Statistica Sinica Preprint N[illegible]S-2016-0188	
Title	On aggregate dimen[illegible]on reduction
Manuscript ID	SS-2016-0188
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202016.0188
Complete List of Authors	Qin Wang Xiangrong Yin Bing Li and Zhihui T[illegible]
Corresponding Author	Qin W[illegible]
E-mail	qw[illegible]3@[illegible]
Notice: Accepted version subje[illegible]Eng[illegible]h editing.	

Statistica Sinica

On Aggregate Dimension Reduction

Qin Wang, Xiangrong Yin, Bing Li and Zhihui Tang

Virginia Commonwealth University, University of Kentucky

Penn State University and PP

Abstract: We propose a dimension reduction method based on aggregation of localized estimators. The dual process of localization and aggregation helps 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 data sets with large volume and high dimension. We demonstrate the efficacy of the method via simulation and real data applications.

Key words and phrases: Central Subspace; k-Nearest Neighbor; Sliced Inverse Regression.

1. Introduction

Suppose Y is a univariate response and $\mathbf{X}$ is a p-dimensional vector of continuous predictors. In its full generality, the goal of regression is to infer about the conditional distribution of Y given $\mathbf{X}$. However, because

of the curse of dimensionality (Bellman, 1961), regression with large p can be difficult in practice. The basic idea of *sufficient dimension reduction* (SDR; Li (1991); Cook (1998)) is to replace the predictor vector by its projection on to a low-dimensional subspace without losing information on the conditional distribution of $Y \mid \mathbf{X}$, and without assuming any specific model for $Y \mid \mathbf{X}$.

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 conditional on $\mathbf{P}_{\mathcal{S}}\mathbf{X}$, where $\mathbf{P}_{\mathcal{S}}$ is the projection on to $\mathcal{S}$. 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 et al. (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 denoted by $d_{Y|\mathbf{X}}$.

A main class of estimators of the central subspace is based on inverse conditional moments such as $E(\mathbf{X} \mid Y)$ and $\mathrm{Var}(\mathbf{X} \mid Y)$. This includes sliced inverse regression (SIR; Li, 1991), sliced average variance estimation (SAVE; Cook and Weisberg (1991)), hybrids of the two (Ye and Weiss, 2003), parametric inverse regression (Bura and Cook, 2001a), sliced average third moment (Yin and Cook, 2003), contour regression (Li et al.,

2005), minimum discrepancy approach (Cook and Ni, 2005) and directional regression (Li and Wang, 2007), among others.

Sliced inverse regression is the first general dimension reduction method and has generated intense interests since its introduction. Many extensions and refinements ensued. Hsing and Carroll (1992), Li and Ng (1995) and Zhu and Fang (1996) studied the asymptotic properties of SIR estimates and its variations. Schott (1994), Velilla (1998) and Bura and Cook (2001) introduced asymptotic inference procedures to determine the dimension of the subspace estimated by 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 SIR and maximal correlation. Hsing (1999) used nearest-neighbor method to develop a variation of SIR that is applicable to multivariate responses. Naik and Tsai (2000) compared the performance of SIR with partial least squares in the context of single-index model. Cook and Critchley (2000) showed that dimension reduction methods in general and 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 SIR by replacing inverse conditional mean $E(\mathbf{X} \mid Y)$ with parametric regression or basis expansion. Li et al. (2004) proposed a cluster-

based estimation to mitigate the effect of nonlinearity on the predictors with the focus on single index models. Zhu et al. (2006) studied asymptotic behavior for SIR when the number of covariates increases with sample size. Recently, Wu et al. (2010) developed an extension by replacing the global average with the local average for each data point to alleviate the issue of degenerate solutions. SIR has found wide applications in diverse fields such as computer vision (Ling et al., 2003, 2005), and biological sciences (Chiaromonte and Martinelli, 2002; Bura and Pfeiffer, 2008; Li, Li, 2004).

In this paper we develop an aggregate dimension reduction 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 these local spaces to recover $\mathcal{S}_{Y|\mathbf{X}}$. The dual process of localization and aggregation brings two benefits. First, since any differentiable function is approximately linear locally, we no longer need to impose the linearity assumption on the conditional mean of the predictors, as required by SIR. Second, it leads to 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 will be rigorously formulated and developed

at the population level in Section 3. In sections 4 and 5 we provide the estimation procedures of localized SIR using k-nearest neighborhood, and discuss various issues involved in the estimation. Simulation studies and two real data examples are presented in sections 6 and 7. Some conclusion remarks are made in Section 8. All proofs are relegated to the Appendix published as online supplementary materials.

2. Principle of finite aggregation

Aggregate dimension reduction 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 explain 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 whose columns form an orthonormal basis of the central subspace. SIR and many other dimension reduction methods require the following *linearity condition* on $\mathbf{X}$:

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

Under this assumption, the random vector $E(\mathbf{X} \mid Y) - E(\mathbf{X})$ 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). Since $\mathbf{B}$ is unknown, this condition is often assumed to

hold for all $p \times d$ matrices, which is equivalent to requiring $\mathbf{X}$ to 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 $\mathbf{m}(\mathbf{u}) = E(\mathbf{X} \mid \mathbf{B}^T\mathbf{X} = \mathbf{u})$ is differentiable, $E(\mathbf{X} \mid \mathbf{B}^T\mathbf{X})$ can be reasonably well approximated by a linear function of $\mathbf{B}^T\mathbf{X}$.

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

$$Y = \beta^T \mathbf{X} + 0.2\varepsilon,$$

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

To construct local dimension reduction spaces, assume $(\mathbf{X}, Y)$ has a

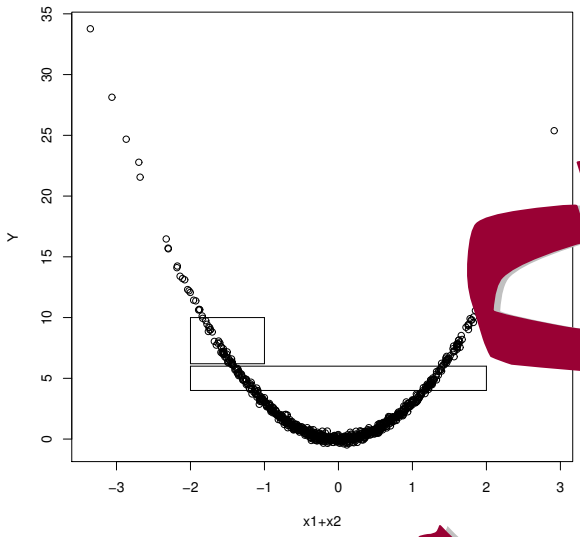

Figure 1: A symmetric model that cannot be estimated by the global SIR

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; that is, $\Omega_{\mathbf{X}} = \{\mathbf{x} : p(\mathbf{x}) > 0\}$ and $\Omega_Y = \{y : g(y) > 0\}$. For convenience, we assume that the support of $(\mathbf{X}, Y)$ 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

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

Let G be any open set in $\Omega_{\mathbf{X}}$. Let $(\mathbf{X}_G, Y_G)$ be defined as $(\mathbf{X}, Y)$ re-

stricted on the set G; that is, for any Borel set $A \subseteq \Omega_{\mathbf{X}} \times \Omega_Y$ one has

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

This defining relation uniquely determines the [illegible]ities and conditional densities of the localized random pair $(\mathbf{X}_G, Y_G)$, a[illegible]n by the fo[illegible] 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 $f_G(\mathbf{x}, y) = f(\mathbf{x}, y)/P(\mathbf{X} \in G)$, $(\mathbf{x}, y) \in G \times \Omega_Y$;*

2. *the marginal density of $\mathbf{X}_G$ is $p_G(\mathbf{x}) = p(\mathbf{x})/P(\mathbf{X} \in G)$, $\mathbf{x} \in G$;*

3. *the conditional density of $Y_G \mid \mathbf{X}_G$ is $h_G(y \mid \mathbf{x}) = h(y \mid \mathbf{x})$, $(\mathbf{x}, y) \in G \times \Omega_Y$;*

4. *the marginal density of Y_G is*

$$q_G(y) = \frac{1}{P(\mathbf{X} \in G)} \int_G f(\mathbf{x}, y) d\mathbf{x}, \quad y \in \Omega_Y.$$

The proof is simple and thus omitted. An important point of this proposition is that the conditional densities of $Y_G \mid \mathbf{X}_G$ and $Y \mid \mathbf{X}$ 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 $\mathcal{S}_{Y|\mathbf{X}}$, since any local relation between Y_G and $\mathbf{X}_G$ must be a part of the global relation between Y and $\mathbf{X}$. In the meantime any relation existing between Y and $\mathbf{X}$ globally must be reflected in some local area G. In fact, more is true — we only need a *finite* collection 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, to be called the Fundamental Aggregation Principle, plays a fundamental role for our method. It guarantees that we can patch together 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-reduction 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 $\dot{h}(y \mid \mathbf{x}) = \partial h(y \mid \mathbf{x})/\partial \mathbf{x}$. Consider the matrices

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

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 span($\boldsymbol{\beta}_G$) = span($\mathbf{H}_G$) and span($\mathbf{B}_G$) = 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 the localized SIR (LSIR) estimator. In an important special case, this space has dimension no greater than that of $\mathbf{X}$. Let $\boldsymbol{\Sigma}_G$ denote the variance matrix of $\mathbf{X}_G$

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

Note that this matrix is of the order $O(\|G\|^2)$ as $\|G\| \to 0$. Let $\bar{G}$ denote the closure of G and $\mathbf{P}_{\boldsymbol{\beta}_G}$ be the projection on to 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 $y \in \Omega_Y$, $g(y) > 0$, $h(y \mid \mathbf{x})$ 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,

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

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

The proof of Theorem 2 is in the Appendix.

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

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

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

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

The proof of Theorem 3 is in the Appendix.

The intuition behind this further reduction of bias is that the leading term of an integral of a centered cubic function over a spherical region is 0. From this theorem we see that the bias of local SIR is four orders of magnitude smaller than the bias of the corresponding global estimate. The bias is surprisingly small, especially if we compare with the population bias of the kernel estimator of a density. Let K be a symmetric 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 $O(\rho)$. The similar bias applies also to the kernel regression setting. This comparison implies that localized dimension reduction has a smaller bias than kernel density estimation or kernel regression. In other words, even in a nonparametric setting where no elliptical distribution is imposed on $\mathbf{X}$, it is still beneficial to first perform dimension reduction before nonparametric regression.

Now let us consider the special case where

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

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

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

Note that

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

This is a matrix of rank 1 unless $\dot{\phi}(\boldsymbol{\mu}_G) = \mathbf{0}$. We summari[illegible] as the following proposition.

Proposition 2. *Suppose $h(y \mid \mathbf{x})$ is of the form [illegible] where h_1 is differentiable with respect to ϕ and ϕ is [illegible] with respect to $\mathbf{x}$. Moreover, suppose $\partial h_1(Y_G, \phi)/\partial\phi$ is squa[illegible]. Then* $\mathrm{span}(\boldsymbol{\beta}_G)$ *has dimension at most 1. That is, [illegible] or of magnitude $O(\|G\|^2)$, the local [illegible]l subspace $\mathcal{S}$ [illegible] dimension at most 1.*

This proposition suggests that if we are interested in finding the central subspace, then we only need to estimate one direction for each local region. That 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 aggregate dimension reduction (ADR), using k-nearest neighbor (kNN) as the localizing mechanism and partial inverse regression as the local dimension reduction estimator. Properties of nearest neighbor estimates have been extensively studied in nonparametric regression and pattern recognition. For example, Hastie et al. (2001).

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

We first describe the estimation procedure at the population level. By Proposition 3.1 under condition (3.3), each local central subspace contains at most 1 direction if we ignore an error of size $\|G\|^2$. This motivates us to employ a two-slice scheme for 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 value 1 if $Y \in J_{G1}$ and 2 if $Y \in J_{G2}$. By the discussion in Section 3, we have, approximately,

$$\text{span}\{\text{Var}[E(\mathbf{X}_G \mid \Delta_G)]\} \subseteq \boldsymbol{\Sigma}_G \mathcal{S}_{\quad} \tag{4.4}$$

Let $\pi_G = P(\Delta_G = 1)$, and $\boldsymbol{\zeta}_{Gu} = E(\mathbf{X}_G \mid \Delta_G = u) - E(\mathbf{X}_G)$ for u Noticing the relation $\pi_G \boldsymbol{\zeta}_{G1} + (1-\pi_G)\boldsymbol{\zeta}_{G2} = \mathbf{0}$, we can rewrite the conditional variance in (4.4) as

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

This is a matrix of rank at most 1.

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

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

where q is a fixed integer. Note that $\boldsymbol{\eta}_G$ is simply the projection of $\boldsymbol{\Sigma}_G^{-1}\boldsymbol{\zeta}$ on to the column space of $\mathbf{R}_G$. Cook et al. (2007) show that the subspace $\text{span}(\mathbf{R}_G)$ is strictly increasing when q increases, and argue that it often grows large enough 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 $\mathcal{S}_{Y_G|\mathbf{X}_G}$. 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 whose meaning and choice are discussed in the next section.

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

1. For each $s = 1, \ldots, n$, let G_s be the set that includes the k nearest $\mathbf{X}_j$'s to $\mathbf{X}_s$ in terms of the Euclidean distance $\|\mathbf{X}_j - \mathbf{X}_s\|$. Note that G_s contains $k + 1$ elements since we do not count $\mathbf{X}_s$ as 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's. Let n_{su}, $u = 1, 2$ be cardinality of the set $\{j : \mathbf{X}_j \in G_s, Y_j \in J_{su}\}$, and $n_s = n_{s1} + n_{s2}$. Let

$$\bar{\mathbf{X}}_{G_{s1}} = \frac{1}{n_{s1}} \sum \mathbf{X}_j I(\mathbf{X}_j \in G_s, Y_j \in J_{s1}), \quad \bar{\mathbf{X}}_{G_{s2}} = \frac{1}{n_{s2}} \sum \mathbf{X}_j I(\mathbf{X}_j \in G_s, Y_j \in J_{s2})$$

and

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

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

It is well known that every kernel based estimate can be introduced from the above choice of the neighborhood in high dimensional input space with finite sample. Since the Euclidean distance measure implies that the input features are homogeneous or isotropic, an immediate remedy would be to use an adaptive metric. Inspired by the work of Hastie and Tibshirani (1996), here we propose a refined estimation where the neighborhoods are elongated along less relevant directions and constricted along those influential ones. After obtaining a basis of 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 will use a p-dimensional ellipsoid with which 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 $\mathbf{X}_s$ as in the step 1 of the above algorithm will be replaced by

$$d_{js}^2 = \|\hat{\mathbf{B}}_{(0)}^T(\mathbf{X}_j - \mathbf{X}_s)\|^2 + \kappa_{(0)}\|(\mathbf{X}_j - \mathbf{X}_s)\|^2$$
$$= (\mathbf{X}_j - \mathbf{X}_s)^T[\hat{\mathbf{B}}_{(0)}\hat{\mathbf{B}}_{(0)}^T + \kappa_{(0)}\mathbf{I}_p](\mathbf{X}_j - \mathbf{X}_s), \quad (4.5)$$

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

Our method differs from Li (1991) where k-nearest neighborhood is applied to multivariate Y to form slice. It is also different from the IMAVE procedure of Xia et al. (2002), in that the latter requires the validity condition.

5. Tuning parameters

In this section we discuss how to choose the various turning parameters in the estimation algorithm described in Section 4, which include the estimation of the structural dimension d, the choices of the weighting function t,

the order q in partial inverse regression, as well as the softening parameter κ in the adaptive nearest neighborhood selection. An appropriate justification of these choices rely on the asymptotic properties of ADR, which are beyond the scope of the present paper, and will be carried out in a separate study. Inevitably, the following recommendations are heuristic in nature. In our extensive numerical experiments, we perform sensitivity analysis on the recommended choices of these tuning parameters and our results showed reasonably stable estimation.

We recommend two choices for t. A natural choice is $t(\omega_G) \equiv 1$. From the discussion in Section 3, $\boldsymbol{\zeta}_G$ 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 1 we let the sliced means be used to determine the relative importance of each neighborhood. A second choice of t is

$$t(\hat{\omega}_G) = \begin{cases} \|\hat{\boldsymbol{\zeta}}_G\|^{-2} & \hat{\omega}_G > c \\ 0 & \hat{\omega}_G \leq c. \end{cases} \tag{5.6}$$

This weighting function introduces a hard thresholding according to the magnitude of $\|\hat{\boldsymbol{\zeta}}\|$; it throws away those neighborhoods with small sliced means. Moreover, when a sliced mean is large enough, its magnitude is

no longer included in the estimation. Based on our experience the second choice seems to work better. We choose threshold c according to a percentage δ of sample size. That is, we choose $\delta \times 100\%$ of neighborhoods with 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 et al. (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 eigenvalues of matrix $\hat{\mathbb{R}}_{G_s}\hat{\mathbb{R}}^T_{G_s}$ 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 $\{\hat{\mathcal{S}}^{(j)}_{d^*},\ j = 1, \ldots, n_b\}$ through bootstrapping, where n_b is the number of bootstrap samples. The distances between $\hat{\mathcal{S}}_{d^*}$ and its bootstrap versions $\{\hat{\mathcal{S}}^{(j)}_{d^*},\ j = 1, \ldots, n_b\}$ can be used to assess the variability of the estimated subspace for $d = d^*$, which in turn can be used to infer the structural dimension. Intuitively, $\hat{\mathcal{S}}_{d^*} \subseteq \mathcal{S}_{Y|\mathbf{X}}$ when $d^* \leq d$. But 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 $\mathcal{S}_{Y|\mathbf{X}}$. Since $\tilde{\mathcal{S}}$ can be arbitrary, we expect to see larger variability of $\hat{\mathcal{S}}_{d^*}$ with its bootstrap versions, compared to when $d^* \leq d$. Therefore, the

structural dimension d can be estimated as the largest d^* that produces a stable estimator.

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

6. Simulation studies

In this section, we evaluate the performance of aggregate dimension reduction by simulation. For comparison purposes, several existing methods were also evaluated in the simulation studies, including sliced inverse regression (SIR), sliced average variance estimation (SAVE), principal Hessian directions (PHD), minimum average variance estimation (MAVE), and sliced regression (SR). The correlation coefficient q (Hotelling, 1936; Ye and Weiss, 2003) is 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 $0 \leq \rho_d \leq \cdots \leq \rho_1 \leq 1$ are the eigenvalues of matrix $\hat{\mathbf{B}}^T(\mathbf{B}\mathbf{B}^T)\hat{\mathbf{B}}$. The larger the q is, the closer $\mathcal{S}(\hat{\mathbf{B}})$ is to $\mathcal{S}(\mathbf{B})$. We chose the Gaussian

kernel and its corresponding optimal bandwidth for MAVE and SR. A rule-of-thumb choice $k = 4p$ was used for our proposed aggregate approach, including kNN sliced inverse regression (kNNSIR) and adaptive kNN sliced inverse regression (a-kNNSIR where adaptive distance (4.1) will be used). More refined ways to choose k, such as cross-validation, could be used at greater computational expense. For each parameter setting, 200 simulation replications were conducted.

The following 4 models were used in the numerical studies:

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: $Y = \text{sign}(\beta_1^T X + \epsilon_1)\log(|\beta_2^T X + 4 + \epsilon_2|)$,

Model 4: $Y = (\beta_1^T X)(\beta_2^T X + 1) + (\beta_2^T X + 2)^3 + 0.5\epsilon$.

All the above models were studied extensively in sufficient dimension reduction literature. For all four models, $X \sim N_p(0, \Sigma)$, independent of the error. Gaussian noises ϵ, ϵ_1 and ϵ_2. The covariance matrix $\Sigma = (\sigma_{ij})_{p\times p}$ where $\sigma_{ij} = \rho^{|i-j|}$, $\rho = 0.5$ 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 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 $\beta_3 = (0, 0, 0, 1, 1, 0, \ldots, 0)^T$.

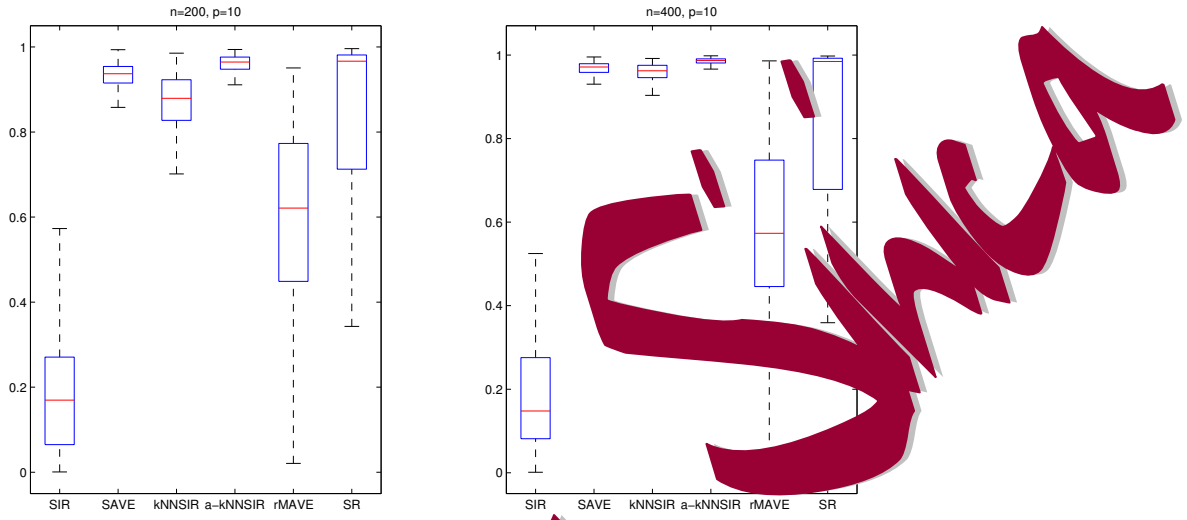

Figure 2: Comparison of estimation accuracy with Model 1

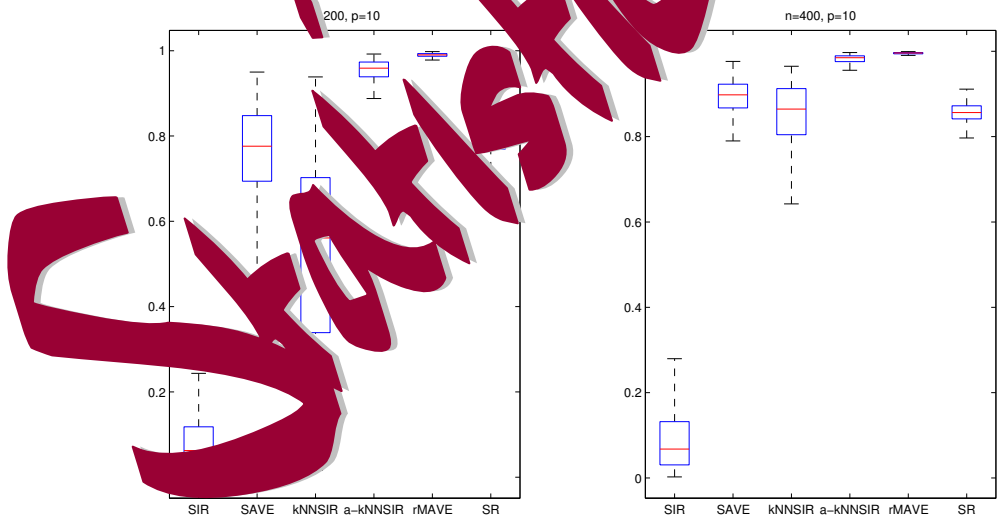

Figure 3: Comparison of estimation accuracy with Model 2

Figures 2 – 5 show the comparisons of the performance among the afore-

mentioned methods. We can have the following observations from these graphical summaries. First, the proposed aggregate SDR, adaptive kNN-SIR, significantly improves the performance of the original inverse regression methods and is broadly comparable with the forward regression approaches (MAVE and SR). Secondly, through localization, adaptive kNN-SIR can overcome the drawback of missing symmetric patterns of the original SIR such as in models 1 and 2. Thirdly, when $\mathcal{S}_{Y|\mathbf{X}}$ is completely contained in the mean regression function $E(Y \mid \mathbf{X})$, MAVE stands out as the best method without surprise while our proposed aggregate kNNSIR is the close second as in models 2 and 4. But when $\mathcal{S}_{Y|\mathbf{X}}$ spans beyond the mean function as in models 1 and 3, a-kNNSIR clearly outperforms MAVE. Finally, larger sample sizes are needed to provide good estimation with the increase of the dimension d. Zhu et al. (2010) studied model 4 ($d = 3$) and showed that n needs to be increased to at least 200 in order for the estimation accuracy of SIR to be comparable when $p \leq 20$. In our numerical study, the proposed a-kNNSIR and MAVE are the only two methods with good performance for moderate sample sizes. It is well known that the computation burden increases significantly with the increase of n and p for forward regression methods (MAVE and SR), while our proposed aggregate inverse regression approach is more computationally efficient since no numerical optimization

was involved. This was also confirmed in our simulation studies.

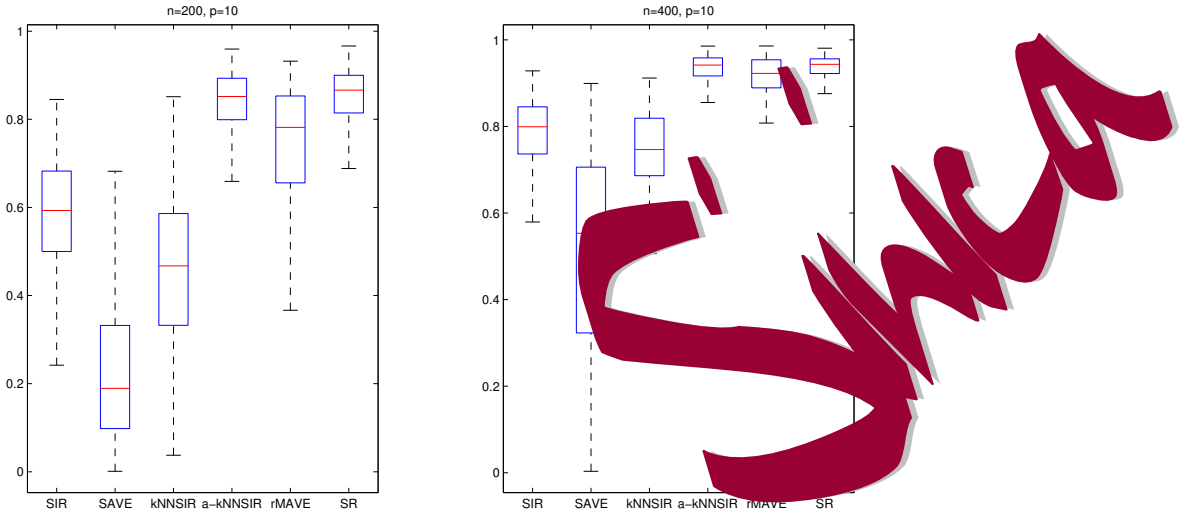

Figure 4: Comparison of estimation accuracy with Model 3

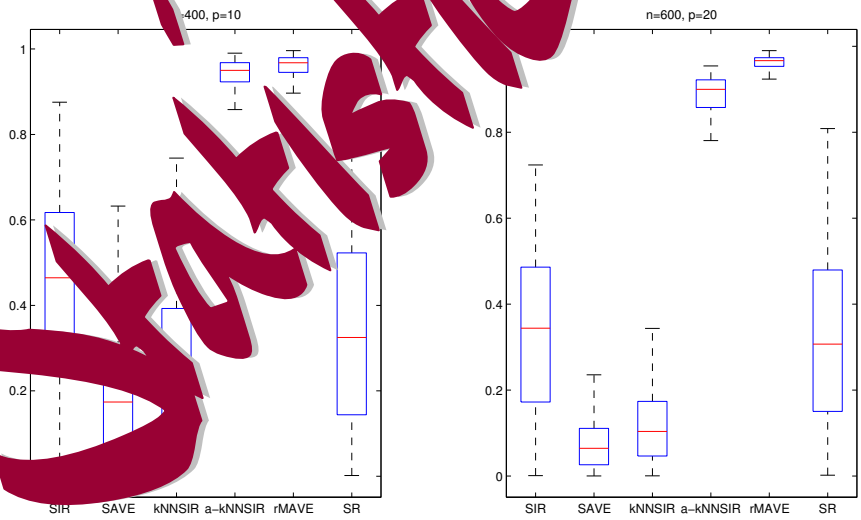

Figure 5: Comparison of estimation accuracy with Model 4

Next, we estimated the structural dimension d using the adopted boot-

strap procedure. In all the 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 of $d^* = 1, 2, \ldots, p-1$, 500 bootstrap samples were drawn and the median of the distances between $\hat{\mathcal{S}}_{d^*}$ and its bootstrap versions $\{\hat{\mathcal{S}}_{d^*}^{(j)},\ j = 1, \ldots, 500\}$ were calculated. Figure 6 shows the dimension variability plots (Zhu and Zeng, 2006) for models 1-4. A substantial large variability showed up when $d^* > d$. Out of 100 samples with $n = 400$ and $p = 10$, the accuracy of correctly estimated d is 99%, 94%, [illegible] 84% for models 1-4, respectively.

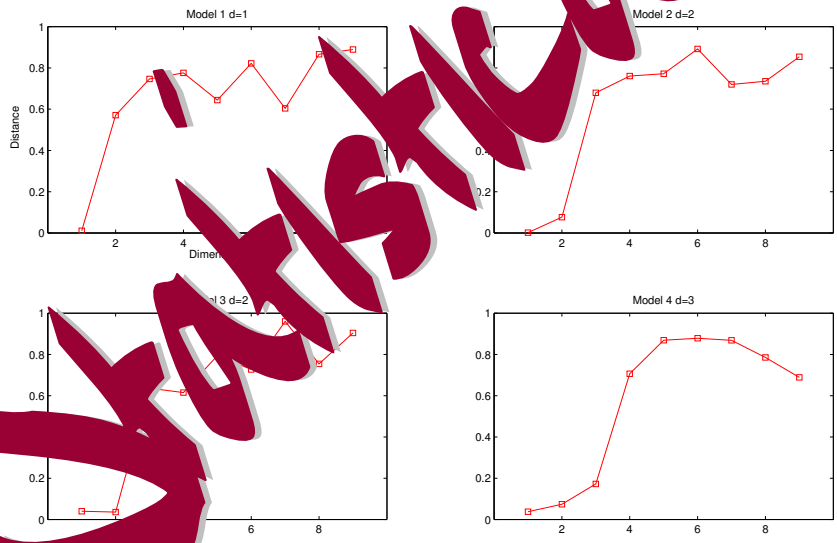

Figure 6: Bootstrap estimation of dimension ($n = 400$ and $p = 10$)

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 set concerning [illegible]ation between the ozone levels and various environmental variables [illegible]reiman and Fr[illegible]n (1985). The data contain 330 observations, with each observation cons[illegible]r of 9 variables: ozone concentration, height, inversion heigh[illegible]e, inversion temperature, humidity, pressure, visi[illegible]ity, and wind speed, where ozone concentration is treated as the response a[illegible] other 8 variables are treated as predictors. Fo[illegible] ease [illegible]terpr[illegible]ation, all predictors were standardized separately. This data [illegible] has [illegible]nalyzed by several authors. See, for example, Li (1992) [illegible]n[illegible] and Li (2004).

SIR identifies one si[illegible]fica[illegible] direction. After a closer investigation of the residua[illegible] fro[illegible] t[illegible]e [illegible]tic fit, Li (1992) argued a second significant [illegible]sa[illegible] nd PHD can recover this direction. Cook and Li (2004) also identifi[illegible] the first direction using IHT method (Inverse Hessian Transfor[illegible]), but argued the estimate of dimension d which is different from different testing methods, leaving some uncertainty.

In our application, the dimension variability plot, shown in Figure 7-

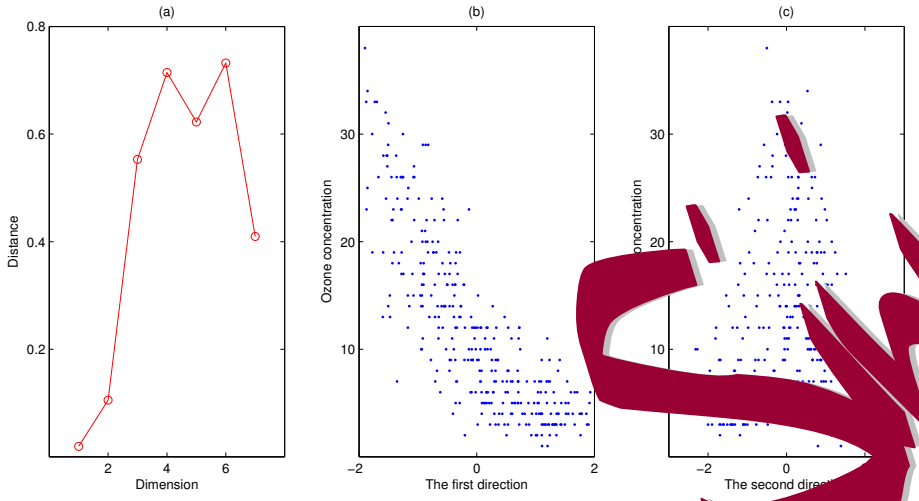

Figure 7: Analysis on Ozone data: (a) dimension variability pl ter-plots of response vs. the two estimated direction

(a), suggested $\hat{d} = 2$. Figure 7-(b)(c) showed pattern identified by our method. Inte estingly, our SIR successfully recovers the two significant component h SIR and PHD, without fitting a detailed model as in y uncertainty on estimating d as in and Li (200

n data

This data set d 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

Table 1: The predictors and the estimated directions of 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		
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 spendi	−0.12	−0.30
x_7	perce of faculty w r gree	0.03	−0.03
x_8	student/facul r	0.13	0.46
x_9	percent o hc nate	0.04	0.07
x_1	instru l ex nditure per student	0.12	−0.26
x_{11}	gr at ate	0.04	−0.60

the nu cations received (y) by 557 private institutions with full time undergraduate student body less than 10,000. The predictors used in our analysis are listed in Table 1. Again for the ease of interpretation, all predictors were standardized separately.

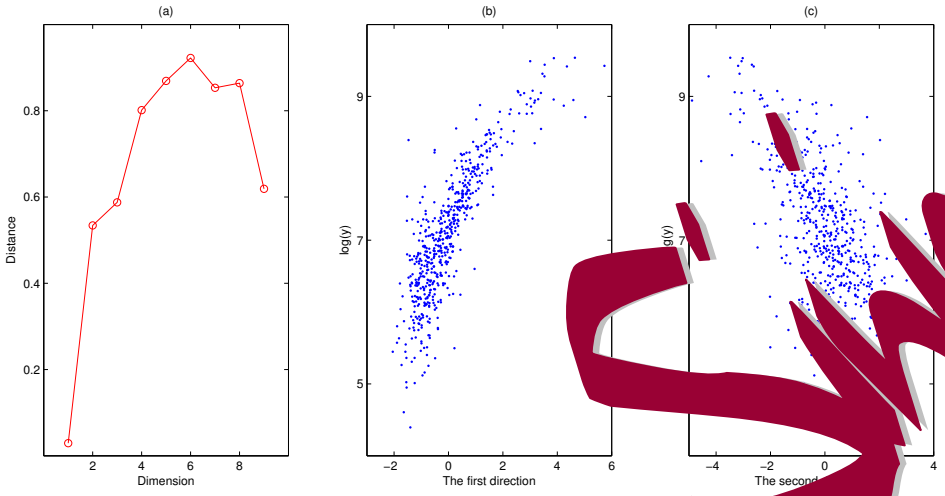

Figure 8: Analysis on College admission data: (a) dimension variability plot, (b-c) scatterplots of response vs. the two estimated directions.

The dimension variability plot in Figure 8(a) suggests at most 3 dimensions. It also indicates that the variability for the second and third directions may not be very strong as their variability is much larger than the first one. Situations like this can often happen in practice as real data may have big noise and weak signal, which makes the determination of the structural dimension less obvious. Nevertheless, we further look at the coefficients and marginal plots for the first three directions. In the end, we decide to keep the first two directions since no good interpretation can be found for the third direction. We also applied SIR to this data set. The asymptotic test also suggested $d = 3$. 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 by our method, we can interpret the first direction as the 'size' factor since it is dominated by x_1, the number of full time undergraduates. 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), we see that in general the number of applications increases with the size of the institution's student body, with this increasing trend tapering off towards the end. Figure 8 (c) shows more students would apply the institutions with high academic quality, meaning high graduation rate, high instructional expenditure and small student/faculty ratio.

8. Discussion

In this article, we proposed an aggregate approach to estimate the central subspace and explored this idea through adaptive kNN sliced inverse regression. We showed that a class of new local dimension reduction approach can be developed under this localization framework. Our new method is not to replace the original SIR. Instead, we developed an alternative approach so that the simplicity of SIR can be extended further.

There are still several open questions that need further study, such as

the asymptotic properties of the proposed estimators and the extension to big data setting. For the study of asymptotic properties, the most related one in the global sense is the paper by Hsing and Carroll (19[illegible]) where the estimator from two-slice approach was shown to be r[illegible]ot-n consisten[illegible] However, due to the use of local approximation, ou[illegible]cal inverse co[illegible]itio[illegible] covariance matrix does not have the closed form as [illegible] in [illegible] and Carroll (1992). Since the k-nearest-neighbor estimation can be tre[illegible]d as a special kernel method, our proposed localization-aggreg[illegible]pproach is similar, in spirit, to the kernel based Outer [illegible]uct of Gradients (OPG) estimation (Xia et al., 2002). Considering the [illegible]allenges and difficulties, we decide to leave it for a sepa[illegible]te st[illegible] On[illegible]eferee brings our attention to extension to big data settin[illegible] lar[illegible] n and/or large p. When the volume n is huge, th[illegible] moderate and $n > p$, we propose to [illegible]ment the *loc[illegible]ati[illegible]aggre[illegible]tion* approach together with 'leveraging based subsa[illegible] (Ma et al., 2015). The case, where $n < p$, or even $n << p$, [illegible] ore challenging. We adopt the *sequential dimension reduction* parad[illegible] roposed 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.

Acknowledgements We would like to thank Co-Editor Professor Zhil-

iang Ying, Associate Editor and two anonymous referees for their careful reading and constructive comments, which led to substantial improvement in the manuscript.

References

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

Breiman, L. and J. H. Friedman (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 R. D. Cook (2001a). Estimating the structural dimension of regressions via parametric inverse regression. *Journal of the Royal Statistical Society Series B 63*, 393–410.

Bura, E. and R. Pfeiffer (2003). Extending sliced inverse regression: the weighted chi-squared test. *Journal of the American Statistical Association 96*, 996–1003.

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

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

REFERENCES

Chiaromonte, F. and J. A. Martinelli (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*. New York: Wiley.

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

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

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

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

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

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

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

Fung, W. K., X. He, L. Liu, and P. Shi (2002). Dimension reduction based on canonical

correlation. *Statistica Sinica 12*, 1093–1113.

Hastie, T. and R. Tibshirani (1996). Discriminant adaptive nearest neighbor classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence 18*(6), 607–616.

Hastie, T., R. Tibshirani, and J. H. Friedman (2001). *The elements of statistical learning : data mining, inference, and prediction*. Springer series in statistics. New York: Springer.

Hotelling, H. (1936). Relations between two sets of variates. *Biometrika*.

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

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

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

Li, B. and S. Wang (2007). On directional regression for dimension reduction. *Journal of the American Statistical Association 102*(479), 997–1008.

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

Li, K. C. (1991). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association 86*(414), 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., R. D. Cook, and C. J. Nachtsheim (2004). Cluster-based estimation for sufficient dimension reduction. *Computational Statistics and Data Analysis 47*, 175–193.

Li, L., R. D. Cook, and C. L. Tsai (2007). Partial inverse regression. *Biometrika 94*, 615–62

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

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

Ma, P., M. W. Mahoney, and [illegible] (2[illegible]). A statistical perspective on algorithmic leveraging. *Journal of Machine Learning Research 16*, 861–911.

[illegible] and C. L[illegible] (20[illegible]). Partial least squares estimator for single-index models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 62*, 763–771.

Schott, J[illegible]. 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 X. Yin (2011). Estimation of inverse mean: An orthogonal series approach. *Computational Statistics and Data Analysis 55*(4), 1656–1664.

Wu, Q., F. Liang, and S. Mukherjee (2010). Localized sliced inverse regression. *Journal of Computational and Graphical Statistics 19*(4), 843–860.

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

Ye, Z. and R. E. Weiss (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 R. D. Cook (2003). Estimating central subspace via inverse third moment. *Biometrika 90*, 113–125.

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

Yin, X., B. Li and R. D. Cook (2008). Successive direction extraction for estimating the central subspace in a multiple-index regression. *Journal of Multivariate Analysis 99*(8), 1733–1757.

Zhu, L. X. and K. T. Fang (1996). Asymptotics for kernel estimate of sliced inverse regression. *The Annals of Statistics 24*(3), 1053–1068.

Zhu, L. X., B. Q. Miao, and H. Peng (2006). On sliced inverse regression with high-dimensional

covariates. *Journal of the American Statistical Association 101*, 630–643.

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

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

Department of Statistical Sciences and Operations Research, Virginia Commonwealth University

E-mail: qwang3@vcu.edu

Department of Statistics, The University of Kentucky

E-mail: yinxiangrong@uky.edu

Department of Statistics, Pennsylvania State University

E-mail: bing@stat.psu.edu

Pharmaceutical Product Development, LLC.

E-mail: lilylilytang@gmail.com