Published in final edited form as:
Int Stat Rev. 2013 April ; 81(1): . doi:10.1111/j.1751-5823.2012.00182.x.

A Review on Dimension Reduction

Yanyuan Ma[1] and **Liping Zhu**[2]

[1]Department of Statistics, Texas A&M University, College Station, TX 77843, USA, ma@stat.tamu.edu

[2]School of Statistics and Management, Shanghai University of Finance and Economics, Key Laboratory of Mathematical Economics (SUFE), Ministry of Education, Shanghai 200433, China, zhu.liping@mail.shufe.edu.cn

Summary

Summarizing the effect of many covariates through a few linear combinations is an effective way of reducing covariate dimension and is the backbone of (sufficient) dimension reduction. Because the replacement of high-dimensional covariates by low-dimensional linear combinations is performed with a minimum assumption on the specific regression form, it enjoys attractive advantages as well as encounters unique challenges in comparison with the variable selection approach. We review the current literature of dimension reduction with an emphasis on the two most popular models, where the dimension reduction affects the conditional distribution and the conditional mean, respectively. We discuss various estimation and inference procedures in different levels of detail, with the intention of focusing on their underneath idea instead of technicalities. We also discuss some unsolved problems in this area for potential future research.

Keywords

Dimension reduction; double robustness; efficiency bound; estimating equation; linearity condition; sliced inverse regression; sufficient dimension reduction

1 Description of the Dimension Reduction Problem

Accompanying the advancement of sciences and technologies, scientific data has the tendency of growing in both size and complexity. One characteristic of such complexity is the sheer amount of available covariates, which makes it difficult to detect the dependence between a response variable and the collection of the covariates.

To reduce the problem of many covariates to one with a few covariates, mainly two different approaches exist in the statistical literature. One is variable selection, where the researchers believe that among all the available covariates, only a few are truly related to the response, all others are redundant and have no real explanatory effect. Literature in variable selection has experienced an explosion recently and we do not further elaborate in this article. The second approach to reduce the number of covariates is the so called (sufficient) dimension reduction, which is the focus of this review here. In contrast to the variable selection approach, dimension reduction approach assumes that the response variable relates to only a few linear combinations of the many covariates. Thus, it could happen that all the covariates have explanatory effect, but the effect is only represented in a few linear combinations. It is the goal of dimension reduction to identify these few linear combinations.

To set notations, we use $Y \in \mathbb{R}$ to represent the response variable, and $\mathbf{x} = (X_1, \ldots, X_p)^{\mathrm{T}} \in \mathbb{R}^p$ to represent the covariates. A main class of dimension reduction problems concern the

distribution of Y conditional on **x**, written as pr($Y \le y \mid \mathbf{x}$). If there exists a $p \times d$ matrix β such that

$$\mathrm{pr}(Y \le y|\mathbf{x}) = \mathrm{pr}(Y \le y|\beta^{\mathrm{T}}\mathbf{x}), \text{for all} y \in \mathbb{R}, \quad (1)$$

that is, conditional on **x**, the distribution of Y is the same as that conditional on $\beta^{\mathrm{T}}\mathbf{x}$, then, as far as the relation between Y and **x** is concerned, the p-dimensional covariates **x** can be replaced by the d-dimensional linear combinations $\beta^{\mathrm{T}}\mathbf{x}$. The dimension reduction model (1) is closely related to the effective dimension reduction model in Li (1991). It implies that Y is statistically independent of **x** when $\beta^{\mathrm{T}}\mathbf{x}$ is given (Cook, 1998), and the replacement of **x** by $\beta^{\mathrm{T}}\mathbf{x}$ retains all the dependence information of Y on **x**. In fact, Zeng & Zhu (2010) formally proved the equivalence of (1), the effective dimension reduction model of Li (1991) and the independence model of Cook (1998). If we can find a so called loading matrix β satisfying (1), we then convert a problem with a p-dimensional covariate to one with a d-dimensional covariate. Typically, d is much smaller than p, hence we achieve the goal of dimension reduction. It is easy to recognize that β is not identifiable, hence the goal of dimension reduction is to identify the central subspace (Cook 1994, 1998), defined as the column space of β which satisfies (1) with the smallest number of columns d. Central subspace is typically denoted by $\mathcal{S}_{Y|\mathbf{x}}$ in the dimension reduction literature.

While the central subspace problem concerns the conditional distribution of Y on **x** hence provides a complete picture of their relation, sometimes one might be only interested in certain aspects of the dependence of Y on **x**. For example, one might be only concerned about the mean of Y conditional on **x**, expressed by $E(Y \mid \mathbf{x})$. In such case, usually a weaker assumption

$$E(Y|\mathbf{x}) = E(Y|\beta^{\mathrm{T}}\mathbf{x}) \quad (2)$$

is made, where β is still a $p \times d$ matrix with d typically much smaller than p. This is referred to as a central mean subspace problem. In comparison to the central subspace assumption, the central mean subspace assumption only concerns the conditional mean function of the response, and it says that the dependence of the conditional mean of Y on **x** is completely described by the dependence of the conditional mean of Y on the d linear combinations $\beta^{\mathrm{T}}\mathbf{x}$. Thus, to study the relationship between the mean of Y and **x**, it suffices to study the relationship between the mean of Y and $\beta^{\mathrm{T}}\mathbf{x}$, which is a multivariate mean regression problem with d, instead of p, covariates. Thus, as long as we find a loading matrix β, we then reduce the p-covariate mean regression problem to a d-covariate mean regression problem. Similar to β in model (1), β in model (2) is also not identifiable. Thus, our interest is in estimating the central mean subspace (Cook & Li, 2002), which is defined as the column space of β satisfying (2) with the smallest number of columns d. Central mean subspace is typically denoted by $\mathcal{S}_{E(Y|\mathbf{x})}$.

Yin & Cook (2002) generalized the idea of the central mean subspace and defined the central k-th moment subspace, which is the column space of a $p \times d$ matrix β with the smallest number of columns d satisfying

$$E(Y^j|\mathbf{x}) = E(Y^j|\beta^{\mathrm{T}}\mathbf{x}), \text{for} j = 1, \ldots, k. \quad (3)$$

Central k-th moment subspace is typically written as $\mathcal{S}^{(k)}_{Y|\mathbf{x}}$. To estimate the conditional variance when **x** is high dimensional, Zhu & Zhu (2009a) introduced the notion of central

variance subspace, defined as the column space of a $p \times d$ matrix β with the smallest number of columns d satisfying

$$\mathrm{var}(Y|\mathbf{x}) = E\{\mathrm{var}(Y|\mathbf{x})|\beta^{\mathrm{T}}\mathbf{x}\}. \quad (4)$$

Central variance subspace is usually denoted as $\mathcal{S}_{\mathrm{var}(Y|\mathbf{x})}$. If we are concerned with some other characteristic of the dependence of Y on $\mathbf{x}$, we can develop the corresponding dimension reduction subspace as well. To deliver the most essential messages, in the following, we focus our discussion on the central subspace $\mathcal{S}_{Y|\mathbf{x}}$ and central mean subspace $\mathcal{S}_{E(Y|\mathbf{x})}$ mainly.

2 The Current Approaches in Literature

The description of the dimension reduction problems in Section 1 indicates that finding the column space of β with the smallest dimension d is the target of the dimension reduction literature. In other words, dimension reduction is considered as a problem of estimating a space, instead of the more classic statistical problem of estimating parameters. Except for some degenerated cases, the smallest subspace usually exists and is uniquely defined (Cook, 2004). Another useful observation is that if we set $\mathbf{z} = \mathbf{V}^{-1}(\mathbf{x} - \mathbf{u})$ for any symmetric invertible matrix $\mathbf{V}$ and any p-dimensional vector $\mathbf{u}$, then $\mathrm{pr}(Y|\beta^{\mathrm{T}}\mathbf{x}) = \mathrm{pr}\{Y|(\mathbf{V}\beta)^{\mathrm{T}}\mathbf{z}\}$, $E(Y^j|\beta^{\mathrm{T}}\mathbf{x}) = E\{Y^j|(\mathbf{V}\beta)^{\mathrm{T}}\mathbf{z}\}$ and $\mathrm{var}(Y|\beta^{\mathrm{T}}\mathbf{x}) = \mathrm{var}\{Y|(\mathbf{V}\beta)^{\mathrm{T}}\mathbf{z}\}$. Thus it follows immediately that $\mathcal{S}_{Y|\mathbf{x}} = \mathbf{V}^{-1}\mathcal{S}_{Y|\mathbf{z}}$, $\mathcal{S}_{E(Y|\mathbf{x})} = \mathbf{V}^{-1}\mathcal{S}_{E(Y|\mathbf{z})}$, $\mathcal{S}^{(k)}_{Y|\mathbf{x}} = \mathbf{V}^{-1}\mathcal{S}^{(k)}_{Y|\mathbf{z}}$ and $\mathcal{S}_{\mathrm{var}(Y|\mathbf{x})} = \mathbf{V}^{-1}\mathcal{S}_{\mathrm{var}(Y|\mathbf{z})}$. This is the celebrated invariance property coined by Cook (1998). This property ensures that centring and normalizing the covariates $\mathbf{x}$ does not change the nature of the dimension reduction problem. Therefore for simplicity, we assume without loss of generality that the covariates $\mathbf{x}$ have mean zero and the identity variance–covariance matrix $\mathbf{I}_p$ in our subsequent exposition.

Generally speaking, there are mainly three classes of methods available in the literature for estimating the column space of β: the inverse regression based methods, the non-parametric methods and the semiparametric methods, respectively. We review these three classes of methods for estimating $\mathcal{S}_{Y|\mathbf{x}}$ and $\mathcal{S}_{E(Y|\mathbf{x})}$ in detail. The methods for estimating $\mathcal{S}^{(k)}_{Y|\mathbf{x}}$ and $\mathcal{S}_{\mathrm{var}(Y|\mathbf{x})}$ are similar hence are omitted.

2.1 Inverse Regression Based Methods

Inverse regression methods form the oldest class of dimension reduction methods and are still under active development currently, see Adragni & Cook (2009) for a comprehensive and detailed review on this class of methods. The origin of inverse regression is an ingenious idea of reversing the relation between the response variable and the covariates (Li, 1991; Li & Duan, 1991). In this class, instead of considering distributions or expectations of functions of Y conditional on $\mathbf{x}$, which suffers the well-known curse of dimensionality when $\mathbf{x}$ is high dimensional, the methods consider expectations of functions of $\mathbf{x}$ conditional on Y, which is suddenly a low dimensional problem because Y is univariate. The inverse regression based methods often rely on some additional assumptions on the covariates to link the low dimensional problem and the original high dimensional problem. These additional assumptions include the linearity condition:

$$E(\mathbf{x}|\beta^{\mathrm{T}}\mathbf{x}) = \mathbf{P}\mathbf{x}, \quad (5)$$

which is indispensable for almost all inverse regression based methods. For some methods in this class, the constant variance condition is also needed, which requires

$$\mathrm{cov}(\mathbf{x}|\beta^{\mathrm{T}}\mathbf{x})=\mathbf{Q}. \quad (6)$$

In the above displays (5) and (6), $\mathbf{P} = \beta(\beta^{\mathrm{T}}\beta)^{-1}\beta^{\mathrm{T}}$, $\mathbf{Q} = \mathbf{I}_p - \mathbf{P}$, and β is a basis matrix of the corresponding dimension reduction subspaces.

For the inverse regression based methods to work, assumptions (5) and (6) only need to hold at the true value of β. But β is unknown. Hence to facilitate the practical verification of these conditions, they are often strengthened to hold for all possible β. If the linearity condition holds for all possible β, then (5) implies that $\mathbf{x}$ has an elliptically contoured distribution (Eaton, 1986). Hall & Li (1993) showed that the linearity condition (5) always offers a good approximation to the reality when p diverges to infinity while d remains fixed. If both conditions (5) and (6) are assumed to hold for all possible β, it implies that $\mathbf{x}$ has a multivariate normal distribution. If the covariates $\mathbf{x}$ do not satisfy these two conditions, useful treatment includes transformation (Box & Cox, 1964) and reweighting (Cook & Nachtsheim, 1994). Note that the presence of any categorical variable will violate the elliptical or normal distributional assumption and void the utility of transformation and reweighting techniques, hence the strengthened version of the linearity condition is not suitable in this case. Because these conditions are not always satisfied, and are not easy to check in practice, Li & Dong (2009) and Dong & Li (2010) relaxed the linearity condition to requiring $E(\mathbf{x} \mid \beta^{\mathrm{T}}\mathbf{x})$ to be a polynomial function of $\beta^{\mathrm{T}}\mathbf{x}$, while retaining the constant variance condition (6).

For estimating $\mathcal{S}_{Y|\mathbf{x}}$, the most representative method in the inverse regression class is the sliced inverse regression (SIR, Li, 1991). It uses the eigenvectors associated with the d non-zero eigenvalues of the matrix $\Lambda_{\mathrm{SIR}} \overset{\mathrm{def}}{=} \mathrm{cov}\{E(\mathbf{x} \mid Y)\}$ to recover $\mathcal{S}_{Y|\mathbf{x}}$. In order for SIR to work, the linearity condition (5) is critical, as will be illustrated more clearly in Section 2.3. Inspired by SIR, Zhu & Fang (1996) proposed kernel inverse regression and Fung *et al.* (2002) proposed canonical correlation (CANCOR) analysis. Observing that SIR fails to identify some symmetric patterns, Cook & Weisberg (1991) developed sliced average variance estimation (SAVE) using the eigenvectors associated with the d non-zero eigenvalues of the matrix $\Lambda_{\mathrm{SAVE}} \overset{\mathrm{def}}{=} E\{\mathbf{I}_p - \mathrm{cov}(\mathbf{x} \mid Y)\}^2$. Cook & Lee (1999) proved that SAVE is more comprehensive than SIR, yet SIR is more efficient than SAVE. To combine the advantages of both SIR and SAVE, Zhu *et al.* (2007) suggested using a hybrid of SIR and SAVE through a convex combination $\alpha\Lambda_{\mathrm{SIR}} + (1 - \alpha)\Lambda_{\mathrm{SAVE}}$ for some $0 \le \alpha \le 1$, and Li & Wang (2007) proposed direction regression (DR) with the kernel matrix defined by $\Lambda_{\mathrm{DR}} \overset{\mathrm{def}}{=} E[\{2\mathbf{I}_p - \mathbf{A}(Y, \tilde{Y})\}^2]$, $\mathbf{A}(Y, \tilde{Y}) = E\{(\mathbf{x} - \mathbf{x}\tilde{})(\mathbf{x} - \mathbf{x}\tilde{})^{\mathrm{T}} \mid Y, \tilde{Y}\}$, and $(\mathbf{x}\tilde{}, \tilde{Y})$ is an independent copy of $(\mathbf{x}, Y)$. One can use either the slicing estimation or the kernel regression method to estimate the kernel matrices of SIR, SAVE, and DR. Yet how to select an optimal number of slices or how to decide an optimal bandwidth remains unknown in the literature. To avoid selecting the tuning parameters such as the number of slices in slicing estimation and the bandwidth in kernel regression, Zhu *et al.* (2010a) introduced the idea of the discretization–expectation method which improves the estimation accuracy of the slicing estimation, and Zhu *et al.* (2010d) proposed a family of cumulative slicing estimation in parallel to the development of SIR, SAVE, and DR. Li *et al.* (2005) proposed contour regression to identify the central subspace $\mathcal{S}_{Y|\mathbf{x}}$. These methods extend the inverse regression idea and are promising in recovering the central subspace.

The inverse regression methods are also used in a similar way to identify the central mean subspace $\mathcal{S}_{E(Y|\mathbf{x})}$. Li & Duan (1991) proved that the column space of the ordinary least squares (OLS) is a subspace of $\mathcal{S}_{E(Y|\mathbf{x})}$ if the covariates $\mathbf{x}$ satisfy the linearity condition (5), which can be viewed as the first attempt in this aspect. Li (1992) and Cook & Li (2002) later

proposed the principal Hessian directions (PHD) and the iterative Hessian transformations methods, which are able to recover $\mathcal{S}_{E(Y|\mathbf{x})}$ if both (5) and (6) are satisfied.

Overall, it is fair to say that inverse regression methods are elegant, mysterious at first glance and relatively simple to implement, but are restricted by the moment conditions such as (5), (6) or their variations. It is also worth mentioning that through assuming the conditional normality of $\mathbf{x}$ on Y, Cook & Forzani (2008) proposed a likelihood based method. Because a fully parametric model is assumed on $\mathbf{x}$ given Y, the joint distribution of ($\mathbf{x}$, Y) can now be fully specified as long as the marginal distribution of Y is obtained. Thus, the inherent difficulty of the dimension reduction problem is assumed away.

The inverse regression based methods can be readily adapted to accommodate regressions with multivariate response. In general the identification of the joint central subspace has been developed along three main streams. The first is to generalize the slicing methodology of Li (1991) through dividing the multivariate response data into several hypercubes. See, for instance, Aragon (1997), Hsing (1999), and Setodji & Cook (2004). The second is to recover the joint central subspace from marginal central subspaces. Some important work in this class includes Cook & Setodji (2003), Saracco (2005), and Yin & Bura (2006). The third approach utilizes one-dimensional projections of response variables. For example, Li *et al.* (2003) chose to project the multivariate response onto a few "optimal" projections, while Li *et al.* (2008) and Zhu *et al.* (2010e) considered to replace the multivariate responses with their random projections to preserve the integrity of the entire joint central subspace.

2.2 Non-Parametric Methods

The idea of non-parametric estimation for dimension reduction is relatively new. Conceptually, this class of estimators estimates the column space of β through minimizing a criterion that describes the fit of the dimension reduction models (1) and (2) to the observed data. Because the criterion inevitably involves unknown distributions or regression functions, non-parametric estimation is always involved. To facilitate our subsequent illustration in this subsection, we further require β to be an orthogonal matrix, that is, $\beta^{\mathrm{T}}\beta = \mathbf{I}_d$. For notational simplicity, even with the additional orthonormal requirement, we still retain the same notation β.

The most original non-parametric method is the minimum average variance estimation method (MAVE, Xia *et al.* 2002) for recovering the central mean subspace $\mathcal{S}_{E(Y|\mathbf{x})}$. The MAVE method is based on a very simple localizing idea used frequently in semiparametric/non-parametric estimation. Specifically, one can recast the dimension reduction model (2) into an equivalent but more familiar multiple-index model

$$Y = m(\beta^{\mathrm{T}}\mathbf{x}) + \varepsilon,$$

where m is an unspecified regression function, and the unobserved error term ε satisfies $E(\varepsilon \mid \mathbf{x}) = 0$. From a Taylor expansion, $m(\beta^{\mathrm{T}}\mathbf{x}_j) \approx m(\beta^{\mathrm{T}}\mathbf{x}_0) + m'(\beta^{\mathrm{T}}\mathbf{x}_0)(\beta^{\mathrm{T}}\mathbf{x}_j - \beta^{\mathrm{T}}\mathbf{x}_0)$ for $\mathbf{x}_j$ close to $\mathbf{x}_0$, where $m' \stackrel{\text{def}}{=} \partial m(\beta^{\mathrm{T}}\mathbf{x})/\partial(\mathbf{x}^{\mathrm{T}}\beta)$ denotes the first derivative of m. With the m function being an unknown smooth function, we localize the usual ordinary least square criterion to obtain $\sum_{i=1}^{n}\{Y_i - a_0 - \mathbf{b}_0^{\mathrm{T}}(\beta^{\mathrm{T}}\mathbf{x}_i - \beta^{\mathrm{T}}\mathbf{x}_0)\}^2 w_i$, where $w_i = K_h(\beta^{\mathrm{T}}\mathbf{x}_i - \beta^{\mathrm{T}}\mathbf{x}_0)$ and $K_h(\cdot) = K(\cdot/h)/h^d$ is a d-dimensional kernel function scaled by the bandwidth h. Minimizing this criterion subject to the constraint $\beta^{\mathrm{T}}\beta = \mathbf{I}_d$ then yields an estimate of a_0, $\mathbf{b}_0$, β. Considering that a common β is shared by different $\mathbf{x}_0$ values, a natural compromise is to set $\mathbf{x}_0$ to be the observed predictor values $\mathbf{x}_1, \ldots, \mathbf{x}_n$, and obtain the corresponding a_j, $\mathbf{b}_j$ and β jointly. This leads to minimizing

$$\sum_{j=1}^{n}\sum_{i=1}^{n}\{Y_i - a_j - \mathbf{b}_j^{\mathrm{T}}(\beta^{\mathrm{T}}\mathbf{x}_i - \beta^{\mathrm{T}}\mathbf{x}_j)\}^2 w_{ij}$$

with respect to a_j's, $\mathbf{b}_j$'s, and β, where $w_{ij} = K_h(\beta^{\mathrm{T}}\mathbf{x}_i - \beta^{\mathrm{T}}\mathbf{x}_j)$. The minimizer $\hat{\beta}$ obtained from the above minimization can be used as an estimated basis matrix of $\mathcal{S}_{E(Y|\mathbf{x})}$. Differing from the aforementioned inverse regression based methods, the MAVE method guarantees to recover $\mathcal{S}_{E(Y|\mathbf{x})}$ if d is correctly specified. This property is named exhaustiveness, first introduced by Li *et al.* (2005). We point out a technique detail that in performing the non-parametric estimation to recover the unknown link function, MAVE has used local linear smoothing, while in the literature, there has been related work in single-index and multiple-index model framework, where kernel method is used (Ichimura, 1993). The local linear choice reduces boundary bias which leads to computational advantages.

To estimate the central subspace $\mathcal{S}_{Y|\mathbf{x}}$, a useful observation is that $E\{K_b(Y - y) \mid \mathbf{x}\}$ converges to the density of Y conditional on $\mathbf{x}$ as the bandwidth b shrinks to zero. Since the central subspace model (1) implies that the conditional density of Y given $\mathbf{x}$ is identical to the conditional density of Y given $\beta^{\mathrm{T}}\mathbf{x}$, Xia (2007) adapted the idea of MAVE by regarding $K_b(Y - y)$ as a response variable and proposed the density based MAVE (dMAVE) procedure targeting at recovering $\mathcal{S}_{Y|\mathbf{x}}$. Specifically, for any fixed y, the MAVE procedure can be adapted to minimizing

$$\sum_{j=1}^{n}\sum_{i=1}^{n}\{K_b(Y_j - y) - a_j - \mathbf{b}_j^{\mathrm{T}}(\beta^{\mathrm{T}}\mathbf{x}_i - \beta^{\mathrm{T}}\mathbf{x}_j)\}^2 w_{ij}$$

with respect to a_j, $\mathbf{b}_j$, and β, where β is an orthonormal matrix and $w_{ij} = K_h(\beta^{\mathrm{T}}\mathbf{x}_i - \beta^{\mathrm{T}}\mathbf{x}_j)$. The above minimization is carried out for an arbitrary y, and selecting y to be the observed response values yields minimizing

$$\sum_{k=1}^{n}\sum_{j=1}^{n}\sum_{i=1}^{n}\{K_b(Y_i - Y_k) - a_{jk} - \mathbf{b}_{jk}^{\mathrm{T}}(\beta^{\mathrm{T}}\mathbf{x}_i - \beta^{\mathrm{T}}\mathbf{x}_j)\}^2 w_{ij},$$

the most basic form of dMAVE. More elaborative forms of dMAVE are described in Xia (2007) and some variations such as sliced regression (SR) are proposed in Wang & Xia (2008). Both dMAVE and SR methods inherit the merit of MAVE in the sense that they can recover the entire $\mathcal{S}_{Y|\mathbf{x}}$ exhaustively if d is correctly specified. Built on similar ideas of MAVE, Hernández & Velilla (2005) proposed to minimize a criterion function which involves kernel density estimation, yet their approach can only be applied to classification problems. Yin & Cook (2005) and Yin *et al.* (2008) proposed a method to recover the dimension reduction space via minimizing another criterion based on Kullback–Leibler distance.

None of the non-parametric methods requires the linearity condition (5) or the constant variance condition (6). Because of the implicit usage of smoothing, however, a common assumption of these non-parametric methods is that all the covariates are continuous. Consequently, as soon as one of the covariates is categorical, the description described above can break down. This issue is likely fixable through more careful handling of the smoothing procedure, although it will further increase the computational complexity of the methods. Overall, comparing with the inverse regression methods, non-parametric methods

are conceptually more intuitive while computationally more complex. A review of both the inverse regression based methods and the non-parametric methods can be found in Yin (2010).

2.3 Semiparametric Methods

The semiparametric methods (Ma & Zhu, 2012a) use a geometric tool in Bickel *et al.* (1993) and Tsiatis (2006) to derive the complete family of influence functions, and subsequently obtain the complete class of estimators. This treatment enables to eliminate the moment conditions (5) and (6) required for the inverse regression methods and the continuity condition required for the non-parametric methods, hence is more flexible. The treatment also further reveals the underlying rationale behind various inverse regression methods and their interrelation, which in turn helps understanding these methods. The concept of influence function was originally introduced by Hampel (1968) in his PhD thesis as a tool to study robustness of an estimator, and has been used by Prendergast (2005, 2007) in the dimension reduction framework to analyse the statistical properties of various inverse regression methods. However, the semiparametric approach of Ma & Zhu (2012a) uses influence function in a completely different way. Instead of deriving the influence function of an existing method, Ma & Zhu (2012a) constructs a class of influence functions directly, then forms the estimators and studies their properties based on the influence functions.

The basic idea of semiparametric methods is very simple. It involves writing down the likelihood of one observation, recognizing that the essential problem of dimension reduction is equivalent to that of parameter estimation in the presence of nuisance components, and taking advantage of the influence function structure to avoid estimating all the nuisance components.

For example, to estimate the central subspace $\mathscr{S}_{Y|\mathbf{x}}$ in model (1), the likelihood of one observation is $\eta_1(\mathbf{x})\eta_2(Y, \beta^{\mathrm{T}}\mathbf{x})$, where η_1, η_2 are the marginal and conditional density functions and are viewed as nuisance, and the interest is solely on β. Thus, the geometric approach readily yields the unnormalized influence function class as

$$\{\mathbf{f}(Y,\mathbf{x}) - E(\mathbf{f}|\beta^{\mathrm{T}}\mathbf{x}, Y) : E(\mathbf{f}|\mathbf{x}) = E(\mathbf{f}|\beta^{\mathrm{T}}\mathbf{x}) \forall \mathbf{f}\}.$$

Some more explicit members in this class are for example

$$\sum_{i=1}^{k}\{\mathbf{g}_i(Y,\beta^{\mathrm{T}}\mathbf{x}) - E(\mathbf{g}_i|\beta^{\mathrm{T}}\mathbf{x})\}\{\alpha_i(\mathbf{x}) - E(\alpha_i|\beta^{\mathrm{T}}\mathbf{x})\}$$

for any $\mathbf{g}_i$, $\boldsymbol{\alpha}_i$, or simply

$$\{\mathbf{g}(Y,\beta^{\mathrm{T}}\mathbf{x}) - E(\mathbf{g}|\beta^{\mathrm{T}}\mathbf{x})\}\{\alpha(\mathbf{x}) - E(\alpha|\beta^{\mathrm{T}}\mathbf{x})\} \quad (7)$$

for any $\mathbf{g}$, $\boldsymbol{\alpha}$. The double centring through subtracting $E(\mathbf{g} \mid \beta^{\mathrm{T}}\mathbf{x})$ from $\mathbf{g}$ and subtracting $E(\boldsymbol{\alpha} \mid \beta^{\mathrm{T}}\mathbf{x})$ from $\boldsymbol{\alpha}$ is beneficial because it provides a double robustness property. This means that the mis-specification of either $E(\mathbf{g} \mid \beta^{\mathrm{T}}\mathbf{x})$ or $E(\boldsymbol{\alpha} \mid \beta^{\mathrm{T}}\mathbf{x})$ can be tolerated and the resulting estimation is still consistent. In particular, mis-specifying $E(\mathbf{g} \mid \beta^{\mathrm{T}}\mathbf{x})$ while providing additional assumptions (such as the linearity condition (5) and the constant variance condition (6) when $\boldsymbol{\alpha}$ is linear or quadratic) to enable the calculation of $E(\boldsymbol{\alpha} \mid \beta^{\mathrm{T}}\mathbf{x})$ is the foundation under the inverse regression class.

The simplest illustration can be obtained from the SIR example. In (7), we select $\mathbf{g}(Y, \beta^{\mathrm{T}}\mathbf{x}) = E(\mathbf{x} \mid Y)$ and $\alpha(\mathbf{x}) = \mathbf{x}^{\mathrm{T}}$. The linearity condition yields $E(\alpha \mid \beta^{\mathrm{T}}\mathbf{x}) = \mathbf{x}^{\mathrm{T}}\mathbf{P}$. Subsequently, the double robustness allows us to mis-specify $E(\mathbf{g} \mid \beta^{\mathrm{T}}\mathbf{x}) = \mathbf{0}$. Simple algebra shows that (7) then becomes $\Lambda_{\mathrm{SIR}}\mathbf{Q} = \mathbf{0}$, which is equivalent to SIR following some linear algebra manipulation. Different choices of $\mathbf{g}$ and α lead to different inverse regression methods, see Ma & Zhu (2012a) for details of more such exercises. The importance of this derivation is that the rationale behind the inverse regression methods is now clear. The true reason that this class of methods work is the double robustness property, which allows the linearity and/or constant variance condition alone to guarantee the consistency of the whole estimating equation. Understanding the underlying driving force is crucial for further relaxing these assumptions. For example, assuming a polynomial instead of a linear form of $E(\mathbf{x} \mid \beta^{\mathrm{T}}\mathbf{x})$ leads to the relaxed methods of Li & Dong (2009) and Dong & Li (2010). Furthermore, without assuming any structure on $E(\alpha \mid \beta^{\mathrm{T}}\mathbf{x})$, but instead, estimating this quantity through the data, will then completely eliminate the linearity and constant variance conditions and provide a generalized inverse regression estimator class.

The estimation of the central mean subspace $\mathcal{S}_{E(Y|\mathbf{x})}$ parallels that of the central subspace $\mathcal{S}_{Y|\mathbf{x}}$. The likelihood of one observation is $\eta_1(\mathbf{x})\eta_2\{Y - m(\beta^{\mathrm{T}}\mathbf{x}), \mathbf{x}\}$, where η_1, η_2 have the same meaning as before, but with the additional constraint that $\int \varepsilon\eta_2(\varepsilon, \mathbf{x})d\mu(\varepsilon) = 0$. The corresponding unnormalized influence function family is

$$[\{Y - E(Y|\beta^{\mathrm{T}}\mathbf{x})\}\{\alpha(\mathbf{x}) - E(\alpha|\beta^{\mathrm{T}}\mathbf{x})\}:\forall\alpha].$$

This is similar to the central subspace analysis, except that the freedom of choosing $\mathbf{g}$ is lost to the sole choice of $\mathbf{g}(Y) = Y$. Comparing the model specification, (1) assumes more structure than (2), hence it is no surprise that the estimator class for $\mathcal{S}_{Y|\mathbf{x}}$ is richer than that for $\mathcal{S}_{E(Y|\mathbf{x})}$. Similarly, various choices of α leads to various inverse regression estimators, and estimating $E(\alpha \mid \beta^{\mathrm{T}}\mathbf{x})$ instead of assuming it leads to the relaxation of the linearity and constant variance conditions.

From the general theory of semiparametrics, the non-parametric methods, provided that they offer root-n consistent estimation, also have their corresponding influence functions, and should correspond to members in the semiparametric class. Establishing a clear link requires theoretical analysis of the non-parametric estimators and the derivation of the corresponding influence functions, which are not yet available in the dimension reduction literature.

Overall, semiparametric approach offers a birds' eye view to dimension reduction. It allows a better understanding of the various methods and their connection, and enables the relaxation of some routine assumptions. The approach further provides tools for deriving new estimation and inference procedures, as is elaborated in Section 3.

3 Inference Issue

Because dimension reduction is generally viewed as a problem of estimating a space, inference is strikingly left out of the main stream research in this field. Except for some sporadic analysis growing out of particular needs of specific problems (Chen & Li, 1998; Cook, 2004; Li *et al.*, 2010), no systematic studies have been devoted to inference. This has resulted in estimation of $\mathcal{S}_{Y|\mathbf{x}}$ or $\mathcal{S}_{E(Y|\mathbf{x})}$ without knowing the variability, and certainly has made it impossible to perform a fair comparison between different estimation procedures.

The underlying difficulty of inference or lack of it is due to the concept of space estimation, instead of the more classical parameter estimation in the dimension reduction problems.

None of the existing statistical methods apply if we are overwhelmed by the idea of space estimation. On the other hand, the spaces $\mathcal{S}_{Y|\mathbf{x}}$ and $\mathcal{S}_{E(Y|\mathbf{x})}$ can be characterized by parameters contained in β. Although different parametrization of β may lead to the same space, for the space estimation purpose, these differences are uninteresting. In other words, as long as we make up our mind to stick with only one particular parametrization strategy, then the issue of space estimation is totally equivalent to that of parameter estimation. This will subsequently make inference more or less standard. We now summarize various different parametrizations and investigate the inference issue subsequently.

3.1 Various Parametrizations

The most widely used parametrization for both $\mathcal{S}_{Y|\mathbf{x}}$ and $\mathcal{S}_{E(Y|\mathbf{x})}$ is $\beta \in \mathbb{R}^{p\times d}$, where each column of β has length 1 and the d columns of β are mutually orthogonal to each other. This certainly does not uniquely match one space with one β matrix, since both β and $\beta\mathbf{A}$ satisfy the constraints and yield the same space. Here $\mathbf{A}$ is an arbitrary $d \times d$ orthogonal matrix. For this reason, when this parametrization is used, not only additional constraints $\beta^{\mathrm{T}}\beta = \mathbf{I}_d$ needs to be respected, but one also has to take into account that different βs may actually correspond to the same space. Therefore, this parametrization can be viewed as a non-thorough conversion from space to parameter. However, in practice, because most numerical procedures look for roots or minimizers in a local neighbourhood, while $\beta\mathbf{A}$s, with different orthogonal matrices $\mathbf{A}$, form a discrete set of matrices, hence the one-to-many mapping issue hardly ever poses a real problem. The constraint of $\beta^{\mathrm{T}}\beta = \mathbf{I}_d$ sometimes offer simplification in computations, although it leads to difficulty in analysing the estimation variability of $\hat{\beta}$, because the domain of β is no longer an open set.

A second parametrization is motivated by the aim of representing the spaces $\mathcal{S}_{Y|\mathbf{x}}$ and $\mathcal{S}_{E(Y|\mathbf{x})}$ without additional constraints on the parameters (Li & Dong, 2009). For $d = 1$, it parametrizes through requiring the k-th element of β to be $\sin(\theta_1) \ldots \sin(\theta_{k-1}) \cos(\theta_k)$ for $k < p$ and the p-th element $\sin(\theta_1) \ldots \sin(\theta_{p-2}) \sin(\theta_{p-1})$. Here $\theta_1, \ldots, \theta_{p-1}$ are values in $[0, 2\pi)$ and are otherwise without constraints. This parametrization is found to bring great computational challenge hence is not further explored for $d > 1$ cases in the literature.

A third parametrization avoids additional constraints in a different way from the second one. Recognizing that $\beta\mathbf{A}$ and β yield the same space as long as $\mathbf{A}$ has full rank, Ma & Zhu (2012b) directly requires the upper $d \times d$ submatrix of β to be the identity matrix $\mathbf{I}_d$. This leaves the lower $(p - d) \times d$ submatrix of β completely free and provides a one-to-one mapping to the spaces. Similar parametrization was already proposed in Ma & Genton (2010) in a different context, and caution needs to be taken since the parametrization does require the first d components of the p covariates to be non-dismissable. Otherwise, the d rows in $\mathbf{I}_d$ will not be always cumulated in the upper part of β. Because with a suitable ordering of the covariates, this parametrization offers a one-to-one mapping between the parameters and spaces without any additional constraints, it is convenient for further investigating inference issues. Since $(p - d)d$ free parameters are needed in this parametrization, it also becomes explicit that $\mathcal{S}_{Y|\mathbf{x}}$ and $\mathcal{S}_{E(Y|\mathbf{x})}$ are effectively objects with $(p - d)d$ degrees of freedom. Both the second and third parametrization can be viewed as a direct parametrization of any point of the Grassmann manifold (Cook *et al.,* 2010), which is defined as the collection of all d-dimensional subspaces of the p-dimensional real space.

3.2 Inference and Efficient Estimation

Having obtained the unnormalized influence function for estimating $\mathcal{S}_{Y|\mathbf{x}}$ and $\mathcal{S}_{E(Y|\mathbf{x})}$ in Section 2.3, adopting the third parametrization in Section 3.1, inference is more or less straightforward. Generally speaking, the raw form of the estimating equation system often contains more than $(p - d)d$ equations. We first use a generalized method of moments

(GMM) procedure to reduce the number of equations to $(p-d)d$. Denoting the resulting estimating equation $\sum_{i=1}^{n}\mathbf{f}(Y_i,\mathbf{x}_i;\hat{\beta})=\mathbf{0}$, a standard Taylor expansion then yields

$$\mathbf{0}=n^{-1/2}\sum_{i=1}^{n}\mathbf{f}(Y_i,\mathbf{x}_i;\beta)+E\left\{\frac{\partial\mathbf{f}(Y_i,\mathbf{x}_i;\beta)}{\partial\mathrm{vecl}(\beta)^{\mathrm{T}}}\right\}\sqrt{n}\{\mathrm{vecl}(\hat{\beta})-\mathrm{vecl}(\beta)\}+o_p(1),$$

and inference is therefore straightforward. Here, vecl(β) or vecl($\hat{\beta}$) is the concatenation of the lower $(p-d)\times d$ block of β or $\hat{\beta}$ (remember the upper $d\times d$ block of β or $\hat{\beta}$ is the identity matrix $\mathbf{I}_d$).

With the availability of influence function family and inference tools, a curious issue is whether or not one can obtain efficient estimators. This is where the central subspace estimation problem splits from the central mean subspace estimation problem, because the latter turns out to be almost impossible due to the high dimensionality of $\mathbf{x}$.

Let us start with the easier problem of estimating $\mathcal{S}_{Y|\mathbf{x}}$ efficiently. Ma & Zhu (2012b) derived the efficient score for estimating β as

$$\mathbf{S}_{\mathrm{eff}}(Y,\mathbf{x},\beta^{\mathrm{T}}\mathbf{x},\eta_2)=\mathrm{vecl}[\{\mathbf{x}-E(\mathbf{x}|\beta^{\mathrm{T}}\mathbf{x})\}\partial\log\{\eta_2(Y,\beta^{\mathrm{T}}\mathbf{x})\}/\partial(\mathbf{x}^{\mathrm{T}}\beta)].$$

Hypothetically, the efficient estimator can be obtained through solving

$$\sum_{i=1}^{n}\mathbf{S}_{\mathrm{eff}}(Y_i,\mathbf{x}_i,\beta^{\mathrm{T}}\mathbf{x}_i,\eta_2)=\mathbf{0}.$$

However, $\mathbf{S}_{\mathrm{eff}}$ is not readily implementable because it contains the unknown quantities $E(\mathbf{x}\mid\beta^{\mathrm{T}}\mathbf{x})$ and $\partial\{\log\eta_2(Y,\beta^{\mathrm{T}}\mathbf{x})\}/\partial(\mathbf{x}^{\mathrm{T}}\beta)$.

To circumvent this problem, we can estimate η_2 and its first derivative based on $(Y_i,\beta^{\mathrm{T}}\mathbf{x}_i)$, for $i=1,\ldots,n$ by modifying the idea of the "double-kernel" local linear smoothing method studied in Fan *et al.* (1996). Let the resulting estimators be $\hat{\eta}_2(\cdot)$ and $\hat{\eta}_2'(\cdot)$. The estimation of $E(\mathbf{x}\mid\beta^{\mathrm{T}}\mathbf{x})$ is more straightforward and we denote the estimator by $\hat{E}(\mathbf{x}\mid\beta^{\mathrm{T}}\mathbf{x})$. Inserting these estimated quantities into $\mathbf{S}_{\mathrm{eff}}$, the resulting estimator will indeed yield efficient estimation for $\mathcal{S}_{Y|\mathbf{x}}$ (Ma & Zhu, 2012b).

The analysis appears to be similar for estimating $\mathcal{S}_{E(Y|\mathbf{x})}$. The efficient score function in this case is

$$\mathbf{S}_{\mathrm{eff}}(\varepsilon,\mathbf{x},\beta^{\mathrm{T}}\mathbf{x},m)=\mathrm{vecl}\left\{\frac{\varepsilon}{E(\varepsilon^2|\mathbf{x})}\left(\mathbf{x}-\frac{E[\mathbf{x}\{E(\varepsilon^2|\mathbf{x})\}^{-1}|\beta^{\mathrm{T}}\mathbf{x}]}{E[\{E(\varepsilon^2|\mathbf{x})\}^{-1}|\beta^{\mathrm{T}}\mathbf{x}]}\right)\frac{\partial m(\beta^{\mathrm{T}}\mathbf{x})}{\partial(\mathbf{x}^{\mathrm{T}}\beta)}\right\}.$$

To obtain the efficient estimator, it seems one just needs to insert various estimated quantities for $\partial m(\beta^{\mathrm{T}}\mathbf{x})/\partial(\mathbf{x}^{\mathrm{T}}\beta)$, $E(\varepsilon^2\mid\mathbf{x})$, $E[\{E(\varepsilon^2\mid\mathbf{x})\}^{-1}\mid\beta^{\mathrm{T}}\mathbf{x}]$, and $E[\mathbf{x}\{E(\varepsilon^2\mid\mathbf{x})\}^{-1}\mid\beta^{\mathrm{T}}\mathbf{x}]$. However, estimating $E(\varepsilon^2\mid\mathbf{x})$ involves a high dimensional problem and is the very same problem that motivated the original central mean subspace model. Hence, without additional assumptions, the efficient estimation of $\mathcal{S}_{E(Y|\mathbf{x})}$ is practically very hard to achieve because of the difficulty in estimating $E(\varepsilon^2\mid\mathbf{x})$ when $\mathbf{x}$ is high-dimensional.

A useful compromise here is to seek local efficiency, in which case one can, for example, set $E(\varepsilon^2 \mid \mathbf{x})$ to a constant or other plausible variance function. Because other than the quantity $E(\varepsilon^2 \mid \mathbf{x})$, all the non-parametric estimation involved in the efficient estimation is of lower dimension and can be handled rather easily, one can calculate the locally efficient estimation via setting a model for $E(\varepsilon^2 \mid \mathbf{x})$ and estimating all other quantities non-parametrically. This will in general still lead to a consistent estimation of $\mathcal{S}_{E(Y|\mathbf{x})}$. In addition, when the model for $E(\varepsilon^2 \mid \mathbf{x})$ happens to be true, then an efficient estimation will be obtained.

3.3 The Role of Linearity/Constant Variance Condition

Among the possible drawbacks of assuming the linearity condition (5) and the constant variance condition (6), the loss of estimation efficiency is certainly the last that one would expect. On the contrary, the typical conjecture would be just the opposite. One would usually suspect that giving up these conditions will incur a price of variability inflation. However, numerical experiments repeatedly support the counter-intuitive phenomenon of improved efficiency through giving up using the conditions (5) and (6) even when they hold, and the ability of performing inference finally allows a formal investigation of this issue.

In this review article we leave out the derivation and proof of this surprising phenomenon. One can refer to Ma & Zhu (2012c) for technical details. Even with the proof, to understand this phenomenon intuitively is not simple at all. One way to help understanding is to refer to Henmi & Eguchi (2004), where using a geometric tool, they explained a similar discovery in treating missing data via inverse probability weighting and provided some general insight. The problem here is unfortunately much more complex than the one they considered and their result or explanation does not apply and cannot be adapted. However, Henmi & Eguchi (2004) at least made one thing clear. In general estimation procedures, using a true quantity to replace an estimated quantity does not necessarily increase or decrease the variability of the estimation of the parameter of interest. The variance of an estimator could go either way, and in our case, using a known form increased the variability. Loosely speaking, when using inverse regression methods, assuming condition (5) and/or (6), $E(\alpha \mid \beta^{\mathrm{T}}\mathbf{x})$ is simply set to its true value when α is a linear or quadratic function of $\mathbf{x}$. Compared with estimating $E(\alpha \mid \beta^{\mathrm{T}}\mathbf{x})$ non-parametrically, this practice eliminates a variance component. This variance component turns out to be in a direction that is beneficial to the variability of estimating β, which is why assuming the conditions would bring an eventual efficiency loss.

This discovery appears somewhat detrimental to the commonly used linearity and constant variance conditions. Because if these conditions are falsely assumed, the inverse regression methods are not consistent, and if they are correctly assumed, they cause inflated variation. In addition, linearity and constant variance conditions are completely irrelevant for the efficient estimation purpose, as is clear from Section 3.2. In fact, these conditions are useful only when α in (7) is taken to be linear or quadratic. In this aspect, if we had chosen to use some other α, we would simply assume a form for $E(\alpha \mid \beta^{\mathrm{T}}\mathbf{x})$ and carry out the subsequent derivation. However, it is also important to recognize the practical usefulness of these conditions. Computationally, these conditions have enabled great implementation simplification of the most familiar inverse regression estimators.

4 Determining the Dimension

Up till now, we have assumed the structural dimension d to be known. In practice, deciding d is not a simple task. In this section we review the literature on how to decide d.

When the linearity condition (5) and/or the constant variance condition (6) are satisfied, the inverse regression methods listed in Section 2.1 formulate the problem of estimating $\mathcal{S}_{Y|\mathbf{x}}$ and $\mathcal{S}_{E(Y|\mathbf{x})}$ into an eigen-decomposition problem. Therefore, the determination of d

becomes a problem of determining the number of non-zero eigenvalues of the corresponding matrix, termed kernel matrix in the literature. Let Λ be the kernel matrix of a specific inverse regression based dimension reduction method. For example, $\Lambda_{SIR} = \text{cov}\{E(\mathbf{x} \mid Y)\}$ in SIR (Li, 1991), $\Lambda_{SAVE} = E\{\mathbf{I}_p - \text{cov}(\mathbf{x} \mid Y)\}^2$ in SAVE (Cook & Weisberg, 1991) and $\Lambda_{PHD} = E\{Y - E(Y)\}\{\mathbf{x} - E(\mathbf{x})\}\{\mathbf{x} - E(\mathbf{x})\}^{\mathrm{T}}$ in PHD (Li, 1992). Since the structural dimension d is the number of non-zero eigenvalues of Λ in these methods, and the data provide an estimator $\hat{\Lambda}$, thus any sensible way of deciding the number of non-zero eigenvalues through using $\hat{\Lambda}$ will yield a reasonable way of deciding d. To this end, there are mainly four approaches in the literature: the sequential test method, the bootstrap method, the BIC type criterion and the sparse eigen-decomposition method.

To implement the sequential test, one tests a series of null hypotheses that Λ has l, from $p - 1$ down to 1, non-zero eigenvalues. The test procedure stops when the first time a null hypothesis is not rejected. In the context of deciding d for the central subspace $\mathscr{S}_{Y|\mathbf{x}}$, Li (1991) proposed a sequential chi-square test for the SIR method by assuming that $\mathbf{x}$ is normally distributed. Schott (1994), Velilla (1998), Bura & Cook (2001), Cook & Yin (2001), and Cook & Ni (2005) proposed more general versions to relax the normality assumption. To determine the dimension d of the central mean subspace $\mathscr{S}_{E(Y|\mathbf{x})}$, Li (1992), Cook & Li (2004) designed several sequential tests for principal Hessian directions (Li, 1992) and iterative Hessian transformation (Cook & Li, 2002). A nice summary of the sequential test in various inverse regression methods is given by Bura & Yang (2011). In general, the dimension $\hat{d}$ obtained from these sequential tests is not consistent due to the presence of the type-I error. When the dimension p is large, such test procedures are computationally inefficient and the cumulative type-I error may not be ignorable. Moreover, it relies heavily on the asymptotic normality of the estimator of the kernel matrix, and requires plugging in an estimate of the asymptotic variance. Due to these theoretical and implementation issues, the sequential test method is usually studied case-by-case for each dimension reduction method in the literature.

Ye & Weiss (2003) proposed to decide d through a bootstrap procedure. Their intuition is that, at the true reduced dimension d, the distance between the estimated central (mean) subspace obtained from the bootstrap sample and that from the original sample must have the smallest variability under suitable distance measure. Therefore, choosing d that corresponds to the smallest estimated variability, assuming no or minimal bias, is a reasonable procedure. Zhu & Zeng (2006) modified the bootstrap procedure through using a different distance measure. The bootstrap procedure estimates the dimension in a data-driven manner, but it is computationally intensive (Zeng, 2008). In addition, the consistency of the bootstrap procedure is still unestablished in the literature.

For the SIR method, Zhu *et al.* (2006) imitated the classical BIC procedure and proposed a BIC type criterion. The estimated dimension obtained from the BIC type criterion is consistent if the estimator of the relevant kernel matrix converges. However, deciding the amount of penalty through a data-driven procedure is usually difficult in practice. Zhu & Zhu (2007) suggested a classical rate $\log n$ penalty, while Luo *et al.* (2009) used a cutoff value for the ratio of eigenvalues to decide d, both are somewhat ad hoc. Empirical studies suggest that the performance of the BIC type procedure can vary a great deal when different penalties are used, unless the sample size is very large.

Zhu *et al.* (2010c) proposed a sparse eigen-decomposition (SED) strategy. The main idea of the SED method is to cast the spectral decomposition problem resulting from an inverse regression method into a least squares problem. They then imposed an adaptive LASSO penalty (Tibshirani, 1996; Zou, 2006) to shrink some small estimated eigenvalues to zero directly. Different from the aforementioned methods, the SED method can estimate $\mathscr{S}_{Y|\mathbf{x}}$ or

$\mathcal{S}_{E(Y|\mathbf{x})}$ and the structural dimension d simultaneously. The minimization can be solved effectively by the LARS algorithm (Efron *et al.,* 2004). Therefore, it is computationally attractive. In addition, the resulting estimators of the eigenvalues are consistent, hence $\hat{d}$ is consistent.

The link to matrix eigen-decomposition is completely lost as soon as the estimation methods deviate from the inverse regression type. In the non-parametric methods we reviewed in Section 2.2, d is often determined through a leave-one-out cross-validation procedure. Specifically, at any fixed working dimension k, the average leave-one-out prediction error is calculated to yield CV(k), and subsequently, the k value that yields the smallest CV(k) is selected as $\hat{d}$. For estimating $\mathcal{S}_{E(Y|\mathbf{x})}$, one can easily formulate

$$\mathrm{CV}(k)=n^{-1}\sum_{i=1}^{n}\{Y_i-\hat{\mathrm{m}}_{-i}(\hat{\beta}_{-i}^{\mathrm{T}}\mathbf{x}_i)\}^2$$

calculated at a fixed working dimension, where the subscript "$-i$" stands for the usual estimate of the corresponding component with the i-th observation left out (Xia *et al.*, 2002). For estimating $\mathcal{S}_{Y|\mathbf{x}}$, a similar formulation

$$\mathrm{CV}(k)=n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{n}\{K_b(Y_i-Y_j)-\hat{\eta}_{2,-i}(Y_j,\hat{\beta}_{-i}^{\mathrm{T}}\mathbf{x}_i)\}^2$$

can be used.

Cross-validation in the non-parametric methods requires the estimation of m in the central mean subspace case and η_2 in the central subspace case, which are the quantities often avoided to estimate in the semiparametric methods we reviewed in Section 2.3. Thus, Ma & Zhu (2012a) proposed a bootstrap procedure to determine d in the semiparametric class. This procedure adapts the original idea of Ye & Weiss (2003) and Dong & Li (2010) in the inverse regression class.

For two arbitrary random vectors **u** and **v**, define

$$r^2(\mathbf{u},\mathbf{v})=\frac{1}{k}\sum_{i=1}^{k}\lambda_i,$$

where $\lambda_1, \ldots, \lambda_k$ are the non-zero eigenvalues of

$$\{\mathrm{var}(\mathbf{u})\}^{-1/2}\mathrm{cov}(\mathbf{u},\mathbf{v}^{\mathrm{T}})\{\mathrm{var}(\mathbf{v})\}^{-1}\mathrm{cov}(\mathbf{v},\mathbf{u}^{\mathrm{T}})\{\mathrm{var}(\mathbf{u})\}^{-1/2}.$$

For any working dimension k, let $\hat{\boldsymbol{\beta}}_k$ be the semiparametric estimate based on the original sample, and $\hat{\boldsymbol{\beta}}_{k,b}$ be the same estimate based on the b-th bootstrap sample, $b = 1, \ldots, B$. Then the bootstrap method determines d through choosing the k value that maximizes the average r^2 value

$$\frac{1}{B}\sum_{b=1}^{B}r^2(\hat{\beta}_k^{\mathrm{T}}\mathbf{x},\hat{\beta}_{k,b}^{\mathrm{T}}\mathbf{x}).$$

The rational behind the bootstrap method is the following. First, we note that r^2 is a kind of correlation measure. For any working dimension k equal to the true dimension d, $\hat{\beta}_k$, and $\hat{\beta}_{k,b}$ both estimate the true dimension reduction space hence are largely identical, thus the average r^2 is expected to be close to 1. For k larger than d, $\hat{\beta}_k$, and $\hat{\beta}_{k,b}$ contain too much freedom hence would have to contain some additional random directions that are contributed purely by the randomness of the sample. Thus, their corresponding average r^2 value would be diluted by the random directions hence would decrease. On the other hand, when k is smaller than d, not enough freedom is included in β, hence $\hat{\beta}_k$ and $\hat{\beta}_{k,b}$ will both only be the best approximation to the true β. This best approximation is purely decided by the data randomness and will be different when the random sample is different, hence the corresponding average r^2 value will also decrease.

5 The Issue of High Dimension

In the classic statistical literature, the dimension p of the covariates $\mathbf{x}$ is typically regarded as fixed. To study statistical properties in situations where p is fairly large in comparison with the sample size n, it is more convenient to consider the case where p diverges with n at some rate. In the dimension reduction context, Zhu *et al.* (2006) established the consistency for SIR when $p = o(n^{1/4})$ under the Frobenius norm. That is, they showed that $|\hat{\Lambda}_{\mathrm{SIR}} - \Lambda_{\mathrm{SIR}}|_2 = o_p(1)$ as long as $p = o(n^{1/4})$ and the linearity condition (5) is true. Zhu *et al.* (2010d) obtained the consistency of the cumulative slicing estimation when $p = o(n^{1/2})$ under the Frobenius norm. We suspect that under the Frobenius norm, all the consistent methods mentioned in Section 2 for a fixed p actually remain consistent as long as $p = o(n^{1/2})$. We also believe that $p = o(n^{1/2})$ is the largest possible dimension to retain the consistency, if no additional assumption is made on the correlation structure of $\mathbf{x}$.

Recent trend in statistics is to study the situation when the covariate dimensionality p is even larger than the sample size n. This poses new challenges to dimension reduction. For example, even the original normalization of the covariates $\mathbf{x}$ is no longer possible because the sample version of $\Sigma = \mathrm{var}(\mathbf{x})$ would be singular. Similarly, in many inverse regression based methods, the kernel matrices involve the computation of quantities such as Σ^{-1}. However, a consistent estimator of Σ^{-1} is typically not available when p is larger than n. To implement the inverse regression methods in Section 2.1, a useful strategy is to combine the idea of partial least squares with the inverse regression based methods. This avoids the operation of matrix inversion which is not possible if p is larger than n. See, for example, Naik & Tsai (2000), Li *et al.* (2007), Cook *et al.* (2007), Zhu & Zhu (2009b), and Zhu *et al.* (2010b). This strategy is also useful when the covariates $\mathbf{x}$ are highly correlated such that Σ is a singular matrix.

Another strategy to handle ultrahigh dimensionality is to utilize the sparsity principle in the dimension reduction context, that is, we assume that only a moderate portion of the covariates are truly relevant to the response variable and need to be reduced further. This corresponds to assuming that most rows in β are simply zeros. In the context of linear regression, Tibshirani (1996), Fan & Li (2001), Candes & Tao (2007) proposed, respectively LASSO, SCAD, and Dantzig selector to penalize some small sample coefficients to zero to yield sparse solutions. One advantage of these methods is that these regularization methods can estimate the magnitude of these non-zero coefficients and screen out those zero coefficients simultaneously. If we only intend to select important covariates in the linear regression context, we can borrow the idea of sure independence screening (SIS) procedure (Fan & Lv, 2008). The SIS procedure ranks the importance of each covariate using the magnitude of the Pearson correlation coefficients. In the screening stage, they only selected the covariates associated with large squared correlation coefficients. They demonstrated that the SIS procedure has the sure screening property, that is, all important predictors will

survive after the screening procedure if the sample size n is sufficiently large. With the normality assumption, their results can be applied to the central mean subspace model (2). Following the idea of Fan & Lv (2008), Zhu *et al.* (2012) proposed sure independence ranking and screening (SIRS) procedure to select important predictors from the central subspace model (1). They proved that the SIRS method has the ranking consistency property, that is, all important predictors will be ranked above the unimportant ones in an asymptotic sense. The ranking consistency property immediately implies the sure screening property provided that an ideal cut-off is available.

The aforementioned procedures all require that the covariates satisfy the linearity condition (5). However, how to handle ultrahigh dimensionality is still unknown when the linearity condition (5) is violated. Research along this line is in much demand.

Acknowledgments

The authors thank the editor's interest in dimension reduction which initiated this review. Ma's work was supported by the National Science Foundation (DMS-1206693 and DMS-1000354) and the National Institute of Neurological Disorders and Stroke (R01-NS073671). Zhu's work was supported by National Natural Science Foundation of China (11071077), Innovation Program of Shanghai Municipal Education Commission (13ZZ055) and Pujiang Project of Science and Technology Commission of Shanghai Municipality (12PJ1403200).

References

Adragni K, Cook RD. Sufficient dimension reduction and prediction in regressions. Philos. Trans. R. Soc., A. 2009; 367:4385–4405.

Aragon Y. A Gauss implementation of multivariate sliced inverse regression. Comput. Statist. 1997; 12:355–372.

Bickel, PJ.; Klaassen, CAJ.; Ritov, Y.; Wellner, JA. Efficient and Adaptive Estimation for Semiparametric Models. Baltimore: The Johns Hopkins University Press; 1993.

Box GEP, Cox DR. An analysis of transformations (with discussion). J. R. Stat. Soc. Ser. B. 1964; 26:211–252.

Bura E, Cook RD. Extending sliced inverse regression: the weighted Chi-squared test. J. Amer. Statist Assoc. 2001; 96:996–1003.

Bura E, Yang J. Dimension estimation in sufficient dimension reduction: a unifying approach. J. Multivariate Anal. 2011; 102:130–142.

Candes E, Tao T. The Dantzig selector: statistical estimation when p is much larger than n (with discussion). Ann. Statist. 2007; 35:2313–2404.

Chen CH, Li KC. Can SIR be as popular as multiple linear regression? Statist. Sinica. 1998; 8:289–316.

Cook RD. On the interpretation of regression plots. J. Amer. Statist. Assoc. 1994; 89:177–189.

Cook, RD. Regression Graphics: Ideas for Studying Regressions through Graphics. New York: Wiley; 1998.

Cook RD. Testing predictor contributions in sufficient dimension reduction. Ann. Statist. 2004; 32:1062–1092.

Cook RD, Forzani L. Principal fitted components for dimension reduction in regression. Statist. Sci. 2008; 23:485–501.

Cook RD, Lee H. Dimension reduction in binary response regression. J. Amer. Statist. Assoc. 1999; 94:1187–1200.

Cook RD, Li B. Dimension reduction for conditional mean in regression. Ann. Statist. 2002; 30:455–474.

Cook RD, Li B. Determining the dimension of iterative Hessian transformation. Ann. Statist. 2004; 32:2501–2531.

Cook RD, Li B, Chiaromonte F. Dimension reduction in regression without matrix inversion. Biometrika. 2007; 94:569–584.

Cook RD, Li B, Chiaromonte F. Envelope models for parsimonious and efficient multivariate liner regression. Statist. Sinica. 2010; 20:927–1010.

Cook RD, Nachtsheim CJ. Reweighting to achieve elliptically contoured covariates in regression. J. Amer. Statist. Assoc. 1994; 89:592–599.

Cook RD, Ni L. Sufficient dimension reduction via inverse regression: a minimum discrepancy approach. J. Amer. Statist. Assoc. 2005; 100:410–428.

Cook RD, Weisberg S. Discussion of “Sliced inverse regression for dimension reduction”. J. Amer. Statist. Assoc. 1991; 86:28–33.

Cook RD, Setodji CM. A model-free test for reduced rank in multivariate regression. J. Amer. Statist. Assoc. 2003; 98:340–351.

Cook RD, Yin X. Dimension reduction and visualization in discriminant analysis (with discussion). Aust. N. Z. J. Stat. 2001; 43:147–199.

Dong Y, Li B. Dimension reduction for non-elliptically distributed predictors: second-order moments. Biometrika. 2010; 97:279–294.

Eaton ML. A characterization of spherical distributions. J. Multivariate Anal. 1986; 34:439–446.

Efron B, Hastie T, Johnstone I, Tibshirani R. Least Angle Regression. Ann. Statist. 2004; 32:407–499.

Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. J. Amer. Statist. Assoc. 2001; 96:1348–1360.

Fan J, Lv J. Sure independence screening for ultrahigh dimensional feature space (with discussion). J. R. Stat. Soc. Ser. B. 2008; 70:849–911.

Fan J, Yao Q, Tong H. Estimation of conditional densities and sensitivity measures in nonlinear dynamical systems. Biometrika. 1996; 83:189–196.

Fung WK, He X, Liu L, Shi P. Dimension reduction based on canonical correlation. Statist. Sinica. 2002; 12:1093–1113.

Hall P, Li KC. On almost linearity of low dimensional projection from high dimensional data. Ann. Statist. 1993; 21:867–889.

Hampel, FR. Thesis. Berkeley: University of California; 1968. Contributions to the Theory of Robust Estimation.

Henmi M, Eguchi S. A paradox concerning nuisance parameters and projected estimating functions. Biometrika. 2004; 91:929–941.

Hernández A, Velilla S. Dimension reduction in nonparametric kernel discriminant analysis. J. Comput. Graph. Statist. 2005; 14:847–866.

Hsing T. Nearest neighbor inverse regression. Ann. Statist. 1999; 27:697–731.

Ichimura H. Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. J. Econometrics. 1993; 58:71–120.

Li B, Dong Y. Dimension reduction for non-elliptically distributed predictors. Ann. Statist. 2009; 37:1272–1298.

Li B, Wang S. On directional regression for dimension reduction. J. Amer. Statist. Assoc. 2007; 102:997–1008.

Li B, Wen SQ, Zhu LX. On a projective resampling method for dimension reduction with multivariate responses. J. Amer. Statist. Assoc. 2008; 103:1177–1186.

Li B, Zha H, Chiaromonte F. Contour regression: a general approach to dimension reduction. Ann. Statist. 2005; 33:1580–1616.

Li G, Zhu LP, Zhu LX. Confidence region for the direction in semi-parametric regressions. J. Multivariate Anal. 2010; 101:1364–1377.

Li LX, Cook RD, Tsai CL. Partial inverse regression. Biometrika. 2007; 94:615–625.

Li KC. Sliced inverse regression for dimension reduction. J. Amer. Statist. Assoc. 1991; 86:316–327.

Li KC. On principal Hessian directions for data visualization and dimension reduction: another application of Stein’s lemma. J. Amer. Statist. Assoc. 1992; 87:1025–1039.

Li KC, Aragon Y, Shedden K, Agnan CT. Dimension reduction for multivariate response data. J. Amer. Statist. Assoc. 2003; 98:99–109.

Li KC, Duan N. Regression analysis under link violation. Ann. Statist. 1991; 17:1009–1052.

Luo R, Wang H, Tsai CL. Contour projected dimension reduction. Ann. Statist. 2009; 37:3743–3778.

Ma Y, Genton MG. Explicit semiparametric estimators for generalized linear latent variable models. J. R. Stat. Soc. Ser. B. 2010; 72:475–495.

Ma Y, Zhu L. A semiparametric approach to dimension reduction. J. Amer. Statist. Assoc. 2012a; 107(497):168–179.

Ma Y, Zhu L. Efficient estimation in sufficient dimension reduction. Ann. Statist. 2012b http://www.stat.tamu.edu/ma/papers/mz5.pdf.

Ma Y, Zhu L. Efficiency loss caused by linearity condition in dimension reduction. Biometrika. 2012c http://www.stat.tamu.edu/ma/papers/mz4.pdf.

Naik P, Tsai CL. Partial least squares estimator for single-index models. J. R. Stat. Soc. Ser. B. 2000; 62:763–771.

Prendergast LA. Influence functions for sliced inverse regression. Scand. J. Statist. 2005; 32:385–404.

Prendergast LA. Implications of influence function analysis for sliced inverse regression and sliced average variance estimation. Biometrika. 2007; 94:585–601.

Saracco J. Asymptotics for pooled marginal slicing estimator based on SIR approach. J. Multivariate Anal. 2005; 96:117–135.

Schott JR. Determining the dimensionality in sliced inverse regression. J. Amer. Statist. Assoc. 1994; 89:141–148.

Setodji CM, Cook RD. *K*-means inverse regression. Technometrics. 2004; 46:421–429.

Tibshirani R. Regression shrinkage and selection via the Lasso. J. R. Stat. Soc. Ser. B. 1996; 58:267–288.

Tsiatis, AA. Semiparametric Theory and Missing Data. New York: Springer; 2006.

Velilla S. Assessing the number of linear components in a general regression problem. J. Amer. Statist. Assoc. 1998; 93:1088–1098.

Wang H, Xia Y. Sliced regression for dimension reduction. J. Amer. Statist. Assoc. 2008; 103:811–821.

Xia Y. A constructive approach to the estimation of dimension reduction directions. Ann. Statist. 2007; 35:2654–2690.

Xia Y, Tong H, Li WK, Zhu LX. An adaptive estimation of dimension reduction space (with discussion). J. R. Stat. Soc. Ser. B. 2002; 64:363–410.

Ye Z, Weiss RE. Using the bootstrap to select one of a new class of dimension reduction methods. J. Amer. Statist. Assoc. 2003; 98:968–979.

Yin, X. Sufficient dimension reduction in regression. In: Cai, TT.; Shen, X., editors. The Analysis of High-dimensional Data. Vol. Chapter 9. New Jersey: World Scientific; 2010.

Yin X, Bura E. Moment based dimension reduction for multivariate response regression. J. Statist. Plann. Inference. 2006; 136:3675–3688.

Yin X, Cook RD. Dimension reduction for the conditional *k*th moment in regression. J. R. Stat. Soc. Ser. B. 2002; 64:159–175.

Yin X, Cook RD. Direction estimation in single-index regressions. Biometrika. 2005; 92:371–384.

Yin X, Li B, Cook RD. Successive direction extraction for estimating the central subspace in a multiple-index regression. J. Multivar. Anal. 2008; 99:1733–1757.

Zeng P. Determining the dimension of the central subspace and central mean subspace. Biometrika. 2008; 95:469–479.

Zeng P, Zhu Y. An integral transform method for estimating the central mean and central subspaces. J. Multivariate Anal. 2010; 101:271–290.

Zhu LP, Li L, Li R, Zhu LX. Model-free feature screening for ultrahigh dimensional data. J. Amer. Statist. Assoc. 2011; 106(496):1464–1475.

Zhu LP, Wang T, Zhu LX, Férre L. Sufficient dimension reduction through discretization-expectation estimation. Biometrika. 2010a; 97:295–304.

Zhu LP, Yin X, Zhu LX. Dimension reduction for correlated data: an alternating inverse regression. J. Comput. Graph. Statist. 2010b; 19:887–899.

Zhu LP, Yu Z, Zhu LX. A sparse eigen-decomposition estimation in semi-parametric regression. Comput. Statist. Data Anal. 2010c; 54:976–986.

Zhu LP, Zhu LX. On kernel method for sliced average variance estimation. J. Multivariate Anal. 2007; 98:970–991.

Zhu LP, Zhu LX. Dimension-reduction for conditional variance in regressions. Statist. Sinica. 2009a; 19:869–883.

Zhu LP, Zhu LX. On distribution-weighted partial least squares with diverging number of highly correlated predictors. J. R. Stat. Soc. Ser. B. 2009b; 71:525–548.

Zhu LP, Zhu LX, Feng Z. Dimension reduction in regressions through cumulative slicing estimation. J. Amer. Statist. Assoc. 2010d; 105:1455–1466.

Zhu LP, Zhu LX, Wen SQ. On dimension reduction in regressions with multivariate responses. Statist. Sinica. 2010e; 20:1291–1307.

Zhu LX, Fang KT. Asymptotics for the kernel estimates of sliced inverse regression. Ann. Statist. 1996; 24:1053–1067.

Zhu LX, Miao B, Peng H. On sliced inverse regression with high-dimensional covariates. J. Amer. Statist. Assoc. 2006; 101:630–643.

Zhu LX, Ohtaki M, Li YX. On hybrid methods of inverse regression based algorithms. Comput. Statist. Data Anal. 2007; 51:2621–2635.

Zhu Y, Zeng P. Fourier methods for estimating the central subspace and the central mean subspace in regression. J. Amer. Statist. Assoc. 2006; 101:1638–1651.

Zou H. The adaptive Lasso and its oracle properties. J. Amer. Statist. Assoc. 2006; 101:1418–1429.