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 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, Sant’Anna School of Advanced Studies, Piazza Martiri della Libertà 33, Pisa, Italy 56127.
¶We gratefully acknowledge Drs. Subir Ghosh and Zhiwei Zhang of University of California Riverside and Drs. Bing Li, Yanyuan Ma, Matthew Reimherr and Lingzhou Xue of Penn State for their useful comments, and Dr. Nikolay V. Balashov of Penn State for clarifications on the Ozone data application.
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

In scientific fields as diverse as biology, medicine, public health, sociology and economics, high-dimensional data are increasingly common and regression problems involve large numbers of covariates. Methods to reduce the dimension of the covariate vector are critical in these settings; 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, e.g. 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 Covariate Information Matrix (CIM) approach to SDR.

Let $Y \in \mathbb{R}$ denote the response and $X \in \mathbb{R}^p$ the covariate vector. Intuitively, the CIM corresponds to the (expected) Fisher Information Matrix for the regression density $f(y | x)$, treating the observed $X = x$ as the “parameter”. Its eigen-decomposition identifies linear combinations of the covariates that are *most informative* of the response; the eigen-vectors capture the *Central Subspace* (CS; [Cook](1994) [1996] [1998]), i.e. the smallest subspace preserving full regression information, and 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 in terms of two Density Information Matrices (DIM’s) (see Hui and Lindsay (2010) and Lindsay and Yao (2012)); the DIM for the marginal covariate density \( f(x) \), and that for the inverse regression density \( f(x | y) \) (see Li (1991), Cook and Weisberg (1991), Wang and Xia (2008)) averaged over the marginal response density \( f(y) \).

We use the \( f^2 \) 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 (IR; Cook and Ni (2005)), Simple Contour Regression (SCR; Li et al. (2005)), Fourier estimation (Fourier; Zhu and Zeng (2006)), and Sliced Regression (SR; Wang and Xia (2008)). Many of these methods require assumptions on the distribution of \( X \). For example, SIR requires the so-called linearity condition and SAVE requires both the linearity condition and the constant conditional variance condition (see Section 2.2). 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 \( X \) that are sufficient solely for the mean regression function \( \text{E}[Y|X] \) – ignoring potential heteroscedasticity of \( Y \). Importantly, recent SDR developments do address these limitations. For instance, Xia (2007) extend MAVE proposing a density-based procedure which targets the whole CS. Also, Ma and Zhu (2012) propose a semiparametric method based on the complete family of influence functions which does not require assumptions on the distribution of \( X \) and uncovers interesting connections among the many SDR approaches that use inverse...
regression. Other recent developments in SDR include Li and Dong (2009), Dong and Li (2010), Luo et al. (2009), Zhu et al. (2010a), Yin and Li (2011), and Ma et al. (2013); see Ma and Zhu (2013) for a review and further references.

Our CIM approach does not require assumptions on $X$, recovers the CS exhaustively and, like other SDR methods based on eigen-decompositions, ranks the projected covariates in order of importance. Interestingly, for CIM this order reflects a rigorously defined regression information notion. The remainder of the article is organized as follows. Section 2 provides background on the DIM and on SDR. Section 3 presents the CIM, its application to SDR, and its implementation. Section 4 describes a bootstrap-based diagnostic approach for estimating the dimension of the CS built upon ideas in Ye and Weiss (2003). Section 5 contains simulation results on the performance of our approach in comparison to 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), connect it to the Fisher Information Matrix for a location parameter, and list some prior applications. We also review some key concepts in SDR.

2.1 Density and Fisher Information Matrices

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

$$\mathbb{J}_f = \int U_f(x)U_f(x)^T f(x)dx = \int \frac{\nabla_x f(x)\nabla_x f(x)^T}{f(x)} dx.$$  \hfill (1)

In the following we often use the notations $\mathbb{J}_f$ and $\mathbb{J}_X$ interchangeably. The DIM has a direct connection to the traditional Fisher Information Matrix. In the parametric location family $f(x - \theta), \theta \in \mathbb{R}^p$, the score vector for $\theta$ at $x$ is defined as $U_{\theta}(x) = \nabla_\theta \log f(x - \theta)$ and the Fisher Information Matrix for $\theta$ as

$$\mathbb{F}_\theta = \int U_{\theta}(x)U_{\theta}(x)^T f(x - \theta)dx.$$  \hfill (2)

Since $\nabla_\theta f(x - \theta) = -\nabla_x f(x - \theta)$, we have $U_{\theta}(x) = -U_f(x - \theta)$. Also, in the context of the location family, $X$ and $X + \theta$ carry the same information for any $\theta$. Thus for the location family densities, the Fisher Information Matrix for $\theta = 0$ coincides with the DIM for $f$ as expressed in (1): $\mathbb{F}_0 = \int U_f(x)U_f(x)^T f(x)dx = \mathbb{J}_f$. While this connection provides an important intuition on the DIM, our developments in this article, like those in [Lindsay and Yao (2012)], are not limited to parametric location families. The density $f$ is not restricted (except for standard regularity conditions) and is considered fully unknown to begin with; as we will see, we will estimate it nonparametrically. 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 $f(x)$ or $X$. Much like the covariance matrix, it conveys useful information about the random vector and it has important applications. For instance, let $Z = \Sigma^{-1/2}X$ with $\Sigma_Z = \mathbb{I}_p$ (the identity matrix). The eigen-decomposition of $\mathbb{J}_Z$ 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 (Lauritzen (1996), Jordan (1998)). Here, we extend the use of the DIM for dimension reduction of covariates in a regression setting.

2.2 Basic Notions on Sufficient Dimension Reduction

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

Some regression applications focus exclusively on the mean regression function $E[Y|X]$. A subspace $S$ is called a mean dimension reduction subspace (Cook and Li, 2002) if $Y \independent E[Y|X] | P_S X$. When the intersection of all subspaces satisfying this requirement also satisfies it, it is called the **Central Mean Subspace** (CMS; Cook and Li (2002)) and denoted by $S_{E[Y|X]}$. Again, by construction, the CMS is the smallest subspace satisfying the requirement; $\tilde{d} = \dim(S_{E[Y|X]})$ is the minimal dimension necessary to capture $E[Y|X]$. 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 does preserve the above requirements. We
assume existence throughout this article. Also note that \( \mathcal{S}_{E|Y|X} \subseteq \mathcal{S}_{Y|X} \); i.e. the space capturing the mean regression function is contained in the space capturing \( Y|X \) in its entirety. Consequently, \( \tilde{d} \leq d \).

Next, we list two conditions on the distribution of \( X \) that are used by popular SDR methods such as SIR, SAVE and PHD. For simplicity of notation, assume \( Var[X] = I_p \).

The linearity condition imposes that

\[
E[X \mid B^T X] = P_B X, \tag{3}
\]

and the constant conditional variance condition that

\[
Var[X \mid B^T X] = Q_B, \tag{4}
\]

where \( B \) is any \( p \times d \) basis matrix of \( \mathcal{S}_{Y|X} \), \( P_B = B(B^T B)^{-1}B^T \) and \( Q_B = I_p - P_B \). If \( X \) has an elliptically contoured multivariate distribution, (3) holds for any projection space (Eaton, 1986). If \( X \) is multivariate Gaussian, both (3) and (4) hold for any projection space (Ma and Zhu, 2013).

### 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 \( X = x \) 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 in terms of 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 \( X \in \mathbb{R}^p \), we adopt a formulation similar to the Fisher Information for a parameter. Consider the regression density \( f(y \mid x) \), assume it satisfies the standard regularity conditions in likelihood analysis, and think of \( X = x \) as its “parameter”. The score vector for \( x \) at \( Y = y \) will be \( U_x(y) = \nabla_x \log f(y \mid x) \) and the Fisher Information Matrix for \( x \) will be

\[
F_x = \int U_x(y)U_x(y)^T f(y \mid x)dy. \tag{5}
\]

\( F_x \) tells us how much Fisher information about \( X = x \) is contained in a single observation of \( Y = y \). This represents a local measure, as it depends on \( x \), and is unconventional in that the “parameter” is observed, not unknown. Nevertheless, it provides a natural way of assessing how sensitive the distribution of \( Y \mid X \) is to changes in \( X \). Now, let \( f(x) \) be the marginal covariate density. We define the Covariate Information Matrix (CIM) as the expected value of (5) with respect to \( f(x) \):

\[
C_x = \int F_x f(x)dx. \tag{6}
\]

This is akin to introducing a prior distribution on our “parameter” \( x \) and computing a “Bayesian version” of the Fisher information matrix. In the Supplementary Material, we provide an example distinguishing traditional Fisher information and covariate information (Section 2) as well as additional insights on the CIM (Sections 3 and 4).
3.2 Properties of the CIM

Here we describe important properties of the CIM through some propositions (proofs are provided in 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 $\mathbf{C}_x = \Gamma_x \Lambda_x \Gamma_x^T$ be the eigen-decomposition of the CIM, with eigenvalues $\lambda_1 \geq \lambda_2 \ldots \geq \lambda_p \geq 0$. Exactly $d = \text{dim}(S_{Y|x})$ of the eigenvalues are $> 0$, and the space spanned by the $d$ corresponding eigenvectors $\gamma_1, \ldots, \gamma_d$ coincides with $S_{Y|x}$.

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 $\tilde{X} = \Gamma_x^T X$. Other SDR methods, while guaranteed to recover directions inside the CS, are not guaranteed to exhaust it. Like in other SDR methods exploiting eigen-decompositions, projected covariates are ordered by importance based on eigenvalues. In the case of CIM though, these reflect a rigorously defined notion of regression information; the magnitude of $\lambda_j$ captures the informational contribution of $\tilde{X}_j$ to the regression; that is, to $Y|x$. 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 $\mathbf{C}_{Ax+a} = A^{-T} \mathbf{C}_x A^{-1}$, 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, $\tilde{X}$ has a diagonal CIM; $\mathbf{C}_{\tilde{X}} = \Gamma_{\tilde{X}}^T (\Gamma_x \Lambda_x \Gamma_x^T) \Gamma_{\tilde{X}} = \Lambda_{\tilde{X}}$. This is appealing because it means that the projected covariates provide independent informational contributions to the regression. We can go further and implement two stages of “whitening” as in Hui and Lindsay (2010): first standardize to $Z = \Sigma_{X}^{-1/2} X$ (recall we assume $E[X] = 0$ without loss of generality), then take the eigen-decomposition of $\mathbf{C}_Z$ and
form $\tilde{Z} = \Gamma_{Z}^{T}Z$. This has a diagonal CIM $C_{Z} = \Lambda_{Z}$ and also a diagonal covariance $Var[\tilde{Z}] = \Gamma_{Z}^{T}Var[Z]\Gamma_{Z} = \Gamma_{Z}^{T}I_{p}\Gamma_{Z} = I_{p}$. Thus, yet more appealingly, $\tilde{Z}$ comprises projected covariates that provide independent informational contributions and are uncorrelated. In summary, performing SDR translates into taking the first $d$ coordinates of the transformed vector $\tilde{Z} = \Gamma_{Z}^{T}Z = \Gamma_{Z}^{T}\Sigma_{X}^{-1/2}X$. In principle, this requires the inverse square root of the matrix $\Sigma_{X}$ and the eigen-decomposition of the matrix $C_{X}\Sigma_{X}^{-1/2}$. However, our next result shows that our target transformation $\Gamma_{Z}^{T}\Sigma_{X}^{-1/2}X$, or in other words the directions $\Sigma_{X}^{-1/2}X\Gamma_{Z}$, can be obtained by-passing the standardization stage, and thus the computation of $\Sigma_{X}^{-1/2}$.

**Proposition 3.3.** The directions $G = \Sigma_{X}^{-1/2}X\Gamma_{Z}$ correspond to the right-side eigenvectors of the matrix $C_{X}\Sigma_{X}^{-1/2}$ in non-increasing order of eigen-values.

In practice this allows us to form $\tilde{Z}$ also when the sample size is not large enough to reliably estimate the precision matrix $\Sigma_{X}^{-1}$, and to avoid computation of the matrix square root. We need reliable estimates of $\Sigma_{X}$, $C_{X}$, and of course $d$. For $\Sigma_{X}$, a natural choice is the sample covariance matrix. Estimation of $C_{X}$ and practical implementation are discussed in Sections 3.3, 3.4 and 3.5 and estimation of $d$ is discussed in Section 4. Before moving on we note that, even if the structural dimension of the regression were not selected properly, the first few coordinates of $\tilde{Z}$ 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 $f(x)$ of the covariate vector $X$, $J_{X}$, which is defined as in (1). The other is associated with the inverse regression of $X$ on $Y$; that is, $X|Y$ captured by the conditional densities $f(y|x) = f(x|y)$ as $Y = y$ varies in its range.
The logic of inverse regression has been extensively used in SDR (e.g., Li (1991), Cook and Weisberg (1991), Wang and Xia (2008)); for any given $y$, we consider the sample space score vector of $f(y)(x)$, $U_{f(y)}(x) = \nabla_x \log f(y)(x)$, form $\mathbb{J}_{X|Y=y} = \int U_{f(y)}(x)U_{f(y)}(x)^T f(y)(x)dx$, and take its expectation with respect to the marginal response density $f(y)$ to obtain

$$\mathbb{J}_{X|Y} = \int \mathbb{J}_{X|Y=y} f(y)dy. \quad (7)$$

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 $X$ is in fact the difference between the information on $X$ in the conditional density $f(y)(x)$ averaged over $y$, and the information on $X$ in the marginal density $f(x)$. 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 $C_X = \mathbb{J}_{X|Y} - \mathbb{J}_X$.

Interestingly, this highlights one way in which our approach differs from many others; existing inverse regression SDR methods use only $X \mid Y$, while we use both $X \mid Y$ and the marginal distribution of $X$. If we assume that the distribution of $X$ is spherically symmetric, then $\mathbb{J}_X$ is equidiagonal (Lindsay and Yao (2012)) and the eigen-decompositions of $C_X$ and $\mathbb{J}_{X|Y}$ are equivalent. However, if the marginal distribution of $X$ 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 (3), which holds if $X$ is elliptically contoured. SAVE relies on both the linearity and the constant conditional variance condition (4), which both hold if $X$ is Gaussian. CIM does not rely on such assumptions and accounts for the marginal of $X$, whichever its nature, through Proposition 3.4. Thus, our approach is fundamentally different and more robust.
3.4 The CIM When the Response is Discrete or Categorical

As noted near Equation (7), the CIM can easily be adapted to the case of a discrete or 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

$$J_X|Y = \sum_{j=1}^{m} \pi_j J_X|Y = y_j, \quad (8)$$

and use the eigen-decomposition of $C_X = J_X|Y - J_X$ to identify the projective covariates that are most informative in discriminating the $m$ sub-populations. In the Supplementary Material, we provide some insight 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 is used by most SDR methods based on inverse regression, e.g. SIR (Li, 1991), SAVE (Cook and Weisberg, 1991) and SR (Wang and Xia, 2008). The number of slices represents a tuning parameter.

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 $C_X$ and $\Sigma_X$. For the latter, it is natural to use the sample covariance matrix $\hat{\Sigma}_X$, whereas the main hurdle lies in estimating $C_X$. This, based on Proposition 3.4, can be turned into estimation of $J_X$ and $J_X|Y$.

**Estimation of $J_X$:** Because of the second equality in (1), we need to estimate $J_X = \int \frac{\nabla f(x) \cdot \nabla^T f(x)}{f(x)} dx$. 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 $J_X$ by $J_S$, the DIM for a surrogate $S$ with density $f_{(2)}(s) = \frac{f^2(s)}{\int f^2(x)dx}$. While very similar to $J_X$ in practice, $J_S$ has an explicit form when $f(x)$ is estimated using $\hat{f}_H(x) = \frac{1}{n|H|} \phi_p(x - x_i; 0, H^2)$, where $\phi_p(\cdot; 0, H^2)$ is the p-variate Gaussian density with mean $0$ and covariance $H^2$. Hui and Lindsay (2010) argued that $f_{(2)}$ preserves topological features of $f$, such as the locations of “peaks” and “valleys”. In fact the transformation increases estimation robustness by enhancing “peaks” and attenuating “valleys” of $f$, thus assigning less weights to outliers. An extended discussion of this surrogate approach, which also proved satisfactory in our simulation study, is provided in [Lindsay and Yao (2012)].

**Estimation of $J_{X|Y}$:** If $Y$ is discrete or categorical, we use (8), estimating $J_{X|Y=y_j}$ and $\pi_j$ for each $j = 1, 2, \ldots, m$. As $\hat{\pi}_j$ we take the sample proportion of observations with $Y = y_j$. To estimate $J_{X|Y=y_j}$, we use the f2 method of computation applied only to $X$ observations with $Y = y_j$. If $Y$ is continuous, we discretize $Y$ 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 $X$-observations. Thus, likewise other SDR methods that employ slicing (e.g. SIR [Li 1991] and SR [Wang and Xia 2008]), the CIM approach is not affected by response outliers and is more robust than SDR methods like MAVE [Xia et al., 2002] and PHD [Li 1992] which 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, CIM performs better with smaller \( L \) (thus more observations per slice) and \( L \) becomes less relevant as the sample size increases. In heteroscedastic settings, 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 of our applications. Another important tuning parameter is the bandwidth \( H \) of the Gaussian kernel density used in the \( f_2 \) method. Hui and Lindsay (2010) argued that, for sufficiently large sample sizes, the quest for informative projections using the \( f_2 \) method is not very sensitive to the choice of \( H \). They recommended the rule of thumb proposed in (Bowman and Foster, 1993), which sets \( H = (4/(p + 2))^{1/(p+4)} \Sigma X n^{-1/p+4} \). This is what we used in our implementation.

**Computational burden:** SDR methods like SIR, SAVE and PHD identify projective covariates through the eigen-decomposition of matrices and are computationally light. At the other end of the spectrum, methods like MAVE require nonparametric mean regression function estimation involving computationally heavy local linear smoothing. CIM is somewhere in between; it uses an eigen-decomposition, and nonparametric estimation of marginal and conditional (inverse regression) covariate densities. Note however, the density estimation in CIM is performed quite efficiently through the \( f_2 \) method. Table 2 (Section 5) shows that on simulated data CIM runs faster than MAVE and SR, but slower than SIR, SAVE and PHD.

### 4 Estimating the Structural Dimension \( d \)

Many methods have been proposed to estimate the dimension 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 (2013)], this 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 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. Here, 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 straightforwardly estimate $d$. The performance of this approach, along with that of the CIM as a means to estimate the CS, is documented 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 orthogonal. We also propose a complementary measure, the squared null trace correlation
\( R^2_o(S_1, S_2) = R^2(S_1^\perp, S_2^\perp) = \frac{1}{p-q} \text{tr}(Q_{S_1} Q_{S_2}), \) 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.** \( R^2_o(S_1, S_2) = 1 - \frac{q}{p-q}(1 - R^2(S_1, S_2)). \)

\( R^2(S_1, S_2) \) and \( R^2_o(S_1, S_2) \) are positively and linearly related, and we have \( R^2(S_1, S_2) = R^2_o(S_1, S_2) \) if \( q = \frac{p}{2} \). Proposition 4.1 also leads to two lower bounds; namely

\[
R^2(S_1, S_2) \geq \frac{q - (p - q)}{q} = \frac{2q - p}{q}, \quad R^2_o(S_1, S_2) \geq \frac{(p - q) - q}{p - q} = \frac{p - 2q}{p - q}.
\]

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 \( R^2(S_1, S_2) \) must go to 1 if \( q \) approaches \( p \) from the left, and \( R^2_o(S_1, S_2) \) must go to 1 if \( q \) approaches 0 from the right – proving a 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 \( S \) be a subspace of \( \mathbb{R}^p \) of dimension \( d \leq p \), and \( S_1 \) and \( S_2 \) be two subspaces of \( S \), each comprising \( q \leq d \) random directions within \( S \). Then \( E[R^2(S_1, S_2)] = \frac{q}{d} \).

The expected similarity increases linearly with \( q \), reaching 1 when \( q \) reaches \( d \) and \( S_1 = S_2 = S \). Using Proposition 4.1, we have \( E[R^2_o(S_1, S_2)] = 1 - \frac{q}{p-q}(1 - E[R^2(S_1, S_2)]) = 1 - \frac{q}{p-q} \frac{d-q}{d} \) 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 $\hat{S}_q$ as the span of the right eigenvectors of $(\tilde{J}_\mathbf{X}\mathbf{Y} - \tilde{J}_\mathbf{X})\hat{\Sigma}_\mathbf{X}$ with the $q$ largest eigen-values.

- Draw $j = 1, \ldots, B$ bootstrap samples; on each estimate $\hat{S}_q^{(j)}$, and compute $R_{q}^{2(j)} = R^2(\hat{S}_q, \hat{S}_q^{(j)})$, $R_{o,q}^{2(j)} = R^2(\hat{S}_q, \hat{S}_q^{(j)})$ 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 $\bar{R}_{q}^{2} R_{o,q}^{2}$.

$\bar{R}_{q}^{2}$, $\bar{R}_{o,q}^{2}$ and $\bar{R}_{q}^{2} R_{o,q}^{2}$ all measure stability of the estimation of $S_q$. As argued in Ye and Weiss (2003) and Zhu and Zeng (2006), when $q < d$ we estimate one among infinitely many subspaces within $S_{\mathbf{Y}|\mathbf{X}}$, resulting in low stability. $\bar{R}_{q}^{2}$, $\bar{R}_{o,q}^{2}$ and $\bar{R}_{q}^{2} R_{o,q}^{2}$ are small. When $q$ is close to $d$, stability increases and the three quantities grow; in particular, for $q = d$, $\hat{S}_q$ and $\hat{S}_q^{(j)}$, $j = 1, \ldots, B$, all estimate the same $S_{\mathbf{Y}|\mathbf{X}}$ 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 directions to $S_{\mathbf{Y}|\mathbf{X}}$. 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 $\bar{R}_{q}^{2} R_{o,q}^{2}$ 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 $\hat{d}$ as the dimension where $\bar{R}_{q}^{2} R_{o,q}^{2}$ 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 give 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 $(\hat{J}_X|Y - \hat{J}_X)\hat{\Sigma}_X$ 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 (e.g. Li (1992), Li et al. (2005), Zhu and Zeng (2006) and Zhu et al. (2010a)). We consider a $p = 10$ dimensional $X$ and three specifications for its distribution:

(a) *Independent*: $X \sim N_p(0_p, I_p)$,

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

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

Note that (a) and (b) satisfy both the linearity (3) and the constant conditional variance (4) 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 five model specifications, all with a \(d = 2\) dimensional CS:

1. \(Y = \cos(2\beta_1^T X) - \cos(\beta_2^T X) + \sigma \epsilon; \beta_1 = (1, 0, \ldots, 0)^T, \beta_2 = (0, 1, 0, \ldots, 0)^T.\)

2. \(Y = (\beta_1^T X)^2 + \beta_2^T X + \sigma \epsilon; \beta_1, \beta_2 \text{ as in (1)}.\)

3. \(Y = \beta_1^T X + (\beta_2^T X) \sigma \epsilon; \beta_1 = (1, 1, 1, 0, \ldots, 0)^T, \beta_2 = (0, \ldots, 0, 1, 1, 1, 1)^T.\)

4. \(Y = \beta_1^T X + 0.1 \beta_2^T X + (\beta_2^T X) \sigma \epsilon; \beta_1, \beta_2 \text{ as in (3)}.\)

5. \(Y = 3 \sin^2(\beta_1^T X/4) + (1 + (\beta_2^T X)^2) \sigma \epsilon; \beta_1 = (1, 1, 1, 0, \ldots, 0)^T, \beta_2 = (1, 0, 0, 0, 1, 3, 0, 0, 0, 0)^T.\)

In all cases \(\epsilon \mid X \text{ and } \epsilon \sim \mathcal{N}(0, 1).\) Due to space limitations, we report only results for the homoscedastic model (2) and the heteroscedastic model (5) in the main text (results for all other models are in the Supplementary Material). Note that \(\beta_1\) and \(\beta_2\) comprise more non-0 coordinates in model (5) than in (2). Also, such coordinates are mutually exclusive in model (2), whereas they share \(X_1\) in model (5). The CS (Central Subspace) and the CMS (Central Mean Subspace) coincide in model (2) while the CS has one more direction than the CMS in model (5), which methods targeting the CMS will miss. Finally, the mean function in model (2) has one \(Y\)-symmetric term out of two, and in model (5) the mean and variance functions are both \(Y\)-symmetric; these symmetries can hinder SIR.

For homoscedastic models (1) and (2), we define the signal-to-noise ratio as \(SNR(\sigma) = \frac{\text{Var}(E[Y|X])}{\sigma^2}\), where \(\sigma^2\) is the variance of the error term. For heteroscedastic models (3) - (5), instead of a typical signal-to-noise ratio, we use \(\sigma\) to “modulate” the ratio between signal in the mean component and signal in the variance component; we define \(SNR^\ast(\sigma) = \frac{\text{Var}(h_M(X))}{\text{Var}(h_V(X, \sigma))}\) (with an abuse of notation, we henceforth use \(SNR\)). For example, in model (5), \(h_M(X) = 3 \sin^2(\beta_1^T X/4)\) and \(h_V(X, \sigma) = (1 + (\beta_2^T X)^2) \sigma\). For homoscedastic models,
we consider $\sigma$'s generating SNR's around 10, 5, 2.5 and 1. For heteroscedastic ones, we considered SNR's 1 or less. Finally, for all combinations of covariate distributions, models and SNR's listed above, we consider two different sample sizes; \( n = 200 \) and 400.

In addition to our CIM approach, we run SIR (Li 1991), SAVE (Cook and Weisberg 1991), PHD (Li 1992), MAVE (Xia et al. 2002), and SR (Wang and Xia 2008). We also add to the comparison a benchmark, where instead of estimating the CS with a given method, we generate random subspaces of prescribed dimensions. SDR methods that require slicing, including ours, are run with different number of slices \( L \). Table 1 displays performance of all methods for the three covariate distributions, models (2) and (5), and both sample sizes, but only for selected SNR's and only for \( L = 5 \). Results for all scenarios considered are reported in the Supplementary Material. Entries in the Table represent trace correlations (the square-root of the quantity defined in Section 4.1) between estimated and true CS, averaged over 200 independently simulated data sets; the closer to 1, the better the performance. We also show dimension estimation plots (Section 4.2) for models (2) and (5). Plots for all scenarios, and alternative boxplot versions, are also provided in the Supplementary Material.

For model (2), all covariate distributions, both sample sizes and both SNR's, the best performer is MAVE. This is expected since CS and CMS coincide. But CIM's performance is very close to MAVE's, and better than all other methods except for the non-linear covariates cases where, somewhat surprisingly, SIR does slightly better than CIM. Also SR performs close to CIM in these cases. For the more complex model (5), SIR has very poor performance due to symmetries, and MAVE and PHD also do not do well because they target the CMS only. CIM and SAVE are clearly the best performer. Interestingly, SAVE has a slight edge at lower sample sizes, while CIM dominates at larger ones. More generally, while the performance of all methods increases with larger samples, the improvement is
particularly marked for CIM – likely due to improved kernel density estimation.

Figure 1 shows dimension estimation plots for CIM with $L = 5$ applied to model (2) ($\sigma = 0.55$, $SNR \approx 10$; panel (a)) and model (5) ($\sigma = 0.2$, ‘$SNR’ \approx 0.02$; panel (b)), Independent $X$ and $n = 400$. The dimensions are correctly identified; $\bar{R}_q^2$, $\bar{R}_{o,q}^2$, and $\bar{R}_q^2 \bar{R}_{o,q}^2$ have peaks at $q = 2$. Note how $\bar{R}_q^2 \bar{R}_{o,q}^2$ (dashed line) peaks only at $q = 2$, while $\bar{R}_q^2$ and $\bar{R}_{o,q}^2$ must grow again to approach 1 as $q$ goes towards $q = 10$ and $q = 0$ respectively. As expected, the “peaking” behavior of $\bar{R}_q^2 \bar{R}_{o,q}^2$ at $q = 2$ is more pronounced for larger SNRs and is maintained across varying $L$’s used in CIM (Supplementary Material; to give a sense of the variability associated with these curves, in the Supplement we also provide alternative diagnostics where we draw boxplots of $R_q^{(j)}$, $R_{o,q}^{(j)}$ and $R_q^{(j)} R_{o,q}^{(j)}$ obtained from 500 bootstrap samples).

In summary, CIM competes closely with the best SDR methods for homoscedastic models, and outperforms existing SDR methods for heteroscedastic models in most scenarios considered. Also, our dimension estimation plots appear effective in both homoscedastic and heteroscedastic scenarios. Finally, Table 2 reports computation time (in seconds) required by different methods to generate $d = 2$ output directions for models (2) and (5) using MATLAB 7.11.0.584 (R2010b) software on a laptop with 2.30GHz Intel(R) Core(TM) i5-2410M CPU. As expected, in terms of computational burden CIM is faster than methods like MAVE and SR (which involve nonparametric estimation of mean regression functions) and slower than methods like SIR, SAVE and PHD (which use only an eigen-decomposition, without nonparametric density estimation).

6 Applications to Real Data

In this section we apply the CIM approach and other SDR methods to two real datasets.
6.1 Wine Recognition Data

This UCI machine learning repository (https://archive.ics.uci.edu/ml/datasets/Wine) 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, we apply CIM with $L = 3$ slices (the response classes). The eigenvalues are $32.76, 9.23, 2.57, 1.73, 1.36, 1.00, 0.79, 0.44, 0.37, 0.36, 0.27, 0.17$, 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 $\bar{R}_q^2$ and $\bar{R}_q^2 \bar{R}^2_{o,q}$ occur at the transition from $q = 2$ to $q = 3$ (boxplot version in the Supplementary Material). The two leading CIM directions are dominated by Flavonoids, Color intensity and Proline:

\[
\hat{\beta}_1 = (0.03, 0.16, 0.15, 0.06, -0.19, -0.57, 0.02, -0.11, 0.50, -0.10, -0.31, -0.47)^T
\]
\[
\hat{\beta}_2 = (-0.24, -0.16, -0.19, -0.04, 0.05, 0.02, -0.02, -0.11, 0.14, -0.53, 0.14, 0.07, -0.74)^T.
\]

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. Interestingly, SIR and SR give results very similar to
CIM on this data (trace correlations between the CS estimates are $\sim 0.98$). 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 ppbm; parts per hundred million), 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 Supplementary Material). However, we follow Li (1992) and perform SDR without correcting for autocorrelation.

After standardizing the predictor vector, we apply CIM with $L = 5$. The eigenvalues are 3.83, 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 $\bar{R}^2_q R^2_{o,q}$ occur at the transition from $q = 2$ to $q = 3$ (boxplot version in the Supplementary Material).
The two leading CIM directions are:

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

The first is mainly driven by \textit{IBHT}, with substantial contributions by other covariates. The second is almost exclusively driven by \textit{VDHT}.

Figure 4 shows the association of the response with the two leading CIM directions. 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 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 $L = 3, 8, 10$ but decisively to $\hat{d} = 2$ for $L = 5$ (see Supplementary Material). MAVE and PHD also 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 slicing-based methods, the trace correlations between the CS estimated by CIM and those estimated by SIR, SAVE, SR, MAVE and PHD are, respectively 0.744, 0.972, 0.886, 0.724 and 0.766; CIM results are clearly closer to SAVE and SR than to SIR, MAVE and PHD. Trace correlations for first directions only are, respectively, 0.971, 0.997, 0.967, 0.953 and 0.843, suggesting that the first direction is correctly and similarly identified by all methods. Those for second directions only are, respectively, 0.317, 0.943, 0.734, 0.288 and 0.489, suggesting that SAVE and SR also catch the second direction found by CIM, but SIR, PHD and MAVE do not.
7 Concluding Remarks

In this article, we describe 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, carry independent regression information, and are naturally ordered based on their information contributions.

Some popular SDR methods leverage structure in the inverse regression \( X \mid Y \) and utilize simple eigen-decompositions – thus being computationally light. These methods though 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. CIM does not require the assumptions of the former; in fact, in addition to leveraging the structure in \( X \mid Y \), it explicitly accounts for structure in the distribution of \( X \). At the same time, CIM is computationally lighter than the latter; it does involve both an eigen-decomposition and kernel density (marginal and inverse) estimation for the covariate \( X \), but it uses the fast, explicit and robust f2 method of computation introduced in Hui and Lindsay (2010). We are currently investigating both computational burden and quality of CS estimation using the f2 method for much larger \( p \) than the ones in our current simulation study and real data applications.

Expanding on Ye and Weiss (2003), we propose a bootstrap-based diagnostic tool for estimating the dimension of the CS. We quantifying 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 use simulations and real data to show the competitive performance of CIM compared to other SDR methods, and the effectiveness
of our structural dimension diagnostics.

CIM involves two tuning parameters; 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 bandwidth, we rely on a rule of thumb recommended in Hui and Lindsay (2010). Regarding the number of slices, for sample sizes $n = 200 - 400$, our simulations suggest that the choice of $L$ between 3 and 5 usually works well. However, 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 and effects of both tuning parameters in CIM.

An 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 $w(x)$ (e.g. a kernel) can be used in place of the overall covariate density $f(x)$ in Equation (6) to define a local version of the matrix. Another practically relevant extension would concern very high dimensional, under-sampled data. Proposition 3.3 avoids inverting and taking the square root of $\hat{\Sigma}_X$. However, estimation quality deteriorates if $n$ is not large relative to $p$ – and is very poor if $p \gg n$ (Ledoit and Wolf (2004)). It would thus 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 have already been introduced for other SDR methods (e.g. Li (2007), Li and Yin (2008), Wang and Yin (2008), Li and Nachtsheim (2006), Wang and Zhu (2013)). 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. Designing a CIM-like approach to handle the reduction of
categorical covariates is another avenue for future research. We also plan to extend our performance comparisons to Directional Regression (Li and Wang, 2007) and semi-parametric SDR methods (Ma and Zhu, 2012, 2014).

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. A CRAN R package is in preparation.

References


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.


503–510.


reduction, projection pursuit, independent component analysis, and more. *Canadian


Figure 1: Dimension estimation plots for CIM with $L = 5$ slices (500 bootstrap replicates). (a) **Model (2)** with $\sigma = 0.55$ ($SNR \approx 10$) and (b) **Model (5)** with $\sigma = 0.2$ ($SNR \approx 0.02$). In both cases $X$ are Independent and $n = 400$.

Figure 2: Dimension estimation plots for CIM (500 bootstrap replicates). (a) **Wine Recognition Data** ($L = 3$) and (b) **Ozone Data** ($L = 5$).
Figure 3: 2D projections of the Wine Recognition Data on: (a) the CS estimated via CIM ($L = 3$) and (b) a Random plane.

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’.
Table 1: Mean (std. deviation) of trace correlation ($R$) in 200 repetitions for $n = 200$ and 400, Models (2) ($\sigma = 0.55$, SNR $\approx 10$ and $\sigma = 1.05$, SNR $\approx 2.75$) and (5) ($\sigma = 0.2$, ‘SNR’ $\approx 0.020 - 0.035$), and Independent (Ind.), Correlated (Corr.), and Non-linear (NonL) $X$. For slice-based methods, $L = 5$. Highest $R$ in each row is boldfaced.

<table>
<thead>
<tr>
<th>$\sigma$ (sample size)</th>
<th>X</th>
<th>SIR (5)</th>
<th>SAVE (5)</th>
<th>PHD</th>
<th>MAVE</th>
<th>SR (5)</th>
<th>CIM (5)</th>
<th>Random</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model (2)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$n = 200$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ind.</td>
<td>0.722 (0.044)</td>
<td>0.802 (0.091)</td>
<td>0.730 (0.045)</td>
<td><strong>0.995 (0.002)</strong></td>
<td>0.860 (0.132)</td>
<td>0.956 (0.021)</td>
<td>0.429 (0.120)</td>
<td></td>
</tr>
<tr>
<td>Corr.</td>
<td>0.766 (0.073)</td>
<td>0.765 (0.089)</td>
<td>0.716 (0.041)</td>
<td><strong>0.991 (0.004)</strong></td>
<td>0.901 (0.113)</td>
<td>0.930 (0.038)</td>
<td>0.424 (0.136)</td>
<td></td>
</tr>
<tr>
<td>NonL</td>
<td>0.788 (0.062)</td>
<td>0.692 (0.042)</td>
<td>0.690 (0.021)</td>
<td><strong>0.987 (0.029)</strong></td>
<td>0.763 (0.079)</td>
<td>0.753 (0.090)</td>
<td>0.437 (0.129)</td>
<td></td>
</tr>
<tr>
<td>$n = 400$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ind.</td>
<td>0.733 (0.048)</td>
<td>0.964 (0.029)</td>
<td>0.733 (0.042)</td>
<td><strong>0.998 (0.001)</strong></td>
<td>0.943 (0.108)</td>
<td>0.984 (0.006)</td>
<td>0.437 (0.128)</td>
<td></td>
</tr>
<tr>
<td>Corr.</td>
<td>0.818 (0.077)</td>
<td>0.926 (0.056)</td>
<td>0.732 (0.047)</td>
<td><strong>0.996 (0.002)</strong></td>
<td>0.980 (0.046)</td>
<td>0.977 (0.010)</td>
<td>0.439 (0.116)</td>
<td></td>
</tr>
<tr>
<td>NonL</td>
<td>0.837 (0.053)</td>
<td>0.701 (0.029)</td>
<td>0.698 (0.009)</td>
<td><strong>0.996 (0.002)</strong></td>
<td>0.780 (0.089)</td>
<td>0.785 (0.096)</td>
<td>0.423 (0.115)</td>
<td></td>
</tr>
<tr>
<td>Model (5)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$n = 200$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ind.</td>
<td>0.709 (0.043)</td>
<td>0.725 (0.058)</td>
<td>0.727 (0.049)</td>
<td><strong>0.979 (0.010)</strong></td>
<td>0.790 (0.120)</td>
<td>0.905 (0.057)</td>
<td>0.420 (0.120)</td>
<td></td>
</tr>
<tr>
<td>Corr.</td>
<td>0.720 (0.063)</td>
<td>0.693 (0.058)</td>
<td>0.719 (0.052)</td>
<td><strong>0.967 (0.019)</strong></td>
<td>0.827 (0.116)</td>
<td>0.855 (0.084)</td>
<td>0.436 (0.132)</td>
<td></td>
</tr>
<tr>
<td>NonL</td>
<td>0.752 (0.070)</td>
<td>0.674 (0.047)</td>
<td>0.681 (0.028)</td>
<td><strong>0.932 (0.080)</strong></td>
<td>0.728 (0.070)</td>
<td>0.720 (0.072)</td>
<td>0.423 (0.130)</td>
<td></td>
</tr>
<tr>
<td>$n = 400$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ind.</td>
<td>0.791 (0.080)</td>
<td>0.769 (0.077)</td>
<td>0.733 (0.044)</td>
<td><strong>0.985 (0.007)</strong></td>
<td>0.944 (0.078)</td>
<td>0.953 (0.019)</td>
<td>0.404 (0.128)</td>
<td></td>
</tr>
<tr>
<td>Corr.</td>
<td>0.808 (0.062)</td>
<td>0.693 (0.020)</td>
<td>0.695 (0.011)</td>
<td><strong>0.985 (0.012)</strong></td>
<td>0.772 (0.092)</td>
<td>0.747 (0.081)</td>
<td>0.430 (0.133)</td>
<td></td>
</tr>
<tr>
<td>Model (5)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$n = 200$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ind.</td>
<td>0.427 (0.130)</td>
<td>0.882 (0.073)</td>
<td>0.637 (0.118)</td>
<td>0.675 (0.086)</td>
<td>0.640 (0.188)</td>
<td>0.580 (0.078)</td>
<td>0.422 (0.134)</td>
<td></td>
</tr>
<tr>
<td>Corr.</td>
<td>0.351 (0.119)</td>
<td>0.804 (0.075)</td>
<td>0.529 (0.117)</td>
<td>0.617 (0.101)</td>
<td>0.605 (0.199)</td>
<td>0.796 (0.089)</td>
<td>0.432 (0.135)</td>
<td></td>
</tr>
<tr>
<td>NonL</td>
<td>0.595 (0.091)</td>
<td>0.728 (0.088)</td>
<td>0.489 (0.133)</td>
<td>0.649 (0.100)</td>
<td>0.730 (0.136)</td>
<td><strong>0.794 (0.094)</strong></td>
<td>0.441 (0.127)</td>
<td></td>
</tr>
<tr>
<td>$n = 400$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ind.</td>
<td>0.415 (0.134)</td>
<td>0.957 (0.030)</td>
<td>0.684 (0.095)</td>
<td>0.698 (0.084)</td>
<td>0.796 (0.134)</td>
<td><strong>0.957 (0.034)</strong></td>
<td>0.440 (0.140)</td>
<td></td>
</tr>
<tr>
<td>Corr.</td>
<td>0.355 (0.132)</td>
<td>0.918 (0.035)</td>
<td>0.556 (0.125)</td>
<td>0.626 (0.085)</td>
<td>0.783 (0.156)</td>
<td><strong>0.927 (0.029)</strong></td>
<td>0.426 (0.130)</td>
<td></td>
</tr>
<tr>
<td>NonL</td>
<td>0.648 (0.068)</td>
<td>0.787 (0.087)</td>
<td>0.501 (0.124)</td>
<td>0.665 (0.103)</td>
<td>0.867 (0.118)</td>
<td><strong>0.893 (0.070)</strong></td>
<td>0.418 (0.123)</td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Computation time in seconds (averaged over 200 runs) to generate $d = 2$ output directions for Models (2) and (5) using MATLAB 7.11.0.584 (R2010b) in a 2.30GHz Intel(R) Core(TM) i5-2410M CPU.

<table>
<thead>
<tr>
<th>Model (2): Ind. $X$, $n = 400$, $\sigma = 0.55$</th>
<th>SIR (5)</th>
<th>SAVE (5)</th>
<th>PHD</th>
<th>MAVE</th>
<th>SR (5)</th>
<th>CIM (5)</th>
<th>Random</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0031</td>
<td>0.0017</td>
<td>0.013</td>
<td>3.0079</td>
<td>2.7933</td>
<td>1.6681</td>
<td>0.0007</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model (5): Ind. $X$, $n = 400$, $\sigma = 0.2$</th>
<th>SIR (5)</th>
<th>SAVE (5)</th>
<th>PHD</th>
<th>MAVE</th>
<th>SR (5)</th>
<th>CIM (5)</th>
<th>Random</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0072</td>
<td>0.0025</td>
<td>0.0122</td>
<td>5.218</td>
<td>2.6707</td>
<td>1.6362</td>
<td>0.0006</td>
<td></td>
</tr>
</tbody>
</table>