Biometrics 67, 896–905
September 2011
DOI: 10.1111/j.1541-0420.2010.01533.x

Prediction-Based Structured Variable Selection through the Receiver Operating Characteristic Curves

Yuanjia Wang,[1,*] Huaihou Chen,[1] Runze Li,[2] Naihua Duan,[3] and Roberto Lewis-Fernández[4]

[1]Department of Biostatistics, Mailman School of Public Health, Columbia University, New York, New York 10032, U.S.A.

[2]Department of Statistics and The Methodology Center, The Pennsylvania State University at University Park, Pennsylvania 16802-2111, U.S.A.

[3]Department of Psychiatry and Department of Biostatistics, Columbia University, New York, New York 10032, U.S.A.

[4]Department of Psychiatry, New York State Psychiatric Institute, Columbia University, New York, New York 10032, U.S.A.

[*]*email:* yuanjia.wang@columbia.edu

Summary. In many clinical settings, a commonly encountered problem is to assess accuracy of a screening test for early detection of a disease. In these applications, predictive performance of the test is of interest. Variable selection may be useful 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: there are several root questions followed by their stem questions. The stem questions will only be asked after a subject has answered the root question. It is therefore unreasonable to select a model that only contains stem variables but not its root variable. In this work, we propose methods to perform variable selection with structured variables when predictive accuracy of a diagnostic test is the main concern of the analysis. We take a linear combination of individual variables to form a combined test. We then maximize a direct summary measure of the predictive performance of the test, the area under a receiver operating characteristic curve (AUC of an ROC), subject to a penalty function to control for overfitting. Since maximizing empirical AUC of the ROC of a combined test is a complicated nonconvex problem (Pepe, Cai, and Longton, 2006, *Biometrics* **62,** 221–229), we explore the connection between the empirical AUC and a support vector machine (SVM). We cast the problem of maximizing predictive performance of a combined test as a penalized SVM problem and apply a reparametrization to impose the hierarchical structure among variables. We also describe a penalized logistic regression variable selection procedure for structured variables and compare it with the ROC-based approaches. We use simulation studies based on real data to examine performance of the proposed methods. Finally we apply developed methods to design a structured screener to be used in primary care clinics to refer potentially psychotic patients for further specialty diagnostics and treatment.

Key words: Area under the curve; Disease screening; Hierarchical variable selection; ROC curve; Support vector machine.

1. Introduction

Screening tests are applied in many clinical settings for early detection of a disease. The goal of a screening test is to detect a disease condition as early as possible. Subjects screened positive will be referred for more definitive diagnostic tests. Statistical problems arising from such practices include how to assess the accuracy of the test and how to design tests with adequate sensitivity and specificity. In this article, we develop prediction-based structured variable selection procedures in order to develop a new disease screener based on an existing screener.

This work was motivated by the development of an improved psychosis screener described later. A study of Psychosis Screening Questionnaire (PSQ; Bebbington and Nayani, 1995) and other variables as a medical test for psychosis detection in primary care clinics in Latino population was conducted at the New York State Psychiatric Institute (Lewis-Fernández, 2003). The PSQ originally designed in Britain was found to have poor performance in a Latino population (Lewis-Fernández, 2003). The goal of the study is to develop an improved screener to more accurately detect psychosis in low-income Latino primary care patients by selecting variables from the PSQ and other surveys. Based on the score from the newly designed screener, a subject visiting a primary care clinic classified as positive will be referred to a psychiatrist for further diagnosis and treatment while a subject classified as negative will not be referred.

It is important to identify important variables in PSQ and other surveys to construct a new screener. This is a variable selection problem with predicting a gold standard psychosis outcome as the goal. The PSQ is a screener with a hierarchical structure among variables. All the variables are grouped into five domains. In each domain, there is a root variable and several stem variables. The stem questions will only be asked after a subject has answered the root question. Readers interested in questions or items in PSQ can consult Bebbington

and Nayani (1995). The statistical problem is to select root variables and stem variables to best predict the gold standard disease status. If one ignores the hierarchical structure in the PSQ and performs a variable selection procedure treating all variables as unrelated, it is possible that a model containing stem variables but not their root variables will be chosen. However, such a model is not interpretable and therefore not admissible.

Several variable selection procedures for structured predictors have been proposed in the literature. Huang et al. (2009) applied a group bridge penalty to perform both the group level and within-group individual level variable selection. Yuan, Joseph, and Zou (2009) proposed methods for variable selection that obey a certain hierarchical rule by imposing inequality constraints to the selection procedure. Wang et al. (2009) proposed penalized Cox regression analysis for hierarchical variables by reparametrization where the variables are selected at the group level first and then at the within-group individual level, and all variables within a group are treated as exchangeable. Methods aforementioned may not be directly applied to our setting in that since the prediction accuracy of a diagnostic test is of primary concern, the likelihood function may not be the optimal loss function to use in a variable selection procedure because it does not directly relate to the prediction performance of a diagnostic test (Pepe, 2005; Pepe et al., 2006). Moreover, the variables in a group in the PSQ study may not be exchangeable in the sense that there is a leading root variable followed by stem variables.

To assess prediction accuracy of a medical test, true positive and false-negative rates are two popular indices. From a continuous test score Y, one can define a binary test under a threshold c as: $Y \geqslant c$: positive, and $Y < c$: negative. The receiver operating characteristic (ROC) curve is the entire collection of possible true positives and false positives with different thresholds. A summary index of a medical test performance can then be defined as the area under an ROC curve (AUC of an ROC), which is equivalent to the probability that test results from a randomly selected pair of diseased and nondiseased subjects are correctly ordered, that is, the diseased subject has a higher score than the nondiseased subject. Pepe (2005) and Pepe et al. (2006) showed that when using a combined linear test as decision rule, the ROC-based approach may outperform the likelihood-based approach in terms of prediction performance. On the one hand, there may exist variables that have large odds ratios in terms of association, but contribute little in terms of prediction. On the other hand, it is possible that when prediction is of interest, allowing some variables with weaker association to stay in a model may improve prediction accuracy (Pinsky, 2005). Therefore prediction and association are two distinct goals, which deserve to be treated separately.

Although AUC for an ROC curve is a widely used measure of predictive performance, other measures may be useful when a decision has to be made on the threshold of a diagnostic test. For example, Briggs and Zaretzki (2008) proposed the Skill Plot, which is a plot of the estimated skill score versus the threshold and is directly related to a decision maker who must use a diagnostic test. The Skill Plot can allow different weights for false positives and false negatives through a constant representing the loss. In contrast to ROC curves, using Skill Plot one can easily identify the optimal threshold. In the comments to Briggs and Zaretzki (2008), Hand (2008) pointed out that AUC for the ROC curve uses a weighting scheme of the false positives and false negatives derived empirically from the data. However, in practice, the importance of misclassifying a case as noncase or vice versa may not be determined solely by the data itself, instead it may come from external information with practical consideration of the impact of the misclassification. In these situations, measures other than AUC of an ROC curve such as Skill Plot or alternative ROC utility function may be used.

When there are only a few variables involved, Pepe et al. (2006) considered maximizing the empirical AUC of a linear combination of the variables (up to two variables in their application) that do not exhibit hierarchial structure. To be specific, they proposed to maximize the empirical AUC defined in (1) in Section 2.1 by a simple grid search to obtain coefficients of the linear combination. However, since the empirical AUC is not a continuous function, it is difficult to optimize the empirical AUC for a large number of variables and high-dimensional grid search is not computationally feasible. Overfitting or hierarchical structure was not considered in Pepe et al. (2006).

In this article, we develop a new variable selection procedure with hierarchical structure among variables. We use a linear combination of variables as a combined test to construct a screener for early disease detection. Compared with penalized least squares or likelihood methods in the literature, the newly proposed variable selection procedure is to maximize the empirical AUC of an ROC curve subject to a penalty function that controls for over-fitting, which is suitable when prediction accuracy of the combined test is of primary interest. Due to complexity and nonconvexity in maximizing the empirical AUC of an ROC (Pepe et al., 2006), we utilize the connection between the empirical AUC and a support vector machine (SVM; Brefeld and Scheffer, 2005), and cast the problem of maximizing prediction performance of a combined test with hierarchical structure among individual variables as a penalized SVM problem and reparametrize coefficients of the variables to impose the hierarchical rule. As an alternative, a penalized logistic regression variable selection procedure is considered for structured variables and compared with the ROC based approaches. We examine performance of the proposed methods by Monte Carlo simulation studies based on real data. We further illustrate the proposed procedures to design a structured screener to be used in primary care clinics to refer potentially psychotic patients for specialty diagnostics and treatment.

The rest of this article is organized as follows. In Section 2, we proposed a penalized AUC method and a penalized logistic regression approach for hierarchically structured variable selection, and develop an inference procedure through bootstrap. In Section 3, we present our simulation studies. We give an empirical analysis of a real data example in Section 4. In Section 5, we present concluding remarks and discussions.

2. Methods

2.1 *Penalized SVM for Maximizing AUC of an ROC Curve with Unrelated Variables*

When accuracy of a medical test is of interest, in a variable selection procedure it is desirable to maximize an objective function that is directly related to the performance of

prediction or classification of the test: the AUC of an ROC curve. We find a linear combination of variables or individual test scores that can maximize the empirical AUC of an ROC, which is a consistent estimator of the true AUC under suitable conditions (Pepe, 2003). This ROC-based approach is applicable to retrospective designs (e.g., case-control studies) and is nonparametric in the sense that it does not assume a known link function between the outcome and the variables, which is in contrast to a generalized linear model-based approach.

First consider the simple case where the predictor variables are not structured. Denoted by D the disease status of a subject diagnosed by a gold standard with one indicating diseased and zero indicating disease free. Let n^+ be the number of subjects with $D = 1$, and n^- the number of subjects with $D = 0$. Denoted by x_{ik} the observation of the kth variable on the ith subject, or the kth individual test score on this subject. We consider the following linear decision rule to form a combined test

$$L_\beta(x_i) = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip},$$

where we will fix $\beta_1 = 1$, and do not include an intercept. This is because the linear decision rules are scale invariant and location-shift invariant in computing the AUC of an ROC curve, therefore a decision rule of $L_\beta(x_i) > c$ is equivalent to a decision rule of $\theta_0 + \theta_1 L_\beta(x_i) > c_1$. Let x_i^+, $i = 1, \ldots, n^+$ be the observations of the diseased subjects, and x_i^-, $i = 1, \ldots, n^-$ be the observations of the disease-free subjects. For a given threshold c, a linear decision rule classifies a subject as diseased when $L_\beta(X) \geqslant c$, and classifies a subject as disease free when $L_\beta(X) < c$. The ROC curve plots the true positive rate versus the false positive rate for all possible c's. The area under an ROC curve denoted by AUC is used to summarize the performance of a diagnostic test. It is equivalent to the probability that in a randomly selected pair of subjects the diseased subject will have the combined score higher than the disease-free subject, that is,

$$AUC(\beta) = \Pr\big(L_\beta\big(x_i^+\big) > L_\beta\big(x_j^-\big)\big).$$

It is evident that this probability can be consistently estimated by its empirical version,

$$\widehat{AUC}(\beta) = \frac{1}{n^+ n^-} \sum_{i=1}^{n^+} \sum_{j=1}^{n^-} \big[I\big(L_\beta\big(x_i^+\big) > L_\beta\big(x_j^-\big)\big) + 0.5 I\big(L_\beta\big(x_i^+\big) = L_\beta\big(x_j^-\big)\big)\big], \quad (1)$$

where $I(\cdot)$ is an indicator function. When the variables used to construct the combined test are given, the coefficients for the combined test can be obtained by

$$\widehat{\beta} = \arg\max_\beta \widehat{AUC}(\beta).$$

The above estimator was shown to be consistent and asymptotically normal by relating to a maximum rank correlation estimator with given predictors (Han, 1987). In many applications, however, it is unknown which variables should be chosen to construct a combined test. To select variables that contribute to the prediction while controlling for over-fitting when there is a large number of variables, we penalize the objective function (1) and solve for the coefficients β by

$$\hat{\beta} = \arg\max_\beta \left[\widehat{AUC}(\beta) - \sum_{k=1}^{p} p_\lambda(|\beta_k|) \right], \quad (2)$$

where $\lambda > 0$ is a tuning parameter and $p_\lambda(\cdot)$ is a penalty function such as ridge penalty or SCAD function (Fan and Li, 2001). The tuning parameter can be selected by a data-driven procedure such as generalized approximate cross-validation (Wahba, Lin, and Zhang, 2000).

Unfortunately maximizing the empirical AUC in (1) is a difficult nonconvex problem (Pepe et al., 2006) and the indicator function in the above objective function is not differentiable. Ma and Huang (2005, 2007) used a sigmoid function to approximate an indicator function in computing the empirical AUC. Here, we introduce a support vector machine (SVM)-based approach to approximate the empirical AUC of an ROC curve and cast the optimization problem (2) as a penalized SVM problem. To see this connection, we first briefly review the regular SVM. An SVM is used to perform classification by constructing a hyperplane that optimally separates the data into two categories. To be specific, a regular SVM with a linear decision rule is to solve

$$\begin{aligned} &\arg\min_\beta \left[\frac{C}{2} \sum_{i=1}^{n} \varepsilon_i + \frac{1}{2} ||\beta||^2 \right] \\ &\text{subject to} \quad y_i(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}) \geqslant 1 - \varepsilon_i, \\ &\qquad\qquad \varepsilon_i \geqslant 0, \quad i = 1, \ldots, n. \end{aligned} \quad (3)$$

The above problem is equivalent to penalizing a hinge loss function subject to an L_2 penalty (Hastie, Tibshirani, and Friedman, 2001, p. 380), that is,

$$\arg\min_\beta \sum_{i=1}^{n} [1 - y_i(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})]_+ + \lambda ||\beta||^2, \quad (4)$$

where the subscript $_+$ denotes the positive part of a function. Zhang et al. (2006) proposed to replace the L_2 penalty in the objective function in (4) with other penalty functions such as the L_1 penalty and the SCAD penalty to achieve variable selection with SVM.

Brefeld and Scheffer (2005) showed that an approximation of the solution to (2) can be found through the following support vector machine allowing for margins between x_i^+ and x_j^-:

$$\begin{aligned} &\arg\min_\beta \left[\frac{1}{n^+ n^-} \sum_{i=1}^{n^+} \sum_{j=1}^{n^-} \varepsilon_{ij} + \sum_{k=1}^{p} p_\lambda(|\beta_k|) \right] \\ &\text{subject to} \quad L_\beta(x_i^+) - L_\beta(x_j^-) \geqslant 1 - \varepsilon_{ij}, \\ &i = 1, \ldots, n^+, \quad j = 1, \ldots, n^-, \\ &\varepsilon_{ij} \geqslant 0, i = 1, \ldots, n^+, \quad j = 1, \ldots, n^-. \end{aligned} \quad (5)$$

To connect the two optimization problems, note that from the constraint $L_\beta(x_i^+) - L_\beta(x_j^-) \geqslant 1 - \varepsilon_{ij}$ it follows that $\varepsilon_{ij} \geqslant 1$ when $L_\beta(x_i^+) - L_\beta(x_j^-) \leqslant 0$. Since $\varepsilon_{ij} \geqslant 0$, under the constraints we have that $\sum_{i=1}^{n^+} \sum_{j=1}^{n^-} \varepsilon_{ij}$ is greater than or equal to the number of pairs that satisfy $L_\beta(x_i^+) - L_\beta(x_j^-) \leqslant 0$, which is also the number of pairs that violates $L_\beta(x_i^+) \geqslant L_\beta(x_j^-)$. Consequently, by solving the SVM problem (5), one minimizes the number of pairs that violates the correct rank

order in computing an empirical AUC. Therefore the algorithm corresponds to a minimax rule that finds a model with maximum lower bound of the empirical AUC subject to a penalty function $p_\lambda(\cdot)$ (see Brefeld and Scheffer, 2005).

Comparing (5) with (3), we see that the optimization problem (5), referred to as penalized ROC-SVM throughout this article, is equivalent to a regular SVM (3) with input variables $w_{ij} = x_i^+ - x_j^-$ and outcome variables $y_{ij} = 1$ (Brefeld and Scheffer, 2005). By this equivalence, the geometric interpretation of the penalized ROC-SVM with squared penalty is to find the hyperplane that passes through the origin that will maximize the margin from the points $x_i^+ - x_j^-$ to the plane, or equivalently, to ensure that the signed distance from the points $x_i^+ - x_j^-$ are at least $C(1 - \xi_i)$. Since when converting to the regular SVM all the response variables y_{ij} are positive, misclassification occurs when $\beta^T w_{ij} = \beta^T(x_i^+ - x_j^-) < 0$.

The penalty function $p_\lambda(\cdot)$ can be a LASSO (Tibshirani, 1996) or a SCAD function (Fan and Li, 2001) when sparse solution is desirable. The computation of a regular penalized SVM with various penalty functions was described in Zhang et al. (2006) and Becker et al. (2009). The main algorithm involves a Newton linear programming SVM algorithm (NLPSVM; Fung and Mangasarian, 2004). By transforming the penalized ROC-SVM to a regular penalized SVM, these algorithms developed for the latter can be directly used.

2.2 *Handling Structured Variables: SROC-SVM*

In some applications, there is a hierarchical structure among variables being considered with root variables and stem variables. For example, in the PSQ questionnaire, a stem question will only be asked after its root question is asked. For the ith subject, let $x_{i1}, \ldots, x_{id}$ be the ith subject's root variables, and x_{ikj}, $j = 1, \ldots, J_k$, be the same subject's stem variables following the kth root variable. Denoted by $\alpha_1, \ldots, \alpha_d$ and β_{kj} the corresponding coefficients. The combined linear score for a subject is then

$$L_{\alpha,\beta}(x_i) = \sum_{k=1}^{d} \alpha_k x_{ik} + \sum_{k=1}^{d} \sum_{j=1}^{J_k} \beta_{kj} x_{ikj},$$

where we fix $\alpha_1 = 1$ due to the scale invariance of a linear decision rule in computing the empirical AUC.

To implement the group structure, we use a similar strategy as in Wang et al. (2009). To be specific, we apply a reparametrization that will enforce the hierarchical structure of the variables, that is, we let

$$\beta_{kj} = \alpha_k \gamma_{kj}, \quad k = 1, \ldots, d, \quad j = 1, \ldots, J_k. \tag{6}$$

In the reparametrization (6), γ_{kj} measures the deviation of each stem variable from its root variable. On the one hand, the coefficients for the stem variables will be nonzero if $\alpha_k \neq 0$ and $\gamma_{kj} \neq 0$. In other words, whenever a stem variable is selected to enter a model, its root variable will also be selected because $\alpha_k \neq 0$. On the other hand, when a root variable is selected ($\alpha_k \neq 0$), its stem variables still have the flexibility of dropping out of the model by having $\gamma_{kj} = 0$. Therefore, the reparametrization (6) allows for variable selection both at the root level and at the stem level. The linear scoring system with the reparametrized parameters is now

$$L_{\alpha,\gamma}(x_i) = \sum_{k=1}^{d} \alpha_k x_{ik} + \sum_{k=1}^{d} \sum_{j=1}^{J_k} \alpha_k \gamma_{kj} x_{ikj},$$

where we again fix $\alpha_1 = 1$. The variable with the coefficient one is a baseline variable and the coefficients of the other variables are relative to this variable. The hierarchical penalized ROC-SVM is to solve

$$\begin{aligned} \arg\min_{\alpha,\gamma} \Bigg[& \frac{1}{n^+ n^-} \sum_{i=1}^{n^+} \sum_{j=1}^{n^-} \varepsilon_{ij} + \sum_{k=1}^{d} p_{\lambda_1}(|\alpha_k|) \\ & + \sum_{k=1}^{d} \sum_{l=1}^{J_k} p_{\lambda_2}(|\gamma_{kl}|) \Bigg] \\ \text{subject to } & L_{\alpha,\gamma}(x_i^+) - L_{\alpha,\gamma}(x_j^-) \geqslant 1 - \varepsilon_{ij}, \\ & i = 1, \ldots, n^+, \quad j = 1, \ldots, n^-, \\ & \varepsilon_{ij} \geqslant 0, \quad i = 1, \ldots, n^+, \quad j = 1, \cdots, n^-, \end{aligned}$$

where $\lambda_1 > 0$ and $\lambda_2 > 0$ are tuning parameters. Here we penalize the root variables and the stem variables separately to allow for flexibility. We solve this optimization problem by the following iterative procedure to obtain the structured ROC-SVM estimates, denoted as SROC-SVM:

(1) Given α_k, $k = 1, \ldots, d$, and define $\tilde{x}_{ijk} = \alpha_k x_{ikj}$, and $\widetilde{L}_\gamma(x_i) = \sum_{k=1}^{d} \sum_{j=1}^{J_k} \gamma_{kj} \tilde{x}_{ikj}$, where we fix $\gamma_{11} = 1$. Solve for γ by

$$\begin{aligned} \arg\min_{\gamma} \Bigg[& \frac{1}{n^+ n^-} \sum_{i=1}^{n^+} \sum_{j=1}^{n^-} \varepsilon_{ij} + \sum_{k=1}^{d} \sum_{l=1}^{J_k} p_{\lambda_2}(|\gamma_{kl}|) \Bigg] \\ \text{subject to } & \widetilde{L}_\gamma(x_i^+) - \widetilde{L}_\gamma(x_j^-) \geqslant 1 - \varepsilon_{ij}, \\ & i = 1, \ldots, n^+, \quad j = 1, \ldots, n^-, \\ & \varepsilon_{ij} \geqslant 0, \quad i = 1, \ldots, n^+, \quad j = 1, \ldots, n^-, \end{aligned}$$

which is a penalized SVM problem.

(2) Given γ_{kj}, $k = 1, \ldots, d$, $j = 1, \ldots, J_k$, and define $\tilde{x}_{ik} = x_{ik} + \sum_{j=1}^{J_k} \gamma_{kj} x_{ikj}$, and $\widetilde{L}_\alpha(x_i) = \sum_{k=1}^{d} \alpha_k \tilde{x}_{ik}$, where $\alpha_1 = 1$. Solve for α by

$$\begin{aligned} \arg\min_{\alpha} \Bigg[& \frac{1}{n^+ n^-} \sum_{i=1}^{n^+} \sum_{j=1}^{n^-} \varepsilon_{ij} + \sum_{k=1}^{d} p_{\lambda_1}(|\alpha_k|) \Bigg] \\ \text{subject to } & \widetilde{L}_\alpha(x_i^+) - \widetilde{L}_\alpha(x_j^-) \geqslant 1 - \varepsilon_{ij}, \\ & i = 1, \ldots, n^+, \quad j = 1, \ldots, n^-, \\ & \varepsilon_{ij} \geqslant 0, \quad i = 1, \ldots, n^+, \quad j = 1, \ldots, n^-, \end{aligned}$$

which is another penalized SVM problem.

The above two steps in the SROC-SVM can be transformed to a regular penalized SVM as introduced in Section 2.1. We use algorithms in Zhang et al. (2006) and Becker et al. (2009) developed for regular penalized SVM and iterate between steps (1) and (2) until convergence is reached. The tuning parameters are selected by generalized approximate cross-validation (Becker et al., 2009). Since the linear scoring is identifiable up to a scale in computing AUC of an ROC, the above iterative procedure implies there is a root variable and a stem variable with similar coefficients selected

as baseline variables. The coefficients estimated for other variables are relative to the baseline variables.

2.3 *A Penalized Hierarchical Logistic Regression Approach*

When the association between variables and a dichotomous outcome is of interest, logistic regression is commonly used to model the association. There is an analogous penalized logistic regression approach for structured variables, which we describe here and we use to compare with the SROC-SVM in the simulations. To enforce the relationship between a root variable and a stem variable in our problem, we apply the same reparametrization as in (6) to the coefficients of the stem variables. A logistic regression under this reparametrization specifies

$$\text{logit}Pr(D_i = 1 \,|\, x_i) = \alpha_0 + \sum_{k=1}^{d} \alpha_k x_{ik} + \sum_{k=1}^{d} \sum_{j=1}^{J_k} \alpha_k \gamma_{kj} x_{ikj}.$$

To control for over-fitting and obtain sparse solution when the number of variables is large, we penalize the likelihood under the logistic regression model. To be specific, we obtain the coefficients of the root and the stem variables by iteratively solving the following two penalized likelihood problems:

(1) Given α_k, $k = 1, \ldots, d$ and define $\tilde{\alpha}_0 = \alpha_0 + \sum_{k=1}^{d} \alpha_k x_{ik}$, and $\tilde{x}_{ijk} = \alpha_k x_{ikj}$, solve for γ by minimizing

$$-\frac{1}{n} \sum_i \left[D_i \log \left(\frac{\exp\left(\tilde{\alpha}_0 + \sum_{k,j} \gamma_{kj} \tilde{x}_{ikj}\right)}{1 + \exp\left(\tilde{\alpha}_0 + \sum_{k,j} \gamma_{kj} \tilde{x}_{ikj}\right)} \right) + (1 - D_i) \log \left(\frac{1}{1 + \exp\left(\tilde{\alpha}_0 + \sum_{k,j} \gamma_{kj} \tilde{x}_{ikj}\right)} \right) \right] + \sum_{k=1}^{d} \sum_{j=1}^{J_k} p_{\lambda_2}(|\gamma_{kj}|);$$

(2) Given γ_{kj}, $k = 1, \ldots, d$, $j = 1, \ldots, J_k$ and define $\tilde{x}_{ik} = x_{ik} + \sum_{j=1}^{J_k} \gamma_{kj} x_{ikj}$, we solve for α by minimizing

$$-\frac{1}{n} \sum_i \left[D_i \log \left(\frac{\exp\left(\alpha_0 + \sum_k \alpha_k \tilde{x}_{ik}\right)}{1 + \exp\left(\alpha_0 + \sum_k \alpha_k \tilde{x}_{ik}\right)} \right) + (1 - D_i) \log \left(\frac{1}{1 + \exp\left(\alpha_0 + \sum_k \alpha_k \tilde{x}_{ik}\right)} \right) \right] + \sum_{k=1}^{d} p_{\lambda_1}(|\alpha_k|),$$

where $p_\lambda(\cdot)$ is an LASSO or a SCAD penalty function. Zou and Li (2008) provided fast one-step solutions to the above optimization problems, which can be applied here. We use five-fold cross-validation to select tuning parameters in each step. A linear decision rule is constructed with the parameters estimated from the logistic regression, that is, $L_{\hat{\alpha},\hat{\gamma}}(x_i) = \sum_{k=1}^{d} \widehat{\alpha}_k x_{ik} + \sum_{k=1}^{d} \sum_{j=1}^{J_k} \widehat{\alpha}_k \widehat{\gamma}_{kj} x_{ikj}$ to compute the empirical AUC.

2.4 *Inference Procedures via Bootstrap*

Here we discuss how to obtain inference for the estimated AUC of an ROC curve. It is well known that if one uses the data to obtain parameters of a model and then uses the same data to estimate measures of prediction performance or model fit (for example, prediction error or AUC of an ROC), the prediction accuracy will be over-estimated (Efron, 1986). An honest estimate of the classification performance should be assessed using independent data. In practice, however, such an independent data is usually not available. We propose to evaluate the estimated AUC through the following bootstrap procedure:

Step 1. Generate the bth copy of the bootstrap sample with size n from the observed data.

Step 2. Partition the bootstrap sample into a training set of size $n_1 = 2n/3$ and a testing set of size $n_2 = n/3$.

Step 3. Fit the model using data in the training set to obtain SROC-SVM. Use the estimated coefficients to compute the linear decision rule and the AUC of the ROC curve using data in the testing set.

Step 4. To avoid getting a large AUC by chance with a "lucky partitioning," repeat the random partition in steps 2 and 3 m times, and use the mean AUC of ROC across repetitions as the estimated AUC of the bth bootstrap sample.

Step 5. Repeat the steps 1 through 4 $\mathcal{B}$ times to obtain the bootstrap distribution of $\widehat{AUC}$.

By this procedure, we can obtain a distribution and the confidence interval of the estimated AUC. When the association between outcome and predictor in the final model is also of interest, to obtain a confidence interval for the estimated coefficients, one can use nonparametric bootstrap. For prospective studies, one takes independent bootstrap samples with replacement. For retrospective studies, one takes bootstrap samples for cases and controls separately. For each bootstrap sample, one can obtain the SROC-SVM and report the confidence interval based on the empirical quantiles of the bootstrapped estimates.

3. Simulations

To investigate performance of the proposed methods, we conducted the following two simulation studies. The first study simulates unstructured variables and the second study simulates structured variables.

3.1 *Simulation Study I*

The binary outcomes Y were generated from the generalized linear model

$$Pr(Y = 1|X) = g(X^T \beta), \tag{7}$$

where g is a link function and $\beta = (1.5, 0, 0, 0, 2, 0, 0, 0, 0, 0)^T$. We considered four different models. In models 1 and 3, $g(u) = \exp(u)/(1 + \exp(u))$ is the logit-link function. In models 2

Table 1
Independent structure: AUC and coefficients; Training set $n = 100$, testing set $n = 50$

Method		X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
True relative coeff.	AUC	1	0	0	0	1.33	0	0	0	0	0
Model 1	0.910*										
ROC-SVM	0.908	1.00	0.01	−0.01	0.01	1.27	0.00	0.00	0.01	−0.01	0.01
Logistic	0.905	1.03	0.02	−0.01	0.00	1.33	0.00	0.01	−0.01	0.00	0.00
SVM	0.900	1.00	0.02	−0.01	−0.01	1.31	0.02	0.00	−0.01	0.01	−0.01
Model 2	0.889*										
ROC-SVM	0.871	1.00	−0.01	0.01	0.01	1.34	−0.01	0.01	0.00	−0.01	0.01
Logistic	0.862	0.69	−0.01	0.02	0.00	0.93	0.00	0.00	0.00	0.01	0.01
SVM	0.858	1.00	0.00	0.02	0.01	1.41	0.00	0.01	0.01	−0.01	0.01
Model 3	0.856*										
ROC-SVM	0.847	1.00	−0.02	0.01	0.02	1.29	−0.03	0.02	−0.01	0.02	0.01
Logistic	0.814	0.91	−0.01	0.01	0.03	1.29	−0.05	0.01	−0.02	0.03	0.00
SVM	0.829	1.00	−0.01	0.01	0.01	1.31	−0.01	0.00	0.03	0.02	0.00
Model 4	0.888*										
ROC-SVM	0.878	1.00	0.06	−0.01	−0.02	1.45	0.01	0.00	−0.03	0.01	0.01
Logistic	0.845	1.13	0.19	0.05	0.04	1.92	0.13	0.11	0.07	0.07	0.17
SVM	0.860	1.00	0.08	−0.02	0.02	1.41	0.03	0.03	0.01	−0.01	0.05

*True AUC estimated from datasets with sample size $n = 500$ based on 200 replications.

and 4, $g(u) = 1/(1 + \exp(-u/2))$ for $u < 0$ and $g(u) = 1/(1 + \exp(-2u))$ for $u \geqslant 0$. Such model was also used in Ma and Huang (2007). For models 1 and 2, the variables $(x_1, \ldots, x_{10})$ were generated from a multivariate normal distribution with mean zero, variance one, and an AR-1 correlation with $\rho = 0.5$. For models 3 and 4, the variables were first generated from multivariate normal distribution as in models 1 and 2, then we dichotomized $x_2, \ldots, x_{10}$ by taking each variable as one if it was greater than or equal to 0 and otherwise 0. We kept x_1 as a continuous variable. The sample size for the training dataset was 100 and for the testing dataset was 50.

Three methods with SCAD penalty were compared: the ROC-SVM, penalized logistic regression, and regular penalized SVM. For each copy of the simulation data, we used the bootstrap procedure described in Section 2.4 to compute an honest estimate of the AUC. Due to computational burden of simulations, here we used $\mathcal{B} = 100$ bootstrap samples. In the real data analysis, we used $\mathcal{B} = 1000$ bootstrap samples. We then reported the mean AUC averaged across all simulation repetitions to compare different methods. For both the ROC-SVM and the SVM, x_1 was chosen as the baseline variable. Table 1 summarizes the mean AUC and the mean estimated coefficients based on 200 replications.

From Table 1 we can see that the ROC-SVM has the highest AUC among the three methods in all of the models regardless of whether the logistic link is correctly or incorrectly specified. The penalized logistic regression and SVM perform similarly in terms of AUC. However, we can see that when the link function is misspecified as a logistic link in models 2 and 4, the estimated coefficients from the penalized logistic regression are severely biased. In contrast, ROC-SVM and SVM are robust to misspecification of the link function due to their nonparametric nature: in all the models they yield a consistent estimate for the relative coefficients of the predictors.

To compare sparsity of the fitted models by different methods, we summarize measures of model complexity and other features as in Zou and Li (2008). In Tables 2 and 4, the column indexed as "C" is the mean number of variables with nonzero coefficients correctly estimated to be nonzero. The column indexed as "IC" is the mean number of variables with zero as coefficients incorrectly estimated as nonzero in the model. The column "Under-fit" is the proportion of models that miss some of the non-noise variables, the column "Correct-fit" is the proportion of models that correctly select the exact subset of the non-null variables, and "Over-fit" is the proportion of models that include some noise variables.

Table 2
Independent structure: model sparsity; training set $n = 100$, testing set $n = 50$

	No. of zeros		Proportion of		
Method	C	IC	Under-fit	Correct-fit	Over-fit
Model 1					
ROC-SVM	1.990	0.185	0.010	0.860	0.130
Logistic	1.975	0.480	0.025	0.695	0.280
SVM	1.960	0.355	0.040	0.720	0.240
Model 2					
ROC-SVM	1.995	0.220	0.005	0.820	0.175
Logistic	1.885	0.455	0.115	0.615	0.270
SVM	1.940	0.580	0.055	0.580	0.365
Model 3					
ROC-SVM	1.940	0.435	0.060	0.675	0.265
Logistic	1.885	0.780	0.110	0.460	0.430
SVM	1.815	0.365	0.165	0.580	0.255
Model 4					
ROC-SVM	1.975	0.435	0.025	0.685	0.290
Logistic	1.790	0.875	0.195	0.405	0.400
SVM	1.920	0.900	0.080	0.440	0.480

It can be seen from Table 2 that the ROC-SVM has the highest proportion of selecting the exact subset of variables in all simulation scenarios. For all methods, the proportion

Table 3
Hierarchical structure: AUC and coefficients

Method True coeff.	AUC	X1 1	X1a 1	X1b 0	X1c 0	X2 1	X2a 1.5	X2b 0	X2c 0	X3 0	X3a 0	X3b 0	X3c 0
Model 1	0.909*												
SROC-SVM	0.881	1	1	0.05	−0.03	1.00	1.44	−0.01	−0.01	0.04	0.02	−0.01	0.00
Slogistic	0.849	1.15	0.99	0.03	−0.01	1.14	1.52	−0.05	0.03	0.00	−0.05	0.01	0.00
Logistic	0.840	1.01	1.02	0.05	−0.03	0.95	1.61	−0.07	0.02	−0.01	−0.01	−0.01	0.02
SVM	0.836	1	1.03	0.03	−0.02	0.95	1.66	−0.07	0.03	−0.03	0.01	−0.02	0.03
Model 2	0.897*												
SROC-SVM	0.864	1	1	0.01	0.00	0.99	1.48	0.02	0.00	0.019	0.00	−0.01	0.00
Slogistic	0.822	0.88	0.61	0.03	−0.04	0.83	0.81	0.05	0.06	−0.07	−0.04	0.01	−0.02
Logistic	0.811	0.74	0.71	0.05	−0.04	0.68	1.08	0.04	0.06	−0.02	0.02	−0.01	0.03
SVM	0.807	1	1.09	0.04	−0.05	0.95	1.82	0.02	0.01	−0.03	0.02	−0.06	−0.04
Model 3	0.882*												
SROC-SVM	0.849	1	1	0.04	−0.04	1.13	1.71	0.00	−0.10	0.05	0.02	−0.03	0.01
Slogistic	0.801	1.29	0.91	0.02	−0.01	1.27	1.45	−0.04	−0.05	0.05	−0.13	0.06	−0.04
Logistic	0.757	1.01	1.02	0.05	−0.03	0.95	1.61	−0.07	0.02	−0.01	−0.01	−0.01	0.02
SVM	0.743	1	1.15	0.06	−0.03	0.95	1.69	0.01	−0.05	0.00	−0.06	0.00	0.03
Model 4	0.903*												
SROC-SVM	0.881	1	1	0.12	0.03	1.12	1.76	0.08	−0.01	0.05	0.01	0.00	−0.01
Slogistic	0.805	1.35	0.74	0.25	0.04	1.51	1.26	0.13	0.09	0.02	0.02	0.05	−0.05
Logistic	0.734	1.32	0.73	0.30	0.08	1.34	1.40	0.14	0.11	0.01	0.15	0.13	0.02
SVM	0.728	1	1.06	0.02	0.05	0.96	1.66	−0.03	−0.04	−0.05	0.07	0.13	0.00

*True AUC estimated from datasets with a large sample size ($n = 500$) based on 200 replications.

Table 4
Hierarchical structure: model sparsity

	No. of zeros		Proportion of		
Method	C	IC	Under-Fit	Correct-Fit	Over-Fit
Model 1					
SROC-SVM	4.000	0.395	0.000	0.685	0.315
Slogistic	3.785	1.215	0.145	0.315	0.540
Logistic	3.615	1.090	0.345	0.225	0.430
SVM	3.475	1.275	0.395	0.145	0.460
Model 2					
SROC-SVM	4.000	0.330	0.000	0.705	0.295
Slogistic	3.655	1.505	0.255	0.240	0.505
Logistic	3.355	1.295	0.495	0.155	0.350
SVM	3.650	1.965	0.295	0.135	0.570
Model 3					
SROC-SVM	3.935	0.310	0.045	0.745	0.210
Slogistic	3.575	1.215	0.245	0.245	0.510
Logistic	3.695	1.480	0.285	0.175	0.540
SVM	3.615	1.650	0.355	0.155	0.490
Model 4					
SROC-SVM	3.970	0.220	0.015	0.790	0.195
Slogistic	3.470	0.640	0.335	0.335	0.330
Logistic	3.505	0.795	0.440	0.240	0.320
SVM	3.680	1.525	0.260	0.180	0.560

of choosing the correct model is higher in the models with continuous outcomes (models 1 and 2) than the models with binary outcomes (models 3 and 4). With the same type of outcome, all methods perform better on models with all continuous predictor variables (models 1 and 2) compared to models with predominantly binary predictor variables (models 3 and 4). Moreover, the proportion of choosing an under-fitted model using ROC-SVM is less than or equal to 0.06, hence is negligible in all of the four simulation settings.

3.2 *Simulation Study II*

Parallel to the unrelated variables model in the preceding section, we simulated four similar models with hierarchically structured variables. The outcomes were generated from (7) with $\beta = (1, 1, 0, 0, 1, 1.5, 0, 0, 0, 0, 0, 0)^T$. There were three groups of predictors with each group having one root variable and three stem variables. For all the models, variables within each group were generated from multivariate normal distribution with mean zero, variance one, and an AR-1 correlation between variables with $\rho = 0.5$. Variables in distinct groups were generated independently. For models 3 and 4, we kept two variables as continuous predictors while we dichotomized all the other variables as one if they were greater than or equal to zero, and otherwise as zero. We specified the first variable in each group as the root variable and the following three variables as the stem variables. This simulation setting is close to the real data analyzed in Section 4.

Four methods with SCAD penalty were compared: hierarchical ROC-SVM, hierarchical penalized logistic regression, regular penalized logistic regression, and regular penalized SVM. The total sample size was 150 with 100 as the training set and 50 as the testing set. The AUCs were obtained using the similar bootstrap procedure described in Sections 2.4 and 3.1. For fair comparison, for the methods that treat variables as unrelated (regular penalized logistic regression and regular penalized SVM), when a root variable is not selected in the final model then its stem variables will be automatically dropped. Table 3 summarizes the analysis results based on 200 replications.

From this set of simulations, we see that the SROC-SVM and hierarchical logistic regression that take into account the structure among variables have higher AUC than the regular penalized logistic regression and SVM that ignore the hierarchical structure. The improvement in AUC for SROC-SVM is nonignorable compared to the penalized logistic regression and regular SVM: the improvement is about 4% for models 1 and 2 and about 10% for models 3 and 4. Therefore when there are binary variables involved as in our data analysis example, the advantage of SROC-SVM is more notable. Analogous to the models with unrelated variables in the preceding section, when the link function is not a logistic function, the penalized logistic regression with or without accounting for the hierarchical structure lead to biased estimates of the coefficients. Again, SROC-SVM and SVM are robust to the choice of link function.

We report the sparsity of the fitted models by different methods in this simulation setting in Table 4. From this table, we can see that the SROC-SVM and hierarchical logistic regression accounting for the structure among variables are more likely to choose correct models and less likely to choose under-fitted models. The SROC-SVM clearly outperforms the other three methods. The proportion of SROC-SVM choosing an under-fitted model is negligible, which is in contrast to the higher proportion using the other three methods. This also explains why they have a smaller AUC compared to the SROC-SVM.

4. A Real Data Example

In this section, we apply the proposed methods to the PSQ data introduced in Section 1. In this study, there were 77 Latino subjects recruited and administered PSQ and other surveys along with a gold standard psychosis diagnostic interview (DSM-IV SCID; First et al., 1998). The PSQ was developed by Bebbington and Nayani (1995) and studied in multiple clinical populations in the primary care clinics as well as in our sample. The screener contains variables grouped into five domains: hypomania, thought interference, persecution, strange experiences, and hallucinosis. The hierarchical structure was chosen for two reasons: First, to start with questions that were deliberately vague as a gentle introduction, so as to avoid "over-incisive" initial questions about psychosis and thereby decrease subjects' desire to continue. Second, to streamline the instrument so that subjects who responded negatively to the initial question could be given a negative answer to later items. The questions thus act as a kind of a funnel, with gentle initial questions getting progressively sharper. The questionnaire requires both kinds of questions in order to entice subjects in, but then reduce the number of likely false positives, which is known to be high for psychosis screeners.

Outcomes collected on the PSQ are binary variables. In addition, there are survey questions such as "how many days have you been off from work/school" with continuous outcomes. All the legitimate combinations of variables in each domain are listed in Table 5. Combinations with only stem variables but not their root variables are not allowed. In the analysis, we standardized the continuous variables by their standard deviations. The hierarchical dependence in the PSQ questions measured by the correlation between the root question and their stem questions ranges from 0.52 to 0.83.

We first apply the ROC-SVM and penalized logistic regression without considering structure among the variables. The AUCs of the ROC for each root question alone were 0.382, 0.453, 0.440, 0.541, 0.699, and 0.691. Since Q5 has the largest AUC, we set it as the baseline variable for the ROC-SVM. The ROC-SVM chooses variables Q1, Q3b, Q4a, Q5, Q5a, and Q6 with coefficients $\widehat{\beta} = (0.56, 1.26, 1.11, 1, 0.94, 0.35)^T$ and AUC = 0.841. The penalized logistic regression chooses variables Q1, Q3a, Q3b, Q5, Q5a, and Q6 with coefficients $\widehat{\beta} = (3.30, 9.27, 8.26, 4.28, 4.80, 1.05)^T$ and AUC = 0.859. The AUCs of the ROC-SVM and penalized logistic regression that did not consider structure among the variables were 0.841 and 0.859, which were not very low. However, none of them gives a legitimate model, since they choose some stem variables without their root variables. For example, ROC-SVM chooses Q3b without its root variable Q3.

We next apply SROC-SVM to the data to accommodate the hierarchical structure of the PSQ. Based on results from the ROC-SVM and penalized logistic regression, we can see that Q5 and Q5a have large effects with similar coefficients, thus we choose Q5 and Q5a as the baseline variables so that their coefficients are set to be one. In the final model, variables Q1, Q3, Q3b, Q5, Q5a, and Q6 are chosen with estimated coefficients $\hat{\beta} = (0.68, 0.22, 0.72, 1, 1, 0.12)^T$. To see whether the procedure selects variables that have good predictive performance, we bootstrapped 1000 times from the original data set and split the data into a training set of the size $2n/3$ and a testing set of size $n/3$. Applying the procedure in Section 2.4, we obtain a mean $\widehat{AUC} = 0.876$, and a 95% confidence interval of $\widehat{AUC}$ of [0.636, 1] based on the bootstrap sample quantiles. In practice, an AUC of the ROC between 0.7 and 0.9 is considered to be useful or having a good predictive performance (Swets, 1988). This range covers the estimated AUC of our final model.

Table 5
Structure of the variables in PSQ of the psychosis data analysis

Domain	Variables involved	Legitimate combinations
PSQ G1	Q1, Q1a, Q1b	NULL, {Q1}, {Q1 Q1a}, {Q1 Q1b}, {Q1 Q1a Q1b}
PSQ G2	Q2, Q2a	NULL, {Q2}, {Q2 Q2a}
PSQ G3	Q3, Q3a, Q3b	NULL, {Q3}, {Q3 Q3a}, {Q3 Q3b}, {Q3 Q3a Q3b}
PSQ G4	Q4, Q4a	NULL, {Q4}, {Q4 Q4a}
PSQ G5	Q5, Q5a	NULL, {Q5}, {Q5 Q5a}
Others	Q6	NULL, {Q6}

5. Discussion

We have developed prediction-based structured variable selection procedures through penalizing empirical AUC of an ROC. The computational issues were resolved by using a connection between the empirical AUC and the penalized support vector machine. The proposed ROC-SVM is applicable to unstructured variables. To account for certain hierarchical structure (e.g., weak hierarchical dependence, Yuan and Lin, 2007) among variables, we applied an appropriate reparametrization and proposed SROC-SVM. Here we illustrate our methods assuming each root variable has several stem variables. It is straightforward to accommodate variables that do not have stem variables and therefore are unrelated with other variables. The proposed methods are illustrated for binary disease outcomes. Extensions can easily be made to accommodate continuous or survival outcomes. For example, Obuchowski (2006) generalizes the usual dichotomous ROC analysis to continuous outcomes. Heagerty and Zheng (2005) proposed ROC curves for evaluating predictive accuracy for survival outcomes. These methods can be implemented for the proposed SROC-SVM without complication.

A limitation of the ROC-based procedures is that one will choose baseline variables before the analysis. Our simulations (results not shown) suggest that the model accuracy and AUC of the ROC curve is not sensitive to the choice of the baseline variables, as long as they are indeed predictive of the outcome. However, if uninformative predictors are forced to enter the model as baseline variables, the performance of the methods can be severely compromised. This phenomena is consistent with that reported in Ma and Huang (2005, 2007). In practice, to avoid choosing an unpredictive variable as baseline, we can compute AUC using each individual predictor and the outcome and choose the one with the highest AUC as the baseline variable.

The asymptotics of the coefficients estimated from the ROC-SVM and SROC-SVM appear to be complicated. When the variables are given a priori, the consistency and asymptotic distribution are known (Han, 1987). For penalized SVM with unrelated variables, Koo et al. (2008) showed that the solution of a regular penalized SVM converges to a minimizer of an appropriate loss function. Future research on this topic is needed.

Here, we turn the AUC-based penalized SVM problem to a regular penalized SVM that allows algorithms developed for the latter to be directly used. In general, the SVM algorithms converge fast. However, the computation burden increases with the sample size and the number of candidate variables. In such cases, more efficient SVM algorithms proposed in Calders and Jaroszewicz (2007) are useful.

Acknowledgements

Yuanjia Wang's research is supported by a National Institutes of Health grant, AG031113-01A2. Runze Li's research was supported by a National Science Foundation grant DMS 0348869, and National Institutes of Health grants R21 DA024260 and P50 DA-10075. The research of Roberto Lewis-Fernández is supported by the National Alliance for Research on Schizophrenia and Depression (NARSAD), NIH grant MH077226, the Bureau of Cultural Competence of the New York State Office of Mental Health, and institutional funds from the New York State Psychiatric Institute. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIDA or the NIH.

References

Bebbington, P. and Nayani, T. (1995). The Psychosis Screening Questionnaire. *International Journal of Methods in Psychiatric Research* **5,** 11–19.

Becker, N., Werft, W., Toedt, G., Lichter, P., and Benner, A. (2009). Penalized SVM: A R-package for feature selection SVM classification. *Bioinformatics* **25,** 1711–1712.

Brefeld, U. and Scheffer, T. (2005). AUC maximizing support vector learning. In *Proceedings of the 22nd International Conference on Machine Learning-Workshop on ROC Analysis in Machine Learning*, Bonn, Germany.

Briggs, M. and Zaretzki, R. (2008). The skill plot: A graphical technique for evaluating continuous diagnostic tests. *Biometrics* **63,** 250–261.

Calders, T. and Jaroszewicz, S. (2007). Efficient AUC Optimization for Classification. *Proceedings of the 11th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD)*, pp. 42–53.

Efron, B. (1986). How biased is the apparent error rate of a prediction rule? *Journal of the American Statistical Association* **81,** 461–470.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96,** 1348–1360.

First, M. B., Spitzer, R. L., Gibbon, M., and Williams, J. (1998). Structured Clinical Interview for DSM-IV Axis I Disorders, Patient Edition (SCID-I/P, Version 2.0, 9/98 revision), Biometrics Research Department, New York State Research Institute.

Fung, G. and Mangasarian, O. L. (2004). A feature selection Newton method for support vector machine classification. *Computational Optimization and Applications* **28,** 185–202.

Han, A. K. (1987). Non-parametric analysis of a generalized regression model. The maximum rank correlation estimator. *Journal of Economics* **35,** 303–316.

Hand, D. (2008). On Briggs and Zaretzki: The Skill Plot: A graphical technique for evaluating continuous diagnostic tests. *Biometrics* **63**, 259.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer.

Heagerty, P. J. and Zheng, Y. (2005). Survival model predictive accuracy and ROC curves. *Biometrics* **61,** 92–105.

Huang, J., Ma, S., Xie, H., and Zhang, C. (2009). A group bridge approach for variable selection. *Biometrika* **96,** 339–355.

Koo, K., Lee, Y., Kim, Y., and Park, C. (2008). A Bahadur representation of the linear support vector machine. *Journal of Machine Learning Research* **9,** 1343–1368.

Lewis-Fernández, R. (2003). Assessing psychosis screeners among underserved urban primary care patients. In *Proceedings of the 15th Annual Scientific Symposium*, National Alliance for Research on Schizophrenia and Depression (NARSAD), October 17, 2003.

Ma, S. and Huang, J. (2005). Regularized ROC method for disease classification and biomarker selection with microarray data. *Bioinformatics* **21,** 4356–4362.

Ma, S. and Huang, J. (2007). Combining multiple markers for classification using ROC. *Biometrics* **63,** 751–757.

Obuchowski, N. A. (2006). An ROC-type measure of diagnostic accuracy when the gold standard is continuous-scale. *Statistics in Medicine* **25,** 481–493.

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

Pepe, M. S. (2005). Evaluating technologies for classification and prediction in medicine. *Statistics in Medicine* **24,** 3687–3696.

Pepe, M. S., Cai, T., and Longton, G. (2006). Combining predictors for classification using the area under the receiver operating characteristic curve. *Biometrics* **62,** 221–229.

Pinsky, P. (2005). Scaling of true and apparent ROC AUC with number of observations and number of variables. *Communications in Statistics: Simulation and Computation* **34,** 771–781.

Swets, J. A. (1988). Measuring the accuracy of diagnostic systems. *Science* **240,** 1285–1293.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B* **58,** 267–288.

Wahba, G., Lin, Y., and Zhang, H. (2000). GACV for support vector machines, or, another way to look at margin-like quantities. In *Advances in Large Margin Classifiers*, A. J. Smola, P. Bartlett, B. Schölkopf, and D. Schurmans (eds), 297–309. Cambridge, Massachusetts: MIT Press.

Wang, S., Nan, B., Zhou, N., and Zhu, J. (2009). Hierarchically penalized Cox regression with grouped variables. *Biometrika* **96,** 307–322.

Yuan, M. and Lin, Y. (2007). An efficient variable selection approach for analyzing designed experiments. *Technometrics* **49,** 430–439.

Yuan, M., Joseph, R., and Zou, H. (2009). Structured variable selection and estimation. *Annals of Applied Statistics* **3,** 1738–1757.

Zhang, H., Ahn, J., Lin, X., and Park, C. (2006). Gene selection using support vector machine with non-convex penalty. *Bioinformatics* **22,** 88–95.

Zou, H. and Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *Annals of Statistics* **36,** 1509–1533.

Received May 2010. *Revised September* 2010.
Accepted September 2010.