Published in final edited form as:

J Am Stat Assoc. 2015 December 1; 110(512): 1515–1527. doi:10.1080/01621459.2014.970687.

Groupwise Dimension Reduction via Envelope Method

Zifang Guo¹, Lexin Li², Wenbin Lu², and Bing Li³

¹Merck & Co., Inc

²North Carolina State University

³Pennsylvania State University

Abstract

The family of sufficient dimension reduction (SDR) methods that produce informative combinations of predictors, or indices, are particularly useful for high dimensional regression analysis. In many such analyses, it becomes increasingly common that there is available a priori subject knowledge of the predictors; e.g., they belong to different groups. While many recent SDR proposals have greatly expanded the scope of the methods' applicability, how to effectively incorporate the prior predictor structure information remains a challenge. In this article, we aim at dimension reduction that recovers full regression information while preserving the predictor group structure. Built upon a new concept of the direct sum envelope, we introduce a systematic way to incorporate the group information in most existing SDR estimators. As a result, the reduction outcomes are much easier to interpret. Moreover, the envelope method provides a principled way to build a variety of prior structures into dimension reduction analysis. Both simulations and real data analysis demonstrate the competent numerical performance of the new method.

Keywords

Central subspace; Direct sum envelope; Groupwise dimension reduction; Multiple-index models; Sliced inverse regression

1 Introduction

High dimensional regression analysis now emerges in almost every branch of science and business. Consider a global surface temperature reconstruction study as an example (Mann et al., 2008). The data contain 1,209 climate proxies yearly recorded over the past 2,000 years (with missing values). An important goal of the climate study is to better understand climate variability and to assess whether the recent temperature rise is anomalous. As instrumental climate records only became available starting the mid 19th century onwards, those proxy series were used to reconstruct global surface temperatures for the past. The temperature reconstruction model was built upon the proxies and the instrumental records of the actual surface temperatures from 1850 to 1998 A.D. This is a high dimensional regression problem with some typical features. First, there is a relatively large number of

¹Address for correspondence: Zifang Guo, 351 N Sumneytown Pike, North Wales, PA 19454. zifang.guo@merck.com.

predictors ($p = 1, 209$) but only a limited sample size ($n < 200$). Second, those proxies naturally fall into several groups, including tree composites, tree rings, ice cores, cave deposits, lake sediments, and historical documentations. Availability of such domain knowledge is becoming increasingly common in modern data collections, and is expected to facilitate statistical analysis in various ways.

Dimension reduction methods are particularly useful for this type of data analysis, by first producing low dimensional but informative summary features, and then feeding them into downstream analysis. A classical and widely used dimension reduction method is principal component analysis (PCA). PCA reduces the dimension of predictors by constructing linear combinations of the predictors that maximize the variation among the predictors. Takane et al. (1995) proposed constrained PCA to incorporate prior group information. However, a major limitation of the PCA is that the analysis is based solely on the variation in the predictors, without taking into account of the response information.

The family of sufficient dimension reduction (SDR) methods overcome this limitation by seeking linear combinations of the predictors that maximize the variation involving the response (Li, 1991; Cook and Weisberg, 1991; Cook, 1998). For a response variable Y and a p -dimensional predictor vector X , SDR seeks a unique population quantity called the central subspace that contains all information in X that is useful for predicting Y . For that reason, the reduction is termed “sufficient” as it loses no information of the conditional distribution of $Y|X$. Later, Cook and Li (2002) weakened the requirement by seeking a quantity called the central mean subspace that contains all information in the conditional mean $E(Y|X)$. Since the proposal of SDR, there have been many methods developed for both the central subspace and the central mean subspace (Li and Duan, 1989; Li, 1991; Cook and Weisberg, 1991; Li, 1992; Xia et al., 2002; Li and Wang, 2007, among many others). Sliced inverse regression (SIR; Li, 1991) is probably the most well known and commonly used method in this family, while a large number of new developments keep refining the methodology continuously. Most notably, recent proposals of Wang and Xia (2008), Li and Dong (2009), and Ma and Zhu (2012, 2013) relaxed the so-called linearity condition, a key restriction of the family of SDR methods on the predictor distributions, and thus have greatly expanded the scope of applicability of SDR.

Nevertheless, there remain some hurdles for the SDR methods, especially for handling predictors that follow some natural structures. The first hurdle is about interpretation of the SDR outcomes. Linear combinations of predictors are often interpreted as *indices*, and single and multiple-index models have received wide applications ranging from computational biology to econometrics (Ichimura, 1993; Hardle et al., 1993). If the predictors follow some natural structure, say, they belong to different groups, attaining combinations of predictors *by groups* are more sensible than indiscriminately combining all predictors together. The outcomes are also much easier to interpret this way. For instance, in the aforementioned temperature study, it is more desirable to obtain summary indices for individual groups and then reach an interpretation such as that the surface temperature is significantly associated with ice cores but not with lake sediments. The second hurdle concerns how to effectively utilize prior subject knowledge in the SDR framework, as its availability is becoming increasingly common. Regularization via various penalty functions

is a prevailing approach to utilize prior information in frequentist modeling. In the context of SDR, there have been works using lasso penalty and its variants to incorporate the sparsity type of prior information (Li, 2007; Wang and Yin, 2008; Bondell and Li, 2009). However, how to use appropriate regularization to incorporate the group or more general form of predictor structure remains largely untapped in SDR.

In this article, we propose a groupwise sufficient dimension reduction method that aims to recover full regression information, while preserving the pre-specified group structure. The core idea of our method is to enclose the subspace spanned by any classical dimension reduction estimator by the direct sum of a number of subspaces that correspond to the given group structure. The smallest such direct sum of subspaces is called the *direct sum envelope*. In this way, the prior structural information is imposed upon a classical SDR estimator, and consequently, the groupwise dimension reduction procedure can be adapted with most existing SDR methods. Moreover, the way we incorporate the prior information through the proposed direct sum envelope provides a systematic and alternative strategy to the traditional penalty function based regularization. The envelope method, in a broad view, is another way of regularization, and in that sense, it complements the penalty function approach and enriches the tactics of regularization. In general, we find the envelope procedure a principled and powerful scheme to incorporate the prior structural information in dimension reduction analysis.

The theory and method of direct sum envelope developed in the article are naturally related to several existing dimension reduction and nonparametric regression methods in the literature. First, it is closely relevant to the gMAVE method introduced by Li et al. (2010) but also distinctive in several ways. Our envelope method aims at the central subspace while gMAVE focuses on the central mean subspace. Moreover, gMAVE utilizes the structure information via a special direct sum operator inside an optimization formulation, and is uniquely tied with the minimum average variance estimator (Xia et al., 2002, MAVE) for dimension reduction. In contrast, our envelope method is not restricted to any specific SDR estimator and can be applied in conjunction with many of them. Second, we will later show that the constrained and assembled SIR introduced by Naik and Tsai (2005) and Li (2009) are special cases of the direct sum envelope method. We will give the formal condition, i.e., when the predictors are uncorrelated across the groups, to characterize this relation. Third, the direct sum envelope can also be regarded as a generalization of the partial dimension reduction methods introduced by Chiaromonte et al. (2002) and Li et al. (2003), where a subset of predictors need to be shielded from reduction. In particular, it can be used as a tool for partial dimension reduction when the blocked-out predictors are continuous — a situation that cannot be handled by most earlier partial dimension reduction methods. Finally, groupwise dimension reduction also expands the scope of the single and multiple-index models in nonparametric regression (Yin et al., 2008; Xia, 2008). In a sense, what we are doing in groupwise dimension reduction is to group the predictors into several single or multiple indices according to the prior domain knowledge in the data, and thus take the grouping into account in a regression analysis.

The rest of the article is organized as follows. We develop the notion of groupwise central subspace in Section 2 and direct sum envelope in Section 3, which are the theoretical

foundation of groupwise sufficient dimension reduction. We then show that the envelope can be constructed through optimization, at the population level in Section 4, and at the sample level in Section 5. In Section 6, we investigate some special cases of our envelope method. We conduct intensive simulations in Section 7, and the real data analysis in Section 8. We offer a rigorous justification for some key concepts in Section 9, and conclude the paper in Section 10. All proofs are relegated to the Appendix.

2 Groupwise dimension reduction subspace

In this section, we first briefly review the framework of sufficient dimension reduction. We then develop the key notion of the *groupwise central subspace* that aims to preserve full regression information while conforming to the group structure. We first present the new concept in an intuitive fashion under some special circumstances. We defer a more comprehensive and rigorous formulation to Section 9. Throughout this paper, we adopt the following notations. For subspaces $\mathcal{S}_1, \dots, \mathcal{S}_k$, the notation $\mathcal{S}_1 + \dots + \mathcal{S}_k$ stands for the subspace $\{\mathbf{v}_1 + \dots + \mathbf{v}_k : \mathbf{v}_1 \in \mathcal{S}_1, \dots, \mathbf{v}_k \in \mathcal{S}_k\}$. When $\mathcal{S}_1, \dots, \mathcal{S}_k$ are orthogonal to one another, we denote this sum by $\mathcal{S}_1 \oplus \dots \oplus \mathcal{S}_k$, and refer to it as the *direct sum* of subspaces. For matrices $\mathbf{A}_1, \dots, \mathbf{A}_k$, the notation $\mathbf{A}_1 \oplus \dots \oplus \mathbf{A}_k$ means the block-diagonal matrix with diagonal blocks $\mathbf{A}_1, \dots, \mathbf{A}_k$ and will be referred to as the direct sum of matrices. Note that the symbol $\oplus$ has different meanings when it is used with subspaces or with matrices.

In SDR, we seek a proper subspace $\mathcal{S}$ of $\mathbb{R}^p$ such that $\text{var}(Y|P_{\mathcal{S}}\mathbf{X})$ is minimized, where $P_{\mathcal{S}}$ represents the projection on to $\mathcal{S}$. This can be formulated more generally as finding a subspace $\mathcal{S}$ of $\mathbb{R}^p$ such that Y is independent of $\mathbf{X}$ given $P_{\mathcal{S}}\mathbf{X}$, i.e., $Y \perp\!\!\!\perp \mathbf{X} | P_{\mathcal{S}}\mathbf{X}$. Under very mild conditions, Yin et al. (2008) showed that the following implication is true.

Assumption 1

For any subspaces $\mathcal{S}$ and $\mathcal{S}'$ of $\mathbb{R}^p$,

$$Y \perp\!\!\!\perp \mathbf{X} | P_{\mathcal{S}}\mathbf{X}, \quad Y \perp\!\!\!\perp \mathbf{X} | P_{\mathcal{S}'}\mathbf{X} \quad \Rightarrow \quad Y \perp\!\!\!\perp \mathbf{X} | P_{\mathcal{S} \cap \mathcal{S}'}\mathbf{X},$$

We will assume it holds throughout this article. Then under this assumption, the *intersection* of all above subspaces $\mathcal{S}$ also satisfies the conditional independence. This intersection is then called the *central subspace*, is denoted by $\mathcal{S}_{Y|\mathbf{X}}$, and contains all information in $\mathbf{X}$ that is useful for predicting Y . Cook and Li (2002) introduced a weaker criterion than conditional independence for SDR by requiring instead $E(Y|\mathbf{X}) = E(Y|P_{\mathcal{S}}\mathbf{X})$. Any subspace $\mathcal{S}$ that satisfies this contains all information in $E(Y|\mathbf{X})$, and the intersection of all such subspaces is called the *central mean subspace*, denoted by $\mathcal{S}_{E(Y|\mathbf{X})}$. Compared to PCA, the SDR methods explicitly incorporate the information in both the predictors and response.

Next, we introduce the predictor group structure into SDR. Let $\mathcal{S}_1, \dots, \mathcal{S}_g \subset \mathbb{R}^p$ denote an orthogonal decomposition of $\mathbb{R}^p$, such that

$$\mathbb{R}^p = \mathcal{S}_1 \oplus \dots \oplus \mathcal{S}_g. \quad (1)$$

This is a mathematical abstraction of dividing the predictors into several groups. For instance, if a 5-dimensional predictor $\mathbf{X}$ is divided into two groups (X_1, X_2, X_3) and (X_4, X_5) , then $\mathcal{S}_1 = \text{span}(\boldsymbol{\gamma}_1) = \text{span}(\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3)$ and $\mathcal{S}_2 = \text{span}(\boldsymbol{\gamma}_2) = \text{span}(\mathbf{e}_4, \mathbf{e}_5)$, where $\mathbf{e}_i$ is a 5-dimensional unit vector with the i th element equals to 1 and others equal to 0, and $\boldsymbol{\gamma}_\ell$ denotes a basis matrix of $\mathcal{S}_\ell$, $\ell = 1, 2$. We emphasize that the group structure is assumed *known* in this article, and as such $\boldsymbol{\gamma}_\ell$'s are known. Moreover, although our results hold for more general scenario of $\boldsymbol{\gamma}_\ell$'s, we assume in this article $\boldsymbol{\gamma}_\ell$'s always take the form of the *column blocks* of the identity matrix $\mathbf{I}_p$, so that $(\boldsymbol{\gamma}_1, \dots, \boldsymbol{\gamma}_g) = \mathbf{I}_p$. This causes no loss of generality, as $\boldsymbol{\gamma}_\ell$'s are known, but it greatly simplifies the presentation and facilitates comprehension of our method.

Following a parallel construction as the central subspace, we then define the notion of the *groupwise central subspace*, which is the main object of interest in our inquiry.

Definition 1

If there are subspaces $\mathcal{T}_\ell \subseteq \mathcal{S}_\ell$ for $\ell = 1, \dots, g$ such that

$$Y \perp\!\!\!\perp \mathbf{X} | P_{\mathcal{T}_1} \mathbf{X}, \dots, P_{\mathcal{T}_g} \mathbf{X},$$

then we call $\mathcal{T}_1 \oplus \dots \oplus \mathcal{T}_g$ a *groupwise dimension reduction subspace* with respect to $\{\mathcal{S}_1, \dots, \mathcal{S}_g\}$.

Definition 2

The *groupwise central subspace* with respect to a given orthogonal decomposition $\{\mathcal{S}_1, \dots, \mathcal{S}_g\}$ is the intersection of all groupwise dimension reduction subspaces relative to that orthogonal decomposition, provided that the intersection is itself a groupwise dimension reduction subspace. We denote this subspace as $\mathcal{S}_{Y|\mathbf{X}}(\mathcal{S}_1, \dots, \mathcal{S}_g)$.

The above Definition assumes that the intersection of all groupwise dimension reduction subspaces is itself a groupwise dimension reduction subspace. Its proof is given in Section 9. Here we comment that, because any groupwise dimension reduction subspace is also a dimension reduction subspace, we have $\mathcal{S}_{Y|\mathbf{X}} \subseteq \mathcal{S}_{Y|\mathbf{X}}(\mathcal{S}_1, \dots, \mathcal{S}_g)$. In other words, given the groupwise central subspace, it is possible to perform further dimension reduction on $\mathbf{X}$ if the group structure of $\mathbf{X}$ is disregarded, and imposing the group structure limits the maximum reduction one can perform. However, if the group structure is to be preserved, the groupwise central subspace is the *smallest* subspace that contains the full regression information in the conditional distribution of $Y|\mathbf{X}$.

We next present the proposed concept of $\mathcal{S}_{Y|\mathbf{X}}(\mathcal{S}_1, \dots, \mathcal{S}_g)$ in coordinates, i.e., its basis representation. This would potentially facilitate the understanding of its structure as well as its estimation. By construction, we have

$$\mathcal{S}_{Y|\mathbf{X}}(\mathcal{S}_1, \dots, \mathcal{S}_g) = \mathcal{T}_1^* \oplus \dots \oplus \mathcal{T}_g^*$$

for some subspaces $\mathcal{T}_1^* \subseteq \mathcal{S}_1, \dots, \mathcal{T}_g^* \subseteq \mathcal{S}_g$, where the dimensions of $\mathcal{S}_\ell$ and $\mathcal{T}_\ell^*$ are p_ℓ and d_ℓ , respectively. By Definition, $p_1 + \dots + p_g = p$, and $d_1 + \dots + d_g = d$, the dimension of $\mathcal{S}_{Y|X}(\mathcal{S}_1, \dots, \mathcal{S}_g)$. Letting $\gamma_\ell \in \mathbb{R}^{p \times p_\ell}$ denote a basis matrix of $\mathcal{S}_\ell$, $\text{span}(\gamma_\ell) = \mathcal{S}_\ell$, then there is a matrix $\beta_\ell \in \mathbb{R}^{p_\ell \times d_\ell}$ of full column rank such that $\text{span}(\gamma_\ell \beta_\ell) = \mathcal{T}_\ell^*$. In other words, the basis of each subspace $\mathcal{T}_\ell^*$ that consists of the groupwise central subspace can be *separated* into two parts: the γ_ℓ 's which represent group information and the β_ℓ 's which are the true underlying directions within each group. We note that, in groupwise dimension reduction, β_ℓ 's are the target of estimation, while γ_ℓ 's are not, as they are known as prior knowledge.

We also consider some models in specific functional forms. They serve both as concrete examples of what groupwise dimension reduction aims at, and as a demonstration of the close association between $\mathcal{S}_{Y|X}(\mathcal{S}_1, \dots, \mathcal{S}_g)$ and the more familiar single and multi-index models. Specifically, consider the previous example where $\mathbf{X} \in \mathbb{R}^5$ is divided into two groups: $\{X_1, X_2, X_3\}$ and $\{X_4, X_5\}$, and $\gamma_1 = (\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3)$, $\gamma_2 = (\mathbf{e}_4, \mathbf{e}_5)$. Suppose the true regression model is one of the following:

$$Y = \exp(X_1 + X_2 - X_3) + (X_4 + X_5)^2 + \varepsilon;$$

$$Y = \exp(X_1 + X_2 - X_3) + (X_4 + X_5)^2 \times \varepsilon,$$

where ε is the random error. In both cases, each group of predictors contribute to the model through a single linear combination of the members in the group; i.e., $d_1 = d_2 = 1$, and $\beta_1 = (1, 1, -1)^\top$, $\beta_2 = (1, 1)^\top$. It follows that $\mathcal{T}_1^* = \text{span}(\gamma_1 \beta_1) = \text{span}\{(1, 1, -1, 0, 0)^\top\}$ and $\mathcal{T}_2^* = \text{span}(\gamma_2 \beta_2) = \text{span}\{(0, 0, 0, 1, 1)^\top\}$. Again, γ_1 and γ_2 are known, whereas β_1 and β_2 are to be estimated. We also note that, although in the above models, $d_1 = d_2 = 1$, it does not have to be so, as d_ℓ could be greater than one, resulting in multiple indices within a single group.

3 Direct sum envelope

We next develop the notion of the *direct sum envelope* of a generic random matrix $\mathbf{U}$, and show its connection with groupwise dimension reduction. Our goal is to find a smallest subspace with the direct sum structure that *encloses* the random subspace $\text{span}(\mathbf{U})$, where $\mathbf{U}$ takes the role of an SDR estimator at the population level. This concept is motivated by the Kronecker envelope of Li et al. (2010), which is the smallest subspace that conforms to a tensorial structure while enclosing an estimator of the central subspace. The direct sum envelope serves a similar purpose as the Kronecker envelope: they both build structures into sufficient dimension reduction. We again begin with an intuitive Definition of the direct sum envelope, and defer its rigorous justification to Section 9.

Definition 3

For a $p \times r$ random matrix $\mathbf{U}$ and an orthogonal decomposition $\{\mathcal{S}_1, \dots, \mathcal{S}_g\}$ of $\mathbb{R}^p$, let

$$\mathfrak{F} = \{\mathcal{T}_1 \oplus \cdots \oplus \mathcal{T}_g : \mathcal{T}_1 \subseteq \mathcal{S}_1, \dots, \mathcal{T}_g \subseteq \mathcal{S}_g, \text{span}(\mathbf{U}) \subseteq \mathcal{T}_1 \oplus \cdots \oplus \mathcal{T}_g \text{ almost surely}\}.$$

The intersection of all members of $\mathfrak{F}$ if itself is a member of $\mathfrak{F}$ is called the direct sum envelope of $\mathbf{U}$ with respect to $\{\mathcal{S}_1, \dots, \mathcal{S}_g\}$, and is denoted by $\mathcal{E}^\oplus(\mathbf{U}|\mathcal{S}_1, \dots, \mathcal{S}_g)$.

Again, the requirement that the intersection of all members of $\mathfrak{F}$ is a member of $\mathfrak{F}$ is a mild condition, and will be rigorously justified in Section 9. We note that, $\mathcal{E}^\oplus(\mathbf{U}|\mathcal{S}_1, \dots, \mathcal{S}_g)$ uniquely exists, and is the smallest subspace that both encloses $\text{span}(\mathbf{U})$ and conforms to the orthogonal decomposition $\{\mathcal{S}_1, \dots, \mathcal{S}_g\}$.

Next, we establish a connection between the direct sum envelope and groupwise sufficient dimension reduction. Through this connection, one can estimate the groupwise central subspace via the direct sum envelope of an estimator of the central subspace. Intuitively, if the column space of the random matrix $\mathbf{U}$ lies almost surely in a dimension reduction central subspace, then its direct sum envelope should be closely linked to the groupwise central subspace. The next theorem summarizes this relation, and is the foundation of our groupwise dimension reduction estimation.

Theorem 1

Under Assumption 1, if there is a random vector or matrix $\mathbf{U} \in \mathbb{R}^{p \times r}$ such that $\text{span}(\mathbf{U}) \subseteq \mathcal{S}_{Y|X}$ almost surely, then

$$\mathcal{E}^\oplus(\mathbf{U}|\mathcal{S}_1, \dots, \mathcal{S}_g) \subseteq \mathcal{S}_{Y|X}(\mathcal{S}_1, \dots, \mathcal{S}_g). \quad (2)$$

Moreover, the equality holds if and only if $\text{span}(\mathbf{U})$ does not lie in a proper subspace of $\mathcal{S}_{Y|X}(\mathcal{S}_1, \dots, \mathcal{S}_g)$ almost surely.

The first part of the theorem states that if there exists a random vector or matrix $\mathbf{U}$ whose column space is contained in the central subspace $\mathcal{S}_{Y|X}$ almost surely, then its direct sum envelope $\mathcal{E}^\oplus(\mathbf{U}|\mathcal{S}_1, \dots, \mathcal{S}_g)$ must be contained in the groupwise central subspace. This suggests we can build up an estimate for $\mathcal{S}_{Y|X}(\mathcal{S}_1, \dots, \mathcal{S}_g)$ from any classical estimate of $\mathcal{S}_{Y|X}$ that is based on such random vectors or matrices. Such random element $\mathbf{U}$ exists for most of the SDR methods based on inverse conditional moments. For instance, consider some typical central subspace estimators, for sliced inverse regression (Li, 1991, SIR), $\mathbf{U}(Y) = \Sigma^{-1}E(\mathbf{X}|Y)$ is a $p \times 1$ random vector; for directional regression (Li and Wang, 2007, DR), $\mathbf{U}(Y, \tilde{Y}) = \Sigma^{-1}\{2\Sigma - E[(\mathbf{X} - \tilde{\mathbf{X}})(\mathbf{X} - \tilde{\mathbf{X}})^\top | Y, \tilde{Y}]\}\Sigma^{-1/2}$ is a $p \times p$ random matrix, where $\tilde{Y}$ is an independent copy of Y .

The second part of the theorem (after “Moreover”) implies that, if $\text{span}(\mathbf{U})$ lies in a proper subspace of $\mathcal{S}_{Y|X}$ almost surely, which happens to SIR under some circumstances, then $\mathcal{E}^\oplus(\mathbf{U}|\mathcal{S}_1, \dots, \mathcal{S}_g)$, too, will lie in a proper subspace of $\mathcal{S}_{Y|X}(\mathcal{S}_1, \dots, \mathcal{S}_g)$. In that case the direct sum envelope cannot fully recover the central subspace. This part also implies that, if $\text{span}(\mathbf{U}) = \mathcal{S}_{Y|X}$ almost surely, then $\mathcal{E}^\oplus(\mathbf{U}|\mathcal{S}_1, \dots, \mathcal{S}_g)$ occupies the whole groupwise central subspace. This property is called *exhaustiveness* (Li, Zha, and Chiaromonte, 2005). With

such a property the targeted subspace can be estimated fully. For example, since DR is an exhaustive estimator for the central subspace, its direct sum envelope is an exhaustive estimator for the groupwise central subspace.

4 Population-level objective function

Given the connection between the direct sum envelope and groupwise dimension reduction, the next question is how to estimate the envelope. We first develop a population-level objective function whose minimizer gives the desired direct sum envelope, and then the corresponding sample-level algorithms in the next section. The next lemma will be useful.

Lemma 1

Let $\mathbf{A}$ and $\mathbf{B}$ be two matrices in $\mathbb{R}^{p \times d}$ of full column rank, $d \leq p$. Let $\mathbf{W}$ be a generic random vector. Suppose there exist $\mathbb{R}^{d \times r}$ -valued function $\mathbf{f}(\mathbf{W})$ and $\mathbf{g}(\mathbf{W})$ such that $\mathbf{A}\mathbf{f}(\mathbf{W}) = \mathbf{B}\mathbf{g}(\mathbf{W})$ almost surely, where $r > 0$ is an arbitrary integer. Furthermore, suppose $E[\mathbf{f}(\mathbf{W})\mathbf{f}^\top(\mathbf{W})]$ is nonsingular. Then $\text{span}(\mathbf{A}) = \text{span}(\mathbf{B})$.

The next theorem provides a way to estimate the direct sum envelope by minimizing an objective function. In the following, $\|\cdot\|$ denotes the Frobenius matrix norm. $\mathbf{W}$ is a general random vector: for SIR, $\mathbf{W} = Y$, and for DR, $\mathbf{W} = (Y, \tilde{Y})^\top$.

Theorem 2

Suppose that the elements of $\mathbf{U} \in \mathbb{R}^{p \times r}$ and $\mathbf{W}$ have finite variances and $\mathbf{U}$ is measurable with respect to $\sigma(\mathbf{W})$. Consider the objective function

$$L(\mathbf{b}_1, \dots, \mathbf{b}_g, \mathbf{f}) = E \|\Sigma^{1/2} \mathbf{U} - \Sigma^{1/2} (\mathbf{b}_1 \oplus \dots \oplus \mathbf{b}_g) \mathbf{f}(\mathbf{W})\|^2. \quad (3)$$

where $\mathbf{b}_\ell \in \mathbb{R}^{p \times d_\ell}$, $\mathbf{f}$ is a function from the support of $\mathbf{W}$ to $\mathbb{R}^{d \times r}$ with finite second moment, $d = d_1 + \dots + d_g$, and $\oplus$ is the direct sum between matrices. Let

$(\mathbf{b}_1^*, \dots, \mathbf{b}_g^*, \mathbf{f}^*) = \arg \min L(\mathbf{b}_1, \dots, \mathbf{b}_g, \mathbf{f})$. Then

$$\text{span}(\mathbf{b}_1^* \oplus \dots \oplus \mathbf{b}_g^*) = \mathcal{E}^\oplus(\mathbf{U} | \mathcal{S}_1, \dots, \mathcal{S}_g).$$

Theorem 2 implies that the minimizer of the objective function (3) always exists. The minimizer is unique up to the following equivalence class:

$\{(\mathbf{b}_1^* \oplus \dots \oplus \mathbf{b}_g^*)(\mathbf{a}_1 \oplus \dots \oplus \mathbf{a}_g), (\mathbf{a}_1 \oplus \dots \oplus \mathbf{a}_g)^{-1} \mathbf{f}^*\}$, where $\mathbf{a}_\ell$ are arbitrary nonsingular matrices of dimensions $d_\ell \times d_\ell$.

Equation (3) summarizes the essence of our procedure for constructing direct sum envelopes from a variety of conventional SDR estimators. To gain further insights into this procedure, we now outline how (3) incorporates SIR, a well-known SDR estimator, to form a direct sum envelope, and how it can be reformulated as a constrained eigenvalue problem. For simplicity, assume $E(\mathbf{X}) = \mathbf{0}$, $\text{var}(\mathbf{X}) = \mathbf{I}_p$. Let $\boldsymbol{\eta}$ denote a basis matrix of $\mathcal{S}_{Y|X}$. Then, under the assumption that $E(\mathbf{X}|\boldsymbol{\eta}^\top \mathbf{X})$ is linear in $\mathbf{X}$, we have $\text{span}\{\text{var}[E(\mathbf{X}|\mathbf{Y})]\} \subseteq \mathcal{S}_{Y|X}$. Again, for

simplicity, assume that SIR is exhaustive, so that $\text{span}\{\text{var}[E(\mathbf{X}|\mathbf{Y})]\} = \mathcal{S}_{Y|X}$. This relation implies that $\boldsymbol{\eta}$ can be found by maximizing

$$\text{trace}\{\mathbf{B}^\top \text{var}[E(\mathbf{X}|\mathbf{Y})]\mathbf{B}\} \quad (4)$$

among all semiorthogonal matrices in $\mathbb{R}^{p \times d}$, where d is the dimension of $\mathcal{S}_{Y|X}$. Alternatively, it is easy to see that $\mathbf{B}$ maximizes (4) if and only if it minimizes

$$E\|E(\mathbf{X}|\mathbf{Y}) - \mathbf{B}\mathbf{B}^\top E(\mathbf{X}|\mathbf{Y})\|^2 \quad (5)$$

This minimization can be solved by iterative least-squares where in each iteration one of the $\mathbf{B}$ s is treated as fixed. Thus, the iterative least squares problem (5) and the eigenvalue problem (4) are equivalent. For groupwise dimension reduction, we use the least squares version with an imposed direct sum structure. In fact, (3) in this case reduces to

$$E\|\mathbf{X} - (\mathbf{b}_1 \oplus \cdots \oplus \mathbf{b}_g)(\mathbf{b}_1 \oplus \cdots \oplus \mathbf{b}_g)^\top E(\mathbf{X}|\mathbf{Y})\|^2, \quad (6)$$

which is to be minimized among all $(\mathbf{b}_1, \dots, \mathbf{b}_g)$, with each $\mathbf{b}_\ell$ being a semiorthogonal matrix in $\mathbb{R}^{p_\ell \times d_\ell}$. In the general scheme of Theorem 2, $E(\mathbf{X}|\mathbf{Y})$ plays the role of $\mathbf{U}$, $\mathbf{Y}$ plays the role of $\mathbf{W}$, and $(\mathbf{b}_1 \oplus \cdots \oplus \mathbf{b}_g)^\top E(\mathbf{X}|\mathbf{Y})$ plays the role of $\mathbf{f}(\mathbf{W})$. Parallel to the classical case, minimizing (6) is equivalent to maximizing

$$\text{trace}\{(\mathbf{b}_1 \oplus \cdots \oplus \mathbf{b}_g)^\top \text{var}[E(\mathbf{X}|\mathbf{Y})](\mathbf{b}_1 \oplus \cdots \oplus \mathbf{b}_g)\}.$$

This alternative presentation shows that the groupwise dimension reduction procedure considered in this paper solves a similar eigenvalue problem as their conventional counterparts, except that a direct sum structure is embedded in the optimization.

Next we investigate how to minimize the objective function (3). In the following we write $\mathbf{b} = (\text{vec}(\mathbf{b}_1)^\top, \dots, \text{vec}(\mathbf{b}_g)^\top)^\top$ and $\mathbf{f}(\mathbf{w}) = (\mathbf{f}_1(\mathbf{w})^\top, \dots, \mathbf{f}_g(\mathbf{w})^\top)^\top$, where $\mathbf{f}_\ell(\mathbf{w})$ is a matrix of dimension $d_\ell \times r$ consisting of the $(\sum_{i=1}^{\ell-1} d_i + 1)$ th through the $(\sum_{i=1}^{\ell-1} d_i + d_\ell)$ th row of the $d \times r$ dimensional matrix $\mathbf{f}(\mathbf{w})$. Furthermore, let

$$\mathbf{V}_1(\mathbf{w}) = \text{vec}\{\boldsymbol{\Sigma}^{1/2} \mathbf{U}(\mathbf{w})\} \in \mathbb{R}^{pr \times 1},$$

$$\mathbf{V}_2(\mathbf{w}) = (\mathbf{f}_1(\mathbf{w})^\top \otimes (\boldsymbol{\Sigma}^{1/2} \boldsymbol{\gamma}_1), \dots, \mathbf{f}_g(\mathbf{w})^\top \otimes (\boldsymbol{\Sigma}^{1/2} \boldsymbol{\gamma}_g)) \in \mathbb{R}^{pr \times (p_1 d_1 + \cdots + p_g d_g)},$$

$$\mathbf{V}_3 = \mathbf{I}_r \otimes \{\boldsymbol{\Sigma}^{1/2}(\mathbf{b}_1 \oplus \cdots \oplus \mathbf{b}_g)\} \in \mathbb{R}^{pr \times dr}.$$

The next result gives the population-level solution.

Proposition 1

The minimization of $L(\mathbf{b}_1, \dots, \mathbf{b}_g, \mathbf{f})$ can be broken down into the following two steps:

1. Given fixed $\mathbf{f}(\mathbf{W})$, the minimizer $(\mathbf{b}_1, \dots, \mathbf{b}_g)$ of $L(\mathbf{b}_1, \dots, \mathbf{b}_g, \mathbf{f})$ is

$$\mathbf{b} = \{E[\mathbf{V}_2^\top(\mathbf{W})\mathbf{V}_2(\mathbf{W})]\}^{-1}E[\mathbf{V}_2^\top(\mathbf{W})\mathbf{V}_1(\mathbf{W})].$$

2. Given fixed $(\mathbf{b}_1, \dots, \mathbf{b}_g)$, the minimizer $\mathbf{f}$ of $L(\mathbf{b}_1, \dots, \mathbf{b}_g, \mathbf{f})$ is

$$\text{vec}\{\mathbf{f}(\mathbf{w})\} = (\mathbf{V}_3^\top \mathbf{V}_3)^{-1} \mathbf{V}_3^\top \mathbf{V}_1(\mathbf{w}).$$

It is seen that the optimization can be divided into iterative least squares steps, and as such is very fast. The orthogonal constraints on $\mathbf{b}$'s are imposed at the end of iteration, which is common in the literature (Li, 2007; Wang and Yin, 2008). Our numerical experiences also suggested that the algorithm often converges to the global optimum, regardless of the choice of starting values, while in practice, using multiple starting values is recommended.

5 Sample-level algorithms

In this section we provide the numerical algorithms of estimating the direct sum envelope given independent and identically distributed (i.i.d.) samples $(\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_n, Y_n)$. Without loss of generality, we assume $\mathbf{X}$ is centered at 0; that is, $E(\mathbf{X}) = 0$. The sample covariance matrix Σ of the predictors $\mathbf{X}$ is estimated by $\hat{\Sigma} = n^{-1} \sum_{i=1}^n (\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})^\top$.

We offer details of two groupwise estimation algorithms based upon two representative central subspace estimators, SIR (Li, 1991) and DR (Li and Wang, 2007), and abbreviate them by gSIR and gDR, respectively. Toward that end, we first slice the response Y using a partition $J_1, \dots, J_s$ of the support of Y , and replace Y by its discretized version $hI(Y \in J_h)$. Correspondingly, the functions $\mathbf{f}(\mathbf{W})$, $\mathbf{V}_1(\mathbf{W})$, and $\mathbf{V}_2(\mathbf{W})$ become functions of h at the sample level, and are denoted as $\mathbf{f}(h)$, $\mathbf{V}_1(h)$, and $\mathbf{V}_2(h)$, respectively.

5.1 Algorithm for gSIR

We first summarize the algorithm of gSIR, where $\mathbf{W} = Y$ and $r = 1$.

Step 1. Obtain an initial estimate of $(\mathbf{b}_1, \dots, \mathbf{b}_g)$.

Step 2. Given the current estimate of $(\mathbf{b}_1, \dots, \mathbf{b}_g)$, update $\mathbf{f}(h)$, $h = 1, \dots, s$, as:

$$\text{vec}\{\hat{\mathbf{f}}(h)\} = (\hat{\mathbf{V}}_3^\top \hat{\mathbf{V}}_3)^{-1} \hat{\mathbf{V}}_3^\top \hat{\mathbf{V}}_1(h),$$

where n_h is the number of observations in partition J_h , and

$$\hat{\mathbf{U}}(h) = \hat{\Sigma}^{-1} E_n(\mathbf{X} | Y \in J_h) = \hat{\Sigma}^{-1} n_h^{-1} \sum_{i=1}^n \mathbf{X}_i I(Y_i \in J_h),$$

$$\hat{\mathbf{V}}_1(h) = \text{vec}\{\hat{\Sigma}^{1/2}\hat{\mathbf{U}}(h)\}, \quad \text{and} \quad \hat{\mathbf{V}}_3 = \mathbf{I}_r \otimes \{\hat{\Sigma}^{1/2}(\hat{\mathbf{b}}_1 \oplus \dots \oplus \hat{\mathbf{b}}_g)\}.$$

Step 3. Given the current estimate of $f(h)$, update $(\mathbf{b}_1, \dots, \mathbf{b}_g)$ as:

$$\hat{\mathbf{b}} = \left\{ \sum_{h=1}^s (n_h/n) \hat{\mathbf{V}}_2(h)^\top \hat{\mathbf{V}}_2(h) \right\}^{-1} \left\{ \sum_{h=1}^s (n_h/n) \hat{\mathbf{V}}_2(h)^\top \hat{\mathbf{V}}_1(h) \right\},$$

$$\text{where } \hat{\mathbf{V}}_2(h) = \left(\hat{\mathbf{f}}_1(h)^\top \otimes (\hat{\Sigma}^{1/2} \boldsymbol{\gamma}_1), \dots, \hat{\mathbf{f}}_g(h)^\top \otimes (\hat{\Sigma}^{1/2} \boldsymbol{\gamma}_g) \right), h = 1, \dots, s.$$

Step 4. Return to Step 2 and iterate, until the following objective function converges:

$$\sum_{h=1}^s (n_h/n) \|\hat{\Sigma}^{1/2}\hat{\mathbf{U}}(h) - \hat{\Sigma}^{1/2}(\hat{\mathbf{b}}_1 \oplus \dots \oplus \hat{\mathbf{b}}_g) \hat{\mathbf{f}}(h)\|^2.$$

5.2 Algorithm for gDR

The algorithm for gDR is similar to gSIR, except that the presence of $(\tilde{\mathbf{X}}, \tilde{Y})$ produces some extra terms. In this case, $r = p$ and $\mathbf{W} = (Y, \tilde{Y})^\top$. As given in Li and Wang (2007), for $h, k =$

$$\begin{aligned} & E\{(\mathbf{X} - \tilde{\mathbf{X}})(\mathbf{X} - \tilde{\mathbf{X}})^\top | Y \in J_h, \tilde{Y} \in J_k\} \\ &= E(\mathbf{X}\mathbf{X}^\top | Y \in J_h) \\ &\quad - E(\mathbf{X} | Y \in J_h) E(\tilde{\mathbf{X}}^\top | \tilde{Y} \in J_k) \\ &\quad - E(\tilde{\mathbf{X}} | \tilde{Y} \in J_k) E(\mathbf{X}^\top | Y \in J_h) \\ &\quad + E(\tilde{\mathbf{X}}\tilde{\mathbf{X}}^\top | \tilde{Y} \in J_k). \end{aligned}$$

1, ..., s, we have . Its sample estimate is given by

$$\begin{aligned} \hat{D}(h, k) &= E_n\{(\mathbf{X} - \tilde{\mathbf{X}})(\mathbf{X} - \tilde{\mathbf{X}})^\top | Y \in J_h, \tilde{Y} \in J_k\} \\ &= E_n(\mathbf{X}\mathbf{X}^\top | Y \in J_h) \\ &\quad - E_n(\mathbf{X} | Y \in J_h) E_n(\tilde{\mathbf{X}}^\top | \tilde{Y} \in J_k) \\ &\quad - E_n(\tilde{\mathbf{X}} | \tilde{Y} \in J_k) E_n(\mathbf{X}^\top | Y \in J_h) \\ &\quad + E_n(\tilde{\mathbf{X}}\tilde{\mathbf{X}}^\top | \tilde{Y} \in J_k). \end{aligned}$$

. We then summarize gDR as follows.

Step 1. Obtain an initial estimate of $(\mathbf{b}_1, \dots, \mathbf{b}_g)$.

Step 2. Given the current estimate of $(\mathbf{b}_1, \dots, \mathbf{b}_g)$, update $f(h, k)$, $h, k = 1, \dots, s$, as:

$$\text{vec}\{\hat{\mathbf{f}}(h, k)\} = (\hat{\mathbf{V}}_3^\top \hat{\mathbf{V}}_3)^{-1} \hat{\mathbf{V}}_3^\top \hat{\mathbf{V}}_1(h, k),$$

$$\text{where } \hat{\mathbf{U}}(h, k) = \hat{\Sigma}^{-1} \{2\hat{\Sigma} - \hat{D}(h, k)\} \hat{\Sigma}^{-1/2},$$

$$\hat{\mathbf{V}}_1(h, k) = \text{vec}\{\hat{\Sigma}^{1/2}\hat{\mathbf{U}}(h, k)\}, \quad \text{and} \quad \hat{\mathbf{V}}_3 = \mathbf{I}_r \otimes \{\hat{\Sigma}^{1/2}(\hat{\mathbf{b}}_1 \oplus \dots \oplus \hat{\mathbf{b}}_g)\}.$$

Step 3. Given the current estimate of $f(h, k)$, $h, k = 1, \dots, s$, update $(\mathbf{b}_1, \dots, \mathbf{b}_g)$ as:

$$\hat{\mathbf{b}} \left\{ \sum_{h=1}^s \sum_{k=1}^s \frac{n_h n_k}{n^2} \hat{\mathbf{V}}_2(h, k)^\top \hat{\mathbf{V}}_2(h, k) \right\}^{-1} \left\{ \sum_{h=1}^s \sum_{k=1}^s \frac{n_h n_k}{n^2} \hat{\mathbf{V}}_2(h, k)^\top \hat{\mathbf{V}}_1(h, k) \right\},$$

$$\text{where } \hat{\mathbf{V}}_2(h, k) = \left\{ \hat{\mathbf{f}}_1(h, k)^\top \otimes (\hat{\Sigma}^{1/2} \gamma_1), \dots, \hat{\mathbf{f}}_g(h, k)^\top \otimes (\hat{\Sigma}^{1/2} \gamma_g) \right\}.$$

Step 4. Return to Step 2 and iterate, until the following objective function converges:

$$\sum_{h=1}^s \sum_{k=1}^s \frac{n_h n_k}{n^2} \left\| \hat{\Sigma}^{1/2} \hat{\mathbf{U}}(h, k) - \hat{\Sigma}^{1/2} (\hat{\mathbf{b}}_1 \oplus \dots \oplus \hat{\mathbf{b}}_g) \hat{\mathbf{f}}(h, k) \right\|^2.$$

5.3 Order determination

In the above estimation procedures, we have assumed that dimensions $d_1, d_2, \dots, d_g$ of $\mathcal{T}_1^* \oplus \dots \oplus \mathcal{T}_g^*$ are known. However, these numbers are rarely available in practice and need to be estimated. Motivated by Zhu et al. (2006); Wang and Yin (2008), we suggest using a Bayesian Information Criterion (BIC) type method to determine these dimensions. Let $L(d_{w1}, d_{w2}, \dots, d_{wg})$ be the minimum value of the objective function obtained from the minimization procedure described above when the working dimensions are chosen as $\{d_{w1}, d_{w2}, \dots, d_{wg}\}$. Then the BIC-type criterion is in the following form:

$$\text{BIC}(d_{w1}, d_{w2}, \dots, d_{wg}) = L(d_{w1}, d_{w2}, \dots, d_{wg}) + C_n \log(n) p_w,$$

where C_n is a scaling factor and $p_w = \sum_{\ell=1}^g p_\ell d_{w\ell} + r \sum_{l=1}^g d_{w\ell}$ is the number of unknown parameters corresponding to $\{d_{w1}, d_{w2}, \dots, d_{wg}\}$. The estimates $\{\hat{d}_1, \hat{d}_2, \dots, \hat{d}_g\}$ are obtained by minimizing this criterion. We used $C_n = 0.05 n^{-1/2} [L(1, 1, \dots, 1)]^{1/2}$, and the results are presented in Section 7.3. Optimal choice of C_n requires further investigation.

6 Special cases

In the literature two methods have been proposed that target the groupwise central subspace: the constrained inverse regression (CIR; Naik and Tsai, 2005) and the assembled SIR (aSIR; Li, 2009). Later Li, Li, and Zhu (2010) showed that the two are equivalent. Let $\mathbf{X}^{(\ell)} = \gamma_\ell^\top \mathbf{X}$ denote the predictors within group ℓ . The idea of CIR and aSIR is simple: to perform dimension reduction for each group of predictors $\mathbf{X}^{(\ell)}$ separately. Take SIR as an example, the aSIR estimator is essentially based on $\Sigma_\ell^{-1} E(\mathbf{X}^{(\ell)} | Y)$, where $\Sigma_\ell = \text{var}(\mathbf{X}^{(\ell)})$, $\ell = 1, \dots, g$. aDR is similar, except that a DR estimator is obtained within each group. Next we show that CIR and aSIR are special cases of the direct sum envelope approach.

Theorem 3

Suppose, in addition to the assumptions in Theorem 2, the following assumptions hold: (a) $\mathbf{U} = \Sigma^{-1} E(\mathbf{X} | Y)$; (b) $\Sigma = \Sigma_1 \oplus \dots \oplus \Sigma_g$. Then the minimizer $(\mathbf{b}_1, \dots, \mathbf{b}_g; \mathbf{f}_1(Y), \dots, \mathbf{f}_g(Y))$ of the objective function (3) satisfies

$$\mathbf{b}_\ell \mathbf{f}_\ell(Y) = \Sigma_\ell^{-1} E(\gamma_\ell^\top \mathbf{X} | Y), \quad \ell = 1, \dots, g.$$

We immediately see that, the minimizer of (3) consists of exactly the population form of SIR applied to the predictors $\mathbf{X}^{(\ell)}$ in group ℓ . In other words, the direct sum envelope approach reduces to CIR and aSIR. under the special circumstance that the predictors from different groups are *uncorrelated*, i.e., $\Sigma = \Sigma_1 \oplus \dots \oplus \Sigma_g$. On the other hand, we note that such an uncorrelated assumption can be both restrictive and difficult to verify in practice.

In sum, the key difference between the new and existing estimators is that, for the envelope approach, the data from all groups are employed for groupwise dimension reduction *jointly*; whereas for CIR and aSIR, dimension reduction is performed within groups *separately*. The joint procedure helps to avoid population bias and estimation efficiency.

7 Simulations

In this section we conduct simulations to investigate the empirical performance of gSIR and gDR under a wide variety of scenarios, and compare with aSIR and aDR. To measure estimation accuracy, the median and median absolute deviation of vector correlation coefficients (Hotelling, 1936) over 100 replications are reported in all the tables below. Note the vector correlation is between 0 and 1, and a larger correlation implies a better estimate. For both gSIR and gDR, the aSIR estimates were used as starting values in our simulations.

7.1 Groupwise index-models

Example 1: A nonlinear heteroscedastic model—We first consider a model with nonlinear relation between the response and predictors and heteroscedastic variance.

$$Y = \exp(0.23\beta_1^\top \mathbf{V}_1) + 0.5\sin(0.2\pi\beta_2^\top \mathbf{V}_2) + 0.01(4 + 0.1\beta_1^\top \mathbf{V}_1 + \beta_3^\top \mathbf{V}_3)^2 \varepsilon, \quad (7)$$

where $\mathbf{X} = (\mathbf{V}_1^\top, \mathbf{V}_2^\top, \mathbf{V}_3^\top)^\top$ is a $p = 20$ dimensional random predictor vector, which follows a multivariate normal distribution with mean zero and a covariance whose diagonal elements are 1 and off-diagonal elements are $\rho = 0.5$, i.e., a compound symmetric covariance matrix. The dimensions of $\mathbf{V}_1$, $\mathbf{V}_2$ and $\mathbf{V}_3$ are $p_1 = 10$, $p_2 = 5$ and $p_3 = 5$, respectively. The error ε is standard normal and is independent of $\mathbf{X}$. The regression coefficients are $\beta_1 = (1, -1, 0, \dots, 0)^\top$, $\beta_2 = (1, 0, \dots, 0)^\top$, and $\beta_3 = (0, \dots, 0, 1, -1)^\top$. We vary the sample size n as 100, 200, 400, 800 and 1200. The set of working dimensions are fixed at the true values (1,1,1) for all four methods. We will examine the dimension selection issue in Section 7.3. Table 1 reports the simulation results for this model. It is clearly seen that gSIR and gDR perform better than aSIR and aDR for β_1 and β_3 . Note that estimation of β_3 is particularly challenging because it only appears in the variance component. In fact, the performance of aSIR and aDR in estimating β_3 does not improve as n increases. For β_2 , all four methods provide similar results, especially for a large n . This is partly due to the fact that the regression model related to β_2 is simple and thus the estimation is relatively easy.

Example 2: Effect of dependence among predictors—Next we examine the Effect of increased dependence among predictors using the following regression model:

$$Y = \text{sign}(\beta_1^\top V_1 + \varepsilon_1) \log(|\beta_2^\top V_2 + 5 + 2\varepsilon_2|), \quad (8)$$

where $\mathbf{X} = (V_1^\top, V_2^\top)^\top$ is a $p = 20$ dimensional predictor vector. The distribution of $\mathbf{X}$ is the same as the first example except that ρ now takes 3 different values 0.2, 0.5, 0.8. The dimensions for the two groups V_1, V_2 are $p_1 = 10$ and $p_2 = 10$. The regression coefficients are $\beta_1 = (1, 1, 1, 0, \dots, 0)^\top$ and $\beta_2 = (1, -1, -1, 0, \dots, 0)^\top$. Since our experiments with several sample sizes from 100 to 1200 yield similar results, we only report the results for $n = 800$ in Table 2. As shown in this table, the performance of all four methods deteriorates as ρ increases. However, the impact of an increased ρ is much smaller on gSIR and gDR than on aSIR and aDR, especially for β_2 . It has been noted in Li, Li, and Zhu (2010) that the assembled methods are unbiased only when the predictors from different groups are conditionally independent given the response Y ; that is, $V_1 \perp\!\!\!\perp V_2 | Y$. Therefore, we expect the performance of aSIR and aDR to deteriorate quickly as ρ increases, which is exactly what we observe in Table 2. On the other hand, although the increase of ρ also affects the accuracy of gSIR and gDR, which is inevitable due to the loss of independent information, the impact is much smaller because gSIR and gDR remain unbiased.

Example 3: Groupwise multiple-index model—In all above models, the true dimension for each group is 1; that is, $d_\ell = 1$. We next consider a case where $d_\ell > 1$:

$$Y = (2 + \beta_{11}^\top V_1)^2 + \exp(0.8\beta_{12}^\top V_1) + 2\beta_{21}^\top V_2 + 2\sin(0.3\pi\beta_{22}^\top V_2) + 0.1\varepsilon, \quad (9)$$

where $\mathbf{X} = (V_1^\top, V_2^\top)^\top$ is 10-dimensional. The rest of setup is the same as model (7). The predictors V_1 and V_2 each has dimension 5. The regression coefficients are $\beta_{11} = (1, 1, 0, 0, 0)^\top$, $\beta_{12} = (0, 0, 0, 1, -1)^\top$, $\beta_{21} = (1, 1, 0, 0, 0)^\top$, and $\beta_{22} = (0, 1, 1, 0, 0)^\top$. Under this setup, the true dimension for both groups is $d_\ell = 2$. We vary the sample size n through $n = 100, 200, 400, 800$ and 1200. Table 3 summarizes the results. Again gSIR and gDR outperform aSIR and aDR, especially when the sample size is large.

7.2 Model with no or misspecified group structure

Example 4: Model with no group structure—Next we examine the potential Effect of imposing a group structure in dimension reduction estimation when there is no underlying group. Consider the model:

$$Y = \exp(0.2\beta^\top \mathbf{X}) + 0.1\varepsilon, \quad (10)$$

where $\mathbf{X} \in \mathbb{R}^{20}$ but has no group partition. The rest of setup is the same as model (7). We artificially group $\mathbf{X}$ into V_1, V_2, V_3 , with dimension $p_1 = 10, p_2 = 5$ and $p_3 = 5$, respectively.

We take $\beta = (\beta_1^\top, \beta_2^\top, \beta_3^\top)^\top$, where $\beta_1 = (1, -1, 0, \dots, 0)^\top$, $\beta_2 = (1, 0, \dots, 0)^\top$, and $\beta_3 = (0, \dots, 0, 1, -1)^\top$. We perform gSIR and gDR as if there were 3 groups of predictors. To

evaluate the estimates, we first obtain $\hat{\beta} = (\hat{\beta}_1 \oplus \hat{\beta}_2 \oplus \hat{\beta}_3)$, then report the vector correlation

coefficient between β and its projection onto $\text{span}(\hat{\beta})$, which is given by $\hat{\beta}(\hat{\beta}^\top \hat{\beta})^{-1} \hat{\beta}^\top \beta$. The performance of gSIR and gDR are compared with ordinary SIR and DR with no group partition. Table 4 summarizes the results. As expected, the performance of gSIR and gDR are slightly worse than those of SIR and DR for a small n . However, the difference diminishes as n increases and becomes essentially negligible for $n = 800$.

Example 5: Misspecified group structure—Next we investigate the consequence of imposing an incorrect group structure in estimation. Toward that goal, we revisit model (7) in Example 1, and employ three different group specifications in gSIR and gDR, one being correct and the other two incorrect. Recall that the correct specification is $V_1 = (X_1, \dots, X_{10})^\top$, $V_2 = (X_{11}, \dots, X_{15})^\top$, and $V_3 = (X_{16}, \dots, X_{20})^\top$. In the first incorrect group specification, we mistakenly assign X_2 in V_1 to V_2 , and assign X_{12} in V_2 to V_1 . The rest assignments remain unchanged. In the second incorrect specification, we assign X_3 in V_1 to V_2 , and X_{12} to V_1 . Note that the corresponding regression coefficient is -1 for X_2 , and 0 for X_3 and X_{12} , and consequently, we expect the first type of misspecification would yield a less accurate estimate than the second. The evaluation criterion is similar as in Example 4, except that $\beta = (\beta_1 \oplus \beta_2 \oplus \beta_3)$. Table 5 summarizes the results. We observed that, when the group structure is misspecified on an important variable such as in the first misspecification structure, the performance of groupwise DR estimation deteriorates, as expected. However, when the group structure is misspecified on unimportant variables as in the second misspecification structure, the performance is barely affected.

7.3 Partial dimension reduction and dimension estimation

Example 6: Partial dimension reduction—Groupwise dimension reduction can also be viewed as a generalization of *partial dimension reduction*. It was originally introduced (Chiaromonte et al., 2002; Li et al., 2003) to handle reduction of continuous predictors when categorical ones are also present. Our proposal then extends partial dimension reduction to a broader scenario when a subset of predictors, continuous or discrete, are to be shielded from reduction. Specifically, suppose X is divided into $V_1 \in \mathbb{R}^{p_1}$ and $V_2 \in \mathbb{R}^{p_2}$, and one is interested in reducing the dimension of V_2 alone, keeping V_1 intact. This problem can be cast into the groupwise dimension reduction framework by forcing β_1 to be the identity matrix I_{p_1} and seeking $\beta_2 \in \mathbb{R}^{p_2 \times d_2}$ for some $d_2 < p_2$. Consider the model:

$$Y = V_1 + \beta^\top V_2 + V_1(\beta^\top V_2) + 0.5\epsilon, \quad (11)$$

where $X = (V_1, V_2^\top)^\top$, V_1 is a scalar, $V_2 \in \mathbb{R}^{10}$. We set β as $(1, -1, 0, \dots, 0)^\top$, and $n = 100, 200, 400, 800, 1200$. The goal is to exclude V_1 from, but also use it to *assist* dimension reduction. Table 6 shows that gSIR and gDR clearly outperform SIR and DR, indicating that using the information in V_1 could facilitate dimension reduction estimation in V_2 .

Example 7: Dimension estimation—Finally, we evaluate the BIC-type criterion in Section 5.3 for dimension estimation. We illustrate with gSIR, and revisit models (7), (8) and (9). For each simulated dataset, all possible sets of dimensions are ranked such that their corresponding BIC values are in ascending order. We then compute the proportions out of

100 replications of selecting the correct set of dimensions within the top one, two and three choices, respectively, according to the above ranking. Table 7 reports the results, showing that the criterion works reasonably well, especially for a large sample size.

8 Application

We analyzed the temperature reconstruction dataset introduced in Section 1. The goal is to reconstruct the past temperatures based upon the proxies that were yearly recorded since B.C. and the instrumental temperature records from 1850 to 1998 A.D. There were originally 1,209 proxies, with many missing values. For simplicity, we only included those proxies with less than 10% missing proportions back to 999 A.D. in our analysis. The resulting 116 proxies belong to six distinct groups, including tree composites, tree rings, ice cores, cave deposits, lake sediments, and historical records and documentation. The sample size is 149; the response is the instrumental temperature anomalies; the predictors are the proxies.

We applied gSIR to this dataset with $d_1 = \dots = d_6 = 1$ as working dimensions. We did not go to higher dimensions partly because the sample size is only modest compared to the number of predictors. We evaluated the groupwise dimension reduction analysis in two ways. First, we applied gSIR to the full data, obtained six summary indices, one for each group, and then fitted a generalized additive model (Hastie and Tibshirani, 1990, GAM). We show the estimated component smooth functions of the GAM model versus the derived indices in Figure 1, and report the p -values for the six functions as 0.126, 3.58×10^{-4} , 7.68×10^{-6} , 2.04×10^{-11} , 0.151 and 0.220, respectively. Based on those p -values, the proxy groups tree rings, ice cores, and cave deposits are significantly associated with the temperature. One may also employ a penalized regression to select important proxy groups. The message here is that, thanks to Effective groupwise dimension reduction, sophisticated statistical analysis is made feasible given the derived summary features.

The second evaluation concerns the predictive capability. Toward that end, we report the mean squared error of the predicted temperatures based on 5-fold cross-validation over 100 random data splits in Figure 2. We compared four estimates, aSIR, gSIR, the GAM fit based on classical principal components analysis (PCA), and the GAM fit based on an assembled version of PCA (aPCA). It is seen from the boxplot that gSIR outperforms its competitors in this example, as it achieves the smallest median prediction error. Its estimation variability is larger than PCA type methods, though this is not surprising, as it is computationally more complicated than PCA. Together, these results suggest that the proposed groupwise dimension reduction method could be useful in improving both prediction accuracy and interpretability in a high dimensional regression analysis.

9 Justification of Definitions 2 and 3

Recall that Definition 2 and Definition 3 were made under assumptions of the form “the intersection of all members of a class is a member of that class”. For Definition 2, the class in question is all groupwise dimension reduction subspaces; for Definition 3, it is all the subspaces that contain U and conform to the group structure. While such statements are technical, they nevertheless constitute the foundation of the theory and deserve to be

rigorously justified. Another motivation of our rigorous treatment in this section is that justifications of this type arise frequently in dimension reduction, and, interestingly, a part of it has never been stated explicitly in the literature. Thus the development in this section also fills a gap in the classical theory of sufficient dimension reduction.

We first introduce three notions. In particular, $\mathfrak{C}$ is the collection of all direct sum of subspaces that conform to the orthogonal decomposition $\{\mathcal{S}_1, \dots, \mathcal{S}_g\}$.

$$\mathfrak{C} = \{\mathcal{T}_1 \oplus \dots \oplus \mathcal{T}_g : \mathcal{T}_1, \dots, \mathcal{T}_g \text{ are subspaces of } \mathbb{R}^p \text{ such that } \mathcal{T}_1 \subseteq \mathcal{S}_1, \dots, \mathcal{T}_g \subseteq \mathcal{S}_g\},$$

$$\mathfrak{G} = \{\mathcal{T}_1 \oplus \dots \oplus \mathcal{T}_g \in \mathfrak{C} : Y \perp\!\!\!\perp X|P_{\mathcal{T}_1}X, \dots, P_{\mathcal{T}_g}X\},$$

$$\mathfrak{F} = \{\mathcal{S} \in \mathfrak{C} : \text{span}(\mathbf{U}) \subseteq \mathcal{S} \text{ almost surely}\}.$$

Then the conditions in Definitions 2 and 3 can be restated as

$$\cap\{\mathcal{G} : \mathcal{G} \in \mathfrak{G}\} \in \mathfrak{G}, \cap\{\mathcal{F} : \mathcal{F} \in \mathfrak{F}\} \in \mathfrak{F}. \quad (12)$$

Theorem 4

Under Assumption 1, the two statements in (12) hold.

As will be seen in the Appendix, the proof of this theorem consists of two steps. The first is to show that $\mathfrak{G}$ and $\mathfrak{F}$ are π -systems; that is, the intersection of two members of $\mathfrak{G}$ (or $\mathfrak{F}$) is a member of $\mathfrak{G}$ (or $\mathfrak{F}$). The second is to show that the intersection of any collection of subspaces of $\mathbb{R}^p$ can be written as a finite intersection. These two steps are also needed to establish the existence of the central subspace (Cook, 1994) and the central mean subspace (Cook and Li, 2002). However, to the best of our knowledge, the second assertion is never stated explicitly in the dimension reduction literature.

10 Discussion

In this article, we have proposed a groupwise dimension reduction method that aims to preserve both full regression information and predictor group structures. Incorporation of such prior structural information is achieved via a new notion of direct sum envelope, and results in much improved interpretability.

The present work hinges upon the envelope idea to incorporate the structural information, and as such can be viewed as a continuation of a series of recent developments on envelope models. It further proves that they are a rich and versatile class that can build a variety of structures into data analysis. For instance, Cook et al. (2007, 2010) used the Σ -envelope to build a sparse eigenstructure into covariance matrices involved in dimension reduction and regression. Li et al. (2010) developed the Kronecker envelope to build a tensorial structure into dimension reduction to handle matrix or array-valued predictors. Likewise, the direct sum envelope introduced here builds a group structure into dimension reduction to incorporate domain knowledge in the predictors. It is conceivable that the envelope scheme can be further expanded to incorporate many other structures that naturally arise in the contemporary complex data in science and business.

There are several possible avenues to extend the present work. First, in this article, we treat the predictor group information as subject knowledge and completely known a priori. Such prior knowledge is often available in applications as the temperature study in Section 8 and those in Li, Li, and Zhu (2010). However, in certain applications, group structure may not be known, and need to be inferred based on statistical features of the predictors and the data; see for instance, Bondell and Reich (2008). It is also possible that the prior group information may be inaccurate. For those scenarios, we speculate some iterative strategy between inferring group structure and estimation under a given structure would be useful. Second, in scenarios where the number of predictors exceeds the sample size, gSIR or gDR is not directly applicable, as they require a SIR or DR estimator of the central subspace as a base (U) while conventional SIR or DR cannot directly handle $n < p$. On the other hand, the direct sum envelope scheme can be combined with those SDR estimators that can handle $n < p$, for instance, Cook et al. (2007); Li et al. (2007). Finally, we illustrate groupwise dimension reduction with SIR and DR, both of which require the linearity condition of elliptical symmetry on the predictor distribution. As such this condition is inherited by gSIR and gDR as well. Nevertheless, the envelope scheme can again be united with some recent SDR estimators that relax the linearity condition, e.g., Ma and Zhu (2012, 2013). Actually a strength of our direct sum envelope proposal is its exibility to be capable of mixing with many existing SDR estimators, as long as they are based on random vectors or matrices in estimating the central subspace. We are currently investigating some of the above extensions as part of our future research.

Acknowledgments

The authors are grateful to three referees and an Associate Editor for their many useful comments and suggestions, which have helped to greatly improve on an earlier manuscript. The work was partially supported by NSF Grant DMS-1106668 (Li, L.), NIH/NCI Grant R01-CA140632 (Lu, W.), and NSF Grant DMS-1106815 (Li, B.).

References

- Bondell HD, Li L. Shrinkage inverse regression estimation for model-free variable selection. *Journal of the Royal Statistical Society. Series B.* 2009; 71:287–299.
- Bondell HD, Reich BJ. Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with OSCAR. *Biometrics.* 2008; 64:115–123. 322–323. [PubMed: 17608783]
- Chiaromonte F, Cook RD, Li B. Sufficient dimension reduction in regressions with categorical predictors. *The Annals of Statistics.* 2002; 30:475–497.
- Cook R, Li B. Dimension reduction for conditional mean in regression. *The Annals of Statistics.* 2002:455–474.
- Cook R, Weisberg S. Discussion of sliced inverse regression for dimension reduction. *Journal of the American Statistical Association.* 1991; 86:328–332.
- Cook RD. Using dimension-reduction subspaces to identify important inputs in models of physical systems. *Proceedings of the Section on Physical and Engineering Sciences.* 1994:18–25.
- Cook, RD. *Regression graphics: Ideas for studying regressions through graphics.* New York: Wiley; 1998.
- Cook RD, Li B, Chiaromonte F. Dimension reduction in regression without matrix inversion. *Biometrika.* 2007; 94:569–584.
- Cook RD, Li B, Chiaromonte F. Envelope models for parsimonious and efficient multivariate linear regression (with discussion). *Statistica Sinica.* 2010; 20:927–1010.

- Hardle W, Hall P, Ichimura H. Optimal smoothing in single-index models. *The Annals of Statistics*. 1993; 21:157–178.
- Hastie, TJ.; Tibshirani, RJ. Generalized additive models. London: Chapman & Hall; 1990.
- Hotelling H. Relations between two sets of variables. *Biometrika*. 1936:321–327.
- Ichimura H. Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. *Journal of Econometrics*. 1993; 58:71–120.
- Li B, Cook RD, Chiaromonte F. Dimension reduction for conditional mean in regressions with categorical predictors. *The Annals of Statistics*. 2003; 31:1636–1668.
- Li B, Dong Y. Dimension reduction for nonelliptically distributed predictors. *The Annals of Statistics*. 2009; 37:1272–1298.
- Li B, Kim MK, Altman N. On dimension folding of matrix- or array-valued statistical objects. *The Annals of Statistics*. 2010; 38:1094–1121.
- Li B, Wang S. On directional regression for dimension reduction. *Journal of the American Statistical Association*. 2007; 102:997–1008.
- Li B, Zha H, Chiaromonte F. Contour regression: a general approach to dimension reduction. *The Annals of Statistics*. 2005; 33:1580–1616.
- Li K-C. Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*. 1991; 86:316–327.
- Li K-C. On principal hessian directions for data visualization and dimension reduction: Another application of Stein's lemma. *Journal of the American Statistical Association*. 1992; 87:1025–1039.
- Li K-C, Duan N. Regression analysis under link violation. *The Annals of Statistics*. 1989:1009–1052.
- Li L. Sparse Sufficient dimension reduction. *Biometrika*. 2007; 94:603–613.
- Li L. Exploiting predictor domain information in Sufficient dimension reduction. *Computational Statistics and Data Analysis*. 2009; 53:2665–2672.
- Li L, Cook RD, Tsai C-L. Partial inverse regression. *Biometrika*. 2007; 94:615–625.
- Li L, Li B, Zhu L-X. Groupwise dimension reduction. *Journal of the American Statistical Association*. 2010; 105:1188–1201.
- Ma Y, Zhu L. A semiparametric approach to dimension reduction. *Journal of the American Statistical Association*. 2012; 107:168–179. [PubMed: 23828688]
- Ma Y, Zhu L. Efficient estimation in Sufficient dimension reduction. *The Annals of Statistics*. 2013; 41:250–268. [PubMed: 24058219]
- Mann ME, Zhang Z, Hughes MK, Bradley RS, Miller SK, Rutherford S, Ni F. Proxy-based reconstructions of hemispheric and global surface temperature variations over the past two millennia. *Proceedings of the National Academy of Sciences*. 2008; 105:13252–13257.
- Naik PA, Tsai C-L. Constrained inverse regression for incorporating prior information. *Journal of the American Statistical Association*. 2005; 100:204–211.
- Takane Y, Kiers HAL, de Leeuw JD. Component analysis with different sets of constraints on different dimensions. *Psychometrika*. 1995; 56:97–120.
- Wang H, Xia Y. Sliced regression for dimension reduction. *Journal of the American Statistical Association*. 2008; 103:811–821.
- Wang Q, Yin X. A nonlinear multi-dimensional variable selection method for high dimensional data: Sparse MAVE. *Computational Statistics and Data Analysis*. 2008; 52:4512–4520.
- Xia Y. A multiple-index model and dimension reduction. *Journal of the American Statistical Association*. 2008; 103:1631–1640.
- Xia Y, Tong H, Li WK, Zhu L-X. An adaptive estimation of dimension reduction space. *Journal of The Royal Statistical Society Series B*. 2002; 64:363–410.
- Yin X, Li B, Cook RD. Successive direction extraction for estimating the central subspace in a multiple-index regression. *Journal of Multivariate Analysis*. 2008; 99:1733–1757.
- Zhu L, Miao BQ, Peng H. On sliced inverse regression with high dimensional covariates. *Journal of the American Statistical Association*. 2006; 101:630–643.

Appendix: proofs

Proof of Theorem 1

By construction,

$$\mathcal{S}_{Y|X}(\mathcal{S}_1, \dots, \mathcal{S}_g) = \cap \{ \mathcal{S} \in \mathfrak{C} : \mathcal{S} \supseteq \mathcal{S}_{Y|X} \}, \quad (13)$$

$$\mathcal{E}^\oplus(\mathbf{U} | \mathcal{S}_1, \dots, \mathcal{S}_g) = \cap \{ \mathcal{S} \in \mathfrak{C} : \mathcal{S} \supseteq \text{span}(\mathbf{U}) \text{ almost surely} \}.$$

If $\text{span}(\mathbf{U}) \subseteq \mathcal{S}_{Y|X}$ almost surely, then $\mathcal{S} \supseteq \mathcal{S}_{Y|X} \Rightarrow \mathcal{S} \supseteq \text{span}(\mathbf{U})$ almost surely. Hence

$$\{ \mathcal{S} \in \mathfrak{C} : \mathcal{S} \supseteq \mathcal{S}_{Y|X} \} \subseteq \{ \mathcal{S} \in \mathfrak{C} : \mathcal{S} \supseteq \text{span}(\mathbf{U}) \text{ almost surely} \}.$$

Hence the intersection of the collection of sets on the left contains the intersection of the collection of sets on the right. So, by (13), $\mathcal{E}^\oplus(\mathbf{U} | \mathcal{S}_1, \dots, \mathcal{S}_g) \subseteq \mathcal{S}_{Y|X}(\mathcal{S}_1, \dots, \mathcal{S}_g)$.

Note that the (random) subspace $\text{span}(\mathbf{U})$ is contained in $\mathcal{S}_{Y|X}$ but does not almost surely lie in a proper subspace of $\mathcal{S}_{Y|X}$ if and only if $\text{span}\{E(\mathbf{U}\mathbf{U}^\top)\} = \mathcal{S}_{Y|X}$. However,

$$\mathcal{S} \supseteq \text{span}(\mathbf{U}) \text{ almost surely} \Rightarrow \mathcal{S} \supseteq \text{span}\{E(\mathbf{U}\mathbf{U}^\top)\}.$$

Thus we have $\mathcal{E}^\oplus(\mathbf{U} | \mathcal{S}_1, \dots, \mathcal{S}_g) \supseteq \mathcal{S}_{Y|X}(\mathcal{S}_1, \dots, \mathcal{S}_g)$, as desired.

Proof of Lemma 1

Multiply both sides of the equality, $\mathbf{A}\mathbf{f}(\mathbf{W}) = \mathbf{B}\mathbf{g}(\mathbf{W})$, from the right by $\mathbf{f}^\top(\mathbf{W})$, and then take expectation, one obtains $\mathbf{A}E[\mathbf{f}(\mathbf{W})\mathbf{f}^\top(\mathbf{W})] = \mathbf{B}E[\mathbf{g}(\mathbf{W})\mathbf{f}^\top(\mathbf{W})]$. By assumption, the matrix on the left has full column rank. This implies $E[\mathbf{g}(\mathbf{W})\mathbf{f}^\top(\mathbf{W})]$ is nonsingular. Hence the columns of $\mathbf{A}$ and $\mathbf{B}$ span the same subspace in $\mathbb{R}^p$.

Proof of Theorem 2

Write $\mathcal{E}^\oplus(\mathbf{U} | \mathcal{S}_1, \dots, \mathcal{S}_g) = \text{span}(\gamma_1 \mathbf{b}_1^0) \oplus \dots \oplus \text{span}(\gamma_g \mathbf{b}_g^0) = \text{span}(\mathbf{b}_1^0 \oplus \dots \oplus \mathbf{b}_g^0)$. Because $\text{span}(\mathbf{U})$ is contained in $\mathcal{E}^\oplus(\mathbf{U} | \mathcal{S}_1, \dots, \mathcal{S}_g)$ almost surely, there is a random matrix $\mathbf{V} \in \mathbb{R}^{p \times r}$ such that $\mathbf{U} = \text{span}(\mathbf{b}_1^0 \oplus \dots \oplus \mathbf{b}_g^0) \mathbf{V}$. Because $\text{span}(\mathbf{U})$ cannot lie in a proper subspace of $\mathcal{E}^\oplus(\mathbf{U} | \mathcal{S}_1, \dots, \mathcal{S}_g)$, we must have $E(\mathbf{V}\mathbf{V}^\top) > 0$. Also, because $\mathbf{U}$ is measurable with respect to $\sigma(\mathbf{W})$, $\mathbf{V}$ can be written as $\boldsymbol{\phi}(\mathbf{W})$ for some measurable function $\boldsymbol{\phi}$. In sum, we have

$$\mathbf{U} = \text{span}(\mathbf{b}_1^0 \oplus \dots \oplus \mathbf{b}_g^0) \boldsymbol{\phi}(\mathbf{W}),$$

where $E[\boldsymbol{\phi}(\mathbf{W})\boldsymbol{\phi}^\top(\mathbf{W})] > 0$. It follows that $\boldsymbol{\Sigma}^{1/2} \mathbf{U} = \boldsymbol{\Sigma}^{1/2} (\mathbf{b}_1^0 \oplus \dots \oplus \mathbf{b}_g^0) \boldsymbol{\phi}(\mathbf{W})$. So $L(\mathbf{b}_1, \dots, \mathbf{b}_g, \mathbf{f})$ reaches its minimum at 0 within the range of $(\mathbf{b}_1, \dots, \mathbf{b}_g, \mathbf{f})$. Therefore any minimizer

$(b_1^*, \dots, b_g^*, f^*)$ must satisfy $\Sigma^{1/2}U = \Sigma^{1/2}(b_1^* \oplus \dots \oplus b_g^*)f^*(W)$ almost surely. Consequently

$$(b_1^0 \oplus \dots \oplus b_g^0)\phi(W) = (b_1^* \oplus \dots \oplus b_g^*)f^*(W) \text{ almost surely.}$$

Now apply Lemma 1 to complete the proof.

Proof of Proposition 1

We first rewrite the objective function (3) as

$$L(b_1, \dots, b_g, f) = E \|\text{vec}(\Sigma^{1/2}U) - \text{vec}\{\Sigma^{1/2}(b_1 \oplus \dots \oplus b_g)f(W)\}\|^2.$$

The first assertion is proved by noting that

$$\text{vec}\{\Sigma^{1/2}(b_1 \oplus \dots \oplus b_g)f(W)\} = \sum_{\ell=1}^g \text{vec}\{\Sigma^{1/2}\gamma_\ell b_\ell f_\ell(W)\} = \sum_{\ell=1}^g \{f_\ell(W)^\top \otimes (\Sigma^{1/2}\gamma_\ell)\} \text{vec}(b_\ell) = V_2(W)b.$$

Thus, for a fixed $f(W)$, the objective can be rewritten as $E\|V_1(W) - V_2(W)b\|^2$, whose minimizer over $b \in \mathbb{R}^{p_1 d_1 + \dots + p_g d_g}$ is as given in step 1 of Proposition 1.

The second assertion is obtained by noting that, for each fixed w in the support of W , $f(w)$ is the minimizer of

$$\begin{aligned} E\{\|\text{vec}(\Sigma^{1/2}U) - \text{vec}\{\Sigma^{1/2}(b_1 \oplus \dots \oplus b_g)\}f\|^2 | W=w\} \\ = \|\text{vec}\{\Sigma^{1/2}U(w)\} - \text{vec}\{\Sigma^{1/2}(b_1 \oplus \dots \oplus b_g)\}f(w)\|^2 \\ = \|\text{vec}\{\Sigma^{1/2}U(w)\} - (I_r \otimes \{\Sigma^{1/2}(b_1 \oplus \dots \oplus b_g)\}) \text{vec}\{f(w)\}\|^2. \end{aligned}$$

Thus, for a fixed b , the object function can be rewritten as $\|V_1(w) - V_3 z\|^2$ over $z \in \mathbb{R}^{dr}$, whose solution is given in step 1 of Proposition 1.

Proof of Theorem 3

Under conditions (a) and (b), the random vector $U = \Sigma^{-1}E(X|Y)$ can be rewritten as

$$U(Y) = (\Sigma_1^{-1} \oplus \dots \oplus \Sigma_g^{-1})E(X|Y) = (E(X^\top \gamma_1 | Y) \Sigma_1^{-1}, \dots, E(X^\top \gamma_g | Y) \Sigma_g^{-1})^\top. \quad (14)$$

Under these conditions we also have

$$\begin{aligned} \Sigma^{1/2}(b_1 \oplus \dots \oplus b_g)f(Y) &= (\Sigma_1^{1/2} \oplus \dots \oplus \Sigma_g^{1/2})(b_1 \oplus \dots \oplus b_g)f(Y) \\ &= (\Sigma_1^{1/2} b_1 \oplus \dots \oplus \Sigma_g^{1/2} b_g)f(Y) = (f_1^\top(Y) b_1^\top \Sigma_1^{1/2}, \dots, f_g^\top(Y) b_g^\top \Sigma_g^{1/2})^\top. \end{aligned} \quad (15)$$

Substitute (14) and (15) into the objective function (3), which reduces to $L(\mathbf{b}, \mathbf{f}) = E\|\mathbf{Y}\|^2$, where

$$\Delta(Y) = \begin{pmatrix} \Sigma_1^{-1/2} E(\gamma_1^\top \mathbf{X} | Y) - \Sigma_1^{1/2} \mathbf{b}_1 f_1(Y) \\ \vdots \\ \Sigma_g^{-1/2} E(\gamma_g^\top \mathbf{X} | Y) - \Sigma_g^{1/2} \mathbf{b}_g f_g(Y) \end{pmatrix}.$$

Consequently we can rewrite (3) as

$$L(\mathbf{b}, \mathbf{f}) = \sum_{\ell=1}^g E \|\Sigma_\ell^{-1/2} E(\gamma_\ell^\top \mathbf{X} | Y) - \Sigma_\ell^{1/2} \mathbf{b}_\ell f_\ell(Y)\|^2,$$

which achieves its minimum 0 if and only if the equalities $\mathbf{b}_\ell f_\ell(Y) = \Sigma_\ell^{-1} E(\gamma_\ell^\top \mathbf{X} | Y)$ hold.

Proof of Theorem 4

1. Proof of $\cap\{\mathcal{G} : \mathcal{G} \in \mathfrak{G} \in \mathfrak{G}\}$

We first show that $\mathfrak{G}$ is a π -system. Let $\mathcal{S}$ and $\mathcal{S}'$ be two members of $\mathfrak{G}$. Then there exist subspaces $\mathcal{T}_1, \dots, \mathcal{T}_g$ and $\mathcal{T}'_1, \dots, \mathcal{T}'_g$ such that: (a) $\mathcal{S} = \mathcal{T}_1 \oplus \dots \oplus \mathcal{T}_g$, $\mathcal{S}' = \mathcal{T}'_1 \oplus \dots \oplus \mathcal{T}'_g$; (b) $Y \perp \mathbf{X} | P_{\mathcal{S}} \mathbf{X}$, $Y \perp \mathbf{X} | P_{\mathcal{S}'} \mathbf{X}$. By the assumption in Theorem 4, $Y \perp \mathbf{X} | P_{\mathcal{S} \cap \mathcal{S}'} \mathbf{X}$. By Lemma 1 of Li, Li, and Zhu (2010),

$$\mathcal{S} \cap \mathcal{S}' = (\mathcal{T}_1 \cap \mathcal{T}'_1) \oplus \dots \oplus (\mathcal{T}_g \cap \mathcal{T}'_g). \quad (16)$$

Hence $\mathcal{S} \cap \mathcal{S}'$ is a member of $\mathfrak{G}$. Note that this also implies the intersection of any finite collection of subspaces in $\mathfrak{G}$ is a member of $\mathfrak{G}$.

Next, we show that any collection $\mathfrak{A}$ of subspaces of $\mathbb{R}^p$ contains a finite sub-collection $\mathfrak{A}_0 \subseteq \mathfrak{A}$ such that

$$\cap\{\mathcal{S} : \mathcal{S} \in \mathfrak{A}\} = \cap\{\mathcal{S} : \mathcal{S} \in \mathfrak{A}_0\}. \quad (17)$$

Let $\mathfrak{B} = \{\mathcal{S}^\perp : \mathcal{S} \in \mathfrak{A}\}$. Because $\text{span}\{\mathcal{U} : \mathcal{U} \in \mathfrak{B}\}$ is a finite-dimensional space, it contains a finite sub-collection $\mathfrak{B}_0$ such that $\text{span}\{\mathcal{U} : \mathcal{U} \in \mathfrak{B}\} = \text{span}\{\mathcal{U} : \mathcal{U} \in \mathfrak{B}_0\}$. Write $\mathfrak{B}_0$ as $\{\mathcal{V}_1^\perp, \dots, \mathcal{V}_m^\perp\}$, where each $\mathcal{V}_\ell \in \mathfrak{A}$. Then, for any $\mathcal{S} \in \mathfrak{A}$, $\mathcal{S}^\perp$ is a subset of $\mathcal{V}_1^\perp + \dots + \mathcal{V}_m^\perp$. We now show that

$$\cap\{\mathcal{S} : \mathcal{S} \in \mathfrak{A}\} = (\mathcal{V}_1^\perp + \dots + \mathcal{V}_m^\perp)^\perp.$$

If $\mathbf{v}$ is a member of the left-hand side, then $\mathbf{v} \perp \mathcal{S}^\perp$ for all $\mathcal{S} \in \mathfrak{A}$. Hence $\mathbf{v} \perp \mathcal{V}_1^\perp + \dots + \mathcal{V}_m^\perp$. Thus $\mathbf{v}$ is a member of the right-hand side. If $\mathbf{v}$ is a member of the right-hand side, then $\mathbf{v} \perp \mathcal{V}_1^\perp + \dots + \mathcal{V}_m^\perp$. Because every $\mathcal{S}^\perp$ with $\mathcal{S} \in \mathfrak{A}$ is a subset of

$\mathcal{V}_1^\perp + \dots + \mathcal{V}_m^\perp$, $\mathbf{v}$ is orthogonal to $\mathcal{S}^\perp$ for each $\mathcal{S} \in \mathfrak{A}$. Hence $\mathbf{v}$ is a member of the left-hand side. However, it is easy to see that

$$(\mathcal{V}_1^\perp + \dots + \mathcal{V}_m^\perp)^\perp = \bigcap_{i=1}^m \mathcal{V}_i.$$

Let $\mathfrak{A}_0 = \{\mathcal{V}_1, \dots, \mathcal{V}_m\}$ to prove (17).

Now the theorem follows because $\bigcap \{\mathcal{G} : \mathcal{G} \in \mathfrak{G}\}$ is equal to the intersection of a finite sub-collection of $\mathfrak{G}$ which, by the fact that $\mathfrak{G}$ is a π -system, must be a member of $\mathfrak{G}$.

2. Proof of $\bigcap \{\mathcal{F} : \mathcal{F} \in \mathfrak{F}\}$

In the light of (17) it suffices to show that $\mathfrak{F}$ is a π -system. Let $\mathcal{S}$ and $\mathcal{S}'$ be two members of $\mathfrak{F}$. Then there exist subspaces $\mathcal{T}_1, \dots, \mathcal{T}_g$ and $\mathcal{T}'_1, \dots, \mathcal{T}'_g$ such that: (a) $\mathcal{S} = \mathcal{T}_1 \oplus \dots \oplus \mathcal{T}_g$, $\mathcal{S}' = \mathcal{T}'_1 \oplus \dots \oplus \mathcal{T}'_g$; (b) $\text{span}(\mathbf{U}) \subseteq \mathcal{S}$ and $\text{span}(\mathbf{U}) \subseteq \mathcal{S}'$ almost surely. Hence $\text{span}(\mathbf{U}) \subseteq \mathcal{S} \cap \mathcal{S}'$ almost surely, which, together with (16), implies that $\mathcal{S} \cap \mathcal{S}'$ is a member of $\mathfrak{F}$.

Figure 1. Temperature reconstruction analysis: estimated functions for derived indices of the six proxy groups.

Figure 2.
Boxplot of the prediction error based on 5-fold cross-validation.

Simulation results for Model (7). Median and median absolute deviation (in parentheses) of vector correlation coefficient over 100 replications are reported.

Table 1

Method	$n = 100$	$n = 200$	$n = 400$	$n = 800$	$n = 1200$
β_1	gSIR 0.890 (0.071)	0.962 (0.023)	0.987 (0.006)	0.995 (0.003)	0.996 (0.002)
	aSIR 0.885 (0.048)	0.928 (0.023)	0.950 (0.012)	0.957 (0.007)	0.961 (0.005)
	gDR 0.854 (0.086)	0.942 (0.026)	0.979 (0.010)	0.991 (0.005)	0.995 (0.003)
	aDR 0.837 (0.090)	0.913 (0.035)	0.943 (0.018)	0.956 (0.009)	0.962 (0.005)
β_2	gSIR 0.944 (0.044)	0.986 (0.011)	0.994 (0.005)	0.997 (0.002)	0.998 (0.001)
	aSIR 0.955 (0.032)	0.982 (0.014)	0.991 (0.006)	0.996 (0.003)	0.998 (0.002)
	gDR 0.933 (0.059)	0.974 (0.023)	0.989 (0.008)	0.996 (0.003)	0.998 (0.002)
	aDR 0.920 (0.062)	0.972 (0.024)	0.988 (0.010)	0.995 (0.005)	0.997 (0.003)
β_3	gSIR 0.441 (0.354)	0.422 (0.299)	0.620 (0.279)	0.821 (0.170)	0.870 (0.100)
	aSIR 0.357 (0.276)	0.199 (0.187)	0.204 (0.196)	0.139 (0.115)	0.118 (0.094)
	gDR 0.338 (0.276)	0.399 (0.290)	0.510 (0.354)	0.625 (0.321)	0.683 (0.260)
	aDR 0.370 (0.286)	0.342 (0.279)	0.330 (0.278)	0.239 (0.184)	0.186 (0.152)

Table 2Simulation results for Model (8) with $n = 800$.

	Method	$\rho = 0.2$	$\rho = 0.5$	$\rho = 0.8$
β_1	gSIR	0.988 (0.005)	0.977 (0.014)	0.931 (0.029)
	aSIR	0.988 (0.006)	0.976 (0.014)	0.927 (0.034)
	gDR	0.983 (0.010)	0.966 (0.014)	0.874 (0.076)
	aDR	0.981 (0.010)	0.967 (0.019)	0.896 (0.051)
β_2	gSIR	0.988 (0.006)	0.980 (0.010)	0.955 (0.021)
	aSIR	0.986 (0.008)	0.881 (0.055)	0.701 (0.124)
	gDR	0.984 (0.007)	0.972 (0.016)	0.897 (0.051)
	aDR	0.977 (0.009)	0.931 (0.034)	0.831 (0.072)

Table 3

Simulation results for Model (9).

Method	$n = 100$	$n = 200$	$n = 400$	$n = 800$	$n = 1200$
β_1	gSIR	0.502 (0.300)	0.561 (0.277)	0.757 (0.232)	0.879 (0.090)
	aSIR	0.435 (0.328)	0.499 (0.245)	0.622 (0.223)	0.649 (0.190)
	gDR	0.526 (0.320)	0.558 (0.296)	0.776 (0.243)	0.883 (0.103)
	aDR	0.441 (0.272)	0.528 (0.283)	0.662 (0.237)	0.715 (0.165)
β_2	gSIR	0.422 (0.319)	0.571 (0.299)	0.771 (0.206)	0.868 (0.107)
	aSIR	0.430 (0.298)	0.580 (0.331)	0.712 (0.228)	0.766 (0.142)
	gDR	0.326 (0.284)	0.536 (0.410)	0.728 (0.205)	0.883 (0.109)
	aDR	0.493 (0.306)	0.635 (0.299)	0.745 (0.230)	0.825 (0.119)

Table 4

Simulation results for Model (10).

Method	$n = 100$	$n = 200$	$n = 400$	$n = 800$	$n = 1200$
gSIR	0.907 (0.055)	0.980 (0.011)	0.995 (0.002)	0.998 (0.001)	0.999 (0.001)
SIR	0.980 (0.008)	0.991 (0.003)	0.996 (0.001)	0.998 (0.001)	0.999 (0.000)
gDR	0.882 (0.059)	0.965 (0.017)	0.988 (0.005)	0.996 (0.002)	0.997 (0.001)
DR	0.960 (0.015)	0.984 (0.006)	0.992 (0.002)	0.996 (0.001)	0.998 (0.001)

Table 5

Simulation results for Model (7) with correct and incorrect group specification.

Method	Group Specification	<i>n</i> = 100	<i>n</i> = 200	<i>n</i> = 400	<i>n</i> = 800	<i>n</i> = 1200
gSIR	Correct	0.319 (0.285)	0.399 (0.272)	0.602 (0.276)	0.814 (0.172)	0.863 (0.101)
	Incorr1	0.104 (0.124)	0.210 (0.145)	0.324 (0.162)	0.444 (0.098)	0.470 (0.062)
	Incorr2	0.317 (0.294)	0.401 (0.275)	0.603 (0.275)	0.817 (0.166)	0.863 (0.099)
gDR	Correct	0.249 (0.243)	0.356 (0.256)	0.487 (0.318)	0.617 (0.316)	0.676 (0.261)
	Incorr1	0.101 (0.104)	0.170 (0.135)	0.292 (0.187)	0.322 (0.184)	0.360 (0.119)
	Incorr2	0.256 (0.255)	0.358 (0.254)	0.467 (0.358)	0.614 (0.302)	0.682 (0.264)

Table 6

Simulation results for Model (11).

Method	$n = 100$	$n = 200$	$n = 400$	$n = 800$	$n = 1200$
gSIR	0.924 (0.051)	0.973 (0.014)	0.990 (0.005)	0.994 (0.003)	0.997 (0.002)
aSIR	0.911 (0.044)	0.951 (0.022)	0.964 (0.012)	0.972 (0.005)	0.975 (0.003)
gDR	0.913 (0.060)	0.961 (0.020)	0.985 (0.006)	0.993 (0.005)	0.995 (0.002)
aDR	0.877 (0.080)	0.947 (0.027)	0.974 (0.011)	0.985 (0.006)	0.987 (0.003)

Table 7

Proportions of selecting the correct set of dimensions within the top 1, 2 and 3 choices using the BIC-type dimension estimation criterion over 100 replications.

Models	Choices	<i>n</i> =100	<i>n</i> =200	<i>n</i> =400	<i>n</i> =800	<i>n</i> =1200
(7)	Top 1	0.29	0.38	0.57	0.59	0.59
	Top 2	0.49	0.62	0.81	0.85	0.92
	Top 3	0.62	0.72	0.90	0.97	0.97
(8)	Top 1	0.67	0.93	0.98	0.99	1.00
	Top 2	0.90	0.99	1.00	1.00	1.00
	Top 3	0.99	1.00	1.00	1.00	1.00
(9)	Top 1	0.56	0.60	0.65	0.63	0.69
	Top 2	0.76	0.84	0.88	0.89	0.95
	Top 3	0.87	0.91	0.92	0.97	0.98