Annual Review of Statistics and Its Application

Principal Components, Sufficient Dimension Reduction, and Envelopes

R. Dennis Cook

School of Statistics, University of Minnesota, Minneapolis, Minnesota 55455, USA; email: dennis@stat.umn.edu

Annu. Rev. Stat. Appl. 2018. 5:533–59

First published as a Review in Advance on December 8, 2017

The *Annual Review of Statistics and Its Application* is online at statistics.annualreviews.org

https://doi.org/10.1146/annurev-statistics-031017-100257

Keywords

central subspace, dimension reduction for covariance matrices, probabilistic principal components, principal fitted components, sliced average variance estimation, sliced inverse regression

Abstract

We review probabilistic principal components, principal fitted components, sufficient dimension reduction, and envelopes, arguing that at their core they are all based on variations of the conditional independence argument that Fisher used to develop his fundamental concept of sufficiency. We emphasize the foundations of the methods. Methodological details, derivations, and examples are included when they convey the flavor and implications of basic concepts. In addition to the main topics, this review covers extensions of probabilistic principal components, the central subspace and central mean subspace, sliced inverse regression, sliced average variance estimation, dimension reduction for covariance matrices, and response and predictor envelopes.

1. INTRODUCTION

Interpreted broadly, dimension reduction has always been a bedrock of statistical thought, as illustrated by the title of a nineteenth-century article by Edgeworth (1884), "On the reduction of observations." Edgeworth's general notion was to reduce a series of observations to a relatively few statistics that hold its essential information. Fisher (1922) formalized that idea by introducing the concept of sufficiency. For Fisher, the reductive process started with the presumption that the data D arose from an unknown member of a known family of distributions indexed by parameter θ that represents the quantity to be estimated. A statistic $t(D)$ is deemed to be sufficient for θ if it contains all the information about θ that is available from D. More specifically, t is sufficient for θ if

$$D|(\theta, t) \sim D|t, \qquad 1.$$

where $\sim$ means equal in distribution. Although Fisher's notion of sufficiency provides a theoretical cornerstone for the reduction of observations that, for a time, stimulated considerable discussion, its use in applications can be problematic because of the starting requirement of a known parametric family of distributions (Stigler 1973). Nevertheless, Equation 1 provides a touchstone for the class of dimension-reduction methods that we consider in this review.

In Section 2 we review principal component analysis (PCA), which is perhaps the most widely used dimension-reduction method across the applied sciences. Its beginnings are usually associated with the work of Pearson (1901) and Hotelling (1933), but it can be traced back at least to Adcock (1878). Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ denote independent copies of a p-dimensional random vector $\mathbf{X}$. PCA is used often to construct from $\mathbf{X}$ a $p^* \times 1$ vector of proxies $\mathbf{X}^*$ with $p^* < p$ that can serve as a substitute for $\mathbf{X}$ without (much) loss of relevant information. This is different from the notion of reduction used by Edgeworth and Fisher. Sufficiency is concerned with reducing the size n of the data to a few sufficient statistics, whereas the goal of PCA is to reduce the dimension p of each vector-valued observation rather than the sample size. There are other differences as well. PCA is fundamentally an ad hoc method that does not require a parametric family, and thus the notion of preservation of information is elusive when viewed against the requirement of sufficiency (Equation 1). Instead, PCA takes a leap of faith and essentially equates information with variation, PCA reductions being designed to preserve much of the variation in the original data, as described in Section 2. Nevertheless, PCA has been used successfully in may applications. Cavalli-Sforza et al. (1994) used PCA to produce acclaimed continental maps summarizing human genetic variation. When used in the analysis of microarray data, principal components have been called eigengenes (Alter et al. 2000). Tipping & Bishop (1999) produced the first probability model that, when estimated with maximum likelihood methods, leads to PCA. This is a landmark in the history of PCA because it provided a context that shows when PCA will likely work well, points to new methodology, and leads back to a version of Fisher's sufficiency criterion (Equation 1). In largely unrecognized work, Chiaromonte (1996, 1997) proposed a general reductive framework that subsumes much of the basic structure proposed by Tipping & Bishop (1999).

In Section 3, we turn to dimension reduction for the regression of a univariate response Y on a vector of p predictors $\mathbf{X} \in \mathbb{R}^p$ based on independent observations $(Y_i, \mathbf{X}_i)$, $i = 1, \ldots, n$. The informal goal here is to reduce the dimension of the predictor vector while retaining all or most of its information about the response. This is like the goal in PCA, except now the information to be retained is relatively unambiguous. Nevertheless, there is a long history of uneven attempts to use PCA as method for reducing $\mathbf{X}$ (Cook 2007). It is now generally recognized that PCA is not a reliable method for predictor reduction in regression because it ignores the response (but see Artemiou & Li 2009 for a different perspective). Instead, contemporary methods have adopted approaches grounded in Fisher's original notion (Equation 1).

Two distinct approaches for reducing the dimension of $\mathbf{X}$ have emerged over the past few decades. Both strive to construct proxy predictors as linear combinations of the original predictors, $\mathbf{X} \mapsto \boldsymbol{\eta}^T \mathbf{X}$, where $\boldsymbol{\eta} \in \mathbb{R}^{p \times d}$ and $d \leq p$, with the property that

$$Y \perp\!\!\!\perp \mathbf{X} \mid \boldsymbol{\eta}^T \mathbf{X}, \qquad 2.$$

where $\perp\!\!\!\perp$ means independent. Under this specification, if we know the proxy predictors $\boldsymbol{\eta}^T \mathbf{X}$, no more information about Y is available from $\mathbf{X}$. One approach, identified broadly by the use of sparsity as a driving constraint, is based on the notion that only $d \ll p$ predictors are relevant to the regression, and is driven by the goal of identifying those predictors. In this scenario, the columns of the matrix $\boldsymbol{\eta}$ in Expression 2 are limited to orthogonal vectors, each with a single nonzero element. Although there are contexts where sparsity is required as part of the overarching science, some seem to view sparsity as akin to a natural law: If you are faced with a high-dimensional regression then naturally it must be sparse. Others have seen sparsity as the only recourse. The statistics community has embraced sparsity as a principle for the development of solutions to nearly any problem in high dimensions. Nevertheless, there are signs that the field of statistics is now seeking other paradigms for dealing with high-dimensional problems.

The focus of the discussion in Section 3 is on sufficient dimension reduction (SDR), which is the second of the two approaches to dimension reduction and which derives its name from the similarity between Expression 2 and Fisher's driving condition (Equation 1). That is, Expression 2 is equivalent to the condition $Y \mid \mathbf{X} \sim Y \mid \boldsymbol{\eta}^T \mathbf{X}$, which is in the spirit of Equation 1. SDR studies have been concerned largely with regressions in which many predictors contribute information on the response and the matrix $\boldsymbol{\eta}$ in Expression 2 is unconstrained, although there are SDR methods that allow for sparsity. SDR and sparse methods are also distinguished by their starting points: Sparse methods are largely model-based, whereas SDR methods are largely model-free. The first SDR methods were sliced inverse regression (SIR) (Li 1991) and sliced average variance estimation (SAVE) (Cook & Weisberg 1991). These methods, along with others, will be described in Section 3. It will be seen there that popular SDR methods are related to PCA and, like PCA, they have found important application in a variety of contexts. For instance, Naik et al. (2000) described how to use SIR (Section 3.2.1) (Li 1991) in marketing studies, Chiaromonte & Martinelli (2002) proposed its use for the analysis of microarray data, and Roley & Newman (2008) used inverse regression estimation (Section 3.2.4.2) (Cook & Ni 2005, 2007) as a preprocessor for predicting Eurasian watermilfoil invasions in Minnesota. There is currently a synergy developing between SDR and the theory of active subspaces, which is used for dimension reduction in computer modeling of complex systems (Constantine 2015).

Diagnostic methods developed in the 1970s and 1980s (Cook & Weisberg 1982, Atkinson 1985) can be effective for model construction and model criticism when the number of predictors is small by today's standards, say in the teens or twenties. But they become unwieldy when there are many more predictors. SDR began with the goal of constructing a low-dimensional projective view of the data that contains all of the regression information without the need to prespecify a parametric model (Cook 1998b). If such a view could be constructed then it could be used as a diagnostic and as an aid to model development. It was seen as high dimensional because it was designed for regressions in which there were too many predictors for standard methods of the day. It was not developed originally to handle contemporary high-dimensional problems in which $n \ll p$, although there are now SDR methods that address such settings.

In Section 4, we turn to envelopes, a relatively new approach to dimension reduction that was introduced by Cook et al. (2007), developed for the multivariate linear model by Cook et al. (2010), and extended to general statistical contexts by Cook & Zhang (2015b). Envelopes can be seen as descendants of SDR. SDR is largely model-free, and envelopes can be applied in

model-based or model-free contexts. Nearly all SDR methods are for predictor reduction in univariate (single-response) regressions, but envelopes can be used for response reduction in multivariate (multi-response) regressions. Most importantly, envelope methods can result in massive efficiency gains relative to standard methods, gains that are equivalent to increasing the sample size many times over.

To introduce one fundamental idea underlying envelopes, consider the regression of a univariate response Y on predictors $\mathbf{X} \in \mathbb{R}^p$, with the goal of reducing the dimension of $\mathbf{X}$ by replacing it with $d < p$ proxy predictors $\boldsymbol{\eta}^T\mathbf{X}$, where $\boldsymbol{\eta} \in \mathbb{R}^{p\times d}$. Under the SDR criterion (Expression 2), $\boldsymbol{\eta}^T\mathbf{X}$ must capture all of the information that $\mathbf{X}$ has about Y. Letting $(\boldsymbol{\eta}, \boldsymbol{\eta}_0) \in \mathbb{R}^{p\times p}$ be nonsingular, Expression 2 can be stated equivalently as $Y \perp\!\!\!\perp \boldsymbol{\eta}_0^T\mathbf{X} \,|\, \boldsymbol{\eta}^T\mathbf{X}$. The proxy predictors can be difficult to estimate under this setup if there is strong marginal dependence between Y and $\boldsymbol{\eta}_0^T\mathbf{X}$ or between $\boldsymbol{\eta}^T\mathbf{X}$ and $\boldsymbol{\eta}_0^T\mathbf{X}$. Envelopes get around this potential difficulty by adding a requirement to Expression 2:

$$Y \perp\!\!\!\perp \boldsymbol{\eta}_0^T\mathbf{X} \,|\, \boldsymbol{\eta}^T\mathbf{X} \text{ and } \boldsymbol{\eta}^T\mathbf{X} \perp\!\!\!\perp \boldsymbol{\eta}_0^T\mathbf{X}. \quad 3.$$

The added condition $\boldsymbol{\eta}^T\mathbf{X} \perp\!\!\!\perp \boldsymbol{\eta}_0^T\mathbf{X}$ forces the proxy predictors to be independent of the remaining part of $\mathbf{X}$. The conditions in Expression 3 are equivalent to requiring that $(Y, \boldsymbol{\eta}^T\mathbf{X}) \perp\!\!\!\perp \boldsymbol{\eta}_0^T\mathbf{X}$, which tells us that Y must be marginally independent of $\boldsymbol{\eta}_0^T\mathbf{X}$ and that $\boldsymbol{\eta}_0^T\mathbf{X}$ cannot furnish any information about Y by virtue of an association with $\boldsymbol{\eta}^T\mathbf{X}$. Analyses under the conditions in Expression 3 tend to be sharper than those under Expression 2 because $\boldsymbol{\eta}_0^T\mathbf{X}$ is independent of both Y and $\boldsymbol{\eta}^T\mathbf{X}$. Additionally, Expression 3 can be applied to achieve dimension reduction in traditional parametric models, whereas Expression 2 alone does not allow for much progress in such settings.

Dimension reduction interpreted loosely is a huge area. Many ad hoc methods have been developed by the computer science and machine learning communities (see, e.g., Burges 2009), and it seems that new methods are being proposed every day. The material in this review reflects a subset of the dimension-reduction literature that can be motivated in a manner similar to Fisher's original concept of sufficiency.

2. PRINCIPAL COMPONENT ANALYSIS

2.1. Classical Principal Components

Beginning with a multivariate sample consisting of n p-dimensional vectors $\mathbf{X}_1, \ldots, \mathbf{X}_n$ that were observed as independent copies of a random vector $\mathbf{X} \in \mathbb{R}^p$, how can we construct a d-dimensional proxy sample $\boldsymbol{\Gamma}^T\mathbf{X}_1, \ldots, \boldsymbol{\Gamma}^T\mathbf{X}_n$, $\boldsymbol{\Gamma} \in \mathbb{R}^{p\times d}$ with $d < p$, that in some sense preserves the relevant information in the original data? PCA addresses this question by roughly equating variation with information and then determining the proxy variables to successively maximize variation. Let $\boldsymbol{\gamma}_j$ denote the jth column of $\boldsymbol{\Gamma} = (\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_d)$ so the d proxy variables for the ith observation $\mathbf{X}_i$ are $\boldsymbol{\Gamma}^T\mathbf{X}_i = (\boldsymbol{\gamma}_1^T\mathbf{X}_i, \ldots, \boldsymbol{\gamma}_d^T\mathbf{X}_i)^T$, $i = 1, \ldots, n$, and let $\boldsymbol{\Sigma}_\mathbf{X} = \text{var}(\mathbf{X})$. Then the first proxy variable is chosen to maximize the variation of $\boldsymbol{\gamma}_1^T\mathbf{X}$ by selecting $\boldsymbol{\gamma}_1 = \arg\max_{\ell\in\mathbb{R}^p} \ell^T\boldsymbol{\Sigma}_\mathbf{X}\ell/\ell^T\ell$. Thus, $\boldsymbol{\gamma}_1$ is the eigenvector of $\boldsymbol{\Sigma}_\mathbf{X}$ corresponding to its largest eigenvalue λ_1, and the variance of the first proxy variable is $\text{var}(\boldsymbol{\gamma}_1^T\mathbf{X}) = \lambda_1$. Subsequent proxy variables are constructed in the same way subject to the constraint that $\boldsymbol{\gamma}_k$ is orthogonal to $\boldsymbol{\gamma}_j$ for $j = 1, \ldots, k-1$. In short, the proxy variables $(\boldsymbol{\gamma}_1^T\mathbf{X}, \ldots, \boldsymbol{\gamma}_p^T\mathbf{X})$ are constructed from the eigenvectors $\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_p$ of $\boldsymbol{\Sigma}_\mathbf{X}$ with corresponding eigenvalues $\lambda_1, \ldots, \lambda_p$. The first d proxy variables $(\boldsymbol{\gamma}_1^T\mathbf{X}, \ldots, \boldsymbol{\gamma}_d^T\mathbf{X})$ are the principal components, and the corresponding $\boldsymbol{\gamma}_j$s are often referred to as principal component directions.

The rationale is the same for sample principal components $(\hat{\boldsymbol{\gamma}}_1^T\mathbf{X}, \ldots, \hat{\boldsymbol{\gamma}}_d^T\mathbf{X})$, where the sample directions $\hat{\boldsymbol{\gamma}}_j$ are obtained by using the sample covariance matrix $\widehat{\boldsymbol{\Sigma}}_\mathbf{X}$ constructed from $\mathbf{X}_1, \ldots, \mathbf{X}_n$ in place of $\boldsymbol{\Sigma}_\mathbf{X}$. Scree plots are a popular tool for choosing d, but other methods are available as

well. A comprehensive review of PCA is available from Jolliffe (2002), and Cook (2007) provides a brief historical perspective.

2.2. Probabilistic Principal Components and Some Generalizations

Although there have been many successful applications of PCA, there also was a feeling that they involved a bit of luck, because there was not yet a widely accepted understanding of its statistical operating characteristics. That situation changed when Tipping & Bishop (1999) introduced the first formal probability model that, when estimated by maximum likelihood, gives rise to principal components.

Consider the following model for $\mathbf{X}$:

$$\mathbf{X} = \boldsymbol{\mu} + \boldsymbol{\Theta}\boldsymbol{\theta}\boldsymbol{\nu} + \boldsymbol{\varepsilon}, \tag{4.}$$

where $\boldsymbol{\mu} \in \mathbb{R}^p$, $\boldsymbol{\Theta} \in \mathbb{R}^{p\times d}$ is a nonstochastic semiorthogonal matrix, $\boldsymbol{\theta} \in \mathbb{R}^{d\times d}$ with $d \leq p$ is nonstochastic and unconstrained, the error $\boldsymbol{\varepsilon} \sim N(0, \boldsymbol{\Delta})$, and the $d \times 1$ latent random vector $\boldsymbol{\nu} \sim N(0, \mathbf{I}_d)$ is independent of $\boldsymbol{\varepsilon}$. Neither $\boldsymbol{\Theta}$ nor $\boldsymbol{\theta}$ is identifiable because $\boldsymbol{\Theta}\boldsymbol{\theta} = (\boldsymbol{\Theta}\mathbf{O})(\mathbf{O}^T\boldsymbol{\theta})$ for any orthogonal matrix $\mathbf{O} \in \mathbb{R}^{d\times d}$, resulting in an equivalent model. Similarly, there is no loss of generality in assuming $\text{var}(\boldsymbol{\nu}) = \mathbf{I}_d$. We think of $\boldsymbol{\nu}$ in this model (Equation 4) as representing variation that is caused by latent extrinsic factors, and $\boldsymbol{\varepsilon}$ represents intrinsic variation that would be present if the extrinsic factors were held fixed. In this formulation, the relevant information is deemed to be the part of $\mathbf{X}$ that is affected by the extrinsic factors $\boldsymbol{\nu}$ and the dimension-reduction goal is to extract that part of $\mathbf{X}$. Under Equation 4 (Cook & Forzani 2008b, theorem 2.1),

$$\mathbf{X} | (\boldsymbol{\nu}, \boldsymbol{\Theta}^T\boldsymbol{\Delta}^{-1}\mathbf{X}) \sim \mathbf{X} | \boldsymbol{\Theta}^T\boldsymbol{\Delta}^{-1}\mathbf{X}, \tag{5.}$$

which is in the spirit of Expression 1. Consequently, $\boldsymbol{\Theta}^T\boldsymbol{\Delta}^{-1}\mathbf{X}$ is the part of $\mathbf{X}$ affected by $\boldsymbol{\nu}$ and is the proxy we would like to estimate, presuming that the intrinsic variation is generally of little relevance. As we observe only $\mathbf{X}$, which is normally distributed with mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\Sigma}_\mathbf{X} = \boldsymbol{\Theta}\boldsymbol{\theta}\boldsymbol{\theta}^T\boldsymbol{\Theta}^T + \boldsymbol{\Delta}$, $\boldsymbol{\Theta}$ and $\boldsymbol{\Delta}$ are confounded and it is not possible to estimate $\boldsymbol{\Theta}^T\boldsymbol{\Delta}^{-1}\mathbf{X}$ without bringing in additional information. For instance, if we require $\boldsymbol{\Delta}$ to be a diagonal matrix, then Equation 4 reduces to a traditional factor analysis model, where $\boldsymbol{\nu}$ is a vector of common factors (see, for example, Lawley & Maxwell 1971). The focus of this review is on versions of Equation 4 that lead to principal components.

2.2.1. Probabilistic principal components. In developing probabilistic principal components, Tipping & Bishop (1999) used a simplification of Equation 4 with

$$\boldsymbol{\Delta} = \delta^2\mathbf{I}_p. \tag{6.}$$

In this case, the proxy variables are $\boldsymbol{\Theta}^T\mathbf{X}$. Let $(\boldsymbol{\Theta}, \boldsymbol{\Theta}_0^T) \in \mathbb{R}^{p\times p}$ be an orthogonal matrix so $\boldsymbol{\Theta}_0^T\mathbf{X}$ represents the part of $\mathbf{X}$ that is independent of $\boldsymbol{\nu}$. The semiorthogonal matrix $\boldsymbol{\Theta}$ is still not identifiable but $\mathcal{T} := \text{span}(\boldsymbol{\Theta})$, the column space of $\boldsymbol{\Theta}$, is identifiable and its maximum likelihood estimator is $\text{span}(\hat{\boldsymbol{\gamma}}_1, \ldots, \hat{\boldsymbol{\gamma}}_d)$, the span of the first d eigenvectors of $\widehat{\boldsymbol{\Sigma}}_\mathbf{X}$. Letting $\hat{\boldsymbol{\Theta}} = (\hat{\boldsymbol{\gamma}}_1, \ldots, \hat{\boldsymbol{\gamma}}_d)$, the estimated proxy variables are then the d principal components $\hat{\boldsymbol{\Theta}}^T\mathbf{X} = (\hat{\boldsymbol{\gamma}}_1^T\mathbf{X}, \ldots, \hat{\boldsymbol{\gamma}}_d^T\mathbf{X})^T$. This solution from the Tipping–Bishop model is not unique, as any full rank transformation $\hat{\boldsymbol{\Theta}}^T\mathbf{X} \mapsto \mathbf{A}\hat{\boldsymbol{\Theta}}^T\mathbf{X}$ is an equivalent solution because the likelihood produces only an estimator of $\mathcal{T}$ and not a particular basis. This may be important in graphical studies, as the principal components might not give the best visualization of the underlying extrinsic variation.

The form of the intrinsic variation $\boldsymbol{\Delta} = \delta^2\mathbf{I}_p$ says that once we account for the extrinsic variation, only isotropic variation remains. Principal components have been used in face recognition to

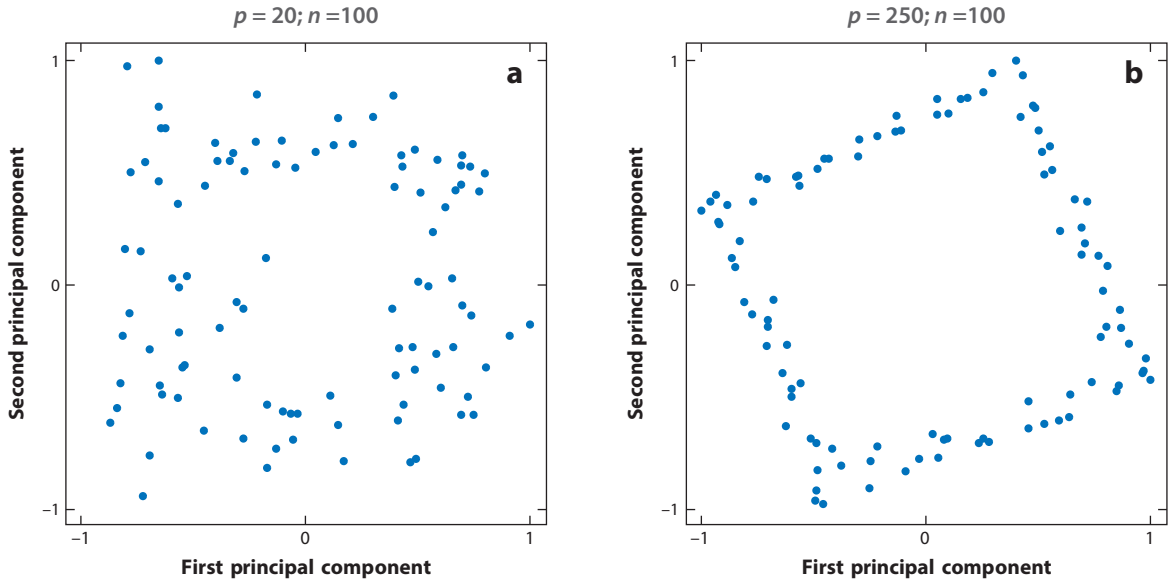

Figure 1

Graphical illustration of probabilistic principal components.

construct eigenface representations of individual images. Beginning with a set of n $r \times c$ grayscale images normalized to have eyes and mouths aligned, the images are then vectorized so each face is represented as a vector $\mathbf{X}$ of length $p = rc$. The mean $\boldsymbol{\mu}$ in Equation 4 represents the average face, and $\boldsymbol{\Theta\theta\nu}$ models the extrinsic variation of individual faces from the average. The error $\boldsymbol{\varepsilon}$ reflects the intrinsic variation in grayscale measurements across images of the same face. The assumption of isotropic variation $\boldsymbol{\Delta} = \delta^2\mathbf{I}_p$ might be reasonable in this case. But it might not be reasonable when the individual coordinates of $\mathbf{X}$ are different measurements (e.g. weight, length, time). In that case, some prefer to standardize each coordinate X_k of $\mathbf{X}$ marginally as $X_k \mapsto (X_k - \bar{X}_k)/[\widehat{\text{var}}(X_k)]^{1/2}$ so each variable has equal marginal variation in the sample. However, there seems to be no theoretical justification for this operation, which is taken mostly as an act of faith, because Expression 5 shows that the proper scaling of $\mathbf{X}$ is $\boldsymbol{\Delta}^{-1}\mathbf{X}$, not $\text{diag}^{-1/2}(\boldsymbol{\Sigma}_\mathbf{X})\mathbf{X}$.

The following simulation example may help to fix ideas and illustrate that the normality of $\boldsymbol{\nu}$ is not essential for useful results. Observations on $\mathbf{X}$ were generated as $\mathbf{X}_i = \mathbf{A}\boldsymbol{\omega}_i + \boldsymbol{\varepsilon}_i, i = 1, \ldots, 100$, where $\boldsymbol{\varepsilon} \sim N(0, \mathbf{I}_p)$, the $\boldsymbol{\omega}_i$s were sampled uniformly from the boundary of the square $[-1, 1]^2$, and the elements of the $p \times 2$ matrix $\mathbf{A}$ were sampled independently from a standard normal distribution. This generating model can be written in the form of Equation 4 by normalizing $\mathbf{A}$ and $\boldsymbol{\omega}$ to give $\boldsymbol{\Theta} = \mathbf{A}(\mathbf{A}^T\mathbf{A})^{-1/2}$, $\boldsymbol{\theta} = (\mathbf{A}^T\mathbf{A})^{1/2}\text{var}^{1/2}(\boldsymbol{\omega})$, and $\boldsymbol{\nu} = \text{var}^{-1/2}(\boldsymbol{\omega})\boldsymbol{\omega}$. The sampling used to construct $\boldsymbol{\Theta}$ is for convenience only; the model is still conditional on its value. The latent vector $\boldsymbol{\nu}$ was chosen to be nonnormal to illustrate the potential robustness of PCA to the assumption that $\boldsymbol{\nu}$ is normally distributed. If PCA works as predicted, then a plot of the sample principal components should recover the square. **Figure 1*a*** shows a plot of the first two principal components for $p = 20$. We see that there is only a hint of the underlying structure. In **Figure 1*b***, $p = 250$, and now the square is clear. It does not align with the coordinate axes because the method is designed to estimate only the subspace $\mathcal{T}$ with isotropic errors. Intuition about why principal components are apparently working in this example can be found in the work of Johnstone & Lu (2009), who studied the asymptotic behavior of the first principal component direction under Equation 4 with

isotropic errors. Basically, if the signal, as measured by a norm of $\boldsymbol{\theta}^T\boldsymbol{\theta}$, continues to grow as p increases with d fixed, the effectiveness of PCA will increase with p.

In the Tipping–Bishop model (Equation 6), the proxy variables are $\boldsymbol{\Theta}^T\mathbf{X}$ and the essential problem is to estimate $\mathcal{T}$. In the next three sections, we present generalizations of Equation 6 that have this same structure.

2.2.2. Minor components. In the Tipping-Bishop model, $\text{var}(\boldsymbol{\Theta}^T\mathbf{X}) = (\boldsymbol{\theta}\boldsymbol{\theta}^T + \delta^2\mathbf{I}_d)$ and $\text{var}(\boldsymbol{\Theta}_0^T\mathbf{X}) = \delta^2\mathbf{I}_{p-d}$. In consequence, the average variation of the extrinsic components of $\mathbf{X}$ will be larger than the average variation of the intrinsic components of $\mathbf{X}$, which partly explains why the proxy variables are always principal components. In some studies it might be useful to consider relaxing this property.

Suppose we model

$$\boldsymbol{\Delta} = \delta^2\boldsymbol{\Theta}\boldsymbol{\Theta}^T + \delta_0^2\boldsymbol{\Theta}_0\boldsymbol{\Theta}_0^T. \qquad 7.$$

The desired reduction is still as given in Expression 5 and the proxy vector is still $\boldsymbol{\Theta}^T\mathbf{X}$, but now there is no implied relation between $\text{var}(\boldsymbol{\Theta}^T\mathbf{X})$ and $\text{var}(\boldsymbol{\Theta}_0^T\mathbf{X})$. In consequence, proxy variables may no longer be principal components. It was shown by Chen (2010) using maximum likelihood estimation of $\mathcal{T}$ that the proxy variables are now either the first d principal components $(\hat{\boldsymbol{\gamma}}_1^T\mathbf{X}, \ldots, \hat{\boldsymbol{\gamma}}_d^T\mathbf{X})$ or the last d minor components $(\hat{\boldsymbol{\gamma}}_{p-d+1}^T\mathbf{X}, \ldots, \hat{\boldsymbol{\gamma}}_p^T\mathbf{X})$.

2.2.3. Extreme components. Suppose next that we expand the model for $\boldsymbol{\Delta}$ as

$$\boldsymbol{\Delta} = \boldsymbol{\Theta}\boldsymbol{\Omega}\boldsymbol{\Theta}^T + \delta_0^2\boldsymbol{\Theta}_0\boldsymbol{\Theta}_0^T, \qquad 8.$$

where $\boldsymbol{\Omega} \in \mathbb{R}^{d\times d}$ is positive definite and the proxy vector is again $\boldsymbol{\Theta}^T\mathbf{X}$. In the $\boldsymbol{\Delta}$ models given in Equations 6 and 7, the extrinsic variation is isotropic, $\text{var}(\boldsymbol{\Theta}^T\boldsymbol{\varepsilon}) = \delta^2\mathbf{I}_d$. But in Equation 8, the extrinsic variation is anisotropic, $\text{var}(\boldsymbol{\Theta}^T\boldsymbol{\varepsilon}) = \boldsymbol{\Omega}$. This difference has a substantial effect on the proxy variables. Welling et al. (2004) and Chen (2010) showed that the maximum likelihood estimator of $\mathcal{T}$ is the span of the first k sample principal component directions and the last $d-k$ sample minor component directions for some $k \in \{0, 1, \ldots, d\}$, and thus the proxy variables will be of the form $(\hat{\boldsymbol{\gamma}}_1^T\mathbf{X}, \ldots, \hat{\boldsymbol{\gamma}}_k^T\mathbf{X}, \hat{\boldsymbol{\gamma}}_{p-(d-k)+1}^T\mathbf{X}, \ldots, \hat{\boldsymbol{\gamma}}_p^T\mathbf{X})$. In this case, k must be selected by maximizing the likelihood.

2.2.4. Moot components. In the models considered so far, the proxy variables $\boldsymbol{\Theta}^T\mathbf{X}$ are independent of the nonproxy variables $\boldsymbol{\Theta}_0^T\mathbf{X}$, which forces useful clarity on the solution. In addition, each of the models asked that the variation of $\boldsymbol{\Theta}^T\mathbf{X}$ or $\boldsymbol{\Theta}_0^T\mathbf{X}$ be isotropic. The most general version of $\boldsymbol{\Delta}$ that we consider keeps the requirement that $\boldsymbol{\Theta}^T\mathbf{X} \perp\!\!\!\perp \boldsymbol{\Theta}_0^T\mathbf{X}$ but no longer requires isotropic variation for either component:

$$\boldsymbol{\Delta} = \boldsymbol{\Theta}\boldsymbol{\Omega}\boldsymbol{\Theta}^T + \boldsymbol{\Theta}_0\boldsymbol{\Omega}_0\boldsymbol{\Theta}_0^T, \qquad 9.$$

where $\boldsymbol{\Omega}_0 \in \mathbb{R}^{(p-d)\times(p-d)}$ is positive definite and $\text{span}(\boldsymbol{\Delta}^{-1}\boldsymbol{\Theta}) = \mathcal{T}$, so the proxy variables are still $\boldsymbol{\Theta}^T\mathbf{X}$. However, in this case, the estimation of $\text{span}(\boldsymbol{\Theta})$ is problematic because the likelihood in $\mathcal{T}$ is maximized by the span of any subset of d component directions. In other words, the likelihood equally supports all subsets of d components from $\{\hat{\boldsymbol{\gamma}}_1^T\mathbf{X}, \ldots, \hat{\boldsymbol{\gamma}}_p^T\mathbf{X}\}$. Equation 9 then represents a limit on what can be achieved by the line of reasoning initiated by Tipping & Bishop (1999).

2.2.5. Berkeley guidance study. In this section we illustrate principal and minor components using data from the Berkeley Guidance Study (Tuddenham & Snyder 1954), which involved

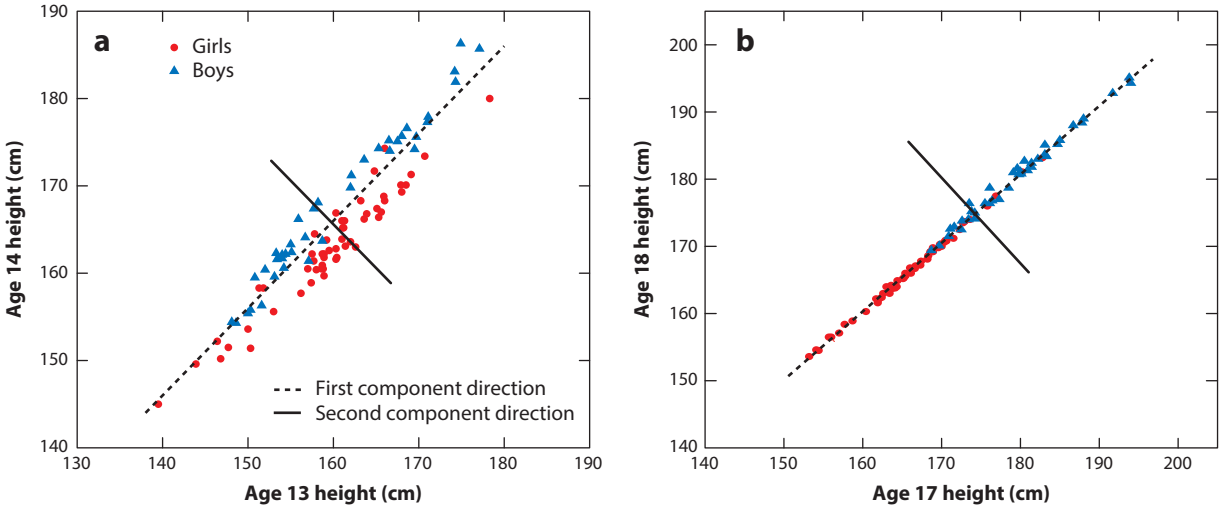

Figure 2

Berkeley guidance study. The two plots give graphical illustrations of the performance of principal component reduction.

monitoring the growth of children born in Berkeley, California, between 1928 and 1929. Data on $n = 93$ children, 39 boys and 54 girls, are available. To facilitate visualization, we selected $p = 2$ variables, the heights of children at ages 13 and 14, and took $d = 1$. Although gender is known, we imagine studying these heights without that knowledge, taking gender as the latent extrinsic factor v. Because $d = 1$, the covariance matrices for extreme (Equation 8) and moot (Equation 9) components reduce to that for minor components (Equation 7).

A plot of the data along with the first (principal) and second (minor) component directions is shown in **Figure 2*a***. Reduction using the first principal component, constructed by projecting the data onto the first principal component direction, would confound the gender information. Reduction using the minor component, constructed by projecting the data onto the minor component direction, would effectively extract the part of $\mathbf{X}$ relating to gender. **Figure 2*b*** gives a plot of heights of children at ages 17 and 18. In this case, reduction by the first principal component would capture the differences due to gender. Although we have imagined a binary latent variable to facilitate visualization, the moment structure (Equation 7) still holds, so normality of the latent variable is again not essential for useful results.

2.3. Principal Fitted Components

As a transition to regression in Section 3, we now turn to the settings where $\mathbf{X}$ is a vector of predictors that we would like to reduce before performing a regression with univariate response Y. Such reduction might be useful in practice because it could mitigate the effects of collinearity, facilitate model specification by allowing visualization of the regression in low dimensions, and provide a relatively small set of predictors on which to base prediction or interpretation. In this section we review predictor reduction based on adapting the formulation described at the outset of Section 2.2.

In Equation 4, $\boldsymbol{\nu}$ is a random vector that is intended to capture extrinsic latent structure that contributes to the variation in $\mathbf{X}$. The extrinsic structure of interest in regression stems from

the response, so we are specifically interested in extracting the part of $\boldsymbol{\nu}$ that is associated with Y. Because the observed responses are known, we can adopt a regression formulation starting with the conditional mean of Equation 4: $E(\mathbf{X}\,|\,Y) = \boldsymbol{\mu} + \boldsymbol{\Theta}\boldsymbol{\theta}E(\boldsymbol{\nu}\,|\,Y)$. We next model $E(\boldsymbol{\theta}\boldsymbol{\nu}\,|\,Y) = \boldsymbol{\beta}\mathbf{f}(Y)$, where $\mathbf{f} \in \mathbb{R}^r$ is a known user-specified vector-valued function of the response and $\boldsymbol{\beta} \in \mathbb{R}^{d\times r}$ with $r \geq d$, leading to the model

$$\mathbf{X}_i = \boldsymbol{\mu} + \boldsymbol{\Theta}\boldsymbol{\beta}\mathbf{f}(Y_i) + \boldsymbol{\varepsilon}_i,\ i = 1, \ldots, n. \tag{10.}$$

This model can be seen as a targeted form of Equation 4 that is conditioned on the responses Y_i, still assuming that $\boldsymbol{\varepsilon} \sim N(0, \boldsymbol{\Delta})$, although here $\boldsymbol{\Delta}$ need not be the same as in Equation 4. Methods for selecting $\mathbf{f}$ were discussed by Cook (2007), Cook & Forzani (2008b), and Adragni (2009). There is considerable flexibility in the choice of $\mathbf{f}$ as a model for $E(\boldsymbol{\theta}\boldsymbol{\nu}\,|\,Y)$: As long as the $d \times r$ matrix of correlations between the elements of $E(\boldsymbol{\theta}\boldsymbol{\nu}\,|\,Y)$ and $\mathbf{f}(Y)$ has rank d, the methods described below produce $\sqrt{n}$-consistent estimators (Cook & Forzani 2008b, theorem 3.5). Essentially, the approximation $\mathbf{f}(Y)$ has to be sufficiently correlated with its target.

The proxy variables under Equation 10 are $\boldsymbol{\Theta}^T\boldsymbol{\Delta}^{-1}\mathbf{X}$; that is,

$$\mathbf{X}\,|\,(Y, \boldsymbol{\Theta}^T\boldsymbol{\Delta}^{-1}\mathbf{X}) \sim \mathbf{X}\,|\,\boldsymbol{\Theta}^T\boldsymbol{\Delta}^{-1}\mathbf{X}. \tag{11.}$$

As with principal components, the maximum likelihood estimator of span$(\boldsymbol{\Delta}^{-1}\boldsymbol{\Theta})$ depends on any additional constraints on $\boldsymbol{\Delta}$. Let $\widehat{\boldsymbol{\Sigma}}_{\text{fit}}$ and $\widehat{\boldsymbol{\Sigma}}_{\text{res}}$ denote the sample covariance matrix of the fitted and residual vectors from the ordinary least squares fit of $\mathbf{X}$ on $\mathbf{f}(Y)$. If we set $\boldsymbol{\Delta} = \delta^2\mathbf{I}_p$, as in the Tipping–Bishop formulation, then span$(\boldsymbol{\Delta}^{-1}\boldsymbol{\Theta}) = \mathcal{T}$ and the maximum likelihood estimator of $\mathcal{T}$ is the span of the first d principal component directions $\hat{\boldsymbol{\gamma}}_{1,\text{fit}}, \ldots, \hat{\boldsymbol{\gamma}}_{d,\text{fit}}$ from $\widehat{\boldsymbol{\Sigma}}_{\text{fit}}$ (Cook 2007, Cook & Forzani 2008b). The corresponding reductions $\hat{\boldsymbol{\gamma}}_{j,\text{fit}}^T\mathbf{X}$ are called principal fitted components. Similarly, there are variations that correspond to the $\boldsymbol{\Delta}$-structures discussed in Sections 2.2.2–2.2.4. But principal fitted components do not get stuck at moot components, because information is coming from the conditional distributions of $\mathbf{X}\,|\,Y$ and is not restricted to the marginal distribution of $\mathbf{X}$. In consequence, it is possible to estimate span$(\boldsymbol{\Delta}^{-1}\boldsymbol{\Theta})$ without restricting the error covariance matrix $\boldsymbol{\Delta}$. With $\boldsymbol{\Delta} > 0$ unrestricted, the maximum likelihood estimator of span$(\boldsymbol{\Delta}^{-1}\boldsymbol{\Theta})$ is $\widehat{\boldsymbol{\Sigma}}_{\text{res}}^{-1/2}$ times the span of the first d principal component directions of $\widehat{\boldsymbol{\Sigma}}_{\text{res}}^{-1/2}\widehat{\boldsymbol{\Sigma}}_{\text{fit}}\widehat{\boldsymbol{\Sigma}}_{\text{res}}^{-1/2}$ (Cook & Forzani 2008b, corollary 3.4). The principal fitted components for a general $\boldsymbol{\Delta}$ can then be described as the first d principal fitted components from the ordinary least squares fit of $\widehat{\boldsymbol{\Sigma}}_{\text{res}}^{-1/2}\mathbf{X}$ on $\mathbf{f}$, operationally assuming $\boldsymbol{\Delta} = \delta^2\mathbf{I}_p$.

Our discussion of principal fitted components with general $\boldsymbol{\Delta}$ has so far required $\widehat{\boldsymbol{\Sigma}}_{\text{res}} > 0$ and thus we need $n > p$. Versions for high-dimensional regressions in which we may have $n < p$ were studied by Cook et al. (2012). Their treatment allows for a variety of different estimators, depending on the assumed form of $\boldsymbol{\Delta}$. They obtained particularly promising results using the sparse permutation-invariant covariance estimator of Rothman et al. (2008) to estimate $\boldsymbol{\Delta}$.

3. SUFFICIENT DIMENSION REDUCTION

Consider the following standard regression scenarios, where in each case Y and $\mathbf{X}$ are jointly distributed, the error ϵ if present is independent of $\mathbf{X}$ and has mean 0 and finite variance, and the $\boldsymbol{\beta}$s are all vectors:

1. Linear regression: $Y = \alpha + \boldsymbol{\beta}^T\mathbf{X} + \epsilon$
2. Nonlinear regression: $Y = \alpha + \alpha_1 \exp(-\boldsymbol{\beta}^T\mathbf{X}) + \epsilon$
3. Logistic regression: $\text{logit}(p) = \alpha + \boldsymbol{\beta}^T\mathbf{X}$

4. Cox model with hazard function $\lambda(t|\mathbf{X}) = \lambda_0(t)\exp(\boldsymbol{\beta}^T\mathbf{X})$
5. Heteroscedastic linear regression: $Y = \alpha + \boldsymbol{\beta}_1^T\mathbf{X} + \sigma(\boldsymbol{\beta}_2^T\mathbf{X})\epsilon$
6. Nonlinear logistic regression: $\text{logit}(p) = \alpha + f(\boldsymbol{\beta}_1^T\mathbf{X}, \boldsymbol{\beta}_2^T\mathbf{X})$, where f is a known function

These models might look very different, but they share a common feature. Models 1–4 each depend on a single linear combination of the predictors $\boldsymbol{\beta}^T\mathbf{X}$ that carries all of the information about the regression that is available from $\mathbf{X}$. Obtaining a good estimator of $\boldsymbol{\beta}$ typically requires information about the model. In contrast, SDR pursues estimation of span($\boldsymbol{\beta}$) without knowledge of the model, so models 1–4 are indistinguishable from the SDR perspective (Expression 2). Scenarios 5 and 6 are more general because they depend on two linear combinations of the predictors that carry all of the information about the regression that is available from $\mathbf{X}$. Again, estimating $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ requires information about the specific regression, whereas SDR methods allow estimation of span($\boldsymbol{\beta}_1, \boldsymbol{\beta}_2$) without such knowledge. The overarching goal of SDR is to estimate the fewest proxy variables $\boldsymbol{\eta}^T\mathbf{X}$ that can serve as a substitute for $\mathbf{X}$ in the regression without loss of information and without prespecifying a parametric model. In models 1–4, $\boldsymbol{\eta}$ is any basis for span($\boldsymbol{\beta}$), and in models 5 and 6, $\boldsymbol{\eta}$ is any basis for span($\boldsymbol{\beta}_1, \boldsymbol{\beta}_2$).

In the following sections, we expand on these informal ideas and discuss basic methods of estimation, beginning by giving a definition of our estimative target, the central subspace (Cook 1994b).

3.1. The Central Subspace

SDR is related in spirit to the principal component methods discussed previously, although its beginnings predate those of probabilistic principal components. As described in the preamble to this section, the informal goal of SDR is to estimate the fewest proxy variables $\boldsymbol{\eta}^T\mathbf{X}$ that satisfy Expression 2, without prespecifying a parametric model. However, Expression 2 holds if and only if $Y \perp\!\!\!\perp \mathbf{X}|(\boldsymbol{\eta}\mathbf{A})^T\mathbf{X}$, where $\mathbf{A}$ is a $d \times d$ nonsingular matrix. This indicates that without additional structure, $\boldsymbol{\eta}$ is not identifiable, but span($\boldsymbol{\eta}$) is identifiable, and so Expression 2 really serves to characterize a subspace and not a set of coordinates. In the next definition, we restate and expand Expression 2 in terms of a subspace $\mathcal{S}$ (Cook 2007).

Definition 1. A projection $\mathbf{P}_\mathcal{S} : \mathbb{R}^p \mapsto \mathcal{S} \subseteq \mathbb{R}^p$ onto a q-dimensional subspace $\mathcal{S}$ is a sufficient linear reduction if it satisfies at least one of the following three conditions:

1. Inverse reduction, $\mathbf{X}|(Y, \mathbf{P}_\mathcal{S}\mathbf{X}) \sim \mathbf{X}|\mathbf{P}_\mathcal{S}\mathbf{X}$
2. Forward reduction, $Y|\mathbf{X} \sim Y|\mathbf{P}_\mathcal{S}\mathbf{X}$
3. Joint reduction, $Y \perp\!\!\!\perp \mathbf{X}|\mathbf{P}_\mathcal{S}\mathbf{X}$

The subspace $\mathcal{S}$ is then called a dimension-reduction subspace.

Each of the three conditions in this definition conveys the idea that the projected predictor $\mathbf{P}_\mathcal{S}\mathbf{X}$ carries all the information that $\mathbf{X}$ has about Y. Condition 1 is the same as Expression 2, and Condition 1 is the basis for principal fitted components (Expression 11). The three conditions are equivalent when $(Y, \mathbf{X})$ has a joint distribution, which we assume throughout our discussion of SDR. We are then free to determine a reduction inversely or jointly and pass it to the forward regression without additional structure. Proxy predictors can be constructed by selecting a basis $\boldsymbol{\eta}$ for $\mathcal{S}$ and then forming $\boldsymbol{\eta}^T\mathbf{X}$. A sufficient summary plot of Y versus $\boldsymbol{\eta}^T\mathbf{X}$ is often an effective diagnostic and tool for guiding the regression (Cook 1994a, 1996, 1998b). The model selection stage could be bypassed in favor of a nonparametric method to estimate $E(Y \mid \boldsymbol{\eta}^T\mathbf{X})$ (Adragni & Cook 2009), which may be reasonable because in practice q is often inferred to be small. The equivalence of

the three conditions in Definition 1 holds also when $\mathbf{P}_{\mathcal{S}}\mathbf{X}$ is replaced with a general nonlinear reduction (Cook 2007), but in this review we confine discussion mostly to linear reductions.

Most SDR methodology is based on the inverse regression of $\mathbf{X}$ on Y. The rationale for this derives from the equivalence of conditions 1 and 2, which implies that a sufficient linear reduction determined from $\mathbf{X} \mid Y$ can be passed to the forward regression $Y \mid \mathbf{X}$ without specifying the marginal distribution of Y or the conditional distribution of $Y \mid \mathbf{X}$. In many regressions, the response is one-dimensional and the number of predictors is in the teens, twenties, or more. Regressions can be challenging in such cases, when they do not come equipped with a prespecified model. In contrast, the inverse regression $\mathbf{X} \mid Y$ combines p one-dimensional regressions, which are easier to handle.

If $\mathcal{S}$ is a dimension-reduction subspace and $\mathcal{S} \subseteq \mathcal{S}_1$, then $\mathcal{S}_1$ is also a dimension-reduction subspace. Within the class of linear reductions, we would like to find the smallest dimension-reduction subspace (Cook 1994b, 1998b).

Definition 2. The intersection of all dimension-reduction subspaces, when it is itself a dimension-reduction subspace, is called the central subspace, $\mathcal{S}_{Y|\mathbf{X}}$.

The central subspace does not always exist, but it does so under mild regularity conditions (Cook 1998b, Yin et al. 2008). It then satisfies all three conditions of Definition 1 and becomes a well-defined target for SDR studies. The central subspace turns out to be an effective construct and over the past 25 years much work has been devoted to methods for estimating it. Most methods require the so-called linearity and constant covariance conditions on the marginal distribution of the predictors:

- Linearity condition: $E(\mathbf{X} \mid \boldsymbol{\eta}^T \mathbf{X})$ is a linear function of $\boldsymbol{\eta}^T \mathbf{X}$, and
- Constant covariance condition: $\text{var}(\mathbf{X} \mid \boldsymbol{\eta}^T \mathbf{X})$ is nonstochastic,

where we now reserve $\boldsymbol{\eta} \in \mathbb{R}^{p \times d}$ to represent a semiorthogonal basis matrix for $\mathcal{S}_{Y|\mathbf{X}}$ with $d = \dim(\mathcal{S}_{Y|\mathbf{X}})$. These conditions are required only at a basis for $\mathcal{S}_{Y|\mathbf{X}}$ and not over all $\boldsymbol{\eta}$. Diaconis & Freedman (1984) and Hall & Li (1993) showed that almost all projections of high-dimensional data are approximately normal, which is often used as partial justification for these conditions. Of the two conditions, the linearity condition seems to be used most often. It holds if and only if (Cook 1998b, proposition 4.2)

$$E(\mathbf{X} \mid \boldsymbol{\eta}^T \mathbf{X}) - E(\mathbf{X}) = \mathbf{P}^T_{\mathcal{S}_{Y|\mathbf{X}}(\boldsymbol{\Sigma}_\mathbf{X})}(\mathbf{X} - E(\mathbf{X})), \tag{12.}$$

where $\boldsymbol{\Sigma}_\mathbf{X} = \text{var}(\mathbf{X})$ and $\mathbf{P}_{\mathcal{S}_{Y|\mathbf{X}}} = \boldsymbol{\eta}(\boldsymbol{\eta}^T \boldsymbol{\Sigma}_\mathbf{X} \boldsymbol{\eta})^{-1} \boldsymbol{\eta}^T \boldsymbol{\Sigma}$ is the projection onto $\mathcal{S}_{Y|\mathbf{X}}$ in the $\boldsymbol{\Sigma}_\mathbf{X}$ inner product. Cook (1998b) and Li & Wang (2007) provide additional discussion.

Although $\mathcal{S}_{Y|\mathbf{X}}$ is now a widely pursued inferential target in SDR studies, it is hard to estimate fully, even with the linearity and constant covariance conditions, because the dependence of Y on $\mathbf{X}$ could be manifested in any conditional moment. Without further restrictive conditions, most SDR methods are guaranteed to estimate only a subspace $\mathcal{S}_m \subseteq \mathcal{S}_{Y|\mathbf{X}}$, where the subscript m is a temporary stand-in for the particular method. The progression of SDR methods is partly a reflection of the general desire to find methods that estimate a larger share of $\mathcal{S}_{Y|\mathbf{X}}$ under conditions that are weak relative to preceding methods. Authors have frequently avoided complications that arise in theoretical developments when $\mathcal{S}_m$ could be a proper subset of $\mathcal{S}_{Y|\mathbf{X}}$ by assuming the coverage condition $\mathcal{S}_m = \mathcal{S}_{Y|\mathbf{X}}$ for the method m under development. This coverage condition generally implies constraints on the regression that may be hard to delineate fully.

3.2. Methods for Estimating the Central Subspace

3.2.1. Sliced inverse regression. In this section we sketch the rationale for SIR as a method for estimating $\mathcal{S}_{Y|\mathbf{X}}$, outline the methodology and give an example to convey the flavor of SDR.

SIR is driven by condition 3 of Definition 1 and the linearity condition:

$$\begin{aligned} E(\mathbf{X}|Y) - E(\mathbf{X}) &= E[E(\mathbf{X}|Y, \mathbf{P}_{\mathcal{S}_{Y|\mathbf{X}}}\mathbf{X})|Y] - E(\mathbf{X}) \\ &= E[E(\mathbf{X}|\mathbf{P}_{\mathcal{S}_{Y|\mathbf{X}}}\mathbf{X})|Y] - E(\mathbf{X}) \\ &= E\{\mathbf{P}^T_{\mathcal{S}_{Y|\mathbf{X}}(\boldsymbol{\Sigma}_\mathbf{X})}[\mathbf{X} - E(\mathbf{X})]|Y\} \\ &= \mathbf{P}^T_{\mathcal{S}_{Y|\mathbf{X}}(\boldsymbol{\Sigma}_\mathbf{X})}[E(\mathbf{X}|Y) - E(\mathbf{X})]. \end{aligned} \quad 13.$$

The first equality is just an expansion of the conditional expectation $E(\mathbf{X}|Y)$. The second follows from condition 3 of Definition 1, and the third follows from the linearity condition and Equation 12. The fourth equality is just simplification. As the value y of Y varies in its sample space, the centered conditional expectations $E(\mathbf{X}|Y = y) - E(\mathbf{X})$ form a one-dimensional manifold in $\mathbb{R}^p$. The final result (Equation 13) shows that the manifold lies in $\boldsymbol{\Sigma}_\mathbf{X}\mathcal{S}_{Y|\mathbf{X}}$; that is, $E(\mathbf{X}|Y) - E(\mathbf{X}) \in \boldsymbol{\Sigma}_\mathbf{X}\mathcal{S}_{Y|\mathbf{X}}$. Let

$$\mathbf{M} = \text{var}[E(\mathbf{X}|Y)] = E\{[E(\mathbf{X}|Y) - E(\mathbf{X})][E(\mathbf{X}|Y) - E(\mathbf{X})]^T\}.$$

Then $\mathcal{S}_{\text{SIR}} := \boldsymbol{\Sigma}_\mathbf{X}^{-1}\text{span}(\mathbf{M}) \subseteq \mathcal{S}_{Y|\mathbf{X}}$, where in our general notation $\mathcal{S}_m = \mathcal{S}_{\text{SIR}}$. This says that we can find a subset $\mathcal{S}_{\text{SIR}}$ of $\mathcal{S}_{Y|\mathbf{X}}$ in the population by taking the span of the relative eigenvectors ℓ from

$$\mathbf{M}\ell = \lambda\boldsymbol{\Sigma}_\mathbf{X}\ell, \quad 14.$$

with nonzero eigenvalues λ. In practice, the variance $\boldsymbol{\Sigma}_\mathbf{X}$ can be well estimated by using its sample version $\widehat{\boldsymbol{\Sigma}}_\mathbf{X}$ when the sample size n is sufficiently large. But how do we estimate the multivariate inverse regression function $E(\mathbf{X}|Y)$? Li (1991) proposed to estimate it by first replacing Y with a discrete version $\tilde{Y}$ constructed by slicing the range of Y into h contiguous nonoverlapping intervals called slices, so $\tilde{Y} = j$ if Y is in the jth slice, $j = 1, \ldots, h$. This works in the population because $\mathcal{S}_{\tilde{Y}|\mathbf{X}} \subseteq \mathcal{S}_{Y|\mathbf{X}}$, with equality when h is sufficiently large. In other words, it is typically sufficient to study the regression of $\mathbf{X}$ on $\tilde{Y}$ instead of the regression of $\mathbf{X}$ on Y. Now, $E(\mathbf{X}|\tilde{Y} = j)$ can be estimated by averaging the predictor vectors that are concomitants of the discrete responses with $\tilde{Y} = j$. We call this average $\bar{\mathbf{X}}_j$. Letting n_j denote the number of responses in slice j, $f_j = n_j/n$ and $\widehat{\mathbf{M}} = \sum_{j=1}^h f_j(\bar{\mathbf{X}}_j - \bar{\mathbf{X}})(\bar{\mathbf{X}}_j - \bar{\mathbf{X}})^T$, we can estimate $\mathcal{S}_{\text{SIR}}$ as $\text{span}(\hat{\ell}_1 \ldots, \hat{\ell}_d)$, where the $\hat{\ell}_j$s are the principal d relative eigenvectors from $\widehat{\mathbf{M}}\ell = \lambda\widehat{\boldsymbol{\Sigma}}_\mathbf{X}\ell$. The SIR (proxy) predictors are then $\hat{\ell}_1^T\mathbf{X}, \ldots, \hat{\ell}_d^T\mathbf{X}$. Let $\mathbf{X}_{jk}$ denote the predictor associated with the kth response in the jth slice. Cook (2004) showed that the SIR estimator of $\mathcal{S}_{Y|\mathbf{X}}$ could be found also by minimizing over semiorthogonal matrices $\mathbf{B} \in \mathbb{R}^{p\times d}$ and vectors $\mathbf{C}_j \in \mathbb{R}^p$ the least squares objective function

$$F_d(\mathbf{B}, \mathbf{C}) = \sum_{j=1}^h \sum_{k=1}^{n_j} \|\widehat{\boldsymbol{\Sigma}}_\mathbf{X}^{-1/2}(\mathbf{X}_{jk} - \bar{\mathbf{X}}) - \mathbf{B}\mathbf{C}_j\|^2, \quad 15.$$

where $\mathbf{B} \in \mathbb{R}^{p\times d}$ represents a basis for $\boldsymbol{\Sigma}_\mathbf{X}^{1/2}\mathcal{S}_{Y|\mathbf{X}}$ and $\mathbf{C}_j \in \mathbb{R}^p$ represents the corresponding coordinate vector. Standard methods for selecting d were discussed by Li (1991), Cook (1998b) and Bura & Cook (2001).

The SIR method relies on slicing the response. Alternatively, a model could be postulated for the regression of $\mathbf{X}$ on Y, which leads back to principal fitted components in Section 2.3. In that context, $\mathcal{S}_{Y|\mathbf{X}} = \text{span}(\boldsymbol{\Delta}^{-1}\boldsymbol{\Theta})$. Further, Cook & Forzani (2008b, section 4) showed in their study of principal fitted components that SIR provides the maximum likelihood estimator of $\mathcal{S}_{Y|\mathbf{X}}$ when the response is categorical and the conditional distribution of $\mathbf{X}|Y$ is multivariate normal.

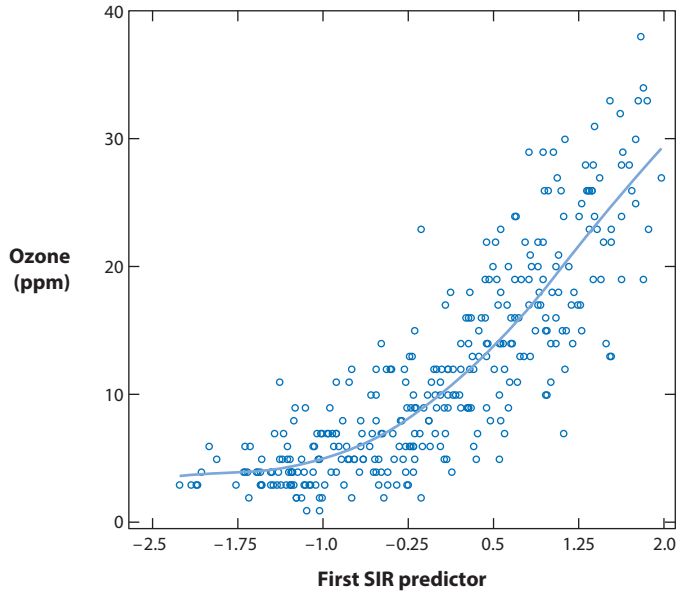

Figure 3

Sliced inverse regression (SIR) summary plot of the response versus the first SIR predictor for the ozone data.

3.2.2. Ozone data. To illustrate the flavor of SDR, we use a widely available dataset on ozone concentration around Upland, CA (Brieman & Friedman 1985). The response is ozone concentration (ppm) and there are eight predictors (temperature, humidity, . . .). Application of SIR with eight slices led to the conclusion that $d = 1$, so it is inferred that a single linear combination of the predictors carries the essential information about the response. **Figure 3** shows a plot of ozone versus the first SIR predictor, $\hat{\ell}_1^T \mathbf{X}$. The line on the plot is a fitted quartic polynomial, so the implied model over the range of the SIR predictor is

$$Y = \alpha + \beta_1(\hat{\ell}_1^T \mathbf{X}) + \beta_2(\hat{\ell}_1^T \mathbf{X})^2 + \beta_3(\hat{\ell}_1^T \mathbf{X})^3 + \beta_4(\hat{\ell}_1^T \mathbf{X})^4 + \epsilon.$$

However, basing inference on this model is likely to be optimistic because it does not recognize the uncertainty in $\hat{\ell}_1$. Instead, we can refit, assuming that $\beta_1 \neq 0$, using the single-index model

$$Y = \alpha + \boldsymbol{\alpha}^T \mathbf{X} + \gamma_2(\boldsymbol{\alpha}^T \mathbf{X})^2 + \gamma_3(\boldsymbol{\alpha}^T \mathbf{X})^3 + \gamma_4(\boldsymbol{\alpha}^T \mathbf{X})^4 + \epsilon,$$

where $\gamma_j = \beta_j/\beta_1^j$, $j = 2, 3, 4$. This model can now be fitted and studied using traditional methods for nonlinear regression.

3.2.3. Sliced average variance estimation. From the discussion in Section 3.2.1 we know that under the linearity condition, the relative eigenvectors of Equation 14 with nonzero eigenvalues are all in $\mathcal{S}_{Y|\mathbf{X}}$. However, these eigenvectors may not span $\mathcal{S}_{Y|\mathbf{X}}$, resulting in a loss of structural information. For an extreme case, suppose that we have three independent standard normal predictors, $\mathbf{X} = (X_1, X_2, X_3)^T$, and that $Y = X_1^2 + \epsilon$ with $\mathbf{X} \perp\!\!\!\perp \epsilon$. Then $E(\mathbf{X} \mid Y) = 0$ and, although $\mathcal{S}_{Y|\mathbf{X}} = \text{span}(1, 0, 0)^T$ is one-dimensional, SIR will estimate the origin, which is of course no help. If we modify the regression to $Y = (X_1 + a)^2 + \epsilon$, where $a \neq 0$, then SIR will give $\mathcal{S}_{Y|\mathbf{X}}$ in the population and will work well in practice if $|a|$ is sufficiently large, but may not do so when $|a|$ is small. SAVE was developed in response to these limitations, and it works well in regressions with strong curvature where SIR fails or has low sensitivity.

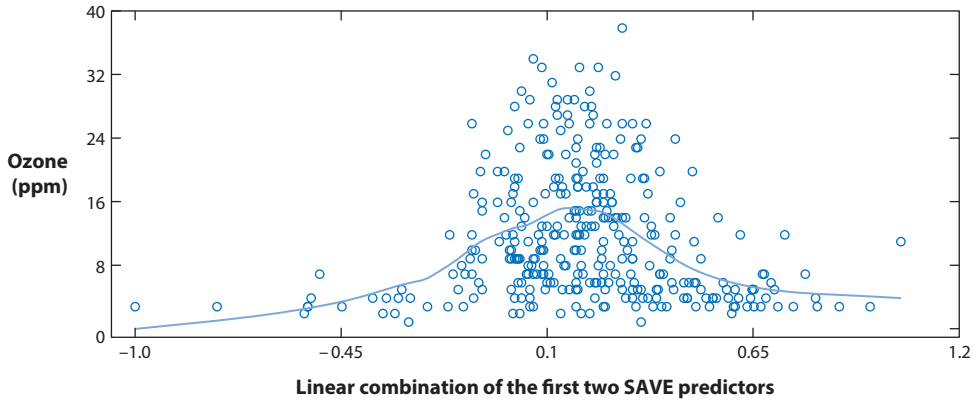

Figure 4

One projective view of the three-dimensional sliced average variance estimation (SAVE) summary plot for the ozone data. The line on the plot is a nonparametric estimate of its mean function. The figure shows the structure that was found by SAVE but missed by sliced inverse regression (SIR) in the analysis leading to **Figure 3**.

Like SIR, SAVE is based on the sliced responses $\tilde{Y}$. Let $\boldsymbol{\Sigma}_{\mathbf{X}|\tilde{Y}} = \text{var}(\mathbf{X} \mid \tilde{Y})$ and let $\mathbf{D}(\tilde{Y}) = \boldsymbol{\Sigma}_{\mathbf{X}} - \boldsymbol{\Sigma}_{\mathbf{X}|\tilde{Y}}$. Then, under the linearity and constant covariance conditions, $\text{span}\{\mathbf{D}(\tilde{Y})\} \subseteq \boldsymbol{\Sigma}_{\mathbf{X}}\mathcal{S}_{Y|\mathbf{X}}$ (Cook & Weisberg 1991, Cook 2000). Let $\mathbf{U} = E(\mathbf{D}^2(\tilde{Y}))$. Then $\mathcal{S}_{\text{SAVE}} := \boldsymbol{\Sigma}_{\mathbf{X}}^{-1}\text{span}\{\mathbf{U}\} \subseteq \mathcal{S}_{Y|\mathbf{X}}$. As in Equation 14, we arrive at a relative eigenvector representation for the SAVE subspace.

$$\mathbf{U}\ell = \lambda\boldsymbol{\Sigma}_{\mathbf{X}}\ell, \qquad 16.$$

and thus the eigenvectors of $\mathbf{U}$ relative to $\boldsymbol{\Sigma}_{\mathbf{X}}$ with nonzero eigenvalues are all in $\mathcal{S}_{Y|\mathbf{X}}$. The sample version of Equation 16 uses the sample version $\widehat{\boldsymbol{\Sigma}}_{\mathbf{X}}$ of $\boldsymbol{\Sigma}_{\mathbf{X}}$ and $\widehat{\mathbf{U}} = \sum_{j=1}^{h} f_j(\widehat{\boldsymbol{\Sigma}}_{\mathbf{X}} - \widehat{\boldsymbol{\Sigma}}_{\mathbf{X}|j})^2$ for $\mathbf{U}$, giving the sample eigen-equations $\widehat{\mathbf{U}}\ell = \lambda\widehat{\boldsymbol{\Sigma}}_{\mathbf{X}}\ell$. The relative eigenvectors $\hat{\ell}_j$ with the largest d eigenvalues from the sample eigen-equations are then used to form the SAVE (proxy) predictors $\hat{\ell}_1^T\mathbf{X}, \ldots, \hat{\ell}_d^T\mathbf{X}$. The dimension of $\mathcal{S}_{\text{SAVE}}$ can be estimated by using permutation tests (Cook 2000, Cook & Yin 2001), chi-squared tests (Shao et al. 2007), or the general SDR procedure developed by Bura & Yang (2011). Shao et al. (2007) also provided conditions under which SAVE satisfies the coverage condition.

Application of SAVE to the ozone data introduced in Section 3.2.2 indicated that $\mathcal{S}_{Y|\mathbf{X}}$ is two- and possibly three-dimensional. One two-dimensional projection of a three-dimensional plot of Y versus the first two SAVE predictors was very similar to the SIR plot in **Figure 3**. **Figure 4** shows a contrasting projection that supports multiple dimensions. The nonparametric estimate of the mean function shown on the plot is nearly symmetric, and this explains why SIR could not find a second direction.

3.2.4. Other methods. SIR and SAVE are the first two methods for estimating at least a portion of the central subspace, and they are still popular as initial choices for dimension reduction. Nevertheless, there are now many more methods available for estimating the central subspace. These newer methods serve to remove the need for the linearity and constant covariance conditions, improve estimation of the central subspace, and expand the scope of SDR methods to cover other statistical problems like time series, functional, and high-dimensional data. In this section we mention a few of these methods.

3.2.4.1. Linearity and constant covariance conditions. The linearity and constant covariance conditions are generally seen as mild, but they are nevertheless essentially uncheckable and thus can be worrisome in application. Several methods that mitigate the need for them are available.

Cook & Nachtsheim (1994) proposed a method for reweighting the predictor vectors so that asymptotically the weighted vectors follow a user-specified distribution that satisfies the linearity and constant covariance conditions. Often, simple power transformations of the predictors can result in a vector of transformation $\mathbf{X}^{(\lambda)} = (X_j^{(\lambda_j)})$ that is close to normal (Cook & Weisberg 1999, section 13.1). Mai & Zou (2015) extended this idea to nonparametric data-driven transformations to multivariate normality and demonstrated good performance for their proposal.

Li & Dong (2009) got around the need for the linearity condition by developing methods via the relationship $E(\mathbf{X}|Y) = E[E(\mathbf{X}|\mathbf{P}_S\mathbf{X})|Y]$ which does not require the linearity condition. Any subspace that satisfies this relation is called a solution space, and the intersection of all solution spaces is called the central solution space, which became their inferential target. If the linearity condition holds in addition, then the central solution space is equal to $\mathcal{S}_{\text{SIR}}$.

Ma & Zhu (2012) proposed a novel approach to dimension-reduction problems by casting them in a semiparametric estimation framework. They were able to remove dependence on the linearity and constant covariance conditions from existing methodology like SIR and SAVE, at the cost of an additional semiparametric regression.

3.2.4.2. Alternative methods of estimation. SIR and SAVE are based on estimators of the first two inverse moments $E(\mathbf{X}\,|\,Y)$ and $\text{var}(\mathbf{X}\,|\,Y)$. If $\text{var}(\mathbf{X}\,|\,Y)$ is constant or varies little, then methods such as SIR or principal fitted components that are based on the first inverse moment should do well. If $\text{var}(\mathbf{X}\,|\,Y)$ varies significantly, then methods like SAVE may be needed. In this section, we describe some of the alternative estimators that are also based on the first two inverse moments. As a class, these methods can be expected to yield useful results in most practical settings.

Let $\hat{\boldsymbol{\xi}}$ be the $p \times h$ matrix with columns $\hat{\boldsymbol{\xi}}_j = \widehat{\boldsymbol{\Sigma}}_{\mathbf{X}}^{-1}(\bar{\mathbf{X}}_j - \bar{\mathbf{X}})$ $(j = 1, \ldots, h)$ and let $\boldsymbol{\xi}$ denote its population version. We know from the discussion of Section 3.2.1 that under the linearity condition $\text{span}(\boldsymbol{\xi}) = \mathcal{S}_{\text{SIR}} \subseteq \mathcal{S}_{Y|\mathbf{X}}$ and in consequence the estimation problem can be framed as minimizing a norm of $\hat{\boldsymbol{\xi}} - \mathbf{BC}$, where $\mathbf{B}$ represents a semiorthogonal basis matrix for $\mathcal{S}_{\text{SIR}}$ and the columns of $\mathbf{C}$ represents coordinates. The columns of $\hat{\boldsymbol{\xi}}$ are dependent and have different variances, so the choice of norm will affect the estimator. Cook & Ni (2005, 2007) studied estimators obtained by minimizing an objective function of the form

$$F_d(\mathbf{B}, \mathbf{C}\,|\,\mathbf{R}_n, \mathbf{V}_n) = [\text{vec}(\hat{\boldsymbol{\xi}}\mathbf{R}_n) - \text{vec}(\mathbf{BC})]^T \mathbf{V}_n [\text{vec}(\hat{\boldsymbol{\xi}}\mathbf{R}_n) - \text{vec}(\mathbf{BC})],$$

where the choice of the $p \times l$ matrix $\mathbf{R}_n$ and the $pl \times pl$ positive definite matrix $\mathbf{V}_n$ determine the estimator. They described how to choose $\mathbf{R}_n$ and $\mathbf{V}_n$ to get the optimal estimator from this class and also showed that SIR corresponds to the suboptimal choice $F_d(\mathbf{B}, \mathbf{C})$ given in Equation 15. This method was extended for use with censored data by Nadkarni et al. (2011).

Li & Wang (2007) proposed a method called directional regression. They showed that it satisfies the coverage condition under mild constraints, and they argued that it is more accurate than or competitive with all of the previous proposals based on the first two inverse moments. Cook & Forzani (2009) derived the maximum likelihood estimator of $\mathcal{S}_{Y|\mathbf{X}}$ when $\mathbf{X}\,|\,Y$ is normally distributed and both $E(\mathbf{X}\,|\,Y)$ and $\text{var}(\mathbf{X}\,|\,Y)$ are allowed to vary with the response. Their estimators are asymptotically most efficient under normality and were shown to perform well in nonnormal settings. Bura & Forzani (2015) extended this treatment to elliptically contoured predictors, showing that when $\mathbf{X}\,|\,Y$ is nonnormal and elliptical, linear SDR methods must necessarily miss a part of $\mathcal{S}_{Y|\mathbf{X}}$, and that a sufficient reduction of $\mathbf{X}$ has both linear and quadratic components. Bura

et al. (2016) developed SDR methods for regressions in which $\mathbf{X} \mid Y$ follows a known multivariate exponential family. Like Bura & Forzani (2015), their reductions need not be linear. Their methodology may be particularly effective when the predictors are conditionally independent, but it becomes more difficult when a dependence model is required (Cook & Li 2009).

3.2.4.3. Nonlinear reduction. The restriction to linear reduction $\mathbf{P}_{\mathcal{S}_{Y|\mathbf{X}}}\mathbf{X}$ is another potential limitation. Lee et al. (2013) extended the foundations of SDR to allow for nonlinear reduction. Their development hinged on going back to basics to carefully define quantities analogous to the central subspace. For instance, they defined an SDR sigma-field as a sub sigma-field $\mathcal{G}$ of the sigma-field generated by $\mathbf{X}$ so that $Y \perp\!\!\!\perp \mathbf{X} \mid \mathcal{G}$. The unique minimal SDR sigma-field is called the central sigma-field. Their formulation also connects SDR to the classical notions of completeness and minimal sufficiency. This breakthrough, like that of Ma & Zhu (2012), opened new frontiers in dimension reduction that promise further significant advances.

3.2.4.4. Testing predictors, and sparse and high-dimensional sufficient dimension reduction. Although most of the SDR literature is on estimating $\mathcal{S}_{Y|\mathbf{X}}$ when $n \gg p$, there are also advances in testing hypotheses about $\mathcal{S}_{Y|\mathbf{X}}$, sparse estimation of $\mathcal{S}_{Y|\mathbf{X}}$, and estimating $\mathcal{S}_{Y|\mathbf{X}}$ when n is not large relative to p. We mention a few of those advances in this section.

Partition $\mathbf{X}$ into two sets of predictors $\mathbf{X}_1$ and $\mathbf{X}_2$ and consider testing the hypothesis that $Y \perp\!\!\!\perp \mathbf{X}_2 \mid \mathbf{X}_1$. In linear regression, tests of this hypothesis reduce to the usual tests on coefficients of $\mathbf{X}_2$, but here there is no model for $Y \mid \mathbf{X}$ and thus the test must be implemented via $\mathcal{S}_{Y|\mathbf{X}}$. Let $\mathcal{H} = \text{span}[(0, \mathbf{I}_r)^T]$. Then this hypothesis can be expressed equivalently as $\mathbf{P}_{\mathcal{H}}\mathcal{S}_{Y|\mathbf{X}} = \mathcal{O}_p$, where $\mathcal{O}_p$ denotes the origin in $\mathbb{R}^p$. Cook (2004) proposed a theory for this type of test with general user-specified $\mathcal{H}$ and provided a specific implementation for SIR using its least squares objective function (Equation 15). Shao et al. (2007) provided corresponding tests for SAVE.

Li (2007) developed sparse versions of several SDR methods, including SIR and SAVE. Although his methods require $n \gg p$ and are coordinate dependent, they can achieve variable selection and dimension reduction simultaneously. Chen et al. (2010) developed a coordinate-independent approach that enabled them to incorporate many model-free and model-based SDR approaches into a unified framework to implement variable selection within SDR. Li & Yin (2008) proposed a regularized version of SIR based on its least-squares formulation (Equation 15). Their method achieves simultaneous predictor reduction and predictor selection when $n < p$. This is a promising approach, but its theoretical properties are unknown. Wu & Li (2011) studied asymptotic properties of SDR with a diverging number of predictors subject to $p = o(n/\log n)$. Under a novel abundant high-dimensional setting that allows $p > n$, Cook et al. (2012) studied the asymptotic behavior of methods based on the framework of principal fitted components. Although they found estimators with good convergence rates, their development does not include variable selection. Yin & Hilafu (2015) and Hilafu & Yin (2017) developed sequential methods for SDR and variable selection that allow $p > n$. They incorporated well-established SDR methods and demonstrated successful application in high-dimensional data analysis, although it is relatively challenging to establish asymptotic properties.

3.3. Specialized Sufficient Dimension Reduction Methodology

The central subspace is designed to capture all directions in the predictor space that have an impact on the response, but there is also specialized SDR methodology for the conditional mean and the conditional variance.

3.3.1. Central mean subspace. In some regressions, there might be a need to perform dimension reduction to understand the conditional mean $E(Y \mid \mathbf{X})$ only, leaving the variance function and higher moments aside. Dimension reduction for the conditional mean was introduced by Cook & Li (2002) via the following definition.

> **Definition 3.** If a subspace $\mathcal{S} \subseteq \mathbb{R}^p$ has the property that $Y \perp\!\!\!\perp E(Y \mid \mathbf{X}) \mid \mathbf{P}_{\mathcal{S}}\mathbf{X}$, then $\mathcal{S}$ is a mean dimension-reduction subspace for the regression of Y on $\mathbf{X}$.

If the intersection of all mean dimension-reduction subspaces is itself a mean dimension-reduction subspace, it is called the central mean subspace. There are a variety of methods for estimating the central mean subspace, including principal Hessian directions (Li 1992, Cook 1998a) and iterative Hessian transformations (Cook & Li 2002). Other methods include estimating the central mean subspace for time series (Park et al. 2009), Fourier methods (Zhu & Zeng 2006), extensions to multivariate regressions (Zhu & Wei 2015), locally efficient estimators (Ma & Zhu 2014), and adaptive estimators (Xia et al. 2002).

3.3.2. Dimension reduction for covariance matrices. We describe in this section an SDR approach to the problem of comparing covariance matrices $\mathbf{\Sigma}_g > 0, g = 1, \ldots, h$, of a random vector $\mathbf{X} \in \mathbb{R}^p$ observed in each of h subpopulations. Tests for equality and proportionality (Muirhead 2005, Flury 1988, Jensen & Madsen 2004) may be helpful, but more intricate methods are needed when such simple characterizations are inadequate. Perhaps the most common methods are based on modeling the spectral decompositions of the $\mathbf{\Sigma}_g$s to connect their eigenvalues and eigenspaces across subpopulations (Flury 1987, Schott 1999, Boik 2002). Although these approaches can be useful for dimension reduction, their motivation seems to rest not primarily in statistical reasoning, but with the convenience of spectral algebra.

Cook & Forzani (2008a) proposed a different approach based on SDR rationale. Let n_g denote the number of observations from population g and let $\mathbf{S}_g = n_g \widehat{\mathbf{\Sigma}}_g$, where $\widehat{\mathbf{\Sigma}}_g$ is the sample version of $\mathbf{\Sigma}_g$. The goal is now to estimate the smallest collection of proxy predictors $\boldsymbol{\alpha}^T \mathbf{X}$, $\boldsymbol{\alpha} \in \mathbb{R}^{p \times q}$ with $q \leq p$ so that $\boldsymbol{\alpha}^T (\mathbf{X} - E(\mathbf{X} \mid g))$ accounts for all differences in $\mathbf{\Sigma}_g$ across subpopulations. Specifically, we require that for any two populations j and k,

$$\mathbf{S}_j \mid (\boldsymbol{\alpha}^T \mathbf{S}_j \boldsymbol{\alpha}, n_j = m) \sim \mathbf{S}_k \mid (\boldsymbol{\alpha}^T \mathbf{S}_k \boldsymbol{\alpha}, n_k = m). \quad 17.$$

This condition implies that, apart from differences in sample size, the quadratic reduction $\boldsymbol{\alpha}^T \mathbf{S} \boldsymbol{\alpha}$ is sufficient to account for all differences in the covariance matrices. Cook & Forzani (2008a) develop corresponding methodology by assuming that the sample sum of squares matrices $\mathbf{S}_j$ follow independent Wishart distributions. In that case they show that a central subspace exists, because the intersection $\mathcal{C}$ of all subspaces span($\boldsymbol{\alpha}$) that satisfy Expression 17 also satisfies Expression 17. When the Wishart assumption fails, the methodology estimates the central mean subspace for covariance matrices. They also give methodology for testing hypotheses of the form $\mathbf{P}_{\mathcal{H}}\mathcal{C} = \mathcal{O}_p$, where $\mathcal{H}$ is a user-specified subspace, and draw sharp contrasts between their proposal and the spectral approaches to comparing covariance matrices.

4. ENVELOPES

The SDR methods discussed in the previous sections may place constraints on the marginal distribution of $\mathbf{X}$, but they are model-free in the sense that they do not require a prespecified parametric model for $Y \mid \mathbf{X}$. This is advantageous during model development, but it affords little help once a model is adopted. For instance, if we are using the linear model stated in the preamble

to Section 3, then $\mathcal{S}_{Y|\mathbf{X}} = \text{span}(\boldsymbol{\beta})$ and thus SDR offers no progress because it leads back to the estimation of $\boldsymbol{\beta}$. Envelopes can be viewed as specialized forms of SDR that are applicable in model-based analyses and can also be used to improve efficiency in model-free SDR studies. We begin our review with response envelopes (Cook et al. 2010) in the context of multivariate linear regression, which is reviewed briefly in Section 4.1, and later turn to predictors envelopes and other constructions.

4.1. The Multivariate Linear Model

Consider the multivariate linear regression of a response vector $\mathbf{Y} \in \mathbb{R}^r$ on a nonstochastic vector of predictors $\mathbf{X} \in \mathbb{R}^p$ based on n independent observations $(\mathbf{Y}_i, \mathbf{X}_i)$, $i = 1, \ldots, n$. We assume the predictors have been centered so that $\bar{\mathbf{X}} = 0$. The linear model for this regression can be represented in vector form as

$$\mathbf{Y} = \boldsymbol{\alpha} + \boldsymbol{\beta}\mathbf{X} + \boldsymbol{\varepsilon}, \tag{18.}$$

where $\boldsymbol{\alpha} \in \mathbb{R}^r$ is a vector of intercepts, $\boldsymbol{\beta} \in \mathbb{R}^{r\times p}$ is an unconstrained matrix of regression coefficients and $\boldsymbol{\varepsilon} \sim N_r(0, \boldsymbol{\Sigma})$. The errors are assumed to be normal to facilitate exposition; this assumption is not needed for success of the methodology. If $\mathbf{X}$ is stochastic, we still condition on its observed values, because it is ancillary under the model of Equation 18. The ordinary least squares estimator $\mathbf{B}$ of $\boldsymbol{\beta}$ under Equation 18 can be constructed by doing r separate univariate linear regressions, one for each element of $\mathbf{Y}$ on $\mathbf{X}$. The coefficients from the jth regression then form the jth row of $\mathbf{B}$, $j = 1, \ldots, r$. If some elements of $\boldsymbol{\beta}$ are linked over rows, then closed-form expressions for the maximum likelihood estimators of $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ may not be possible. We will use the unconstrained model (Equation 18) in this review, although envelopes can be used in conjunction with other versions as well (Cook & Zhang 2015b).

4.2. Response Envelopes

The motivation for response envelopes comes from allowing for the possibility that there are linear combinations of the response vector whose distribution is invariant to changes in the nonstochastic predictor vector. We refer to such linear combinations of $\mathbf{Y}$ as X-invariants. If X-invariants exist, then formally allowing for them in the model (Equation 18) can result in substantial reduction in estimative variation. The envelope model arises by reparameterizing the multivariate linear model (Equation 18) in terms of the smallest subspace $\mathcal{E}$ of $\mathbb{R}^r$ with the properties that, for all relevant $\mathbf{x}_1$ and $\mathbf{x}_2$,

$$(i)\ \mathbf{Q}_\mathcal{E}\mathbf{Y} \,|\, (\mathbf{X} = \mathbf{x}_1) \sim \mathbf{Q}_\mathcal{E}\mathbf{Y} \,|\, (\mathbf{X} = \mathbf{x}_2) \text{ and } (ii)\ \mathbf{P}_\mathcal{E}\mathbf{Y} \perp\!\!\!\perp \mathbf{Q}_\mathcal{E}\mathbf{Y} \,|\, \mathbf{X}, \tag{19.}$$

where $\mathbf{P}_\mathcal{E}$ is the projection onto $\mathcal{E}$ and $\mathbf{Q}_\mathcal{E} = \mathbf{I}_r - \mathbf{P}_\mathcal{E}$. These properties serve to identify the X-invariant part $\mathbf{Y}$. Condition i stipulates that the marginal distribution of $\mathbf{Q}_\mathcal{E}\mathbf{Y}$ must be unaffected by changes in $\mathbf{X}$. It holds if and only if $\text{span}(\boldsymbol{\beta}) \subseteq \mathcal{E}$. Condition ii requires that $\mathbf{Q}_\mathcal{E}\mathbf{Y}$ be unaffected by changes in $\mathbf{X}$ through an association with $\mathbf{P}_\mathcal{E}\mathbf{Y}$, and it holds if and only if $\text{cov}(\mathbf{P}_\mathcal{E}\mathbf{Y}, \mathbf{Q}_\mathcal{E}\mathbf{Y} \mid \mathbf{X}) = \mathbf{P}_\mathcal{E}\boldsymbol{\Sigma}\mathbf{Q}_\mathcal{E} = 0$. Conditions i and ii together imply that any dependence of $\mathbf{Y}$ on $\mathbf{X}$ must be concentrated in $\mathbf{P}_\mathcal{E}\mathbf{Y}$, the part of $\mathbf{Y}$ that is material to the regression, and $\mathbf{Q}_\mathcal{E}\mathbf{Y}$ is X-invariant and thus immaterial to the regression. The conditions in Expression 19 imply also the first SDR requirement given in Definition 1 with the roles of $\mathbf{X}$ and $\mathbf{Y}$ interchanged: $\mathbf{Q}_\mathcal{E}\mathbf{Y} \mid \mathbf{X} \sim \mathbf{Q}_\mathcal{E}\mathbf{Y} \mid (\mathbf{X}, \mathbf{P}_\mathcal{E}\mathbf{Y}) \sim \mathbf{Q}_\mathcal{E}\mathbf{Y} \,|\, \mathbf{P}_\mathcal{E}\mathbf{Y}$. Consequently, we could have reached the same point by starting with the first SDR requirement and adding additional structure to arrive at Expression 19.

The next two definitions, which do not require the model in Equation 18, formalize the construction of an envelope in general.

Definition 4. A subspace $\mathcal{R} \subseteq \mathbb{R}^r$ is said to be a reducing subspace of $\mathbf{M} \in \mathbb{S}^{r\times r}$ if $\mathcal{R}$ decomposes $\mathbf{M}$ as $\mathbf{M} = \mathbf{P}_\mathcal{R}\mathbf{M}\mathbf{P}_\mathcal{R} + \mathbf{Q}_\mathcal{R}\mathbf{M}\mathbf{Q}_\mathcal{R}$. If $\mathcal{R}$ is a reducing subspace of $\mathbf{M}$, we say that $\mathcal{R}$ reduces $\mathbf{M}$.

This definition is equivalent to that used by Cook et al. (2010). It is common in the literature on invariant subspaces and functional analysis, although the underlying notion of reduction differs from the usual understanding in statistics. Here it is used to guarantee condition *ii* of Expression 19, because the decomposition holds if and only if $\mathbf{P}_\mathcal{R}\mathbf{M}\mathbf{Q}_\mathcal{R} = 0$.

Definition 5. Let $\mathbf{M} \in \mathbb{S}^{r\times r}$ and let $\mathcal{B} \subseteq \text{span}(\mathbf{M})$. Then the $\mathbf{M}$-envelope of $\mathcal{B}$, denoted by $\mathcal{E}_\mathbf{M}(\mathcal{B})$, is the intersection of all reducing subspaces of $\mathbf{M}$ that contain $\mathcal{B}$.

Definition 5 (Cook et al. 2010) is central to the development of envelope methodology. It formalizes the construction of the smallest subspace that satisfies the conditions in Expression 19 by asking for intersection of all subspaces that envelop span($\boldsymbol{\beta}$) and thus satisfy condition *i* from among those that satisfy condition *ii*. Suppose that $\mathbf{M}$ has $q \leq r$ distinct eigenvalues with projections onto the corresponding eigenspaces represented by $\mathbf{P}_k$, $k = 1, \ldots, q$. Then, as shown by Cook et al. (2010),

$$\mathcal{E}_\mathbf{M}(\mathcal{B}) = \sum_{k=1}^{q} \mathbf{P}_i\mathcal{B}. \qquad 20.$$

This result shows that $\mathcal{B}$ is in fact enveloped by using the eigenspaces of $\mathbf{M}$.

Back to the model in Equation 18, let $\mathcal{B} = \text{span}(\boldsymbol{\beta})$ and let $\mathbf{P}_\mathcal{E}$ denote the projection onto $\mathcal{E}_\mathbf{\Sigma}(\mathcal{B})$. The model (Equation 18) can be parameterized in terms of $\mathcal{E}_\mathbf{\Sigma}(\mathcal{B})$ by using a basis. Let $u = \dim(\mathcal{E}_\mathbf{\Sigma}(\mathcal{B}))$ and let $(\mathbf{\Gamma}, \mathbf{\Gamma}_0) \in \mathbb{R}^{r\times r}$ be an orthogonal matrix with $\mathbf{\Gamma} \in \mathbb{R}^{r\times u}$, $\text{span}(\mathbf{\Gamma}) = \mathcal{E}_\mathbf{\Sigma}(\mathcal{B})$, and $\text{span}(\mathbf{\Gamma}_0) = \mathcal{E}^\perp_\mathbf{\Sigma}(\mathcal{B})$. Then the envelope model can be written as

$$\mathbf{Y} = \boldsymbol{\alpha} + \mathbf{\Gamma}\boldsymbol{\eta}\mathbf{X} + \boldsymbol{\varepsilon},\ \mathbf{\Sigma} = \mathbf{\Gamma}\mathbf{\Omega}\mathbf{\Gamma}^T + \mathbf{\Gamma}_0\mathbf{\Omega}_0\mathbf{\Gamma}_0^T. \qquad 21.$$

The coefficient vector $\boldsymbol{\beta} = \mathbf{\Gamma}\boldsymbol{\eta}$, where $\boldsymbol{\eta} \in \mathbb{R}^{u\times p}$ carries the coordinates of $\boldsymbol{\beta}$ relative to the basis matrix $\mathbf{\Gamma}$, and $\mathbf{\Omega} \in \mathbb{R}^{u\times u}$ and $\mathbf{\Omega}_0 \in \mathbb{R}^{(r-u)\times(r-u)}$ are positive definite matrices. It is straightforward to see that this model satisfies Expression 19.

Although Equation 21 depends on several parameters, these are not normally of interest in their own right but are regarded as stepping stones to efficient estimation of $\boldsymbol{\beta}$ and $\mathbf{\Sigma}$. The maximum likelihood estimator of $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}} = \mathbf{P}_{\hat{\mathcal{E}}}\mathbf{B}$, the projection of $\mathbf{B}$ onto the estimated envelope. It is $\sqrt{n}$-consistent and asymptotically normal when the errors have finite fourth moments and the predictors satisfy mild regularity conditions. The residual bootstrap is a useful method for estimating standard errors. The asymptotic covariance matrix of $\hat{\boldsymbol{\beta}}$ under normality was given by Cook et al. (2010). Standard errors can be obtained using this result or the residual bootstrap, and the dimension of the envelope can be selected using an information criterion, cross validation, or a holdout sample. Model averaging over the envelope dimension can also be used to estimate $\boldsymbol{\beta}$ (Eck & Cook 2017).

To illustrate how envelope methodology works, consider a multivariate regression with $r = 2$ responses and a single binary predictor $X \in \{0, 1\}$. We parameterize the model so that $\boldsymbol{\alpha} = E(\mathbf{Y} \mid X = 0)$ and the two coordinates of $\boldsymbol{\beta} = E(\mathbf{Y} \mid X = 1) - E(\mathbf{Y} \mid X = 0)$ are the differences between the population means. Standard methodology will likely perform well when the two conditional

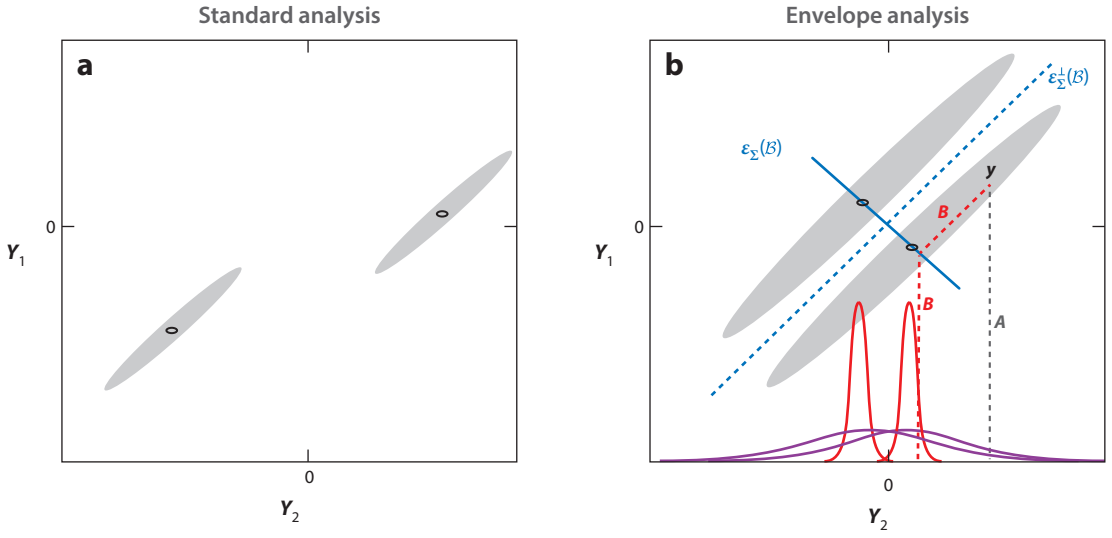

Figure 5

Schematic illustration of regressions with $r = 2$ responses and a single binary predictor. The ellipses in each panel represent the conditional densities of $\mathbf{Y} \mid X$. (*a*) Setting where a standard analysis may work well. (*b*) Setting where an envelope analysis offers substantial gains. The red and purple curves represent marginal densities of Y_2 with and without the envelope. Lines *A* and *B* are illustrative projection paths giving rise to the purple (*A*) and red (*B*) densities.

distributions $\mathbf{Y} \mid (X = 0)$ and $\mathbf{Y} \mid (X = 1)$ are well separated, as illustrated in **Figure 5*a***. However, standard analyses tend to lose efficiency when the distributions are close, as illustrated in **Figure 5*b***. For instance, to infer about the second coordinate $\beta_2 = E(Y_2 \mid X = 1) - E(Y_2 \mid X = 0)$ of $\boldsymbol{\beta}$, a standard likelihood-based analysis begins by projecting the data onto the horizontal axis, as illustrated by path *A* in **Figure 5*b***. Projecting many points produces two densities as represented by the purple curves along the horizontal axis. These densities overlap substantially, so it may take a large sample size to infer that $\beta_2 \neq 0$. An envelope analysis proceeds differently. The envelope, which from Equation 20 is a one-dimensional subspace in this illustration, aligns with the second eigenvector of $\boldsymbol{\Sigma}$ as shown in **Figure 5*b***. The distributions are the same along $\mathcal{E}_{\boldsymbol{\Sigma}}^{\perp}(\mathcal{B})$ and all differences lie in the direction of $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B})$. Envelope methodology makes use of this fact and begins by projecting data onto the envelope to remove the X-invariant variation and then projecting onto the horizontal axis for inference on β_2, resulting in the two marginal envelope distributions shown in red in **Figure 5*b***. The red envelope distributions are well separated and much sharper than the purple distributions, which reflects the efficiency gain from an envelope analysis. There are two noteworthy caveats on **Figure 5*b***. First, because response is two-dimensional, the one-dimensional envelope has to align with one of the eigenvectors of $\boldsymbol{\Sigma}$ for a nontrivial illustration. This is not required in higher dimensions, as shown by Equation 20. Second, the illustration is based on the true envelope. The envelope needs to be estimated in practice, which would have the effect of causing it to wobble in **Figure 5*b***. That wobble produces increased variation of the red envelope distributions. However, regardless of the degree of wobble, the asymptotic variance of the envelope estimators will not exceed the asymptotic variance of the standard estimators, which is reflected by the purple distributions (Cook et al. 2010). In other words, the envelope estimator will always do at least as well as the standard maximum likelihood estimator asymptotically. Examples of data that follow the schematic representation in **Figure 5** were given by Su & Cook (2011), Cook & Zhang (2015b), and others.

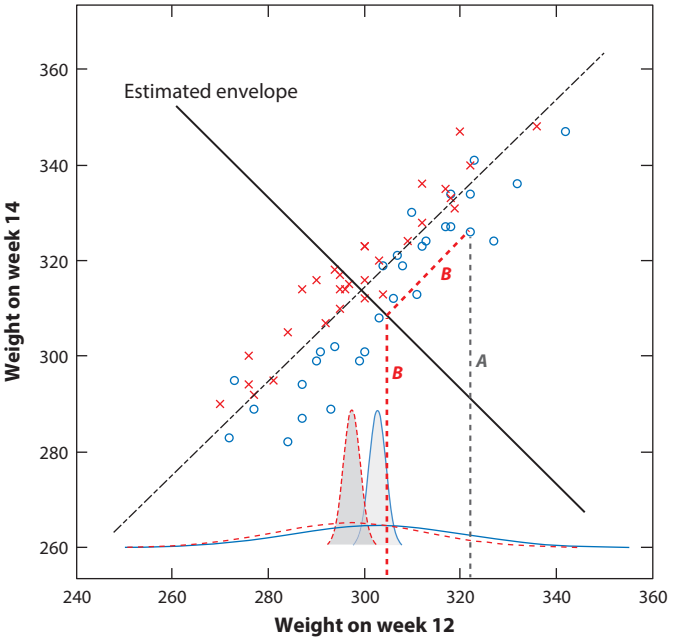

Figure 6

Estimated response envelope for the cattle data. The interpretation of the components in the plot follow that of **Figure 5*b***. The red dashed curves represent the week 12 densities of the projected male weights, indicated by the red crosses, with and without the envelope. The blue solid curves represent the week 12 densities of the projected female weights with and without envelopes. Abbreviations: *A*, representative standard projection path; *B*, representative envelope projection path. Adapted from Cook & Zhang (2015b).

A randomized experiment was conducted to compare two treatments for controlling roundworm in cattle (Kenward 1987). The treatments were randomly assigned to 60 cows, with 30 cows per treatment. Their weights in kilograms were recorded at the beginning of the study prior to treatment application and at two-week intervals thereafter, with the final measurement after a one-week interval. A standard analysis of the regression of the $r = 10$ responses on the binary indicator for treatments failed to show any notable differences between the treatments, but an envelope analysis indicated that $u = 1$ and provided strong evidence that there are treatment differences at week 10 that persisted through the end of the study (Cook & Zhang 2015b). Because there are $r = 10$ responses, an overall graphical construction like that shown in **Figure 5*b*** is not possible. However, the marked plot for week 12 weight versus week 14 weight given in **Figure 6** suggests a clear difference in weights and exhibits the envelope structure represented schematically in **Figure 5*b***. An envelope analysis of these bivariate data again indicates that $u = 1$, leading to envelope standard errors that are approximately 5.7 times smaller than those from the standard model and highly significant coefficient estimates. This means that we would need approximately $n = 2{,}000$ observations for a standard analysis to yield the standard errors that an envelope analysis did with $n = 60$ observations. This level of reduction is commensurate with that for the full data.

4.3. Predictor Envelopes

Following the general development of Cook et al. (2013), we next turn to predictor reduction in Equation 18, where $\mathbf{X}$ is now assumed to be a normal random vector with mean $\boldsymbol{\mu}_{\mathbf{X}}$ and variance

$\boldsymbol{\Sigma}_{\mathbf{X}}$. As with response envelopes, this normality assumption is not needed but is imposed to facilitate discussion. The rationale for predictor envelopes comes from allowing for the possibility that only certain linear combinations of the predictors can impact the distribution of the response vector. This possibility can be introduced into the model of Equation 18 by defining $\mathcal{E}$ to be the smallest subspace with the properties that

$$(i)\ \mathbf{Y} \perp\!\!\!\perp \mathbf{X} \,|\, \mathbf{P}_{\mathcal{E}}\mathbf{X},\ (ii)\ \mathbf{P}_{\mathcal{E}}\mathbf{X} \perp\!\!\!\perp \mathbf{Q}_{\mathcal{E}}\mathbf{X}. \tag{22.}$$

Condition *i* is the same as condition 3 of Definition 1, and it alone is satisfied when $\mathcal{E} = \mathcal{S}_{Y|\mathbf{X}} = \text{span}(\boldsymbol{\beta}^T)$. Condition *ii* forces $\mathcal{S}_{Y|\mathbf{X}} \subseteq \mathcal{E}$, so $\mathcal{E}$ must be a dimension-reduction subspace. It follows from these requirements that predictor envelopes are governed by $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{X}}}(\mathcal{B}')$, which denotes the smallest reducing subspace of $\boldsymbol{\Sigma}_{\mathbf{X}}$ that contains $\mathcal{B}' = \text{span}(\boldsymbol{\beta}^T)$. Let $q = \dim(\mathcal{B}')$ and let $\boldsymbol{\Phi} \in \mathbb{R}^{p\times q}$ denote a semiorthogonal basis matrix for $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{X}}}(\mathcal{B}')$. Because $\mathcal{B}' \subseteq \mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{X}}}(\mathcal{B}')$, we have $\boldsymbol{\beta}^T = \boldsymbol{\Phi}\boldsymbol{\eta}$ for coordinate matrix $\boldsymbol{\eta} \in \mathbb{R}^{q\times r}$. We then rewrite Equation 18 as

$$\mathbf{Y} = \boldsymbol{\alpha} + \boldsymbol{\eta}^T\boldsymbol{\Phi}^T(\mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}}) + \boldsymbol{\varepsilon} \ \text{ with } \ \boldsymbol{\Sigma}_{\mathbf{X}} = \boldsymbol{\Phi}\boldsymbol{\Delta}\boldsymbol{\Phi}^T + \boldsymbol{\Phi}_0\boldsymbol{\Delta}_0\boldsymbol{\Phi}_0^T, \tag{23.}$$

where $\boldsymbol{\Delta} > 0$ and $\boldsymbol{\Delta}_0 > 0$. In contrast to response envelopes, here there are no envelope constraints placed on $\boldsymbol{\Sigma}$. Estimation can now proceed via maximum likelihood and, like response envelopes, the gain in efficiency of the envelope estimator of $\boldsymbol{\beta}$ over the usual maximum likelihood estimator can be substantial.

Cook et al. (2013) showed that there is a connection between partial least squares regression and predictor envelopes. In particular, partial least squares provides a moment-based $\sqrt{n}$-consistent estimator of $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{X}}}(\mathcal{B}')$. They also demonstrated that the likelihood-based estimator from Equation 23 typically does much better than the partial least squares estimator, even when the predictors and errors are not normal. Partial least squares regression has evolved as a core method in chemometrics (Wold et al. 2001), neuroimaging (Krishnan et al. 2011), and operations management (Rönkkö et al. 2016). Envelopes can be seen as a superior replacement for partial least squares in these and many other applications.

4.4. Extensions

Envelope methodology has been extended and adapted for application in several different settings related to the model in Equation 18. Cook & Zhang (2015c) developed methods for simultaneous reduction of both the responses and predictors. Su & Cook (2011) developed partial response envelopes for settings in which selected columns of $\boldsymbol{\beta}$ are of special interest, and Su & Cook (2013a) extended them to comparison of several multivariate populations with different covariance matrices. Envelopes can serve also as supplemental constructions to enhance existing methodology. Cook et al. (2015a) adapted envelopes to improve estimation in reduced rank regression.

Envelope methodology is not generally invariant or equivariant under rescaling of the response or predictors, and for this reason it tends to work best when the responses or predictors are in the same units. Su & Cook (2013b) and Cook & Su (2016) developed scale invariant versions of response and predictor envelopes. Their predictor scaling methodology includes estimating scales for partial least squares regression. Still in the context of Equation 18, Su et al. (2016) extended response envelopes to allow for sparsity and $n < r$, and Khare et al. (2017) developed a Bayesian version of response envelopes. Li & Zhang (2017) extended the model of Equation 18 to allow for array-valued responses, emphasizing application in neuroimaging. Park et al. (2017) developed groupwise envelope models for imaging genetic analysis.

It can happen that $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B}) = \mathbb{R}^r$, in which case there is no response reduction and the envelope estimators reduce to the maximum likelihood estimator from Equation 18. This implies that the

distribution of all linear combinations of $\mathbf{Y}$ are affected by changes in the predictor vector, a conclusion that might be useful in some studies. Nevertheless, efficiency gains might still be possible by using inner envelopes, defined as the largest reducing subspace of $\boldsymbol{\Sigma}$ that is contained within $\mathcal{B}$. This line of reasoning, along with corresponding methodology, was developed by Su & Cook (2012).

So far our review has been in the context of the model shown by Equation 18 or close to it. Cook & Zhang (2015b) extended the idea of envelopes to rather general statistical constructs. Let $\widehat{\boldsymbol{\theta}} \in \mathbb{R}^r$ be an asymptotically normal estimator of $\boldsymbol{\theta}$, $\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \to N(0, \boldsymbol{\Sigma}(\boldsymbol{\theta}, \boldsymbol{\tau}))$, where the asymptotic covariance matrix can depend on the parameter being estimated, $\boldsymbol{\theta}$, and perhaps additional nuisance parameters $\boldsymbol{\tau}$. The estimator $\widehat{\boldsymbol{\theta}}$ could be maximum likelihood, moment, robust, or any estimator as long as it is asymptotically normal. As with the multivariate linear model, the estimator $\widehat{\boldsymbol{\theta}}$ can often be improved by projecting it onto an estimator of $\mathcal{E}_{\boldsymbol{\Sigma}(\boldsymbol{\theta},\boldsymbol{\tau})}(\text{span}(\boldsymbol{\theta}))$.

5. COMPUTING

The dimension-reduction methods emphasized in Sections 2 and 3 are available in the `R` packages `DR` (Weisberg 2002) and `ldr` (Adragni & Raim 2014). Several methods are also available in the MATLAB package `LDR` (Cook et al. 2009). An `R` package for high-dimensional principal fitted components is available from Rothman (2013).

A MATLAB toolbox (Cook et al. 2015b) is available for fitting most of the envelope types discussed in Section 4.4. These envelopes all require optimization of objective functions that are nonconvex and defined on a Grassmannian, and thus are difficult to optimize without good starting values. Cook & Zhang (2015a) developed a sequential one-direction-at-a-time algorithm that is only weakly dependent on starting values and faster than brute force Grassmann optimization. Cook et al. (2016) took a different tack and developed $\sqrt{n}$-consistent starting values for non-Grassmann optimization of a common class of objective functions. The MATLAB code `envlp` is based on the latter algorithm. Martin et al. (2016) developed an R package called `ManifoldOptim` that calls the C++ library ROPTLIB for optimization over Riemannian manifolds. They give examples showing how `ManifoldOptim` can be used to optimize envelope objective functions.

DISCLOSURE STATEMENT

The author is not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

The author thanks Liliana Forzani and the reviewers for helpful comments on an earlier version of this article.

LITERATURE CITED

Adcock RJ. 1878. A problem in least squares. *Analyst* 5:53–54

Adragni KP. 2009. *Some basis functions for principal fitted components*. Work. Pap., Dep. Biostat., Univ. Ala. at Birmingham. **http://userpages.umbc.edu/~kofi/reprints/BasisFunctions.pdf**

Adragni KP, Cook RD. 2009. Sufficient dimension reduction and prediction in regression. *Philos. Trans. R. Soc. Lond. A* 367:4385–405

Adragni KP, Raim AM. 2014. `ldr`: An `R` software package for likelihood-based sufficient dimension reduction. *J. Stat. Softw.* 61:1–21

Alter O, Brown P, Botstein D. 2000. Singular value decomposition for genome-wide expression data processing and modeling. *PNAS* 97:10101–6

Artemiou A, Li B. 2009. On principal components and regression: a statistical explanation of a natural phenomenon. *Stat. Sinica* 19:1557–66
Atkinson AC. 1985. *Plots, Transformations and Regression*. Oxford, UK: Oxford Univ. Press
Boik RJ. 2002. Spectral models for covariance matrices. *Biometrika* 89:159–82
Brieman L, Friedman J. 1985. Estimating optimal transformations for multiple regression and correlation. *J. Am. Stat. Assoc.* 80:580–98
Bura E, Cook RD. 2001. Extending sliced inverse regression: the weighted chi-square test. *J. Am. Stat. Assoc.* 96:996–1003
Bura E, Duarte S, Forzani L. 2016. Sufficient reductions in regressions with exponential family inverse predictors. *J. Am. Stat. Assoc.* 111:1313–29
Bura E, Forzani L. 2015. Sufficient reductions in regressions with elliptically contoured inverse predictors. *J. Am. Stat. Assoc.* 110:420–34
Bura E, Yang J. 2011. Dimension estimation in sufficient dimension reduction: a unifying approach. *J. Multivariate Anal.* 102:130–42
Burges CJC. 2009. Dimension reduction: a guided tour. *Found. Trends Mach. Learn.* 2:275–365
Cavalli-Sforza L, Menozzi P, Piazza A. 1994. *The History and Geography of Human Genes*. Princeton, NJ: Princeton Univ. Press
Chen X. 2010. *Sufficient dimension reduction and variable selection*. PhD Thesis, Univ. Minn.
Chen X, Zou C, Cook RD. 2010. Coordinate-independent sparse sufficient dimension reduction and variable selection. *Ann. Stat.* 38:3696–723
Chiaromonte F. 1996. *A reduction paradigm for multivariate laws*. PhD Thesis, Univ. Minn.
Chiaromonte F. 1997. A reduction paradigm for multivariate laws. In *L_1-Statistical Procedures and Related Topics*, ed. Y Dodge, pp. 229–40. Hayward, CA: Inst. Math. Stat.
Chiaromonte F, Martinelli J. 2002. Dimension reduction strategies for analyzing global gene expression data with a response. *Math. Biosci.* 176:123–44
Constantine PG. 2015. *Active Subspaces: Emerging Ideas for Dimension Reduction in Parameter Studies*. Philadelphia: Soc. Ind. Appl. Math.
Cook RD. 1994a. On the interpretation of regression plots. *J. Am. Stat. Assoc.* 89:177–89
Cook RD. 1994b. Using dimension-reduction subspaces to identify important inputs in models of physical systems. *1994 Proc. Sect. Phys. Eng. Sci.*, pp. 18–25. Alexandria, VA: Am. Stat. Assoc.
Cook RD. 1996. Graphics for regressions with a binary response. *J. Am. Stat. Assoc.* 91:983–92
Cook RD. 1998a. Principal Hessian directions revisited. *J. Am. Stat. Assoc.* 93:84–94
Cook RD. 1998b. *Regression Graphics: Ideas for Studying Regressions Through Graphics*. New York: Wiley
Cook RD. 2000. SAVE: a method for dimension reduction and graphics in regression. *Commun. Stat. Theory Methods* 29:2109–21
Cook RD. 2004. Testing predictor contributions in sufficient dimension reduction. *Ann. Stat.* 32:1062–92
Cook RD. 2007. Fisher Lecture: dimension reduction in regression. *Stat. Sci.* 22:1–26
Cook RD, Forzani L. 2008a. Covariance reducing models: an alternative to spectral modelling of covariance matrices. *Biometrika* 95:799–812
Cook RD, Forzani L. 2008b. Principal fitted components for dimension reduction in regression. *Stat. Sci.* 23:485–501
Cook RD, Forzani L. 2009. Likelihood-based sufficient dimension reduction. *J. Am. Stat. Assoc.* 104:197–208
Cook RD, Forzani L, Rothman AJ. 2012. Estimating sufficient reductions of the predictors in abundant high-dimensional regressions. *Ann. Stat.* 40:353–84
Cook RD, Forzani L, Su Z. 2016. A note on fast envelope estimation. *J. Multivariate Anal.* 150:42–54
Cook RD, Forzani L, Tomassi DR. 2009. `ldr`: a package for likelihood-based sufficient dimension reduction. *J. Stat. Softw.* 39:1–20
Cook RD, Forzani L, Zhang X. 2015a. Envelopes and reduced-rank regression. *Biometrika* 102:439–56
Cook RD, Helland IS, Su Z. 2013. Envelopes and partial least squares regression. *J. R. Stat. Soc. B* 75:851–77
Cook RD, Li B. 2002. Dimension reduction for the conditional mean in regression. *Ann. Stat.* 30:455–74
Cook R, Li L. 2009. Dimension reduction in regressions with exponential family predictors. *J. Comput. Graph. Stat.* 18:774–91

Cook RD, Li B, Chiaromonte F. 2007. Dimension reduction in regression without matrix inversion. *Biometrika* 94:569–84
Cook RD, Li B, Chiaromonte F. 2010. Envelope models for parsimonious and efficient multivariate linear regression. *Stat. Sinica* 20:927–60
Cook RD, Nachtsheim C. 1994. Reweighting to achieve elliptically contoured covariates in regression. *J. Am. Stat. Assoc.* 89:592–99
Cook RD, Ni L. 2005. Sufficient dimension reduction via inverse regression. *J. Am. Stat. Assoc.* 100:410–28
Cook RD, Ni L. 2007. A robust inverse regression estimator. *Stat. Probab. Lett.* 77:343–49
Cook RD, Su Z. 2016. Scaled predictor envelopes and partial least-squares regression. *Technometrics* 58:155–65
Cook RD, Su Z, Yang Y. 2015b. `envlp`: A MATLAB toolbox for computing envelope estimators in multivariate analysis. *J. Stat. Softw.* 62:1–20
Cook RD, Weisberg S. 1982. *Residuals and Influence in Regression*. London: Chapman and Hall
Cook RD, Weisberg S. 1991. Sliced inverse regression for dimension reduction: comment. *J. Am. Stat. Assoc.* 86:328–32
Cook RD, Weisberg S. 1999. *Applied Regression Including Computing and Graphics*. New York: Wiley
Cook RD, Yin X. 2001. Special invited paper: dimension reduction and visualization in discriminant analysis (with discussion). *Aust. N. Z. J. Stat.* 43:147–99
Cook RD, Zhang X. 2015a. Algorithms for envelope estimation. *J. Comput. Graph. Stat.* 25:284–300
Cook RD, Zhang X. 2015b. Foundations for envelope models and methods. *J. Am. Stat. Assoc.* 110:599–611
Cook RD, Zhang X. 2015c. Simultaneous envelopes for multivariate linear regression. *Technometrics* 57:11–25
Diaconis P, Freedman D. 1984. Asymptotics of graphical projection pursuit. *Ann. Stat.* 12:793–815
Eck DJ, Cook RD. 2017. Weighted envelope estimation to handle variability in model selection. arXiv:1701.00856 [stat.ME]
Edgeworth FY. 1884. On the reduction of observations. *Philos. Mag.* 17:135–41
Fisher RA. 1922. On the mathematical foundations of theoretical statistics. *Philos. Trans. R. Soc. A* 222:309–68
Flury B. 1987. Two generalizations of the common principal component model. *Biometrika* 74:59–69
Flury B. 1988. *Common Principal Components and Related Multivariate Models*. New York: Wiley
Hall P, Li KC. 1993. On almost linearity of low dimensional projections from high dimensional data. *Ann. Stat.* 21:867–89
Hilafu H, Yin X. 2017. Sufficient dimension reduction and variable selection for large-*p*-small-*n* data with highly correlated predictors. *J. Comput. Graph. Stat.* 26:26–34
Hotelling H. 1933. Analysis of a complex statistical variable into principal components. *J. Educ. Psychol.* 24:417–41
Jensen ST, Madsen J. 2004. Estimation of proportional covariance matrices in the presence of certain linear restrictions. *Ann. Stat.* 32:219–32
Johnstone IM, Lu AY. 2009. On consistency and sparsity for principal components analysis in high dimensions. *J. Am. Stat. Assoc.* 104:689–93
Jolliffe IT. 2002. *Principal Component Analysis*. New York: Springer
Kenward MG. 1987. A method for comparing profiles of repeated measurements. *J. R. Stat. Soc. C* 36:296–308
Khare K, Pal S, Su Z. 2017. A Bayesian approach for envelope models. *Ann. Stat.* 45:196–222
Krishnan A, Williams LJ, McIntosh AR, Abdi H. 2011. Partial least squares (PLS) methods for neuroimaging: a tutorial and review. *NeuroImage* 56:455–75
Lawley DN, Maxwell AE. 1971. *Factor Analysis as a Statistical Method*. New York: Elsevier
Lee KY, Li B, Chiaromonte F. 2013. A general theory for non-linear sufficient dimension reduction: formulation and estimation. *Ann. Stat.* 41:221–49
Li B, Dong Y. 2009. Dimension reduction for nonelliptically distributed predictors. *Ann. Stat.* 37:1272–98
Li B, Wang S. 2007. On directional regression for dimension reduction. *J. Am. Stat. Assoc.* 102:997–1008
Li KC. 1991. Sliced inverse regression for dimension reduction (with discussion). *J. Am. Stat. Assoc.* 86:316–42
Li KC. 1992. On principal Hessian directions for data visualization and dimension reduction: another application of Stein's lemma. *J. Am. Stat. Assoc.* 87:1025–39
Li L. 2007. Sparse sufficient dimension reduction. *Biometrika* 94:603–13
Li L, Yin X. 2008. Sliced inverse regression with regularizations. *Biometrics* 64:124–31

Li L, Zhang X. 2017. Parsimonious tensor response regression. *J. Am. Stat. Assoc.* 112:1131–46
Ma Y, Zhu L. 2014. On estimation efficiency of the central mean subspace. *J. R. Stat. Soc. B* 76:885–901
Ma Y, Zhu LA. 2012. Semiparametric approach to dimension reduction. *J. Am. Stat. Assoc.* 107:168–79
Mai Q, Zou H. 2015. Nonparametric variable transformation in sufficient dimension reduction. *Technometrics* 57:1–10
Martin S, Raim AM, Huang W, Adragni KP. 2016. `ManifoldOptim`: An `R` interface to the ROPTLIB library for Riemannian manifold optimization. arXiv:1612.03930 [stat.CO]
Muirhead RJ. 2005. *Aspects of Multivariate Statistical Theory*. New York: Wiley
Nadkarni NV, Zhao Y, Kosorok MR. 2011. Inverse regression estimation for censored data. *J. Am. Stat. Assoc.* 106:178–90
Naik PA, Hagerty MR, Tsai CL. 2000. A new dimension reduction approach for data-rich marketing environments: sliced inverse regression. *J. Mark. Res.* 37:88–101
Park JH, Sriram TN, Yin X. 2009. Central mean subspace for time series. *J. Comput. Graph. Stat.* 18:717–30
Park Y, Su Z, Hhu H. 2017. Groupwise envelope models for imaging genetic analysis. *Biometrics* **http://dx.doi.org/10.1111/biom.12689**
Pearson K. 1901. On lines and planes of closest fit to systems of points in space. *Philos. Mag.* 2:559–72
Roley SS, Newman RM. 2008. Predicting Eurasian watermilfoil invasions in Minnesota. *Lake Reservoir Manag.* 24:361–69
Rönkkö M, McIntosh CN, Antonakis J, Edwards JR. 2016. Partial least squares path modeling: time for some serious second thoughts. *J. Oper. Manag.* 47–48:9–27
Rothman AJ. 2013. abundant. R software package for abundant regression and high-dimensional principal fitted components. **https://www.r-pkg.org/pkg/abundant**
Rothman AJ, Bickel PJ, Levina E, Zhu J. 2008. Sparse permutation invariant covaraince estimation. *Electron. J. Stat.* 2:494–515
Schott JR. 1999. Partial common principal component subspaces. *Biometrika* 86:899–908
Shao Y, Cook RD, Weisberg S. 2007. Marginal tests with sliced average variance estimation. *Biometrika* 94:285–96
Stigler SM. 1973. Studies in the history of probability and statistics. XXXII: Laplace, Fisher and the discovery of the concept of sufficiency. *Biometrika* 60:439–45
Su Z, Cook RD. 2011. Partial envelopes for efficient estimation in multivariate linear regression. *Biometrika* 98:133–46
Su Z, Cook RD. 2012. Inner envelopes: efficient estimation in multivariate linear regression. *Biometrika* 99:687–702
Su Z, Cook RD. 2013a. Estimation of multivariate means with heteroscedastic errors using envelope models. *Stat. Sinica* 23:213–30
Su Z, Cook RD. 2013b. Scaled envelopes: scale-invariant and efficient estimation in multivariate linear regression. *Biometrika* 100:939–54
Su Z, Zhu G, Chen X, Yang Y. 2016. Sparse envelope model: estimation and response variable selection in multivariate linear regression. *Biometrika* 103:579–93
Tipping ME, Bishop CM. 1999. Probabilistic principal component analysis. *J. R. Stat. Soc. B* 61:611–22
Tuddenham RD, Snyder MM. 1954. Physical growth of California boys and girls from birth to age 18. *Univ. Calif. Publ. Child Dev.* 1:183–364
Weisberg S. 2002. Dimension reduction regression in R. *J. Stat. Softw.* 7:1–22
Welling M, Williams C, Agakov F. 2004. Extreme component analysis. In *Advances in Neural Information Processing Systems 16*, ed. S Thrun, SK Saul, B Schölkopf, pp. 137–44. Cambridge, MA: MIT Press
Wold S, Sjöström M, Eriksson L. 2001. PLS-regression: a basic tool of chemometrics. *Chemom. Intell. Lab. Syst.* 58:109–30
Wu Y, Li L. 2011. Asymptotic properties of sufficient dimension reduction with a diverging number of predictors. *Stat. Sinica* 21:707–30
Xia Y, Tong H, Li W, Zhu LX. 2002. An adaptive estimation of dimension reduction space. *J. R. Stat. Soc. Ser. B* 64:363–410
Yin X, Hilafu H. 2015. Sequential sufficient dimension reduction for large *p*, small *n* problems. *J. R. Stat. Soc. Ser. B* 77:879–92

Yin X, Li B, Cook RD. 2008. Successive direction extraction for estimating the central subspace in a multiple-index regression. *J. Multivar. Anal.* 99:1733–57

Zhu L, Wei Z. 2015. Estimation and inference on central mean subspace for multivariate response data. *Comput. Stat. Data Anal.* 92:68–83

Zhu Y, Zeng P. 2006. Fourier methods for estimating the central subspace and the central mean subspace in regression. *J. Am. Stat. Assoc.* 101:1638–51

Annual Review of Statistics and its Application

Volume 5, 2018

Contents

Errata

An online log of corrections to *Annual Review of Statistics and Its Application* articles may be found at http://www.annualreviews.org/errata/statistics