Statistics and Its Interface Volume 2 (2009) 503–511

A minimum discrepancy approach to multivariate dimension reduction via k-means inverse regression*

Xuerong Meggie Wen†, C. Messan Setodji and Akim Adekpedjou

We proposed a new method to estimate the intra-cluster adjusted central subspace for regressions with multivariate responses. Following Setodji and Cook (2004), we made use of the k-means algorithm to cluster the observed response vectors. Our method was designed to recover the intra-cluster information and outperformed previous method with respect to estimation accuracies on both the central subspace and its dimension. It also allowed us to test the predictor effects in a model-free approach. Simulation and a real data example were given to illustrate our methodology.

AMS 2000 subject classifications: Primary 62G08, 62H12; secondary 62H30.
Keywords and phrases: Multivariate Regression, Dimension Reduction, Central Subspaces, Intra-cluster Information, k-means Clustering.

1. INTRODUCTION

Sufficient dimension reduction (SDR; Cook 1994, Cook 1998) focuses on reducing the dimension of the predictors in a regression context without assuming any parametric model. The basic idea is to replace the predictors $\mathbf{X} \in \mathbb{R}^p$ with a lower dimensional projection $P_{\mathcal{S}}\mathbf{X}$ onto a subspace $\mathcal{S} \subseteq \mathbb{R}^p$ without the loss of information on the original regression $\mathbf{Y}|\mathbf{X}$, where $P_{(.)}$ stands for a projection operator with respect to the standard inner product, and $\mathbf{Y} \in \mathbb{R}^r$, $r \geq 1$.

Regressions with multivariate responses are wildly used in many areas such as chemometrics, econometrics, financial engineering and psychometrics. But for most usage, model assumptions such as a linear model are commonly made in order to analyze the relationship between outcomes and predictors. Any possibility of reduction of the predictors greatly reduces the burden of positing a model structure at the onset of an analysis, because with only one response and one predictor for example, a plot of those two variables will inform about such structure. Over the past decades, methods such as principal component analysis (Massy, 1965), partial least squares (Helland, 1989, 1990) and canonical correlation (Hotelling, 1935, 1936) have been used for variable reduction as well as the broader notion of sufficient dimension reduction (Li and Duan, 1989; Cook, 1994). Following Setodji and Cook (2004), we propose a model-free dimension reduction method for multivariate regressions under the notion of multivariate central subspaces (Setodji and Cook, 2004).

The multivariate central subspace (CS), denoted by $\mathcal{S}_{\mathbf{Y}|\mathbf{X}}$ (Cook 1994, Setodji and Cook 2004) is the smallest subspace $\mathcal{S}$ of $\mathbb{R}^p$ such that, for the projection $P_{\mathcal{S}}\mathbf{X}$ of $\mathbf{X}$ on $\mathcal{S}$,

$$\mathbf{Y} \perp\!\!\!\perp \mathbf{X} \mid P_{\mathcal{S}}\mathbf{X}, \tag{1.1}$$

which is uniquely defined and coincides with the intersection of all those subspaces satisfying (1.1) under some mild conditions (Cook, 1998). It guarantees that the regression of $\mathbf{Y}$ on $\mathbf{X}$ can be replaced by the regression of $\mathbf{Y}$ on the smaller dimension random variable $P_{\mathcal{S}}\mathbf{X}$ without any loss of information and any model assumption. We assume that central subspace exists throughout this article and we will define $d = \dim(\mathcal{S}_{\mathbf{Y}|\mathbf{X}})$.

In the past decade, many methods have been developed to estimate $\mathcal{S}_{\mathbf{Y}|\mathbf{X}}$ when the response is univariate, but there is relatively little methodology available when the response is multivariate. In the univariate case where $Y \in \mathbb{R}^1$, Li (1991) proposed for estimation the method of sliced inverse regression (SIR) that is based on the inverse mean $\mathrm{E}(\mathbf{X}|Y)$ which is easily estimable when Y is categorical, but when the outcome is continuous, it is replaced by a dichotomized version $\tilde{Y}$ with the range of Y partitioned into a fixed number (h) of slices. Except for a few methods such as (constrained) CANCOR (Fung, He, Liu and Shi, 2002; Zhou and He, 2008) which used the B-spline basis functions generated for the response variable, and kernel-based methods (Zhu and Fang, 1996; Zhu and Zhu, 2007); slicing a continuous response is a common practice in the field of inverse-regression dimension reduction. Similar dichotomization can be used for multivariate responses. However, when the dimensionality of the response vector increases, the usual slicing methods may not work well due to the curse of dimensionality. To deal with such problem, Aragon (1997) proposed marginal slicing, where the slices are obtained through a user-specified

*We thank the editor and an anonymous referee for their insightful suggestions and comments which lead to great improvement of an earlier draft.
†Corresponding author.

scalar function of $\mathbf{Y}$, for example, the first principal component of $\mathbf{Y}$. Setodji and Cook (2004) introduced a new way of performing the slicing to obtain $\tilde{\mathbf{Y}} \in \mathbb{R}^1$. They adopted the k-means algorithm (Hartigan 1975) to cluster the observed response vectors, and assume the inverse distribution function of $\mathbf{X}|\tilde{\mathbf{Y}}$ well approximates the true distribution of $\mathbf{X}|\mathbf{Y}$ when the number of clusters is sufficiently large. Their method widens the SDR methodology to regressions with multivariate responses without worrying about the number of clusters (slices) increasing exponentially.

However, whether using the standard slicing method or the k-means clustering, all these methods ignored intra-cluster (slice) information, which could be substantial under some circumstances. For univariate response $Y \in \mathbb{R}^1$, Cook and Ni (2006) recently pointed out that intra-slice information was lost when converting a continuous response to a categorical one, and they developed new methodology for recovering intra-slice information in univariate regressions. In this paper we extend the new dimension reduction method to multivariate responses, which is designed to recover intra-cluster information when at least one response is continuous. Our method also allows us to test predictor effects easily, a simple chi-squared distribution can be used in inference methods for d and it has optimal properties to be described later.

The rest of this paper is organized as follows. We first give a brief review of the previous method (*k-means inverse regression estimation*, KIR) proposed by Setodji and Cook (2004) in Section 2. In Section 3 we propose our new method, which we call *generalized k-means inverse regression estimation* (GM.KIRE). Its asymptotics are also discussed. Section 4 is dedicated to the discussion of testing predictor effects under the context of GM.KIRE. The performances of GM.KIRE and KIR are compared via simulation studies, and a real data analysis is also discussed in Section 5. Brief conclusions are given in Section 6.

2. K-MEANS INVERSE REGRESSION ESTIMATION

Li's 1991 SIR method is based on the notion that if one can find a matrix M such that $\text{Span}(M) \simeq \mathcal{S}_{\mathbf{Y}|\mathbf{X}}$ then $\mathcal{S}_{\mathbf{Y}|\mathbf{X}}$ can be estimated as the span of the left singular vectors of the sample estimate $\hat{M}$ of M whose singular values are inferred to be non-zero. He posited that if the linearity condition, defined by $\text{E}(\mathbf{X}|\boldsymbol{\rho}^T\mathbf{X})$ being a linear function of $\boldsymbol{\rho}^T\mathbf{X}$ where the columns of $\boldsymbol{\rho} \in \mathbb{R}^{p\times d}$ is a basis for $\mathcal{S}_{\mathbf{Y}|\mathbf{X}}$, is satisfied, then the inverse mean space $\mathcal{S}_{\text{E}(\mathbf{X}|\mathbf{Y})} = \text{Span}\{\text{E}(\mathbf{X}|\mathbf{Y}) - \text{E}(\mathbf{X}), \mathbf{Y} \in \Omega_y\}$ where Ω_y is the support of $\mathbf{Y}$ provides a good estimate of $\mathcal{S}_{\mathbf{Y}|\mathbf{X}}$. Further more, he showed that $\mathcal{S}_{\text{E}(\mathbf{X}|\mathbf{Y})} = \text{Span}[\text{Var}\{\text{E}(\mathbf{X}|\mathbf{Y})\}]$ and thus $\text{Var}\{\text{E}(\mathbf{X}|\mathbf{Y})\}$ could be used as the matrix M. The linearity condition, an assumption only about the marginal distribution of $\mathbf{X}$ is commonly used in dimension reduction methods and holds for elliptically contoured predictors (Eaton 1986). Additionally, Hall and Li (1993) showed that as the number of predictors p increases, the linearity condition holds to a reasonable approximation in many problems. See Wen and Cook (2007) for more discussions on this topic.

When $\mathbf{Y}$ is categorical $\text{Var}\{\text{E}(\mathbf{X}|\mathbf{Y})\}$ can be estimated directly, but for continuous responses, $\mathbf{Y}$ is categorized by slicing. Let $\boldsymbol{\Sigma} = \text{Cov}(\mathbf{X})$ be the covariance matrix of $\mathbf{X}$, and let $J_s(\mathbf{Y}) = I\{\tilde{\mathbf{Y}} = s\}$ be the dichotomized random variable taking value 1 when the values of $\mathbf{Y}$ fall into the s slice and 0 if not, with $s = 1, \ldots, h$, h is the total number of slices. We will also define $\boldsymbol{\xi}_s = \boldsymbol{\Sigma}^{-1}\text{Cov}(\mathbf{X}, J_s)$ and $f_s = \Pr(J_s = 1)$ the probability of an observation falling in slice s. Under the linearity condition, $\boldsymbol{\xi}_s \in \mathcal{S}_{\mathbf{Y}|\mathbf{X}}$ and the matrix M with columns formed by $\boldsymbol{\xi}_s$, $s = 1, \ldots, h$ will well approximate the central subspace $\mathcal{S}_{\mathbf{Y}|\mathbf{X}}$. K-means inverse regression (KIR) is an extension of SIR where simple slices are replaced by clusters obtained from the k-means algorithm.

Now starting with a random sample $(\mathbf{X}_i, \mathbf{Y}_i)$, $i = 1, \ldots, n$, on $(\mathbf{X}, \mathbf{Y})$, the estimate $\hat{\boldsymbol{\xi}}_s$ of $\boldsymbol{\xi}_s$ can be estimated as the $\mathbf{X}$ coefficients from the ordinary least squares fit of $J_s(\mathbf{Y}_i)$ on $\mathbf{X}_i$, including an intercept.

Besides the SIR/KIR algorithm described above, Cook and Ni (2005) showed that, if one assumes $d = dim(\mathcal{S}_{\mathbf{Y}|\mathbf{X}})$ the dimension of the central subspace known, a basis of the $\mathcal{S}_{\mathbf{Y}|\mathbf{X}}$ can be obtained by minimizing the minimum discrepancy function.

$$F_d^{SIR}(\mathbf{B}, \mathbf{C}) \quad = \quad \sum_{s=1}^{h} (\hat{f}_s\hat{\boldsymbol{\xi}}_s - \mathbf{B}\mathbf{C}_s)^T \hat{f}_s^{-1}\hat{\boldsymbol{\Sigma}}(\hat{f}_s\hat{\boldsymbol{\xi}}_s - \mathbf{B}\mathbf{C}_s),$$

where $\mathbf{B} \in \mathbb{R}^{p\times d}$, $\mathbf{C}_s \in \mathbb{R}^d$, $\hat{f}_s$ is the fraction of the sample points in slice s, $s = 1, \ldots, h$, and $\hat{\boldsymbol{\Sigma}}$ is the sample version of $\boldsymbol{\Sigma}$. The value of $\mathbf{B}$ that minimizes such function F_d^{SIR} provides such an estimate. Here and in what follows we use $\hat{F}$ to denote the value of an objective function minimized over its arguments. For instance, $\hat{F}_d^{SIR} = F_d^{SIR}(\hat{\mathbf{B}}, \hat{\mathbf{C}})$ and $\text{Span}(\hat{\mathbf{B}})$ provides a consistent estimate of $\mathcal{S}_{\mathbf{Y}|\mathbf{X}}$ where $n\hat{F}_m^{SIR}$ can be used to test hypothesis $d = m$ vs $d > m$. In the next section, we present a generalized version of KIR.

3. GENERALIZED K-MEANS INVERSE REGRESSION ESTIMATION

3.1 GM.KIRE

In section 2, as we can see it, KIR used only the inverse mean information $\text{E}(\mathbf{X}|\tilde{\mathbf{Y}})$ within a slice but ignored the valuable information between slices or clusters (intra-slice or intra-cluster information). When $\mathbf{Y}$ is continuous, the loss of information could be substantial. An extreme case is when $h = 1$, $\boldsymbol{\xi}_s = 0$, suggesting no information is left. Cook and Ni (2006) developed a method to recover the intra-slice information when the response Y is univariate. In this section, we propose a new method which recovers this intra-cluster information for multivariate $\mathbf{Y}$.

Letting $Y^{(k)}$ be the k-th coordinate of $\mathbf{Y}$, $k = 1, \ldots, r$, and because of the subsequent cluster arguments, without loss of generality, we assume that all these $Y^{(k)}$ are continuous. Define

$$\boldsymbol{\beta}_{ks} = \boldsymbol{\Sigma}^{-1}\mathrm{Cov}(\mathbf{X}, Y^{(k)}J_s(\mathbf{Y})), \tag{3.2}$$

where $s = 1, \ldots, h$ and h is the total number of clusters. Hence, within each cluster s, we have r vectors: $\boldsymbol{\beta}_{1s}, \ldots, \boldsymbol{\beta}_{rs}$. Under the linearity condition, we have $\boldsymbol{\beta}_{ks} \in \mathcal{S}_{\mathbf{Y}|\mathbf{X}}$. We will then assume that $\mathcal{S}_{\mathbf{Y}|\mathbf{X}} = \mathrm{Span}(\boldsymbol{\beta}_{11}, \ldots, \boldsymbol{\beta}_{r1}, \ldots, \boldsymbol{\beta}_{1h}, \ldots, \boldsymbol{\beta}_{rh})$, a *coverage condition* commonly used. See Cook and Ni (2005) and Wen and Cook (2007) for more discussion on the coverage condition. For $\boldsymbol{\beta} = (\boldsymbol{\beta}_{11}, \ldots, \boldsymbol{\beta}_{rh})$, we can always find a matrix $\boldsymbol{\gamma} \in \mathbb{R}^{d\times rh}$ such that $\boldsymbol{\beta} = \boldsymbol{\rho}\boldsymbol{\gamma}$, where the columns of $\boldsymbol{\rho} \in \mathbb{R}^{p\times d}$ form a basis for the multivariate central subspace $\mathcal{S}_{\mathbf{Y}|\mathbf{X}}$.

Following Cook and Ni (2006), we decompose $\boldsymbol{\beta}_{ks}$ into two parts:

$$\boldsymbol{\beta}_{ks} = f_s\boldsymbol{\Sigma}^{-1}\mathrm{Cov}(\mathbf{X}, Y^{(k)}|J_s = 1) + f_sE(Y^{(k)}|J_s = 1)\boldsymbol{\xi}_s. \tag{3.3}$$

Equation 3.3 shows us how the intra-cluster information is being recovered for each continuous variable $Y^{(k)}$ with GM.KIRE, by using the the intra-cluster covariance between $Y^{(k)}$ and $\mathbf{X}$. Let $\hat{\boldsymbol{\beta}}_{ks}$ be the $\mathbf{X}$ coefficients from the ordinary least squares fits of $Y_i^{(k)}J_s(Y_i^{(k)})$ on $\mathbf{X}_i$ respectively, including an intercept. Letting $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_{11}, \ldots, \hat{\boldsymbol{\beta}}_{rh})$, we then estimate $(\boldsymbol{\rho}, \boldsymbol{\gamma})$ by minimizing the following quadratic discrepancy function:

$$(\,\mathrm{vec}(\hat{\boldsymbol{\beta}}) - \mathrm{vec}(\mathbf{BC}))^T\mathbf{V}_n(\,\mathrm{vec}(\hat{\boldsymbol{\beta}}) - \mathrm{vec}(\mathbf{BC})), \tag{3.4}$$

where $\mathbf{B} \in \mathbb{R}^{p\times d}$, $\mathbf{C} \in \mathbb{R}^{d\times(rh)}$, and $\mathbf{V}_n > 0$. Different choices of $\mathbf{V}_n$ will specify different estimation methods, while the inverse of the asymptotic covariance matrix of $\sqrt{n}(\,\mathrm{vec}(\hat{\boldsymbol{\beta}}) - \mathrm{vec}(\boldsymbol{\beta}))$ is the "optimal" choice (Ferguson 1958; Shapiro 1985).

Define $\boldsymbol{\epsilon} = (\varepsilon_{11}, \ldots, \varepsilon_{r1}, \ldots, \varepsilon_{rh})^T$ where the elements ε_{ks}, $k = 1, \ldots, r$; $s = 1, \ldots, h$, are the population residuals from the ordinary least squares fit of $Y^{(k)}J_s(\mathbf{Y})$ on $\mathbf{X}$. The asymptotic distribution necessary to select $\mathbf{V}_n$ optimally is given in the following Lemma. The proof is similar to that of Theorem 1 of Cook and Ni (2005) and hence is omitted.

Lemma 1. *Assume that the data* $(\mathbf{Y}_i, \mathbf{X}_i)$, $i = 1, \ldots, n$, *are a simple random sample of* $(\mathbf{Y}, \mathbf{X})$ *with finite fourth moments. Then*

$$\sqrt{n}(\,\mathrm{vec}(\hat{\boldsymbol{\beta}}) - \mathrm{vec}(\boldsymbol{\rho}\boldsymbol{\gamma})) \xrightarrow{\mathcal{D}} \mathrm{Normal}(0, \boldsymbol{\Gamma}) \tag{3.5}$$

where $\boldsymbol{\Gamma} = \mathrm{Cov}(\,\mathrm{vec}(\boldsymbol{\Sigma}^{-1}\{\mathbf{X} - \mathrm{E}(\mathbf{X})\}\boldsymbol{\epsilon}^T)) \in \mathbb{R}^{(prh)\times(prh)}$.

Letting $\hat{\boldsymbol{\Gamma}}$ be a consistent estimate of $\boldsymbol{\Gamma}$, our new method is obtained by minimizing

$$F_d^{\mathrm{gm}}(\mathbf{B}, \mathbf{C}) = (\,\mathrm{vec}(\hat{\boldsymbol{\beta}}) - \mathrm{vec}(\mathbf{BC}))^T\hat{\boldsymbol{\Gamma}}^{-1}(\,\mathrm{vec}(\hat{\boldsymbol{\beta}}) - \mathrm{vec}(\mathbf{BC})), \tag{3.6}$$

which is an application of (3.4) with $\mathbf{V}_n = \hat{\boldsymbol{\Gamma}}^{-1}$. The estimate of $\mathcal{S}_{\mathbf{Y}|\mathbf{X}}$ constructed by minimizing (3.6) is called the *generalized multivariate k-means inverse regression estimation* (GM.KIRE) estimator.

Since $\mathbf{X}$ and $\boldsymbol{\epsilon}$ are uncorrelated, we can rewrite $\boldsymbol{\Gamma}$ as $\mathrm{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T \otimes \boldsymbol{\Sigma}^{-1}\{\mathbf{X} - \mathrm{E}(\mathbf{X})\}\{\mathbf{X} - \mathrm{E}(\mathbf{X})\}^T\boldsymbol{\Sigma}^{-1}]$. One choice of $\hat{\boldsymbol{\Gamma}}$ can be constructed straightforwardly by substituting sample versions for the population moments in $\boldsymbol{\Gamma}$:

$$\hat{\boldsymbol{\Gamma}} = \sum_{i=1}^{n}\frac{1}{n}\{\hat{\boldsymbol{\epsilon}}_i\hat{\boldsymbol{\epsilon}}_i^T \otimes \hat{\boldsymbol{\Sigma}}^{-1}(\mathbf{X}_i - \bar{\mathbf{X}}_\bullet)(\mathbf{X}_i - \bar{\mathbf{X}}_\bullet)^T\hat{\boldsymbol{\Sigma}}^{-1}\}.$$

Let $\boldsymbol{\Delta}_{\mathrm{gm}} \equiv (\boldsymbol{\nu}^T \otimes I_p, I_{rh} \otimes \boldsymbol{\rho})$, which is the Jacobian matrix

$$\boldsymbol{\Delta} = \left(\frac{\partial\,\mathrm{vec}(\mathbf{BC})}{\partial\,\mathrm{vec}(\mathbf{B})}, \frac{\partial\,\mathrm{vec}(\mathbf{BC})}{\partial\,\mathrm{vec}(\mathbf{C})}\right)$$

evaluated at $(\mathbf{B} = \boldsymbol{\rho}, \mathbf{C} = \boldsymbol{\nu})$. The associated asymptotic properties of GM.KIRE are given in the following theorem. The proof is structurally similar to that of Theorem 2 in Cook and Ni (2005) and is given in the appendix.

Theorem 1. *Assume that the data* $(\mathbf{Y}_i, \mathbf{X}_i)$, $i = 1, \ldots, n$, *are a simple random sample of* $(\mathbf{Y}, \mathbf{X})$ *with finite fourth moments. Let* $\hat{F}_d^{\mathrm{gm}}$ *be the minimum value of* (3.6), *and let*

$$(\hat{\boldsymbol{\rho}}, \hat{\boldsymbol{\nu}}) = \arg_{\mathbf{B},\mathbf{C}}\min F_d^{\mathrm{gm}}(\mathbf{B}, \mathbf{C}).$$

Then,

1. *The estimate* $\mathrm{vec}(\hat{\boldsymbol{\rho}}\hat{\boldsymbol{\nu}})$ *is asymptotically efficient, and*

$$\sqrt{n}(\,\mathrm{vec}(\hat{\boldsymbol{\rho}}\hat{\boldsymbol{\nu}}) - \mathrm{vec}(\boldsymbol{\rho}\boldsymbol{\nu})) \xrightarrow{\mathcal{D}} \mathrm{Normal}(0, \boldsymbol{\Delta}_{\mathrm{gm}}(\boldsymbol{\Delta}_{\mathrm{gm}}^T\boldsymbol{\Gamma}_{\mathrm{gm}}^{-1}\boldsymbol{\Delta}_{\mathrm{gm}})^-\boldsymbol{\Delta}_{\mathrm{gm}}^T).$$

2. $n\hat{F}_d^{\mathrm{gm}}$ *has an asymptotic chi-squared distribution with degrees of freedom* $(p - d)(rh - d)$.

Theorem 1 provides a basis for inference about $\mathcal{S}_{\mathbf{Y}|\mathbf{X}}$. In particular, the second conclusion can be used to test the hypothesis $d = m$ vs. $d > m$, rejecting if $n\hat{F}_m^{\mathrm{gm}}$ exceeds a selected quantile of the chi-squared distribution with $(p - m)(rh - m)$ degrees of freedom. A useful property of GM.KIRE is that $n\hat{F}_d^{\mathrm{gm}}$ follows an asymptotic chi-squared distribution, while the corresponding statistics for KIR is distributed as a linear combination of independent chi-squared random variables. Asymptotic efficiency means that the estimator has minimum asymptotic variance among all choices of the possible $\mathbf{V}_n$'s in (3.4).

3.2 Clustering

Since clustering the multivariate responses plays the same role as slicing a univariate response, previous research and

comments on the impact of slicing on the SDR methods applies to clustering too. How to select h optimally is an open question in SDR research.

Zhu and Ng (1995) and Li and Zhu (2007) studied the asymptotic behavior of SIR and SAVE (sliced average variance estimation; another well-known method in SDR) respectively. They showed that SIR is relatively insensitive to the choice of the total number of slices (clusters) h, under some regularity conditions, the asymptotic normality of SIR holds with h ranging from $\sqrt{n}$ to $\frac{n}{2}$; while SAVE is much more sensitive to h. Li and Zhu (2007) proved that when the response Y is continuous, SAVE cannot be $\sqrt{n}$ consistent even when each slice contains a fixed number of data points that do not depend on the sample size n.

Empirical studies suggested that the number of slices (clusters) h is a tuning parameter much like the tuning parameter encountered in the smoothing literature (Li 1987; Härdle et al., 1988). As Cook and Forzani (2009) concluded: "h doesn't matter much, provided that it is large enough to allow estimation of d and that there are sufficient observations per slice to estimate the intraslice parameters ...". Our experience indicates that good results are often obtained by choosing h to be somewhat larger than $d+1$, trying a few different values of h as necessary. Choosing h very much larger than d should generally be avoided since small sample performance of the data may not agree with the asymptotic findings with few observations within each slice. Wang and Xia (2008) gave the same advice.

3.3 Computing and generalized inverse

The alternating least squares algorithm for inverse regression estimation (Cook and Ni, 2005; Ruhe and Wedin, 1980) can be adapted for the minimization of $\hat{F}_d^{\mathrm{gm}}$. Since the form of the inner-product matrix in (3.6) depends on $\mathcal{S}_{\mathbf{Y}|\mathbf{X}}$, we could adapt the iterative algorithm to reduce the variability of this inner-product matrix. Here is a sketch of this idea. First, obtain $\mathrm{Span}(\hat{\boldsymbol{\rho}})$, an estimate of $\mathcal{S}_{\mathbf{Y}|\mathbf{X}}$, via the alternating least squares method. Second, update the inner-product matrix using $\mathrm{Span}(\hat{\boldsymbol{\rho}})$. Then, re-run the alternating least squares algorithm to update $\mathrm{Span}(\hat{\boldsymbol{\rho}})$ applying this new inner-product matrix. Carroll and Ruppert (1988) recommended at least two cycles but we suggest the use of at least a three-cycle iterative computation algorithm for GM.KIRE.

The inverse of the inner-product matrix $\hat{\boldsymbol{\Gamma}} \in \mathbb{R}^{prh\times prh}$ became very unstable when the total number of clusters (slices) and the dimensions of the responses (predictors) are too big, and the consequent computation results could be time-consuming and misleading. Instead, we use a Moore-Penrose generalized inverse of $\hat{\boldsymbol{\Gamma}}$ to replace the regular inverse matrix in our algorithm. Both simulation studies and real-data analysis presented in Section 5 suggest that this approach works well and deserves further studies for all sufficent dimension reduction methods taking the minimum discrepancy approaches.

4. TESTING PREDICTOR EFFECTS

GM.KIRE is a new dimension reduction method to deal with multivariate responses. It follows the minimum discrepancy approach introduced by Cook and Ni (2005). In addition to provide more accurate estimates for both the central subspace and its dimension, the quadratic objective functions used in the minimum discrepancy approach we adopted for GM.KIRE, enables us to test conditional independence hypotheses. We consider testing hypotheses of the forms: $\mathbf{Y} \perp\!\!\!\perp P_{\mathcal{H}}\mathbf{X} \mid (Q_{\mathcal{H}}\mathbf{X})$, where $\mathcal{H}$ is an l-dimensional user-selected subspace of the predictor space, $Q_{\mathcal{H}} = I_p - P_{\mathcal{H}}$. This is equivalent to testing hypotheses $P_{\mathcal{H}}\mathcal{S}_{\mathbf{Y}|\mathbf{X}} = \mathcal{O}_p$ under the usual linearity and coverage conditions (Cook and Ni, 2005).

Since the marginal predictor hypothesis $P_{\mathcal{H}}\mathcal{S}_{\mathbf{Y}|\mathbf{X}} = \mathcal{O}_p$ does not require specification of d, equivalently, we test the hypothesis $\boldsymbol{H}^T\boldsymbol{\beta} = 0$, where $\boldsymbol{H} \in \mathbb{R}^{p\times l}$ is an orthonormal basis for $\mathcal{H}$. It follows from Theorem 1 that a Wald test statistic of the form

$$T^{\mathrm{bm}}(\mathcal{H}) = n\,\mathrm{vec}(\boldsymbol{H}^T\hat{\boldsymbol{\beta}})^T\{(I_{rh}\otimes\boldsymbol{H}^T)\hat{\boldsymbol{\Gamma}}_{\mathrm{gm}}(I_{rh}\otimes\boldsymbol{H})\}^{-1} \times \mathrm{vec}(\boldsymbol{H}^T\hat{\boldsymbol{\beta}}) \tag{4.7}$$

can be used to test a marginal predictor hypothesis. Following Theorem 1 and Slutsky's theorem, it can be shown that, under the hypothesis, (4.7) is distributed asymptotically as a chi-squared random variable with degrees of freedom of $l(rh)$.

Let $\boldsymbol{H}_0 \in \mathbb{R}^{p\times(p-l)}$ be an orthonormal basis for $\mathrm{Span}(Q_{\mathcal{H}})$, then the joint dimension-predictor hypothesis $P_{\mathcal{H}}\mathcal{S}_{\boldsymbol{\beta}} = \mathcal{O}_p$ and $d = m$, is equivalent to $\boldsymbol{\beta} = Q_{\mathcal{H}}\boldsymbol{\beta} = Q_{\mathcal{H}}\boldsymbol{\rho}\boldsymbol{\nu} = \boldsymbol{H}_0\boldsymbol{\rho}_0\boldsymbol{\nu}$, where $\boldsymbol{\rho}_0$ contains the coordinates of $\boldsymbol{\rho}$ represented in terms of the basis $\boldsymbol{H}_0$. We then can fit under the joint hypothesis by minimizing the following constrained optimal discrepancy function

$$F_{m,H}^{\mathrm{gm}}(\mathbf{B},\mathbf{C}) = (\,\mathrm{vec}(\hat{\boldsymbol{\beta}}) - \mathrm{vec}(\boldsymbol{H}_0\mathbf{B}\mathbf{C}))^T \times \hat{\boldsymbol{\Gamma}}_{\mathrm{gm}}^{-1}(\,\mathrm{vec}(\hat{\boldsymbol{\beta}}) - \mathrm{vec}(\boldsymbol{H}_0\mathbf{B}\mathbf{C})) \tag{4.8}$$

over $\mathbf{B} \in \mathbb{R}^{(p-l)\times m}$ and $\mathbf{C} \in \mathbb{R}^{m\times(rh)}$. Values of $\mathbf{B}$ and $\mathbf{C}$ that minimize (4.8) provide estimates of $\boldsymbol{\rho}_0$ and $\boldsymbol{\nu}$. Following the result of Theorem 1, we know that, under the null hypothesis, the statistic $n\hat{F}_{d,H}^{\mathrm{gm}}$ has an asymptotic chi-squared distribution with degrees of freedom $(p-d)(rh-d)+dl$.

Finally, when d is specified, or when inference on d using marginal dimension tests results in a firm estimate, we might consider the conditional hypothesis $P_{\mathcal{H}}\mathcal{S}_{\boldsymbol{\beta}} = \mathcal{O}_p$ given d. The difference in minimum discrepancies

$$T^{\mathrm{bc}}(\mathcal{H}|d) = n\hat{F}_{d,H}^{\mathrm{gm}} - n\hat{F}_d^{\mathrm{gm}} \tag{4.9}$$

is used to test a conditional predictor hypothesis. Under the null hypothesis, $T^{\mathrm{bc}}(\mathcal{H}|d)$ has an asymptotic chi-squared

distribution with degrees of freedom ld. It can be shown that the conditional predictor test statistic and the marginal dimension statistic are asymptotically independent (interested readers could refer to Wen and Cook (2007) for a more detailed discussion).

5. SIMULATION RESULTS AND DATA ANALYSIS

In this section, we report simulation results to support our theoretical conclusions regarding GM.KIRE. The performance of GM.KIRE and KIR were compared regarding estimation accuracies and actual testing levels.

5.0.0.1. Inverse model

5.1 Estimation accuracy

We now consider an inverse regression model:

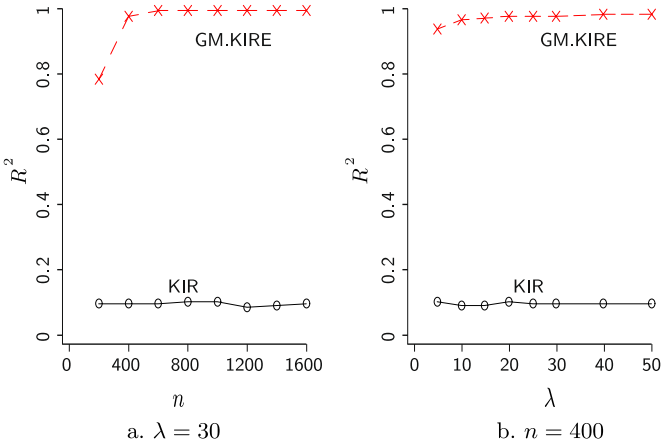
$$\begin{aligned} Y_1 &= \lambda(Z_1 + Z_2) + (Z_1 - Z_2) \\ Y_2 &= \lambda(Z_1 + Z_2) - (Z_1 - Z_2), \\ Y_3 &= \frac{1}{Y_1} + Y_2 + 0.1 * Z_3, \\ Y_4 &= \frac{1}{Y_2} + Y_1 + 0.1 * Z_4 \end{aligned}$$

$\mathbf{X} \in \mathbb{R}^{10}$ and $\mathbf{X} = \boldsymbol{\alpha}(Y_1 - Y_2) + \boldsymbol{\epsilon}$, where Z_i, $i = 1, \ldots, 4$ are independent standard normal random variates, $\boldsymbol{\alpha} = (0, \ldots, 0, 1)^T \in \mathbb{R}^{10}$, $\boldsymbol{\epsilon} \perp\!\!\!\perp (Z_1, \ldots, Z_4)^T$ is a 10-dimensional standard normal vector, $\lambda \in \mathbb{R}^1$ is a constant. Letting $\mathbf{Y} = (Y_1, Y_2, Y_3, Y_4)$, we now have $\mathcal{S}_{\mathbf{Y}|\mathbf{X}} = \mathcal{S}_{\boldsymbol{\alpha}}$. In this model, the responses are strongly correlated to each other, and the noises become bigger with the increment of λ. A similar model was discussed by Wen and Cook (2008).

We used $h = 6$ clusters for all simulation runs. The estimation accuracy of the central subspace is measured by the R^2 values from the regression of X_{10} on the first estimated sufficient predictor. As shown in Figure 1a, GM.KIRE definitely won over KIR with $\lambda = 30$. For example, when $n = 600$, the average of R^2 from $1,000$ replications was 0.997 from GM.KIRE, and 0.097 from KIR. Also, the R^2 from GM.KIRE exceeded the R^2 from KIR 99.9% of the time. Shown in Figure 1b are the results from $1,000$ simulation runs at various values of λ with fixed sample size $n = 400$. As we can see, the changes in λ did not affect the performances of GM.KIRE much. It always shows strong advantages over KIR: the average R^2's are always greater than 0.939 for GM.KIRE; while they are about 0.1 for KIR.

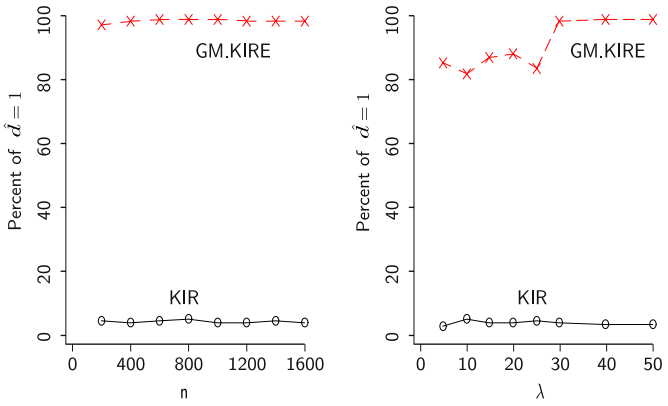
We then compared the performances of GM.KIRE and KIR concerning the estimation of d, which is the dimension of the central subspace. The standard approach in SDR is to test a sequence of hypotheses $H_0 : d = m$ versus $H_a : d > m$, with m incremented by 1 until the hypothesis is not rejected. At which point $\hat{d}$ is the last value of m tested.

The percentages of correct estimates $\hat{d} = 1$ from $1,000$ replications versus varying sample sizes and λ at testing



a. $\lambda = 30$ b. $n = 400$

Figure 1. Estimation accuracy: average R^2 versus n and λ.



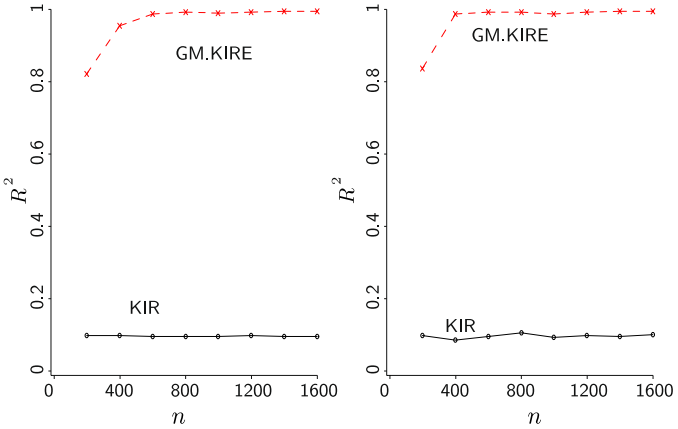
a. $\lambda = 30$ b. λ with $n = 400$

Figure 2. Percentage of runs in which $\hat{d} = 2$ versus n and λ.

level $\alpha = 0.05$ are also studied. With $\lambda = 30$, Figure 2a shows clear advantages of GM.KIRE over KIR. Regardless the sample size, GM.KIRE always made over 95% correct decisions, while the performance of KIR was much worse. As shown in Figure 2b, GM.KIRE also performed much better than KIR at various values of λ with fixed sample size $n = 400$. It responded much faster to the increment of λ.

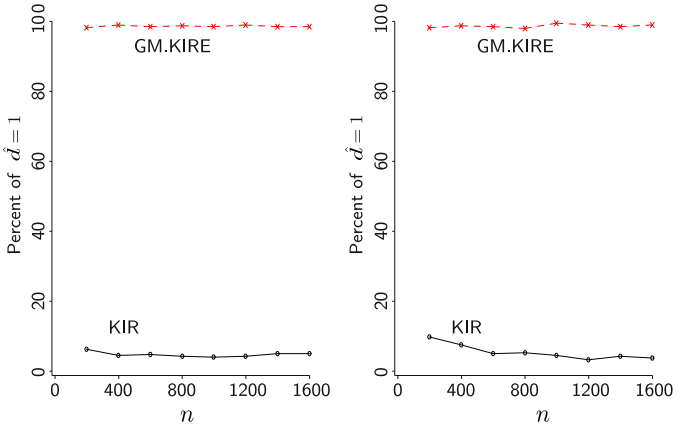
Figures 3 and 4 showed the performances of GM.KIRE and KIR with different clusters. As we can see, both GM.KIRE and KIR are quite robust to the change of clusters when h lies in our recommended range (say from 4 to 10), though KIR does perform a little bit better with more clusters when a sample size is small. Overall, our simulation results suggest that there can be substantial differences between the GM.KIRE and KIR estimators of d, and GM.KIRE outperformed KIR all the time. There are clear advantages due to the incorporation of the intra-cluster information.

To show the advantage of analyzing the multivariate response simultaneously over pooling marginal dimension reduction estimates, we also considered four univariate dimen-



a. $\lambda = 30$, $h = 4$ b. $\lambda = 30$, $h = 8$

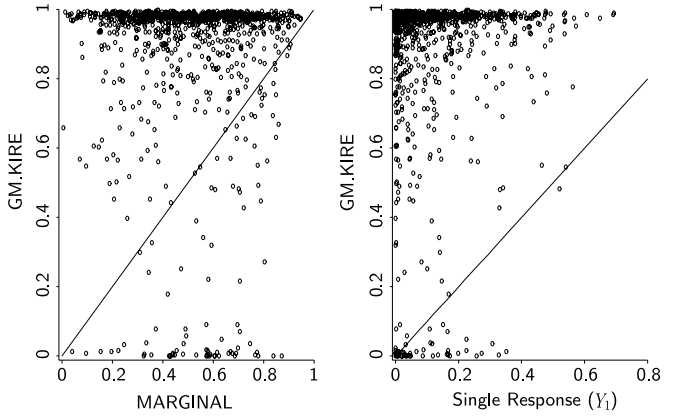
Figure 3. Estimation accuracy: average R^2 versus n with 4 and 8 clusters.



a. $\lambda = 30$, $h = 4$ b. $\lambda = 30$, $h = 8$

Figure 4. Percentage of runs in which $\hat{d} = 2$ versus n with 4 and 8 clusters.

sion reduction for regressions $Y_1|\mathbf{X}$, $Y_2|\mathbf{X}$, $Y_3|\mathbf{X}$ and $Y_4|\mathbf{X}$. The method of the covariance inverse regression estimator (CIRE, Cook and Ni, 2007) was used to obtain the marginal central subspaces. Figure 5 (a) showed the scatterplot of the R^2 values from the regression of X_{10} on the first estimated sufficient predictor from our method versus those from the regression of X_{10} on the first estimated sufficient predictors from the four marginal dimension reduction models. With $n = 600$ and $\lambda = 30$, our method won over the pooled CIRE method about 90.3% of the times, which is reasonable since the method pooling marginal estimates ignored the information contained in the relation between $Y_i|\mathbf{X}$ and $Y_j|\mathbf{X}$ when $i \neq j$. Figure 5 (b) compared our method with the marginal univariate dimension reduction estimate when Y_1 was the single response, where our method gave a better result about 96% of the times. Similar results were obtained with the other three marginal univariate dimension reduction estimates.



a. Pooling CIRE b. CIRE with Y_1 as response.

Figure 5. Estimation accuracy: R^2 from 1,000 simulation runs with $n = 600$, $\lambda = 30$ and 6 clusters.

Table 1. P-values of dimension tests for Minneapolis School Data.

	$d = 0$	$d = 1$	$d = 2$
GM.KIRE	0	.791	1.00
KIR	0	.038	.628

5.2 Data analysis

Following Cook and Setodji (2003) and Yin and Bura (2006), we use data on the performance of students in $n = 63$ Minneapolis Schools to illustrate our methodology. Our response consists of $r = 4$ percentages of students in a school scoring above (A) and below (B) average on standardized fourth and sixth grade reading comprehension tests, $\mathbf{Y} = (P_{A4}, P_{B4}, P_{A6}, P_{B6})^T$). Eight predictors were used to characterize a school: (1) the pupil teacher ratio (PT), (2) the percentage of children receiving Aid to Families with Dependent Children (AFDC), (3) the percentage of children not living with both biological parents (BP), (4) the percentage of adults in the school area who completed high school (HS), (5) the percentage of persons in the area below the federal poverty level (PL), (6) the percentage of minority students (MIN), (7) the percentage of mobility (MOB), and (8) the percentage of students attending school regularly (ATT).

As suggested by Yin and Bura (2006), the square root of all percentages predictors were used to remove the effects of possible outlying points or heteroskedasticity. We then applied our method (GM.KIRE) and KIR to the multivariate regression of $\mathbf{Y}$ versus the 8-dimensional predictor $\mathbf{X}$. We used four clusters. Table 1 provides the p-values of the asymptotic tests for $d = 0, 1$ and 2 using our method and KIR(Setodji and Cook, 2004). GM.KIRE inferred that the multivariate central subspace is 1-dimensional comparing to the 2-dimensional subspace that KIR inferred. Figure 6 suggests that the relationship between the second estimated direction from KIR and the four responses

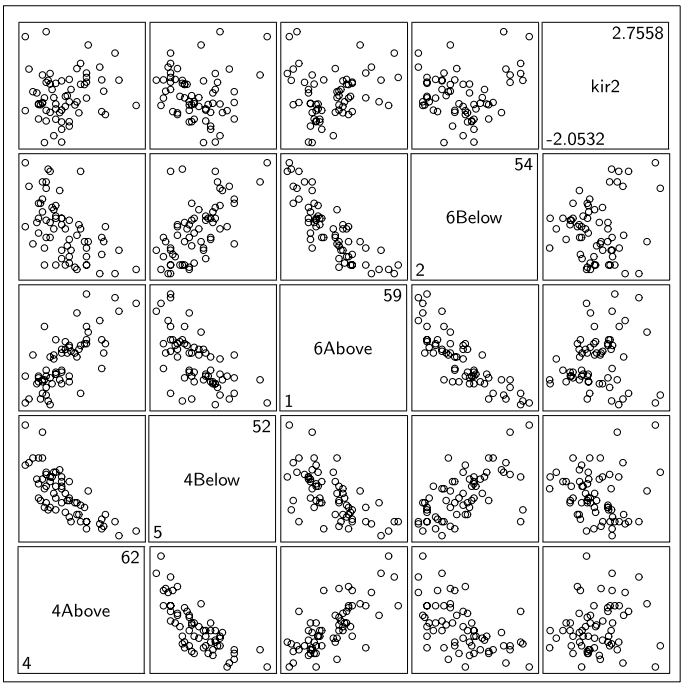


Figure 6. Scatterplot matrix of $\mathbf{Y}$ *versus the second KIR predictor.*

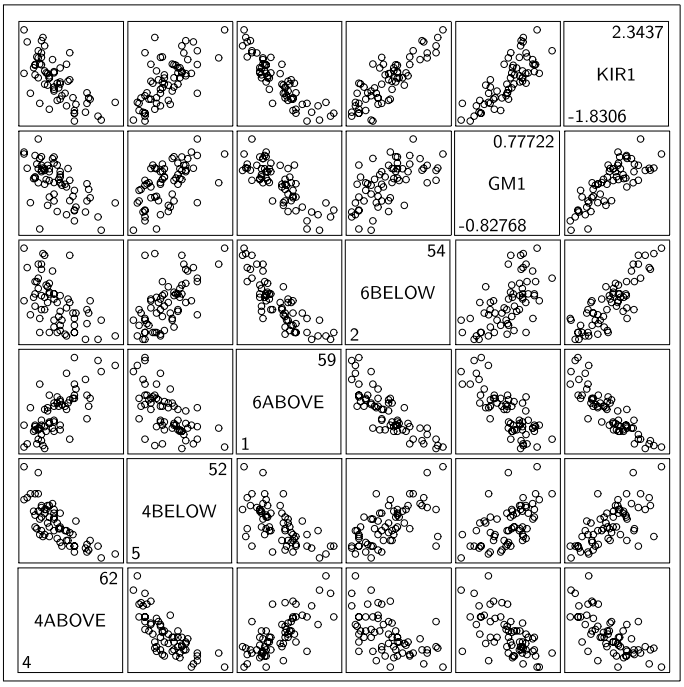


Figure 7. Scatterplot matrix of $\mathbf{Y}$ *versus the first KIR and GM.KIRE predictors.*

are relatively weak. So we expect that a second direction might be spurious. Instead, Figure 7 showed that both first estimated directions from the two methods had a strong relationship with the four responses. We then concluded that the multivariate central subspace is one-dimensional.

An important advantage of our proposed method is that it allows us to test the predictor effects without assuming a model or specifying the structural dimension. For this example, we used marginal predictor tests as the basis of a model-free backward elimination procedure (Li et al., 2005). As shown in Table 2, pupil teacher ratio and poverty level were screened out using 5% tests, leaving 6 predictors for further analysis. We next re-estimated the dimension of the regression based on these six predictors. Using GM.KIRE, we again inferred that only one direction is required.

Table 2. P-values from marginal predictor tests for Minneapolis School data

Predictor	Step 1	Step 2	Step 3
AFDC	0.000	0.000	0.000
Attend	0.002	0.000	0.000
BP	0.001	0.000	0.000
HS	0.000	0.000	0.000
Minority	0.000	0.000	0.000
Mobility	0.000	0.002	0.000
Poverty	0.272	0.583	deleted
PT-ratio	0.627	deleted	

6. DISCUSSION

We proposed a new dimension reduction method for regressions with multivariate responses. Our method is designed to recover the intra-cluster information when the responses are continuous. Simulations showed that our proposed method is superior to those based on SIR because it provide better estimates of the structural dimension of the regression when one has large enough sample size, improves the estimation accuracy of the central subspace, and allows prior screening of predictors.

When the conditional mean function $\mathrm{E}(\mathbf{Y}|\mathbf{X})$ is of special interest, the inquiry of SDR is restricted to the *central mean subspace*, the intersection of all subspaces $\mathcal{S}$ satisfying $\mathbf{Y} \perp\!\!\!\perp \mathrm{E}(\mathbf{Y}|\mathbf{X})|P_{\mathcal{S}}\mathbf{X}$. Or equivalently, the intersection of all subspaces $\mathcal{S}$ satisfying the conditional independent condition

$$\mathrm{E}(\mathbf{Y}|\mathbf{X}) \perp\!\!\!\perp \mathbf{X}|P_{\mathcal{S}}\mathbf{X}.$$

Cook and Li (2002) investigated possible approaches to inferring about the central mean subspace when Y is univariate. Cook and Setodji (2003), Li et al. (2003), Yoo and Cook (2007) studied estimation methods for the multivariate central mean subspace. Their works are not applicable to the estimation of the central subspace and are beyond the scope of this paper.

APPENDIX

Proof of Theorem 1. The proof of Theorem 1 hinges on Shapiro's (1986) results on asymptotics of over-parameterized discrepancy functions and two supplemental lemmas (Cook and Ni, 2005). The discrepancy functions that Shaprio considered are

$$H(\boldsymbol{\tau}_n, g(\boldsymbol{\theta})) = (\boldsymbol{\tau}_n - g(\boldsymbol{\theta}))^T \mathbf{V} (\boldsymbol{\tau}_n - g(\boldsymbol{\theta})),$$

where $\boldsymbol{\tau}_n$ is an asymptotically normal estimate of the population value $g(\boldsymbol{\theta}_0)$, and $\mathbf{V}$ is a known inner product matrix.

The following setting makes it clear that $H_d(\mathbf{B}, \mathbf{C})$ is in the form of Shapiro's discrepancy function H:

$$\begin{aligned} \boldsymbol{\theta} &= \begin{pmatrix} \mathrm{vec}(\mathbf{B}) \\ \mathrm{vec}(\mathbf{C}) \end{pmatrix} \in \mathbb{R}^{d\times(p+rh)} \\ g(\boldsymbol{\theta}) &= \mathrm{vec}(\mathbf{BC}) \in \mathbb{R}^{p\times rh} \\ \boldsymbol{\tau}_n &= \mathrm{vec}(\hat{\boldsymbol{\beta}}) \\ g(\boldsymbol{\theta}_0) &= \mathrm{vec}(\boldsymbol{\rho}\boldsymbol{\nu}) \end{aligned}$$

where $\boldsymbol{\rho} \in \mathbb{R}^{p\times d}$ is in general a basis for $\mathcal{S}_{\boldsymbol{\xi}}$ and $\boldsymbol{\nu} \in \mathbb{R}^{d\times rh}$. Following from Shapiro (1986), we then have $\mathrm{vec}(\hat{\boldsymbol{\rho}}\hat{\boldsymbol{\nu}})$ of $H_d(\mathbf{B}, \mathbf{C})$ is asymptotically efficient with

$$\begin{aligned} &\sqrt{n}(\,\mathrm{vec}(\hat{\boldsymbol{\rho}}\hat{\boldsymbol{\nu}}) - \,\mathrm{vec}(\boldsymbol{\beta}\boldsymbol{\nu})) \\ &\quad \xrightarrow{\mathcal{D}} \mathrm{Normal}(0, \boldsymbol{\Delta}_{\mathrm{gm}}(\boldsymbol{\Delta}_{\mathrm{gm}}^T \mathbf{V} \boldsymbol{\Delta}_{\mathrm{gm}})^- \boldsymbol{\Delta}_{\mathrm{gm}}), \end{aligned}$$

which leads to the conclusion 1 of Theorem 1. And $n\hat{H}$ has an asymptotic chi-squared distribution with degrees of freedom $p \times (rh) - \mathrm{rank}(\boldsymbol{\Delta}_{\mathrm{gm}})$, where

$$\begin{aligned} \mathrm{rank}(\boldsymbol{\Delta}_{\mathrm{gm}}) &= \mathrm{rank}(\boldsymbol{\nu}^T \otimes \mathbf{Q}_{\boldsymbol{\beta}}, I_{rh} \otimes \boldsymbol{\beta}) \\ &= d \times (p-d) + d \times (rh) \\ &= d(p + rh - d). \end{aligned}$$

Therefore, the degrees of freedom are $p\times(rh) - d(p+rh-d) = (p-d)(rh-d)$. Thus, conclusion 2 is proved. □

Received 14 November 2008

REFERENCES

Aragon, Y. (1997). A Gauss implementation of multivariate sliced inverse regression. Computational Statistics, **12**, 355–372.

Carroll, R. and Ruppert, D. (1988). Transformation and Weighting in Regression. Chapman and Hall.

Cook, R. D. (1994). Using dimension-reduction subspaces to identify important inputs in models of physical systems. In Proceedings of Section on Physical and Engineering Sciences, pp. 18–25. Alexandria, VA: American Statistical Association.

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

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

Cook, R. D. and Setodji, C. M. (2003). A model-free test for reduced rank in multivariate regression. Journal of the American Statistical Association, **98**, 340–351.

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

Cook, R. D. and Ni, L. (2006). Using intra slice covariances for improved estimation of the central subspace in regression. Biometrika, **93**, 65–74.

Cook, R. D. and Forzani, L. (2009). Likelihood-based sufficient dimension reduction. Journal of the American Statistical Association, **104**, 197–208.

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

Furgerson, T. (1958). A method of generating best asymptotically normal estimates with application to the estimation of bacterial densities. Annals of Mathematical Statistics, **29**, 1046–1062.

Fung, W. K., He, X., Liu, L. and Shi, P. D. (2002). Dimension reduction based on canonical correlations. Statistica Sinica, **12**, 1093–1114.

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

Härdle, W., Hall, P. and Marron, S. (1988). How far are automatically chosen regression smoothing parameters from their optimum. Journal of the American Statistical Association, **83**, 86–101.

Hartigan, J. A. (1975). Clustering Algorithms. Wiley, New York.

Helland, I. S. (1989). On the structure of partial least squares regression. Communications in Statistics, B, **17**, 581–607.

Helland, I. S. (1990). Partial least squares regression and statistical models. Scandinavian Journal of Statistics, **17**, 97–114.

Hotelling, H. (1935). The most predictable criterion. Journal of Educational Psychology, **26**, 139–142.

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

Li, L., Cook, R. D. and Nachtsheim, C. J. (2005). Model-free variable selection. Journal of the Royal Statistical Society, Ser. B **67**, 285–300.

Li, B., and Wang, S. (2007) On Directional Regression for Dimension Reduction. Journal of the American Statistical Association, **102**, 997–1008.

Li, K. C. (1987). Asymptotic optimality for C_p, C_L, cross-validation and generalized cross-validation: discrete index set. The Annals of Statistics, **15**, 958–975.

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

Li, K. C., Aragon, Y., Shedden, K. and Agnan, C. T. (2003). Dimension reduction for multivariate response data. Journal of the American Statistical Association, **98**, 99–109.

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

Li, Y. and Zhu, L. X. (1995). Asymptotics for sliced average variance estimation. Annals of Statistics, **35**, 41–69.

Massy, W. (1965). Principal components regression with exploratory statistical research. Journal of the American Statistical Association, **60**, 234–246.

Ruhe, A. and Wedin P. A. (1980). Algorithms for separable nonlinear least squares problems. SIAM Review, **22**, 318–337.

Setodji, C. M. and Cook, R. D. (2004). K-means inverse regression. Technometrics , **46**, 421–429.

Shapiro, A. (1986). Asymptotic theory of overparameterized structural model. Journal of the American Statistical Association, **81**, 142–149.

Wang, H. and Xia, Y. (2008). Sliced regression for dimension reduction. Journal of the American Statistical Association, **103**, 811–821.

Wen, X. and Cook, R. D. (2007). Optimal sufficient dimension reduction in regressions with categorical predictors. Journal of Statistical Planning and Inference, **137**, 1961–1978.

Wen, X. and Cook, R. D. (2009). New Approaches to Model-free Dimension Reduction for Bivariate Regression. Journal of Statistical Planning and Inference, **139**, 734–748.

Yin, X. and Bura, E. (2006). Moment-based dimension reduction for multivariate response regression. Journal of Statistical Planning and Inference, **136**, 3675-3688.
Yoo, P. and Cook, R.D. (2007). Optimal sufficient dimension reduction for the conditional mean in multivariate regression. Biometrika, **94**, 231–242.
Zhu, L. X. and Fang, K. T. (1996). Asymptotics for kernel estimate of sliced inverse regression. The Annals of Statistics, **24**, 1053–1068.
Zhou, J. and He, X. (2008). Dimension reduction based on constrained canonical correlation and variable filtering. Annals of Statistics, **36**, 1649–1668.
Zhu, L. X. and Ng, K. W. (1995). Asymptotics of sliced inverse regression. Statistica Sinica, **5**, 727–736.
Zhu, L. X., Ohtaki, M. and Li, Y. (2007). On hybrid methods of inverse regression-based algorithms. Computational Statistics and Data Analysis, **51**, 2621–2635.
Zhu, L. P. and Zhu, L. X. (2007). On kernel method for sliced average variance estimation. Journal of Multivariate Analysis, **98**, 970–991.

Xuerong Meggie Wen
Department of Mathematics and Statistics
Missouri University of Science and Technology
MO 65409, U.S.A.
E-mail address: wenx@mst.edu

C. Messan Setodji
The Rand Coorporation, Pittsburgh
PA 15213, U.S.A.

Akim Adekpedjou
Department of Mathematics and Statistics
Missouri University of Science and Technology
MO 65409, U.S.A.