned}
E(\boldsymbol{\Lambda}_S) \ =\ & c_1(\omega^D)^t\left[I - (K^D)^{\frac{1}{2}}H^D_S(K^D)^{-\frac{1}{2}}\right]\boldsymbol{\delta}_1 \\
& + c_2(\omega^{\bar{D}})^t\left[I - (K^{\bar{D}})^{\frac{1}{2}}H^{\bar{D}}_S(K^{\bar{D}})^{-\frac{1}{2}}\right]\boldsymbol{\delta}_2, \\
Var(\boldsymbol{\Lambda}_S) \ =\ & \tau_0^2 + c_1^2(\omega^D)^t(K^D)^{\frac{1}{2}}H^D_S(K^D)^{\frac{1}{2}}\omega^D \\
& + c_2^2(\omega^{\bar{D}})^t(K^{\bar{D}})^{\frac{1}{2}}H^{\bar{D}}_S(K^{\bar{D}})^{\frac{1}{2}}\omega^{\bar{D}},
\end{aligned}$$

and the MSE of $\widehat{g}_S(\mathbf{Z}_0)$ is

$$\begin{aligned}
\gamma(S) \ =\ & \tau_0^2 + c_1^2(\boldsymbol{\omega}^D)^t\left(I - (K^D)^{\frac{1}{2}}H^D_S(K^D)^{-\frac{1}{2}}\right)\boldsymbol{\delta}_1\boldsymbol{\delta}_1^t \\
& \times\left(I - (K^D)^{-\frac{1}{2}}H^D_S(K^D)^{\frac{1}{2}}\right)\boldsymbol{\omega}^D \\
& + c_1^2(\boldsymbol{\omega}^D)^t(K^D)^{\frac{1}{2}}H^D_S(K^D)^{\frac{1}{2}}\boldsymbol{\omega}^D \\
& + c_2^2(\boldsymbol{\omega}^{\bar{D}})^t\left(I - (K^{\bar{D}})^{\frac{1}{2}}H^{\bar{D}}_S(K^{\bar{D}})^{-\frac{1}{2}}\right)\boldsymbol{\delta}_2\boldsymbol{\delta}_2^t \\
& \times\left(I - (K^{\bar{D}})^{-\frac{1}{2}}H^{\bar{D}}_S(K^{\bar{D}})^{\frac{1}{2}}\right)\boldsymbol{\omega}^{\bar{D}} \\
& + c_2^2(\boldsymbol{\omega}^{\bar{D}})^t(K^{\bar{D}})^{\frac{1}{2}}H^{\bar{D}}_S(K^{\bar{D}})^{\frac{1}{2}}\boldsymbol{\omega}^{\bar{D}},
\end{aligned}$$

where $\tau_0^2 = c_1^2(\frac{\partial g}{\partial \boldsymbol{\theta}_1})^t(J^D_{00})^{-1}\frac{\partial g}{\partial \boldsymbol{\theta}_1} + c_2^2(\frac{\partial g}{\partial \boldsymbol{\theta}_2})^t(J^{\bar{D}}_{00})^{-1}\frac{\partial g}{\partial \boldsymbol{\theta}_2}$.

The MSE $\gamma(S)$ of $\widehat{g}_S(\mathbf{Z}_0)$ is still unknown. But it can be consistently estimated after obtaining the corresponding estimators $\widehat{K}^D, \widehat{K}^D_S, \widehat{H}^D_S, \widehat{K}^{\bar{D}}, \widehat{K}^{\bar{D}}_S$, $\widehat{H}^{\bar{D}}_S$, $\widehat{\omega}^D$, $\widehat{\omega}^{\bar{D}}$, and $\widehat{\boldsymbol{\delta}}_k$ $(k = 1, 2)$ under the full model. Furthermore, we can get values of these estimators at points $(\widehat{\boldsymbol{\theta}}^0_1, \boldsymbol{\eta}^0_1)$ and $(\widehat{\boldsymbol{\theta}}^0_2, \boldsymbol{\eta}^0_2)$, where $\widehat{\boldsymbol{\theta}}^0_1$ and $\widehat{\boldsymbol{\theta}}^0_2$ are the MLEs of $\boldsymbol{\theta}^0_1$ and $\boldsymbol{\theta}^0_2$ under either narrow model or full model.

Finally, we obtain the FIC with focus on AUC at a covariate vector $\mathbf{Z}_0$ as follows:

$$\begin{aligned}
\text{(2.2)}\quad \text{FIC} \ =\ & (\widetilde{\psi}_{\text{full}} - \widetilde{\psi}_S)^2 + 2\left[\widehat{c}_1^2(\widehat{\boldsymbol{\omega}}^D_S)^t\widehat{K}^D_S(\widehat{\boldsymbol{\omega}}^D_S)\right. \\
& \left. + \widehat{c}_2^2(\widehat{\boldsymbol{\omega}}^{\bar{D}}_S)^t\widehat{K}^{\bar{D}}_S(\widehat{\boldsymbol{\omega}}^{\bar{D}}_S)\right],
\end{aligned}$$

where $\widehat{\boldsymbol{\omega}}^D_S = \pi_S\widehat{\boldsymbol{\omega}}^D$, $\widehat{\boldsymbol{\omega}}^{\bar{D}}_S = \pi_S\widehat{\boldsymbol{\omega}}^{\bar{D}}$, $\widehat{c}_1 = \sqrt{1+\frac{n_2}{n_1}}$, $\widehat{c}_2 = \sqrt{1+\frac{n_1}{n_2}}$, $\widetilde{\psi}_{\text{full}} = \widehat{c}_1(\widehat{\boldsymbol{\omega}}^D)^t\widehat{\boldsymbol{\delta}}_1 + \widehat{c}_2(\widehat{\boldsymbol{\omega}}^{\bar{D}})^t\widehat{\boldsymbol{\delta}}_2$,

and $\widetilde{\psi}_S = \widehat{c}_1(\widehat{\boldsymbol{\omega}}^D)^t(\widehat{K}^D)^{\frac{1}{2}}\widehat{H}_S^D(\widehat{K}^D)^{-\frac{1}{2}}\widehat{\boldsymbol{\delta}}_1 + \widehat{c}_2(\widehat{\boldsymbol{\omega}}^{\bar{D}})^t(\widehat{K}^{\bar{D}})^{\frac{1}{2}}\widehat{H}_S^{\bar{D}}(\widehat{K}^{\bar{D}})^{-\frac{1}{2}}\widehat{\boldsymbol{\delta}}_2$.

This FIC can be used to do variable selection in ROC regression model. For a ROC regression model with q covariates, there are 2^q sub-models between the full model and the narrow model. Under each sub-model, the value of FIC can be calculated at covariates $\mathbf{Z}_0$ by using (2.2). With the FIC focused on AUC, the sub-model with the smallest FIC value is selected as the best sub-model. In other words, we select the set of the most important covariates which has the biggest effect on AUC estimation.

As an example, we apply Theorem 1 to the linear regression model (2.1) considered by Faraggi and Reiser (2002).

In the variable selection, We always keep the intercepts and the variances parameter (i.e., $\boldsymbol{\theta}_k = (\eta_{k,0}, \sigma_k)$) in the models, while explanatory variables (i.e., $\boldsymbol{\eta}_k = (\eta_{k,1}, \cdots, \eta_{k,q})^t$) need to be selected. At $\mathbf{Z}_0$, $AUC(\mathbf{Z}_0)$ is

$$g(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\eta}_1, \boldsymbol{\eta}_2|\mathbf{Z}_0) = \Phi\left(\frac{\mu^D - \mu^{\bar{D}}}{\sqrt{\sigma_1^2 + \sigma_2^2}}\right),$$

where $\mu^D = \eta_{1,0}X_0 + \boldsymbol{\eta}_1^t\mathbf{Z}_0$, $\mu^{\bar{D}} = \eta_{2,0}X_0^{\bar{D}} + \boldsymbol{\eta}_2^t\mathbf{Z}_0$, and $\Phi(\cdot)$ is the distribution function of $N(0,1)$.

Under the linear regression models (2.1), the information matrix evaluated at null points $(\boldsymbol{\theta}_k^0, \boldsymbol{\eta}_k^0)$, $k = 1, 2$, with $\boldsymbol{\eta}_1^0 = 0$ and $\boldsymbol{\eta}_2^0 = 0$, for diseased and non-diseased samples are

$$\frac{1}{\sigma^2}\begin{pmatrix} \frac{1}{n_1}\sum_{i=1}^{n_1}\frac{3\varepsilon_{0,i}^D}{\sigma_1^2} - 1 & \frac{1}{n_1}\sum_{i=1}^{n_1}\frac{2\varepsilon_{0,i}^D}{\sigma_1} & \frac{1}{n_1}\sum_{i=1}^{n_1}\frac{2\varepsilon_{0,i}^D}{\sigma_1}\mathbf{Z}_i^D \\ \frac{1}{n_1}\sum_{i=1}^{n_1}\frac{2\varepsilon_{0,i}^D}{\sigma_1} & \frac{1}{n_1}\sum_{i=1}^{n_1}X_i^D & \frac{1}{n_1}\sum_{i=1}^{n_1}\mathbf{Z}_i^D \\ \frac{1}{n_1}\sum_{i=1}^{n_1}\frac{2\varepsilon_{0,i}^D}{\sigma_1}\mathbf{Z}_i^D & \frac{1}{n_1}\sum_{i=1}^{n_1}\mathbf{Z}_i^D & \frac{1}{n_1}\sum_{i=1}^{n_1}(\mathbf{Z}_i^D)(\mathbf{Z}_i^D)^t \end{pmatrix},$$

$$\frac{1}{\sigma_2^2}\begin{pmatrix} \frac{1}{n_2}\sum_{j=1}^{n_2}\frac{3\varepsilon_{0,j}^{\bar{D}}}{\sigma_2^2} - 1 & \frac{1}{n_2}\sum_{j=1}^{n_2}\frac{2\varepsilon_{0,j}^{\bar{D}}}{\sigma_2} & \frac{1}{n_2}\sum_{j=1}^{n_2}\frac{2\varepsilon_{0,i}}{\sigma_2}\mathbf{Z}_j^{\bar{D}} \\ \frac{1}{n_2}\sum_{j=1}^{n_2}\frac{2\varepsilon_{0,j}^{\bar{D}}}{\sigma_2} & \frac{1}{n_2}\sum_{j=1}^{n_2}X_j^{\bar{D}} & \frac{1}{n_2}\sum_{j=1}^{n_2}\mathbf{Z}_j^{\bar{D}} \\ \frac{1}{n_2}\sum_{j=1}^{n_2}\frac{2\varepsilon_{0,j}^{\bar{D}}}{\sigma_2}\mathbf{Z}_j^{\bar{D}} & \frac{1}{n_2}\sum_{j=1}^{n_2}\mathbf{Z}_j^{\bar{D}} & \frac{1}{n_2}\sum_{j=1}^{n_2}(\mathbf{Z}_j^{\bar{D}})(\mathbf{Z}_j^{\bar{D}})^t \end{pmatrix},$$

respectively, where $\varepsilon_{0,i}^D = Y_i^D - \eta_{1,0}X_i^D$, and $\varepsilon_{0,j}^{\bar{D}} = Y_j^{\bar{D}} - \eta_{2,0}X_j^{\bar{D}}$ with $\boldsymbol{\eta}_1^0 = 0$ and $\boldsymbol{\eta}_2^0 = 0$ at the null points.

The partial derivatives of g, which are included in $\boldsymbol{\omega}^D$ and $\boldsymbol{\omega}^{\bar{D}}$, evaluated at the null points are

$$\left(\frac{\partial g(\cdot|\mathbf{Z}_0)}{\partial\sigma_1}, \frac{\partial g(\cdot|\mathbf{Z}_0)}{\partial\eta_{1,0}}, \frac{\partial g(\cdot|\mathbf{Z}_0)}{\partial\boldsymbol{\eta}_1}\right)$$
$$= \phi\left(\frac{\mu^D - \mu^{\bar{D}}}{\sqrt{\sigma_1^2 + \sigma_2^2}}\right)\frac{1}{\sqrt{\sigma_1^2 + \sigma_2^2}}\left(-\frac{(\mu^D - \mu^{\bar{D}})\sigma_1}{\sigma_1^2 + \sigma_2^2}, 1, \mathbf{Z}_0^t\right),$$

$$\left(\frac{\partial g(\cdot|\mathbf{Z}_0)}{\partial\sigma_2}, \frac{\partial g(\cdot|\mathbf{Z}_0)}{\partial\eta_{2,0}}, \frac{\partial g(\cdot|\mathbf{Z}_0)}{\partial\boldsymbol{\eta}_2}\right)$$
$$= \phi\left(\frac{\mu^D - \mu^{\bar{D}}}{\sqrt{\sigma_1^2 + \sigma_2^2}}\right)\frac{1}{\sqrt{\sigma_1^2 + \sigma_2^2}}\left(-\frac{(\mu^D - \mu^{\bar{D}})\sigma_2}{\sigma_1^2 + \sigma_2^2}, -1, -\mathbf{Z}_0^t\right),$$

where $\phi(\cdot)$ is the density function of $N(0,1)$.

The explicit formula of the FIC for the linear regression models can be derived based on Theorem 1 by using the above information matrix and the derivatives of g.

3. FIC BASED ON PLACEMENT VALUE MODELS

It is well known that the ROC curve can be viewed as the probability distribution of the placement value defined by $U = 1 - F^{\bar{D}}(Y) = P(Y^{\bar{D}} > Y)$ (Hanley and Haijian-Tilaki, 1997; Pepe and Cai, 2004). The placement value U is a transformation of Y that standardizes the distribution in the reference (non-diseased) population. It can be interpreted as the proportion of the reference population with values larger than Y. The distribution of the placement value in the non-disease (reference) population $U^{\bar{D}} = 1 - F^{\bar{D}}(Y^{\bar{D}})$ is $U(0,1)$ by definition, and $U^D = 1 - F^{\bar{D}}(Y^D) = P(Y^{\bar{D}} > Y^D)$ is the placement value in the diseased population. The distribution of U^D measures the separation between the diseased and non-diseased populations. If setting $u = 1 - F_{\bar{D}}(y)$ as a false positive rate, $ROC(u) = P(U^D < u)$. The corresponding AUC can be expressed as $AUC = E(1 - U^D)$.

Accounting for covariates can improve the diagnostic accuracy of a test. In this section, we consider the following placement-value model (see also Pepe and Cai, 2004):

$$H_{\boldsymbol{\alpha}}(U^D) = -\eta_0 X^D - \boldsymbol{\eta}^t\mathbf{Z}^D + \varepsilon, \tag{3.3}$$

where $U^D = 1 - F^{\bar{D}}(Y^D)$, X^D can be the intercept term or the variable always included in the model,ε has a specified distribution $g_{\boldsymbol{\gamma}}$ with parameter vector $\boldsymbol{\gamma}$, $H_{\boldsymbol{\alpha}}$ is an increasing function with parameter vector $\boldsymbol{\alpha}$, and $\boldsymbol{\eta} = (\eta_1, \cdots, \eta_q)^t$. The covariate-specific AUC at $\mathbf{Z}_0 = (Z_{0,1}, \cdots, Z_{0,q})$ is

$$g(\eta_0, \boldsymbol{\eta}^t, \boldsymbol{\alpha}, \boldsymbol{\gamma}|\mathbf{Z}_0) = E(1 - U^D|\mathbf{Z}_0).$$

Our goal is to select a set of important covariates from $(Z_{0,1}, \cdots, Z_{0,q})$ for model (3.3).

Let $\{(U_i^D, \mathbf{Z}_i^D), i = 1, \cdots, n_1\}$ be i.i.d. variables with the density function $f(U^D|\boldsymbol{\theta}, \boldsymbol{\eta}, \mathbf{Z}^D)$, where $\boldsymbol{\theta} = (\eta_0, \boldsymbol{\alpha}, \boldsymbol{\gamma})^t$ and $\boldsymbol{\eta} = (\eta_1, \cdots, \eta_q)$. The full model includes all available covariates, while the narrow model is the one with $\boldsymbol{\eta}^0 = 0$. The true model for the placement-value takes the form $f(U^D|\boldsymbol{\theta}^0, \boldsymbol{\eta}^0 + \boldsymbol{\delta}/\sqrt{n_1}, \mathbf{Z}^D)$, where $\boldsymbol{\delta} = (\delta_1, \cdots, \delta_q)^t$. Under these model assumptions, the covariate-specific AUC at $\mathbf{Z}_0 = (Z_{0,1}, \cdots, Z_{0,q})$ is $g(\mathbf{Z}_0) = g(\boldsymbol{\theta}^0, \boldsymbol{\eta}^0 + \boldsymbol{\delta}/\sqrt{n_1}|\mathbf{Z}_0)$. Let $(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\eta}}_S)$ be the MLE of $(\boldsymbol{\theta}, \boldsymbol{\eta}_S)$ under a sub-model S, then the covariate-specific AUC estimator under the sub-model is $\widehat{g}_S(\mathbf{Z}_0) = g(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\eta}}_S|\mathbf{Z}_0)$.

Under sub-model S, the information matrix and its inverse at $(\boldsymbol{\theta}^0, \boldsymbol{\eta}^0)$ are

$$J_S = \begin{pmatrix} J_{00} & J_{01,S} \\ J_{10,S} & J_{11,S} \end{pmatrix} \quad \text{and} \quad J_S^{-1} = \begin{pmatrix} J^{00,S} & J^{01,S} \\ J^{10,S} & J^{11,S} \end{pmatrix}.$$

The above matrices without subscript S are those corresponding to the full model. Denote $K = J^{11}$, $K_S = J^{11,S}$, $H_S = K^{-\frac{1}{2}}\pi_S^t K_S \pi_S K^{-\frac{1}{2}}$, $\boldsymbol{\omega} = J_{10}J_{00}^{-1}\frac{\partial g}{\partial \boldsymbol{\theta}} - \frac{\partial g}{\partial \boldsymbol{\eta}}$, where the partial derivatives are evaluated at the null point $(\boldsymbol{\theta}^0, \boldsymbol{\eta}^0)$. Similar to Theorem 1, we have the following theorem for the placement value model.

Theorem 2. *Assume that sub-model S is fitted for the placement value. Then, $\sqrt{n_1}(\widehat{g}_S(\mathbf{Z}_0) - g(\mathbf{Z}_0))$ converges in distribution to*

$$\begin{aligned} \boldsymbol{\Pi}_S &= (\frac{\partial g(\cdot|\mathbf{Z}_0)}{\partial \boldsymbol{\theta}})^t (J_{00})^{-1} M \\ &\quad + \boldsymbol{\omega}^t \left[\boldsymbol{\delta} - K^{\frac{1}{2}} H_S K^{-\frac{1}{2}} (\boldsymbol{\delta} + W) \right], \end{aligned}$$

where $M \sim N(0, J_{00})$, $W \sim N(0, K)$, M and W are independent.

The proof of Theorem 2 is relegated to the Appendix. From Theorem 2, it can be verified that $E(\boldsymbol{\Pi}_S) = \boldsymbol{\omega}^t \left(I - K^{\frac{1}{2}} H_S K^{-\frac{1}{2}} \right) \boldsymbol{\delta}$ and $Var(\boldsymbol{\Pi}_S) = (\frac{\partial g}{\partial \boldsymbol{\theta}})^t J_{00}^{-1} \frac{\partial g}{\partial \boldsymbol{\theta}} + \boldsymbol{\omega}^t K^{\frac{1}{2}} H_S K^{\frac{1}{2}} \boldsymbol{\omega}$. We can derive the FIC with focus on AUC at covariate $\mathbf{Z}_0$ as follows:

$$\text{FIC} = (\tilde{\phi}_{\text{full}} - \tilde{\phi}_S)^2 + 2(\widehat{\boldsymbol{\omega}}_S)^t \widehat{K}_S \widehat{\boldsymbol{\omega}}_S, \tag{3.4}$$

where $\widehat{\boldsymbol{\omega}}_S = \pi_S \widehat{\boldsymbol{\omega}}$, $\tilde{\phi}_{\text{full}} = \widehat{\boldsymbol{\omega}}^t \widehat{\boldsymbol{\delta}}_{\text{full}}$, and $\tilde{\phi}_S = \widehat{\boldsymbol{\omega}}^t \widehat{K}^{\frac{1}{2}} \widehat{H}_S \widehat{K}^{-\frac{1}{2}} \widehat{\boldsymbol{\delta}}_{\text{full}}$. Using (3.4), we will choose the sub-model with the smallest value of FIC at covariate $\mathbf{Z}_0$ among all the possible candidate models.

As an example, let's consider the following model for the placement value:

$$\Phi^{-1}(U^D|\mathbf{Z}^D) = -\eta_0 X^D - \boldsymbol{\eta}^t \mathbf{Z}^D + \varepsilon, \tag{3.5}$$

where $\boldsymbol{\eta} = (\eta_1, \cdots, \eta_q)^t$, ε follows $N(0, \sigma^2)$.

Under model (3.5), $\boldsymbol{\theta} = (\eta_0, \sigma^2)$, and the density function of U^D can be expressed as

$$f(U^D|\mathbf{Z}^D) = \frac{1}{\sigma} \exp\{-\frac{\varepsilon^2}{2\sigma^2} + \frac{[\Phi^{-1}(U^D|\mathbf{Z}^D)]^2}{2}\}.$$

Then the information matrix evaluated at the null points $(\boldsymbol{\theta}^0, \boldsymbol{\eta}^0)$ is

$$\frac{1}{\sigma^2} \begin{pmatrix} \frac{1}{n_1}\sum_{i=1}^{n_1} \frac{3\varepsilon_{0,i}}{\sigma^2} - 1 & \frac{1}{n_1}\sum_{i=1}^{n_1} \frac{2\varepsilon_{0,i}}{\sigma} & \frac{1}{n_1}\sum_{i=1}^{n_1} \frac{2\varepsilon_{0,i}}{\sigma} \mathbf{Z}_i^D \\ \frac{1}{n_1}\sum_{i=1}^{n_1} \frac{2\varepsilon_{0,i}}{\sigma} & \frac{1}{n_1}\sum_{i=1}^{n_1} X_i^D & \frac{1}{n_1}\sum_{i=1}^{n_1} \mathbf{Z}_i^D \\ \frac{1}{n_1}\sum_{i=1}^{n_1} \frac{2\varepsilon_{0,i}}{\sigma} \mathbf{Z}_i^D & \frac{1}{n_1}\sum_{i=1}^{n_1} \mathbf{Z}_i^D & \frac{1}{n_1}\sum_{i=1}^{n_1} (\mathbf{Z}_i^D)(\mathbf{Z}_i^D)^t \end{pmatrix}.$$

where $\varepsilon_{0,i} = \Phi^{-1}(U_i^D) + \eta_0 X_i^D$ with $\boldsymbol{\eta}^0 = 0$ at the null point.

For given $\mathbf{Z}_0$, denote $Q_0 = \frac{1}{2}\left([\Phi^{-1}(U^D|\mathbf{Z}_0)]^2 - \frac{\varepsilon_0^2}{\sigma^2} \right)$. Then the derivatives of g, which are included in $\boldsymbol{\omega}$, evaluated at the null points $(\boldsymbol{\theta}^0, \boldsymbol{\eta}^0)$ are

$$\begin{aligned} \frac{\partial g}{\partial \sigma}(\cdot|\mathbf{Z}_0) &= \int (1 - U^D)(-\frac{1}{\sigma^2} + \frac{1}{\sigma^4}\varepsilon_0^2) \exp(Q_0) dU^D, \\ \frac{\partial g}{\partial \eta_0}(\cdot|\mathbf{Z}_0) &= \int (1 - U^D)(-\frac{1}{\sigma^3})\varepsilon_0 \exp(Q_0) dU^D, \\ \frac{\partial g}{\partial \boldsymbol{\eta}}(\cdot|\mathbf{Z}_0) &= \int (1 - U^D)(-\frac{1}{\sigma^3})\varepsilon_0 \mathbf{Z}_0^D \exp(Q_0) dU^D. \end{aligned}$$

Therefore, we can derive the explicit formula of FIC for placement value model by using (3.4).

4. SIMULATION STUDIES

In this section, based on the placement value model, we conduct simulation studies to evaluate the finite sample performances of the AIC, BIC and FIC in terms of the Mean Square Error (MSE) and the Mean Absolute Deviation (MAD) of the estimators for AUC index.

For the diseased sample, the AIC and BIC under a sub-model S can be expressed as (See Hjort and Claeskens, 2003):

$$\text{AIC}_S^D = -\widehat{\boldsymbol{\delta}}_{\text{full}} (K^D)^{-\frac{1}{2}} H_S^D (K^D)^{-\frac{1}{2}} \widehat{\boldsymbol{\delta}}_{\text{full}} + 2|S|, \tag{4.6}$$

$$\text{BIC}_S^D = -\widehat{\boldsymbol{\delta}}_{\text{full}} (K^D)^{-\frac{1}{2}} H_S^D (K^D)^{-\frac{1}{2}} \widehat{\boldsymbol{\delta}}_{\text{full}} + log(n)|S|, \tag{4.7}$$

respectively, where $|S|$ is the number of elements in S.

Based on expressions (4.6) and (4.7), we choose the models with the smallest AIC and BIC value as the best one. Using the FIC criteria, we choose the model with the smallest FIC value focused on AUC as the best one. In simulation studies, we compare performances of the AIC, BIC, and FIC criteria through comparing the estimates $\widehat{AUC}(\mathbf{Z}_0)$ of AUC at the given covariates $\mathbf{Z}_0$, the MSE and MAD of $\widehat{AUC}(\mathbf{Z}_0)$ over M=1000 simulation runs under each simulation setting, where $\text{MSE}(\mathbf{Z}_0) = \frac{1}{M}\sum_{m=1}^{M}(\widehat{AUC}_m(\mathbf{Z}_0) - AUC(\mathbf{Z}_0))^2$, $\text{MAD}(\mathbf{Z}_0) = \frac{1}{M}\sum_{m=1}^{M}|\widehat{AUC}(\mathbf{Z}_0)_m - AUC(\mathbf{Z}_0)|$, and $\widehat{AUC}(\mathbf{Z}_0)_m$ is the estimate for $AUC(\mathbf{Z}_0)$ based on the m-th simulated sample.

We use the following placement value model in examples 1–5:

$$\Phi^{-1}(U^D|\mathbf{Z}^D) = -\eta_0 X^D - \boldsymbol{\eta}^t \mathbf{Z}^D + \varepsilon,$$

where $\varepsilon \sim N(0, \sigma^2)$. The simulated data are generated from the models with different simulation settings.

Example 1: We set $X^D = 1$. The q dimension covariates $\mathbf{Z}^D$ are generate from $\mathbf{Z}^D \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu} = (1, \cdots, 1)$, and the covariance matrix $\boldsymbol{\Sigma} = (\Sigma_{ij})$ with $\Sigma_{ii} = \rho^{|i-j|}$, $1 \le i \ne j \le q$. The correlation coefficient ρ is chosen to be 0. We choose $\boldsymbol{\theta} = (\eta_0, \sigma) = (0.8, 0.1)$, and $\boldsymbol{\eta} = (0.5, 0.3, 0.2, 0)$ with $q = 4$. The diseased sample size is $n_1 = 300$.

Example 2: The model is the same as that in example 1 except that the sample size is $n_1 = 500$;

Example 3: The model is the same as that in example 1 except that the sample size is $n_1 = 1000$;

Example 4: The model is the same as that in example 1 except that $\boldsymbol{\eta} = (0.5, 0.3, 0, 0, 0, 0, 0.4, 0)$ with $q = 8$;

Example 5: To consider the robustness of the proposed method, we consider a case in which the error term doesn't follow the normal distribution, but the simulation is still conducted under the assumption that the error follows the normal distribution. The placement values are generated from the following model:

$$\Phi^{-1}(U^D|\mathbf{Z}^D) = -\eta_{\mathbf{0}}\mathbf{X^D} - \boldsymbol{\eta}^{\mathbf{t}}\mathbf{Z^D} + \varepsilon,$$

where $\mathbf{X}^D = (1, \xi)^t$ with $\xi \sim N(0, 1)$, $\mathbf{Z}^D$ are generated from the same distribution as that in example 1. The coefficients $\eta_{\mathbf{0}} = (\mathbf{0.2}, \mathbf{0.1})$ and $\eta = (\mathbf{0.5}, \mathbf{0}, \mathbf{0.3}, \mathbf{0.2}, \mathbf{0}, \mathbf{0})$. The true distribution of error term is $\varepsilon \sim 0.1t(3)$, where $t(3)$ is a t-distribution with 3 degree of freedom.

For given $\mathbf{Z}_0$, AUC can be expressed as $AUC(\mathbf{Z}_0) = g(\boldsymbol{\theta}^0, \boldsymbol{\eta}^0 + \boldsymbol{\delta}/\sqrt{n_1}|\mathbf{Z}_0) = E\left(1 - U^D|\mathbf{Z}_0\right)$. Using the simulated data from the true placement value models described in examples 1–5, we estimate AUCs at 100 different covariates $\mathbf{Z}_0$ and the corresponding MSE($\mathbf{Z}_0$)'s and MAD($\mathbf{Z}_0$)'s under the selected models by using AIC, BIC and FIC over M=1000 simulation runs, respectively.

Figures 1–5 display the results for AUC, MSE amd MAD comparisons by using the AIC, BIC and FIC. From these figures, we can see that the true AUC is varying with $\mathbf{Z}_0$, and the estimates of AUC based on FIC are much closer to the true AUC than the AIC and BIC based estimates. Figures 1–5 show that the MSE($\mathbf{Z}_0$) and the MAD($\mathbf{Z}_0$) based on the FIC selected models are smaller than those based on the AIC and BIC selected models in most cases considered here, which indicates that the FIC has better finite sample performances than the AIC and BIC in variable selection of placement value model.

In examples 1–5, we also consider cases with $\rho = 0.5$ and $\rho = 0.8$, the simulation results are similar to those with $\rho = 0$. To save space, the figures with $\rho = 0.5$ and 0.8 are put in the supplemental file of this article: http://intlpress.com/site/pub/pages/journals/items/sii/content/vols/0010/0002/s001.

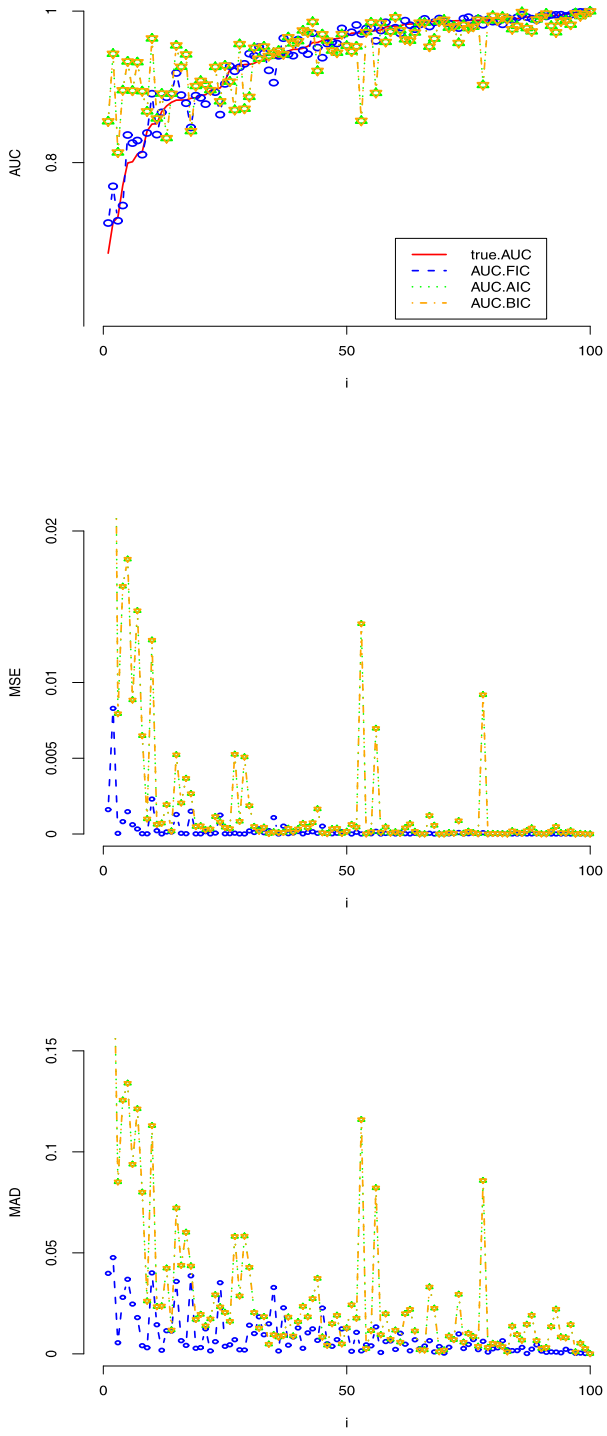
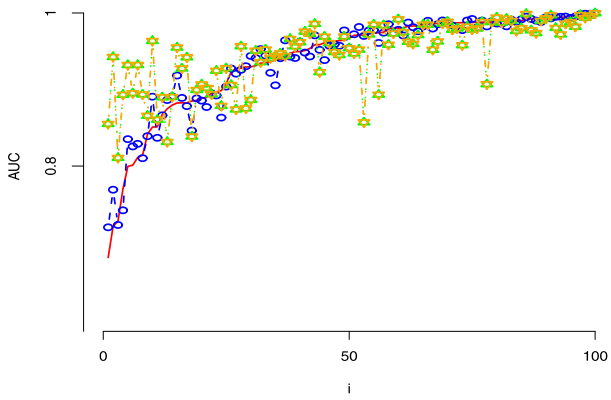


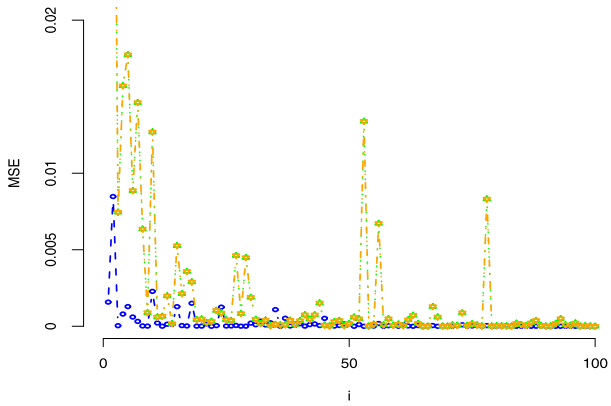
Figure 1. Example 1: Comparison with $n_1 = 300$ and $\rho = 0$.

5. ANALYSIS OF THE AUDIOLOGY DATA

In this section, the audiology data from the DPOAE test described in Section 1 are used to evaluate the diagnostic accuracy of the test with the hearing device. The study involved 107 hearing impaired and 103 normally hearing subjects who were examined at three frequencies (f) and three intensity (L) settings of the DPOAE device. The effect of severity of hearing impairment is also of interest. An audiometric threshold can be yielded at each setting. If the audiometric threshold is greater than 20 dB HL, the disease variable $D = 1$; otherwise $D = 0$. Each subject was tested in only one ear. The test result is the negative signal to noise ratio, −SNR.

In this dataset, the covariates to be selected are $Z_1 = frequency$ HZ/100, $Z_2 = intensity$ dB/10, and $Z_3 = (hearing\ threshold - 20)$dB/10. Z_1 takes three values: 10.01, 14.16, and 20.02. Z_2 also takes three values: 5.5,





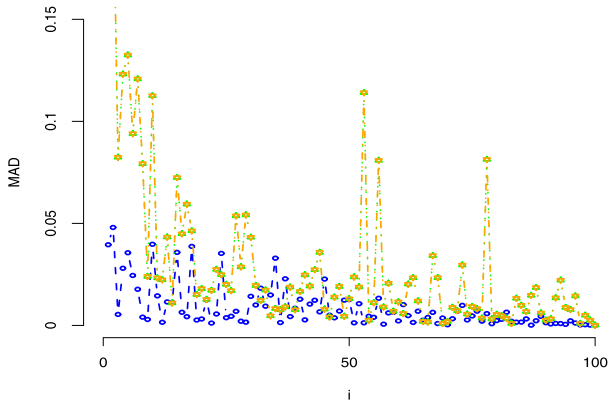
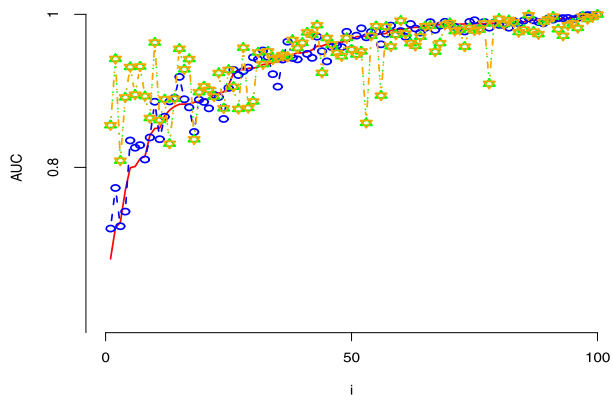
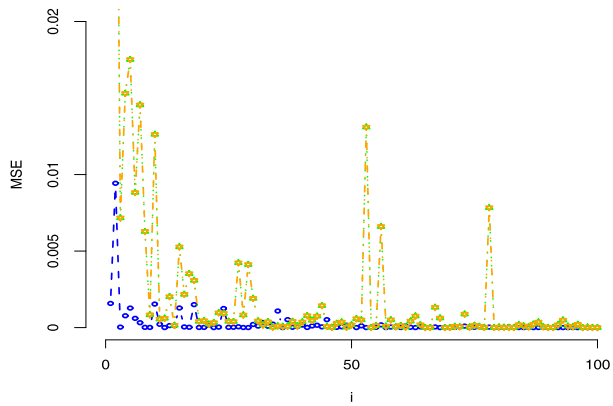


Figure 2. Example 2: Comparison with $n_1 = 500$ *and* $\rho = 0$.





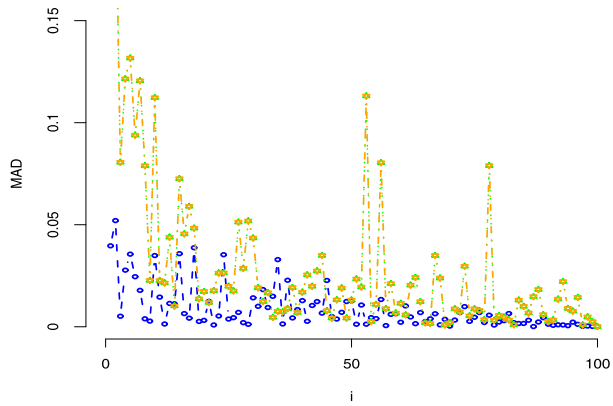


Figure 3. Example 3: Comparison with $n_1 = 1000$ *and* $\rho = 0$.

6.0, and 6.5. Z_3 is a centered continuous variable. We denote the 25% sample quantile of Z_3 as $Z_{3,1}$, the 50% sample quantile of Z_3 as $Z_{3,2}$, and the 75% sample quantile of Z_3 as $Z_{3,3}$. In order to encourage the model selection, we incorporate two-way interaction terms, i.e. $\mathbf{Z}^D = (Z_1, Z_2, Z_3, Z_1Z_2, Z_1Z_3, Z_2Z_3)$. We select three covariate vectors $\mathbf{Z_0}$: $(10.01, 5.5, Z_{3,1}, 10.01*5.5, 10.01*Z_{3,1}, 5.5*Z_{3,1})$ (case i), $(14.16, 6.0, Z_{3,2}, 14.16 * 6.0, 14.16 * Z_{3,2}, 6.0 * Z_{3,2})$ (case ii), and $(20.02, 6.5, Z_{3,3}, 20.02 * 6.5, 20.02 * Z_{3,3}, 6.5 * Z_{3,3})$ (case iii) as the specific values of the covariates to illustrate the proposed FIC method. The placement value model (3.5) is used to fit this DPOAE dataset.

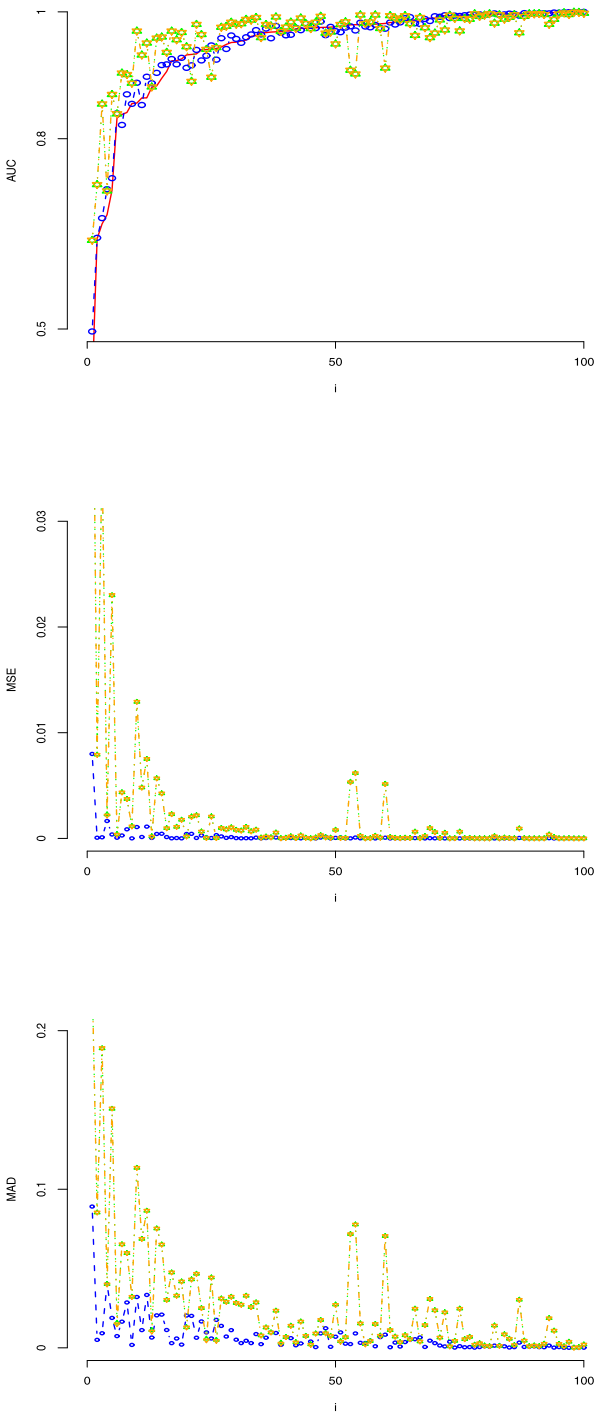


Figure 4. Example 4: Comparison with $n_1 = 300$, $\rho = 0$, and $q = 8$.

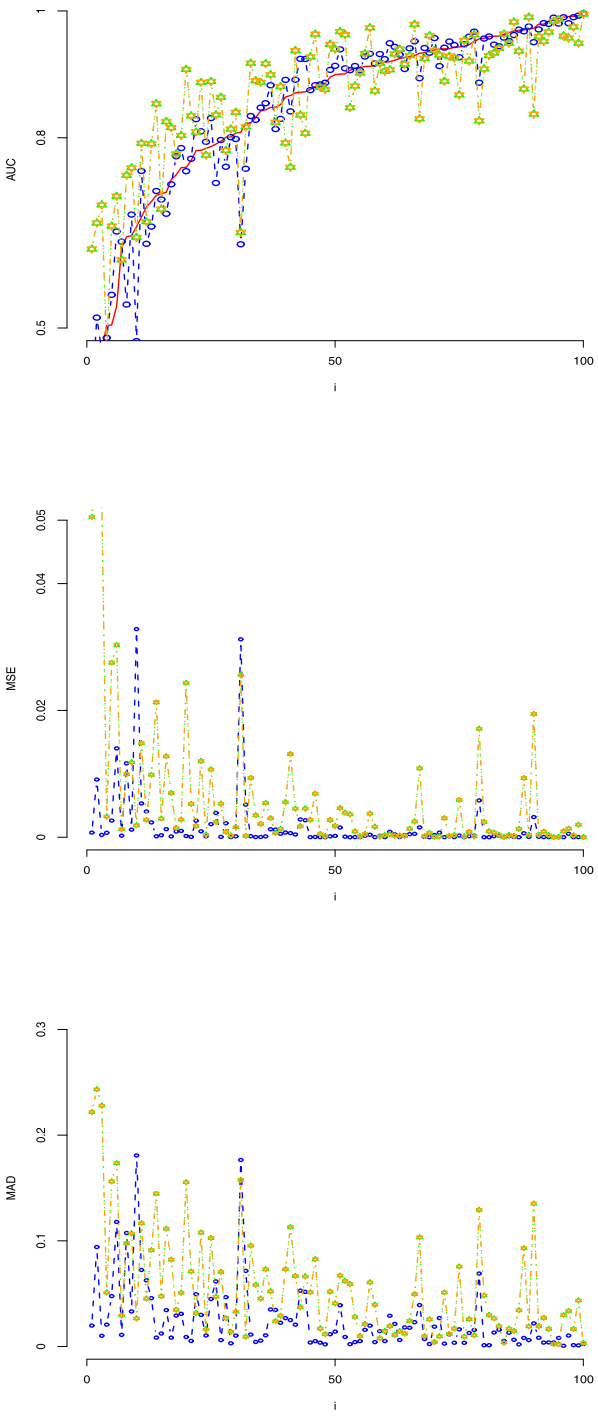


Figure 5. Example 5: Comparison with $n_1 = 300$, $\rho = 0$, $q = 6$ and $\varepsilon \sim 0.1t(3)$.

Tables 1–2 show the models selected by using the AIC, BIC and FIC with focus on AUC, and estimates for the model parameters as well as predictions of $AUC(\mathbf{Z_0})$ at the three given covariate vectors of $\mathbf{Z_0}$. Based on the AIC (or BIC) selected model, we estimate the AUC value at the three given covariates (denoted as (i), (ii), and (iii)). Based on the FIC selected models, we also can estimate the AUC values at the three given covariates. It is shown in Table 1 that both the AIC and BIC methods select the same final model. In Table 2, we can see that FIC method selects different variables at different $\mathbf{Z}_0$'s. The estimated AUC values (which are 0.9438, 0.9639, 0.9914) based on the selected placement

Table 1. Variable selection in placement value model for the DPOAE data based on AIC and BIC criteria

		based on AIC	based on BIC
sub-model		0 0 1 1 0 1	0 0 1 1 0 1
$\hat{\boldsymbol{\eta}}$		(0, 0, 0.1707, 0.0566, 0, 0.0238)	(0, 0, 0.1707, 0.0566, 0, 0.0238)
$\hat{\boldsymbol{\theta}}$		(1.4027,0.4817)	(1.4027, 0.4817)
$\widehat{\text{AUC}}$	i	0.9278	0.9278
	ii	0.9539	0.9539
	iii	0.9847	0.9847

Table 2. Variable selection in placement value model for the DPOAE data based on FIC criteria

			based on FIC	
	sub-model	$\hat{\boldsymbol{\eta}}$	$\hat{\boldsymbol{\theta}}$	$\widehat{\text{AUC}}$
i	0 0 1 0 1 1	(0, 0, 0.1696, 0 0.0201, -0.0238)	(1.4196, 0.5318)	0.9438
ii	0 0 1 0 0 1	(0, 0, 0.1703, 0, 0, -0.0238)	(1.4089, 0.5318)	0.9639
iii	0 0 1 1 1 0	(0.0282, 0.0708, -0.0055, 0)	(1.3982, 0.5318)	0.9914

value model are closer to 1 and higher than those (0.9278 0.9539 0.9847) based on the AIC and BIC selected model. These results show that the final model selected by the FIC results in higher covariate-specific AUC estimates than the AIC and BIC selected model in this example.

6. DISCUSSION

In this paper, we have discussed how to select important covariates in ROC analysis with focus on improving diagnostic accuracy of a test. In diagnostic testing, instead of predicting test measurements of the test conducted on new individuals, we are interested in the diagnostic accuracy of the test to distinguish diseased subjects from non-diseased subjects. The classical variable selection criteria such as AIC, BIC may not be suitable for this purpose. Claeskens and Hjort (2003) claimed that the variable selector should focus on the accuracy of the parameter of interest, and proposed the FIC criteria. Since AUC is the most popular summary index of the discriminatory accuracy of a test, we have considered variable selection with focus on AUC based on the placement value model. From our simulation studies and a real data analysis, we observe that the proposed FIC performs better than the AIC and BIC in placement value model selection. Therefore, we recommend the FIC based model selection method for placement value models in ROC analysis in estimation of a covariate-specific AUC.

APPENDIX: PROOFS

Proof of Theorem 1.

After some algebra, we can further simplify A_S^D and B_S^D in Lemma 1 as follows

$$\begin{aligned} A_S^D &= (J_{00}^D)^{-1}J_{01}^D(I-(K^D)^{\frac{1}{2}}H_S^D(K^D)^{-\frac{1}{2}})\boldsymbol{\delta}^1 \\ &\quad +(J_{00}^D)^{-1}M^D-(J_{00}^D)^{-1}J_{01}^D(K^D)^{\frac{1}{2}}H_S^D(K^D)^{-\frac{1}{2}}W^D \\ B_S^D &= K_S^D\pi_S(K^D)^{-1}(\boldsymbol{\delta}^1+W^D). \end{aligned}$$

Under sub-model S, the density is $f_1(y^D|\boldsymbol{\theta}_1,\boldsymbol{\eta}_{1,S},\mathbf{Z}^D)$. By Taylor expansion at null points $(\boldsymbol{\theta}_1^0,\boldsymbol{\eta}_{1,S}^0)$ and $(\boldsymbol{\theta}_2^0,\boldsymbol{\eta}_{2,S}^0)$, we get that

$$\begin{aligned} &\sqrt{n_1+n_2}\left(\widehat{g}_S(\widehat{\boldsymbol{\theta}}_1,\widehat{\boldsymbol{\eta}}_{1,S},\widehat{\boldsymbol{\theta}}_2,\widehat{\boldsymbol{\eta}}_{2,S})\right. \\ &\qquad \left.-g(\boldsymbol{\theta}_1^0,\boldsymbol{\eta}_{1,S}^0+\frac{\boldsymbol{\delta}^1}{\sqrt{n_1}},\boldsymbol{\theta}_2^0,\boldsymbol{\eta}_{2,S}^0+\frac{\boldsymbol{\delta}^2}{\sqrt{n_2}})\right) \\ =\ & \frac{\sqrt{n_1+n_2}}{\sqrt{n_1}}\left(\frac{\partial g(\boldsymbol{\theta}_1^0,\boldsymbol{\eta}_{1,S}^0,\boldsymbol{\theta}_2^0,\eta_{2,S}^0)}{\partial\boldsymbol{\theta}_1}\sqrt{n_1}(\widehat{\boldsymbol{\theta}}_1-\boldsymbol{\theta}_1^0)\right. \\ &\left.+\frac{\partial g(\boldsymbol{\theta}_1^0,\boldsymbol{\eta}_{1,S}^0,\boldsymbol{\theta}_2^0,\eta_{2,S}^0)}{\partial\boldsymbol{\eta}_{1,S}}\sqrt{n_1}(\widehat{\boldsymbol{\eta}}_{1,S}-\boldsymbol{\eta}_{1,S}^0)\right) \\ &+\frac{\sqrt{n_1+n_2}}{\sqrt{n_2}}\left(\frac{\partial g(\boldsymbol{\theta}_1^0,\boldsymbol{\eta}_{1,S}^0,\boldsymbol{\theta}_2^0,\boldsymbol{\eta}_{2,S}^0)}{\partial\boldsymbol{\theta}_2}\sqrt{n_2}(\widehat{\boldsymbol{\theta}}_2-\boldsymbol{\theta}_2^0)\right. \\ &\left.+\frac{\partial g(\boldsymbol{\theta}_1^0,\boldsymbol{\eta}_{1,S}^0,\boldsymbol{\theta}_2^0,\boldsymbol{\eta}_{2,S}^0)}{\partial\boldsymbol{\eta}_{2,S}}\sqrt{n_2}(\widehat{\boldsymbol{\eta}}_{2,S}-\boldsymbol{\eta}_{2,S}^0)\right) \\ &-\frac{\sqrt{n_1+n_2}}{\sqrt{n_1}}\frac{\partial g(\boldsymbol{\theta}_1^0,\boldsymbol{\eta}_{1,S}^0,\boldsymbol{\theta}_2^0,\boldsymbol{\eta}_{2,S}^0)}{\partial\boldsymbol{\eta}_{1,S}}\boldsymbol{\delta}^1 \\ &-\frac{\sqrt{n_1+n_2}}{\sqrt{n_2}}\frac{\partial g(\boldsymbol{\theta}_1^0,\boldsymbol{\eta}_{1,S}^0,\boldsymbol{\theta}_2^0,\boldsymbol{\eta}_{2,S}^0)}{\partial\boldsymbol{\eta}_{2,S}}\boldsymbol{\delta}^2 \\ =\ & c_1(\frac{\partial g}{\partial\boldsymbol{\theta}_1})^t(J_{00}^D)^{-1}M^D+c_2(\frac{\partial g}{\partial\boldsymbol{\theta}_2})^t(J_{00}^{\bar{D}})^{-1}M^{\bar{D}} \\ &+c_1(\boldsymbol{\omega}^D)^t\left[\boldsymbol{\delta}^1-(K^D)^{1/2}H_S^D(K^D)^{-1/2}(\boldsymbol{\delta}^1+W^D)\right] \\ &+c_2(\boldsymbol{\omega}^{\bar{D}})^t\left[\boldsymbol{\delta}^2-(K^{\bar{D}})^{1/2}H_S^{\bar{D}}(K^{\bar{D}})^{-1/2}(\boldsymbol{\delta}^2+W^{\bar{D}})\right]. \end{aligned}$$

Denote

$$\begin{aligned} W^D &= J^{10,D}M^D+J^{11,D}N^D \\ &= K^D(N^D-J_{10}^D(J_{00}^D)^{-1}M^D), \end{aligned}$$

where $M^D\sim N_2(0,J_{00}^D)$. It is easy to get that $E(M^D)=0$,

$E(W^D) = E(J^{10,D}M^D + J^{11,D}N^D) = 0$, and

$$\begin{aligned} E(W^D M^D) &= E(K^D(N^D - J_{10}^D(J_{00}^D)^{-1}M^D)M^D) \\ &= K^D\left(J_{10}^D - J_{10}^D(J_{00}^D)^{-1}J_{00}^D\right) = 0, \\ E(W^D(W^D)^t) &= E\left[(K^D(N^D - J_{10}^D(J_{00}^D)^{-1}M^D)\right. \\ &\qquad \left.(K^D(N^D - J_{10}^D(J_{00}^D)^{-1}M^D)^t\right] \\ &= K^D\left(J_{11}^D - J_{10}^D(J_{00}^D)^{-1}J_{10}^D\right)(K^D)^t \\ &= K^D. \end{aligned}$$

Then Theorem 1 follows from the independence between W^D and M^D, and $W^D \sim N(0, K^D)$.

Theorem 2 can be proved by following the proof of Theorem 1. Theorem 2 can also be proved by direct use of Claeskens and Hjort (2003)'s method.

Received 16 December 2015

REFERENCES

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. *2nd International Symposium on Information Theory*, Petrov B. N. and Caski F., Eds, 267–281. MR0483125

Claeskens, G. and Hjort, N. L. (2003). The focused information criterion. *Journal of the American Statistical Association*, 98, 900–916. MR2041482

Dodd, L. E. and Pepe, M. S. (2003). Semi-parametric regression for the area under the receiver operating characteristic curve. *Journal of the American Statistical Association*, 98, 409–417. MR1995717

Dorfman, D. D., Berbaum, K. S., and Metz, C. E. (1992). Receiver operating characteristic analysis: generalization to the population of readers and patients with the jackknife method. *Statistics in Radiology*, 27, 723–731.

Faraggi, D. and Reiser, B. (2002). Estimation of the area under the ROC curve. *Statistics in Medicine*, 21, 3093–3106.

Hanley, J. A. and Hajian-Tilaki, K. O. (1997). Sampling variability of nonarametric estimates of the areas under receiver operating characteristic curves: An update. *Academic Radiology*, 4, 49–58.

Hjort, N. L. and Claeskens, G. (2003). Frequentist model average estimators. *Journal of the American Statistical Association*, 98, 879–899. MR2041481

Hjort, N. L. and Claeskens, G. (2006). Focused information criteria and model averaging for the Cox hazard regression model. *Journal of the American Statistical Association*, 101, 1449–1464. MR2279471

Obuchowski, N. A. (1995). Multireader, multimodality receiver operating characteristic curve studies: Hypothesis testing and sample size estimation using analysis of variance approach with dependent observations. *Academic Radiology*, 2(Suppl 1), 22–29.

Pepe, M. S. (2003). The Statistical Evaluation of Medical Tests for Classification and Prediction. New York: Oxford University Press. MR2260483

Pepe, M. S. and Cai, T. (2004). The analysis of placement values for evaluating discriminatory measures. *Biometrics*, 60, 528–535. MR2067011

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6, 461–464. MR0468014

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van der, L. (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society, Ser B*, 64, 583–639. MR1979380

Stover, L., Gorga, M. P., Neely, S. T., and Montoya, D. (1996). Toward optimizing the clinical utility of distortion product otoacoustic emission measurements. *Journal of the Acoustical Society of America*, 100, 956–967.

Swets, J. A. and Pickett, R. M. (1982). Evaluation of Diagnostic Systems Methods from Signal Detection Theory. Academic Press.

Thompson, M. L. and Zucchini, W. (1989). On the statistical analysis of ROC curves. *Statistics in Medicine*, 8, 1277–1290.

Wang, Y., Chen, H., Li, R., Duan, N., and Lewis-Fernádez, R. (2011). Prediction-based structured variable selection through the receiver operating characteristic curves. *Biometrics*, 67, 896–905. MR2829264

Zhou, X. H., Obuchowski, N. A., and McClish, D. K. (2002). Statistical Methods in Diagnostic Medicine. New York: John Wiley & Sons. MR1915698

Baoying Yang
Department of Statistics
College of Mathematics
Southwest Jiaotong University
China
E-mail address: yangbaoying@home.swjtu.edu.cn

Xin Huang
Division of Public Health Sciences
Fred Hutchinson Cancer Research Center
Seattle, WA
USA
E-mail address: watsonxhuang@gmail.com

Gengsheng Qin
Department of Mathematics and Statistics
Georgia State University
USA
E-mail address: gqin@gsu.edu