Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: http://www.tandfonline.com/loi/uasa20

Covariate Information Matrix for Sufficient Dimension Reduction

Weixin Yao, Debmalya Nandy, Bruce G. Lindsay & Francesca Chiaromonte

To cite this article: Weixin Yao, Debmalya Nandy, Bruce G. Lindsay & Francesca Chiaromonte (2018): Covariate Information Matrix for Sufficient Dimension Reduction, Journal of the American Statistical Association, DOI: 10.1080/01621459.2018.1515080

To link to this article: https://doi.org/10.1080/01621459.2018.1515080

Covariate Information Matrix for Sufficient Dimension Reduction

Weixin Yao[*],
1337 Olmsted Hall, Department of Statistics,
University of California, Riverside, Riverside, CA 92521
Debmalya Nandy[†], **Bruce G. Lindsay**[‡] and **Francesca Chiaromonte**[§¶]
325 Thomas Building, Department of Statistics,
Penn State University, University Park, PA 16802

Abstract

Building upon recent research on the applications of the Density Information Matrix (DIM), we develop a tool for Sufficient Dimension Reduction (SDR) in regression problems called Covariate Information Matrix (CIM). CIM exhaustively identifies the Central Subspace (CS) and provides a rank ordering of the reduced covariates in terms of their regression information. Compared to other popular SDR methods, CIM does not require distributional assumptions on the covariates, or estimation of the mean regression function. CIM is implemented via eigen-decomposition of a matrix estimated with a previously developed efficient nonparametric density estimation

[*]W. Yao is the corresponding author and was supported by NSF grant DMS-1461677.

[†]D. Nandy and W. Yao contributed equally to this work and are jointly first-authors of this article.

[‡]During the development of this work, B. G. Lindsay passed away due to an illness. We honor the contributions of our dear friend, generous mentor and brilliant colleague.

[§]F. Chiaromonte, B. G. Lindsay and D. Nandy were supported by NSF grant DMS-1407639. F. Chiaromonte is also affiliated with the Institute of Economics and EMbeDS, Sant'Anna School of Advanced Studies, Piazza Martiri della Libertà 33, Pisa, Italy 56127.

[¶]We gratefully acknowledge Drs. Subir Ghosh and Zhiwei Zhang of the University of California Riverside, and Drs. Bing Li, Yanyuan Ma, Matthew Reimherr and Lingzhou Xue of Penn State, for their useful comments. We acknowledge Dr. Nikolay V. Balashov of Penn State for clarifications on the Ozone data application. We are also grateful to Drs. Yanyuan Ma (Penn State), Liping Zhu (Shanghai University) and Peng Zeng (Auburn University) for generously sharing codes respectively for the semiparametric, dMAVE and Fourier Sufficient Dimension Reduction methods. Our special thanks go to Dr. Yingcun Xia (National University of Singapore) for sharing code on an extension of the SR method still in preparation, and for his insightful responses to several questions. Ge Zhao, an advisee of Dr. Ma in the Statistics graduate program at Penn State, was instrumental in adapting FORTRAN code for our simulations and real data application. Finally, we are in debt to anonymous reviewers and editors whose comments helped us greatly improve our work.

technique. We also propose a bootstrap-based diagnostic plot for estimating the dimension of the CS. Results of simulations and real data applications demonstrate superior or competitive performance of CIM compared to that of some other SDR methods.

Keywords: Fisher information matrix; Density information matrix; Central subspace; Sufficient dimension reduction; Nonparametric density estimation; Bootstrap.

1 Introduction

Contemporary applications of regression, such as those in biology, medicine, public health, sociology, and economics, almost always involve a multiplicity of predictor variables. Whether the covariate vector comprises tens, hundreds or thousands of variables, methods to reduce its dimension are critical; after reduction, parametric and nonparametric regression modeling techniques, as well as graphical diagnostics, are more effective and easier to handle. The literature on dimension reduction in regression includes several approaches, such as Projection Pursuit Regression (Friedman and Stuetzle, 1981), Principal Component Regression (Hotelling, 1957; Kendall, 1957) and Sufficient Dimension Reduction (SDR) (see Li (1991) and references below). This article introduces a novel approach to SDR based on the new introduced *Covariate Information Matrix* (CIM).

Let $Y \in \mathbb{R}$ denote the response and $\mathbf{X} \in \mathbb{R}^p$ the covariate vector. Intuitively, the CIM corresponds to the (expected) Fisher Information Matrix for the regression density $f(y \mid \mathbf{x})$, treating the observed $\mathbf{X} = \mathbf{x}$ as the "parameter". Its eigen-decomposition identifies linear combinations of the covariates that are *most informative* on the response; the eigen-vectors capture the *Central Subspace* (CS; Cook (1994, 1996, 1998)), that is the smallest covariate-subspace preserving full regression information. The eigen-values rank the reduced covariates in terms of such information, providing guidance on how many to use for subsequent analysis. The CIM can also be written as the difference between two *Density Information Matrices* (DIM's) (see Hui and Lindsay (2010) and Lindsay and Yao (2012)): the DIM for the inverse regression density $f(\mathbf{x} \mid y)$ (see Li (1991), Cook and Weisberg (1991), Wang and Xia (2008)), and that for the marginal covariate density $f(\mathbf{x})$. We

use the *f2 method of computation*, a reliable and computationally efficient nonparametric density estimation technique (Hui and Lindsay, 2010), to estimate the DIM's, and hence, the CIM.

In the last 25 years, several methods have been proposed to perform SDR, e.g. *Sliced Inverse Regression* (SIR; Li (1991)), *Sliced Average Variance Estimation* (SAVE; Cook and Weisberg (1991)), *Principal Hessian Directions* (PHD; Li (1992)), *Minimum Average Variance Estimation* (MAVE; Xia et al. (2002)), *Inverse Regression* (Cook and Ni (2005)), *Simple Contour Regression* (Li et al. (2005)), *Fourier estimation* (Fourier, hereafter; Zhu and Zeng (2006)), and *Sliced Regression* (SR; Wang and Xia (2008)). Many of these methods require assumptions on the distribution of $\mathbf{X}$. For example, SIR requires the so-called *linearity condition* and SAVE requires both the *linearity condition* and the *constant conditional variance condition* (see Section 2.1). In addition, some SDR methods do not guarantee an exhaustive estimation of the CS (Zhu and Zeng, 2006; Wang and Xia, 2008). For example, SIR cannot capture directions in the covariate space along which Y varies symmetrically, and PHD and MAVE focus on linear combinations of $\mathbf{X}$ that are sufficient solely for the mean regression function $\mathrm{E}[Y|\mathbf{X}]$ – ignoring potential heteroscedasticity of Y. Importantly, there exist SDR developments that address these limitations. For instance, Xia (2007) extended MAVE to dMAVE, which considers conditional density functions and thus targets the whole CS. Also, Ma and Zhu (2012) proposed a semiparametric method for efficient estimation of the CS based on the complete family of influence functions, which does not require assumptions on the distribution of $\mathbf{X}$ and unveils interesting connections among many SDR approaches that use inverse regression. Other recent developments in SDR include Li and Wang (2007), Li and Dong (2009), Dong and Li (2010), Luo et al. (2009), Zhu et al. (2010a), Yin and Li (2011), Ma and Zhu (2013a), Ma and Zhu (2014), Wang et al. (2015), and Xue et al. (2018). See Ma and Zhu (2013b) for a review and further references.

Our CIM approach does not require strong assumptions on the distribution of $\mathbf{X}$, such as the linearity condition or the constant conditional variance condition. It recovers the CS exhaustively based on an eigen-decomposition, and it ranks the projected covariates based

on the magnitude of the corresponding eigenvalues. Interestingly, for CIM this order reflects a rigorously defined Fisher information as contained in the conditional regression density. The remainder of the article is organized as follows. Section 2 provides background on the DIM and SDR. Section 3 presents the CIM, its application to SDR, and its implementation. Section 4 describes a bootstrap-based diagnostic for estimating the dimension of the CS, built upon ideas in Ye and Weiss (2003). Section 5 contains simulation results on the performance of CIM in comparison to some popular SDR methods and on our diagnostic for dimension estimation. Section 6 presents real data analyses and Section 7 some concluding remarks. Proofs and additional details are given in the Supplementary Material.

2 Relevant Background

In this section, we describe the *Density Information Matrix* (DIM), which has close connections to the methods proposed in Section 3, and list some of its prior applications. We also review some key concepts in Sufficient Dimension Reduction.

Let $\mathbf{X} \in \mathbb{R}^p$ be a random vector with density $f(\mathbf{x})$ satisfying standard regularity conditions. Assume finite first and second order moments; $E[\mathbf{X}] = \mathbf{0}$ (without loss of generality) and $Var[\mathbf{X}] = \Sigma_{\mathbf{X}}$. Following Hui and Lindsay (2010) and Lindsay and Yao (2012), we define the sample space score vector for f at $\mathbf{x}$ as $U_f(\mathbf{x}) = \nabla_{\mathbf{x}} \log f(\mathbf{x})$ and the *Density Information Matrix* (DIM) for f as

$$\mathbb{J}_f = \int U_f(\mathbf{x}) U_f(\mathbf{x})^T f(\mathbf{x}) d\mathbf{x} = \int \frac{\nabla_{\mathbf{x}} f(\mathbf{x}) \nabla_{\mathbf{x}} f(\mathbf{x})^T}{f(\mathbf{x})} d\mathbf{x}. \tag{1}$$

In what follows, we often use the notations $\mathbb{J}_f$ and $\mathbb{J}_{\mathbf{X}}$ interchangeably. Note that Papaioannou and Ferentinos (2005) already introduced $\mathbb{J}_f$ for a univariate density, calling it the *Fisher information number*. The DIM is a matrix characterizing $f(\mathbf{x})$ or $\mathbf{X}$. Much like the covariance matrix, it conveys useful information about the random vector and has important applications. For instance, let $\mathbf{Z} = \Sigma_{\mathbf{X}}^{-1/2} \mathbf{X}$ with $\Sigma_{\mathbf{Z}} = \mathbb{I}_p$ (the identity matrix of order p). The eigen-decomposition of $\mathbb{J}_{\mathbf{Z}}$ identifies directions of strongest departure from

normality – thus efficiently solving the otherwise computationally burdensome problem of Projection Pursuit (Friedman and Tukey, 1974); see Hui and Lindsay (2010) for more details. Lindsay and Yao (2012) further extended the applications of the DIM to Independent Component Analysis (Jutten and Herault, 1991; Comon, 1994; Hyvärinen and Oja, 2000), the assessment of spherical symmetry of densities (Bartlett, 1934; Hartman and Wintner, 1940), and Markov networks utilized in graphical models (Jordan, 1998; Lauritzen, 1996). Here, we extend the use of the DIM for sufficient dimension reduction in a regression setting.

2.1 Basic Notions on Sufficient Dimension Reduction

Back to a regression problem with response Y and covariate vector $\mathbf{X} \in \mathbb{R}^p$, a subspace $\mathcal{S}$ of $\mathbb{R}^p$ is called a dimension reduction subspace of the regression, i.e. of the conditional distribution of $Y|\mathbf{X}$, if $Y \perp\!\!\!\perp \mathbf{X}|P_{\mathcal{S}}\mathbf{X}$ ($\perp\!\!\!\perp$ indicates independence and $P_{(\cdot)}$ the orthogonal projection operator in the standard inner product). When the intersection of all dimension reduction subspaces also satisfies this requirement, it is called the *Central Subspace* (CS; Cook (1994, 1996, 1998)) and denoted by $\mathcal{S}_{Y|\mathbf{X}}$. By construction, the CS is the smallest dimension reduction subspace. $d = dim(\mathcal{S}_{Y|\mathbf{X}})$, the minimal dimension necessary to capture $Y|\mathbf{X}$, is called the *structural dimension* of the regression.

Some regression applications focus exclusively on the mean regression function $\mathrm{E}[Y|\mathbf{X}]$. A subspace $\mathcal{S}$ is called a mean dimension reduction subspace (Cook and Li, 2002) if $Y \perp\!\!\!\perp \mathrm{E}[Y|\mathbf{X}]|P_{\mathcal{S}}\mathbf{X}$. When the intersection of all mean dimension reduction subspaces also satisfies this requirement, we call it the *Central Mean Subspace* (CMS; Cook and Li (2002)), denoted by $\mathcal{S}_{\mathrm{E}[Y|\mathbf{X}]}$. Again, by construction, the CMS is the smallest mean dimension reduction subspace. $\tilde{d} = dim(\mathcal{S}_{E[Y|\mathbf{X}]})$ is the minimal dimension necessary to capture $\mathrm{E}[Y|\mathbf{X}]$. As shown in Cook (1998) and Yin et al. (2008), under mild assumptions, the CS and the CMS exist – in the sense that intersecting subspaces do preserve the above requirements. We assume this existence throughout this article. Also note that $\mathcal{S}_{\mathrm{E}[Y|\mathbf{X}]} \subseteq \mathcal{S}_{Y|\mathbf{X}}$, that is the space capturing the mean regression function is contained in the space capturing $Y|\mathbf{X}$ in its entirety. Consequently, $\tilde{d} \leq d$.

Next, we list two conditions on the distribution of $\mathbf{X}$ that are used by popular SDR methods, such as SIR, SAVE and PHD. For notational simplicity, we assume $Var[\mathbf{X}] = \mathbb{I}_p$. The *linearity condition* imposes that

$$\mathrm{E}[\mathbf{X} \mid B^T\mathbf{X}] = P_B\mathbf{X}, \tag{2}$$

and the *constant conditional variance condition* that

$$Var[\mathbf{X} \mid B^T\mathbf{X}] = Q_B, \tag{3}$$

where B is any $p \times d$ basis matrix of $\mathcal{S}_{Y|\mathbf{X}}$, $P_B = B(B^TB)^{-1}B^T$ and $Q_B = \mathbb{I}_p - P_B$. If $\mathbf{X}$ has an elliptically contoured multivariate distribution, (2) holds for any projection space (Eaton, 1986). If $\mathbf{X}$ is multivariate Gaussian, both (2) and (3) hold for any projection space (Ma and Zhu, 2013b).

3 The Covariate Information Matrix

In this section we introduce the *Covariate Information Matrix* (CIM) of a regression as an expected Fisher Information Matrix where the observed covariate vector $\mathbf{X} = \mathbf{x}$ plays the role of "parameter". We show that its eigen-decomposition can be used to identify the CS and prove some important properties. Crucially, we rewrite the CIM as the difference between two DIM's and, based on this formulation, describe the implementation of our SDR approach.

3.1 Definition of the CIM

To capture covariate information in the regression of $Y \in \mathbb{R}$ on $\mathbf{X} \in \mathbb{R}^p$, we adopt a formulation similar to the Fisher Information for a parameter. Consider the regression density $f(y \mid \mathbf{x})$, assume it satisfies the standard regularity conditions in likelihood analysis, and think of $\mathbf{X} = \mathbf{x}$ as its "parameter". The score vector for $\mathbf{x}$ at $Y = y$ will be $U_{\mathbf{x}}(y) =$

$\nabla_{\mathbf{x}} \log f(y \mid \mathbf{x})$ and the Fisher Information Matrix for $\mathbf{x}$ will be

$$\mathbb{F}_{\mathbf{x}} = \int U_{\mathbf{x}}(y) U_{\mathbf{x}}(y)^T f(y \mid \mathbf{x}) dy. \tag{4}$$

$\mathbb{F}_{\mathbf{x}}$ tells us how much Fisher information about $\mathbf{X} = \mathbf{x}$ is contained in $f(y|\mathbf{x})$. This represents a local measure, as it depends on $\mathbf{x}$, and is unconventional in that the "parameter" is observed, not unknown. Nevertheless, it provides a natural way of assessing how sensitive the distribution of $Y|\mathbf{X}$ is to changes in $\mathbf{X}$. Now, let $f(\mathbf{x})$ be the marginal covariate density. We define the *Covariate Information Matrix* (CIM) as the expected value of (4) with respect to $f(\mathbf{x})$:

$$\mathbb{C}_{\mathbf{X}} = \mathrm{E}[\mathbb{F}_{\mathbf{X}}] = \int \mathbb{F}_{\mathbf{x}} f(\mathbf{x}) d\mathbf{x}. \tag{5}$$

This is akin to introducing a prior distribution on our "parameter" $\mathbf{x}$ and computing a "Bayesian version" of the Fisher information matrix. Next, we provide two examples to clarify the concept of CIM.

Example 3.1. *Let $Y \in \{0, 1\}$ be a binary response and $\mathbf{X} \in \mathbb{R}^p$ a vector of continuous covariates. Consider the logistic regression model:*

$$\Pr(Y = 1|\mathbf{x}) = p_{\mathbf{x}} = \frac{\exp(\boldsymbol{\beta}^T \mathbf{x})}{1 + \exp(\boldsymbol{\beta}^T \mathbf{x})}, \quad \boldsymbol{\beta} \in \mathbb{R}^p.$$

The traditional parametric Fisher information matrix for $\boldsymbol{\beta}$ at a fixed $\boldsymbol{X} = \mathbf{x}$ is $\mathbf{x}\mathbf{x}^T \cdot p_{\mathbf{x}}(1 - p_{\mathbf{x}})$. However, the Fisher information matrix for observed "parameter" $\mathbf{x}$ is $\mathbb{F}_{\mathbf{x}} = \boldsymbol{\beta}\boldsymbol{\beta}^T \cdot p_{\mathbf{x}}(1-p_{\mathbf{x}})$. This matrix tells us about the Fisher information on the (fixed) $\mathbf{x}$ contained within $f(y|\mathbf{x})$ at $\mathbf{x}$. The CIM in this case is $\mathbb{C}_{\boldsymbol{X}} = E[\mathbb{F}_{\boldsymbol{X}}] = \boldsymbol{\beta}\boldsymbol{\beta}^T \cdot E[p_{\boldsymbol{X}}(1-p_{\boldsymbol{X}})]$, which is of rank 1 with the non-null eigen-vector proportional to $\boldsymbol{\beta}$. Therefore, the CIM indicates that there is just one linear combination of $\mathbf{X}$, namely $\boldsymbol{\beta}^T\mathbf{X}$, containing information about Y.

Example 3.2. *Let $f(y|\mathbf{x})$ be the density of a normal distribution with mean $\boldsymbol{\beta}^T\mathbf{x}$ and*

variance σ^2*. Then the Fisher information matrix for "parameter"* $\mathbf{x}$ *in* $f(y|\mathbf{x})$ *is given by*

$$\mathbb{F}_{\mathbf{x}} = \int [y - \mathbf{x}^T\boldsymbol{\beta}]\boldsymbol{\beta}\boldsymbol{\beta}^T[y - \mathbf{x}^T\boldsymbol{\beta}]/\sigma^4 \; f(y|\mathbf{x})dy = \boldsymbol{\beta}\boldsymbol{\beta}^T/\sigma^2.$$

Since $\mathbb{F}_{\mathbf{x}}$ *does not depend on* $\mathbf{x}$*, the CIM is* $\mathbb{C}_{\boldsymbol{X}} = \mathbb{F}_{\mathbf{x}}$*, which is again of rank* 1 *with the non-null eigen-vector proportional to* $\boldsymbol{\beta}$*. This means there is just one linear combination,* $\boldsymbol{\beta}^T\mathbf{X}$*, carrying information about* Y*. In the next section, we show that the eigen-decomposition of* $\mathbb{C}_{\boldsymbol{X}}$ *can recover all informative linear combinations of* $\boldsymbol{X}$*, and can be used to perform SDR. Also note that, in accordance with our intuition for the Fisher information about a parameter in a density,* $\mathbb{F}_{\mathbf{x}}$ *is inversely proportional to* σ^2 *– which is the variance of* $Y \mid \mathbf{x}$*.*

3.2 Properties of the CIM

Here we describe important properties of the CIM through some propositions (proofs are provided in Section 1 of the Supplementary Material). First, we link its eigen-decomposition to the CS. Note that, by construction, the CIM is non-negative definite.

Proposition 3.1. *Let* $\mathbb{C}_{\boldsymbol{X}} = \Gamma_{\boldsymbol{X}}\Lambda_{\boldsymbol{X}}\Gamma_{\boldsymbol{X}}^T$ *be the eigen-decomposition of the CIM, with eigenvalues* $\lambda_1 \geq \lambda_2 \ldots \geq \lambda_p \geq 0$*. Exactly* $d = dim(\mathcal{S}_{Y|\boldsymbol{X}})$ *of the eigenvalues are* > 0*, and the space spanned by the* d *corresponding eigenvectors,* $\gamma_1, \ldots, \gamma_d$*, coincides with* $\mathcal{S}_{Y|\boldsymbol{X}}$*.*

This means that the eigen-decomposition of the CIM recovers the CS exhaustively, identifying the minimal sufficient set of projected covariates as the first d coordinates of the vector $\tilde{\mathbf{X}} = \Gamma_{\mathbf{X}}^T\mathbf{X}$. Many other SDR methods, while guaranteed to recover directions inside the CS, are not guaranteed to exhaust it. In addition, the ordered directions produced by the CIM based on eigenvalues reflect their average Fisher information content, and thus, a rigorously defined notion of regression information; the magnitude of λ_j captures the informational contribution of $\tilde{X}_j$ to the regression; that is, to $Y|\mathbf{X}$. Our next result describes the effects of full-rank affine transformations on the CIM.

Proposition 3.2. *Let* A *be a full-rank* $p \times p$ *matrix and* $a \in \mathbb{R}^p$*. Then* $\mathbb{C}_{A\boldsymbol{X}+a} = A^{-T}\,\mathbb{C}_{\boldsymbol{X}}\,A^{-1}$*, where* A^{-T} *indicates the transpose of* A^{-1}*.*

Because of this proposition, we can work with any convenient affine transformation of the covariate vector. For instance, $\tilde{\mathbf{X}}$ has a diagonal CIM; $\mathbb{C}_{\tilde{\mathbf{X}}} = \Gamma_{\mathbf{X}}^T(\Gamma_{\mathbf{X}}\Lambda_{\mathbf{X}}\Gamma_{\mathbf{X}}^T)\Gamma_{\mathbf{X}} = \Lambda_{\mathbf{X}}$. We can go further and implement two stages of "whitening" to $\mathbf{X}$ as in Hui and Lindsay (2010): first standardize $\mathbf{X}$ to $\mathbf{Z} = \Sigma_{\mathbf{X}}^{-1/2}\mathbf{X}$ (recall we assume $\mathrm{E}[\mathbf{X}] = \mathbf{0}$ without loss of generality), then take the eigen-decomposition of $\mathbb{C}_{\mathbf{Z}}$ and form $\tilde{\mathbf{Z}} = \Gamma_{\mathbf{Z}}^T\mathbf{Z}$. This has a diagonal CIM $\mathbb{C}_{\tilde{\mathbf{Z}}} = \Lambda_{\mathbf{Z}}$ and also a diagonal covariance $Var[\tilde{\mathbf{Z}}] = \Gamma_{\mathbf{Z}}^T Var[\mathbf{Z}]\Gamma_{\mathbf{Z}} = \Gamma_{\mathbf{Z}}^T\mathbb{I}_p\Gamma_{\mathbf{Z}} = \mathbb{I}_p$. This is appealing because it means that the projected covariates in $\tilde{\mathbf{Z}}$ provide uncorrelated informational contributions to the regression with a diagonal covariate information matrix. In summary, performing SDR translates into taking the first d coordinates of the transformed vector $\tilde{\mathbf{Z}} = \Gamma_{\mathbf{Z}}^T\mathbf{Z} = \Gamma_{\Sigma_{\mathbf{X}}^{-1/2}\mathbf{X}}^T\Sigma_{\mathbf{X}}^{-1/2}\mathbf{X}$. In principle, this requires the inverse square root of the matrix $\Sigma_{\mathbf{X}}$ and the eigen-decomposition of the matrix $\mathbb{C}_{\Sigma_{\mathbf{X}}^{-1/2}\mathbf{X}}$. However, our next result shows that our target transformation $\Gamma_{\Sigma_{\mathbf{X}}^{-1/2}\mathbf{X}}^T\Sigma_{\mathbf{X}}^{-1/2}$, or in other words the directions $\Sigma_{\mathbf{X}}^{-1/2}\Gamma_{\Sigma_{\mathbf{X}}^{-1/2}\mathbf{X}}$, can be obtained by-passing the standardization stage, and thus the computation of $\Sigma_{\mathbf{X}}^{-1/2}$.

Proposition 3.3. *The directions $G = \Sigma_{\boldsymbol{X}}^{-1/2}\Gamma_{\Sigma_{\boldsymbol{X}}^{-1/2}\boldsymbol{X}}$ correspond to the right-side eigenvectors of the matrix $\mathbb{C}_{\boldsymbol{X}}\Sigma_{\boldsymbol{X}}$ in non-increasing order of eigenvalues.*

In practice, this allows us to form $\tilde{\mathbf{Z}}$ also when the sample size is not large enough to reliably estimate the precision matrix $\Sigma_{\mathbf{X}}^{-1}$, and to avoid computation of the matrix square root. We need reliable estimates of $\Sigma_{\mathbf{X}}$, $\mathbb{C}_{\mathbf{X}}$, and of course d. For $\Sigma_{\mathbf{X}}$, a natural choice is the sample covariance matrix. Estimation of $\mathbb{C}_{\mathbf{X}}$ and its practical implementation are discussed in Sections 3.3, 3.4, and 3.5. Estimation of d is discussed in Section 4. Before moving on, notice that, even if the structural dimension of the regression were not selected accurately, the first few coordinates of $\tilde{\mathbf{Z}}$ do represent the *most informative* projected covariates and those that carry the largest portion of explanatory power with respect to the response.

3.3 The CIM in Terms of Covariates-related DIMs

Here we rewrite the CIM as a difference between two Density Information Matrices (DIMs). One is the DIM for the marginal density $f(\mathbf{x})$ of the covariate vector $\mathbf{X}$, $\mathbb{J}_{\mathbf{X}}$, which is defined as in (1). The other is associated with the *inverse regression* of $\mathbf{X}$ on Y; that

is, $\mathbf{X}|Y$ captured by the conditional densities $f^{(y)}(\mathbf{x}) = f(\mathbf{x}|y)$ as $Y = y$ varies in its range. The logic of inverse regression has been extensively used in SDR, for example in Li (1991), Cook and Weisberg (1991), and Wang and Xia (2008), among others; for any given y, we consider the sample space score vector of $f^{(y)}(\mathbf{x})$, $U_{f^{(y)}}(\mathbf{x}) = \nabla_{\mathbf{x}} \log f^{(y)}(\mathbf{x})$, form $\mathbb{J}_{\mathbf{X}|Y=y} = \int U_{f^{(y)}}(\mathbf{x}) U_{f^{(y)}}(\mathbf{x})^T f^{(y)}(\mathbf{x}) d\mathbf{x}$, and take its expectation with respect to the marginal response density $f(y)$ to obtain

$$\mathbb{J}_{\mathbf{X}|Y} = \int \mathbb{J}_{\mathbf{X}|Y=y} f(y) dy. \tag{6}$$

Note that this definition is easily adapted to the case of a discrete or a categorical Y, replacing integration with the appropriate sum. The next result states that the CIM for the regression of Y on $\mathbf{X}$ is in fact the difference between the *density information* on $\mathbf{X}$ in the conditional density $f^{(y)}(\mathbf{x})$ averaged over y and that in the marginal density $f(\mathbf{x})$. As we will see in Section 3.5, this is key for the practical implementation of our approach.

Proposition 3.4. *The CIM can be written as* $\mathbb{C}_{\boldsymbol{X}} = \mathbb{J}_{\boldsymbol{X}|Y} - \mathbb{J}_{\boldsymbol{X}}$.

This proposition highlights one way in which our approach differs from many others: existing inverse regression SDR methods use only $\mathbf{X} \mid Y$, while we use both $\mathbf{X} \mid Y$ and the marginal distribution of $\mathbf{X}$. If we assume that the distribution of $\mathbf{X}$ is spherically symmetric, then $\mathbb{J}_{\mathbf{X}}$ is equidiagonal (Lindsay and Yao, 2012) and the eigen-decompositions of $\mathbb{C}_{\mathbf{X}}$ and $\mathbb{J}_{\mathbf{X}|Y}$ are equivalent. However, if the marginal distribution of $\mathbf{X}$ is more complex, the two eigen-decompositions differ. This remark connects with the conditions required by some SDR methods: SIR (Li, 1991) and PHD (Li, 1992) rely on the linearity condition (2), which holds if $\mathbf{X}$ is elliptically contoured. SAVE relies on both the linearity and the constant conditional variance condition (3), both of which hold if $\mathbf{X}$ is Gaussian. CIM does not rely on such assumptions and accounts for the marginal of $\mathbf{X}$, whichever its nature, through Proposition 3.4. Thus, our approach is fundamentally different and more general.

3.4 The CIM When the Response is Discrete or Categorical

As noted near Equation (6), the CIM can be easily adapted to the case of a discrete or a categorical Y. Broadly speaking, in these cases the CIM approach can be viewed as a form of discriminant analysis among the sub-populations defined by the levels of Y. In terms of Proposition 3.4, if $Y \in \{y_1, \ldots, y_m\}$ with $Pr(Y = y_j) = \pi_j$, $j = 1, \ldots, m$, we rewrite

$$\mathbb{J}_{\mathbf{X}|Y} = \sum_{j=1}^{m} \pi_j \mathbb{J}_{\mathbf{X}|Y=y_j}, \tag{7}$$

and use the eigen-decomposition of $\mathbb{C}_{\mathbf{X}} = \mathbb{J}_{\mathbf{X}|Y} - \mathbb{J}_{\mathbf{X}}$ to identify the projected covariates that are most informative in discriminating the m sub-populations. In the Supplementary Material, we provide some insights on the CIM approach in terms of discrimination between two densities corresponding to a binary Y. In the next section, we will see that even when Y is continuous, in practice we discretize *slicing* its range as to create sub-populations. This strategy of slicing is used by most SDR methods based on inverse regression, such as SIR (Li, 1991), SAVE (Cook and Weisberg, 1991) and SR (Wang and Xia, 2008). The number of slices represents a *tuning parameter* (see next section).

3.5 Implementation of the CIM Approach

So far our discussion has been at the population level. Based on Proposition 3.3, to implement our approach we need to estimate $\mathbb{C}_{\mathbf{X}}$ and $\Sigma_{\mathbf{X}}$. For the latter, it is natural to use the sample covariance matrix $\hat{\Sigma}_{\mathbf{X}}$, whereas the main hurdle lies in estimating $\mathbb{C}_{\mathbf{X}}$. Based on Proposition 3.4, this issue can be turned into estimation of $\mathbb{J}_{\mathbf{X}}$ and $\mathbb{J}_{\mathbf{X}|Y}$.

Estimation of $\mathbb{J}_{\mathbf{X}}$: Because of the second equality in (1), we need to estimate $\mathbb{J}_{\mathbf{X}} = \int \frac{\nabla_{\mathbf{x}} f(\mathbf{x}) \cdot \nabla_{\mathbf{x}}^T f(\mathbf{x})}{f(\mathbf{x})} d\mathbf{x}$. Even using a kernel density estimate, this integral will not have an explicit form due to the density in the denominator. As an alternative to computing the integral numerically with a simulation-based technique, Hui and Lindsay (2010) proposed the fast and explicit *f2 method of computation* which slightly alters the information problem replacing $\mathbb{J}_{\mathbf{X}}$ by $\mathbb{J}_{\mathbf{S}}$, the DIM for a surrogate $\mathbf{S}$ with density $f_{(2)}(\mathbf{s}) = \frac{f^2(\mathbf{s})}{\int f^2(\mathbf{x}) d\mathbf{x}}$. While very

similar to $\mathbb{J}_{\mathbf{X}}$ in practice, $\mathbb{J}_{\mathbf{S}}$ has an *explicit* form when $f(\mathbf{x})$ is estimated using $\hat{f}_H(\mathbf{x}) = \sum_{i=1}^{n} \frac{1}{n|H|}\phi_p(\mathbf{x} - \mathbf{x}_i\,;\,\mathbf{0}, H^2)$, where $\phi_p(\cdot\,;\,\mathbf{0}, H^2)$ is the p-variate Gaussian density with mean $\mathbf{0}$ and covariance matrix H^2. See Section 4 of the Supplementary Material for more details on the computation of $\mathbb{J}_{\mathbf{S}}$. Hui and Lindsay (2010) argued that $f_{(2)}$ preserves topological features of f, such as the locations of "peaks" and "valleys". An extended discussion of this surrogate approach, which also proved satisfactory in our simulation study, is provided in Lindsay and Yao (2012).

Estimation of $\mathbb{J}_{\mathbf{X}|Y}$: If Y is discrete or categorical, we use (7), estimating $\mathbb{J}_{\mathbf{X}|Y=y_j}$ and π_j for each $j = 1, 2, \ldots, m$. As $\hat{\pi}_j$, we take the sample proportion of observations with $Y = y_j$. To estimate $\mathbb{J}_{\mathbf{X}|Y=y_j}$, we use the f2 method of computation applied only to $\mathbf{X}$ observations with $Y = y_j$. If Y is continuous, we discretize it slicing its range; slices are usually formed as to have $\hat{\pi}_j \approx \frac{1}{L}$ for each j, where L is the number of slices, and we proceed with the f2 method of computation within each slice. Notably, in this fashion, we do not use the actual observed values of the continuous Y; rather, we use their order to partition the $\mathbf{X}$-observations. Thus, likewise other SDR methods that employ slicing, such as SIR (Li, 1991) and SR (Wang and Xia, 2008), the CIM approach is not affected by response outliers. Therefore, CIM is more robust in this regard compared to SDR methods like MAVE (Xia et al., 2002) and PHD (Li, 1992) that use the actual Y values.

Tuning parameters: Choosing the *number of slices* L in inverse regression SDR methods is recognized as a challenge (Wang and Xia, 2008). Our simulations show that for moderate sample size ($n = 200$ and 400), the performance of the CIM approach does vary with the number of slices used. In homoscedastic settings (that is, regressions with additive homoscedastic errors), CIM performs better with smaller L (thus more observations per slice) and L becomes less relevant as the sample size increases. In heteroscedastic settings (that is, regressions with additive heteroscedastic errors), performance is best with moderate L at all sample sizes considered (see Section 5 and Supplementary Material). Based on our empirical experience, $L = 3 - 5$ slices work well for most applications. Another important tuning parameter is the *bandwidth matrix*H of the Gaussian kernel density used in the f2 method. Hui and Lindsay (2010) argued that, for sufficiently large sample sizes, the quest

for informative projections using the f2 method is not very sensitive to the choice of H. They recommended the rule of thumb proposed in Bowman and Foster (1993), which sets $H = (4/(p+2))^{1/(p+4)} \hat{\Sigma}_{\mathbf{X}}^{1/2} n^{-1/p+4}$. This is what we used in our implementation.

Computational burden: SDR techniques like SIR (Li, 1991), SAVE (Cook and Weisberg, 1991), and PHD (Li, 1992) identify projective covariates through the computationally light eigen-decomposition of kernel matrices. The Fourier method (Zhu and Zeng, 2006) is also computationally light; it combines kernel matrices obtained using a family of transformations of the response Y and Fourier transforms of the gradients of their conditional mean functions. MAVE (Xia et al., 2002) estimates the conditional mean function with an expensive local linear smoothing. dMAVE (Xia, 2007) estimates the conditional density using a similarly heavy approach, with "double kernel" local linear smoothing. SR (Wang and Xia, 2008) slices the range of Y, and then utilizes nonparametric local linear smoothing similar to MAVE within each slice. Finally, the semiparametric method (Ma and Zhu, 2013a) exploits a family of influence functions with geometrical techniques described in (Bickel et al., 1998) and (Tsiatis, 2007). The implementation relies on an initial $\sqrt{n}$-consistent estimator, and utilizes nonparametric estimation of a conditional density, its derivative, and a conditional mean function – eventually solving an estimating equation. Notably, the semiparametric procedure involves the tuning of several bandwidths (see also Sections 5 and 7).

The computational burden of CIM lies somewhere in between those for the comparatively "lighter" (SIR, SAVE, PHD, and Fourier) and the "heavier" methods (SR, MAVE, dMAVE, and semiparametric) – CIM uses an eigen-decomposition and nonparametric estimation of marginal and conditional (inverse regression) covariate densities. However, the density estimation in CIM is performed quite efficiently with the f2 method described in Hui and Lindsay (2010). Table 2 in Section 5 illustrates the computational burden of several methods on simulated data.

4 Estimating the Structural Dimension d

To date, several methods have been proposed in SDR literature to estimate the dimension d of the CS (or the CMS). For inverse regression methods such as SIR (Li, 1991) and SAVE (Cook and Weisberg, 1991), as well as for our CIM, one needs to estimate the number of positive eigen-values of a particular non-negative definite matrix. As reviewed in Ma and Zhu (2013b), this task has been tackled with a variety of approaches including sequential tests (e.g. Bura and Yang (2011)), information criteria (e.g. Zhu et al. (2012)), sparse eigen-decomposition techniques (e.g. Zhu et al. (2010b)), and bootstrap-based techniques (e.g. Ye and Weiss (2003)). For methods such as MAVE (Xia et al., 2002), which employ nonparametric estimation of the mean regression function, d is estimated by minimizing leave-one-out cross-validation prediction error. Sequential tests do not provide a consistent estimate of d due to the type-I error, and pose other theoretical and implementation related concerns. Information criteria require an appropriate choice of penalty. Sparse eigen-decomposition techniques estimate d and the CS simultaneously, converting the eigen-decomposition of an inverse regression method into a least-squares problem and imposing an adaptive LASSO penalty (Tibshirani, 1996; Zou, 2006). Bootstrap-based techniques, though computationally expensive (Zeng, 2008), are entirely data-driven and intuitively appealing. In our article, expanding on ideas in Ye and Weiss (2003), we use the bootstrap to quantify *stability* in estimating the CS with various "working" structural dimensions, and propose a diagnostic plot that allows one to easily estimate d. We document the performance of this diagnostic plot, along with that of the CIM as a means to estimate the CS, via simulations and data applications in Sections 5 and 6.

4.1 Squared Trace Correlation and Its Properties

Let $\mathcal{S}_1$ and $\mathcal{S}_2$ be two subspaces of $\mathbb{R}^p$, both of dimension $q \leq p$. The *squared trace correlation* between them is $R^2(\mathcal{S}_1, \mathcal{S}_2) = \frac{1}{q}\text{tr}(P_{\mathcal{S}_1} P_{\mathcal{S}_2})$, where $P_{(\cdot)}$ indicates the orthogonal projection onto the argument space. This measures similarity between the two subspaces, reaching its maximum 1 if they coincide, and its minimum 0 if they are or-

thogonal. We also propose a complementary measure, the *squared null trace correlation* $R_o^2(\mathcal{S}_1, \mathcal{S}_2) = R^2(\mathcal{S}_1^\perp, \mathcal{S}_2^\perp) = \frac{1}{p-q}\text{tr}(Q_{\mathcal{S}_1} Q_{\mathcal{S}_2})$, where $Q_{(\cdot)}$ indicates the orthogonal projection onto the orthogonal complement of the argument space. This too reaches 1 if the orthogonal complements, and thus the subspaces, coincide, and 0 in the case of orthogonality. The next result establishes a rigorous relationship between the two quantities.

Proposition 4.1. $R_o^2(\mathcal{S}_1, \mathcal{S}_2) = 1 - \frac{q}{p-q}(1 - R^2(\mathcal{S}_1, \mathcal{S}_2))$.

$R^2(\mathcal{S}_1, \mathcal{S}_2)$ and $R_o^2(\mathcal{S}_1, \mathcal{S}_2)$ are positively and linearly related, and we have $R^2(\mathcal{S}_1, \mathcal{S}_2) = R_o^2(\mathcal{S}_1, \mathcal{S}_2)$ if $q = \frac{p}{2}$. Proposition 4.1 also leads to two lower bounds, namely

$$R^2(\mathcal{S}_1, \mathcal{S}_2) \geq \frac{q - (p-q)}{q} = \frac{2q - p}{q} \quad , \quad R_o^2(\mathcal{S}_1, \mathcal{S}_2) \geq \frac{(p-q) - q}{p-q} = \frac{p - 2q}{p-q}.$$

The first bound is meaningful when $q \geq \frac{p}{2}$, and the second when $q \leq \frac{p}{2}$. Intuitively, they capture the fact that if q is large (small) enough, the two subspaces (their orthogonal complements) cannot be orthogonal and the squared correlation (null correlation) cannot reach 0. The bounds also show us that $R^2(\mathcal{S}_1, \mathcal{S}_2)$ must go to 1 if q approaches p from the left, and $R_o^2(\mathcal{S}_1, \mathcal{S}_2)$ must go to 1 if q approaches 0 from the right – proving the conjecture regarding the trace correlation in Ye and Weiss (2003). The next result quantifies the expected similarity of two random subspaces within a given space.

Proposition 4.2. *Let $\mathcal{S}$ be a subspace of $\mathbb{R}^p$ of dimension $d \leq p$, and $\mathcal{S}_1$ and $\mathcal{S}_2$ be two subspaces of $\mathcal{S}$, each comprising $q \leq d$ random directions within $\mathcal{S}$. Then $E[R^2(\mathcal{S}_1, \mathcal{S}_2)] = \frac{q}{d}$.*

The expected similarity increases linearly with q, reaching 1 when q reaches d and $\mathcal{S}_1 = \mathcal{S}_2 = \mathcal{S}$. Using Proposition 4.1, we have $\text{E}[R_o^2(\mathcal{S}_1, \mathcal{S}_2)] = 1 - \frac{q}{p-q}(1 - \text{E}[R^2(\mathcal{S}_1, \mathcal{S}_2)]) = 1 - \frac{q}{p-q}\frac{d-q}{d}$ which also reaches 1 when q reaches d.

4.2 Bootstrap Scheme and a Novel Diagnostic Plot

Based on the above discussion, for each "working" structural dimension $q = 1, \ldots, p-1$ (omitting the uninteresting $q = 0$ and p), we:

- estimate a q-dimensional informative subspace $\hat{\mathcal{S}}_q$ as the span of the right eigen-vectors of $(\hat{\mathbb{J}}_{\mathbf{X}|Y} - \hat{\mathbb{J}}_{\mathbf{X}})\hat{\Sigma}_{\mathbf{X}}$ with the q largest eigen-values.

- draw $j = 1, \ldots, B$ bootstrap samples using each of which we estimate $\hat{\mathcal{S}}_q^{(j)}$ and compute $R_q^{2(j)} = R^2(\hat{\mathcal{S}}_q, \hat{\mathcal{S}}_q^{(j)})$, $R_{o,q}^{2(j)} = R_o^2(\hat{\mathcal{S}}_q, \hat{\mathcal{S}}_q^{(j)})$ and their product $R_q^{2(j)} R_{o,q}^{2(j)}$.

- average over bootstrap samples to form $\bar{R}_q^2 = \frac{1}{B}\sum_{j=1}^{B} R_q^{2(j)}$, $\bar{R}_{o,q}^2 = \frac{1}{B}\sum_{j=1}^{B} R_{o,q}^{2(j)}$ and $\overline{R_q^2 R_{o,q}^2} = \frac{1}{B}\sum_{j=1}^{B} R_q^{2(j)} R_{o,q}^{2(j)}$.

$\bar{R}_q^2$, $\bar{R}_{o,q}^2$ and $\overline{R_q^2 R_{o,q}^2}$ all measure "stability" of the estimation of $\mathcal{S}_q$. As argued in Ye and Weiss (2003) and Zhu and Zeng (2006), when $q < d$ we estimate one among infinitely many subspaces within $\mathcal{S}_{Y|\mathbf{X}}$, resulting in low stability. Therefore, $\bar{R}_q^2$, $\bar{R}_{o,q}^2$ and $\bar{R}_q^2 \bar{R}_{o,q}^2$ are small. When q is close to d, stability increases and the three quantities grow; in particular, for $q = d$, $\hat{\mathcal{S}}_q$ and $\hat{\mathcal{S}}_q^{(j)}$, $j = 1, \ldots, B$, all estimate the same $\mathcal{S}_{Y|\mathbf{X}}$ and the three quantities peak with values close to 1. When $q > d$, stability is low again as we estimate one among infinitely many subspaces formed adding irrelevant direction(s) to $\mathcal{S}_{Y|\mathbf{X}}$. This translates again in small values for the three quantities. However, the curve described by $\bar{R}_q^2$ must grow again to approach 1 when q moves right of d and towards p from the left. Similarly, the curve described by $\bar{R}_{o,q}^2$ must grow again to approach 1 as q moves left of d and towards 0 from the right. In contrast, the proposed measure $\overline{R_q^2 R_{o,q}^2}$ is not bound to grow again moving away from d on either side, and peaks exclusively at $q = d$. Our proposal is to plot all three curves on the same display and identify $q = \hat{d}$ as the dimension where $\overline{R_q^2 R_{o,q}^2}$ has its highest value, and $\bar{R}_q^2$ and $\bar{R}_{o,q}^2$ reach values similar to the ones they approach near p and 0, respectively (where they have their "technical" maxima). Examples of such *dimension estimation plots* are in Sections 5 and 6. In the Supplementary Material, we also include alternative versions in which, for each q, we draw boxplots of bootstrap values instead of averages. Note that, similar to what is argued in Ye and Weiss (2003), if $(\hat{\mathbb{J}}_{\mathbf{X}|Y} - \hat{\mathbb{J}}_{\mathbf{X}})\hat{\Sigma}_{\mathbf{X}}$ has a few very dominant eigen-values, $\bar{R}_q^2$ may be high (close to 1) also for $q < d$; however, this quantity is still expected to have a notable drop at the transition from $q = d$ to $q = d + 1$. Some plots in Sections 5 and 6 do in fact illustrate this behavior.

5 Simulation Study

To assess the performance of the CIM approach and compare it to other SDR methods, we create simulation scenarios combining different covariate distributions, models to generate the response, signal-to-noise ratios (SNR's) and sample sizes. These expand upon scenarios used already in the SDR literature, such as Li (1992), Li et al. (2005), Zhu and Zeng (2006), Zhu et al. (2010a), and Wang et al. (2015). Even though CIM and other methods can be used on much larger problems, for simplicity, here we consider a $p = 10$-dimensional $\mathbf{X}$. We provide three specifications for its distribution:

(a) *Independent:* $\mathbf{X} \sim \mathcal{N}_p(\mathbf{0}_p, I_p)$,

(b) *Correlated:* $\mathbf{X} \sim \mathcal{N}_p(\mathbf{0}_p, \Sigma_{\mathbf{X}})$, where the (i, j)-th element of $\Sigma_{\mathbf{X}}$ is $0.5^{|i-j|}$, and

(c) *Non-linear:* Generate $\mathbf{X} \sim \mathcal{N}_p(\mathbf{0}_p, \Sigma_{\mathbf{X}})$, then replace the 3^{rd} and 4^{th} coordinates with $X_3 = |X_1 + X_2| + |X_1|\epsilon_1$ and $X_4 = (X_1 + X_2)^2 + |X_2|\epsilon_2$, where ϵ_1 and ϵ_2 are independently drawn from $\mathcal{N}(0, 1)$.

Note that (a) and (b) satisfy both the linearity (2) and the constant conditional variance (3) conditions. However, in (c), both conditions are violated – potentially hindering the performance of methods, such as SIR, SAVE, and PHD. We form the response based on eight model specifications, each one with a $d = 2$ dimensional CS:

(1) $Y = \cos(2\boldsymbol{\beta}_1^T\mathbf{X}) - \cos(\boldsymbol{\beta}_2^T\mathbf{X}) + \sigma\epsilon$; $\boldsymbol{\beta}_1 = (1, 0, \ldots, 0)^T$, $\boldsymbol{\beta}_2 = (0, 1, 0, \ldots, 0)^T$.

(2) $Y = (\boldsymbol{\beta}_1^T\mathbf{X})^2 + \boldsymbol{\beta}_2^T\mathbf{X} + \sigma\epsilon$; $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2$ as in (1).

(3) $Y = \boldsymbol{\beta}_1^T\mathbf{X} + (\boldsymbol{\beta}_2^T\mathbf{X})\sigma\epsilon$; $\boldsymbol{\beta}_1 = (1, 1, 1, 1, 0, \ldots, 0)^T$, $\boldsymbol{\beta}_2 = (0, \ldots, 0, 1, 1, 1, 1)^T$.

(4) $Y = \boldsymbol{\beta}_1^T\mathbf{X} + 0.1\,\boldsymbol{\beta}_2^T\mathbf{X} + (\boldsymbol{\beta}_2^T\mathbf{X})\,\sigma\epsilon$; $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2$ as in (3).

(5) $Y = 3\sin^2(\boldsymbol{\beta}_1^T\mathbf{X}/4) + (1 + (\boldsymbol{\beta}_2^T\mathbf{X})^2)\,\sigma\epsilon$;
$\boldsymbol{\beta}_1 = (1, 1, 1, 0, \ldots, 0)^T$, $\boldsymbol{\beta}_2 = (1, 0, 0, 0, 1, 3, 0, 0, 0, 0)^T$.

(6) $Y \in \{0,1,2,3\}$ formed as $Y = I(\beta_1^T\mathbf{X} + \sigma\epsilon > 1) + 2\, I(\beta_2^T\mathbf{X} \cdot \sigma\epsilon > 1)$;
$\boldsymbol{\beta}_1 = (1,1,1,1,0,\ldots,0)^T$, $\boldsymbol{\beta}_2 = (0,\ldots,0,1,1,1,1)^T$.

(7) $Y \in \{0,1,2\}$ formed as $Y = I(-2 < Y_0 < 2) + 2\, I(Y_0 \geq 2)$,
where $Y_0 = 2(\beta_1^T\mathbf{X}) + 2\exp(\beta_2^T\mathbf{X})\,\sigma\epsilon$;
$\boldsymbol{\beta}_1 = (1,2,0,\ldots,0,2)^T/3$, $\boldsymbol{\beta}_2 = (0,0,3,4,0,\ldots,0)^T/5$.

(8) $Y \in \{0,1,2\}$ formed as $Y = I(-2 < Y_0 < 2) + 2\, I(Y_0 \geq 2)$,
where $Y_0 = 2(\beta_1^T\mathbf{X})^2 + 2\exp(\beta_2^T\mathbf{X})\sigma\epsilon$;
$\boldsymbol{\beta}_1, \boldsymbol{\beta}_2$ as in (7).

In all cases $\epsilon \perp\!\!\!\perp \mathbf{X}$ and $\epsilon \sim \mathcal{N}(0,1)$. Note that Y is a continuous variable in models (1)-(5), and a discrete variable in models (6)-(8). Among the models with continuous Y, Models (1) and (2) have a simple homoscedastic error term added to the mean function, whereas models (3)-(5) have a more complex, heteroscedastic structure – the error term is still additive with mean 0, but its variance depends on a linear combination of the covariates. Models (6)-(8) also belong to the heteroscedastic case in the sense that $Var(Y|\mathbf{x})$ depends on $\mathbf{x}$ (through $\boldsymbol{\beta}_j^T\mathbf{x}, j = 1,2$). The Central Subspace (CS) and the Central Mean Subspace (CMS) coincide in models (1), (2), (4), and (6)-(8). The CS has one more direction than the CMS in models (3) and (5), which will be missed by methods targeting the CMS. Finally, the mean function of model (1) has two Y-symmetric terms, that of model (2) has one, and both the mean and the variance functions of model (5) are Y-symmetric. These symmetries are expected to impede the performance of SIR.

We consider the ratio $SNR = \frac{Var(E(Y|\mathbf{X}))}{E(Var(Y|\mathbf{X}))}$. For homoscedastic models with continuous response (models (1) and (2)), this produces a traditional *Signal-to-Noise Ratio*: $SNR(\sigma) = \frac{Var(E(Y|\mathbf{X}))}{\sigma^2}$, where σ^2 is the variance of the error term. However, we compute SNR also for heteroscedastic models with both continuous and discrete response (models (3)-(5) and (6)-(8), respectively). Here both the numerator and the denominator comprise a signal (also the denominator depends on $\boldsymbol{\beta}^T\mathbf{X}$), and the ratio benchmarks the signal in the mean to that in the variance – a kind of *"Mean Signal"-to-"Variance Signal" Ratio*. In all cases, including the heteroscedastic ones, the parameter σ is used to modulate SNR values.

For homoscedastic models, we consider σ's generating SNR's around $10, 5, 2.5$ and 1. For heteroscedastic ones, we consider SNR's both above and below 1 (see Tables S3-S8 in the Supplement). For all simulation scenarios, we consider sample sizes between $n = 200$ and 400.

We compared our CIM approach to SIR (Li, 1991), SAVE (Cook and Weisberg, 1991), SR (Wang and Xia, 2008), PHD (Li (1992); residual-based), MAVE (Xia et al. (2002); "refined" MAVE), dMAVE (Xia, 2007), Fourier (Zhu and Zeng, 2006) and the semiparametric method (Ma and Zhu, 2013a). Methods that require slicing, including CIM, were run with different number of slices L. We also added to the comparison a benchmark, where instead of estimating the CS with a given method, we generated random subspaces of prescribed dimensions.

About the semiparametric method, estimation of a 2-dimensional CS requires tuning 4 bandwidth parameters on a case-by-case basis, without any general tuning guidelines. A poor choice of bandwidths can cause the algorithm to reach unsatisfactory local optima, producing unreliable results and performance comparisons. Among the models with continuous Y (i.e. (1)-(5)), we were able to implement good tuning and obtain reliable results only for a small subset of the simulation scenarios relative to models (1) and (2) – we omit these from the main text, and report them in the Supplementary Material (Tables S9 and S10). For models with discrete Y (i.e. (6)-(8)) we did not run the semiparametric method because available code does not readily generalize to these cases.

Table 1 displays selected results. These concern performance of the methods for models (2) and (5) with all three covariate distribution settings and $L = 5$, and model (8) with *Independent* $\mathbf{X}$ only. Only a few σ values (SNR's) are shown, with sample sizes $n = 200$ and 400. Complete results are reported in the Supplementary Material (Tables S1-S10). Entries in the Tables represent *trace correlations* (the square-root of the quantity defined in Section 4.1) between the estimated and the true CS, averaged over 200 independently simulated data sets; the closer the values are to 1, the better the performance.

For model (2), the best performer is MAVE closely followed by dMAVE – the extended version of MAVE that can estimate the full CS (recall that the CS and the CMS coincide

here). However, CIM is quite competitive, similar in performance to the Fourier method, and better than all other methods in most scenarios. For the more complex model (5), SIR has very poor performance due to symmetries, and MAVE and PHD do not do well because they target the CMS only. Here the best performer is clearly dMAVE. But CIM is again competitive, similar in performance to SAVE, and substantially better than all other methods. Notably, CIM performs especially well for $n = 400$; while all methods improve with larger samples, the gain is particularly marked for SR and CIM – likely due to improved local linear smoothing and kernel density estimation, respectively. For model (8) with discrete Y, CIM performs best in all scenarios. In particular, CIM performs substantially better than MAVE and dMAVE, for large values of σ. Other models with discrete Y ((6) and (7); see Tables S6-S7 in the Supplementary Material) confirmed these findings. We note that, to implement the Fourier method on these models, we used adjustments similar to those in the discrete response example of Zhu and Zeng (2006). For PHD and MAVE, we simply treated the response as an ordinal variable.

Next, we demonstrate the performance of the dimension estimation plots using our CIM approach (Section 4.2). Again due to space limitations, we show results for selected scenarios relative to models (2) and (5). Specifically, Figure 1 shows dimension estimation plots using CIM with $L = 5$, applied to model (2) with $\sigma = 0.55$ ($SNR \approx 9.9$; panel (a)), and to model (5) with $\sigma = 0.2$ ($SNR \approx 0.02$; panel (b)). In both cases we consider *Independent* $\mathbf{X}$ and $n = 400$. The dimensions are correctly identified; $\bar{R}^2_q, \bar{R}^2_{o,q}$, and $\overline{R^2_q R^2_{o,q}}$ have peaks at $q = 2$. Note that $\overline{R^2_q R^2_{o,q}}$ (dashed line) peaks only at $q = 2$, while $\bar{R}^2_q$ and $\bar{R}^2_{o,q}$ must grow again to approach 1 as q goes towards $q = 10$ and $q = 0$, respectively. Plots for additional scenarios, and corresponding alternative boxplot versions which give a sense of the variability associated with these curves, are provided in the Supplementary Material (Figures S1-S15). As expected, the "peaking" behavior of $\overline{R^2_q R^2_{o,q}}$ at $q = 2$ is more pronounced for larger SNR's and is maintained across varying L's used in CIM.

Finally, we compare methods in terms of computational burden. Table 2 reports computation time (in seconds) required to generate $d = 2$ output directions for models (2), (5), and (8), all with *Independent* $\mathbf{X}$, $n = 400$, and $\sigma = 0.55, 0.2$ and 4, respectively (see

also Table S11 in the Supplementary Material). SIR, SAVE, SR, PHD, MAVE, dMAVE and CIM were implemented in MATLAB R2017a (version 9.2.0.556344) and run on a laptop with $2.70 - 2.90$ GHz Intel(R) Core(TM) i7-7500U CPU, 31.8 GB usable RAM, and Windows 10 Education OS. The Fourier method was implemented in R version 3.4.2 with platform x86 64-w64- mingw32/x64 (64-bit) using C functions, and run on the same laptop. The semiparametric method was implemented in FORTRAN 90 and run on a Penn State cluster. As expected (see Section 3.5), CIM is slower than SIR, SAVE, PHD and the Fourier method (which uses only eigen-decompositions and other inexpensive steps) and faster than MAVE, dMAVE, SR and the semiparametric method (which involve nonparametric estimation of one or more among conditional mean functions, conditional densities and their gradients).

In summary, we find that CIM competes closely with the best performing SDR methods in both homoscedastic and heteroscedastic scenarios with continuous Y, and can outperform all methods considered in our simulation study in scenarios with discrete Y. Notably, it does so with a substantially lower computational cost compared to that of the existing best methods. Our dimension estimation plots also appear effective in both homoscedastic and heteroscedastic scenarios.

6 Applications to Real Data

In this section, we apply the CIM approach and other SDR methods to two real datasets. Similar to the simulated data considered in Section 5, and to many real data applications used in prior literature, these comprise an order of ten covariates.

6.1 Wine Recognition Data

This UCI machine learning repository data set has been widely used to demonstrate machine learning and statistical methods (e.g., Coomans et al. (1992)). It comprises a categorical response with three classes indicating types of wine cultivars, and $p = 13$ quantitative covariates representing wine constituents determined via a chemical analysis. These are

(1) Alcohol, (2) Malic acid, (3) Ash, (4) Alkalinity of ash, (5) Magnesium, (6) Total phenols, (7) Flavonoids, (8) Nonflavonoid phenols, (9) Proanthocyanins, (10) Color intensity, (11) Hue, (12) OD280/OD315 of diluted wines, and (13) Proline. The total sample size is $n = 178$, with 59 observations belonging to class 1, 71 to class 2, and 48 to class 3. Our purpose is to find linear combinations of the covariates which are most informative in predicting wine cultivars.

After standardizing the covariate vector to have mean zero and identity covariance matrix, we apply CIM with $L = 3$ slices (no. of response-classes). The eigenvalues are $\mathbf{32.76}, \mathbf{9.23}, 2.57, 1.73, 1.36, 1.00, 0.79, 0.44, 0.37,\ 0.36, 0.27, 0.17$, and 0.06. Prominence of the first two, which jointly capture $\sim 82\%$ of the covariate information on the response, suggests $\hat{d} = 2$. The dimension estimation plot (Figure 2(a)) confirms this choice; the largest drops in $\bar{R}^2_q$ and $\overline{R^2_q R^2_{o,q}}$ occur at the transition from $q = 2$ to $q = 3$ (for the boxplot version, see Figure S16 in the Supplementary Material). The two leading CIM directions are dominated by *Flavonoids, Color intensity* and *Proline*:

$$\hat{\boldsymbol{\beta}}_1 = (.03, .16, 0, .15, .06, -.19, -.\mathbf{57}, .02, -.11, .\mathbf{50}, -.10, -.31, -.\mathbf{47})^T$$
$$\hat{\boldsymbol{\beta}}_2 = (-.24, -.16, -.19, -.04, .05, .02, -.02, -.11, .14, -.\mathbf{53}, .14, .07, -.\mathbf{74})^T.$$

The projection of the data on their plane (the estimated CS) is shown in Figure 3(a), along with a random 2D projection for benchmarking in Figure 3(b). The three response classes are very nicely separated. In addition to CIM, we run other slice-based methods, viz. SIR, SAVE and SR, on this dataset (see Figure S17 in the Supplementary Material). Other methods do not have readily implementable codes for categorical responses. Interestingly, SIR and SR give results very similar to CIM on this data (trace correlations between the CS estimates are ≈ 0.98). Note that SIR here works rather well even though the covariate vector is clearly not elliptical. SAVE on the other hand performs poorly (results not shown).

6.2 Ozone Data

This 'mlbench' R-package dataset was used in Breiman and Friedman (1985) as well as in the SDR literature (see Li (1992)). It comprises a continuous response, the atmospheric ozone concentration in the Los Angeles Upland basin measured daily (maximum one-hour average mixing ratio by volume in parts per hundred million (pphm)), and $p = 8$ quantitative covariates representing meteorological features; namely: (1) *SBTP*: the Sandburg (CA) air force base temperature (in °F), (2) *IBHT*: the inversion base height (in ft.) at the Los Angeles International Airport (LAX), (3) *DGPG*: the pressure gradient (in mm Hg) from Daggett to LAX, (4) *VSTY*: the visibility (in miles) at LAX, (5) *VDHT*: the Vandenburg 500 millibar pressure height (in m), (6) *HMDT*: the humidity (in percent) at LAX, (7) *IBTP*: the inversion base temperature (in °F) at LAX, and (8) *WDSP*: the wind speed (in mph) at LAX. The dataset covers $n = 330$ days in 1976 considering only complete observations (no missing values). Note that, as expected, the response shows marked autocorrelation (see Figure S18 in the Supplementary Material). However, we follow Li (1992) and perform SDR without correcting for autocorrelation.

After standardizing the covariate vector to have mean zero and identity covariance matrix, we apply CIM with $L = 5$. The eigenvalues are $\mathbf{3.83}, \mathbf{1.37}, 0.50, 0.37, 0.27, 0.21, 0.19$, and 0.18. The first two capture $\sim 75\%$ of the covariate information on the response, suggesting $\hat{d} = 2$. The dimension estimation plot (Figure 2(b)) confirms this choice; also for this data the largest drops in $\bar{R}^2_q$ and $\overline{R^2_q R^2_{o,q}}$ occur at the transition from $q = 2$ to $q = 3$ (for the boxplot version, see Figure S20 in the Supplementary Material). The two leading CIM directions are:

$$\hat{\boldsymbol{\beta}}_1 = (-.25, .\mathbf{82}, -.34, .22, -.22, -.24, -.10, -.01)^T$$
$$\hat{\boldsymbol{\beta}}_2 = (.01, .02, -.27, .05, .\mathbf{90}, -.33, -.04, -.09)^T.$$

The first is driven by *IBHT*, but with substantial contributions by several other covariates. The second is strongly driven by *VDHT*, with only a couple of other covariates contributing non-negligibly.

Figure 4 shows the association of the response with the two leading CIM directions using $L = 5$ slices (see Figure S33 in the Supplementary Material for similar plots with different L's). Ozone concentration has a strong, curved but asymmetric mean relationship with the first projected variable, and a weaker, curved and symmetric mean relationship with the second – accompanied by marked heteroscedasticity. CIM estimation of the CS and the structural dimension using different L's in the range $3-10$ produces similar results (see Figures S19-S24 in the Supplementary Material). However, other SDR methods produce less satisfactory and seemingly less robust results on this data. Using our dimension estimation plots, SIR leads to $\hat{d} = 1$ for L in $3 - 10$, and SAVE to $\hat{d} = 1$ for $L = 8$ and 10 but decisively to $\hat{d} = 2$ for $L = 3$ and 5 (see Figures S25-S32 in the Supplementary Material). Based on association measures among subspaces discussed next, MAVE and PHD also seem to settle on $\hat{d} = 1$. Not surprisingly, the second direction is harder to detect for SIR, which misses symmetric effects on the mean, as well as for MAVE and PHD, which target only the CMS and miss effects on the variance. If we fix $\hat{d} = 2$ and use $L = 5$ for all slice-based methods, the trace correlations between the CS estimated by CIM and those estimated by SIR, SAVE, SR, PHD, MAVE, dMAVE, Fourier, and the semiparametric method are respectively $0.744, 0.972, 0.886, 0.664, 0.724, 0.760, 0.944$, and 0.772. The semiparametric method was initialized using the output of dMAVE, and we were able to tune the bandwidths (see also Section 5). CIM results are clearly closer to SAVE, SR, and Fourier than to the rest. Trace correlations for the first directions only are, respectively, $0.971, 0.997, 0.966, 0.843, 0.953, 0.938, 0.974$, and 0.499, suggesting that the first direction is correctly and similarly identified by all methods except for the semiparametric one. Those for the second directions only are, respectively, $0.317, 0.943, 0.734, 0.128, 0.288, 0.389, 0.897$, and 0.485, suggesting that SAVE, SR, and Fourier also catch the second direction found by CIM, but SIR, PHD, MAVE, dMAVE and the semiparametric method do not. Additional results using CIM with different numbers of slices are presented in Tables S12-S15 of the Supplementary Material.

7 Concluding Remarks

In this article, we described a new tool for Sufficient Dimension Reduction (SDR); the Covariate Information Matrix (CIM). Our proposal builds upon the novel and appealing use of information matrices in Hui and Lindsay (2010) and Lindsay and Yao (2012), exhaustively identifies the Central Subspace (CS) of a regression, and produces reduced covariates that are uncorrelated with diagonal density information, and are naturally ordered based on their regression information contributions.

Some popular SDR methods leverage structure in the inverse regression $\mathbf{X} \mid Y$ and utilize simple eigen-decompositions – thus being computationally light. Their reduced covariates are usually ordered by eigenvalue size, but such ordering is not as meaningful as the one produced by CIM. In addition, many of these methods require distributional assumptions on the covariates. Other SDR methods avoid such assumptions, often at the price of much heavier computation – e.g., requiring local nonparametric estimation of mean regression functions. Furthermore, with techniques such as SR, MAVE, dMAVE and the semiparametric method, the subspace of interest needs to be re-estimated for any specification of its dimension d – imposing an additional computational burden in applications where d itself is to be estimated on the data. On a different note, based on our personal communication with Dr. Yingcun Xia, Dr. Xia's group is developing a new SDR method based on the modification over SR and dMAVE. Please see Dr. Xia's personal website for more information.

CIM does not require "linearity" or the "constant conditional variance" condition for the covariate vector $\mathbf{X}$; in fact, in addition to leveraging the structure in $\mathbf{X} \mid Y$, it explicitly accounts for structure in the distribution of $\mathbf{X}$. At the same time, CIM is computationally light. In addition to an eigen-decomposition, it does involve kernel density estimation (marginal and inverse) for $\mathbf{X}$ – but this is accomplished with the fast, explicit and robust f2 method of computation introduced in Hui and Lindsay (2010). Also, CIM does not need to be re-run for different d's. In terms of tuning parameters, CIM relies on the number of slices used to reconstruct the structure of $\mathbf{X} \mid Y$ when the response is continuous, and

the bandwidth used for kernel density estimation in the f2 method. For the latter, we employ a rule of thumb recommended in Hui and Lindsay (2010). For the former, our simulations suggest choices of L between 3 and 5 for sample sizes $n = 200 - 400$. The underlying structural dimension in our simulations is $d = 2$, but we note that in CIM (like in SR and unlike in SIR) the choice of L does not constrain the rank of the matrix and, thus, the "reconstructable" structural dimension. Finding a data-driven, optimal number of slices is admittedly a delicate issue for all SDR methods that employ slicing, and more exploration is warranted on the role of both tuning parameters in CIM. However, our approach remains operational and performs satisfactorily under a range of reasonable tuning parameter choices.

Building upon ideas in Ye and Weiss (2003), we also proposed a bootstrap-based diagnostic tool for estimating the dimension d of the CS. We quantify stability in estimating both the CS and the corresponding null space; the product of these two measurements provides a diagnostic with an easier to interpret "peaking" behavior.

We used simulations and real data to show the competitive performance of CIM compared to other SDR methods, and the effectiveness of our structural dimension diagnostics. Interestingly, in discrete response scenarios, CIM outperformed popular methods such as MAVE (Xia et al., 2002) and dMAVE (Xia, 2007) whose performance was excellent with a continuous response. We are also in the process of investigating both computational burden and quality of CS estimation using the f2 method for larger p's (say, several tens or hundreds) than the ones in our current simulation study and real data applications.

Relatedly, an important extension of our work would concerns *very high dimensional, under-sampled data*. Proposition 3.3 theoretically allows one to avoid inverting and taking the square root of $\Sigma_{\mathbf{X}}$. However, the quality of estimation of $\Sigma_{\mathbf{X}}$ deteriorates if n is not large relative to p, and is very poor if $p \gg n$ (Ledoit and Wolf, 2004). Moreover, because of our choice of H along with the use of the surrogate $\mathbb{J}_{\mathbf{S}}$, the current CIM implementation does still require inversion of $\hat{\Sigma}_X$ – though this may be circumvented with an appropriate alternative bandwidth matrix H. In light of these considerations, it would be interesting to develop screening techniques rooted in the same information framework underlying Density

Information Matrices (Hui and Lindsay, 2010; Lindsay and Yao, 2012) and the CIM itself. In addition to screening, very high dimensional settings may warrant the use of penalization to implement sparse estimation of the CIM and of the CS. These ideas have already been introduced for other SDR methods, e.g. in Li (2007), Li and Yin (2008), Wang and Yin (2008), Li and Nachtsheim (2006) and Wang and Zhu (2013), to name a few.

Another interesting extension of our work would be an *adaptive CIM*, which would exploit local structure in the data building different reduced covariates in different regions of the covariate space. Technically, a local weighting density $w(\mathbf{x})$ (e.g. a kernel) can be used in place of the overall covariate density $f(\mathbf{x})$ in Equation (5) to define a local version of the matrix.

Finally, while CIM, like other SDR methods based on slicing, is applicable regardless of the nature of the response, the covariates are always assumed to be continuous. Projections and linear combinations are meaningful only for such variables, but methods have been developed to perform SDR in regression that comprise also *categorical covariates* (see for instance, Chiaromonte et al. (2002); Wen and Cook (2007)). Using the CIM approach on regressions with a mix of continuous and categorical covariates is outside the scope of this article, but an extension along the lines of the *partial SDR* of Chiaromonte et al. (2002) is certainly conceivable and worth pursuing.

Supplementary Material and Codes

A Supplementary Material file for this article is available online, containing proofs and more details on our analyses. CIM was implemented in MATLAB and plots were produced in R; codes are available from the authors upon request. We also intend to publish a CRAN package in near future.

References

Bartlett, M. S. (1934). The vector representation of a sample. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 30, pages 327–340. Cambridge Univ Press.

Bickel, P. J., Klaassen, C. A., Bickel, P. J., Ritov, Y., Klaassen, J., Wellner, J. A., and Ritov, Y. (1998). *Efficient and adaptive estimation for semiparametric models*, volume 2. Springer New York.

Bowman, A. and Foster, P. (1993). Adaptive smoothing and density-based tests of multivariate normality. *Journal of the American Statistical Association*, 88(422):529–537.

Breiman, L. and Friedman, J. H. (1985). Estimating optimal transformations for multiple regression and correlation. *Journal of the American statistical Association*, 80(391):580–598.

Bura, E. and Yang, J. (2011). Dimension estimation in sufficient dimension reduction: a unifying approach. *Journal of Multivariate Analysis*, 102(1):130–142.

Chiaromonte, F., Cook, R. D., and Li, B. (2002). Sufficient dimension reduction in regressions with categorical predictors. *Annals of Statistics*, pages 475–497.

Comon, P. (1994). Independent component analysis, a new concept? *Signal processing*, 36(3):287–314.

Cook, R. and Weisberg, S. (1991). Discussion of a paper by kc li. *J. Am. Statist. Assoc*, 86:328–32.

Cook, R. D. (1994). On the interpretation of regression plots. *Journal of the American Statistical Association*, 89(425):177–189.

Cook, R. D. (1996). Graphics for regressions with a binary response. *Journal of the American Statistical Association*, 91(435):983–992.

Cook, R. D. (1998). *Regression Graphics: Ideas for Studying Regressions Through Graphics*, volume 318. John Wiley & Sons.

Cook, R. D. and Li, B. (2002). Dimension reduction for conditional mean in regression. *Annals of Statistics*, pages 455–474.

Cook, R. D. and Ni, L. (2005). Sufficient dimension reduction via inverse regression: A minimum discrepancy approach. *Journal of the American Statistical Association*, 100:410–428.

Coomans, D., Aeberhard, S., and de Vel, O. (1992). Comparison of classifiers in high dimensional settings. Technical report, Technical Report 92-02, Dept. of Computer Science and Dept. of Mathematics and Statistics, James Cook University of North Queensland.

Dong, Y. and Li, B. (2010). Dimension reduction for non-elliptically distributed predictors: second-order methods. *Biometrika*, 97(2):279–294.

Eaton, M. L. (1986). A characterization of spherical distributions. *Journal of Multivariate Analysis*, 20(2):272–276.

Friedman, J. H. and Stuetzle, W. (1981). Projection pursuit regression. *Journal of the American statistical Association*, 76(376):817–823.

Friedman, J. H. and Tukey, J. W. (1974). A projection pursuit algorithm for exploratory data analysis.

Hartman, P. and Wintner, A. (1940). On the spherical approach to the normal distribution law. *American Journal of Mathematics*, 62(1):759–779.

Hotelling, H. (1957). The relations of the newer multivariate statistical methods to factor analysis. *British Journal of Statistical Psychology*, 10(2):69–79.

Hui, G. and Lindsay, B. G. (2010). Projection pursuit via white noise matrices. *Sankhya B*, 72(2):123–153.

Hyvärinen, A. and Oja, E. (2000). Independent component analysis: algorithms and applications. *Neural networks*, 13(4):411–430.

Jordan, M. I. (1998). *Learning in graphical models*, volume 89. Springer Science & Business Media.

Jutten, C. and Herault, J. (1991). Blind separation of sources, part i: An adaptive algorithm based on neuromimetic architecture. *Signal processing*, 24(1):1–10.

Kendall, M. (1957). A course in multivariate analysis.

Lauritzen, S. L. (1996). *Graphical models*, volume 17. Clarendon Press.

Ledoit, O. and Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of multivariate analysis*, 88(2):365–411.

Li, B. and Dong, Y. (2009). Dimension reduction for nonelliptically distributed predictors. *The Annals of Statistics*, pages 1272–1298.

Li, B. and Wang, S. (2007). On directional regression for dimension reduction. *Journal of the American Statistical Association*, 102(479):997–1008.

Li, B., Zha, H., and Chiaromonte, F. (2005). Contour regression: a general approach to dimension reduction. *Annals of statistics*, pages 1580–1616.

Li, K.-C. (1991). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86(414):316–327.

Li, K.-C. (1992). On principal hessian directions for data visualization and dimension reduction: another application of stein's lemma. *Journal of the American Statistical Association*, 87(420):1025–1039.

Li, L. (2007). Sparse sufficient dimension reduction. *Biometrika*, 94(3):603–613.

Li, L. and Nachtsheim, C. J. (2006). Sparse sliced inverse regression. *Technometrics*, pages 503–510.

Li, L. and Yin, X. (2008). Sliced inverse regression with regularizations. *Biometrics*, 64(1):124–131.

Lindsay, B. G. and Yao, W. (2012). Fisher information matrix: A tool for dimension reduction, projection pursuit, independent component analysis, and more. *Canadian Journal of Statistics*, 40(4):712–730.

Luo, R., Wang, H., Tsai, C.-L., et al. (2009). Contour projected dimension reduction. *The Annals of Statistics*, 37(6B):3743–3778.

Ma, Y. and Zhu, L. (2012). A semiparametric approach to dimension reduction. *Journal of the American Statistical Association*, 107(497):168–179.

Ma, Y. and Zhu, L. (2013a). Efficient estimation in sufficient dimension reduction. *Annals of statistics*, 41(1):250.

Ma, Y. and Zhu, L. (2013b). A review on dimension reduction. *International Statistical Review*, 81(1):134–150.

Ma, Y. and Zhu, L. (2014). On estimation efficiency of the central mean subspace. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(5):885–901.

Papaioannou, T. and Ferentinos, K. (2005). On two forms of fisher's measure of information. *Communications in Statistics-Theory and Methods*, 34(7):1461–1470.

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

Tsiatis, A. (2007). *Semiparametric theory and missing data*. Springer Science & Business Media.

Wang, H. and Xia, Y. (2008). Sliced regression for dimension reduction. *Journal of the American Statistical Association*, 103(482):811–821.

Wang, Q. and Yin, X. (2008). A nonlinear multi-dimensional variable selection method for high dimensional data: Sparse mave. *Computational Statistics & Data Analysis*, 52(9):4512–4520.

Wang, Q., Yin, X., and Critchley, F. (2015). Dimension reduction based on the hellinger integral. *Biometrika*, 102(1):95–106.

Wang, T. and Zhu, L. (2013). Sparse sufficient dimension reduction using optimal scoring. *Computational Statistics & Data Analysis*, 57(1):223–232.

Wen, X. and Cook, R. D. (2007). Optimal sufficient dimension reduction in regressions with categorical predictors. *Journal of Statistical Planning and Inference*, 137(6):1961–1978.

Xia, Y. (2007). A constructive approach to the estimation of dimension reduction directions. *The Annals of Statistics*, pages 2654–2690.

Xia, Y., Tong, H., Li, W., and Zhu, L.-X. (2002). An adaptive estimation of dimension reduction space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):363–410.

Xue, Y., Wang, Q., and Yin, X. (2018). A unified approach to sufficient dimension reduction. *Journal of Statistical Planning and Inference*.

Ye, Z. and Weiss, R. E. (2003). Using the bootstrap to select one of a new class of dimension reduction methods. *Journal of the American Statistical Association*, 98(464):968–979.

Yin, X. and Li, B. (2011). Sufficient dimension reduction based on an ensemble of minimum average variance estimators. *The Annals of Statistics*, pages 3392–3416.

Yin, X., Li, B., and Cook, R. D. (2008). Successive direction extraction for estimating the central subspace in a multiple-index regression. *Journal of Multivariate Analysis*, 99(8):1733–1757.

Zeng, P. (2008). Determining the dimension of the central subspace and central mean subspace. *Biometrika*, 95(2):469–479.

Zhu, L., Miao, B., and Peng, H. (2012). On sliced inverse regression with high-dimensional covariates. *Journal of the American Statistical Association*.

Zhu, L., Wang, T., Zhu, L., and Ferré, L. (2010a). Sufficient dimension reduction through discretization-expectation estimation. *Biometrika*, 97(2):295–304.

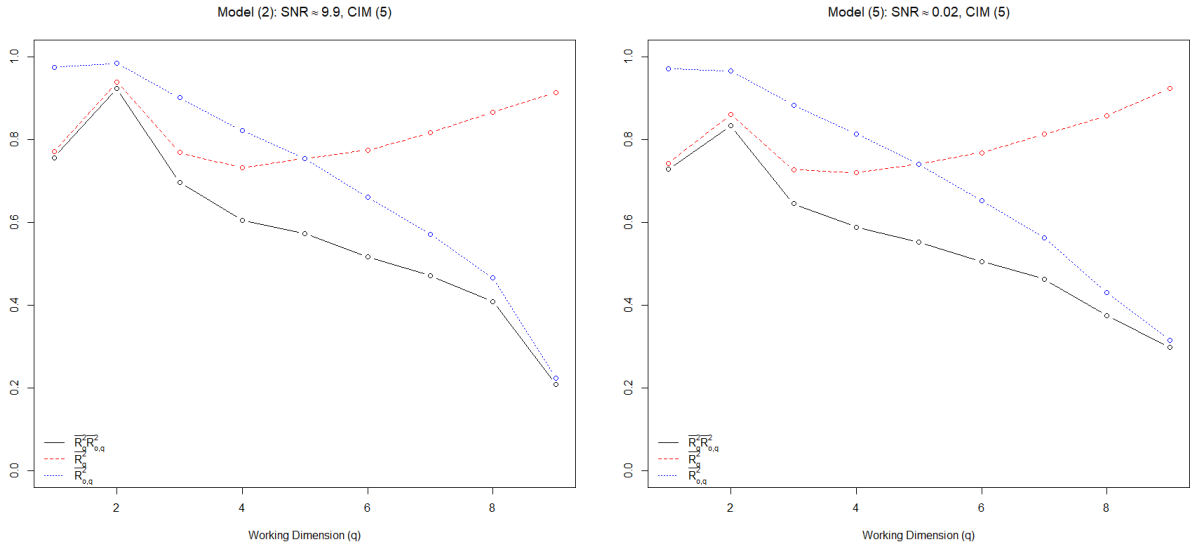
Zhu, L.-P., Yu, Z., and Zhu, L.-X. (2010b). A sparse eigen-decomposition estimation in semiparametric regression. *Computational Statistics & Data Analysis*, 54(4):976–986.

Zhu, Y. and Zeng, P. (2006). Fourier methods for estimating the central subspace and the central mean subspace in regression. *Journal of the American Statistical Association*, 101(476):1638–1651.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429.

Table 1: Mean (std. deviation) of trace correlation (R) in 200 repetitions for $n = 200$ and 400 in Models (2), (5), and (8) with *Independent* (Ind.), *Correlated* (Corr.), and *Non-linear* (NonL) $\mathbf{X}$. No. of slices $L = 5$ for SIR, SAVE, SR, and CIM in Models (2) and (5), and $L = 3$ naturally for Model (8) with discrete Y. Highest R in each row is **boldfaced**.

σ (Sample Size)	X	SIR	SAVE	SR	PHD	MAVE	dMAVE	Fourier	CIM	Random
Model (2) - Homoscedastic, Continuous Y										
$\sigma = 0.55$ (n = 200)	**Ind.**	0.722 (0.044)	0.802 (0.091)	0.860 (0.132)	0.729 (0.048)	**0.995** (0.002)	0.990 (0.004)	0.959 (0.022)	0.956 (0.021)	0.429 (0.126)
	Corr.	0.766 (0.073)	0.765 (0.089)	0.901 (0.113)	0.720 (0.043)	**0.991** (0.004)	0.982 (0.008)	0.927 (0.038)	0.930 (0.038)	0.424 (0.136)
	NonL	0.788 (0.062)	0.692 (0.042)	0.763 (0.079)	0.776 (0.105)	**0.987** (0.029)	0.947 (0.072)	0.820 (0.055)	0.753 (0.090)	0.437 (0.129)
$\sigma = 0.55$ (n = 400)	**Ind.**	0.733 (0.048)	0.964 (0.029)	0.943 (0.108)	0.733 (0.042)	**0.998** (0.001)	0.995 (0.002)	0.984 (0.007)	0.984 (0.006)	0.437 (0.128)
	Corr.	0.818 (0.077)	0.926 (0.056)	0.980 (0.046)	0.731 (0.045)	**0.996** (0.002)	0.992 (0.003)	0.966 (0.015)	0.977 (0.010)	0.439 (0.116)
	NonL	0.837 (0.053)	0.701 (0.029)	0.780 (0.089)	0.789 (0.118)	**0.996** (0.002)	0.986 (0.022)	0.847 (0.044)	0.785 (0.096)	0.423 (0.115)
Model (5) - Heteroscedastic, Continuous Y										
$\sigma = 0.2$ (n = 200)	**Ind.**	0.427 (0.130)	0.882 (0.073)	0.642 (0.189)	0.645 (0.099)	0.675 (0.086)	**0.893** (0.099)	0.741 (0.073)	0.870 (0.078)	0.422 (0.134)
	Corr.	0.351 (0.119)	0.804 (0.075)	0.603 (0.199)	0.543 (0.108)	0.617 (0.101)	**0.863** (0.100)	0.670 (0.068)	0.796 (0.089)	0.432 (0.135)
	NonL	0.595 (0.091)	0.728 (0.088)	0.730 (0.136)	0.560 (0.113)	0.649 (0.100)	**0.904** (0.060)	0.657 (0.075)	0.794 (0.094)	0.441 (0.127)
$\sigma = 0.2$ (n = 400)	**Ind.**	0.415 (0.134)	0.957 (0.030)	0.796 (0.134)	0.700 (0.088)	0.698 (0.084)	**0.980** (0.041)	0.798 (0.081)	0.957 (0.034)	0.440 (0.140)
	Corr.	0.355 (0.132)	0.918 (0.035)	0.783 (0.156)	0.589 (0.111)	0.626 (0.085)	**0.974** (0.017)	0.740 (0.081)	0.927 (0.029)	0.426 (0.130)
	NonL	0.648 (0.068)	0.787 (0.087)	0.867 (0.118)	0.573 (0.110)	0.665 (0.103)	**0.952** (0.035)	0.690 (0.047)	0.893 (0.070)	0.418 (0.123)
Model (8), Heteroscedastic, Discrete Y										
$\sigma = 3$ (n = 200)	**Ind.**	0.699 (0.055)	0.643 (0.101)	0.726 (0.105)	0.619 (0.108)	0.576 (0.152)	0.766 (0.121)	0.736 (0.081)	**0.823** (0.090)	0.439 (0.129)
$\sigma = 3$ (n = 400)	**Ind.**	0.718 (0.049)	0.762 (0.095)	0.786 (0.125)	0.681 (0.048)	0.681 (0.119)	0.894 (0.088)	0.821 (0.093)	**0.924** (0.045)	0.429 (0.122)
$\sigma = 4$ (n = 200)	**Ind.**	0.686 (0.047)	0.594 (0.130)	0.686 (0.088)	0.598 (0.095)	0.541 (0.146)	0.660 (0.145)	0.710 (0.070)	**0.776** (0.102)	0.426 (0.134)
$\sigma = 4$ (n = 400)	**Ind.**	0.718 (0.045)	0.723 (0.100)	0.760 (0.103)	0.658 (0.066)	0.633 (0.136)	0.819 (0.113)	0.783 (0.085)	**0.896** (0.065)	0.432 (0.119)
$\sigma = 5$ (n = 200)	**Ind.**	0.689 (0.056)	0.556 (0.133)	0.676 (0.075)	0.542 (0.128)	0.520 (0.144)	0.599 (0.148)	0.702 (0.069)	**0.727** (0.096)	0.440 (0.121)
$\sigma = 5$ (n = 400)	**Ind.**	0.713 (0.047)	0.685 (0.095)	0.736 (0.090)	0.622 (0.078)	0.579 (0.140)	0.741 (0.132)	0.767 (0.082)	**0.856** (0.076)	0.436 (0.126)

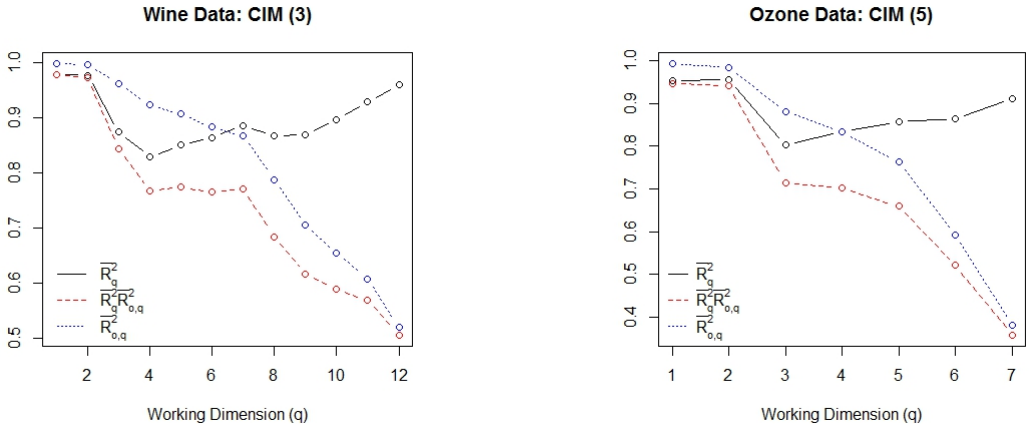


(a) (b)

Figure 1: Dimension estimation plots using **CIM** with $L = 5$ slices (500 bootstrap replicates). (a) **Model (2)** with $\sigma = 0.55$ ($SNR \approx 9.9$) and (b) **Model (5)** with $\sigma = 0.2$ ($SNR \approx 0.02$). In both cases the covariates in $\mathbf{X}$ are *Independent* and $n = 400$.

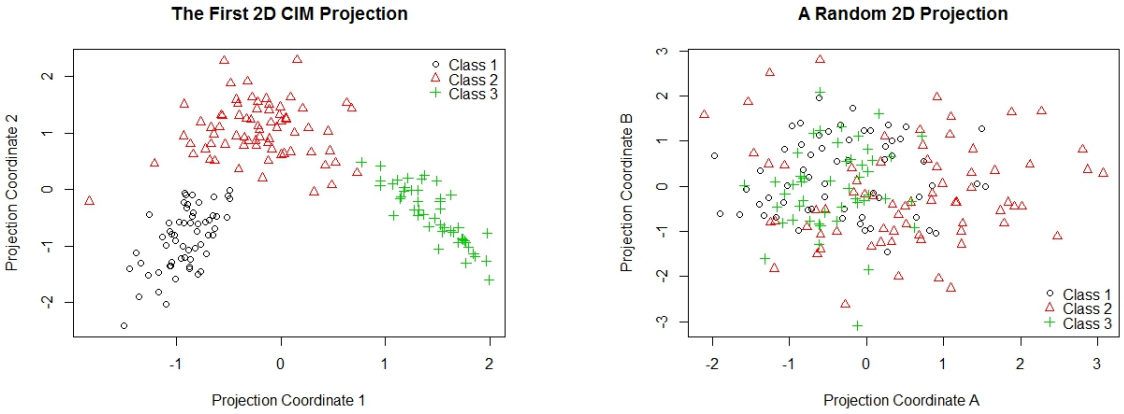
Table 2: Computation time in seconds (average of 200 runs; SD in parentheses) to generate $d = 2$ output directions for Models (2), (5), and (8) with $n = 400$, Independent $\mathbf{X}$, and specified σ. Number of slices $L = 5$ for SIR, SAVE, SR, and CIM in Models (2) and (5), and $L = 3$ naturally for Model (8) with discrete Y.

SIR	**SAVE**	**SR**	**PHD**	**MAVE**	**dMAVE**	**Fourier**	**Semiparametric**	**CIM**
Model (2): Homoscedastic, Continuous $Y(\sigma = 0.55)$								
5e-04	6e-04	0.8499	0.0026	1.1625	12.1979	0.0127	1.1711	0.5253
(3e-04)	(2e-04)	(0.0433)	(4e-04)	(0.0944)	(0.9517)	(0.00889)	(0.07786)	(0.0351)
Model (5): Heteroscedastic, Continuous $Y(\sigma = 0.20)$								
5e-04	6e-04	0.8358	0.0026	2.0753	12.8684	0.0131	6.7422	0.5290
(4e-04)	(2e-04)	(0.0141)	(3e-04)	(0.0534)	(2.0811)	(0.0082)	(1.8080)	(0.0096)
Model (8): Heteroscedastic, Discrete $Y(\sigma = 4)$								
6e-04	8e-04	0.8202	0.0026	2.0446	17.5507	0.0132	XX	0.5906
(5e-04)	(3e-04)	(0.0149)	(5e-04)	(0.0229)	(0.1781)	(0.0088)		(0.0084)



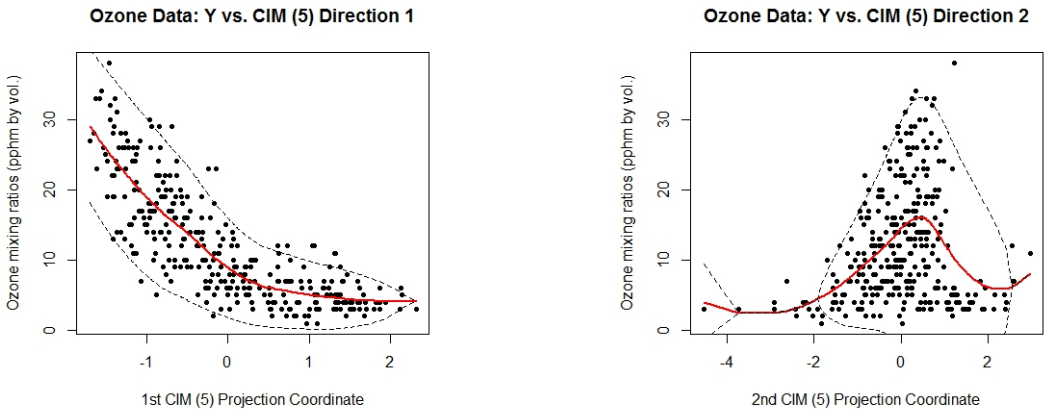
(a) (b)

Figure 2: Dimension estimation plots for **CIM** (500 bootstrap replicates). (a) **Wine Recognition Data** ($L = 3$) and (b) **Ozone Data** ($L = 5$).



(a) (b)

Figure 3: 2D projections of the **Wine Recognition Data** on: (a) the CS estimated via **CIM** ($L = 3$) and (b) a **Random** plane.



(a) (b)

Figure 4: Scatterplots of ozone concentration against the first (a) and second (b) leading directions estimated via **CIM** ($L = 5$). Solid lines are LOESS smooths; dashed lines around them represent 95% prediction bands obtained using the CRAN package 'msir'.