HHS Public Access

Author manuscript

Biometrics. Author manuscript; available in PMC 2019 May 31.

Published in final edited form as:

Biometrics. 2017 March ; 73(1): 220–231. doi:10.1111/biom.12556.

Sufficient dimension reduction for censored predictors

Diego Tomassi[1], **Liliana Forzani**[1], **Efstathia Bura**[2], and **Ruth Pfeiffer**[3,*]

[1]Instituto de Matemática Aplicada del Litoral and Facultad de Ingeniería Química, CONICET and UNL, Santa Fe, Argentina

[2]Department of Statistics, George Washington University, Washington, DC, U.S.A

[3]Biostatistics Branch, National Cancer Institute, Bethesda, MD 20892, U.S.A

Summary

In many molecular epidemiologic and clinical investigations multiple biomarkers are measured simultaneously using high throughput technologies. The analysis of multi-marker data poses two challenges. First, the fairly high dimensionality of the correlated markers makes modeling and variable selection challenging. Second, censoring of marker measurements due to lower and upper limits of detection needs to be accommodated in the analysis. To directly address the question how censored markers relate to a particular outcome we propose several approaches of varying computational complexity to analyzing censored predictors using likelihood-based sufficient dimension reduction (SDR) methods. We extend the theory and the likelihood-based SDR framework in two ways: (a) we accommodate censored predictors directly in the likelihood, and (b) we incorporate variable selection in the likelihood via a penalty term. We find linear combinations that contain all the information from correlated markers, i.e. are sufficient for modeling and prediction of an outcome variable, while accounting for left and right censoring of the markers. These methods apply to any type of outcome, including continuous and categorical outcomes and are efficient. Careful evaluations and comparisons of the methods using data from a study conducted to evaluate the associations of 51 inflammatory markers and lung cancer risk, and in simulations show that explicit accounting for the censoring in the likelihood set-up can lead to appreciable gains in efficiency and prediction accuracy.

Keywords

Informative missingness; Limits of detection; Missing data; Penalized likelihood; Shrinkage

1. Introduction

New technologies allow investigators to measure multiple biomarkers simultaneously for research on disease etiology, diagnosis, and outcomes following diagnosis. A recent example is the development of a marker panel to study the impact of chronic inflammation on cancer risk. Chronic inflammation and immune dysregulation are now recognized as important etiologic factors for many cancers (see e.g. Mantovani et al., 2008). To comprehensively

* pfeiffer@mail.nih.gov.

evaluate a wide range of markers to elucidate pathways involved in carcinogenesis, investigators at the National Cancer Institute and various companies designed a multiplex immune panel capable of measuring up to 79 analytes simultaneously. This panel has been used in several epidemiologic investigations of cancer, including one on the risk of lung cancer (Shiels et al., 2013). Our work was motivated by a study conducted to replicate and expand on the associations of inflammatory markers and lung cancer risk found in Shiels et al. (2013) using the same marker panel, and to assess the potential of an ”inflammation score” for lung cancer risk prediction (Shiels et al., 2015).

The analysis of multimarker panel data poses two challenges. First, the fairly high dimensionality of the correlated markers makes modeling and variable selection challenging. Second, censoring of marker measurements due to lower and upper limits of detection needs to be accommodated in the analysis to avoid inconsistent or inefficient results. For example, in the original lung cancer study serum levels of 68 inflammation markers were analyzed. The percentage of values below the lowest limit of detection (LLOD) was < 25% for 38 markers, 25% to 50% 5 markers, 50% to 75% for 7 markers, and 75% to 90% for 18 markers. Five markers had both, lower and upper limits of detection. For analyses, marker levels were categorized, where the choices of category cut-points were dependent on the amount of censored data. For example, markers with < 25% of individuals below the LLOD were categorized into quartiles and markers with 75% to 90% of individuals below the LLOD were categorized as ”undetectable” and ”detectable”. Using these categories, eleven markers were associated with lung cancer risk in marginal logistic regression models (Shiels et al., 2013). However, categorization of continuous markers can lead to a loss of information and thus loss of power to detect associations. Categorization also distorts correlations among the markers which limits joint analysis of all the markers and the ability to fully interpret any findings.

Only a few approaches have been proposed to combine information from multiple correlated markers that are also left and/or right censored. For a binary outcome, Dong et al. (2014) estimated the moments of correlated biomarkers with LLODs separately in the two outcome groups assuming multivariate censored normal distributions, and used the estimated moments to combine the markers into a linear score for prediction. In contrast, the focus of etiologic research is features of the model itself, such as odds ratios from logistic regression. One approach to handle censored values is to impute them, then analyze the data using standard regression models, and combine results from multiple imputed datasets to accommodate the uncertainty due to imputation (Rubin, 1987). As censored data are not missing at random (see, e.g. Rubin, 1987), the missing data mechanism needs to be modeled. General software packages for multiple imputation, e.g. *mice* in *R* (van Buuren and Groothuis-Oudshoorn, 2011), do not allow the specification of censored distributions. Lee, Kong, and Weissfeld (2012) used Gibbs sampling to impute left censored marker values assuming multivariate normality of the markers, and also allowed the marker mean to depend on covariates or the outcome. The imputed marker values were then included in a regression model and the additional variability from the imputation accommodated in the variance estimation (Rubin, 1987). However, this approach is computationally challenging with a large number of correlated markers, and in simulations and the data example the authors used only two and three markers, respectively. Another limitation of multiple

imputation is that it may not yield fully efficient results and methods for variable selection are limited (Chen and Wang, 2013).

To directly address the question how censored markers relate to a particular outcome we propose several approaches of varying computational complexity to analyzing censored predictors using sufficient dimension reduction (SDR) approaches, specifically likelihood-based SDR (Cook and Forzani, 2009, 2008) assuming that the markers given the response are jointly normal. We extend the theory and the likelihood-based SDR framework in two ways: (a) we accommodate censored predictors directly in the likelihood, and (b) we incorporate variable selection in the likelihood via a penalty term. We find linear combinations that contain all the information in correlated markers, i.e. are sufficient, for modeling and prediction of an outcome variable, while accounting for left and right censoring of the markers. The methods we develop are appealing as they apply generally to any type of outcome, including continuous and categorical outcomes, and are efficient.

The rest of the paper is organized as follows. After a brief overview of SDR, and specifically likelihood-based SDR (Section 2), we introduce likelihood-based SDR for censored data and propose an EM algorithm for computation (Section 3). In Section 4 we apply the methods to inflammatory markers from the lung cancer replication study and comprehensively compare the results from various SDR approaches to those obtained using multiple imputations. We assess the performance of our methods extensively in simulations in Section 5.

2. Background: sufficient dimension reduction (SDR)

We briefly introduce linear SDR methodology to provide the context for SDR for regression or classification with censored predictors.

2.1 Overview of dimension reduction approaches

We are interested in inferring the relationship between a univariate response variable Y and a covariate vector $\mathbf{Z} = (Z_1, \ldots, Z_p)^T \in \mathbb{R}^p$. When p is large, modeling is challenging as it is difficult to visualize how Y changes as a function of the covariates. The goal of dimension reduction is to reduce the complexity of the regression/classification problem. In particular, Sufficient Dimension Reduction (SDR) (Cook, 1998) aims to find a function $\mathbf{R}: \mathbb{R}^p \rightarrow \mathbb{R}^d$ with $d \leq p$, that contains the same information about Y as $\mathbf{Z}$. That is, $F(Y|\mathbf{Z}) = F(Y|\mathbf{R}(\mathbf{Z}))$, where $F(\cdot|\cdot)$ is the conditional distribution function of Y given the second argument. This version of dimension reduction is called *sufficient* because $\mathbf{R}(\mathbf{Z})$ replaces the predictor vector $\mathbf{Z}$ without any loss of information on Y.

With a few exceptions (e.g. Fukumizu, Bach, and Jordan, 2004), mostly linear sufficient transformations, $\mathbf{R}(\mathbf{Z}) = \boldsymbol{\alpha}^T\mathbf{Z}$ with $\boldsymbol{\alpha} \in \mathbb{R}^{p\times d}$, have been studied and used in SDR methodology, (e.g. Li, 1991; Cook and Forzani, 2008). The reduction $\boldsymbol{\alpha}^T\mathbf{Z}$ is not unique since for any invertible matrix $\mathbf{A} \in \mathbb{R}^{d\times d}$, $F(Y|\boldsymbol{\alpha}^T\mathbf{Z}) = F(Y|\mathbf{A}\boldsymbol{\alpha}^T\mathbf{Z})$, and therefore the parameter of interest is not $\boldsymbol{\alpha}$ per se but the span of its columns. In the sequel $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_d)$ denotes a basis of the linear subspace spanned by the columns of $\boldsymbol{\alpha}$, and $\mathbf{R}(\mathbf{Z}) = \boldsymbol{\alpha}^T\mathbf{Z} = (\boldsymbol{\alpha}_1^T\mathbf{Z}, \ldots, \boldsymbol{\alpha}_d^T\mathbf{Z})$ a *linear sufficient reduction* for the regression of Y on $\mathbf{Z}$. The dimension d of $\boldsymbol{\alpha}$ is called the *structural dimension* of the regression of Y on $\mathbf{Z}$, and can take

on any value in the set $\{0, 1, \ldots, p\}$. When $d < p$ the complexity of the regression is reduced. A detailed exposition of linear SDR methodology is given in Cook (1998).

Originally, SDR methods estimated $\boldsymbol{\alpha}$ using functions of moments of the conditional distribution of $\mathbf{Z}|Y$, i.e. using inverse regression (IR). IR has the advantage that modeling p predictors given a univariate response is much simpler than the forward modeling of Y as a function of p variables. A few examples include Sliced Inverse Regression (SIR; Li, 1991), Sliced Average Variance Estimation (SAVE; Cook and Weisberg, 1991), and Directional Regression (DR; Li and Wang, 2007). These methods require different conditions on the marginal distribution of the predictors to hold to yield sufficient reductions.

Recently, likelihood-based IR has been proposed, which assumes that $\mathbf{Z}|Y$ has a specific distribution or belongs to a family of distributions (Cook, 2007; Cook and Forzani, 2009, 2008). It derives from the fact that

$$Y \mid \mathbf{Z} \overset{d}{=} Y \mid \mathbf{R}(\mathbf{Z}) \quad \text{iff} \quad \mathbf{Z} \mid (\mathbf{R}(\mathbf{Z}), Y) \overset{d}{=} \mathbf{Z} \mid \mathbf{R}(\mathbf{Z}). \quad (1)$$

The equivalence in (1) means that if one treats Y as a parameter and finds a *sufficient statistic* $\mathbf{R}(\mathbf{Z})$ for Y using the distribution of $\mathbf{Z}|Y$, then $\mathbf{R}(\mathbf{Z})$ is also a *sufficient reduction* for $\mathbf{Z}$ in the forward regression of Y on $\mathbf{Z}$. It also reveals that the intrinsic dimension of the regression of Y on $\mathbf{Z}$ is the dimension of the sufficient statistic for Y in the inverse model $\mathbf{Z}|Y$. This result, which first appeared in Cook (2007), yields a powerful tool for obtaining sufficient reductions in regression and is the basic premise of likelihood-based IR which has two main features: (a) the reduction is *sufficient*, in contrast to moment-based approaches; and (b) the maximum likelihood estimates (MLEs) of the reduction can be obtained, which are thus optimal under the true model with respect to efficiency.

2.2 Likelihood-based SDR

We focus on two likelihood-based SDR methods to find the efficient estimator of the linear reductions of the predictors: Principal Fitted Components (PFC, Cook and Forzani, 2008) and Likelihood Acquired Directions (LAD, Cook and Forzani, 2009). Before we discuss the application and extension of these methods to censored data we first summarize them.

Both LAD and PFC require that $\mathbf{Z}|Y$ be normally distributed to find the *sufficient* reduction and its MLE. Both methods yield consistent estimators of the sufficient reduction when the normality assumption is relaxed, and conditions are placed on the first two moments of the predictors (Section 4 and Section 3.2 of Cook and Forzani, 2009, 2008, respectively). The two methods differ in that for PFC, $\mathbf{Z}$ depends on Y only through its conditional mean, whereas in LAD the dependence extends to the conditional variance. Consider the following multivariate model for the inverse regression of $\mathbf{Z}$ on Y

$$\mathbf{Z}_y := \mathbf{Z} \mid (Y = y) = \boldsymbol{\mu}_y + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N_p(\mathbf{0}, \boldsymbol{\Delta}_y), \quad (2)$$

where $\boldsymbol{\mu}_Y \in \mathbb{R}^{p\times 1}$ and $\text{var}(\mathbf{Z}|Y=y) = \boldsymbol{\Delta}_y$ in LAD, or $\text{var}(\mathbf{Z}|Y=y) = \boldsymbol{\Delta}$ in PFC. The conditional mean $\boldsymbol{\mu}_y$, as well as $\boldsymbol{\Delta}_y$, are unknown functions of y, with $\boldsymbol{\Delta}_y$ and $\boldsymbol{\Delta}$ positive definite and otherwise unconstrained.

Let $\boldsymbol{\mu} = E(\boldsymbol{\mu}_y)$ and $\boldsymbol{\Delta} = E(\boldsymbol{\Delta}_y)$. The goal is to find a sufficient reduction $\mathbf{R}(\mathbf{Z}) = \boldsymbol{\alpha}^T\mathbf{Z}$ with $\boldsymbol{\alpha} \in \mathbb{R}^{p\times d}$ and $\text{rank}(\boldsymbol{\alpha}) = d < p$. Under model (2), Theorem 1 of Cook and Forzani (2009) determines that

$$\text{span}(\boldsymbol{\alpha}) = \boldsymbol{\Delta}^{-1}\text{span}(\boldsymbol{\mu}_Y - \boldsymbol{\mu}, Y \in S_Y) \cup \text{span}(\boldsymbol{\Delta}_Y^{-1} - \boldsymbol{\Delta}^{-1}, Y \in S_Y), \quad (3)$$

where S_Y is the sample space of Y. Notice that when $\boldsymbol{\Delta}_Y$ does not depend on Y, $\text{span}(\boldsymbol{\Delta}_Y^{-1} - \boldsymbol{\Delta}^{-1})$ is an empty set and the shifted and scaled conditional mean $\boldsymbol{\mu}_Y$ contains all the information about the reduction $\boldsymbol{\alpha}$. When $\boldsymbol{\Delta}_Y$ is not constant, (3) states that to recover $\text{span}(\boldsymbol{\alpha})$, the contribution of the inverse regression variance is also needed. If orthogonal directions to $\boldsymbol{\Delta}^{-1}\text{span}(\boldsymbol{\mu}_Y - \boldsymbol{\mu})$ lie in the $\text{span}(\boldsymbol{\Delta}_Y^{-1} - \boldsymbol{\Delta}^{-1})$, these directions together with $\boldsymbol{\Delta}^{-1}\text{span}(\boldsymbol{\mu}_Y - \boldsymbol{\mu})$ comprise a sufficient reduction. From (3) we can rewrite the parameters as (Cook and Forzani, 2009, Proposition 1)

$$\boldsymbol{\mu}_Y - \boldsymbol{\mu} = \boldsymbol{\Delta}\boldsymbol{\alpha}\boldsymbol{\nu}_Y \text{ and } \boldsymbol{\Delta}_Y - \boldsymbol{\Delta} = \boldsymbol{\Delta}\boldsymbol{\alpha}\mathbf{T}_Y\boldsymbol{\alpha}^T\boldsymbol{\Delta}, \quad (4)$$

where $\boldsymbol{\nu}_Y \in \mathbb{R}^{d\times 1}$, $d = \dim(\boldsymbol{\Delta}^{-1}\text{span}(\boldsymbol{\mu}_Y - \boldsymbol{\mu}, Y \in S_Y) \cup \text{span}(\boldsymbol{\Delta}_Y^{-1} - \boldsymbol{\Delta}^{-1}, Y \in S_Y))$ with $\text{E}(\boldsymbol{\nu}_Y) = 0$ and $\mathbf{T}_Y \in \mathbb{R}^{d\times d}$ with $\text{E}(\mathbf{T}_Y) = 0$.

To facilitate the optimization of the likelihood for $\mathbf{Z}|Y$ with respect to the parameters of interest in (2), based on Proposition 2 in Cook and Forzani (2009), an alternative way to write model (2) under (4) is

$$\boldsymbol{\alpha}^T\mathbf{Z} \mid Y \sim N(\boldsymbol{\alpha}^T\boldsymbol{\nu}_Y + \boldsymbol{\alpha}^T\boldsymbol{\Delta}\boldsymbol{\alpha}\boldsymbol{\nu}_Y, \boldsymbol{\alpha}^T\boldsymbol{\Delta}_Y\boldsymbol{\alpha}), \quad (5)$$

$$\boldsymbol{\alpha}_0^T\mathbf{Z} \mid (\boldsymbol{\alpha}^T\mathbf{Z}, Y) \sim N(\mathbf{H}\boldsymbol{\alpha}^T\mathbf{Z} + (\boldsymbol{\alpha}_0^T - \mathbf{H}\boldsymbol{\alpha}^T)\boldsymbol{\mu}, \mathbf{D}), \quad (6)$$

where $\boldsymbol{\nu}_Y \in \mathbb{R}^{d\times 1}$, $\mathbf{D} = (\boldsymbol{\alpha}_0^T\boldsymbol{\Delta}^{-1}\boldsymbol{\alpha}_0)^{-1}$, $\mathbf{H} = (\boldsymbol{\alpha}_0^T\boldsymbol{\Delta}\boldsymbol{\alpha})(\boldsymbol{\alpha}^T\boldsymbol{\Delta}\boldsymbol{\alpha})^{-1}$, and $\boldsymbol{\alpha}_0 \in \mathbb{R}^{p\times(p-d)}$ is the semi-orthogonal complement of $\boldsymbol{\alpha}$. Under the PFC assumption of constant variance, $\boldsymbol{\Delta}_Y$ is replaced by $\boldsymbol{\Delta}$ in (5).

For a random sample $(Y_i, \mathbf{Z}_i)$, $i = 1, \ldots, n$, and using (5) and (6), the likelihood is

$$
\begin{aligned}
L_d(\mathbf{Z} \mid Y = y) = & -\frac{np}{2}\log 2\pi - \frac{n}{2}\log \mid \mathbf{D} \mid -\frac{1}{2}\sum_y n_y \log \mid \boldsymbol{\alpha}^T \boldsymbol{\Delta}_y \boldsymbol{\alpha} \mid \\
& -\frac{1}{2}\sum_y n_y [\boldsymbol{\alpha}^T(\tilde{\mathbf{z}}_y - \boldsymbol{\mu} - \boldsymbol{\Delta}\boldsymbol{\alpha}\boldsymbol{\nu}_y)]^T (\boldsymbol{\alpha}^T \boldsymbol{\Delta}_y \boldsymbol{\alpha})^{-1} [\boldsymbol{\alpha}^T(\tilde{\mathbf{z}}_y - \boldsymbol{\mu} - \boldsymbol{\Delta}\boldsymbol{\alpha}\boldsymbol{\nu}_y)] \\
& -\frac{1}{2}\sum_y n_y (\tilde{\mathbf{z}}_y - \boldsymbol{\mu})^T \mathbf{K}\mathbf{D}^{-1}\mathbf{K}^T (\tilde{\mathbf{z}}_y - \boldsymbol{\mu}) - \frac{1}{2}\sum_y n_y \mathrm{tr}\left(\boldsymbol{\alpha}^T \tilde{\boldsymbol{\Delta}}_y \boldsymbol{\alpha} (\boldsymbol{\alpha}^T \boldsymbol{\Delta}_y \boldsymbol{\alpha})^{-1}\right) \\
& -\frac{1}{2}\sum_y n_y \mathrm{tr}\left(\mathbf{K}\mathbf{D}^{-1}\mathbf{K}^T \tilde{\boldsymbol{\Delta}}_y\right),
\end{aligned}
\tag{7}
$$

where n_y denotes the sample size, $\tilde{\mathbf{z}}_y$ is the sample mean and $\tilde{\boldsymbol{\Delta}}_y$ is the corresponding sample covariance matrix in the outcome group $Y = y$ (see (10) in Cook and Forzani, 2009). The matrices $\mathbf{D}$ and $\mathbf{H}$ are defined in the text below equation (6) and $\mathbf{K} = \boldsymbol{\alpha}_0 - \boldsymbol{\alpha}\mathbf{H}^T$. Under LAD, the parameters in (7) are $\boldsymbol{\theta}_{LAD} = (\boldsymbol{\alpha}, \boldsymbol{\mu}, \boldsymbol{\nu}_y, \mathbf{D}, \mathbf{H}, \boldsymbol{\Delta}, \boldsymbol{\alpha}^T\boldsymbol{\Delta}_y\boldsymbol{\alpha})$, and under PFC ($\boldsymbol{\Delta}_Y = \boldsymbol{\Delta}$) $\boldsymbol{\theta}_{PFC} = (\boldsymbol{\alpha}, \boldsymbol{\mu}, \boldsymbol{\nu}_y, \mathbf{D}, \mathbf{H}, \boldsymbol{\alpha}^T\boldsymbol{\Delta}\boldsymbol{\alpha})$. To compute the MLE $\hat{\boldsymbol{\alpha}}$ of the sufficient reduction $\boldsymbol{\alpha}^T\mathbf{Z}$, maximizing (7) for all the parameters $\boldsymbol{\theta}_{LAD}$ or $\boldsymbol{\theta}_{PFC}$ is required.

Remark—The above likelihood can be modified to accommodate continuous Y. For PFC, this is done by using a linear model for $\boldsymbol{\nu}_Y$, $\boldsymbol{\nu}_Y = \boldsymbol{\beta}\mathbf{f}_Y$, where $\boldsymbol{\beta} \in \mathbb{R}^{d \times r}$, $d \leq r$ has rank d and $\mathbf{f}_Y \in \mathbb{R}^r$ is a known vector-valued function of the response with $\mathrm{E}(\mathbf{f}_Y) = 0$ (see Adragni and Cook (2009) for details). The choice of $\mathbf{f}_Y$ can be guided by the p plots of Z_j versus Y, $j = 1, \ldots, p$, as described e.g. in Chapter 10 of Cook (1998). For LAD, the sample space S_Y for a continuous Y is divided into H non-overlapping slices $S_1, \ldots, S_H$ and then modeled based on equation (2), where $\boldsymbol{\mu}_h = E(\mathbf{Z} \mid Y \in S_h)$ and $\boldsymbol{\Delta}_y = \mathrm{var}(\mathbf{Z} \mid Y \in S_h)$, $h = 1, \ldots, H$ are the within slice moments.

3. Likelihood-based SDR for censored predictors

We present several approaches for applying likelihood-based SDR to censored data and then extend the theory and the likelihood framework of Section 2.2 in two ways: (a) we accommodate censored predictors directly, and (b) we incorporate variable selection in the LAD likelihood via a penalty term.

3.1 Sufficient dimension reduction for censored predictors

As in Section 2, Y denotes the response variable and $\mathbf{Z} = (Z_1, \ldots, Z_p)^T \in \mathbb{R}^p$ the markers (covariates) that relate to Y through equation (2). However, instead of $\mathbf{Z}$, we only observe a censored version $\mathbf{X} = (X_1, \ldots, X_p)^T$, defined component-wise as

$$
X_j = \begin{cases} a_j & \text{if } Z_j \leq a_j, \\ Z_j & \text{if } Z_j \in (a_j, b_j) \\ b_j & \text{if } Z_j \geq b_j, \end{cases} \tag{8}
$$

where a_j and b_j, denote the *known* lower and upper limits of detection for marker $j = 1, \ldots, p$.

Theorem 1—*Assume* $\mathbf{Z}$ *and* Y *are related through model* (2) *and* $\mathbf{X}$ *are the censored observations based on* $\mathbf{Z}$ *defined in* (8). *If* $\boldsymbol{\alpha}$ *satisfies* (3), *i.e.* $\mathbf{R}(\mathbf{Z}) = \boldsymbol{\alpha}^T\mathbf{Z}$ *is the sufficient reduction for the regression of* Y *on* $\mathbf{Z}$, *then* $\mathbf{R}(\mathbf{X}) = \boldsymbol{\alpha}^T\mathbf{X}$ *is a sufficient reduction for the regression of* Y *on* $\mathbf{X}$.

Proof: From Section 2.2, $\mathbf{R}(\mathbf{Z}) = \boldsymbol{\alpha}^T\mathbf{Z}$ with $\boldsymbol{\alpha}$ given by (3) is such that $\mathbf{Z}|(\mathbf{R}(\mathbf{Z}), Y)$ does not depend on Y, and therefore $\mathbf{X}|(\mathbf{R}(\mathbf{X}), Y)$ does not depend on Y. In fact, for any realization $\mathbf{x}$ of $\mathbf{X}$, let $S(\mathbf{x})$ be the set of possible values of $\mathbf{Z}$ given by (8). Then

$$f(\mathbf{X} = \mathbf{x} \mid \mathbf{R}(\mathbf{X}) = \mathbf{R}(\mathbf{x}), Y) = f(\mathbf{Z} \in S(\mathbf{x}) \mid \mathbf{R}(\mathbf{Z}) \in \mathbf{R}(S(\mathbf{x})), Y)$$
$$= f(\mathbf{Z} \in S(\mathbf{x}) \mid \mathbf{R}(\mathbf{Z}) \in \mathbf{R}(S(\mathbf{x}))) = f(\mathbf{X} = \mathbf{x} \mid \mathbf{R}(\mathbf{X}) = \mathbf{R}(\mathbf{x})).$$

Following Cook (2007), the above equality implies that $\mathbf{R}(\mathbf{X}) = \boldsymbol{\alpha}^T\mathbf{X}$ is a sufficient reduction for the regression of Y on $\mathbf{X}$, i.e. the coefficients in the linear combinations of the reduction of the censored $\mathbf{X}$ are the same as the coefficients of $\mathbf{Z}$, if the latter were observed.

3.2 PFC and LAD for censored predictors

We propose several approaches to estimating the sufficient reduction for censored predictors that we compare extensively using the inflammation marker data example and in simulations.

PFC and LAD—The simplest approach is to apply standard PFC or LAD to the observed $\mathbf{X}$ given by (8). Under model (2) for $\mathbf{Z}$, the censored observations $\mathbf{X}$ satisfy moment conditions that ensure consistent estimation of $\mathbf{R}(\mathbf{X}) = \boldsymbol{\alpha}^T\mathbf{X}$ based on LAD and PFC (Cook and Forzani, 2009). However, the resulting estimates $\hat{\boldsymbol{\alpha}}$ are not efficient, as they are no longer MLEs.

PFC and LAD with moments computed under censoring—A somewhat improved approach is to replace the moment estimates in (7) with those estimated under censoring, e.g. using the algorithm proposed by Lee and Scott (2012). We call this approach cmLAD (cmPFC) for ”censored moments LAD (PFC)”. It also leads to consistent but not efficient estimates $\hat{\boldsymbol{\alpha}}$.

PFC and LAD applied to data with censored values imputed—We estimated $\hat{\boldsymbol{\mu}}_y$, $\hat{\boldsymbol{\Delta}}_y$ using a censored normal distribution separately in the groups defined by Y and then created ten imputed datasets by imputing censored values z_{io} for study subject i from the conditional normal distribution $z_{io}|z_{im}$, with parameters derived from $\hat{\boldsymbol{\mu}}_y$, $\hat{\boldsymbol{\Delta}}_y$. We analyzed the imputed data sets using PFC and LAD (”MI-LAD” and ”MI-PFC”).

PFC and LAD likelihood for censored predictors—Here we extend the theory for PFC and LAD estimation to censored predictors (cPFC and cLAD) and obtain MLEs of $\boldsymbol{\alpha}$.

All coordinates of $\mathbf{Z}$ that fall outside the detectable range, $\prod_{j=1}^{p}(a_j, b_j)$, are censored and their exact values are unknown. For a given sample $(Y_i, \mathbf{X}_i)$, $i = 1, \ldots, n$, each vector $\mathbf{X}_i$ may have different censoring patterns. We thus index the censored (missing) and uncensored (observed) coordinates by m_i and o_i, respectively, so that the related random vector $\mathbf{Z}_i$ can be partitioned into $\mathbf{Z}_i = [\mathbf{Z}_{m_i}^T, \mathbf{Z}_{o_i}^T]^T$, where $\mathbf{Z}_{m_i} = (Z_i^{(j)}, j = 1, \ldots, |\, m_i \,|)$ denotes the censored and $\mathbf{Z}_{o_i} = (\mathbf{Z}_i^{(j)}, j = 1, \ldots, |\, o_i \,|)$ denotes the uncensored components of $\mathbf{Z}_i$. We do not re-arrange the vector $\mathbf{Z}_i$ in this pattern, this representation is used merely for notational convenience. The distribution of the observed vector $\mathbf{X}_i | Y_i$ for sample i, accounting for the censoring, is

$$f(\mathbf{X}_i \mid Y_i) = \int_{\mathcal{X}_{c_{m_i}}} f(\mathbf{Z}_{o_i}, \mathbf{Z}_{m_i} \mid Y_i) d\mathbf{Z}_{m_i} = f(\mathbf{Z}_{o_i} \mid Y_i) \int_{\mathcal{X}_{c_{m_i}}} f(\mathbf{Z}_{m_i} \mid \mathbf{Z}_{o_i}, Y_i) d\mathbf{Z}_{m_i} \quad (9)$$

where the integration is only over the censored coordinates, and $\mathcal{X}_{c_{mi}}$ denotes the corresponding integration range, $\mathcal{X}_{c_{m_i}} = \Pi_{k \in m_i}(-\infty, a_k]^{I(Z_{mik} = a_k)}(b_k, \infty]^{I(Z_{mik} = b_k)}$. To estimate $\boldsymbol{\alpha}$ we need to maximize the observed likelihood

$$L_d(\mathbf{X} \mid Y = y) = \prod_y \prod_{i=1}^{n_y} \left[\frac{1}{(2\pi \mid \boldsymbol{\Delta}_y \mid)^{p/2}} \int_{\mathcal{X}_{c_{m_i}}} \exp\left(-\frac{1}{2}(\mathbf{z}_{iy} - \boldsymbol{\mu}_y)^T \boldsymbol{\Delta}_y^{-1} (\mathbf{z}_{iy} - \boldsymbol{\mu}_y) \right) d\mathbf{Z}_{m_i} \right]$$

$$= \prod_y \prod_{i=1}^{n_y} \int_{\mathcal{X}_{c_{m_i}}} L_d(\mathbf{Z}_i \mid Y_i) d\mathbf{Z}_{m_i} = \prod_y \prod_{i=1}^{n_y} \int_{\mathcal{X}_{c_{m_i}}} L_d(\mathbf{Z}_i \mid Y_i) d\mathbf{Z}_{m_i}$$

(10)

where $\boldsymbol{\mu}_y$ and $\boldsymbol{\Delta}_y$ satisfy conditions (4).

3.3 Estimation of the reduction: the cPFC and cLAD algorithms

To avoid dealing with the integral form of (9), we employ an EM algorithm similar to Lee and Scott (2012) to iteratively maximize the auxiliary function

$$Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)}) = \sum_y \sum_{i=1}^{n_y} E_{\mathbf{Z}_{m_i y} \mid \mathbf{Z}_{o_i y}, \mathcal{X}_{c_{m_i}}; \boldsymbol{\theta}^{(t)}} \ell_d(\mathbf{Z}_i \mid Y_i = y), \quad (11)$$

where $\ell_d(\mathbf{Z}_i|Y_i) = \log L_d(\mathbf{Z}_i|Y_i)$ is the joint log-likelihood of the observed and censored components of $\mathbf{Z}_i|Y_i$ under the model specified by (7) and $\boldsymbol{\theta}^{(t)} = (\boldsymbol{\alpha}^{(t)}, \boldsymbol{\mu}^{(t)}, \boldsymbol{\nu}_y^{(t)}, \mathbf{D}^{(t)}, \mathbf{H}^{(t)}, \boldsymbol{\alpha}^{(t)^T} \boldsymbol{\Delta}_y^{(t)} \boldsymbol{\alpha}^{(t)})$ denotes the estimate of $\boldsymbol{\theta}$ at iteration t of the algorithm. The EM algorithm has the following steps:

Initialization—we start with values $\boldsymbol{\theta}^{(0)} = \boldsymbol{\theta}_{\text{ncLAD}}(\boldsymbol{\alpha}^{(0)}, \boldsymbol{\mu}^{(0)}, \boldsymbol{\nu}_y^{(0)}, \mathbf{D}^{(0)}, \mathbf{H}^{(0)}, \boldsymbol{\Delta}^{(0)}, \boldsymbol{\alpha}^{(0)^T} \boldsymbol{\Delta}_y^{(0)} \boldsymbol{\alpha}^{(0)})$, obtained by maximizing (7) with the first and second moments computed under censoring using the algorithm in Lee and Scott (2012).

E-step—For fixed $\boldsymbol{\theta}^{(t)}$ the auxiliary function $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ in (11) is given by (7), with the first and second sample moments replaced by

$$\tilde{\mathbf{z}}_y = \frac{1}{n_y} \sum_{i=1}^{n_y} \mathrm{E}_{\mathbf{Z}_{m_i y} \mid \mathbf{Z}_{o_i y}, \mathcal{X}_{c_{m_i}}; \boldsymbol{\theta}^{(t)}} \{\mathbf{z}_{iy}\}, \quad (12)$$

$$\tilde{\boldsymbol{\Delta}}_y = \frac{1}{n_y} \sum_{i=1}^{n_y} \mathrm{E}_{\mathbf{Z}_{m_i y} \mid \mathbf{Z}_{o_i y}, \mathcal{X}_{c_{m_i}}; \boldsymbol{\theta}^{(t)}} \{(\mathbf{z}_{iy} - \tilde{\mathbf{z}}_y)(\mathbf{z}_{iy} - \tilde{\mathbf{z}}_y)^T\}, \quad (13)$$

These expectations are computed following the approach of Lee and Scott (2012) (see details in Appendix A in the Supplemental Material).

M-step—To maximize $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ with respect to $\boldsymbol{\theta}$, we partially maximize it sequentially with respect to $\boldsymbol{\nu}_y$, $\boldsymbol{\mu}$, $\boldsymbol{\alpha}\boldsymbol{\Delta}_y\boldsymbol{\alpha}$, $\mathbf{H}$ and $\mathbf{D}$, considering $\boldsymbol{\alpha}$ fixed and then compute $\hat{\boldsymbol{\alpha}}$. Finding $\hat{\boldsymbol{\alpha}}$ is a Grassmann manifold optimization problem. A Grassmann manifold $\mathcal{G}_{(d,p)}$ is defined as the subspace of $\mathbb{R}^p$ with dimension $d \leq p$ that corresponds to a hyperplane through the origin (Edelman et al., 1999). More specifically, to obtain $\hat{\boldsymbol{\alpha}}$ we proceed as follows.

Estimating $\hat{\boldsymbol{\alpha}}$ for cLAD: To find $\hat{\boldsymbol{\alpha}}$ for cLAD, we maximize the log likelihood function

$$-\frac{np}{2}(1 + \log 2\pi) + \frac{n}{2} \log |P_{\boldsymbol{\alpha}} \tilde{\boldsymbol{\Sigma}} P_{\boldsymbol{\alpha}}|_0 - \frac{n}{2} \log |\tilde{\boldsymbol{\Sigma}}| - \frac{1}{2} \sum_y n_y \log |P_{\boldsymbol{\alpha}} \tilde{\boldsymbol{\Delta}}_y P_{\boldsymbol{\alpha}}|_0, \quad (14)$$

over $\text{span}(\boldsymbol{\alpha}) \in \mathcal{G}_{(d,p)}$. In (14), $P_{\boldsymbol{\alpha}} = \boldsymbol{\alpha}(\boldsymbol{\alpha}^T\boldsymbol{\alpha})^{-1}\boldsymbol{\alpha}^T$ denotes the projection operator onto $\text{span}(\boldsymbol{\alpha})$, $|\mathbf{A}|_0$ the product of the non-zero eigenvalues of the positive semi-definite symmetric matrix $\mathbf{A}$, and

$$\widetilde{\boldsymbol{\Delta}} = \frac{1}{n}\sum_y n_y \widetilde{\boldsymbol{\Delta}}_y \quad \text{and} \quad \widetilde{\boldsymbol{\Sigma}} = \widetilde{\boldsymbol{\Delta}} + \widetilde{\mathbf{M}}, \quad (15)$$

where

$$\widetilde{\mathbf{M}} = \frac{1}{n}\sum_y n_y (\widetilde{\mathbf{z}}_y - \widetilde{\mathbf{z}})(\widetilde{\mathbf{z}}_y - \widetilde{\mathbf{z}})^T \quad \text{and} \quad \widetilde{\mathbf{z}} = \frac{1}{n}\sum_y n_y \widetilde{\mathbf{z}}_y . \quad (16)$$

Further details on the M-step and on estimating $\boldsymbol{\alpha}$ are provided in Appendices B and C, respectively.

Estimating $\hat{\boldsymbol{\alpha}}$ for cPFC: To find $\hat{\boldsymbol{\alpha}}$ for cPFC, i.e. when $\boldsymbol{\Delta}_Y = \boldsymbol{\Delta}$, is much simpler. The last term of (14) is replaced by $-\frac{n}{2}\log \mid P_{\boldsymbol{\alpha}}\widetilde{\boldsymbol{\Delta}}P_{\boldsymbol{\alpha}} \mid_0$. and an explicit solution is given by span($\boldsymbol{\alpha}$) = $\hat{\boldsymbol{\Delta}}^{-1/2}$ span($\boldsymbol{\gamma}$), where $\boldsymbol{\gamma}$ are the first d eigenvectors of $\hat{\boldsymbol{\Delta}}^{-1/2}\,\hat{\boldsymbol{\Sigma}}\hat{\boldsymbol{\Delta}}^{-1/2}$.

Once $\hat{\boldsymbol{\alpha}}$ is obtained, the maximum likelihood estimates for the other parameters are:

$$\begin{aligned}
\widetilde{\boldsymbol{\Delta}} &= \left[\widetilde{\boldsymbol{\Sigma}}^{-1} + \hat{\boldsymbol{\alpha}}\left(\hat{\boldsymbol{\alpha}}^T\widetilde{\boldsymbol{\Delta}}\hat{\boldsymbol{\alpha}}\right)^{-1}\hat{\boldsymbol{\alpha}}^T - \hat{\boldsymbol{\alpha}}\left(\hat{\boldsymbol{\alpha}}^T\widetilde{\boldsymbol{\Sigma}}\hat{\boldsymbol{\alpha}}\right)^{-1}\hat{\boldsymbol{\alpha}}^T\right]^{-1},\\
\widehat{\boldsymbol{\Delta}}_y &= \widehat{\boldsymbol{\Delta}} + \widehat{\boldsymbol{\Delta}}\hat{\boldsymbol{\alpha}}\left(\hat{\boldsymbol{\alpha}}^T\widetilde{\boldsymbol{\Delta}}\hat{\boldsymbol{\alpha}}\right)^{-1}\hat{\boldsymbol{\alpha}}^T(\widetilde{\boldsymbol{\Delta}}_y - \widehat{\boldsymbol{\Delta}})\hat{\boldsymbol{\alpha}}\left(\hat{\boldsymbol{\alpha}}^T\widetilde{\boldsymbol{\Delta}}\hat{\boldsymbol{\alpha}}\right)^{-1}\hat{\boldsymbol{\alpha}}^T\widehat{\boldsymbol{\Delta}},\\
\widehat{\boldsymbol{\Sigma}} &= \widehat{\boldsymbol{\Delta}} + \frac{1}{n}\sum_y n_y(\hat{\boldsymbol{\mu}}_y - \boldsymbol{\mu})(\hat{\boldsymbol{\mu}}_y - \boldsymbol{\mu})^T\\
\hat{\boldsymbol{\mu}} &= \frac{1}{n}\sum_y n_y\widetilde{\mathbf{z}}_y, \text{ and } \hat{\boldsymbol{\mu}}_y = \hat{\boldsymbol{\mu}} + \widehat{\boldsymbol{\Delta}}\hat{\boldsymbol{\alpha}}\hat{\boldsymbol{\nu}}_y .
\end{aligned}$$

After the algorithm has converged, the final estimate $\boldsymbol{\theta}^{(T)}$ yields the MLE of the model parameters and the reduced predictors are estimated by $\hat{\boldsymbol{\alpha}}^T\mathbf{X}$.

3.4 Variable selection

In addition to reducing the dimension of the markers $\mathbf{X}$, it is desirable to identify the variables associated with the outcome Y and remove irrelevant and redundant ones when computing linear combinations. This is important for etiologic research, as in the lung cancer study, to make results more interpretable and facilitate replication and translation of any findings to clinical settings. It is also important for prediction purposes, as the accuracy of a classifier may be diminished if it includes many noisy variables.

To identify the predictors that are conditionally independent of Y, we follow a proposal by Chen et al. (2010) and incorporate a group lasso type penalty (Yuan and Lin, 2006) into the log-likelihood (10)

$$L_d(\mathbf{X} \mid Y) - \lambda \sum_{j=1}^{p} \|\boldsymbol{\alpha}_j\|_2, \quad (17)$$

where $\boldsymbol{\alpha}_j$, the jth row of $\boldsymbol{\alpha}$, corresponds to all the coefficients for the jth predictor X_j, and $\|\cdot\|_2$ denotes the Euclidian norm. This penalty term exploits the non-differentiability of $\|\boldsymbol{\alpha}_j\|_2$ at $\boldsymbol{\alpha}_j = 0$, setting whole rows $\boldsymbol{\alpha}_j$ exactly to zero. Thus X_j does not contribute to the projection $\boldsymbol{\alpha}^T\mathbf{X}$ and can be discarded for modeling Y. The sparsity of the solution is determined by the tuning parameter λ. Chen et al. (2010) show that the penalty in (17) is coordinate independent and has the oracle property.

When the EM algorithm is used for estimation, the penalty term should be added to the auxiliary function Q in (11). As choosing an optimal value of λ in each EM iteration is computationally challenging, we first estimate the non-regularized parameters for any of the likelihood-based methods, and then select variables in an additional step. For the PFC based methods this is done following the approach in Chen et al. (2010). For the LAD based methods however, optimization of (17) is computationally difficult.We thus find a regularized estimator based on a simpler convex approximation to $L_d(\mathbf{X}|Y)$ based on a trace operator. Let $\mathbf{S} = \log(\hat{\boldsymbol{\Delta}}^{-1/2}\hat{\boldsymbol{\Sigma}}\hat{\boldsymbol{\Delta}}^{-1/2}) - n^{-1}\sum_y n_y \log(\hat{\boldsymbol{\Delta}}^{-1/2}\hat{\boldsymbol{\Delta}}_y\hat{\boldsymbol{\Delta}}^{-1/2})$, with $\hat{\boldsymbol{\Sigma}}$ and $\hat{\boldsymbol{\Delta}}_y$ estimated using the non-penalized algorithm for the model of choice (e.g. LAD, mcLAD, or cLAD). It can be shown (Cook et al., 2014) that under model (4) the maximizer $\boldsymbol{\gamma}_*$ of the population version of the objective function $J_d(\mathbf{X}|Y) = \mathrm{tr}(\boldsymbol{\gamma}^T\mathbf{S}\boldsymbol{\gamma})$ is the same as the maximizer $\boldsymbol{\alpha}_*$ of the population version of (10), in the sense that $\mathrm{span}(\boldsymbol{\alpha}_*) \equiv \boldsymbol{\Delta}^{-1/2}\mathrm{span}(\boldsymbol{\gamma}_*)$, and $L_d(\mathbf{X}|Y)_{|\boldsymbol{\alpha}_*} = J_d(\mathbf{X}|Y)_{|\boldsymbol{\gamma}_*}$, where $\boldsymbol{\Delta} = E(\boldsymbol{\Delta}_y)$. The solution $\boldsymbol{\gamma}_*$ that maximizes $\mathrm{tr}(\boldsymbol{\gamma}^T\mathbf{S}\boldsymbol{\gamma})$ is easily computed as it is given by the d leading eigenvectors of $\mathbf{S}$. Though the equivalence between (14) and the trace approximation holds in the population, in our experience in any finite sample the trace-based estimates $\hat{\boldsymbol{\gamma}}$ are indeed very close to $\hat{\boldsymbol{\alpha}}$. Thus, instead of (17) we maximize

$$J_d(\mathbf{X} \mid Y) - \lambda \sum_{j=1}^{p} \|(\widehat{\boldsymbol{\Delta}}^{-1/2}\boldsymbol{\gamma})_j\|_2 = J_d(\mathbf{X} \mid Y) - \rho(\mathbf{V}), \quad (18)$$

with $\mathbf{V} = \hat{\boldsymbol{\Delta}}^{-1/2}\boldsymbol{\gamma}$.

The solution to (18) is computed using an iterative procedure based on quadratic local approximations of $\rho(\mathbf{V})$ described in detail in Chen et al. (2010). In each iteration irrelevant variables are removed, until the subspace spanned by $\boldsymbol{\gamma}^{(t)}$ does not change from one iteration to the next. In the initial step, $\boldsymbol{\gamma}^{(0)}$, the non-penalized solution for (14) serves as a starting value and MLE of the covariance matrices estimated by the chosen reduction method (e.g. LAD, cmLAD, cLAD) are used to compute $\mathbf{S}$. In the t-th iteration we use that the matrix derivative of $\partial\rho/\partial\mathbf{V}$ is approximated around $\tilde{\mathbf{V}}^{(t)} = \hat{\boldsymbol{\Delta}}^{-1/2}\boldsymbol{\gamma}^{(t)}$ by $\frac{\partial\rho}{\partial\mathbf{V}} \approx \mathrm{diag}(1/\|\tilde{\mathbf{v}}_1^{(t)}\|, \ldots, 1/\|\tilde{\mathbf{v}}_p^{(t)}\|)\mathbf{V} := \mathbf{G}^{(t)}\mathbf{V}$, where $\mathbf{v}_i$ denotes the i-th row vector of $\mathbf{V}$, and

thus after a second order Taylor expansion, $\rho(\mathbf{V}) \approx \frac{1}{2}\mathrm{tr}\{\mathbf{V}^T\mathbf{G}^{(t)}\mathbf{V}\} + C_0$, where the constant C_0 does not depend on $\mathbf{V}$. Finding the minimizer of (18) in the k-th step of the iteration can thus be done by finding

$$\hat{\boldsymbol{\gamma}}^{(t+1)} = \arg\min \left\{ \boldsymbol{\gamma}^T \left(\frac{1}{2}\widehat{\boldsymbol{\Delta}}^{-1/2}\mathbf{G}^{(t)}\widehat{\boldsymbol{\Delta}}^{-1/2} - \widehat{\mathbf{S}} \right)\boldsymbol{\gamma} \right\}. \quad (19)$$

The iteration is stopped when the angle between the subspace spanned by $\boldsymbol{\gamma}^{(t)}$ and $\boldsymbol{\gamma}^{(t-1)}$ is less than a given tolerance level ε. The regularization parameter λ in (18) is chosen using a BIC type criterion, following (Chen et al., 2010).

Algorithm 1 (Supplemental Material) summarizes the procedure to select the subset of variables that are truly associated with the outcome Y.

4. Data Example: Analysis of Inflammation Markers

We illustrate the methods using data from a lung cancer study conducted to independently replicate findings from Shiels et al. (2013), and to identify further associations of serum inflammation markers with lung cancer risk (Shiels et al., 2015). In addition to providing biologic insights into lung carcinogenesis, assessing the utility of an ”inflammation score” based on marker combinations for risk stratification is of interest.

Study subjects were 526 lung cancer cases diagnosed in the Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial and 625 matched controls (Shiels et al., 2015). After excluding 11 markers that performed poorly in Shiels et al. (2013), levels of 51 inflammation markers were measured in serum (Supplemental Table 1). We excluded 4 markers with poor performance characteristics, leaving 47 markers with 0–75% censoring for analysis.

We first estimated associations of the markers with binary case-control status, $Y = 0$ and $Y = 1$. Information on all 47 markers was available for 509 cases and 606 controls. We also assessed marker associations with smoking status in three categories (never, former, current; $Y = 0, 1, 2$) in controls, and we used data on 146 never, 247 former and 213 current smokers.

For each outcome, we estimated the sufficient reduction using cLAD and cPFC, by applying LAD and PFC directly to the data ignoring the censoring, by cmLAD and cmPCF, with moments computed under censoring and based on imputed data (MI-LAD and MI-PFC) (see Section 3.2) with the intrinsic dimension d inferred using AIC.

4.1 Results for marker selection

For the binary outcome (lung cancer status), the estimated dimension of the sufficient reduction was $d = 6$ for all LAD based methods, and $d = 1$ for all PFC based approaches. Note, however, that for binary Y PFC methods can identify at most a single dimension. Table 1 shows the markers selected by each method. cLAD selected ten markers, LAD and MI-LAD seven, and cmLAD selected eight markers. Among them, five markers were

identified by all LAD based methods (Table 1). On the other hand, cPFC and cmPFC selected two markers, while PFC and MI-PFC selected the same two and one additional marker. The two markers selected by cPCF were also identified by cLAD. Several markers, including *CXCL9MIG*, *CRP* and *IL1RA*, that were associated with lung cancer risk in Shiels et al. (2013) were also identified in this replication study by some or all of our methods. All methods selected *CXCL9MIG*. With the exception of LAD and MI-LAD all methods identified *CRP*. However, only cLAD and cmLAD identified *IL1RA* as an associated marker.

For the smoking outcome, we estimated $\hat{d} = 8$ for cLAD and cmLAD, and $d = 9$ for LAD and MI-LAD, while all PFC based methods estimated $\hat{d} = 2$. cLAD identified ten markers, LAD thirteen, cmLAD eleven and MI-LAD twelve. Seven markers were selected by all LAD based methods. cPFC, cmPFC and MI-PFC identified eight markers, and PFC seven and the methods had six selected markers in common. Of the ten markers selected by cLAD, six were also selected by cPCF. All methods identified *IL7*, *IP10*, *SAP*, *TARC* and *TNFβ*.

As *CRP* and *CXCL9MIG* were only associated with lung cancer risk but not smoking, it strengthes the evidence for their etiologic role in lung cancer development rather than through inflammation caused by smoking.

4.2 Predictive performance

To assess the utility of an ”inflammation score” for lung cancer risk prediction, we studied the performance of a prediction model based on the d linear predictors $\hat{\boldsymbol{\alpha}}^T\mathbf{X}$, where $\hat{\boldsymbol{\alpha}}$ was estimated by the various methods. The prediction rules $\psi(\hat{\boldsymbol{\alpha}}^T\mathbf{X})$ we used were quadratic discriminant analysis (QDA) and polytomous or binary logistic regression where the d linear combinations were included as main effects. The predictive performance of the proposed methods was also compared to that of SAVE, a standard second moment-based IR method (Cook and Weisberg, 1991). SAVE estimates were computed both using the measured predictors (SAVE) and using moments computed under censoring (cmSAVE). For the LAD based methods and SAVE we used fixed $d = 1$, 2 and $d = 5$ and for the PFC based methods we let $d = 1$ or $d = 1$, 2 (for the smoking outcome), and the optimal d selected by AIC. We also assessed the performance of a classifier that included all $\mathbf{X}$ directly in QDA or logistic models and after multiple imputation using moment estimates computed for censored data as in Lee and Scott (2012).

To quantify predictive performance, we estimated the prediction error $E_P[I\{Y \neq \psi(\hat{\boldsymbol{\alpha}}^T\mathbf{X})\}]$, based on twenty-fold cross validation. For the binary outcome we also computed the AUC, the area under the receiver operating characteristic (ROC) curve, that can be expressed as the probability that the scalar predictive score for a randomly selected case exceeds that for a randomly selected control (see Pepe (2004), page 67). Of two prediction models the one with the larger AUC has ”better” predictive performance. The predicted probability $\hat{P}(Y = 1|\mathbf{X})$ obtained with each prediction rule was used as a score in the AUC computation.

For binary Y, the prediction error for cLAD was 0.25 for QDA with the optimal $\hat{d} = 6$, and 0.35 for logistic regression with the $\hat{d} = 6$; for categorical Y it was 0.26 for QDA with $\hat{d} = 6$ and 0.31 for logistic regression. cmLAD had only slightly worse predictive performance

than cLAD. The prediction errors for LAD with the optimal d and cPCF were similar for QDA for binary Y and much higher than other LAD based methods (0.41 for LAD and 0.40 for cPFC). For logistic regression the differences between methods were somewhat less pronounced, with prediction errors for LAD based methods ranging from 0.35 for cLAD to 0.40 for LAD and from 039 for cPFC to 0..43 for cmPFC (Table 2). Using all the predictors directly in QDA resulted in higher prediction errors than using the predictors after imputation, while logistic regression yielded similar predictive performance. This difference was most pronounced for binary Y with QDA as the classifier, where the prediction error using the measured $\mathbf{X}$ was 0.60, while it was 0.47 after imputation of censored values. Similar patterns were observed for the categorical smoking outcome, however, the difference in predictive performance between the LAD or PFC based methods was less pronounced than for binary Y. Again, cLAD with estimated dimension and QDA as the classifier had the lowest prediction error (0.26) of all methods. All likelihood based dimension reduction methods led to better predictive performance than SAVE based methods or using the predictors.

cLAD had the highest AUC= 0.676, i.e. the best discriminatory performance, for $\hat{d} = 6$, when the $\hat{\boldsymbol{\alpha}}^T\mathbf{X}$ predictors were used in a logistic regression model followed by cPFC with an AUC=0.651 (Table 6). When QDA was used as the classifier, cPFC had the highest AUC, 0.644, and cLAD with $\hat{d} = 6$ resulted in the second highest value, $AUC = 0.641$. While for all PFC based methods the AUC was between 0.60 and 0.65 for $d = 1$, LAD based methods had noticeably improved discriminatory performance when more dimensions were used, e.g. for cLAD with QDA and $d = 2$ AUC=0.591, while it was 0.631 for the optimal $\hat{d} = 6$. Both SAVE, applied to the censored predictors directly or cmSAVE had worse performance than using the censored predictors directly in QDA or logistic regression or after imputing $\mathbf{X}$. Overall, using $\hat{\boldsymbol{\alpha}}\mathbf{X}$ in logistic models produced higher AUC values than using QDA as the classifier. (Table 3)

5. Simulation study

To study the impact of the number of predictors p, the dimension of the central subspace d, the amount of censoring (%C) and the sample size, n_y, in each group defined by Y on the performance of the methods, we conducted simulations based on censored data generated from normal populations for binary and categorical Y. The mean and covariance parameters depend on Y through (4) and were chosen to yield correlations between predictors ranging from 0.2 to 0.35, similar to the real data example. We also assessed robustness of the methods to violations of normality.

We let $p = 10$, 20 and 30, $d = 1$ and $d = 3$, and $n_y = 100$ and $n_y = 500$. For each setting we assumed 10% and 30% censoring for all markers. All results in the tables are means over 200 repetitions based on the same setting. For all calculations d was assumed to be known.

Similar to the data example, we compared the performance of cLAD and cPFC to that of applying standard LAD, PFC and SAVE naively to the censored measurements, and that of cmLAD, cmPFC, and cmSAVE that use moments computed accounting for the censoring.

5.1 Accuracy of estimates of α and the moments

First we assessed how well $\boldsymbol{\alpha}$ and the moments in each group defined by Y are estimated. As the distance between the true and estimated spaces span($\boldsymbol{\alpha}$) and span($\hat{\boldsymbol{\alpha}}$) (e.g. Li and Wang, 2007) we used $\|P_{\hat{\boldsymbol{\alpha}}} - P_{\boldsymbol{\alpha}}\|_F$, where $\|.\|$ is the Frobenius norm and $P_{\boldsymbol{\alpha}}$ is the orthogonal projection onto span($\boldsymbol{\alpha}$). The differences in the estimated moments from the truth are measured by the Euclidean and Frobenius norms, respectively as $\|\hat{\boldsymbol{\mu}}_Y - \boldsymbol{\mu}_Y\|/\|\boldsymbol{\mu}_Y\|$, $\|\hat{\boldsymbol{\Delta}}_Y - \boldsymbol{\Delta}_Y\|_F/\|\boldsymbol{\Delta}_Y\|_F$.

Table 4 shows results for $n_y = 100$ with $p = 20$, and for $n_y = 500$ with $p = 30$ markers with true dimension $d = 3$ when $\boldsymbol{\Delta}_0 \neq \boldsymbol{\Delta}_1$. cLAD was substantially better in estimating $\boldsymbol{\alpha}$ and all moments than all other methods, including cPFC. The improvement was 35% compared to cPFC even for $n_y = 100$ with $p = 20$. The difference in performance between cLAD and cPFC was more pronounced for larger p and larger values of n/p. The percent of censoring had little impact on these findings, with only slightly better results obtained for 10% censoring for all methods than for 30% censoring. The improvement in estimating $\boldsymbol{\alpha}$ using cLAD compared to LAD was 40% or greater for all settings, while the improvement in estimating $\boldsymbol{\alpha}$ using cPFC instead of PFC was slightly lower at 35 – 40%. The improvement in estimation based on cLAD compared to cmLAD was around 20%. Applying SAVE to the censored predictors yielded worse results than applying any other method, while cmSAVE resulted in better performance than LAD and PFC directly applied to **X**. As findings were qualitatively similar when $d = 1$, these results are presented in Supplemental Tables.

When $\boldsymbol{\Delta}_Y = \boldsymbol{\Delta}$, as expected, cPFC performed somewhat better in estimating $\boldsymbol{\alpha}$ and the moments of the distributions than cLAD. Otherwise the performance of the methods was similar to those in Table 6, and are presented in Supplemental Tables.

5.2 Predictor selection

Here we first generated data for binary Y from multivariate normal models with $\boldsymbol{\Delta}_0 \neq \boldsymbol{\Delta}_1$. We then also assessed the impact of violations to normality by generating $\mathbf{Z}|Y$ from a mixture, $0.85N(\boldsymbol{\mu}_{ya}, \boldsymbol{\Delta}_{ya}) + 0.15N(\boldsymbol{\mu}_{yb}, \boldsymbol{\Delta}_{yb})$. The moments $(\boldsymbol{\mu}_{ya}, \boldsymbol{\Delta}_{ya})$ were the same as for the normal model, and $\boldsymbol{\mu}_{yb} = \boldsymbol{\mu}_{ya} + \tau\mathbf{1}$, to perturb the right tails of the corresponding marginals. The covariance matrices $\boldsymbol{\Delta}_{yb}$ were chosen to have the same correlations as $\boldsymbol{\Delta}_{ya}$, but with smaller variances. Only left censoring was used in this scenario.

To study the ability of regularized versions of cLAD, LAD, cmLAD, cPFC, PFC and cmPFC to identify predictors truly associated with the outcome we computed the average number of true positives (TPs) as

$$TP = \frac{1}{I}\sum_{i=1}^{I}\frac{\sum_{k=1}^{p} I(X_{ik}\ \text{correctly selected})}{\sum_{k=1}^{p} I(X_{ik}\ \text{selected})},$$

where $I = 500$ was the number of repetitions using the same setting.

cLAD and cmLAD followed by cPFC had the highest TP rate among all methods (Table 5). Even with 30% censoring, TP was greater than 84%, 78%, 77% and 75% for cLAD, cmLAD, cPFC and cmPFC, respectively. LAD and had a somewhat higher TP than PFC, but both were lower than 75% for all settings we studied. For all methods the TP rate decreased as censoring increased. For cLAD the TP was reduced by 5% when 30% of the data were censored compared to 10% for all choices of n_y, p and d.

While the TP rates were slightly lower when $\mathbf{Z}$ was generated from a mixture of normal distributions, the overall pattern of performance of the methods was similar to the normal settings. However, censoring had a stronger impact, e.g. for $n_y = 500$, $p = 30$ and $d = 1$ with 10% censoring the TP rate for cLAD was 0.808, while it was 0.781 and 0.759 for 20% and 30% censoring, respectively. The reduction in performance as the percent of censoring increased was less pronounced for cLAD and cmLAD than for LAD.

5.3 Predictive performance

We assessed the predictive performance of the proposed methods for binary and categorical Y for $n_y = 500$, $p = 30$, $d = 2$ and 30% censoring. The basis matrix $\boldsymbol{\alpha}$ was obtained by letting $\tilde{\boldsymbol{\alpha}} = (\tilde{\boldsymbol{\alpha}}_1^T \mathbf{0}_{d\times 10})^T$, with the entries in $\tilde{\boldsymbol{\alpha}}_1$ randomly sampled from {0, 1} and then applying orthonormalization. We used cLAD, LAD, cmLAD, cPFC, PFC and cmPFC with regularization for variable selection, and QDA and polytomous or binary logistic regression as prediction rules, with the d linear combinations included as main effects. The prediction error was estimated using five-fold cross-validation. We also computed the prediction error for an "oracle" procedure, using the true $\boldsymbol{\alpha}$ and uncensored predictors $\mathbf{Z}$.

First we generated data from normal distributions with means and covariances given in (4), with a different covariance matrix for each outcome group. For the binary outcome Y, the prediction error for the oracle procedure was 0.091 for both QDA and logistic regression (Table 6). Using cLAD resulted in only slightly higher prediction errors, 0.121 for QDA and 0.097 for logistic regression, followed by cmLAD and cPFC. LAD and PFC had prediction errors around 20% for both QDA and logistic regression for binary outcomes. The performance of all methods was slightly worse for categorical outcomes for all methods. However, the improvement in prediction based on cLAD was more pronounced than for binary Y. E.g. cLAD with QDA had a 56% and 34% better prediction error compared to LAD and cmLAD with QDA, respectively. For categorical Y, using polytomous logistic regression models resulted in a slightly greater improvement in prediction over QDA than seen for binary Y. The performance of methods in estimation of $\hat{\boldsymbol{\alpha}}$ corresponded closely to the prediction performance, with cLAD providing the best estimation of the subspace, followed by cmLAD, cPFC, cmPFC then LAD and PFC.

We also assessed the performance of the methods when $\mathbf{Z}|Y$ were generated from a mixture $0.85\mathcal{N}(\boldsymbol{\mu}_{ya}, \boldsymbol{\Delta}_{ya}) + 0.15\mathcal{N}(\boldsymbol{\mu}_{yb}, \boldsymbol{\Delta}_{yb})$, with parameters set as in Section 5.2 and only left censoring. The "oracle" results were obtained by using the true $\boldsymbol{\alpha}$ and the uncensored predictors $\mathbf{Z}$ for classification, but assuming that $\mathbf{Z}|Y$ arose from a single normal density.

In this setting all methods had higher prediction errors than for normal data, including the oracle, however, the relative performance of the methods was similar to the normal case for the prediction error and also for the estimation of $\boldsymbol{\alpha}$ (Table 6). cLAD had a 30% higher rate than the oracle, followed by cmLAD, cPFC, cmPFC and LAD and PFC.

6. Discussion

In molecular epidemiology, measurements of biomarkers often fall outside an assay's lower or upper limits of detection (LODs) due to technological limitations of the measurement process, leading to censored observations. Numerous approaches are available to model single censored measurements (see, e.g. Dinse et al., 2014). When the number of markers is small, multiple imputation has been proposed (Lee et al., 2012), but in addition to the computational burden, procedures for variable selection for imputed data are limited.

Here we extended likelihood-based sufficient dimension reduction methods, particularly Principal Fitted Components (PFC, Cook and Forzani, 2008) and Likelihood Acquired Directions (LAD, Cook and Forzani, 2009), to regression or classification with censored predictors. We compared the performance of the full likelihood approaches (cLAD, cPFC), to applying LAD/PFC directly to the censored data without further accounting for the censoring, to cmLAD/cmPFC, that use moments estimated under censoring and to MI-LAD/MI-PFC that applied LAD/PFC after imputing censored values. While all methods provide consistent estimates of the sufficient reduction, as moment conditions are satisfied even for the censored data, only cLAD and cPFC are fully efficient. We also extended the coordinate-independent sparse estimation (CISE) algorithm proposed by Chen et al. (2010) to cLAD to combine dimension reduction with variable selection. Our method thus allows to parsimoniously describe the data structure and discover scientifically interesting features.

When we analyzed inflammation markers in relation to risk of lung cancer and smoking status, all LAD-based approaches resulted in fairly high estimates of dimension d, which was six or more for case-control status, and 8 or higher for the smoking outcome. This indicates, not surprisingly, that the relationship between the markers and outcomes is complex, and simple modeling may result in a loss of power to detect associations and for prediction. For binary case-control status, the AUC for cLAD was approximately 7% higher than the AUC for a score that was computed using all markers after multiple imputations and 6% higher than the AUC based on a score from LAD applied directly to the censored predictors, which is a substantial improvement in discriminatory performance. This is an important finding, as multiple imputation is a popular approach for dealing with missing data.

In our simulations the improvement in estimating the central subspace and the moments of the distributions using cLAD was substantial in all settings that had different within-group covariance matrices. The small gain in using cPFC over cLAD when the within-group covariance matrices were the same does not warrant the possible trade off in robustness.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank M. Shiels and A. Chaturvedi for access to the data and helpful comments.

References

Adragni KP, Cook RD. 2009; Sufficient dimension reduction and prediction in regression. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences. 367:43–85.

Chen Q, Wang S. 2013; Variable selection for multiply-imputed data with application to dioxin exposure study. Statistics in Medicine. 32:3646–3659. [PubMed: 23526243]

Chen X, Zou C, Cook RD. 2010; Coordinate-independent sparse sufficient dimension reduction and variable selection. The Annals of Statistics. 38:3696–3723.

Cook, R. Regression Graphics. Wiley; New York: 1998.

Cook R. 2007; Fisher lecture: Dimension reduction in regression (with discussion). Statistical Science. 22:1–26.

Cook R, Forzani L. 2008; Principal fitted components in regression. Statistical Science. 23:485–501.

Cook R, Forzani L. 2009; Likelihood-based sufficient dimension reduction. Journal of the American Statistical Association. 104:197–208.

Cook R, Forzani L, Tomassi D. 2014A note on trace-based approximations for sufficient dimension reduction.

Cook R, Weisberg S. 1991; Discussion of sliced inverse regression. Journal of the American Statistical Association. 86:328–332.

Dinse G, Jusko T, Ho L, Annam K, Graubard B, Hertz-Picciotto I, Miller F, Gillespie B, Weinberg C. 2014; Accommodating measurements below a limit of detection: A novel application of cox regression. American Journal of Epidemiology. 179:1018–1024. [PubMed: 24671072]

Dong T, Liu CC, Petricoin EF, Tang LL. 2014; Combining markers with and without the limit of detection. Statistics in Medicine. 33:1307–1320. [PubMed: 24132938]

Edelman A, Arias T, Smith S. 1999; The geometry of algorithms with orthogonality constraints. SIAM Journal of Matrix Analysis and Applications. 20:303–353.

Fukumizu K, Bach FR, Jordan MI. 2004; Dimensionality reduction for supervised learning with reproducing kernel hilbert spaces. Journal of Machine Learning Research. 5:73–99.

Lee G, Scott C. 2012; Em algorithms for multivariate gaussian mixture models with truncated and censored data. Computational Statistics and Data Analysis. 56:2816–2829.

Lee M, Kong L, Weissfeld L. 2012; Multiple imputation for left-censored biomarker data based on gibbs sampling method. Statistics in Medicine. 31:1838–1848. [PubMed: 22359320]

Li B, Wang S. 2007; On directional regression for dimension reduction. Journal of the American Statistical Association. 102:997–1008.

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

Mantovani A, Allavena P, Sica A, Balkwill F. 2008; Cancer-related inflammation. Nature. 454:436–444. [PubMed: 18650914]

Pepe, MS. The statistical evaluation of medical tests for classification and prediction. 2004.

Rubin, DB. Multiple Imputation for Nonresponse in Surveys (Wiley Series in Probability and Statistics). Wiley; 1987.

Shiels M, Katki H, Hildesheim A, Pfeiffer R, Engels E, Williams M, Kemp T, Caporaso N, Pinto L, Chaturvedi A. 2015Circulating inflammation markers, risk of lung cancer and utility for risk stratification. Journal of the National Cancer Institute.

Shiels M, Pfeiffer R, Hildesheim A, Engels E, Kemp T, Park JH, Katki H, Koshiol J, Shelton G, Caporaso N, Pinto L, Chaturvedi A. 2013; Circulating inflammation markers and prospective risk for lung cancer. Journal of the National Cancer Institute. 105:1871–1880. [PubMed: 24249745]

van Buuren S, Groothuis-Oudshoorn K. 2011; mice: Multivariate imputation by chained equations in R. Journal of Statistical Software. 45:1–67.

Yuan M, Lin Y. 2006; Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society: Series B. 68:49–67.

Table 1

Markers selected by at least one method using penalized likelihood. The regularization parameter was set using AIC

Marker	case-control status (binary *Y*)							
	cLAD	**LAD**	**cmLAD**	**MI-LAD**	**cPFC**	**PFC**	**cmPFC**	**MI-PFC**
*X*6*CKINE*	●	●	●	●				
*BCA*1								
CRP	●		●		●	●	●	●
CTACK	●	●	●	●				
*CXCL*11*ITAC*								
*CXCL*9*MIG*	●	●	●	●	●	●	●	●
EGF								
*FGF*2	●		●					
GRO		●		●				
*IL*1*RA*	●		●					
*IL*7								
*IP*10								
MDC								
*MIP*1*B*		●		●				
*MIP*1*D*								
SAP								
SEGFR	●	●	●	●				
SILRII								
*SIL*6*R*								
*SV EGFR*2	●	●	●	●				
TARC	●					●		●
TNFB	●							

Marker	smoking status (categorical *Y*)							
	cLAD	**LAD**	**cmLAD**	**MI-LAD**	**cPFC**	**PFC**	**cmPFC**	**MI-PFC**
*X*6*CKINE*								
*BCA*1		●	●	●				
CRP								

Marker	smoking status (categorical Y)							
	cLAD	LAD	cmLAD	MI-LAD	cPFC	PFC	cmPFC	MI-PFC
CTACK	●	●	●	●				
CXCL11ITAC	●		●	●				
CXCL9MIG		●		●				
EGF		●			●	●	●	●
FGF2	●	●	●	●	●		●	●
GRO								
IL1RA	●		●	●				
IL7	●	●	●	●	●	●	●	●
*IP*10	●	●	●	●	●	●	●	●
MDC		●		●	●		●	●
MIP1B								
MIP1D		●						
SAP	●	●	●	●	●	●	●	●
SEGFR	●							
SILRII						●		
SIL6R		●						
SV EGFR2								
TARC	●	●	●	●	●	●	●	●
TNFB	●	●	●	●	●	●	●	●

Table 2

Prediction error for cLAD, standard LAD, cPFC and standard PFC based on 20-fold cross validation for the lung cancer data for case control status (binary Y), and for the categorical smoking outcome Y with H = 3 categories

outcome *Y*		**prediction error based on**									
		QDA					**logistic regression**				
		d					d				
		1	2	5	AIC best	*p*	1	2	5	AIC best	*p*
Binary	cLAD	.42	.39	.26	.25 (d=6)		.42	.39	.36	.35 (d=6)	
	LAD	.48	.44	.42	.41 (d=6)		.48	.48	.40	.40 (d=6)	
	cmLAD	.43	41	.30	.29 (d=6)		.43	43	.39	.36 (d=6)	
	MI-LAD	.45	.42	.40	.38 (d=6)		.43	.42	39	.37 (d=6)	
	cPFC	.40					.39				
	PFC	.46					.43				
	cmPFC	.43					.44				
	MI-PFC	.44					.43				
	SAVE	.49	.48	.47			.48	.48	.47		
	cmSAVE	.48	.47	.45			.47	.46	.45		
	X					.59					.46
	MI-**X**					.47					.45
Categorical	cLAD	.55	.43	.30	.26 (d=8)		.55	.39	.31	.31(d=8)	
	LAD	.60	.50	.36	.34 (d=9)		.59	.45	.36	.36 (d=9)	
	cmLAD	.58	.48	.34	.30 (d=8)		.56	.42	.33	.32 (d=8)	
	MI-LAD	.58	.48	.35	.30 (d=9)		.56	.43	.33	.32 (d=9)	
	cPFC	.42	.38				.39	.34			
	PFC	.48	.40				.44	.37			
	cmPFC	.46	.39				.44	.35			
	MI-PFC	.46	.41				.44	.35			
	SAVE	.64	60	.58			.61	.56	.54		
	cmSAVE	.62	.58	.55			.59	.58	.52		
	X					.54					.53

outcome *Y*	prediction error based on			
	QDA		logistic regression	
	MI-X			
	.50		.48	

Table 3

AUC for various algorithms based on 5-fold cross validation for the lung cancer data for case control status (binary Y)

	***AUC* based on**									
	QDA					**logistic regression**				
	d					d				
	1	2	5	AIC best	*p*	1	2	5	AIC best	*p*
cLAD	.582	.591	.633	.641 (d=6)		.619	.652	.671	.676 (d=6)	
LAD	.531	.540	.578	.613 (d=6)		.552	.574	.609	.619 (d=6)	
cmLAD	.564	.576	.603	.627 (d=6)		.588	.613	.636	.647 (d=6)	
MI-LAD	.561	.578	.591	.622 (d=6)		.595	.619	.633	.639 (d=6)	
cPFC	.644					.649				
PFC	.603					.615				
cmPFC	.618					.627				
MI-PFC	.611					.621				
SAVE	.519	.522	.543	.		.512	.519	.540		
cmSAVE	.528	.533	.560			.531	.537	.568		
X					.568					.587
MI-**X**					.586					.605

Table 4

Evaluations of the parameter estimation of cLAD, LAD, cmLAD, cPFC, PFC, and cmPFC for $\boldsymbol{\Delta}_0 \neq \boldsymbol{\Delta}_1$. All numbers are means over 200 repetitions for each value of n_y, the sample size in each group defined by Y, C the % of censoring, p, the number of predictors and d, the true dimension of the model.

n_y, p, d	method	%C	$\frac{\|P_{\hat{\alpha}} - P_{\alpha}\|}{\|P_{\alpha}\|}$	$\frac{\|\hat{\boldsymbol{\Delta}}_0 - \boldsymbol{\Delta}_0\|}{\|\boldsymbol{\Delta}_0\|}$	$\frac{\|\hat{\boldsymbol{\Delta}}_1 - \boldsymbol{\Delta}_1\|}{\|\boldsymbol{\Delta}_1\|}$	$\frac{\|\hat{\mu}_0 - \mu_0\|}{\|\mu_0\|}$	$\frac{\|\hat{\mu}_1 - \mu_1\|}{\|\mu_1\|}$
100,20,3	cLAD	10	0.411	0.313	0.321	0.108	0.093
	LAD		0.913	0.552	0.592	0.193	0.188
	cmLAD		0.534	0.391	0.403	0.142	0.137
	cPFC		0.594	0.420	0.440	0.161	0.149
	PFC		0.962	0.579	0.589	0.224	0.219
	cmPFC		0.816	0.497	0.510	0.190	0.184
	SAVE		0.982				
	cmSAVE		0.898				
500,30,3	cLAD	10	0.331	0.312	0.320	0.156	0.159
	LAD		0.846	0.555	0.572	0.263	0.275
	cmLAD		0.412	0.369	0.379	0.189	0.198
	cPFC		0.501	0.427	0.442	0.201	0.214
	PFC		0.913	0.613	0.651	0.291	0.310
	cmPFC		0.622	0.479	0.500	0.247	0.256
	SAVE		0.946				
	cmSAVE		0.798				
100,20,3	cLAD	30	0.429	0.333	0.341	0.123	0.106
	LAD		0.966	0.601	0.632	0.267	0.243
	cmLAD		0.581	0.429	0.453	0.170	0.158
	cPFC		0.649	0.442	0.461	0.178	0.167
	PFC		0.993	0.653	0.669	0.311	0.306
	cmPFC		0.852	0.538	0.552	0.236	0.204
	SAVE		0.998				
	cmSAVE		0.932				

n_y, p, d	method	%C	$\frac{\|P_{\widehat{\alpha}} - P_{\alpha}\|}{\|P_{\alpha}\|}$	$\frac{\|\widehat{\boldsymbol{\Delta}}_0 - \boldsymbol{\Delta}_0\|}{\|\boldsymbol{\Delta}_0\|}$	$\frac{\|\widehat{\boldsymbol{\Delta}}_1 - \boldsymbol{\Delta}_1\|}{\|\boldsymbol{\Delta}_1\|}$	$\frac{\|\widehat{\mu}_0 - \mu_0\|}{\|\mu_0\|}$	$\frac{\|\widehat{\mu}_1 - \mu_1\|}{\|\mu_1\|}$
500,30,3	cLAD	30	0.360	0.3389	0.347	0.181	0.189
	LAD		0.922	0.613	0.631	0.314	0.338
	cmLAD		0.465	0.402	0.419	0.229	0.235
	cPFC		0.550	0.462	0.479	0.240	0.251
	PFC		0.987	0.673	0.701	0.375	0.400
	cmPFC		0.663	0.499	0.525	0.293	0.302
	SAVE		0.996				
	cmSAVE		0.854				

Table 5

True positive (TP) rate, i.e. proportion of correctly selected predictors, for $\boldsymbol{\Delta}_0 \neq \boldsymbol{\Delta}_1$. All numbers are means over 200 repetitions for each value of n_y, the sample size in each group defined by Y, C the % of censoring, p, the number of predictors and d, the true dimension of the model.

n, p, d	***%C***	**cLAD**	**LAD**	**cmLAD**	**cPFC**	**PFC**	**cmPFC**
Normal data							
100,20,1	10	.884	.752	.878	.828	.714	.798
2		.868	.738	.852	.814	.712	.786
500,30,1		.854	.718	.826	.836	.798	.822
2		.862	.728	.824	.828	.720	.818
100,20,1	20	.868	.672	.846	.798	.658	.778
2		.848	.704	.824	.802	.660	.772
500,30,1		.842	.686	.810	.826	.762	.776
2		.850	.736	.816	.838	.732	.782
100,20,1	30	.832	.636	.804	.768	.588	.752
2		.816	.618	.798	.752	.580	.736
500,30,1		.818	.616	.804	.798	.654	.766
2		.822	.660	.776	.804	.668	.760
Non-normal data (mixture model)							
n, p, d	*%C*	cLAD	LAD	cmLAD	cPFC	PFC	cmPFC
100,20,1	10	.845	.711	.829	.806	.709	.780
2		.819	.674	.799	.803	.675	.777
500,30,1		.808	.689	.753	.732	.644	.713
2		.811	.677	.759	.718	.631	.712
100,20,1	20	.814	.665	.792	.771	.640	.755
2		.778	.669	.761	.748	.658	.736
500,30,1		.781	.637	.736	.706	.621	.695
2		.786	.619	.718	.704	.609	.683
100,20,1	30	.762	.637	.741	.732	.615	.701
2		.753	.634	.739	.706	.621	.687

n, *p*, *d*	%C	cLAD	LAD	cmLAD	cPFC	PFC	cmPFC
500,30,1		.758	.617	.716	.699	.603	.641
	2	.743	.598	.694	.688	.575	.678

Table 6

Prediction error for cLAD, LAD, cmLAD, cPFC, PFC and cmPFC based on five-fold cross validation

outcome Y		**prediction error based on**		
		QDA	**logistic regression**	$\|\|P_a - \hat{P}_a\|\|/\|\|P_a\|\|$
Normal data				
binary	cLAD	.121	.097	.458
	LAD	.194	.173	.816
	cmLAD	.138	.132	.569
	cPFC	.153	.152	.653
	PFC	.218	.211	.937
	cmPFC	.171	.169	.711
	oracle	.091	.091	
categorical	cLAD	.126	.112	.394
	LAD	.287	.258	.843
	cmLAD	.184	.169	.515
	cPFC	.183	.167	.582
	PFC	.314	.288	.901
	cmPFC	.237	.225	.744
	oracle	.108	.109	
Non-normal data (mixture of normals)				
binary	cLAD	.219	.197	.541
	LAD	.272	.251	.896
	cmLAD	.246	.228	.517
	cPFC	.236	.224	.602
	PFC	.291	.286	.958
	cmPFC	.262	.253	.863
	oracle	.167	.149	
categorical	cLAD	.206	.183	.466
	LAD	.273	.249	.887
	cmLAD	.231	.214	.601
	cPFC	.241	.216	.624
	PFC	.280	.266	.903
	cmPFC	.256	.228	.859
	oracle	.158	.126	