Moment Based Dimension Reduction for Multivariate Response Regression

Xiangrong Yin* Efstathia Bura†

January 20, 2005

Abstract

Dimension reduction aims to reduce the complexity of a regression without requiring a pre-specified model. In the case of multivariate response regressions, covariance-based estimation methods for the k-th moment based dimension reduction subspaces circumvent slicing and nonparametric estimation so that they are readily applicable to multivariate regression settings. In this article, the covariance-based method developed by Yin and Cook (2002) for univariate regressions is extended to multivariate response regressions and a new method is proposed. Simulated and real data examples illustrating the theory are presented.

Key Words: Central k-th moment subspace, Central mean subspaces, Permutation tests

*Department of Statistics, 204 Statistics Building, University of Georgia, Athens, GA 30602.
†Department of Statistics, George Washington University, Washington, DC 20052.

1 Introduction

Let $\mathbf{Y}$ be a $q \times 1$ response vector, and $\mathbf{X}$ be a $p \times 1$ vector of predictors. The classical multivariate regression model is of the form

$$\mathbf{Y} = \boldsymbol{\alpha} + \mathbf{B}^T\mathbf{X} + \boldsymbol{\epsilon}, \tag{1}$$

where $\boldsymbol{\alpha}$ is a $q \times 1$ vector of intercepts, $\mathbf{B}$ is $p \times q$ matrix of unknown regression coefficients, and $\boldsymbol{\epsilon}$ is an error vector that is independent of $\mathbf{X}$, has mean 0 and constant variance.

The *dimension* of the regression problem, i.e. how many predictors and in what form they should be used in the model without losing relevant information on $\mathbf{Y}$, has always been at the epicenter of regression analysis. For multivariate regressions, Anderson (1951), in an attempt to formulate the notion of the dimension of the regression, introduced reduced-rank regression. Reduced-rank regression hypothesizes that $\text{rank}(\mathbf{B}) < \min(q, p)$. Such models are often used when there is need to reduce the number of parameters in (1). They have wide applicability in fields such as chemometrics (Frank and Friedman 1993), psychometrics (Anderson and Rubin 1956), and econometrics (Velu, Reinsel and Wichern 1986) among others. Under normality of the error terms, the asymptotic distributions of estimators of the coefficient matrix and asymptotic rank tests are available (Anderson 1951; Izenman 1975; Reinsel and Velu 1998, Ch. 2; Schmidli 1995, Ch. 4). Bura and Cook (2003) developed a theory for estimating the rank of the subspace spanned by the coefficient matrix without requiring normal errors and non-constant variances. However, the developed methodology in its entirety requires that the model structure of the multivariate regression be correct.

Cook and Setodji (2003) developed a test for estimating the rank in multivariate

regression when the functional form of the model is not correct imposing conditions only on the distribution of the predictor vector $\mathbf{X}$. For example, their method can accommodate the following situation: $q = 2$, $Y_1 = \alpha_1 + \mathbf{b}_1^T\mathbf{X} + \sigma_1(\mathbf{b}_1^T\mathbf{X})\epsilon_1$, $Y_2 = \alpha_2 + g_2(\mathbf{b}_2^T\mathbf{X}) + \epsilon_2$, where $(\mathbf{b}_1, \mathbf{b}_2)$ form $\mathbf{B}$, $\sigma_1(\mathbf{b}_1^T\mathbf{X})$ is a possibly non-constant variance function, the ϵ's are i.i.d. error terms with mean 0 and variance 1, and the nonlinear function g_2 can be either known or unknown. Their work, which to some degree is parallel to Li, Cook and Chiaromonte (2003), can be thought of as an extension of univariate OLS under missing link (Li and Duan, 1989) to multivariate OLS.

In general, the extension of existing dimension reduction methods such as SIR (Sliced inverse regression, Li 1991) and SAVE (Cook and Weisberg 1991) for univariate response regressions to their multivariate analogs is conceptually straightforward. On the other hand, the efficacy of the methodology based on nonparametric estimation of the inverse curves, encompassing SIR, SAVE and their variations (e.g. Schott, 1994, Velilla 1998), may be hindered by the curse of dimensionality. There has been only one other attempt to our knowledge (Bura and Cook, 2001) to tackle the same problem in multivariate regression. Their *parametric inverse regression* algorithm accommodates multivariate regressions naturally by utilizing multivariate linear models to estimate the inverse regression curves.

In this paper we consider the general multivariate regression problem without specifying a link function. We require that $F(\mathbf{Y}|\mathbf{X}) = F(\mathbf{Y}|\boldsymbol{\eta}^T\mathbf{X})$ where $\boldsymbol{\eta} = (\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_d)$ is a $p \times d$ matrix of full rank, and F is a cdf function. When $d < p$ the dimension of the problem is reduced. We focus on dimension reduction for the mean, variance and moment functions in multivariate regression and generalize and extend Cook and

Setodji's (2003) method. Cook and Li (2002) and Yin and Cook (2002) introduced the *k-th moment dimension reduction subspace(DRS)* and the *central k-th moment dimension reduction subspace(DRS)* for univariate response regressions. We adapt ideas from both these references to multivariate regression using marginal moments methods. One major advantage of the marginal moment based dimension reduction methodology is that it uses moment estimates of moments of functions of the response and predictor vectors without any recourse to nonparametric estimation and the ensuing subjective choice of slices (i.e. window width).

In Section 2 we introduce the concept of *the central k-th moment subspace* $\mathcal{S}^{(k)}_{\mathbf{Y}|\mathbf{X}}$ in multivariate regression and study its basic properties. In this setup, the main goal is to estimate the fewest linear combinations $\boldsymbol{\eta}_1^T\mathbf{X}, \ldots, \boldsymbol{\eta}_d^T\mathbf{X}$, $d \leq p$, with the property that the first k moments of $\mathbf{Y}|\mathbf{X}$ and $\mathbf{Y}|\boldsymbol{\eta}^T\mathbf{X}$ are the same, where $\boldsymbol{\eta} = (\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_d)$ is a $p \times d$ matrix of full rank. Thus, $\boldsymbol{\eta}^T\mathbf{X}$ contains all the information about the first k conditional moments of $\mathbf{Y}|\mathbf{X}$. No model for $\mathbf{Y}|\mathbf{X}$ is either required or assumed.

Regression analysis most often focuses on the conditional mean and variance of the response given the predictors. We propose two methods to estimate the directions in $\mathcal{S}^{(2)}_{\mathbf{Y}|\mathbf{X}}$ in Section 3. A small simulation study is presented in Section 4 and in Section 5 the proposed method is applied to the Minneapolis School data. The last section contains a discussion of our findings. All proofs are placed in the Appendix.

We assume throughout that the data $(\mathbf{Y}_i^T, \mathbf{X}_i^T)$, $i = 1, \ldots, n$, are i.i.d. observations on $(\mathbf{Y}^T, \mathbf{X}^T)$ with finite moments. The notation $\mathbf{U} \perp\!\!\!\perp \mathbf{V}|\mathbf{Z}$ means that the random vectors $\mathbf{U}$ and $\mathbf{V}$ are independent given any value for the random vector $\mathbf{Z}$. Subspaces will be denoted by $\mathcal{S}$, and $\mathcal{S}(\mathbf{B})$ signifies the subspace of $\mathcal{R}^t$ spanned by the columns of

the $t \times u$ matrix $\mathbf{B}$.

2 k-th Moment Dimension Reduction Subspaces

When considering conditional moments, dimension reduction hinges upon finding a $p \times d$ matrix $\boldsymbol{\eta}$, $d \leq p$, so that the random vector $\boldsymbol{\eta}^T \mathbf{X}$ contains all the information about $\mathbf{Y}$ that is available from $\mathrm{E}(\mathbf{Y}|\mathbf{X})$,$\mathrm{Var}(\mathbf{Y}|\mathbf{X})$,..., $M^{(k)}(\mathbf{Y}|\mathbf{X})$, where $M^{(k)}(\mathbf{Y}|\mathbf{X})$ for $k \geq 2$ is the centered k-th conditional moment,

$$M^{(k)}(\mathbf{Y}|\mathbf{X}) = \mathrm{E}((\mathbf{Y} - \mathrm{E}(\mathbf{Y}|\mathbf{X})) \otimes \cdots \otimes (\mathbf{Y} - \mathrm{E}(\mathbf{Y}|\mathbf{X})) \otimes (\mathbf{Y} - \mathrm{E}(\mathbf{Y}|\mathbf{X}))^T|\mathbf{X}).$$

The symbol $\otimes$ indicates the Kronecker product that appears $k-1$ times in the definition. For notational convenience, we let $M^{(1)}(\mathbf{Y}|\mathbf{X})$ stand for $\mathrm{E}(\mathbf{Y}|\mathbf{X})$.

Definition 1 *If*

$$\mathbf{Y} \perp\!\!\!\perp \{M^{(1)}(\mathbf{Y}|\mathbf{X}), ..., M^{(k)}(\mathbf{Y}|\mathbf{X})\}|\boldsymbol{\eta}^T \mathbf{X},$$

then $\mathcal{S}(\boldsymbol{\eta})$ is called a k-th moment dimension reduction subspace (DRS) for the regression of $\mathbf{Y}$ on $\mathbf{X}$.

This definition derives from Cook and Li (2002) and Cook and Setodji (2003) when $k = 1$, and Yin and Cook (2002) when $k > 1$. The next proposition generalizes Proposition 1 of Yin and Cook (2002) and Cook and Setodji (2003). It provides equivalent conditions for the conditional independence used in Definition 1. It is proven in the Appendix.

Proposition 1 *Let $g_j(\mathbf{Y}) = \mathbf{Y} \otimes \cdots \otimes \mathbf{Y} \otimes \mathbf{Y}^T$ with the Kronecker product appearing $j-1$ times. The following statements are equivalent:*

(i) $\mathbf{Y} \perp\!\!\!\perp \{M^{(1)}(\mathbf{Y}|\mathbf{X}), ..., M^{(k)}(\mathbf{Y}|\mathbf{X})\}|\boldsymbol{\eta}^T\mathbf{X}$.

(ii) *COV*$(g_j(\mathbf{Y}), M^{(j)}(\mathbf{Y}|\mathbf{X})|\boldsymbol{\eta}^T\mathbf{X}) = 0$, *for* $j = 1, ..., k$.

(iii) $M^{(j)}(\mathbf{Y}|\mathbf{X})$ *is a function of* $\boldsymbol{\eta}^T\mathbf{X}$. *Equivalently,* $\mathrm{E}(g_j(\mathbf{Y})|\mathbf{X})$ *is a function of* $\boldsymbol{\eta}^T\mathbf{X}$ *for* $j = 1, ..., k$.

(iv) *COV*$(g_j(\mathbf{Y}), f(\mathbf{X})|\boldsymbol{\eta}^T\mathbf{X}) = 0$ *for* $j = 1, ..., k$ *and any function* $f(\mathbf{X})$.

The *central k-th moment subspace (CKMS)*, denoted by $\mathcal{S}^{(k)}_{\mathbf{Y}|\mathbf{X}}$, is defined to be the intersection over all k-th moment DRSs if the intersection itself is a DRS. If it exists, it is the smallest k-th moment DRS. The existence of the CKMS can be guaranteed under various mild conditions. For example, if the support of $\mathbf{X}$ is open and convex, then the CKMS and the central subspace exist (see Cook, 1998, p. 108–only the univariate response case is presented but its extension to multivariate response regressions is straightforward). For most data sets existence is not a crucial practical issue and thus we assume it throughout the rest of this article. The following relations hold:

$$\mathcal{S}^{(1)}_{\mathbf{Y}|\mathbf{X}} \subseteq \cdots \subseteq \mathcal{S}^{(k)}_{\mathbf{Y}|\mathbf{X}}.$$

As the focus in multivariate regression is on the first two moments of the conditional distribution of the response given the predictor, the central subspace $\mathcal{S}^{(2)}_{\mathbf{Y}|\mathbf{X}}$ is of particular interest.

Let $\mathbf{V} = \mathbf{A}^T\mathbf{X}$ for some invertible matrix $\mathbf{A}$. Then $\mathcal{S}^{(k)}_{\mathbf{Y}|\mathbf{V}} = \mathbf{A}^{-1}\mathcal{S}^{(k)}_{\mathbf{Y}|\mathbf{X}}$. Consequently, $\mathcal{S}^{(k)}_{\mathbf{Y}|\mathbf{X}} = \boldsymbol{\Sigma}^{-1/2}_{\mathbf{X}}\mathcal{S}^{(k)}_{\mathbf{Y}|\mathbf{Z}}$ which implies that there is no loss of generality in standardizing $\mathbf{X}$ to have mean zero and identity covariance matrix. For the majority of the subsequent developments we work in terms of the standardized predictor $\mathbf{Z}$.

3 Methodology and Estimation

In this section we develop methods for finding directions in the central k-th moment subspace. Although the extension of the central k-th moment subspace from the univariate to the multivariate response regression is straightforward, estimation methods such as sliced inverse regression (SIR, Li, 1991) and sliced average variance estimate (SAVE, Cook and Weisberg, 1991) may not be efficiently extended. In the univariate response case, by switching the roles of response and predictors the problem becomes manageable as we can then slice the univariate response to obtain simple nonparametric estimates of the inverse regressions. In the multivariate case, however, such a device does not result in p univariate regressions and the curse of dimensionality may remain an estimation hurdle. Still, dimension reduction can be carried out by employing multivariate response versions of marginal methods such as COV_k (Yin and Cook, 2002) and PHD (Li, 1992).

3.1 Population Structure

The following proposition, an extension of the univariate response version of Yin and Cook (2002), indicates how to find vectors in $\mathcal{S}^{(k)}_{\mathbf{Y}|\mathbf{Z}}$ using the covariance of $\mathbf{Z}$ and a polynomial in $\mathbf{Y}$. Its proof is given in the appendix.

Proposition 2 *Let $\boldsymbol{\gamma}$ be a basis for $\mathcal{S}^{(k)}_{\mathbf{Y}|\mathbf{Z}}$, and assume that $\mathrm{E}(\mathbf{Z}|\boldsymbol{\gamma}^T\mathbf{Z})$ is linear in $\boldsymbol{\gamma}^T\mathbf{Z}$. If $f^{(j)}(\mathbf{Y}) = \mathbf{Y}^T \otimes \cdots \otimes \mathbf{Y}^T$ with the Kronecker product appearing $j-1$ times, then for all $j = 1, ..., k$,*

$$\mathcal{S}(\mathrm{E}(\mathbf{Z}f^{(j)}(\mathbf{Y}))) \subset \mathcal{S}^{(k)}_{\mathbf{Y}|\mathbf{Z}}.$$

The following example highlights the potential usefulness of this proposition.

Example 1: Let $z_1, z_2, z_3, \epsilon_1, \epsilon_2$ be independent standard normal random variables. Let $\mathbf{Z} = (z_1, z_2, z_3)^T$ and $\mathbf{Y} = (y_1, y_2)^T$ with $y_1 = z_1 + z_1 z_2 + \epsilon_1$ and $y_2 = 1 + z_1 + e^{z_2}\epsilon_2$. The structural dimension of this bivariate regression is two. Observe that

$$\mathrm{E}(\mathbf{Z}\mathbf{Y}^T) = \begin{pmatrix} 1 & 1 \\ 0 & 0 \\ 0 & 0 \end{pmatrix},$$

but

$$\mathrm{E}(\mathbf{Z}\mathbf{Y}^T \otimes \mathbf{Y}^T) = \begin{pmatrix} 0 & 1 & 1 & 2 \\ 2 & 1 & 1 & 2e^2 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

We note that while OLS finds one direction, the method of Proposition 2 finds the second when $k = 2$. However, methods based on Proposition 2 are designed to find linear structure using the mean, the variance and higher order momemts of the regression function. They are analogous to the kernel matrix of the SIR dimension reduction methodology and hence they fail when the mean function is even and symmetric as in the following example.

Example 2: Let $z_1, z_2, z_3, \epsilon_1, \epsilon_2$ be independent standard normal random variables. Let $\mathbf{Z} = (z_1, z_2, z_3)^T$ and $\mathbf{Y} = (y_1, y_2)^T$ with $y_1 = z_1^2 + \epsilon_1$ and $y_2 = z_1^2 + z_2^2 + \epsilon_2$. The structural dimension of this regression is also two. Observe that Proposition 2 detects nothing since all kernel matrices are null. To amend this, a method similar to univariate response PHD method can be used as in Proposition 3 below, deriving from Theorem 2

of Cook and Li (2002) and Yin and Cook (2003), to handle the symmetric and even mean function case.

Proposition 3 *Let $\boldsymbol{\gamma}$ be a basis for $\mathcal{S}^{(k)}_{\mathbf{Y}|\mathbf{Z}}$, and assume that $\mathrm{E}(\mathbf{Z}|\boldsymbol{\gamma}^T\mathbf{Z})$ is linear in $\boldsymbol{\gamma}^T\mathbf{Z}$. If $f^{(j)}(\mathbf{Y}) = \mathbf{Y}^T \otimes \cdots \otimes \mathbf{Y}^T - \mathrm{E}(\mathbf{Y}^T \otimes \cdots \otimes \mathbf{Y}^T)$, where the Kronecker product appears $j-1$ times, and is uncorrelated with Var$(\mathbf{Z}|\boldsymbol{\gamma}^T\mathbf{Z})$, then for all $j = 1, ..., k$,*

$$\mathcal{S}(\mathrm{E}[\mathbf{Z}\mathbf{Z}^T \otimes f^{(j)}(\mathbf{Y})]) \subset \mathcal{S}^{(k)}_{\mathbf{Y}|\mathbf{Z}}.$$

It may be of interest to note that this choice for $f^{(1)}(\mathbf{Y})$ leads to a multivariate version of PHD.

Example 2 continue: The method of Proposition 3 finds all directions:

$$\mathrm{E}[\mathbf{Z}\mathbf{Z}^T \otimes (\mathbf{Y}^T - \mathrm{E}(\mathbf{Y}^T))] = \begin{pmatrix} 2 & 2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

3.2 Methodology

Here we present methods for estimating $\mathcal{S}^{(2)}_{\mathbf{Y}|\mathbf{Z}}$, the subspace spanned by the conditional mean and conditional variance.

Define the population kernel matrix

$$\mathbf{K}_{21} = (\mathrm{E}(\mathbf{Z}\mathbf{Y}^T), \mathrm{E}(\mathbf{Z}\mathbf{Y}^T \otimes \mathbf{Y}^T)).$$

Proposition 2 implies $\mathcal{S}(\mathbf{K}_{21}) \subseteq \mathcal{S}^{(2)}_{\mathbf{Y}|\mathbf{Z}}$. Cook and Setodji (2003) used $\mathrm{E}(\mathbf{Z}\mathbf{Y}^T)$ as a multivariate version of OLS for estimating part of the central mean subspace. The

kernel matrix $\mathbf{K}_{21}$ we propose generalizes and extends their method for finding part of the central second moment subspace.

Also, define another population kernel matrix

$$\mathbf{K}_{22} = \mathrm{E}[\mathbf{Z}\mathbf{Z}^T \otimes [\mathbf{Y}^T - \mathrm{E}(\mathbf{Y}^T)]].$$

Proposition 3 yields $\mathcal{S}(\mathbf{K}_{22}) \subseteq \mathcal{S}^{(1)}_{\mathbf{Y}|\mathbf{Z}}$. Hence, for $i = 1, 2$, the subspace spanned by the left singular vectors of $\mathbf{K}_{2i}$ corresponding to its non-zero singular values arranged in descending order is a subspace of $\mathcal{S}^{(2)}_{\mathbf{Y}|\mathbf{Z}}$. The matrix $\mathbf{K}_{22}$ is the extended version of PHD (Li 1992) to multivariate response regression.

For application in practice, let $\hat{\mathbf{K}}_{2i}$ be the moment estimate of $\mathbf{K}_{2i}$. Then, if $d = \dim(\mathcal{S}(\mathbf{K}_{2i}))$, the subspace spanned by the left singular vectors of $\hat{\mathbf{K}}_{2i}$ corresponding to the d largest singular values is consistent for $\mathcal{S}^{(2)}_{\mathbf{Y}|\mathbf{Z}}$. Inference methods about d are discussed in section 3.3.

The response is usually standardized in order to avoid numerical instability. The corresponding kernel matrices with $\mathbf{Y}$ centered and scaled are given by

$$\mathbf{K}_{21c} = (\mathrm{E}(\mathbf{Z}\mathbf{W}^T), \mathrm{E}(\mathbf{Z}\mathbf{W}^T \otimes \mathbf{W}^T)),$$

and

$$\mathbf{K}_{22c} = \mathrm{E}[\mathbf{Z}\mathbf{Z}^T \otimes [\mathbf{W}^T - \mathrm{E}(\mathbf{W}^T)]$$

where $\mathbf{W} = \mathbf{V}(\mathbf{Y})^{-1/2}(\mathbf{Y} - \mathrm{E}(\mathbf{Y}))$ and $\mathbf{V}(\mathbf{Y}) = diag(\mathrm{Var}(Y_i))$.

It follows logically that one would consider using the kernel matrix $\mathrm{E}[\mathbf{Z}\mathbf{Z}^T \otimes (\mathbf{Y}^T \otimes \mathbf{Y}^T - \mathrm{E}(\mathbf{Y}^T \otimes \mathbf{Y}^T))]$ as well, since the subspace spanned by the columns of this matrix is also in $\mathcal{S}^{(2)}_{\mathbf{Y}|\mathbf{X}}$. However, in practice $\mathrm{E}(\mathbf{Z}\mathbf{Y}^T)$ and $\mathrm{E}(\mathbf{Z}\mathbf{Y}^T \otimes \mathbf{Y}^T)$ do better at detecting

linear trends or odd functions, while $\mathrm{E}[\mathbf{Z}\mathbf{Z}^T \otimes (\mathbf{Y}^T - \mathrm{E}(\mathbf{Y}^T))]$ is better at revealing symmetric trends or even functions. Using $\mathrm{E}(\mathbf{Z}\mathbf{Z}^T \otimes (\mathbf{Y}^T \otimes \mathbf{Y}^T - \mathrm{E}(\mathbf{Y}^T \otimes \mathbf{Y}^T)))$ will most likely yield directions that overlap with the ones already found. Also, with the response being a vector, the high dimension of this matrix may perhaps weaken its usefulness.

3.3 Estimation Procedure

Let $\hat{\mathbf{\Sigma}}_{\mathbf{X}}$ denote the usual moment estimate of $\mathbf{\Sigma}_{\mathbf{X}}$, and let us define the standardized observations

$$\hat{\mathbf{Z}}_i = \hat{\mathbf{\Sigma}}_{\mathbf{X}}^{-1/2}(\mathbf{X}_i - \bar{\mathbf{X}}),\ i = 1, \ldots, n,$$

where $\bar{\mathbf{X}}$ is the sample mean of the predictor vector. Let $\hat{\mathbf{V}}(\mathbf{Y})$ denote the moment estimate of $\mathbf{V}(\mathbf{Y}) = diag(\mathrm{Var}(Y_i))$, and define the standardized responses by

$$\hat{\mathbf{W}}_i = \hat{\mathbf{V}}(\mathbf{Y})^{-1/2}(\mathbf{Y}_i - \bar{\mathbf{Y}}),\ i = 1, \ldots, n,$$

where $\bar{\mathbf{Y}}$ is the sample mean of the response vector.

The sample kernel matrices corresponding to $\mathbf{K}_{21c}$ and $\mathbf{K}_{22c}$ are given by

$$\hat{\mathbf{K}}_{21c} = (\frac{1}{n}\sum_{i=1}^{n} \hat{\mathbf{Z}}_i \hat{\mathbf{W}}_i^T, \frac{1}{n}\sum_{i=1}^{n} \hat{\mathbf{Z}}_i \hat{\mathbf{W}}_i^T \otimes \hat{\mathbf{W}}_i^T),$$

and

$$\hat{\mathbf{K}}_{22c} = \frac{1}{n}\sum_{i=1}^{n} \hat{\mathbf{Z}}_i \hat{\mathbf{Z}}_i^T \otimes \hat{\mathbf{W}}_i^T.$$

Let $\hat{\mathbf{K}}_{2c}$ be either $\hat{\mathbf{K}}_{21c}$ or $\hat{\mathbf{K}}_{22c}$, $\hat{s}_1 \geq \hat{s}_2 \geq \cdots \geq \hat{s}_r$ be the singular values of $\hat{\mathbf{K}}_{2c}$, and $\hat{l}_1, ..., \hat{l}_r$ be the corresponding left singular vectors, where $r = \min\{p, q + q^2\}$. Then $\hat{\mathcal{S}}(\mathbf{K}_{2c}) = \mathcal{S}(\hat{\mathbf{K}}_{2c}) = \mathcal{S}(\hat{l}_1, ...\hat{l}_d)$.

The estimation of d is carried out via a sequential testing procedure suggested by Li (1991) that uses the test statistic

$$\hat{\Lambda}_m = n \sum_{j=m+1}^{r} \hat{s}_j^2.$$

Starting at $m = 0$, we compare $\hat{\Lambda}_m$ to the percentage points of its distribution under the hypothesis $d = m$ and determine the p-value p_m. If p_m is larger than a pre-set cut-off value there is not sufficient information to contradict the null hypothesis. If it is smaller, we conclude that $d > m$, increment m by 1 and repeat the procedure. The estimate $\hat{d} = m$ follows when p_{m-1} is smaller than the cut-off, implying that $d > m-1$, while p_m is larger. The estimate of $\mathcal{S}(\mathbf{K}_{2c})$ is then given by $\mathcal{S}\{\hat{l}_1, ..., \hat{l}_{\hat{d}}\}$. These vectors can be back transformed to $\hat{\boldsymbol{\eta}}_j$ in the original scale of the predictors.

Either the asymptotic distribution of $\hat{\Lambda}_d$ or a nonparametric alternative is required to implement this procedure in practice. Herein we adapt the permutation test suggested by Cook and Weisberg (1991) and further developed by Cook and Yin (2001) and Yin and Cook (2002) in the univariate response case. We assume without loss of generality that the kernel matrix $\mathbf{K}$ is a $p \times p$ positive semidefinite symmetric matrix. Starting with a non-symmetric kernel $\mathbf{A}$, for example, we can set $\mathbf{K} = \mathbf{A}\mathbf{A}^T$. The test statistic is then

$$\hat{\Lambda}_m = n \sum_{j=m+1}^{p} \hat{\lambda}_j,$$

where the $\hat{\lambda}_j$ are the eigenvalues of $\hat{\mathbf{K}}$. Note that $\mathcal{S}^{(k)}_{\mathbf{Y}|\mathbf{X}} = \boldsymbol{\Sigma}_{\mathbf{X}}^{-1/2}\mathcal{S}^{(k)}_{\mathbf{Y}|\mathbf{Z}}$ and $\mathcal{S}^{(k)}_{\mathbf{W}|\mathbf{Z}} = \mathcal{S}^{(k)}_{\mathbf{Y}|\mathbf{Z}}$, thus the dimensionality does not change under standardized variables.

Let $\mathbf{U} = (\mathbf{u}_j)$ denote the $p \times p$ matrix of eigenvectors $\mathbf{u}_j$ of the population kernel matrix $\mathbf{K}$, let $d = \dim(\mathcal{S}(\mathbf{K}))$ and assume that $\mathcal{S}(\mathbf{K}) = \mathcal{S}_{\mathbf{W}|\mathbf{Z}}$, the central subspace of

$\mathbf{W}$ given $\mathbf{Z}$.

Consider testing the hypothesis that $d \leq m$ versus $d > m$. Partition $\mathbf{U} = (\mathbf{U}_1, \mathbf{U}_2)$ where $\mathbf{U}_1$ is $p \times m$ so that under the null hypothesis $\mathcal{S}_{\mathbf{W}|\mathbf{Z}} \subseteq \mathcal{S}(\mathbf{U}_1)$. Based on Proposition 3 (Yin and Cook 2002) we can gain information on d by testing the null hypothesis $(\mathbf{W}, \mathbf{U}_1^T\mathbf{Z}) \perp\!\!\!\perp \mathbf{U}_2^T\mathbf{Z}$ as follows:

Step 1. Calculate $\hat{\mathbf{A}}$ to be either $\hat{\mathbf{K}}_{21c}$ or $\hat{\mathbf{K}}_{22c}$, and then compute $\hat{\mathbf{K}}$. Calculate $\hat{\Lambda}_m$ for $m = 0, ..., p-1$.

Step 2. For any fixed m, we randomly permute $(\hat{\mathbf{W}}_i, \hat{\mathbf{U}}_1^T\hat{\mathbf{Z}}_i)$ over $i = 1, ..., n$, and recalculate $\hat{\Lambda}_m$ from the permuted data. Repeat this a large number of times to obtain the null distribution for $\hat{\Lambda}_m$ (we use 1000 permutations in the application sections that follow).

Step 3. Calculate the precentage of values of $\hat{\Lambda}_m$ in step 2 greater than $\hat{\Lambda}_m$ in step 1. This is the p-value of the test.

Step 4. Repeat Steps 2 and 3 for all $m = 0...., p-1$ to get the p-values for all p tests for dimension.

The p-values generated above are then compared to a pre-selected significance level (usually, .05) sequentially. The estimated dimension is d, if $p_m < .05$ for all $m = 0, ..., d-1$, and $p_d > .05$.

4 Simulation Study

The simulation study is based on examples 1 and 2 in section 3.1. The regressions in both examples are two-dimensional. The simulation study illustrates the potential

usefulness of the two kernel matrices.

Tables 1 and 2 present estimates of the power and size of the test of dimension for the regressions in examples 1 and 2, respectively. The power and level estimates were computed as the proportion of p-values smaller than the nominal 5% alpha level. The p-values were computed by a permutation test as described in Section 3.3. One thousand permutations were used throughout. All calculations were carried out in R using the package *dr* (Weisberg, 2003). The code we developed for computing the kernel matrices is available upon request from the authors.

We remind the reader that the test statistic based on the $\hat{\mathbf{K}}_{21c}$ kernel is expected to correctly estimate the dimension for example 1, whereas the test statistic using the kernel $\mathbf{K}_{22c}$ should estimate correctly the dimension of the regression in example 2. Also, the former should fail in the case of example 2.

Table 1: Example 1, permutation test results for $\hat{\mathbf{K}}_{21c}$. D is number of data sets generated, n is the sample size, the power entries are in 2nd and 3rd columns while the last column contains the level.

	$d = 0$ vs $d \geq 1$	$d = 1$ vs $d \geq 2$	$d = 2$ vs $d \geq 3$
$n = 100, D = 1000$	1.000	.904	.057
$n = 200, D = 1000$	1.000	.988	.067
$n = 400, D = 1000$	1.000	1.00	.057

It can be seen from Table 1 that the test correctly estimates the dimension to be two about 90% of the time with the level being very close to the nominal 5% for a sample size of 100. If the sample size is increased the power for the second dimension tends to

one.

Table 2: Example 2, permutation test results. The number of data sets generated is denoted by D, n is the sample size, the power entries are in 2nd and 3rd columns while the last column contains the level.

	Kernel	$d = 0$ vs $d \geq 1$	$d = 1$ vs $d \geq 2$	$d = 2$ vs $d \geq 3$
$n = 100, D = 1000$	$\mathbf{K}_{22c}$	1.000	.992	.231
$n = 400, D = 500$	$\mathbf{K}_{22c}$	1.000	1.000	.148
$n = 100, D = 1000$	$\mathbf{K}_{21c}$	.986	.337	.041
$n = 400, D = 500$	$\mathbf{K}_{21c}$	1.000	.364	.048

Table 2 shows that the test based on $\hat{\mathbf{K}}_{22c}$ has an inflated level which decreases as the sample size increases. The estimation of $\hat{\mathbf{K}}_{22c}$ requires the computation of a large number of quantities which appears to adversely affect the accuracy of the test. It does though attain maximum power for both dimensions 0 and 1.

The test based on $\hat{\mathbf{K}}_{21c}$ estimates the dimension to be one. This may seem surprising at first glance as we expected $\hat{\mathbf{K}}_{21c}$ to fail in this example. By visually examining the simulated data, we observed that they were not exactly symmetric due to the variation induced by simulation. This lack of symmetry misled the $\hat{\mathbf{K}}_{21c}$ kernel so that the existence of a linear trend, even though weak, resulted in 1D estimated structure. To further investigate this phenomenon, we considered several simulated data for example 2 and proceeded to remove the points that appeared to affect the symmetry the most; that is, we eliminated points with large absolute values of either y_1 or y_2. For these reduced data, the test based on $\hat{\mathbf{K}}_{21c}$ estimated the dimension to be zero in agreement with the

developed theory. Should the sample size increase drastically, the lack of symmetry will become much less pronounced and $\hat{\mathbf{K}}_{21c}$ will fail to detect any directions in $\mathcal{S}^{(2)}_{\mathbf{Y}|\mathbf{Z}}$.

5 Minneapolis Elementary School Data

To illustrate the estimation of the dimension of a multivariate regression using the $\hat{\mathbf{K}}_{21c}$ kernel matrix, we use data on the performance of students in $n = 63$ Minneapolis Schools obtained from Cook (1998, p. 216). The $m = 4$ dimensional response vector $\mathbf{Y}$ consists of the percentages $\pi_{(\cdot)}$ of students in a school scoring above (A) and below (B) average on standardized fourth and sixth grade reading comprehension tests, $\mathbf{Y} = (\pi_{A4}, \pi_{B4}, \pi_{A6}, \pi_{B6})^T$. The percentage of students scoring about average on the test is not used in our illustration since the sum of this with A and B is 100%. Eight predictors ($p = 8$) were used to characterize a school: (1) the percentage of children receiving Aid to Families with Dependent Children (AFDC), (2) the percentage of children not living with biological parents, (3) the percentage of persons in the area below the federal poverty level, (4) percent of adults in the school area who completed high school, (5) percent of minority students, (6) percent of mobility, (7) percent of students attending school regularly, and (8) the pupil-teacher ratio.

We began the analysis by inspecting the scatterplot matrix of the eight predictors. The square-root of all percentages resulted in roughly linear scatterplots that do not exhibit either outlying points or heteroskedasticity; that is, the two assumptions for the proposed methology to apply are satisfied. We next turned to the multivariate regression of $\mathbf{Y}$ on the square-roots of the seven percentage predictors and the pupil-teacher ratio.

Table 3: Test results for the Minneapolis School Data

d	$\hat{\Lambda}$	p-value
0D vs 1D	361.54	0.000
1D vs 2D	201.65	0.074
2D vs 3D	130.27	0.179
3D vs 4D	79.82	0.373

The transformed predictor vector is denoted by $\mathbf{X}$.

We constructed the test statistics $\hat{\Lambda}_k$, $k = 1, \ldots, 7$ using the $\hat{\mathbf{K}}_{21c}$ kernel matrix. One thousand perturbations were used. The results of the perturbation test are given in Table 3.

The p-value for testing the null of 1D was 0.074. This is not significant at level 0.05 so we inferred that $d = 1$ and that a single linear combination X_0 of the predictors carries the information that $\mathbf{X}$ has to furnish about $\mathbf{Y}$. If this conclusion is reasonable then the scatterplot matrix of the four responses versus X_0 contains all the information in the eight original regressors for modeling the response vector. This scatterplot matrix can be seen in Figure 1. The plots suggest that quadratic regressions in X_0 capture the mean tendency in the responses except possibly for π_{B4} (indicated as Y4Below in the plot) for which a simple linear regression may be adequate.

Notably, the test using the $\hat{\mathbf{K}}_{22c}$ kernel matrix failed to identify the dimension. That is, it estimated the dimension to be eight, the number of predictors, resulting in no reduction. As there was no symmetry apparent in these data, the method of Proposition 3 is not helpful.

6 Discussion

According to both Propositions 2 and 3, general polynomials of $\mathbf{Y}$ in $f^{(k)}(\mathbf{Y})$ can be used as they are equivalent to the choice we made in this paper. Although we focus on $\mathcal{S}^{(k)}_{\mathbf{Y}|\mathbf{X}}$ with $k = 2$, as the primary interest in regression is limited to the first two moments of the regression curve, one can accordingly construct the corresponding kernel matrices for larger values of k, thus considering higher marginal moments between $\mathbf{Y}$ and $\mathbf{X}$. A sufficiently large sample size would be required in order for the method to be able to recover the true dimension for a large k, especially for a multivariate response.

Outliers can easily affect moment methods. In particular, in the case of a multivariate response with a large kernel matrix, one would be advised to remove outliers or "dominating" values before applying these methods. We suggest that one use the kernel $\mathbf{K}_{22}$ as complementary to $\mathbf{K}_{21}$ unless data either of large size or obvious symmetry about the mean are available.

As an aside, let us note that the methods of this article, as well as all dimension reduction methods, are applicable as model diagnostics by replacing the response $\mathbf{Y}$ with any residual vector $\mathbf{r}$.

Appendix: Justifications

Proof of Proposition 1:

$(i) \Rightarrow (ii)$, $(iii) \Rightarrow (i)$ and $(iv) \Rightarrow (ii)$ are immediate.

$\boldsymbol{(ii) \Rightarrow (iii)}$:

Suppose that for $i = 1, ..., k-1$, $M^{(i)}(Y|\mathbf{X})$ are functions of $\boldsymbol{\eta}^T\mathbf{X}$ where $k > 1$. By

expanding $M^{(k)}(\mathbf{Y}|\mathbf{X})$, we have

$$M^{(k)}(\mathbf{Y}|\mathbf{X}) = \mathrm{E}(g_k(\mathbf{Y})|\mathbf{X}) + g(\mathbf{X}),$$

where $g(\mathbf{X})$ is a function of $M^{(i)}(\mathbf{Y}|\mathbf{X})$ for $i = 1, ..., k-1$, hence $g(\mathbf{X})$ is a function of $\boldsymbol{\eta}^T\mathbf{X}$. We next show that $M^{(k)}(\mathbf{Y}|\mathbf{X})$ is a function of $\boldsymbol{\eta}^T\mathbf{X}$. This also implies that $\mathrm{E}(g_k(\mathbf{Y})|\mathbf{X})$ is a function of $\boldsymbol{\eta}^T\mathbf{X}$. The proof proceeds by first using (ii),

$$\mathrm{COV}(g_k(\mathbf{Y}), M^{(k)}(\mathbf{Y}|\mathbf{X})|\boldsymbol{\eta}^T\mathbf{X}) = 0,$$

to show that

$$\mathrm{E}[\mathrm{E}(g_k(\mathbf{Y})|\mathbf{X})M^{(k)}(\mathbf{Y}|\mathbf{X})|\boldsymbol{\eta}^T\mathbf{X}] = \mathrm{E}[\mathrm{E}(g_k(\mathbf{Y})|\mathbf{X})|\boldsymbol{\eta}^T\mathbf{X}]\mathrm{E}(M^{(k)}(\mathbf{Y}|\mathbf{X})|\boldsymbol{\eta}^T\mathbf{X}).$$

Next, by assumption, $g(\mathbf{X})$ is a function of $\boldsymbol{\eta}^T\mathbf{X}$. We then have

$$\mathrm{E}[g(\mathbf{X})M^{(k)}(\mathbf{Y}|\mathbf{X})|\boldsymbol{\eta}^T\mathbf{X}] = \mathrm{E}[g(\mathbf{X})|\boldsymbol{\eta}^T\mathbf{X}]\mathrm{E}(M^{(k)}(\mathbf{Y}|\mathbf{X})|\boldsymbol{\eta}^T\mathbf{X}).$$

Adding the last two equations, gives

$$\mathrm{E}[M^{(k)}(\mathbf{Y}|\mathbf{X})M^{(k)}(\mathbf{Y}|\mathbf{X})|\boldsymbol{\eta}^T\mathbf{X}] = \mathrm{E}[M^{(k)}(\mathbf{Y}|\mathbf{X})|\boldsymbol{\eta}^T\mathbf{X}]\mathrm{E}(M^{(k)}(\mathbf{Y}|\mathbf{X})|\boldsymbol{\eta}^T\mathbf{X}).$$

Therefore

$$\mathrm{E}[(M^{(k)}(\mathbf{Y}|\mathbf{X}))^2|\boldsymbol{\eta}^T\mathbf{X}] = (\mathrm{E}(M^{(k)}(\mathbf{Y}|\mathbf{X})|\boldsymbol{\eta}^T\mathbf{X}))^2.$$

The latter yields that $\mathrm{Var}(M^{(k)}(\mathbf{Y}|\mathbf{X})|\boldsymbol{\eta}^T\mathbf{X}) = 0$. Thus, given $\boldsymbol{\eta}^T\mathbf{X}$, $M^{(k)}(\mathbf{Y}|\mathbf{X})$ is a constant, it follows that $(ii) \Rightarrow (iii)$.

$(iii) \Rightarrow (iv)$:

$$
\begin{aligned}
&\mathrm{COV}(g_k(\mathbf{Y}), f(\mathbf{X})|\boldsymbol{\eta}^T\mathbf{X}) \\
&= \mathrm{E}(g_k(\mathbf{Y})f(\mathbf{X})|\boldsymbol{\eta}^T\mathbf{X}) - \mathrm{E}(g_k(\mathbf{Y})|\boldsymbol{\eta}^T\mathbf{X})\mathrm{E}(f(\mathbf{X})|\boldsymbol{\eta}^T\mathbf{X}) \\
&= \mathrm{E}[\mathrm{E}(g_k(\mathbf{Y})f(\mathbf{X})|\mathbf{X})|\boldsymbol{\eta}^T\mathbf{X}] - \mathrm{E}(g_k(\mathbf{Y})|\boldsymbol{\eta}^T\mathbf{X})\mathrm{E}(f(\mathbf{X})|\boldsymbol{\eta}^T\mathbf{X}) \\
&= \mathrm{E}(g_k(\mathbf{Y})|\mathbf{X})\mathrm{E}[f(\mathbf{X})|\boldsymbol{\eta}^T\mathbf{X}] - \mathrm{E}(\mathbf{Y}^k|\mathbf{X})\mathrm{E}(f(\mathbf{X})|\boldsymbol{\eta}^T\mathbf{X}) \\
&= 0.
\end{aligned}
$$

The third equality follows from the condition (iii). □

For the two proofs that follow, let $P_{\boldsymbol{\gamma}}$ denote the projection operator for $\mathcal{S}(\boldsymbol{\gamma})$ with respect to the usual inner product.

Proof of Proposition 2: Using Proposition 1, and the linearity condition for any $j = 1, ..., k$,

$$
\begin{aligned}
\mathrm{E}(\mathbf{Z}f^{(j)}(\mathbf{Y})) &= \mathrm{E}[\mathrm{E}(\mathbf{Z}f^{(j)}(\mathbf{Y})|\mathbf{Z})] \\
&= \mathrm{E}[\mathbf{Z}\mathrm{E}(f^{(j)}(\mathbf{Y})|\boldsymbol{\gamma}^T\mathbf{Z})] \\
&= \mathrm{E}[\mathrm{E}(\mathbf{Z}|\boldsymbol{\gamma}^T\mathbf{Z})\mathrm{E}(f^{(j)}(\mathbf{Y})|\boldsymbol{\gamma}^T\mathbf{Z})] \\
&= P_{\boldsymbol{\gamma}}\mathrm{E}(\mathbf{Z}f^{(j)}(\mathbf{Y})).
\end{aligned}
$$

□

Proof of Proposition 3: Using Proposition 1, the linearity and uncorrelated condi-

tions for any $j = 1, ..., k$,

$$
\begin{aligned}
& \mathrm{E}[\mathbf{Z}\mathbf{Z}^T \otimes f^{(j)}(\mathbf{Y})] \\
= \; & \mathrm{E}[\mathrm{E}(\mathbf{Z}\mathbf{Z}^T \otimes f^{(j)}(\mathbf{Y})|\mathbf{Z})] \\
= \; & \mathrm{E}[\mathbf{Z}\mathbf{Z}^T \otimes \mathrm{E}(f^{(j)}(\mathbf{Y})|\boldsymbol{\gamma}^T\mathbf{Z})] \\
= \; & \mathrm{E}[\mathrm{E}(\mathbf{Z}\mathbf{Z}^T|\boldsymbol{\gamma}^T\mathbf{Z}) \otimes \mathrm{E}(f^{(j)}(\mathbf{Y})|\boldsymbol{\gamma}^T\mathbf{Z})] \\
= \; & \mathrm{E}[(\mathrm{Var}(\mathbf{Z}|\boldsymbol{\gamma}^T\mathbf{Z}) + P_{\boldsymbol{\gamma}}\mathbf{Z}\mathbf{Z}^T P_{\boldsymbol{\gamma}}) \otimes \mathrm{E}(f^{(j)}(\mathbf{Y})|\boldsymbol{\gamma}^T\mathbf{Z})] \\
= \; & \mathrm{E}[(\mathrm{Var}(\mathbf{Z}|\boldsymbol{\gamma}^T\mathbf{Z}) \otimes \mathrm{E}(f^{(j)}(\mathbf{Y})|\boldsymbol{\gamma}^T\mathbf{Z})] \\
& +\mathrm{E}[(P_{\boldsymbol{\gamma}}\mathbf{Z}\mathbf{Z}^T P_{\boldsymbol{\gamma}}) \otimes \mathrm{E}(f^{(j)}(\mathbf{Y})|\boldsymbol{\gamma}^T\mathbf{Z})] \\
= \; & \mathrm{E}[\mathrm{E}(\mathrm{Var}(\mathbf{Z}|\boldsymbol{\gamma}^T\mathbf{Z}) \otimes f^{(j)}(\mathbf{Y})|\boldsymbol{\gamma}^T\mathbf{Z})] \\
& +\mathrm{E}[\mathrm{E}((P_{\boldsymbol{\gamma}}\mathbf{Z}\mathbf{Z}^T P_{\boldsymbol{\gamma}}) \otimes f^{(j)}(\mathbf{Y})|\boldsymbol{\gamma}^T\mathbf{Z})] \\
= \; & \mathrm{E}(\mathrm{Var}(\mathbf{Z}|\boldsymbol{\gamma}^T\mathbf{Z}) \otimes f^{(j)}(\mathbf{Y})) \\
& +\mathrm{E}((P_{\boldsymbol{\gamma}}\mathbf{Z}\mathbf{Z}^T P_{\boldsymbol{\gamma}}) \otimes f^{(j)}(\mathbf{Y})) \\
= \; & P_{\boldsymbol{\gamma}}\mathrm{E}[\mathbf{Z}\mathbf{Z}^T \otimes f^{(j)}(\mathbf{Y})]P_{\boldsymbol{\gamma}} \otimes I.
\end{aligned}
$$

□

Acknowledgments

We would like to thank the referees whose comments led to many improvements. The second author was supported in part by the National Foundation Grant DMS-0204563. She would also like to thank the Department of Statistics of the London School of Economics where most of her sabbatical leave was spent.

References

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

Anderson, T. W., Rubin, H., 1956. Statistical inference in factor analysis. In: J. Neyman (Ed.), Proceedings of the third Berkeley Symposium on Mathematical Statistics and Probability, University of California Press, Berkeley, pp. 111–150.

Bura, E., Cook, R. D., 2001. Estimating the structural dimension of regressions via parametric inverse regression. J. Roy. Statist. Soc. Ser. B, 63, 393-410.

Bura, E., Cook, R. D., 2003. Rank estimation in reduced rank regression. J. Multivariate Anal. 87, 159–176.

Cook, R. D., 1998. Regression Graphics: Ideas for studying regressions through graphics. Wiley, New York.

Cook, R. D., Li, B., 2002. Dimension reduction for the conditional mean in regression. Ann. Statist. 30, 455–474.

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

Cook, R. D., Weisberg, S., 1991. Discussion of Li (1991). J. Amer. Statist. Assoc. 86, 328–332.

Cook, R. D., Yin, X., 2001. Dimension reduction and visualization in discriminant analysis (with discussion). Aus. & New Zealand J. Statist. Vol. 43, No. 2,

147–199.

Frank, I. E., Friedman, J. H., 1993. A statistical view of some chemometrics regression tools (with discussion). Technometrics, 35, 109–148.

Izenman, A. L., 1975. Reduced-rank regression for multivariate linear models. J. Multivariate Anal. 5, 248–264.

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, 1025–1039.

Li, B., Cook, R. D., Chiarmonte, F., 2003. Dimension reduction for the conditional mean in regressions with categorical predictors. Ann. Statist. 31, 1636–1668.

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

Reinsel, G. C., Velu, R. P., 1998. Multivariate reduced-rank regression. Springer, New York.

Schmidli, H., 1995. Reduced-rank regression. Physica, Berlin.

Schott, J. R., 1994. Determining the dimensionality in sliced inverse regression. J. Amer. Statist. Assoc. 89, 141–148.

Weisberg, S., 2003. *The dr package: Methods for dimension reduction for regression*. http://www.r-project.org.

Velilla, S., 1998. Assessing the number of linear components in a general regression problem. J. Amer. Statist. Assoc. 93, 1088–1098.

Velu, R. P., Reinsel, G. C., Wichern, D. W., 1986. Reduced rank models for multiple time series. Biometrika, 73, 105–118.

Yin, X., Cook, R. D., 2002. Dimension reduction for the conditional k-th moment in regression. J. Roy. Statist. Soc. Ser. B, 64, pp. 159-175.

Yin, X., Cook, R. D., 2003. Dimension reduction via marginal high moments in regression. In preparation.

Figure 1: Scatterplot matrix of the four responses and X_0.