Published in final edited form as:

J Am Stat Assoc. 2014 ; 109(506): 574–589. doi:10.1080/01621459.2013.858630.

Variable Selection in Nonparametric Classification via Measurement Error Model Selection Likelihoods

L. A. Stefanski, Yichao Wu, and Kyle White

North Carolina State University

Abstract

Using the relationships among ridge regression, LASSO estimation, and measurement error attenuation as motivation, a new measurement-error-model-based approach to variable selection is developed. After describing the approach in the familiar context of linear regression, we apply it to the problem of variable selection in nonparametric classification, resulting in a new kernel-based classifier with LASSO-like shrinkage and variable-selection properties. Finite-sample performance of the new classification method is studied via simulation and real data examples, and consistency of the method is studied theoretically. Supplementary materials for the paper are available online.

key words and Phrases

Attenuation; Bayes rule; Binary regression; Convolution; Discriminant analysis; Kernel discriminant analysis; LASSO; Linear regression; Maximum likelihood rule; Model selection; Ridge regression

1 Introduction

Adopting a measurement-error-model-based approach to variable selection, we propose a nonparametric, kernel-based classifier with LASSO-like shrinkage and variable-selection properties. The use of measurement error model (MEM) ideas to implement variable selection is new and provides a different way of thinking about variable selection, with potential for applications in other nonparametric variable selection problems. Thus we also describe what we call *measurement error selection likelihoods* in addition to our main methodological results on variable selection in nonparametric classification.

Compared to parametric classification, variable selection for nonparametric classification methods is in its infancy. Our research helps fill that gap with a sparsity-seeking kernel method, *sparse kernel discriminant analysis* (SKDA), obtained by implementing the MEM-

L. A. Stefanski is Drexel Professor of Statistics, Department of Statistics, North Carolina State University, Raleigh, NC 27695. stefanski@ncsu.edu. Yichao Wu is Associate Professor, Department of Statistics, North Carolina State University, Raleigh, NC 27695. wu@stat.ncsu.edu. Kyle White is Doctoral Student, Department of Statistics, North Carolina State University, Raleigh, NC 27695. kwhite3@ncsu.edu.

6 Supplementary Materials

Theoretical Complements. The file VarSelViaMEM.Supp.pdf contains the derivation of the equivalence of LASSO and $\hat{L}_{SEL_1}$ - MESSO; the distributional result showing how to generate sparse classification data; and proofs of the asymptotic results (Lemmas 1 and 2, Theorem 1, and Corollary 1).

based approach to variable selection. SKDA is kernel-based with a familiar form, but with a bandwidth parameterization and selection strategy that results in variable selection. We provide additional background and introductory material at the start of the main methodological section on classification, Section 4.

In response to a suggestion by the Associate Editor we end this section with remarks relating to the potential benefits of approaching variable selection problems via measurement error models. The fact that a version of our MEM-based approach to variable selection applied to linear regression results in LASSO estimation is of independent interest. Knowing different paths to the same result and the relationships among them, enhances understanding even when it does not lead to new methods. However, our approach is more than just another LASSO computational algorithm. It is a useful conceptualization and generalization of the LASSO that has potential to suggest new variable selection methods.

Penalizing parameters is not always intuitive simply because it is not always the case that variables enter a model through easily intuited parameters; as with nonparametric models and algorithmic fitting methods. However, it is always possible to intuit the case that a variable has (a lot of) measurement error in it. Admittedly, turning the idea that a variable contains measurement error into a variable selection method may require additional, creative modeling, and perhaps extensive computing power to simulate the measurement error process when analytical expressions are not possible. However, the key point is that the MEM-based approach provides another way for researchers to think about variable selection in nonstandard problems. Although variable selection in nonparametric classification is the primary methodological contribution of the paper, the idea of approaching variable selection via measurement error modeling may have broader impact because of the possibility of adapting the strategy to other problems not readily handled by traditional penalty approaches.

The connections between measurement error attenuation, shrinkage, and selection underlying the new approach to variable selection are discussed in Section 2. The new approach is illustrated in the context of linear regression in Section 3. The main results on variable selection in nonparametric classification are in Section 4, which includes performance assessment via simulation studies, applications to two data sets, and asymptotic results. Concluding remarks appear in Section 5 and technical details in the online supplementary materials.

2 Variable Selection and Measurement Error

We now describe the connection between measurement error and variable selection that is used to derive the nonparametric classification selection method studied in Section 4.

2.1 Attenuation, Shrinkage, and Selection

We begin with the connections between measurement error attenuation, ridge regression, and LASSO estimation in linear models $\mathbf{Y}_{n \times 1} = \mathbf{X}\boldsymbol{\beta}_{p \times 1} + \boldsymbol{\varepsilon}$.

Measurement error attenuation—Measurement error attenuation is usually introduced in the case of simple linear regression in terms of ‘true’ $\{Y_i, X_i\}$, and measured $\{Y_i, W_i\}$ data, $i = 1, \dots, n$, where $W_i = X_i + \sigma_U Z_i$, and $Z_i \stackrel{\text{iid}}{\sim} N(0, 1)$ with the Z_i independent of (X_i, Y_i) (Fuller, 1987; Carroll et al., 2006; Buonaccorsi, 2010). The least squares slope of Y on W , $\hat{\beta}_{\text{ATTEN}} = s_{yw}/s_w^2 \xrightarrow{P} \sigma_{yw}/\sigma_w^2 = (\sigma_x^2 + \sigma_U^2)^{-1} \sigma_{yx}$. Attenuation results from the inequality $\sigma_x^2 + \sigma_U^2 > \sigma_x^2$. The multi-predictor version of this result for the error model $\mathbf{W} = \mathbf{X} + \mathcal{D}_{\{\sigma_U\}} \mathbf{Z}$, with $\mathcal{D}_{\{\sigma_U\}} = \text{diag}(\sigma_{U,1}, \dots, \sigma_{U,p})$ and $\mathbf{Z} \sim N(\mathbf{0}, \mathbf{I})$ is

$$\hat{\beta}_{\text{ATTEN}} = \hat{\mathbf{V}}_W^{-1} \hat{\mathbf{V}}_{WY} \approx \left(\hat{\mathbf{V}}_X + \mathcal{D}_{\{\sigma_U^2\}} \right)^{-1} \hat{\mathbf{V}}_{XY} = \left(\hat{\mathbf{V}}_X + \mathcal{D}_{\{1/\sigma_U^2\}}^{-1} \right)^{-1} \hat{\mathbf{V}}_{XY}. \quad (1)$$

Notes. i) The approximation “ $\approx$ ” in (1) is valid in large samples because

$\hat{\mathbf{V}}_W^{-1} \hat{\mathbf{V}}_{WY} - \left(\hat{\mathbf{V}}_X + \mathcal{D}_{\{\sigma_U^2\}} \right)^{-1} \hat{\mathbf{V}}_{XY}$ converges in probability to $\mathbf{0}$. **ii)** We use $1/\sigma_U^2$ to denote componentwise division, i.e., $1/\sigma_U^2 = (1/\sigma_{U,1}^2, \dots, 1/\sigma_{U,p}^2)^T$. **iii)** We include the right-hand-side expression because many of our results are better understood in terms of measurement error *precisions* (or square-root precisions). **iv)** We use $\mathcal{D}_{\{\mathbf{v}\}}$ to denote a diagonal matrix with vector $\mathbf{v}$ on its diagonal. **v)** For any vector $\mathbf{z}$, $|\mathbf{z}|$ denotes the vector of componentwise absolute values. Similarly, $\mathbf{z}^m$ denotes the vector of componentwise m^{th} powers. **vi)** Finally, $\hat{\mathbf{V}}$ denotes a sample variance/covariance matrix of the subscripted variables.

Ridge shrinkage—Ridge regression (Hoerl and Kennard, 1970) is much studied and well known. Thus we simply note that after scaling the ridge parameters $\mathbf{v}$ it has the form

$$\hat{\beta}_{\text{RIDGE}} = \left(\hat{\mathbf{V}}_X + \mathcal{D}_{\{\mathbf{v}\}} \right)^{-1} \hat{\mathbf{V}}_{XY} = \left(\hat{\mathbf{V}}_X + \mathcal{D}_{\{1/\mathbf{v}\}}^{-1} \right)^{-1} \hat{\mathbf{V}}_{XY}. \quad (2)$$

LASSO shrinkage and selection—The LASSO has an attenuation-like representation that follows from the iterated-ridge computation algorithm; see Tibshirani (1996) and Section 3.2. The iterated ridge solution to the penalty form of the LASSO optimization is

$$\hat{\beta}_{\text{LASSO}} = \underset{\beta_1, \dots, \beta_p}{\text{argmin}} \left\{ \frac{1}{n-1} \|\mathbf{Y} - \mathbf{X}\beta\|^2 + \eta \sum_{j=1}^p |\beta_j| \right\} \quad (3)$$

$$= \left(\mathbf{I} + \mathcal{D}_{\{|\hat{\beta}_{\text{LASSO}}|/\eta\}} \hat{\mathbf{V}}_X \right)^{-1} \mathcal{D}_{\{|\hat{\beta}_{\text{LASSO}}|/\eta\}} \hat{\mathbf{V}}_{XY} \quad (4)$$

$$= \left(\hat{\mathbf{V}}_X + \mathcal{D}_{\{|\hat{\beta}_{\text{LASSO}}|/\eta\}}^{-1} \right)^{-1} \hat{\mathbf{V}}_{XY}. \quad (5)$$

Notes: **i)** Expressions (4) and (5) are not computational solutions to the LASSO optimization problem, but rather equations satisfied by the solution. **ii)** Both are included to show that the inversion of a singular matrix in (5) when some components of $\hat{\beta}_{\text{LASSO}}$ are zero is not necessary; however, we use expressions like that in (5) for their compactness.

Comparing equations (1), (2), and (5) shows that $\hat{\beta}_{\text{ATTEN}}$, $\hat{\beta}_{\text{RIDGE}}$, and $\hat{\beta}_{\text{LASSO}}$ exhibit the same *form* of attenuation, with equality in the case that

$$\frac{1}{\sigma_{U,j}^2} = \frac{1}{\nu_j} = \frac{|\hat{\beta}_{\text{LASSO},j}|}{\eta}. \quad (6)$$

Measurement error attenuation results from uncontrollable and undesirable measurement error in the predictors, whereas ridge and LASSO attenuation is analyst-controlled to manipulate sampling properties of the estimators. In Section 2.3 we show that measurement error attenuation can also be controlled to accomplish shrinkage and selection. For another interesting example of creatively using measurement error to impart favorable sampling properties on estimators in a different context, see Seo and Lindsay (2011).

2.2 Oracle Heuristics

We motivate our measurement error approach to variable selection using a toy example, imagining an oracle's solution to a simple challenge. Suppose that the regression of Y on predictors $X_1, \dots, X_5$ satisfies

$$E(Y|X_1, X_2, X_3, X_4, X_5) = E(Y|X_2, X_5).$$

The nullity of X_1, X_3 , and X_4 is unknown to a statistician, but not to the oracle. However, suppose the oracle agrees to play a game wherein s/he must add a nonzero, total amount of measurement error, $(\sigma_{U,1}Z_1, \sigma_{U,2}Z_2, \sigma_{U,3}Z_3, \sigma_{U,4}Z_4, \sigma_{U,5}Z_5)$, to the predictors $(X_1, X_2, X_3, X_4, X_5)$ in a manner that minimizes the loss of their collective predictive power ("a nonzero, total amount of measurement error" means that $\sum_j \sigma_{U,j}^2 > 0$).

Because the oracle knows that

$$\begin{aligned} E(Y|X_1, X_2, X_3, X_4, X_5) &= E(Y|X_2, X_5) \Rightarrow E(Y|X_1 \\ &+ \sigma_{U,1}Z_1, X_2, X_3 + \sigma_{U,3}Z_3, X_4 \\ &+ \sigma_{U,4}Z_4, X_5) = E(Y|X_2, X_5), \end{aligned}$$

the oracle's solution is to set $\sigma_{U,2} = \sigma_{U,5} = 0$ and let $\sigma_{U,1}$, $\sigma_{U,3}$, and $\sigma_{U,4}$ be nonzero. A statistician seeing the oracle's solution would know the identity of the null and informative predictors. In the next section we show that this oracle game can be mimicked using a (pseudo-profile) likelihood to replace the omniscience of the oracle, and constraints to 'force' the addition of measurement error to the data.

2.3 MEM Selection Likelihoods

We now give a general description of the construction and use of measurement error model (MEM) selection likelihoods. As an illustration of the approach, we then implement it for linear models in Section 3, where it is shown to lead to the LASSO for linear models. This equivalency lends credibility to the new approach in the sense that it can be viewed as a generalization of the LASSO that reduces to the LASSO in the case of linear regression. In Section 4 we implement the approach in the context of nonparametric classification obtaining the new method, *sparse kernel discriminant analysis* (SKDA). The latter is the main methodological contribution of the paper. However, Section 4 also demonstrates the utility of the measurement error approach to variable selection for identifying and deriving variable selection strategies in nonparametric/nonstandard models.

Denote the data as $(Y_i, \mathbf{X}_i)$, $i = 1, \dots, n$, and let $(Y, \mathbf{X})$ represent a generic observation. The MEM selection likelihood construction proceeds in four basic steps:

1. Start with an assumed ‘true’ likelihood for $(\mathbf{X}_i, Y_i)$, $i = 1, \dots, n$, denoted $L_{\text{TRUE}}(\boldsymbol{\theta})$ where $\boldsymbol{\theta}$ could be finite (parametric) or infinite dimensional (nonparametric).
2. Construct the associated MEM likelihood under the ‘false’ assumption that the components of $\mathbf{X}$ are measured with independent error. That is, assume that $\mathbf{W}$ is observed in place of $\mathbf{X}$ where $\mathbf{W}|\mathbf{X} \sim N\left(\mathbf{X}, \mathcal{D}_{\{\sigma_U^2\}}\right)$ with $\sigma_U^2 = (\sigma_{U,1}^2, \dots, \sigma_{U,p}^2)$. The resulting likelihood depends on $\boldsymbol{\theta}$ and σ_U^2 and is denoted $L_{\text{MEM}}(\boldsymbol{\theta}, \sigma_U^2)$. Note that even though $L_{\text{MEM}}(\boldsymbol{\theta}, \sigma_U^2)$ is derived under a measurement error model assumption, it is calculated from the error-free data $(\mathbf{X}_i, Y_i)$, $i = 1, \dots, n$.
3. Replace $\boldsymbol{\theta}$ in $L_{\text{MEM}}(\boldsymbol{\theta}, \sigma_U^2)$ with an estimate $\hat{\boldsymbol{\theta}}$ resulting in the pseudo-profile likelihood $\hat{L}_{\text{pMEM}}(\sigma_U^2) = L_{\text{MEM}}(\hat{\boldsymbol{\theta}}, \sigma_U^2)$. Note that $\hat{\boldsymbol{\theta}}$ is an estimator for $\boldsymbol{\theta}$ calculated from the observed data without regard to the ‘false’ measurement error assumption, e.g., $\hat{\boldsymbol{\theta}}$ could be the maximum likelihood estimator from $L_{\text{TRUE}}(\boldsymbol{\theta})$.
4. Reexpress the pseudo-profile likelihood $\hat{L}_{\text{pMEM}}(\sigma_U^2)$ in terms of precision (or square-root precision) $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_p)$ where $\lambda_j = 1/\sigma_{U,j}^2$ (or $\lambda_j = 1/\sigma_{U,j}$), resulting in the MEM selection likelihood $\hat{L}_{\text{SEL}}(\boldsymbol{\lambda})$.

$\hat{L}_{\text{SEL}}(\boldsymbol{\lambda})$ is maximized subject to: $\lambda_j > 0$, $j = 1, \dots, p$; and $\sum_j \lambda_j \leq \tau$. Setting the tuning parameter $\tau < \infty$ in the latter constraint ensures that the harmonic mean of the measurement error variances (or standard deviations) is $p/\tau > 0$. This is how the approach ‘forces’ measurement error into the likelihood. Maximizing $\hat{L}_{\text{SEL}}(\boldsymbol{\lambda})$ subject to the constraint ensures that the measurement error is distributed to predictors optimally in the sense of diminishing the likelihood the least. The idea is that null predictors tend to be assigned large measurement error variances, whereas informative predictors tend to be assigned relatively small ones. Thus maximizing $\hat{L}_{\text{SEL}}(\boldsymbol{\lambda})$ plays the role of the oracle. The constrained maximum can have some $\lambda_j = 0$ for small τ , indicating that the data are consistent with

predictor X_j containing infinite measurement error, and therefore is uninformative. Note that $\hat{\lambda}_j=0 \Rightarrow \hat{\sigma}_{U,j}^2=1/0=\infty$, see (6).

Standardization was not mentioned previously. Doing so facilitates interpretation, numerical calculation, and comparison with other estimators, thus we assume henceforth that predictor variables are centered and scaled to unit variance. Viewed from the measurement error perspective, standardization ensures that the measurement error variances referred to in the development in Section 3, $\sigma_{U,1}^2, \dots, \sigma_{U,p}^2$ are all on the same scale.

The four steps broadly define the approach. In particular applications it may make sense to restrict to a partial/pseudo likelihood as illustrated in the next section. Regardless, the resulting “likelihood” will generally not be a true likelihood; however, we use the term accepting the abuse of terminology in exchange for brevity. In the next section we illustrate the approach in the familiar context of linear regression, which also serves to justify the approach since we show it leads to the LASSO. In Section 4 we show that the approach leads to a useful new variable selection method for kernel-based classification rules.

3 Linear Model Selection Likelihoods

3.1 Linear Regression

We start with the assumption that $(Y, \mathbf{X})$ is multivariate normal with mean $(\mu_Y, \boldsymbol{\mu}_X^T)^T$ and partitioned variance matrix $(V_Y, V_{YX}, V_{XY}, V_X)$. Under the ‘false’ assumption that the predictor is measured with error, the distribution is multivariate normal with the same mean, but with partitioned variance matrix $(V_Y, V_{YX}, V_{XY}, V_X + \mathcal{D}_{\{\sigma_U^2\}})$. Thus

$$L_{\text{TRUE}}(\boldsymbol{\theta}) = \prod_i \Psi \left(Y_i, \mathbf{X}_i; (\mu_Y, \boldsymbol{\mu}_X^T)^T, (V_Y, V_{YX}, V_{XY}, V_X) \right),$$

$$L_{\text{MEM}}(\boldsymbol{\theta}, \sigma_U^2) = \prod_i \Psi \left(Y_i, \mathbf{X}_i; (\mu_Y, \boldsymbol{\mu}_X^T)^T, (V_Y, V_{YX}, V_{XY}, V_X + \mathcal{D}_{\{\sigma_U^2\}}) \right),$$

where $\Psi(y, \mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Omega})$ is the multivariate normal density with mean $\boldsymbol{\mu}$ and variance matrix $\boldsymbol{\Omega}$, and $\boldsymbol{\theta} = (\mu_Y, \boldsymbol{\mu}_X, V_Y, V_{YX}, V_{XY}, \text{vech}(V_X))$. We set $\hat{\boldsymbol{\theta}}$ equal to the sample moments of $(Y_i, \mathbf{X}_i)$, $i = 1, \dots, n$, i.e., $\hat{\boldsymbol{\theta}} = (\hat{\mu}_Y, \hat{\boldsymbol{\mu}}_X, \hat{V}_Y, \hat{V}_{YX}, \hat{V}_{XY}, \text{vech}(\hat{V}_X))$.

Then with $\hat{\boldsymbol{\beta}}(\boldsymbol{\lambda}) = (\hat{\mathbf{V}}_X + \mathcal{D}_{\{\boldsymbol{\lambda}^m\}}^{-1})^{-1} \hat{\mathbf{V}}_{XY}$, where $\boldsymbol{\lambda}^m = (\lambda_1^m, \dots, \lambda_p^m)$, and $m = 1$ for $\lambda_j = 1/\sigma_{U,j}^2$ and $m = 2$ for $\lambda_j = 1/\sigma_{U,j}$, it follows that

$$\hat{L}_{\text{SEL}}(\boldsymbol{\lambda}) \propto \hat{L}_{\text{SEL}_1} \hat{L}_{\text{SEL}_2} \hat{L}_{\text{SEL}_3}, \quad (7)$$

where

$$\begin{aligned}\hat{L}_{\text{SEL}_1} &= \left\{ \hat{\mathbf{V}}_Y - \hat{\mathbf{V}}_{XY}^T \hat{\boldsymbol{\beta}}(\boldsymbol{\lambda}) \right\}^{-n/2}, \\ \hat{L}_{\text{SEL}_2} &= \exp \left(-\frac{n}{2} \left[\frac{\hat{\mathbf{V}}_Y - \hat{\mathbf{V}}_{XY}^T \hat{\boldsymbol{\beta}}(\boldsymbol{\lambda}) + \left\{ \hat{\boldsymbol{\beta}}(\boldsymbol{\lambda})^T \hat{\mathbf{V}}_X \hat{\boldsymbol{\beta}}(\boldsymbol{\lambda}) - \hat{\mathbf{V}}_{XY}^T \hat{\boldsymbol{\beta}}(\boldsymbol{\lambda}) \right\}}{\hat{\mathbf{V}}_Y - \hat{\mathbf{V}}_{XY}^T \hat{\boldsymbol{\beta}}(\boldsymbol{\lambda})} \right] \right), \\ \hat{L}_{\text{SEL}_3} &= \left[\det \left\{ \left(\hat{\mathbf{V}}_X + \mathcal{D}_{\{\lambda^m\}}^{-1} \right)^{-1} \right\} \right]^{n/2} \exp \left[-\frac{n}{2} \text{tr} \left\{ \hat{\mathbf{V}}_X \left(\hat{\mathbf{V}}_X + \mathcal{D}_{\{\lambda^m\}}^{-1} \right)^{-1} \right\} \right].\end{aligned}\quad (8)$$

Factoring $\hat{L}_{\text{SEL}}$ as in (7) facilitates discussion of the linear model in Section 3.2.

Although $\left(\hat{\mathbf{V}}_X + \mathcal{D}_{\{\lambda^m\}}^{-1} \right)^{-1}$ is more succinct, when a component of $\boldsymbol{\lambda}$ is 0, we use either of the equivalent expressions, $(\mathbf{I} + \mathcal{D}_{\{\lambda^m\}} \mathbf{V}_X)^{-1} \mathcal{D}_{\{\lambda^m\}}$ or $(\mathbf{I} + \mathbf{V}_X \mathcal{D}_{\{\lambda^m\}})^{-1}$ (compare to (4) and (5)). These identities imply that

$$\begin{aligned}\hat{\boldsymbol{\beta}}(\boldsymbol{\lambda}) &= \left(\hat{\mathbf{V}}_X + \mathcal{D}_{\{\lambda^m\}}^{-1} \right)^{-1} \hat{\mathbf{V}}_{XY} = \mathcal{D}_{\{\lambda^m\}} \left(\mathbf{I} + \hat{\mathbf{V}}_X \mathcal{D}_{\{\lambda^m\}} \right)^{-1} \hat{\mathbf{V}}_{XY} \\ &= \left(\mathbf{I} + \mathcal{D}_{\{\lambda^m\}} \hat{\mathbf{V}}_X \right)^{-1} \mathcal{D}_{\{\lambda^m\}} \hat{\mathbf{V}}_{XY},\end{aligned}\quad (9)$$

from which it is apparent that $\lambda_j = 0$ implies that the j^{th} component of $\hat{\boldsymbol{\beta}}(\boldsymbol{\lambda}) = 0$. The simple identity $\mathcal{D}_{\{v_1\}} \mathbf{v}_2 = \mathcal{D}_{\{v_2\}} \mathbf{v}_1$ shows that $\hat{\boldsymbol{\beta}}(\boldsymbol{\lambda}) = \mathcal{D}_{\{\Delta\}} \boldsymbol{\lambda}^m$ where $\mathcal{D}_{\{\Delta\}} = (\mathbf{I} + \mathbf{V}_X \mathcal{D}_{\{\lambda^m\}})^{-1} \mathbf{V}_X \hat{\mathbf{V}}_{XY}$. Under common distributional assumptions on $Y_1, \dots, Y_n$, the components of $\hat{\boldsymbol{\beta}}(\boldsymbol{\lambda})$ are nonzero almost surely. In this case it follows that $\boldsymbol{\lambda}^m = \mathcal{D}_{\{\Delta\}}^{-1} \hat{\boldsymbol{\beta}}(\boldsymbol{\lambda})$ almost surely, and thus the j^{th} component of $\hat{\boldsymbol{\beta}}(\boldsymbol{\lambda}) = 0$ if and only if $\lambda_j = 0$.

We now study MEM selection likelihood method for linear regression in the case that λ_j is parameterized in terms of precisions ($m = 1$). The main results in the following section are: i) maximizing $\hat{L}_{\text{SEL}_1}$ is a convex optimization problem that yields an estimator equivalent to the LASSO; ii) maximizing the full-conditional likelihood $\hat{L}_{\text{SEL}_1} \hat{L}_{\text{SEL}_2}$ with $m = 1$ is not equivalent to the LASSO, and in terms of variance per shrinking bias, can be less variable than the LASSO; and iii) maximizing the full likelihood $\hat{L}_{\text{SEL}_1} \hat{L}_{\text{SEL}_2} \hat{L}_{\text{SEL}_3}$ can never result in $\hat{\lambda}_j = 0$ and thus is not an option for variable selection (but is an option for shrinkage only). We use the acronym MESSO (Measurement Error Shrinkage and Selection Operator) to identify estimators derived from the selection likelihoods in Section 3.1.

3.2 Relationship to LASSO

In the supplemental Appendix Section A-7.1 we prove that constrained maximization of $\hat{L}_{\text{SEL}_1}$ in (8) results in LASSO solution paths as functions of τ . The proof establishes convexity of $\sigma^2(\boldsymbol{\lambda}) = (\hat{L}_{\text{SEL}_1})^{-2/n}$ and equivalence of KKT conditions. Note that although the MESSO selection likelihood in Section 3 was derived starting with an assumption of multivariate normality, the equivalency to the LASSO does not depend on that assumption. We now consider the relationship of LASSO to other MESSO versions obtained by maximizing the full-conditional likelihood, $\hat{L}_{\text{SEL}_1} \hat{L}_{\text{SEL}_2}$, and the full likelihood $\hat{L}_{\text{SEL}_1} \hat{L}_{\text{SEL}_2} \hat{L}_{\text{SEL}_3}$ for $\lambda_j = 1/\sigma_j^2$ and $\lambda_j = 1/\sigma_j$. A detailed comparison is beyond the scope of this paper. However, we make a few brief, relevant observations. Figure 1 displays plots of relative

deviations from average total prediction error variance versus total squared prediction error bias from a small simulation study comparing LASSO and the three MESSO estimators when $\lambda_j = 1/\sigma_j^2$. That is, the plot displays $\{\text{TV-Ave(TV)}\}/\text{Ave(TV)}$ versus TB, where: $\text{TB} = \{\hat{E}(\hat{\beta}) - \beta\}^T \Omega \{\hat{E}(\hat{\beta}) - \beta\}$; $\text{TV} = \text{tr}\{\hat{\text{Var}}(\hat{\beta})\Omega\}$; Ω is the predictor covariance matrix; Ave(TV) is TV averaged over the three unique estimators; and $\hat{E}$ and $\hat{\text{Var}}$ denote Monte Carlo expectation and variance. The plot allows differentiation of the estimators based on total variance for equivalent total squared bias. Data were simulated according to a design in Tibshirani (1996): $Y_i = \mathbf{X}_i^T \beta + \sigma \varepsilon_i$, $i = 1, \dots, n$, where $\mathbf{X}_i$ are iid $N(\mathbf{0}_{p \times 1}, \text{AR1}(\rho))$, ε_i are iid $N(0, 1)$, with $n = 20$, $p = 8$, $\beta = (3, 1.5, 0, 0, 2, 0, 0, 0)^T$, $\rho = .5$, and $\sigma = 3$. This model calls for a mix of selection and shrinkage. Figure 1 shows that the LASSO is less variable (as measured by $\text{tr}\{\hat{\text{Var}}(\hat{\beta})\Omega\}$) for small total squared bias (as measured by $\{\hat{E}(\hat{\beta}) - \beta\}^T \Omega \{\hat{E}(\hat{\beta}) - \beta\}$), but not for larger values of TB. For reference, we obtained $\text{TB} = 1.27$ for the LASSO estimator tuned using 10-fold cross validation, which corresponds very closely to the value of TB for which the black-dashed and red-blue curves intersect. Thus in a region of total squared bias of known practical importance, the LASSO and the full-conditional MESSO estimator have similar total variance.

Examination of L_{SEL_3} in (8) reveals that if any $\lambda_j = 0$, then the determinant term is zero and the maximizer of the full likelihood $L_{\text{SEL}_1} L_{\text{SEL}_2} L_{\text{SEL}_3}$ never contains zero elements, i.e., selection is not possible. Finally, when λ is parameterized as square-root precisions, $\lambda_j = 1/\sigma_j$, the MESSO optimization is no longer convex, and results in greater selection.

Our comparison to the LASSO is not because we consider MESSO to be a competitor to the LASSO for linear models, but rather to establish the viability of the new approach in a familiar context. In the next section we show that it leads to a promising method of variable selection for nonparametric classification.

4 Nonparametric Classification

Classification has a long and important history in statistics. Several methods are widely used: Fisher's linear discriminant analysis (LDA, Fisher, 1936); logistic regression (McCullagh and Nelder, 1989); support vector machine (Vapnik, 1995; Cristianini and Shawe-Taylor, 2000); and boosting (Freund and Schapire, 1997). LDA is the estimated Bayes rule under within-class normality and equal covariance matrices (Anderson, 2003), whereas with unequal covariances the estimated Bayes rule is quadratic discriminant analysis (QDA).

A nonparametric Bayes rule obtained by replacing conditional densities in the theoretical Bayes rule with kernel density estimates (Wand and Jones, 1995) is known as kernel density-based discriminant analysis (KDA, Cooley and MacEachern, 1998; Ghosh et al., 2006). See Duda et al. (2000) and Hastie et al. (2001) for recent reviews of KDA.

With multiple predictors, variable selection is important. Some predictors may not carry useful information and their inclusion can deteriorate performance of an estimated classification rule. Bickel and Levina (2004) studied how dimension affects performance

and showed that LDA behaves like random guessing when the ratio of p to n is large. Thus motivated, they proposed an independence rule by using a diagonal variance/covariance matrix. Fan and Fan (2008) selected a set of important predictors via thresholding, and applied the independence rule to the selected predictors, calling the method an *annealed independence rule*. Mai et al. (2012) proposed variable selection for LDA by formulating it as a regression problem. Cai and Liu (2011) studied a direct estimation approach for variable selection in LDA. For other methods of variable selection for classification, see Hall et al. (2009), Witten and Tibshirani (2009), Shao et al. (2011) and references therein.

4.1 Selection Likelihoods for Classification

The true model is defined via the conditional distributions $\mathbf{X} | (Y = 0) \sim f_0(\mathbf{x})$, $\mathbf{X} | (Y = 1) \sim f_1(\mathbf{x})$, and the marginal probability $P(Y = 1) = \pi_1 = 1 - \pi_0$. It follows that

$$P(Y=1|\mathbf{X}=\mathbf{x})=P_{\text{TRUE}}(\mathbf{x};\pi_0,\pi_1,f_0,f_1)=\pi_1 f_1(\mathbf{x})/(\pi_0 f_0(\mathbf{x})+\pi_1 f_1(\mathbf{x})) \\ =\left\{1+\frac{\pi_0}{\pi_1}\frac{f_0(\mathbf{x})}{f_1(\mathbf{x})}\right\}^{-1}.$$

Under the ‘false’ measurement error model assumption $\mathbf{W}|\mathbf{X}\sim N\left(\mathbf{X},\mathcal{D}_{\{\sigma_{U,j}^2\}}\right)$, the conditional distributions are, using $\phi()$ to denote the standard normal density,

$$\mathbf{W}|(Y=k)\sim f_k(\mathbf{x},\boldsymbol{\lambda})=\int\cdots\int\prod_{j=1}^p\frac{1}{\sigma_{U,j}}\varphi\left(\frac{x_j-t_j}{\sigma_{U,j}}\right)f_k(\mathbf{t})dt_1\cdots dt_p.$$

It follows that the MEM selection likelihood is determined by

$$P(\mathbf{x};\pi_0,\pi_1,f_0,f_1,\boldsymbol{\lambda})=\left\{1+\frac{\pi_0}{\pi_1}R(\mathbf{x};f_0,f_1,\boldsymbol{\lambda})\right\}^{-1}, \quad (10)$$

where:

$$R(\mathbf{x};f_0,f_1,\boldsymbol{\lambda})=\frac{\int\cdots\int\prod_{j=1}^p\lambda_j^c\varphi\left(\lambda_j^c(x_j-t_j)\right)f_0(\mathbf{t})dt_1\cdots dt_p}{\int\cdots\int\prod_{j=1}^p\lambda_j^c\varphi\left(\lambda_j^c(x_j-t_j)\right)f_1(\mathbf{t})dt_1\cdots dt_p} \\ =\frac{\int\cdots\int\prod_{j=1}^p\varphi\left(\lambda_j^c(x_j-t_j)\right)dF_0(\mathbf{t})}{\int\cdots\int\prod_{j=1}^p\varphi\left(\lambda_j^c(x_j-t_j)\right)dF_1(\mathbf{t})};$$

$c = 1/2$ when $\lambda_j = 1/\sigma_{U,j}^2$, and $c = 1$ when $\lambda_j = 1/\sigma_{U,j}$ respectively; $F_k()$ is the distribution corresponding to $f_k()$; and the common factor $\prod_{j=1}^p\lambda_j^c$ is deleted in the final ratio. For our work on classification we take $c = 1$.

Replacing π_1 with $\pi_1 = \bar{Y}$, π_0 with $\pi_0 = 1 - \pi_1$, and F_0 and F_1 with their nonparametric MLEs (empirical distribution functions) in $P(\mathbf{x}; \pi_0, \pi_1, f_0, f_1, \boldsymbol{\lambda})$ in (10) results in

$$\hat{P}(\mathbf{x}; \boldsymbol{\lambda}) = \left\{ 1 + \frac{\hat{\pi}_0}{\hat{\pi}_1} R(\mathbf{x}; \boldsymbol{\lambda}) \right\}^{-1}, \quad (11)$$

where

$$\begin{aligned} \hat{R}(\mathbf{x}; \boldsymbol{\lambda}) &= \frac{n_0^{-1} \sum_{i: Y_i=0}^n \prod_{j=1}^p \varphi(\lambda_j^c(x_j - X_{j,i}))}{n_1^{-1} \sum_{i: Y_i=1}^n \prod_{j=1}^p \varphi(\lambda_j^c(x_j - X_{j,i}))} \\ &= \frac{n_0^{-1} \sum_{i: Y_i=0}^n \prod_{j: \lambda_j \neq 0}^p \varphi(\lambda_j^c(x_j - X_{j,i}))}{n_1^{-1} \sum_{i: Y_i=1}^n \prod_{j: \lambda_j \neq 0}^p \varphi(\lambda_j^c(x_j - X_{j,i}))}, \end{aligned} \quad (12)$$

with $n_k = n\pi_k$. The second expression in (12) is displayed to emphasize the fact that for every $\lambda_j = 0$ there is a common factor $\varphi(0)$ in the numerator and denominator of $R(\mathbf{x}; \boldsymbol{\lambda})$ that cancels, in which case $R(\mathbf{x}; \boldsymbol{\lambda})$ does not depend on $X_{j,1}, \dots, X_{j,n}$.

Because Y_i is binary a suitable MEM selection likelihood for this model is

$$\hat{L}_{\text{SEL}}(\boldsymbol{\lambda}) = \sum_{i=1}^n Y_i \log(\hat{P}(\mathbf{X}_i; \boldsymbol{\lambda})) + (1 - Y_i) \log(1 - \hat{P}(\mathbf{X}_i; \boldsymbol{\lambda})),$$

resulting in the optimization problem

$$\begin{aligned} \max_{\lambda_1, \lambda_2, \dots, \lambda_p} \quad & \sum_{i=1}^n Y_i \log(\hat{P}(\mathbf{X}_i; \boldsymbol{\lambda})) + (1 - Y_i) \log(1 - \hat{P}(\mathbf{X}_i; \boldsymbol{\lambda})) \\ \text{subject to} \quad & \sum_{j=1}^p \lambda_j = \tau; \quad \lambda_j \geq 0, \quad j=1, 2, \dots, p; \end{aligned} \quad (13)$$

Maximizing $L_{\text{SEL}}(\boldsymbol{\lambda})$ subject to the constraints $\lambda_j \geq 0, j=1, \dots, p$, and $\sum_j \lambda_j = \tau$ results in some $\lambda_j = 0$ when τ is small enough. In this case $R(\mathbf{x}; \boldsymbol{\lambda})$ is a ratio of kernel density estimates of only those components of $\mathbf{X}$ corresponding to nonzero λ_j . That is, $\lambda_j = 0$ implies that the j^{th} predictor is selected out of the model, just as in the linear model. Finally we note that (11) corresponds to the estimated Bayes rule classifier. For an estimated maximum likelihood classifier set $\pi_0 = \pi_1 = 1/2$. The tuning parameter τ can be selected by minimizing classification error, or the likelihood of an independent tuning set, or via cross validation. In our numerical work we used cross validation.

4.2 Multicategory SKDA

SKDA is readily extended to multicategory classification. In multicategory classification with K classes, we label the response as $1, 2, \dots, K$. The training data set is $\{(\mathbf{X}_i, Y_i) : i=1, 2, \dots, n\}$ with $\mathbf{X}_i \in \mathbb{R}^p$ and $Y_i \in \{1, 2, \dots, K\}$.

With a slight abuse of notation define $\tilde{f}_k(\mathbf{x}; \boldsymbol{\lambda}) = \sum_{i: y_i = k} \prod_{j=1}^p K(\lambda_j(x_j - X_{ij}))/n_k$ and $\hat{\pi}_k = n_k/n$, $k = 1, 2, \dots, K$ with n_k being the number of observations in class k and $\hat{P}_k(\mathbf{x}; \boldsymbol{\lambda}) = \hat{\pi}_k \tilde{f}_k(\mathbf{x}; \boldsymbol{\lambda}) / (\sum_{m=1}^K \hat{\pi}_m \tilde{f}_m(\mathbf{x}; \boldsymbol{\lambda}))$. The log-likelihood is $\sum_{i=1}^n \log(\hat{P}_{y_i}(\mathbf{X}_i; \boldsymbol{\lambda}))$ and the multicategory SKDA (MSKDA) requires solving

$$\begin{aligned} & \max_{\lambda_1, \lambda_2, \dots, \lambda_p} \sum_{i=1}^n \log(\hat{P}_{y_i}(\mathbf{X}_i; \boldsymbol{\lambda})) \\ & \text{subject to } \sum_{j=1}^p \lambda_j = \tau, \quad \lambda_j \geq 0, \quad j=1, 2, \dots, p. \end{aligned} \quad (14)$$

The MSKDA rule for $\mathbf{X} = \mathbf{x}$ is given by $\arg\max_k \hat{P}_k(\mathbf{x}; \hat{\boldsymbol{\lambda}})$. As defined $\hat{P}_k(\mathbf{x}; \boldsymbol{\lambda})$ yields the estimated Bayes rule classifier. For the maximum likelihood classifier set all $\hat{\pi}_k = 1/K$.

4.3 Simulation Studies

In this section, we evaluate the performance of SKDA and MSKDA via simulation, comparing them with LDA, sparse LDA (SLDA, Mai et al., 2012), and KDA (resp. MKDA, the multicategory extension of KDA). For the binary case, we also compare with the SCAD penalized logistic regression (Fan and Li, 2001) and SCAD penalized support vector machine (SVM, Zhang, 2006). Note that only LDA does not depend on a tuning parameter. The other methods require the determination of a single tuning parameter. In their most general form, KDA and MKDA each require the estimation of p tuning parameters (bandwidths). In our numerical work we replace the p bandwidths with a single common bandwidth. This simplification is necessary given the large dimensions studied.

We determine tuning parameters using 10-fold cross validation coupled with a grid search. We use classification error as the selection criterion for SLDA as done in Mai et al. (2012) and SCAD SVM, and the log-likelihood as the selection criterion for the kernel-based methods and SCAD logistic regression.

We included π_k , the prior probability of class k , and its estimator $\hat{\pi}_k$ in the proposed methods and their theoretical development for generality. However in the simulation examples, we consider simulation settings with equal prior probabilities for all classes and thus we set all $\hat{\pi}_k$ to be $1/K$. The same simplification is used in the real data examples, thus we study the maximum likelihood classification rule rather than the Bayes rule.

Our simulation study design requires generating multivariate observations of which only a subset of the (correlated) variables appear in the Bayes classifier. In the supplemental Appendix Section A-7.2 we give a proposition explaining how to accomplish the data generation.

4.3.1 Two-group, Normal Linear Discriminant Model—Our first simulation study is two-group classification for which LDA and SLDA are specifically designed: $\mathbf{X}_{p \times 1} | (Y = k) \sim \text{MVN}_p(\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$ for $k = 0, 1$, $p = 10$ predictors, and $Y \sim \text{Bernoulli}(0.5)$. The means were set as $\boldsymbol{\mu}_0 = (-1, 1, \mathbf{0}_{8 \times 1}^T)^T$ and $\boldsymbol{\mu}_1 = (1, -1, \mathbf{0}_{8 \times 1}^T)^T$, and $\boldsymbol{\Sigma}$ was the equicorrelation matrix with

parameter ρ , for levels of ρ given by 0, 0.3, and 0.6. A test set of size 20,000 was fixed to evaluate the performance of the classification methods. Training sets of size $n = 200$ are used in the simulation, and 10-fold cross validation was used to select the tuning parameters for KDA, SKDA, SLDA, SCAD-Logistic, and SCAD-SVM. Results for 100 simulated data sets are in Table 1 and Figure 2.

Table 1 gives predictor selection frequencies by SKDA, SLDA, SCAD logistic regression, and SCAD SVM over 100 repetitions. Note that, only X_1 and X_2 are important for this example according to Proposition 1 in Section A–7.2. Table 1 shows that the new nonparametric method SKDA performs competitively in terms of variable selection, even though the data generation strongly favors LDA and SLDA.

We plot the two important predictors, X_1 versus X_2 , for one training sample in the left panels of Figure 2 with classes represented by blue “○” and red “+.” Box plots of test errors over the independent test set across 100 repetitions are in the right panels. The top, middle, and bottom rows correspond to $\rho = 0, 0.3$ and 0.6 , respectively. As the setting is ideal for LDA, and SLDA, and only marginally less so for SCAD-Logistic and SCAD-SVM, these methods perform very well with SLDA having a slight edge. KDA neither exploits the marginal normality of the data, nor does it do variable selection, and thus performs poorly. However, although SKDA does not exploit the marginal normality, its performance in terms of classification error and selection is competitive.

4.3.2 Two-Group Mixtures of Equicorrelated Normals—This is another binary example with $Y \sim \text{Bernoulli}(0.5)$. In this example, we consider an AR(1) type correlation $\Sigma = (\sigma_{ij})$ with $\sigma_{ij} = \rho^{|i-j|}$. We partition predictors with $p_1 = 2$ and $p_2 = p - 2$. Predictors are generated as $\mathbf{X} | (Y = 0) \sim N(\mathbf{0}, \Sigma)$ and

$\mathbf{X} | (Y = 1) \sim 0.5N((\theta_1^T \sum_{11}, \theta_1^T \sum_{12})^T, \sum) + 0.5N(-(\theta_1^T \sum_{11}, \theta_1^T \sum_{12})^T, \sum)$ with $\theta_1 = (3, -3)^T$. The ‘0’ class is multivariate normal, while the ‘1’ class is a mixture of two multivariate normal distributions. Thus the true classification rule is nonlinear and LDA and SLDA are misspecified in this case. Yet it can be checked that the true classification rule depends only on the first two predictors X_1 and X_2 (according to Proposition 1 in Section A–7.2). Training sets of size 200 and a test set of size 20,000 are used. We set $p = 10$ and consider three different ρ values 0, 0.3, and 0.6. Results over 100 repetitions are reported in Table 2 and Figure 3 in a similar way as the previous example.

Table 2 indicates that the two important predictors X_1 and X_2 are selected by SKDA for all 100 repetitions while unimportant predictors are selected infrequently. SLDA, SCAD logistic regression, and SCAD SVM fail to select the two important predictors for most repetitions. The poor performance of these methods is not unexpected in light of the data generation model. Variables X_1 and X_2 from one simulated training data set are plotted in the left panels of Figure 3 for three values of ρ : 0 (top), 0.3 (middle), and 0.6 (bottom). There is no good way to linearly separate the two classes well. Test error results are consistent as shown in the right panels of Figure 3. SKDA performs well whereas LDA, SLDA, SCAD logistic regression, and SCAD SVM perform like random guessing.

High dimensional case: We repeated the study changing only $p = 10$ to $p = 50$. Boxplots of the test errors over 100 repetitions are shown in right panels of Figure 4 for different methods and three values of ρ : 0 (top), 0.3 (middle), and 0.6 (bottom). Out of 100 repetitions, the frequency of X_1 and X_2 and average frequency of $X_3, \dots, X_{50}$ being selected by SKDA and SLDA are reported in Table 3.

The true classifier in this example is nonlinear, and thus LDA, SLDA, SCAD logistic regression, and SCAD SVM are misspecified. SKDA is nonparametric and thus more flexible. hence it performs well at both variable selection and classification accuracy.

To address a review comment, we increase the dimension further to $p = 100$. For the case with $\rho = 0$, the SKDA selects the important predictors X_1 and X_2 for all 100 repetitions and the average selection frequency for the other 98 unimportant predictors is 5.2959. The average testing error for the SKDA is 0.0426 with a standard deviation of 0.0095. We expect a similar performance for the cases with $\rho = 0.3$ and $\rho = 0.6$.

4.3.3 Three-group, Normal Linear Discriminant Model—This is a 3-class example with $p = 10$ and an AR(1) correlation $\Sigma = (\sigma_{ij})$ with $\sigma_{ij} = \rho^{|i-j|}$. We partition predictors with $p_1 = 2$ and $p_2 = p - 2$. Data are generated in two steps: generate Y uniformly over $\{1, 2, 3\}$. Conditional on Y , predictors are generated as $\mathbf{X} | (Y = 1) \sim N(\mathbf{0}, \Sigma)$;

$\mathbf{X} | (Y = 1) \sim N(\mathbf{0}, \Sigma)$; $\mathbf{X} | (Y = 2) \sim N((\theta_{12}^T \sum_{11}, \theta_{12}^T \sum_{12})^T, \Sigma)$ with $\theta_{12} = (2, 2\sqrt{3})^T$;

$\mathbf{X} | (Y = 3) \sim N((\theta_{13}^T \sum_{11}, \theta_{13}^T \sum_{12})^T, \Sigma)$ with $\theta_{13} = (-12, 2\sqrt{3})^T$. According to Proposition 1 in Section A-7.2, the Bayes classification rule depends only on X_1 and X_2 .

For this 3-class example, training sets of size 300 and a test set of size 30,000 are used. As it is not clear how to extend SLDA to multicategory case, we only compare MSKDA to LDA and MKDA in multicategory examples. Results over 100 repetitions are reported in Table 4 and Figure 5. In particular, Table 4 reports the frequency of each predictor selected by MSKDA. It shows that MSKDA gives very good variable selection performance. Note that, in this example, each class is represented by a multivariate normal distribution. Thus LDA is correctly specified and should do well in terms of classification error. This is confirmed by results in Figure 5. MSKDA reduces the test error in comparison to MKDA as a result of its variable selection. Because MSKDA performs variable selection very well, its performance is similar to that of LDA even though it is nonparametric.

To address a reviewer's query about the performance of the SKDA and MSKDA when some predictors have discrete distributions, we reran the $\rho = 0$ case above replacing the informative continuous predictor X_2 with the 'new' informative discrete predictor X_2 generated conditionally on Y according to $X_2 = 3(Y - 1) + U$ where U has a discrete uniform distribution on the set $\{1, 2, 3, 4, 5\}$, and replacing the noninformative continuous X_7 with the 'new' noninformative discrete $X_7 \sim \text{Poisson}(4)$. The selection frequencies for variables $X_1, \dots, X_{10}$ over 100 simulated data sets were 100, 100, 0, 1, 0, 1, 0, 1, 0, 1, respectively (compare to the top row of Table 4). A training sample and boxplots of test errors for this discrete case are displayed in Figure 6 (compare to the top row of Figure 5). These results are promising although far from a thorough study of SKDA and MSKDA with discrete

predictors. The latter will likely prove fruitful just as the study of discrete predictors in the framework of kernel regression without variable selection has been; see Racine and Li (2004) and Li and Racine (2007) and the reference therein.

4.3.4 Three-Group, Mixtures of Equicorrelated Normals—This is another 3-class example with $p = 10$ and an AR(1) correlation $\Sigma = (\sigma_{ij})$ with $\sigma_{ij} = \rho^{|i-j|}$. We also partition predictors with $p_1 = 2$ and $p_2 = p - 2$. Data are generated in two steps: generate Y uniformly over $\{1, 2, 3\}$. Conditional on $Y = k$, predictors are generated as

$\mathbf{X}|(Y=k) \sim 0.5N((\theta_{1k}^T \sum_{11}, \theta_{1k}^T \sum_{12})^T, \sum) + 0.5N(-(\theta_{1k}^T \sum_{11}, \theta_{1k}^T \sum_{12})^T, \sum)$, a mixture of two multivariate Gaussian distributions, for $k = 1, 2, 3$. Here

$\theta_{11} = (5, 0_9^T)^T$, $\theta_{12} = (2.5, 5\sqrt{3}/2, 0_8^T)^T$, and $\theta_{13} = (2.5, -5\sqrt{3}/2, 0_8^T)^T$. According to Proposition 1 in Section A-7.2, the Bayes classification rule depends only on X_1 and X_2 .

Training sets of size 300 and a test set of size 30,000 are used. Results over 100 repetitions are in Table 5 and Figure 7 in the same format as the previous example. For this data generation model, LDA is badly misspecified. Consequently LDA performs close to random guessing as seen in by Figure 7. MSKDA improves upon MKDA in terms of classification error due to its variable selection ability.

4.4 Illustrations with Real Data

4.4.1 Wisconsin Breast Cancer Data—We apply SKDA to the Wisconsin Breast Cancer Data (WBCD) containing information on 569 patients, including a binary indicator of tumor malignancy (malignant/benign) and thirty purported predictors.¹ We standardized each predictor to mean zero and variance one. In each repetition, we randomly select 300 observations as the training set and the remaining 269 are used as the test set to report performance by calculating the classification error for each method. Ten-fold cross validation was used for estimating tuning parameters. Test error results over 40 repetitions are reported in Table 6 and displayed in Figure 8. SKDA had a test error comparable to LDA, KDA, SCAD logistic regression, and SCAD SVM but a smaller test error than SLDA.

The average number of predictors selected by SKDA, SLDA, SCAD logistic regression, and SCAD SVM are 3.8, 11.0, 6.63, and 8.3, respectively. Predictor selection frequencies for SLDA and SKDA for this example are given in Table 7. SKDA selects far fewer predictors yet is competitive in terms of classification error.

4.4.2 Chemical Signatures—The proliferation of natural gas wells due to economically feasible welling technologies has heightened awareness and concern about emissions from these wells and their impacts on human health and the environment. On-ground, off-site, remote emissions measuring technology will play a key role in obtaining quality data for assessing emissions from oil and gas production sites. We illustrate the application and performance of SKDA using data collected during the development of a remote measuring technology (Thoma et al., 2012). The data contain measurements of 38 chemical compounds

¹Available at the UCI Machine Learning Repository <http://archive.ics.uci.edu/ml/>

recorded at each of $n = 170$ sites. Secondary to the remote-sensing method development research, the data were clustered using k -means clustering with $k = 2$ (and using all 38 compounds) ostensibly aligning with well type (*dry*, *wet*). We are not concerned with the accuracy or interpretation of this initial clustering. Rather for our illustration we take the classification as given and address the question of whether a smaller subset of the 38 compounds explains it. Because the study is still in progress, we identify the chemical compounds only as $X_1, \dots, X_{38}$.

The heat map in Figure 9 illustrates the data and classification. The figure is too small for the well labels, 1, ..., 170, to be legible on the heat map plot. However, it is sufficient to know that the first 50 columns correspond to the " $Y = 0$ " group, and that the last 120 columns correspond to the " $Y = 1$ " group. Chemical compound labels, identified here only as $X_1, \dots, X_{38}$, are legible on the right vertical axis.

To assess the relative performance of methods on this data set we used 40 random splits of the data into training and test sets of sizes 136 and 34 respectively. The training data were analyzed using eight-fold cross validation to select tuning parameters where required. Average test errors (with standard errors) are reported in Table 8.

In addition to having the lowest test error, the collection of forty models selected by SKDA clearly suggests that two variables, X_2 and X_5 , provide the bulk of the discriminatory power, whereas the collection of models selected by SLDA is far more diverse and less suggestive with regard to possible sparsity in a good classification rule.

4.5 Consistency of SKDA

We now establish consistency of SKDA. Below we establish the notation used for the statements of the results below, and in the proofs in the appendix Section A-7.3.

Partition $\mathbf{X}$ into important, $\mathbf{X}_{\mathcal{I}}$, and unimportant, $\mathbf{X}_{\mathcal{U}}$, components. Without loss of generality, assume that $\mathbf{X} = (\mathbf{X}_{\mathcal{I}}^T, \mathbf{X}_{\mathcal{U}}^T)^T$. Denote the conditional density function of $\mathbf{X}$ given $Y = k$ by $f_k(\mathbf{x}) \equiv f_k(\mathbf{x}_{\mathcal{I}}, \mathbf{x}_{\mathcal{U}})$ and the conditional probability of $Y = 1$ given $\mathbf{X} = \mathbf{x}$ by

$P(\mathbf{x}) \equiv P(\mathbf{x}_{\mathcal{I}}, \mathbf{x}_{\mathcal{U}}) = \frac{\pi_1 f_1(\mathbf{x}_{\mathcal{I}}, \mathbf{x}_{\mathcal{U}})}{\pi_0 f_0(\mathbf{x}_{\mathcal{I}}, \mathbf{x}_{\mathcal{U}}) + \pi_1 f_1(\mathbf{x}_{\mathcal{I}}, \mathbf{x}_{\mathcal{U}})}$. Note that $\mathbf{X}_{\mathcal{U}}$ being unimportant means

$$P(\mathbf{x}_{\mathcal{I}}, \mathbf{x}_{\mathcal{U}}) = P(\tilde{\mathbf{x}}_{\mathcal{I}}, \tilde{\mathbf{x}}_{\mathcal{U}}) \text{ as long as } \tilde{\mathbf{x}}_{\mathcal{I}} = \mathbf{x}_{\mathcal{I}}. \quad (15)$$

Note further that this partition is not unique since $\mathcal{I} \cup \{j\}$ and $\mathcal{U} \setminus \{j\}$ is another partition of important and unimportant predictors for any $j \in \mathcal{U}$ as long as $\mathcal{I}$ and $\mathcal{U}$ satisfy (15). To ensure uniqueness, assume that $\mathcal{I}$ has minimal cardinality among all such partitions.

Extending the partition notation to $P(\tilde{\mathbf{x}}) \equiv P(\tilde{\mathbf{x}}; \boldsymbol{\lambda}) = \pi_1 \hat{f}_1(\tilde{\mathbf{x}}) / (\pi_0 \hat{f}_0(\tilde{\mathbf{x}}) + \pi_1 \hat{f}_1(\tilde{\mathbf{x}}))$, we define $\tilde{P}(\mathbf{x}_{\mathcal{I}}, \mathbf{x}_{\mathcal{U}}) \equiv P(\mathbf{x}_{\mathcal{I}}, \mathbf{x}_{\mathcal{U}}; \boldsymbol{\lambda}_{\mathcal{I}}, \boldsymbol{\lambda}_{\mathcal{U}}) = \pi_1 \hat{f}_1(\mathbf{x}_{\mathcal{I}}, \mathbf{x}_{\mathcal{U}}) / (\pi_0 \hat{f}_0(\mathbf{x}_{\mathcal{I}}, \mathbf{x}_{\mathcal{U}}) + \pi_1 \hat{f}_1(\mathbf{x}_{\mathcal{I}}, \mathbf{x}_{\mathcal{U}}))$, and

$$\ell(\boldsymbol{\lambda}_{\mathcal{I}}, \boldsymbol{\lambda}_{\mathcal{U}}) = \sum_{i=1}^n \left\{ y_i \log(\tilde{P}(\mathbf{x}_{i\mathcal{I}}, \mathbf{x}_{i\mathcal{U}}; \boldsymbol{\lambda}_{\mathcal{I}}, \boldsymbol{\lambda}_{\mathcal{U}})) + (1 - y_i) \log(1 - \tilde{P}(\mathbf{x}_{i\mathcal{I}}, \mathbf{x}_{i\mathcal{U}}; \boldsymbol{\lambda}_{\mathcal{I}}, \boldsymbol{\lambda}_{\mathcal{U}})) \right\}.$$

For any $\lambda = (\lambda_1, \lambda_2, \dots, \lambda_d)^T$, denote $\mathcal{A}(\lambda) = \{j : \lambda_j > 0\}$. For any vector $\mathbf{x}$ and set $\mathcal{A}$, denote $\mathbf{x}_{\mathcal{A}}$ to be the subvector of $\mathbf{x}$ with indices in $\mathcal{A}$. We use this notation often. For example, $\mathbf{s}_{\mathcal{B}}$ and $\mathbf{c}_{\mathcal{B}}$ are subvectors of $\mathbf{s}$ and $\mathbf{c}$ with indices in $\mathcal{B}$, and so on. By slight abuse of notation, let $f_k(\mathbf{x}_{\mathcal{A}})$ be the conditional density function of $\mathbf{X}_{\mathcal{A}}$ given $Y = k$ for $k = 0, 1$. For a symmetric kernel $K(\cdot)$, denote $\mu_2 = \int t^2 K(t) dt$ and $\nu_0 = \int K^2(t) dt$.

Lemma 1: Consider an index subset $\mathcal{A} \subset \{1, 2, \dots, d\}$. If $\lambda_j \rightarrow \infty$ as $\min_k \{n_k\} \rightarrow \infty$ when $j \in \mathcal{A}$ and $\lambda_j = 0$ otherwise, and $\max_k \{\prod_{j \in \mathcal{A}} \lambda_j / n_k\} \rightarrow 0$, then

$\hat{f}_k(\mathbf{x}_{\mathcal{A}}) = \frac{1}{n_k} \sum_{i: y_i = k} \prod_{j \in \mathcal{A}} \lambda_j K(\lambda_j (x_j - x_{ij}))$ is a consistent estimator of $f_k(\mathbf{x}_{\mathcal{A}})$ for $k = 0, 1$.

Asymptotically, the bias and variance are given by

$$E\hat{f}_k(\mathbf{x}_{\mathcal{A}}) - f_k(\mathbf{x}_{\mathcal{A}}) = \frac{\mu_2}{2} \sum_{j \in \mathcal{A}} f_{k,jj}^{(2)}(\mathbf{x}_{\mathcal{A}}) / (\lambda_j^2) + o(\sum_{j \in \mathcal{A}} \lambda_j^{-2}) \text{ and}$$

$$\text{Var}(\hat{f}_k(\mathbf{x}_{\mathcal{A}})) = f_k(\mathbf{x}_{\mathcal{A}}) \prod_{j \in \mathcal{A}} (\nu_0 \lambda_j) / n_k + o(\frac{1}{n_k} \prod_{j \in \mathcal{A}} \lambda_j).$$

Lemma 1 shows that the estimator for the conditional density is consistent when all nonzero elements of λ diverge to infinity when $n \rightarrow \infty$. We establish next that the estimator of the conditional density is not consistent when some elements of λ converges to finite positive numbers while $n \rightarrow \infty$.

Lemma 2: Consider subsets $\mathcal{A}$ and $\mathcal{B}$ of $\{1, 2, \dots, d\}$. If $\lambda_j \rightarrow \infty$ as $n \rightarrow \infty$ when $j \in \mathcal{A}$, $\lambda_j \rightarrow c_j$ for some constant $c_j > 0$ when $j \in \mathcal{B}$, and $\lambda_j = 0$ otherwise,

$\hat{f}_k(\mathbf{x}_{\mathcal{A}}, \mathbf{x}_{\mathcal{B}}) = \frac{1}{n_k} \sum_{i: y_i = k} \prod_{j \in \mathcal{A} \cup \mathcal{B}} \lambda_j K(\lambda_j (x_j - x_{ij}))$ is not a consistent estimator of $f_k(\mathbf{x}_{\mathcal{A}}, \mathbf{x}_{\mathcal{B}})$

for $k = 0, 1$. Asymptotically, we have

$$\begin{aligned} E\hat{f}_k(\mathbf{x}_{\mathcal{A}}, \mathbf{x}_{\mathcal{B}}) &= \int \mathbf{s}_{\mathcal{B}} \left(\prod_{j \in \mathcal{B}} K(s_j) \right) f_k(\mathbf{x}_{\mathcal{A}}, \mathbf{x}_{\mathcal{B}} \\ &\quad + \mathcal{D}_{\{\mathbf{c}_{\mathcal{B}}\}} \mathbf{s}_{\mathcal{B}}) d\mathbf{s}_{\mathcal{B}} + \frac{\mu_2}{2} \sum_{j \in \mathcal{A}} \frac{1}{\lambda_j^2} \int \mathbf{s}_{\mathcal{B}} \left(\prod_{j \in \mathcal{B}} K(s_j) \right) f_{k,jj}^{(2)}(\mathbf{x}_{\mathcal{A}}, \mathbf{x}_{\mathcal{B}} \\ &\quad + \mathcal{D}_{\{\mathbf{c}_{\mathcal{B}}\}} \mathbf{s}_{\mathcal{B}}) d\mathbf{s}_{\mathcal{B}} + o(1 \\ &\quad + \sum_{j \in \mathcal{A}} \lambda_j^{-2}) \end{aligned} \quad \text{and}$$

$$\text{Var}(f_k(\mathbf{x}_{\mathcal{A}})) = \prod_{j \in \mathcal{B}} (\lambda_j) \left(\int \mathbf{s}_{\mathcal{B}} \left(\prod_{j \in \mathcal{B}} K(s_j) \right) f_k(\mathbf{x}_{\mathcal{A}}, \mathbf{x}_{\mathcal{B}} + \mathcal{D}_{\{\mathbf{c}_{\mathcal{B}}\}} \mathbf{s}_{\mathcal{B}}) d\mathbf{s}_{\mathcal{B}} \right) \prod_{j \in \mathcal{A}} (\nu_0 \lambda_j) / n_k + o(1 + \frac{1}{n_k} \prod_{j \in \mathcal{A}} \lambda_j).$$

The extra 1 in the little- o terms is due to the assumption that $\lambda_j \rightarrow c_j > 0$ for $j \in \mathcal{B}$.

Theorem 1: Assume that the domain $\mathcal{X}$ is compact and $\tau \rightarrow \infty$ and $\tau^d/n \rightarrow 0$ as $n \rightarrow \infty$ in (13). Then the optimizer $\hat{\lambda} = (\hat{\lambda}_1, \hat{\lambda}_2, \dots, \hat{\lambda}_d)^T$ is such that $\hat{\lambda}_j \rightarrow \infty$ for $j \in \mathcal{I}$ and $\hat{\lambda}_j \rightarrow 0$ for $j \in \mathcal{U}$.

Corollary 1: Assume that the domain $\mathcal{X}$ is compact and $\tau \rightarrow \infty$ and $\tau^d/n \rightarrow 0$ as $n \rightarrow \infty$ in (14). Then the optimizer $\hat{\lambda} = (\hat{\lambda}_1, \hat{\lambda}_2, \dots, \hat{\lambda}_d)^T$ satisfies that $\hat{\lambda}_j \rightarrow \infty$ for $j \in \mathcal{I}$ and $\hat{\lambda}_j \rightarrow 0$ for $j \in \mathcal{U}$.

Note. Because (13) and (14) are not convex in λ establishing asymptotic properties other than consistency is more involved and we leave that for future work.

5 Summary

We introduced a new approach to variable selection that adapts naturally to nonparametric models. The method is derived from a measurement error model likelihood, and exploits the fact that a variable containing a lot of measurement error is not useful for modeling. When the approach is applied to linear regression, the resulting new method is closely related to the LASSO, yielding a novel interpretation for the LASSO and lending credibility to the new MEM selection likelihood approach. When the approach is used for variable selection in nonparametric classification, a new method is obtained, *sparse kernel discrimination analysis* (SKDA). SKDA was shown to be competitive with existing methods in the case that the true classifier is linear (the case for which the existing methods were derived), and generally superior in cases where the true classifier is nonlinear.

In Section 3.2 we showed that for linear regression with λ_j parameterized in terms of precisions, MESSO optimization is convex and amenable to efficient computation algorithms. For other cases that of interest MESSO optimization is not convex. For the simulation results and examples in this paper we used the MATLAB constrained minimization function *fmincon*.² We have also had good results with coordinate descent (modified to handle the sum-to- τ constraint) and a standard BFGS iteration after reparametrization to eliminate constraints, e.g., setting $\lambda_j = \tau \eta_j^2 / (\eta_1^2 + \dots + \eta_p^2)$. A web search of “*parallel coordinate descent*” returns numerous recent technical reports describing promising modifications of coordinate descent that take advantage of parallel processing. We are investigating the use of these algorithms to enable MESSO to be routinely used with even larger-dimension problems than those studied in Section 4.3.2 ($p = 50, 100$).

In 1986 R. Prentice wrote “There appears to be a dirth [sic] of realistic covariate measurement error models that lead to explicit forms ...” for distributions of the observed data in a measurement error model (Prentice, 1986). The same is true today and thus the construction of a MEM Selection Likelihood as defined in Section 2.3 is not always straightforward. A popular solution to this dilemma in the MEM literature is *regression calibration* (Carroll et al., 2006). It provides a ready-made (approximate) solution to the construction of the MEM likelihoods (Step 2 of the algorithm in Section 2.3). Our initial attempts at using regression calibration to implement our MEM-based variable selection in generalized linear models are promising. Additionally, unlike in 1986, it is now feasible to compute likelihoods even when explicit forms are not available. Thus we anticipate that a second route to realizing the full potential of the MEM-based approach to variable selection will likely entail numerical (Monte Carlo) evaluation of MEM likelihoods.

Acknowledgments

The authors thank the entire review team and especially the Associate Editor and Editor for suggestions and comments that lead to substantial improvements in the paper. K. White was funded by NIH training grant

²MATLAB and Statistics Toolbox Release 2012b, The MathWorks, Inc., Natick, Massachusetts, United States.

T32HL079896 and NSF grant DMS-1055210; Y. Wu by NSF grant DMS-1055210 and NIH/NCI grant R01-CA149569; and L. Stefanski by NIH grants R01CA085848 and P01CA142538, and NSF grant DMS-0906421.

References

- Anderson, TW. An Introduction to Multivariate Statistical Analysis. 3. New Jersey: Wiley-Interscience; 2003.
- Bazaraa, MS.; Sherali, HDMSC. Nonlinear Programming: Theory and Algorithms. New Jersey: Wiley-Interscience; 2006.
- Bickel PJ, Levina E. Some theory for Fisher's linear discriminant function, 'naive Bayes', and some alternatives when there are many more variables than observations. *Bernoulli*. 2004; 10:989–1010.
- Buonaccorsi, JP. Measurement Error: Models, Methods, and Applications. Chapman and Hall/CRC; New York: 2010.
- Cai T, Liu W. A Direct Estimation Approach to Sparse Linear Discriminant Analysis. *Journal of the American Statistical Association*. 2011; 106:1566–1577.
- Carroll, RJ.; Ruppert, D.; Stefanski, LA.; Crainiceanu, C. Measurement Error in Nonlinear Models: A Modern Perspective. 2. Chapman and Hall; London: 2006.
- Cooley CA, MacEachern SN. Classification via Kernel Product Estimators. *Biometrika*. 1998; 85:823–833.
- Cristianini, N.; Shawe-Taylor, J. An introduction to support vector machines and other kernel-based learning methods. Cambridge University Press; 2000.
- Duda, R.; Hart, P.; Stork, DG. Pattern Classification. New York: Wiley; 2000.
- Fan J, Fan Y. High dimensional classification using features annealed independence rules. *The Annals of Statistics*. 2008; 36:2605–2637.
- Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of American Statistical Association*. 2001; 96:1348–1360.
- Fisher RA. The use of multiple measurements in taxonomic problems. *Annals Eugen*. 1936; 7:179–188.
- Freund Y, Schapire R. A decision theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*. 1997; 55:119–139.
- Fuller, WA. Measurement Error Models. John Wiley; New York: 1987.
- Ghosh AK, Chaudhuri P, Sengupta D. Classification Using Kernel Density Estimates. *Technometrics*. 2006; 48:120–132.
- Hall P, Titterton DM, Xue JH. Median-Based Classifiers for High-Dimensional Data. *Journal of the American Statistical Association*. 2009; 104:1597–1608.
- Hansen BE. Uniform Convergence Rates for Kernel Estimation with Dependent Data. *Econometric Theory*. 2008; 24:726–748.
- Hastie, T.; Tibshirani, R.; Friedman, J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer-Verlag; New York: 2001.
- Hoerl AE, Kennard RW. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*. 1970; 12:55–67.
- Li, Q.; Racine, J. Nonparametric Econometrics: Theory and Practice. Princeton University Press; 2007.
- Mai Q, Zou H, Yuan M. A Direct Approach to Sparse Discriminant Analysis in Ultra-high Dimensions. *Biometrika*. 2012 to appear.
- McCullagh, P.; Nelder, J. Generalized Linear Models. Chapman & Hall/CRC; 1989.
- Prentice R. Binary Regression Using an Extended Beta-Binomial Distribution, With Discussion of Correlation Induced by Covariate Measurement Errors. *Journal of the American Statistical Association*. 1986; 81:321–327.
- Racine J, Li Q. Nonparametric estimation of regression functions with both categorical and continuous data. *Journal of Econometrics*. 2004; 119:99–130.
- Seo B, Lindsay BG. A universally consistent modification of maximum likelihood. *Statistica Sinica*. 2011 preprint. 10.5705/ss.2011.255

- Shao J, Wang Y, Deng X, Wang S. Sparse Linear Discriminant Analysis With High Dimensional Data. *The Annals of Statistics*. 2011; 39:1241–1265.
- Thoma, ED.; Squier, BC.; Olson, D.; Eisele, AP.; DeWees, JM.; Segall, RR.; Amin, MS.; Modrak, M. Assessment of methane and voc emissions from select upstream oil and gas production operations using remote measurements, interim report on recent survey studies. Proceedings of 105th Annual Conference of the Air & Waste Management Association, Control No. 2012-A-21-AWMA; 2012. p. 298-312.
- Tibshirani R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*. 1996; 58:267–288.
- Vapnik, V. *The Nature of Statistical Learning Theory*. New York: Springer; 1995.
- Wand, MP.; Jones, MC. *Kernel Smoothing*. London: Chapman & Hall/CRC; 1995.
- Witten D, Tibshirani R. Covariance-Regularized Regression and Classification for High Dimensional Problems. *Journal of the Royal Statistical Society, Ser B*. 2009; 71:615–636.
- Zhang HH. Variable selection for support vector machines via smoothing spline ANOVA. *Statistica Sinica*. 2006; 16:659–674.

Figure 1.

Plot of relative deviations from the average total variance versus total squared bias. LASSO (red-dashed); MESSO estimator from $\hat{L}_{SEL1}$ (solid blue); full-conditional MESSO estimator from $\hat{L}_{SEL1} \hat{L}_{SEL2}$ (black dashed); and the full-unconditional MESSO estimator from $\hat{L}_{SEL1} \hat{L}_{SEL2} \hat{L}_{SEL3}$ (green dot-dash). Because of their equivalency the $\hat{L}_{SEL1}$ -MESSO and LASSO lines overlap and appear as a single alternating red/blue dashed line.

Figure 2.

Plots of typical training samples and boxplots of test errors for data generated with $\rho = 0$ (top), 0.3 (middle), and 0.6 (bottom), from Section 4.3.1.

Figure 3.

Plot of one training sample (left) and boxplots of test errors (right) for the study in Section 4.3.2 with $p = 10$ and three values of ρ : 0 (top), 0.3 (middle), and 0.6 (bottom).

Figure 4.

Plot of one training sample (left) and boxplots of test errors (right) for the study in Section 4.3.2 with $p = 50$ and three values of ρ : 0 (top), 0.3 (middle), and 0.6 (bottom).

Figure 5.
Plot of one training sample (left) and boxplots of test errors (right) for the study in Section 4.3.3.

Figure 6.

Plot of a training sample (left) and boxplots of test errors (right) for the discrete predictor simulation study in Section 4.3.3.

Figure 7. Plot of one training sample (left) and boxplots of test errors (right) for the study in Section 4.3.4.

Figure 8.
Boxplots of test errors for the WBCD example

Figure 9.
Chemical data heat map. First 50 columns (left to right) are the “ $Y = 0$ ” group. Columns within groups randomly ordered. Rows identify compounds $X_1, \dots, X_{38}$.

Table 1
Predictor selection frequencies for SKDA and SLDA for the study in Section 4.3.1.

ρ		Frequency									
		X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
0	SKDA	100	100	10	14	6	15	11	9	9	6
	SLDA	100	100	1	1	3	1	2	2	3	3
	SCADlogist	100	100	13	11	10	10	7	12	12	11
	SCADsvm	100	100	13	14	11	15	18	12	9	10
0.3	SKDA	100	100	7	15	16	10	10	7	8	10
	SLDA	100	100	2	0	0	1	1	0	0	0
	SCADlogist	100	100	10	12	7	5	6	10	11	14
	SCADsvm	100	100	6	6	5	5	9	5	7	9
0.6	SKDA	100	100	16	12	9	8	6	14	13	10
	SLDA	100	100	0	0	0	0	0	0	0	0
	SCADlogist	100	100	0	0	1	0	2	1	0	3
	SCADsvm	100	100	0	2	1	1	0	2	0	3

Table 2
Predictor selection frequencies for SKDA and SLDA for the study in Section 4.3.2.

ρ		Frequency									
		X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
0	SKDA	100	100	9	9	8	6	14	10	7	8
	SLDA	0	0	5	4	3	5	6	6	4	8
	SCADlogist	0	0	8	10	9	7	12	11	8	11
	SCADsvm	38	38	50	54	55	56	59	56	53	57
0.3	SKDA	100	100	3	1	3	5	8	6	4	5
	SLDA	0	0	4	5	5	3	6	6	6	9
	SCADlogist	4	0	4	11	10	5	8	11	12	9
	SCADsvm	46	39	52	52	51	54	56	60	60	60
0.6	SKDA	100	100	6	0	4	0	1	2	3	4
	SLDA	2	1	4	5	7	3	4	3	7	6
	SCADlogist	4	6	6	7	9	6	10	4	13	7
	SCADsvm	51	54	46	52	52	53	56	56	60	59

Table 3

Frequency of X_1 and X_2 and the average frequency of $X_3, \dots, X_{50}$ being selected by SKDA and SLDA

ρ	Method	(Average) Frequency			
		X_1	X_2	$X_3, \dots, X_{50}$	
0	SKDA	100	100	2.5	
	SLDA	1	0	4.2	
	SCADlogist	0	0	4.8	
	SCADsvm	31	23	44.6	
0.3	SKDA	100	100	1.5	
	SLDA	1	0	4.4	
	SCADlogist	1	1	4.6	
	SCADsvm	28	37	41.8	
0.6	SKDA	100	100	6.0	
	SLDA	1	2	2.1	
	SCADlogist	2	0	2.2	
	SCADsvm	47	43	45.7	

Table 4

Predictor selection frequencies for MSKDA for the study in Section 4.3.3.

ρ	Frequency									
	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
0	100	100	1	1	1	0	1	1	0	0
0.3	100	100	0	0	0	1	0	0	0	0
0.6	100	100	1	0	0	0	0	0	0	1

Table 5

Predictor selection frequencies for MSKDA for the study in Section 4.3.4.

ρ	Frequency									
	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
0	100	100	0	0	0	0	0	0	0	0
0.3	100	100	0	0	0	0	0	0	0	0
0.6	100	100	0	0	1	0	0	0	1	3

Table 6

Test error summary for 40 random splits of the WBCD data.

	LDA	SLDA	KDA	SKDA	SCADlogit	SCADsym
Test Error Average	0.049	0.130	0.043	0.045	0.045	0.038
Test Error Std Dev	0.011	0.020	0.010	0.011	0.014	0.010

Table 7

Predictor selection frequencies for SLDA and SKDA for the WBCD data example.

X_j	SLDA	SKDA	SCADlogist	SCADsvm
X_1	5	0	0	0
X_2	14	3	4	4
X_3	0	0	0	0
X_4	1	0	0	0
X_5	2	0	0	0
X_6	18*	0	0	1
X_7	2	0	2	7
X_8	30*	6	15*	24*
X_9	9	0	0	1
X_{10}	22	0	1	2*
X_{11}	26*	7	35*	17*
X_{12}	10	0	2	2
X_{13}	2	1	0	0
X_{14}	21*	0	0	0
X_{15}	32*	0	1	4*
X_{16}	19	0	3*	4
X_{17}	12*	0	0	0
X_{18}	4	0	0	0
X_{19}	10	0	0	0
X_{20}	7	0	7	5
X_{21}	38*	40*	32*	40*
X_{22}	25*	37*	36*	37*
X_{23}	1	0	4	6
X_{24}	9*	0	8*	1
X_{25}	14	14*	6	29*
X_{26}	1	0	0	0
X_{27}	18*	2	9*	14*
X_{28}	32*	28*	25*	24
X_{29}	30*	13	7	27*
X_{30}	26*	0	2	0

Asterisks indicate variables selected using the entire data set.

Table 8

Test error summary for 40 random splits of the chemical signature data.

	LDA	SLDA	KDA	SKDA	SCADlogit	SCADsvm
Test Error Average	0.15	0.38	0.17	0.02	0.31	0.02
Test Error Std Dev	0.06	0.07	0.06	0.03	0.03	0.03

Table 9

Predictor selection frequencies for SLDA and SKDA for the chemical signature data.

X_j	SLDA	SKDA	SCADlogist	SCADsvm
X_1	26*	3	0	0
X_2	40*	40*	40*	40*
X_3	28*	0	0	0
X_4	29*	8	2	24*
X_5	23*	33*	15*	18
X_6	13*	0	0	0
X_7	19*	0	0	0
X_8	21*	0	0	0
X_9	25*	0	0	0
X_{10}	20*	0	0	0
X_{11}	21	0	0	1
X_{12}	14*	0	0	0
X_{13}	10	1	0	0
X_{14}	4	0	0	0
X_{15}	16*	0	0	0
X_{16}	10	0	0	0
X_{17}	19*	2	0	1
X_{18}	15	0	0	0
X_{19}	10*	0	0	0
X_{20}	11	0	0	0
X_{21}	22*	0	0	0
X_{22}	3	0	0	0
X_{23}	10*	0	0	0
X_{24}	23*	0	0	0
X_{25}	11*	0	0	0
X_{26}	9	0	0	0
X_{27}	10	0	0	0
X_{28}	19*	0	0	0
X_{29}	23*	0	0	0
X_{30}	14*	0	0	0
X_{31}	24*	0	0	0
X_{32}	20	0	0	1
X_{33}	17*	5	0	4
X_{34}	22*	0	0	2
X_{35}	20*	3	1	10
X_{36}	18*	0	0	0

X_j	SLDA	SKDA	SCADlogist	SCADsvm
X_{37}	14*	0	0	0
X_{38}	16*	0	0	0

Asterisks indicate variables selected using the entire data set.