Statistica Sinica **28** (2018), 1415-1436
doi:https://doi.org/10.5705/ss.202016.0056

# ON THE SECOND-ORDER INVERSE REGRESSION METHODS FOR A GENERAL TYPE OF ELLIPTICAL PREDICTORS

Wei Luo

*Baruch College*

*Abstract:* In sufficient dimension reduction, the second-order inverse regression methods, such as the principal Hessian directions and directional regression, commonly require the predictor to be normally distributed. In this paper, we introduce a type of elliptical distributions called the quadratic variance ellipticity family, which covers and approximates a variety of commonly seen elliptical distributions, with the normal distribution as a special case. When the predictor belongs to this family, we study the properties of the second-order inverse regression methods and adjust them accordingly to preserve consistency. When the dimension of the predictor is sufficiently large, we show the consistency of the conventional methods, which strengthens a previous result in Li and Wang (2007). Simulation studies and data analysis are conducted to illustrate the effectiveness of the adjusted methods.

*Key words and phrases:* Central mean subspace, central subspace, directional regression, principal Hessian directions, quadratic variance ellipticity family.

## 1. Introduction

Sufficient dimension reduction has attracted much attention in the recent decades, for its capability of condensing large dimensional data while preserving the relative information. For a $p$-dimensional predictor $X$ and a response variable $Y$, it aims to find a lower-dimensional $\beta^{\mathsf{T}}X$ that satisfies

$$Y \perp\!\!\!\perp X \,|\, \beta^{\mathsf{T}}X, \tag{1.1}$$

where "$\perp\!\!\!\perp$" means independence and $\beta \in \mathbb{R}^{p\times d}$ for some $d < p$. $\beta^{\mathsf{T}}X$ is then a sufficient statistic, and using it in place of $X$ will not cause information loss in subsequent analysis. For identifiable parametrization and dimension reduction effectiveness, Cook (1998) introduced the central subspace as the linear space spanned by columns of $\beta$ in (1.1), with smallest possible dimension $d$. This space exists under fairly general conditions (Yin, Li and Cook (2008)), and is commonly denoted by $\mathcal{S}_{Y|X}$.

When statistical interest is on the conditional mean $E(Y|X)$, sufficient dimension reduction is adjusted to find $\beta^{\mathsf{T}}X$ that satisfies

$$E(Y|X) \perp\!\!\!\perp X \,|\, \beta^{\mathsf{T}}X, \tag{1.2}$$

so that only those linear combinations of $X$ that affect $E(Y|X)$ will be estimated. In this case, Cook and Li (2002) defined the central mean subspace as the identifiable parameter, and denoted it by $\mathcal{S}_{E(Y|X)}$. For ease of presentation, we do not distinguish it from the central subspace, and refer to both as the central dimension reduction subspace, unless otherwise pointed out. We assume its dimension $d$ to be known a priori. Because $\beta$ in (1.1) and (1.2) can be easily adjusted for linear transformations of $X$, we assume a standardized $X$ that has zero mean and identity covariance matrix $I_p$ throughout.

To estimate the central dimension reduction subspace, a main stream of methods in the literature can be characterized as first introducing a symmetric matrix parameter $M$, called the kernel matrix, and then estimating the column space of $M$ by the leading eigenvectors of its sample analog $\hat{M}$, where "leading" means that the corresponding eigenvalues have the greatest absolute value. When the column space of $M$ is contained in the central dimension reduction subspace, these methods estimate a subspace of the latter and are called Fisher consistent; when the two spaces further coincide, these methods are called exhaustive. Because the moments of $X$ given $Y$ are used in constructing all the kernel matrices, these methods are commonly called the inverse regression methods. Depending on the order of the moments, they can be further categorized as first-order methods, including ordinary least regression (OLS; Li and Duan (1989)) and sliced inverse regression (SIR; Li (1991)), and second-order methods, including principal Hessian directions (pHd; Li (1992)), sliced average variance estimation (SAVE; Cook and Weisberg (1991)), and directional regression (DR; Li and Wang (2007)). The kernel matrices of these methods are listed in Table 1. Because $\hat{M}$ can be easily constructed by slicing $Y$ and using sample moments, its form is omitted.

To achieve Fisher consistency, first-order methods require the linearity condition

$$E(X|\beta^{\mathsf{T}}X) = \beta(\beta^{\mathsf{T}}\beta)^{-1}\beta^{\mathsf{T}}X \text{ for any } \beta \in \mathbb{R}^{p\times d}, \tag{1.3}$$

which, by Li (1991), is equivalent to an elliptical distribution on $X$ and thus is quite mild. The exhaustiveness of these methods requires additional constraints on the model structure and may be infeasible in practice.

Second-order methods, in particular SAVE and directional regression, achieve

Table 1. Inverse regression methods and their kernel matrices. In column 2, var$(\cdot)$ means the covariance matrix of the distribution, and $(\tilde{X}, \tilde{Y})$ is an independent copy of $(X, Y)$.

| methods | kernel matrix | target space |
|---|---|---|
| OLS | $M_{OLS} = E(XY)E^{\mathsf{T}}(XY)$ | $\mathcal{S}_{E(Y\|X)}$ |
| SIR | $M_{SIR} = E\{E(X\|Y)E^{\mathsf{T}}(X\|Y)\}$ | $\mathcal{S}_{Y\|X}$ |
| pHd | $M_{pHd} = E[XX^{\mathsf{T}}\{Y - E(Y)\}]$ | $\mathcal{S}_{E(Y\|X)}$ |
| SAVE | $M_{SAVE} = E[\{I_p - \mathrm{var}(X\|Y)\}^2]$ | $\mathcal{S}_{Y\|X}$ |
| DR | $M_{DR} = E[2I_p - E\{(X - \tilde{X})(X - \tilde{X})^{\mathsf{T}}\|Y, \tilde{Y}\}]^2$ | $\mathcal{S}_{Y\|X}$ |

exhaustiveness without specifying any model structure (Li and Wang (2007)). In addition to the linearity condition (1.3), their Fisher consistency requires the constant variance condition:

$$\mathrm{var}(X|\beta^{\mathsf{T}}X) = I_p - \beta(\beta^{\mathsf{T}}\beta)^{-1}\beta^{\mathsf{T}}, \text{ for any } \beta \in \mathbb{R}^{p\times d}, \tag{1.4}$$

which, together with (1.3), is equivalent to a multivariate normal distribution on $X$ (Cook and Weisberg (1991)).

When (1.3) is violated, generalization of the inverse regression methods has been studied by such authors as Li and Dong (2009), Dong and Li (2010), and Ma and Zhu (2012). In particular, Dong and Li (2010) considered the case when (1.3) is violated but (1.4) is satisfied. The case when (1.3) is satisfied but (1.4) is violated, which amounts to an elliptical but non-normal distribution of $X$, has been overlooked in the literature. This case can occur if the data contain many outliers, or if the data are clearly bounded, or if they do not form a convex hull, in which only the first-order methods are applicable. Because the second-order methods enjoy richer theoretical properties, it will benefit researchers if these methods can be adjusted to also achieve consistency under the ellipticity of $X$ alone.

The adjustment of the second-order inverse regression methods is also worth investigation, if transformations on $X$ that only preserve its ellipticity are of interest when conducting sufficient dimension reduction. This occurs, for example, in Luo, Wang and Tsai (2009) and in Dong, Yu and Zhu (2015), both of which enhance the stableness of sample moments when the magnitude of $X$ has a heavy-tailed distribution. Luo, Wang and Tsai (2009) project $X$ onto the unit sphere in $\mathbb{R}^p$, and adjust SAVE and directional regression by changing $I_p$ in the kernel matrices to $\tau(Y)I_p$, where $\tau(Y)$ is the median eigenvalue of $\mathrm{var}(X|Y)$. This adjustment requires $p > 2d$. Whether it is consistent for other transformations remains open. Dong, Yu and Zhu (2015) transform the magnitude of $X$ to have range between zero and one, but the corresponding adjustment of the

second-order inverse regression methods has not been studied.

In this paper, we study the properties of the second-order inverse regression methods and how they can be adjusted when $X$ belongs to a subfamily of elliptical distributions called the quadratic variance ellipticity family. The adjustment adopts a simple form, and preserves the estimation consistency of the conventional methods. The subfamily of elliptical distributions, characterized by the form of $\mathrm{var}(X|\beta^{\mathsf{T}}X)$, includes, and approximates, a variety of commonly seen cases. When $X$ is in this family and has a sufficiently large dimension, we show the consistency of the conventional second-order methods.

In the rest of the article, we introduce the quadratic variance ellipticity family in Section 2, then address pHd in Section 3 and directional regression in Section 4 accordingly. In Section 5 we show consistency for (the conventional) directional regression in high-dimensional cases. A summary of implementation is given in Section 6, and simulation studies and a data example are presented in Sections 7 and 8, respectively, to illustrate the effectiveness of the adjusted methods.

## 2. The Quadratic Variance Ellipticity Family

Here is notation used throughout. With $\|\cdot\|$ the usual Euclidean norm for real vectors, we denote $p^{1/2}X/\|X\|$ by $U$ and $p^{-1/2}\|X\|$ by $R$, so that $X = UR$. Since $X$ is elliptically distributed and standardized, $p^{-1/2}U$ is uniformly distributed on the unit sphere in $\mathbb{R}^p$ and is independent of $R$, and $E(R^2) = 1$. Let $\beta$ be an arbitrary matrix in $\mathbb{R}^{p\times d}$ and $\gamma$ be an arbitrary vector in $\mathbb{R}^p$ such that $(\beta, \gamma)$ is orthonormal. We use $\mathcal{S}(\beta)$ to denote the column space of $\beta$, with orthogonal complement $\mathcal{S}^{\perp}(\beta)$, and use $\Pi(\mathcal{S}(\beta))$, or simply $\Pi(\beta)$, to denote the corresponding projection matrix. Let $\{1,\ldots,p\}^d$ be the collection of subsets of $\{1,\ldots,p\}$ with cardinality $d$. For any $i \in \{1,\ldots,p\}$ and $A \in \{1,\ldots,p\}^d$, let $X_i$ denote the $i$th component of $X$, and $X_A$ denote the components of $X$ indexed by $A$. Likewise, for any $j \in \{1,\ldots,d\}$ and $B \subset \{1,\ldots,d\}$, $\beta_i$ is the $i$th column of $\beta$, and $\beta_B$ the set of columns of $\beta$ indexed by $B$. For a kernel matrix $M \in \mathbb{R}^{p\times p}$, let $\lambda_1(M),\ldots,\lambda_p(M)$ be the eigenvalues of $M$ according to the descending order of their absolute values, and $\beta(M)$ be the array of eigenvectors following the same order. When equality exists in the eigenvalues, we allow arbitrariness in $\beta(M)$. We use $\hat{M}_n$, or simply $\hat{M}$, to denote the sample analog of $M$ based on a random sample of size $n$. The $\sqrt{n}$-consistency of $\hat{M}$ is commonly satisfied in the literature, and is assumed throughout. We define $\{\lambda_1(\hat{M}),\ldots,\lambda_p(\hat{M})\}$ and $\beta(\hat{M})$ similarly.

By the ellipticity of $X$, $\mathrm{var}(\gamma^{\mathsf{T}}X|\beta^{\mathsf{T}}X)$ is a symmetric function with respect to the origin in $\mathbb{R}^d$. Therefore, it is natural to be generalized from a constant in (1.4) to a quadratic function with the linear term being zero. That is, there exists $a_\gamma \in \mathbb{R}$ and $B_\gamma \in \mathbb{R}^{d\times d}$ such that

$$\mathrm{var}(\gamma^{\mathsf{T}}X|\beta^{\mathsf{T}}X) = a_\gamma + (X^{\mathsf{T}}\beta)B_\gamma(\beta^{\mathsf{T}}X). \tag{2.1}$$

We call this condition the quadratic variance condition and the induced distribution family the quadratic variance ellipticity family. Although adopting a parametric assumption, this family covers many commonly seen cases. In particular, it reduces to the multivariate normal distribution if $B_\gamma$ in (2.1) is zero. We give two more examples.

**Example 1.** Let $R$ be degenerate at 1, then $X$ is uniformly distributed on the sphere centered at the origin and with radius $\sqrt{p}$. This distribution is considered in Luo, Wang and Tsai (2009) for its advantage in not generating outliers. After simple calculation, we have,

$$\mathrm{var}(\gamma^{\mathsf{T}}X|\beta^{\mathsf{T}}X) = \frac{(p - \|\beta^{\mathsf{T}}X\|^2)}{(p-d)}. \tag{2.2}$$

**Example 2.** Kotz (1975) introduced the $p$-dimensional Pearson Type II distribution, with density function with respect to the Lebesque measure

$$f(x) = \{(2m+p+2)\pi\}^{-p/2}\frac{\Gamma(p/2+m+1)}{\Gamma(m+2)}\left(1 - \frac{x^{\mathsf{T}}x}{2m+p+2}\right)^m$$

on $\{x \in \mathbb{R}^p : x^{\mathsf{T}}x \le 2m+p+2\}$ and zero elsewhere, $m > -1$ being the shape parameter. As discussed in Chapter 6 of Johnson (2013), this family is closely related to the univariate Beta distribution. In particular, it reduces to the uniform distribution in the ball centered at the origin and with radius $\sqrt{p+2}$, if $m$ is zero. Johnson (2013) has

$$\mathrm{var}(\gamma^{\mathsf{T}}X|\beta^{\mathsf{T}}X) = \frac{(2m+p+2-\|\beta^{\mathsf{T}}X\|^2)}{(2m+p-d+2)}.$$

The support of a distribution in this family can coincide with the real space, or can be bounded and do or do not form a convex hull. The family excludes such distributions as the mixture normal and the multivariate t. However, simulation study suggests that it can still approximate these distributions, if transformed according to Dong, Yu and Zhu (2015). More details can be found in Section 7.

The coefficients in (2.1) typically have a simple form because of the ellipticity of $X$. As discussed in Luo, Wang and Tsai (2009), inverse regression methods become infeasible if $X$ contains heavy tails. Therefore, we assume the existence

of $E(R^4)$ throughout the article.

**Theorem 1.** *The coefficients in* (2.1) *are* $a_\gamma = 1 - d\delta$ *and* $B_\gamma = \delta I_d$*, where*

$$\delta = \frac{p\,E(R^4) - (p+2)}{p\,(d+2)\,E(R^4) - d(p+2)}; \tag{2.3}$$

$\delta$ *increases with* $E(R^4)$ *and is in the interval* $[-1/(p-d), 1/(d+2))$.

The lower bound of $\delta$ is reached at the uniform distribution on a sphere. It remains questionable whether the upper bound can be lowered using the constraint (2.1). We leave this for future investigation.

Although the components of an elliptical distribution are uncorrelated, their squared values may not be. Here the correlation between the squared components can be measured by $\delta$, or equivalently $E(R^4)$, where $E(R^4) = (p+2)/p$ from the multivariate normal corresponds to zero correlation and serves as the null value. When $E(R^4)$ exceeds this value, which occurs if $R$ has heavy tails and the data tend to have more outliers, the squared components are positively correlated. This conforms to the fact that if an observation is an outlier in one component, then it also tends to be in the other components. When $E(R^4)$ is below the null value, the squared components of the distribution are negatively correlated. Because $R^2$ has a small variance in this case, the negative correlation can also be explained by the constraint that the sum of the squared components, which equals $R^2$, is nearly constant.

Given a sample, we can estimate $\delta$ using (2.3), where $E(R^4)$ is replaced by its sample analog. Alternatively, we can first linearly regress $X_i^2$ on $(1, \|X_A\|^2)$, where $(i, A)$ runs through $\{1, \ldots, p\}^{d+1}$, and then average the resulting coefficient for $\|X_A\|^2$. An omitted simulation study has shown that the second estimator, denoted as $\hat{\delta}$, slightly outperforms the first. Thus we adopt it hereafter. Its $\sqrt{n}$-consistency is clear.

## 3. Adjusted pHd for Non-Normal Predictor

The consistency of pHd is based on the fact that when $X$ is normally distributed, the eigenvectors of the kernel matrix associated with nonzero eigenvalues are contained in the central mean subspace. When $X$ belongs to the quadratic variance ellipticity family, the central mean subspace can still be recovered by a set of eigenvectors of the same kernel matrix. However, these eigenvectors cannot be identified by nonzero eigenvalues.

**Theorem 2.** *If* $X$ *is in the quadratic variance ellipticity family, then there exists*

*$A_0 \in \{1, \ldots, p\}^d$ and $\beta(M_{pHd})$ such that $\beta_{A_0}(M_{pHd})$ is an orthonormal basis of the central mean subspace. In addition, for any $j \notin A_0$, we have*

$$\lambda_j(M_{pHd}) = \delta \sum_{i \in A_0} \lambda_i(M_{pHd}), \tag{3.1}$$

*which implies $1 \in A_0$ if $p > 2d$ and $M_{pHd}$ is nonzero.*

The uniqueness of $A_0$ as the solution to (3.1), which implies nonzero kernel matrix $M_{pHd}$, requires the eigenvalues associated with the central mean subspace to differ from the rest. In other words, $M_{pHd}$ must distinguish the central mean subspace from its orthogonal complement, in the sense that each eigenspace belongs to one of these two spaces, but does not intersect with both. When the predictor is normally distributed, this condition is equivalent to the exhaustiveness of pHd, which requires the regression function $E(Y|X)$ to be asymmetric with respect to the origin in $\mathbb{R}^p$ in any direction. For simplicity, we adopt this condition throughout unless otherwise specified.

When $X$ has an elliptical but non-normal distribution, the nonzero $\delta$ makes the kernel matrix of pHd no longer of low rank unless the signals in the central mean subspace accumulate to zero, which is rare. Nonetheless, if $d$ is considerably smaller than $p$, then at least the eigenvector(s) corresponding to the leading eigenvalue lie in the central mean subspace, which means that pHd is still capable of recovering the corresponding direction(s). This is especially useful if the central mean subspace is one-dimensional.

In general, to ensure the consistency of pHd, the data must be restricted so that either $\delta$ is negligible, which requires the predictor to be nearly normally distributed so that its components are weakly dependent, or, $p > 2d$ and the eigenvalues associated with the central mean subspace are similar to each other. The latter requires the signal strength to vary negligibly with different directions in the central mean subspace. In particular, the leading signal cannot dominate the weakest one. If neither restriction is satisfied, then pHd must be adjusted by choosing the right set of eigenvectors.

When $X$ is uniformly distributed on a sphere, Luo, Wang and Tsai (2009) assumed $p > 2d$ and adjusted SAVE and directional regression. Their adjustment can be easily parallelized for pHd by picking those eigenvalues that differ the most from the median eigenvalue. Here we adopt a natural criterion based on Theorem 2, which does not require a relationship between $p$ and $d$: over all the elements in $\{1, \ldots, p\}^d$, we choose $\hat{A}$ to minimize

$$\textstyle\sum_{j \notin A} |\lambda_j(\hat{M}_{pHd}) - \hat{\delta} \sum_{i \in A} \lambda_i(\hat{M}_{pHd})|, \tag{3.2}$$

and select $\beta_{\hat{A}}(\hat{M}_{pHd})$ as the basis which spans an estimate of the central mean subspace. When $p > 2d$, we further restrict the candidate sets to those that contain 1. The selection consistency of this criterion is shown in the following, with the $\sqrt{n}$-consistency of the adjusted pHd as a natural consequence.

**Theorem 3.** *Suppose $A_0$ is the unique solution to* (3.1). *Then as $n \to \infty$, the $\hat{A}$ that minimizes* (3.2) *satisfies $P(\hat{A} = A_0) \to 1$, and*

$$\|\Pi(\beta_{\hat{A}}(\hat{M}_{pHd})) - \Pi(\mathcal{S}_{E(Y|X)})\| = O_P(n^{-1/2}).$$

When the solution to (3.1) is not unique but at least $M_{pHd}$ is nonzero, we speculate that there always exists $k < d$ such that, by using $k$ instead of $d$ in (2.3), the solution to (3.1) exists in $\{1, \ldots, p\}^k$ and is unique, in which case a proper subset of $A_0$ is identified and the adjusted pHd recovers a proper subspace of the central mean subspace. We leave such details to future research.

## 4. Adjusted Directional Regression for Non-Normal Predictor

Because both SAVE and directional regression involve the square of the second-order inverse moment, the detail is slightly more complicated when we parallelize the discussion in the previous section for these methods. Here we focus on directional regression as arguments for SAVE can be developed similarly.

The normality of $X$ is needed in directional regression, as it guarantees the identity between the central subspace and the column space of the kernel matrix. This identity fails when the normality is relaxed to the quadratic variance ellipticity family. However, the central subspace can still be recovered by a set of eigenvectors of the kernel matrix.

**Theorem 4.** *If $X$ belongs to the quadratic variance ellipticity family, then there exists $A_0 \in \{1, \ldots, p\}^d$ and $\beta(M_{DR})$ such that $\beta_{A_0}(M_{DR})$ is an orthonormal basis of the central subspace. Further, for any $j \notin A_0$,*

$$\lambda_j(M_{DR}) = -2d^2\delta^2 + 2\delta^2 E\{\textstyle\sum_{i \in A_0} E(\|\beta_i(M_{DR})^{\mathsf{T}} X\|^2 | Y)\}^2, \qquad (4.1)$$

*which implies $1 \in A_0$ if $p > 2d$ and $M_{DR}$ is nonzero.*

Similar to (3.1), the uniqueness of $A_0$ as the solution to (4.1) requires certain condition on $M_{DR}$, generally subtler than that for pHd. Nonetheless, when the predictor is normally distributed, the condition is equivalent to the exhaustiveness of directional regression, which is fairly general; see Li and Wang (2007). We adopt it throughout.

If the predictor is non-normally distributed, then the kernel matrix of directional regression is not of reduced rank. When interest is only on the leading

signal of the data, directional regression can still be applied as long as the central subspace is of considerably lower dimension. In general, the method requires either the predictor not to severely deviate from normality, or the signal strength not to vary dramatically with the directions in the central subspace. Otherwise, it will miss the weakest signals. Accordingly, we adjust the method by picking the eigenvectors associated with the set of eigenvalues that minimizes

$$\textstyle\sum_{j\notin A} |\lambda_j(\hat{M}_{DR}) + 2d^2\hat{\delta}^2 - 2\hat{\delta}^2\hat{E}\{\sum_{i\in A}\hat{E}(\|\beta_i^{\mathsf{T}}(\hat{M}_{DR})X\|^2|Y)\}^2| \tag{4.2}$$

over $\{1,\ldots,p\}^d$, where $\hat{E}(\cdot)$ is the sample analog used in directional regression to estimate the true moment. If $p > 2d$, then we further restrict the candidate sets to those that contain 1. The selection procedure is consistent, which also implies the $\sqrt{n}$-consistency of the adjusted method.

**Theorem 5.** *Suppose $A_0$ is the unique solution to* (4.1). *Then as $n \to \infty$, $\hat{A}$ that minimizes* (4.2) *satisfies $P(\hat{A} = A_0) \to 1$, and*

$$\|\Pi(\beta_{\hat{A}}(\hat{M}_{DR})) - \Pi(\mathcal{S}_{Y|X})\| = O_P(n^{-1/2}).$$

## 5. Consistency of Directional Regression in High-Dimensional Cases

The estimation in the last two sections requires an exhaustive search over certain subsets of $\{1,\ldots,p\}$, computationally challenging when $p$ is large. We argue that, in a certain sense, for all large $p$ the adjustment is redundant for directional regression, as the method is consistent. The same arguments can be applied to pHd and SAVE.

Our result is closely related to the work of Diaconis and Freedman (1984) and Hall and Li (1993), who demonstrated the approximate normality of high-dimensional $X$ when projected onto a low-dimensional space. Following them, let $\{(X^{(p)} \in \mathbb{R}^p, Y^{(p)} \in \mathbb{R}) : p \in \mathbb{N}\}$ be a series of random vectors in which each $X^{(p)}$ is standardized and belongs to the distribution family (2.1). For each $p \in \mathbb{N}$, let $\delta^{(p)}$ be defined as in Theorem 1 for $X^{(p)}$, and let $\mathcal{S}(\beta^{(p)})$ be the central subspace for $(X^{(p)}, Y^{(p)})$ with dimension $d^{(p)}$. We assume that

$$E\left(\frac{\|X^{(p)}\|^4}{p^2}\right) \to 1 \quad \text{as } p \to \infty, \tag{5.1}$$

which is satisfied if each $X^{(p)}$ follows one of the distributions mentioned in Section 2. More discussion about this assumption can be found in Diaconis and Freedman (1984) and Hall and Li (1993).

Based on Hall and Li (1993), Li and Wang (2007) argued that directional regression is approximately consistent as $p$ increases to infinity. However, the

approximation rate was unclear. Under the quadratic variance condition (2.1), this rate is infinity, in the sense that the consistency of the method holds exactly for sufficiently large $p$.

**Theorem 6.** *Suppose that, (a) $d^{(p)} = O(1)$; (b) there exists $r > 0$ such that*

$$\frac{\min\{\beta_i^{(p)\mathsf{T}} M_{DR} \beta_i^{(p)}, i = 1, \ldots, d^{(p)}\}}{\max\{\beta_i^{(p)\mathsf{T}} M_{DR} \beta_i^{(p)}, i = 1, \ldots, d^{(p)}\}} \geq r. \tag{5.2}$$

*Then for sufficiently large $p$, the eigenvectors of $M_{DR}$ associated with the greatest $d^{(p)}$ eigenvalues span the central subspace.*

Condition (a) regulates the sparsity of the signal from the predictor, and condition (b) regulates the signal strength variation as the direction varies within the central subspace. These conditions can be relaxed if a convergence order is known for $\{\delta^{(p)}\}$. For example, if each $\delta^{(p)}$ is zero or negative, which occurs in all the examples in Section 2, then Theorem 1 indicates that $\delta^{(p)} = O((p - d^{(p)})^{-1})$. Hence the consistency result for directional regression still holds if we require $d^{(p)}$ and $r^{(p)}$, $r^{(p)}$ denoting the left-hand side of (5.2), to satisfy $d^{(p)} = o(p(r^{(p)})^{1/2})$. In particular, we now allow $d^{(p)}$, the dimension of the central subspace, to grow to infinity with $p$.

## 6. Summary of Implementation

We summarize the implementation of the adjusted pHd and the adjusted directional regression.

Step 0. If $X$ is not standardized, do so by using $\hat{\Sigma}_X^{-1/2}\{X - \hat{E}(X)\}$, in which $\hat{\Sigma}_X$ and $\hat{E}(X)$ are the sample covariance matrix and the sample mean of $X$.

Step 1. For each $i \in \{1, \ldots, p\}$ and $A \in \{1, \ldots, p\}^d$ such that $i \notin A$, linearly regress $X_i^2$ on $(1, \|X_A\|^2)$. Let $b_{i,A}$ be the corresponding coefficient for $\|X_A\|^2$. Estimate $\delta$ by $\hat{\delta}$, the average value of $\{b_{i,A} : (i, A) \in \{1, \ldots, p\}^{d+1}\}$.

Step 2. Estimate the kernel matrices $M_{pHd}$ by $\hat{M}_{pHd}$, and $M_{DR}$ by $\hat{M}_{DR}$; see Li (1992) and Li and Wang (2007) for detail.

Step 3. For pHd, select the index set $\hat{A}$ of eigenvalues that corresponds to the central mean subspace by minimizing (3.2), where $\hat{\delta}$ is derived in Step 1; for directional regression, select the $\hat{A}$ that corresponds to the central subspace by minimizing (4.2) with the same $\hat{\delta}$. The adjusted pHd estimates the central mean subspace by $\mathcal{S}(\hat{\Sigma}_X^{-1/2}\beta_{\hat{A}}(\hat{M}_{pHd}))$, and the adjusted directional regression estimates the central subspace by $\mathcal{S}(\hat{\Sigma}_X^{-1/2}\beta_{\hat{A}}(\hat{M}_{DR}))$.

## 7. Simulation Studies

We illustrate the finite-sample effectiveness of the adjusted pHd and the adjusted directional regression, using simulation models and predictors that follow various elliptical distributions. The conventional pHd and directional regression are included in the comparison as references.

We generated $X$ sequentially from a normal distribution, the uniform distribution on a sphere in $\mathbb{R}^p$, the uniform distribution in a ball in $\mathbb{R}^p$, and the Pearson Type II distribution with $m = 1$. The last distribution has a ball-shaped support, but it typically generates a ring-shaped sample. Following Dong, Yu and Zhu (2015), we generated $X$ by applying the transformation $X\|X\|/(1 + \|X\|^2)$ to both the multivariate t-distribution with 5 degrees of freedom, and the mixture normal distribution $\alpha Z + (1 - \alpha)(3Z)$, where $Z$ is normal with zero mean and $\alpha$ is an independent Bernoulli random variable with $E(\alpha) = 0.5$. Again, we further standardized $X$ so that it has zero mean and identity covariance matrix.

Among these distributions, the last two violate the quadratic variance condition (2.1). To see the severity of the violation at each dimension $d$ from 1 to $p - 1$, for each $i$ in $\{1, \ldots, p\}$ and each $A$ in $\{1, \ldots, p\}^d$ that excludes $i$, we used $X_i^2$ as the response and $(1, \|X_A\|^2)$ as the predictor, and conducted a goodness-of-fit test with linear regression as the reduced model and polynomial regression of degree 3 as the full model. The p-value was then averaged over all pairs of $(i, A)$. To further stabilize the result, we generated 500 samples and recorded the average of the averaged p-values. For reference, we repeated the process for the four other distributions of $X$. $n$ was set at 500 for a reasonable choice to control the sensitivity of the test, and $p$ was set at 6, 10, and 20, sequentially. The results are depicted in Figure 1 and summarized in Tables 2 and 3.

The large p-values for the first four distributions in Figure 1 show that the quadratic variance condition (2.1) is satisfied in these distributions. They also indicate an approximation of (2.1) at small $d$ in the last two distributions. For each fixed small $d$, such an approximation improves as $p$ grows, as supported by the generally increasing p-values in Tables 2 and 3. The improvement of the approximation suggests a potentially wide application of (2.1) in the elliptical distribution family in high-dimensional cases. As $d$ increases towards $p$, the approximation fails no matter how large $p$ is.

Based on these observations, we applied the inverse regression methods to Models I–IV below, with $X$ following each of the six distributions and the independent error term $\varepsilon$ following $N(0, 0.5^2)$.

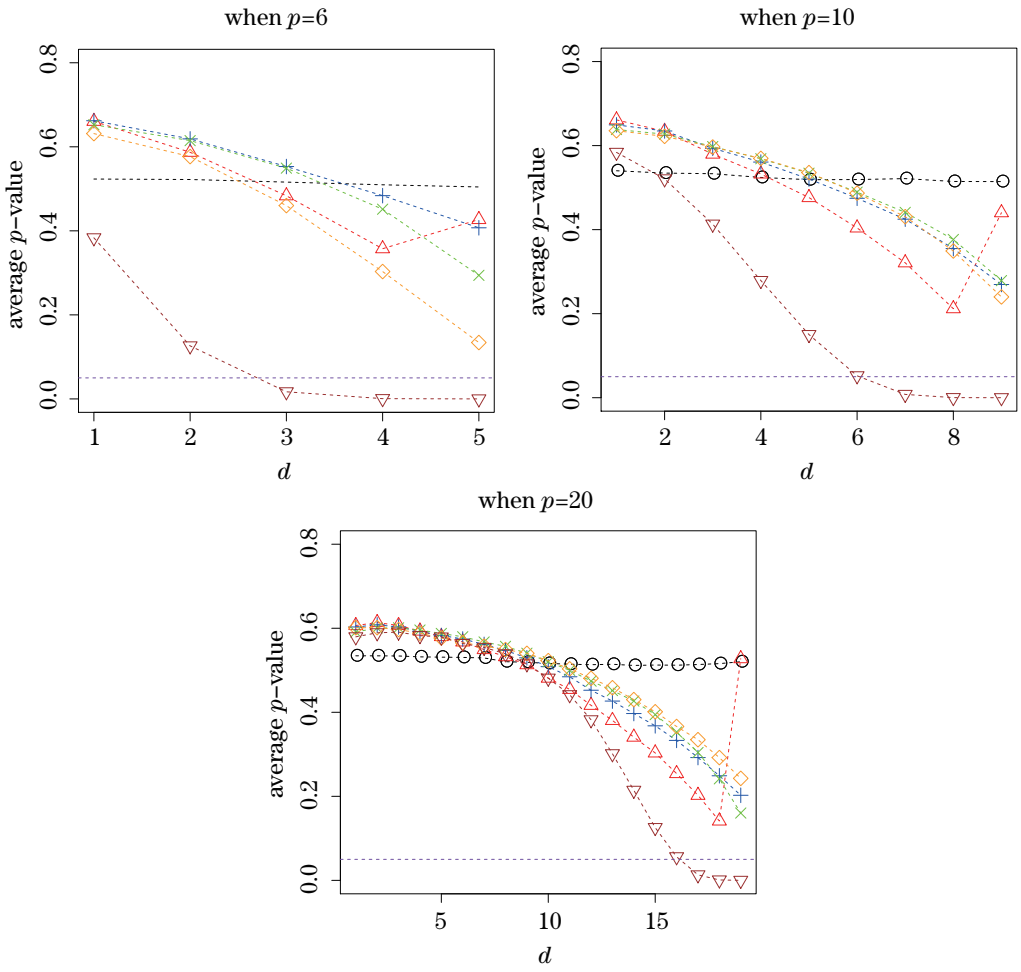


Figure 1. Approximation of (2.1) in the distributions. In the upper-left, upper-right, and bottom panels, $p = 6$, 10, and 20, respectively. In each panel, the x-axis is $d$ that runs from 1 to $p - 1$, and the y-axis is the averaged p-value for the goodness-of-fit test. "$\circ$" stands for the normal distribution, "$\triangle$" for the uniform distribution on a sphere, "$+$" for the uniform distribution in a ball, "$\times$" for Pearson Type II distribution with $m = 1$, "$\diamond$" for the transformed t-distribution, and "$\bigtriangledown$" for the transformed mixture normal distribution. The dashed line stands for the 0.05 threshold.

I. $Y = X_1^2 + X_2^2 + \varepsilon$;

II. $Y = X_1^2 + 3X_2^2 + \varepsilon$;

III. $Y = X_1 \sin(X_1) + X_2 \text{sign}(\varepsilon)$;

IV. $Y = 0.4(X_1 + X_2 + X_3)^2 + |X_1 + X_{p-1} + 3X_p|^{1/2} + 0.4\,\varepsilon$.

Table 2. The average p-value for testing the quadratic form of $\mathrm{var}(\gamma^{\mathsf{T}}X|\beta^{\mathsf{T}}X)$, when $X$ follows the transformed multivariate $t$-distribution.

| $p$ \ $d$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| 6 | 0.631 | 0.577 | 0.460 | 0.303 | 0.134 | – | – | – | – |
| 10 | 0.635 | 0.622 | 0.597 | 0.570 | 0.536 | 0.487 | 0.431 | 0.349 | 0.240 |
| 20 | 0.601 | 0.601 | 0.597 | 0.588 | 0.577 | 0.566 | 0.558 | 0.549 | 0.540 |

Table 3. The average p-value for testing the quadratic form of $\mathrm{var}(\gamma^{\mathsf{T}}X|\beta^{\mathsf{T}}X)$, when $X$ follows the transformed multivariate mixture normal distribution.

| $p$ \ $d$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| 6 | 0.383 | 0.126 | 0.017 | 0.001 | 0.000 | – | – | – | – |
| 10 | 0.584 | 0.521 | 0.413 | 0.279 | 0.150 | 0.052 | 0.008 | 0.000 | 0.000 |
| 20 | 0.580 | 0.589 | 0.591 | 0.582 | 0.579 | 0.564 | 0.554 | 0.539 | 0.515 |

A variation of Model III has been used in Li (1992), and Model IV has been used in Li and Wang (2007). In all the models, the central dimension reduction subspace is low dimensional - in Model III, the central mean subspace is 1-dimensional and the central subspace is 2-dimensional; in the other models the two spaces coincide and are 2-dimensional. Because Model I is invariant to rotations in $(X_1, X_2)$, the two eigenvalues associated with the central dimension reduction subspace are equal. Thus we expect that as long as $p > 4$, the conventional pHd without eigenvalue selection is consistent in Models I and III, and the conventional directional regression is consistent in Model I.

To compare the adjusted and conventional methods, we let $p = 6, 10$, sequentially. For the kernel matrices to be consistently estimated, we let $n = 100, 200$, respectively. For each pair of $(n, p)$, 500 samples were independently generated from each model and each distribution of $X$. Given an estimate $\mathcal{S}(\hat{\beta})$, we measured its distance from the central dimension reduction subspace, denoted as $\mathcal{S}(\beta^{(0)})$, by

$$m(\mathcal{S}(\hat{\beta}), \mathcal{S}(\beta^{(0)})) = \|\Pi(\hat{\beta}) - \Pi(\beta^{(0)})\|.$$

The performances of the adjusted and the conventional pHd are recorded in Tables 4 and 5, with respect to different values of $(n, p)$; likewise, the performances of the adjusted and the conventional directional regression are recorded in Tables 6 and 7.

Because the same models are equipped with different distributions of $X$, a

Table 4. Comparison between the adjusted and conventional pHd when $n = 100$ and $p = 6$. In each cell, the number on the top (bottom) is the sample mean (standard deviation) of the distance between the estimates and the central mean subspace. Column "Normal" stands for cases with $X$ normal, "Unif-S" for the uniform distribution on a sphere, "Unif-B" for the uniform distribution in a ball, "Pearson" for the Pearson Type II distribution with $m = 1$, "DYZ-$t_5$" for the Dong, Yu and Zhu's transformation on the multivariate t-distribution, and "DYZ-MN" for the same transformation on the mixture normal distribution.

| | Adjusted pHd | | | | Conventional pHd | | | |
|---|---|---|---|---|---|---|---|---|
| | I | II | III | IV | I | II | III | IV |
| Normal | 0.536 | 0.773 | 0.603 | 0.828 | 0.555 | 0.773 | 0.603 | 0.852 |
| | (0.160) | (0.310) | (0.302) | (0.319) | (0.151) | (0.310) | (0.302) | (0.333) |
| Unif-S | 0.334 | 0.482 | 0.457 | 0.486 | 0.322 | 10.41 | 0.457 | 10.41 |
| | (0.099) | (0.209) | (0.264) | (0.273) | (0.088) | (0.049) | (0.264) | (0.110) |
| Unif-B | 0.362 | 0.563 | 0.484 | 0.563 | 0.348 | 10.35 | 0.484 | 10.32 |
| | (0.094) | (0.290) | (0.261) | (0.328) | (0.087) | (0.257) | (0.261) | (0.282) |
| Pearson | 0.368 | 0.579 | 0.476 | 0.588 | 0.368 | 10.27 | 0.476 | 10.19 |
| | (0.094) | (0.274) | (0.254) | (0.326) | (0.094) | (0.348) | (0.254) | (0.318) |
| DYZ-$t_5$ | 0.391 | 0.603 | 0.512 | 0.634 | 0.381 | 10.08 | 0.512 | 10.14 |
| | (0.117) | (0.284) | (0.278) | (0.330) | (0.099) | (0.453) | (0.278) | (0.439) |
| DYZ-MN | 0.438 | 0.645 | 0.556 | 0.744 | 0.444 | 0.684 | 0.556 | 0.699 |
| | (0.118) | (0.246) | (0.254) | (0.300) | (0.114) | (0.277) | (0.254) | (0.276) |

Table 5. Comparison between the adjusted and conventional pHd when $n = 200$ and $p = 10$. The abbreviations and other specifications are the same as described in the legend of Table 4.

| | Adjusted pHd | | | | Conventional pHd | | | |
|---|---|---|---|---|---|---|---|---|
| | I | II | III | IV | I | II | III | IV |
| Normal | 0.571 | 0.813 | 0.521 | 0.857 | 0.571 | 0.813 | 0.521 | 0.873 |
| | (0.103) | (0.253) | (0.180) | (0.265) | (0.103) | (0.253) | (0.180) | (0.276) |
| Unif-S | 0.375 | 0.560 | 0.403 | 0.540 | 0.375 | 10.29 | 0.403 | 10.34 |
| | (0.065) | (0.166) | (0.130) | (0.166) | (0.065) | (0.331) | (0.130) | (0.276) |
| Unif-B | 0.420 | 0.600 | 0.434 | 0.596 | 0.420 | 10.13 | 0.434 | 10.21 |
| | (0.070) | (0.188) | (0.140) | (0.198) | (0.070) | (0.426) | (0.140) | (0.387) |
| Pearson | 0.419 | 0.589 | 0.433 | 0.583 | 0.419 | 10.01 | 0.433 | 10.11 |
| | (0.077) | (0.169) | (0.122) | (0.172) | (0.077) | (0.445) | (0.122) | (0.428) |
| DYZ-$t_5$ | 0.422 | 0.612 | 0.451 | 0.619 | 0.422 | 10.02 | 0.451 | 10.07 |
| | (0.075) | (0.188) | (0.170) | (0.209) | (0.075) | (0.427) | (0.170) | (0.444) |
| DYZ-MN | 0.493 | 0.751 | 0.512 | 0.794 | 0.493 | 0.751 | 0.512 | 0.805 |
| | (0.083) | (0.215) | (0.159) | (0.248) | (0.083) | (0.215) | (0.159) | (0.252) |

comparison between the columns within each table and each inverse regression method cannot provide a guideline on which transformation of $X$ should be

Table 6. Comparison between the adjusted and conventional directional regression when $n = 100$ and $p = 6$. In each cell, the number on the top (bottom) is the sample mean (standard deviation) of the distance between the estimates and the central subspace. "DR" stands for directional regression. Other abbreviations and specifications are the same as described in the legend of Table 4.

| | Adjusted DR | | | | Conventional DR | | | |
|---|---|---|---|---|---|---|---|---|
| | I | II | III | IV | I | II | III | IV |
| Normal | 0.509 | 0.622 | 0.843 | 0.544 | 0.509 | 0.622 | 0.843 | 0.544 |
| | (0.148) | (0.265) | (0.292) | (0.215) | (0.148) | (0.265) | (0.292) | (0.215) |
| Unif-S | 0.639 | 0.836 | 0.976 | 0.841 | 0.856 | 1.23 | 0.996 | 10.05 |
| | (0.284) | (0.333) | (0.325) | (0.315) | (0.409) | (0.271) | (0.328) | (0.366) |
| Unif-B | 0.537 | 0.841 | 0.906 | 0.721 | 0.605 | 0.986 | 0.905 | 0.799 |
| | (0.208) | (0.344) | (0.331) | (0.316) | (0.313) | (0.375) | (0.317) | (0.353) |
| Pearson | 0.499 | 0.788 | 0.874 | 0.656 | 0.518 | 0.913 | 0.869 | 0.737 |
| | (0.192) | (0.315) | (0.319) | (0.279) | (0.230) | (0.368) | (0.328) | (0.363) |
| DYZ-$t_5$ | 0.501 | 0.768 | 0.827 | 0.617 | 0.505 | 0.815 | 0.827 | 0.647 |
| | (0.184) | (0.337) | (0.302) | (0.273) | (0.189) | (0.362) | (0.302) | (0.294) |
| DYZ-MN | 0.495 | 0.564 | 0.844 | 0.531 | 0.495 | 0.564 | 0.844 | 0.531 |
| | (0.130) | (0.215) | (0.284) | (0.207) | (0.130) | (0.215) | (0.284) | (0.207) |

Table 7. Comparison between the adjusted and conventional directional regression when $n = 200$ and $p = 10$. The abbreviations and other specifications are the same as described in the legends of Table 4 and Table 6.

| | Adjusted DR | | | | Conventional DR | | | |
|---|---|---|---|---|---|---|---|---|
| | I | II | III | IV | I | II | III | IV |
| Normal | 0.496 | 0.588 | 0.807 | 0.509 | 0.496 | 0.588 | 0.807 | 0.509 |
| | (0.099) | (0.166) | (0.243) | (0.138) | (0.099) | (0.166) | (0.243) | (0.138) |
| Unif-S | 0.474 | 0.755 | 0.758 | 0.602 | 0.474 | 0.881 | 0.828 | 0.639 |
| | (0.091) | (0.269) | (0.233) | (0.193) | (0.091) | (0.357) | (0.234) | (0.266) |
| Unif-B | 0.464 | 0.687 | 0.755 | 0.547 | 0.464 | 0.766 | 0.755 | 0.569 |
| | (0.116) | (0.265) | (0.230) | (0.177) | (0.116) | (0.317) | (0.230) | (0.232) |
| Pearson | 0.455 | 0.675 | 0.744 | 0.533 | 0.455 | 0.673 | 0.804 | 0.531 |
| | (0.086) | (0.279) | (0.210) | (0.186) | (0.086) | (0.285) | (0.226) | (0.188) |
| DYZ-$t_5$ | 0.461 | 0.662 | 0.833 | 0.541 | 0.461 | 0.696 | 0.833 | 0.552 |
| | (0.085) | (0.245) | (0.248) | (0.171) | (0.085) | (0.299) | (0.248) | (0.175) |
| DYZ-MN | 0.498 | 0.561 | 0.823 | 0.506 | 0.498 | 0.561 | 0.823 | 0.506 |
| | (0.093) | (0.142) | (0.216) | (0.127) | (0.093) | (0.142) | (0.216) | (0.127) |

made when applying sufficient dimension reduction. Thus such a comparison can only be interpreted qualitatively; that is, using the normal distribution as the reference, if an inverse regression method does not perform dramatically worse when applied to a non-normal distribution of $X$, then it is considered effective for that distribution. In this sense, these tables suggest that the adjusted pHd

and the adjusted directional regression improve the conventional methods, in the sense that their effectiveness only requires the quadratic variance condition (2.1) on $X$.

Theoretically, the adjusted and conventional methods should coincide when pHd is applied to Models I and III, or when directional regression is applied to Model I, or when both methods are applied to a normal $X$. This can be verified in the tables up to minor random errors, indicating the effectiveness of the proposed eigenvalue selection criterion. An exception is that, when $p = 6$ and $X$ is uniformly distributed on a sphere or in a ball, the improvement by using the adjusted directional regression is significant in Model I. By carefully examining which eigenvectors we selected in each sample, the improvement is due to the non-negligible error in estimating the kernel matrix at the current sample size, which makes the sample eigenvalues corresponding to the central subspace possibly not the largest. From Table 7, this possibility vanishes as $n$ increases.

Table 7 suggests that the adjusted and conventional directional regressions nearly coincide in most cases when $p = 10$. From an omitted simulation study, we observed the same phenomenon for the adjusted pHd when $p = 20$. This is expected from Theorem 6: when $X$ belongs to the quadratic variance ellipticity family with sufficiently large dimension $p$, the conventional second-order inverse regression methods are consistent and the proposed adjustments are redundant.

## 8. A Data Example

We applied the adjusted directional regression to the pen-based recognition of handwritten digits data set, available in the UCI machine learning repository at `https://archive.ics.uci.edu/ml/machine-learning-databases/pendigits/`; see also Li and Wang (2007). Forty-four writers were collected in the data set, each of whom was asked to write 250 randomly generated digits, ranging from 0 to 9, on a tablet. The coordinate information of each written digit was recorded and converted into a 16-dimensional predictor. Interest is in classifying digits based on the coordinate information and, for this purpose, the data set was divided into a training set, which consists of the first thirty writers, and a testing set, which consists of the other fourteen. Since the digits 1 and 7 are sometimes difficult to distinguish in practice, we focused on the classification of these two digits in the training set, which includes 779 and 778 observations, respectively.

We assumed the predictor to be elliptically distributed. From an omitted exploratory data analysis, most components of the predictor have heavy-tailed

distributions, which would adversely affect the stableness of the sample moments. To address this issue, we applied the transformation in Dong, Yu and Zhu (2015), which preserves the ellipticity of the predictor. The transformation also enhances the plausibility of the quadratic variance condition (2.1). For example, taking $d = 3$ as the working dimension and applying the goodness-of-fit test in the simulation studies, the averaged p-value was less than 0.01 before the transformation and 0.25 after. Thus, it is reasonable to apply the adjusted directional regression to the transformed data.

Following Luo and Li (2016), we assumed that $d < p(\log p)^{-1}$, so $d \leq 5$. For each working dimension $d$ in this range, we found that the adjusted and conventional directional regressions coincided when $p$ was relatively large, which again conformed to Theorem 6. Thus, we can apply the conventional directional regression in the data set.

We applied the ladle estimator (Luo and Li, 2016) to determine that $d = 3$. The scatter plots for the reduced predictor and its first component are shown in the lower-left and upper-left panels of Figure 2, respectively. Following Li and Wang (2007), we also include the response in the latter to enhance the visualization. Clearly, the first component of the reduced predictor represents the direction in which the groups mean of the original predictor differ, and the second and third components represent those in which the group covariance matrices differ. As the two digits are clearly distinguished in the scatter plot, the reduced predictor from directional regression is an appropriate classifier.

For reference, we also applied SIR and SAVE. The corresponding scatter plots are shown in the upper-right and lower-right panels of Figure 2, respectively. Because the response is binary, SIR can only recover an univariate reduced predictor. From the scatter plots in the upper panels, this predictor has the same effect as the first component from directional regression on differentiating the group means of the original predictor. Thus directional regression is more comprehensive than SIR. On the other hand, SAVE is effective for the data set without adjustment and produces similar results to directional regression, as the corresponding scatter plots are nearly identical up to rotations. This similarity can also be seen from the fact that the correlation matrix between the two sets of reduced predictors has determinant 0.988.

## Appendix

*Proof of Theorem* 1. By the ellipticity of $X$, the diagonal elements of $B_\gamma$ are

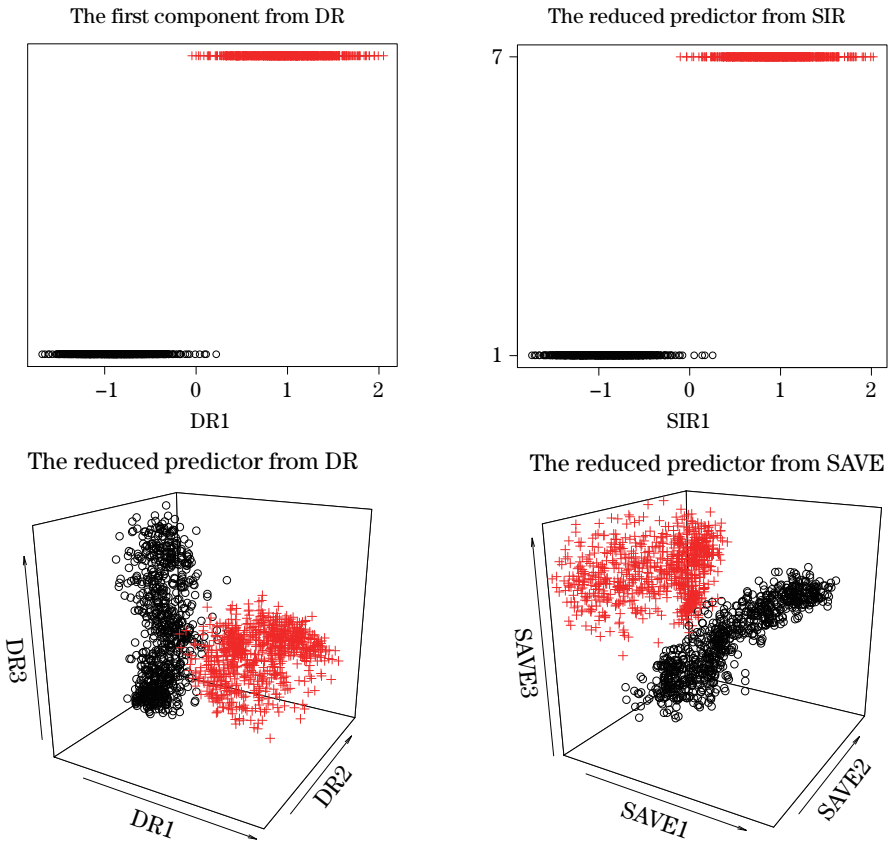


Figure 2. The vertical axis in the upper panels represents the binary response. The horizontal axis in the upper-left panel represents the reduced predictor from SIR; that in the upper-right panel represents the first component of the reduced predictor from directional regression. The axes in the lower-left panel represent the three components of the reduced predictor from directional regression, as do those in the lower-right panel for SAVE. ($\circ$, $+$) represent observations with digits 1 and 7 correspondingly.

equal, as well as its off-diagonal elements. We denote them by $\delta$ and $\rho$ accordingly. Since $E(\gamma^{\mathsf{T}}X|\beta^{\mathsf{T}}X) = \gamma^{\mathsf{T}}\beta(\beta^{\mathsf{T}}\beta)^{-1}\beta^{\mathsf{T}}X = 0$, we have $\operatorname{var}(\gamma^{\mathsf{T}}X|\beta^{\mathsf{T}}X) = E[(\gamma^{\mathsf{T}}X)^2|\beta^{\mathsf{T}}X]$. By the symmetry of $X$, we have

$$0 = E\{(\gamma^{\mathsf{T}}X)^2(\beta_1^{\mathsf{T}}X)(\beta_2^{\mathsf{T}}X)\} = E\{E[(\gamma^{\mathsf{T}}X)^2|\beta^{\mathsf{T}}X](\beta_1^{\mathsf{T}}X)(\beta_2^{\mathsf{T}}X)\}$$
$$= a_\gamma E\{(\beta_1^{\mathsf{T}}X)(\beta_2^{\mathsf{T}}X)\} + 2\rho E\{(\beta_1^{\mathsf{T}}X)^2(\beta_2^{\mathsf{T}}X)^2\}.$$

Again by the symmetry of $X$, $E[(\beta_1^{\mathsf{T}}X)(\beta_2^{\mathsf{T}}X)] = 0$, which implies that $\rho = 0$. Thus $B_\gamma$ has the form $\delta I_p$. We further have

$$1 = E(\gamma^{\mathsf{T}}X)^2 = a_\gamma + d\delta,$$

$$E(X_1^2X_2^2) = E\{(\gamma^{\mathsf{T}}X)^2(\beta_1^{\mathsf{T}}X)^2\} = a_\gamma + \delta\{E(X_1^4) + (d-1)E(X_1^2X_2^2)\},$$

which together imply that

$$\delta = \frac{E(X_1^2X_2^2) - 1}{E(X_1^4) + (d-1)E(X_1^2X_2^2) - d}. \tag{A.1}$$

By the symmetry of $X$,

$$E(\|X\|^4) = pE(X_1^4) + p(p-1)E(X_1^2X_2^2) = p^2E(R^4).$$

Writing $U$ as $(U_1, \ldots, U_p)^{\mathsf{T}}$, then $E(X_1^4)/E(X_1^2X_2^2) = E(U_1^4)/E(U_1^2U_2^2)$, which is an invariant of $R$. If $X$ is normally distributed, then this ratio is 3. Thus $E(X_1^2X_2^2) = E(R^4)p/(p+2)$ and (A.1) can be written in the form of (2.3). To see that $\delta$ increases with $E(R^4)$, by Jensen's inequality, $E(R^4) \geq E^2(R^2) = 1$. Writing $E(R^4)$ as $t$, we have

$$\frac{\partial\delta}{\partial t} \propto p\{p\,(d+2)\,t - d(p+2)\} - p(d+2)\{p\,t - (p+2)\} = 2p(p+2) > 0.$$

Thus $\delta$ reaches the minimum value $-1/(p-d)$ at $t = 1$, and converges to the least upper bound $1/(d+2)$ if we (hypothetically) allow $t$ to tend to infinity.

*Proof of Theorem* 2. Let $\lambda_i(M_{pHd})$ be $\lambda_i$ and $Y - E(Y)$ be $Y_c$, where "c" means "centered", and assume that the central mean subspace is spanned by the first $d$ columns of $I_p$, denoted by $\{e_1, \ldots, e_d\}$. By the ellipticity of $X$, for any $i = 1, \ldots d$ and $j = d+1, \ldots, p$, we have $E(X_iX_jY_c) = E[X_iX_jE(Y_c|X_1, \ldots, X_d)] = E[X_iE(X_j|X_1, \ldots, X_d)E(Y_c|X_1, \ldots, X_d)] = 0$. Thus

$$M_{pHd}e_i = E(XX_iY_c) = (E(X_1X_iY_c), \ldots, E(X_dX_iY_c), 0, \ldots, 0)^{\mathsf{T}}$$

and $M_{pHd}e_i \in \mathcal{S}(e_1, \ldots, e_d)$, which means that $\mathcal{S}_{E(Y|X)}$ can be spanned by a set of eigenvectors of $M_{pHd}$. To show (3.1), we can equivalently show that for any unit-length $\gamma \in \mathcal{S}^{\perp}(\beta_{A_0}(M_{pHd}))$, $\gamma^{\mathsf{T}}M_{pHd}\gamma = \delta\sum_{i\in A_0}\lambda_i$. Again by the ellipticity of $X$, we only need to show this when $\gamma^{\mathsf{T}}X = X_p$. By Theorem 1,

$$\begin{aligned}\gamma^{\mathsf{T}}M_{pHd}\gamma &= E[X_p^2Y_c] = E\{E(X_p^2|X_1, \ldots, X_d)Y_c\} = E\{(1 - d\delta + \delta\textstyle\sum_{i=1}^d X_i^2)Y_c\} \\ &= \delta\textstyle\sum_{i=1}^d E(X_i^2Y_c) = \delta\sum_{i\in A_0}\lambda_i.\end{aligned}$$

To show that $1 \in A_0$ when $p > 2d$ and $M_{pHd} \neq \mathbf{0}$, the latter implying $\lambda_1 \neq 0$, it suffices to show that $|\delta\sum_{i\in A_0}\lambda_i| < |\lambda_1|$. This is straightforward, since the left-hand side is bounded from above by $d|\delta||\lambda_1|$ and, by Theorem 1, $|\delta| < 1/d$ in this case.

*Proof of Theorem* 3. We denote $\beta(M_{pHd})$ by $\beta$ and $\lambda_i(M_{pHd})$ by $\lambda_i$ for each $i = 1, \ldots, p$, and denote $\hat{\beta}$ similarly. By Theorem 2, $\beta_{A_0}$ spans the central mean subspace. Let $f$ be the mapping such that for any $A \in \{1, \ldots, p\}^d$,

$$f(A) = \sum_{j\notin A} |\lambda_j - \delta \sum_{i\in A} \lambda_i|.$$

Then $f \geq 0$ and the uniqueness of the solution to (3.1) is equivalent to the uniqueness of the minimizer of $f$. Thus $f(A) = 0$ if and only if $A = A_0$. For any set $A$, denote (3.2) as $\hat{f}(A)$. Since $\hat{\delta}$ is a $\sqrt{n}$-consistent estimator of $\delta$, by Li (1992) and Zhao, Krishnaiah and Bai (1986), $\hat{f}(A) = f(A) + O_P(n^{-1/2})$. Therefore, $\hat{f}$ is minimized at $A_0$ in probability. To see the $\sqrt{n}$-consistency of the adjusted pHd, by Zhao, Krishnaiah and Bai (1986) again, $\mathcal{S}(\hat{\beta}_{A_0})$ is a $\sqrt{n}$-consistent estimator of $\mathcal{S}(\beta_{A_0})$. Therefore, for any $\{c_n : n \in \mathbb{N}\} \in \mathbb{R}$ such that $c_n \to \infty$,

$$\begin{aligned}
&P(\|\Pi(\hat{\beta}_{\hat{A}}) - \Pi(\beta_{A_0})\| > c_n n^{-1/2}) \\
&= P(\|\Pi(\hat{\beta}_{A_0}) - \Pi(\beta_{A_0})\| > c_n n^{-1/2} | \hat{A} = A_0) P(\hat{A} = A_0) \\
&\quad + P(\|\Pi(\hat{\beta}_{\hat{A}}) - \Pi(\beta_{A_0})\| > c_n n^{-1/2} | \hat{A} \neq A_0) P(\hat{A} \neq A_0) \\
&\leq P(\|\Pi(\hat{\beta}_{A_0}) - \Pi(\beta_{A_0})\| > c_n n^{-1/2}) + P(\hat{A} \neq A_0) = o(1).
\end{aligned}$$

This completes the proof.

*Proof of Theorem* 4. Here we assume that the central subspace is spanned by the first $d$ columns of the identity matrix $I_p = \{e_1, e_2, \ldots, e_p\}$. By Li and Wang (2007), $M_{DR}$ can be alternatively written as

$$\begin{aligned}
M_{DR} = {} & 2E\{E^2(XX^{\mathsf{T}}|Y)\} + 2E^2\{E(X|Y)E^{\mathsf{T}}(X|Y)\} \\
& + 2E\{E^{\mathsf{T}}(X|Y)E(X|Y)\}E\{E(X|Y)E^{\mathsf{T}}(X|Y)\} - 2I_p.
\end{aligned}$$

By the ellipticity of $X$, for any $i = 1, \ldots d$ and $j = d+1, \ldots, p$, we have $E[X_j(1, X_i)|Y] = E[E(X_j|X_1, \ldots, X_d)(1, X_i)|Y] = 0$. Thus

$$\begin{aligned}
e_j^{\mathsf{T}} M_{DR} e_i = {} & 2\textstyle\sum_{l=1}^p E\{E(X_j X_l|Y)E(X_l X_i|Y)\} \\
& + 2\textstyle\sum_{l=1}^p E\{E(X_j|Y)E(X_l|Y)\}E\{E(X_l|Y)E(X_i|Y)\} \\
& + 2\textstyle\sum_{l=1}^p E\{E^2(X_l|Y)\}E\{E(X_j|Y)E(X_i|Y)\} - 2e_j^{\mathsf{T}} e_i \\
= {} & 0,
\end{aligned}$$

and $M_{DR} e_i \in \mathcal{S}(e_1, \ldots, e_d)$, which means that $\mathcal{S}_{Y|X}$ can be spanned by a set of eigenvectors of $M_{DR}$. To show (4.1), we can equivalently show that for any unit-length $\gamma \in \mathcal{S}_{Y|X}^{\perp}$, $\gamma^{\mathsf{T}} M_{DR} \gamma$ is equal to the right-hand side of (4.1). Again by the ellipticity of $X$, we only need to show this when $\gamma^{\mathsf{T}} X = X_p$. For any $j = d+1, \ldots, p-1$, we have $E(X_p X_j|Y) = E\{E(X_p X_j|X_1, \ldots, X_d)|Y\} = 0$. Since $E(X_p^2|Y) = E\{E(X_p^2|X_1, \ldots, X_d)|Y\}$, together with Theorem 1,

$$\begin{aligned}
\gamma^{\mathsf{T}} M_{DR} \gamma &= 2\textstyle\sum_{l=1}^p E\{E^2(X_p X_l|Y)\} - 2 \\
&= 2E\{E^2(X_p^2|Y)\} - 2
\end{aligned}$$

$$
\begin{aligned}
&= 2E[E^2\{E(X_p^2|X_1,\ldots,X_d)|Y\}] - 2\\
&= 2E\{E^2(1-d\delta+\delta\textstyle\sum_{i=1}^d X_i^2|Y)\} - 2\\
&= 2(1-d\delta)^2 + 4\delta(1-d\delta)\textstyle\sum_{i=1}^d E(X_i^2)\\
&\quad + 2\delta^2\textstyle\sum_{k=1}^d\sum_{l=1}^d E\{E(X_k^2|Y)E(X_l^2|Y)\} - 2\\
&= -2d^2\delta^2 + 2\delta^2\textstyle\sum_{k=1}^d\sum_{l=1}^d E\{E(X_k^2|Y)E(X_l^2|Y)\}.
\end{aligned}
$$

To show that $1 \in A_0$ when $p > 2d$ and $M_{DR} \neq \mathbf{0}$, the latter implying $\lambda_1 \neq 0$, it suffices to show that $|\gamma^{\mathsf{T}} M_{DR}\gamma| < |\lambda_1|$. Since $M_{DR}$ is positive semi-definite (Li and Wang (2007)), this is equivalent to that $\gamma^{\mathsf{T}} M_{DR}\gamma < \lambda_1$. For each $k = 1,\ldots,d$, we have

$$\lambda_1 \geq e_k^{\mathsf{T}} M_{DR} e_k \geq 2\textstyle\sum_{j=1}^d E\{E^2(X_kX_j|Y)\} - 2 \geq 2E\{E^2(X_k^2|Y)\} - 2.$$

Since for each $l = 1,\ldots,d$, $E\{E(X_k^2|Y)E(X_l^2|Y)\} \leq \sum_{i=k,l} E\{E^2(X_i^2|Y)\}/2$, we have

$$e_p^{\mathsf{T}} M_{DR} e_p \leq -2d^2\delta^2 + d^2\delta^2(\lambda_1+2) = d^2\delta^2\lambda_1.$$

By Theorem 1, $|\delta| < 1/d$ in this case. Thus for any unit-length $\gamma \in \mathcal{S}_{Y|X}^{\perp}$, $\gamma^{\mathsf{T}} M_{DR}\gamma < \lambda_1$, which means that any eigenvector associated with $\lambda_1$ must be in $\mathcal{S}_{Y|X}$. Hence $1 \in A_0$.

*Proof of Theorem* 5. The proof is similar to that of Theorem 3, and is omitted.

*Proof of Theorem* 6. Since $E\{(R^{(p)})^4\} \to 1$, we have $\delta^{(p)} \to 0$. By condition (a), $d^{(p)} < 2p$ for all large $p$, thus $\max\{\beta_i^{(p)\mathsf{T}} M_{DR}\beta_i^{(p)}, i = 1,\ldots,d^{(p)}\} = \lambda_1^{(p)}$. From the proof of Theorem 4 we have that, for any unit-length $\gamma \in \mathcal{S}_{Y|X}^{\perp}$,

$$\gamma^{\mathsf{T}} M_{DR}\gamma \leq [d^{(p)}]^2[\delta^{(p)}]^2\lambda_1^{(p)} = o(\lambda_1^{(p)}).$$

Condition (b) then implies the conclusion of the theorem.

# References


Cook, R. D. (1998). *Regression Graphics*. Wiley, New York.

Cook, R. D. and Li, B. (2002). Dimension reduction for conditional mean in regression. *The Annals of Statistics* **30**, 455–474.

Cook, R. D. and Weisberg, S. (1991). Discussion of "Sliced inverse regression for dimension reduction". *Journal of the American Statistical Association* **86**, 316–342.

Diaconis, P. and Freedman, D. (1984). Asymptotics of graphical projection pursuit. *The Annals of statistics* **12**, 793–815.

Dong, Y. and Li, B. (2010). Dimension reduction for non-elliptically distributed predictors: second-order methods. *Biometrika* **97**, 279–294.

Dong, Y., Yu, Z. and Zhu, L. (2015). Robust inverse regression for dimension reduction. *Journal of Multivariate Analysis* **134**, 71–81.

Hall, P. and Li, K.-C. (1993). On almost linearity of low dimensional projections from high dimensional data. *The Annals of Statistics* **21**, 867–889.

Johnson, M. E. (2013). *Multivariate Statistical Simulation: A Guide to Selecting and Generating Continuous Multivariate Distributions*. John Wiley & Sons.

Li, B. and Dong, Y. (2009). Dimension reduction for nonelliptically distributed predictors. *The Annals of Statistics* **37**, 1272–1298.

Li, B. and Wang, S. (2007). On directional regression for dimension reduction. *Journal of the American Statistical Association* **35**, 2143–2172.

Li, K. C. (1992). On principal Hessian directions for data visualization and dimension reduction: another application of Stein's lemma. *Journal of the American Statistical Association* **87**, 1025–1039.

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

Li, K. C. and Duan, N. (1989). Regression analysis under link violation. *The Annals of Statistics* **17**, 1009–1052.

Luo, R., Wang, H. and Tsai, C.-L. (2009). Contour projected dimension reduction. *The Annals of Statistics* **37**, 3743–3778.

Luo, W. and Li, B. (2016). Combining eigenvalues and variation of eigenvectors for order determination. *Biometrika*, accepted.

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

Yin, X., Li, B. and Cook, R. D. (2008). Successive direction extraction for estimating the central subspace in a multiple-index regression. *Journal of Multivariate Analysis* **99**, 1733–1757.

Zhao, L. C., Krishnaiah, P. R. and Bai, Z. D. (1986). On detection of the number of signals in presence of white noise. *Journal of Multivariate Analysis* **20**, 1–25.

Paul H. Chook Department of Information Systems and Statistics, Baruch College, New York, NY 10010, USA.

E-mail: wei.luo@baruch.cuny.edu

(Received January 2016; accepted February 2017)