Statistica Sinica **30** (2020), 1027-1048
doi:https://doi.org/10.5705/ss.202016.0188

ON AGGREGATE DIMENSION REDUCTION

Qin Wang, Xiangrong Yin, Bing Li and Zhihui Tang

The University of Alabama, University of Kentucky
Penn State University and PPDI

Abstract: We propose a dimension-reduction method based on the aggregation of localized estimators. The dual process of localization and aggregation helps to mitigate the bias due to the symmetry in the predictor distribution, and achieves exhaustive estimation of the dimension-reduction space. This approach does not involve numerical optimization or the inversion of large matrices, resulting in a fast and stable algorithm suited for processing large, high-dimensional data sets. We demonstrate the efficacy of our method via simulation and real-data applications.

Key words and phrases: Central subspace, k-nearest neighbor, sliced inverse regression.

1. Introduction

Suppose that Y is a univariate response and $\mathbf{X}$ is a p-dimensional vector of continuous predictors. In its full generality, the goal of a regression is to infer the conditional distribution of Y, given $\mathbf{X}$. However, because of the curse of dimensionality (Bellman (1961)), regressions with large p can be difficult, in practice. The basic idea of a *sufficient dimension reduction* (SDR; Li (1991); Cook (1998)) is to replace the predictor vector by its projection onto a low-dimensional subspace, without losing information on the conditional distribution of $Y|\mathbf{X}$, and without assuming any specific model for $Y|\mathbf{X}$.

In mathematical terms, a sufficient dimension-reduction space is a subspace $\mathcal{S}$ of $\mathbb{R}^p$, such that Y and $\mathbf{X}$ are independent, conditioning on $\mathbf{P}_{\mathcal{S}}\mathbf{X}$, where $\mathbf{P}_{\mathcal{S}}$ is a projection onto $\mathcal{S}$. The intersection of all such $\mathcal{S}$ if itself satisfies the above independent condition is called the *central subspace*, and is denoted by $\mathcal{S}_{Y|\mathbf{X}}$. As shown in Cook (1998) and Yin, Li and Cook (2008), under very mild conditions, the central subspace exists and is the smallest and unique dimension-reduction space. The dimension of $\mathcal{S}_{Y|\mathbf{X}}$ is called the structural dimension, and is denoted by $d_{Y|\mathbf{X}}$.

A widely used class of estimators of the central subspace is based on inverse conditional moments, such as $E(\mathbf{X}|Y)$ and $\mathrm{Var}(\mathbf{X}|Y)$. This includes methods such as the sliced inverse regression (SIR; Li (1991)) and sliced average variance estimation (SAVE; Cook and Weisberg (1991)), and hybrids of the two (Ye and Weiss (2003)), as well as the parametric inverse regression (Bura and Cook (2001a)), sliced average third moment (Yin and Cook (2003)), contour regression (Li, Zha and Chiaromonte (2005)), minimum discrepancy approach (Cook and Ni (2005)), and directional regression (Li and Wang (2007)), among others.

The sliced inverse regression was the first general dimension-reduction method, and has been widely researched, with many subsequent extensions and refinements. Hsing and Carroll (1992), Zhu and Ng (1995), and Zhu and Fang (1996) studied the asymptotic properties of the SIR estimator and its variations. Schott (1994), Velilla (1998), and Bura and Cook (2001b) introduced asymptotic inference procedures to determine the dimension of the subspace estimated by the SIR. Following Cook and Weisberg (1991), Cook and Yin (2001) developed a permutation testing procedure to determine this dimension. Chen and Li (1998) studied the relation between the SIR and a maximal correlation. Hsing (1999) used the nearest-neighbor method to develop a variation of the SIR that is applicable to multivariate responses. Naik and Tsai (2000) compared the performance of the SIR with that of the partial least squares in the context of a single-index model. Cook and Critchley (2000) showed that dimension-reduction methods, in general, and the SIR in particular, can be useful for identifying outliers and regression mixtures. Bura and Cook (2001a), Fung et al. (2002), Bura (2003), and Wang and Yin (2011) further expanded the scope of the SIR by replacing the inverse conditional mean $E(\mathbf{X}|Y)$ with a parametric regression or basis expansion. Li, Cook and Nachtsheim (2004) proposed a cluster-based estimation to mitigate the effect of nonlinearity on the predictors, focusing on single-index models. Zhu, Miao and Peng (2006) studied the asymptotic behavior of the SIR when the number of covariates increases with the sample size. Recently, Wu, Liang and Mukherjee (2010) developed an extension by replacing the global average with the local average for each data point, thus alleviating the issue of degenerate solutions. The SIR has found wide application in diverse fields such as computer vision (Ling, Yin and Bhandarkar (2003); Ling et al. (2005)) and the biological sciences (Chiaromonte and Martinelli (2002); Bura and Pfeiffer (2003); Li and Li (2004)).

In this study we develop an aggregate dimension-reduction (ADR) procedure. The theoretical basis of this method is that the central subspace $\mathcal{S}_{Y|\mathbf{X}}$ can always be decomposed into *finitely many* local dimension-reduction spaces, and that we can aggregate the local spaces to recover $\mathcal{S}_{Y|\mathbf{X}}$. The dual process of localization and aggregation brings two benefits. First, because all differential functions are approximately linear locally, we no longer need to impose a strong linearity assumption on the conditional mean of the predictors, as required by the SIR. Second, it leads to an exhaustive estimation of the central subspace $\mathcal{S}_{Y|\mathbf{X}}$.

We outline the main ideas and benefits of localized dimension reductions in Section 2. These ideas are rigorously formulated and developed at the population level in Section 3. In Sections 4 and 5, we provide estimation procedures for the localized SIR using the k-nearest neighborhood, and discuss various issues involved in the estimation, respectively. Simulation studies and two real-data examples are presented in Sections 6 and 7, respectively. Section 8 concludes the paper. All proofs are relegated to the Appendix, published as online Supplementary Material.

2. Principle of Finite Aggregation

ADR consists of performing ordinary sufficient dimension reduction over a number of local regions in the predictor sample space, and then aggregating the results to recover the global dimension reduction subspace. We first present the two benefits of this dual process in concrete terms. Let $\mathbf{B} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_d)$ be a $p \times d$ matrix, the columns of which form an orthonormal basis of the central subspace. The SIR and many other dimension-reduction methods require the following *linearity condition* on $\mathbf{X}$:

$$E(\mathbf{X}|\mathbf{B}^T\mathbf{X}) \text{ is a linear function of } \mathbf{B}^T\mathbf{X}. \tag{2.1}$$

Under this assumption, the random vector $E(\mathbf{X}|Y) - E(\mathbf{X})$ is contained almost surely in $\boldsymbol{\Sigma}_{\mathbf{X}}\mathcal{S}_{Y|\mathbf{X}}$, where $\boldsymbol{\Sigma}_{\mathbf{X}}$ denotes the covariance matrix of $\mathbf{X}$ (Li (1991)). Because $\mathbf{B}$ is unknown, this condition is often assumed to hold for all $p \times d$ matrices, which is equivalent to requiring that $\mathbf{X}$ have an elliptically contoured distribution (Eaton (1986)), an assumption that seems too strong for many applications. However, if we restrict $\mathbf{X}$ to a relatively small region, then, as long as the function $\mathbf{m}(\mathbf{u}) = E(\mathbf{X}|\mathbf{B}^T\mathbf{X} = \mathbf{u})$ is differentiable, $E(\mathbf{X}|\mathbf{B}^T\mathbf{X})$ can be reasonably well approximated by a linear function of $\mathbf{B}^T\mathbf{X}$.

The second benefit is that it overcomes a well-known drawback of SIR. That is, if the distribution of $\mathbf{X}$ given Y is symmetric about $E(\mathbf{X})$ along certain directions of $\mathbf{X}$, then the random vector $E(\mathbf{X}|Y) - E(\mathbf{X})$ vanishes along those directions, and consequently cannot provide any information about those directions. For example, consider the model

$$Y = 3(\beta^T\mathbf{X})^2 + 0.2\varepsilon,$$

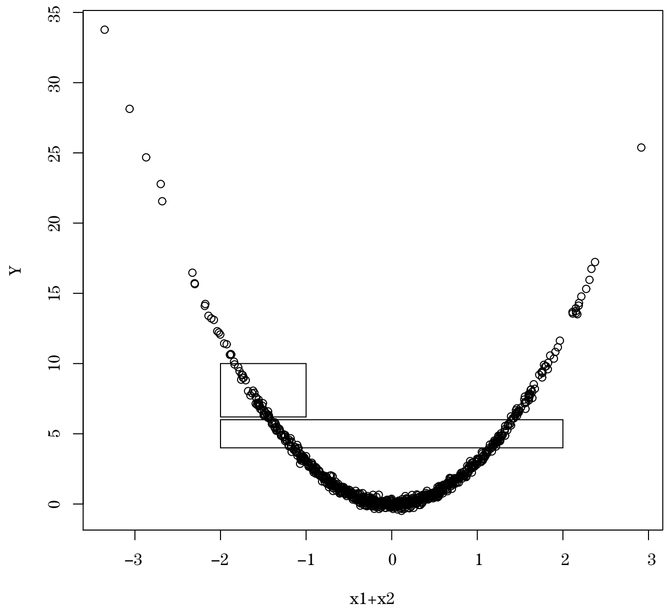

Figure 1. A symmetric model that cannot be estimated by a global SIR.

where $\boldsymbol{\beta} = (1, 1, 0, \ldots, 0)'$, $\varepsilon \sim N(0, 1)$, $\varepsilon \perp\!\!\!\perp \mathbf{X}$, and $\mathbf{X} \sim N(0, \mathbf{I}_{10})$. Although the linearity condition (2.1) is satisfied, the random vector $E(\mathbf{X}|Y) - E(\mathbf{X})$ is degenerate at $\mathbf{0}$, which does not tell us anything about $\boldsymbol{\Sigma}_{\mathbf{X}}\mathcal{S}_{Y|\mathbf{X}}$, even though it does belong to $\boldsymbol{\Sigma}_{\mathbf{X}}\mathcal{S}_{Y|\mathbf{X}}$. This situation is illustrated in Figure 1, where $E(\mathbf{X}|Y) - E(\mathbf{X})$ in the longer rectangle vanishes. However, if we restrict $\mathbf{X}$ to a local region, as indicated by the shorter rectangle, then $E(\mathbf{X}|Y) - E(\mathbf{X})$ does not vanish.

To construct local dimension-reduction spaces, assume $(\mathbf{X}, Y)$ has a joint density $f(\mathbf{x}, y)$. Let $p(\mathbf{x})$, $g(y)$, and $h(y|\mathbf{x})$ denote the marginal density of $\mathbf{X}$, the marginal density of Y, and the conditional density of Y given $\mathbf{X} = \mathbf{x}$, respectively. Let $\Omega_{\mathbf{X}}$ and Ω_Y be the support of $\mathbf{X}$ and Y, respectively; that is, $\Omega_{\mathbf{X}} = \{\mathbf{x} : p(\mathbf{x}) > 0\}$, $\Omega_Y = \{y : g(y) > 0\}$. For convenience, assume that the support of f is the Cartesian product $\Omega_{\mathbf{X}} \times \Omega_Y$. Though this assumption is not crucial for our subsequent analysis, it does help to simplify the discussion. In summary, we assume

$$\Omega_{\mathbf{X},Y} = \{(\mathbf{x}, y) : f(\mathbf{x}, y) > 0\} = \{(\mathbf{x}, y) : p(\mathbf{x}) > 0, g(y) > 0\} = \Omega_{\mathbf{X}} \times \Omega_Y. \quad (2.2)$$

Let G be any open set in $\Omega_{\mathbf{X}}$. Let $(\mathbf{X}_G, Y_G)$ be defined as $(\mathbf{X}, Y)$ restricted on the set G; that is, for any Borel set $A \subseteq \Omega_{\mathbf{X}} \times \Omega_Y$, we have

$$P[(\mathbf{X}_G, Y_G) \in A] = \frac{P[(\mathbf{X}, Y) \in A \cap (G \times \Omega_Y)]}{P[(\mathbf{X}, Y) \in G \times \Omega_Y]}$$
$$= \frac{P[(\mathbf{X}, Y) \in A \cap (G \times \Omega_Y)]}{P(\mathbf{X} \in G)}. \tag{2.3}$$

This defining relation uniquely determines the densities and conditional densities of the localized random pair $(\mathbf{X}_G, Y_G)$, as given by the following proposition.

Proposition 1. *Suppose that* $(\mathbf{X}_G, Y_G)$ *is defined by* (2.3). *Then:*

1. *The joint density of* $(\mathbf{X}_G, Y_G)$ *is* $f_G(\mathbf{x}, y) = f(\mathbf{x}, y)/P(\mathbf{X} \in G), \quad (\mathbf{x}, y) \in G \times \Omega_Y$;
2. *The marginal density of* $\mathbf{X}_G$ *is* $p_G(\mathbf{x}) = p(\mathbf{x})/P(\mathbf{X} \in G), \quad \mathbf{x} \in G$;
3. *The conditional density of* $Y_G|\mathbf{X}_G$ *is* $h_G(y|\mathbf{x}) = h(y|\mathbf{x}), \quad (\mathbf{x}, y) \in G \times \Omega_Y$;
4. *The marginal density of* Y_G *is*
$$g_G(y) = \frac{1}{P(\mathbf{X} \in G)} \int_G f(\mathbf{x}, y) d\mathbf{x}, \quad y \in \Omega_Y.$$

The proof is simple, and thus is omitted. An important point of this proposition is that the conditional densities of $Y_G|\mathbf{X}_G$ and $Y|\mathbf{X}$ coincide over the cylinder $G \times \Omega_Y$. The central subspace of Y_G versus $\mathbf{X}_G$, $\mathcal{S}_{Y_G|\mathbf{X}_G}$ is called the *local central subspace* for the neighborhood G. Intuitively, any direction in a local central subspace $\mathcal{S}_{Y_G|\mathbf{X}_G}$ must also belong to the global central subspace $\mathcal{S}_{Y|\mathbf{X}}$, because any local relation between Y_G and $\mathbf{X}_G$ must be part of the global relation between Y and $\mathbf{X}$. At the same time, any relation existing between Y and $\mathbf{X}$ globally must be reflected in some local area G. In fact, we only need a *finite* number of local central subspaces to recover the global central subspace.

Theorem 1. *Suppose* $\Omega_{\mathbf{X}}$ *is an open set in* $\mathbb{R}^p$. *Then, there exist a finite number of open sets, say* $G_1, \ldots, G_m$, *in* $\Omega_{\mathbf{X}}$, *such that* $\mathcal{S}_{Y|\mathbf{X}} = \text{span}\{\mathcal{S}_{Y_{G_i}|\mathbf{X}_{G_i}} : i = 1, \ldots, m\}$.

This theorem, which we refer to as the finite aggregation principle, plays a fundamental role in our method: it guarantees that we can join a finite number of local central subspaces to recover the global central subspace. The proof of Theorem 1 is given in the Appendix.

3. Bias-reducing Effect of Localization

Let $\|G\|$ denote the "diameter" of an open set G in $\Omega_{\mathbf{X}}$, in the sense that

$$\|G\| = \sup\{\|\mathbf{x} - \mathbf{x}'\| : \mathbf{x} \in G, \mathbf{x}' \in G\}.$$

Let $\boldsymbol{\mu}_G = E(\mathbf{X}_G)$ and $\dot{h}(y|\mathbf{x}) = \partial h(y|\mathbf{x})/\partial \mathbf{x}$. Consider the matrices

$$\mathbf{H}_G = E[\dot{h}(Y_G|\boldsymbol{\mu}_G)\dot{h}^T(Y_G|\boldsymbol{\mu}_G)] \quad \text{and} \quad \mathbf{H}^*_G = E[\dot{h}(Y_G|\mathbf{X}_G)\dot{h}^T(Y_G|\mathbf{X}_G)].$$

From a result of Zhu and Zeng (2006), it can be deduced that

$$\text{span}(\mathbf{H}_G) \subseteq \text{span}(\mathbf{H}^*_G) = \mathcal{S}_{Y_G|\mathbf{X}_G}.$$

Let $\boldsymbol{\beta}_G$ and $\mathbf{B}_G$ be matrices of full column rank, such that $\text{span}(\boldsymbol{\beta}_G) = \text{span}(\mathbf{H}_G)$ and $\text{span}(\mathbf{B}_G) = \text{span}(\mathbf{H}^*_G)$. We show that (i) if $\|G\|$ is small, then, approximately, $\boldsymbol{\beta}_G$ and $\mathbf{B}_G$ share the same column space; (ii) the shared column space is approximately the local central subspace; (iii) the latter can be approximated by a localized SIR; and (iv) in an important special case, this space has dimension no more than one. Let $\boldsymbol{\Sigma}_G$ denote the variance matrix of $\mathbf{X}_G$:

$$\int_G (\mathbf{x} - \boldsymbol{\mu}_G)(\mathbf{x} - \boldsymbol{\mu}_G)^T p_G(\mathbf{x}) d\mathbf{x}.$$

Note that this matrix is of order $O(\|G\|^2)$ as $\|G\| \to 0$. Let $\bar{G}$ denote the closure of G, and let $\mathbf{P}_{\boldsymbol{\beta}_G}$ be the projection onto $\text{span}(\boldsymbol{\beta}_G)$. That is,

$$\mathbf{P}_{\boldsymbol{\beta}_G} = \boldsymbol{\beta}_G(\boldsymbol{\beta}_G^T\boldsymbol{\beta}_G)^{-1}\boldsymbol{\beta}_G^T.$$

Theorem 2. *Suppose that, for a fixed $y \in \Omega_Y$, $g(y) > 0$, $h(y|\mathbf{x})$ is twice differentiable with respect to $\mathbf{x}$ on $\bar{G}$, and the second derivatives are bounded on $\bar{G}$. Then, as $\|G\| \to 0$, and almost everywhere on Ω_Y,*

$$\left|\boldsymbol{\Sigma}_G^{-1}[E(\mathbf{X}_G|y) - E(\mathbf{X}_G)] - \mathbf{P}_{\boldsymbol{\beta}_G}\boldsymbol{\Sigma}_G^{-1}[E(\mathbf{X}_G|y) - E(\mathbf{X}_G)]\right|_{\mathcal{F}} = O(\|G\|), \quad (3.1)$$

where $|A|_{\mathcal{F}}$ denotes the Frobenius norm of a matrix A.

The proof of Theorem 2 is provided in the Appendix.

Note that the relation given in (3.1) tells us that, except for an error of magnitude $O(\|G\|^2)$, the local SIR vector, $\|G\|\boldsymbol{\Sigma}_G^{-1}[E(\mathbf{X}_G|y) - E(\mathbf{X}_G)]$, belongs to the central subspace. In other words, the bias due to the nonlinearity of $E(\mathbf{X}_G|\boldsymbol{\beta}_G^T\mathbf{X}_G)$ is two orders of magnitude smaller than the bias of the global inverse mean $\boldsymbol{\Sigma}^{-1}[E(\mathbf{X}|y)-E(\mathbf{X})]$. In fact, if we assume slightly stronger regularity conditions, this bias can be further reduced by two orders of magnitude.

Theorem 3. *Suppose that in addition to the conditions in Theorem 2, $h(y|\mathbf{x})$ has a bounded third derivative with respect to $\mathbf{x}$, $p(\mathbf{x})$ has a bounded first derivative on $\bar{G}$, and G is an open ball in $\Omega_{\mathbf{X}}$. Then, as $\|G\| \to 0$,*

$$\left|\boldsymbol{\Sigma}_G^{-1}[E(\mathbf{X}_G|y) - E(\mathbf{X}_G)] - \mathbf{P}_{\boldsymbol{\beta}_G}\boldsymbol{\Sigma}_G^{-1}[E(\mathbf{X}_G|y) - E(\mathbf{X}_G)]\right|_{\mathcal{F}} = O(\|G\|^3), \quad (3.2)$$

where $|A|_{\mathcal{F}}$ denotes the Frobenius norm of a matrix A.

The proof of Theorem 3 is provided in the Appendix.

The intuition behind this further reduction in the bias is that the leading term of an integral of a centered cubic function over a spherical region is zero. From this theorem, we see that the bias of the local SIR is four orders of magnitude smaller than that of the corresponding global estimate. This bias is surprisingly small, especially if we compare it with the population bias of the kernel estimator of a density. Let K be a symmetric kernel density, and ϕ be a density to be estimated, with ρ being the bandwidth. Then, it is known that

$$\int \frac{1}{\rho^p} K\left(\frac{\mathbf{x}-\mathbf{a}}{\rho}\right) \phi(\mathbf{x})\, d\mathbf{x} = \phi(\mathbf{a}) + O(\rho^2).$$

Here, ρ corresponds roughly to $\|G\|$ in our problem. If we use asymmetric K, then the error is $O(\rho)$. A similar bias applies to the kernel regression setting. This comparison indicates that the bias of a localized dimension reduction is smaller than those of a kernel density estimation and kernel regression. In other words, even in a fully nonparametric setting in which no elliptical distribution assumption is imposed on $\mathbf{X}$, it is still beneficial to perform a dimension reduction before conducting a nonparametric regression.

Now, let us consider the special case where

$$h(y|\mathbf{x}) = h_1[y, \phi(\mathbf{x})], \tag{3.3}$$

with some function ϕ from $\mathbb{R}^p$ to $\mathbb{R}$. For example, the location model $Y = \phi(\mathbf{X})+\varepsilon$ and the scale model $Y = \phi(\mathbf{X})\varepsilon$ belong to this category. Then,

$$\dot{h}(y|\boldsymbol{\mu}_G) = \frac{\partial h_1[y, \phi(\boldsymbol{\mu}_G)]}{\partial \phi}\, \dot{\phi}(\boldsymbol{\mu}_G).$$

Note that

$$\mathbf{H}_G = E\left\{\frac{\partial h_1[Y_G, \phi(\boldsymbol{\mu}_G)]}{\partial \phi}\right\}^2 \dot{\phi}(\boldsymbol{\mu}_G)\dot{\phi}^T(\boldsymbol{\mu}_G).$$

This is a matrix of rank one unless $\dot{\phi}(\boldsymbol{\mu}_G) = \mathbf{0}$. We summarize this result in the following proposition.

Proposition 2. *Suppose $h(y|\mathbf{x})$ is of the form given in* (3.3) *where h_1 is differentiable with respect to ϕ, and ϕ is differentiable with respect to $\mathbf{x}$. Moreover, suppose $\partial h_1(Y_G, \phi)/\partial \phi$ is square integrable. Then,* $\mathrm{span}(\boldsymbol{\beta}_G)$ *has dimension at most one. That is, ignoring an error of magnitude $O(\|G\|^2)$, the local central subspace $\mathcal{S}_{Y_G|\mathbf{X}_G}$ has dimension at most one.*

This proposition suggests that if we are interested in finding the central subspace, then we need only to estimate one direction for each local region.

That is, it is sufficient to discretize Y_G into binary variables for each G, which is important, because there are fewer observations in a local region.

4. Estimation

In this section, we introduce an estimation procedure for ADR that uses a k-nearest neighbor (kNN) localizing mechanism, and a partial inverse regression as the local dimension-reduction estimator. The properties of nearest neighbor estimators have been studied extensively in the nonparametric regression and pattern recognition literature; see, for example, Hastie, Tibshirani and Friedman (2001).

One of the main problems we need to solve when designing an estimation procedure is how to handle the inversion of $\hat{\mathbf{\Sigma}}_G$, the sample estimate of the local covariance matrix of predictor $\mathbf{X}$. This is especially important in the context of a localized dimension reduction, because the relevant sample size is the number of observations within each neighborhood, which is much smaller than the total sample size n required for a global dimension-reduction estimator, such as the SIR. We solve this problem using the partial inverse regression scheme proposed by Li, Cook and Tsai (2007) and Cook, Li and Chiaromonte (2007).

We first describe the estimation procedure at the population level. By Proposition 2, under condition (3.3), each local central subspace contains at most one direction if we ignore an error of size $\|G\|^2$. This motivates us to employ the following two-slice scheme for the inverse regression. Divide the support of Y_G (which, under assumption (2.2), is the same as Ω_Y) into two intervals, J_{G1} and J_{G2}, and let Δ_G be a Bernoulli random variable that takes the value one if $Y \in J_{G1}$ and two if $Y \in J_{G2}$. From the discussion in Section 3, we have, approximately,

$$\text{span}\{\text{Var}[E(\mathbf{X}_G|\Delta_G)]\} \subseteq \mathbf{\Sigma}_G \mathcal{S}_{Y_G|\mathbf{X}_G}. \tag{4.1}$$

Let $\pi_G = P(\Delta_G = 1)$ and $\boldsymbol{\zeta}_{Gu} = E(\mathbf{X}_G|\Delta_G = u) - E(\mathbf{X}_G)$, for $u = 1, 2$. Noting the relation $\pi_G \boldsymbol{\zeta}_{G1} + (1 - \pi_G)\boldsymbol{\zeta}_{G2} = \mathbf{0}$, we can rewrite the conditional variance in (4.1) as

$$\text{Var}[E(\mathbf{X}_G|\Delta_G)] = \pi_G \boldsymbol{\zeta}_{G1}\boldsymbol{\zeta}_{G1}^T + (1 - \pi_G)\boldsymbol{\zeta}_{G2}\boldsymbol{\zeta}_{G2}^T = \frac{\pi_G}{1 - \pi_G}\boldsymbol{\zeta}_{G1}\boldsymbol{\zeta}_{G1}^T.$$

This is a matrix of rank at most one.

An obvious way to recover the local central subspace $\mathcal{S}_{Y_G|\mathbf{X}_G}$ is to use $\mathbf{\Sigma}_G^{-1}\boldsymbol{\zeta}_G$. However, because k may be close to or even smaller than p, a direct sample estimate of the full inverse of $\mathbf{\Sigma}_G$ is either unstable or nonexistent. To avoid this difficulty, let

$$\mathbf{R}_{G} = (\boldsymbol{\zeta}_{G}, \boldsymbol{\Sigma}_{G}\boldsymbol{\zeta}_{G}, \ldots, \boldsymbol{\Sigma}_{G}^{q-1}\boldsymbol{\zeta}_{G}), \quad \boldsymbol{\eta}_{G} = \mathbf{R}_{G}\left(\mathbf{R}_{G}{}^{T}\boldsymbol{\Sigma}_{G}\mathbf{R}_{G}\right)^{-1}\mathbf{R}_{G}{}^{T}\boldsymbol{\zeta}_{G},$$

where $1 \leq q < p$. Note that $\boldsymbol{\eta}_{G}$ is simply the projection of $\boldsymbol{\Sigma}_{G}^{-1}\boldsymbol{\zeta}$ onto the column space of $\mathbf{R}_{G}$. Cook, Li and Chiaromonte (2007) show that the subspace span$(\mathbf{R}_{G})$ is strictly increasing when q increases, arguing that it often grows sufficiently large to contain the central subspace (in our context, $\mathcal{S}_{Y_G|\mathbf{X}_G}$), for reasonably small q. It is easy to see that when this occurs, $\boldsymbol{\eta}_{G}$ becomes a member of $\mathcal{S}_{Y_G|\mathbf{X}_G}$. Thus, we use $\boldsymbol{\eta}_{G}$ in place of $\boldsymbol{\Sigma}_{G}^{-1}\boldsymbol{\zeta}_{G}$ as the local dimension-reduction estimate.

To combine directions from each neighborhood, let $t : [0, \infty) \to [0, \infty)$ be a nondecreasing function, and

$$\omega_{G} = \frac{\pi_{G}}{1 - \pi_{G}}\boldsymbol{\zeta}_{G1}^{T}\boldsymbol{\zeta}_{G1}.$$

Define the matrix

$$\mathbf{V} = \sum t(\omega_{G})\boldsymbol{\eta}_{G}\boldsymbol{\eta}_{G}^{T},$$

where the summation is a collection of neighborhoods, and t is a weighting function, the meaning and choice of which are described in the next section.

We now summarize the sample-level algorithm for ADR. Let $\{(\mathbf{X}_i, Y_i)$, for $i = 1, \ldots, n\}$, be a sample from $(\mathbf{X}, Y)$. The algorithm assumes that the structural dimension d is known; the estimation of d is discussed in the next section.

1. For each $s = 1, \ldots, n$, let G_s be the set that includes the k nearest $\mathbf{X}_j$ to $\mathbf{X}_s$ in terms of the Euclidean distance $\|\mathbf{X}_j - \mathbf{X}_s\|$. Note that G_s contains $k + 1$ elements because we do not count $\mathbf{X}_s$ among these k points.

2. Divide the set $\{Y_j : \mathbf{X}_j \in G_s\}$ into two intervals, J_{s1} and J_{s2}, each containing roughly the same number of Y_j. Let n_{su}, for $u = 1, 2$, be the cardinality of the set $\{j : \mathbf{X}_j \in G_s, Y_j \in J_{su}\}$ and $n_s = n_{s1} + n_{s2}$. Let

$$\bar{\mathbf{X}}_{G_s1} = \frac{1}{n_{s1}}\sum \mathbf{X}_j I(\mathbf{X}_j \in G_s, Y_j \in J_{s1}), \quad \bar{\mathbf{X}}_{G_s} = \frac{1}{n_s}\sum \mathbf{X}_j I(\mathbf{X}_j \in G_s),$$

and

$$\hat{\boldsymbol{\zeta}}_{G_s} = (\bar{\mathbf{X}}_{G_s1} - \bar{\mathbf{X}}_{G_s}), \quad \hat{\omega}_{G_s} = \left(\frac{n_{s1}}{n_{s2}}\right)\|\bar{\mathbf{X}}_{G_s1} - \bar{\mathbf{X}}_{G_s}\|^2.$$

3. Compute

$$\hat{\mathbb{R}}_{G_s} = \left(\hat{\boldsymbol{\zeta}}_s, \hat{\boldsymbol{\Sigma}}_{G_s}\hat{\boldsymbol{\zeta}}_{G_s}, \ldots, \hat{\boldsymbol{\Sigma}}_{G_s}^{q-1}\hat{\boldsymbol{\zeta}}_{G_s}\right) \text{ and } \hat{\boldsymbol{\eta}}_{G_s} = \hat{\mathbb{R}}_{G_s}(\hat{\mathbb{R}}_{G_s}^{T}\hat{\boldsymbol{\Sigma}}_{G_s}\hat{\mathbb{R}}_{G_s})^{-1}\hat{\mathbb{R}}_{G_s}^{T}\hat{\boldsymbol{\zeta}}_{G_s}.$$

4. Use the first d eigenvectors of the matrix $\hat{\mathbf{V}} = \sum_{s=1}^{m} t(\hat{\omega}_{G_s})\hat{\boldsymbol{\eta}}_{G_s}\hat{\boldsymbol{\eta}}_{G_s}^{T}$ as the estimate of a basis for the global central subspace $\mathcal{S}_{Y|\mathbf{X}}$.

It is well known that a severely biased estimate can be introduced as a result of the above choice of k-nearest neighborhood in a high-dimensional input space with finite samples. Because the Euclidean distance measure implies that the input features are homogeneous or isotropic, an immediate remedy would be to use a locally adaptive metric. Inspired by the work of Hastie and Tibshirani (1996), we propose a refined estimation in which the neighborhoods are elongated along less relevant directions, and constricted along more influential directions. After obtaining a basis for the global central subspace $\mathcal{S}_{Y|X}$ (say, $\hat{\mathbf{B}}_0$) from the above-mentioned algorithm, instead of a p-dimensional ball as the k-nearest neighborhood, we use a p-dimensional ellipsoid to shrink the neighborhoods in directions orthogonal to $\hat{\mathbf{B}}_0$ and to elongate those parallel to this initial estimate. More specifically, the distance between $\mathbf{X}_j$ and $\mathbf{X}_s$ in step 1 of the above algorithm is replaced by

$$\begin{aligned} d_{js}^2 &= \|\hat{\mathbf{B}}_{(0)}^T(\mathbf{X}_j - \mathbf{X}_s)\|^2 + \kappa_{(0)}\|(\mathbf{X}_j - \mathbf{X}_s)\|^2 \\ &= (\mathbf{X}_j - \mathbf{X}_s)^T[\hat{\mathbf{B}}_{(0)}\hat{\mathbf{B}}_{(0)}^T + \kappa_{(0)}\mathbf{I}_p](\mathbf{X}_j - \mathbf{X}_s), \end{aligned} \tag{4.2}$$

where $\kappa_{(0)}$ is a small "softening" parameter used to control the shrinkage and elongation along different directions. An iterative estimation can be implemented until a certain convergence criterion is met.

Our method differs from that of Hsing (1999), who applies a k-nearest neighborhood to multivariate Y to avoid slicing. It is also different from the IMAVE procedure of Xia et al. (2002), in that the latter requires a linearity condition.

5. Tuning Parameters

In this section, we discuss how to choose the various tuning parameters for the estimation algorithm described in Section 4. As such, we estimate the structural dimension d, and choose the weighting function t, the order q for the partial inverse regression, and the softening parameter κ for the adaptive nearest neighborhood selection. An appropriate justification of these choices relies on the asymptotic properties of ADR; this is beyond the scope of this study, and thus is left to future research. Inevitably, the following recommendations are heuristic in nature. In extensive numerical experiments, we performed sensitivity analyses on the recommended choices of these tuning parameters, with our results showing reasonably stable estimations.

We recommend two choices for t. A natural choice is $t(\omega_{G}) \equiv 1$. From the discussion in Section 4, $\hat{\boldsymbol{\zeta}}_{G}$ are approximately aligned with the local central subspace. Thus, if a neighborhood is in a region in which there is no significant change in Y, then $\|\hat{\boldsymbol{\zeta}}_G\|$ tends to be small. By setting t equal to one, we let the sliced means themselves determine the relative importance of each neighborhood. A second choice of t is

$$t(\hat{\omega}_G) = \begin{cases} \|\hat{\boldsymbol{\zeta}}_G\|^{-2} & \hat{\omega}_G > c, \\ 0 & \hat{\omega}_G \leq c. \end{cases} \tag{5.1}$$

This weighting function introduces a hard thresholding according to the magnitude of $\|\hat{\boldsymbol{\zeta}}\|$, discarding those neighborhoods with small sliced means. Moreover, when a sliced mean is sufficiently large, its magnitude is no longer included in the estimation. Based on our experience, the second choice seems to work better. We choose the threshold c according to a percentage δ of the sample size. That is, we choose $\delta \times 100\%$ of neighborhoods with the highest $\hat{\omega}_G$. The choice $\delta = 0.5$ works well in our simulation experiments.

To choose q_{G_s}, we use the threshold recommended by Li, Cook and Tsai (2007),

$$q_{G_s} = \sum_{j=1}^{p-1} I\left(\frac{r_j(G_s)}{r_{j+1}(G_s)} > \alpha_0\right),$$

where $r_1(G_s) \geq \cdots \geq r_p(G_s)$ are the eigenvalues of the matrix $\hat{\mathbb{R}}_{G_s}\hat{\mathbb{R}}_{G_s}^T$, and α_0 is taken to be 1.5. Following Hastie and Tibshirani (1996), we choose $\kappa_{(0)} = 1/3$ in our numerical studies.

To estimate the structural dimension d, we adopt the bootstrap procedure proposed in Ye and Weiss (2003) and Zhu and Zeng (2006). Let $\hat{\mathcal{S}}_{d^*}$ be an estimate of $\mathcal{S}_{Y|\mathbf{X}}$ for a fixed d^*. We can get a set of bootstrap-estimated $\{\hat{\mathcal{S}}_{d^*}^{(j)},\ j = 1, \ldots, n_b\}$ by bootstrapping, where n_b is the number of bootstrap samples. The distances between $\hat{\mathcal{S}}_{d^*}$ and its bootstrap version $\{\hat{\mathcal{S}}_{d^*}^{(j)},\ j = 1, \ldots, n_b\}$ can be used to assess the variability of the estimated subspace at $d = d^*$, which, in turn, can be used to infer the structural dimension d. Intuitively, $\hat{\mathcal{S}}_{d^*} \subseteq \mathcal{S}_{Y|\mathbf{X}}$ when $d^* \leq d$. However, when $d^* > d$, $\hat{\mathcal{S}}_{d^*} = \mathcal{S}_{Y|\mathbf{X}} \oplus \tilde{\mathcal{S}}$, where $\tilde{\mathcal{S}}$ is a $(d^* - d)$-dimensional subspace orthogonal to $\mathcal{S}_{Y|\mathbf{X}}$. Because $\tilde{\mathcal{S}}$ can be arbitrary, we expect to see greater variability in $\hat{\mathcal{S}}_{d^*}$, with its bootstrap versions, than when $d^* \leq d$. Therefore, the structural dimension d can be estimated as the largest d^* that produces a stable estimator.

Finally, we set the number of observations in each neighborhood as $2p \leq k \leq 4p$. This choice is reasonable only when p is considerably smaller than n.

6. Simulation Studies

In this section, we evaluate the performance of ADR using simulations. For comparison purposes, several existing methods were also evaluated in the simulation studies, including the SIR, sliced average variance estimation (SAVE), principal Hessian directions (PHD), minimum average variance estimation (MAVE), and sliced regression (SR). The vector correlation coefficient q (Hotelling (1936); Ye and Weiss (2003)) was used to measure the estimation accuracy. Let $\mathbf{B}$ be an orthonormal basis of the central subspace, and $\hat{\mathbf{B}}$ be an estimate of the orthonormal basis. Then, the vector correlation coefficient

$$q = \sqrt{||\hat{\mathbf{B}}^T(\mathbf{B}\mathbf{B}^T)\hat{\mathbf{B}}||} = \sqrt{\prod_{i=1}^{d} \rho_i^2},$$

where $0 \le \rho_d \le \cdots \le \rho_1 \le 1$ are the eigenvalues of the matrix $\hat{\mathbf{B}}^T(\mathbf{B}\mathbf{B}^T)\hat{\mathbf{B}}$. As q increases, $\mathcal{S}(\hat{\mathbf{B}})$ becomes closer to $\mathcal{S}(\mathbf{B})$. We chose the Gaussian kernel and its corresponding optimal bandwidth for the MAVE and SR. A rule-of-thumb choice of $k = 4p$ was used for our proposed aggregate approach, including the kNN sliced inverse regression (kNNSIR) and the adaptive kNN sliced inverse regression (a-kNNSIR, where the adaptive distance given in (4.2) is used). Note that more refined ways of choosing k, such as cross-validation, can be used, but at greater computational expense. For each parameter setting, 200 simulation replications were conducted.

The following four models were used in the numerical study:

Model 1: $Y = \exp\{(\beta^T X)^2 + \epsilon\}$,

Model 2: $Y = \cos(2\beta_1^T X) - \cos(\beta_2^T X) + 0.2\epsilon$,

Model 3: $Y = \text{sign}(\beta_1^T X + \epsilon_1)\log(|\beta_2^T X + 3 + \epsilon_2|)$,

Model 4: $Y = (\beta_1^T X)(\beta_2^T X + 2) + (\beta_3^T X + 2)^3 + 0.5\epsilon$.

All of these models have been studied extensively in the literature on sufficient dimension reduction. In all four models, $X \sim N_p(0,\ \Sigma)$, independent of standard Gaussian noises ϵ, ϵ_1, and ϵ_2. The covariance matrix $\Sigma = (\sigma_{ij}) = (\rho^{|i-j|})$, where $\rho = 0.5$ in Models 1-3 and $\rho = 0$ in Model 4. In Model 1, $\beta = (1, 0.5, 1, 0, \ldots, 0)^T$. In Model 2, $\beta_1 = (1, 0, \ldots, 0)^T$ and $\beta_2 = (0, 1, 0, \ldots, 0)^T$. In Model 3, $\beta_1 = (1, 1, 1, 1, 0, \ldots, 0)^T$, $\beta_2 = (0, \ldots, 0, 1, 1, 1, 1)^T$, and the function sign$(\cdot)$ takes the value 1 or -1, depending on the sign of the argument. In Model 4, $\beta_1 = (1, 0, \ldots, 0)^T$, $\beta_2 = (0, 1, 1, 0, \ldots, 0)^T$, and $\beta_3 = (0, 0, 0, 1, 1, 0, \ldots, 0)^T$.

In Figures 2 – 5, we compare the performance of the aforementioned methods. The results are as follows. First, the proposed aggregate SDR, adaptive

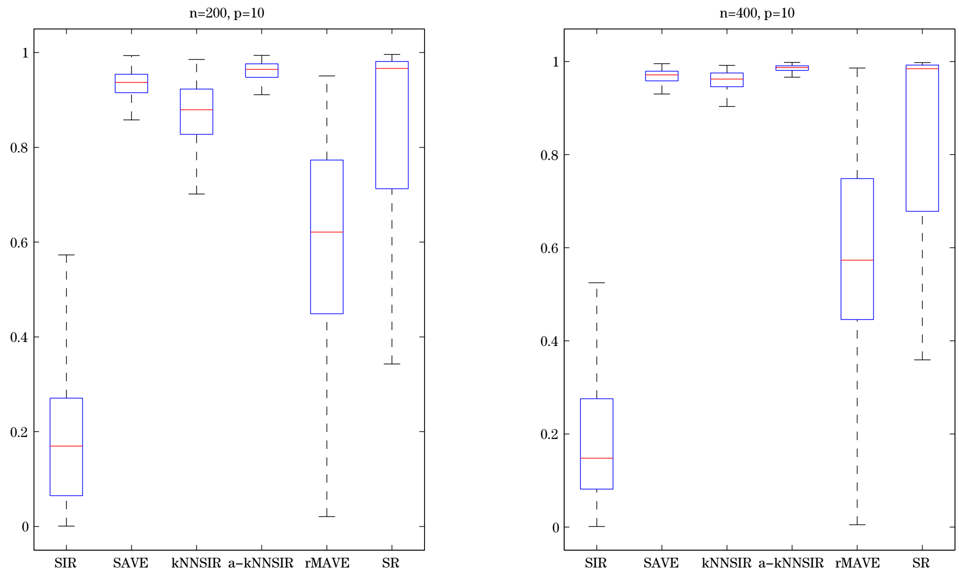

Figure 2. Comparison of estimation accuracy with Model 1.

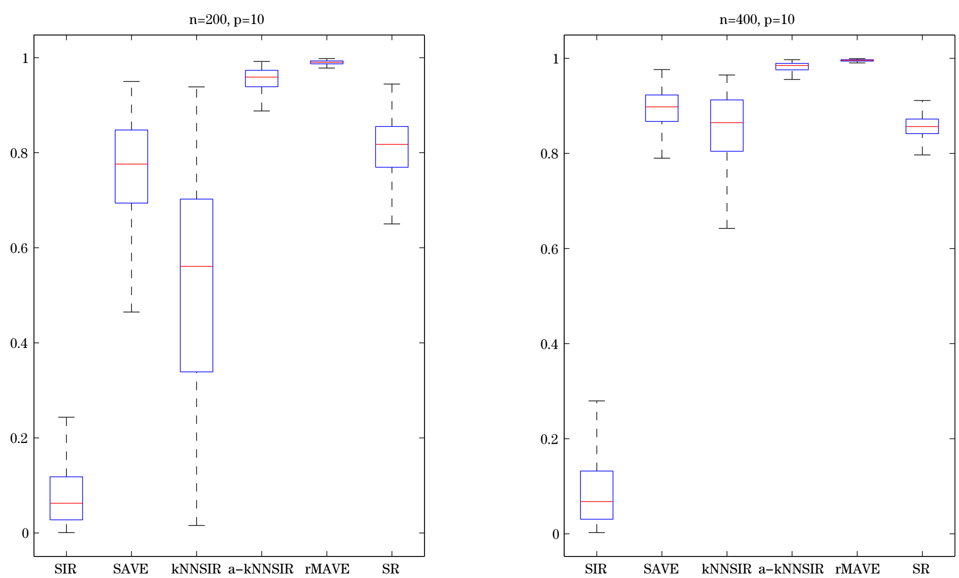

Figure 3. Comparison of estimation accuracy with Model 2.

kNN-SIR, significantly improves the performance of the original inverse regression methods, and is broadly comparable with the forward regression approaches (MAVE and SR). Second, through localization, the adaptive kNN-SIR overcomes

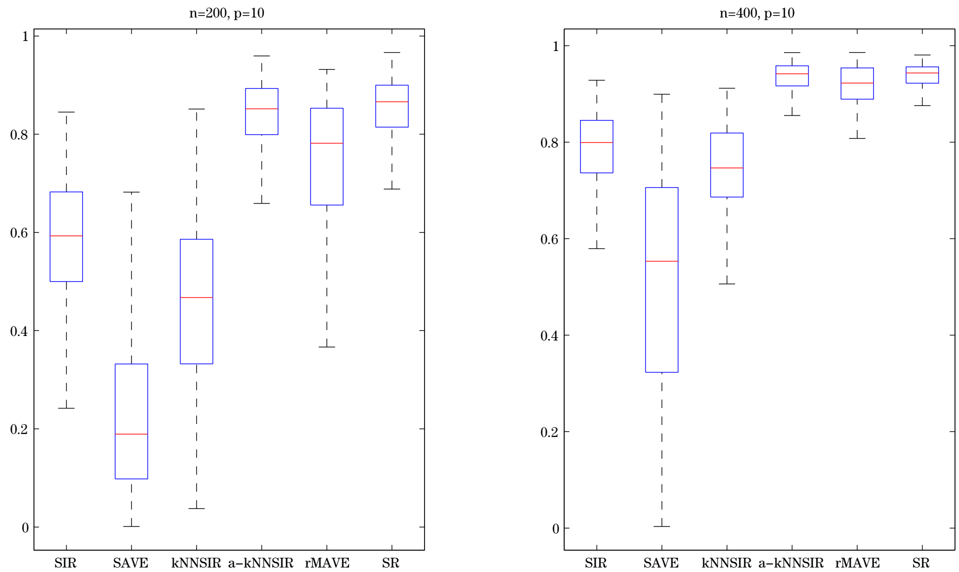

Figure 4. Comparison of estimation accuracy with Model 3.

the drawback of missing symmetric patterns in the original SIR, as shown in Models 1 and 2. Third, when $\mathcal{S}_{Y|\mathbf{X}}$ is completely contained in the mean regression function $E(Y|\mathbf{X})$, the MAVE stands out as the best method, which is expected, with the proposed a-kNNSIR a close second, shown in Models 2 and 4. However, when $\mathcal{S}_{Y|\mathbf{X}}$ spans beyond the mean function, as in Models 1 and 3, the a-kNNSIR clearly outperforms the MAVE. Finally, larger sample sizes are needed to provide a good estimation with an increase of the dimension d. Zhu, Miao and Peng (2006) studied Model 4 ($d = 3$), showing that n needs to be increased to 3,200 in order for the estimation accuracy of the SIR to be acceptable when $p \leq 20$. In our numerical study, the proposed a-kNNSIR and the MAVE are the only two methods that show good performance for moderate sample sizes. It is well known that the computational burden increases significantly with n and p for the forward regression methods (MAVE and SR). In contrast, the proposed aggregate inverse regression approach is more computationally efficient, because no numerical optimization is required. This is confirmed by the results of our simulation studies.

Next, we estimated the structural dimension d using the adopted bootstrap procedure. In all numerical studies, we used $1 - q$ as the distance measure to assess the variability between $\hat{\mathcal{S}}_{d^*}$ and its bootstrap versions. For each $d^* = 1, 2, \ldots, p-1$, 500 bootstrap samples were drawn, and the median of the distances

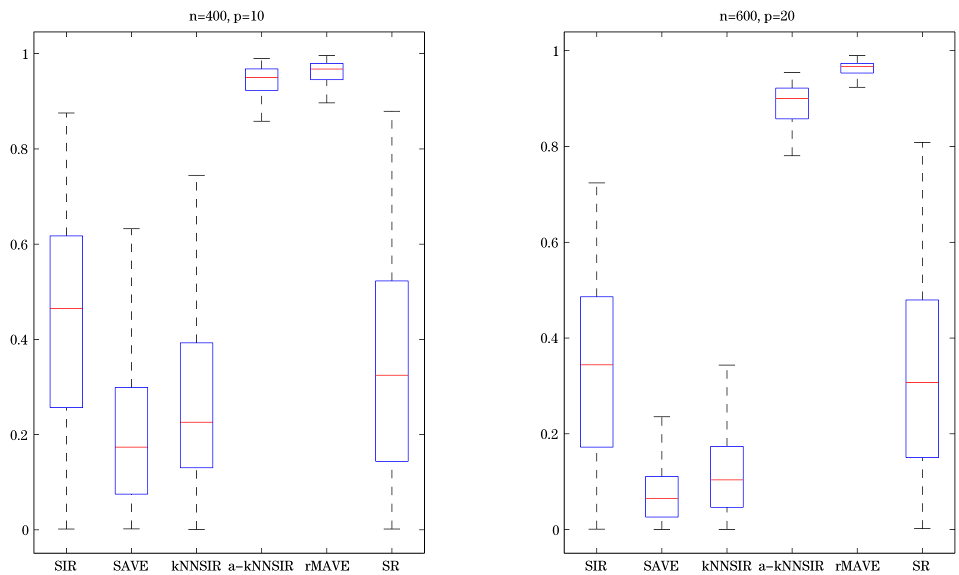

Figure 5. Comparison of estimation accuracy with Model 4.

between $\hat{\mathcal{S}}_{d^*}$ and its bootstrap versions $\{\hat{\mathcal{S}}_{d^*}^{(j)},\ j = 1, \ldots, 500\}$ was calculated. Figure 6 shows the dimension variability plots (Zhu and Zeng (2006)) for Models 1-4. As expected, large variability is evident when $d^* > d$. Of of 100 samples with $n = 400$ and $p = 10$, the accuracy of correctly estimating d is 99%, 94%, 99%, and 84% for Models 1-4, respectively.

7. Real-data Analyses

7.1. Ozone data

In this section, we investigate the performance of the proposed aggregate SIR when it is applied to real data on the relations between ozone levels and various environmental variables Breiman and Friedman (1985). The data contain 330 observations, with each observation consisting of nine variables: ozone concentration, height, inversion height, temperature, inversion temperature, humidity, pressure, visibility, and wind speed. Here, ozone concentration is treated as the response, and the other eight variables are treated as predictors. For ease of interpretation, all predictors are standardized separately. This data set has been analyzed by several authors. See, for example, Li (1992) and Cook and Li (2004).

The SIR identifies one significant direction. After a closer investigation of

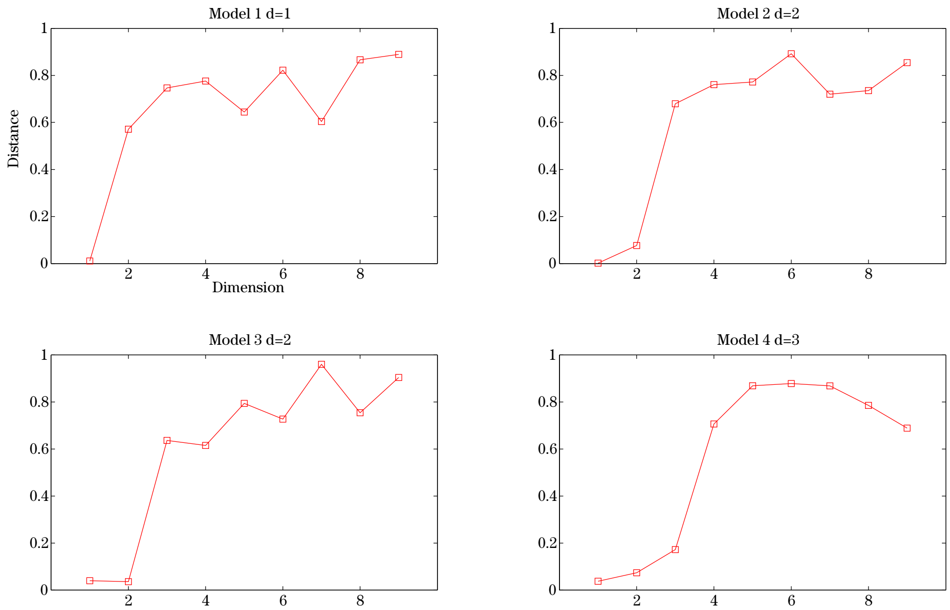

Figure 6. Bootstrap estimation of dimension ($n = 400$ and $p = 10$).

the residual from the quadratic fit, Li (1992) argued that a second significant component is necessary, and that the PHD can recover this direction. Cook and Li (2004) also identified the first direction using an inverse Hessian transformation (IHT). However, their estimate of the dimension d differs from that of others, leaving some uncertainty.

In our application, the dimension variability plot, shown in Figure 7 (a), suggests $\hat{d} = 2$. Figure 7 (b)(c) shows the pattern identified by our method. Interestingly, our proposed a-kNNSIR successfully recovers the two significant components identified by the SIR and PHD, without fitting a detailed model, as in Li (1992), and without the uncertainty associated with estimating d evident in Cook and Li (2004).

7.2. College admission data

This data set was used in the 1995 Data Analysis Exposition, sponsored by the American Statistical Association. It is also included in the textbook, "An introduction to statistical learning with applications in R" (James et al. (2013)), and the associated R package ISLR. We are interested in predicting the number of applications received (y) by 557 private institutions that have a full-time undergraduate student body of less than 10,000. The predictors used in

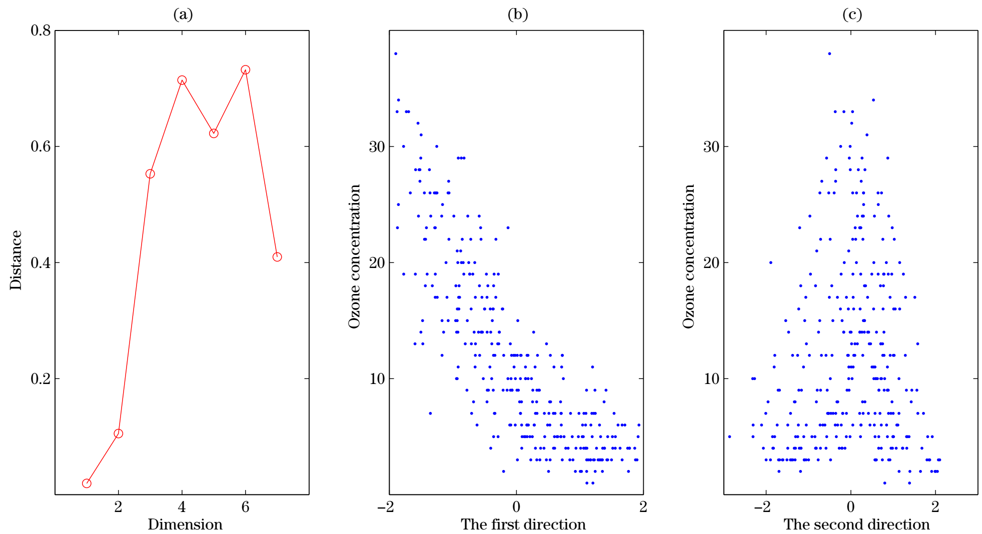

Figure 7. Analysis of ozone data: (a) dimension variability plot, (b–c) scatterplots of response vs. the two estimated directions.

Table 1. The predictors and the estimated directions for the college admission data.

	Predictor	$\hat{\beta}_1$	$\hat{\beta}_2$
x_1	number of full time undergraduates	0.91	0.06
x_2	number of part time undergraduates	0.00	−0.38
x_3	out-of-state tuition	0.34	−0.25
x_4	room and board costs	0.06	−0.21
x_5	estimated book costs	−0.04	−0.03
x_6	estimated personal spending	−0.12	−0.30
x_7	percent of faculty with terminal degree	0.03	−0.03
x_8	student/faculty ratio	0.13	0.46
x_9	percent of alumni who donate	0.04	0.07
x_{10}	instructional expenditure per student	0.12	−0.26
x_{11}	graduation rate	0.04	−0.60

our analysis are listed in Table 1. Again, for ease of interpretation, all predictors were standardized separately.

The dimension variability plot in Figure 8 (a) suggests at most three dimensions. It also indicates that the prediction ability for the second and third directions may not be very strong, because their variability is much larger than that of the first direction. Situations such as this can often happen in practice, because real data may include significant noise and weak signals, which makes determining the structural dimension less obvious. Nevertheless, we further con-

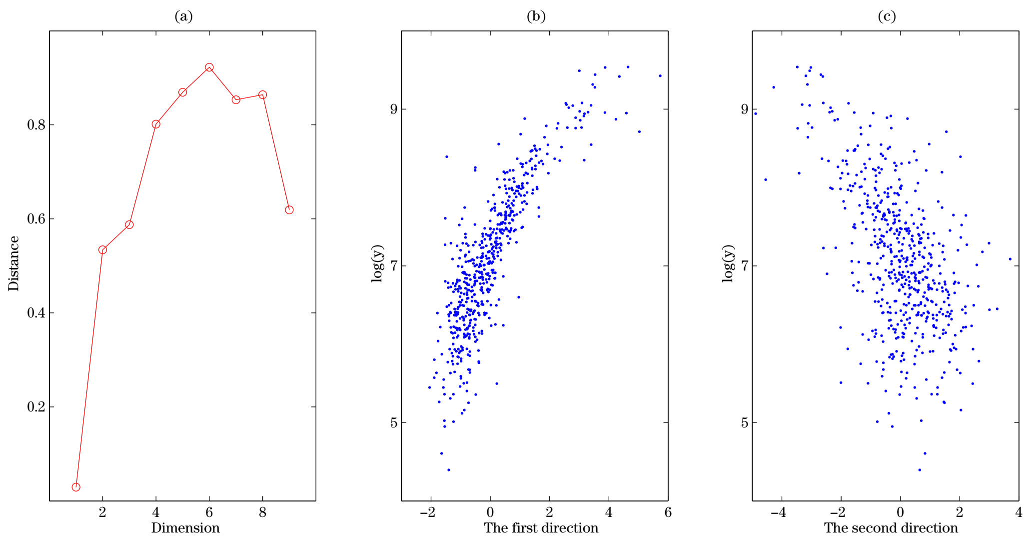

Figure 8. Analysis on College admission data: (a) dimension variability plot, (b-c) scatterplots of response vs. the two estimated directions.

sider the coefficients and marginal plots for the first three directions. Finally, we retained the first two directions, because no meaningful interpretation was available for the third direction. We also applied the SIR to this data set, with the asymptotic test also suggesting $d = 3$. The first direction is dominated by x_1, the number of full-time undergraduates, but the second and the third directions are not that clear. From the estimated directions $\hat{\beta}_1$ and $\hat{\beta}_2$ in Table 1 using our method, we can interpret the first direction as a "size" factor, because it is dominated by x_1. The second direction can be seen as an "academic quality" factor, which includes x_8 (student/faculty ratio), x_{10} (instructional expenditure per student), and x_{11} (the graduation rate). In Figure 8 (b), in general, the number of applications increases with the size of the institution's student body, with this increasing trend tapering off toward the end. Figure 8 (c) shows that more students apply to institutions with higher academic quality, meaning high graduation rate, high instructional expenditure, and a small student/faculty ratio.

8. Discussion

We have proposed an aggregate approach for estimating the central subspace, which we illustrated using an adaptive kNN sliced inverse regression. We believe that a class of new local-dimension reduction methods can be developed under this localization framework. Our new method does not seek to replace the original SIR. Instead, we have developed an alternative approach so that the simplicity of the SIR can be extended further.

There are still several open questions that need further study, including those related to the asymptotic properties of the proposed estimators and an extension to a big data setting. To study these asymptotic properties, the most related work, in the global sense, is the study of Hsing and Carroll (1992), who show that the estimator from the two-slice approach is root-n consistent. However, owing to the use of a local approximation, our local inverse conditional covariance matrix does not have the closed form of equation (1.2) in Hsing and Carroll (1992). Because the k-nearest-neighbor estimation can be treated as a special kernel method, our proposed localization-aggregation approach is similar, in spirit, to the kernel-based outer product of gradients (OPG) estimation (Xia et al. (2002)). Overcoming these challenges and difficulties is left to future research. A referee brought our attention to extending the method to a big data setting, with large n and/or large p. When the volume n is huge, the dimension p is moderate, and $n > p$, we propose implementing the *localization-aggregation* approach together with "leveraging based subsampling" (Ma, Mahoney and Yu (2015)). The case, where $n < p$, or even $n \ll p$, is clearly more challenging. We adopt the *sequential dimension-reduction* paradigm proposed by Yin and Hilafu (2015) to sidestep the curse of dimensionality. Such an investigation is currently under way by our team, and our preliminary results are very promising.

Supplementary Material

The online Supplementary Material provides the proofs of Theorems 1–3 in the paper.

Acknowledgments

We would like to thank Co-Editor Professor Zhiliang Ying, the Associate Editor, and two anonymous referees for their careful reading and constructive comments.

References

Bellman, R. (1961). *Adaptive Control Processes*. Princeton University Press, Princeton, NJ.

Breiman, L. and Friedman, J. H. (1985). Estimating optimal transformations for multiple regression and correlation. (with discussion). *Journal of the American Statistical Association*

80, 580–598.

Bura, E. (2003). Using linear smoothers to assess the structural dimension of regressions. *Statistica Sinica* **13**, 143–162.

Bura, E. and Cook, R. D. (2001a). Estimating the structural dimension of regressions via parametric inverse regression. *Journal of the Royal Statistical Society Series B (Statistical Methodology)* **63**, 393–410.

Bura, E. and Cook, R. D. (2001b). Extending sliced inverse regression: the weighted chi-squared test. *Journal of the American Statistical Association* **96**, 996–1003.

Bura, E. and Pfeiffer, R. (2003). Graphical methods for class prediction using dimension reduction techniques on dna microarray data. *Bioinformatics* **19**, 1252–1258.

Chen, C. and Li, K. C. (1998). Can sir be as popular as multiple linear regression? *Statistica Sinica* **8**, 289–316.

Chiaromonte, F. and Martinelli, J. A. (2002). Dimension reduction strategies for analyzing global gene expression data with a response. *Mathematical Biosciences* **176**, 123–144.

Cook, R. D. (1998). *Regression Graphics: Ideas for Studying Regressions Through Graphics*. Wiley, New York.

Cook, R. D. and Critchley, F. (2000). Identifying regression outliers and mixtures graphically. *Journal of the American Statistical Association* **95**, 781–794.

Cook, R. D. and Li, B. (2004). Determining the dimension of iterative hessian transformation. *The Annals of Statistics* **32**, 2501–2531.

Cook, R. D., Li, B. and Chiaromonte, F. (2007). Dimension reduction without matrix inversion. *Biometrika* **94**, 569–584.

Cook, R. D. and Ni, L. (2005). Sufficient dimension reduction via inverse regression. *Journal of the American Statistical Association* **100**, 410–428.

Cook, R. D. and Weisberg, S. (1991). Sliced inverse regression for dimension reduction: Comment. *Journal of the American Statistical Association* **86**, 328–332.

Cook, R. D. and Yin, X. (2001). Dimension reduction and visualization in discriminant analysis (with discussion). *Australian & New Zealand Journal of Statistics* **43**, 147–199.

Eaton, M. L. (1986). A characterization of spherical distributions. *Journal of Multivariate Analysis* **20**, 272–276.

Fung, W. K., He, X., Liu, L. and Shi, P. (2002). Dimension reduction based on canonical correlation. *Statistica Sinica* **12**, 1093–1113.

Hastie, T. and Tibshirani, R. (1996). Discriminant adaptive nearest neighbor classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **18**, 607–616.

Hastie, T., Tibshirani, R. and Friedman, J. H. (2001). *The Elements of Statistical Learning : Data Mining, Inference, and Prediction*. Springer series in statistics. Springer, New York.

Hotelling, H. (1936). Relations between two sets of variates. *Biometrika* **28**, 321–377.

Hsing, T. (1999). Nearest neighbor inverse regression. *The Annals of Statistics* **27**, 697–731.

Hsing, T. and Carroll, R. J. (1992). An asymptotic theory of sliced inverse regression. *The Annals of Statistics* **20**, 1040–1061.

James, G., Witten, D., Hastie, T. and Tibshirani, R. (2013). *An Introduction to Statistical Learning with Applications in R*. Springer texts in statistics. Springer-Verlag, New York.

Li, B. and Wang, S. (2007). On directional regression for dimension reduction. *Journal of the American Statistical Association* **102**, 997–1008.

Li, B., Zha, H. and Chiaromonte, F. (2005). Contour regression: a general approach to dimension reduction. *The Annals of Statistics* **33**, 1580–1616.

Li, K. C. (1991). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association* **86**, 316–327.

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, L., Cook, R. D. and Nachtsheim, C. J. (2004). Cluster-based estimation for sufficient dimension reduction. *Computational Statistics and Data Analysis* **47**, 175–193.

Li, L., Cook, R. D. and Tsai, C. L. (2007). Partial inverse regression. *Biometrika* **94**, 615–625.

Li, L. and Li, H. (2004). Dimension reduction methods for microarrays with application to censored survival data. *Bioinformatics* **20**, 3406–3412.

Ling, Y., Bhandarkar, S., Yin, X. and Lu, Q. (2005). Saveface and sirface: Appearance-based recognition of faces and facial expressions. *Proceeding of the IEEE International Conference on Image Processing ICIP*, 466–469.

Ling, Y., Yin, X. and Bhandarkar, S. (2003). Sirface vs fisherface: recognition using class specific linear projection. *Proceeding of the IEEE international conference on Image Processing ICIP*, 885–888.

Ma, P., Mahoney, M. W. and Yu, B. (2015). A statistical perspective on algorithmic leveraging. *Journal of Machine Learning Research* **16**, 861–911.

Naik, P. and Tsai, C. L. (2000). Partial least squares estimator for single-index models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **62**, 763–771.

Schott, J. R. (1994). Determining the dimensionality in sliced inverse regression. *Journal of the American Statistical Association* **89**, 141–148.

Velilla, S. (1998). Assessing the number of linear components in a general regression problem. *Journal of the American Statistical Association* **93**, 1088–1098.

Wang, Q. and Yin, X. (2011). Estimation of inverse mean: An orthogonal series approach. *Computational Statistics and Data Analysis* **55**, 1656–1664.

Wu, Q., Liang, F. and Mukherjee, S. (2010). Localized sliced inverse regression. *Journal of Computational and Graphical Statistics* **19**, 843–860.

Xia, Y., Tong, H., Li, W. and Zhu, L. (2002). An adaptive estimation of dimension reduction space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **64**, 363–410.

Ye, Z. and Weiss, R. E. (2003). Using the bootstrap to select one of a new class of dimension reduction methods. *Journal of the American Statistical Association* **98**, 968–979.

Yin, X. and Cook, R. D. (2003). Estimating central subspace via inverse third moment. *Biometrika* **90**, 113–125.

Yin, X. and Hilafu, H. (2015). Sequential sufficient dimension reduction for large p, small n problems. *Journal of Royal Statistical Society, Series B (Statistical Methodology)* **77**, 879–892.

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.

Zhu, L. X. and Fang, K. T. (1996). Asymptotics for kernel estimate of sliced inverse regression. *The Annals of Statistics* **24**, 1053–1068.

Zhu, L. X., Miao, B. Q. and Peng, H. (2006). On sliced inverse regression with high-dimensional covariates. *Journal of the American Statistical Association* **101**, 630–643.

Zhu, L. X. and Ng, K. W. (1995). Asymptotics of sliced inverse regression. *Statistica Sinica* **5**, 727–736.

Zhu, Y. and Zeng, P. (2006). Fourier methods for estimating the central subspace and the central mean subspace in regression. *Journal of the American Statistical Association* **101**, 1638–1651.

Department of Information Systems, Statistics, and Management Science, The University of Alabama, Tuscaloosa, AL 35487, USA.

E-mail: qwang57@ua.edu

Department of Statistics, The University of Kentucky, Lexington, KY 40506, USA.

E-mail: yinxiangrong@uky.edu

Department of Statistics, Penn State University, State College, PA 16801, USA.

E-mail: bing@stat.psu.edu

Pharmaceutical Product Development, LLC, Wilmington,NC, USA.

E-mail: lilylilytang78@gmail.com

(Received April 2016; accepted July 2018)