A model free approach to combining biomarkers

Ruth M. Pfeiffer*[1] and **Efstathia Bura**[2]

[1] Biostatistics Branch, National Cancer Institute, 6120 Executive Blvd, EPS/8030, Bethesda, MD 20892

[2] Department of Statistics, George Washington University, Washington, DC 20052

Summary

For most diseases, single biomarkers do not have adequate sensitivity or specificity for practical purposes. We present an approach to combine several biomarkers into a composite marker score without assuming a model for the distribution of the predictors. Using sufficient dimension reduction techniques, we replace the original markers with a lower-dimensional version, obtained through linear transformations of markers that contain sufficient information for regression of the predictors on the outcome. We combine the linear transformations using their asymptotic properties into a scalar diagnostic score via the likelihood ratio statistic. The performance of this score is assessed by the area under the receiver-operator characteristics curve (ROC), a popular summary measure of the discriminatory ability of a single continuous diagnostic marker for binary disease outcomes. An asymptotic chi-squared test for assessing individual biomarker contribution to the diagnostic score is also derived.

Key words: Dimension Reduction; AUC; SAVE; SIR; Singular value decomposition; NHANES III.

1 Introduction

Much research effort is devoted to searching for predictors, also called markers, that may aid in diagnosis of disease or physical impairment. Although a number of criteria have been advocated for evaluating screening tests (Gastwirth, 1987), much attention has been devoted to discriminatory accuracy, which is

* Corresponding author: Ruth M. Pfeiffer e-mail: pfeiffer@mail.nih.gov, Phone: +001 301 594 7832, Fax: +001 301 402 0081

defined in terms of test sensitivity and specificity. Ideally one would obtain a single marker with very high specificity and sensitivity. However, such high performance markers are yet to be found for many diseases. Strategies for combining information from multiple diagnostic predictors are needed, since a combination may provide a better tool for diagnosis or screening applications than any single marker on its own. For binary disease outcomes, McIntosh and Pepe (2002) showed that if the joint distribution of all the markers in a panel in both the diseased (case) and the non-diseased (control) populations is known, the likelihood ratio (LR) statistic provides the most powerful means of combining the markers into a scalar diagnostic score for discriminating between the diseased and non-diseased states. However, the LR approach requires knowledge of the joint marker distribution and may be sensitive to violations of the distributional assumptions. Many markers, including biochemical measurements based on serum, are continuous, and specifying their joint distribution can be challenging.

In this paper, an approach to combining multiple continuous predictors is proposed that provides optimal or near optimal results in many settings without requiring the predictor distribution of cases and controls be known. We start with a sufficient number of linear combinations (projections) of markers obtained from two sufficient dimension reduction (SDR) methods, Sliced Inverse Regression (SIR; Li 1999) and Sliced Average Variance Estimation (SAVE; Cook and Weisberg, 1991). Neither SIR nor SAVE require the specification of a model for the relationship between the markers and the outcome. These linear combinations, which retain the discrimination potential of the original predictors, are independent and approximately normally distributed. If more than one linear combination is needed for optimal discrimination, we use this fact to further combine them into a scalar diagnostic score for discrimination via the LR statistic in section 3. If a single linear combination contains sufficient discriminatory information, we use it directly to classify individuals. The key advantage of using the SIR or SAVE projections in the LR statistic, instead of the markers directly, is that except for two moment conditions, no other distributional assumption for the markers is needed.

The LR statistic is optimal for assessing many aspects of discrimination, including the expected misclassification rate (Anderson, 1984, Chap. 6), and the discriminatory accuracy, measured by the area under the receiver-operator characteristics (ROC) curve. The ROC curve plots sensitivity against (1-specificity) (true vs. false positivity) for all thresholds that can be used to define ”test positive.” Two diagnostic tests can be compared by calculating the difference between the areas under their two ROC curves (AUCs), with the larger area corresponding to the ”better” test. McIntosh and Pepe (2002) showed that the ROC curve for the LR statistic is uniformly above the ROC curve for all other scalar functions of the markers, and thus maximizes the AUC. Assuming multivariate normality of the markers, several authors have derived diagnostic scores consisting of a single linear combination of the marker values by directly maximizing the AUC (Su and Liu, 1993), or the partial area under the curve (McGlish, 1989), i.e. sensitivity over a range of specificities (Liu, Schisterman and Zhu, 2005). In section 3 we show that under multivariate normality, our approach using SIR, a first order moment method, results in the same linear combination that Su and Liu (1993) obtained, and our approach using SAVE captures all the discriminatory information in the LR statistic. The methods developed in this paper performed well compared to LR, even with non-normally distributed data.

We also propose a test statistic to assess which of the original markers are not statistically significant contributors to the SIR or SAVE predictors, in Section 4. This test overcomes a second limitation of using the LR statistic to combine markers, namely that it does not allow the assessment of the contributions of individual predictors to discrimination.

Following the presentation of background material (Section 2) and methods (Sections 3 and 4), we carry out simulations to assess the performance of the proposed scalar discrimination scores based on SAVE and SIR and of the test for marker contributions to the diagnostic score (Section 5). Two data examples are given in Section 6, and we conclude with a discussion in Section 7.

2 A Brief Description of SIR and SAVE

2.1 Notation and definition of SIR and SAVE

Let Y be a response variable and $X = (X_1, \ldots, X_p)^T \in R^p$, a vector of continuous markers. In sufficient dimension reduction methods via inverse regression, X is replaced with a lower-dimensional projection $P_S X$ without loss of information about the conditional distribution of Y given X, i.e. $F(Y|X) = F(Y|P_S X)$, and without requiring a pre-specified model for $Y|X$. $F(\cdot|\cdot)$ is the conditional distribution function of Y given the second argument. P_S is the orthogonal projection onto the vector space S in the usual inner product. The intersection of all subspaces $S^* \subseteq R^p$ that satisfy $F(Y|X) = F(Y|P_{S^*}X)$ is denoted by $S_{Y|X}$ and called the central dimension reduction subspace (Cook, 1998). The dimension $s = \dim(S_{Y|X})$, referred to as the structural dimension of the regression of Y on X, can take on any value in the set $\{0, 1, \ldots, p\}$. When $s < p$, the structural dimension of the regression is smaller than the number of predictors. If $\eta = (\eta_1, \ldots, \eta_s)$ is a basis for $S_{Y|X}$, then $P_\eta X$, or equivalently, the s linear combinations $\eta^T X = (\eta_1^T X, \ldots, \eta_s^T X)$, contain all the information in X about Y. A detailed exposition of sufficient dimension reduction methods is provided in Cook (1998).

The estimation of $S_{Y|X}$ is based on finding a matrix Ω_x, called a kernel matrix, so that span$(\Omega_x) \subseteq S_{Y|X}$. For numerical stability, and without loss of generality, the standardized predictors $Z = \Sigma_x^{-1/2}(X - \mathrm{E}(X))$ are commonly used to compute the kernel matrix. First moment methods such as SIR (Li, 1991) use kernel matrices computed from first moments, whereas second moment methods, such as SAVE (Cook and Weisberg, 1991), use second moment based kernels. SAVE is the most comprehensive sufficient dimension reduction method. For SIR, the kernel matrix is $\Omega_z = \mathrm{cov}(\mathrm{E}(Z|Y))$, and for the original X predictors it is $\Omega_x = \Sigma_x^{-1/2}\Omega_z$. For binary outcome Y, the SIR kernel matrix is equivalent to

$$\Omega_{SIR} = (\mathrm{E}(Z|Y = 0) - \mathrm{E}(Z|Y = 1)), \tag{1}$$

a $p \times 1$ vector (Cook and Lee, 1999). The SAVE kernel matrix is $\Omega_z = \mathrm{E}(I - \mathrm{cov}(Z|Y))^2$ (Cook and Weisberg, 1991), and for the original X predictors is $\Omega_x = \Sigma_x^{1/2}\Omega_z H$, where

$$H = \begin{pmatrix} \Sigma_x^{1/2} & \mathbf{0} \\ 0 & 1 \end{pmatrix}. \tag{2}$$

For binary outcomes Y, Cook and Lee (1999) showed that the SAVE kernel matrix is equivalent to the $p \times (p+1)$ matrix

$$\Omega_{SAVE} = (\mathrm{E}(Z|Y=0) - \mathrm{E}(Z|Y=1), \mathrm{cov}(Z|Y=0) - \mathrm{cov}(Z|Y=1)). \tag{3}$$

For SIR, the condition needed for span$(\Omega_x) \subseteq S_{Y|X}$ is that $\mathrm{E}(X|\gamma^T X)$ be linear in $\gamma^T X$. For SAVE, in addition, it is required that the conditional variance cov$(X|\gamma^T X)$ be constant. Both conditions refer to the marginal distribution of X and are trivially satisfied when X is normally distributed. The linearity condition is satisfied by all elliptically contoured distributions. Both conditions can be empirically checked by considering the scatterplot matrix of the predictors, i.e. the matrix of all the pairwise scatter plots of $X_1, \ldots, X_p$. Linearity of $\mathrm{E}(X|\gamma^T X)$ implies that the scatterplots look random or linear, and homoscedasticity implies that there are no pronounced fluctuations in data density.

For binary outcomes Y, and when $X|Y$ is normally distributed, SAVE always estimates the entire central dimension reduction subspace $S_{Y|X}$ resulting in no loss of information (Cook and Lee, 1999).

2.2 Projecting X onto span(Ω_{SIR}) or span(Ω_{SAVE}) to construct independent predictors X^*

For binary outcomes Y, the projection of the original markers X onto span(Ω_{SIR}) is found by simply computing $X_1^* = \Sigma_x^{-1}(\mathrm{E}(X|Y=1) - \mathrm{E}(X|Y=0))X$. When $X|Y$ is normally distributed and cov$(X|Y=0) =$ cov$(X|Y=1)$, SIR and Linear Discriminant Analysis (LDA) estimate the same discriminant linear combination of the predictors. However, SIR is more general than LDA, as it does not require cov$(X|Y=0) =$ cov$(X|Y=1)$. SIR also yields the same linear combination as logistic regression models that include the predictors as main effects (see Cook, 1998, Prop. 8.1). In binary regression, both SIR and LDA estimate at most a single direction in $S_{Y|X}$ and may miss information that could improve discriminatory accuracy.

For SAVE finding the projections on span(Ω_{SAVE}) is slightly more involved. Let $d = \text{rank}(\Omega_{SAVE})$. The singular value decomposition of Ω_{SAVE} is

$$\Omega_{SAVE} = \mathrm{U}_z^T \begin{pmatrix} \mathrm{D}_z & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \mathrm{R}_z, \tag{4}$$

where the orthogonal matrix $\mathrm{U}_z^T = (\mathrm{U}_{z1}, \mathrm{U}_{z0})$ is $p \times p$ with U_{z1}: $p \times d$, U_{z0}: $p \times (p - d)$, $\mathrm{D}_z = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_d)$ is a diagonal matrix of the descending singular values of Ω_{SAVE}, $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_d > 0$, and $\mathrm{R}_z^T = (\mathrm{R}_{z1}, \mathrm{R}_{z,0})$ is orthogonal with R_{z1}: $(p+1) \times d$, R_{z0}: $(p+1) \times (p - d + 1)$. The d left singular vectors $\mathrm{U}_{z1} = (U_1^{(1)}, \ldots, U_d^{(1)})$ of Ω_{SAVE} that correspond to its d non-zero singular values $\lambda_1 \geq \ldots \geq \lambda_d$ span Ω_{SAVE}. The SAVE predictors $(Z_1^*, \ldots, Z_d^*) = (U_1^{(1)T} Z, \ldots, U_d^{(1)T} Z)$ are the projections of Z onto span(Ω_{SAVE}) and contain part or all the information contained in the predictor vector Z for regressing Y on Z. The SAVE predictors on the X scale are $(X_1^*, \ldots, X_d^*) = (\Sigma_x^{-1/2} U_1^{(1)T} X, \ldots, \Sigma_x^{-1/2} U_d^{(1)T} X)$.

In implementing either SAVE or SIR, the conditional moments are replaced by the corresponding sample moments, $\bar{x}_y$ and $\widehat{\Sigma}_{x|y}$ for $y = 0, 1$, to yield $\hat{\Omega}_{SAVE} = (\widehat{\Sigma}_x^{-1/2}(\widehat{\Sigma}_{x|1} - \widehat{\Sigma}_{x|0})\widehat{\Sigma}_x^{-1/2}, \widehat{\Sigma}_x^{-1/2}(\bar{x}_1 - \bar{x}_0))$, and $\hat{\Omega}_{SIR} = (\widehat{\Sigma}_x^{-1/2}(\bar{x}_1 - \bar{x}_0))$. To infer the rank d, a test statistic based on the singular values of $\widehat{\Omega}_{SAVE}$ or $\widehat{\Omega}_{SIR}$ is typically used. Cook and Lee (1999) propose a test statistic for rank(Ω_{SAVE}) for binary Y and show that it has an asymptotic weighted chi-squared distribution under the linearity and constant variance condition. Approximate p-values can be obtained, for example, using a result by Wood (1989). The test statistic for rank(Ω_{SIR}) has an asymptotic chi-squared distribution for normal X (Li, 1991), and an asymptotic weighted chi-squared distribution under the linearity condition (Bura and Cook, 2001). In both cases, the estimation is carried out by sequentially testing $H_0 : d = k$ against $H_a : d > k$, starting at

$k = 0$, which corresponds to independence of Y and X, and adding unit increments to k until H_0 cannot be rejected at a pre-set α level.

3 Combining diagnostic markers using SIR and SAVE

Before we show how to derive a diagnostic score from the SIR and SAVE predictors, we define the area under the ROC curve, and discuss its relation to the LR statistic.

3.1 Assessing discriminatory accuracy of a diagnostic score via the AUC

As mentioned in the introduction, the discriminatory power of a diagnostic test for binary outcomes can be quantified by the ROC curve. For a threshold c, define ”test positive” for the continuous score T as $T \geq c$, with corresponding true and false positive fractions $TPF(c) = P(T \geq c|Y = 1)$ and $FPF(c) = P(T \geq c|Y = 0)$, respectively. The ROC curve is $ROC(.) = \{(FPF(c), TPF(c)), c \in (-\infty, \infty)\}$ or alternatively, with $t = FPF(c)$, $ROC(.) = \{(t, ROC(t), t \in (0, 1)\}$ (see also Pepe, 2003, page 67). The most widely used summary measure for the ROC curve is the area under the curve (AUC), $AUC = \int_0^1 ROC(t)dt$, with values varying between 0.5 and 1, where 1 corresponds to perfect discriminatory accuracy of the test and 0.5 to no discriminatory ability. The AUC can also be expressed as the probability that the scalar diagnostic score for a randomly selected case T_1, exceeds that for a randomly selected control T_0, i.e. $AUC = \mathrm{P}(T_1 > T_0)$. The AUC is invariant with respect to any monotonic transformation of T and can be estimated by the two-sample Mann-Whitney U-statistic.

McIntosh and Pepe (2002) showed that among all possible functions of the predictor vector X, the likelihood ratio function $LR(X) = \mathrm{P}(X|Y = 1)/\mathrm{P}(X|Y = 0)$ maximizes the AUC. Thus, if $\mathrm{P}(X|Y)$ is known for $Y = 0$ and $Y = 1$, the optimal composite marker score is given by the LR function.

3.2 Combining diagnostic markers using SIR

For binary outcomes, the SIR single linear combination of the markers, or the SAVE linear combination if rank(Ω_{SAVE}) is estimated to be one, can be used directly as a scalar diagnostic score, T. However, for SAVE a scalar needs to be derived if rank(Ω_{SAVE}) is estimated to be larger than one.

3.3 Relation of SIR to existing results on linear combinations of markers

For conditionally normally distributed $X|Y \sim MVN(\mu_{x|Y}, \Sigma_{x|Y})$,$Y = 0, 1$, Su and Liu (1993) found a linear marker combination by maximizing

$$AUC(a) = \mathrm{P}(a^T X_1 > a^T X_0) = \Phi\Big\{ \frac{a^T(\mu_{x|1} - \mu_{x|0})}{(a^T \Sigma_x a)^{1/2}} \Big\},$$

as a function of a, where Φ denotes the standard normal cumulative distribution function. The maximizer is $a^* = c\Sigma_x^{-1}(\mu_{x|1} - \mu_{x|0})$, with $\Sigma_x = \Sigma_{x|0} + \Sigma_{x|1}$ and c any arbitrary constant. The SIR linear combination (section 2) $X_1^* = \Sigma_x^{-1}(\mu_{x|1} - \mu_{x|0})X$ is proportional to the linear combination obtained by Su and Liu (1993), and consequently maximizes the AUC when the markers are normally distributed among cases and controls.

3.4 Combining diagnostic markers using SAVE

For conditionally normally distributed markers $X|Y \sim MVN(\mu_{x|Y}, \Sigma_{x|Y})$, the likelihood ratio satisfies

$$\log LR(X) \propto \frac{1}{2} X^T(\Sigma_{x|0}^{-1} - \Sigma_{x|1}^{-1})X + X^T(\Sigma_{x|1}^{-1}\mu_{x|1} - \Sigma_{x|0}^{-1}\mu_{x|0}).$$

The LR statistic for multivariate normal predictors is thus fully characterized by $(\Sigma_{x|0}^{-1} - \Sigma_{x|1}^{-1})$ and $(\Sigma_{x|1}^{-1}\mu_{x|1} - \Sigma_{x|0}^{-1}\mu_{x|0})$, which also determine Ω_{SAVE}. In consequence, the SAVE predictors completely capture the discriminatory information contained in the LR statistic. The following algorithm uses this fact to propose a scalar diagnostic score for the SAVE predictors.

1. We first estimate the dimension d of $\hat{\Omega}_{\text{SAVE}}$, $\hat{d}$, and then the linear combinations $(X_1^*, \ldots, X_{\hat{d}}^*) = (\hat{\Sigma}_x^{-1/2}\hat{U}_1^{(1)T}X, \ldots, \hat{\Sigma}_x^{-1/2}\hat{U}_{\hat{d}}^{(1)T}X)$ that constitute the estimated $\hat{d}$ SAVE predictors.

2. **For many data sets, most projections are approximately Gaussian** (Diaconis and Freedman, 1984; Hall and Li, 1993). We thus assume that the estimated SAVE predictors are approximately normally distributed, $X^* \sim N(\mu^*, \Sigma^*)$, regardless of the distribution of the original predictor vector X. By construction, the SAVE predictors are orthogonal, and thus independent with $\Sigma^* = \text{diag}(\sigma_i^*)$.

3. A diagnostic scalar score T for the SAVE predictors $X_1^*, \ldots, X_{\hat{d}}^*$ is the LR statistic based on the product of $\hat{d}$ independent univariate normal distributions. That is,

$$T(X^*) = LR(X^*) = \frac{\text{P}(X^*|Y=1)}{\text{P}(X^*|Y=0)} \approx \frac{\prod_{j=1}^{\hat{d}} \phi(X_j^*; \mu_{1j}^*, \sigma_{1j}^*)}{\prod_{j=1}^{\hat{d}} \phi(X_j^*; \mu_{0j}^*, \sigma_{0j}^*)}$$

where ϕ denotes the univariate normal density function and parameters are estimated by the first two sample moments of X^*.

When $\hat{d} = 1$, both SIR and SAVE can be used to obtain a linear combination of marker values. However, when the two populations differ only in their means, SIR is more accurate as it requires the estimation of fewer parameters.

4 Assessing marker contributions to the SAVE or SIR predictors

Variable selection is important when building classifiers, especially when the dimension of input variables is large, as the accuracy of a classifier that includes too many noise variables may not be satisfactory. We evaluate the contributions of markers to the composite score T by assessing their contribution to the SAVE or SIR linear combinations. For example, if marker X_k does not contain any information for discriminating cases and controls, then $\text{U}_{1k}^{(1)}, \ldots, \text{U}_{dk}^{(1)}$, i.e. its contribution to the SAVE predictors $(X_1^*, \ldots, X_d^*) = (\Sigma_x^{-1/2} U_1^{(1)T} X, \ldots, \Sigma_x^{-1/2} U_d^{(1)T} X)$, should be zero.

Let $\widehat{\Omega}_{SAVE} = \widehat{\Omega}_z = \widehat{\Sigma}_z^{1/2} \widehat{\Omega}_x \widehat{H}^{-1}$, where $\widehat{\Sigma}_x$, $\widehat{H}$ and $\widehat{\Omega}_x$ are the sample moment estimates of Σ_x, H and of $\Omega_x = (\text{E}(X|Y=0) - \text{E}(X|Y=1), \text{cov}(X|Y=0) - \text{cov}(X|Y=1))$. The singular value decomposition (SVD) of $\widehat{\Omega}_z$ is

$$\widehat{\Omega}_z = \widehat{\mathrm{U}}_z^T \begin{pmatrix} \widehat{\mathrm{D}}_{z1} & 0 \\ 0 & \widehat{\mathrm{D}}_{z0} \end{pmatrix} \widehat{\mathrm{R}}_z \tag{5}$$

with $\widehat{\mathrm{U}}_z^T = (\widehat{\mathrm{U}}_{z1}, \widehat{\mathrm{U}}_{z0})$, $\widehat{\mathrm{R}}_z^T = (\widehat{\mathrm{R}}_{z1}, \widehat{\mathrm{R}}_{z0})$, where the partition conforms to the SVD of Ω_z in (4). The diagonal matrices with the decreasing singular values of $\widehat{\Omega}_z$ are $\widehat{\mathrm{D}}_{z1} = \mathrm{diag}(\hat{\lambda}_1, \hat{\lambda}_2, \ldots, \hat{\lambda}_d)$ of order $d \times d$, and $\widehat{\mathrm{D}}_{z0} = \mathrm{diag}(\hat{\lambda}_{d+1}, \hat{\lambda}_{d+2}, \ldots, \hat{\lambda}_p)$ of order $(p-d) \times (p-d)$.

The asymptotic distribution of $\widehat{\mathrm{U}}_{z1} = (\hat{\mathrm{U}}_1^{(1)}, \ldots, \hat{\mathrm{U}}_d^{(1)})$ is

$$n^{1/2}\ \mathrm{vec}(\widehat{\mathrm{U}}_{z1} - \mathrm{U}_{z1}) \Longrightarrow N_{pd}(0, \Sigma_U), \tag{6}$$

where $\Sigma_U = (\mathrm{D}_z^{-1}\mathrm{R}_{z1}^T H^{-1} \otimes \Sigma_x^{-1/2}) V_x (H^{-1}\mathrm{R}_{z1}\mathrm{D}_z^{-1} \otimes \Sigma_x^{-1/2})$. The definition of V_x is provided in the appendix and details on its derivation can be found in Bura and Pfeiffer (2007). We assess the contribution of any given subset of the individual predictors $X_1, \ldots, X_p$ to the SAVE predictors $X_1^*, \ldots, X_d^*$ via testing a hypothesis of the form $C\mathrm{vec}(\mathrm{U}_{z1}) = 0$ for some $r \times pd$ matrix C of zeroes and ones. In this way we circumvent technical difficulties relating to the multiplicity of singular values. The rank r of C equals the number of the elements of U_{z1} set to zero.

From (6) we obtain

$$n^{1/2}\ C\ \mathrm{vec}(\widehat{\mathrm{U}}_{z1} - \mathrm{U}_{z1}) \Longrightarrow N_r(0, C\Sigma_U C^T) \tag{7}$$

A general Wald's type test for H_0: $C\mathrm{vec}(\mathrm{U}_{z1}) = 0$ is given next. Let C be an $r \times pd$ matrix of rank r, $\theta = \mathrm{vec}(C\mathrm{vec}(\mathrm{U}_{z1}))$ and $\widehat{\theta} = \mathrm{vec}(C\mathrm{vec}(\widehat{\mathrm{U}}_{z1}))$, both $rpd \times 1$ vectors. Then when $\theta = 0$ the following holds (see Bura and Pfeiffer, 2007):

a. If V_x is positive definite,

$$\widehat{T} = n\ \hat{\theta}^T (C\widehat{\Sigma}_U C^T)^{-1}\hat{\theta} \Longrightarrow \chi^2(r) \tag{8}$$

where $\widehat{\Sigma}_U$ is a consistent estimate of Σ_U.

b. If V_x is positive semidefinite with rank$(V_x) \geq r$ and rank$(\widehat{V}_x) \xrightarrow{p}$ rank(V_x),

$$\widehat{T} = n\, \hat{\theta}^T (C\widehat{\Sigma}_U C^T)^+ \hat{\theta} \Longrightarrow \chi^2(r) \tag{9}$$

where $\widehat{\Sigma}_U^+$ is a consistent estimate of Σ_U^+.

The notation $^+$ denotes the Moore-Penrose generalized matrix inverse. A test of marker contribution to the SIR linear combination is given in detail in Bura and Pfeiffer (2007).

5 Simulation Studies

In this simulation study the discriminatory performance of several composite diagnostic scores was assessed with respect to the AUC. We compared scores built using SIR, SAVE with estimated number of linear predictors $\hat{d}$ and with $d = 2$ and $d = 4$, to that of a score built by combining the markers directly in the LR statistic (McIntosh and Pepe, 2002). We also computed the AUC for the linear marker combination estimated from fitting logistic regression models that included the markers as main effects, and for some simulations, all pairwise marker interaction terms and quadratic terms. To obtain unbiased estimates of the AUC, we split the sample and used one half to estimate d and build the predictors, and the other half to estimate the AUC. In practical applications, v-fold cross validation may be more appropriate.

5.1 Results for normal markers

We generated samples of $p = 10, 20$ or 30 markers $X = (X_1, \ldots, X_p)$ from two normal populations, $(X|Y = i) \sim MVN(A\omega_i, A\Omega_i A^T), \quad i = 0, 1$. The parameters were chosen so that X satisfied a q-dimensional ($q \leq p$) discrimination subspace model (DSM(q)), that confined differences in p predictors to a linear subspace of dimension q based on the O'Neill model as described in Flury, Boukai and Flury (1997). For $Y = 0$, we let $\omega_0 = (0, \ldots, 0) \in R^p$, and $\Omega_0 = I_p$, the identity matrix, and for $Y = 1$, $\omega_1 = \sum_{j=1}^q \delta_j e_j$ and $\Omega_1 = I_p + \sum_{j=1}^q \tau_j e_j^T e_j$, for some $\delta \in$ R and $\tau > -1$, where e_j denotes the j-th unit basis vector. The matrix $A = (a_{ij})$ had elements $a_{ii} = 1$ and $a_{ij} = 0.5$. (Simulations with $a_{ij} = 0.1$ or $a_{ij} = 0.25$ gave very similar results and are not shown here).

For a model with $p = 20$ markers and the same mean and covariance matrices for both groups, $(X|Y = i) \sim MVN(A\omega, A\Omega A^T), \quad i = 0, 1$, the AUC estimates (with standard errors) were 0.51 (0.01) for LR, 0.51 (0.01) for SAVE with estimated dimension $\hat{d}$, 0.51 (0.01) for SAVE with fixed $d = 2$, 0.51 (0.01) for SAVE with fixed $d = 4$, 0.52 (0.01) for SIR and 0.52 (0.01) for logistic regression. All methods thus performed well when there was no difference in the distribution of the predictors between the two groups.

Table 1 presents means in 100 simulations for a $DMS(2)$ (dimension 2) model with $\delta_1 = 0.1, \delta_2 = 1$, for several values of τ_1 and τ_2 for $p = 10, 20$ and 30 markers. For 200 cases and controls and $\tau_1 = 1$ and $\tau_2 = 2$, so that the covariance matrices for both groups are fairly close, all methods resulted in similar AUC values for $p = 10, 20$ and $p = 30$ markers, with the score based on SAVE with $\hat{d}$ corresponding to a somewhat lower AUC than the score based on SAVE with $d = 2$ or $d = 4$. When τ_1 and τ_2 were 2 and 5 respectively, for $p = 10, 20$ and $p = 30$, and sample sizes of 200 and 500 cases and controls the SAVE scores with $\hat{d}, d = 2$ and $d = 4$ yielded AUC values close to the optimal LR-obtained AUCs, which were significantly higher then the AUCs of SIR or logistic regression scores. When the τ parameters were 5 and 10, the SAVE composite scores corresponded to AUC values of 0.89, compared to AUCs of around 0.63 for logistic regression and SIR, even for sample sizes of 200 cases and controls, as both these methods were unable to detect the second dimension arising from the rather pronounced differences in the covariance matrices of the predictors. We have already noted that for normally distributed markers, SIR and logistic regression produce the same linear combination, which results in identical predictive performance in Table 1.

5.2 Results for markers for cases simulated from mixtures of normals

The robustness of SAVE to non-normality was studied by drawing the markers for the cases from a two-component multivariate mixture model that had mean vectors with opposite signs, $(X|Y = 1) \sim 0.5MVN(A\omega_1, A\Omega_1 A^T) + 0.5MVN(-A\omega_1, A\Omega_1 A^T)$ and $(X|Y = 0) \sim MVN(0, A\Omega_0 A^T)$. Under this model the overall means among cases and controls were zero, but the dimension of the central subspace was equal to two. The results in Table 2 indicated that LR and SAVE performed very well; however, there was no discriminatory power in the linear combination obtained by SIR or in using the markers in a logistic regression model as there were no differences in marker means between cases and controls. For $p = 10$ markers, we also fitted logistic regression models that contained all pairwise interaction terms between the markers, which resulted in a total of 55 coefficients and thus could not be stably estimated with 200 cases and controls. For 500 cases and controls this model had very similar discriminatory performance to the SAVE scores. For $p = 20$ or $p = 30$ markers, however, including all pairwise interaction terms in addition to main effects in a logistic regression model was not feasible for sample sizes available in practice.

5.3 Results for markers simulated from multivariate t-distributions with three degrees of freedom

For distributions with heavy tails, such as Cauchy, low dimensional projections of high dimensional data can fail to be normally distributed (Diaconis and Freedman, 1984). We thus also studied the robustness of our procedure by simulating markers for the cases and controls from a multivariate t-distribution with three degrees of freedom. This distribution has heavy tails compared to the multivariate normal distribution, and also meets the moment conditions of SIR and SAVE. The parameters were chosen such that $d = 2$. The results in Table 3 indicate that SAVE performed very well compared to LR computed using the multivariate t-distribution with estimated moments and known degrees of freedom. However, their performance depends on how pronounced the second dimension is. For example, for $\tau_1 = 1.0$ and $\tau_2 = 2,0$ the second dimension did not add much to the discrimination, and LR and SAVE yielded similar results to SIR and logistic regression. However, when the second dimension was more pronounced ($\tau_1 = 2, \tau_2 = 5$ and $\tau_1 = 5, \tau_2 = 10$), i.e. the difference in the two groups was mostly characterized by difference in the two covariance matrices, the performance of LR and SAVE was superior to the performance of the first order methods.

5.4 Results for markers simulated from multivariate log-normal distributions

We also simulated data from a multivariate log-normal distribution to assess the performance of our procedure with highly skewed predictors (Table 4). The performance of SAVE was somewhat inferior to the performance of the LR statistic, computed with the correct distribution and estimated parameters. For example, for $\tau_1 = 2, \tau_2 = 5$ the AUC for LR was 0.69, while it was only 0.60 for SAVE with estimated dimension. However, this difference was not statistically significant. A log-transformation would yield normal data, and optimal performance for SAVE.

5.5 Testing which markers contribute significantly to prediction

To assess the performance of the Wald test, we first generated $p - 1$ predictors from a DSM(2) model and added the non-informative $X_p \sim N(0, 1)$ for both cases and controls. We tested the null hypothesis that X_p did not contribute to the SAVE predictors. Table 3 shows the results for various sample sizes and parameter choices for 500 simulations with known SAVE dimension. The nominal α level of 5% was included in all confidence intervals around the observed proportion of rejections of the null hypothesis (Table 4). To assess the power of the Wald test, we repeated the simulation with all p predictors generated from a DSM(2) model. Under this scenario the test with known d had 30% power for $\delta_1 = 0.1, \delta_2 = 0.1$ and $\tau_1 = 0$ and $\tau_2 = 2$ for $p = 10$ and $p = 20$ markers, and 62% for $p = 10$ and 68% for $p = 20$ when τ_1 and τ_2 were changed to 0.2 and 0.5, and 100% for $\tau_1 = 2$ and $\tau_2 = 5$ for both p. For $\delta_1 = 1$ and $\delta_2 = 0.1$ the power was 100% for all choices of τ we studied. Not surprisingly, better separation of the two groups in the two-dimensional subspace translated into higher power of the test.

The formulation of the test statistic assumes that the rank d of $\hat{\Omega}_{SAVE}$ is known. When d was estimated, and the difference in the covariance matrices between the groups was not large, the observed size of the test was somewhat larger than the nominal 5%, for example for $\delta_1 = 1, \delta_2 = 0.1, \tau_1 = 1$, and $\tau_2 = 1$ with 200 cases and controls, the observed size was 0.08 with 95% confidence interval $[0.05, 0.1]$. The size of the test was correct however, when d could be estimated accurately, for example, for large sample sizes, or large differences between covariance matrices.

6 Data examples

6.1 Fisher's iris data

We first analyzed a data set that has been extensively studied in the classification literature. The data consist of a sample of 150 irises, of three different types, Iris-setosa, Iris-verginica and Iris-versicolor. The predictors are petal length, petal width, sepal length, and sepal width. The lengths are measured in millimeters. As we focused on binary prediction, we combined Iris-setosa and Iris-virginica into a single group ($Y = 0$) of 100 observations and aimed to discriminate it from the 50 Iris-versicolor ($Y = 1$) observations. Visual inspection of the scatterplot matrix of the predictors revealed no pronounced violations of the linearity and constant variance assumptions so that both SIR and SAVE can be used.

In this example, when all the data were analyzed, the estimated dimension was $\hat{d} = 2$.

In an ideal setting an independent data set would be available for the purposes of assessing the performance of our predictor. As we did not have an independent test set, we used 10-fold cross validation to obtain unbiased estimates for the AUC. That is, we randomly assigned each of the observations in the two groups to one of 10 partitions, such that the partitions are size 5 for the $Y = 1$ and size 10 for the $Y = 0$ group. Subsequently we used all but one of the partitions to estimate the dimension of SAVE, built the SAVE, SIR and LR predictors, and evaluated the performance of our predictor on the remaining partition, labeled the test set. The AUC was computed for each of the 10 test sets and then averaged.

The AUC values for the SAVE composite scores (with standard errors in parenthesis) for $\hat{d}$, and fixed $d = 1$ and $d = 2$ were $0.85(0.15)$, $0.63(0.05)$ and $0.99(0.006)$, respectively, and $0.79(0.11)$ for SIR. Thus the SAVE composite scores with $\hat{d}$ as well as SAVE with $d = 2$ resulted in a 10% and 20% improvement in discriminatory accuracy compared to a single linear combination, either from SIR or SAVE with $d = 1$. This improvement in AUC values was statistically significant for the SAVE composite scores with $d = 2$.

The crossvalidated AUC for LR could not be computed, as the covariance matrices were numerically singular due to high correlations between sepal length and sepal width ($r = 0.96$) and between petal width and sepal width ($r = 0.82$).

We also attempted to predict class membership using logistic regression including all main effects and first order interaction terms. However, the algorithms to maximize the likelihood did not converge for these models, probably because sample sizes were small and there was high collinearity between the markers and corresponding interaction terms.

We tested for the individual contributions of the four predictors separately based on all the iris data and obtained the following values of the Wald chi-square test statistic: 4.55 for petal length, 49.73 for petal width, 119.63 for sepal length, and 57.86 for sepal width. Using 6.24 as the critical value, which is the $\chi_1^2(0.0125)$ quantile to control for multiple testing, we found that petal width, sepal length, and sepal width were contributing statistically significantly to the composite marker score. At $\alpha = 0.05$, all predictors were significant contributors.

6.2 NHANES data

To illustrate the proposed methods on a biomarker example, we used data from the publicly available Third National Health and Nutrition Examination Survey (NHANES III). The aim was to classify men age 40 years or older into two groups, heavy drinkers and abstainers, using nine serum biomarkers that had previously been shown to be associated with alcoholism (Eckhardt et al, 1981). 17]. The markers were the transformed values of blood glucose, hematocrit, sodium, chloride, phosphorus, uric acid, blood urea nitrogen, alkaline phosphatase, and albumin. We log-transformed all predictors, with the exception of blood glucose, which we transformed using a power transformation, x^λ, with $\lambda = -1.5$, and albumin, which we used on the original scale. These transformations yielded approximately symmetric distributions. As the heavy drinkers were significantly older than the abstainers, and age is known to influence the values of the biomarkers, we included age as a predictor. The abstaining group consisted of 327 men who reported to have had no more than twelve drinks in their lifetime, and the heavy drinking group was comprised of 338 men who reported to have had five or more drinks a day for more than one hundred days during the year before the interview.

We checked the SIR and SAVE assumptions by visually considering the scatterplot matrix of the predictors. The plots appeared to be random or linear, and there were no pronounced fluctuations in data density. Seven individuals exhibited outlying marker values and were removed from the analysis.

We used 10-fold cross validation (as described in the Iris data example) to obtain unbiased estimates of the AUC. That is, roughly 290 abstainers and 300 heavy drinkers were used to estimate the dimension of SAVE and to build the SAVE, SIR, logistic regression and LR predictors. The LR predictor was constructed assuming a multivariate normal distribution of the marker data in each group. The remaining observations were used to estimate the AUC. The AUC values presented are the averages over the ten partitions into training and test sets. For all ten training sets, the dimension estimate was $\hat{d} = 1$. The AUC values for the SAVE composite scores (with standard errors in parenthesis) for $\hat{d}$, and fixed $d = 1$ and $d = 2$ were $0.82(0.12)$, $0.82(0.12)$, and $0.81(0.13)$, respectively, and for SIR and logistic regression the AUC values were $0.82(0.14)$ and $0.82(0.13)$, respectively. The AUC for LR was only $0.54(0.03)$. The reason for the poor performance of LR was that the log transformed markers were not normally distributed, while the moment conditions for SAVE, as assessed by scatterplots, were met. This example highlights the difficulty of specifying joint marker distributions, which is precisely why it is precarious to use the LR on the original or even transformed data. The six panels of Figure 1 show the QQ plots for the SIR and SAVE predictors in heavy drinkers and abstainers for one specific choice of the training set, indicating that each of the projections follows a normal distribution.

Using all the NHANES data, the dimension of SAVE was estimated to be $d = 1$. When we applied the test statistic to each of the predictors separately based on all the NHANES data, we obtained the following values of the Wald chi-square test statistic: 0.09 for blood glucose, 2.32 for hematocrit, 0.31 for sodium, 182.04 for age, 0.31 for chloride, 5.05 for phosphorus, 29.23 for uric acid, 141.46 for blood urea nitrogen, 10.49 for alkaline phosphatase, and 9.71 for albumin. Using 7.88 as the critical value, the $\chi^2_1(0.005)$ quantile to control for multiple testing, we found that age, uric acid, blood urea nitrogen and alkaline phosphatase were contributing statistically significantly to the composite marker score. At $\alpha = 0.05$, phosphorus would also be a significant contributor.

In a logistic regression model with all transformed markers and age as main effects, the following markers contributed significantly at the $\alpha = 0.05$ level: age (logistic regression coefficient: 3.68, 95%CI: [3.26, 4.09]), phosphorus (logistic regression coefficient: -1.61, 95%CI:[-2.45,-0.77]), uric acid (logistic regression coefficient: -1.90, 95%CI:[-2.23, -1.57]), blood urea nitrogen (logistic regression coefficient: 1.98, 95%CI:[1.79, 2.16]), alkaline phosphatase (logistic regression coefficient: 0.69, 95%CI:[0.45, 0.93]) and albumin (logistic regression coefficient: -0.47, 95%CI:[-0.63, -0.30]). In this example, there was no evidence for higher than one-dimensional structure in the data, a conclusion that is also supported by the very similar performance of SIR and SAVE. Even though the SAVE composite scores did not increase the AUC in this example, the SAVE analysis yields the additional insight that a single linear combination of markers extracts all the available discriminatory information.

7 Discussion

We introduce two sufficient dimension reduction methods, Sliced Inverse Regression (SIR) and Sliced Average Variance Estimation (SAVE) to combine biomarkers into a diagnostic score. Our approach entails identifying a sufficient number of linear combinations of markers that can be combined into a diagnostic score via the likelihood ratio statistic for independent normal projections. This approach extends previous approaches (LDA; Su and Liu, 1993) to combining biomarkers as it does not require multivariate normality of the markers themselves, but only their projections. As most lower-dimensional projections of multivariate data have been shown to be approximately normal (Diaconis and Freedman, 1984, Hall and Li, 1993), this assumption is plausible. Our approach is computationally very tractable, even for large numbers of markers, and also offers a way to check which markers contribute significantly to prediction.

Key assumptions on the linearity of expectations and the constancy of the variances can be easily checked by diagnostic plots.

Pepe and Thompson (2002) and Pepe, Cai and Longton (2006) also relaxed the assumption of multivariate normality and computed a linear marker combination that maximizes a distribution-free estimate of the AUC. While their approach is attractive, since it can be adapted to maximize the partial area under the curve and to incorporate covariates, it is computationally difficult when one wishes to combine more than two markers. Ma and Huang (2007) extended the work by Pepe, Cai and Longton (2006) by maximizing the sigmoid approximation of the empirical AUC, which is computationally more affordable while preserving the asymptotic efficiency of the AUC estimator. These procedures cannot be fully efficient compared to parametric LR when the dimension of the central dimension reduction subspace exceeds one, however.

Our procedure is equivalent to the LR statistic, which requires knowing the joint marker distribution (McIntosh and Pepe, 2002), when the marker data arise from multivariate normal distributions. An important consideration is how well our procedure performs for non-normal data. Diaconis and Freedman (1984) found that projections may be non-normal for distributions with heavy tails, such as Cauchy, for which most projections (suitably scaled) are Cauchy. We therefore studied mixtures of normals and multivariate t-distributions. We found in simulations that SAVE performed comparably to LR in these cases (Tables 2 and 3). In retrospect, the good performance for multivariate-t data is not surprising, because it is an elliptically contoured distribution. Some loss in AUC for our procedure was detected when data were simulated from multivariate log-normal distributions compared to the LR statistic (Table 4). We therefore recommend that transformations be performed, before using SIR or SAVE procedures, to make the data more normal or symmetric. Of course, with log-normal data, a log transformation would yield normally distributed data, for which our procedures would be optimal.

Logistic regression with main effects performed comparably to SIR for multivariate normal data (Table 1). Including all first order interaction and quadratic terms in logistic models can recapture some of the discriminatory accuracy as illustrated by data on ten markers from a DSM(2) model, where all the discriminatory information is contained in a two dimensional subspace (Table 2). However, limited sample size or large numbers of markers make it impractical to include all first order interactions and quadratic terms in a logistic model. If the structural dimension is larger than two, the logistic regression model would have to include all three and higher order interaction terms in order to have similar discriminatory performance to SAVE, which would require an unrealistically large sample size in most applications.

As an alternative to using the normal LR statistic on SAVE projections, one could also substitute the SAVE projections into a logistic model as suggested in Cook and Lee (1999). However, in unreported simulations we found that using logistic regression was substantially less efficient than using our normal likelihood procedure, a finding that resembles an observation by Efron (1975). The use of the LR may be equivalent to the risk score when all parameters are known (McIntosh and Pepe, 2002), but not when parameters need to be estimated.

We proposed an asymptotically chi-squared distributed test statistic to assess contributions of markers to the diagnostic score. If a marker does not have a significant contribution to the SAVE or SIR linear combinations, it should be omitted from the diagnostic panel. Cook (2004) proposed an asymptotic weighted chi-squared test for variable contribution in directions derived by SIR. Our test is general; there are no restrictions on either the predictor or the response distribution, which can also be continuous. Our test procedure can be applied to test the significance of a predictor contribution in any linear combination of the predictors whose coefficients are computed by a singular value decomposition of a kernel matrix. The technique can be used with either SIR or SAVE.

The asymptotic framework for the test of marker contribution assumes that the dimension of SAVE is known. In simulations, the size of the test tended to be larger than the nominal level when the dimension was estimated and the second dimension of the DSM(2) subspace was not well separated. That can lead to inclusion of markers that do not truly contribute to the diagnostic accuracy of the score. Future work will entail adjustments of the test for estimating the dimension and comparing the performance of our test to other procedures to assess the importance of a variable for prediction, such as a model-free variable selection methodology recently proposed in Li, Cook and Nachtsheim (2005). In addition, a comprehensive comparison of our method with other available non-parametric methods (for example, Zhang et al, 1999) will be part of future work.

In summary, we have developed methods to combine markers for two class discrimination that can provide greater discriminatory accuracy as assessed by the AUC than standard methods, such as LDA or logistic regression when discriminatory features extend beyond shifts in location. Software written in *R* for finding marker or predictor combinations based on SAVE and SIR and testing marker contributions is available from the first author.

Acknowledgment

The second author was supported in part by NSF grant DMS-0204563. We thank Barry Graubard and Mitchell Gail for helpful discussions and the referees for insightful comments that led to improvements of the paper.

Appendix

Letting $f = \mathrm{P}(Y = 1)$, $V_{x11|j} = \mathrm{var}[\,\mathrm{vec}(X - \mu_{x|j})(X - \mu_{x|j})^T|Y = j]$, and $V_{x12|j} = \mathrm{cov}[\,\mathrm{vec}(X - \mu_{x|j})(X - \mu_{x|j})^T, X|Y = j]$ for $j = 0, 1$, $V_x = f^{-1/2}\ V_{x|1} + (1 - f)^{-1/2}\ V_{x|0}$, with

$$V_{x|j} = \begin{pmatrix} V_{x11|j} & V_{x12|j} \\ V_{x12|j}^T & \Sigma_{x|j} \end{pmatrix}.$$

References

Anderson, T. W. (1984). *An Introduction to Multivariate Statistical Analysis, 2nd Edition*. New York: Wiley-Interscience.

Bura, E. and Cook, R. D. (2001). Extending SIR: The Weighted Chi-Square Test, *Journal of the American Statistical Association* **96**, 996–1003.

Bura, E. and Pfeiffer, R.M. (2007). On the distribution of the left singular vectors of a random matrix and its applications. *Statistics and Probability Letters*, in press.

Cook, R. D. (1998). *Regression Graphics: Ideas for studying regressions through graphics*. New York: Wiley.

Cook, R. D. (2004). Testing predictor contributions in sufficient dimension reduction. *Annals of Statistics* **32**, 1062–1092.

Cook, R. D. and Lee, H. (1999). Dimension reduction in regressions with a binary response. *Journal of the American Statistical Association* **94**, 1187–1200.

Cook, R.D. and Weisberg, S. (1991). Discussion of Li. *Journal of the American Statistical Association* **86**, 328–332

Diaconis, P. and Freedman, D. (1984). Asymptotics of graphical projection pursuit. *Annals of Statistics* **12**, 793–815.

Eckhardt, M.J ., Ryback, R.S., Rawlings, R. R. and Graubard, B. I. (1981). Biochemical diagnosis of alcoholism - a test of the discriminating capabilities of γ-Glutamyl Transpeptidase and mean corpuscular volume. *Journal of the American Medical Association* **246**, 2707–2710.

Efron, B. (1975). Efficiency of logistic regression compared to normal discriminant analysis. *Journal of the Statistical Association* **70**, 892–898.

Flury, L., Boukai, B. and Flury, B.D. (1997) The discrimination subspace model. *Journal of the American Statistical Association* **92**, 758-766.

Gastwirth, J.L. (1987). The Statistical Precision of Medical Screening Procedures: Application to Polygraph and AIDS Antibodies Test Data. *Statistical Science* **2**, 213-238.

Hall, P. and Li, K. C. (1993). On almost linearity of low dimensional projections from high dimensional data. *Annals of Statistics* **21**, 867–889.

Li, K. C. (1991). Sliced inverse regression for dimension reduction (with discussion). *Journal of the American Statistical Association* **86**, 316–342.

Li, L., Cook, R. D. and Nachtsheim, C. J. (2005). Model-free variable selection. *Journal of the Royal Statistical Society, Series B* **67**, 285–299.

Liu, A. Y., Schisterman, E. F. and Zhu, Y. (2005). On linear combinations of biomarkers to improve diagnostic accuracy. *Statistics in Medicine* **24**, 37–47.

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

McGlish, D. K. (1989). Analysing a portion of the ROC curve. *Medical Decision Making* ;**9**, 190–195.

McIntosh, M. W. and Pepe, M. S. (2002). Combining several screening tests: Optimality of the risk score. *Biometrics* **58**, 657–664.

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

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

Pepe, M. S. and Thompson, M. H. (2000). Combining diagnostic test results to increase accuracy. *Biostatistics*; **1**, 123–140.

Su, J. Q. and Liu, J. S. (1993). Linear Combinations of Multiple Diagnostic Markers. *Journal of the American Statistical Association* **88**, 1350–1355.

Wood, A.T.A. (1989). An F approximation to the distribution of a linear combination of Chi-squared variables, *Communications in Statistics, Simulation & Computation* **18**, 1439–1456.

Zhang, Z., Barnhill, S.D., Zhang, H., Xu, F., Yu, Y., Jacobs, I., Woolas, R.P., Berchuck, A., Madyastha, K.R., and Bast, R.C. Jr. (1999). Combination of multiple serum markers using an artificial neural network to improve specificity in discriminating malignant from benign pelvic masses. *Gynecologic Oncolology* **73**, 56-61.

Table 1: Mean AUC values over 100 simulations for p predictors from the DSM(2) multivariate normal models, $(X|Y=0) \sim MVN(0, AA'), (X|Y=1) \sim MVN(A\omega_1, A\Omega_1 A')$, with $(A)_{ii} = 1, (A)_{ij} = 0.5$, $\omega_1 = \sum_{j=1}^{2} \delta_j e_j$, $\Omega_1 = I_p + \sum_{j=1}^{2} \tau_j e_j' e_j$, where e_j denotes the j-th unit basis vector. Standard errors are given below the mean estimates. The column labeled $\hat{d}$ gives the mean AUC when number of SAVE predictors is estimated.

p	N_0, N_1^*	$\tau_1, \tau_2, \delta_1, \delta_2^*$	LR		S_{SAVE}		S_{SIR}	Logistic
			est	$\hat{d}$	$d = 2$	$d = 4$		Regression
10	200, 200	1.0,2.0,1.0,0.1	0.75	0.69	0.76	0.75	0.70	0.70
			0.03	0.05	0.03	0.02	0.03	0.03
20	200, 200	1.0,2.0,1.0,0.1	0.69	0.65	0.70	0.71	0.69	0.69
			0.03	0.04	0.05	0.04	0.03	0.03
30	200, 200	1.0,2.0,1.0,0.1	0.65	0.62	0.65	0.66	0.68	0.68
			0.03	0.03	0.05	0.04	0.03	0.03
10	500, 500	1.0,2.0,1.0,0.1	0.77	0.77	0.78	0.77	0.71	0.71
			0.02	0.02	0.02	0.02	0.02	0.02
20	500, 500	1.0,2.0,1.0,0.1	0.73	0.73	0.77	0.76	0.71	0.71
			0.02	0.05	0.02	0.02	0.02	0.02
30	500, 500	1.0,2.0,1.0,0.1	0.70	0.68	0.75	0.74	0.70	0.70
			0.01	0.05	0.02	0.02	0.02	0.02
10	200, 200	2.0,5.0,1.0,0.1	0.81	0.82	0.83	0.82	0.68	0.68
			0.02	0.03	0.02	0.02	0.03	0.03
20	200, 200	2.0,5.0,1.0,0.1	0.77	0.76	0.81	0.79	0.67	0.67
			0.02	0.04	0.02	0.02	0.03	0.03
30	200, 200	2.0,5.0,1.0,0.1	0.73	0.73	0.79	0.77	0.65	0.65
			0.02	0.03	0.03	0.03	0.03	0.03
10	500, 500	2.0,5.0,1.0,0.1	0.83	0.83	0.83	0.83	0.69	0.69
			0.01	0.01	0.01	0.01	0.02	0.02
20	500, 500	2.0,5.0,1.0,0.1	0.81	0.83	0.83	0.82	0.68	0.68
			0.01	0.01	0.01	0.01	0.02	0.02
30	500, 500	2.0,5.0,1.0,0.1	0.78	0.81	0.83	0.81	0.67	0.67
			0.02	0.04	0.02	0.02	0.02	0.02
10	200, 200	5.0,10.0,1.0,0.1	0.88	0.89	0.89	0.88	0.63	0.63
			0.02	0.02	0.02	0.02	0.03	0.03
20	200, 200	5.0,10.0,1.0,0.1	0.85	0.88	0.88	0.87	0.61	0.61
			0.02	0.03	0.02	0.02	0.03	0.03
30	200, 200	5.0,10.0,1.0,0.1	0.82	0.82	0.88	0.86	0.59	0.59
			0.02	0.04	0.02	0.02	0.03	0.03
10	500, 500	5.0,10.0,1.0,0.1	0.89	0.89	0.89	0.89	0.64	0.64
			0.01	0.01	0.01	0.01	0.02	0.02
20	500, 500	5.0,10.0,1.0,0.1	0.87	0.89	0.89	0.88	0.63	0.63
			0.01	0.01	0.01	0.01	0.02	0.02
30	500, 500	5.0,10.0,1.0,0.1	0.86	0.89	0.89	0.88	0.62	0.62
			0.01	0.01	0.01	0.01	0.02	0.02

* size of training set

Table 2: Mean AUC values for 10 predictors from multivariate normal mixtures, $(X|Y = 0) \sim MVN(0, AA')$ and $(X|Y = 1) \sim 0.5MVN(A\omega_1, A\Omega_1 A')+ 0.5MVN(-A\omega_1, A\Omega_1 A')$, with $(A)_{ii} = 1, (A)_{ij} = 0.5$, $\omega_1 = \sum_{j=1}^{2} \delta_j e_j$, $\Omega_1 = I_p + \sum_{j=1}^{2} \tau_j e_j' e_j$, where e_j denotes the j-th unit basis vector. Standard errors are given below the mean estimates.

p	N_0, N_1	$\tau_1, \tau_2, \delta_1, \delta_2,$	LR	S_{SAVE}			S_{SIR}	Logistic	Logistic
			est	$\hat{d}^*$	$d = 2$	$d = 4$		Reg^1	Reg^2
10	500, 500	2,5,1.0,0.1	0.82	0.83	0.83	0.82	0.51	0.51	0.81
			0.01	0.01	0.01	0.01	0.01	0.01	0.01
10	500, 500	5,10,1.0,0.1	0.89	0.90	0.90	0.89	0.51	0.51	0.88
			0.01	0.01	0.01	0.01	0.01	0.01	0.01
10	1000, 1000	5,10,1.0,0.1	0.90	0.90	0.90	0.90	0.51	0.51	0.89
			0.01	0.01	0.01	0.01	0.01	0.01	0.01
20	200, 200	0,2.0,1.0,0.1	0.72	0.81	0.79	0.76	0.53	0.53	NA^3
			0.03	0.02	0.03	0.03	0.02	0.02	
20	500, 500	0,2.0,1.0,0.1	0.76	0.83	0.82	0.80	0.52	0.52	NA
			0.02	0.02	0.02	0.02	0.01	0.01	
20	200, 200	2.0,5.0,1.0,0.1	0.76	0.81	0.82	0.79	0.52	0.52	NA
			0.03	0.03	0.03	0.03	0.02	0.02	
20	200, 200	5.0,10.0,1.0,0.1	0.85	0.89	0.89	0.87	0.52	0.52	NA
			0.02	0.02	0.02	0.02	0.02	0.02	
30	500, 500	5.0,10.0,1.0,0.1	0.86	0.89	0.89	0.88	0.52	0.52	NA
			0.01	0.01	0.01	0.01	0.01	0.01	

[*] d denotes the number of SAVE predictors used in the diagnostic score

[1] Logistic regression with main effects terms

[2] Logistic regression with main effects terms and all first order interaction terms and quadratic terms

[3] NA means not available because there were too many parameters to fit

Table 3: Mean AUC values over 100 simulations for p predictors from the a multivariate t-distribution with 3 degrees of freedom, $(X|Y=0) \sim MVT(0, AA', df=3), (X|Y=1) \sim MVT(A\omega_1, A\Omega_1 A', df=3)$, with $(A)_{ii} = 1, (A)_{ij} = 0.5$, $\omega_1 = \sum_{j=1}^{2} \delta_j e_j$, $\Omega_1 = I_p + \sum_{j=1}^{2} \tau_j e_j' e_j$, where e_j denotes the j-th unit basis vector. Standard errors are given below the mean estimates. The column labeled $\hat{d}$ gives the mean AUC when number of SAVE predictors is estimated.

p	N_0, N_1^*	$\tau_1, \tau_2, \delta_1, \delta_2^*$	LR	S_{SAVE}			S_{SIR}	Logistic
			est	$\hat{d}$	$d=2$	$d=4$		Regression
10	500, 500	1.0,2.0,1.0,0.1	0.71	0.67	0.68	0.69	0.67	0.67
			0.02	0.05	0.06	0.03	0.02	0.02
20	500, 500	1.0,2.0,1.0,0.1	0.67	0.63	0.63	0.63	0.64	0.64
			0.02	0.05	0.05	0.04	0.02	0.02
30	500, 500	1.0,2.0,1.0,0.1	0.64	0.62	0.62	0.60	0.64	0.64
			0.02	0.04	0.04	0.03	0.02	0.02
10	500, 500	2.0,5.0,1.0,0.1	0.78	0.76	0.76	0.76	0.64	0.64
			0.02	0.03	0.04	0.02	0.02	0.02
20	500, 500	2.0,5.0,1.0,0.1	0.75	0.71	0.72	0.73	0.64	0.64
			0.02	0.05	0.05	0.03	0.02	0.02
30	500, 500	2.0,5.0,1.0,0.1	0.72	0.69	0.69	0.71	0.62	0.62
			0.02	0.05	0.04	0.03	0.02	0.02
10	500, 500	5.0,10.0,1.0,0.1	0.85	0.82	0.83	0.83	0.60	0.61
			0.02	0.04	0.02	0.02	0.02	0.02
20	500, 500	5.0,10.0,1.0,0.1	0.83	0.79	0.81	0.81	0.59	0.59
			0.02	0.06	0.03	0.02	0.02	0.02
30	500, 500	5.0,10.0,1.0,0.1	0.81	0.79	0.79	0.80	0.58	0.58
			0.02	0.06	0.04	0.02	0.02	0.02

* size of training set

Table 4: Mean AUC values over 100 simulations for p predictors from the a multivariate log normal distribution, $(\log(X)|Y = 0) \sim MVN(0, AA', df = 3), (\log(X)|Y = 1) \sim MVN(A\omega_1, A\Omega_1 A', df = 3)$, with $(A)_{ii} = 1, (A)_{ij} = 0.5$, $\omega_1 = \sum_{j=1}^{2} \delta_j e_j$, $\Omega_1 = I_p + \sum_{j=1}^{2} \tau_j e_j' e_j$, where e_j denotes the j-th unit basis vector. Standard errors are given below the mean estimates. The column labeled $\hat{d}$ gives the mean AUC when number of SAVE predictors is estimated.

p	N_0, N_1^*	$\tau_1, \tau_2, \delta_1, \delta_2^*$	LR	S_{SAVE}			S_{SIR}	Logistic
			est	$\hat{d}$	$d = 2$	$d = 4$		Regression
10	500, 500	2.0,5.0,1.0,0.1	0.69	0.60	0.62	0.63	0.60	0.61
			0.02	0.07	0.04	0.03	0.04	0.02
20	500, 500	2.0,5.0,1.0,0.1	0.68	0.60	0.60	0.57	0.57	0.58
			0.02	0.04	0.03	0.03	0.03	0.05
30	500, 500	2.0,5.0,1.0,0.1	0.67	0.61	0.61	0.60	0.56	0.57
			0.02	0.03	0.02	0.02	0.03	0.06
10	500, 500	5.0,10.0,1.0,0.1	0.69	0.62	0.63	0.62	0.55	0.55
			0.02	0.07	0.03	0.03	0.02	0.02
20	500, 500	5.0,10.0,1.0,0.1	0.69	0.62	0.59	0.58	0.55	0.55
			0.02	0.02	0.03	0.03	0.03	0.03
30	500, 500	5.0,10.0,1.0,0.1	0.68	0.61	0.59	0.57	0.53	0.54
			0.02	0.03	0.02	0.02	0.02	0.02

* size of training set

Table 5: Performance of Wald test to assess the contribution of maker p based on 500 repetitions

$(X_1, \ldots, X_{p-1}) \sim DSM(2), X_p \sim N(0,1)$

$H_0 : X_p$ does not contribute to the diagnostic score, H_0 is true

p	N_0, N_1	$\tau_1, \tau_2, \delta_1, \delta_2,$	Fraction of rejections (95%CI)
10	200,200	2.0, 1.0, 1.0, 0.1	0.07 (0.04, 0.10)
10	200,200	2.0, 5.0, 1.0, 0.1	0.06 (0.04, 0.08)
10	500,500	2.0, 1.0, 1.0, 0.1	0.06 (0.04, 0.08)
10	500,500	2.0, 5.0, 1.0, 0.1	0.05 (0.04, 0.07)
15	500,500	2.0, 5.0, 1.0, 0.1	0.05 (0.03, 0.07)
20	500,500	2.0, 1.0, 1.0, 0.1	0.05 (0.03, 0.07)
20	500,500	2.0, 5.0, 1.0, 0.1	0.05 (0.03, 0.07)

$(X_1, \ldots, X_p) \sim DSM(2)$, H_0 is false

$H_0 : X_p$ does not contribute to the diagnostic score

10	200,200	0.2, 0.5, 0.1, 0.1	0.62 (0.59, 0.65)
10	200,200	0.2, 0.5, 0.5, 0.1	0.98 (0.97, 0.99)
10	200,200	0.2, 0.5, 1.0, 0.1	1.00 (1.00, 1.00)
10	200,200	2.0, 5.0, 0.1, 0.1	1.00 (1.00, 1.00)
10	500,500	0.2, 0.5, 0.1, 0.1	0.70 (0.67, 0.72)
10	500,500	0.0, 2.0, 0.1, 0.1	0.35 (0.26, 0.44)
20	500,500	0.0, 2.0, 0.1, 0.1	0.30 (0.21, 0.39)
20	500,500	0.2, 0.5, 0.1, 0.1	0.68 (0.59, 0.77)
20	500,500	0.2, 0.5, 0.3, 0.1	0.98 (0.95, 1.00)
20	500,500	0.2, 0.5, 0.5, 0.1	1.00 (1.00, 1.00)