Statistica Sinica **19** (2009), 1119-1136

MODEL FREE MULTIVARIATE REDUCED-RANK REGRESSION WITH CATEGORICAL PREDICTORS

Claude Messan Setodji and Lexin Li

RAND Corporation and North Carolina State University

Abstract: Cook and Setodji (2003) introduced the notion of model-free reduced-rank in multivariate regression. However, they only focused on continuous predictors. In this article, we propose an extension of model-free multivariate reduced-rank regression to incorporate a mixture of continuous and categorical predictors. A test for reduced-rank is proposed that requires no parametric model specification. Simulations and a data analysis are provided to demonstrate its effectiveness.

Key words and phrases: Central mean subspace, dimension reduction, multivariate reduced-rank regression.

1. Introduction

Consider a multivariate response $\boldsymbol{Y} \in \mathbb{R}^r$, a vector $\boldsymbol{X} \in \mathbb{R}^p$ of p continuous predictors, and a vector $\boldsymbol{W} \in \mathbb{R}^q$ of q categorical predictors. Regression analysis hinges on describing the conditional distribution of $\boldsymbol{Y}|(\boldsymbol{X}, \boldsymbol{W})$ by a parsimonious parametric model, whereas the attention is often concentrated on the conditional mean function $\mathrm{E}(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{W})$. Commonly used is the classical multivariate linear model of the form

$$\boldsymbol{Y} = \boldsymbol{\alpha} + \boldsymbol{B}^\tau \boldsymbol{X} + \boldsymbol{C}^\tau \boldsymbol{W} + \boldsymbol{\delta} \tag{1.1}$$

where $\boldsymbol{\alpha}$ is an $r \times 1$ vector of intercepts, $\boldsymbol{B}$ and $\boldsymbol{C}$ are $p \times r$ and $q \times r$ matrices of unknown regression coefficients, and $\boldsymbol{\delta}$ is an error vector that is independent of the predictors $(\boldsymbol{X}, \boldsymbol{W})$ with zero mean and constant variance.

When $\mathrm{rank}(\boldsymbol{B}^\tau, \boldsymbol{C}^\tau) < \min(r, p+q)$, (1.1) is called a multivariate reduced-rank linear model, since only a reduced number of linear combinations of the predictors is needed to convey full information about $\mathrm{E}(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{W})$. Such reduction is often appealing from both estimation and interpretation points of view, especially when there are a fair number of predictors under examination. Multivariate reduced-rank linear model has been studied extensively by Anderson (1951), Izenman (1975) and Reinsel and Velu (1998), among others. In particular, Anderson (1951) considered a version of (1.1), where $\boldsymbol{X}$ has a reduced-rank coefficient matrix $\boldsymbol{B}$, while $\boldsymbol{W}$ has a full rank coefficient matrix $\boldsymbol{C}$, and the reduction is only possible for $\boldsymbol{X}$.

A linear model such as (1.1) may not always be appropriate, since the true functional predictor-response relation may not be known a priori, and the response may depend on the predictors in a complicated manner. Moreover, (1.1) is restricted by the assumption of no interaction between the continuous and categorical predictors. Cook and Setodji (2003) proposed a test for the multivariate reduced-rank regression without positing any parametric model. However, their method focused only on the case where all predictors are continuous. Li, Cook and Chiaromonte (2003) considered a mixture of continuous and categorical predictors, but they only dealt with a univariate response. When the components of the multivariate response are correlated, a univariate analysis of individual responses is often less informative than a multivariate treatment.

The ability to handle both a multivariate response and a mixture of predictors is often desirable in applications. One example is the Berkeley Guidance Study. Tuddenham and Snyder (1954) investigated the physical growth of 136 white children, 66 boys and 70 girls, born in Berkeley, California. Weight (WT) and height (HT) were measured for all children at years 2, 9, and 18, while leg circumference (LG) and strength (ST) were measured at years 9 and 18. In the following, we denote a measurement V taken at age t by V_t. To study the effect of early life trajectories on aging, researchers were interested in modelling the weight and height of a child at age 18 based on the measurements taken at earlier years, including HT_2, HT_9, WT_2, WT_9, LG_9, and ST_9. The two response variables, (HT_{18}, ST_{18}), were found to be highly correlated, a correlation of 0.7, suggesting that a multivariate analysis might provide better inference than a univariate analysis of the two responses separately. Moreover there is a natural expectation of a difference in the growth of boys and girls; then it would be important to include gender as a predictor in addition to other predictors.

In this article we extend Cook and Setodji (2003) and Li, Cook and Chiaromonte (2003) to develop a multivariate reduced-rank regression with a mixture of continuous and categorical predictors, meanwhile imposing no parametric model such as (1.1). The proposed method could potentially be useful in the exploratory stage of analysis prior to model specification. It could also lead to construction of low-dimensional summary plots that contain all regression information, thus facilitating the subsequent model specification. In this article we use a univariate $W \in \{1, \ldots, c\}$ to denote the categorical variable. It can represent a single qualitative predictor like gender, or a combination of several qualitative predictors like gender and race; any multivariate categorical variable can be reparameterized into this form. For this reason, dimension reduction here is focused on the continuous predictors only. The rest of the article is organized as follows. Section 2 introduces the concept of the central partial mean subspace that will serve as a basis for our methodology development. A model-free reduced-rank regression

estimation and a dimension test with mixture of predictors based on ordinary least square (OLS) are developed in Section 3. An application to the Berkeley Guidance Study is presented in Section 4, and simulations are presented in Section 5. Moreover, we discuss extension of the proposed methodology to other estimators of the central mean subspace in Section 6, and conclude the paper in Section 7. All technical proofs are relegated to an online supplementary appendix available at http://www.stat.sinica.edu.tw/statistica.

2. Central Partial Mean Subspace

2.1. Multivariate central partial mean subspace

When the focus of a regression analysis is only on the conditional mean $\mathrm{E}(\boldsymbol{Y}|\boldsymbol{X})$, Cook and Li (2002) developed the notion of the central mean subspace, denoted by $\mathcal{S}_{\mathrm{E}(\boldsymbol{Y}|\boldsymbol{X})}$, which is defined as the intersection of all subspaces $\mathcal{S}$ in $\mathbb{R}^p$ satisfying that $\boldsymbol{Y} \perp\!\!\!\perp \mathrm{E}(\boldsymbol{Y}|\boldsymbol{X})|P_{\mathcal{S}}X$, where $\perp\!\!\!\perp$ denotes independence and $P_{\mathcal{S}}$ is the orthogonal projection onto $\mathcal{S}$. Such a subspace uniquely exists under minor conditions, and is the minimum subspace that contains all information about $\boldsymbol{Y}$ that is available through $\mathrm{E}(\boldsymbol{Y}|\boldsymbol{X})$. Cook and Setodji (2003) extended the concept of the central mean subspace to a multivariate response, whereas Li, Cook and Chiaromonte (2003) extended it to incorporate a categorical predictor W by introducing the central partial mean subspace. Stemming from these concepts, we next define the central partial mean subspace for regression of the multivariate response $\boldsymbol{Y}$ given a mixture of quantitative and qualitative predictors $\boldsymbol{X}$ and W.

The *central partial mean subspace* (CPMS), denoted by $\mathcal{S}^{(W)}_{\mathrm{E}(\boldsymbol{Y}|\boldsymbol{X})}$, is defined as the intersection of all subspaces $\mathcal{S}$ of $\mathbb{R}^p$ satisfying

$$\boldsymbol{Y} \perp\!\!\!\perp E(\boldsymbol{Y}|\boldsymbol{X}, W)|(\boldsymbol{\eta}^{\tau}\boldsymbol{X}, W), \tag{2.1}$$

where $\boldsymbol{\eta}$ denotes a basis of $\mathcal{S}$. $\mathcal{S}^{(W)}_{\mathrm{E}(\boldsymbol{Y}|\boldsymbol{X})}$ uniquely exists under minor conditions (Cook (1998)) and is assumed to exist here. Following this definition, (2.1) indicates that $(\boldsymbol{\eta}^{\tau}\boldsymbol{X}, W)$ contains all information that the predictors $(\boldsymbol{X}, W)$ have to furnish about the conditional mean $\mathrm{E}(\boldsymbol{Y}|\boldsymbol{X}, W)$ and, as such, one can replace $\boldsymbol{X}$ with $\boldsymbol{\eta}^{\tau}\boldsymbol{X}$ to characterize $\mathrm{E}(\boldsymbol{Y}|\boldsymbol{X}, W)$ and lose no information about the conditional mean.

The next proposition, which is a multivariate version of Proposition 2.1 of Li, Cook and Chiaromonte (2003), gives alternative ways to characterize the central partial mean subspace.

Proposition 1. *Condition* (2.1) *is equivalent to either of*

(i) $\mathrm{Cov}(\boldsymbol{Y}, \mathrm{E}(\boldsymbol{Y}|\boldsymbol{X}, W)|\boldsymbol{\eta}^{\tau}\boldsymbol{X}, W) = 0$,

(ii) $\mathrm{E}(\boldsymbol{Y}|\boldsymbol{X}, W) = \mathrm{E}(\boldsymbol{Y}|\boldsymbol{\eta}^{\tau}\boldsymbol{X}, W)$.

The proposition suggests that $\boldsymbol{\eta}^{\tau}\boldsymbol{X}$ is sufficient for the mean function if and only if $\boldsymbol{Y}$ and $\mathrm{E}(\boldsymbol{Y}|\boldsymbol{X}, W)$ are uncorrelated within each subpopulation determined by W, or equivalently, $\mathrm{E}(\boldsymbol{Y}|\boldsymbol{X}, W)$ depends on $\boldsymbol{X}$ only through $\boldsymbol{\eta}^{\tau}\boldsymbol{X}$.

2.2. Decomposition of multivariate CPMS

Let Y_k denote the kth univariate response variable of $\boldsymbol{Y}$, $k = 1, \ldots, r$, and let $\boldsymbol{X}_w$ and Y_{k_w} denote random variables distributed as $\boldsymbol{X}|(W = w)$ and $Y_k|(W = w)$, respectively, for $w = 1, \ldots, c$. In the univariate response case ($r = 1$), Li, Cook and Chiaromonte (2003) showed that the CPMS can be derived from the integration of all the central mean subspaces obtained from each subpopulation of W. That is, for each $k = 1, \ldots, r$, $\mathcal{S}^{W}_{\mathrm{E}(Y_k|\boldsymbol{X})} = \bigoplus_{w=1}^{c} \mathcal{S}_{\mathrm{E}(Y_{k_w}|\boldsymbol{X}_w)}$. Here $\oplus$ indicates direct sum between subspaces: $S_1 \oplus S_2 = \{s_1 + s_2; s_1 \in S_1, s_2 \in S_2\}$. In principle, the subpopulation central mean subspaces $\mathcal{S}_{\mathrm{E}(Y_{k_w}|\boldsymbol{X}_w)}$ can overlap in any fashion, but always add up to $\mathcal{S}^{W}_{\mathrm{E}(Y_k|\boldsymbol{X})}$. For the multivariate response case, Cook and Setodji (2003) showed that the central mean subspace of $\boldsymbol{Y}$ given $\boldsymbol{X}$ can be obtained as the direct sum of the central mean subspaces of all the univariate coordinates Y_k given $\boldsymbol{X}$, i.e., $\mathcal{S}_{\mathrm{E}(\boldsymbol{Y}|\boldsymbol{X})} = \bigoplus_{k=1}^{r} \mathcal{S}_{\mathrm{E}(Y_k|\boldsymbol{X})}$. Combining these observations, we obtain the following.

Proposition 2.

$$\mathcal{S}^{(W)}_{\mathrm{E}(\boldsymbol{Y}|\boldsymbol{X})} = \bigoplus_{w=1}^{c} \mathcal{S}_{\mathrm{E}(\boldsymbol{Y}_w|\boldsymbol{X}_w)} = \bigoplus_{k=1}^{r} \mathcal{S}^{(W)}_{\mathrm{E}(Y_k|\boldsymbol{X})} = \bigoplus_{k=1}^{r}\bigoplus_{w=1}^{c} \mathcal{S}_{\mathrm{E}(Y_{k_w}|\boldsymbol{X}_w)}.$$

This proposition provides a way to develop an estimator of the CPMS through estimation of the individual central mean subspaces $\mathcal{S}_{\mathrm{E}(Y_{k_w}|\boldsymbol{X}_w)}$. A thorough treatment of estimation of $\mathcal{S}^{(W)}_{\mathrm{E}(\boldsymbol{Y}|\boldsymbol{X})}$ is presented in Sections 3 and 6.

2.3. Connection with multivariate linear reduced-rank model

In this section we examine some specific models to help fix the ideas of the central partial mean subspace. Returning to the multivariate linear model (1.1), one assumes that

$$Y_k = \alpha_k + \boldsymbol{b}_k^{\tau}\boldsymbol{X} + c_k W + \delta_k, \quad \text{for } k = 1, \ldots, r,$$

where $\boldsymbol{B} = (\boldsymbol{b}_1, \ldots, \boldsymbol{b}_r)$, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_r)^{\tau}$, $\boldsymbol{C} = (c_1, \ldots, c_r)^{\tau}$, and $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_r)^{\tau}$. Here W is treated as a univariate variable. That is, the mean of each coordinate $\mathrm{E}(Y_k|\boldsymbol{X}, W)$ is a linear function in $\boldsymbol{X}$ and $\boldsymbol{W}$. For this model, it is

straightforward to verify that $\mathcal{S}_{\mathrm{E}(Y_{k_w}|\boldsymbol{X}_w)} = \mathrm{Span}(\boldsymbol{b}_k)$ for all w's, and $\mathcal{S}^{(W)}_{\mathrm{E}(\boldsymbol{Y}|\boldsymbol{X})} = \mathrm{Span}(\boldsymbol{B}) = \bigoplus_{k=1}^{r} \mathrm{Span}(\boldsymbol{b}_k)$.

Note that (1.1) does not allow for interaction between $\boldsymbol{X}$ and W, which may be restrictive in practice. Taking the Berkeley Guidance Study as an example, it seems more reasonable to believe that different covariate effects on aging may exist for boys and girls. The structure of $\mathcal{S}^{(W)}_{\mathrm{E}(\boldsymbol{Y}|\boldsymbol{X})}$, on the contrary, does permit interactions, e.g.,

$$Y_k = \alpha_k + \sum_{w=1}^{c} \boldsymbol{b}_{k_w}^{\mathsf{T}} \boldsymbol{X} \mathbb{1}_{(W=w)} + c_k W + \delta_k, \quad \text{for } k = 1, \ldots, r,$$

where $\mathbb{1}$ denotes the indicator function and different covariate effects $\boldsymbol{b}_{k_w}$ could be estimated for different groups in W. In this model, $\mathcal{S}_{\mathrm{E}(Y_{k_w}|\boldsymbol{X}_w)} = \mathrm{Span}(\boldsymbol{b}_{k_w})$, and $\mathcal{S}^{(W)}_{\mathrm{E}(\boldsymbol{Y}|\boldsymbol{X})} = \bigoplus_{k=1}^{r} \bigoplus_{w=1}^{c} \mathrm{Span}(\boldsymbol{b}_{k_w})$.

Actually the structure of $\mathcal{S}^{(W)}_{\mathrm{E}(\boldsymbol{Y}|\boldsymbol{X})}$ permits more general regression forms than the ones mentioned above, for instance,

$$Y_k = \alpha_k + c_k W + \sum_{w=1}^{c} \left(\boldsymbol{b}_{k_w}^{\mathsf{T}} \boldsymbol{X} \mathbb{1}_{(W=w)} + \sigma_{k_w}(\boldsymbol{b}_{k_w}^{\mathsf{T}} \boldsymbol{X}) \delta_{k_w} \right),$$

where the subpopulation variance function σ_{k_w} can depend on the linear combinations of $\boldsymbol{X}$ that serves as the mean function. It also covers models such as

$$Y_k = \alpha_k + \sum_{w=1}^{c} f_{k_w}\left(\boldsymbol{b}_{k_w}^{\mathsf{T}} \boldsymbol{X} \mathbb{1}_{(W=w)} \right) + \delta_k,$$

where the conditional mean $\mathrm{E}(Y_k|\boldsymbol{X}, W)$ can depend on the terms $(\boldsymbol{b}_{k_1}^{\mathsf{T}}\boldsymbol{X}, \ldots, \boldsymbol{b}_{k_c}^{\mathsf{T}}\boldsymbol{X})$ and W in a nonlinear fashion. In the above examples, we always have $\mathcal{S}^{(W)}_{\mathrm{E}(\boldsymbol{Y}|\boldsymbol{X})} = \bigoplus_{k=1}^{r} \bigoplus_{w=1}^{c} \mathrm{Span}(\boldsymbol{b}_{k_w})$.

3. Estimation of Reduced-rank Regression with Mixture Predictors

3.1. Estimation of CPMS via OLS

Proposition 2suggests that the CPMS can be derived from the central mean subspace for each univariate response coordinate within each subpopulation implied by W. We next develop an estimator of the CPMS and the associated test for dimension.

For $\boldsymbol{\Sigma}_w = \mathrm{Var}(\boldsymbol{X}_w)$, assumed positive definite, and $\boldsymbol{\sigma}_{k_w} = \mathrm{Cov}(Y_{k_w}, \boldsymbol{X}_w)$, let $\boldsymbol{\beta}_{k_w} = \boldsymbol{\Sigma}_w^{-1}\boldsymbol{\sigma}_{k_w}$ denote the population ordinary least squares vector regressing Y_{k_w} on $\boldsymbol{X}_w$, and let $\boldsymbol{\beta} = (\boldsymbol{\beta}_{1_1}, \ldots, \boldsymbol{\beta}_{r_1}, \ldots, \boldsymbol{\beta}_{1_c}, \ldots, \boldsymbol{\beta}_{r_c})$. We assume the following conditions:

(C.1) $\mathrm{E}\,(\boldsymbol{X}_w|\boldsymbol{\gamma}^\intercal \boldsymbol{X}_w = \boldsymbol{\nu})$ is a linear function of $\boldsymbol{\nu}$ for any $\boldsymbol{\gamma} \in \mathbb{R}^p$;

(C.2) $\mathcal{S}_{\mathrm{E}\,(Y_{k_w}|\boldsymbol{X}_w)} \subseteq \mathrm{Span}(\boldsymbol{\beta})$ for $k = 1, \ldots, r$ and $w = 1, \ldots, c$.

The first condition (C.1) is often called the linearity condition, and is a common assumption imposed by most sufficient dimension reduction methods. Li and Duan (1989) showed that, under this condition, $\mathrm{Span}(\boldsymbol{\beta}) \subseteq \mathcal{S}_{\mathrm{E}\,(\boldsymbol{Y}|\boldsymbol{X})}$. It is typically viewed as a mild restriction since it involves only the marginal distribution of the predictors, and it holds to a reasonable approximation as the number of predictors increases (Hall and Li (1993)). It may also be induced by predictor transformation, re-weighting (Cook and Nachtsheim (1994)), or clustering (Li, Cook and Nachtsheim (2004)). Condition (C.2) is satisfied when the distribution of $Y_{k_w}|\boldsymbol{X}_w$ can be summarized through at most one linear combination of the predictors. Even so, (C.2) permits the distribution of $\boldsymbol{Y}|\boldsymbol{X}$ to depend on multiple linear combinations of the predictors. In general, the conditions (C.1) and (C.2) are flexible enough to cover many regression structures. For instance, they hold for all the models discussed in Section 2.3.

Assuming (C.1) and (C.2), the following relation can be derived from Proposition 2.

$$\mathcal{S}^{(W)}_{\mathrm{E}\,(\boldsymbol{Y}|\boldsymbol{X})} = \mathrm{Span}(\boldsymbol{\beta}) = \mathrm{Span}(\boldsymbol{\beta}_{1_1}, \ldots, \boldsymbol{\beta}_{r_1}, \ldots, \boldsymbol{\beta}_{1_c}, \ldots, \boldsymbol{\beta}_{r_c}).$$

Consequently, we propose to use $\mathrm{Span}(\hat{\boldsymbol{\beta}}) = \mathrm{Span}(\hat{\boldsymbol{\beta}}_{1_1}, \ldots, \hat{\boldsymbol{\beta}}_{r_1}, \ldots, \hat{\boldsymbol{\beta}}_{1_c}, \ldots, \hat{\boldsymbol{\beta}}_{r_c})$ to construct an estimate of $\mathcal{S}^{(W)}_{\mathrm{E}\,(\boldsymbol{Y}|\boldsymbol{X})}$, where $\hat{\boldsymbol{\beta}}_{k_w}$ is the usual OLS sample estimate. We next develop a formal test to determine the rank of $\boldsymbol{\beta}$.

3.2. Reduced-rank dimension test

As implied by the discussion in Section 2.3, the rank of $\boldsymbol{B}$ in the multivariate reduced-rank linear model (1.1) is the dimension of the central partial mean subspace. As a consequence, we can infer the rank of the model (1.1) through a formal test of the dimension of $\mathcal{S}^{(W)}_{\mathrm{E}\,(\boldsymbol{Y}|\boldsymbol{X})}$. In addition, the dimension test we develop here applies to more general structures of the reduced-rank models than (1.1), as suggested in Section 2.3.

Specifically, we consider the hypotheses

$$H_0 : \mathrm{rank}(\boldsymbol{\beta}) = m \quad \text{versus} \quad H_A : \mathrm{rank}(\boldsymbol{\beta}) > m. \tag{3.1}$$

We repeat the test for a series of values of m from 0 to $p-1$, and we take the minimum m such that H_0 is not rejected as an estimate of $\mathrm{rank}(\boldsymbol{\beta}) = \dim(\mathcal{S}^{(W)}_{\mathrm{E}\,(\boldsymbol{Y}|\boldsymbol{X})})$. To help derive the large sample test for (3.1), we first note that the rank of $\boldsymbol{\beta}$ remains unchanged if it is multiplied by a full rank matrix, or its columns

are multiplied by nonzero scalars. We thus introduce the following transformation. Let $a_w = Pr(W = w)^{1/2} > 0$ denote the square root of the probability of being in a subpopulation w defined by W. Write $\boldsymbol{\beta}_w = (\boldsymbol{\beta}_{1_w}, \ldots, \boldsymbol{\beta}_{r_w})$, and let $\boldsymbol{\beta}^* = (a_1\boldsymbol{\beta}_1, \ldots, a_c\boldsymbol{\beta}_c)$. Next define $\boldsymbol{\Sigma}_\bullet = \sum_{w=1}^c a_w^2\boldsymbol{\Sigma}_w = \mathrm{E}(\boldsymbol{\Sigma}_W)$, and assume it is positive definite. Compute the residual ε_{k_w} from the population OLS fit of the kth response variable Y_{k_w} on $\boldsymbol{X}_w$ within the subpopulation w, $\varepsilon_{k_w} = (Y_{k_w} - \mathrm{E}(Y_{k_w})) - \boldsymbol{\beta}_{k_w}^\intercal(\boldsymbol{X}_w - \mathrm{E}(\boldsymbol{X}_w))$, and write $\boldsymbol{\varepsilon}_w = (\varepsilon_{1_w}, \ldots, \varepsilon_{r_w})^\intercal$, $\boldsymbol{\Omega}_w = \mathrm{Var}(\boldsymbol{\varepsilon}_w) > 0$, and define $\boldsymbol{\Omega} = \mathrm{diag}(\boldsymbol{\Omega}_1, \ldots, \boldsymbol{\Omega}_c)$ as the block diagonal matrix with diagonal blocks $\boldsymbol{\Omega}_w$. Then the null hypothesis H_0 in (3.1) is equivalent to

$$H_0' : \mathrm{rank}(\boldsymbol{\Sigma}_\bullet^{1/2}\boldsymbol{\beta}^*\boldsymbol{\Omega}^{-1/2}) = m. \tag{3.2}$$

Given n i.i.d. sample observations of $(\boldsymbol{Y}, \boldsymbol{X}, W)$, $\{(\boldsymbol{Y}_{i_w}, \boldsymbol{X}_{i_w}) : i = 1, \ldots, n_w\}$, with $n = \sum_{w=1}^c n_w$, the population quantities $\boldsymbol{\Sigma}_\bullet$, $\boldsymbol{\beta}^*$, and $\boldsymbol{\Omega}$ in (3.2) can be estimated by their usual sample counterparts $\hat{\boldsymbol{\Sigma}}$, $\hat{\boldsymbol{\beta}}^*$ and $\hat{\boldsymbol{\Omega}}$, respectively. Since the rank of a matrix corresponds to its number of nonzero singular values, a natural test statistic for (3.2) is

$$\hat{\Lambda}_m = \sum_{i=m+1}^{\min(p,rc)} \hat{\lambda}_i,$$

where $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_p \geq 0$ are the ordered eigenvalues of the $p \times p$ matrix

$$n(\hat{\boldsymbol{\Sigma}}_\bullet^{1/2}\hat{\boldsymbol{\beta}}^*\hat{\boldsymbol{\Omega}}^{-1/2})(\hat{\boldsymbol{\Sigma}}_\bullet^{1/2}\hat{\boldsymbol{\beta}}^*\hat{\boldsymbol{\Omega}}^{-1/2})^\intercal.$$

We next derive the asymptotic distribution of the test statistic $\hat{\Lambda}_m$ under H_0. Let $d = \mathrm{rank}(\boldsymbol{\beta})$ and consider the singular value decomposition

$$\boldsymbol{\Sigma}_\bullet^{1/2}\boldsymbol{\beta}^*\boldsymbol{\Omega}^{-1/2} = (\boldsymbol{\Gamma}_0\, \boldsymbol{\Gamma}) \begin{pmatrix} \boldsymbol{D} & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \boldsymbol{\Psi}_0^\intercal \\ \boldsymbol{\Psi}^\intercal \end{pmatrix},$$

where $(\boldsymbol{\Gamma}_0, \boldsymbol{\Gamma})$ and $(\boldsymbol{\Psi}_0, \boldsymbol{\Psi})$ are orthogonal matrices with dimensions $p \times p$ and $rc \times rc$, and $\boldsymbol{D}$ is a $d \times d$ diagonal matrix with positive diagonal elements, $\boldsymbol{\Gamma}$ has dimension $p \times (p-d)$ and $\boldsymbol{\Psi}$ has dimension $rc \times (rc-d)$. Define the standardized predictor ${}_w = \boldsymbol{\Sigma}_w^{-1/2}(\boldsymbol{X}_w - \mathrm{E}(\boldsymbol{X}_w))$, and a random vector $\boldsymbol{T}_w = \mathrm{vec}({}_w\varepsilon_w^\intercal)$ where vec is a matrix operator that stacks all the columns of a matrix to a vector. Next, define

$$\boldsymbol{\Delta} = \sum_{w=1}^c [(\Psi_w^\intercal\boldsymbol{\Omega}_w^{-1/2}) \otimes (\boldsymbol{\Gamma}^\intercal\boldsymbol{\Sigma}_\bullet^{1/2}\boldsymbol{\Sigma}_w^{-1/2})](\mathrm{E}(\boldsymbol{T}_w\boldsymbol{T}_w^\intercal))[(\boldsymbol{\Omega}_w^{-1/2}\boldsymbol{\Psi}_w) \otimes (\boldsymbol{\Sigma}_w^{-1/2}\boldsymbol{\Sigma}_\bullet^{1/2}\boldsymbol{\Gamma})],$$

where $\otimes$ indicates a kronecker product, and for $w \in \{1, \ldots, c\}$, the $\boldsymbol{\Psi}_w$'s are the $r \times (rc-d)$ row matrices of $\boldsymbol{\Psi}$ so that $\boldsymbol{\Psi}^\intercal = (\boldsymbol{\Psi}_1^\intercal, \ldots, \boldsymbol{\Psi}_c^\intercal)$. The next proposition gives the asymptotic distribution of the test statistic $\hat{\Lambda}_d$.

Proposition 3. *Assuming that all moments involved in* $\boldsymbol{\Delta}$ *are finite, then as* $n_w \to \infty$ *for* $w = 1, \ldots, c$,

$$\hat{\Lambda}_d = \sum_{i=d+1}^{\min(p,rc)} \hat{\lambda}_i \xrightarrow{\mathcal{L}} \sum_{i=1}^{(p-d)(rc-d)} \alpha_i K_i$$

where $\alpha_1, \ldots, \alpha_{(p-d)(rc-d)}$ *are the eigenvalues of the matrix* $\boldsymbol{\Delta}$, *and* $K_1, \ldots, K_{(p-d)(rc-d)}$ *are independent identically distributed random variables with a* χ^2_1 *distribution.*

Proposition 3 suggests that one can get the p-value of the test from a weighted chi-squared distributions, where the weights α_i's can be estimated consistently from a sample version of $\boldsymbol{\Delta}$. Alternatively, instead of calculating the p-value for the distribution of $\hat{\Lambda}_m$, one may also use Satterthwaite's (1941) chi-squared approximation. See Bentler and Xie (2000), Cook and Setodji (2003), and Li, Cook and Chiaromonte (2003) for more details about the approximate chi-squared test.

In some regression problems, the asymptotic distribution of $\hat{\Lambda}_m$ can be simplified to a single chi-squared rather than a linear combination of chi-squared distributions. The next corollary gives conditions for such a simplification.

Corollary 1. *Suppose:*

(a) $\mathrm{Var}(\boldsymbol{T}_w) = \mathrm{E}(\boldsymbol{\varepsilon}_w \boldsymbol{\varepsilon}_w^\intercal) \otimes \mathrm{E}(\boldsymbol{\varepsilon}_w^\intercal \boldsymbol{\varepsilon}_w)$, *for all* $w = 1, \ldots, c$, *or*

(b) $\boldsymbol{Y}_w | \boldsymbol{X}_w$ *follows a location regression, i.e.,* $\boldsymbol{Y}_w \perp\!\!\!\perp \boldsymbol{X}_w | \mathrm{E}(\boldsymbol{Y}_w | \boldsymbol{X}_w)$, *and* $\boldsymbol{X}_w$ *is normally distributed.*

Then $\boldsymbol{\Delta} = \sum_{w=1}^{c} [\Psi_w^\intercal \Psi_w] \otimes [\boldsymbol{\Gamma}^\intercal \boldsymbol{\Sigma}_\bullet^{1/2} \boldsymbol{\Sigma}_w^{-1} \boldsymbol{\Sigma}_\bullet^{1/2} \boldsymbol{\Gamma}]$. *If, in addition,* $\boldsymbol{\Sigma}_\bullet = \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \cdots = \boldsymbol{\Sigma}_c$ *then* $\hat{\Lambda}_d \xrightarrow{\mathcal{L}} \chi^2_{(p-d)(rc-d)}$.

Note that condition (a) is satisfied when any of the following conditions are met (Cook and Setodji (2003)): $\mathrm{Cov}(\varepsilon_{k_w} \varepsilon_{j_w}, \boldsymbol{\varepsilon}_w^\intercal) = 0$ for $j, k = 1, \ldots, r$ and $w = 1, \ldots, c$, or $\boldsymbol{Y}_w | \boldsymbol{X}_w$ follows a location regression and $d = 0$. Moreover, the condition of equal predictor covariance matrix $\boldsymbol{\Sigma}_w$ across all subpopulations may be appropriate in some applications, for instance, in cases where W denotes a treatment that is randomly assigned to all the experimental units.

4. Application to the Berkeley Guidance Study

We revisit the Berkeley Guidance Study data introduced in Section 1 to illustrate aspects of data analysis of the proposed methods. In this analysis, the response vector consists of the height and strength of the children at age 18, $\boldsymbol{Y} = (HT_{18}, ST_{18})$, and the predictors include HT_2, HT_9, WT_2, WT_9, LG_9, ST_9 ($\boldsymbol{X}$), plus gender (W).

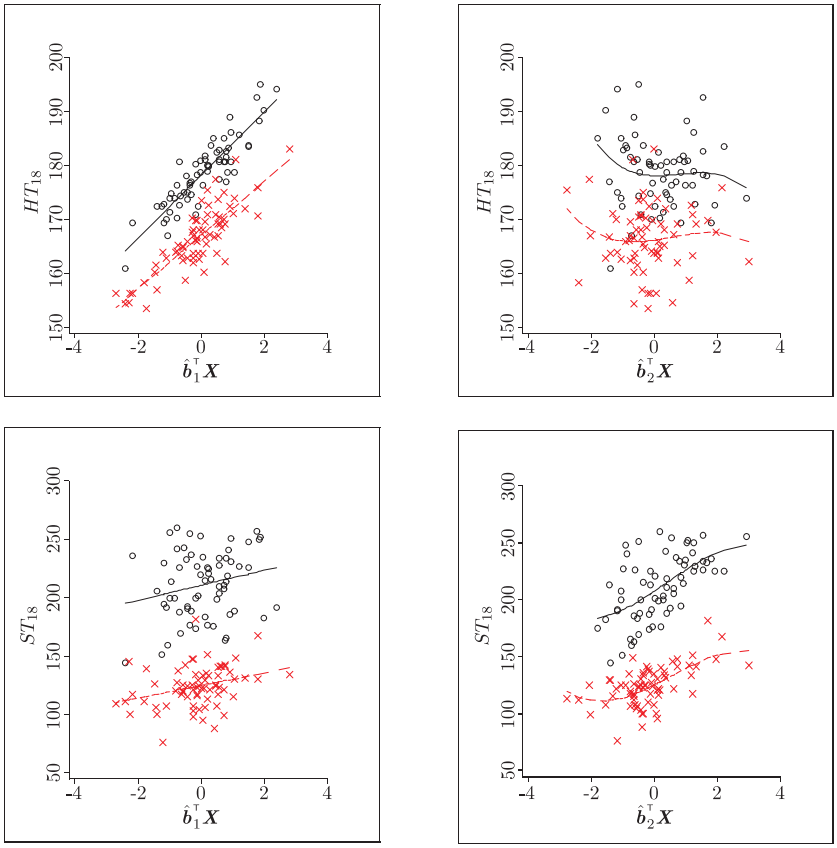


Figure 4.1. Summary plot of the responses versus the sufficient predictors marked by gender. Boys (o), girls (+).

To ensure condition (C.1), we first transformed the predictors height and leg circumference measurements to log scale, while the rest of the predictors remained untransformed. Such a transformation leads to an approximate multivariate normal distribution, and as such the linearity condition is met. For the second condition (C.2), we examined the dimension of the subgroup central mean subspace $\mathcal{S}_{\mathrm{E}(Y_{k_w}|\boldsymbol{X}_w)}$ based on the dimension test of sliced inverse regression (Li (1991)) within each gender group. In all cases the dimension was concluded to be no greater than one, suggesting that (C.2) holds.

We then constructed the test statistics $\hat{\Lambda}_m$ for $m = 1, \ldots, 4$, and referenced those statistics to the weighted chi-squared distribution as in Proposition 3. The resulting p-values were 0, 0, 0.32 and 0.56, respectively. We also tried with the simplified chi-squared reference distribution as in Corollary 1, yielding the p-values 0, 0, 0.34, and 0.58. We thus concluded that only two linear combinations of $\boldsymbol{X}$ are needed to characterize the conditional mean $\mathrm{E}(\boldsymbol{Y}|\boldsymbol{X}, W)$. After

Table 4.1. Estimated mean and standard deviation (in parenthesis) from a linear model with interactions on the Berkeley Guidance Study.

Model predictors Boys vs. Girls	Model for ST_{18} ($R^2 = 0.87$) -80.3 (3.2)		Model for HT_{18} ($R^2 = 0.88$) -11.2 (0.5)	
	with boys	with girls	with boys	with girls
$\hat{\boldsymbol{b}}_1^\top \boldsymbol{X}$ interaction	9.4 (2.3)	4.1 (2.2)	5.9 (0.4)	5.1 (0.4)
$\hat{\boldsymbol{b}}_2^\top \boldsymbol{X}$ interaction	18.0 (2.3)	9.2 (2.2)	0.2 (0.4)	-0.9 (0.4)

standardizing the coordinates of $\boldsymbol{X}$ to have sample standard deviation one, we obtained the linear combinations

$$\begin{aligned}\hat{\boldsymbol{b}}_1^\top \boldsymbol{X} &= -0.06\log(HT_2) + 0.97\log(HT_9) + 0.10WT_2 - 0.22WT_9 - 0.07\log(LG_9)\\ &\quad +0.003ST_9,\\ \hat{\boldsymbol{b}}_2^\top \boldsymbol{X} &= 0.12\log(HT_2) - 0.57\log(HT_9) - 0.15WT_2 + 0.38WT_9 - 0.03\log(LG_9)\\ &\quad +0.700ST_9.\end{aligned}$$

Figure 4.1 shows the summary plot of $\boldsymbol{Y} = (HT_{18}, ST_{18})$ versus $\hat{\boldsymbol{b}}_1^\top \boldsymbol{X}$ and $\hat{\boldsymbol{b}}_2^\top \boldsymbol{X}$. The plot sustains the common belief that boys and girls have different growth trajectories, and thus there exists interaction between $\boldsymbol{X}$ and W. Secondly, both response variables exhibit a strong linear correlation with the first summary variable $\hat{\boldsymbol{b}}_1^\top \boldsymbol{X}$, while there is a hint of a possibly quadratic relation with $\hat{\boldsymbol{b}}_2^\top \boldsymbol{X}$. These observations would facilitate subsequent model formulation.

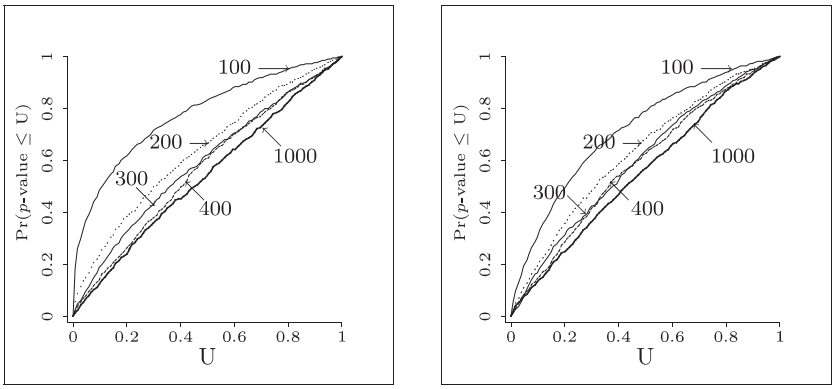
To complete the analysis, we fitted a simple linear model for each response variable given the two new predictors $\hat{\boldsymbol{b}}_1^\top \boldsymbol{X}$ and $\hat{\boldsymbol{b}}_2^\top \boldsymbol{X}$ interacted with gender. Both models yielded R^2 equal to about 0.9, indicating a good model fit to the data. The results are summarized in Table 4.1 where it can be seen, for instance, on average the strength of boys is about 80kg greater than that of girls, while boys are about 11cm taller than girls at age 18; the strength of teenagers increases at a rate of 9kg for boys and 4 kg for girls for every one unit increase in the score of $\hat{\boldsymbol{b}}_1^\top \boldsymbol{X}$. Similar interpretation can be obtained for the two response variables given $\hat{\boldsymbol{b}}_1^\top \boldsymbol{X}$ and $\hat{\boldsymbol{b}}_2^\top \boldsymbol{X}$ for boys and girls, respectively. We also examined the linear model with a quadratic term of $\hat{\boldsymbol{b}}_2^\top \boldsymbol{X}$ for both response variables, but there was no significant gain with this additional term.

5. Finite Sample Performance

5.1. Reduced-rank dimension test

We first examine the finite sample performance of the chi-squared tests in Proposition 3 and Corollary 1. We start with the following model, with $p = 5$ and $r = 4$,

$$Y_1 = (\boldsymbol{b}_1^\top \boldsymbol{X})^2 \mathbb{1}_{(\mathrm{W}=0)} + \boldsymbol{b}_2^\top \boldsymbol{X}\, \mathbb{1}_{(\mathrm{W}=1)} + (\boldsymbol{b}_1^\top \boldsymbol{X})\delta_1,$$



a. Chi-squared test of Corollary 1 b. Weighted chi-squared test of Proposition 3

Figure 5.2. Estimated Pr(p-value $\leq$ U) for model (5.1) with sample size $n = 100, 200, 300, 400$ and $1,000$.

$$
\begin{aligned}
Y_2 &= \boldsymbol{b}_2^{\intercal}\boldsymbol{X}\mathbb{I}_{(\mathrm{W}=0)} + \boldsymbol{b}_1^{\intercal}\boldsymbol{X}\mathbb{I}_{(\mathrm{W}=1)} + \delta_2,\\
Y_3 &= \boldsymbol{b}_1^{\intercal}\boldsymbol{X}\mathbb{I}_{(\mathrm{W}=0)} + \delta_3,\\
Y_4 &= Y_2^2 + \boldsymbol{b}_1^{\intercal}\boldsymbol{X}\mathbb{I}_{(\mathrm{W}=1)} + (\boldsymbol{b}_2^{\intercal}\boldsymbol{X})\delta_4,
\end{aligned} \tag{5.1}
$$

where $\boldsymbol{X}$ follows a multivariate standard normal distribution, W is a Bernoulli random variable with probability of success 0.4, $\delta_1, \ldots, \delta_4$ are standard normal errors independent of $\boldsymbol{X}$ and W, and $\boldsymbol{b}_1$, $\boldsymbol{b}_2$ are two vectors in $\mathbb{R}^p$. First we set $\boldsymbol{b}_1 = \boldsymbol{b}_2 = (1,1,1,1,1)^{\intercal}$, and so the central partial mean subspace is $\mathcal{S}^{(W)}_{\mathrm{E}(\boldsymbol{Y}|\boldsymbol{X})} = \mathrm{Span}(\boldsymbol{b}_1)$ with dimension $d = 1$. Since W is independent of $\boldsymbol{X}$, we have $\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}_1$. In addition, since $\boldsymbol{Y}_w|X_w$ follows a location regression and $\boldsymbol{X}_w$ is normally distributed, the chi-squared test in Corollary 1can be readily applied.

For each simulated data, $\hat{\Lambda}_1$ was computed and the corresponding p-value was derived. This procedure was repeated 1,000 times to obtain an estimate of the distribution of the p-value for $\hat{\Lambda}_1$. If the distribution of $\hat{\Lambda}_1$ is well approximated by a chi-squared with $(p-1)(rc-1)$ degrees of freedom, then it is expected to have Pr(p-values $\leq u$) $\approx u$ for $0 \leq u \leq 1$. The empirical cdf's of the p-values for the sample sizes $n = 100, 200, 300, 400$, and 1,000 are shown in Figure 5.2a. For a small sample size as $n = 100$, the empirical cdf was notably curved. However, it improved substantially when the sample size increased, and the actual level quickly approached the true nominal level. The actual level of the nominal 5% test was 9.8% for $n = 200$, 6.8% for $n = 400$, and 5.6% for $n = 1,000$. We also considered a model with $\boldsymbol{b}_1 = (1,1,1,1,1)^{\intercal}$ and $\boldsymbol{b}_2 = (1,-1,-1,0,0)^{\intercal}$, which yields a reduced-rank model with dimension $d = 2$. The resulting plot was similar to Figure 5.2a, and thus is not shown. Numerically, the observed level of

the nominal 5% test was about 12.3% when $n = 200$, and dropped to 6.4% for $n = 400$ and 5.9% for $n = 1,000$.

We next examined a model where the chi-squared test of Corollary 1 no longer applied, but the weighted chi-squared test of Proposition 3 was still applicable. Specifically, we set W as a Bernoulli with probability of success equal to 0.4. $X_j = \{u_j W + v_j(1-W)\}T_j$, $j = 1, \ldots, 5$, where T_j's were independent standard normal random variables, $u_1 = 0.3$, $v_1 = 1.2$; $u_2 = 1.3$, $v_2 = 0.2$; $u_3 = 0.5$, $v_3 = 0.6$; $u_4 = 0.6$, $v_4 = 1.2$; and $u_5 = 1.1$, $v_5 = 0.5$. The rest of data generation remained the same as in (5.1). In this setup, X depends on W and $\boldsymbol{\Sigma}_w$ differs across different levels of W. We examined both a $d = 1$ model and a $d = 2$ model as before, and Figure 5.2b shows the empirical cdf's of the p-values obtained from the weighted chi-squared test of $\hat{\Lambda}_2$ for the $d = 2$ case. Similar qualitative patterns as Figure 5.2a were observed, indicating that the weighted chi-squared test worked quite well too.

We have also examined the effect of the dimension r of Y and p of X on the performance of the proposed test. As anticipated, a smaller number r or p yielded a more accurate actual level for the test. For brevity, those results are not reported here. In summary we conclude, based on our simulations, that the actual level of the proposed chi-squared test is sufficiently close to the nominal level to be practically useful, provided that there is a reasonable amount of, say, about 200 or more sample observations.

5.2. Estimation of basis of CPMS

We next evaluated accuracy of basis estimation for the central partial mean subspace. We employed the squared multiple correlation coefficient (Li (1991, p.318)) as the evaluation criterion,

$$\boldsymbol{R}^2(\hat{\boldsymbol{b}}_j) = \max_{\boldsymbol{\gamma} \in \mathrm{Span}\{\boldsymbol{b}_1, \ldots, \boldsymbol{b}_d\}} \frac{(\hat{\boldsymbol{b}}_j^{\top} \boldsymbol{\Sigma} \boldsymbol{\gamma})^2}{(\hat{\boldsymbol{b}}_j^{\top} \boldsymbol{\Sigma} \hat{\boldsymbol{b}}_j) \times (\boldsymbol{\gamma}^{\top} \boldsymbol{\Sigma} \boldsymbol{\gamma})}, \quad \text{for} \quad j = 1, \ldots, d,$$

where $\boldsymbol{\Sigma} = \mathrm{Var}(\boldsymbol{X})$, $(\boldsymbol{b}_1, \ldots, \boldsymbol{b}_d)$ denotes a basis of $\mathcal{S}^{(W)}_{\mathrm{E}(\boldsymbol{Y}|\boldsymbol{X})}$ and $(\hat{\boldsymbol{b}}_1, \ldots, \hat{\boldsymbol{b}}_d)$ are the corresponding sample estimates.

We first examined a bivariate response model with $d = 1$:

$$\begin{aligned} Y_1 &= \delta_1 \exp\{W(X_1 + X_2) - (1-W)(X_1 + X_2)^2\}, \\ Y_2 &= \frac{X_1 + X_2}{1 + \exp(X_1 + X_2 + W)} + \delta_2, \end{aligned} \tag{5.2}$$

where $\boldsymbol{X}$ and W were generated the same way as in the weighted chi-squared test example in Section 5.1. For this model, $\mathcal{S}^{(W)}_{\mathrm{E}(\boldsymbol{Y}|\boldsymbol{X})} = \mathrm{Span}(\boldsymbol{b}_1)$, with $\boldsymbol{b}_1 =$

Table 5.2. Estimated mean and standard deviation (in parentheses) of the squared multiple correlation coefficient.

	Model (5.2)	Model (5.3)	
Sample size	$\boldsymbol{R}^2(\hat{\boldsymbol{b}}_1)$	$\boldsymbol{R}^2(\hat{\boldsymbol{b}}_1)$	$\boldsymbol{R}^2(\hat{\boldsymbol{b}}_2)$
100	0.900 (0.103)	0.965 (0.038)	0.803 (0.241)
300	0.958 (0.039)	0.991 (0.007)	0.933 (0.088)
500	0.972 (0.023)	0.995 (0.004)	0.961 (0.061)

$(1, 1, 0, \ldots, 0)^\tau$. The first response component Y_1 is heteroscedastic, while the second component Y_2 depends on $\boldsymbol{b}_1^\tau X$ in a nonlinear fashion. Table 5.2 reports the mean and the standard deviation of $\boldsymbol{R}^2(\hat{\boldsymbol{b}}_1)$ out of 1,000 data replications. The average $\boldsymbol{R}^2$ was above 0.9 for all sample sizes, indicating that the reduced-rank method produced really good estimate of the basis of the CPMS.

We next considered a bivariate response model with $d = 2$:

$$\begin{aligned} Y_1 &= \{X_1 + (1 - W)X_2\}(X_1 + X_2 + 1) + \delta_1, \\ Y_2 &= \frac{X_1 + (1 - W)X_2}{0.5 + (X_1 + 2X_2 + 1)^2} + \delta_2, \end{aligned} \tag{5.3}$$

where the rest of data generation was the same as in the previous example. For this setup, $\mathcal{S}^{(W)}_{\mathrm{E}(\boldsymbol{Y}|\boldsymbol{X})} = \mathrm{Span}(\boldsymbol{b}_1, \boldsymbol{b}_2)$, with $\boldsymbol{b}_1 = (1, 0, \ldots, 0)^\tau$ and $\boldsymbol{b}_2 = (0, 1, 0, \ldots, 0)^\tau$. Table 5.2 reports the mean and the standard deviation of $\boldsymbol{R}^2(\hat{\boldsymbol{b}}_1)$ and $\boldsymbol{R}^2(\hat{\boldsymbol{b}}_2)$, respectively, based on 1,000 replications. Again, the proposed method is seen to estimate the true basis pretty accurately.

6. Extended Estimation of Reduced-rank Regression

6.1. Multivariate partial iterative Hessian transformation

In Section 3, estimation of the central partial mean subspace was developed based on the ordinary least squares estimator. The same idea can also be generalized to other estimators of the central mean subspace. Representative estimators in this category include principal Hessian directions (PHD, Li (1992) and Cook (1998)), iterative Hessian transformation (IHT, Cook and Li (2002, 2004)), and minimum average variance estimation (MAVE, Xia, Tong, Li and Zhu (2002)). Among them, PHD requires an additional constant variance condition besides the linearity condition, and is generally insensitive to linear trends in regression. MAVE further relaxes the linearity condition, but it involves high-dimensional nonparametric smoothing and may be computationally intensive. IHT combines OLS and PHD in estimating the central mean subspace, requires only the linearity condition, and also avoids nonparametric smoothing. For these reasons,

in this section, we extend IHT to multivariate reduced-rank regression with a mixture of continuous and categorical predictors.

Let $\boldsymbol{\beta}_{k_w} = \mathrm{Var}\,(Y_{k_w})^{-1}\boldsymbol{\Sigma}_w^{-1}\boldsymbol{\sigma}_{k_w}$ denote the standardized version of the OLS vector of regressing Y_{k_w} on $_w$, and let $\boldsymbol{H}_{k_w} = \mathrm{E}\,[\mathrm{Var}\,(Y_{k_w})^{-1}\{Y_{k_w} - \mathrm{E}\,(Y_{k_w}) - \boldsymbol{\beta}_{k_w}^{\intercal}{}_w\}{}_w{}^{\intercal}{}_w]$. Following Cook and Li (2002, 2004), and assuming the linearity condition (C.1), IHT estimates the central mean subspace by a matrix consisting of iteratively transformed $\boldsymbol{\beta}_{k_w}$ by $\boldsymbol{H}_{k_w}$, i.e.,

$$\mathrm{Span}\Big(\boldsymbol{\beta}_{k_w}, \boldsymbol{H}_{k_w}\boldsymbol{\beta}_{k_w}, \boldsymbol{H}_{k_w}^2\boldsymbol{\beta}_{k_w}, \ldots, \boldsymbol{H}_{k_w}^{p-1}\boldsymbol{\beta}_{k_w}\Big) \subseteq \mathcal{S}_{\mathrm{E}\,(Y_{k_w}|\,w)}.$$

Consequently one can estimate $\mathcal{S}^{(W)}_{\mathrm{E}\,(\boldsymbol{Y}|\,)}$ by combining IHT estimates of $\mathcal{S}_{\mathrm{E}\,(Y_{k_w}|\,w)}$ for each component of $\boldsymbol{Y}$ and at each level of W. Define $\boldsymbol{M}_{k_w} = (\boldsymbol{\beta}_{k_w}, \boldsymbol{H}_{k_w}\boldsymbol{\beta}_{k_w}, \boldsymbol{H}_{k_w}^2\boldsymbol{\beta}_{k_w}, \ldots, \boldsymbol{H}_{k_w}^{p-1}\boldsymbol{\beta}_{k_w})$, and $\boldsymbol{M} = (a_1\boldsymbol{M}_{1_1}, \ldots, a_1\boldsymbol{M}_{r_1}, \ldots, a_c\boldsymbol{M}_{1_c}, \ldots, a_c\boldsymbol{M}_{r_c})$, where $a_w = Pr(W = w)^{1/2}$ as defined before. Then, following Proposition 2, under the linearity condition (C.1) and the usually imposed coverage condition, we have

$$\mathcal{S}^{(W)}_{\mathrm{E}\,(\boldsymbol{Y}|\,)} = \mathrm{Span}(\boldsymbol{M}).$$

We next derive a test statistic for estimating the rank of $\boldsymbol{M}$ obtained through IHT. The development parallels that in Section 3.2 for OLS-based estimation. Consider the following hypotheses

$$H_0 : \mathrm{rank}(\boldsymbol{M}) = m \quad \text{versus} \quad H_A : \mathrm{rank}(\boldsymbol{M}) > m. \tag{6.1}$$

Letting $\hat{\boldsymbol{M}}$ denote the sample estimate of $\boldsymbol{M}$ by substituting corresponding sample estimates of a_w, $\boldsymbol{\beta}_{k_w}$ and $\boldsymbol{H}_{k_w}$ in $\boldsymbol{M}$, we construct the test statistic for (6.1) as $\tilde{\Lambda}_m = \sum_{i=m+1}^{p}\tilde{\lambda}_i$, where $\tilde{\lambda}_1 \geq \cdots \geq \tilde{\lambda}_p \geq 0$ are the ordered eigenvalues of the $p \times p$ matrix $n\hat{\boldsymbol{M}}\hat{\boldsymbol{M}}^{\intercal}$. Next consider the singular value decomposition

$$\boldsymbol{M} = \begin{pmatrix}\tilde{\boldsymbol{\Gamma}}_0\ \tilde{\boldsymbol{\Gamma}}\end{pmatrix}\begin{pmatrix}\boldsymbol{D} & 0\\ 0 & 0\end{pmatrix}\begin{pmatrix}\tilde{\boldsymbol{\Psi}}_0^{\intercal}\\ \tilde{\boldsymbol{\Psi}}^{\intercal}\end{pmatrix},$$

where $(\tilde{\boldsymbol{\Gamma}}_0, \tilde{\boldsymbol{\Gamma}})$ and $(\tilde{\boldsymbol{\Psi}}_0, \tilde{\boldsymbol{\Psi}})$ are orthogonal matrices with dimensions $p \times p$ and $prc \times prc$, and $\boldsymbol{D}$ is a $d \times d$ diagonal matrix with positive diagonal elements, $\tilde{\boldsymbol{\Gamma}}$ has dimension $p \times (p - d)$ and $\tilde{\boldsymbol{\Psi}}$ has dimension $prc \times (prc - d)$. We partition $\tilde{\boldsymbol{\Psi}} = (\tilde{\boldsymbol{\Psi}}_1^{\intercal}, \ldots, \tilde{\boldsymbol{\Psi}}_c^{\intercal},)^{\intercal}$, where $\tilde{\boldsymbol{\Psi}}_w$ has dimension $pr \times (prc - d)$ for $w = 1, \ldots, c$.

We further introduce the following notations: for $k = 1 \ldots, r$ and $w = 1, \ldots, c$,

$$\xi_{1(k_w)} = {}_wY_{k_w} - \boldsymbol{\beta}_{k_w} - \frac{1}{2}({}_w{}_w^{\intercal} - \mathbf{I}_p)\boldsymbol{\beta}_{k_w} - \frac{1}{2}(Y_{k_w}^2 - 1)\boldsymbol{\beta}_{k_w},$$

$$\xi_{i(k_w)} = \Big\{\varepsilon_{k_w}(\,{}_w\overset{\top}{w} - \mathbf{I}_p) - \boldsymbol{H}_{k_w} - \frac{1}{2}(\,{}_w\overset{\top}{w} - \mathbf{I}_p)\boldsymbol{H}_{k_w} - \frac{1}{2}\boldsymbol{H}_{k_w}(\,{}_w\overset{\top}{w} - \mathbf{I}_p) - \frac{1}{2}(Y_{k_w}^2 - 1)\boldsymbol{H}_{k_w}\Big\}\boldsymbol{H}_{k_w}^{i-2}\boldsymbol{\beta}_{k_w}, \quad i = 2, \ldots, p,$$

$$\xi_{k_w} = (\xi_{1(k_w)}^\top, \xi_{2(k_w)}^\top, \ldots, \xi_{p(k_w)}^\top)^\top, \qquad \xi_w = (\xi_{1_w}^\top, \xi_{2_w}^\top, \ldots, \xi_{r_w}^\top)^\top, \quad \text{and}$$

$$\boldsymbol{G}_{k_w} = \begin{pmatrix} \mathbf{I}_p & 0 \ldots & 0 & 0 \\ \boldsymbol{H}_{k_w} & \mathbf{I}_p \ldots & 0 & 0 \\ \vdots & \ddots & & \vdots \\ \boldsymbol{H}_{k_w}^{p-1} & \ldots & \boldsymbol{H}_{k_w} & \mathbf{I}_p \end{pmatrix}, \qquad \boldsymbol{G}_w = \begin{pmatrix} \boldsymbol{G}_{1_w} & 0 & \ldots & 0 \\ 0 & \boldsymbol{G}_{2_w} \ldots & & 0 \\ \vdots & & \ddots & 0 \\ 0 & 0 & \ldots & \boldsymbol{G}_{r_w} \end{pmatrix}.$$

The next proposition gives the asymptotic distribution of the test statistic $\tilde{\Lambda}_d$, again a weighted chi-squared distribution under the null hypothesis.

Proposition 4. *Define*

$$\tilde{\boldsymbol{\Delta}} = \sum_{w=1}^{c} (\tilde{\boldsymbol{\Psi}}_w \otimes \tilde{\boldsymbol{\Gamma}})^\top \boldsymbol{G}_w \mathrm{E}\,(\xi_w \xi_w^\top) \boldsymbol{G}_w^\top (\tilde{\boldsymbol{\Psi}}_w \otimes \tilde{\boldsymbol{\Gamma}}),$$

and assume that all moments involved in $\tilde{\boldsymbol{\Delta}}$ *are finite. Then, as* $n_w \to \infty$,

$$\tilde{\Lambda}_d = \sum_{i=d+1}^{p} \tilde{\lambda}_i \xrightarrow{\mathcal{L}} \sum_{i=1}^{(p-d)(prc-d)} \alpha_i K_i,$$

where $\alpha_1, \ldots, \alpha_{(p-d)(prc-d)}$ *are the eigenvalues of the matrix* $\tilde{\boldsymbol{\Delta}}$*, and* $K_1, \ldots, K_{(p-d)(prc-d)}$ *are independent identically distributed random variables with a* χ_1^2 *distribution.*

Comparing the above partial IHT estimator with the OLS-based estimator, the new method further relaxes the condition (C.2). That is, the distribution of $Y_{k_w}|\boldsymbol{X}_{k_w}$ can depend on more than one linear combinations of the predictors including the ones not covered by the OLS estimates. On the other hand, we also note that a relatively large sample size is required for the IHT-based method since it involves an estimation of a $(p-d)(prc-d) \times (p-d)(prc-d)$ matrix $\tilde{\boldsymbol{\Delta}}$. To help fix ideas, consider an example with $r = 4$ response variables, $p = 5$ predictors, and a binary $c = 2$ categorical variable. Besides, suppose the dimension of the CPMS is $d = 2$. Given this setup, one would need to estimate a 114×114 matrix $\tilde{\boldsymbol{\Delta}}$ to carry out the proposed IHT-based rank test. By contrast, the OLS-based rank test involves only a 18×18 matrix $\boldsymbol{\Delta}$ as given in Proposition 3. For this reason, if the sample size of the data is not large, and if the data supports (C.2) as in the case of the Berkeley Guidance Study, we would recommend the OLS-based approach.

6.2. Asymptotically efficient estimation

The methods described so far all hinge on the spectral decomposion of a $p \times h$ matrix $\boldsymbol{\theta}$ satisfying $\mathrm{Span}(\boldsymbol{\theta}) = \mathcal{S}^{(W)}_{\mathrm{E}(\boldsymbol{Y}|)}$. For the OLS-based method, $\boldsymbol{\theta} = \boldsymbol{\beta}$ with $h = rc$; for the IHT-based method, $\boldsymbol{\theta} = \boldsymbol{M}$ with $h = prc$. Following the recent development of Cook and Ni (2005), one may further construct an asymptotically efficient estimator of $\mathcal{S}^{(W)}_{\mathrm{E}(\boldsymbol{Y}|)}$, by minimizing a quadratic discrepancy function,

$$F(\boldsymbol{\eta}, \boldsymbol{\gamma}) = \left\{\mathrm{vec}(\hat{\boldsymbol{\theta}}) - \mathrm{vec}(\boldsymbol{\eta}\boldsymbol{\gamma})\right\}^{\intercal} \hat{\boldsymbol{V}} \left\{\mathrm{vec}(\hat{\boldsymbol{\theta}}) - \mathrm{vec}(\boldsymbol{\eta}\boldsymbol{\gamma})\right\}, \tag{6.2}$$

over a $p \times d$ matrix $\boldsymbol{\eta}$ and a $d \times h$ matrix $\boldsymbol{\gamma}$. Here $\hat{\boldsymbol{\theta}}$ is a usual $\sqrt{n}$-consistent estimator of $\boldsymbol{\theta}$, and $\hat{\boldsymbol{V}}$ is a $\sqrt{n}$-consistent estimator of $\boldsymbol{\Lambda}^{-1}$, with $\boldsymbol{\Lambda}$ denoting the asymptotic covariance matrix of $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$. The next proposition summarizes the construction of an asymptotically efficient estimator of $\mathcal{S}^{(W)}_{\mathrm{E}(\boldsymbol{Y}|)}$.

Proposition 5. *Consider the optimization in* (6.2). *For OLS-based estimation,* $\boldsymbol{\theta} = \boldsymbol{\beta}$ *and* $\boldsymbol{\Lambda} = \mathrm{diag}\{\mathrm{E}(\boldsymbol{T}_w \boldsymbol{T}_w^{\intercal}), w = 1, \ldots, c\}$; *for IHT-based estimation,* $\boldsymbol{\theta} = \boldsymbol{M}$ *and* $\boldsymbol{\Lambda} = \mathrm{diag}\{\boldsymbol{G}_w \mathrm{E}(\xi_w \xi_w^{\intercal}) \boldsymbol{G}_w^{\intercal}, w = 1, \ldots, c\}$. *Let* $(\hat{\boldsymbol{\eta}}, \hat{\boldsymbol{\gamma}}) = \arg\min_{\boldsymbol{\eta}, \boldsymbol{\gamma}} F(\boldsymbol{\eta}, \boldsymbol{\gamma})$ *and* $\hat{F} = F(\hat{\boldsymbol{\eta}}, \hat{\boldsymbol{\gamma}})$. *Then,*

(a) $\sqrt{n}\{\mathrm{vec}(\hat{\boldsymbol{\eta}}\hat{\boldsymbol{\gamma}}) - \mathrm{vec}(\boldsymbol{\eta}\boldsymbol{\gamma})\}$ *converges asymptotically to a normal distribution with mean 0 and covariance matrix* $\Delta_\theta(\Delta_\theta^{\intercal}\boldsymbol{\Lambda}^{-1}\Delta_\theta)^{-}\Delta_\theta^{\intercal}$, *where* $\Delta_\theta = (\boldsymbol{\gamma}^{\intercal} \otimes \mathbf{I}_p, \mathbf{I}_h \otimes \boldsymbol{\eta})$;

(b) $n\hat{F}$ *follows asymptotically a* $\chi^2_{(p-d)(h-d)}$ *distribution;*

(c) *Span*$(\hat{\boldsymbol{\eta}})$ *is a consistent and asymptotically efficient estimator of Span*$(\boldsymbol{\theta}) = \mathcal{S}^{(W)}_{\mathrm{E}(\boldsymbol{Y}|)}$.

This proposition is a direct consequence of Theorem 2 in Cook and Ni (2005), and thus its proof is omitted here. It is also similar in spirit to Wen and Cook (2007) and Yoo and Cook (2007) in the context of multivariate response regression. Based on Proposition 5, we note the following. First, the asymptotically efficient estimator can be constructed for both the OLS-based and IHT-based approaches, which share a common framework and corresponding specific forms of $\boldsymbol{\Lambda}$ as given in the proposition. Secondly, the conclusion in Proposition 5 (b) allows one to conduct an alternative test to determine the dimension of the CPMS. See also Cook and Ni (2005) for more detailed discussion on asymptotically efficient estimation.

7. Conclusion

In this article we have developed a dimension-reduction-based rank test for the multivariate reduced-rank regression with presence of both quantitative and

qualitative predictors. The proposed methods apply to a wide class of reduced-rank models permitting interactions, nonlinear means, and heteroscedastic variances. The methods are mostly useful at the outset of an analysis, by allowing visualization of regressions in low-dimensional projections, and providing a relatively small set of composite predictors for subsequent model formulation. The strategy proposed in this article is applicable to most central mean subspace estimators, though we only implemented it with OLS and IHT-based approaches. Asymptotically efficient estimation is also briefly discussed.

Acknowledgement

The authors are grateful to the Editor, an associate editor, and two referees for their constructive comments and suggestions, which have greatly improved the paper. Li's research was partially supported by National Science Foundation grant DMS 0706919.

References

Anderson, T. W. (1951). Estimating linear restrictions on regression coefficients for multivariate normal distribution. *Ann. Math. Statist.* **22**, 327-351.

Bentler, P. M. and Xie, J. (2000). Corrections to test statistics in principal Hessian direction. *Statist. Probab. Lett.* **47**, 381-389.

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

Cook, R. D. and Li, B. (2002). Dimension reduction for the conditional mean in regression. *Ann. Statist.* **30**, 455-474.

Cook, R. D. and Li, B. (2004). Determining the dimension of iterative Hessian transformation. *Ann. Statist.* **32**, 2501-2531.

Cook, R. D. and Nachtsheim, C. J. (1994). Re-weighting to achieve elliptically contoured covariates in regression. *J. Amer. Statist. Assoc.* **89**, 592-600.

Cook, R. D. and Ni, L. (2005). Sufficient dimension reduction via inverse regression: a minimum discrepancy approach. *J. Amer. Statist. Assoc.* **100**, 410-428.

Cook, R. D. and Setodji, C. M. (2003). A model-free test for reduced rank in multivariate regression. *J. Amer. Statist. Assoc.* **98**, 340–351.

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

Izenman, A. L. (1975). Reduced-rank regression for multivariate linear model. *J. Multivariate Anal.* **5**, 248-264.

Li, B., Cook, R. D. and Chiaromonte, F. (2003). Dimension reduction for the conditional mean in regressions with categorical predictors. *Ann. Statist.* **31**, 1636-1668.

Li, L., Cook, R. D. and Nachtsheim, C. J. (2004). Cluster-based estimation for sufficient dimension reduction. *Comput. Statist. Data Anal.* **47**, 175-193.

Li, K. C. (1991). Sliced inverse regression for dimension reduction (with discussion). *J. Amer. Statist. Assoc.* **86**, 316-342.

Li, K. C. (1992). On principal Hessian directions for data visualization and dimension reduction: Another application of Stein's lemma. *J. Amer. Statist. Assoc.* **87**, 1025-1040.

Li, K. C. and Duan, N. (1989), Regression analysis under link violation. *Ann. Statist.* **17** , 1009–1052.

Reinsel, G. C. and Velu, R. P. (1998). *Multivariate Reduced-rank Regression*. Springer, New York.

Satterthwaite, F. E. (1941). Synthesis of variance. *Psychometrika* **6**, 309-316.

Tuddenham, R. D. and Snyder, M. M. (1954). Physical growth of California boys and girls from birth to eighteen years. *University of California Publications in Child Development* **1**, 183-364.

Wen, X. and Cook, R. D. (2007). Optimal sufficient dimension reduction in regressions with categorical predictors. *J. Statist. Plann. Inference* **137**, 1961-1978.

Xia, Y., Tong, H., Li, W. K. and Zhu, L. X. (2002). An adaptive estimation of dimension reduction space. *J. Roy. Statist. Soc. Ser. B* **64**, 363-410.

Yoo, J. K. and Cook, R. D. (2007). Optimal sufficient dimension reduction for the conditional mean in multivariate regression. *Biometrika* **94**, 231-242.

RAND, 1776 Main St, P.O Box 2138, Santa Monica, CA 90407 U.S.A.

E-mail: setodji@rand.org

Department of Statistics, North Carolina State University, Box 8203, Raleigh, NC 27695 U.S.A.

E-mail: li@stat.ncsu.edu

(Received March 2007; accepted July 2008)