Journal of the American Statistical Association
Publication details, including instructions for authors and subscription information:
http://amstat.tandfonline.com/loi/uasa20

Latent Supervised Learning

Susan Wei [a] & Michael R. Kosorok [a]
[a] Department of Statistics and Operations Research , University of North Carolina at Chapel Hill , Chapel Hill , NC , 27599
Accepted author version posted online: 15 Jul 2013.Published online: 27 Sep 2013.

To cite this article: Susan Wei & Michael R. Kosorok (2013) Latent Supervised Learning, Journal of the American Statistical Association, 108:503, 957-970, DOI: 10.1080/01621459.2013.789695

To link to this article: http://dx.doi.org/10.1080/01621459.2013.789695

Latent Supervised Learning

Susan WEI and Michael R. KOSOROK

This article introduces a new machine learning task, called latent supervised learning, where the goal is to learn a binary classifier from *continuous* training labels that serve as surrogates for the unobserved class labels. We investigate a specific model where the surrogate variable arises from a two-component Gaussian mixture with unknown means and variances, and the component membership is determined by a hyperplane in the covariate space. The estimation of the separating hyperplane and the Gaussian mixture parameters forms what shall be referred to as the change-line classification problem. We propose a data-driven sieve maximum likelihood estimator for the hyperplane, which in turn can be used to estimate the parameters of the Gaussian mixture. The estimator is shown to be consistent. Simulations as well as empirical data show the estimator has high classification accuracy.

KEY WORDS: Classification and clustering; Glivenko–Cantelli classes; Sieve maximum likelihood estimation; Sliced inverse regression; Statistical learning.

1. INTRODUCTION

This article introduces a new machine learning task, latent supervised learning. The goal is to learn a *binary* classifier from *continuous* training labels. The term latent describes the hidden underlying relationship between the surrogate and the unobserved class label. This latency structure manifests in many real-world applications. Take for instance the world of clinical trials, where it is common to show a direct clinical benefit to a *surrogate* marker rather than a real clinical endpoint (Fleming 2005). An example of a surrogate marker is a continuous measurement such as blood pressure. The corresponding real clinical endpoint might be a binary indicator of death. Using a surrogate variable to guide classification, latent supervised learning directly targets the setting where clearly labeled training data are unavailable.

In this way, latent supervised learning bridges the gap between unsupervised and supervised learning. In the former, data are unlabeled and the goal is simply to discover useful classes of items. This is also known as clustering, see Jain, Murty, and Flynn (1999) for a review. On the other hand, supervised learning, see Hastie, Tibshirani, and Friedman (2003) for an overview, seeks to derive a function from labeled training data. Such a function is called a classifier if the label is discrete or a regression function if the label is continuous. There are instances, however, when carefully trained data are difficult or too costly to obtain. In such cases, supervised learning is infeasible and latent supervised learning provides a preferable alternative to clustering if a clearly generalizable classification rule is desired.

This article studies a specific problem in latent supervised learning, which shall be referred to as the change-line classification problem. The surrogate variable arises from a Gaussian mixture distribution with unknown parameters where the latent structure between the surrogate and the component class label is determined by an unknown hyperplane in the covariate space. We propose a data-driven sieve maximum likelihood estimator (MLE) to estimate the hyperplane. Importantly, the classification of future objects solely depends on the separating hyperplane. This makes the method generalizable and advantageous in situations where the surrogate variable may not be available for future data.

The estimator is shown to be consistent. Its accuracy is demonstrated on simulated data. Three health-related datasets are used to illustrate its applicability. Two of the datasets are accompanied by binary outcome variables. For these, the subgroups estimated by the method will be compared to the ones given by the binary outcome variable. The data-driven sieve estimator is able to achieve, without using the binary training labels, classification accuracy comparable to that of logistic regression, a fully supervised procedure. For the third dataset where there is no binary outcome variable available, an interpretation of the subgroups discovered is offered.

The article is organized as follows. In the next section, the model is formally defined. In Section 3, related work is discussed. In Section 4, a variety of existing "off-the-shelf" statistical methods are examined and the caveats of using each is addressed. Section 5 presents the methodology. The consistency of the estimator is established in Section 6. The issue of model checking and diagnostics is discussed in Section 7. Simulations in Section 8 compare the method to other competitors. Applications to real-world datasets are presented in Section 9. The article ends with a discussion in Section 10. Some additional supporting material including proofs of results and data preprocessing steps are given in the Appendix.

2. THE MODEL

The setup of the problem is as follows. Let the covariate $X \in \mathbb{R}^d$ be related to the surrogate variable $Y \in \mathbb{R}$ in the following manner:

$$Y = \mu_{1,0} 1\left\{\omega_0^T X - \gamma_0 \geq 0\right\} + \mu_{2,0} 1\left\{\omega_0^T X - \gamma_0 < 0\right\} + \epsilon, \tag{1}$$

Susan Wei is Doctoral Student, Department of Statistics and Operations Research, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599 (Email: *susanwe@live.unc.edu*). Michael R. Kosorok is Professor and Chair, Department of Biostatistics and Professor, Department of Statistics and Operations Research, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599 (Email: *kosorok@unc.edu*). The first author was funded through the National Science Foundation Graduate Fellowship and the National Institutes of Health (NIH) grant T32 GM067553-05S1. The second author was funded in part by the NIH grant CA142538. We thank Editor Xuming He, the Associate Editor, and two anonymous referees for their helpful comments that led to a significantly improved article.

Journal of the American Statistical Association
September 2013, Vol. 108, No. 503, Theory and Methods
DOI: 10.1080/01621459.2013.789695

where the means $\mu_{1,0}, \mu_{2,0} \in \mathbb{R}$ are unknown, and

$$\epsilon \sim N\left(0, \sigma_{1,0}^2 1\left\{\omega_0^T X - \gamma_0 \geq 0\right\} + \sigma_{2,0}^2 1\left\{\omega_0^T X - \gamma_0 < 0\right\}\right),$$

where the variances $\sigma_{1,0}^2, \sigma_{2,0}^2 \in \mathbb{R}^+$ are also unknown. The relationship between the means and variances is allowed to be arbitrary as long as the equations $\mu_{1,0} = \mu_{2,0}$ and $\sigma_{1,0}^2 = \sigma_{2,0}^2$ are not simultaneously true. The sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ is observed iid from Model (1). The estimation of ω_0 and γ_0 is of primary interest.

3. RELATED WORK

The model considered here was first described in Kang's Ph.D. thesis, see Kang (2011). Kang proposed an estimator for the special case $p = 2$. The procedure involved first enumerating all linear hyperplanes in $\mathbb{R}^2$ that separate the sample of data $x_1, \ldots, x_n$ into two groups. Then the hyperplane that maximizes the likelihood is taken to be the estimate. A procedure enumerating all hyperplanes splitting the data for $\mathbb{R}^3$ or higher does not seem to be generalizable from the procedure for $\mathbb{R}^2$. Thus, an extension to $\mathbb{R}^3$ or beyond based on this technique appears difficult.

It was also Kang who coined the term "change-line classification." This is likely a reference to the well-studied topic of change-point problems, see Carlstein, Müller, and Siegmund (1994) for an overview. The relationship to the present model can be seen as follows. In its simplest form, the change-point model assumes the following structure:

$$Y = \alpha_0 1_{X \leq \zeta_0} + \beta_0 1_{X > \zeta_0} + \epsilon,$$

where ϵ is a normally distributed error term. The parameter of interest is ζ_0, the change-point. Model (1) encompasses this basic change-point model; set $\mu_{1,0} = \alpha_0, \mu_{2,0} = \beta_0, \sigma_{1,0}^2 = \sigma_{2,0}^2$, $\omega_0 = 1$ and $\gamma_0 = \zeta_0$ to see this. Model (1) is a generalization of the basic change-point model in two ways: (a) no restrictions are placed on the relationship between $\sigma_{1,0}^2$ and $\sigma_{2,0}^2$ and (b) the search for a change-point is generalized to a change-hyperplane. These generalizations in turn require a whole new set of tools.

4. OFF-THE-SHELF SOLUTIONS

This section provides motivation for the necessity of a new methodology to solve the change-line classification problem. This is addressed by first considering several "off-the-shelf" statistical methods. Discussion of the caveats of the application of each to the change-line classification problem follows.

4.1 Linear Regression

A simple regression of Y on X could be used for the change-line classification problem. However, under Model (1),

$$E(Y|X) = \mu_{1,0} 1\{\omega^T X - \gamma \geq 0\} - \mu_{2,0} 1\{\omega^T X - \gamma < 0\}.$$

This is not linear in X and thus linear regression is unlikely to perform well.

4.2 SIR

The more sophisticated procedure sliced inverse regression (SIR) assumes that there exists a lower-dimensional projection of the covariates X that explains all that needs to be known about the surrogate variable Y (Li 1991). Formally, the model stipulates

$$Y = f(\beta_1 X, \beta_2 X, \ldots, \beta_k X, \epsilon),$$

where the β's are unknown and f is an arbitrary unknown function.

The implementation of SIR will now be described in detail as it will play a key role in the proposed methodology. For simplicity, assume the covariate X has been standardized to have mean zero and identity covariance. In the first step of SIR, the range of Y is partitioned into H (not necessarily equal) slices $\{I_1, \ldots, I_H\}$. Let $\hat{m}_h$ be the sample mean of the covariates in the hth slice, that is,

$$\hat{m}_h = \frac{\sum_{i=1}^n X_i 1\{Y_i \in I_h\}}{\sum_{i=1}^n 1\{Y_i \in I_h\}}.$$

The kth largest eigenvector (eigenvector corresponding to the kth largest eigenvalue) of the weighted covariance matrix $\sum_{h=1}^H |I_h| \hat{m}_h \hat{m}_h'$ is taken to be an estimate of β_k. To estimate ω_0 in the change-line estimation problem, set $k = 1$ and apply SIR. It will seen later in Section 5 that a direct application of SIR under Model (1) is often sensitive to noise in the data and can have poor performance even when the sample size is moderately large.

4.3 EM

The methods described thus far focus on modeling the relationship between the covariate X and the surrogate variable Y. Also each method produces an estimate of ω_0 only. An entirely different line of approach is to first estimate the binary labels $1\{\omega_0^T X_i - \gamma_0 \geq 0\}$ for each $i = 1, \ldots, n$ and then apply a standard binary linear classification method, such as the support vector machine (SVM), to estimate ω_0 and γ_0. This approach requires that the binary labels first be estimated with a high degree of accuracy.

One possible way to estimate these binary labels is the expectation-maximization (EM) algorithm. The data arising from Model (1) is a Gaussian mixture with unknown parameters. The EM algorithm more directly targets the estimation of the parameters $\mu_{1,0}, \mu_{2,0}, \sigma_{1,0}^2, \sigma_{2,0}^2$ but can do a poor job of estimating the actual class membership labels $1\{\omega_0^T x_i - \gamma_0 \geq 0\}$.

4.4 Clustering

Another possibility is to use clustering methods to estimate the binary labels. The cluster membership can then be used as training labels in a binary linear classifier such as SVM. A basic clustering algorithm such as k-means clustering with $k = 2$ can be performed on the Y space. This, however, entirely ignores the information in the covariate X and the resulting clusters may not be sensible when viewed in the covariate space. Another approach, clustering on the (X, Y) space to estimate the binary labels, has the drawback that the dimension of the covariate space is usually higher than the one-dimensional surrogate variable Y, but a standard clustering algorithm will weigh them equally.

In Section 8, simulations are performed to compare the proposed methodology to each of the methods above. The results suggest the new methodology is generally more accurate for the change-line classification problem than any of these "off-the-shelf" methods.

5. METHODOLOGY

The estimation of ω_0 in Model (1) uses a sieve maximum likelihood approach. A sieve is a sequence of approximating spaces that grows dense as the sample size increases (Grenander 1981). Maximization is carried out over these approximating spaces rather than the full parameter space. Traditionally, the method of sieves has been used in nonparametric maximum likelihood estimation. There, sieves are either (a) deterministic or (b) random but not data dependent. See Geman and Hwang (1982) for examples of the former and Shen, Shi, and Wong (1999) for the latter.

The proposed sieve estimation procedure is unique in that the sieve is constructed using the observed data. The construction begins with a data-driven sieve that is based on the information in the covariate X. Next the sieve is "boosted" by incorporating information from the surrogate variable Y.

5.1 The Likelihood

The expression of the likelihood function is described here. Let $\theta(\omega, \gamma)$ be the collected nuisance parameters

$$\theta(\omega, \gamma) := \left(\mu_1(\omega, \gamma), \mu_2(\omega, \gamma), \sigma_1^2(\omega, \gamma), \sigma_2^2(\omega, \gamma)\right),$$

where

$$\mu_1(\omega, \gamma) := E(Y|\omega^T X - \gamma \geq 0) \quad \text{and}$$
$$\mu_2(\omega, \gamma) := E(Y|\omega^T X - \gamma < 0)$$

and

$$\sigma_1^2(\omega, \gamma) := \text{var}(Y|\omega^T X - \gamma \geq 0) \quad \text{and}$$
$$\sigma_2^2(\omega, \gamma) := \text{var}(Y|\omega^T X - \gamma < 0).$$

The log-likelihood of the data under Model (1) as a function of (ω, γ) is given by

$$L_n(\omega, \gamma, \theta(\omega, \gamma)) = -\frac{1}{2}\sum_{i=1}^{n}\left[\log(2\pi\sigma^2(x_i, \omega, \gamma)) + \frac{(y_i - \mu(x_i, \omega, \gamma))^2}{\sigma^2(x_i, \omega, \gamma)}\right], \tag{2}$$

where

$$\mu(x, \omega, \gamma) = (\mu_1(\omega, \gamma) - \mu_2(\omega, \gamma))1\{\omega^T x - \gamma \geq 0\} + \mu_2(\omega, \gamma) \tag{3}$$

and

$$\sigma^2(x, \omega, \gamma) = \left(\sigma_1^2(\omega, \gamma) - \sigma_2^2(\omega, \gamma)\right)1\{\omega^T x - \gamma \geq 0\} + \sigma_2^2(\omega, \gamma). \tag{4}$$

A natural estimate for $\theta(\omega, \gamma)$ is

$$\hat{\theta}_n(\omega, \gamma) := \left(\hat{\mu}_1(\omega, \gamma), \hat{\mu}_2(\omega, \gamma), \hat{\sigma}_1^2(\omega, \gamma), \hat{\sigma}_2^2(\omega, \gamma)\right), \tag{5}$$

where the estimated means are given by

$$\hat{\mu}_1(\omega, \gamma) = \frac{\sum_{i=1}^n y_i 1\{\omega^T x_i - \gamma \geq 0\}}{\sum_{i=1}^n 1\{\omega^T x_i - \gamma \geq 0\}} \quad \text{and}$$
$$\hat{\mu}_2(\omega, \gamma) = \frac{\sum_{i=1}^n y_i 1\{\omega^T x_i - \gamma < 0\}}{\sum_{i=1}^n 1\{\omega^T x_i - \gamma < 0\}}$$

and the estimated variances are given by

$$\hat{\sigma}_1^2(\omega, \gamma) = \frac{\sum_i (y_i - \hat{\mu}_1(\omega, \gamma))^2 1\{\omega^T x_i - \gamma \geq 0\}}{\sum_i 1\{\omega^T x_i - \gamma \geq 0\}}$$

and

$$\hat{\sigma}_2^2(\omega, \gamma) = \frac{\sum_i (y_i - \hat{\mu}_2(\omega, \gamma))^2 1\{\omega^T x_i - \gamma < 0\}}{\sum_i 1\{\omega^T x_i - \gamma < 0\}}.$$

Let $\mathbb{S}^p$ denote the unit sphere in $\mathbb{R}^p$. The likelihood L_n is maximized over a sieve $\hat{\Omega}_n \subset \mathbb{S}^p$ using the plug-in estimate $\hat{\theta}_n(\omega, \gamma)$. Let $\hat{\Gamma}_n(\omega) \subset \mathbb{R}$ be the set of γ's such that $\hat{\theta}_n(\omega, \gamma)$ is well defined. The sieved estimator is

$$\left(\hat{\omega}_n^s, \hat{\gamma}_n^s\right) := \min \underset{\omega \in \hat{\Omega}_n, \gamma \in \hat{\Gamma}_n(\omega)}{\arg\max} L_n(\omega, \gamma, \hat{\theta}_n(\omega, \gamma)), \tag{6}$$

where min arg max denote the smallest arg max. This is necessary since there is a whole interval of γ's that maximize the likelihood. The next two sections describe the construction of the sieve $\hat{\Omega}_n$.

5.2 The Simple Sieve

The simple sieve is based on the mean difference (MD) discrimination rule applied to the covariates x. The MD, also known as the nearest centroid method [see Chapter 1 of Scholkopf and Smola (2001)], is a forerunner to the shrunken nearest centroid method of Tibshirani et al. (2002). It is based on the class sample mean vectors, denoted by $\bar{x}^+$ and $\bar{x}^-$. A new data vector is assigned to the positive (negative) class if it is closer to $\bar{x}^+$ ($\bar{x}^-$). Thus the MD discrimination method results in a separating hyperplane with normal vector $\bar{x}^+ - \bar{x}^-$. The simple sieve consists of MD directions formed in the following manner:

1. Partition the covariate space X into K regions. Let $S_k \subset \{1, \ldots, n\}$ be the index set for region k.
2. Let $\mathcal{P}_k$ denote the collection of partitions of the set S_k into two parts. For $P \in \mathcal{P}_k$, let P_1 and P_2 be the parts of the partition, that is, $P_1 \cup P_2 = S_k$ and $P_1 \cap P_2 = \emptyset$.
3. For each $P \in \bigcup_k \mathcal{P}_k$, calculate the MD direction $\omega^{\text{MD}}(P)$—the vector connecting the centroids of the two classes $\{X_i : i \in P_1\}$ and $\{X_i : i \in P_2\}$,

$$\omega^{\text{MD}}(P) = \frac{\bar{X}_{P_1} - \bar{X}_{P_2}}{||\bar{X}_{P_1} - \bar{X}_{P_2}||},$$

where $\bar{X}_{P_1}$ and $\bar{X}_{P_2}$ are the sample means of X's in P_1 and P_2, respectively.

K-means clustering can be used for the first step to obtain a partition of the covariate space. If K-means returns clusters that are very large, sample a manageable portion of the cluster. The parameter K should be chosen to ensure the cardinality of the sieve is not too big. Setting K to be roughly $n/10$ works well in practice. This choice results in the sieve having approximately $\sum_{k=1}^K 2^{|S_k|} = n2^{10}/10$ elements, which grows linearly in n and is quite manageable computationally.

5.3 Incorporating the Surrogate Variable

To incorporate the information from the surrogate variable Y to boost the simple sieve, the SIR procedure is applied to the bivariate $(Y, 1\{\omega^T X - \gamma \geq 0\})$; henceforth this will be referred to as the modified SIR. First, slice the range of Y into H (not necessarily equal) slices $\{I_1, \ldots, I_H\}$. Next, standardize X to have mean zero and unit covariance: $\tilde{X} = \hat{\Sigma}_{xx}^{-1/2}(X_i - \bar{X})$, for $i = 1, \ldots, n$, where $\bar{X}$ and $\hat{\Sigma}_{xx}$ are the sample mean and sample covariance matrix of X, respectively. Let $\hat{m}_{h,1}(\omega, \gamma)$ be the average of the $\tilde{X}$'s in the hth slice that are above the hyperplane $\omega^T x - \gamma \geq 0$,

$$\hat{m}_{h,1}(\omega, \gamma) = \frac{\sum_{i=1}^n \tilde{X}_i 1\{Y_i \in I_h\} 1\{\omega^T X_i - \gamma \geq 0\}}{\sum_{i=1}^n 1\{Y_i \in I_h\} 1\{\omega^T X_i - \gamma \geq 0\}}$$

and analogously for below the hyperplane

$$\hat{m}_{h,2}(\omega, \gamma) = \frac{\sum_{i=1}^n \tilde{X}_i 1\{Y_i \in I_h\} 1\{\omega^T X_i - \gamma < 0\}}{\sum_{i=1}^n 1\{Y_i \in I_h\} 1\{\omega^T X_i - \gamma < 0\}}.$$

The quantities $\hat{m}_{h,1}(\omega, \gamma)$ and $\hat{m}_{h,2}(\omega, \gamma)$ are sample versions of $E(\tilde{X}|Y \in I_h, \omega^T X - \gamma \geq 0)$ and $E(\tilde{X}|Y \in I_h, \omega^T X - \gamma < 0)$, respectively. The theoretical expectations will show variation along the direction ω_0 under Model (1). The direction along which the points $\hat{m}_{h,1}$ and $\hat{m}_{h,2}$ exhibit the most variation is found using a weighted principal components analysis (PCA). The $d \times d$ weighted covariance matrix, expressed in terms of ω and γ, is given by

$$\hat{V}_n(\omega, \gamma) = \sum_{h=1}^{H} (|I_{h,1}(\omega, \gamma)| \hat{m}_{h,1}(\omega, \gamma) \hat{m}_{h,1}(\omega, \gamma)' + |I_{h,2}(\omega, \gamma)| \hat{m}_{h,2}(\omega, \gamma) \hat{m}_{h,2}(\omega, \gamma)'), \quad (7)$$

where

$$|I_{h,1}(\omega, \gamma)| = \sum_{i=1}^n 1\{Y_i \in I_h\} 1\{\omega^T X_i - \gamma \geq 0\}$$

and

$$|I_{h,2}(\omega, \gamma)| = \sum_{i=1}^n 1\{Y_i \in I_h\} 1\{\omega^T X_i - \gamma < 0\}.$$

The weights in the PCA are chosen so that $V(\omega, \gamma)$, the population version of $\hat{V}_n(\omega, \gamma)$, has ω_0 as its largest eigenvector.

Let $\hat{v}_n(\omega, \gamma)$ be the largest eigenvector of $\hat{V}_n(\omega, \gamma)$. It is the direction along which $\hat{m}_{h,1}(\omega, \gamma)$ and $\hat{m}_{h,2}(\omega, \gamma)$ show maximal variation. The boosted sieve $\hat{\Omega}_n$ is a result of applying $\hat{v}_n$ to the simple sieve of MD directions:

$$\hat{\Omega}_n := \left\{ \hat{v}_n(\omega^{\mathrm{MD}}(P), \gamma^{\mathrm{MD}}(P)) \hat{\Sigma}_{xx}^{-1/2} : P \in \bigcup_{k=1}^{K} \mathcal{P}_k \right\}. \quad (8)$$

The term $\gamma^{\mathrm{MD}}(P)$ is the intercept that maximizes the likelihood given $\omega^{\mathrm{MD}}(P)$ and the term $\hat{\Sigma}_{xx}^{-1/2}$ is necessary to transform the estimate back to the original scale.

Experience indicates the proposed method is not sensitive to the choice of H, the number of slices, and setting $H = n/10$ works well in most applications.

5.4 Illustrative Example

The modified SIR procedure described in the previous section is very similar to the original SIR procedure. The main difference is that the subgroup structure is taken into account in the former. Note that in SIR, all terms $\tilde{X}_i \tilde{X}_j'$ are included in the covariance matrix, whereas in the modification, only terms where X_i and X_j lie on the same side of the hyperplane $\omega^T X - \gamma = 0$ are included. This additional restriction helps reduce the noise that can arise from aggregating across subgroups. To illustrate the noise issue, the performance of SIR is examined by studying a simple toy example. Set the parameters in Model (1) to the following:

$$n = 100, \quad d = 3, \quad \omega_0 = \left(\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}}, 0\right), \quad \gamma_0 = \frac{1}{4},$$
$$(\mu_{1,0}, \sigma_{1,0}^2) = (0, 4), \quad (\mu_{2,0}, \sigma_{2,0}^2) = (4, 1),$$
$$X \sim N(0, I_3).$$

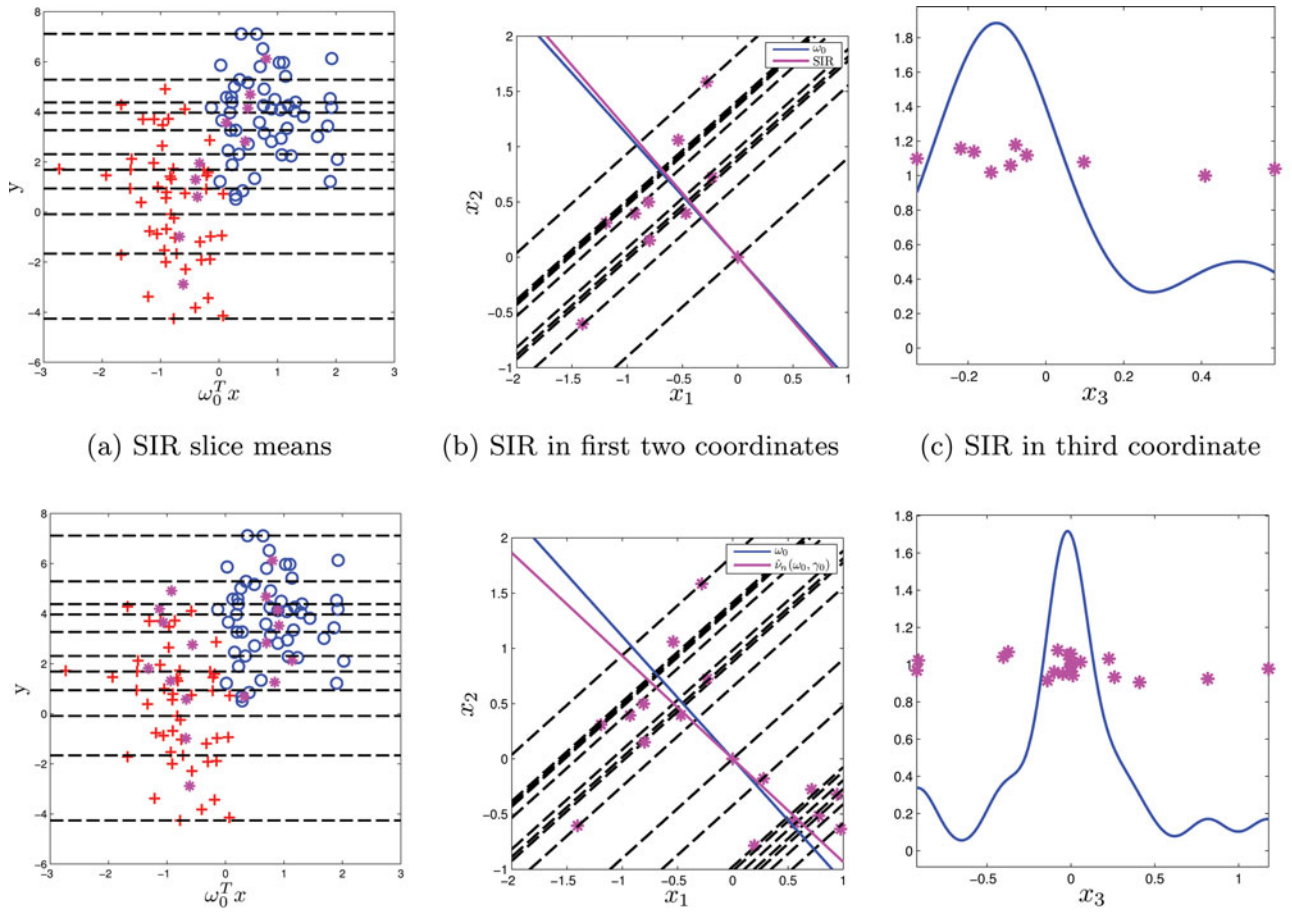
Note that the third component of ω_0 is 0 and thus the third dimension contains no information on the subgroup structure. Despite the overlap between the distributions $N(0, 4)$ and $N(4, 1)$, the surrogate variable clearly has valuable information for guiding classification.

The number of slices H is set to $n/10$ in both the modified and the original SIR procedure. The top row in Figure 1 examines various aspects of the original SIR estimator for this toy dataset. Figure 1(a) plots the projection of x onto the true direction ω_0 against the surrogate variable y. The circle and plus symbols correspond to the true subgroup membership. The asterisks in Figure 1(a) represent the sample means $\hat{m}_h$ within each slice whose boundaries are delineated by the horizontal dashed lines. The slice means exhibit variation along the ω_0 direction moving across the slices. Figure 1(b) shows the positions of the sample means $\hat{m}_h$ in the first two coordinates. The SIR estimate is compared to the true ω_0 direction. The distance between them in the first two coordinates is 0.0545. Figure 1(c) shows the distribution of the slice means in the third coordinate. The slice means are not centered at zero despite ω_0 being zero in the third coordinate. This suggests the SIR estimate will be inaccurate in the third coordinate. Indeed, the distance between the SIR estimate and ω_0 in the third coordinate is 0.2718, much higher than in the first two coordinates combined. Thus, although SIR is accurate in the first two coordinates, it is inaccurate in the third coordinate.

Next the performance of the modified SIR procedure on this toy example is examined. The second row in Figure 1 is as in the top row except the asterisks now represent the sample means $\hat{m}_{h,1}(\omega_0, \gamma_0)$ and $\hat{m}_{h,2}(\omega_0, \gamma_0)$ for $h = 1, \ldots, H$. The distance between $\hat{v}_n(\omega_0, \gamma_0)$ and ω_0 is 0.0888 in the first two coordinates, which is larger than the distance between the SIR estimate and ω_0. However, the accuracy in the third coordinate is a significant improvement over SIR. Figure 1(f) shows that the slice means $\hat{m}_{h,1}(\omega_0, \gamma_0)$ and $\hat{m}_{h,2}(\omega_0, \gamma_0)$ in the third coordinate are centered at zero. The distance between $\hat{v}_n(\omega_0, \gamma_0)$ and ω_0 in the third coordinate is found to be 0.11. Thus, overall across all three dimensions, $\hat{v}_n(\omega_0, \gamma_0)$ is more accurate than the SIR estimate.

6. CONSISTENCY

In this section, M-estimation theory is used to establish the consistency of the data-driven sieved MLE $(\hat{\omega}_n^s, \hat{\gamma}_n^s)$. Let P denote the probability measure of $Z = (X, Y)$ under Model (1). Define the empirical measure to be $\mathbb{P}_n = n^{-1} \sum_{i=1}^n \delta_{Z_i}$ where δ_z is the measure that assigns mass 1 at z and zero elsewhere.



(a) SIR slice means (b) SIR in first two coordinates (c) SIR in third coordinate

(d) Modified SIR slice means (e) Modified SIR in first two coordinates (f) Modified SIR in third coordinate

Figure 1. Toy example illustrating the differences between SIR and the proposed method of incorporating the surrogate variable described in Section 5.3. The estimate $\hat{\nu}_n(\omega_0, \gamma_0)$ is less accurate than the SIR estimate in the first two dimensions but a better overall estimate across all three dimensions. (a) SIR slice means, (b) SIR in first two coordinates, (c) SIR in third coordinate, (d) Modified SIR slice means, (e) Modified SIR in first two coordinates, (f) Modified SIR in third coordinate. The online version of this figure is in color.

For a measurable function f, let $\mathbb{P}_n f = n^{-1} \sum_{i=1}^n f(Z_i)$ be the expectation of f under the measure $\mathbb{P}_n$ and $Pf = \int f \, dP$ the expectation under P. Using the empirical processes notation described above, the likelihood expression in Equation (2) can be rewritten as

$$M_n(\omega, \gamma, \theta(\omega, \gamma)) = \mathbb{P}_n m_{\omega,\gamma,\theta(\omega,\gamma)},$$

where

$$m_{\omega,\gamma,\theta(\omega,\gamma)}(x, y) = -\log(\sigma^2(x, \omega, \gamma)) - \frac{(y - \mu(x, \omega, \gamma))^2}{\sigma^2(x, \omega, \gamma)}. \tag{9}$$

Note that the constant $1/2$ and the $\log 2\pi$ terms have been dropped as they do not affect the maximization. The following assumptions are needed:

(A1) The intercept γ_0 is known to lie in a bounded interval $[a, b]$.

(A2) The univariate random variable $\omega_0^T X$ has a strictly bounded and positive density f over $[a, b]$ with $P(\omega_0^T X < a) > 0$ and $P(\omega_0^T X > b) > 0$.

(A3) $\mu_{1,0} = \mu_{2,0}$ and $\sigma_{1,0}^2 = \sigma_{2,0}^2$ are not simultaneously true.

(A4) The surrogate variable Y has finite first and second moments, that is, $EY < \infty$ and $EY^2 < \infty$.

(A5) For any $b \in \mathbb{R}^p$, the conditional expectation $E(b^T X | \omega_0^T X)$ is linear in $\omega_0^T X$.

(A6) The covariate X has a continuous distribution.

The interval $[a, b]$ in (A1) may be estimated from the data by first calculating the direction of maximal variation of the sample covariates X and next considering the range of the resulting projections. The second assumption is satisfied for most continuous distributions of X whose support includes $[a, b]$. The third assumption ensures that the Gaussian mixture parameters are well defined. Assumption A4 is reasonable for most surrogate variables in practice. A5 is a key assumption in Li (1991) and is satisfied when the distribution of X is Gaussian or, more generally, elliptically symmetric. Finally, Assumption A6 is necessary to guarantee the semicontinuity of the function $M(\omega, \gamma, \theta(\omega, \gamma))$. Certain of these assumptions are for mathematical convenience and may be stronger than necessary. For instance, the last assumption requiring the covariate X to have a continuous distribution is quite stringent and may be relaxed at the cost of more complicated proofs. The proposed method is

later applied to real datasets in Section 9 that contain categorical covariates and the method is seen to perform well despite this.

Theorem 1. Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be iid from Model (1). Under (A1)–(A6), the data-driven sieved MLE defined in (6) using the boosted sieve in (8) is consistent for the true parameters (ω_0, γ_0).

Proof of Theorem 1. Following theorem 14.1 (Argmax Theorem) in Kosorok (2008), the following will be established to show consistency: (a) the sequence $(\hat{\omega}_n^s, \hat{\gamma}_n^s)$ is uniformly tight; (b) the map $(\omega, \gamma) \mapsto M(\omega, \gamma, \theta(\omega, \gamma))$ is upper semicontinuous with a unique maximum at (ω_0, γ_0); (c) uniform convergence of M_n to M over compact subsets K of $\mathbb{S}^p \times [a, b]$, that is,

$$\sup_{(\omega,\gamma)\in K} |M_n(\omega, \gamma, \theta(\omega, \gamma)) - M(\omega, \gamma, \theta(\omega, \gamma))| \to 0$$

in probability; and (d) the estimator "nearly" maximizes the objective function, that is, $\hat{\omega}_n^s$ and $\hat{\gamma}_n^s$ satisfy

$$M_n\left(\hat{\omega}_n^s, \hat{\gamma}_n^s, \theta\left(\hat{\omega}_n^s, \hat{\gamma}_n^s\right)\right) \geq M_n(\omega_0, \gamma_0, \theta(\omega_0, \gamma_0)) - o_P(1).$$

The first condition is easily seen to hold. Since $\hat{\omega}_n^s$ is a unit vector in $\mathbb{R}^p$, it is easy to see $||\hat{\omega}_n^s|| = O_P(1)$. The intercept estimate $\hat{\gamma}_n^s$ lies in the interval $[a, b]$ and is thus uniformly tight.

To check semicontinuity of $M(\omega, \gamma, \theta(\omega, \gamma))$, the conditional expectation of $m_{\omega,\gamma,\theta(\omega,\gamma)}$ given X is first examined. Taking the expectation with respect to the randomness in Y gives

$$\begin{aligned}
&P(m_{\omega,\gamma,\theta(\omega,\gamma)}(X, Y)|X) \\
&\quad = -\log(\sigma^2(X, \omega, \gamma)) - \frac{P\{(Y - \mu(X, \omega, \gamma))^2|X\}}{\sigma^2(X, \omega, \gamma)} \\
&\quad = -\log(\sigma^2(X, \omega, \gamma)) \\
&\qquad - \frac{P\{(Y - \mu(X, \omega, \gamma))^2 1\{\omega^T X - \gamma \geq 0\}|X\}}{\sigma^2(X, \omega, \gamma)} \\
&\qquad - \frac{P\{(Y - \mu(X, \omega, \gamma))^2 1\{\omega^T X - \gamma < 0\}|X\}}{\sigma^2(X, \omega, \gamma)} \\
&\quad = -\log(\sigma^2(X, \omega, \gamma)) \\
&\qquad - \frac{P\{(Y - \mu_1(\omega, \gamma))^2 1\{\omega^T X - \gamma \geq 0\}|X\}}{\sigma^2(X, \omega, \gamma)} \\
&\qquad - \frac{P\{(Y - \mu_2(\omega, \gamma))^2 1\{\omega^T X - \gamma < 0\}|X\}}{\sigma^2(X, \omega, \gamma)} \\
&\quad = -\log(\sigma^2(X, \omega, \gamma)) \\
&\qquad - \frac{P\{(Y - \mu_{1,0} + \mu_{1,0} - \mu_1(\omega, \gamma))^2 1\{\omega^T X - \gamma \geq 0\}|X\}}{\sigma^2(X, \omega, \gamma)} \\
&\qquad - \frac{P\{(Y - \mu_{2,0} + \mu_{2,0} - \mu_2(\omega, \gamma))^2 1\{\omega^T X - \gamma < 0\}|X\}}{\sigma^2(X, \omega, \gamma)} \\
&\quad = -\log(\sigma^2(X, \omega, \gamma)) \\
&\qquad - \frac{\left[\sigma_{1,0}^2 + (\mu_{1,0} - \mu_1(\omega, \gamma))^2\right] 1\{\omega^T X - \gamma \geq 0\}}{\sigma^2(X, \omega, \gamma)} \\
&\qquad - \frac{\left[\sigma_{2,0}^2 + (\mu_{2,0} - \mu_2(\omega, \gamma))^2\right] 1\{\omega^T X - \gamma < 0\}}{\sigma^2(X, \omega, \gamma)}.
\end{aligned}$$

Taking the expectation on both sides (this time with respect to the randomness in X) gives

$$\begin{aligned}
&M(\omega, \gamma, \theta(\omega, \gamma)) \\
&\quad = -\log\left(\sigma_1^2(\omega, \gamma)\right) P1\{\omega^T X - \gamma \geq 0\} \\
&\qquad - \log\left(\sigma_2^2(\omega, \gamma)\right) P1\{\omega^T X - \gamma < 0\} \\
&\qquad - \left\langle \left[\sigma_{1,0}^2 + (\mu_{1,0} - \mu_1(\omega, \gamma))^2\right] P1\{\omega^T X - \gamma \geq 0\} \right\rangle \\
&\qquad \Big/ \left\langle \sigma_1^2(\omega, \gamma) P1\{\omega^T X - \gamma \geq 0\} \right. \\
&\qquad \left. + \sigma_2^2(\omega, \gamma) P1\{\omega^T X - \gamma < 0\} \right\rangle \\
&\qquad - \left\langle \left[\sigma_{2,0}^2 + (\mu_{2,0} - \mu_2(\omega, \gamma))^2\right] P1\{\omega^T X - \gamma < 0\} \right\rangle \\
&\qquad \Big/ \left\langle \sigma_1^2(\omega, \gamma) P1\{\omega^T X - \gamma \geq 0\} \right. \\
&\qquad \left. + \sigma_2^2(\omega, \gamma) P1\{\omega^T X - \gamma < 0\} \right\rangle.
\end{aligned}$$

Since $P1\{\omega^T X - \gamma \leq 0\}$ is nonzero for $(\omega, \gamma) \in \mathbb{S}^p \times [a, b]$, both $\mu_1(\omega, \gamma)$ and $\sigma_1^2(\omega, \gamma)$ are well defined. Next, since X has a continuous distribution by Assumption A6, derivations in Lemma 2 in the Appendix show $\mu_1(\omega, \gamma)$ and $\sigma_1^2(\omega, \gamma)$ are both continuous in (ω, γ). It can be similarly shown that $\mu_2(\omega, \gamma)$ and $\sigma_2^2(\omega, \gamma)$ are continuous and well defined. Thus $M(\omega, \gamma, \theta(\omega, \gamma))$ is upper semicontinuous (in fact continuous) in (ω, γ).

Next the unique maximality of (ω_0, γ_0) is established. The conditional expectation of $(Y - \mu(X, \omega, \gamma))^2$ given X is uniquely minimized when $\mu(X, \omega, \gamma) = E(Y|X)$, that is, when $\omega = \omega_0$ and $\gamma = \gamma_0$. Thus $M(\cdot)$ is uniquely maximized at (ω_0, γ_0).

Establishing the third condition reduces to showing that the individual classes of functions that comprise $\{m_{\omega,\gamma,\theta(\omega,\gamma)}\}$ are Glivenko–Cantelli (GC) with integrable envelopes. Next the fact that sums, differences, products, and compositions of GC classes with integrable envelopes are GC can be used. Lemma 2 in the Appendix provides the proof for this.

Finally the last condition of near maximization is checked. Lemma 3 in the Appendix establishes the existence of a sequence $\omega_n^s \in \hat{\Omega}_n$ that converges to ω_0 and a corresponding sequence of intercept estimates $\gamma_n^s \in [a, b]$ that converges to γ_0. By definition, the sieve estimator $(\hat{\omega}_n^s, \hat{\gamma}_n^s)$ satisfies

$$M_n\left(\hat{\omega}_n^s, \hat{\gamma}_n^s, \hat{\theta}_n\left(\hat{\omega}_n^s, \hat{\gamma}_n^s\right)\right) \geq M_n\left(\omega_n^s, \gamma_n^s, \hat{\theta}_n\left(\omega_n^s, \gamma_n^s\right)\right). \quad (10)$$

Lemma 4 in the Appendix shows that

$$|M_n(\omega_n, \gamma_n, \hat{\theta}_n(\omega, \gamma)) - M_n(\omega_n, \gamma_n, \theta(\omega, \gamma))| \to 0$$

in probability for any sequence $(\omega_n, \gamma_n) \in \mathbb{S}^p \times [a, b]$. Rewriting Equation (10) (by adding and subtracting the same expressions) gives

$$\begin{aligned}
0 \leq\ & M_n\left(\hat{\omega}_n^s, \hat{\gamma}_n^s, \hat{\theta}_n\left(\hat{\omega}_n^s, \hat{\gamma}_n^s\right)\right) - M_n\left(\hat{\omega}_n^s, \hat{\gamma}_n^s, \theta\left(\hat{\omega}_n^s, \hat{\gamma}_n^s\right)\right) \\
&+ M_n\left(\omega_n^s, \gamma_n^s, \theta\left(\omega_n^s, \gamma_n^s\right)\right) - M_n\left(\omega_n^s, \gamma_n^s, \hat{\theta}_n\left(\omega_n^s, \gamma_n^s\right)\right) \\
&+ M_n\left(\hat{\omega}_n^s, \hat{\gamma}_n^s, \theta\left(\hat{\omega}_n^s, \hat{\gamma}_n^s\right)\right) - M_n\left(\omega_n^s, \gamma_n^s, \theta\left(\omega_n^s, \gamma_n^s\right)\right).
\end{aligned}$$

Applying Lemma 4 to the second and the third line above gives

$$M_n\left(\hat{\omega}_n^s, \hat{\gamma}_n^s, \theta\left(\hat{\omega}_n^s, \hat{\gamma}_n^s\right)\right) \geq M_n\left(\omega_n^s, \gamma_n^s, \theta\left(\omega_n^s, \gamma_n^s\right)\right) - o_P(1). \quad (11)$$

Now consider the following decomposition

$$\begin{aligned}
&\left|M_n(\omega_0, \gamma_0, \theta(\omega_0, \gamma_0)) - M_n\left(\omega_n^s, \gamma_n^s, \theta\left(\omega_n^s, \gamma_n^s\right)\right)\right| \\
&\quad \leq |M_n(\omega_0, \gamma_0, \theta(\omega_0, \gamma_0)) - M(\omega_0, \gamma_0, \theta(\omega_0, \gamma_0))| \\
&\qquad + \left|M_n\left(\omega_n^s, \gamma_n^s, \theta\left(\omega_n^s, \gamma_n^s\right)\right) - M\left(\omega_n^s, \gamma_n^s, \theta\left(\omega_n^s, \gamma_n^s\right)\right)\right| \\
&\qquad + \left|M\left(\omega_n^s, \gamma_n^s, \theta\left(\omega_n^s, \gamma_n^s\right)\right) - M(\omega_0, \gamma_0, \theta(\omega_0, \gamma_0))\right|.
\end{aligned}$$

The first two lines go to zero in probability by Lemma 2. The third line goes to zero in probability since M is continuous in

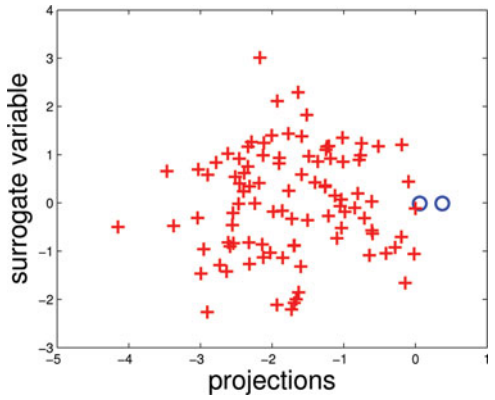


Figure 2. Estimated subgroups when there is actually only one component in the model. The plot here shows that the method gives a reasonable answer when there is only one component. The online version of this figure is in color.

(ω, γ) and (ω_n^s, γ_n^s) converges to (ω_0, γ_0). Thus,

$$\left|M_n(\omega_0, \gamma_0, \theta(\omega_0, \gamma_0)) - M_n\left(\omega_n^s, \gamma_n^s, \theta\left(\omega_n^s, \gamma_n^s\right)\right)\right| \to 0 \quad (12)$$

in probability. Combining Equations (11) and (12) gives

$$\begin{aligned} & M_n\left(\hat{\omega}_n^s, \hat{\gamma}_n^s, \theta\left(\hat{\omega}_n^s, \hat{\gamma}_n^s\right)\right) \\ & \geq M_n\left(\omega_n^s, \gamma_n^s, \theta\left(\omega_n^s, \gamma_n^s\right)\right) - o_P(1) \\ & = M_n(\omega_0, \gamma_0, \theta(\omega_0, \gamma_0)) - \left[M_n(\omega_0, \gamma_0, \theta(\omega_0, \gamma_0))\right. \\ & \quad \left. - M_n\left(\omega_n^s, \gamma_n^s, \theta\left(\omega_n^s, \gamma_n^s\right)\right)\right] - o_P(1) \\ & = M_n(\omega_0, \gamma_0, \theta(\omega_0, \gamma_0)) - o_P(1). \end{aligned}$$

Thus the near-maximization criterion for $(\hat{\omega}_n^s, \hat{\gamma}_n^s)$ is satisfied. □

7. MODEL CHECKING

Model (1) describes the ideal situation where (a) the surrogate variable arises from a two-component Gaussian mixture and (b) component membership is completely determined by a hyperplane in the covariate space. Suppose the number of components in the Gaussian mixture is one, or three or more. The case when the number of components is three or more will not be studied in depth here. In such a case, the proposed estimator is likely to merge two or more similar subgroups, which can be considered a less serious offense than splitting the sample into two subgroups when there is in fact no subgroup structure at all. Fortunately there exist several methods for determining the number of components in a Gaussian mixture. One common approach is to add a penalty function, say based on the Bayesian information criterion, to the main log-likelihood term.

To understand what happens if the proposed method is applied to the setting where there is no subgroups structure at all, consider the following simulation setting. Let $\mu_{1,0} = \mu_{2,0} = 0$ and $\sigma_{1,0}^2 = \sigma_{2,0}^2 = 1$ in Model (1). Let the dimension and sample size be set to $p = 5$ and $n = 100$, respectively. The covariate X is drawn from the standard p-variate Gaussian distribution. The first $p/2$ components of ω_0 are set to $-p^{1/2}$ and the rest to $p^{1/2}$, and the intercept is set to $1/4$. Figure 2 displays the projections onto the sieve estimated direction $\hat{\omega}_n^s$ shifted by the estimated intercept $\hat{\gamma}_n^s$ against the surrogate variable y. The resulting subgroups are indicated by different symbols and are seen to be highly unbalanced as the plus subgroup contains merely two members. This greatly suggests that there is indeed only one component in the model.

Remark 1. Because there are subgroup size constraints in the estimation process, that is, no subgroup of size one or less is allowed, for otherwise the sample variation in that group would be zero, the estimate will never result in two subgroups where one is completely empty.

Another major violation of Model (1) occurs if the separating decision boundary is not linear in x. Consider the following setup: (a) the means and variances are set to $(\mu_{1,0}, \sigma_{1,0}^2) = (0, 1)$ and $(\mu_{1,0}, \sigma_{1,0}^2) = (4, 1)$ and (b) subgroup membership is determined by the quadratic boundary $||x|| \leq 2$. Intuitively, the estimator will seek to pick out one subgroup that arises from a single Gaussian signal, while the other subgroup will be a mixture of the two Gaussian signals. Figure 3 confirms this is indeed the case. The left panel shows the estimated subgroups. The right panel plots the surrogate variable in the circle subgroup, which is clearly bimodal. In general, if the two-component Gaussian mixture assumption is confirmed to hold, then this type of diagnostic suggests the boundary is not linear in x.

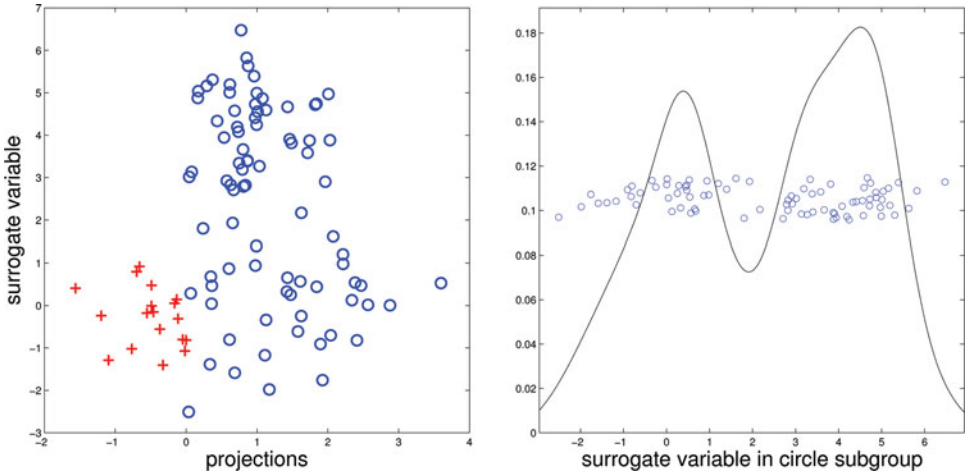


Figure 3. Left panel shows estimated subgroups when the decision boundary is not linear but quadratic. Right panel shows the bimodality of the surrogate variable in the circle subgroup. These plots suggest an easy visual tool to diagnose this type of assumption violation. The online version of this figure is in color.

The two issues discussed above are major departures from Model (1). There are certainly other ways in which the presumed model may not hold—take departures from the normal distribution, for instance. This turns out to be a rather minor issue. For one, there exist many methods to transform a univariate random variable to have an approximate Gaussian distribution. Also, simulations in the next section suggest that the methodology is robust against non-Gaussianity of the surrogate variable.

There is also the question of how to assess whether the surrogate variable approximates well the underlying class label. This is an important, albeit philosophical, issue. In some cases, the selection of an appropriate surrogate variable can be guided by previous studies. When this is not possible, a surrogate variable can be chosen that is interesting in its own right. The binary outcome of interest can be defined a posteriori with respect to the chosen surrogate variable. For instance, the surrogate variable "cholesterol level" is of interest in and of itself. The corresponding binary outcome of interest can then be defined with respect to this choice.

8. SIMULATIONS

The various simulation settings considered are summarized in Table 1. The first setting is called Stochastically Ordered (SO) because the surrogate variable Y is stochastically smaller in subgroup 1. The second setting is Non-Stochastically Ordered (NSO) since subgroup 1 has a smaller mean but a higher variance than subgroup 2. The third setting, denoted by VO for Variance Only, has identical means in the subgroups but different variances. This is a challenging setting because the noise to signal ratio is high. Finally, a setting where the surrogate variable arises from the exponential distribution is considered. This is of interest because many outcome variables related to time can be well approximated by the exponential distribution. Since Model (1) assumes normality for the surrogate variable, this setting also tests how robust the methodology is to distributional violations in Model (1).

The vector of covariates X is distributed as a standard multivariate Gaussian. Two different settings for the direction ω_0 are considered. In the first setting, which shall be referred to as "sparse," all components of ω_0 are set to zero except the first two that are set to $(2^{-1/2}, -2^{-1/2})$. This reflects situations where only a few covariates matter. In the other setting, which shall be referred to as "abundant," the first $p/2$ components of ω_0 are set to $-p^{1/2}$ and the rest to $p^{1/2}$. This reflects situations where all the covariates drive the separation between the two subgroups. The intercept is set to $\gamma_0 = 1/4$, which results in roughly 60/40 split of the data into two subgroups.

Table 1. Description of simulation settings

Simulation setting	Subgroup 1	Subgroup 2
Stochastically Ordered (SO)	$N(0, 1)$	$N(4, 1)$
Non-Stochastically Ordered (NSO)	$N(0, 4)$	$N(4, 1)$
Variance Only (VO)	$N(0, 1)$	$N(0, 4)$
Exponentials (EXP)	exp(1)	exp(10)

NOTE: The subgroups are determined by a hyperplane $\omega^T X - \gamma = 0$ and the distributions of the surrogate variable Y in each subgroup is given.

Table 2. Sparse ω_0, low-dimensional setting. Average norm difference between estimate and ω_0 over 1000 Monte Carlo simulations

Settings	NSO	SO	VO	EXP
Y clustering	0.31 (0.11)	0.25 (0.09)	0.85 (0.31)	0.48 (0.18)
X–Y clustering	0.31 (0.11)	0.25 (0.08)	0.84 (0.33)	0.48 (0.18)
EM	0.32 (0.15)	0.27 (0.13)	0.53 (0.23)	0.33 (0.13)
Regression	0.25 (0.10)	0.19 (0.07)	1.07 (0.26)	0.36 (0.13)
SIR	0.24 (0.09)	0.19 (0.07)	0.49 (0.23)	0.29 (0.12)
Simple sieve	0.22 (0.09)	0.20 (0.08)	0.33 (0.18)	0.24 (0.11)
Proposed method	*0.14* (0.07)	*0.11* (0.05)	*0.30* (0.16)	*0.20* (0.10)

NOTE: The standard error is given in the parentheses. The best estimator (lowest norm difference) is highlighted in italics.

Different ratios of sample size to dimension are considered for the simulations. In the low-dimensional problem, the sample size is set to $n = 100$ and dimension to $p = 5$, and $n = 200$, $p = 25$ for the high dimensional. For the sparse setting, Tables 2 and 3 show the average norm difference between the estimate and the true ω_0 over 1000 Monte Carlo simulations for various settings. The lowest average norm difference is highlighted in italics. Tables 4 and 5 give the corresponding results for the abundant setting.

In addition to the methods in Section 4, a comparison of the proposed methodology will also be made to the simple sieve method. In the simple sieve method, the estimator is the sieve maximum likelihood estimator (MLE) defined in (6) using the simple sieve of MD directions outlined in Section 5.2. The simulations show the proposed method outperforms the other methods in all settings considered here. The "boosting" that comes from incorporating the surrogate variable Y is seen to be crucial; the final sieve estimator offers a significant improvement over the simple sieve estimator in many settings, especially high-dimensional settings. The best competitor appears to be the SIR method though the proposed method outperforms it in every setting considered here, by large margins at times (see for instance the low-dimensional settings). Linear regression performs poorly in the low-dimensional, VO setting. The simple sieve method is consistently among the worst in the high-dimensional settings. The two clustering methods perform very similarly to each other and are decent for the NSO and SO settings, though they perform poorly for the VO and Exp simulations.

Table 3. Sparse ω_0, high-dimensional setting. Average norm difference between estimate and ω_0 over 1000 Monte Carlo simulations

Settings	NSO	SO	VO	EXP
Y clustering	0.52 (0.08)	0.43 (0.06)	1.13 (0.19)	0.75 (0.11)
X–Y clustering	0.52 (0.08)	0.43 (0.06)	1.14 (0.20)	0.75 (0.11)
EM	0.50 (0.10)	0.43 (0.08)	0.78 (0.18)	0.54 (0.09)
Regression	0.45 (0.07)	0.35 (0.06)	1.28 (0.10)	0.61 (0.09)
SIR	0.44 (0.08)	0.34 (0.05)	0.82 (0.19)	0.54 (0.11)
Simple sieve	0.95 (0.11)	0.91 (0.11)	1.01 (0.13)	0.98 (0.12)
Our method	*0.40* (0.08)	*0.31* (0.05)	*0.72* (0.14)	*0.49* (0.10)

NOTE: The standard error is given in the parentheses. The best estimator (lowest norm difference) is highlighted in italics.

Table 4. Abundant ω_0, low-dimensional setting

Settings	NSO	SO	VO	EXP
Y clustering	0.32 (0.12)	0.25 (0.09)	0.85 (0.33)	0.47 (0.17)
X–Y clustering	0.32 (0.12)	0.25 (0.09)	0.84 (0.34)	0.47 (0.17)
EM	0.32 (0.15)	0.26 (0.12)	0.55 (0.23)	0.33 (0.13)
Regression	0.25 (0.10)	0.20 (0.07)	1.05 (0.26)	0.36 (0.13)
SIR	0.24 (0.09)	0.19 (0.07)	0.50 (0.24)	0.29 (0.12)
Simple sieve	0.22 (0.09)	0.21 (0.08)	0.35 (0.18)	0.25 (0.11)
Proposed method	*0.14* (0.07)	*0.11* (0.06)	*0.33* (0.18)	*0.19* (0.10)

NOTE: Average norm difference between estimate and ω_0 over 1000 Monte Carlo simulations. The standard error is given in the parentheses. The best estimator (lowest norm difference) is highlighted in italics.

Table 5. Abundant ω_0, high-dimensional setting

Settings	NSO	SO	VO	EXP
Y clustering	0.52 (0.08)	0.43 (0.06)	1.14 (0.19)	0.75 (0.11)
X–Y clustering	0.52 (0.08)	0.43 (0.06)	1.15 (0.19)	0.75 (0.11)
EM	0.50 (0.09)	0.43 (0.07)	0.78 (0.18)	0.54 (0.08)
Regression	0.45 (0.07)	0.35 (0.06)	1.28 (0.10)	0.61 (0.09)
SIR	0.44 (0.08)	0.34 (0.06)	0.83 (0.19)	0.53 (0.10)
Simple sieve	0.94 (0.12)	0.90 (0.10)	1.01 (0.13)	0.98 (0.12)
Proposed method	*0.41* (0.08)	*0.31* (0.06)	*0.72* (0.13)	*0.50* (0.09)

NOTE: Average norm difference between estimate and ω_0 over 1000 Monte Carlo simulations. The standard error is given in the parentheses. The best estimator (lowest norm difference) is highlighted in italics.

Simulation run times for low-dimensional $n = 100$, $p = 5$ and high-dimensional $n = 200$, $p = 25$ settings are as follows. For the former, 473.125580 sec were needed for 100 Monte Carlo runs, resulting in approximately 5 sec for each individual run. For the latter, 2883.471355 sec were needed for 100 Monte Carlo runs, which gives an approximate run time of half-a-minute for each individual run. The current implementation relies heavily on for-loops in MATLAB. This is known to be computationally slow and the algorithm has great potential to be improved. Finally, the performance of the methods seem quite insensitive to the setting of ω_0.

9. EXAMPLES

The proposed method is applied to three health-related datasets. The first two come from the UCI Machine Learning Repository (Frank and Asuncion 2010). The third was used as an example in chapter 1 of Hastie, Tibshirani, and Friedman (2003) and is available at the book's web site. The full list of variables and preprocessing steps for each dataset are described in the Appendix. The subgroups discovered by the proposed method will be compared to the ones given by the binary variable, if available. For the first two data examples, the method is able to achieve, without using the binary training labels, classification accuracy comparable to logistic regression, a fully supervised procedure. For the third dataset that does not have binary labels, an interpretation for the subgroups discovered by the proposed method is offered.

9.1 Pima Indian Diabetes Dataset

The Pima Indian Diabetes dataset contains information on eight clinical measurements, including a 2-hr insulin measurement, for 768 individuals. It also records whether each individual later developed diabetes. The proposed method will be applied to find a diabetes and nondiabetes subgroup. The corresponding surrogate variable should approximately satisfy the normality assumption in Model (1) and be relevant to the binary event of interest. The 2-hr insulin measurement is a reasonable surrogate for the unobserved binary outcome and was approximately Gaussian. Furthermore, 374 out of the 768 total cases were missing the 2-hr insulin measurement. Since classification in the proposed method is completely determined by a separating hyperplane in the covariate space, it does not make use of the surrogate variable for classifying future objects. Thus the surrogate variable can be a quantity that is difficult to measure or obtain, as is the case here, since it is used only in the learning process.

The projections of the covariates onto the estimated separating hyperplane are shown in the first panel of Figure 4. A smoothed histogram of the 2-hr insulin measurement in each discovered subgroup is shown in the next two panels of Figure 4. There is a bit of departure from Gaussianity here, but it does not seem severe enough to affect the performance of the method. The circle subgroup corresponds well with the individuals who later develop diabetes and the plus subgroup with those who did not.

The classification test error of the proposed method is assessed on an independent test set consisting of the 374 individuals missing the 2-hr insulin measurement. The percentage reported is the misclassification rate on this test set. The error rates of logistic regression and three "off-the-shelf" methods described in Section 4—Y clustering, X–Y clustering, and the EM algorithm—are also examined. The bottom row in Table 6

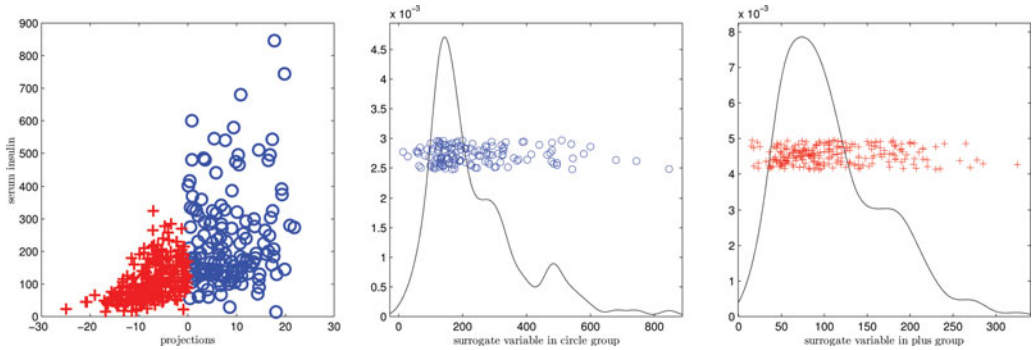


Figure 4. Diabetes dataset. The first panel shows the projections onto the estimated separating hyperplane versus the surrogate variable, 2-hr insulin. The second and third panels show the distribution of the surrogate variable in each of the discovered subgroups. The online version of this figure is in color.

Table 6. Classification accuracy

Dataset	The proposed method	Logistic regression	*Y* clustering	*X–Y* clustering	EM
Heart	0.23 (0.06)	*0.18 (0.04)*	0.40 (0.05)	0.43 (0.09)	0.41 (0.05)
Diabetes	0.27	*0.26*	0.29	0.29	0.30

NOTE: For the Heart dataset, accuracy is measured by 10-fold cross-validation. Standard error across the folds is given in the parentheses. For the Diabetes dataset, accuracy is measured by the test error on a held-out test set of 374 cases who are missing the 2-hr insulin measurement. The italicized text indicates the method with the lowest error rate.

shows the performance of each method for this data example. To make the methods comparable, the surrogate variable used in the proposed method is not included in the logistic regression model. Logistic regression is a rather minor improvement over the proposed method considering that it requires trained labels. The EM, Y clustering, and X-Y clustering are all slightly less accurate than the proposed method.

9.2 Cleveland Heart Disease Dataset

This dataset contains information on heart disease for 297 individuals. There are 13 clinical measurements in addition to the diagnosis, that is, presence/absence of heart disease. The data were collected from the Cleveland Clinic Foundation. The proposed method was applied to find a subgroup with heart disease and a subgroup without. The maximum-heart-rate-achieved variable was chosen as the surrogate variable because it was approximately normally distributed and is correlated to cardiac mortality (Lauer et al. 1999).

The projections of the covariates onto the estimated separating hyperplane are shown in the first panel of Figure 5. A smoothed histogram of the maximum-heart-rate measurement for each discovered subgroup is shown in the last two panels of Figure 5. The Gaussian assumption seems to hold quite well and there is no indication that the two-component structure is incorrect. The plus subgroup corresponds well with the individuals who were diagnosed with heart disease and the circle subgroup with those who were not.

Because the dataset is relatively small, a large independent test set could not be afforded and the 10-fold cross-validation error rate is reported instead. The first row of Table 6 shows the error rates of the proposed method, logistic regression, and three off-the-shelf methods. Unsurprisingly, the logistic regression has the best accuracy because it uses trained labels. The proposed method performs relatively well considering that it does not use labeled data at all. The other methods, EM, *Y* clustering, and *X–Y* clustering, perform quite poorly for this dataset.

9.3 Prostate Cancer Dataset

The Prostate dataset comes from a study that examined the relationship between the level of Prostate-Specific Antigen (PSA) and certain clinical measures in men who were about to receive a radical prostatectomy. The dataset has information on 97 subjects and eight covariate measurements. Using the log PSA (lpsa) as the surrogate variable, the proposed method is applied to find two subgroups that differ in terms of lpsa. There is no binary outcome variable provided in this dataset. However, PSA is known to be associated with more severe grades of prostate cancer, so the binary outcome could be taken to be "more severe" versus "less severe" grades of prostate cancer.

Figure 6 is a scatterplot of the continuous covariates in the Prostate dataset. The subgroups found by the proposed method are displayed as different symbols, with the circle subgroup having higher lpsa. Taking a look at Figure 6, patients with higher lpsa (circle) indeed have higher log cancer volume (lcavol) and log prostate weight (lweight).

Other interesting covariates include the categorical variables seminal vesicle invasion (SVI) and Gleason score (gleason). The presence of SVI generally means a poor outlook for the patient and a high Gleason score means the cancer is more likely to have spread past the prostate. Surprisingly, patients without SVI are split roughly evenly between the subgroups, but those with SVI are entirely from the circle subgroup (higher lpsa), see Figure 7(a). The circle subgroup also has higher Gleason scores on average than the plus subgroup, see Figure 7(b).

10. DISCUSSION

In this article, a new type of machine learning task was introduced called latent supervised learning. This type of learn-

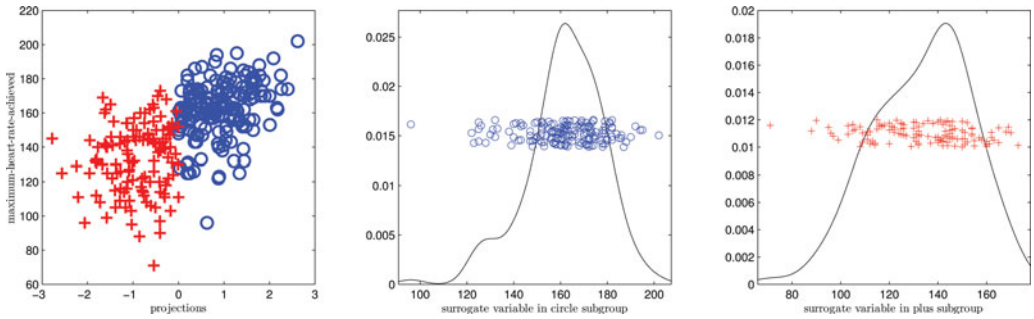


Figure 5. Heart dataset. The first panel shows the projections onto the estimated separating hyperplane versus the surrogate variable, maximum-heart-rate achieved. The second and third panels show the distribution of the surrogate variable in each of the discovered subgroups. The online version of this figure is in color.

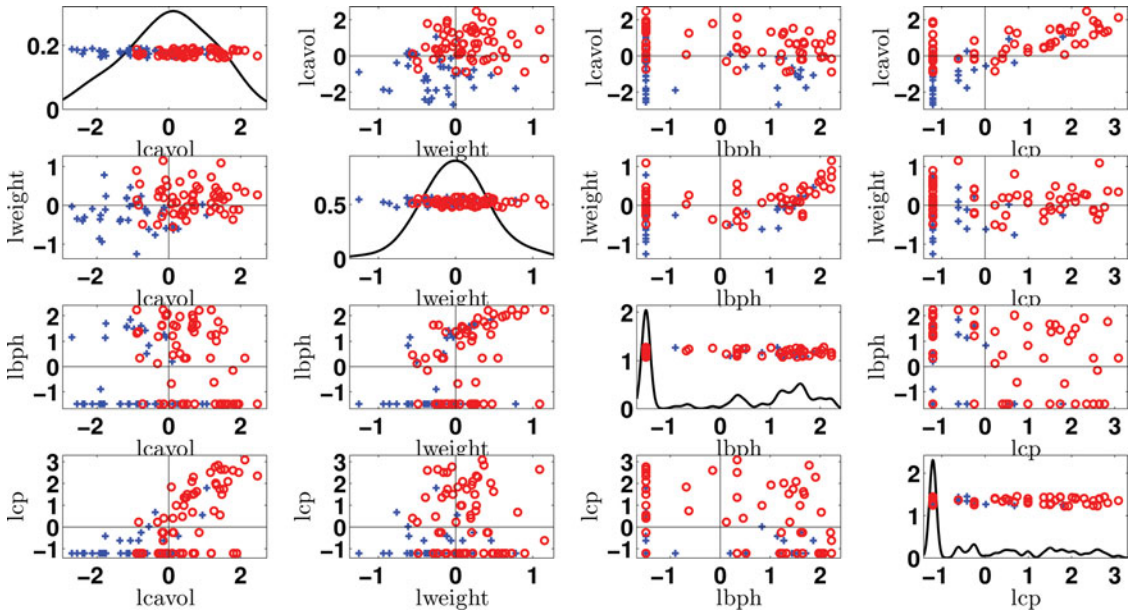
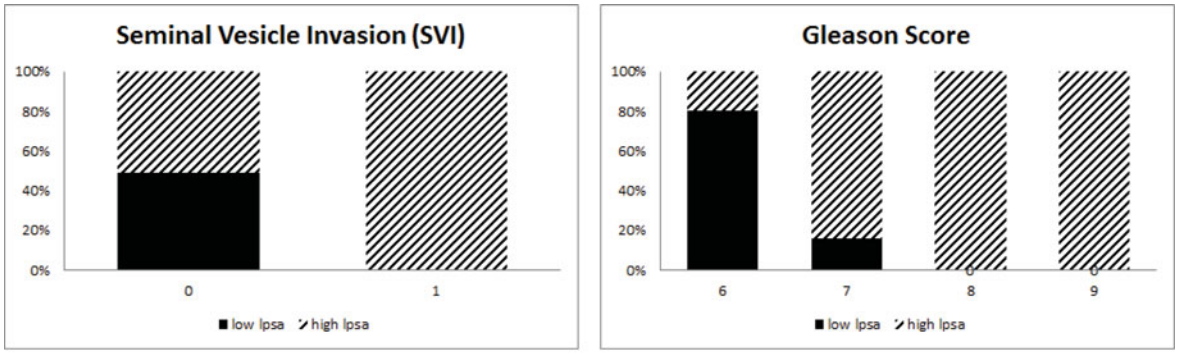


Figure 6. Scatterplot of the continuous covariates in the Prostate dataset. A complete list of the full names of the variables is given in the Appendix. The symbols represent the subgroups found by the proposed method where the circle subgroup has higher lpsa values on average. Note that the circle subgroup has higher log cancer volume (lcavol) and higher log prostate weight (lweight), two variables that are linked to the severity of the cancer. The online version of this figure is in color.

ing represents a paradigm shift away from the conventional assumption that labels are either completely unavailable (as in unsupervised learning) or when available, hard-coded truths (as in supervised learning) to the more realistic idea that labels are actually "fuzzy" in nature. A specific problem in latent supervised learning was studied called the change-line classification problem. The proposed estimator was shown to be accurate on simulated data and provide meaningful and interpretable results on real datasets.

A major challenge to the proposed methodology is high-dimensional data settings. The simulations in Section 8 show that the performance of the proposed method suffers when dimension is increased from 5 to 25. The high-dimensional setting presents various challenges already familiar to modern statisticians. If sparsity is assumed for the coefficients of the normal vector to the separating hyperplane, likelihood penalization is a promising approach. This is an active field of research and many existing techniques can be borrowed for extending the proposed method to high-dimensional settings. Another approach is to improve the construction of the simple sieve. This currently requires training a binary linear classifier given a particular enumeration of the class labels. When the dimension is very high, noise will hamper the accuracy of the candidate directions in the simple sieve. The boosting in the second part of the sieve construction may not be enough to compensate for the effects of the noise. A simple solution is to first reduce the dimension of the covariates using principal component analysis. Directions along which there is very little variation are unlikely to play an important role in the classification rule and thus can be safely discarded.

Other future problems include extending Model (1) to the case where Y is a survival time. It is of direct clinical interest



(a) (b)

Figure 7. Distributions of the subgroups discovered by the proposed method in the Prostate dataset for the categorical variables SVI and gleason. Note that the circle subgroup (higher lpsa) mostly comprises the higher end of the Gleason score and comprises the presence for SVI entirely.

to find subgroups that are homogeneous with respect to some biological characteristic, gene expression for example, but differ with respect to survival. The extension to survival data would be straightforward if the survival times were completely observed but will require careful development in the case of censoring. Work has begun to address the case of right-censored survival data.

Another problem of interest is to extend the change-line *classification* problem studied here to the change-line *regression* problem. In the regression problem, the regression function changes when crossing a separating hyperplane in the covariate space. This could be of interest to fields such as personalized medicine which postulate that different subgroups experience different treatment effects.

As for the theoretical aspects of the proposed methodology, rate of convergence and weak convergence of the estimator remain open problems. A closer look at the simple sieve defined in Section 5.2 and Lemma 3 reveals that the simple sieve is much richer than necessary. However despite this, it does not perform nearly as well as the boosted sieve as evidenced by simulations in Section 8. This suggests that boosting may induce a speed-up of the rate of convergence and hence yield better performance on finite samples.

APPENDIX

A.1. PROOFS

Lemma 1. Let $\Theta_{K_1,K_2} := \{\theta = (\mu_1, \sigma_1^2, \mu_2, \sigma_2^2) : |\mu_1|, |\mu_2| < K_1, \frac{1}{K_2} < \sigma_1^2, \sigma_2^2 < K_2\}$, where $K_1 \in (0, \infty)$ and $K_2 \in (1, \infty)$. Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be iid from Model (1), under Assumptions (A1)—(A6), the class of functions

$$\{m_{\omega,\gamma,\theta}(x, y) : (\omega, \gamma) \in \mathbb{S}^d \times [a, b], \theta \in \Theta_{K_1,K_2}\}$$

is GC.

Proof of Lemma 1. Recall the definition of m:

$$m_{\omega,\gamma,\theta}(x, y) = -\log\left[\left(\sigma_1^2 - \sigma_2^2\right) 1\{\omega^T x - \gamma \geq 0\} + \sigma_2^2\right] - \frac{[y - (\mu_1 - \mu_2)1\{\omega^T x - \gamma \geq 0\} - \mu_2]^2}{\left(\sigma_1^2 - \sigma_2^2\right) 1\{\omega^T x - \gamma \geq 0\} + \sigma_2^2}.$$

Lemma 8.12 in Kosorok (2008) establishes the measurability of the class of indicator functions $\{1(\omega^T x - \gamma \geq 0) : (\omega, \gamma) \in K\}$. Standard Vapnik-Chervonenkis (VC) class arguments then show the class $\{1(\omega^T x - \gamma \geq 0) : (\omega, \gamma) \in K\}$ is GC. The classes

$$\{\mu_j : (\omega, \gamma) \in \mathbb{S}^d \times [a, b], \theta \in \Theta_{K_1,K_2}\}$$

and

$$\left\{\sigma_j^2 : (\omega, \gamma) \in \mathbb{S}^d \times [a, b], \theta \in \Theta_{K_1,K_2}\right\},$$

for $j = 1, 2$ are trivially GC as they are not data dependent. Furthermore, these classes have integrable (in fact finite) envelopes by the definition of Θ_{K_1,K_2}. The preservation result given by corollary 9.27 (i) and (ii) in Kosorok (2008) can now be applied to show the classes

$$\{(\mu_1 - \mu_2)1\{\omega^T x - \gamma \geq 0\} + \mu_2\} \quad \text{with envelope } K_1,$$
$$\left\{\left(\sigma_1^2 - \sigma_2^2\right) 1\{\omega^T x - \gamma \geq 0\} + \sigma_2^2\right\} \quad \text{with envelope } K_2,$$

and

$$\left\{\frac{1}{\left(\sigma_1^2 - \sigma_2^2\right) 1\{\omega^T x - \gamma \geq 0\} + \sigma_2^2}\right\} \quad \text{with envelope } K_2,$$

are GC with finite envelopes. Using corollary 9.27 (iii), we have the class

$$\left\{\log\left(\left(\sigma_1^2 - \sigma_2^2\right) 1\{\omega^T x - \gamma \geq 0\} + \sigma_2^2\right)\right\}$$

that is GC with integrable envelope $\log K_2$. The class

$$\{y : (\omega, \gamma) \in \mathbb{S}^d \times [a, b], \theta \in \Theta_{K_1,K_2}\}$$

is GC simply by the regular Law of Large Numbers. The function $|y|$ is an envelope for this class and is integrable since $E|Y| < \infty$ by (A4). The class

$$\{(y - (\mu_1 - \mu_2)1\{\omega^T x - \gamma \geq 0\} - \mu_2)^2\}$$

has an integrable envelope since $EY^2 < \infty$ by (A4). Using corollary 9.27 (i) and (iii), we can show the class is GC. Applying corollary 9.27 (i) and (ii) one last time gives the desired result that $m_{\omega,\gamma,\theta}$ is itself a GC class. □

Lemma 2. Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be iid from Model (1). Under Assumptions (A1)–(A6), the class of functions

$$\{m_{\omega,\gamma,\theta(\omega,\gamma)} : (\omega, \gamma) \in \mathbb{S}^d \times [a, b]\}$$

is GC.

Proof of Lemma 2. We can apply Lemma 1 directly to show the desired result. First we show there exists some $K_1 \in [0, \infty)$ such that $|\mu_1(\omega, \gamma)|, |\mu_2(\omega, \gamma)| < K_1$ for all $(\omega, \gamma) \in \mathbb{S}^d \times [a, b]$. For the class of functions $\{\mu_1(\omega, \gamma)\}$, we can write

$$\begin{aligned}
\mu_1(\omega, \gamma) &= \frac{P(Y1\{\omega^T X - \gamma \geq 0\})}{P(1\{\omega^T X - \gamma \geq 0\}} \\
&= \frac{1}{P(1\{\omega^T X - \gamma \geq 0\}} P\left(\left\{Y1\left\{\omega_0^T X - \gamma_0 \geq 0\right\}\right.\right. \\
&\quad \left.\left. + Y1\left\{\omega_0^T X - \gamma_0 \geq 0\right\}\right\} 1\{\omega^T X - \gamma \geq 0\}\right) \\
&= \frac{1}{P(1\{\omega^T X - \gamma \geq 0\}} P\left(\left\{\mu_{1,0}1\left\{\omega_0^T X - \gamma_0 \geq 0\right\}\right.\right. \\
&\quad \left.\left. + \mu_{2,0}1\left\{\omega_0^T X - \gamma_0 \geq 0\right\}\right\} 1\{\omega^T X - \gamma \geq 0\}\right) \\
&\leq \max \mu_{1,0}, \mu_{2,0}.
\end{aligned}$$

The above also shows $\mu_1(\omega, \gamma) \geq \min \mu_{1,0}, \mu_{2,0}$ and thus $|\mu_1(\omega, \gamma)| \leq K_1 = \max |\mu_{1,0}|, |\mu_{2,0}|$. Similarly, we can show $|\mu_2(\omega, \gamma)| \leq K_1$.

Next we show there exists some $K_2 \in (0, \infty)$ such that $\frac{1}{K_2} < \sigma_1^2(\omega, \gamma), \sigma_2^2(\omega, \gamma) < K_2$. We have

$$\begin{aligned}
&\sigma_1^2(\omega, \gamma) \\
&= \frac{P\{(Y - \mu_1(\omega, \gamma))^2 1\{\omega^T X - \gamma \geq 0\}\}}{P(1\{\omega^T X - \gamma \geq 0\}} \\
&= \left[P\left\{(Y - \mu_1(\omega, \gamma))^2 1\{\omega^T X - \gamma \geq 0\}\left\{1\left\{\omega_0^T X - \gamma_0 \geq 0\right\}\right.\right.\right. \\
&\quad \left.\left.\left. + 1\left\{\omega_0^T X - \gamma_0 < 0\right\}\right\}\right]\right] \Big/ [P(1\{\omega^T X - \gamma \geq 0\}] \\
&= \left[P\left\{\left(\sigma_{1,0}^2 + (\mu_{1,0} - \mu_1(\omega, \gamma))^2\right) 1\left\{\omega_0^T X - \gamma_0 \geq 0\right\}\right.\right. \\
&\quad \left.\left. \times\ 1\{\omega^T X - \gamma \geq 0\}\right\}\right] \Big/ [P(1\{\omega^T X - \gamma \geq 0\}] \\
&\quad + \left[P\left\{\left(\sigma_{2,0}^2 + (\mu_{2,0} - \mu_2(\omega, \gamma))^2\right) 1\left\{\omega_0^T X - \gamma_0 < 0\right\}\right.\right. \\
&\quad \left.\left. \times\ 1\{\omega^T X - \gamma \geq 0\}\right\}\right] \Big/ [P(1\{\omega^T X - \gamma \geq 0\}].
\end{aligned}$$

Thus we have $c_1 \leq \sigma_1^2(\omega, \gamma) \leq c_2$, where

$$c_1 = \inf_{\omega,\gamma \in K} \min\left\{\sigma_{1,0}^2 + (\mu_{1,0} - \mu_1(\omega, \gamma))^2, \sigma_{2,0}^2 + (\mu_{2,0} - \mu_1(\omega, \gamma))^2\right\}$$

and

$$c_2 = \sup_{\omega,\gamma \in K} \max\left\{\sigma_{1,0}^2 + (\mu_{1,0} - \mu_1(\omega, \gamma))^2, \sigma_{2,0}^2 + (\mu_{2,0} - \mu_1(\omega, \gamma))^2\right\}.$$

Because $|\mu_1(\omega,\gamma)$ is bounded, we have that c_1 and c_2 are both finite. Let K_2 be such that $1/K_2 < c_1$ and $K_2 < c_2$. Then $\frac{1}{K_2} < \sigma_1^2(\omega,\gamma) < K_2$. A similar argument can be applied to $\sigma_2^2(\omega,\gamma)$. □

Lemma 3. Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be iid from Model (1). Under Assumptions (A1)–(A6), there exists a sequence ω_n in $\hat{\Omega}_n$ that converges to ω_0, where $\hat{\Omega}_n$ is the boosted sieve defined in Equation (8). Further the corresponding intercept estimate $\gamma_n \in [a,b]$ is consistent for γ_0.

Proof of Lemma 3. Recall that the sieve $\hat{\Omega}_n$ is populated by boosted MD directions

$$\left\{ \hat{v}_n(\omega^{\mathrm{MD}}(P), \gamma^{\mathrm{MD}}(P))^T \hat{\Sigma}_{xx}^{-1/2} : P \in \bigcup_k \mathcal{P}_k \right\},$$

where $\hat{v}_n$ is the largest eigenvector of $\hat{V}_n$, which was defined in Equation (7) as

$$\hat{V}_n(\omega,\gamma) = \sum_{h=1}^{H} (|I_{h,1}(\omega,\gamma)|\hat{m}_{h,1}(\omega,\gamma)\hat{m}_{h,1}(\omega,\gamma)' + |I_{h,2}(\omega,\gamma)|\hat{m}_{h,2}(\omega,\gamma)\hat{m}_{h,2}(\omega,\gamma)').$$

Let $p_{h,1}(\omega,\gamma) = E1\{Y \in I_h, \omega^T X - \gamma \geq 0\}$ and $p_{h,2}(\omega,\gamma) = E1\{y \in I_h, \omega^T X - \gamma < 0\}$ be the theoretical proportions in each subslice. Let $Z = \Sigma_{xx}^{-1}[X - EX]$ be the standardized covariate and $m_{h,1}(\omega,\gamma) = E[E(Z|Y)|Y \in I_h, \omega^T X - \gamma \geq 0]$ and $m_{h,2}(\omega,\gamma) = E[E(Z|Y)|Y \in I_h, \omega^T X - \gamma < 0]$ be the theoretical means in each subslice. Define the matrix

$$V(\omega,\gamma) = \sum_{h=1}^{H} p_{h,1}(\omega,\gamma) m_{h,1} m_{h,1}(\omega,\gamma)' + \sum_{h=1}^{H} p_{h,2}(\omega,\gamma) m_{h,2}(\omega,\gamma) m_{h,2}(\omega,\gamma)'.$$

It is easy to see that $\hat{V}_n(\omega,\gamma)$ is uniformly consistent for $V(\omega,\gamma)$ over $(\omega,\gamma) \in \mathbb{S}^d \times [a,b]$. By corollary 3.1 in Li (1991) which uses (A5), the largest eigenvector of $V(\omega,\gamma)$ falls in the linear space generated by $\omega_0 \Sigma_{xx}^{1/2}$. Since $\hat{v}_n(\omega,\gamma)$ is consistent for the largest eigenvector of $V(\omega,\gamma)$ and $\hat{\Sigma}_{xx}$ is consistent for Σ_{xx}, we have $\hat{v}_n(\omega,\gamma)^T \hat{\Sigma}_{xx}^{-1/2} \to \omega_0$ uniformly over $(\omega,\gamma) \in \mathbb{S}^d \times [a,b]$. We can find a corresponding intercept estimate that is consistent in the following manner. For a consistent estimator ω_n of ω_0, the corresponding intercept estimate given by

$$\begin{aligned} \gamma_n &= \min \arg\max_{\gamma \in \hat{\Gamma}_n(\omega_n)} L_n(\omega_n, \gamma, \hat{\theta}_n(\omega_n,\gamma)) \\ &= \min \arg\max_{\gamma \in \hat{\Gamma}_n(\omega_n)} M_n(\omega_n, \gamma, \hat{\theta}_n(\omega_n,\gamma)) \end{aligned}$$

is consistent for γ_0. To see this, we invoke the Argmax Theorem in Kosorok (2008) along with the continuity of M to show γ_n converges to the argmax over γ of $M(\omega_0,\gamma,\theta(\omega_0,\gamma))$. By the proof in Theorem 1, however, the argmax of $M(\omega_0,\gamma,\theta(\omega_0,\gamma))$ over γ is γ_0. □

Lemma 4. Let M_n and $\hat{\theta}_n$ be as defined in Section 6 and Equation (5), respectively. Under Assumptions (A1)–(A6), we have

$$\sup_{(\omega,\gamma)\in\mathbb{S}^d\times[a,b]} |M_n(\omega,\gamma,\hat{\theta}_n(\omega,\gamma)) - M_n(\omega,\gamma,\theta(\omega,\gamma))| \to 0 \quad \text{(A.1)}$$

in probability.

Proof. In general, if a class of functions $\mathcal{F}$ is GC then $|P_n f - Pf| \to 0$ in probability uniformly in f varying over $\mathcal{F}$. It is obvious then that $|P_n \hat{f}_n - P\hat{f}_n| \to 0$ in probability for every sequence of random functions $\hat{f}_n$ contained in $\mathcal{F}$. Furthermore if $\hat{f}_n \to f_0$ and the random sequence is dominated so that $P\hat{f}_n \to Pf_0$, then it follows that $P_n\hat{f}_n \to Pf_0$.

In Lemma 2, it was shown that for some K_1, K_2, $\theta(\omega,\gamma) \in \Theta_{K_1,K_2}$ for all $(\omega,\gamma) \in \mathbb{S}^d \times [a,b]$. It follows that there exists a δ-neighborhood around $\theta(\omega,\gamma)$ that lives in $\Theta_{K_1',K_2'}$ for some K_1', K_2' for all $(\omega,\gamma) \in \mathbb{S}^d \times [a,b]$. To see this, set $K_1' = K_1 + \delta$ and let K_2' be such that $K_2 + \delta < K_2'$ and $1/K_2' < 1/K_2 - \delta$. Thus the enlarged class

$$\mathcal{F}^\delta = \{m_{\omega,\gamma,\theta}(x,y) : (\omega,\gamma) \in \mathbb{S}^d \times [a,b], \theta \in \theta^\delta(\omega,\gamma)\}$$

is contained in $\Theta_{K_1',K_2'}$ and is hence GC. By Lemma 5, $\hat{\theta}_n$ is uniformly consistent for θ over $(\omega,\gamma) \in \mathbb{S}^d \times [a,b]$. Then we have $\hat{\theta}_n(\omega,\gamma) \in \theta^\delta(\omega,\gamma)$ for n large enough for all $(\omega,\gamma) \in \mathbb{S}^d \times [a,b]$. This implies the class of functions

$$\{m_{\omega,\gamma,\hat{\theta}_n(\omega,\gamma)}(x,y) : (\omega,\gamma) \in \mathbb{S}^d \times [a,b]\} \subset \mathcal{F}^\delta$$

for n large enough and is hence GC, that is,

$$\sup_{(\omega,\gamma)\in\mathbb{S}^d\times[a,b]} |M_n(\omega,\gamma,\hat{\theta}_n(\omega,\gamma)) - M(\omega,\gamma,\hat{\theta}_n(\omega,\gamma))| \to 0$$

in probability. Then we have by the continuity of M

$$\sup_{(\omega,\gamma)\in\mathbb{S}^d\times[a,b]} |M_n(\omega,\gamma,\hat{\theta}_n(\omega,\gamma)) - M(\omega,\gamma,\theta(\omega,\gamma))| \to 0$$

in probability. Using this and Lemma 2 once again, we have

$$\begin{aligned} &\sup_{(\omega,\gamma)\in\mathbb{S}^d\times[a,b]} |M_n(\omega,\gamma,\hat{\theta}_n(\omega,\gamma)) - M_n(\omega,\gamma,\theta(\omega,\gamma))| \\ &\leq \sup_{(\omega,\gamma)\in\mathbb{S}^d\times[a,b]} |M_n(\omega,\gamma,\hat{\theta}_n(\omega,\gamma)) - M(\omega,\gamma,\theta(\omega,\gamma))| \\ &\quad + \sup_{(\omega,\gamma)\in\mathbb{S}^d\times[a,b]} |M_n(\omega,\gamma,\theta(\omega,\gamma)) - M(\omega,\gamma,\theta(\omega,\gamma))| \\ &= o_P(1) + o_P(1). \end{aligned}$$

Thus we have proven (A.1). □

Lemma 5. Let $\hat{\theta}_n$ be as defined in Equation (5). Under Assumptions (A1)–(A6), $\hat{\theta}_n(\omega,\gamma)$ is uniformly consistent for $\theta(\omega,\gamma)$ over $(\omega,\gamma) \in \mathbb{S}^d \times [a,b]$, that is,

$$\sup_{(\omega,\gamma)\in\mathbb{S}^d\times[a,b]} |\hat{\theta}_n(\omega,\gamma) - \theta(\omega,\gamma)| \to 0$$

in probability.

Proof. We showed in Lemma 1, the classes $\{y : (\omega,\gamma) \in \mathbb{S}^d \times [a,b], \theta \in \Theta_{K_1,K_2}\}$ are GC with an integrable envelope. We also showed the class of indicator functions $1\{\omega^T X - \gamma \geq 0\}$ is GC. We can apply corollary 9.27 (ii) in Kosorok (2008) to see the numerator of $\hat{\mu}_1$, $n^{-1}\sum_{i=1}^n y_i 1\{\omega^T x_i - \gamma \geq 0\}$ converges in probability to $P(Y1\{\omega^T X - \gamma \geq 0\})$ uniformly over $(\omega,\gamma) \in \mathbb{S}^d \times [a,b]$. The denominator $n^{-1}\sum_{i=1}^n 1\{\omega^T x_i - \gamma \geq 0\}$ converges in probability to $P(1\{\omega^T X - \gamma \geq 0\})$, which is bounded away from zero, uniformly over $(\omega,\gamma) \in \mathbb{S}^d \times [a,b]$. Thus, $\hat{\mu}_1(\omega,\gamma)$ converges in probability to $\mu_1(\omega,\gamma)$ uniformly. A similar argument can be applied to $\hat{\mu}_2(\omega,\gamma)$.

The estimated variance $\hat{\sigma}_1^2$ is given by

$$\hat{\sigma}_1^2(\omega,\gamma) = \frac{\sum_i (y_i - \hat{\mu}_1(\omega,\gamma))^2 1\{\omega^T x_i - \gamma \geq 0\}}{\sum_i 1\{\omega^T x_i - \gamma \geq 0\}}.$$

The numerator converges to $P((Y - \mu_1(\omega,\gamma))^2 1\{\omega^T X - \gamma \geq 0\})$, while the denominator converges to $P(1\{\omega^T X - \gamma \geq 0\})$, which is bounded away from zero by assumption. Thus $\hat{\sigma}_1^2(\omega,\gamma)$ converges in probability to $\sigma_1^2(\omega,\gamma)$ uniformly over $(\omega,\gamma) \in \mathbb{S}^d \times [a,b]$. A similar argument can be applied to $\hat{\sigma}_2^2(\omega,\gamma)$. □

A.2. DATA PRE-PROCESSING AND SUMMARY OF DATA FEATURES

The Cleveland Heart Disease Dataset is available at the UCI Machine Learning Repository (Frank and Asuncion 2010). The dataset actually

Appendix A.1. Dataset features

Heart	Diabetes	Prostate
1. Age in years	1. Number of times pregnant	1. log cancer volume (lcavol)
2. 1 = male, 0 = female	2. Plasma glucose concentration	2. log prostate weight (lweight)
3. Chest pain type	3. Diastolic blood pressure	3. age
4. Resting blood pressure	4. Triceps skin fold thickness	4. log of the amount of benign prostatic hyperplasia (lbph)
5. Serum cholesterol	5. Body mass index	5. Seminal vesicle invasion (SVI)
6. Fasting blood sugar indicator	6. Diabetes pedigree function	6. log of capsular penetration (lcp)
7. Resting electrocardiographic results	7. Age (years)	7. Gleason score (gleason)
8. Maximum heart rate achieved		8. Percent of Gleason scores 4 or 5 (pgg45)
9. Exercise induced angina indicator		
10. ST depression induced by exercise relative to rest		
11. Slope of the peak exercise ST segment		
12. Number of major vessels colored by fluoroscopy		
13. 3 = normal, 6 = fixed defect, 7 = reversible defect		

contains 76 features, but most published work seems to focus on the subset listed in Table A.1. There is a feature titled "goal," valued from 0 to 4, corresponding to the degree of heart disease in the patient. The presence of heart disease (1,2,3,4) was combined into a single group versus the absence of heart disease (0).

The Pima Indian Diabetes Dataset is also available at the UCI machine learning repository. The binary variable of interest is whether the patient has diabetes. The dataset also contains information on various clinical measurements, summarized in Table A.1. This is a large dataset with 768 cases. Some minimal preprocessing was necessary as certain cases have missing values encoded by 0 (where a 0 value would actually be biologically impossible). Three-Nearest Neighbors was used to impute the missing values. Also, a large number of cases (about 300) are missing feature 5, the 2-hr serum insulin measurement, which was taken to be the surrogate variable.

The Prostate Cancer dataset was analyzed in Chapter 1 of Hastie, Tibshirani, and Friedman (2003) and is available on the authors' web site. No preprocessing was necessary.

[Received August 2012. Revised January 2013.]

REFERENCES

Carlstein, E., Müller, H., and Siegmund, D. (1994), "Change-Point Problems," *IMS Lecture Notes 23*, Hayward, CA: IMS. [958]

Fleming, T. R. (2005), "Surrogate Endpoints and FDA's Accelerated Approval Process," *Health Affairs*, 24, 67–78. [957]

Frank, A., and Asuncion, A. (2010), *UCI Machine Learning Repository*, Irvine, CA: University of California. [965,969]

Geman, S., and Hwang, C.-R. (1982), "Nonparametric Maximum Likelihood Estimation by the Method of Sieves," *The Annals of Statistics*, 10, 401–414. [959]

Grenander, U. (1981), *Abstract Inference*, New York: Wiley. [959]

Hastie, T., Tibshirani, R., and Friedman, J. H. (2003), *The Elements of Statistical Learning*, Berlin: Springer, corrected edition. [957,965,970]

Jain, A. K., Murty, M. N., and Flynn, P. J. (1999), "Data Clustering: A Review," *ACM Computing Surveys*, 31, 264–323. [957]

Kang, C. (2011), "New Statistical Learning Methods for Chemical Toxicity Data Analysis," Ph.D. thesis, University of North Carolina at Chapel Hill. [958]

Kosorok, M. R. (2008), *Introduction to Empirical Processes and Semiparametric Inference (Springer Series in Statistics)* (1st ed.), New York: Springer. [962,968,969]

Li, K.-C. (1991), "Sliced Inverse Regression for Dimension Reduction," *Journal of the American Statistical Association*, 86, 316–327. [958,961,969]

Lauer, M. S., Francis, G. S., Okin, P. M., Pashkow, F. J., Snader, C. E., and Marwick, T. H. (1999), "Impaired Chronotropic Response to Exercise Stress Testing as a Predictor of Mortality," *JAMA: The Journal of the American Medical Association*, 281, 524–529. [966]

Scholkopf, B., and Smola, A. J. (2001), *Learning With Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*, Cambridge, MA: MIT Press. [959]

Shen, X., Shi, J., and Wong, W. H. (1999), "Random Sieve Likelihood and General Regression Models," *Journal of the American Statistical Association*, 94, 835–846. [959]

Tibshirani, R., Hastie, T., Narasimhan, B., and Chu, G. (2002), "Diagnosis of Multiple Cancer Types by Shrunken Centroids of Gene Expression," *Proceedings of the National Academy of Sciences*, 99, 6567–6572. [959]