2016

The Dual Central Subspaces in dimension reduction

Ross Iaci
Coll William & Mary, Dept Math, Williamsburg, VA 23185 USA

Xiangrong Yin
Univ Kentucky, Dept Stat, Lexington, KY 40536 USA;

Lixing Zhu
Hong Kong Baptist Univ, Dept Math, Kowloon Tong, Hong Kong, Peoples R China

Recommended Citation

Iaci, R., Yin, X., & Zhu, L. (2016). The dual central subspaces in dimension reduction. Journal of Multivariate Analysis, 145, 178-189.

The Dual Central Subspaces in Dimension Reduction

Ross Iaci,* Xiangrong Yin and Lixing Zhu

December 4, 2015

Abstract

Existing dimension reduction methods in multivariate analysis have focused on reducing sets of random vectors into equivalently sized dimensions, while methods in regression settings have focused mainly on decreasing the dimension of the predictor variables. However, for problems involving a multivariate response, reducing the dimension of the response vector is also desirable and important. In this paper, we develop a new concept, termed the Dual Central Subspaces (DCS), to produce a method for simultaneously reducing the dimensions of two sets of random vectors, irrespective of the labels predictor and response. Different from previous methods based on extensions of Canonical Correlation Analysis (CCA), the recovery of this subspace provides a new research direction for multivariate sufficient dimension reduction. A particular model-free approach is detailed theoretically and the performance investigated through simulation and a real data analysis.

Key Words and Phrases: Canonical Correlation Analysis; Dimension reduction; Dual Central Subspaces; Multivariate analysis; Visualization.

1 Introduction

Methods for dimension reduction in multivariate association studies for two sets of random vectors generally focus on reducing the dimensions of both sets of variables, where the role of predictor and response is unimportant, while multivariate regression centers on the dimension reduction of the vector labeled the predictor variables.

A popular method pioneered by Hotelling (1936) for the pairwise extraction of the significant relationships that exist between two random vectors is Canonical Correlation Analysis (CCA). Kettenring (1971, 1985) investigated five measures, extending Hotellings (1936) theory to multiple sets, while Van der Burg & De Leeuw (1983) developed a method termed nonlinear canonical correlation analysis. More recently, many methods advancing this area of research have been proposed, see for example Yin (2004), Yin & Sriram (2008), Iaci et

*Ross Iaci, Department of Mathematics, The College of William and Mary, Williamsburg, VA, 23185. E-mail: riaci@wm.edu. Xiangrong Yin, Department of Statistics, University of Kentucky, Lexington, KY, 40536. Email: yinxiangrong@gmail.com. Lixing Zhu, Department of Mathematics, Hong Kong Baptist University, Kowloon Tong, Hong Kong, China. Email: lzhu@hkbu.hk.

al. (2008) and Iaci et al. (2010) and references therein. Importantly, all of these methods require that the number of coefficient vectors that provide the dimension reduction be equal. While this restriction simplifies the problem, if the number of coefficient vectors that recover the true associations between the random vectors are not equal then this could result in a critical loss of information. Therefore, methods that allow the number of coefficient vectors to be different and thus, provide a sufficient dimension reduction, are crucial in multivariate analysis.

To this end, we introduce the Dual Central Subspaces (DCS), and subsequently provide a new method to estimate these subspaces, which provides a simultaneous sufficient dimension reduction of two multivariate random vectors. That is, our approach provides a dimension reduction of both vectors without requiring the dimensions of the reduction to be equal. To identify the DCS, we consider a higher-order information measure based on the Kullback-Leibler (KL) divergence, rather than extending traditional methods for estimating the Central Subspaces (CS) that recover information from lower moments, such as SIR and SAVE. The KL index was introduced in Iaci et al. (2008) to provide a measure of overall association between random vectors, the main focus of their paper; a more detailed discussion of the differences between the use of the index in both papers is provided below. An advantage, and motivation, for using this information based measure is that it is able to detect both linear and nonlinear relationships that exist between random vectors, which enables a more complete recovery of the DCS while treating both vectors equivalently. Moreover, in using this method no distributional assumptions, except for the existence of the joint density, are required and the estimation of the DCS becomes an optimization problem. The method is directly applicable for random vectors labeled as predictor and response and thus, also provide a powerful tool for dimension reduction in a multivariate regression setting.

Since Li's sliced inverse regression (1991) method, there have been many statistical studies that have focused on dimension reduction in a regression setting. For example, see Cook & Weisberg (SAVE, 1991), Li (pHd, 1992), Yin & Cook (Cov_k, 2002), Xia et al (MAVE, 2002), the seminal papers of Ma & Zhu (2012, 2013a,b) and for a detailed review see Cook (1998b) or Cook & Weisberg (1999). All of these methods consider only a univariate response and thus, dimension reduction is performed only on the predictor variables. A few methods have been developed in a multivariate regression setting, but the dimension reduction is focused only on the predictors; see for example Cook & Setodji (2003), Yin & Bura (2006) and Li et al. (2008). Methods for sufficient dimension reduction, especially with a multivariate response, for example Zhu et al. (2010) and Setodji & Cook (2004), could also be considered to develop a method to identify the DCS, but prefer the flexibility of the information based procedure in this initial work. More recently, Cook et al. (2010) developed an envelope model for multivariate linear regression that not only reduces the dimension of the predictors, but also the noninformative responses in order to obtain a more efficient estimator. While their method and those of others, such as Su & Cook (2011, 2012, 2013), Cook et al. (2013) and Cook & Su (2013), have made significant advances in this area, the focus of these techniques are only on the regression mean function for a specified regression model. The proposed method of Li (2003) for achieving a dimension reduction in a multivariate response regression setting could be considered for developing a method for the identification of the DCS, however the linearity conditions and the exhaustive nature of recovering all the directions using this SIR based method are viewed to be somewhat restrictive. Importantly, procedures based on spectral decompositions, for example SIR and SAVE, and moment based methods in general, have been shown to perform poorly, even under strong conditions like normality, when nonlinear relationships exist between the responses and predictors. To investigate this in the context of estimating the DCS, and further motivate our use of the KL information based method, we use an alternating search procedure to estimate the DCS using the projective resampling SIR procedure of Li et al. (2008) and compare the performance to our method in simulation.

The article is organized as follows. In Section 2.1 we introduce the concept of the DCS, discuss the theoretical properties, and its role in providing a new method for multivariate sufficient dimension reduction. Identification of the DCS and computational aspects of our approach are described in Section 2.2. Simulation studies are performed in Section 3 and, in Section 4, we revisit the Los Angeles County dataset that was initially investigated in Shumway et al. (1988) to gain further insight into the associations that exist between mortality and environmental conditions using our method. Proofs of the presented results and the projective resampling SIR study are provided in the Appendix.

2 Methodology

2.1 The Dual Central Subspaces

In this section, we define the Dual Central Subspaces (DCS) to reduce the dimensions of two sets of variables sufficiently and discuss the relevant properties. Even though contextually each vector may be regarded as the response or predictor, the labeling of the vectors as predictor and response is used only for the convenient exposition of the method. Importantly, this novel concept allows the size of the dimension for which the reduction occurs to vary for each random vector.

Let $\mathcal{S}$ denote a generic subspace, $\mathcal{S}(\mathbf{A}_r)$ represent the r-dimensional subspace in $\mathbf{R}^p$ spanned by the columns of a $p \times r$ full rank matrix $\mathbf{A}$ and finally, let $P_{\mathcal{S}}$ designate the projection onto $\mathcal{S}$ with respect to the usual inner product. Consider two sets of random variables, a $p \times 1$ vector $\mathbf{X}$ and a $q \times 1$ vector $\mathbf{Y}$, the Dimension Reduction Subspace (DRS) for reducing the dimension of $\mathbf{X}$ is defined as the subspace $\mathcal{S}$ such that

$$\mathbf{Y} \perp\!\!\!\perp \mathbf{X} | P_{\mathcal{S}}\mathbf{X}. \tag{1}$$

Here, the notation means that $\mathbf{Y}$ is independent of $\mathbf{X}$ given $P_{\mathcal{S}}\mathbf{X}$, the projection of $\mathbf{X}$ onto the subspace $\mathcal{S}$. The Central Subspace (CS), denoted $\mathcal{S}_{\mathbf{Y}|\mathbf{X}}$, is defined as the intersection of all DRSs, which importantly is also a DRS. Note that, when $q = 1$ and $\mathbf{Y}$ is considered the response, this is equivalent to the CS defined in Cook (1994, 1996, 1998b).

In a multivariate dimension reduction CCA context, it is also necessary to reduce the dimension of $\mathbf{Y}$ sufficiently. To this end, we define the CS of $\mathbf{Y}$, denoted $\mathcal{S}_{\mathbf{X}|\mathbf{Y}}$, by simply interchanging the roles of $\mathbf{X}$ and $\mathbf{Y}$ in the above definition. That is, we define the DRS for the dimension reduction of $\mathbf{Y}$ as the subspace $\mathcal{S}$ such that $\mathbf{X} \perp\!\!\!\perp \mathbf{Y} | P_{\mathcal{S}}\mathbf{Y}$. Again, $P_{\mathcal{S}}$ is the usual projection onto the subspace $\mathcal{S}$ and the CS is defined as the intersection of all

DRSs. Thus, similar to the role of the CS associated with $\mathbf{X}$, the CS of $\mathbf{Y}$ will also play an important role in dimension reduction, and with this subspace the information from $\mathbf{Y}$ can be preserved.

In the sense of reducing the dimensions of both $\mathbf{X}$ and $\mathbf{Y}$, the two sets of variables can be treated equally and thus, we term the subspaces, $\mathcal{S}_{\mathbf{Y}|\mathbf{X}}$ and $\mathcal{S}_{\mathbf{X}|\mathbf{Y}}$, the Dual Central Subspaces (DCS). The dimensions of the respective Central Subspaces are denoted d_x and d_y. In multivariate association studies the roles of predictor and response are interchangeable and thus, recovering the DCS provides a powerful tool for studying the relationships between two multivariate random vectors. Importantly, it is not necessary that the reduced dimensions be paired, $d_x = d_y$, as in standard multivariate association methodologies. Therefore, our definition differs significantly from the usual canonical approaches to study the multivariate association between two random vectors, such as CCA and its extensions. Moreover, requiring paired dimensional subspaces can be quite limiting as illustrated in the following simple example.

Illustrative example 1: Consider the two random vectors $\mathbf{X} = (X_1, X_2)^\top$ and $\mathbf{Y} = (Y_1, Y_2, Y_3)^\top$, where X_1, X_2 and $Y_3 \sim \mathcal{N}(0,1)$. Next, suppose that $Y_1 = (X_1 + X_2)^2 + \epsilon_1$ and $Y_2 = X_1 + X_2 + \epsilon_2$, with error terms $\epsilon_i \sim \mathcal{N}(0,1)$, $i = 1, 2$.

Here, $\mathcal{S}_{\mathbf{Y}|\mathbf{X}} = \text{Span}\{(1,1)^\top\}$ with $d_x = 1$, but $\mathcal{S}_{\mathbf{X}|\mathbf{Y}} = \text{Span}(\mathbf{e}_1, \mathbf{e}_2)$ with $d_y = 2$, where $\mathbf{e}_1 = (1,0,0)^\top$ and $\mathbf{e}_2 = (0,1,0)^\top$. Therefore, any method used to recover the DCS that requires that the reduced dimensions be equal will fail to recover one dimension of $\mathcal{S}_{\mathbf{X}|\mathbf{Y}}$ if $d_x = d_y = 1$. Alternatively, if the dimensions are taken to be $d_x = d_y = 2$, then the dimension of $\mathcal{S}_{\mathbf{Y}|\mathbf{X}}$ will be overestimated and thereby, fail to provide a minimum sufficient dimension reduction.

All of the properties for the Central Subspace in a univariate response regression setting, as in Cook (1998b) for example, hold for the DCS. Moreover, the existence of the DCS can be directly established from Cook (1998b) and Yin et al. (2008) and therefore, assume the existence of this subspace hereafter. Additional properties of the DCS are provided in the following proposition.

Proposition 1 *Let* $\mathbf{A}$ *and* $\mathbf{B}$ *be the bases for* $\mathcal{S}_{\mathbf{Y}|\mathbf{X}}$ *and* $\mathcal{S}_{\mathbf{X}|\mathbf{Y}}$*, respectively. Then, the following three conditions are equivalent:*

(i) $\mathbf{Y} \perp\!\!\!\perp \mathbf{X} | \mathbf{A}^\top \mathbf{X}$ *and* $\mathbf{Y} \perp\!\!\!\perp \mathbf{X} | \mathbf{B}^\top \mathbf{Y}$.

(ii) $\mathbf{Y} \perp\!\!\!\perp \mathbf{X} | \mathbf{A}^\top \mathbf{X}$ *and* $\mathbf{Y} \perp\!\!\!\perp \mathbf{A}^T \mathbf{X} | \mathbf{B}^\top \mathbf{Y}$.

(iii) $\mathbf{B}^\top \mathbf{Y} \perp\!\!\!\perp \mathbf{X} | \mathbf{A}^\top \mathbf{X}$ *and* $\mathbf{Y} \perp\!\!\!\perp \mathbf{X} | \mathbf{B}^\top \mathbf{Y}$.

Corollary 1 $d_x = 0$ *if and only if* $d_y = 0$.

The Proof of Proposition 1 is given in Appendix A.1, while Corollary 1 follows directly by definition. Proposition 1 suggests that methods for dimension reduction can be developed using a two-stage alternating search procedure. That is, first $\mathbf{Y}$ is considered the response and the dimension of $\mathbf{X}$ is reduced. Next, the recovered reduced predictor $\mathbf{A}^\top \mathbf{X}$ can be regarded as the response and the dimension of $\mathbf{Y}$ reduced to identify the transformation $\mathbf{B}^\top \mathbf{Y}$. This alternating search can also be done by initially regarding $\mathbf{X}$ as the response.

Therefore, treating either $\mathbf{Y}$ or $\mathbf{X}$ as the response vector, many of the dimension reduction methods developed in a multivariate regression setting, such as those in Cook & Setodji (2003), Yin & Bura (2006) and Li et al. (2008), could be directly applied in each alternating search. It is likely that such a procedure using these moment based methods would, under strong conditions such as normality, be successful in recovering the spaces of the DCS that correspond to linear associations, but would have difficulty in doing so for those corresponding to nonlinear relationships. Motivated by this, in the next section we propose a new approach to identify the DCS using the Kullback-Leibler (KL) divergence, which treats $\mathbf{Y}$ and $\mathbf{X}$ equivalently and has been shown in Iaci et al. (2008), and references therein, to effectively recover both linear and nonlinear relationships when the dimensions of the reduction are equal, $d_x = d_y$. Different from existing research directions, which require the dimensions to be equal, this novel concept of the DCS emphasizes sufficient dimension reduction, allowing the reduction dimensions to be unequal, which could lead to a new research direction in the study of multivariate association and sufficient dimension reduction.

2.2 Identification of the DCS

Consider the random vectors $\mathbf{X}_{p\times 1}$ and $\mathbf{Y}_{q\times 1}$ and the matrices $\mathbf{A} = \mathbf{A}_{p\times d_x}$ and $\mathbf{B} = \mathbf{B}_{q\times d_y}$ with full ranks d_x and d_y, respectively. Next, let $f\left(\mathbf{A}^\top\mathbf{X}, \mathbf{B}^\top\mathbf{Y}\right)$, $f\left(\mathbf{A}^\top\mathbf{X}\right)$ and $f\left(\mathbf{B}^\top\mathbf{Y}\right)$ denote the joint and marginal densities of the linear transformations $\mathbf{A}^\top\mathbf{X}$ and $\mathbf{B}^\top\mathbf{Y}$. To enable the recovery of the DCS, we consider the KL divergence between the joint and the product of the marginal densities and define the index

$$\mathbf{D}\left(\mathbf{A}, \mathbf{B}\right) = \mathrm{D}_{KL}\big\{f(\mathbf{A}^\top\mathbf{X}, \mathbf{B}^\top\mathbf{Y}) || f(\mathbf{A}^\top\mathbf{X}) f(\mathbf{B}^\top\mathbf{Y})\big\} = \mathrm{E}\left\{\ln\left(\frac{f\left(\mathbf{A}^\top\mathbf{X}, \mathbf{B}^\top\mathbf{Y}\right)}{f\left(\mathbf{A}^\top\mathbf{X}\right) f\left(\mathbf{B}^\top\mathbf{Y}\right)}\right)\right\}, \quad (2)$$

where the expectation is taken with respect to the joint density of $(\mathbf{A}^\top\mathbf{X}, \mathbf{B}^\top\mathbf{Y})$. Also, by definition $\mathbf{D}(\mathbf{A}, \mathbf{B}) \geq 0$, since $f(\mathbf{A}^\top\mathbf{X}, \mathbf{B}^\top\mathbf{Y})$ and $f(\mathbf{A}^\top\mathbf{X})f(\mathbf{B}^\top\mathbf{Y})$ are both density functions; Kullback (1959). Note that, the direct connection of $\mathbf{D}(\mathbf{A}, \mathbf{B})$ to mutual information and thus, marginal and conditional entropy, yields the equivalent forms:

$$\mathbf{D}(\mathbf{A}, \mathbf{B}) = \mathrm{E}_{\mathbf{B}^\top\mathbf{Y}}\left[\mathrm{D}_{KL}\big\{f(\mathbf{A}^\top\mathbf{X}|\mathbf{B}^\top\mathbf{Y}) || f(\mathbf{A}^\top\mathbf{X})\big\}\right] \equiv \mathrm{E}_{\mathbf{A}^\top\mathbf{X}}\left[\mathrm{D}_{KL}\big\{f(\mathbf{B}^\top\mathbf{Y}|\mathbf{A}^\top\mathbf{X}) || f(\mathbf{B}^\top\mathbf{Y})\big\}\right];$$

see Appendix A.3. The index $\mathbf{D}(\mathbf{A}, \mathbf{B})$ can be thought of as a measure of the amount of information lost in projecting the random vectors $\mathbf{X}$ and $\mathbf{Y}$ into subspaces of sizes $d_x \leq p$ and $d_y \leq q$. Importantly, if no information is lost through the projection, that is $\mathbf{D}(\mathbf{A}, \mathbf{B}) = \mathbf{D}(\mathbf{I}_{p\times p}, \mathbf{I}_{q\times q})$, then $\mathbf{Y} \perp\!\!\!\perp \mathbf{X}|\mathbf{A}^\top\mathbf{X}$ and $\mathbf{X} \perp\!\!\!\perp \mathbf{Y}|\mathbf{B}^\top\mathbf{Y}$. This and other properties of the index are provided in the following proposition.

Proposition 2 *Consider the random vectors $\mathbf{X}_{p\times 1}$ and $\mathbf{Y}_{p\times 1}$ and let $\mathcal{S}(\mathbf{A})$ and $\mathcal{S}(\mathbf{B})$ denote the subspaces spanned by the columns of $\mathbf{A}_{p\times k}$ and $\mathbf{B}_{q\times l}$, $k \leq p$ and $l \leq q$, respectively.Then, the following hold:*

(i) If $\mathcal{S}(\mathbf{A}_1) \subseteq \mathcal{S}(\mathbf{A})$ and $\mathcal{S}(\mathbf{B}_1) \subseteq \mathcal{S}(\mathbf{B})$, then $\mathbf{D}(\mathbf{A}_1, \mathbf{B}_1) \leq \mathbf{D}(\mathbf{A}, \mathbf{B})$.

(ii) $\mathbf{D}(\mathbf{A}, \mathbf{B}) = \mathbf{D}(\mathbf{I_{p\times p}}, \mathbf{I_{q\times q}})$ if and only if $\mathbf{Y} \perp\!\!\!\perp \mathbf{X}|\mathbf{A}^\top\mathbf{X}$ and $\mathbf{Y} \perp\!\!\!\perp \mathbf{X}|\mathbf{B}^\top\mathbf{Y}$.

Proofs of Proposition 2 are given in Appendix A.2.

Part (*ii*) of Proposition 2, the motivation for using this index in the context of recovering the DCS to provide a sufficient dimension reduction, suggests that $\mathcal{S}_{\mathbf{X}|\mathbf{Y}}$ and $\mathcal{S}_{\mathbf{Y}|\mathbf{X}}$ can be found by finding the linear transformations $\mathbf{A}^\top\mathbf{X}$ and $\mathbf{B}^\top\mathbf{Y}$ that maximize $\mathbf{D}(\mathbf{A},\mathbf{B})$. That is, the DCS of the random vectors $\mathbf{X}$ and $\mathbf{Y}$ can be recovered by searching iteratively for the coefficient matrices $\mathbf{A}$ and $\mathbf{B}$ such that $\mathbf{A}^\top\mathbf{X}$ and $\mathbf{B}^\top\mathbf{Y}$ extract the largest amount of information by maximizing the KL index in (2), subject to the constraints $\mathbf{A}^\top\Sigma_{\mathbf{X}}\mathbf{A} = \mathbf{I}_{d_x\times d_x}$ and $\mathbf{B}^\top\Sigma_{\mathbf{Y}}\mathbf{B} = \mathbf{I}_{d_y\times d_y}$. These are the usual CCA constraints commonly used in multivariate dimension reduction methodologies. Importantly, the index in (2) is invariant under nonsingular transformations of the vectors $\mathbf{X}$ and $\mathbf{Y}$; see Appendix A.4. Therefore, we can simplify the constraints through the transformations $\mathbf{Z_X} = \Sigma_{\mathbf{X}}^{-1/2}\{\mathbf{X} - \mathrm{E}(\mathbf{X})\}$ and $\mathbf{Z_Y} = \Sigma_{\mathbf{Y}}^{-1/2}\{\mathbf{Y} - \mathrm{E}(\mathbf{Y})\}$, which changes the scale, but not the relationships that exist between the original vectors. In this transformed scale, termed the whitened scale, the constraints are reduced to $\mathbf{A}_{\mathbf{z}}^\top\mathbf{A}_{\mathbf{z}} = \mathbf{I}_{d_x\times d_x}$ and $\mathbf{B}_{\mathbf{z}}^\top\mathbf{B}_{\mathbf{z}} = \mathbf{I}_{d_y\times d_y}$. Transforming the random vectors to have identity dispersion matrices not only eases computation, but also rescales the variables to have equivalent magnitudes, which aids in the interpretation of the loadings of the individual vectors of the coefficient matrices. If $\mathbf{A}_{\mathbf{z}}$ and $\mathbf{B}_{\mathbf{z}}$ are the coefficient matrices in the transformed scale, then the coefficient matrices in the original scale are easily recovered through the transformations $\mathbf{A} = \Sigma_{\mathbf{X}}^{1/2}\mathbf{A}_{\mathbf{z}}$ and $\mathbf{B} = \Sigma_{\mathbf{Y}}^{1/2}\mathbf{B}_{\mathbf{z}}$.

The index in (2) was also proposed in Iaci et al. (2008), not with a focus on dimension reduction, but rather on developing a measure of overall association between multiple sets of random vectors. To this end, the authors noted that, here in the context of two random vectors, if both coefficient matrices, $\mathbf{A}_{p\times p}$ and $\mathbf{B}_{q\times q}$, are nonsingular then $\mathbf{D}(\mathbf{A},\mathbf{B})$ recovers the full amount of information between the vectors; note that, it is not necessary that the coefficient matrices in part (ii) of Proposition 2 be invertible. Next, Proposition 3 of Iaci et al. (2008) showed that $\mathbf{D}(\mathbf{AC}_1,\mathbf{BC}_2) = \mathbf{D}(\mathbf{A},\mathbf{B})$, when $\mathbf{C}_1$ and $\mathbf{C}_2$ are both full rank matrices. Finally, letting $\mathbf{C}_1 = \mathbf{A}^{-1}$ and $\mathbf{C}_2 = \mathbf{B}^{-1}$ so that $\mathbf{D}(\mathbf{A},\mathbf{B}) = \mathbf{D}(\mathbf{AC}_1,\mathbf{BC}_2) = \mathbf{D}(\mathbf{I_{p\times p}},\mathbf{I_{q\times q}})$, the authors used the last index in the equality as an overall measure of association, which importantly, in practice does not require matrix maximization for estimation. A permutation based method was developed to test the null hypothesis that the vectors were independent, $\mathbf{D}(\mathbf{I_{p\times p}},\mathbf{I_{q\times q}}) = 0$, and if rejected, dimension reduction was performed to extract the relationships between the vectors. However, one-dimensional coefficient vectors were estimated successively, as in CCA, to recover the existent relationships, requiring the final dimension of the reduction to be equal, $d_x = d_y$, which would fail to recover relationships when $d_x \neq d_y$ and thereby, fail to provide a sufficient dimension reduction. Note that, the equally dimensioned reduction methods developed in Iaci et al. (2008) can be viewed as an extension of those of Yin (2004) and Yin & Sriram (2008) to multiple sets and groups of multiple sets, respectively.

For comparison, we also applied the projective resampling SIR method of Li et al. (2008). To recover the DCS, $\mathcal{S}_{\mathbf{Y}|\mathbf{X}}$ and $\mathcal{S}_{\mathbf{X}|\mathbf{Y}}$, we apply their method with $\mathbf{Y}$ considered the response and $\mathbf{X}$ the predictor to determine the coefficient matrix $\mathbf{A}$ and then, simply interchange the role of predictor and response to identify $\mathbf{B}$. Note that, SIR requires that the predictors satisfy the linear conditional mean (LCM) condition: $\mathrm{E}(c^\top\mathbf{X}|\mathbf{M}^T\mathbf{X})$ is linear in $\mathbf{M}^\top\mathbf{X}$ for all $c \in \mathbf{R}^p$. For example, the LCM condition is satisfied when the distribution of $\mathbf{X}$ is elliptically contoured. When the LCM condition is violated, SIR is still expected to work well when the relationships between the predictors and response are linear. SIR is also known to fail when symmetric dependencies exist. The projective resampling SIR based method is applied to both simulations of Study 1 in Section 3, where the distributions of the random vectors are not multivariate normal and contain both linear and nonlinear relationships between the vectors. As expected, the results given in Appendix A.5 show that the method is not comparable to our method in recovering the spaces of the DCS when the LCM condition is likely violated and when symmetric relationships exist. This further supports the use of the index in (2) and will be the focus hereafter.

2.3 Estimation of the DCS

A method for estimating the dimensions of the DCS is given in Section 2.4, but here we assume that the dimensions d_x and d_y are known. Let $\{(\mathbf{x}_j, \mathbf{y}_j), j = 1, \ldots, n\}$ denote a random sample from $(\mathbf{X}_{p\times 1}, \mathbf{Y}_{q\times 1})$, then the sample estimate of the index in (2) is given by

$$\widehat{\mathbf{D}}\left(\mathbf{A}, \mathbf{B}\right) = \frac{1}{n}\sum_{j=1}^{n} \ln\left(\frac{\widehat{f}\left(\mathbf{A}^\top \mathbf{x}_j, \mathbf{B}^\top \mathbf{y}_j\right)}{\widehat{f}\left(\mathbf{A}^\top \mathbf{x}_j\right)\widehat{f}\left(\mathbf{B}^\top \mathbf{y}_j\right)}\right), \tag{3}$$

where $\widehat{f}\left(\mathbf{A}^\top \mathbf{x}_j, \mathbf{B}^\top \mathbf{y}_j\right)$, $\widehat{f}\left(\mathbf{A}^\top \mathbf{x}_j\right)$ and $\widehat{f}\left(\mathbf{B}^\top \mathbf{y}_j\right)$ are the kernel density estimates of $f\left(\mathbf{A}^\top \mathbf{X}, \mathbf{B}^\top \mathbf{Y}\right)$, $f\left(\mathbf{A}^\top \mathbf{X}\right)$ and $f\left(\mathbf{B}^\top \mathbf{Y}\right)$, respectively. Specifically, for a given set of coefficient matrices $\mathbf{A} = \left[\mathbf{a}_1, \ldots, \mathbf{a}_{d_x}\right]$ and $\mathbf{B} = \left[\mathbf{b}_1, \ldots, \mathbf{b}_{d_y}\right]$ we use the Gaussian product kernel density estimate,

$$\widehat{f}_n\left(\mathbf{A}^\top \mathbf{x}_i, \mathbf{B}^\top \mathbf{y}_i\right) = \frac{1}{n\prod_{k=1}^{d_x} h_k \prod_{l=1}^{d_y} h_l}\sum_{j=1}^{n}\left(\prod_{k=1}^{d_x} K\left[\left\{\mathbf{a}_k^\top(\mathbf{x}_j - \mathbf{x}_i)\right\}/h_k\right]\prod_{l=1}^{d_y} K\left[\left\{\mathbf{b}_l^\top(\mathbf{y}_j - \mathbf{y}_i)\right\}/h_l\right]\right),$$

where the bandwidth $h_k = \{4/(d_x+2)\}^{1/(d_x+4)}\, s_k n^{-1/(d_x+4)}, k \in \{1, \ldots, d_x\}$, and s_k is the sample standard deviation of $\left\{\mathbf{a}_k^\top \mathbf{x}_i^{(k)}, i = 1, \ldots, n\right\}$. The bandwidth h_l is determined analogously based on the sample observations of $\mathbf{Y}$. The use of Gaussian product kernels for density estimation was suggested by Scott (1992) and Silverman (1986) and were shown to work well in Iaci et al. (2008). The selection of the bandwidth was initially motivated by the results of Yin (2004) and further supported by the results in Iaci & Sriram (2013). The estimates of the matrices that form the bases of the DCS, $\widehat{\mathbf{A}}$ and $\widehat{\mathbf{B}}$, are the solutions of

$$(\widehat{\mathbf{A}}, \widehat{\mathbf{B}}) = \operatorname{argmax}_{\mathbf{A},\mathbf{B}}\ \widehat{\mathbf{D}}\left(\mathbf{A}, \mathbf{B}\right), \tag{4}$$

subject to the sample versions of the population constraints, $\widehat{\mathbf{A}}^\top \widehat{\Sigma}_{\mathbf{X}} \widehat{\mathbf{A}} = \mathbf{I}_{d_x \times d_x}$ and $\widehat{\mathbf{B}}^\top \widehat{\Sigma}_{\mathbf{Y}} \widehat{\mathbf{B}} = \mathbf{I}_{d_y \times d_y}$. Here, $\widehat{\Sigma}_{\mathbf{X}}$ and $\widehat{\Sigma}_{\mathbf{Y}}$ are the sample covariance matrices for $\mathbf{X}$ and $\mathbf{Y}$, respectively. Although matrix maximization was not necessary to provide an overall measure of association in Iaci et al. (2008), the consistency proof was generalized to coefficient matrices with unequal column dimension and so, by Theorem 1 in their paper $(\widehat{\mathbf{A}}, \widehat{\mathbf{B}}) \overset{w.p.1}{\rightarrow} (\mathbf{A}, \mathbf{B})$ as $n \rightarrow$

∞. Due to the invariance property of $\mathbf{D}(\mathbf{A}, \mathbf{B})$, we work in the whitened scale using the sample versions corresponding to the vectors $\mathbf{Z_X}$ and $\mathbf{Z_Y}$. However, for ease in exposition the notation $\mathbf{X}$ and $\mathbf{Y}$ is maintained throughout this section. Motivated by Proposition 2 part (*ii*), we develop an alternating iterative search procedure for estimating the coefficient matrices, which is different from the simultaneous search procedure proposed in Iaci et al. (2008). The algorithm is detailed, with bullets after each step providing additional comments and details pertaining to that step, as follows:

Step 0: Set $l = 0$ and generate an initial guess of the $(p \times d_x)$ and $(q \times d_y)$ coefficient matrices $\widehat{\mathbf{A}}_0$ and $\widehat{\mathbf{B}}_0$, respectively.

- Initial guesses were generated in two different ways. First, N_1 orthogonal matrices in the positive direction, consisting of zeros and ones, are generated at random. Next, additional N_2 orthogonal matrices are randomly generated. The initial guesses $\mathbf{A}_0$ and $\mathbf{B}_0$ are taken to be the pair of these matrices that generate the largest sample information index $\widehat{\mathbf{D}}(\mathbf{A}, \mathbf{B})$ among the $N_1 + N_2$ random matrices. In the simulations $N_1 = N_2 = 50$, or 75, worked well. While many different methods for generating the initial guess were investigated, in general the above hybrid method provided the most consistent results.

Step 1: Hold the matrix $\widehat{\mathbf{B}}_l$ constant and determine $\widehat{\mathbf{A}}_{l+1}$ such that the sample index is maximized. That is, $\widehat{\mathbf{A}}_{l+1} = \text{argmax}_{\mathbf{A}}\ \widehat{\mathbf{D}}(\mathbf{A}, \widehat{\mathbf{B}}_l)$. Next, $\widehat{\mathbf{A}}_{l+1}$ is held constant and $\widehat{\mathbf{B}}_{l+1}$ is the solution, $\widehat{\mathbf{B}}_{l+1} = \text{argmax}_{\mathbf{B}}\ \widehat{\mathbf{D}}(\widehat{\mathbf{A}}_{l+1}, \mathbf{B})$.

Step 2: Set $l \equiv l + 1$ and repeat step 1 until either the user defined maximum number of iterations is reached or the difference between the sample index value at the l^{th} and $(l-1)^{\text{th}}$ step is less than the user defined tolerance, say, 10^{-6}.

Our algorithm performs the maximization at each iteration using the nonlinear constrained minimizer *fmincon* function obtainable in Matlab, which implements a Sequential Quadratic Programming (SQP) method that simultaneously incorporates the nonlinear constraints.

The algorithm given in Iaci et al. (2008) uses the same density based estimation of the sample index in (3), but suggested maximizing $\widehat{\mathbf{D}}\,(\mathbf{A}, \mathbf{B})$ with respect to $\mathbf{A}$ and $\mathbf{B}$ simultaneously to determine $\widehat{\mathbf{A}}$ and $\widehat{\mathbf{B}}$. However to estimate the overall measure of association between random vectors the coefficient matrices are set to the identity and thus, matrix maximization was not implemented. The iterative maximization approach of our algorithm outperformed direct simultaneous maximization in simulation. In addition, an algorithm searching sequentially for the estimated coefficient vectors was considered, but aslo did not perform as well as the algorithm presented here.

2.4 Estimation of the dimensions of the DCS

In practice the pair of true dimensions, (d_x, d_y), are unknown and thus, need to be estimated. To this end, we modify the bootstrap procedures developed in Zhu & Zeng (2006) and Iaci et al. (2010) that were originally inspired by the method developed in Ye & Weiss (2003) for selecting an optimal dimension reduction procedure for regression.

For simplicity, let $\mathbf{A}_{d_x} = \mathbf{A}_{p\times d_x}$ and $\mathbf{B}_{d_y} = \mathbf{B}_{q\times d_y}$ denote the true bases for $\mathcal{S}_{\mathbf{Y}|\mathbf{X}}$ and $\mathcal{S}_{\mathbf{X}|\mathbf{Y}}$, respectively. The dimensions of the DCS are determined by finding the subspaces $\mathcal{S}(\mathbf{A}_k)$ and $\mathcal{S}(\mathbf{B}_l)$ that collectively have the lowest variability, which is expected to occur when $k = d_x$ and $l = d_y$. To this end, let $\mathcal{S}\big(\widehat{\mathbf{A}}_k\big)$ represent an estimate of $\mathcal{S}(\mathbf{A}_k)$ based on the original data, then the variability of the subspace is naturally quantified by calculating a bootstrap estimate of the subspace, denoted $\mathcal{S}\big(\widehat{\mathbf{A}}_k^b\big)$, and then evaluating a distance between the subspaces. The collective variability is determined for every combination of k and l, where $k \le p-1$ and $l \le q-1$, and the pair that yields the least amount of variability is selected as the true dimensions of the DCS. The search is performed in one less dimension then the full dimension because the estimated coefficient matrices are orthonormal bases when $(k,l) = (p,q)$, due to the imposed orthogonality constraints, and thus, there is no variability since $\mathcal{S}\big(\widehat{\mathbf{A}}_p\big) = \mathrm{I}_{\mathbf{p}\times\mathbf{p}} = \mathcal{S}\big(\widehat{\mathbf{A}}_p^b\big)$ and $\mathcal{S}\big(\widehat{\mathbf{B}}_q\big) = \mathrm{I}_{\mathbf{q}\times\mathbf{q}} = \mathcal{S}\big(\widehat{\mathbf{B}}_q^b\big)$. As in Ye & Weiss (2003), we use the squared vector correlation coefficient to measure the distance between the two subspaces.

For the fixed dimensions (k,l), the squared vector correlation coefficient between the orthonormal bases $\widehat{\mathbf{A}}_k$ and $\widehat{\mathbf{A}}_k^b$ is, $q^2_{(\widehat{\mathbf{A}}_k,b)} = |\big(\widehat{\mathbf{A}}_k^b\big)^\top \widehat{\mathbf{A}}_k \widehat{\mathbf{A}}_k^\top \big(\widehat{\mathbf{A}}_k^b\big)| = \prod_{i=1}^k \lambda_i$, where the λ_i are the eigenvalues of $\big(\widehat{\mathbf{A}}_k^b\big)^\top \widehat{\mathbf{A}}_k \widehat{\mathbf{A}}_k^\top \big(\widehat{\mathbf{A}}_k^b\big)$. The statistic $q_{(\widehat{\mathbf{A}}_k,b)}$ is a measure of the correlation between the subspaces spanned by the original and bootstrap estimated coefficient matrices and thus, $0 \le q_{(\widehat{\mathbf{A}}_k,b)} \le 1$ with $q_{(\widehat{\mathbf{A}}_k,b)} = 1$ when the subspaces are equal and $q_{(\widehat{\mathbf{A}}_k,b)} = 0$ when the two subspaces are orthogonal. We calculate $\{1 - q_{(\widehat{\mathbf{A}}_k,b)}\}$ so that smaller values correspond to subspaces with less variability. The collective estimate of the variability of $\mathcal{S}(\mathbf{A}_k)$ and $\mathcal{S}(\mathbf{B}_l)$ is investigated on the average of $\{1 - q_{(\widehat{\mathbf{A}}_k,b)}\}$ and $\{1 - q_{(\widehat{\mathbf{B}}_l,b)}\}$ as $q_{(k,l,b)} = \big[1 - \{q_{(\widehat{\mathbf{A}}_k,b)} + q_{(\widehat{\mathbf{B}}_l,b)}\}/2\big]$.

In practice, we calculate $q_{(k,l,b)}$ for all $b \in \{1,\ldots,B\}$ bootstrap iterations at each combination of (k,l) , where $k \in \{1,\ldots,(p-1)\}$ and $l \in \{1,\ldots,(q-1)\}$, and create a box plot of each measure to visually compare the variability of the subspaces. The dimensions that result in the smallest mean (median) with the least varying box-plot, say (k^*,l^*), is taken to be the estimate of the dimensions of the DCS, $\big(\widehat{d}_x = k^*, \widehat{d}_y = l^*\big)$. If none of the box plots are centered close to zero with low variability then either no relationships exist or the true dimensions of the DCS are p and or q. In the latter case, reducing the dimensions is not beneficial and other techniques to analyze the data in the full space should be investigated.

3 Simulation studies

3.1 Introduction

In this section we investigate various scenarios involving different sample sizes and combinations of linear and nonlinear relationships between sets that contain variables following a variety of distributions. For simplicity, all models are formed for the regression of $\mathbf{Y}$ on $\mathbf{X}$. The simulation results are reported in the whitened scale, but for clarity the notations $\mathbf{X}$ and $\mathbf{Y}$ are retained. In this scale, the population bases of the DCS are in a Grassmann manifold, that is, $\mathbf{A}^\top\mathbf{A} = \mathbf{I}_{d_x}$ and $\mathbf{B}^\top\mathbf{B} = \mathbf{I}_{d_y}$, and so are the respective estimates, $\widehat{\mathbf{A}}$ and $\widehat{\mathbf{B}}$. We quantify the accuracy of the estimated DCS with the following two distance measures between the true and estimated coefficient matrices:

1. *Hotelling's squared vector correlation coefficient:* $\rho^2(\widehat{\mathbf{A}}) = \rho^2(\widehat{\mathbf{A}}, \mathbf{A}) = |\mathbf{A}^\top\widehat{\mathbf{A}}\widehat{\mathbf{A}}^\top\mathbf{A}| = \prod_i^p \lambda_i$, where the λ_i are the eigenvalues of $\mathbf{A}^\top\widehat{\mathbf{A}}\widehat{\mathbf{A}}^\top\mathbf{A}$ and $0 \leq \rho(\widehat{\mathbf{A}}) \leq 1$, as mentioned in Section 2.4.

2. *L_2 distance of the difference between the projection matrices:* $||\widehat{\mathbf{A}}||_2 = ||\widehat{\mathbf{A}}, \mathbf{A}||_2 = ||\mathbf{A}^p - \widehat{\mathbf{A}}^p||_2$, where $\mathbf{A}^p = \mathbf{A}\mathbf{A}^\top$ and $\widehat{\mathbf{A}}^p = \widehat{\mathbf{A}}\widehat{\mathbf{A}}^\top$ are projection matrices. Here, the matrix operator $||\mathbf{M}||_2$ is the standard Euclidean norm, the largest singular value of $\mathbf{M}$.

For 500 repetitions of each simulation model, we calculate the means of both measures, denoted as $\overline{\rho}(\widehat{\mathbf{A}})$ and $\overline{||\widehat{\mathbf{A}}||_2}$, and report the standard errors in the parentheses.

3.2 Estimation accuracy

In the following simulations we design two different studies to investigate the performance of our method in the presence of complicated linear and nonlinear relationships between the random vectors $\mathbf{X}$ and $\mathbf{Y}$ and confuse the relationships further via within set dependence associations between the vector $\mathbf{Y}$ in simulation I of Study 1 and simulation III of Study 2. The simulations can be summarized as follows:

Study 1: We define the multivariate random vectors $\mathbf{X} = (X_1, \ldots, X_5)^\top$ and $\mathbf{Y} = (Y_1, \ldots, Y_4)^\top$, where $X_1 \sim t(15)$, $X_2 \sim t(20)$, $X_3 \sim \Gamma(2,3)$, $X_4 \sim \chi^2_{(2)}$, $X_5 \sim \mathcal{N}(0,1)$ and $\epsilon_j \sim \mathcal{N}(0,1), j = 1,2$. The variables $Y_3 \sim \mathcal{N}(0,1)$ and $Y_4 \sim \chi^2_{(4)}$. In simulation I, the variable $Y_2 \sim \mathcal{N}(0,1)$. The remaining variables are defined in the Study 1 block of Table 1.

Study 2: We define the multivariate random vectors $\mathbf{X} = (X_1, \ldots, X_5)^\top$ and $\mathbf{Y} = (Y_1, \ldots, Y_4)^\top$, where $X_i \sim \mathcal{N}(0,1), i = 1, \ldots, 5$ and $Y_3 \sim \mathcal{N}(0,1)$, $Y_4 \sim \chi^2_{(5)}$. The error terms are $\epsilon_j \sim \mathcal{N}(0, \mathbf{\Sigma}), j = 1,2$, where $\mathbf{\Sigma} = [(2,-1)^T, (-1,1)^\top]$ and $\epsilon_3 \sim \mathcal{N}(0,1)$. The variable $Y_4 \sim \chi^2_{(5)}$ and the remaining variables are defined in the Study 2 block of Table 1.

The simulation models and the corresponding coefficient matrices that span the DCS for each study are given in Table 1. Note that, the variables in both vectors follow a variety of distributions and that we quantify the performance of our method at three different sample sizes. For sample sizes of $n = 100, 200$ and 300, datasets are generated according to the above specifications for each simulation and estimates of the matrices $\widehat{\mathbf{A}}$ and $\widehat{\mathbf{B}}$ that form the bases for the estimated DCS are calculated for 500 repetitions at each sample size. Table 2 gives the estimated mean and standard errors of the vector correlation coefficient and the L_2 distance measure for Simulations I$-$III at each sample size.

In Simulation I, the mean vector correlation coefficients between the true and estimated bases for the DCS are relatively strong with values $\overline{\rho}(\widehat{\mathbf{A}})$ =.9382 and .9332, respectively, when the sample size is the smallest. When the sample size is increased to n = 200 the same correlations rise to .9922 and .9868 and increase to .9962 and .9923 when n =30. The average of the L_2 distances show analogous results, especially for the moderate to high sample sizes, where for example the respective mean distances are .1326 and .0087 when n =30. These results indicate that our method accurately identifies the one-dimensional bases that recover the DCS.

In Simulation II, the mean vector correlation coefficients between the true and estimated bases of the subspaces $\mathcal{S}_{\mathbf{Y}|\mathbf{X}}$ and $\mathcal{S}_{\mathbf{X}|\mathbf{Y}}$ are all near one for each of the sample sizes and all

Simulation	Model	True Coefficient Matrices
	Study 1	
I	$Y_1 = -2Y_2 + \sin(X_1 + X_2) + 0.7\epsilon_1$	$\mathbf{A} = (1,1,0,0,0)^\top$, $\mathbf{B} = (1,2,0,0)^\top$
II	$Y_1 = 4\cos(X_1 + X_2) + 0.3\epsilon_1$ $Y_2 = (X_1 + X_2) + 0.5\epsilon_2$	$\mathbf{A} = (1,1,0,0,0)^\top$ $\mathbf{B} = [(1,0,0,0)^\top, (0,1,0,0)^\top]$
	Study 2	
III	$Y_1 = 4\cos(X_1 + X_3) + 0.3\epsilon_1$ $Y_2 = (X_1 + X_3) + 0.5\epsilon_2$ $Y_3 = Y_4 + X_5 + 0.6\epsilon_3$	$\mathbf{A} = [(1,0,1,0,0)^\top, (0,0,0,0,1)^\top]$ $\mathbf{B} = [(1,0,0,0)^\top, (0,1,0,0)^\top, (0,0,1,-1)^\top]$

Table 1: Simulation models

	n	100		200		300	
		$\overline{\rho}(\cdot)$	$\overline{\vert\vert \cdot \vert\vert}_2$	$\overline{\rho}(\cdot)$	$\overline{\vert\vert \cdot \vert\vert}_2$	$\overline{\rho}(\cdot)$	$\overline{\vert\vert \cdot \vert\vert}_2$
				Study 1			
Sim							
I	$\widehat{\mathbf{A}}$	.9382(.0068)	.2337(.0092)	.9922(.0013)	.1053(.0027)	.9962(.0001)	.0824(.0013)
	$\widehat{\mathbf{B}}$	.9332(.0063)	.2711(.0085)	.9868(.0011)	.1434(.0032)	.9923(.0003)	.1143(.0021)
II	$\widehat{\mathbf{A}}$	.9993(.0000)	.0342(.0006)	.9998(.0000)	.0206(.0003)	.9999(.0000)	.0156(.0003)
	$\widehat{\mathbf{B}}$	.9889(.0019)	.1210(.0028)	.9955(.0001)	.0840(.0015)	.9971(.0000)	.0672(.0012)
				Study 2			
III	$\widehat{\mathbf{A}}$	.9846(.0017)	.1461(.0034)	.9923(.0022)	.0895(.0030)	.9971(.0000)	.0688(.0012)
	$\widehat{\mathbf{B}}$	.8814(.0073)	.3903(.0094)	.9670(.0025)	.2204(.0051)	.9843(.0006)	.1629(.0030)

Table 2: Mean correlations (standard errors) $\overline{\rho(\cdot)}$ and mean distances (standard errors) $\overline{\vert\vert \cdot \vert\vert}_2$.

of the corresponding mean L_2 distances are near zero, as anticipated. These measures show that our method accurately estimates the bases that recover the trigonometric and linear relationships that exist between the random vectors.

In Simulation III, the first two functional relationships have an additional linear relationship defined between Y_3 and X_5 that is further complicated by a within set dependence relationship with Y_4. These added associations have the largest impact at the n =100 sample size. At this sample size the mean correlation decreases slightly from $\overline{\rho}(\widehat{\mathbf{A}})$ =.9993 to $\overline{\rho}(\widehat{\mathbf{A}})$ =.9846, with the largest drop occurring in the estimation of the subspace $\mathcal{S}_{\mathbf{X}|\mathbf{Y}}$, from $\overline{\rho}(\widehat{\mathbf{B}})$ =.9864 to $\overline{\rho}(\widehat{\mathbf{B}})$ =.8814. These correlations quickly rise as the sample size increases to n =200 and 300, which indicates that our procedure performs well in very complicated scenarios.

3.3 Bootstrap method for dimension estimation

A dataset of size n =300 is selected from each of the simulations above to illustrate the bootstrap method of Section 2.4 for detecting the dimensions of the DCS. The bootstrap boxplots using the vector correlation coefficient based on b =250 bootstrap iterations are given in Figure 1 for Simulations I and II of Study 1 and Simulation III of Study 2. We use a boxplot instead of the typically used mean plot so that not only can we see the changes in the center, but also the variability across the dimensions. For each simulation we only perform the bootstrap procedure up to one dimension higher than the true dimensions of the DCS, since the true dimensions are known.

The boxplots corresponding to Simulations I (left panel) and II (middle panel) in Figure 1 have the smallest median near zero and the least variability for the dimensions ($\widehat{d}_x = 1, \widehat{d}_y = 1$), correctly recovering the true dimensions of the DCS. For Simulation III (right panel), clearly the boxplots with the smallest medians and least variability occur when the dimensions are $(1, 2)$ and $(2, 3)$. The difference between the interquatiles of the two boxplots are negligible, which indicates that the same relationship is often recovered in the original and bootstrapped estimated $(1, 2)$ dimensional subspaces. This is further evidenced by more extreme values away from the third quartile for the smaller dimensioned subspaces, which occurs when different relationships are recovered from the bootstrapped dataset. Also, the variability of the boxplot corresponding to the dimensions $(2, 3)$ is smaller, with a standard deviation of 0.1170 for the bootstrapped estimates, compared to 0.1298 for the dimensions $(1, 2)$. Therefore, we conclude that the same relationships are frequently recovered in the smaller dimensioned subspaces and that the higher dimensioned subspaces, $(2, 3)$, correctly estimate the true dimensions of the DCS.

4 LA pollution data

The dataset analyzed here was obtained from a study by Shumway et al. (1988) on the possible effects of temperature and pollution on daily mortality in Los Angeles (LA) County. The complete data consists of 11 series measured daily in LA County during a 10-year period from 1970 to 1979. The three mortality series were extracted from an extensive mortality file

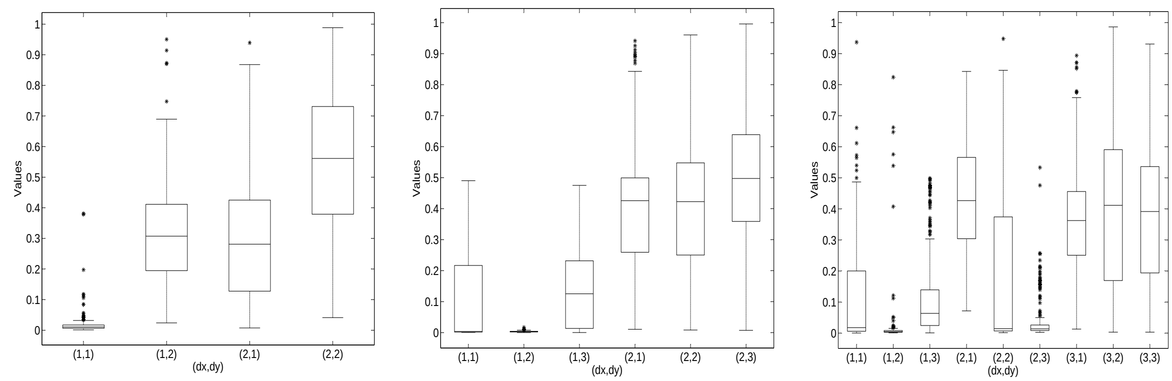

Figure 1: Vector correlation coefficient bootstrap box plots $n = 300$. Study 1: Simulations I (left panel) and II (middle panel); Study 2: III (right panel).

including all deaths of LA residents, nonresidents, and residents in other localities; the two weather series consist of maximum daily temperature and average humidity at Downtown Los Angeles and at four nearby airports; and the six pollutants were measured at six urban monitoring stations in the county. As in Iaci et al. (2010), we subset the data for analysis by using the weekly averages over the time period the data was collected, which yields a dataset of $n = 508$ observations. The following data analysis is performed in the whitened scale, but the notation $\mathbf{X}$ and $\mathbf{Y}$ is maintained throughout the section.

The second paragraph of the introduction in Shumway et al. (1988) reads: "One can generally attempt to answer two separate questions relating to the possible effects of air pollution levels on mortality. The first is that of determining the extent and nature of the association between pollutants and mortality levels in the presence of possible environmental contributors such as weather while taking account of the fact that the observations made over time are inherently correlated." This question was addressed partly by Iaci et al. (2010), who used their Generalized Canonical Analysis (GCA) method to study the multivariate associations between sets of mortality, pollution and weather random variables. More specifically, they used their multivariate dimension reduction method to first reduce the dimensions of the mortality, weather and pollutant vectors and then developed two time series regression models using the dimension reduced mortality vector as the response and the reduced pollutant and weather vectors separately as the predictors.

The second question raised by Shumway et al. (1988) is that of "defining the nature of a dose-response relation for use in predicting levels of mortality as a function of pollution and weather effects." Iaci et al. (2010) note that answering this question comprehensively is challenging and even more challenging if the predictors are correlated. To answer their own question, Shumway et al. (1988) built a nonlinear time series regression model for the response variable total mortality using the predictors, temperature and one of the three pollutants: carbon monoxide, hydrocarbons and particulates. However, they selected both the predictor and response variables, from the respective vectors, for their nonlinear regression models in an exploratory manner.

Here, we propose to provide an answer to both questions in two stages. First, we use our procedure to identify the DCS, with the mortality variables, naturally considered the multivariate response. Next, we project the predictor and response variables into the DCS and then model the projected responses using ordinary least squares regression. To this end, we consider the initial random vectors examined in Shumway et al. (1988). The multivariate response vector considered is $\mathbf{X} = (X_1, X_2, X_3)^\top$, where X_1 = total mortality, X_2 = respiratory mortality and X_3 = cardiovascular mortality. The predictor vector $\mathbf{Y} = (Y_1, \ldots, Y_6)^\top$ consists of the variables: Y_1 = temperature, Y_2 = relative humidity, Y_3 = carbon monoxide levels, Y_4 = hydrocarbon levels, Y_5 = ozone levels and Y_6 = suspended particulates.

Referencing the boxplot in the left panel of Figure 2, the bootstrap method of Section 2.4 estimates the dimension of the DCS to be $(1, 1)$ and the respective estimates are $\widehat{\mathbf{A}} = \widehat{\mathbf{a}}_1 = (0.673,0.074,0.736)^\top$ and $\widehat{\mathbf{B}} = \widehat{\mathbf{b}}_1 = (-0.589,-0.262,0.514,0.458,-0.0916,0.321)^\top$. The loadings for the mortality vector corresponding to the subspace $\mathcal{S}_{\mathbf{Y}|\mathbf{X}}$ provide a weighted average of the variables X_1 = total mortality and X_3 = cardiovascular mortality. For the estimated coefficient vector corresponding to the subspace $\mathcal{S}_{\mathbf{X}|\mathbf{Y}}$, the largest negative weight -0.5886 is placed on the variable Y_1 = temperature, followed by a decreased negative loading of -0.2620 on relative humidity and a negligible weight on ozone levels, Y_5. Next, the variables Y_3 = carbon dioxide levels, Y_4 = hydrocarbon levels and Y_6 = suspended particulates are given relatively equal positive loadings. The variables temperature and ozone levels are known to be strongly correlated, further evidenced here by an insignificant weight placed on ozone when the strongest loading in the negative direction is positioned on the temperature variable and thus, interpret the estimated coefficient vector as a contrast between the *weather* (Y_1, Y_2) and *pollutant* variables (Y_3, Y_4, Y_6).

The plot of the estimated variates $\upsilon_1 = \widehat{\mathbf{a}}_1^\top \mathbf{x}$ vs $\eta_1 = \widehat{\mathbf{b}}_1^\top \mathbf{y}$ in the right panel of Figure 2 indicates that an increasing linear association exists between υ_1 and η_1. The mortality variate and thus, total and cardiovascular mortality rates, escalate in general as the values of $\eta_1 = \widehat{\mathbf{b}}_1^\top \mathbf{y}$ increase. Based on our interpretation of the vector coefficients this occurs when the pollutant variables hydrocarbon, carbon dioxide and ozone levels increase and the weather variables decrease in relation. Note that, in general temperature and relative humidity are inversely related, with higher values of relative humidity occurring for lower temperatures, while high temperatures cause lower relative humidity. Thus, we infer that mortality rates are predominately at the lowest when the variate $\eta_1 = \widehat{\mathbf{b}}_1^\top \mathbf{y} \leq -1$, which corresponds to more extreme temperatures and reduced levels of pollutants. Alternatively, the mortality rates are highest when temperature and relative humidity are at moderate levels relative to high levels of the pollutant variables.

Next, having reduced the dimensions of both vectors, and studied the associations between these transformed variables, positions us to model the transformed mortality variables using multiple linear regression, with the estimated variate $\eta_1 = \widehat{\mathbf{b}}_1^\top \mathbf{y}$ as the predictor. To determine the "best" regression model we performed a stepwise regression for the set of predictor variables, η_1, η_1^2 and η_1^3, and set the level of significance for a variable to enter and remain in the model to be α =0.15. In the first and second steps, η_1 and η_1^3 enter the model, in that order, which is anticipated due to the increasing linear trend between υ_1 and η_1 for much of the range of η_1, which is then followed by a very slight downward trend if the few observations in the opposite direction are ignored. In the third step, η_1^2 enters the regression model, but the p-value=0.096 indicates that it is not significant. However, for the purpose of preserving the hierarchical structure, the final fitted regression model is taken to be $\widehat{\upsilon}_1 = 0.83927\eta_1 + 0.05763\eta_1^2 - 0.04461\eta_1^3$.

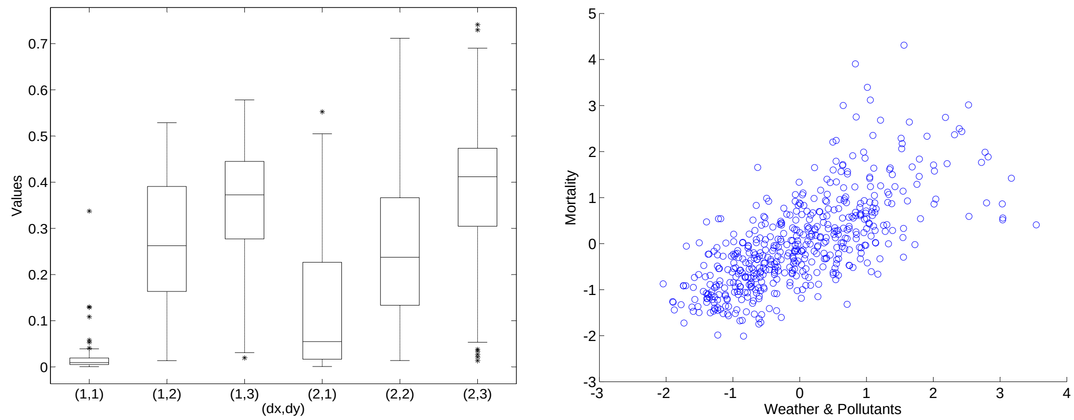

Figure 2: Left panel: Bootstrap boxplots; Right panel: Variate plot of $\upsilon_1 = \widehat{\mathbf{a}}_1^\top \mathbf{x}$ (weather & pollutants) vs $\eta_1 = \widehat{\mathbf{b}}_1^\top \mathbf{y}$ (mortality).

Acknowledgment

We would like to thank the Editor, Associate Editor and a referee for the careful reading of the article and the insightful comments that greatly improved the paper. Iaci's work was supported in part by NSF grant 1309954 and Yin's work was supported in part by NSF grant 1205546.

A Appendix

A.1 Proof of Proposition 1

Due to symmetry, we only need to prove the equivalence of conditions *(i)* and *(ii)*. Let $(\mathbf{A}, \mathbf{A}_0)$ and $(\mathbf{B}, \mathbf{B}_0)$ be orthonormal matrices.

If condition *(i)* holds, then $\mathbf{Y} \perp\!\!\!\perp \mathbf{X} | \mathbf{B}^\top \mathbf{Y}$ is equivalent to $\mathbf{Y} \perp\!\!\!\perp (\mathbf{A}^\top \mathbf{X}, \mathbf{A}_0^\top \mathbf{X}) | \mathbf{B}^\top \mathbf{Y}$, by proposition 4.6 (Cook 1998b), which is also equivalent to $\mathbf{Y} \perp\!\!\!\perp \mathbf{A}_0^\top \mathbf{X} | (\mathbf{A}^\top \mathbf{X}, \mathbf{B}^\top \mathbf{Y})$ and $\mathbf{Y} \perp\!\!\!\perp \mathbf{A}^\top \mathbf{X} | \mathbf{B}^\top \mathbf{Y}$ and thus, condition *(ii)* holds.

Next, if condition *(ii)* holds, then $\mathbf{Y} \perp\!\!\!\perp \mathbf{X} | \mathbf{A}^\top \mathbf{X}$ is equivalent to $\mathbf{Y} \perp\!\!\!\perp (\mathbf{A}^\top \mathbf{X}, \mathbf{A}_0^\top \mathbf{X}) | \mathbf{A}^\top \mathbf{X}$, again by proposition 4.4 (Cook 1998b), which is equivalent to $\mathbf{Y} \perp\!\!\!\perp \mathbf{A}_0^\top \mathbf{X} | \mathbf{A}^\top \mathbf{X}$, however this is equivalent to $\mathbf{A}_0^\top \mathbf{X} \perp\!\!\!\perp (\mathbf{B}^\top \mathbf{Y}, \mathbf{B}_0^\top \mathbf{Y}) | \mathbf{A}^\top \mathbf{X}$. The last condition, by proposition 4.6 (Cook 1998b), implies that $\mathbf{A}_0^\top \mathbf{X} \perp\!\!\!\perp \mathbf{B}_0^\top \mathbf{Y} | (\mathbf{A}^\top \mathbf{X}, \mathbf{B}^\top \mathbf{Y})$, which again by proposition 4.4 (Cook 1998b), implies that $\mathbf{A}_0^\top \mathbf{X} \perp\!\!\!\perp (\mathbf{B}_0^\top \mathbf{Y}, \mathbf{A}^\top \mathbf{X}, \mathbf{B}^\top \mathbf{Y}) | (\mathbf{A}^\top \mathbf{X}, \mathbf{B}^\top \mathbf{Y})$. However, applying proposition 4.5 (Cook, 1998b), the last condition implies that $\mathbf{A}_0^\top \mathbf{X} \perp\!\!\!\perp (\mathbf{B}_0^\top \mathbf{Y}, \mathbf{B}^\top \mathbf{Y}) | (\mathbf{A}^\top \mathbf{X}, \mathbf{B}^\top \mathbf{Y})$, which is equivalent to $\mathbf{A}_0^\top \mathbf{X} \perp\!\!\!\perp \mathbf{Y} | (\mathbf{A}^\top \mathbf{X}, \mathbf{B}^\top \mathbf{Y})$. This last condition, together with the result that $\mathbf{Y} \perp\!\!\!\perp \mathbf{A}^\top \mathbf{X} | \mathbf{B}^\top \mathbf{Y}$, gives that $\mathbf{Y} \perp\!\!\!\perp \mathbf{X} | \mathbf{B}^\top \mathbf{Y}$, again by proposition 4.6 (Cook, 1998b). Therefore, condition 1 holds.

A.2 Proof of Proposition 2

(i) Since, $\mathcal{S}(\mathbf{A}_1) \subseteq \mathcal{S}(\mathbf{A})$ and $\mathcal{S}(\mathbf{B}_1) \subseteq \mathcal{S}(\mathbf{B})$, then $\mathbf{A}_1 = \mathbf{A}\mathbf{C}_1$ and $\mathbf{B}_1 = \mathbf{B}\mathbf{C}_2$ for some matrices $\mathbf{C}_1$ and $\mathbf{C}_2$. Next, a direct extension of the proof of proposition 1 in Yin & Cook (2008) to coefficient matrices gives,

$$\begin{aligned}\mathbf{D}(\mathbf{A},\mathbf{B}) - \mathbf{D}(\mathbf{A}_1,\mathbf{B}) &= \mathrm{E}\big[\ln\big\{f(\mathbf{B}^\top\mathbf{Y}|\mathbf{A}^\top\mathbf{X})/f(\mathbf{B}^\top\mathbf{Y})\big\}\big] - \mathrm{E}\big[\ln\big\{f(\mathbf{B}^\top\mathbf{Y}|\mathbf{A}_1^\top\mathbf{X})/f(\mathbf{B}^\top\mathbf{Y})\big\}\big] \\ &= \mathrm{E}\big[\ln\big\{f(\mathbf{B}^\top\mathbf{Y}|\mathbf{A}^\top\mathbf{X})/f(\mathbf{B}^\top\mathbf{Y})\big\}\big] - \mathrm{E}\big[\ln\big\{f(\mathbf{B}^\top\mathbf{Y}|\mathbf{C}_1^\top\mathbf{A}^\top\mathbf{X})/f(\mathbf{B}^\top\mathbf{Y})\big\}\big] \\ &= \mathrm{E}\big(\mathrm{E}_{\mathbf{B}^\top\mathbf{Y}|\mathbf{A}^\top\mathbf{X}}\big[\ln\big\{f(\mathbf{B}^\top\mathbf{Y}|\mathbf{A}^\top\mathbf{X})/f(\mathbf{B}^\top\mathbf{Y}|\mathbf{C}_1^\top\mathbf{A}^\top\mathbf{X})\big\}\big]\big) \geq 0,\end{aligned}$$

where the last inequality follows from Kullback (1959), since $f(\mathbf{B}^\top\mathbf{Y}|\mathbf{A}^\top\mathbf{X})$ and $f(\mathbf{B}^\top\mathbf{Y}|\mathbf{C}_1^\top\mathbf{A}^\top\mathbf{X})$ are densities. Therefore, $\mathbf{D}(\mathbf{A},\mathbf{B}) \geq \mathbf{D}(\mathbf{A}_1,\mathbf{B})$ and, by interchanging the roles of $\mathbf{X}$ and $\mathbf{Y}$ above, $\mathbf{D}(\mathbf{A},\mathbf{B}) \geq \mathbf{D}(\mathbf{A},\mathbf{B}_1)$. Next, applying this result twice yields: $\mathbf{D}(\mathbf{A},\mathbf{B}) \geq \mathbf{D}(\mathbf{A},\mathbf{B}_1) \geq \mathbf{D}(\mathbf{A}_1,\mathbf{B}_1)$. Further, if $\mathcal{S}(\mathbf{A}_1) = \mathcal{S}(\mathbf{A})$ and $\mathcal{S}(\mathbf{B}_1) = \mathcal{S}(\mathbf{B})$ then $\mathbf{C}_1$ and $\mathbf{C}_2$ are invertible matrices and thus, due to the invariance property, $\mathbf{D}(\mathbf{A},\mathbf{B}) = \mathbf{D}(\mathbf{A}_1,\mathbf{B}_1)$.
(ii) "$\Rightarrow$". Note that,

$$\mathbf{D}(\mathbf{I}_{p\times p},\mathbf{I}_{q\times q}) - \mathbf{D}(\mathbf{A},\mathbf{B}) = \{\mathbf{D}(\mathbf{I}_{p\times p},\mathbf{I}_{q\times q}) - \mathbf{D}(\mathbf{A},\mathbf{I}_{q\times q})\} + \{\mathbf{D}(\mathbf{A},\mathbf{I}_{q\times q}) - \mathbf{D}(\mathbf{A},\mathbf{B})\}\,.$$

By part *(i)* above, the two terms on the right-hand side are nonnegative. However, if the left-hand side is 0, then both of the terms on the right-hand side are 0. This implies that $\mathbf{D}(\mathbf{I}_{p\times p},\mathbf{I}_{q\times q}) - \mathbf{D}(\mathbf{A},\mathbf{I}_{q\times q}) = 0$, which implies that $\mathrm{E}\big[\ln\big\{f(\mathbf{Y}|\mathbf{X})/f(\mathbf{Y}|\mathbf{A}^\top\mathbf{X})\big\}\big] = 0$. Hence, by Kullback (1959), $f(\mathbf{Y}|\mathbf{X}) = f(\mathbf{Y}|\mathbf{A}^\top\mathbf{X})$. That is, $\mathbf{Y} \perp\!\!\!\perp \mathbf{X}|\mathbf{A}^\top\mathbf{X}$. Similarly, we can prove that $\mathbf{Y} \perp\!\!\!\perp \mathbf{X}|\mathbf{B}^\top\mathbf{Y}$.

"$\Leftarrow$". On the other hand, using the above argument in reverse order: if $Y \perp\!\!\!\perp \mathbf{X}|\mathbf{A}^\top\mathbf{X}$, then $\mathbf{D}(\mathbf{I}_{p\times p},\mathbf{I}_{q\times q}) - \mathbf{D}(\mathbf{A},\mathbf{I}_{q\times q}) = 0$ and using $\mathbf{Y} \perp\!\!\!\perp \mathbf{X}|\mathbf{B}^\top\mathbf{Y}$, we have that

$$\begin{aligned}\mathbf{D}(\mathbf{A},\mathbf{I}_{q\times q}) - \mathbf{D}(\mathbf{A},\mathbf{B}) &= \mathrm{E}\big[\ln\{f(\mathbf{Y}|\mathbf{A}^\top\mathbf{X})/f(\mathbf{Y})\}\big] - \mathrm{E}\big[\ln\{f(\mathbf{B}^\top\mathbf{Y}|\mathbf{A}^\top\mathbf{X})/f(\mathbf{B}^\top\mathbf{Y})\}\big] \\ &= \mathrm{E}\big(\mathrm{E}_{\mathbf{Y}|\mathbf{A}^\top\mathbf{X}}\big[\ln\{f(\mathbf{A}^\top\mathbf{X}|\mathbf{Y})/f(\mathbf{A}^\top\mathbf{X}|\mathbf{B}^\top\mathbf{Y})\}\big]\big) \\ &= \mathrm{E}\big(\mathrm{E}_{\mathbf{Y}|\mathbf{A}^\top\mathbf{X}}\big[\ln\{f(\mathbf{A}^\top\mathbf{X}|\mathbf{B}^\top\mathbf{Y})/f(\mathbf{A}^\top\mathbf{X}|\mathbf{B}^\top\mathbf{Y})\}\big]\big) = 0.\end{aligned}$$

Hence, $\mathbf{D}(\mathbf{I}_{p\times p},\mathbf{I}_{q\times q}) - \mathbf{D}(\mathbf{A},\mathbf{B}) = \{\mathbf{D}(\mathbf{I}_{p\times p},\mathbf{I}_{q\times q}) - \mathbf{D}(\mathbf{A},\mathbf{I}_{q\times q})\} + \{\mathbf{D}(\mathbf{A},\mathbf{I}_{q\times q}) - \mathbf{D}(\mathbf{A},\mathbf{B})\} = 0$.

A.3 Proof of equivalent form

The mutual information between the random variables W_1 and W_2 is defined as $\mathrm{I}(W_1,W_2) = \mathrm{E}_{(W_1,W_2)}\left[\ln\left\{f(W_1,W_2)/f(W_1)f(W_2)\right\}\right] = H(W_1) - H(W_1|W_2)$, where $H(W_1) = -\mathrm{E}\left[\ln\left\{f(W_1)\right\}\right]$ and $H(W_1|W_2) = -\mathrm{E}\left[\ln\left\{f(W_1,W_2)/f(W_2)\right\}\right] = -\mathrm{E}_{W_2}\left(\mathrm{E}_{W_1|W_2}\left[\ln\left\{f(W_1|W_2)\right\}\right]\right)$, are the marginal and conditional entropies, respectively. Hence, viewing the index in (2) as a measure of the mutual information between the linear transformed random vectors, $\mathbf{A}^\top\mathbf{X}$ and $\mathbf{B}^\top\mathbf{Y}$, then straightforwardly,

$$
\begin{aligned}
&\mathbf{D}(\mathbf{A},\mathbf{B})\\
&= \mathrm{E}_{\mathbf{B}^\top\mathbf{Y}}\big(\mathrm{E}_{\mathbf{A}^\top\mathbf{X}|\mathbf{B}^\top\mathbf{Y}}\big[\ln\{f(\mathbf{A}^\top\mathbf{X}|\mathbf{B}^\top\mathbf{Y})\}\big]\big) - \mathrm{E}_{\mathbf{B}^\top\mathbf{Y}}\big(\mathrm{E}_{\mathbf{A}^\top\mathbf{X}|\mathbf{B}^\top\mathbf{Y}}\big[\ln\{f(\mathbf{A}^\top\mathbf{X})|\mathbf{B}^\top\mathbf{Y}\}\big]\big)\\
&= \mathrm{E}_{\mathbf{B}^\top\mathbf{Y}}\big(\mathrm{E}_{\mathbf{A}^\top\mathbf{X}|\mathbf{B}^\top\mathbf{Y}}\big[\ln\{f(\mathbf{A}^\top\mathbf{X}|\mathbf{B}^\top\mathbf{Y})/f(\mathbf{A}^\top\mathbf{X})\}\big]\big)\\
&= \mathrm{E}_{\mathbf{B}^\top\mathbf{Y}}\big[\mathrm{D}_{KL}\{f(\mathbf{A}^\top\mathbf{X}|\mathbf{B}^\top\mathbf{Y})||f(\mathbf{A}^\top\mathbf{X})\}\big].
\end{aligned}
$$

A.4 Proof of invariance

Let $\mathbf{W}_1 = \mathbf{C}_1^{-1}\mathbf{X} + a$ and $\mathbf{W}_2 = \mathbf{C}_2^{-1}\mathbf{Y} + b$, where $\mathbf{C}_1$ and $\mathbf{C}_2$ are nonsingular matrices and $a \in \mathbf{R}^p$ and $b \in \mathbf{R}^q$. Letting $\mathbf{D}_{(\mathbf{X},\mathbf{Y})}$ and $\mathbf{D}_{(\mathbf{W}_1,\mathbf{W}_2)}$ differentiate the indices in the $(\mathbf{X},\mathbf{Y})$ and $(\mathbf{W}_1,\mathbf{W}_2)$ supports and utilizing the result in A.3, then

$$
\begin{aligned}
\mathbf{D}_{(\mathbf{X},\mathbf{Y})}(\mathbf{A},\mathbf{B}) &= \mathrm{E}_{\mathbf{B}^\top\mathbf{Y}}\big(\mathrm{E}_{\mathbf{A}^\top\mathbf{X}|\mathbf{B}^\top\mathbf{Y}}\big[\ln\big\{f(\mathbf{A}^\top\mathbf{X}|\mathbf{B}^\top\mathbf{Y})/f(\mathbf{A}^\top\mathbf{X})\big\}\big]\big)\\
&= \mathrm{E}_{\mathbf{B}^\top\mathbf{C}_2(\mathbf{W}_2-b)}\big(\mathrm{E}_{\mathbf{A}^\top\mathbf{X}|\mathbf{B}^\top\mathbf{C}_2(\mathbf{W}_2-b)}\big[\ln\{f(\mathbf{A}^\top\mathbf{X}|\mathbf{B}^\top\mathbf{C}_2(\mathbf{W}_2-b))f(\mathbf{A}^\top\mathbf{X})\}\big]\big)\\
&= \mathrm{E}_{\mathbf{B}^\top\mathbf{C}_2\mathbf{W}_2}\big(\mathrm{E}_{\mathbf{A}^\top\mathbf{X}|\mathbf{B}^\top\mathbf{C}_2\mathbf{W}_2}\big[\ln\big\{f(\mathbf{A}^\top\mathbf{X}|\mathbf{B}^\top\mathbf{C}_2\mathbf{W}_2)/f(\mathbf{A}^\top\mathbf{X})\big\}\big]\big)\\
&= \mathrm{E}_{\mathbf{A}^\top\mathbf{C}_1\mathbf{W}_1}\big(\mathrm{E}_{\mathbf{B}^\top\mathbf{C}_2\mathbf{W}_2|\mathbf{A}^\top\mathbf{C}_1\mathbf{W}_1}\big[\ln\big\{f(\mathbf{B}^\top\mathbf{C}_2\mathbf{W}_2|\mathbf{A}^\top\mathbf{C}_1\mathbf{W}_1)f(\mathbf{B}^\top\mathbf{C}_2\mathbf{W}_2)\big\}\big]\big)\\
&= \mathbf{D}_{(\mathbf{W}_1,\mathbf{W}_2)}(\mathbf{C}_1^\top\mathbf{A},\mathbf{C}_2^\top\mathbf{B}).
\end{aligned}
$$

Note that, if $(\mathbf{A},\mathbf{B}) = \mathrm{argmax}_{(\mathbf{A}^*,\mathbf{B}^*)}\mathbf{D}_{(\mathbf{X},\mathbf{Y})}(\mathbf{A}^*,\mathbf{B}^*)$ then $(\mathbf{C}_1^\top\mathbf{A},\mathbf{C}_2^\top\mathbf{B}) = \mathrm{argmax}_{(\mathbf{A}^*,\mathbf{B}^*)}$ $\mathbf{D}_{(\mathbf{W}_1,\mathbf{W}_2)}$ $(\mathbf{A}^*,\mathbf{B}^*)$. Therefore, for the transformations $\mathbf{W}_1 = \Sigma_{\mathbf{X}}^{-1/2}\{\mathbf{X} - \mathrm{E}(\mathbf{X})\}$ and $\mathbf{W}_2 = \Sigma_{\mathbf{Y}}^{-1/2}\{\mathbf{Y} - \mathrm{E}(Y)\}$, $\mathbf{D}_{(\mathbf{X},\mathbf{Y})}(\mathbf{A},\mathbf{B}) = \mathbf{D}_{(\mathbf{W}_1,\mathbf{W}_2)}(\Sigma_{\mathbf{X}}^{1/2}\mathbf{A},\Sigma_{\mathbf{Y}}^{1/2}\mathbf{B})$ and thus, the index is invariant under this transformation.

A.5 Estimating the DCS using Projective Resampling SIR

This section of the appendix gives the results of applying the projective resampling SIR method of Li et al. (2008), as discussed in Section 2.2, to Simulations I and II of Study 1. Table 3 gives the estimated mean and standard errors of the vector correlation coefficient and the L_2 distance measure, at each sample size. The Monte Carlo sample size for resampling is $m_n = 2000$ and the number of slices is $h = 10$.

Considering the regression of $\mathbf{Y}$ on $\mathbf{X}$, the LCM condition is likely violated due to the non-normal random variables comprising the vector $\mathbf{X}$. In simulation I, the coefficient matrix $\mathbf{A}$ cannot be estimated well due to the nonlinear relationship between the predictor and response, even at the sample size $n = 300$, resulting in a poor estimate of $\mathcal{S}_{\mathbf{Y}|\mathbf{X}}$. In Simulation II the coefficient matrix $\mathbf{A}$ can be estimated through the linear relationship $Y_2 = (X_1 + X_2) + 0.5\epsilon_2$ and thus, $\mathcal{S}_{\mathbf{Y}|\mathbf{X}}$ is recovered accurately.

Next, consider the regression of $\mathbf{X}$ on $\mathbf{Y}$. In Simulation I, even though a nonlinear relationship exists between the predictor and response, since the LCM condition is likely met this method reasonably estimates the coefficient matrix $\mathbf{B}$ and hence, recovers $\mathcal{S}_{\mathbf{X}|\mathbf{Y}}$ well as the sample size increases. In Simulation II, the coefficient vector $\mathbf{b_2} = (0,1,0,0)^\top$ can be estimated through the linear relationship $Y_2 = (X_1 + X_2) + 0.5\epsilon_2$. However, due to the symmetry in the relationship $Y_1 = 4\cos(X_1 + X_2) + 0.3\epsilon_1$, estimation of $\mathbf{b}_1 = (1, 0, 0, 0)^\top$ is problematic, resulting in $\mathcal{S}_{\mathbf{X}|\mathbf{Y}}$ not being recovered accurately.

	n	100		200		300	
		$\overline{\rho}(\cdot)$	$\overline{\|\|\cdot\|\|_2}$	$\overline{\rho}(\cdot)$	$\overline{\|\|\cdot\|\|_2}$	$\overline{\rho}(\cdot)$	$\overline{\|\|\cdot\|\|_2}$
				Study 1			
Sim							
I	$\widehat{\mathbf{A}}$	.6984(.0116)	.6174(.0114)	.8027(.0095)	.5068(.0104)	.8977(.0060)	.3739(.0086)
	$\widehat{\mathbf{B}}$	.9337(.0067)	.2388(.0099)	.9889(.0019)	.1062(.0042)	.9949(.0015)	.0684(.0030)
II	$\widehat{\mathbf{A}}$	.9965(.0002)	.0749(.0016)	.9988(.0000)	.0437(.0003)	.9993(.0000)	.0351(.0007)
	$\widehat{\mathbf{B}}$	.6856(.0115)	.6305(.0114)	.8731(.0065)	.4087(.0098)	.9268(.0049)	.3005(.0087)

Table 3: Mean correlations (standard errors) $\overline{\rho(\cdot)}$ and mean distances (standard errors) $\overline{\|\cdot\|}_2$.

References

Cook, R. D. (1994). Using dimension-reduction subspaces to identify important inputs in models of physical systems. In *1994 Proceedings of the Section on Physical Engineering Sciences*, Washington.

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

Cook, R. D. (1998a). Principal Hessian directions revisited (with discussion). *Journal of the American Statistical Association*, **93**, 84–100.

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

Cook, R. D. and Li, B. (2002). Dimension reduction for the conditional mean in regression. *Ann. Statist.* **30**, 455–474.

Cook, R. D. and Setodji, C. M. (2003). A model-free test for reduced rank in multivariate regression. *Journal of the American Statistical Association*, **98**, 340–351.

Cook, R. D., Helland, I. and Su, Z. (2013). Envelopes and partial least squares regression. *Journal of the Royal Statistical Society, B*, to appear.

Cook, R.D. and Su, Z. (2013). Scaled Envelopes: Scale invariant and efficient estimation in multivariate linear regression. *Biometrika*, **100**, 921-938.

Cook, R. D., Li, B. and Chiaromonte, F. (2010). Envelope models for parsimonious and efficient multivariate linear regression (with discussion). *Statistica Sinica*, **20**, 927-1010.

Hotelling, H. (1936). Relations between two sets of variables. *Biometrika*, **58**, 433-451.

Iaci, R., Yin, X., Sriram, T. N. and Klingenberg, C. P. (2008). An informational measure of association and dimension reduction for multiple sets and groups with applications in morphometric analysis. *Journal of the American Statistical Association*, **103**, 1166-1176.

Iaci, R., Sriram T.N. and Yin, X. (2010). Multivariate Association and Dimension Reduction: A Generalization of Canonical Correlation Analysis, *Biometrics*, **66**, 1107-1118.

Iaci, R. and Sriram, T. N. (2013). Robust multivariate association and dimension reduction using density divergences, *Journal of Multivariate Analysis*, **117**, 281-295.

Kettenring, J. R. (1971). Canonical correlation analysis of several sets of variable. *Biometrika*, **58**. 433-451.

Kettenring, J. R. (1985). Canonical correlation analysis. In *Encyclopedia of Statistical Sciences*, S. Kotz and N. L. Johnson (eds.), New York: John Wiley, 354365.

Kullback, S. (1959). Information Theory and Statistics. New York: John Wiley & Sons, Inc.

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

Li, K. C., Aragon, Y., Shedden, K. and Agnan, T. (2003). Dimension reduction for multivariate response data. *Journal of the American Statistical Association*, **98**, 99-109.

Li, B., Wen, S. and Zhu, L. (2008). On a projective resampling method for dimension reduction with multivariate responses. *Journal of the American Statistical Association*, **103**, 11771186.

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

Ma, Y. and Zhu, P. (2013a). Efficient estimation in sufficient dimension reduction. *The Annals of Statistics*. **41**, 250-268.

Ma, Y. and Zhu, P. (2013b). Efficiency loss caused by linearity condition in dimension reduction. *Biometrika*. **100**, 371-383.

Scott, D. W. (1992). Density Estimation for Statistics and Data Analysis: Theory, Practice and Visualization. Wiley, New York.

Setodji, C.M. and Cook, R.D. (2004). K-means inverse regression. *Technometrics* , **46**, 421-429.

Shumway, R. H., Azari, A. S., and Pawitan, Y. (1988). Modeling mortality fluctuations in Los Angeles as functions of pollution and weather effects. *Environmental Research*, **45**, 224241.

Silverman, B.W. (1986). Density estimation for statistics and data analysis. New York: Chapman & Hall

Su, Z. and Cook, R. D. (2011). Partial envelopes for efficient estimation in multivariate linear regression. *Biometrika*, **98**, 133-146.

Su, Z. and Cook, R. D. (2012). Inner envelopes: Efficient estimation in multivariate linear regression. *Biometrika*, **99**, 687-702.

Su, Z. and Cook, R. D. (2013). Estimation of multivariate means with heteroscedastic errors using envelope models. *Statistica Sinica*, **23**, 213-230.

Van Der Burg, E. and De Leeuw, J. (1983). Non-linear Canonical Correlation. *British Journal of Mathematical and Statistical Psychology*, **36**, 54-80.

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. (2004). Canonical correlation analysis based on information theory. *Journal of Multivariate Analysis*, **91**, 161-176.

Yin, X. and Bura, E. (2006). Dimension reduction for multivariate response in regression. *Journal of Statistical Planning and Inference,* **136**, 3675-3688.

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**, 1733-1757.

Yin, X. and Cook, R. D. (2002). Dimension reduction for the conditional kth moment in regression. *Journal of the Royal Statistical Society, series B*, **64**, 159–175.

Yin, X. and Cook, R. D. (2005). Direction estimation in single-index regressions. *Biometrika*, **92**, 371-384.

Yin, X., and Sriram, T. N. (2008). Common Canonical Variates for Independent Groups Using Information Theory. *Statistica Sinica*, **18**, 335-353.

Zhu, L.P., Zhu, Lixing and Wen, S. Q. On dimension reduction in regression with multivariate responses. *Statistica Sinica*. **20**, 1291-1307.

Zhu, Y. and Zeng, P. (2006). Fourier methods for estimating the Central Subspace and the Central Mean Subspace in regression(PDF). *Journal of the American Statistical Association*, **101**, 1638-1651.