Hindawi Publishing Corporation
Computational and Mathematical Methods in Medicine
Volume 2013, Article ID 436493, 8 pages
http://dx.doi.org/10.1155/2013/436493

Research Article

Variable Selection in ROC Regression

Binhuan Wang

New York University School of Medicine, New York, NY 10016, USA

Correspondence should be addressed to Binhuan Wang; binhuan.wang@nyumc.org

Received 6 August 2013; Accepted 18 September 2013

Academic Editor: Gengsheng Qin

Regression models are introduced into the *receiver operating characteristic* (ROC) analysis to accommodate effects of covariates, such as genes. If many covariates are available, the variable selection issue arises. The traditional induced methodology separately models outcomes of diseased and nondiseased groups; thus, separate application of variable selections to two models will bring barriers in interpretation, due to differences in selected models. Furthermore, in the ROC regression, the accuracy of area under the curve (AUC) should be the focus instead of aiming at the consistency of model selection or the good prediction performance. In this paper, we obtain one single objective function with the group SCAD to select grouped variables, which adapts to popular criteria of model selection, and propose a two-stage framework to apply the focused information criterion (FIC). Some asymptotic properties of the proposed methods are derived. Simulation studies show that the grouped variable selection is superior to separate model selections. Furthermore, the FIC improves the accuracy of the estimated AUC compared with other criteria.

1. Introduction

In modern medical diagnosis or genetic studies, the *receiver operating characteristic* (ROC) curve is a popular tool to evaluate the discrimination performance of a certain biomarker on a disease status or a phenotype. For example, in a continuous-scale test, the diagnosis of a disease is dependent upon whether a test result is above or below a specified cutoff value. Also, genome-wide association studies in human populations aim at creating genomic profiles which combine the effects of many associated genetic variants to predict the disease risk of a new subject with high discriminative accuracy [1]. For a given cutoff value of a biomarker or a combination of biomarkers, the sensitivity and the specificity are employed to quantitatively evaluate the discriminative performance. By varying cutoff values throughout the entire real line, the resulting plot of sensitivity against 1-specificity is a ROC curve. The area under the ROC curve (AUC) is an important one-number summary index of the overall discriminative accuracy of a ROC curve, by taking the influence of all cutoff values into account. Let Y_D be the response of a diseased subject, and let $Y_{\overline{D}}$ be the response of a nondiseased subject; then, the AUC can be expressed as $P(Y_D > Y_{\overline{D}})$ [2]. Pepe [3] and Zhou et al. [4] provided broad reviews on many statistical methods for the evaluation of diagnostic tests.

Traditional ROC analyses do not consider the effect of characteristics of study subjects or operating conditions of the test, so test results may be affected in the way of influencing distributions of test measurements for diseased and/or nondiseased subjects. Additionally, although the number of genes is large, there may be only a small number of them associated with the disease risk or phenotype. Therefore, regression models are introduced into the ROC analysis. Chapter Six in Pepe [3] offered a wonderful introduction to the adjustment for covariates in ROC curves. As reviewed in Rodríguez-Álvarez et al. [5], there are two main methodologies of regression analyses in ROC: (1) "induced" methodology, which firstly models outcomes of diseased and nondiseased subjects separately and then uses these outcomes to induce ROC and AUC and (2) "direct" methodology, which directly models the AUC on all covariates. In this paper, we focus on the induced methodology, to which current model selection techniques may be extended.

If there are many covariates, the variable selection issue arises in terms of the consideration of model interpretation and estimability. There are two main groups of variable selection procedures. One is the best-subset selection associated with criteria such as cross-validation (CV, [6]), generalized cross-validation (GCV, [7]), AIC [8], and BIC [9]. The other is based on regularization methods such as LASSO [10],

SCAD [11], and adaptive LASSO [12], with tuning parameters selected by the same criteria such as CV and BIC. Procedures in the second group have recently become popular because they are stable [13] and applicable for high-dimensional data [14].

So far, not much attention has been drawn on the topic of variable selection in the ROC regression. Two possible reasons may account for this situation. Firstly, if we model outcomes of diseased and nondiseased subjects separately, selected submodels may be different. The difference will result in difficulties in interpretation, because it is natural to expect that the same set of variables contributes to discriminating diseased and nondiseased subjects. Secondly, most current criteria for variable selection procedures focus on the prediction performance or variable selection consistency. However, in the ROC regression, instead of prediction or model selection, our focus is the precision of an estimated AUC, which means that most popular criteria may not be appropriate. Claeskens and Hjort [15] argued that these "one-fit-all" model selection criteria aim at selecting a single model with good overall properties. Alternatively, they developed the focused information criterion (FIC), which focuses on a parameter singled out for interests. The insight behind this criterion is that a model that gives good precision for one estimand may be worse when used in inference for another estimand. Wang and Fang [16] successfully applied the FIC to variable selection in linear models and demonstrated that the FIC exactly improved the estimation performance of singled-out parameters. This "individualized" criterion exactly fits the ROC regression.

The remaining parts of this paper are organized as follows. In Section 2, we rewrite the ROC regression into a grouped variable selection form so that current criteria can be applied. Then, a general two-stage framework with a BIC selector for the group SCAD under the local model assumption is proposed in Section 3. Simulation studies and a real data analysis are given in Sections 4 and 5. A brief discussion is provided in Section 6. All proofs are presented in the Supplement; see Supplementary Materials available online at http://dx.doi.org/10.1155/2013/436493.

2. ROC Regression

In this section, we rewrite the penalized ROC regression with induced methodology into a problem of the grouped variable selection by SCAD. Initially, we require that all covariates be centered at 0 for the consideration of comparability. Also, for notation simplicity, response variables are centered. If not, we can center responses to finish the model selection and then add centers back to evaluate the AUC. By following notations of the local model, which generalizes the commonly used sparsity assumption, homoscedastic regression models for diseased and nondiseased subjects are assumed as follows:

$$\begin{aligned} y_D &= z^T\theta_D + \sigma_D\varepsilon_D = x^T\beta_{D0} + u^T\gamma_D + \sigma_D\varepsilon_D, \\ y_{\overline{D}} &= z^T\theta_{\overline{D}} + \sigma_{\overline{D}}\varepsilon_{\overline{D}} = x^T\beta_{\overline{D}0} + u^T\gamma_{\overline{D}} + \sigma_{\overline{D}}\varepsilon_{\overline{D}}, \end{aligned} \tag{1}$$

where x includes p variables added always, u includes q variables which may or may not be added, $z = (x^T, u^T)^T$, β_{D0} and $\beta_{\overline{D}0}$ are p dimensional vectors, $\gamma_D = \delta_D/\sqrt{n_D}$ and $\gamma_{\overline{D}} = \delta_{\overline{D}}/\sqrt{n_{\overline{D}}}$ are q dimensional vectors with n_D and $n_{\overline{D}}$ as sample sizes for diseased and nondiseased groups, respectively, $\theta_D = (\beta_{D0}^T, \gamma_D^T)^T = (\theta_{D1}, \ldots, \theta_{Dd})^T$ and $\theta_{\overline{D}} = (\theta_{\overline{D}1}, \ldots, \theta_{\overline{D}d})^T$ are $d \triangleq p + q$ dimensional vectors, and ε_D and $\varepsilon_{\overline{D}}$ independently follow $\mathcal{N}(0, 1)$. Especially, if $\delta_D = \delta_{\overline{D}} = \mathbf{0}_q$, a sparse model is given. Then, the AUC given z can be written as

$$\mathrm{AUC}_z = \Pr\left(y_D \geq y_{\overline{D}} \mid z\right) = \Phi\left(\frac{z^T\left(\theta_D - \theta_{\overline{D}}\right)}{\sqrt{\sigma_{\overline{D}}^2 + \sigma_D^2}}\right), \tag{2}$$

where $\Phi(\cdot)$ is the cumulative distribution function of a standard normal distribution. Clearly, the narrow model is $\mathcal{S}_0 = \{1, \ldots, p\}$, including all constant effects β_{D0} and $\beta_{\overline{D}0}$. More details of the local model assumption are provided in the following section.

Assume that observed i.i.d that the samples are $\{(y_{Di}, z_{Di})\}$, $i = 1, \ldots, n_D$, and $\{(y_{\overline{D}j}, z_{\overline{D}j})\}$, $j = 1, \ldots, n_{\overline{D}}$. Instead of selecting separate models, we consider the following single objective function with a group penalty, given a tuning parameter λ:

$$\begin{aligned} Q_\lambda\left(\theta_D, \theta_{\overline{D}}\right) = {} & \frac{1}{2n\sigma_D^2}\sum_{i=1}^{n_D}\left(y_{Di} - z_{Di}^T\theta_D\right)^2 \\ & + \frac{1}{2n\sigma_{\overline{D}}^2}\sum_{j=1}^{n_{\overline{D}}}\left(y_{\overline{D}j} - z_{\overline{D}j}^T\theta_{\overline{D}}\right)^2 \\ & + \lambda\sum_{s=1}^{d} p_\lambda\left(\|\boldsymbol{\theta}_s\|\right), \end{aligned} \tag{3}$$

where $\boldsymbol{\theta}_s = (\theta_{\overline{D}s}, \theta_{Ds})^T$, a 2-dimensional vector, with the sth component θ_{Ds} of θ_D and the sth component $\theta_{\overline{D}s}$ of $\theta_{\overline{D}}$, and $\partial p_\lambda(w)/\partial w = \lambda I(w \leq \lambda) + \max(0, a\lambda - w)I(w > \lambda)/(a-1)$ with $a = 3.7$. More generally, instead of the L_2 norm for $\boldsymbol{\theta}_s$, we can define $\|\boldsymbol{\theta}_s\|_{K_s} = \sqrt{\boldsymbol{\theta}_s^T K_s \boldsymbol{\theta}_s}$ with a positive definite 2×2 matrix W_s. Then, given λ, the minimizer of (3) can be obtained as an estimate of $(\theta_D^T, \theta_{\overline{D}}^T)^T$. The motivation of considering such a penalty on $\boldsymbol{\theta}_s$ jointly rather than separately is that the inclusion or exclusion of the effect of a certain variable should be simultaneous for both diseased and nondiseased groups. It may not be appropriate to include either θ_{Dk} or $\theta_{\overline{D}k}$ in the model only, which will bring troubles in interpretation of the resulting model. This is exactly the motivation of the group LASSO method by Yuan and Lin [17] to handle categorical variables, and the group SCAD by Wang et al. [18] to address spline bases.

Note that there are two separate summations of residual squares in (3). In order to comply with the framework of selecting grouped variables, a modified version of the objective function (3) is required. Let $\otimes$ be the Kronecker product operator. Define $\boldsymbol{\theta} = \theta_{\overline{D}} \otimes (1, 0)^T + \theta_D \otimes (0, 1)^T$,

$\mathbf{z}_{\overline{D}j} = z_{\overline{D}j} \otimes (1,0)^T$, $j = 1,\ldots,n_{\overline{D}}$, and $\mathbf{z}_{Di} = z_{Di} \otimes (0,1)^T$, $i = 1,\ldots,n_D$. In matrix form, we have

$$Y = \left(y_{\overline{D}1},\ldots,y_{\overline{D}n_{\overline{D}}},y_{D1},\ldots,y_{Dn_D}\right)^T,$$
$$Z = \left(\mathbf{z}_{\overline{D}1},\ldots,\mathbf{z}_{\overline{D}n_{\overline{D}}},\mathbf{z}_{D1},\ldots,\mathbf{z}_{Dn_D}\right)^T, \quad (4)$$

where Y is an $n \triangleq n_{\overline{D}} + n_D$ dimensional vector with components y_i, $i = 1,\ldots,n$, and Z is an $n \times 2d$ dimensional matrix. Clearly, there are d grouped variables, and Z can be split into d submatrices $Z = (Z_l,\ldots,Z_d)$, each of which includes two consecutive columns of Z in turn. Similarly, $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^T,\ldots,\boldsymbol{\theta}_d^T)^T$ with $\boldsymbol{\theta}_m = (\theta_{\overline{D}m},\theta_{Dm})^T$, $m = 1,\ldots,d$. Additionally, due to different variances of healthy and diseased subjects, weighted least squares should be applied. Let W be a diagonal matrix, with each diagonal entry

$$W_{ii} = \begin{cases} \sigma_{\overline{D}}^{-2} & \text{if } i = 1,\ldots,n_{\overline{D}}, \\ \sigma_D^{-2} & \text{if } i = n_{\overline{D}}+1,\ldots,n. \end{cases} \quad (5)$$

Then, the objective function (3) is written as

$$Q_\lambda(\boldsymbol{\theta}) = \frac{1}{2n}\left\|Y - \sum_{m=1}^{d} Z_m\boldsymbol{\theta}_m\right\|_W^2 + \lambda\sum_{m=1}^{d} p_\lambda\left(\|\boldsymbol{\theta}_m\|\right). \quad (6)$$

Furthermore, in order to facilitate computation with current R packages, we would define transformed observations by weighting. Simply, put $\widetilde{Y} = W^{1/2}Y$ and $\widetilde{Z}_m = W^{1/2}Z_m$. Therefore,

$$Q_\lambda(\boldsymbol{\theta}) = \frac{1}{2n}\left\|\widetilde{Y} - \sum_{m=1}^{d} \widetilde{Z}_m\boldsymbol{\theta}_m\right\|^2 + \lambda\sum_{m=1}^{d} p_\lambda\left(\|\boldsymbol{\theta}_m\|\right). \quad (7)$$

Finally, the penalized ROC regression (3) has been written into a group SCAD-type problem (7). Then, current model selection criteria, like CV, GCV, AIC, and BIC, can be applied to select a final model. For this specific ROC regression problem, where AUC is the focus, these criteria may not be appropriate. Therefore, as argued by Claeskens and Hjort [15], the FIC can play a role here.

Under the local model assumption, a novel procedure of applying the FIC to the grouped variable selection is developed, which is motivated by Wang and Fang [16]. Briefly speaking, the procedure consists of two steps. Firstly, a narrow model, containing variables added always, is identified through the objective function (7). Secondly, the FIC is applied to select a subgroup of remaining variables. As a consequence, the final model is the combination of variables selected in both two steps. Details are provided in the following section. In terms of FIC, naturally, the focus parameter is the AUC at a given z_0; that is, $\mu(\boldsymbol{\theta}) = \Phi((z_0^T \otimes (-1,1))\boldsymbol{\theta}/\sqrt{\sigma_{\overline{D}}^2 + \sigma_D^2})$ with $\partial\mu/\partial\boldsymbol{\theta} = \phi((z_0^T \otimes (-1,1))\boldsymbol{\theta}/\sqrt{\sigma_{\overline{D}}^2 + \sigma_D^2})(z_0 \otimes (-1,1)^T/\sqrt{\sigma_{\overline{D}}^2 + \sigma_D^2})$.

Later, in simulation studies, the separate variable selection for diseased and nondiseased models will also be utilized to make a comparison. We expect, the group selection is superior to the separate selection.

3. A BIC Selector for Group SCAD under the Local Model Assumption

This section follows notations used in the two fundamental papers of the FIC: Hjort and Claeskens [19] and Claeskens and Hjort [15]. Furthermore, we allow grouped variables, each of which stands for a factor, such as a series of dummy variables coded from a multilevel categorical variable. The starting assumption of the FIC is that some variables are added to the regression model always and the others may or may not be added; that is,

$$y_i = x_i^T\beta_0 + u_i^T\gamma + \varepsilon_i, \quad i = 1,\ldots,n, \quad (8)$$

where x_i includes p variables which are added always, u_i includes q variables which may or may not be added, and $\varepsilon_i \sim \mathcal{N}(0,\sigma_\varepsilon^2)$. Without loss of generality, both x_i and y_i are standardized to remove the intercept term. Furthermore, we assume that x_i actually consists of K factors, that is, $x_i = (x_{i1}^T,\ldots,x_{iK}^T)^T$, and the corresponding $\beta_0 = (\beta_{01}^T,\ldots,\beta_{0K}^T)^T$, with dimensions p_k for each x_{ik} and β_{0k}, $k = 1,\ldots,K$, such that $\sum_{k=1}^K p_k = p$. Similarly, u_i consists of L factors, that is, $u_i = (u_{i1}^T,\ldots,u_{iL}^T)^T$, and the corresponding $\gamma = (\gamma_1^T,\ldots,\gamma_L^T)^T$, with dimensions q_l for each u_{il} and γ_l, $l = 1,\ldots,L$, such that $\sum_{l=1}^L q_l = q$. Let $z_i^T = (x_i^T,u_i^T) = (z_{i1}^T,\ldots,z_{iM}^T)$, with $d \triangleq p+q$ dimensions, and each z_{im} has d_m dimensions, $m = 1,\ldots,M$, such that $M \triangleq K+L$ and $\sum_{m=1}^M d_m = d$. Let $Y = (y_1,\ldots,y_n)^T$, $X = (x_1,\ldots,x_n)^T$, $U = (u_1,\ldots,u_n)^T$, and $Z = (z_1,\ldots,z_n)^T$. For simplicity, assume that the residual variance σ_ε^2 is estimated based on the full model and is not considered as a parameter.

In the literature of the variable selection, in order to show the selection consistency of a variable selection procedure, usually, the true model is assumed to be sparse. Thus, the sparsity assumption plays a critical role in the current model selection literature. Many procedures have been shown to be selection consistent under this sparsity assumption [20]. For example, the SCAD with tuning parameter selected via BIC has been shown to be selection consistent by Wang et al. [21, 22], and Zhang et al. [23].

However, it is questionable or too strict to assume that the true model is sparse. It is more reasonable and flexible to consider the local model (8) with $\theta_{\text{true}}^T = (\beta_0^T,\gamma^T)$ and $\gamma = \gamma_0 + \delta/\sqrt{n}$ as a true model, where $\gamma_0 = \mathbf{0}_q$ for the purpose of variable selection, under which the FIC is developed. This model is close to the sparse model, but it is different from it by $\gamma - \mathbf{0}_q = \delta/\sqrt{n}$. The sparsity assumption, with notations in this paper, is equivalent to assume that $\delta = \mathbf{0}_q$ and $\theta_{\text{true}}^T = (\beta_0^T,\mathbf{0}_q^T)$. Therefore, the local model assumption used here is a natural extension of the sparsity assumption. All "consistency" results obtained in this paper still apply to sparse models with grouped variables.

The FIC centers at the inference on a certain estimand or focus, denoted by $\mu_{\text{true}} = \mu(\theta_{\text{true}})$. It is well known that using a bigger model would typically mean smaller bias but bigger variance. Therefore, the FIC tries to balance the bias and the variance of estimating a certain parameter estimand. To be specific, like what any existing criterion does, among

a possible model range, the FIC starts with a narrow model that includes only variables in x_i and searches over submodels including some factors in u_i. The whole process leads to totally 2^L submodels, one for each subset of $\{1, \ldots, L\}$.

In this framework, various estimators of the focus parameter range from $\widehat{\mu}_{\text{full}} = \mu(\widehat{\beta}_{\text{full}}, \widehat{\gamma}_{\text{full}})$ to $\widehat{\mu}_{\text{narr}} = \mu(\widehat{\beta}_{\text{narr}}, \mathbf{0}_q)$. In general, the FIC attempts to select a subset $\widehat{\mathcal{S}}$ associated with the smallest mean squared error (MSE) of $\widehat{\mu}_{\mathcal{S}} = \mu(\widehat{\beta}_{\mathcal{S}}, \widehat{\gamma}_{\mathcal{S}}, \gamma_{0,\mathcal{S}^c})$, where $\mathcal{S}^c$ is the complement of $\mathcal{S}$ and the subscript $\mathcal{S}$ means a subset of corresponding vectors indexed by $\mathcal{S}$.

3.1. Stage 1: Consistent Selection of the Narrow Model. Once assuming the true model (8) with $\theta_{\text{true}}^T = (\beta_0^T, \gamma^T)$ and $\gamma = \delta/\sqrt{n}$ as well as grouped variables, here arises the first important question regarding whether we can select the narrow model $\mathcal{S}_0 = \{1, \ldots, K\}$ consistently. A similar question has been addressed by Wang and Fang [16], where they considered nongrouped variables. In the following, we show that the group SCAD with a tuning parameter selected via BIC can consistently select the narrow model.

Wang et al. [18] extended the SCAD, proposed by Fan and Li [11], to grouped variables and established its oracle property, following an elegant idea of the group LASSO [17]. The group SCAD generates an estimate via following penalized least squares:

$$\widehat{\theta}_\lambda = \underset{\theta}{\operatorname{argmin}} \left\{ \frac{1}{2n} \|Y - Z\theta\|^2 + \sum_{m=1}^{M} p_\lambda \left(\|\theta_m\| \right) \right\}, \quad (9)$$

where $\theta = (\theta_1^T, \ldots, \theta_M^T)^T$ with d_m-dimensional θ_m, and $p_\lambda(\cdot)$ is defined in the previous section. Let $\widehat{\mathcal{S}}_\lambda = \{m : \widehat{\theta}_{\lambda m} \neq \mathbf{0}_{d_m}\}$ be the selected narrow model for a given λ. With similar arguments in the previous section, the L_2 norm used in the penalty can be replaced by any metric with the form $\|\theta_m\|_{K_m} \triangleq (\theta_m^T K_m \theta_m)^{1/2}$ such that K_m is a symmetric $d_m \times d_m$ positive definite matrix.

Under the local model assumption with no grouped variables, Wang and Fang [16] showed that, with a tuning parameter λ selected via BIC, the SCAD is selection consistent; that is, with probability tending to one, the narrow model can be identified. Similarly, a BIC selector can be defined based on the group SCAD as follows:

$$\widehat{\lambda}_{\text{B}} = \underset{\lambda}{\operatorname{argmin}} \left\{ \log\left(\widehat{\sigma}_\lambda^2\right) + \frac{\text{df}_\lambda \log(n)}{n} \right\}, \quad (10)$$

where $\widehat{\sigma}_\lambda^2 = \|Y - Z\widehat{\theta}_\lambda\|^2/n$ and $\text{df}_\lambda = \sum_{m \in \widehat{\mathcal{S}}_\lambda} d_m$. We expect that the group SCAD is still selection consistent in the sense that $\Pr(\widehat{\mathcal{S}}_{\widehat{\lambda}_{\text{B}}} = \mathcal{S}_0) \to 1$ as $n \to \infty$, provided that $\mathcal{S}_0$ is the narrow model.

Formally, within the framework of FIC, assuming that the local model (8) is the true model and that $\mathcal{S}_0$ is the narrow model, we show the following theorem. Proofs can be found in the Supplement.

Theorem 1. *Under some mild conditions (see the Supplement for details), one has that*

$$\Pr\left(\widehat{\mathcal{S}}_{\widehat{\lambda}_B} = \mathcal{S}_0\right) \longrightarrow 1, \quad \text{as } n \longrightarrow \infty, \quad (11)$$

provided that model (8) *with* $\theta_{true} = (\beta_0^T, \gamma^T)^T$ *and* $\gamma = \delta/\sqrt{n}$ *is the true model.*

Remark 2. If we assume that $\delta = \mathbf{0}_q$, that is, the model is sparse, then Theorem 1 provides a BIC selector for the tuning parameter in the group SCAD, which can consistently identify nonzero effects. In other words, we extend the BIC selector for the SCAD proposed by Wang et al. [21] to the situation with the group SCAD.

Theorem 1 also implies both advantages and disadvantages of the BIC, which have been discussed by Wang and Fang [16]. Briefly speaking, the BIC sacrifices prediction consistency [24] in the sense of filtering all of the variables whose effect sizes are of order $O(1/\sqrt{n})$ to achieve the model selection consistency. The previous theorem provides a data-driven method to consistently specify a narrow model, which is critical before applying FIC. In the following subsection, we suggest a two-stage framework to apply the FIC based upon a narrow model selected via the BIC, in order to recover part of the variables filtered by the BIC.

3.2. Stage 2: FIC. In Stage 1, a narrow model, $\widehat{\mathcal{S}}_0 = \{1, \ldots, \widehat{K}\}$, has been identified via the group SCAD with a tuning parameter selected via BIC. In Stage 2, any subset of $\widehat{\mathcal{S}}_0^c = \{\widehat{K}+1, \ldots, M = \widehat{K} + \widehat{L}\}$ can be added to $\widehat{\mathcal{S}}_0$. A direct application of the FIC proposed by Claeskens and Hjort [15] is not plausible even for moderate size of $\widehat{L}$, because there are $2^{\widehat{L}}$ subsets of $\widehat{\mathcal{S}}_0^c$. Furthermore, the best-subset selection is unstable [13]. Therefore, similar to Wang and Fang [16], without double minimizations through both subsets and tuning parameters proposed by Claeskens [25], we suggest limiting the search domain to those subsets on the solution path from any group regularization procedure such as group LASSO or group SCAD.

With a selected narrow model $\widehat{\mathcal{S}}_0 = \{1, \ldots, \widehat{K}\}$, let $\widetilde{x}_i = (z_{i1}^T, \ldots, z_{i\widehat{K}}^T)^T$, $\widetilde{u}_i = (z_{i,\widehat{K}+1}^T, \ldots, z_{iM}^T)^T$, $\widetilde{\beta} = (\theta_1^T, \ldots, \theta_{\widehat{K}}^T)^T$, and $\widetilde{\gamma} = (\theta_{\widehat{K}+1}^T, \ldots, \theta_M^T)^T$. Then, a solution path is generated from the following group LASSO procedure (or group SCAD):

$$\left(\widehat{\beta}_\tau, \widehat{\gamma}_\tau\right) = \underset{\widetilde{\beta}, \widetilde{\gamma}}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} \frac{\left(y_i - \widetilde{x}_i^T \widetilde{\beta} - \widetilde{u}_i^T \widetilde{\gamma}\right)^2}{2n} + \tau \sum_{l=1}^{\widehat{L}} \|\widetilde{\gamma}_l\| \right\}, \quad (12)$$

where the tuning parameter τ controls the grouped variables included in the subset $\widehat{\mathcal{A}}_\tau = \{l : \widehat{\gamma}_{\tau l} \neq \mathbf{0}_{q_l}\}$. As the tuning parameter τ varies from some large value to 0, $\widehat{\mathcal{A}}_\tau$ increases from an empty set to a "full" set $\{1, \ldots, \widehat{L}\}$. Then, we utilize the FIC to guide the selection of τ in (12) over the resulting $\widehat{\mathcal{A}}_\tau$'s, which consist of a search domain.

Now, Stage 2 of the FIC for a certain focus $\mu_{\text{true}} = \mu(\theta_{\text{true}})$ is summarized as follows. For a given τ, a subset $\widehat{\mathcal{A}}_\tau$ is provided by indices of nonzero factors from (12). Then, based on

the submodel $\mathcal{S} = \widehat{\mathcal{S}}_0 \cup \widehat{\mathcal{A}}_\tau$, the FIC_τ is evaluated according to a formula developed in Claeskens and Hjort [15, formula (3.3)], which is essentially a parametric estimate of the MSE of μ_{true} on a model $\mathcal{S}$. Consequently, τ is selected as

$$\widehat{\tau}_{\mathrm{F}} = \underset{\tau}{\operatorname{argmin}}\, \mathrm{FIC}_\tau, \tag{13}$$

and the final submodel is selected as $\widehat{\mathcal{S}}_{\mathrm{F}} = \widehat{\mathcal{S}}_0 \cup \widehat{\mathcal{A}}_{\widehat{\tau}_{\mathrm{F}}}$.

4. Simulation

Simulated data are generated under models (1) with 0 as intercepts. Moderate sample sizes are set to $n_{\overline{D}} = 50$ and $n_D = 50$, compared with 8 and 20 as numbers of covariates. Three scenarios of parameters are considered in the following:

(1) $\sigma_{\overline{D}} = \sigma_D = 2$, $\beta_D = (1.5, 2, 3)$, $\gamma_{Dj} = (3 - 0.5(j-1))/\sqrt{n_D}$, $j = 1, \ldots, 5$, $\theta_D = (\beta_D^T, \gamma_D^T)^T$, $\beta_{\overline{D}} = (0.5, 1, 2)$, $\gamma_{\overline{D}j} = (1 - 0.2(j-1))/\sqrt{n_{\overline{D}}}$, $j = 1, \ldots, 5$, $\theta_{\overline{D}} = (\beta_{\overline{D}}^T, \gamma_{\overline{D}}^T)^T$, $p = 3$, $q = 5$, $d = 8$;

(2) $\sigma_{\overline{D}} = \sigma_D = 2$, $\beta_D = (1.5, 2, 3)$, $\gamma_{Dj} = (2 - 0.05(j-1))/\sqrt{n_D}$, $j = 1, \ldots, 17$, $\theta_D = (\beta_D^T, \gamma_D^T)^T$, $\beta_{\overline{D}} = (0.5, 1, 2)$, $\gamma_{\overline{D}j} = (1 - 0.05(j-1))/\sqrt{n_{\overline{D}}}$, $j = 1, \ldots, 17$, $\theta_{\overline{D}} = (\beta_{\overline{D}}^T, \gamma_{\overline{D}}^T)^T$, $p = 3$, $q = 17$, $d = 20$;

(3) $\sigma_{\overline{D}} = \sigma_D = 1$, $\theta_{Dj} = 3/2j$, $\theta_{\overline{D}j} = 2/2j$, $j = 1, \ldots, 8$, $d = 8$.

Clearly, the narrow model of the first two settings is $\{1, 2, 3\}$, whereas, for the third one, no clear boundary is specified between big effects and small effects.

Corresponding to each setting, test datasets z_{01}, z_{02}, and z_{03} are selected to generate AUC around 0.6, 0.8, and 0.95 to accommodate low-, moderate-, and high-accuracy cases, respectively. Consider the following:

(1) $z_{01} = (0.2, \ldots, 0.2)^T$, AUC = 0.611; $z_{02} = (0.7, \ldots, 0.7)^T$, AUC = 0.838; $z_{03} = (1.2, \ldots, 1.2)^T$, AUC = 0.955;

(2) $z_{01} = (0.15, \ldots, 0.15)^T$, AUC = 0.613; $z_{02} = (0.45, \ldots, 0.45)^T$, AUC = 0.805; $z_{03} = (0.9, \ldots, 0.9)^T$, AUC = 0.957;

(3) $z_{01} = (0.3, \ldots, 0.3)^T$, AUC = 0.613; $z_{02} = (0.9, \ldots, 0.9)^T$, AUC = 0.806; $z_{03} = (1.8, \ldots, 1.8)^T$, AUC = 0.958.

Besides the proposed two-stage framework (FIC) with group SCAD, for comparison purpose, four popular variable selection criteria, including 5-fold CV, GCV, AIC, and BIC, are also employed. Additionally, the SCAD penalty is applied to diseased and healthy groups separately to show the gain of applying the group SCAD.

Two popular measurements, $\mathrm{MSE} = E(\mu(\widehat{\theta}_{\widehat{\mathcal{S}}_{\mathrm{F}}}) - \mu(\theta_{\text{true}}))^2$ and the mean absolute error (MAE), defined by $E|\mu(\widehat{\theta}_{\widehat{\mathcal{S}}_{\mathrm{F}}}) - \mu(\theta_{\text{true}})|$, are utilized to evaluate the prediction performance of selected models based on different criteria, where $\widehat{\theta}_{\widehat{\mathcal{S}}_{\mathrm{F}}}$ is an estimate of θ based on the final model $\widehat{\mathcal{S}}_{\mathrm{F}}$ selected by a certain selection criterion. Due to the limited range of AUC and skewed distributions of estimates of AUC especially at boundaries, the MAE is supposed to be more appropriate.

Table 1: Model selection performance for group SCAD.

Setting	Method	F-measure (%)
1	CV	71.3
	GCV	72.8
	AIC	70.9
	BIC	77.4
2	CV	66.0
	GCV	66.0
	AIC	66.2
	BIC	67.8

In this paper, a composite measurement, the F-measure, is employed to evaluate the performance of selecting the narrow model among various methods, including commonly used proportions of selecting underfitting, correct, and overfitting models separately. As noted by Lim and Yu [26], a high F-measure means that both false-positive and false-negative rates are low. Define Precision = true positivity, Recall = true discovery and then, F-measure $\triangleq$ (2 · Precision · Recall)/(Precision+Recall). All results are summarized based on 500 repetitions according to simulation settings in Tables 1, 2, and 3.

Table 1 indicates that the BIC has the best performance to identify the narrow model, compared with others. Also, if there are more weak signals, like Setting 2, the performance is not as good as that of Setting 1. This is reasonable, because, with increasing number of variables given the sample size, it is more challenging to filter weak signals, even under the sparsity assumption. From Table 2, we can see that, in all three settings, these five methods perform well. Specifically, for moderate and large AUC cases, the FIC performs slightly better, providing smaller MAE. Additionally, in these cases, the FIC improves the BIC substantially, which once again indicates that the BIC would filter weak signals.

In order to show how we can benefit from applying the grouped variable selection, separate model selections for diseased and healthy subjects are also considered, and results are summarized in Table 3. By comparing Tables 2 and 3, in most cases, the group penalty provides smaller MSE and MAE for every criterion. Due to limited range of the AUC, all MSE and MAE values in Tables 2 and 3 are small, but the group selection can improve separate selections by as high as 25%. It is not surprising to see that, in high AUC situations, differences are small, and separate selections with BIC are better. Possible reasons are the following: (1) there is no much room for an estimated AUC to vary when it is close to 1; (2) separate selections with BIC offer a larger flexibility to obtain a sparse model.

5. Real Data Analysis

In this section, we demonstrate the proposed procedure by the audiology data reported by Stover et al. [27], which has

Table 2: Prediction of AUC at z_{0k} with group SCAD. Size means the number of selected factors, where each factor contains two variables.

Setting	Methods	z_{01}			z_{02}			z_{03}		
		MSE	MAE	Size	MSE	MAE	Size	MSE	MAE	Size
1	CV	0.00345	0.0467	4.72	0.00295	0.0437	4.72	0.00114	0.0255	4.72
	GCV	0.00345	0.0467	4.49	0.00294	0.0432	4.49	0.00114	0.0251	4.49
	AIC	0.00349	0.0468	4.90	0.00291	0.0428	4.90	0.00110	0.0246	4.90
	BIC	0.00335	0.0461	3.62	0.00317	0.0450	3.62	0.00147	0.0278	3.62
	FIC	0.00339	0.0464	4.23	0.00283	0.0426	4.23	0.00108	0.0247	4.23
2	CV	0.00328	0.0461	8.31	0.00279	0.0428	8.31	0.00073	0.0209	8.31
	GCV	0.00339	0.0470	9.43	0.00281	0.0433	9.43	0.00066	0.0206	9.43
	AIC	0.00344	0.0472	12.05	0.00285	0.0434	12.05	0.00064	0.0204	12.05
	BIC	0.00324	0.0458	6.14	0.00328	0.0462	6.14	0.00129	0.0259	6.14
	FIC	0.00327	0.0459	7.97	0.00290	0.0440	7.97	0.00081	0.0224	7.97
3	CV	0.00369	0.0483	6.67	0.00439	0.0535	6.67	0.00199	0.0317	6.67
	GCV	0.00367	0.0483	6.14	0.00436	0.0533	6.14	0.00197	0.0316	6.14
	AIC	0.00369	0.0484	6.36	0.00441	0.0534	6.36	0.00201	0.0318	6.36
	BIC	0.00367	0.0482	5.14	0.00473	0.0549	5.14	0.00247	0.0345	5.14
	FIC	0.00368	0.0483	5.46	0.00451	0.0532	5.49	0.00219	0.0324	5.49

Table 3: Prediction of AUC at z_{0k} with models on diseased and healthy groups separately. Size means the sum of numbers of selected variables in diseased and non-diseased groups.

Setting	Methods	Size	z_{01}		z_{02}		z_{03}	
			MSE	MAE	MSE	MAE	MSE	MAE
1	CV	8.78	0.00384	0.0490	0.00303	0.0439	0.00107	0.0247
	GCV	8.23	0.00383	0.0488	0.00298	0.0432	0.00103	0.0239
	AIC	8.18	0.00383	0.0488	0.00397	0.0432	0.00102	0.0239
	BIC	6.96	0.00383	0.0490	0.00313	0.0447	0.00114	0.0254
2	CV	14.47	0.00483	0.0566	0.00351	0.0481	0.00079	0.0218
	GCV	16.31	0.00553	0.0609	0.00390	0.0515	0.00069	0.0218
	AIC	15.46	0.00545	0.0606	0.00388	0.0516	0.00069	0.0218
	BIC	12.67	0.00514	0.0590	0.00384	0.0502	0.00092	0.0225
3	CV	12.29	0.00405	0.0494	0.00481	0.0558	0.00225	0.0332
	GCV	10.55	0.00403	0.0497	0.00461	0.0541	0.00208	0.0320
	AIC	10.55	0.00402	0.0497	0.00461	0.0541	0.00209	0.0320
	BIC	9.53	0.00403	0.0499	0.00468	0.0550	0.00206	0.0325

been analyzed by Pepe [3, 28]. The dataset contains results of distortion product otoacoustic emissions (DPOAE) test used to diagnose the hearing impairment. There are 208 subjects who were examined at different combinations of three frequencies (f) and three intensities (L) of the DPOAE device. An audiometric threshold can be obtained for each combination. At a particular frequency, if the audiometric threshold is greater than 20 dB HL, an ear was classified as hearing impaired. In the original dataset, there are multiple records for each subject. In this study, we randomly select one record for each subject, and among 208 subjects there are 55 subjects with hearing impairment. The test result is the negative signal-to-noise ratio, −SNR. The covariates used in Dodd and Pepe [29] are z_f = frequency Hz/100, z_L = intensity dB/10, and z_D = (hearing threshold − 20) dB/10. In order to encourage the model selection, we incorporate two-way interaction terms. Quadratic terms are not included due to the high correlation between each variable and its quadratic term. Therefore, z is the centered $(z_f, z_L, z_D, z_f z_L, z_f z_D, z_L z_D)^T$ for each element.

Former studies on this dataset showed that −SNR provided quite high discriminative performance and that z_f had a small effect. In order to avoid specifying inappropriate covariates, we randomly select three centered observations from the whole dataset as focused subjects.

Table 4 shows AUC values of models selected by each method as well as corresponding model sizes. CV, AIC, and GCV tend to select a full model. On the contrary, BIC tends to select a sparse model, only containing z_D. The full model may not provide the largest AUC, because a large model will bring instability and ruin the AUC. As indicated in the table, for the second test point, both BIC and FIC provide a higher AUC than the full model. But a single variable selected by the BIC seems to be too strict. By focusing on the precision of

Table 4: Estimated AUC at three test points. Size means the number of selected factors.

Methods	Test point 1		Test point 2		Test point 3	
	AUC	Size	AUC	Size	AUC	Size
CV	0.971	6	0.916	6	0.982	6
AIC	0.971	6	0.916	6	0.982	6
GCV	0.971	6	0.916	6	0.982	6
BIC	0.949	1	0.957	1	0.944	1
FIC	0.963	3	0.957	2	0.944	1

estimated focus parameter, the FIC provides a customized way to fill the gap: for the first test point, three main effects are selected; for the second one, z_L and z_D are selected; for the third one, only z_D is selected. Based on the precision of estimated AUC, the FIC performs as a compromise, selecting models to generate AUC values in the middle.

6. Discussion

In this paper, we rewrite the model selection problem of the ROC regression into a grouped factor selection form with induced methodology. Also, we develop a two-stage framework to apply the FIC to select a final model with group SCAD under the local model assumption. Specifically, if the true model is sparse, our framework naturally accommodates current model selection criteria. Furthermore, the BIC selector is proved to be model selection consistent if either a sparse or a local model is assumed, in the sense of selecting a sparse model or a narrow model.

Most current model selection criteria aim at the prediction performance or model selection consistency; thus, in the ROC regression where the AUC is a focus parameter, they may not be appropriate. This observation motivates an application of FIC, which is shown to perform well through simulation studies. Therefore, our method has a potential application in genetic studies, where the number of gene arrays is always large, compared with the sample size.

For the direct methodology, the literature based on generalized estimating equations is prosperous, which is motivated by the range $[0, 1]$ of the AUC, similar to the probability of a binary random variable. Our future work will extend the framework developed here to generalized estimating equations and apply it to the ROC regression with the direct methodology.

As discussed by one referee, it is possible that some coefficients are the same for both Y_D and $Y_{\overline{D}}$. As in (1), modeling them separately will increase the degree of freedom in (3), especially when a large number of genes are covariates. If the shrinkage of a coefficient, which is known a priori to be the same in both diseased and healthy groups, is not necessary, then it is natural for the FIC to include it in the narrow model with a single coefficient. By using the proposed objective function, a fused LASSO type of penalty may be applied to obtain such kind of structure, in addition to the group LASSO/SCAD. Friedman et al. [30] provided a note on the group LASSO and the sparse group LASSO, which could shed light on the question here. It will be also an interesting topic in the future.

Conflict of Interests

There is no conflict of interests regarding the publication of this article.

Acknowledgments

The authors would like to thank Dr. Yixin Fang for his invaluable suggestions and generous support which make this paper publishable. They also thank the editor, the associate editor, and the referees for their valuable comments which led to substantial improvements of this paper.

References

[1] N. R. Wray, J. Yang, M. E. Goddard, and P. M. Visscher, "The genetic interpretation of area under the ROC curve in genomic profiling," *PLoS Genetics*, vol. 6, no. 2, Article ID e1000864, 2010.

[2] D. Bamber, "The area above the ordinal dominance graph and the area below the receiver operating characteristic graph," *Journal of Mathematical Psychology*, vol. 12, no. 4, pp. 387–415, 1975.

[3] M. S. Pepe, *The Statistical Evaluation of Medical Tests for Classification and Prediction*, Oxford University Press, New York, NY, USA, 2003.

[4] X. H. Zhou, N. A. Obuchowski, and D. M. McClish, *Statistical Methods in Diagnostic Medicine*, John Wiley & Sons, New York, NY, USA, 2nd edition, 2011.

[5] M. X. Rodríguez-Álvarez, P. G. Tahoces, C. Cadarso-Suárez, and M. J. Lado, "Comparative study of ROC regression techniques-applications for the computer-aided diagnostic system in breast cancer detection," *Computational Statistics and Data Analysis*, vol. 55, no. 1, pp. 888–902, 2011.

[6] M. Stone, "Cross-validatory choice and assessment of statistical predictions," *Journal of the Royal Statistical Society B*, vol. 36, no. 2, pp. 111–147, 1974.

[7] P. Craven and G. Wahba, "Smoothing noisy data with spline functions—estimating the correct degree of smoothing by the method of generalized cross-validation," *Numerische Mathematik*, vol. 31, no. 4, pp. 377–403, 1979.

[8] H. Akaike, "Information theory and an extension of the maximum likelihood principle," in *Proceedings of the 2nd International Symposium Information Theory*, B. N. Petrov and F. Csaki, Eds., pp. 267–281, Akademia Kiado, Budapest, Hungary, 1973.

[9] G. Schwarz, "Estimating the dimension of a model," *Annals of Statistics*, vol. 6, pp. 461–464, 1978.

[10] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society B*, vol. 58, no. 1, pp. 267–288, 1996.

[11] J. Fan and R. Li, "Variable selection via nonconcave penalized likelihood and its oracle properties," *Journal of the American Statistical Association*, vol. 96, no. 456, pp. 1348–1360, 2001.

[12] H. Zou, "The adaptive lasso and its oracle properties," *Journal of the American Statistical Association*, vol. 101, no. 476, pp. 1418–1429, 2006.

[13] L. Breiman, "Heuristics of instability and stabilization in model selection," *Annals of Statistics*, vol. 24, no. 6, pp. 2350–2383, 1996.

[14] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, Springer, New York, NY, USA, 2009.

[15] G. Claeskens and N. L. Hjort, "The focused information criterion," *Journal of the American Statistical Association*, vol. 98, no. 464, pp. 900–916, 2003.

[16] B. Wang and Y. Fang, "On the focused information criterion for variable selection," submitted.

[17] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society B*, vol. 68, no. 1, pp. 49–67, 2006.

[18] L. Wang, G. Chen, and H. Li, "Group SCAD regression analysis for microarray time course gene expression data," *Bioinformatics*, vol. 23, no. 12, pp. 1486–1494, 2007.

[19] N. L. Hjort and G. Claeskens, "Frequentist model average estimators," *Journal of the American Statistical Association*, vol. 98, no. 464, pp. 879–899, 2003.

[20] P. Bühlmann and S. van de Geer, *Statistics for High-Dimensional Data: Methods, Theory and Applications*, Springer, 2011.

[21] H. Wang, R. Li, and C.-L. Tsai, "Tuning parameter selectors for the smoothly clipped absolute deviation method," *Biometrika*, vol. 94, no. 3, pp. 553–568, 2007.

[22] H. Wang, B. Li, and C. Leng, "Shrinkage tuning parameter selection with a diverging number of parameters," *Journal of the Royal Statistical Society B*, vol. 71, no. 3, pp. 671–683, 2009.

[23] Y. Zhang, R. Li, and C.-L. Tsai, "Regularization parameter selections via generalized information criterion," *Journal of the American Statistical Association*, vol. 105, no. 489, pp. 312–323, 2010.

[24] Y. Yang, "Can the strengths of AIC and BIC be shared? A conflict between model indentification and regression estimation," *Biometrika*, vol. 92, no. 4, pp. 937–950, 2005.

[25] G. Claeskens, "Focused estimation and model averaging with penalization methods: an overview," *Statistica Neerlandica*, vol. 66, no. 3, pp. 272–287, 2012.

[26] C. Lim and B. Yu, "Estimation Stability with Cross Validation (ESCV)," http://arxiv.org/abs/1303.3128.

[27] L. Stover, M. P. Gorga, S. T. Neely, and D. Montoya, "Toward optimizing the clinical utility of distortion product otoacoustic emission measurements," *Journal of the Acoustical Society of America*, vol. 100, no. 2, part 1, pp. 956–967, 1996.

[28] M. S. Pepe, "Three approaches to regression analysis of receiver operating characteristic curves for continuous test results," *Biometrics*, vol. 54, no. 1, pp. 124–135, 1998.

[29] L. E. Dodd and M. S. Pepe, "Semiparametric regression for the area under the receiver operating characteristic curve," *Journal of the American Statistical Association*, vol. 98, no. 462, pp. 409–417, 2003.

[30] J. Friedman, T. Hastie, and R. Tibshirani, *A Note on the Group Lasso and a Sparse Group Lasso*, 2010.