Topics on Reduced Rank Methods for Multivariate Regression

by

Ashin Mukherjee

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Statistics)
in The University of Michigan
2013

Doctoral Committee:

Professor Naisyin Wang, Co-Chair
Professor Ji Zhu, Co-Chair
Professor Jionghua Jin
Professor Kerby Shedden

To my Thamma (Grandma)

ACKNOWLEDGEMENTS

Firstly I wish to express my sincerest gratitude to my advisor Prof. Ji Zhu. He introduced me to this topic of multivariate regression and without his constant support, encouragements and invaluable insights this work would not have been possible. I thank him for being patient with me and teaching me the importance of hard work in every walk of life. I feel fortunate to have him as my mentor and the lessons that I learned through this journey will stay with me for the rest of my life.

I am also indebted to my co-advisor Prof. Naisyin Wang for her guidance, suggestions and many helpful discussions that helped me immensely throughout this dissertation process. I feel lucky to have found a collaborator in Dr. Akbar Waljee, his incredible drive and work ethics is a source of constant inspiration. I wish to thank Prof. Kerby Shedden and Prof. Jionghua Jin for being members of my dissertation committee and providing many useful comments.

I would like to thank my many friends here at Michigan, the soccer teams and the friends from ISI dispersed all over India and US they have definitely made my life more colorful. Two persons that require special mention are Nirupam Chakraborty my ever dependable roommate and Bodhisattva Sen who have been nothing short of an caring elder brother. My parents, my brother and my grandma are always in my mind. I would not be here without their sacrifices, continuous support and unconditional love. Their pride in anything that I achieve is the warmest motivation for me. And finally, I would like to express my heartfelt appreciation to Jie Cheng, we started our journey here at Michigan. Her presence has made my life infinitely richer and I cherish her companionship more than anything else.

TABLE OF CONTENTS

CHAPTER

LIST OF FIGURES

Figure

LIST OF TABLES

Table

LIST OF APPENDICES

Appendix

ABSTRACT

Topics in Reduced Rank methods for Multivariate Regression
by
Ashin Mukherjee

Advisors: Professor Ji Zhu and Professor Naisyin Wang

Multivariate regression problems are a simple generalization of the univariate regression problem to the situation where we want to predict $q(>1)$ responses that depend on the same set of features or predictors. Problems of this type is encountered commonly in many quantitative fields. The main goal is to build more accurate and interpretable models that can exploit the dependence structure among the responses and achieve dimension reduction. Reduced rank regression has been an important tool to this end due to its simplicity, computational efficiency and superior predictive performance than much more complex models. In The first two parts of this thesis we investigate certain important practical aspects of the reduced rank regression method such as handling collinearity in the design matrix, selection of optimal rank. The last part focuses on extensions of the reduced rank methods to general functional models.

In Chapter 2 we emphasize that the usual reduced rank regression is vulnerable to high collinearity among the predictor variables as that can seriously distort the singular structure of the signal matrix. To address this we propose the reduced rank ridge regression method that incorporates a ridge penalty in addition to the low rank constraint on the coefficient matrix. Ridge penalty introduces shrinkage which allows us to avoid the singularities when predictors are collinear. We are able to develop a straightforward computational algorithm to solve the optimization problem. We also discuss a novel extension of the reduced rank methodology to the Reproducing Kernel Hilbert Space(RKHS) setting.

Chapter 3 studies the effective degrees of freedom of the reduced rank estimators in the framework of Steins Unbiased Risk Estimation (SURE). A finite sample exact unbiased estimator is derived that admits a closed form solution. This can be used to calculate popular model selection criteria such as BIC, Mallows C_p, GCV which provide a principled way of selecting the optimal rank without going for computationally expensive cross-validation procedures. The results hold in high-dimensional settings where both the feature and response dimension can exceed sample size. The methods developed in this chapter are rather general and can be used for any modeling approach that employs singular value thresholding including reduced rank ridge regression and adaptive nuclear norm thresholding methods.

The final chapter considers a non-parametric extension to the popular reduced rank method for multi-response regression under the high-dimensional setting, where only a few predictors affect the responses. Under the framework of additive models, we propose a two- step procedure that encourages both dimension reduction as well as variable selection. Theoretical properties of the proposed method is investigated including rank selection consistency and error bound. Numerical studies demonstrate that the proposed method outperforms relevant competitors in a broad range of settings. An application to low- density polyethylene reactor data further illustrates the usefulness of the proposed method.

CHAPTER I

Introduction to Multivariate Regression

1.1 Background and Literature Review

Predicting multiple response from the same set of predictors also known as multi-task learning have attracted a lot of attention of late. Problems of this kind is commonly encountered in many quantitative fields such as bio-informatics, chemometrics, gene regulatory networks and so on, where one is interested in simultaneously predicting several responses generated by a single mechanism or experiment. The basic assumption is that many of the responses are highly dependent and therefore appropriate dimension reduction might lead to a more interpretable and accurate predictive model.

Multivariate linear regression is the extension of the classical univariate regression model to the case where we have $q(>1)$ responses and p predictors. We can express the multivariate linear regression model in matrix notation as follows. Let $\mathbf{X}$ denote the $n \times p$ predictor or design matrix, with the i-th row $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip}) \in \mathbb{R}^p$. Similarly the $n \times q$ dimensional response matrix is denoted by $\mathbf{Y}$, where the i-th row is $\mathbf{y}_i = (y_{i1}, y_{i2}, \ldots, y_{iq}) \in \mathbb{R}^q$. The regression parameters are given by the coefficient matrix $\mathbf{B}$ which is of dimension $p \times q$. Note that the k-th column of $\mathbf{B}$ is the regression coefficient vector for regressing the k-th response on the predictors. Let $\mathbf{E}$ denote the $n \times q$ random error matrix with independent entries with mean zero and variance σ^2. Then the multivariate linear regression model is given by

$$\mathbf{Y} = \mathbf{X}\mathbf{B} + \mathbf{E}. \tag{1.1}$$

Note that, this reduces to the classical univariate regression model when $q = 1$. For notational simplicity, we assume that the responses and the predictors are centered, and hence the intercept term can be omitted without any loss of generality. The ordinary least squares approach of estimating $\mathbf{B}$ leads to

$$\widehat{\mathbf{B}}_{\text{ols}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y}. \tag{1.2}$$

The ordinary least squares estimate amounts to performing q separate univariate regressions and completely ignores the multivariate aspect of the problem, where many of the responses might be highly correlated and hence the effective dimensionality can be much smaller than q. Also it is unsuitable for the high-dimensional case where both $p, q > n$. A large number of methods have been proposed to overcome these deficiencies most of which are based on ideas of dimension reduction and tries to find some underlying latent structure. Popular methods include Principal Component Regression (*Massy*, 1965), Partial Least Squares (*Wold*, 1975), Canonical Correlation Analysis (*Hotelling*, 1935). All of these methods can be classified under the larger class of Linear Factor Regression, in which the response Y is regressed against a small number of linearly transformed predictors, often called the factors. The models differ in the way they choose the factors. The estimation proceeds in two steps, transforming the original predictors in the chosen factor space and selecting the number of relevant factors r, often achieved through cross-validation. It is easy to see that as r decreases we are able to achieve greater dimensionality reduction.

Several penalization methods have been proposed in recent years to address the issues of prediction performance and variable selection in multivariate regression. *Turlach et al.* (2005) introduced an ℓ_∞ penalty on the rows of $\mathbf{B}$ to encourage simultaneous variable selection. *Peng et al.* (2009) used a combined penalty function of the form, $\mathcal{J}(\mathbf{B}) = \lambda_1 \sum_{j=1}^{p} \|\mathbf{B}_{j.}\|_1 + \lambda_2 \sum_{j=1}^{p} \|\mathbf{B}_{j.}\|_2$ to identify "master predictors" in genomics studies, where $\mathbf{B}_{j.}$ denotes the j-th row of the coefficient matrix. The first part of the penalty imposes sparsity on entries of $\mathbf{B}$, whereas the second part forces some of the entire rows of $\widehat{\mathbf{B}}$ to be zero, encouraging the selection of "master predictors" that influence many response variables. *Obozinski et al.* (2011) developed asymptotic theory for the ℓ_1/ℓ_2 penalized multivariate regression problem, which can also be thought of as a special case of the joint penalty employed by *Peng et al.* (2009) with $\lambda_1 = 0$. In particular, they prove that the multivariate group lasso penalty recovers the correct row support with high probability in high-dimensional settings.

Yet another line of research focuses on the rank of the regression coefficient matrix. *Anderson* (1951) proposed a class of regression models that restrict the rank of the coefficient matrix to be much smaller than the dimensionality of $\mathbf{B}$, i.e. $rank(\mathbf{B}) \leq r \leq \min\{p, q\}$. This is a quite reasonable assumption in many multivariate regression problems, which can be interpreted as follows: the q responses are related to the p predictors only through r effective linear factors. It results in the following optimization problem

$$\widehat{\mathbf{B}}(r) = \underset{\{\mathbf{B}:\text{rank}(\mathbf{B})\leq r\}}{\text{argmin}} \|\mathbf{Y} - \mathbf{X}\mathbf{B}\|_F^2, \tag{1.3}$$

where $\|.\|_F$ denotes the Frobenius norm of a matrix. Even though the rank penalty makes it a non-convex optimization problem, it admits a closed form solution as we shall see later. *Izenman* (1975) introduced the term reduced rank regression for this class of models and derived the asymptotic distributions and confidence intervals for reduced rank regression estimators. A non-exhaustive list of notable work includes *Rao* (1978), *Davies and Tso* (1982), *Anderson* (1999, 2002a); see *Reinsel and Velu* (1998) or *Izenman* (2008) for a more comprehensive account. Recently, there has been a revival of interest in the reduced rank methods. Instead of restricting the rank, *Yuan et al.* (2007) proposed to put an ℓ_1 penalty on the singular values of $\mathbf{B}$ also known as the nuclear norm. They try to minimize the following constrained least squares criterion

$$\hat{\mathbf{B}} = \arg\min_{\mathbf{B}} \|\mathbf{Y} - \mathbf{X}\mathbf{B}\|_F^2 + \lambda \sum_{j=1}^{\min\{P,Q\}} \sigma_j(\mathbf{B}). \tag{1.4}$$

Where $\sigma_j(\mathbf{B})$ denotes the j-th singular value of $\mathbf{B}$. This constraint encourages sparsity in the singular values of $\mathbf{B}$ and hence the solution $\hat{\mathbf{B}}$ is of lower rank. Though motivated from linear factor regression this approach avoids the explicit choice of the factor space by choosing a clever set of basis functions. The optimization problem in (1.4) is shown to be equivalent to a 2-nd order cone program and the authors use the SDPT3 solver to obtain the solution. SDPT3 can solve conic linear optimization problems over a closed, convex pointed set in a finite-dimensional inner-product space (*Tutuncu et al.*, 2003). Unlike reduced rank regression solution this provides a continuous regularization path. The nuclear-norm penalized least squares criterion encourages sparsity among the singular values to achieve simultaneous rank reduction and shrinkage coefficient estimation (*Neghaban and Wainwright*, 2011; *Lu et al.*, 2012). However, this method is computationally intensive and tends to overestimate the rank (*Bunea et al.*, 2011). *Bunea et al.* (2012) proposed the rank selection criterion extending reduced rank regression to high-dimensional settings, in which rank-constrained estimation was cast as a penalized least squares method with the penalty proportional to the rank of the coefficient matrix, or equivalently, the ℓ_0 norm of its singular values. Under that framework the authors were able to characterize the choice of tuning parameter, which guarantees asymptotic consistency in terms of rank selection. *Chen et al.* (2012) adopted sparsity penalties on singular vectors for reduced rank regression problems that lead to more interpretable models. Very recently *Chen and Huang* (2012) proposed an adaptive nuclear norm penalty on the signal matrix $\mathbf{XB}$ aiming to close the gap between ℓ_0 and ℓ_1 penalties on singular values. The resulting optimization problem admits a closed form solution and enjoys many desirable theoretical properties.

1.2 Outline of the Thesis

In Chapter 2 we emphasize that the low-rank structure of the response matrix $\mathbf{Y}$ can arise from the collinearity in the design matrix $\mathbf{X}$ in addition to the low rank structure of the coefficient matrix $\mathbf{B}$. In that situation reduced rank regression will often end-up selecting a wrong rank due to the distortion of the singular values of $\mathbf{XB}$. To address this we propose the reduced rank ridge regression method that incorporates a ridge penalty in addition to the low rank constraint on the coefficient matrix $\mathbf{B}$. Ridge penalty introduces shrinkage which allows us to avoid the singularities when predictors are collinear. We are able to develop a straightforward computational algorithm to solve the optimization problem. We also discuss a novel extension of the reduced rank methodology to the Reproducing Kernel Hilbert Space(RKHS) setting.

Chapter 3 deals with the all important question about how to choose the optimal rank for a reduced rank method. We study the effective degrees of freedom of a general class of reduced rank estimators in the framework of Stein's Unbiased Risk Estimation (SURE). We derive a finite sample exact unbiased estimator that admits a closed form solution in terms of the singular values or the thresholded singular values of the ordinary least squares estimator. This is of much practical importance as it allows us to use many of the popular model selection criteria such as AIC, BIC, Mallow's C_p to select the optimal rank without going for computationally expensive cross-validation procedures. The results continue to hold in the high-dimensional scenario when both the feature and response dimension exceeds sample size. We are able to obtain some new insights into the empirical behavior of the exact degrees of freedom estimator by comparing it with commonly used naive estimator which is given by the number of free parameters in matrix of low rank.

The final Chapter (4) considers a non-parametric extension of the reduced rank regression for multivariate regression with variable selection. In modern high-dimensional problem it is rather common to encounter sparsity, that is, only a few features carry information regarding the responses. We propose a two step approach that encourage both dimension reduction and variable selection under the framework of additive models. Accurate approximation of the component functions by B-spline basis under appropriate assumptions of smoothness is a key part of the methodological and theoretical aspects. We prove that the proposed method possesses many desirable theoretical properties such as rank selection consistency and error bounds. Simulation studies and real data examples help to illustrate that the proposed method outperforms its relevant competitors in a broad range of settings.

CHAPTER II

Reduced Rank Ridge Regression and its Kernel Extensions

2.1 Introduction

In Chapter I we introduced the multivariate reduced rank regression model in details and also discussed many of its variants and relevant competitors. Most of them impose a rank constraint on the coefficient matrix $\mathbf{B}$ to exploit the approximate low rank structure of the response matrix $\mathbf{Y}$. At this point important to notice that $rank(\mathbf{Y})$ depends on $rank(\mathbf{XB})$ rather than $rank(\mathbf{B})$ alone. For example if the design matrix $\mathbf{X}$ is highly collinear then even if the true coefficient matrix $\mathbf{B}$ is of full rank, we can end up with an approximately low rank $\mathbf{Y}$. The existing methods all depend on the singular structure of the ordinary least squares solution (1.2) to attain the rank reduction. If the design matrix is highly collinear then the ordinary least squares solution is unstable therefore any method that depends on it will also be affected negatively. We propose a combination of the ridge penalty and rank constraint on the coefficient matrix $\mathbf{B}$ to overcome this problem. The ridge penalty helps to ensure that estimate of $\mathbf{B}$ is well-behaved even in the presence of multicollinearity, whereas the rank constraint encourages dimension reduction.

The rest of the Chapter is organized as follows: In Section 2.2 we formally introduce the reduced rank ridge regression model and discuss some of the finer details. Section 2.3 presents numerical examples which include simulation studies comparing the proposed model to relevant competitors as well as some real-data example. We extend the reduced rank approach to the kernel settings in Section 2.4, and show a real data applications. Section 2.5 concludes with a summary and brief discussion.

2.2 Reduced Rank Ridge Regression Model

We propose a regularized estimator for the coefficient matrix $\mathbf{B}$. Two penalties are added to the usual squared error loss. Ridge penalty to ensure that the estimator of $\mathbf{B}$ is well-behaved even in the presence of collinearity among the predictor variables. Rank constraint encourages dimensionality reduction by restricting the rank of $\hat{\mathbf{B}}$. We seek to minimize

$$\hat{\mathbf{B}}(\lambda, r) = \underset{\{\mathbf{B}:\mathrm{rank}(\mathbf{B}) \leq r\}}{\arg\min} \|\mathbf{Y} - \mathbf{X}\mathbf{B}\|_F^2 + \lambda\|\mathbf{B}\|_F^2. \tag{2.1}$$

Where $r \leq \min\{n, p, q\}$. $\|.\|_F^2$ denotes the Frobenius norm for matrices. For each fixed λ we can transform this problem to a Reduced Rank Regression problem on an augmented data set. Define

$$\mathbf{X}^*_{(n+p)\times p} = \begin{pmatrix} \mathbf{X} \\ \sqrt{\lambda}\mathbf{I} \end{pmatrix}; \qquad \mathbf{Y}^*_{(n+p)\times q} = \begin{pmatrix} \mathbf{Y} \\ \mathbf{0} \end{pmatrix}. \tag{2.2}$$

Then it is a matter of simple algebra to notice that the minimization problem in (2.1) is equivalent to the following reduced rank regression problem

$$\hat{\mathbf{B}}(\lambda, r) = \underset{\{\mathbf{B}:\mathrm{rank}(\mathbf{B}) \leq r\}}{\arg\min} \|\mathbf{Y}^* - \mathbf{X}^*\mathbf{B}\|_F^2. \tag{2.3}$$

Now we can use the orthogonal projection property of the ols estimator to decompose the squared error loss function in two parts, $\|\mathbf{Y}^* - \mathbf{X}^*\mathbf{B}\|_F^2 = \|\mathbf{Y}^* - \hat{\mathbf{Y}}^*_R\|_F^2 + \|\hat{\mathbf{Y}}^*_R - \mathbf{X}^*\mathbf{B}\|_F^2$. Here $\hat{\mathbf{Y}}^*_R = \mathbf{X}^*\hat{\mathbf{B}}^*_R$ denotes the Ridge regression estimate which is also same as the ols estimate obtained from the linear model $\mathbf{Y}^* = \mathbf{X}^*\mathbf{B} + \mathbf{E}^*$. Note that the first term do not involve $\mathbf{B}$ hence we get the following equivalent form for the minimization problem (2.3) as

$$\hat{\mathbf{B}}(\lambda, r) = \underset{\{\mathbf{B}:\mathrm{rank}(\mathbf{B}) \leq r\}}{\arg\min} \|\hat{\mathbf{Y}}^*_R - \mathbf{X}^*\mathbf{B}\|_F^2. \tag{2.4}$$

Let us assume that $\hat{\mathbf{Y}}^*_R = \sum_{i=1}^{\tau} \sigma_i u_i v_i^T$ gives the singular value decomposition of $\hat{\mathbf{Y}}^*_R$. σ_i's denote the singular values, u_i and v_i denote the left and right singular vectors of $\hat{\mathbf{Y}}^*_R$ respectively. τ is the rank of $\hat{\mathbf{Y}}^*_R$. Then a fairly elementary result in linear algebra known as the Eckart-Young theorem (*Eckart and Young*, 1936) tells us that the best rank r approximation to $\hat{\mathbf{Y}}^*_R$ in the Frobenius norm is given by

$$\hat{\mathbf{Y}}^*_r = \sum_{i=1}^{r} \sigma_i u_i v_i^T. \tag{2.5}$$

Define, $\underset{q\times q}{\mathbf{P}_r} = \sum_{i=1}^{r} v_i v_i^T$, and let $\hat{\mathbf{B}}(\lambda, r) = \hat{\mathbf{B}}^*_R \mathbf{P}_r$. Clearly $\text{rank}(\hat{\mathbf{B}}(\lambda, r)) \leq r$, since $\text{rank}(\mathbf{P}_r) = r$. And plugging them back in we get

$$\begin{aligned} \mathbf{X}^*\hat{\mathbf{B}}(\lambda, r) &= \mathbf{X}^*\hat{\mathbf{B}}^*_R\mathbf{P}_r = \left(\sum_{i=1}^{\tau} \sigma_i u_i v_i^T\right)\left(\sum_{j=1}^{r} v_j v_j^T\right) \\ &= \sum_{i=1}^{r} \sigma_i u_i v_i^T = \hat{\mathbf{Y}}^*_r. \end{aligned}$$

Hence we are able to show that proposed solution $\hat{\mathbf{B}}(\lambda, r) = \hat{\mathbf{B}}^*_R\mathbf{P}_r$ is the minimizer of the optimization problem (2.1), which is the original reduced rank ridge regression problem that we started with. Writing down explicitly in terms of $\mathbf{X}$, $\mathbf{Y}$, λ and r we get the following

$$\hat{\mathbf{B}}(\lambda, r) = \hat{\mathbf{B}}^*_R\mathbf{P}_r = \left(\mathbf{X}^{*T}\mathbf{X}^*\right)^{-1}\mathbf{X}^{*T}\mathbf{Y}\mathbf{P}_r = \left(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\mathbf{X}^T\mathbf{Y}\mathbf{P}_r \tag{2.6}$$

$$\hat{\mathbf{Y}}(\lambda, r) = \mathbf{X}\left(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\mathbf{X}^T\mathbf{Y}\mathbf{P}_r = \hat{\mathbf{Y}}_\lambda\mathbf{P}_r. \tag{2.7}$$

$\hat{\mathbf{Y}}_\lambda$ in the above equation denotes the multivariate ridge regression estimator for $\mathbf{Y}$ with a penalty parameter λ. This shows that the reduced rank ridge regression is actually projecting $\hat{\mathbf{Y}}_\lambda$ to a r-dimensional space with projection matrix $\mathbf{P}_r$. Here it is important to notice that this is a projection of the rows of $\hat{\mathbf{Y}}_\lambda$ which in general lives in a Q-dimensional space to a lower r-dimensional space. Easy to see that for $r = Q$ we get back the ridge regression solution.

2.2.1 Illustrative Example

To illustrate the issues with Reduced Rank regression we construct a simple toy example. Set $p = q = 3$ and $n = 50$ and let

$$\mathbf{B} = \begin{pmatrix} 1 & 3 & 0 \\ 3 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix} \quad , \quad \mathbf{\Sigma}_X = \begin{pmatrix} 1 & 0.95 & 0 \\ 0.95 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

The first 2 columns of $\mathbf{B}$ are linearly independent and thus it has rank 2. But at the same time we make predictors X_1 and X_2 highly collinear, so that the effective dimension of the response reduces to 1. We simulate $\mathbf{X} \sim N(0, \mathbf{\Sigma}_X)$ and given $\mathbf{X}$, $\mathbf{Y}$ is generated from $\mathbf{Y} \sim N(\mathbf{XB}, 0.25\mathbf{I})$. The eigenvalues of $\mathbf{Y}^T\mathbf{Y}$ comes out to be $\sigma^2 = [1252, 16, 11]$. Hence Reduced Rank regression would select rank to be 1 and seek a rank 1 estimator of $\mathbf{B}$ which is clearly not the case here. This happens because Reduced Rank regression fails to account for the correlated predictors and that is precisely where Reduced Rank Ridge regression improves by adding ridge penalty.

2.2.2 Selection of Tuning Parameters

For the reduced rank ridge regression we propose to choose the tuning parameters (λ, r) using a simple K-fold cross-validation procedure. We first define a grid for (λ, r) note that r can only take values in $\{1, 2, \ldots \min\{n, p, q\}\}$. For each combination of λ and r we evaluate average of validation prediction errors over the K-folds and choose the optimal combination as the one that minimizes this quantity. Notationally

$$(\hat{\lambda}, \hat{r}) = \underset{(\lambda, r)}{\arg\min} \sum_{k=1}^{K} \|\mathbf{Y}^{(k)} - \mathbf{X}^{(k)}\hat{\mathbf{B}}^{(-k)}(\lambda, r)\|_F^2. \tag{2.8}$$

Where $\mathbf{X}^{(k)}$ and $\mathbf{Y}^{(k)}$ denote the predictor and response matrix for the k-th fold, and $\hat{\mathbf{B}}^{(-k)}(\lambda, r)$ denotes the estimated regression coefficient matrix computed leaving out the observations in the k-th fold when using the penalty parameters (λ, r). This would encourage a trade-off between the penalty parameters based on the data. We would look into the choice of tuning parameters more deeply in the simulation studies section.

2.3 Numerical Examples

2.3.1 Simulation Study

We compare the estimation performance of the proposed reduced rank ridge regression method to other multivariate linear regression methods that have been proposed in the literature based on the idea of dimension reduction and borrowing strength from dependent response variables. Methods compared include - Ordinary least squares (OLS); Curd and Whey (CW) procedure developed by Breiman and Friedman with the GCV approach; Reduced Rank Regression (RRR); Multivariate Ridge Regression (MVR) with same tuning parameter for each response; Separate Ridge Regression

(SR); Partial Least Squares (PLS); Principal Component Regression (PCR) and the proposed Reduced Rank Ridge Regression (RRR). For the methods that require a selection of tuning parameter we do so by looking at the prediction error on an independently generated validation set of same size. We measure the performance of various methods by model error following *Breiman and Friedman* (1997). The model error of an estimate $\hat{\mathbf{B}}$ is given by

$$\mathrm{ME}(\hat{\mathbf{B}}) = \mathrm{trace}\left[(\mathbf{B} - \hat{\mathbf{B}})^T \mathbf{\Sigma}_X (\mathbf{B} - \hat{\mathbf{B}})\right]. \tag{2.9}$$

where $\mathbf{B}$ denotes the true coefficient matrix and $\mathbf{\Sigma}_X$ denotes $\mathbb{E}(XX^T)$.

2.3.1.1 Models

In each replication of the simulation study we generate a design matrix $\mathbf{X}_{n\times p}$ with each rows drawn independently from $N(0, \mathbf{\Sigma}_X)$. Where $\mathbf{\Sigma}_X$ has the structure, $\mathbf{\Sigma}_X(i,j) = \rho^{|i-j|}$. We used 3 different levels for the correlation parameter $\rho = [0, 0.5, 0.9]$. To generate the true coefficient matrix $\mathbf{B}_{p\times q}$ we first generate a random $p \times q$ matrix from $N(0,1)$ distribution. The singular values are then replaced with following structures,

- **Model 1** The first half of the singular values are 2 and rest as 0.
- **Model 2** All the singular values are equal to 1
- **Model 3** The largest singular value as 5 and rest 0.

We choose the above mentioned models to ensure that we cover a broad spectrum of rank-deficient situations. Model 2 covers the case of no rank redundancy in the coefficient matrix $\mathbf{B}$ which is the usual multivariate linear regression assumption. Model 3 represents the case for a severe rank deficiency whereas Model 1 is a compromise between these two extreme situations. We analyze each model at different correlation levels between the predictors thus covering most of the possible real scenarios. For each combination of model and correlation we simulate a training and validation set each of size $p = 50, q = 20, n = 100$. And compute each of the estimators described above. The process is repeated 100 times leading to an error-vector of length 100 for each competing method.

All the methods outperforms OLS by a big margin under this settings. PLS and PCR appear quite competitive to RR but fails to perform in the same level as RRR, MVR or SRR. Note that the proposed method RRR dominates all the other methods at

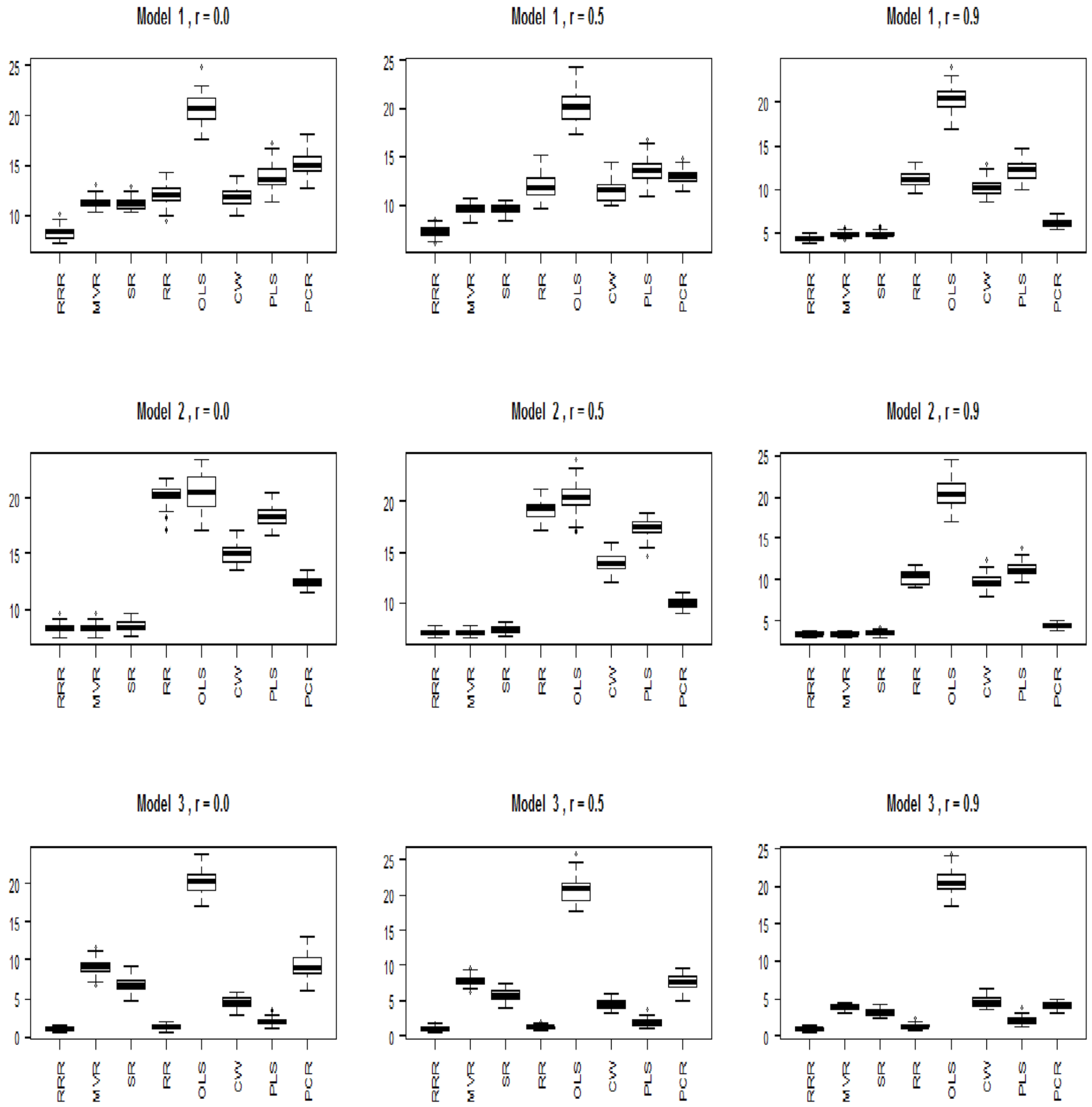

Figure 2.1: Boxplot of ME for each method over 100 replications.

every combination of settings. It is interesting to note that for Model 2 where the true $\mathbf{B}$ had full rank RR does significantly worse than RRR, MVR and SRR for all choices of ρ. Whereas in Model 3 which had the strongest rank deficiency we see that RRR and RR dominates the other methods which also seems intuitive. The biggest advantage of the RRR over only ridge and only rank penalty comes in Model 1 which has nearly half the singular values non-zero. For all three models we see that as the value of ρ increases MVR and SRR tends to catch up with the best method.

To gain further insight, we look at the the singular values of the $\hat{\mathbf{B}}$ for OLS, MVR, RR and RRR method. For this part we use a smaller set-up with $p = 20, q = 8$ and $n = 30$ the singular values of $\mathbf{B}$ are $\sigma = [3, 2, 1.5, 0, 0, 0, 0, 0]$. We plot the singular values over 100 replicates at two extreme correlation levels $\rho = 0.0$, and 0.9.

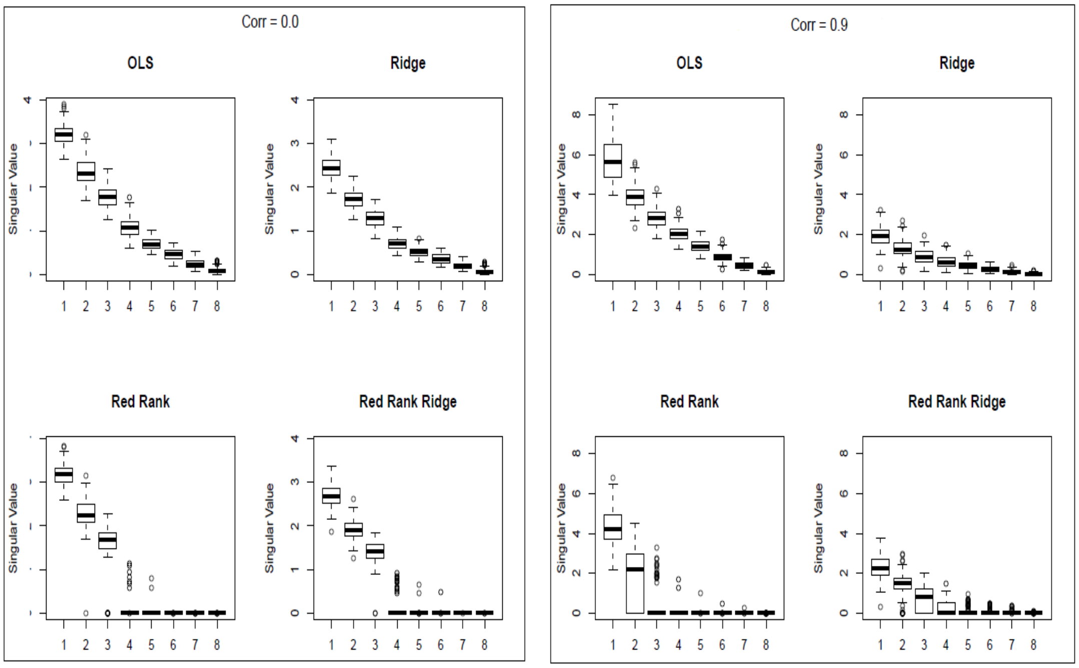

Figure 2.2: Singular values of $\hat{\mathbf{B}}$. Left: $\rho = 0$; Right: $\rho = 0.9$.

For $\rho = 0$ we see that both RR and RRR does a fairly good job of recovering the singular value structure. But as the collinearity among the predictors increases we find that RR most of the times selects 2 or 1 as the rank whereas RRR is able to do a much better job. MVR and OLS fail to achieve any dimension reduction. Similar patterns are observed at other settings as well which we skip for brevity. This clearly illustrates that the trade-off between ridge penalty and the rank constraint is the key that enables us to correctly estimate singular value structure even in presence of serious collinearity.

2.3.2 Application in Chemometrics Example

It is originally from *Skagerberg et al.* (1992). There are $n = 56$ observations with $p = 22$ and $q = 6$. The data is generated from a simulation of a low density tubular polyethylene reactor. The predictor variables consists of 20 temperature measurements at equal distance along the reactor along with the wall temperature and the feed rate. The responses are output characteristics of the polymers produced, namely, *Number avg. molecular weight.*(Y_1), *Weight avg. molecular weight*(Y_2), *Long chain branching*(Y_3), *Short chain branching*(Y_4), *content of vinyl group*(Y_5) and *content of vinyledene group*(Y_6). As the responses were all right skewed we applied log transformation, and finally standardized them. The response correlation is reported below.

$$Corr(Y) = \begin{pmatrix} 1.00 & 0.96 & 0.06 & 0.25 & 0.26 & 0.26 \\ 0.96 & 1.00 & -0.13 & 0.28 & 0.27 & 0.28 \\ 0.06 & -0.13 & 1.00 & -0.50 & -0.48 & -0.48 \\ 0.25 & 0.28 & -0.50 & 1.00 & 0.97 & 0.98 \\ 0.26 & 0.27 & -0.48 & 0.97 & 1.00 & 0.98 \\ 0.26 & 0.28 & -0.48 & 0.98 & 0.98 & 1.00 \end{pmatrix}$$

This shows $\{Y_4, Y_5, Y_6\}$ form a strongly correlated group as does $\{Y_1, Y_2\}$. Y_3 is mildly correlated to the others, which suggests an effective response dimensionality of 3. Average absolute correlation between the predictors is about 0.44 with many of them being very highly correlated. The predictive performance is measured using leave-one-out cross validation. We fit the models based on 55 of the 56 points and predict the left-out point and the procedure is repeated 56 times. Note that we do an 11-fold cross validation within the 55 points to select tuning parameters for the models that have one. We report the prediction error for each response as well as overall average prediction error.

Overall RRR performs the best with MVR being a very close second. The good performance of MVR can also be explained by the fact that many predictors are highly collinear. Comparing columns of RR and RRR, we see that for Y_4, Y_5 and Y_6 RR has much smaller prediction error than RRR but it incurs larger error for Y_1, Y_2 and especially Y_3. Because of the strong correlation structure of the responses, RR

	OLS	CW-gcv	PLS	RR	MVR	RRR
Y_1	0.49	0.49	0.68	0.44	0.15	0.15
Y_2	1.12	0.74	0.90	0.46	0.22	0.22
Y_3	0.53	0.49	0.45	0.65	0.39	0.39
Y_4	0.24	0.18	0.18	0.14	0.26	0.24
Y_5	0.30	0.22	0.26	0.18	0.28	0.27
Y_6	0.28	0.21	0.21	0.16	0.28	0.27
Avg	0.50	0.39	0.45	0.34	0.27	0.26

Table 2.1: Performance comparison for the chemometrics data

concentrates on the heavily correlated group $\{Y_4, Y_5, Y_6\}$, selecting 2 or 1 components most times(out of 56 leave-one-out runs) whereas RRR is able to pick 3 as the optimal dimension with high proportion. So even though it loses a little bit for the highly correlated group overall prediction accuracy is much better.

2.4 Extension to RKHS

Before we go into the details for reduced rank approach in the *Reproducing Kernel Hilbet Space*(RKHS) setting let us first give a very brief introduction to it.

2.4.1 Brief Introduction to RKHS

A Hilbert space is a real/complex inner product space which is complete under the norm induced by the inner product. Examples include $\mathbb{R}^n$ with $\langle x, y\rangle = x^T y$, $\mathbb{L}^2$-space of all square integrable functions on $\mathbb{R}$ with $\langle f, g\rangle = \int_{\mathbb{R}} f(x)g(x)dx$. The reason we are interested in functional spaces is because we would like to fit models like $y = f(x) + \epsilon$ where $f : \mathbb{R}^p \to \mathbb{R}$ to model the data in a much more flexible non-parametric way. $\mathbb{L}_2$ is *too big* for our purpose as it contains too many non-smooth functions. One way to obtain such spaces of smooth functions which allows us to fit a non-parametric functional regression model without explicitly specifying the function f is the RKHS approach.

A positive definite kernel is a function $K(\cdot, \cdot) : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ such that for any $n \geq 1$ and $\{x_1, x_2, \ldots x_n\} \in \mathcal{X}^n$ and $\{a_1, a_2, \ldots, a_n\} \in \mathbb{R}^n$, we have, $\sum_{i=1}^{n}\sum_{i'=1}^{n} a_i a_{i'} K(x_i, x_{i'}) \geq 0$. In other words the gram matrix $\mathbf{K} = [K(x_i, x_{i'})]_{i,i'=1}^n$ is positive definite for all, $\{x_1, x_2, \ldots x_n\} \in \mathcal{X}^n$. For most of our purposes $\mathcal{X} = \mathbb{R}^p$, the space of the predictor variables. It is a well known result (*Wahba*) that given such a kernel we can construct a unique functional Hilbert space $\mathcal{H}$ on $\mathcal{X}$ such that $K(\cdot, \cdot)$ is the inner product in that space and $f(x) = \langle f, K(\cdot, x)\rangle$ for all $f \in \mathcal{H}$ and $x \in \mathcal{X}$ and vice versa.

2.4.2 Kernel Reduced Rank Regression Approach

In the univariate case, given data $\{(x_1, y_1), (x_2, y_2), \ldots (x_n, y_n)\}$, note that $y_i \in \mathbb{R}$ and $x_i \in \mathbb{R}^p$, our objective is to find a function $f \in \mathcal{H}$ that minimizes

$$J_\lambda(f) = \sum_{i=1}^{n}(y_i - f(x_i))^2 + \lambda \|f\|^2_{\mathcal{H}}. \tag{2.10}$$

Where $\|.\|^2_{\mathcal{H}}$ denotes the norm in $\mathcal{H}$. This is introduced to encourage smoothness and to avoid overfitting. Then the Representer Theorem says that any f minimizing (2.10) can be written as

$$f(x) = \sum_{i=1}^{n} \alpha_i K(x, x_i), \quad \text{for} \quad (\alpha_1, \alpha_2, \ldots, \alpha_n) \in \mathbb{R}^n. \tag{2.11}$$

For the multivariate response $y_i \in \mathbb{R}^q$, in the RKHS set-up we want to find $(f_1, f_2, \ldots, f_q) \in \mathcal{H}$ which minimizes a joint loss function defined by

$$J_\lambda(f_1, f_2, \ldots, f_q) = \sum_{j=1}^{q}\sum_{i=1}^{n}\|y_{ij} - f_j(x_i)\|^2 + \lambda \sum_{j=1}^{q}\|f_j\|^2_{\mathcal{H}}. \tag{2.12}$$

Like in the linear case it is fairly easy to see that in absence of any constraint on the functions $(f_1, f_2, \ldots, f_q)$ the above optimization is same as doing q separate single-response kernel ridge regression problem. If we want to exploit the dependence among the responses we need some equivalent way of expressing the reduced rank constraint under the RKHS set-up. The following proposition gives one such way.

Proposition II.1. *Let $\mathcal{H}$ be the RKHS corresponding to a positive-definite kernel $K(\cdot, \cdot) : \mathbb{R}^p \times \mathbb{R}^p \mapsto \mathbb{R}$. Given data $\{(x_1, y_1), (x_2, y_2), \ldots (x_n, y_n)\}$, $y_i \in \mathbb{R}^q$ and $x_i \in \mathbb{R}^p$, we consider the optimization problem*

$$\min_{f_1, f_2, \ldots, f_q \in \mathcal{H}} J_\lambda\left(f_1, f_2, \ldots, f_q\right) \quad \textit{subject to,} \quad \textit{dim}\left(\textit{span}\{f_1, f_2, \ldots, f_q\}\right) \le r. \tag{2.13}$$

Where $1 \le r \le q$ and $J_\lambda\left(f_1, f_2, \ldots, f_q\right)$ is defined as in (15)*. The solution has the* *following representation*

$$f_q(x) = \sum_{i=1}^{n} \alpha_{ij} K(x, x_i), \quad \textit{for } j = 1, 2, \ldots, q, \quad \alpha_{ij} \in \mathbb{R}. \tag{2.14}$$

The constraint $\dim\left(\mathrm{span}\{f_1, \ldots, f_q\}\right) \leq r$ can be viewed as an extension to the rank constraint for linear functions. The only difference being instead of working with linear functions here we are in a general functional space. We defer the proof to the appendix. The next natural step is to find some sufficient conditions under which the rank constraint in (2.13) becomes equivalent to a rank constraint on the coefficient matrix $\mathbf{A} = [\alpha_{ij}]_{n \times q}$. Because that would allow us to extend the reduced rank ridge regression solution developed in Section 2.2 in a natural way to the kernel setting.

Proposition II.2. *If $K(\cdot,\cdot)$ is strictly positive definite and $\{x_1, x_2, \ldots x_n\}$ are distinct then $\dim\left(span\{f_1, f_2, \ldots f_q\}\right) \leq r \Rightarrow rank(\mathbf{A}) \leq r$, where*

$$[f_1, \ldots f_q] = [K(\cdot, x_1), \ldots K(\cdot, x_n)] \underset{n \times q}{\mathbf{A}} \tag{2.15}$$

This proposition translates the reduced rank constraint for RKHS into a simple rank constraint for the coefficient matrix $\mathbf{A}$, under some condition on $K(\cdot,\cdot)$. It is easy to show that Gaussian kernel, $K(x, x') = \exp\left(\frac{-\|x-x'\|^2}{2\sigma^2}\right)$, Laplacian kernel, $K(x, x') = \exp\left(\frac{\|x-x'\|_1}{2\sigma^2}\right)$, Inverse multi-quadratic kernel $K(x, x') = \frac{1}{\sqrt{\|x-y\|^2+c}}$ would satisfy strict positive definiteness. Polynomial kernels in general would not satisfy it because it is essentially an extension to a bigger but finite-dimensional space. But in practice the infinite-dimensional RKHS's are the ones that we would be interested in, so the condition for strict positive definiteness is not very prohibitive.

2.4.3 Extending the Solution

Let us recall the solution to the reduced rank ridge regression problem with penalty parameters (λ, r), derived in Section 2.2. For a given point $x \in \mathbb{R}^p$ (row vector) prediction had the form

$$\hat{Y}_x(\lambda, r) = x\left(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\mathbf{X}^T\mathbf{Y}\mathbf{P}_r.$$

where $\mathbf{P}_r$ was the projection matrix to the space spanned by r principal eigenvectors of $\mathbf{P} = \mathbf{Y}^T\mathbf{X}\left(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\mathbf{X}^T\mathbf{Y}$. Using the matrix inversion lemma we can easily expand the prediction formula in terms of the inner product matrix $\mathbf{X}\mathbf{X}^T$. Then replacing the inner product matrix by the *Gram matrix* $\mathbf{K} = [(K(x_i, x_{i'})]_{i,i'=1}^{n}$ we get

$$\mathbf{Y}^T\mathbf{X}\left(\mathbf{X}^T\mathbf{X}+\lambda\mathbf{I}\right)^{-1}\mathbf{X}^T\mathbf{Y} = \mathbf{Y}^T\mathbf{K}\left(\mathbf{K}+\lambda\mathbf{I}\right)^{-1}\mathbf{Y}, \tag{2.16}$$

$$x\left(\mathbf{X}^T\mathbf{X}+\lambda\mathbf{I}\right)^{-1}\mathbf{X}^T\mathbf{Y} = K(x)\left(\mathbf{K}+\lambda\mathbf{I}\right)^{-1}\mathbf{Y}. \tag{2.17}$$

Note that $K(x) = [K(x, x_1), K(x, x_2), \ldots K(x, x_n)]_{1\times n}$. If we denote the projection matrix to the space spanned by r principal eigenvectors of $\mathbf{Y}^T\mathbf{K}\left(\mathbf{K}+\lambda\mathbf{I}\right)^{-1}\mathbf{Y}$ by $\mathbf{P}_r^K$ then the final prediction for the point $x \in \mathbb{R}^p$ is given by

$$\hat{Y}_x(\lambda, r) = K(x)\left(\mathbf{K}+\lambda\mathbf{I}\right)^{-1}\mathbf{Y}\mathbf{P}_r^K. \tag{2.18}$$

Which is similar to projection of the kernel ridge regression estimator to a constrained space of dimension $\leq r$ as in the linear case.

2.4.4 Simulation Study

In this section we compare the performance of the proposed kernel Reduced Rank Ridge Regression (kernel RRR) with kernel Ridge Regression. We perform the comparison with the choice of two popular choices of kernel function namely, the Gaussian kernel which is strictly positive-definite and thus satisfies the condition of Proposition II.1 and the polynomial kernel which is clearly finite-dimensional and hence does not satisfy the sufficient condition provided in Proposition II.2.

We present the results for $p = 10$, $q = 10$ and $n = 100$, similar results were obtained for other choices of p and q. Rows of the design matrix $\mathbf{X}$ were generated independently from $N(0, \mathbf{I}_P)$. Responses are generated as linear combinations of $m = 10$ basis functions of the form $K(., b_j)$ where $\{b_j : j = 1, 2, \ldots, 10\}$ were generated independently from a multivariate Gaussian distribution. We consider 2 cases,

- **Full Rank Situation** The coefficient matrix is full-rank, i.e., of rank 10.
- **Reduced Rank Situation** The coefficient matrix has rank 5.

The tuning parameters, i.e. (λ, r, σ) were chosen using independently generated validation data sets of same size. In the following figure we present the box-plots of the percentage ratio of MSE of kernel RRR and kernel Ridge Regression over 100 replications of the experiment.

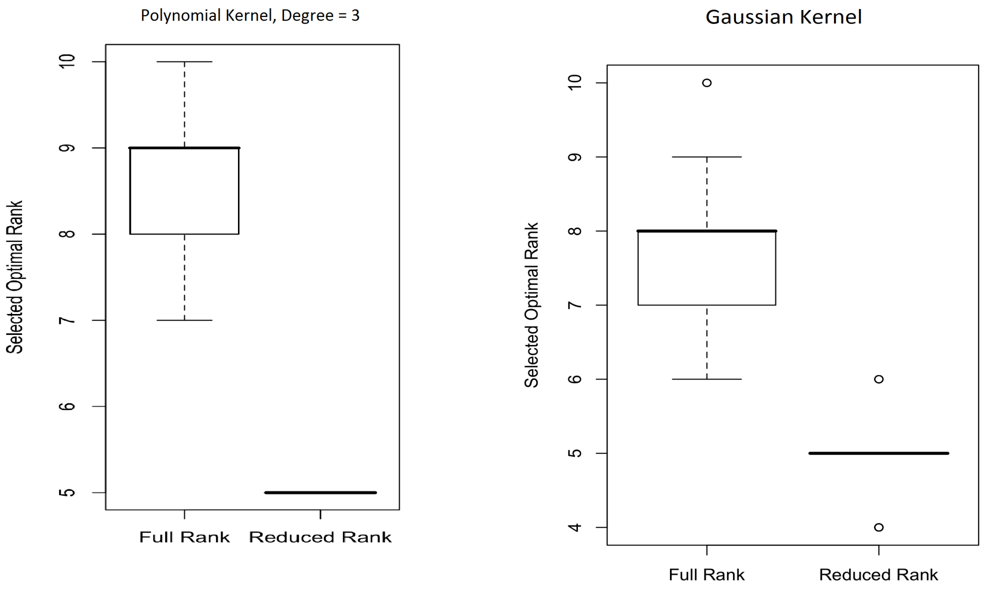

Figure 2.3: Boxplot of selected optimal rank. Left: Polynomial kernel; Right: Gaussian kernel.

As expected we find that kernel Reduced Rank Ridge improves over kernel Ridge significantly when the underlying process is truly low-rank, and even in the full-rank case it performs comparably with kernel Ridge regression. The conclusions hold not only for the Gaussian kernel but for the polynomial kernel as well which as we discussed before does not satisfy the sufficient conditions in Proposition **??**. Also the estimated optimal rank seem to be quite accurate when the underlying functional space is low-rank. Here it is useful to note that if the sample size is too high then the gram matrix for polynomial kernel might become nearly singular causing unstable solutions.

2.4.5 Chemometrics Data Revisited

We apply the kernel RRR on the previously discussed Chemometrics data set and compare its performance against linear RRR and kernel Ridge Regression. We used the popular Gaussian kernel $K(x, x') = \exp\left(\frac{\|x-x'\|^2}{2\sigma^2}\right)$ and the Inverse multi-quadratic kernel $K(x, x') = \frac{1}{\sqrt{\|x-y\|^2+c}}$. Both predictors and responses were standardized for this analysis. An 8 fold-cross-validation is performed to select the tuning parameters, that is (λ, r, σ^2) in case of the Gaussian kernel and (λ, r, c) for the inverse multi-

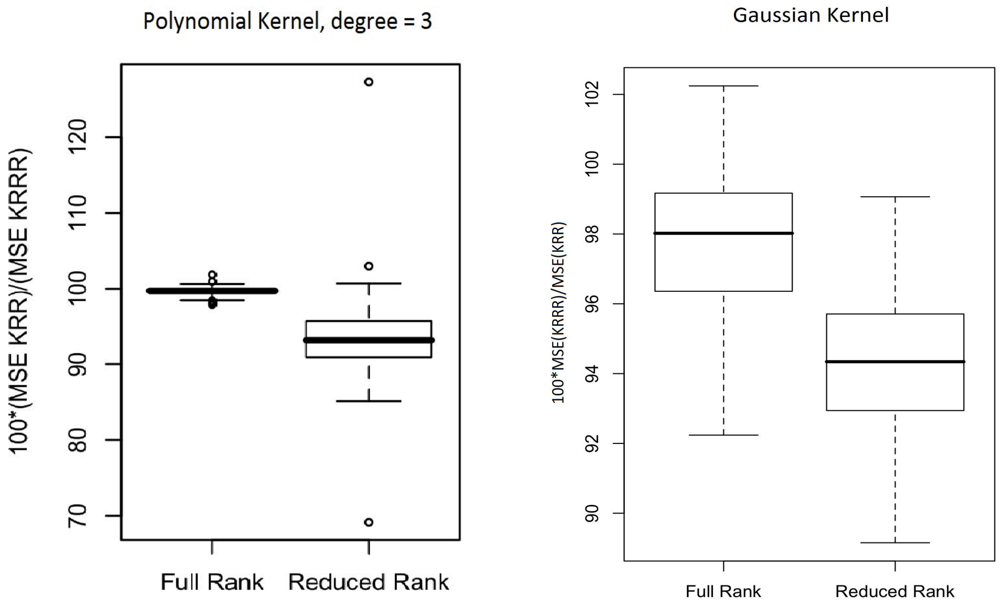

Figure 2.4: Boxplot of ratio of test MSE for kernel RRR. Left: Polynomial kernel; Right: Gaussian kernel.

quadratic kernel.

	Linear RRR	Kernel Ridge	Kernel RRR
Y_1	0.153	0.088	0.087
Y_2	0.250	0.148	0.129
Y_3	0.230	0.113	0.111
Y_4	0.188	0.054	0.044
Y_5	0.205	0.107	0.071
Y_6	0.211	0.070	0.064
Avg	0.206	0.097	0.084

Table 2.2: Performance comparison for Kernel RRR on the Chemometrics data

We used cross-validation error estimate on the hold-out fold to select the tuning parameters. Optimal rank for the kernel RRR which turns out to be 3 for both choices of the kernels as it was for linear Reduced Rank regression implying that the intrinsic dimensionality of the response space is 3. Both choices of the kernel lead to very similar results. Kernel RRR improves by a big margin over the linear RRR, whereas the improvement over kernel ridge regression is less pronounced but still notable for this data set. The Gaussian kernel is able to attain a greater reduction in MSE which is due to the fact that it corresponds to a bigger functional class. The results seem scientifically reasonable since the first two responses namely, Number avg. molecular weight and Weight avg. molecular weight are approximately dependent. Similarly the last 3 responses form a functional group, in the sense that, Short chain branching is an approximate measure of the contents of Vinyl and Vinyledene groups and thus are highly correlated. Long chain branching is negatively correlated to the Short chain branching group.

2.5 Summary and Discussion

We propose Reduced Rank Ridge Regression to produce a low-rank estimator of the regression coefficient matrix $\mathbf{B}$. This is very useful when the responses are highly dependent or there are reasons to believe a latent variable structure among the predictors. Our method accounts for multicollinearity in predictor variables by incorporating a ridge penalty, here it is important to note that both high collinearity in $\mathbf{X}$ and low-rank of the true coefficient matrix $\mathbf{B}$ might lead to the response matrix being rank-deficient and hence it makes sense to apply the penalties jointly and decide the trade-off based on the data. We also extend the reduced rank idea to the RKHS set-up and give some intuition for the meaning of a rank constraint in a functional space.

The solution to the Reduced Rank Ridge Regression problem is obtained as a projection of the Ridge Regression estimator to a constrained space. And hence it is computationally simple. We propose a cross-validation approach to select the tuning parameters. The proposed method was tested in broad variety of simulation settings as well as couple of real data sets. Results are promising and the proposed method is able to outperform relevant competitors under most of the settings. We also apply the kernel RRR on a real data example and it shows some significant improvement over the linear RRR and kernel ridge regression. These applications also helps us understand some statistical insights into the working of the proposed Reduced Rank Ridge Regression method.

CHAPTER III

Degrees of Freedom of Reduced Rank Estimators in Multivariate Regression

3.1 Introduction

One of the main practical challenges of any reduced rank methodology is the selection of the optimal rank. In Chapter II we used a cross-validation technique to select the joint tuning parameters (λ, r). In high-dimensional problems such methods can often prove to be computational bottlenecks. Another approach is to use some well established model selection criteria such as AIC (*Akaike*, 1974) or BIC (*Schwarz*, 1978) and so on. But these approaches require an estimator of model complexity or degrees of freedom. This chapter studies the degrees of freedom of the reduced rank estimators in multivariate linear regression models. The degrees of freedom is a very familiar and one of the most widely used terms in statistics. We utilize it from ANOVA t-tests to various model selection criteria. However, it has been largely overlooked in the reduced rank regression literature except for some heuristic suggestions (*Davies and Tso*, 1982; *Reinsel and Velu*, 1998). For example, the number of free parameters in a $p \times q$ matrix of rank r, given by $r(p + q - r)$ has been suggested as a naive estimate of the degrees of freedom of the reduced rank regression estimator when restricted to rank $r \leq \min\{p, q\}$. More precisely, for an arbitrary design matrix, the number of free parameters should be $(r_x + q - r)r$, where $r_x = \text{rank}(\mathbf{X})$ is the rank of the design matrix (*Bunea et al.*, 2011). Henceforth, we refer to this as the naive estimator of the degrees of freedom of a rank-r model. In this paper, we aim to find a finite-sample unbiased estimator of the degrees of freedom for a general class of reduced rank estimators for the multivariate regression model and investigate its properties. The result covers a significant gap in the literature, as the previously suggested naive estimate lacks both statistical motivation and practical performance.

In a nutshell, the degrees of freedom quantifies the complexity of a statistical modeling procedure (*Hastie and Tibshirani*, 1990). In the case of the univariate linear regression model, it is well-known that the degrees of freedom is the number of estimated parameters, p. However, in general there is no exact correspondence between the degrees of freedom and the number of free parameters in the model (*Ye*, 1998). For example, in the best subset selection for univariate regression (*Hocking and Leslie*, 1967), we search for the best model of size $p_0 \in \{1, 2, \ldots, p\}$ that minimizes the residual sum of squares. The resulting model has p_0 parameters but intuitively the degrees of freedom would be higher than p_0 since the search for the "optimal" subset of size p_0 increases model complexity (*Hastie et al.*, 2009). In other words, for best subset selection the optimal p_0-dimensional subspace that minimizes the residual sum of squares clearly depends on $\mathbf{y}$. Thus the final estimator is highly non-linear in $\mathbf{y}$, which results in the loss of correspondence between degrees of freedom and the number of parameters in the model.

Similar arguments also apply to the reduced rank regression. Instead of searching for best p_0-variables as in the case of best subset selection, here we are searching for best r linear combinations of the predictors that minimize the least squares loss, which should intuitively suggest increased model complexity. Since the optimal rank r-subspace depends on the response matrix $\mathbf{Y}$, the natural correspondence between number of free parameters and degrees of freedom need not hold. This is where reduced rank regression is different from other linear factor regression methods, e.g. principal component regression (*Massy*, 1965). In principal component regression, the factors are principal components of the design matrix $\mathbf{X}$, which do not depend on the response $\mathbf{Y}$, thus the final estimator is still linear in $\mathbf{Y}$.

The rest of the paper is organized as follows. In section 3.2, we review the degrees of freedom in the framework of Stein's unbiased risk estimation (*Stein*, 1981). The reduced rank regression estimator is discussed in detail in section 3.3, additionally, we also introduce a more general class of reduced rank estimators. Sections 3.4, 3.5 and 3.6 contain the main results on our proposed exact unbiased estimator of the degrees of freedom including derivation of a closed form expression, connections to naive degrees of freedom and almost everywhere existence. In section 3.7, we show that the exact unbiased estimator of the degrees of freedom for reduced rank regression methods can be significantly different from the naive estimator through several numerical examples.

We also show that using the exact unbiased estimate of degrees of freedom can lead to gain in prediction accuracy over its heuristic counterpart. In section **??**, we apply the developed method to a genetic association study, and we conclude the paper with a discussion in section 3.9.

3.2 Degrees of freedom

Stein (1981) in his theory of unbiased risk estimation (SURE) first introduced a rigorous definition of the degrees of freedom of a statistical estimation procedure. Later *Efron* (2004) showed that Stein's treatment can be considered as a special case of a more general notion under the assumption of Gaussianity. Assume that we have data of the form $(\mathbf{y}_{n\times 1}, \mathbf{X}_{n\times p})$. Given $\mathbf{X}$, the response originates from the following model $\mathbf{y} \sim (\boldsymbol{\mu}, \sigma^2\mathbf{I})$, where $\boldsymbol{\mu}$ is the true mean that can be a function of $\mathbf{X}$, and σ^2 is the common variance. Then for any estimation procedure $m(\cdot)$ with fitted values $\hat{\boldsymbol{\mu}} = m(\mathbf{X}, \mathbf{y})$, the degrees of freedom of $m(\cdot)$ is defined as

$$df(m) = \sum_{i=1}^{n} cov(\hat{\mu}_i, y_i)/\sigma^2. \tag{3.1}$$

The rationale is that more complex models would try to fit the data better, and hence the covariance between observed and fitted pairs would be higher. This expression is not directly observable except for certain simple cases, for example, when $m(\mathbf{y}) = \mathbf{S}\mathbf{y}$, a linear smoother. In that case, it is not difficult to see that $df(m) = tr(\mathbf{S})$. Stein was able to overcome this hurdle for a special case when $\mathbf{y} \sim N(\boldsymbol{\mu}, \sigma^2 I)$. Using a simple equality for the Gaussian distribution, he proved that as long as the partial derivative $\partial\hat{\mu}_i/\partial y_i$ exists almost everywhere for all $i \in \{1, 2, \ldots, n\}$, the following holds

$$cov(\hat{\mu}_i, y_i) = \sigma^2 \mathbb{E}\left(\frac{\partial \hat{\mu}_i}{\partial y_i}\right).$$

Thus, we have the following unbiased estimator of the degrees of freedom for the fitting procedure $m(\cdot)$

$$\widehat{df} = \sum_{i=1}^{n} \frac{\partial \hat{\mu}_i}{\partial y_i}. \tag{3.2}$$

Using the degrees of freedom definition as in (3.1), *Efron* (2004) employed the covariance penalty approach to prove that the C_p-type statistics (*Mallow*, 1973) is an unbiased estimator of the true prediction error, where

$$C_p(\hat{\boldsymbol{\mu}}) = \frac{1}{n}\|\mathbf{y} - \boldsymbol{\mu}\|^2 + \frac{2df(\hat{\boldsymbol{\mu}})}{n}\sigma^2. \tag{3.3}$$

This reveals the important role played by the degrees of freedom in model assessment. It gives us a principled way of selecting the optimal model without going for computationally expensive methods such as cross-validation, and in certain settings it can offer significantly better prediction accuracy than such methods (*Efron*, 2004). Indeed the degrees of freedom is an integral part of almost every model selection criterion, including Bayesian Information Criterion (BIC) (*Schwarz*, 1978), generalized cross-validation (GCV) (*Golub et al.*, 1979) and so on. Many important works followed that of *Stein* (1981) and *Efron* (2004). For example, *Donoho and Johnstone* (1995) used the SURE theory to derive the degrees of freedom for the soft-thresholding operator in wavelet shrinkage; *Meyer and Woodroofe* (2000) employed this framework to derive the same for shape restricted regression; *Li and Zhu* (2008) also used this set-up to derive an unbiased estimator of the degrees of freedom for penalized quantile regression. *Zou et al.* (2007) applied the SURE theory for the popular regression shrinkage and variable selection method lasso (*Tibshirani*, 1996). This is a challenging problem because of the non-linear nature of lasso solution, which does not admit an analytical solution except for certain special cases. Using sophisticated mathematical analysis, *Zou et al.* (2007) were able to show that the number of non-zero coefficients provides an unbiased estimate of the degrees of freedom for the lasso. This is a result of great practical importance since this allows one to come up with model selection criteria such as C_p and BIC for the lasso without incurring any extra computational cost.

The degrees of freedom for the reduced rank estimators also proves to be a challenging problem because of the non-linearity of the estimator. As we will see shortly, even though it admits a closed-form solution, the solution is highly non-linear depending on singular value decomposition of the least squares solution $\widehat{\mathbf{Y}}$ described in (3.4). In the next several sections, we study the degrees of freedom of a general class of reduced rank estimators in the framework of SURE and propose a finite-sample exactly unbiased estimator. The importance of such an estimator has been emphasized repeatedly by *Shen and Ye* (2002), *Efron* (2004), *Zou et al.* (2007) among others.

To overcome the analytical difficulty in computing the degrees of freedom, *Ye* (1998) and *Shen and Ye* (2002) proposed the generalized degrees of freedom approach, where they evaluate (3.2) numerically, using data perturbation techniques to compute an approximately unbiased estimator of the degrees of freedom. *Efron* (2004) also proposed a bootstrap based idea to arrive at an approximately unbiased estimator of (3.1). Though these kind of simulation based approaches allow us to extend the degrees of freedom approach to many highly non-linear modeling frameworks, they are computationally expensive. Also this type of numerical solutions does not admit any closed form expression making investigation of the theoretical properties an extremely difficult task, thus limiting our insight.

3.3 A class of reduced rank estimators

Recall the multivariate linear regression model as in (4.1). Let $\widehat{\mathbf{Y}}$ be the least squares estimate which admits a singular value decomposition of the form

$$\widehat{\mathbf{Y}} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{+}\mathbf{X}^T\mathbf{Y} = \underset{n\times\bar{r}}{\mathbf{W}}\,\underset{\bar{r}\times\bar{r}}{\mathbf{D}}\,\underset{\bar{r}\times q}{\mathbf{V}^T}, \tag{3.4}$$

here $(\mathbf{A})^{+}$ denotes the Moore-Penrose inverse (*Moore*, 1920; *Penrose*, 1955) of a generic matrix $\mathbf{A}$. Note that this is well defined even when $p, q > n$ or the design matrix $\mathbf{X}$ is of low rank. $\mathbf{W}$ and $\mathbf{V}$ are orthogonal matrices that represent the left and right singular vectors and $\mathbf{D} = diag\{d_i, i = 1, \ldots, \bar{r}\}$ with $d_1 \geq \cdots \geq d_{\bar{r}} > 0$ are the singular values of $\widehat{\mathbf{Y}}$. Without loss of generality we assume that, $rank(\widehat{\mathbf{Y}}) = \bar{r} = \min\{r_x, q\}$, where r_x denotes the rank of the design matrix. We will denote the k-th column of $\mathbf{W}$ and $\mathbf{V}$ by w_k and v_k respectively. Using the Eckart-Young theorem (*Eckart and Young*, 1936), it is not difficult to show that the reduced rank regression estimator for (1.3) can be expressed as

$$\widehat{\mathbf{Y}}(r) = \widehat{\mathbf{Y}}\sum_{k=1}^{r} v_k v_k^T = \mathbf{W}^{(r)}\mathbf{D}^{(r)}\mathbf{V}^{(r)T}, \quad r = 1, \ldots, \bar{r}, \tag{3.5}$$

where $\mathbf{A}^{(r)}$ denotes the first r-columns of a generic matrix $\mathbf{A}$. This rank constrained estimation procedure can also be viewed under a more generalized penalized least squares framework

$$\min_{\mathbf{B}} \left\{ \frac{1}{2}\|\mathbf{Y} - \mathbf{X}\mathbf{B}\|_F^2 + \lambda\mathcal{P}(\mathbf{B}) \right\}, \tag{3.6}$$

in which the penalty is proportional to the rank of the coefficient matrix $\mathbf{B}$, i.e., $\mathcal{P}(B) = rank(B)$ (*Bunea et al.*, 2011). This leads to a hard-thresholding of the singular values of $\widehat{\mathbf{Y}}$. More generally, under the regularized estimation framework

(3.6), a set of reduced-rank estimators may be indexed by the regularization parameter λ, which controls the penalty level and hence the model's complexity. In light of that, we consider a broad class of such reduced-rank estimators defined as

$$\tilde{\mathbf{Y}}(\lambda) = \mathbf{X}\tilde{\mathbf{B}}(\lambda) = \sum_{k=1}^{\bar{r}} s_k(d_k, \lambda) d_k \mathbf{w}_k \mathbf{v}_k^\top = \widehat{\mathbf{Y}} \sum_{k=1}^{\bar{r}} s_k(d_k, \lambda) \mathbf{v}_k \mathbf{v}_k^\top, \tag{3.7}$$

where each $s_k(d_k, \lambda) \in [0, 1]$ is a function of d_k and λ, and they satisfy $s_1(d_1, \lambda) \geq \cdots \geq s_{\bar{r}}(d_{\bar{r}}, \lambda) \geq 0$. To avoid confusion, we may simply write $s_k(d_k, \lambda) = s_k(\lambda) = s_k$. The reduced rank regression/rank selection criterion estimator can be viewed as a special case of this general framework with $s_k(d_k, r) = \mathbb{1}(k \leq r) \in \{0, 1\}$, $r = 1, \ldots, \bar{r}$. Note that the solutions are indexed by the rank constraint r, instead of a continuous penalty parameter λ. This class of estimators has the same set of singular vectors as the reduced rank regression estimator in (3.5), but may have different singular value estimates given by some shrunk or thresholded versions of the estimated singular values from least squares. Such estimators can be obtained from a non-convex singular-value penalization or thresholding operations (*She*, 2009, 2012; *Chen and Huang*, 2012). The class of estimators (3.7) is computationally efficient and possesses many desirable theoretical properties, such as, rank selection consistency and achieving minimax error bound (*Bunea et al.*, 2011; *Chen and Huang*, 2012) under both the classical and the high-dimensional asymptotic regimes. Some examples include the reduced rank regression, rank selection criterion (*Bunea et al.*, 2011), the nuclear norm penalized estimator under an orthogonal design (*Yuan et al.*, 2007), and the adaptive nuclear norm estimator proposed by *Chen and Huang* (2012).

3.4 Degrees of freedom of reduced rank estimators

In the previous section we discussed a broad class of reduced rank estimators covering both hard-thresholding and soft-thresholding of the singular values of $\widehat{\mathbf{Y}}$. Next we apply definition (3.2) to such multivariate regression estimators to estimate the degrees of freedom. To answer that we start by rewriting the multivariate linear regression model (4.1) as follows

$$\underset{nq\times 1}{vec(\mathbf{Y})} = \underset{nq\times pq}{[I_q \otimes \mathbf{X}]} \underset{pq\times 1}{vec(\mathbf{B})} + \underset{nq\times 1}{vec(\mathbf{E})},$$

where $\otimes$ denotes the usual Kronecker product between matrices, and $vec(.)$ stands for the column-wise vectorization operator on a matrix. We will first derive the results for the special case of reduced rank regression estimator (3.5) and later extend it to the general class of model (3.7). Applying definition (3.2) we get

$$\widehat{df}(r) = tr\left\{\frac{\partial vec(\widehat{\mathbf{Y}}(r))}{\partial vec(\mathbf{Y})}\right\}, \quad r = 1, \ldots, \bar{r}, \tag{3.8}$$

where $tr(\cdot)$ denotes the trace operator for a real square matrix. Recall that we assumed $rank(\widehat{\mathbf{Y}}) = \bar{r} = \min\{r_x, q\}$ which is not restrictive in general and does not depend on the dimensions of the problem. Let $\mathbf{X}^\top\mathbf{X} = \mathbf{Q}\mathbf{S}^2\mathbf{Q}^\top$ be the eigen decomposition of $\mathbf{X}^\top\mathbf{X}$, i.e., $\mathbf{Q} \in \mathbb{R}^{p\times r_x}$, $\mathbf{Q}^\top\mathbf{Q} = \mathbf{I}$, and $\mathbf{S} \in \mathbb{R}^{r_x\times r_x}$ is a diagonal matrix with positive diagonal elements. Then, the Moore-Penrose inverse of $\mathbf{X}^\top\mathbf{X}$ can be written as $(\mathbf{X}^\top\mathbf{X})^+ = \mathbf{Q}\mathbf{S}^{-2}\mathbf{Q}^\top$. Define

$$\mathbf{H} = \mathbf{S}^{-1}\mathbf{Q}^\top\mathbf{X}^\top\mathbf{Y}.$$

It then follows that $\mathbf{H} \in \mathbb{R}^{r_x\times q}$ admits an SVD of the form

$$\mathbf{H} = \mathbf{U}\mathbf{D}\mathbf{V}^\top, \tag{3.9}$$

where $\mathbf{U} \in \mathbb{R}^{r_x\times\bar{r}}$, $\mathbf{U}^\top\mathbf{U} = \mathbf{I}$, and $\mathbf{V}$, $\mathbf{D}$ are defined in (3.4). The matrix $\mathbf{H}$ shares the same set of singular values and right singular vectors with $\widehat{\mathbf{Y}}$ in (3.4), as $\mathbf{H}^\top\mathbf{H} = \widehat{\mathbf{Y}}^\top\widehat{\mathbf{Y}} = \mathbf{Y}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^+\mathbf{X}^\top\mathbf{Y}$. Moreover, $\mathbf{H}$ is full rank since $\widehat{\mathbf{Y}}$ is of rank $\bar{r} = \min(r_x, q)$. The matrix $\mathbf{H}$ plays a key role in deriving a simple form of the degrees of freedom as we shall see later. In particular, this construction allows us to avoid singularities arising from $r_x < p$ in the high-dimensional scenario. Simplifying (3.8) using matrix equalities such as $tr(\mathbf{AB}) = tr(\mathbf{BA})$ and $vec(\mathbf{ABC}) = (\mathbf{C}^\top \otimes \mathbf{A})vec(\mathbf{B})$ we obtain our unbiased estimator of the degrees of freedom of reduced rank regression as

$$\widehat{df}(r) = tr\left\{\frac{\partial \mathrm{vec}(\mathbf{U}^{(r)}\mathbf{D}^{(r)}\mathbf{V}^{(r)\top})}{\partial \mathrm{vec}(\mathbf{H})}\right\} = tr\left\{\frac{\partial \mathrm{vec}(\mathbf{H}(r))}{\partial \mathrm{vec}(\mathbf{H})}\right\} = \sum_{i=1}^{r_x}\sum_{j=1}^{q}\frac{\partial h_{ij}(r)}{\partial h_{ij}}, \tag{3.10}$$

where $\mathbf{H}(r) = \mathbf{U}^{(r)}\mathbf{D}^{(r)}\mathbf{V}^{(r)\top} = (h_{ij}(r))_{r_x\times q}$ is the rank r approximation to $\mathbf{H}$. The details of this derivation could be found in the Appendix. For the general class of reduced-rank estimators in (3.7), we have

$$\tilde{\mathbf{Y}}(\lambda) = \mathbf{XQS}^{-1}\mathbf{H}\sum_{k=1}^{\bar{r}} s_k(d_k,\lambda)\mathbf{v}_k\mathbf{v}_k^\top = \mathbf{XQS}^{-1}\mathbf{U}\tilde{\mathbf{D}}(\lambda)\mathbf{V}^\top,$$

where $\tilde{\mathbf{D}}(\lambda) = \text{diag}\{s_k(d_k,\lambda)d_k, k = 1, ..., \bar{r}\}$. Once again using similar matrix algebra we arrive at a simpler expression for the degrees of freedom for the general class of reduced rank models

$$\tilde{df}(\lambda) = tr\left\{\frac{\partial\text{vec}\{\mathbf{U}\tilde{\mathbf{D}}(\lambda)\mathbf{V}^\top\}}{\partial\text{vec}(\mathbf{H})}\right\} = tr\left\{\frac{\partial\text{vec}\{\tilde{\mathbf{H}}(\lambda)\}}{\partial\text{vec}(\mathbf{H})}\right\}, \qquad (3.11)$$

where $\tilde{\mathbf{H}}(\lambda) = \mathbf{U}\tilde{\mathbf{D}}(\lambda)\mathbf{V}^\top$. It is now clear that the problem boils down to determining the divergence of a low-rank approximation of the matrix $\mathbf{H}$ with respect to $\mathbf{H}$ itself. This involves the derivatives of its singular values and singular vectors. Note that the singular values and vectors of a matrix are not only highly non-linear functions of the underlying matrix, they are also discontinuous on certain subsets of matrices (*O'Neil*, 2005). This makes that degrees of freedom calculation for the reduced rank regression is a rather challenging problem. *Stein* (1973) used derivatives of the singular values of a positive semi-difinite matrix to estimate the risk improvement for a class of estimators for the mean of a multivariate Gaussian distribution. *Tsukuma* (2008) also used a similar method to prove minimaxity for Bayes estimators for the mean matrix for a Gaussian distribution. We note that our set-up is very different from the ones considered by *Stein* (1973) and *Tsukuma* (2008). Specifically, we consider a regression setting where the design matrix makes the derivation more challenging. Also as we aim to estimate the degrees of freedom of the model we need the derivatives of both singular values and vectors to compute the right hand side of (3.11). There has also been a considerable amount of work on the smoothness and differentiability of the singular value decomposition of a real matrix in applied mathematics literature; main references include *Magnus and Neudecker* (1998), *O'Neil* (2005) and *de Leeuw* (2007). In view of this, we will proceed in two main steps:

1. Derive the partial derivatives in (3.10) and (3.11) under the condition that $\mathbf{H}$ does not have repeated singular values, i.e., $d_1 > d_2 > \cdots > d_{\bar{r}} > 0$. Use them to obtain an explicit exact unbiased estimator of degrees of freedom.

2. Prove that the set where the partial derivatives do not exist has Lebesgue mea-

sure 0.

The following two sections will address the aforementioned steps respectively and thus will complete the derivation of degrees of freedom estimator for the reduced rank estimators for multivariate regression under the SURE framework.

3.5 Proposed estimator

We start by examining the derivatives of the singular values and singular vectors of a matrix with respect to an entry of the matrix itself. All the proofs are provided in the Appendix.

Theorem III.1. *Suppose* $\mathbf{H}$ *is an* $r_x \times q$ *matrix of rank* q*, with* $r_x \geq q$*. Let its SVD be given by* $\mathbf{H} = \mathbf{UDV}^\top$*, where* $\mathbf{U} \in \mathbb{R}^{r_x \times q}$*,* $\mathbf{U}^\top\mathbf{U} = \mathbf{I}$*,* $\mathbf{V} \in \mathbb{R}^{q \times q}$*,* $\mathbf{V}^\top\mathbf{V} = \mathbf{I}$*, and* $\mathbf{D} = \mathrm{diag}\{d_i, i = 1, ..., q\}$ *with* $d_1 > \cdots > d_q > 0$*. Then for each* $1 \leq i \leq r_x$*,* $1 \leq j \leq q$*, and* $1 \leq k \leq q$*,*

$$\frac{\partial \mathbf{v}_k}{\partial h_{ij}} = -\,(\mathbf{H}^\top\mathbf{H} - d_k^2\mathbf{I})^-(\mathbf{H}^\top\mathbf{Z}^{(ij)} + \mathbf{Z}^{(ij)\top}\mathbf{H})\mathbf{v}_k, \tag{3.12}$$

$$\frac{\partial d_k}{\partial h_{ij}} = \frac{1}{2d_k}\mathbf{v}_k^\top(\mathbf{H}^\top\mathbf{Z}^{(ij)} + \mathbf{Z}^{(ij)\top}\mathbf{H})\mathbf{v}_k, \tag{3.13}$$

where $(\mathbf{H}^\top\mathbf{H} - d_k^2\mathbf{I})^- = \mathbf{V}(\mathbf{D}^2 - d_k^2\mathbf{I})^+\mathbf{V}^\top$ *with* $(\cdot)^+$ *denoting the Moore-Penrose inverse, and* $\mathbf{Z}^{(ij)} = \partial\mathbf{H}/\partial h_{ij}$ *is an* $r_x \times q$ *matrix of zeros with only its* (i,j)*th entry being one.*

Without loss of generality, we have assumed $r_x \geq q$ in the above theorem. When $r_x \leq q$, the same results could be presented for $\mathbf{H}^\top$ with exchanged r_x and q. Theorem III.1 is established from the general results in *de Leeuw* (2007) about the derivatives of a generalized eigen-system. To ensure the derivatives are well-defined, we have assumed that the singular values are distinct. This is merely a restriction for real applications, as the observed singular values rarely coincide, a formal proof is provided in the next section.

It is not immediately clear whether the derived unbiased estimators in (3.10) and (3.11) may admit explicit form. Examining the SVD structure of $\mathbf{H}$ sheds light on this problem. The pairs of singular vectors $(\mathbf{u}_k, \mathbf{v}_k)$ are orthogonal to each other, representing distinct directions in $\mathbb{R}^{r_x \times q}$ without any redundancy. Intuitively, these directions themselves are not distinguishable from each other, and their relative importance or contribution in constituting the matrix $\mathbf{H}$ are entirely revealed by the singular values. This suggests that the complexity of reduced-rank estimation, as reflected by the relative complexity of a low rank approximation $\mathbf{H}(r)$ or $\tilde{\mathbf{H}}(\lambda)$ with respect to $\mathbf{H}$, may only depend on the singular values of the matrix $\mathbf{H}$ and the mechanism of singular-value shrinkage or thresholding. This is the main intuition that motivated the findings for explicit forms of (3.10) and (3.11), which are summarized in the following theorems.

Theorem III.2. *Let $\widehat{\mathbf{Y}}$ be the least squares estimator in* (3.4)*. Let $\bar{r} = rank(\widehat{\mathbf{Y}}) = \min(r_x, q)$ and suppose the singular values of $\widehat{\mathbf{Y}}$ satisfy $d_1 > \cdots > d_{\bar{r}} > 0$. Consider the reduced-rank estimator $\widehat{\mathbf{Y}}(r)$ in* (3.5)*. An unbiased estimator of the effective degrees of freedom is*

$$\widehat{df}(r) = \begin{cases} \max(r_x, q)r + \displaystyle\sum_{k=1}^{r}\sum_{l=r+1}^{q} \frac{d_k^2 + d_l^2}{d_k^2 - d_l^2}, & r < \bar{r}; \\ r_x q, & r = \bar{r}. \end{cases}$$

The results are further generalized to the class of reduced-rank estimators in (3.7). It is worth noting that the weights $s_k(d_k, \lambda)$ are treated as random quantities since they are usually some functions of the singular values.

Theorem III.3. *Let $\widehat{\mathbf{Y}}$ be the least squares estimator in* (3.4)*. Let $\bar{r} = rank(\widehat{\mathbf{Y}}) = \min(r_x, q)$ and suppose the singular values of $\widehat{\mathbf{Y}}$ satisfy $d_1 > \cdots > d_{\bar{r}} > 0$. Consider the reduced-rank estimator $\tilde{\mathbf{Y}}(\lambda)$ in* (3.7)*, and let $\tilde{r} = \tilde{r}(\lambda) = \max\{k : s_k(d_k, \lambda) > 0.\}$. An unbiased estimator of the effective degrees of freedom is*

$$\tilde{df}(\lambda) = \begin{cases} \max(r_x, q)\displaystyle\sum_{k=1}^{\tilde{r}} s_k + \sum_{k=1}^{\tilde{r}}\sum_{l=\tilde{r}+1}^{\bar{r}} \frac{s_k(d_k^2 + d_l^2)}{d_k^2 - d_l^2} + \sum_{k=1}^{\tilde{r}}\sum_{l\neq k}^{\tilde{r}} \frac{d_k^2(s_k - s_l)}{d_k^2 - d_l^2} + \sum_{k=1}^{\tilde{r}} d_k s_k', & \tilde{r} < \bar{r}; \\ \max(r_x, q)\displaystyle\sum_{k=1}^{\tilde{r}} s_k + \sum_{k=1}^{\tilde{r}}\sum_{l\neq k}^{\tilde{r}} \frac{d_k^2(s_k - s_l)}{d_k^2 - d_l^2} + \sum_{k=1}^{\tilde{r}} d_k s_k', & \tilde{r} = \bar{r}. \end{cases}$$

where for simplicity we write $s_k = s_k(d_k, \lambda)$ and $s_k' = \partial s_k(d_k, \lambda)/\partial d_k$.

The explicit formulae presented in the above theorems facilitate further exploration of the behaviors and properties of the degrees of freedom. Consider the unbiased estimator for reduced rank regression in Theorem III.2. It is always true that

$$\widehat{df}(r) \geq \max(r_x, q)r + \sum_{k=1}^{r} \sum_{l=r+1}^{\min(r_x,q)} \frac{d_k^2 + 0}{d_k^2 - 0} = (r_x + q - r)r, \qquad r = 1, ..., \bar{r}. \tag{3.14}$$

This suggests that the proposed estimator is always greater than the naive estimator, i.e., the number of free parameters $(r_x + q - r)r$. Similar to the lasso method in univariate regression problems (*Tibshirani*, 1996; *Zou et al.*, 2007), the reduced-rank estimation can be viewed as a latent factor selection procedure, in which we both construct and search over as many as $\bar{r}$ latent linear factors. Therefore, the increments in the degrees of freedom as shown in (3.14) can be interpreted as the price we have to pay for performing this latent factor selection. For the general methods considered in Theorem III.3, this inequality no longer holds, due to the shrinkage effects induced by the weights $0 \leq s_k \leq 1$. The reduction in the degrees of freedom due to singular-value shrinkage can offset the price paid for searching over the set of latent variables. Therefore, similar to lasso, adaptive singular-value penalization can provide effective control over the model complexity (*Tibshirani and Taylor*, 2011; *Chen and Huang*, 2012).

Although the unbiased estimator and the naive estimator are quite different, some interesting connections can be made. The two estimators are close to each other when they are evaluated at the true underlying rank, especially when the signal is strong relative to the noise level. This phenomenon was also noted in the empirical studies. Suppose the true model rank is $\text{rank}(\mathbf{B}) = r^*$. Intuitively, the $\bar{r} - r^*$ smallest singular values from least squares may be close to zero and are not comparable to the r^* largest ones; using the approximation $d_k \approx 0$, $k = r^* + 1, ..., \bar{r}$, we obtain $\widehat{df}(r^*) \approx (r_x + q - r^*)r^*$. A more rigorous argument can be made from either classical or high-dimensional theoretical perspective. In classical large n settings, under standard assumptions, the consistency of the least squares estimation can be readily established (*Reinsel and Velu*, 1998). Using techniques such as the perturbation expansion of matrices (*Izenman*, 1975), the consistency of $\widehat{\mathbf{Y}}$ implies the consistency of the estimated singular values, i.e., the first r^* estimated singular values converge to their nonzero true counterparts while the rest converge to zero in probability. It follows that

$$\widehat{df}(r^*) \rightarrow_p (r_x + q - r^*)r^* \tag{3.15}$$

in probability as $n \to \infty$. An immediate implication of this result is that for each

$r = 1, ..., \bar{r}$, if we assume the true model is of rank r, then in an asymptotic sense, the number of free parameter, $(r_x + q - r)r$, is the correct degrees of freedom to use. This clearly relates to the error degrees of freedom of the classical asymptotic χ^2 statistic from the likelihood ratio test of $H_0 : \text{rank}(\mathbf{B}) = r$ (*Izenman*, 1975), for each $r = 1, ..., \bar{r}$. In high-dimensional models, non-asymptotic prediction error bounds have been developed for the considered reduced-rank estimation methods, and the minimax convergence rate in fact coincides with the number of free parameters (*Rohde and Tsybakov*, 2011; *Bunea et al.*, 2011; *Chen and Huang*, 2012). These results provide further justification of the proposed unbiased estimator and reveal the limitations, the underlying assumptions and the asymptotic nature of the naive estimator.

The derived formulae also reveal some interesting behaviors of rank reduction. In essence, the reduced-rank methods distinguish the signal from the noise by examining the estimated singular values from least squares estimation: the large singular values more likely represent the signals while the small singular values mostly correspond to the noise (*Bunea et al.*, 2011; *Chen and Huang*, 2012). By rank reduction, we aim to recover the signals exceeding certain noise level. Consider the case when d_k and d_{k+1} are close for some $k = 1, ..., \bar{r} - 1$. It can be argued that the true model rank is unlikely to be k, because the $(k + 1)$th layer and the kth layer are hardly distinguishable. Indeed, this is reflected from the degrees of freedom: for $r = k$, the formula includes a term $(d_k + d_{k+1})/(d_k - d_{k+1})$, which can be excessively large. On the other hand, there is no such term for $r = k + 1$. Consequently, the unbiased estimator of the degrees of freedom may not monotonically increase as the rank r increases, in contrast to the naive estimator. In the above scenario, the estimates for $r = k$ can even be larger than that of $r = k + 1$. This automatically reduces the chance of k being selected as the final rank.

3.6 Existence of partial derivatives almost everywhere

One of the main technical assumptions for Stein's degrees of freedom estimator is that the partial derivatives must exist almost everywhere. Theorem III.1 gives us the condition, $d_1 > d_2 > \ldots > d_{\bar{r}} > 0$ for the existence of the partial derivatives of singular values and singular vectors of $\mathbf{H} \in \mathbb{R}^{r_x \times q}$, where $\{d_i\}_{i=1}^{\bar{r}}$ denote the singular values of $\mathbf{H}$. Also recall that $\bar{r} = \min\{r_x, q\}$. Therefore, to apply Stein's framework we must show that matrices with full rank and non-repeated singular values are *"dense"* in the set of all real matrices of dimension $r_x \times q$. The following theorem gives that result.

Theorem III.1. *Let $\mathbb{R}^{r_x \times q}$ be the space of all real-valued $r_x \times q$ dimensional matrices equipped with the Lebesgue measure μ. Also, let $\mathcal{S} \subseteq \mathbb{R}^{r_x \times q}$ denote the subset of matrices that have full rank and no repeated singular values. Then $\mu(\mathcal{S}) = 1$.*

To prove the theorem, we start with a few definitions and facts from algebraic geometry and matrix analysis.

Definition III.2. An algebraic variety over $\mathbb{R}^k$(or $\mathbb{C}^k$) is defined as the set of points satisfying a system of polynomial equations $\{f_\ell(x_1, x_2, \ldots, x_k) = 0; \quad \ell \in \mathcal{I}\}$.

Here each $f_\ell(\cdot)$ is a polynomial function of its arguments and $\mathcal{I}$ denotes an index set. If at least one of the $f_\ell(\cdot) \not\equiv 0$, then it is called a proper sub-variety. Note that a proper sub-variety must be of dimension less than k and therefore has Lebesgue measure 0 in $\mathbb{R}^k$ (*Allman et al.*, 2009). For a more detailed discussion, we recommend *Hartshorne* (1977) or *Cox et al.* (2007).

Proposition III.3. *(Laub, 2004) Any square symmetric matrix $\mathbf{M} \in \mathbb{R}^{k \times k}$ has at least one repeated eigenvalue if and only if rank $(\mathbf{M} \otimes \mathbf{I}_k - \mathbf{I}_k \otimes \mathbf{M}) < (k^2 - k)$.*

Now we prove the theorem. First we define

$$\begin{aligned}
\mathcal{S}_1 &= \{\mathbf{A} \in \mathbb{R}^{r_x \times q} : \mathbf{A} \text{ has at least one 0 singular value}\}, \\
\mathcal{S}_2 &= \{\mathbf{A} \in \mathbb{R}^{r_x \times q} : \mathbf{A} \text{ has at least one repeated singular value}\}.
\end{aligned}$$

Note that $\mathcal{S}^c = \mathcal{S}_1 \cup \mathcal{S}_2$, thus it is enough to show that $\mu(S_1) = 0$ and $\mu(S_2) = 0$. By definition III.2 and the discussion above it suffices to show that $\mathcal{S}_1$ and $\mathcal{S}_2$ are proper sub-varieties of $\mathbb{R}^{r_x \times q}$. Note that $\mathcal{S}_1$ can be rewritten as follows

$$\mathcal{S}_1 = \{\mathbf{A} \in \mathbb{R}_{r_x \times q} : det(\mathbf{A}^\top \mathbf{A}) = 0\}.$$

Here $det(\cdot)$ denotes the determinant operator for a square matrix. Note that $det(\mathbf{A}^\top \mathbf{A})$ is a non-trivial polynomial in entries of $\mathbf{A}$ and hence $\mathcal{S}_1$ is a proper sub-variety and has Lebesgue measure 0. For $\mathcal{S}_2$ note that if $\mathbf{A} \in \mathbb{R}^{p \times q}$ has at least one repeated singular value, it implies that $\mathbf{A}^T \mathbf{A} \in \mathbb{R}^{p \times q}$ has at least one repeated eigenvalue. Then in view of proposition III.3, $\mathcal{S}_2$ can be reformulated as

$$\mathcal{S}_2 = \left\{\mathbf{A} \in \mathbb{R}_{r_x \times q} : rank\left(\mathbf{A}^\top \mathbf{A} \otimes \mathbf{I}_q - \mathbf{I}_q \otimes \mathbf{A}^\top \mathbf{A}\right) < (q^2 - q)\right\}.$$

This is an algebraic variety since it can be expressed as the solution to all minors of order $\geq (q^2 - q)$ being equal to 0, which are all polynomial equations in the entries of $\mathbf{A}$. Thus, we have shown that, $\mu(\mathcal{S}_1 \cup \mathcal{S}_2) = 0$.

3.7 Simulation studies

In this section, we evaluate the performance of the proposed method by simulation studies. Specifically, we aim to demonstrate two things: 1) the exact unbiased estimator of the degrees of freedom for the reduced rank regression is in general significantly higher than the naive estimator; 2) using the exact estimator of the degrees of freedom enables us to gain prediction accuracy over the naive estimator.

3.7.1 Unbiasedness

In this simulation, we aim to show that the degrees of freedom estimator defined via Theorem III.2 is unbiased and it can be significantly higher than the naive estimator that simply counts the number of free parameters. Here unbiasedness is defined over the error distribution, and we treat $\mathbf{X}$ as a fixed design matrix. We conduct the study at two different parameter settings one for low-dimension and one for high-dimension. Parameters of the setting are as follows

$$\begin{aligned} \text{Setting I} &: \quad n = 100, p = 20, q = 12, r_0 = 6 \\ \text{Setting II} &: \quad n = 40, p = 80, q = 50, r_0 = 10 \end{aligned}$$

where r_0 denotes the true rank of $\mathbf{B}$. Let $\mathbf{\Sigma}$ denote the covariance matrix of the predictor variables, $\mathbf{X}$, and we set $\mathbf{\Sigma}_{jj'} = 0.3^{|j-j'|}$. Rows of the predictor matrix are generated independently from $N_p(\mathbf{0}, \mathbf{\Sigma})$. To control the singular structure of $\mathbf{B}$ through the covariance of signals $\mathbf{XB}$, $\mathbf{B}^T\mathbf{\Sigma}\mathbf{B}$, we take the left singular vectors of $\mathbf{B}$ the same as the eigenvectors of $\mathbf{\Sigma}$, whereas the right singular vectors of $\mathbf{B}$ are generated by orthogonalizing a random standard normal matrix. The difference between successive non-zero singular value of $\mathbf{B}$ is fixed at 2. The error matrix is generated from i.i.d. standard normal distribution. We replicate the process 200 times; note that the design matrix remains fixed. We compare the proposed exact method against the data perturbation technique (*Ye*, 1998) and the Monte-Carlo estimator of the true degrees of freedom which is computed from (3.1). For the data perturbation method, we consider 50 perturbations of the response matrix for each replication to estimate the partial derivatives numerically. We used the choice of 0.1σ for the perturbation size, where σ is the error standard deviation. Ideally we would expect the proposed exact estimator to be fairly close to the data perturbation and Monte-carlo estimator on average. We compare estimators against the naive degrees of freedom estimate namely, $df_n(r) = r(r_x + q - r)$, which denotes the number of free parameters in a $p \times q$ matrix of rank r. Note that the naive estimator does not depend on the data.

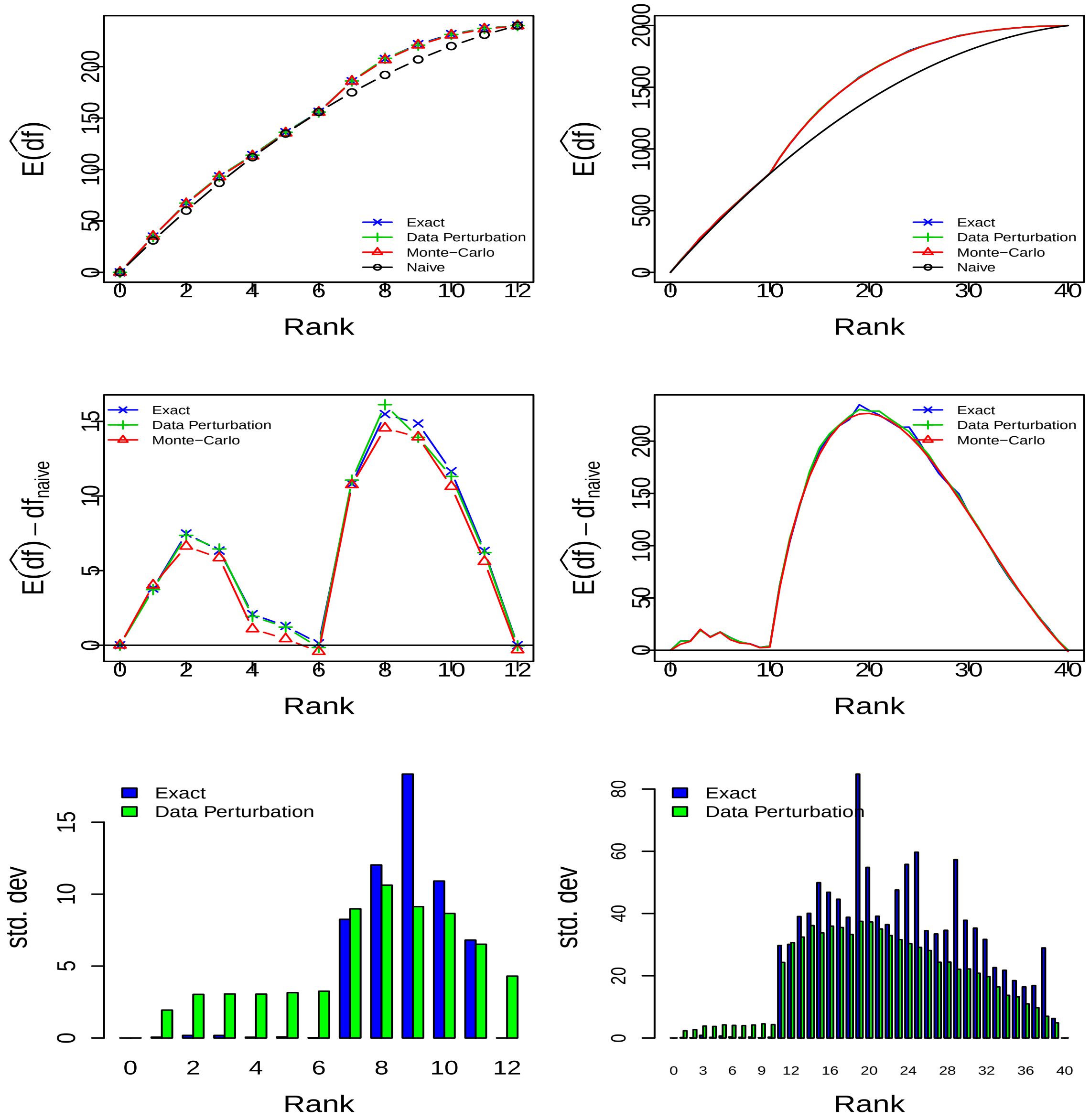

Figure 3.1: Left column: low dimensional setting, right column: high-dimensional setting

On the top row of Figure 3.1 we see that for both high-dimensional and low-dimensional settings the proposed exact method, the data perturbation estimator and the Monte-Carlo estimator are nearly identical; further, they are significantly higher than the naive estimator, as indicated in the middle row of Figure 3.1. The difference is especially large once we go above the correct rank. It also justifies our theoretical intuition that the exact estimators seem to match the naive estimator very closely at the true rank. The bottom panels allow us to get a sense of the variability of the estimation procedures. Standard error for the exact method is orders of magnitudes smaller than that of data perturbation below the true rank but once we go above the true rank the standard errors of the exact estimator becomes drastically higher. This arises from the fact that once we go above the true rank, the singular values of $\widehat{\mathbf{Y}}$ basically correspond to noise, and can be very close to each other. Hence slight perturbations of the data might lead to different singular directions being selected, which implies higher variability in model complexity. This has also been noted by *Ye* (1998), that is, if we are trying to fit pure error components, the degrees of freedom tends to be higher and unstable.

3.7.2 Prediction performance

The previous set of simulations have shown that the exact degrees of freedom estimator can be significantly different from the number of free parameters estimator. Degrees of freedom estimates are commonly used in various model selection criteria. In this subsection, we aim to show that for reduced rank regression, we can gain in prediction accuracy by using the exact degrees of freedom estimator in a model selection criterion instead of the naive estimator. Since our focus is on prediction accuracy, we consider generalized cross-validation(GCV) (*Golub et al.*, 1979) as our model selection criterion. This choice was motivated by the fact that it does not require an estimate for the error variance. Other popular choices such as Mallows C_p (*Mallow*, 1973) require an estimate of error variance which is hard to obtain in high-dimensional settings. In the context of reduced rank regression, the GCV criterion is defined as follows

$$GCV(r) \quad = \quad \frac{nq\|\mathbf{Y} - \widehat{\mathbf{Y}}(r)\|_F^2}{(nq - df(r))^2}.$$

We select the model that minimizes the GCV criterion over $1 \leq r \leq \min\{n, p, q\}$. Once again we choose a low-dimensional and a high-dimensional setting for a comprehensive comparison.

$$\begin{aligned} \text{LD Setting} &: n = 50, p = 12, q = 10, r_0 = 3 \\ \text{HD Setting} &: n = 40, p = 80, q = 50, r_0 = 5 \end{aligned}$$

For each setting we consider two different levels for error variance, namely, $\sigma^2 = 1$ and 4. This allows us to controls the signal to noise ratio defined as $SNR = d_{r_0}(\mathbf{XB})/d_1(\mathbf{E})$. The numerator stands for the smallest non-zero singular value of the signal matrix, a measure of the signal strength, whereas the largest singular value of the error matrix measures the noise strength (*Bunea et al.*, 2011). Correlation among predictor variables is kept at a moderate level of 0.5. The data generation scheme remains the same as before. We fit the optimal model based on GCV with the exact degrees of freedom (GCV(e)) and GCV with the naive degrees of freedom (GCV(n)) and report the following: estimation error $Est = 100\|\mathbf{B} - \widehat{\mathbf{B}}\|_F^2/(pq)$, the prediction error $Pred = 100\|\mathbf{XB} - \mathbf{X}\widehat{\mathbf{B}}\|_F^2/(nq)$ as well as the selected rank. . Table 3.1 summarizes the results. We report the averages over 100 replications and the numbers inside the parenthesis indicate standard error.

Table 3.1: Prediction performance comparison between different model selection criteria

Error Variance and SNR	Performance Measure	LD setting GCV(e)	LD setting GCV(n)	HD setting GCV(e)	HD setting GCV(n)
	Est	1.56(0.4)	1.80(0.8)	3.25(0.5)	3.30(0.5)
$\sigma^2 = 1$, $SNR \approx 1$	Pred	11.95(2.2)	12.97(3.4)	22.89(1.5)	28.28(4.3)
	Rank	3.01(0.1)	3.18(0.4)	4.84(0.4)	5.30(0.5)
	Est	6.00(2.7)	7.47(3.4)	3.77(0.5)	4.00(0.6)
$\sigma^2 = 4$, $SNR \approx 0.5$	Pred	50.64(10.8)	54.31(10.8)	78.48(6.2)	89.93(17.4)
	Rank	2.41(0.6)	2.86(0.6)	4.00(0.0)	4.46(0.6)

We find that using the proposed exact degrees of freedom estimator in GCV criterion performs better in terms of prediction accuracy than its naive counterpart. It has lower average estimation error and prediction error for all the settings. The relative gain is larger for the prediction error. We wish to note that similar results were obtained at other levels of correlation but were excluded to facilitate brevity. For the low-dimensional setting where an estimator of σ^2 is available we also studied the performance of Mallow's C_p criterion and once again the results were very close to the ones reported and therefore excluded. We find that in the settings with moderately high SNR, the naive degrees of freedom estimator tends to overestimate the rank leading to inflated error measures. On the other hand in the low SNR settings often the smallest non-zero singular value have very little explanatory power and therefore selecting a lower rank model enables us to do better in terms of prediction accuracy due to the bias-variance trade-off. As the exact degrees of freedom estimator is usually higher than the naive estimator it penalizes more strictly and selects a simpler model which predicts better. To get a better understanding for the comparison between the two degrees of freedom estimators, we also computed the percentage of pairwise relative gain, which is defined as follows

$$PRG = 100 \times \frac{(Pred(n) - Pred(e))}{Pred(e)}\%,$$

where $Pred(e)$ denotes the prediction error when using exact degrees of freedom estimator in GCV criterion, similarly $Pred(n)$ denotes the prediction error when using the naive degrees of freedom estimator in GCV. Note that these ratios are computed on a per data set basis. As we can see in Figure 3.2, the boxplots tend to stay above zero almost always indicating that the exact degrees of freedom outperforms the naive estimator consistently. Also the relative gain is larger in the high-dimensional scenario.

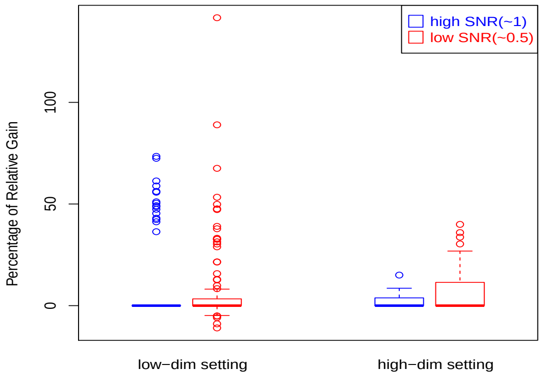

Figure 3.2: Relative gain in prediction error by using the exact estimator of degrees of freedom over the naive estimator

3.8 Analysis of Arabidopsis thaliana data

In this section, we apply the proposed degrees of freedom methodology to fit a reduced rank model to a genetic association data set that was published in *Wille et al.* (2004). This is a microarray experiment aimed at understanding the regulatory control mechanisms between the isoprenoid gene network in Arabidopsis thaliana plant (more commonly known as thale cress or mouse-ear cress). It is known that isoprenoids serve many important biochemical functions in plants. To monitor the gene-expression levels, 118 GeneChip microarray experiments were carried out. The predictors consist of 39 genes from two isoprenoid bio-synthesis pathways namely MVA and MEP, whereas the responses consist of gene-expression of 795 genes from 56 metabolic pathways, many of which are downstream of the two pathways considered as predictors. Thus some of the responses are expected to show significant associations to the predictor genes. To facilitate it further, we select two downstream pathways namely, Caroteniod and Phytosterol as our responses. It has already been proven experimentally that the Carotenoid pathway is strongly attached to the MEP pathway, whereas the Phytosterol pathway is significantly related to the MVA pathway. See *Wille et al.* (2004) and the references therein for a more detailed discussion on the biological aspects. Finally we have 118 observations on $p = 39$ predictors and $q = 36$ responses. All the predictors and responses are log-transformed to reduce the skewness of the data. We also standardize the responses in order to make them comparable.

We split the data set randomly into training and test sets of equal size. The model is fit using the training samples and then we use it to predict on the test set. The performance measure under consideration is the usual mean squared prediction error

$$\text{MSPE} = \frac{2}{nq}\|\mathbf{Y}_{\text{test}} - \widehat{\mathbf{Y}}_{\text{test}}\|_F^2. \tag{3.16}$$

The entire process is repeated 100 times based on random splits to ensure that the results remain robust to the process of splitting. We used Mallow's C_p, GCV and BIC with the exact degrees of freedom and the naive degrees of freedom to select the optimal rank.

The mean squared prediction errors for each method are summarized using the boxplot in Figure 3.3. As we can see, for all three model selection criteria considered, the use of the exact unbiased estimator enables us to outperform the one which uses the naive estimator in terms of prediction accuracy. The relative gain is almost always

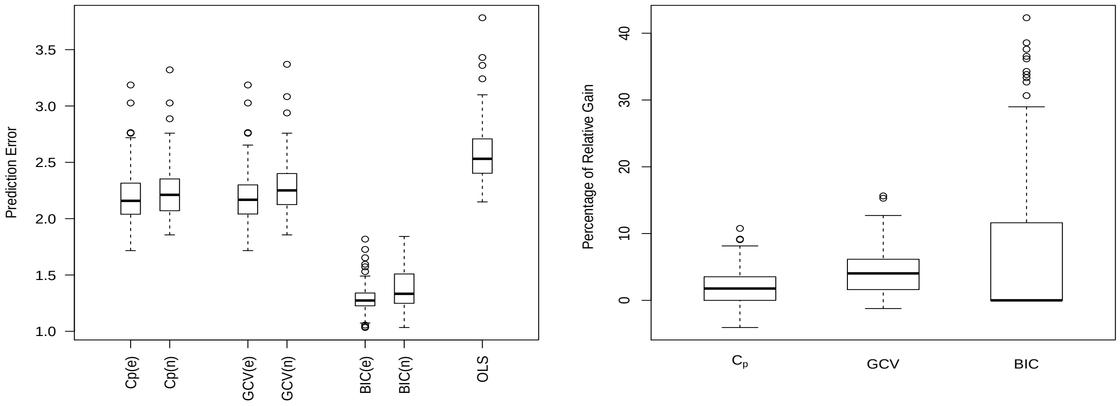

Figure 3.3: Left: boxplot of mean square prediction error of each method over 100 random splits; Right: Relative increase in prediction error for using naive degrees of freedom over the exact degrees of freedom estimator for each model selection criteria.

Table 3.2: Prediction accuracy and rank selection performance for the competing methods on the Arabidopsis thaliana data.

	Cp(e)	Cp(n)	GCV(e)	GCV(n)	BIC(e)	BIC(n)	OLS
Avg(Pred Err)	2.197	2.243	2.192	2.282	1.297	1.387	2.589
Std(Pred Err)	0.250	0.246	0.248	0.246	0.134	0.201	0.282
Mean(Est Rank)	8.760	9.710	8.680	10.520	1.090	1.480	–
Std(Est Rank)	1.15	0.83	1.27	0.97	0.38	0.76	–

positive as we can see from the right panel of the figure. Also among the three model selection criteria BIC appears to be the clear winner in terms of prediction error by selecting a very parsimonious model.

3.9 Concluding remarks

We have proposed an exact unbiased estimator of the degrees of freedom for a general class of reduced rank estimators for the multivariate linear regression model in the framework of SURE. The proposed estimator can be computed explicitly leading to an efficient model selection procedure compared to computationally expensive cross-validation or data-perturbation based methods. The closed form also provides us with some much needed insight regarding the connection between the exact and the naive degrees of freedom estimator. The proposed methodology does not make any assumption regarding the dimensions of the problem or the rank of the design matrix and is very suitable for application to high-dimensional problems $(p, q > n)$ as illustrated via several numerical examples. The methods developed here are quite general and can be extended to other related estimation procedures that employ regularization of the singular values, e.g., reduced rank ridge regression (*Mukherjee and Zhu*, 2011). There are several directions for future research. We have mainly considered the reduced-rank estimators which share the same set of singular vectors with the least squares solution. It would be interesting and challenging to extend the results for other reduced-rank methods, such as, the nuclear-norm penalized regression (*Yuan et al.*, 2007). Since reduced-rank estimation can be more effective when combined with sparse estimation, e.g., selecting latent factors of a sparse subset of original variables, it would be very interesting to extend the methodology to sparse and low-rank models (*Zou et al.*, 2007; *Chen et al.*, 2012; *Bunea et al.*, 2012). Another pressing problem concerns investigating the proposed approach in reduced rank generalized linear models (*Yee and Hastie*, 2003; *Li and Chan*, 2007; *She*, 2012). Finally, as the reduced rank methods are commonly used in multiple time series analysis, the proposed approach can be extended to these settings, including reduced rank models with multiple sets of regressors (*Velu*, 1991) and the co-integration problem (*Anderson*, 2002b).

CHAPTER IV

Non-parametric Reduced Rank Regression with Variable Selection

4.1 Introduction

We have discussed many different methods for reduced rank regression and variable selection in the context of multivariate linear regression in Section I. The next set of novel methodologies seek to combine the attractive features of both these approaches creating a framework for simultaneous rank reduction and variable selection. *Chen et al.* (2012) adopted a sparsity penalty on the singular vectors of the coefficient matrix to obtain more interpretable models. *Bunea et al.* (2012) and *Chen and Huang* (2012) both propose methodologies for joint rank and variable selection approaches for high-dimensional multivariate regression through a group lasso penalty on the coefficient matrix $\mathbf{B}$ with a rank constraint. The methods differ in terms of the optimization algorithm as well as the asymptotic analysis.

The methods reviewed in the preceding paragraphs all assume a linear parametric model. In many modern applications, there is little justification for assuming the predictors have a linear effect on the responses. In many econometric, genetic or experimental studies such as the one described in Section 4.6 it has been demonstrated that the effects are highly non-linear. One way of overcoming this drawback is to assume a non-parametric additive model. The additive form helps to avoid the *curse of dimensionality* and also makes it easy to device penalization approaches for variable selection in high-dimensional settings. A large amount of statistical literature has been dedicated to the analysis of such models of which we describe a few. *Lin and Zhang* (2006) proposed the component selection and smoothing operator (COSSO) method which incorporates an RKHS norm penalty to encourage variable selection in the smoothing-spline ANOVA set-up . The authors derive the rate of convergence of their estimator and show model selection consistency for certain special cases. But their method fails to handle the high-dimensional case where, $p > n$. *Meier et al.* (2009) introduced a sparsity-smoothness penalty to achieve variable selection for non-parametric additive models under high-dimensional setting. Under moderate smoothness assumptions the authors are able to characterize the solution to the optimization problem as a natural cubic spline and develop a computational algorithm based on B-spline basis expansions. But model selection consistency is not guaranteed even in an asymptotic sense. *Ravikumar et al.* (2009) introduced another non-parametric variable selection method called sparse additive models (SpAM). Variable selection is achieved by penalizing the ℓ_2 norm of the functional components. The estimator is computed via sparse backfitting algorithm using non-parametric smoothers. The authors prove model selection consistency but the conditions required for that are not easy to check and it is unclear whether a given set of basis functions would satisfy them. More recently, *Huang et al.* (2010) proposed a two-step adaptive group lasso method for variable selection in non-parametric additive model based on spline approximations. Model selection consistency is established under mild regularity conditions using the fact that the ordinary group lasso estimator in step-I selects the important predictors with a high probability.

Motivated by the flexibility and the versatile nature of the non-parametric regression methods in this paper we propose a framework for reduced rank estimation under such settings. Despite the broad appeal of such a model very little work has been done to extend the reduced rank framework to functional regression. In this paper we propose to model each response as a smooth function which is additive in its arguments. The underlying assumption is that the rank of true functional signal matrix (to be defined in Section 4.2) is much smaller than its dimensions thus a low-rank estimator may have better prediction accuracy and also provide more insight about the data. Variable selection also plays an important role as many of the real world problems that we face today involve high-dimensional features and a large fraction of which may carry little or no information about the responses (sparsity). We propose a two-step low-rank group lasso algorithm to encourage simultaneous dimension reduction as well as variable selection for the non-parametric additive model based on spline approximation of the functional components. The algorithm can be thought of as a non-parametric extension of the rank-constrained group lasso algorithm proposed by *Bunea et al.* (2012) in the linear case. But there are some significant differences due to the spline basis expansion which transforms the variable selection step to selecting sub-matrices in an expanded coefficient matrix. The approximation step also makes theoretical analysis of the proposed methodology much more challenging than the linear case. Appropriate bounds on the spline approximation error proves to be the key to establishing theoretical guarantees such as rank selection consistency and estimation error bounds.

The rest of the paper is organized as follows. Section 4.2 describes the non-parametric additive model in the context of multiple response regression and introduces the spline approximation. Section 4.3 contains the optimization algorithm for obtaining a low-rank solution with variable selection. We also describe the issue of tuning parameter selection in that context. Theoretical results are described in Section 4.4 as well as the required conditions. Detailed proofs are postponed to the appedix. Section 4.5 presents the results of numerical studies comparing the proposed method against relevant competitors on simulated data. Section 4.6 illustrates an application on a chemometrics experiment data. We conclude the paper with a discussion in Section 4.7.

4.2 Model: Low-rank Non-parametric Additive Models

This section introduces the non-parametric model with rank constraints and also describes the crucial spline approximation step. Let us assume that we have n observations on a p-dimensional predictor vector $X = (X_1, X_2, \dots, X_p)$ and q-dimensional response vector $Y = (Y_1, Y_2, \dots, Y_q)$. In matrix notation we have a design matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ and a response matrix $\mathbf{Y} \in \mathbb{R}^{n \times q}$ where each row of $\mathbf{X}$ and $\mathbf{Y}$ correspond to an observation. We assume that the responses are related to the preditors through the following non-parametric additive model

$$y_{ik} = \sum_{j=1}^{p} f_{kj}(x_{ij}) + \epsilon_{ik}, \; i = 1, \dots, n; \; j = 1, \dots, p; \; k = 1, \dots, q, \tag{4.1}$$

where $\epsilon_{ik} \overset{iid}{\sim} N(0, \sigma^2)$ and f_{kj} are the component functions. For identifiability purposes we assume that $\mathbb{E}(f_{kj}(X_j)) = 0$. To keep the notations simple and concise we drop the intercept term from the model. This model can be rewritten in matrix notation as

$$\mathbf{Y} = \mathbf{f}(\mathbf{X}) + \mathbf{E}, \tag{4.2}$$

where $f(x)_{ik} = \sum_{j=1}^{p} f_{kj}(x_{ij})$ and is referred to as the signal matrix. We assume a low-rank structure on the signal matrix, which explains the dependence structure among the responses given the predictors and achieves dimension reduction. To address the high-dimensionality of the predictors we assume that only a few predictors are related with the responses. The following conditions make the assumptions more concrete.

A1. $rank(\mathbf{f}(\mathbf{X})) \ll \min\{n, p, q\}$

A2. Without loss of generality $\exists J \ll p$ such that $f_{kj} \equiv 0 \; \forall j > J$ and $k = 1, 2, \ldots, q$.

Our objective is to build a non-parametric model that achieves simultaneous rank and variable selection leading to a more interpretable and accurate predictive model. Under mild smoothness assumptions on the component functions the signal matrix can be well-approximated by a set of basis functions and therefore rank and variable selection can be performed on that approximate representation. We start by approximating the non-parametric component functions by B-splines much to the spirit of *Huang et al.* (2010). B-spline basis functions are an attractive choice because they are computationally efficient and very flexible. Assume that each $X_j \in [a, b]$ where $-\infty < a < b < \infty$, and let $a = \xi_0 < \xi_1 < \ldots < \xi_{T_n} < \xi_{T_n+1} = b$ be a partition of that interval. Let $T_n = n^{\nu}$ for some $0 < \nu < 0.5$, such that the mesh of the partition defined as $\max_{1 \leq t \leq T_n+1} |\xi_t - \xi_{t-1}| = \mathcal{O}(n^{-\nu})$. Let $\mathcal{S}_n$ be the space of all polynomial splines of degree d on $[a, b]$ with the above partition. Then there exists a normalized B-spline basis $\{\phi_\ell : 1 \leq \ell \leq L_n\}$, where $L_n = T_n + d$ (See *Stone* (1985) or *Schumaker* (1981) for a detailed analysis) such that any $f \in \mathcal{S}_n$ can be written as

$$f(x) = \sum_{\ell=1}^{L_n} \beta_\ell \phi_\ell(x), \qquad x \in [a, b]. \tag{4.3}$$

Under suitable smoothness assumptions on the component functions f_{kj} in (4.1) can be well approximated by functions in $\mathcal{S}_n$ and accordingly the rank and variable selection procedures developed in this paper will be based on that. Let us define the approximating functions as

$$\tilde{f}_{kj}(x) = \sum_{\ell=1}^{L_n} \beta_{kj\ell} \phi_\ell(x), \quad \forall j = 1, \ldots, p; k = 1, \ldots, q. \tag{4.4}$$

Note that we have the extra constraint of $\mathbb{E}(f_{kj}(X_j)) = 0$ to make the model identifiable but the B-spline basis functions need not be mean zero. We adjust for that by centering the basis functions over the sample points. Define

$$\bar{\phi}_{\ell j} = \frac{1}{n}\sum_{i=1}^{n}\phi_\ell(x_{ij}), \quad \psi_{\ell j}(x) = \phi_\ell(x) - \bar{\phi}_{\ell j}. \tag{4.5}$$

The centered functions $\psi_{\ell j}$ act as the new basis functions which ensures identifiability. To keep notations simple and compact let us introduce the basis expanded design and coefficient matrices

$$\underset{n\times pL_n}{\mathbf{\Psi}} = \left[\underset{n\times L_n}{\mathbf{\Psi}_1} \mid \underset{n\times L_n}{\mathbf{\Psi}_2} \mid \cdots \mid \underset{n\times L_n}{\mathbf{\Psi}_p}\right], \quad \underset{pL_n\times q}{\mathbf{B}} = \left[\begin{array}{c} \underset{L_n\times q}{\mathbf{B}_1} \\ \hline \vdots \\ \hline \underset{L_n\times q}{\mathbf{B}_p} \end{array}\right] \tag{4.6}$$

where $\Psi_j(i,\ell) = \psi_{\ell j}(x_{ij})$ and $b_j(j,\ell) = \beta_{kj\ell}$. Note that under assumption **A2**, we have $\mathbf{B}_j \equiv 0$, $\forall j > J$. Using this matrix notation we can reformulate (4.2) as

$$\mathbf{Y} = \mathbf{\Psi}\mathbf{B} + \mathcal{E}(\mathbf{X}) + \mathbf{E}, \tag{4.7}$$

where $\mathcal{E}(\mathbf{X})$ denotes the approximation error in approximating the signal matrix $\mathbf{f}(\mathbf{X})$ by the spline basis expansion. This expansion plays a crucial role in the development of a simultaneous rank and variable selection algorithm for the non-parametric case.

4.3 Method: Non-parametric Low-rank Group Lasso

In this section we propose a two-step rank and variable selection method for the non-parametric additive model using expansion (4.7). The two steps are as follows.

Step I. Rank selection

$$\hat{r} = rank\left(\underset{\mathbf{W}}{\arg\min}\left\{\|\mathbf{Y} - \mathbf{\Psi}\mathbf{W}\|_F^2 + \mu.\text{rank}(\mathbf{W})\right\}\right). \tag{4.8}$$

Step II. Variable selection

$$\widehat{\mathbf{W}} = \underset{\mathbf{W}:\text{rank}(\mathbf{W})\leq\hat{r}}{\arg\min}\left\{\frac{1}{2}\|\mathbf{Y} - \mathbf{\Psi}\mathbf{W}\|_F^2 + \lambda\|\mathbf{W}\|_{F,1}\right\}. \tag{4.9}$$

Where $\|.\|_F$ denotes the Frobenius norm of a matrix and $\|\mathbf{W}\|_{F,1} = \sum_{k=1}^{p}\|\mathbf{W}_k\|_F$

which is a group lasso type penalty on the block of the coefficient matrix corresponding to each predictor. The rank penalty in the first step encourages dimension reduction while the group lasso penalty in the second step ensures that we select a sparse model.

The first step admits a closed form solution based on the singular value decomposition of the generalized least squares solution. Let $d_r(\mathbf{A})$ denote the r-th largest singular values of a generic matrix $\mathbf{A}$. Then the solution to step I is given by the following

$$\hat{r} = \max\left\{r : d_r(\mathbf{\Psi}(\mathbf{\Psi}^\top\mathbf{\Psi})^-\mathbf{\Psi}^\top\mathbf{Y}) > \sqrt{\mu}\right\}, \tag{4.10}$$

here $\mathbf{A}^-$ denotes the Moore-Penrose generalized inverse for a square matrix $\mathbf{A}$. See *Bunea et al.* (2011) for the detailed derivation. For an appropriate choice of tuning parameter μ we can guarantee consistent rank selection as long as the signal is sufficiently strong. In practice the optimal rank is often selected via cross-validation.

The optimization problem for the second step is non-convex due to the rank constraint therefore solving it directly might be challenging. Observe that a rank $\hat{r}$ matrix $\mathbf{W}$ can be decomposed as $\mathbf{W} = \mathbf{U}\mathbf{V}^\top$ where $\mathbf{U} \in \mathbb{R}^{pL_n\times\hat{r}}$ and $\mathbf{V}_{q\times\hat{r}}$ is an orthogonal matrix. Note that there is no constraint on $\mathbf{U}$. With this re-parametrization we can simplify the criterion function (4.9) as

$$\begin{aligned} \|\mathbf{Y} - \mathbf{\Psi}\mathbf{W}\|_F^2 + \lambda\|\mathbf{W}\|_{F,1} &= \|\mathbf{Y} - \mathbf{\Psi}\mathbf{U}\mathbf{V}^\top\|_F^2 + \lambda\|\mathbf{U}\|_{F,1} \\ &= \|\mathbf{Y}\mathbf{V} - \mathbf{\Psi}\mathbf{U}\|_F^2 + \lambda\|\mathbf{U}\|_{F,1}. \end{aligned} \tag{4.11}$$

Now we can solve this optimization problem in an iterative fashion. For a fixed $\mathbf{U}$, optimizing over $\mathbf{V}$ is a least squares problem that admits a closed form solution. On the other hand if we fix $\mathbf{V}$ then the optimization problem over $\mathbf{U}$ is a regular group lasso problem with the groups being given by the p-blocks of size $L_n \times \hat{r}$ in $\mathbf{U}^\top = [\mathbf{U}_1^\top, \mathbf{U}_2^\top, \ldots, \mathbf{U}_p^\top]$. This is a convex optimization problem and can be solved easily using off-the-shelf optimization algorithms. Therefore we can start with some initial guesses for $\mathbf{U}$ and $\mathbf{V}$ and iteratively solve the two sub-problems until the objective function does not change. It can be shown that the algorithm described above possesses monotonic convergence property given any starting point $(\mathbf{U}^0, \mathbf{V}^0)$. That is the cost function is guaranteed to decrease at each iteration.

We can infer the set of effective predictors and the rank of the system from the final solution $\widehat{\mathbf{W}}$. A predictor variable is non-informative for all of the response if the block of the coefficient matrix corresponding to the basis expansion is shrunk to zero. Whereas in the linear case that would be determined by a single row of the coefficient matrix. Given any new predictor vector x we expand it using the b-spline basis functions described in the previous section and use that to make the predictions.

4.4 Theoretical Properties

This section describes some of the desirable theoretical properties of our method that we were able to establish under moderate regularity conditions. We start by describing the smoothness conditions under which the component functions can be well approximated by the B-splines. Let α be a non-negative integer and $\gamma \in (0, 1]$, such that, $\alpha + \gamma > 0.5$. Define the class of functions f whose α-th order derivative exists and satisfies a Lipschitz condition of order γ. Specifically

$$\mathcal{F} = \left\{ f : [a,b] \rightarrow \mathbb{R} : |f^{(\alpha)}(s) - f^{(\alpha)}(t)| \leq |s-t|^{\gamma}, \text{for } s,t \in [a,b] \right\}. \tag{4.12}$$

C1. The component functions $f_{kj} \in \mathcal{F}$ and have $\mathbb{E}(f_{kj}(X_j)) = 0 \ \forall \ k, j$.
C2. X has a continuous density and there exists constants c_1 and c_2 such that the marginal densities g_j for X_j satisfy $0 < c_1 \leq g_j(x) \leq c_2 < \infty$ for $x \in [a, b]$ and $j = 1, 2, \ldots, p$.

These are rather mild assumptions that are common in many non-parametric additive model literature. The next lemma bounds the error $\mathcal{E}(\mathbf{X})$ in approximating the signal matrix $\mathbf{f}(\mathbf{X})$ using centered spline functions as described before.

Lemma IV.1. *Under **C1** and **C2**, $\exists \ \mathbf{B} \in \mathbb{R}^{pL_n \times q}$, such that, $\mathbf{f}(\mathbf{X}) = \mathbf{\Psi}\mathbf{B} + \mathcal{E}(\mathbf{X})$, with*

$$\|\mathcal{E}(\mathbf{X})\|_F^2 \leq C_1 |J| q L_n, \tag{4.13}$$

where C_1 is a constant of proportionality.

The proof adapts *Lemma 1* in *Huang et al.* (2010) to the case of multiple functional responses. This allows us to bound the largest singular value of the approximation error which is a crucial step towards proving our first result on rank selection consistency.

Theorem IV.2. *Assume **C1** and **C2** and let rank*$(\mathbf{f}(\mathbf{X})) = r$. *If* $d_r(\mathbf{f}(\mathbf{X})) > 4\sigma(\sqrt{q}+\sqrt{pL_n}) + 7C_1\sqrt{|J|qL_n}$, *then for an appropriate choice of* μ *we have*

$$\mathbb{P}(\hat{r} \neq r_0) \leq exp(-C_2 pqL_n). \tag{4.14}$$

Note that $\mathbf{X}$ is deterministic and the probability is defined over the distribution of the error $\mathbf{E}$. If $pqL_n \to \infty$ then probability of selecting the wrong rank goes to zero at an exponential rate. The details of the proof are available in the appendix. The non-parametric case is more challenging due to the presence of the approximation error term. Therefore to achieve rank selection consistency we need to have a stronger signal as measured by $d_r(\mathbf{f}(\mathbf{X}))$, because it needs to dominate the sum of the additive error in the model $\mathbf{E}$ and the approximation error $\mathcal{E}(\mathbf{X})$. This results in the extra term given by $7C_1\sqrt{|J|qL_n}$ in the inequality condition for $d_{r_0}(\mathbf{f}(\mathbf{X}))$ when compared to the linear case. We introduce two more moderate regularity conditions which will be crucial to prove next result on error bound.

C3. Let $\mathbf{\Sigma} = \mathbf{\Psi}^\top \mathbf{\Psi}/n$. With a slight abuse of notation define $J = \{1, 2, \ldots, J\}$, the set of effective predictors and $\delta_J > 0$ such that, $\mathbf{\Sigma}$ satisfies

$$trace(\mathbf{A}^\top \mathbf{\Sigma} \mathbf{A}) \geq \delta_J \sum_{j \in J} \|\mathbf{A}_j\|_F^2, \tag{4.15}$$

for any $\mathbf{A} \in \mathbb{R}^{pL_n \times q}$ with blocks $\mathbf{A}_j$ that satisfy $2\sum_{j\in J}\|\mathbf{A}_j\|_F \geq \sum_{j \in J^c}\|\mathbf{A}_j\|_F$.

C4. $\log(\|\mathbf{f}(\mathbf{X})\|_F^2) \leq C_3|J|qL_n$, for some numerical constant C_3.

C3 is a technical condition that can be seen as a version of the restricted eigenvalue property for the multiple response regression problem. See *Bickel et al.* (2008) or *Bühlmann and van de Geer* (2011) for a more detailed discussion on this. **C4** is required to ensure that the estimation error remains bounded even when the rank is estimated wrongly in the first step. Now we present our main result on bounding the estimation error.

Theorem IV.3. *Assume that **C1**, **C2**, **C3** and **C4** hold then under appropriate choice of the tuning parameters* μ *and* λ *the two-step non-parametric low-rank group lasso estimator defined in* (4.9) *satisfies*

$$\mathbb{E}\left(\|\mathbf{f}(\mathbf{X}) - \mathbf{\Psi}\widehat{\mathbf{W}}\|_F^2\right) \lesssim rq + |J|qL_n + r|J|L_n \log(ep). \tag{4.16}$$

The symbol $\lesssim$ stands for proportionality up to certain multiplicative constants. Once again we remind the readers that $\mathbf{X}$ and $\mathbf{\Psi}$ are treated as deterministic quantities and the expectation is taken of the distribution of the errors which is Gaussian by assumption. We also note that both theorems IV.2 and IV.3 hold without any restriction on n, p, q as long as the fairly moderate regularity conditions on component functions and $\mathbf{\Sigma}$ continue to hold. A comparison with the result obtained in Theorem 4 by *Bunea et al.* (2012) shows that we have incurred one extra term ($|J|qL_n$) which arises from the error in approximating the non-parametric functions by the B-spline basis. Also the third term of the right hand side of (4.16) has an extra factor of L_n which can be attributed to the basis expansion leading to groups of size $\hat{r}L_n$. In the linear case we have groups of size $\hat{r}$ in the second step which explains the difference.

Details of the proofs are provided in the appendix. The key differences from the proof of Theorem 4 in *Bunea et al.* (2012) is the handling of the approximation error in the non-parametric case. We control the approximation error using the bound provided in Lemma IV.1 and an application of Mirsky's theorem (*Mirsky*, 1960) that bounds the sum of squared differences between the ordered singular values for an additive perturbation. Rigorous details are provided in the appendix.

4.5 Numerical Studies

This section uses simulated data to evaluate the performance of the joint non-parametric rank and variable selection method. We compare the proposed joint non-parametric rank and variable selection method against separate rank and variable selection methods. By separate method we mean applying the rank selection/variable selection algorithm on the basis expanded design matrix to make it a fair comparisons as the data are generated from non-parametric models.

The component functions are chosen to be, $f_1(x_1) = 2x_1$, $f_2(x_2) = 2(x_2^2 - 3)$, $f_3(x_3) = 0.5x_3^3$, $f_4(x_4) = \sin(\pi x_4/3)$, $f_5(x_5) = e^{x_5} - (e^3 - e^{-3})/6$ and $f_6(x_6) = f_7(x_7) = \cdots = f_p(x_p) = 0$, which implies that the set of informative predictors is given by $J = \{1, 2, 3, 4, 5\}$. Note that the constant shifts are incorporated to make the component functions have zero mean. Let $\mathbf{f}(\mathbf{X}) = (f_j(x_{ij}))$ denote the functional component matrix where $j = 1, 2, \ldots, p$ and $i = 1, 2, \ldots, n$. Given a rank $r \leq \min\{n, p, q\}$ we generate the response matrix as follows

$$\mathbf{Y} = \mathbf{f}(\mathbf{X})\mathbf{B} + \mathbf{E}, \tag{4.17}$$

where $\mathbf{B}$, the coefficient matrix is a $p \times q$ matrix of rank r and $\mathbf{E}$ is the error matrix, such that, $e_{ij} \overset{iid}{\sim} N(0,1)$. To generate the predictors we start by simulating a random matrix $\mathbf{A}_{n\times(p+1)}$ where $a_{ij} \overset{iid}{\sim} U(-3,3)$. Then we set $x_{ij} = (a_{ij} + \rho a_{i(p+1)})/(1+\rho)$ for $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots p$. This construction allows us to introduce correlations among the prediction variables in an efficient manner as $corr(\mathbf{X}_{ij}, \mathbf{X}_{ij'}) = \rho^2/(1+\rho^2)$. In our simulations we consider two particular cases $\rho = 0, \sqrt{1/3}$. $\rho = 0$ correspond to the independent case while $\rho = \sqrt{1/3}$ introduces a mild correlation of 0.25 between each pair of predictors. We also consider two sets of choices for the models dimensions

Low-dimensional Setting(LD): $n = 100$, $p = 25$, $q = 10$, $r = 2$
High-dimensional Setting(HD): $n = 100$, $p = 100$, $q = 20$, $r = 3$

leading to a total of four possible combinations. At each setting we generate a training and validation sets of same sample size. In addition a test data set of size $n_{test} = 10n = 1000$ is generated to assess the prediction performance. A crucial practical issue that we have not discussed yet is the selection of optimal tuning parameters. For Step I we used the prediction error in the validation set to select the rank. For Step II, we considered two different approaches to select the penalty parameter λ. First one was the validation method similar to one described for step I. Although this approach is simple to implement but it often tends to select too many variables (see *Shao* (1993) or *Arlot and Celisse* (2010) for a more comprehensive account). To avoid that we also considered a Bayesian Information Creterion (BIC) (*Schwarz*, 1978) for the group lasso algorithm given by

$$BIC(\lambda) = \log\left(\|\mathbf{Y} - \mathbf{\Psi}\widehat{\mathbf{W}}(\lambda)\|_F^2\right) + df(\lambda).\frac{\log(n\hat{r})}{n\hat{r}}, \tag{4.18}$$

where $\hat{r}$ is the selected rank in Step I, and $df(\lambda)$ denotes the degrees of freedom for the group lasso problem in Step II. The degrees of freedom of the group lasso is quite a non-trivial problem. We used the unbiased estimator of the degrees of freedom of the group lasso problem for a general design matrix proposed by *Vaiter et al.* (2012). We report the results of both the model selection approaches over 60 replication at each setting of the study. Results are summarized using the following measure of accuracy

1. Prediction error(PE) given by $\|\mathbf{Y}_{test} - \mathbf{\Psi}_{test}\widehat{\mathbf{W}}_{train}\|_F^2/(qn_{test})$

2. Average number of selected variables $|\hat{J}|$.

3. Average estimated rank.

4. Proportion of replications where $\hat{J} \supseteq J$ called inclusion (IN).

5. Proportion of replications where $\hat{J} = J$ called correct selection (CS).

Table 4.1: Simulation results comparing three non-parametric multiple response regression methods, with joint rank and variable selection (NP:RR+VS), with variable selection (NP:VS) and with joint rank selection (NP:RR)

			NP:RR+VS					NP:VS				NP:RR	
	ρ		PE	$\|\hat{J}\|$	Rank	IN	CS	PE	$\|\hat{J}\|$	IN	CS	PE	Rank
Low-Dim p=25, q=10, J=5, r=2	0	BIC	1.49	5.16	2.00	1.00	0.84	1.69	11.12	1.00	0.03	3.15	2.00
			(0.2)	(0.4)	(0.0)	(0.0)	(0.4)	(0.2)	(2.8)	(0.0)	(0.2)	(0.4)	(0.0)
		Valid	1.27	8.20	2.00	1.00	0.00	1.63	13.75	1.00	0.00		
			(0.1)	(1.4)	(0.0)	(0.0)	(0.0)	(0.2)	(1.9)	(0.0)	(0.0)		
	$\frac{1}{4}$	BIC	2.52	5.22	2.00	0.95	0.71	2.72	6.43	0.97	0.38	3.26	2.00
			(0.3)	(0.6)	(0.0)	(0.2)	(0.5)	(0.3)	(1.5)	(0.2)	(0.5)	(0.3)	(0.0)
		Valid	2.14	9.11	2.00	1.00	0.00	2.47	11.40	1.00	0.00		
			(0.1)	(1.7)	(0.0)	(0.0)	(0.0)	(0.2)	(2.2)	(0.0)	(0.0)		
High- Dim p=100, q=20, J=5, r=3	0	BIC	1.52	5.03	3.00	1.00	0.97	1.62	24.06	1.00	0.02	7.27	3.00
			(0.2)	(0.2)	(0.0)	(0.0)	(0.2)	(0.2)	(6.0)	(0.0)	(0.12)	(0.3)	(0.0)
		Valid	1.21	35.95	3.00	1.00	0.00	1.61	27.06	1.00	0.00		
			(0.1)	(5.31)	(0.0)	(0.0)	(0.0)	(0.1)	(3.61)	(0.0)	(0.0)		
	$\frac{1}{4}$	BIC	2.01	5.79	2.80	0.75	0.19	2.11	10.50	0.97	0.02	3.51	2.79
			(0.2)	(1.1)	(0.4)	(0.4)	(0.4)	(0.2)	(3.3)	(0.2)	(0.1)	(0.1)	(0.4)
		Valid	1.60	38.50	2.80	0.98	0.00	1.94	23.19	1.00	0.00		
			(0.1)	(5.4)	(0.4)	(0.1)	(0.0)	(0.2)	(4.4)	(0.0)	(0.0)		

Table 4.1 presents the results for all four combinations of settings and the numbers in the parenthesis represent the standard errors. As we can see the joint non-parametric method does better both in terms prediction error as well as variable selection for all combination of settings. The variable selection approach comes out as a close second in terms of prediction error followed by just rank selection which is a distant third. BIC approach fares well in terms of model selection performance for the joint estimation method(NP:RR+VS) especially when the predictors are uncorrelated as reflected by the inclusion(IN) and correct selection(CS) proportions. The validation approach tends to have lower prediction error at the cost of selecting too many unimportant variables. The non-parametric variable selection method (NP:VS) tends to select too many variables even with BIC approach, which reflects the importance of the rank constraint in identifying the true model. Also it is computationally much slower due to the large dimensions after basis expansion. For example, under the high-dimensional setting in the non-parametric variable selection(NP:VS) approach

we are fitting a group lasso problem with $100 \times 7 \times 20 = 14,000$ variables divided into 100 groups of size 140 each on data set with sample size 2000. In our experience the joint approach is often $3 - 5$ times faster in terms of run time due to the dimension reduction achieved by rank selection in Step I.

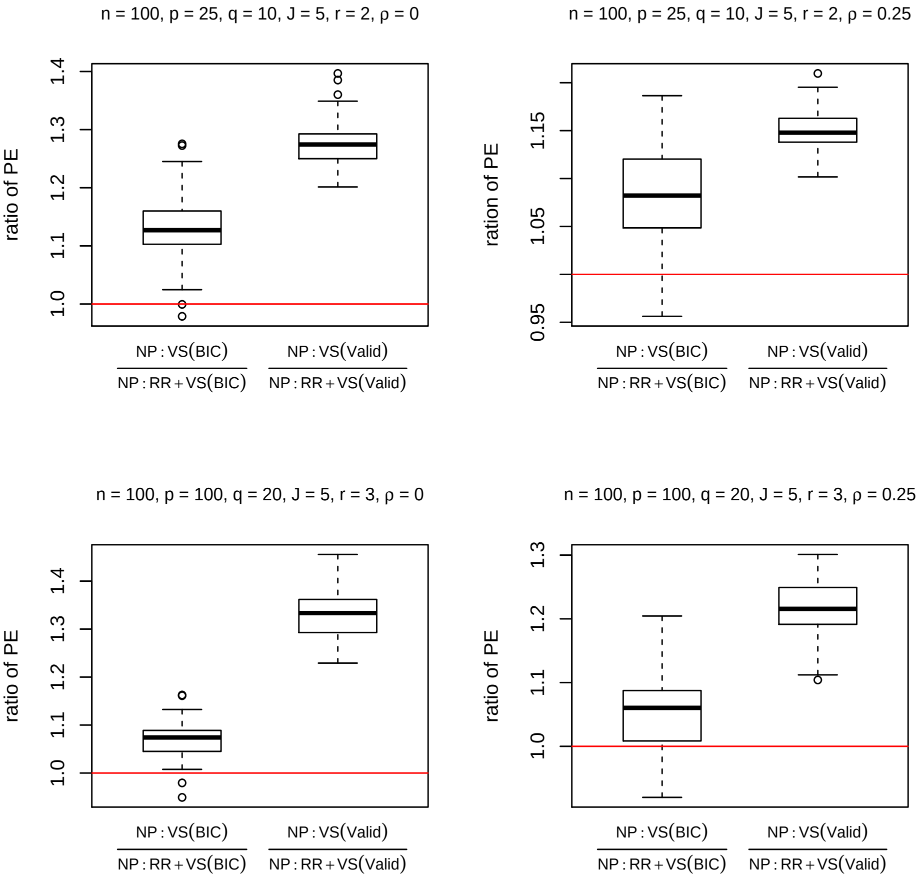

Figure 4.1: Boxplot of ratio of test MSE between NP:VS and NP:RR+VS for both BIC and Validation approach for tuning parameter selection. Top row presents the low dimensional setting while the bottom row presents the high-dimensional setting

Figure 4.1 facilitates pairwise comparison by plotting the ratio of the prediction error (PE) between the pure variable selection method(NP:VS) and joint rank and variable selection approach (NP:RR+VS). We find that the ratio is mostly greater than 1 at every setting which provides a stronger indication of superior prediction performance of the joint method compared to only variable selection.

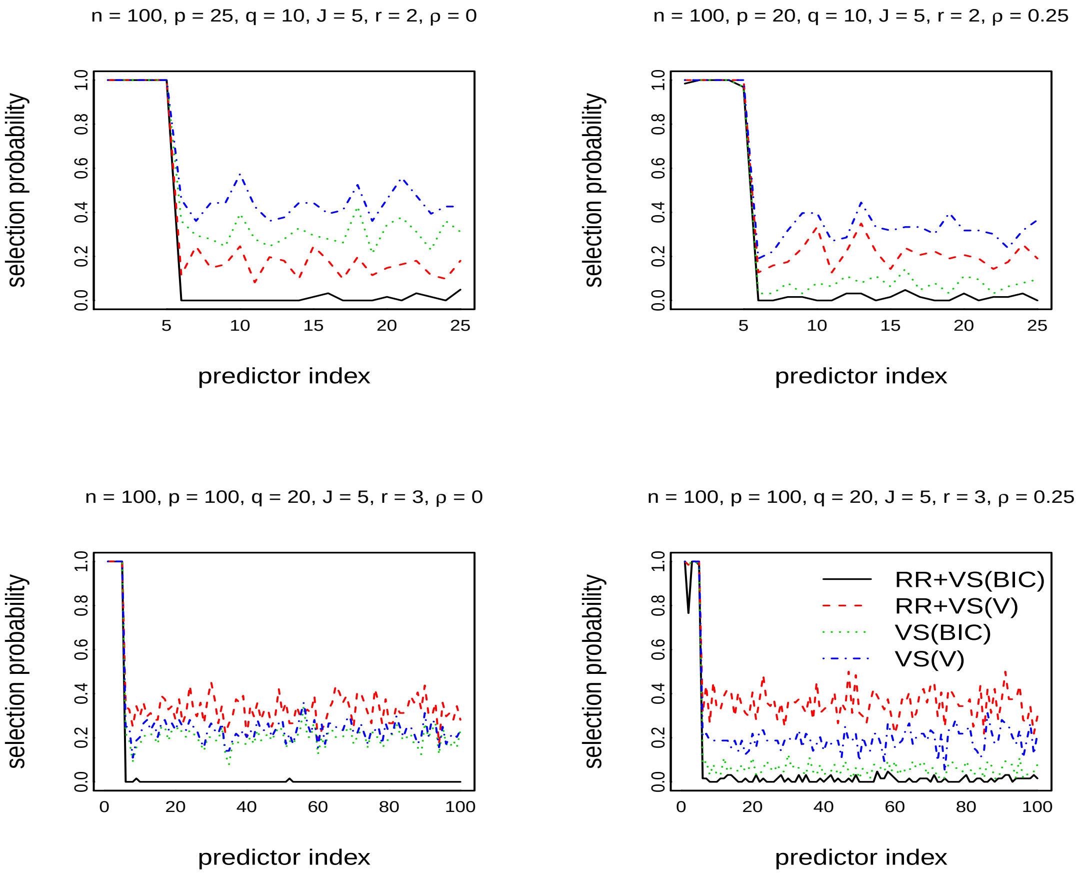

Figure 4.2: For each predictor variable we plot the selection proportion out of the 60 replications on the Y-axis. The top row presents the low dimensional setting while the bottom row presents the high dimensional setting

Figure 4.2 plots the selection proportion for each predictor variable over 60 replications. The four methods are overlaid as a line plot for each setting. Ideally we would like to see high proportion for the first five variables which are the truly informative components and a low selection proportion for the rest of the predictors. The black line corresponding to NP:RR+VS with tuning parameter in Step II selected by BIC is the clear winner at all settings. All the methods select the truly informative variables consistently but for the uninformative variables we find that the other methods tend to pick too many variables. This is consistent with the critique of cross-validation approach in terms of variable selection. The results clearly demonstrate that if the underlying model satisfies the low-rank and sparsity assumptions and is truly non-linear then the proposed method outperforms its closest competitors both in terms of prediction accuracy as well as model selection.

4.6 Application in a Chemometrics study

We apply the methodology proposed in section 3 to a chemometrics data set first published by *Skagerberg et al.* (1992). The data was collected on low-density polyethylene production in tubular reactors under high pressure and temperature. It consists of 6 responses on the properties of the polymer namely, number average molecular weight (M_n), weight average molecular weight (M_w), frequency of long chain branching (LCB), frequency of short chain branching (SCB), content of vinyl groups (VNL) and the content of vinylidene groups (VND) in the polymer chain. These responses determine the quality of the final polymer products such as, tensile and impact strength. Unfortunately, quantities like molecular weight, branching frequencies or double bonds are difficult to measure requiring highly expensive, sophisticated and time consuming techniques, such as nuclear magnetic resonance spectroscopy(NMR). The goal is to find a good predictive model for the responses using routinely collected process variables such as temperature profile along the reactor measured at 20 points along the reactor ($T_1, T_2, \ldots, T_{20}$), solvent flow rate (S) and wall temperature (T_w). Once the predictive model is built it can be used to infer the polymer properties or for the purpose of quality control with investing in expensive and complicated tests. A schematic diagram is presented in Figure 4.3. The left panel shows a section of tubular LDPE reactor used for the data collection. The right panel presents a typical temperature profile within the reactor, that is, $T_1 - T_{20}$. The reaction is highly exothermic reaching a maximum temperature somewhere around the middle of the tube when all the initial solvent are consumed. After that the mixture cools down in a rather linear fashion.

Reduced rank models are relevant for this problem due to the fact that several of the responses are very highly correlated which indicates a possibility of dimension reduction. Existing works that analyzed this data set assumed linear model (*Breiman and Friedman*, 1997; *Mukherjee and Zhu*, 2011). But a simple scatter plot of the responses against the predictors reveal that the relationships are highly non-linear. Therefore a non-parametric approach seems more reasonable. Another interesting aspect of this data set is that the reactor temperatures for points that are close to

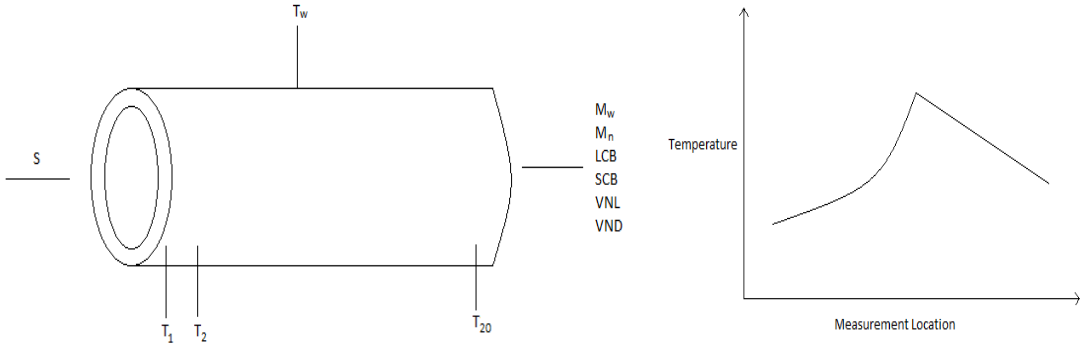

Figure 4.3: Left Panel: Schematic diagram of a section of tubular reactor for the production of LDPE. S denotes the solvent feed rate. $T_1 - T_{20}$ measure the temperature along the reactor at equidistant points and T_w denotes the wall temperature. M_w, M_n, LCB, SCB, VNL and VND denotes the 6 responses. Right Panel: A typical temperature profile in a LDPE reactor (*Skagerberg et al.*, 1992).

each other (e.g. T_i with T_{i-1} and T_{i+1}) might not carry much unique information regarding the responses. This leads us to believe that the reactor temperature at a few key locations would suffice for a good predictive model implying a possibility of variable selection.

The final data set consists of 56 points on 22 predictor variables and 6 responses. The responses are log-transformed and scaled make them comparable to each other. We also scale the predictors to the interval $[-1, 1]$ to introduce the B-spline basis expansion. 9 basis functions were used for the non-parametric method, with knots being placed at $(1/7, 2/7, \ldots, 6/7)$-th quantiles of the empirical distribution for each column of $\mathbf{X}$. We randomly split the data into a training set consisting of 45 observations and a test set of 11 observations. The process is replicated 100 times. The methods that are compared include multivariate least squares (OLS), linear reduced rank regression (LN:RR), linear reduced rank regression with variable selection (LN:RR+VS), non-parametric reduced rank regression (NP:RR), non-parametric joint rank and variable selection (NP:RR+VS) which is the method that we are proposing and non-parametric variable selection for multivariate regression (NP:VS). Note that for the variable selection methods we choose the optimal tuning parameter via BIC defined in (4.18) as it proved to be more competent in variable selection in our simulation studies.

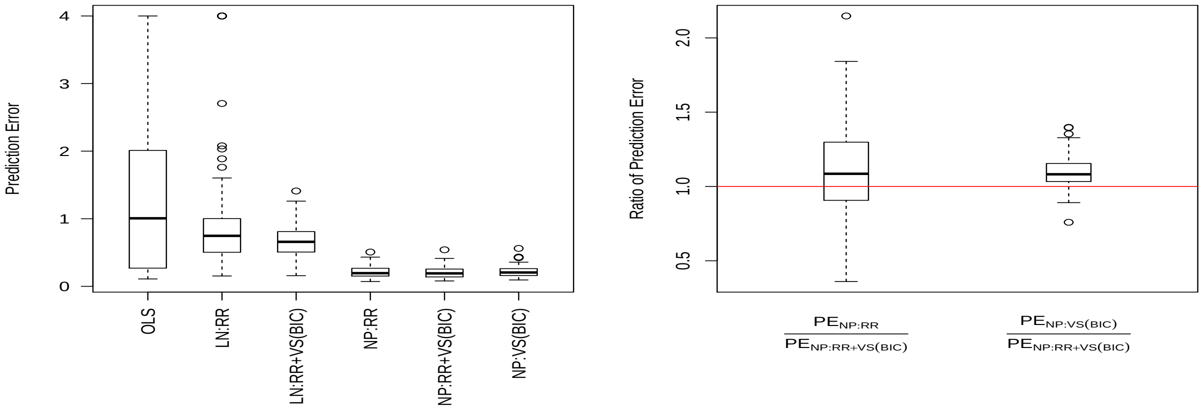

Figure 4.4: Left panel: box plot comparing the test MSE of the competing methods. Right panel: pairwise comparison between the non-parametric methods using ratio of test MSE.

We plot the box plots of the prediction error over 100 replication for each method on the left panel of Figure 4.4. Clearly the non-parametric methods outperform the linear methods in terms of prediction error. This can be explained by non-linearity of the underlying model and poor rank selection performance for the linear methods. We find that the linear approach selects a rank of 1 in more than 90% of the replications which is caused by the extreme collinearity of the design matrix, also noted by *Mukherjee and Zhu* (2011). On the other hand the non-parametric method selects an optimal rank of 3 in all the replication. Scientifically it is well known that the first two responses M_n and M_w are approximately dependent. Similarly the last three responses form a functional group in the sense that SCB is an approximate measure of the contents of VNL and VND. The third response LCB has a moderately negative correlation with the other groups. This leads to a effective dimension of 3 which is consistent with our findings. The Y-axis of the left panel is inflated due to the large variability in prediction error for OLS making it hard for us to compare between the three non-parametric methods. To address that we do a box plot of pairwise ratio of prediction error for (NP:RR)/(NP:RR+VS) and (NP:VS)/(NP:RR+VS) on the right panel. More than 50% points for the first ratio and more than 75% points for the second ratio are above 1 indicating that on average the proposed joint method does better than the individual rank and variable selection methods for the non-parametric approach.

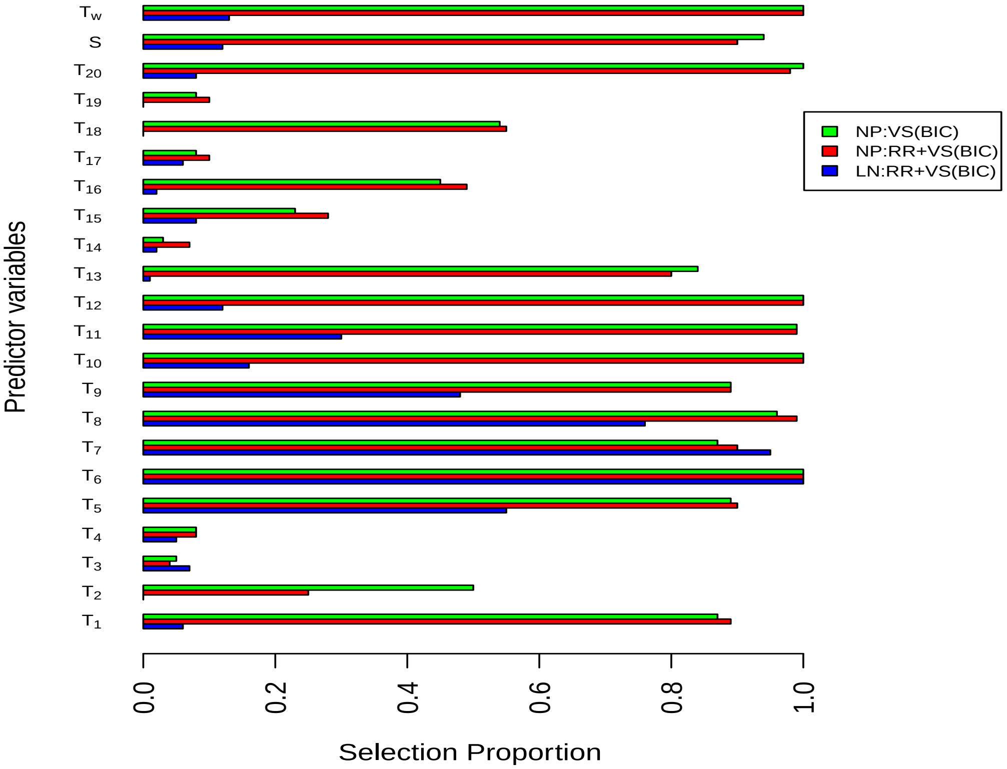

Figure 4.5: Frequency of selection out of 100 replication for each predictor variables, the competing methods are $LN : RR + VS$ (blue), $NP : RR + VS$ (red) and $NP : VS$ (green).

Next we investigate the variable selection performance for the methods with sparsity penalty which are (LN:RR+VS), (NP:VS) and (NP:RR+VS). A grouped bar plot for the selection proportion of each variable out of 100 replications is presented in Figure 4.5. On an average the non-parametric joint rank and variable selection method (NP:RR+VS) selects about 14 out of 22 predictors. A closer look reveals some interesting patterns. Among the reactor temperature profile, the start and the end point, that is, T_1 and T_{20} and the middle portion $T_5 - T_{13}$ gets selected with high proportion as does the two other predictors solvent flow rate (S) and wall temperature (T_w). This is consistent with a typical temperature profile in an LDPE reactor in Figure 4.3. The middle portion of the profile where sharp changes occur corresponds to $T_5 - T_{13}$. The start and the end phases are quite linear therefore might not carry much extra information once T_1 and T_{20} are selected. The linear rank and variable selection method tends to select too few components (5 on average) which contributes to its poor performance. The non-parametric variable selection method possesses a very similar selection pattern to that of the joint non-parametric method and has a competitive prediction error.

4.7 Summary

In this paper we have proposed a two-step joint rank and variable selection method for non-parametric additive modeling of multi-response regression. The method is based on B-spline approximation of the component function under moderate smoothness assumptions. In the first step we select the rank by running a reduced rank regression of the response matrix on the basis expanded design matrix. The second step introduces a group lasso penalty on the coefficient matrix to encourage variable selection. Error bounds and rank selection consistency are established under conditions on signal strength, Gaussian error and regularity of the basis expanded covariance matrix. Simulation studies and an application on low-density polyethylene data set illustrate the superior predictive and model selection performance of the proposed method over relevant competitors.

APPENDICES

APPENDIX A

Representer Theorem for Reduced Rank Estimators in RKHS

Proof of Proposition II.1

Proof. Let $(f_1, f_2, \ldots, f_q)$ be the minimizer to (2.13). Define,

$$\mathcal{F}_K = \text{span}\{K(\cdot, x_i) : i = 1, 2, \ldots n\} \tag{A.1}$$

We can decompose each $f_j = f_j^* + f_j^0$ where f_j^* is the projection of f_j onto $\mathcal{F}_K$ and f_j^0 is the orthogonal to $\mathcal{F}_K$. Then for $j = 1, 2, \ldots q$ and $i = 1, 2, \ldots, n$,

$$\begin{aligned} f_j(x_i) &= \langle f_j^* + f_j^0, K(\cdot, x_i)\rangle = f_j^*(x_i) \\ \|f_j\|_{\mathcal{H}}^2 &= \|f_j^*\|_{\mathcal{H}}^2 + \|f_j^0\|_{\mathcal{H}}^2 \end{aligned}$$

Clearly, $J_\lambda\left(f_1^*, f_2^*, \ldots, f_q^*\right) \leq J_\lambda\left(f_1, f_2, \ldots f_q\right)$ and $\dim\left(\text{span}\{f_1^*, f_2^*, \ldots, f_q^*\}\right) \leq r$ also holds since they are just projection of $(f_1, f_2, \ldots, f_q\})$ to $\mathcal{F}_K$, where $\dim\left(\text{span}\{f_1, \ldots, f_q\}\right) \leq r$ since they are a solution to (16). Thus the solution to (16) can be expressed as,

$$f_j(x) = \sum_{i=1}^{n} \alpha_{iq} K(x, x_i), \quad \text{for } j = 1, 2, \ldots q, \quad \alpha_{ij} \in \mathbb{R} \tag{A.2}$$

□

Proof of Proposition II.2

Proof. If $r = q$ then the result holds vacuously. If $r < q$ then $\exists$ non-trivial linear combinations $\sum_{j=1}^{q} c_j f_j(\cdot) \equiv 0$. Equivalently, we have, $\|\sum_{j=1}^{q} c_j f_j(\cdot)\|_{\mathcal{H}}^2 = 0$.

$$\|\sum_{j=1}^{q} c_j f_j(\cdot)\|_{\mathcal{H}}^2 = 0 \Leftrightarrow c_{q\times 1}^T \mathbf{A}^T \left[(K(x_i, x_{i'})\right]_{i,i'=1}^{n} \mathbf{A} c_{q\times 1} = 0$$

Under the strict positive definiteness assumption on $K(\cdot,\cdot)$ this can only happen if $\mathbf{A}c = 0_{q\times 1} \Leftrightarrow c \in \text{Ker}(\mathbf{A})$, where $\text{Ker}(\mathbf{T})$ for any matrix/linear operator $\mathbf{T}$ denotes its null space. Let us define a map, $T : \mathbf{R}^q \mapsto V = \text{span}\{f_1, f_2, \ldots f_q\}$, where, $T(c) = \sum_{j=1}^{q} c_j f_j(\cdot)$ Then using the Rank-Nullity Theorem and the previous part,

$$\begin{aligned} \dim\left(\text{Ker}(T)\right) + \dim\left(\text{Img}(T)\right) &= q \\ \Rightarrow \quad \dim\left(\text{Ker}(\mathbf{A})\right) + \dim\left(V\right) &= q \\ \Rightarrow \quad \text{rank}(\mathbf{A}) = dim(V) &\leq r \end{aligned}$$

□

APPENDIX B

Derivation of the exact Degrees of Freedom Estimator

Derivation of Equation (3.10)

Note that

$$\begin{aligned}
\widehat{\mathbf{Y}} &= \mathbf{XQS}^{-1}\mathbf{H}, \\
\Longrightarrow \widehat{\mathbf{Y}}(r) &= \mathbf{XQS}^{-1}\mathbf{H}(r), \quad r = 1, \ldots, \bar{r},
\end{aligned}$$

Using the trace identity, $tr(\mathbf{AB}) = tr(\mathbf{BA})$, the equality, $vec(\mathbf{ABC}) = (\mathbf{C}^T \otimes \mathbf{A})vec(\mathbf{B})$ and the chain rule of differentiation we get

$$\begin{aligned}
\widehat{df}(r) &= tr\left\{\frac{\partial vec(\widehat{\mathbf{Y}}(r))}{\partial vec(\mathbf{Y})}\right\} \\
&= tr\left\{\left[\mathbf{I}_q \otimes \mathbf{XQS}^{-1}\right]\left(\frac{\partial vec(\mathbf{H}(r))}{\partial vec(\mathbf{Y})}\right)\right\} \\
&= tr\left\{\left[\mathbf{I}_q \otimes \mathbf{XQS}^{-1}\right]\left(\frac{\partial vec(\mathbf{H}(r))}{\partial vec(\mathbf{H})}\right)\left(\frac{\partial vec(\mathbf{H})}{\partial vec(\mathbf{Y})}\right)\right\} \\
&= tr\left\{\left[\mathbf{I}_q \otimes \mathbf{XQS}^{-1}\right]\left(\frac{\partial vec(\mathbf{H}(r))}{\partial vec(\mathbf{Y})}\right)\left[\mathbf{I}_q \otimes \mathbf{S}^{-1}\mathbf{Q}^\top\mathbf{X}\right]\right\} \\
&= tr\left\{\frac{\partial vec(\mathbf{H}(r))}{\partial vec(\mathbf{Y})}\right\}.
\end{aligned}$$

Proof of Theorem III.1

We acknowledge that the proof of Theorem III.1 is mainly based on the results developed in *de Leeuw* (2007) about the derivatives of a generalized eigensystem. Note that we have assumed $r_x \geq q$, and the same results can be presented for $\mathbf{H}^\top$ when $r_x \leq q$.

Denote $\mathbf{A} = \mathbf{H}^\top\mathbf{H}$, and let $(d^2, \mathbf{v})$ denote a pair of eigenvalue and eigenvector of $\mathbf{A}$. Suppose $\mathbf{A}$ is two times continuously differentiable at θ, e.g., $\theta = h_{ij}$ for any $i = 1, ..., r_x$ and $j = 1, ..., q$. Then the eigenvalues and eigenvectors are also differentiable at θ. From

$$\mathbf{A}\mathbf{v} = d^2\mathbf{v},$$

it follows that

$$\frac{\partial \mathbf{A}}{\partial \theta}\mathbf{v} + \mathbf{A}\frac{\partial \mathbf{v}}{\partial \theta} = d^2\frac{\partial \mathbf{v}}{\partial \theta} + \frac{\partial d^2}{\partial \theta}\mathbf{v},$$

and this gives

$$(\mathbf{A} - d^2\mathbf{I})\frac{\partial \mathbf{v}}{\partial \theta} = -(\frac{\partial \mathbf{A}}{\partial \theta} - \frac{\partial d^2}{\partial \theta}\mathbf{I})\mathbf{v}. \tag{B.1}$$

Premultiplying both sides by $\mathbf{v}^\top$ gives

$$\mathbf{v}^\top(\mathbf{A} - d^2\mathbf{I})\frac{\partial \mathbf{v}}{\partial \theta} = -\mathbf{v}^\top\frac{\partial \mathbf{A}}{\partial \theta}\mathbf{v} + \frac{\partial d^2}{\partial \theta}.$$

It is obvious that the left-hand-side equals to 0, and it then follows that

$$\frac{\partial d}{\partial \theta} = \frac{1}{2d}\mathbf{v}^\top\frac{\partial \mathbf{A}}{\partial \theta}\mathbf{v}. \tag{B.2}$$

Define $(\mathbf{A} - d^2\mathbf{I})^- = \mathbf{V}(\mathbf{D}^2 - d^2\mathbf{I})^+\mathbf{V}^\top$ with $(\cdot)^+$ denoting the Moore-Penrose inverse. Therefore, $(\mathbf{A} - d^2\mathbf{I})^-(\mathbf{A} - d^2\mathbf{I}) = \mathbf{I} - \mathbf{v}\mathbf{v}^\top$ and $(\mathbf{A} - d^2\mathbf{I})^-\mathbf{v} = 0$. Premultiplying both sides of (B.1) by $(\mathbf{A} - d^2\mathbf{I})^-$ gives

$$(\mathbf{I} - \mathbf{v}\mathbf{v}^\top)\frac{\partial \mathbf{v}}{\partial \theta} = -(\mathbf{A} - d^2\mathbf{I})^-\frac{\partial \mathbf{A}}{\partial \theta}\mathbf{v}.$$

From $\mathbf{v}^\top\mathbf{v} = 1$, we know that $\mathbf{v}^\top(\partial \mathbf{v}/\partial \theta) = 0$. It then follows that

$$\frac{\partial \mathbf{v}}{\partial \theta} = -(\mathbf{A} - d^2\mathbf{I})^-\frac{\partial \mathbf{A}}{\partial \theta}\mathbf{v}. \tag{B.3}$$

Define $\mathbf{Z}^{(ij)} = \partial \mathbf{H}/\partial h_{ij}$ be an $r_x \times q$ matrix of zeros with only its (i, j)th entry equalling to one. For any $\theta = h_{ij}$,

$$\frac{\partial \mathbf{A}}{\partial h_{ij}} = \mathbf{H}^\top \mathbf{Z}^{(ij)} + \mathbf{Z}^{(ij)\top}\mathbf{H}. \tag{B.4}$$

The proof is completed by combining the results in (B.2), (B.3) and (B.4).

Proof of Theorem III.2

For simplicity and without loss of generality, we assume $r_x \geq q$. When $r_x \leq q$, one can repeat the same proof using $\mathbf{H}^\top$. When $r = q$, the result $\widehat{df}(q) = r_x q$ holds trivially. So in the following, we consider $r < q$. Consider $\partial \mathbf{H}^{(r)}/\partial h_{ij}$ for any $1 \leq i \leq r_x$, $1 \leq j \leq q$. Because $\mathbf{H}^{(r)} = \mathbf{H}\sum_{k=1}^{r} \mathbf{v}_k \mathbf{v}_k^\top$, by the chain rule, we have

$$\begin{aligned}\frac{\partial \mathbf{H}^{(r)}}{\partial h_{ij}} =& \frac{\partial \mathbf{H}}{\partial h_{ij}} \sum_{k=1}^{r} \mathbf{v}_k \mathbf{v}_k^\top + \mathbf{H}\sum_{k=1}^{r} \frac{\partial \mathbf{v}_k}{\partial h_{ij}} \mathbf{v}_k^\top + \mathbf{H}\sum_{k=1}^{r} \mathbf{v}_k \frac{\partial \mathbf{v}_k^\top}{\partial h_{ij}} \\ =& \mathbf{Z}^{(ij)}\mathbf{V}^{(r)}\mathbf{V}^{(r)\top} - \mathbf{H}\sum_{k=1}^{r}\left\{(\mathbf{H}^\top\mathbf{H} - d_k^2\mathbf{I})^-(\mathbf{H}^\top\mathbf{Z}^{(ij)} + \mathbf{Z}^{(ij)\top}\mathbf{H})\mathbf{v}_k\mathbf{v}_k^\top\right\} \\ &- \mathbf{H}\sum_{k=1}^{r}\left\{\mathbf{v}_k\mathbf{v}_k^\top(\mathbf{H}^\top\mathbf{Z}^{(ij)} + \mathbf{Z}^{(ij)\top}\mathbf{H})(\mathbf{H}^\top\mathbf{H} - d_k^2\mathbf{I})^-\right\}. \end{aligned} \tag{B.5}$$

Consider the first term on the right-hand-side of (B.5). Its (i,j)th entry equals to $\sum_{k=1}^{r} v_{jk}^2$. Therefore, its contribution to the degrees of freedom (3.10) is

$$\sum_{i=1}^{r_x}\sum_{j=1}^{q}\sum_{k=1}^{r} v_{jk}^2 = r_x r, \tag{B.6}$$

because $\sum_{j=1}^{q} v_{jk}^2 = 1$. We know

$$(\mathbf{H}^\top\mathbf{H} - d_k^2\mathbf{I})^- = \sum_{l\neq k}^{q} \frac{1}{d_l^2 - d_k^2}\mathbf{v}_l\mathbf{v}_l^\top.$$

We also have

$$\mathbf{H}^\top\mathbf{Z}^{(ij)} + \mathbf{Z}^{(ij)\top}\mathbf{H} = \begin{pmatrix} & & h_{i1} & & \\ & & \vdots & & \\ h_{i1} & \cdots & 2h_{ij} & \cdots & h_{iq} \\ & & \vdots & & \\ & & h_{iq} & & \end{pmatrix}.$$

Now consider the second term on the right-hand-side of (B.5). After some algebra, its (i, j)th entry can be written as $\mathbf{u}_i^\top \mathbf{D}\mathbf{a}^{(ij)}$, where $\mathbf{a}^{(ij)} \in \mathbb{R}^q$ and

$$a_k^{(ij)} = -\sum_{l \neq k}^{r} \frac{1}{d_k^2 - d_l^2}(v_{jk}v_{jl}\mathbf{h}_i^\top \mathbf{v}_l + v_{jl}^2 \mathbf{h}_i^\top \mathbf{v}_k), \qquad k = 1, ..., q.$$

Similarly, the (i, j)th entry of the third term on the right-hand-side of (B.5) is given by $\mathbf{u}_i^\top \mathbf{D}\mathbf{b}^{(ij)}$, where $\mathbf{b}^{(ij)} \in \mathbb{R}^q$,

$$b_k^{(ij)} = -\sum_{l \neq k}^{q} \frac{1}{d_l^2 - d_k^2}(v_{jk}v_{jl}\mathbf{h}_i^\top \mathbf{v}_s + v_{jl}^2 \mathbf{h}_i^\top \mathbf{v}_k), \qquad k = 1, ..., r,$$

and $b_k^{(ij)} = 0$ for $k = r + 1, ..., q$ whenever $r < q$. Now consider the second and third terms together. Since

$$a_k^{(ij)} + b_k^{(ij)} = \begin{cases} \displaystyle\sum_{l=r+1}^{q} \frac{1}{d_k^2 - d_l^2}(v_{jk}v_{jl}\mathbf{h}_i^\top \mathbf{v}_l + v_{jl}^2 \mathbf{h}_i^\top \mathbf{v}_k) & k = 1, ..., r; \\ \displaystyle\sum_{l=}^{r} \frac{1}{d_l^2 - d_k^2}(v_{jk}v_{jl}\mathbf{h}_i^\top \mathbf{v}_l + v_{jl}^2 \mathbf{h}_i^\top \mathbf{v}_k) & k = r + 1, ..., q. \end{cases}$$

it follows that the contribution from the second and the third term to the degrees of freedom equals

$$
\begin{aligned}
&\sum_{i=1}^{r_x}\sum_{j=1}^{q}\left\{\sum_{k=1}^{r} u_{ik}d_k \sum_{l=r+1}^{q}\frac{1}{d_k^2-d_l^2}(v_{jk}v_{jl}\mathbf{h}_i^\top\mathbf{v}_l + v_{jl}^2\mathbf{h}_i^\top\mathbf{v}_k)\right\}\\
&\quad+\sum_{i=1}^{r_x}\sum_{j=1}^{q}\left\{\sum_{k=r+1}^{q} u_{ik}d_k \sum_{l=1}^{r}\frac{1}{d_l^2-d_k^2}(v_{jk}v_{jl}\mathbf{h}_i^\top\mathbf{v}_l + v_{jl}^2\mathbf{h}_i^\top\mathbf{v}_k)\right\}\\
=&\sum_{i=1}^{r_x}\left\{\sum_{k=1}^{r} u_{ik}d_k \sum_{l=r+1}^{q}\frac{1}{d_k^2-d_l^2}\sum_{j=1}^{q}(v_{jk}v_{jl}\mathbf{h}_i^\top\mathbf{v}_l + v_{jl}^2\mathbf{h}_i^\top\mathbf{v}_k)\right\}\\
&\quad+\sum_{i=1}^{r_x}\left\{\sum_{k=r+1}^{q} u_{ik}d_k \sum_{l=1}^{r}\frac{1}{d_l^2-d_k^2}\sum_{j=1}^{q}(v_{jk}v_{jl}\mathbf{h}_i^\top\mathbf{v}_l + v_{jl}^2\mathbf{h}_i^\top\mathbf{v}_k)\right\}\\
=&\sum_{i=1}^{r_x}\left\{\sum_{k=1}^{r}\sum_{l=r+1}^{q}\frac{d_k}{d_k^2-d_l^2}u_{ik}(\mathbf{h}_i^\top\mathbf{v}_k) + \sum_{k=r+1}^{q}\sum_{l=1}^{r}\frac{d_k}{d_l^2-d_k^2}u_{ik}(\mathbf{h}_i^\top\mathbf{v}_k)\right\}\\
=&\sum_{i=1}^{r_x}\left\{\sum_{k=1}^{r}\sum_{l=r+1}^{q}\frac{d_k}{d_k^2-d_l^2}u_{ik}(\mathbf{h}_i^\top\mathbf{v}_k) + \sum_{k=1}^{r}\sum_{l=r+1}^{q}\frac{d_l}{d_k^2-d_l^2}u_{il}(\mathbf{h}_i^\top\mathbf{v}_l)\right\}\\
=&\sum_{k=1}^{r}\sum_{l=r+1}^{q}\left\{\frac{d_k}{d_k^2-d_l^2}\sum_{i=1}^{r_x}u_{ik}(\mathbf{h}_i^\top\mathbf{v}_k) + \frac{d_l}{d_k^2-d_l^2}\sum_{i=1}^{r_x}u_{il}(\mathbf{h}_i^\top\mathbf{v}_l)\right\}\\
=&\sum_{k=1}^{r}\sum_{l=r+1}^{q}\left\{\frac{d_k}{d_k^2-d_l^2}\mathbf{u}_k^\top\mathbf{H}\mathbf{v}_k + \frac{d_l}{d_k^2-d_l^2}\mathbf{u}_l^\top\mathbf{H}\mathbf{v}_l\right\}\\
=&\sum_{k=1}^{r}\sum_{l=r+1}^{q}\left\{\frac{d_k^2}{d_k^2-d_l^2} + \frac{d_l^2}{d_k^2-d_l^2}\right\}\\
=&\sum_{k=1}^{r}\sum_{l=r+1}^{q}\frac{d_k^2+d_l^2}{d_k^2-d_l^2}.
\end{aligned}
$$

Combining the result in (B.6), the proof is completed.

Proof of Theorem III.3

Again, we assume $r_x \geq q$. When $r_x \leq q$, one can repeat the same proof using $\mathbf{H}^\top$. Recall that $\tilde{\mathbf{H}}(\lambda) = \mathbf{U}\tilde{\mathbf{D}}(\lambda)\mathbf{V}^\top$. Consider $\partial\tilde{\mathbf{H}}(\lambda)/\partial h_{ij}$ for any fixed $\lambda > 0$, $1 \leq i \leq r_x$ and $1 \leq j \leq q$. Denote $\tilde{r} = \tilde{r}(\lambda) = \max\{k : s_k > 0.\}$. Because

$\tilde{\mathbf{H}}(\lambda) = \mathbf{H}\sum_{k=1}^{\tilde{r}} s_k \mathbf{v}_k \mathbf{v}_k^\top$, by the chain rule, we have

$$\begin{aligned}
\frac{\partial \tilde{\mathbf{H}}(\lambda)}{\partial h_{ij}} =& \frac{\partial \mathbf{H}}{\partial h_{ij}} \sum_{k=1}^{\tilde{r}} s_k \mathbf{v}_k \mathbf{v}_k^\top + \mathbf{H}\sum_{k=1}^{\tilde{r}} s_k \frac{\partial \mathbf{v}_k}{\partial h_{ij}} \mathbf{v}_k^\top + \mathbf{H}\sum_{k=1}^{\tilde{r}} s_k \mathbf{v}_k \frac{\partial \mathbf{v}_k^\top}{\partial h_{ij}} + \mathbf{H}\sum_{k=1}^{\tilde{r}} \frac{\partial s_k}{\partial h_{ij}} \mathbf{v}_k \mathbf{v}_k^\top \\
=& \mathbf{Z}^{(ij)} \mathbf{V}^{(\tilde{r})} \mathbf{D}^{(\tilde{r})-1} \tilde{\mathbf{D}}^{(\tilde{r})} \mathbf{V}^{(\tilde{r})\top} \\
& - \mathbf{H}\sum_{k=1}^{\tilde{r}} \left\{ s_k (\mathbf{H}^\top \mathbf{H} - d_k^2 \mathbf{I})^- (\mathbf{H}^\top \mathbf{Z}^{(ij)} + \mathbf{Z}^{(ij)\top}\mathbf{H}) \mathbf{v}_k \mathbf{v}_k^\top \right\} \\
& - \mathbf{H}\sum_{k=1}^{\tilde{r}} \left\{ s_k \mathbf{v}_k \mathbf{v}_k^\top (\mathbf{H}^\top \mathbf{Z}^{(ij)} + \mathbf{Z}^{(ij)\top}\mathbf{H}) (\mathbf{H}^\top \mathbf{H} - d_k^2 \mathbf{I})^- \right\} \\
& + \mathbf{H}\sum_{k=1}^{\tilde{r}} \left\{ s_k' \{ \frac{1}{2d_k} \mathbf{v}_k^\top (\mathbf{H}^\top \mathbf{Z}^{(ij)} + \mathbf{Z}^{(ij)\top}\mathbf{H}) \mathbf{v}_k \} \mathbf{v}_k \mathbf{v}_k^\top \right\},
\end{aligned} \tag{B.7}$$

where $s_k' = \partial s_k / \partial d_k$. Consider the first term on the right-hand-side of (**??**). It can be shown that its (i,j)th entry equals to $\sum_{k=1}^{\tilde{r}} s_k v_{jk}^2$. Therefore, its contribution to the degrees of freedom (3.10) is

$$\sum_{i=1}^{r_x} \sum_{j=1}^{q} \sum_{k=1}^{\tilde{r}} s_k v_{jk}^2 = r_x \sum_{k=1}^{\tilde{r}} s_k, \tag{B.8}$$

because $\sum_{j=1}^{q} v_{jk}^2 = 1$. Similar to the proof of Theorem III.2, the (i,j)th entry of the second and third terms on the right-hand-side of (B.7) can be shown to be

$$\mathbf{u}_i^\top \mathbf{D}(\tilde{\mathbf{a}}^{(ij)} + \tilde{\mathbf{b}}^{(ij)}) \tag{B.9}$$

where $\tilde{\mathbf{a}}^{(ij)} \in \mathbb{R}^q$, $\tilde{\mathbf{b}}^{(ij)} \in \mathbb{R}^q$, and

$$\tilde{a}_k^{(ij)} + \tilde{b}_k^{(ij)} = \begin{cases} \displaystyle\sum_{l \neq k}^{q} \frac{s_k - s_l}{d_k^2 - d_l^2} (v_{jk} v_{jl} \mathbf{h}_i^\top \mathbf{v}_l + v_{jl}^2 \mathbf{h}_i^\top \mathbf{v}_k) & k = 1, ..., \tilde{r}; \\ \displaystyle\sum_{l=1}^{\tilde{r}} \frac{s_l}{d_l^2 - d_k^2} (v_{jk} v_{jl} \mathbf{h}_i^\top \mathbf{v}_l + v_{jl}^2 \mathbf{h}_i^\top \mathbf{v}_k) & k = \tilde{r}+1, ..., q. \end{cases}$$

After some algebra, it follows that the contribution from the second and the third term to the degrees of freedom equals

$$\begin{aligned}
&\sum_{i=1}^{r_x}\sum_{j=1}^{q}\left\{\sum_{k=1}^{\tilde{r}} u_{ik}d_k \sum_{l\neq k}^{q} \frac{s_k - s_l}{d_k^2 - d_l^2}(v_{jk}v_{jl}\mathbf{h}_i^\top \mathbf{v}_l + v_{jl}^2 \mathbf{h}_i^\top \mathbf{v}_k)\right\} \\
&+\sum_{i=1}^{r_x}\sum_{j=1}^{q}\left\{\sum_{k=\tilde{r}+1}^{q} u_{ik}d_k \sum_{l=1}^{\tilde{r}} \frac{s_l}{d_l^2 - d_k^2}(v_{jk}v_{jl}\mathbf{h}_i^\top \mathbf{v}_l + v_{jl}^2 \mathbf{h}_i^\top \mathbf{v}_k)\right\} \\
=&\sum_{k=1}^{\tilde{r}}\sum_{s=\tilde{r}+1}^{q}\left\{\frac{d_k^2(s_k - s_l) + d_l^2 s_k}{d_k^2 - d_l^2}\right\} + \sum_{k=1}^{\tilde{r}}\sum_{l\neq k}^{\tilde{r}}\left\{\frac{d_k^2(s_k - s_l)}{d_k^2 - d_l^2}\right\}. \qquad \text{(B.10)}
\end{aligned}$$

Consider the fourth term on the right-hand-side of (B.7). Note that

$$\mathbf{v}_k^\top(\mathbf{H}^\top \mathbf{Z}^{(ij)} + \mathbf{Z}^{(ij)\top}\mathbf{H})\mathbf{v}_k = 2v_{jk}(\mathbf{v}_k^\top \mathbf{h}_i).$$

The (i, j)th entry of the fourth term is given by

$$\sum_{k=1}^{\tilde{r}} s_k' u_{ik} v_{jk}^2 (\mathbf{v}_k^\top \mathbf{h}_i).$$

It then follows that the contribution of the fourth term to the degrees of freedom equals

$$\begin{aligned}
&\sum_{i=1}^{r_x}\sum_{j=1}^{q}\sum_{k=1}^{\tilde{r}} s_k' u_{ik} v_{jk}^2 (\mathbf{v}_k^\top \mathbf{h}_i) \\
=&\sum_{i=1}^{r_x}\sum_{k=1}^{\tilde{r}} s_k' u_{ik} (\mathbf{v}_k^\top \mathbf{h}_i) \\
=&\sum_{k=1}^{\tilde{r}} s_k' \sum_{i=1}^{r_x} u_{ik} \mathbf{h}_i^\top \mathbf{v}_k \\
=&\sum_{k=1}^{\tilde{r}} d_k s_k'.
\end{aligned}$$

Combining with the results in (B.8) and (B.10), the proof is completed.

APPENDIX C

Theoretical Properties of Non-parametric Joint Rank and Variable Selection method

Proof of Lemma IV.1

Define the space of centered spline functions of degree d as

$$\mathcal{S}_k^0 = \left\{ f_k^0 : f_k^0(x) = \sum_{l=1}^{L_n} \beta_{lk}\psi_{lk}(x),\ (\beta_{1k}, \ldots, \beta_{L_n k}) \in \mathbb{R}^{L_n} \right\}, \quad k = 1, \ldots, p. \tag{C.1}$$

We suppress the dependence on n to keep the notations simple. We assume that the true component functions $\{f_{jk} : j = 1, \ldots, q;\ k = 1, \ldots, p\}$ satisfy conditions **C1** and **C2**. Lemma 1 in (*Huang et al.*, 2010) states that for each fixed j there exist $\tilde{f}_{jk}^0 \in \mathcal{S}_k^0$, such that

$$\sum_{i=1}^{n} \sum_{k=1}^{J} \Big(f_{jk}(\mathbf{X}_{ik}) - \tilde{f}_{jk}^0(\mathbf{X}_{ik}) \Big)^2 \leq C_1 |J| L_n. \tag{C.2}$$

Clearly this holds for each response $j = 1, 2, \ldots, q$ thus we can extend the result to the multiple response scenario. Following the notation introduced in Section 4.2 and using (C.2) we get

$$\|\mathcal{E}(\mathbf{X})\|_F^2 \leq C_1 |J| q L_n. \tag{C.3}$$

Proof of Theorem IV.2

Let $rank(\mathbf{f}(\mathbf{X})) = r$ and $\hat{r}$ denote the selected in Step I (4.8) of non-parametric low-rank group lasso. The solution is characterized by (4.10) for a detailed derivation see (*Bunea et al.*, 2011). Let us denote the projection matrix to the spline basis matrix by $\mathbf{P}_\Psi = \mathbf{\Psi}(\mathbf{\Psi}^\top\mathbf{\Psi})^-\mathbf{\Psi}^\top$, then the solution can be simplified as

$$\mathbf{P}_\Psi\mathbf{Y} = \mathbf{\Psi}\mathbf{B} + \mathbf{P}_\Psi\mathcal{E}(\mathbf{X}) + \mathbf{P}_\Psi\mathbf{E} = \mathbf{\Psi}\mathbf{B} + \tilde{\mathbf{E}}. \tag{C.4}$$

Suppose $\exists\, \delta \in (0,1)$ such that $d_r(\mathbf{\Psi}\mathbf{B}) > (1+\delta)\sqrt{\mu}$ and $d_{r+1}(\mathbf{\Psi}\mathbf{B}) < (1-\delta)\sqrt{\mu}$, then the result in Theorem 2 of (*Bunea et al.*, 2011) implies

$$\mathbb{P}(\hat{r} = r) \geq 1 - \mathbb{P}(d_1(\tilde{\mathbf{E}}) \geq \delta\sqrt{\mu}), \tag{C.5}$$

where the probability is defined over the distribution of the random errors, $\mathbf{E}$. Next we try to bound the largest singular value of $\tilde{\mathbf{E}}$ using the bound from Lemma reflemma1 and largest singular value of $\mathbf{P}_\Psi\mathbf{E}$. We will also make use of the following inequality $d_k(\mathbf{M}_1 + \mathbf{M}_2) \leq d_k(\mathbf{M}_1) + d_1(\mathbf{M}_2)$ for two generic matrices $\mathbf{M}_1$ and $\mathbf{M}_2$.

$$\begin{aligned}
d_1(\tilde{\mathbf{E}}) &\leq d_1(\mathbf{P}_\Psi\mathcal{E}(\mathbf{X})) + d_1(\mathbf{P}_\Psi\mathbf{E}) \\
&\leq \|\mathbf{P}_\Psi\mathcal{E}(\mathbf{X})\|_F + d_1(\mathbf{P}_\Psi\mathbf{E}) \\
&\leq d_1(\mathbf{P}_\Psi\mathbf{E}) + C_1\sqrt{|J|qL_n}.
\end{aligned}$$

Combining this with the inequality (C.5) we have the following

$$\mathbb{P}(\hat{r} = r) \geq 1 - \mathbb{P}(d_1(\mathbf{P}_\Psi\mathbf{E}) \geq \delta\sqrt{\mu} - C_1\sqrt{|J|qL_n}). \tag{C.6}$$

Now we need to bound the largest singular value of the projection of a matrix with i.i.d. Gaussian entries. We use Lemma 3 in (*Bunea et al.*, 2011) to achieve that. Let $\delta = 1/2$ and $\sqrt{\mu} = 2\sigma(\sqrt{q} + \sqrt{pL_n}) + 4C_1\sqrt{|J|qL_n}$. Then we have

$$\mathbb{P}(\hat{r} = r) \geq 1 - exp\left\{-\frac{C_1^2|J|qL_n}{2\sigma^2}\right\}, \tag{C.7}$$

which goes to 0 as $|J|qL_n \to \infty$. To complete the proof we only need to show that for the choices of δ and μ made in the previous part we have $d_r(\mathbf{\Psi}\mathbf{B}) > (1+\delta)\sqrt{\mu}$ and $d_{r+1}(\mathbf{\Psi}\mathbf{B}) < (1-\delta)\sqrt{\mu}$ hold as long as $d_r(\mathbf{f}(\mathbf{X})) \geq 4\sigma(\sqrt{q} + \sqrt{pL_n}) + 7C_1\sqrt{|J|qL_n}$. We again make use of the inequality for the singular value of sum of two matrices using the singular values for the individual matrices. In particular

$$\begin{aligned} d_{r+1}(\mathbf{\Psi B}) &\leq d_{r+1}(\mathbf{f}(\mathbf{X})) + d_1(\mathcal{E}(\mathbf{X})) \leq C_1\sqrt{|J|qL_n} \leq \sigma(\sqrt{q}+\sqrt{pL_n}) + 2C_1\sqrt{|J|qL_n} = \frac{\sqrt{\mu}}{2}, \\ d_r(\mathbf{\Psi B}) &\geq d_r(\mathbf{f}(\mathbf{X})) - C_1\sqrt{|J|qL_n} \geq 3\sigma(\sqrt{q}+\sqrt{pL_n}) + 6C_1\sqrt{|J|qL_n} \geq \frac{3\sqrt{\mu}}{2} \quad \square \end{aligned}$$

Proof of Theorem IV.3

Let us start with the fixed rank case where $\hat{r} = s$ for some $1 \leq s \leq \min\{n, p, q\}$. Define $\widehat{\mathbf{W}}^{(s)}$ to be the solution to the optimization problem in (4.9) when rank is restricted to be s

$$\widehat{\mathbf{W}}^{(s)} = \underset{\mathbf{W}:rank(\mathbf{W})\leq s}{\arg\min} \left\{ \frac{1}{2}\|\mathbf{Y} - \mathbf{\Psi W}\|_F^2 + \lambda\|\mathbf{W}\|_{F,1} \right\}.$$

Let $\hat{J}_s = \{k : \ \|\widehat{\mathbf{W}}_k^{(s)}\| > 0; \ k = 1, \ldots, p\}$ denote the set of selected variables in step II. Define $\tilde{J} = \hat{J} \cup J$, where J denotes the set of truly effective predictors. Also let $\tilde{\mathbf{\Psi}}$ to be the sub-matrix of $\mathbf{\Psi}$ containing the columns in $\tilde{J}$. From the definition of $\widehat{\mathbf{W}}^{(s)}$ it follows that for any $\mathbf{A} \in \mathbb{R}^{n \times pL_n}$ with $rank(\mathbf{A}) \leq s$

$$\begin{aligned} \|\mathbf{Y} - \mathbf{\Psi}\widehat{\mathbf{W}}^{(s)}\|_F^2 + 2\lambda\|\widehat{\mathbf{W}}^{(s)}\|_{F,1} &\leq \|\mathbf{Y} - \mathbf{\Psi A}\|_F^2 + 2\lambda\|\mathbf{A}\|_{F,1} & \text{(C.8)} \\ \|\mathbf{f}(\mathbf{X}) - \mathbf{\Psi}\widehat{\mathbf{W}}^{(s)}\|_F^2 + 2\lambda\|\widehat{\mathbf{W}}^{(s)}\|_{F,1} &\leq \|\mathbf{f}(\mathbf{X}) - \mathbf{\Psi A}\|_F^2 + 2\langle \mathbf{E}, \mathbf{\Psi}(\mathbf{W}^{(s)} - \mathbf{A})\rangle + 2\lambda\|\mathbf{A}\|_{F,1}. \end{aligned}$$

Where $\langle \mathbf{A}, \mathbf{B} \rangle = trace(\mathbf{A}^\top \mathbf{B})$ denotes the inner product corresponding to the Frobenius norm introduced earlier. To keep notations simple we define the following two quantities

$$\widehat{\Delta}_s^2 = \|\mathbf{f}(\mathbf{X}) - \mathbf{\Psi}\widehat{\mathbf{W}}^{(s)}\|_F^2, \quad \widehat{\Delta}^2 = \|\mathbf{f}(\mathbf{X}) - \mathbf{\Psi A}\|_F^2. \qquad \text{(C.9)}$$

Next we bound the inner product using the fact, $\langle \mathbf{A}, \mathbf{B} \rangle \leq \|\mathbf{A}\|_2 \|\mathbf{B}\|_1$. Where $\|\mathbf{A}\|_2 = d_1(\mathbf{A})$ is the operator norm of a matrix and $\|\mathbf{B}\|_1 = \sum d_j(\mathbf{B})$ is the nuclear norm.

$$\begin{aligned} 2\langle \mathbf{E}, \mathbf{\Psi}(\mathbf{W}^{(s)} - \mathbf{A})\rangle &= 2\langle \mathbf{P}_{\tilde{J}}\mathbf{E}, \tilde{\mathbf{\Psi}}(\mathbf{W}^{(s)} - \mathbf{A})\rangle \\ &\leq 2d_1(\mathbf{P}_{\tilde{J}}\mathbf{E})\sqrt{2s}\|\mathbf{\Psi}(\mathbf{W}^{(s)} - \mathbf{A})\|_F \\ &\leq 2d_1(\mathbf{P}_{\tilde{J}}\mathbf{E})\sqrt{2s}(\widehat{\Delta}_s + \Delta) \\ &\leq 16sd_1^2(\mathbf{P}_{\tilde{J}}\mathbf{E}) + \frac{1}{4}(\widehat{\Delta}_s^2 + \Delta^2). \end{aligned}$$

Here we have made use of the inequality $2xy \leq 4x^2 + y^2/4$ to arrive at the last inequality. Plugging it back into (C.8) and collecting terms we have

$$\frac{3}{4}\widehat{\Delta}_s^2 + 2\lambda\|(\widehat{\mathbf{W}}^{(s)} - \mathbf{A})_{J^c}\|_{F,1} \leq \frac{5}{4}\widehat{\Delta}^2 + 16sd_1^2(\mathbf{P}_{\tilde{J}}\mathbf{E}) + 2\lambda\|(\widehat{\mathbf{W}}^{(s)} - \mathbf{A})_J\|_{F,1}. \quad \text{(C.10)}$$

Next we break down the proof into two disjoint cases which makes the proof tractable and easier to follow.

Case I: $\frac{5}{4}\widehat{\Delta}^2 + 16sd_1^2(\mathbf{P}_{\tilde{J}}\mathbf{E}) \leq 2\lambda\|(\widehat{\mathbf{W}}^{(s)} - \mathbf{A})_J\|_{F,1}$

In this case (C.10) implies

$$\|(\widehat{\mathbf{W}}^{(s)} - \mathbf{A})_{J^c}\|_{F,1} \leq 2\|(\widehat{\mathbf{W}}^{(s)} - \mathbf{A})_J\|_{F,1}. \quad \text{(C.11)}$$

Invoking condition **C3** on Σ_Ψ we know that there esists $\delta_J > 0$ such that

$$\|\mathbf{\Psi}(\widehat{\mathbf{W}}^{(s)} - \mathbf{A})_{J^c}\|_F^2 \geq n\delta_J \sum_{k\in J}\|\widehat{\mathbf{W}}_k^{(s)} - \mathbf{A}_k\|_F^2. \quad \text{(C.12)}$$

Again using the condition in Case I we have

$$\begin{aligned}
\frac{3}{4}\widehat{\Delta}_s^2 &\leq 4\lambda\|(\widehat{\mathbf{W}}^{(s)} - \mathbf{A})_J\|_{F,1} \\
&= 2.2\lambda\sum_{k\in J}\|\widehat{\mathbf{W}}_k^{(s)} - \mathbf{A}_k\|_F \\
&\leq 2\left\{\frac{a}{n}\lambda^2|J| + \frac{n}{a}\sum_{k\in J}\|\widehat{\mathbf{W}}_k^{(s)} - \mathbf{A}_k\|_F^2\right\} \\
&\leq 2\left\{\frac{a}{n}\lambda^2|J| + \frac{1}{a\delta_J}\|\Psi\widehat{\mathbf{W}}_k^{(s)} - \Psi\mathbf{A}_k\|_F^2\right\} \\
&\leq 2\left\{\frac{a}{n}\lambda^2|J| + \frac{2}{a\delta_J}(\widehat{\Delta}_s^2 + \Delta^2)\right\} \\
&\leq \frac{16}{n\delta_J}\lambda^2|J| + \frac{1}{2}(\widehat{\Delta}_s^2 + \Delta^2) \quad \text{(taking } a = 8/\delta_J\text{)} \\
\implies \widehat{\Delta}_s^2 &\leq 2\Delta^2 + \frac{64}{n\delta_J}\lambda^2|J| \\
\implies \|\mathbf{f}(\mathbf{X}) - \mathbf{\Psi}\widehat{\mathbf{W}}^{(s)}\|_F^2 &\leq 4\|\mathbf{\Psi}\mathbf{B} - \mathbf{\Psi}\widehat{\mathbf{A}}\|_F^2 + 4\|\mathcal{E}(\mathbf{X})\|_F^2 + \frac{64}{n\delta_J}\lambda^2|J| \quad \text{(C.13)}
\end{aligned}$$

Note thet **B** is the coefficient matrix arising from spline approximation and $\mathcal{E}(\mathbf{X})$ is the approximation error. Even though $\mathbf{f}(\mathbf{X})$ is of rank s, $\mathbf{B}$ can have rank $> s$. Therefore we cannot choose $\mathbf{A} = \mathbf{B}$ and get rid of the first term in (C.13). This term reflects the effect of the approximation error which we will bound in the later part of the proof.

Case II: $\frac{5}{4}\widehat{\Delta}^2 + 16sd_1^2(\mathbf{P}_{\tilde{J}}\mathbf{E}) > 2\lambda\|(\widehat{\mathbf{W}}^{(s)} - \mathbf{A})_J\|_{F,1}$

Plugging it back into (C.10) we get

$$3\widehat{\Delta}_s^2 \leq 10\Delta^2 + 128sd_1^2(\mathbf{P}_{\tilde{J}}\mathbf{E}). \tag{C.14}$$

Taking expectation over the distribution of $\mathbf{E}$ we have

$$3\mathbb{E}(\widehat{\Delta}_s^2) \leq 10\Delta^2 + 128s\mathbb{E}(d_1^2(\mathbf{P}_{\tilde{J}}\mathbf{E})). \tag{C.15}$$

We apply Lemma 8 in (*Bunea et al.*, 2011) to bound the last term of (C.15) which leads to

$$\begin{aligned}
\mathbb{E}(d_1^2(\mathbf{P}_{\tilde{J}}\mathbf{E})) &\leq 6\sigma^2 q + 12\sigma^2 e^{-q/2} + 6\sigma^2\mathbb{E}\left(|\tilde{J}|L_n + |\tilde{J}|\log\frac{ep}{|\tilde{J}|}\right) \\
&\leq 6\sigma^2 q + 12\sigma^2 e^{-q/2} + 12\sigma^2\log(ep)L_n\mathbb{E}(|\tilde{J}|) \\
&\leq 6\sigma^2 q + 12\sigma^2 e^{-q/2} + 12\sigma^2\log(ep)L_n|J| + 12\sigma^2\log(ep)L_n\mathbb{E}(|\hat{J}_s|)
\end{aligned} \tag{C.16}$$

The next step involves bounding $\mathbb{E}(|\hat{J}_s|)$. As $rank(\widehat{\mathbf{W}}^{(s)}) \leq s$ we can write $\widehat{\mathbf{W}}^{(s)} = \widehat{\mathbf{S}}\widehat{\mathbf{V}}_s^\top$ such that $\widehat{\mathbf{V}}_s^\top\widehat{\mathbf{V}}_s = \mathbf{I}_s$. Given $\widehat{\mathbf{V}}_s$, $\widehat{\mathbf{S}}$ is the global optimum of a convex optimization problem and therefore satisfies the following KKT condition

$$\widehat{\mathbf{W}}_k^{(s)} \neq 0 \iff \|\mathbf{\Psi}_k^\top(\mathbf{\Psi}\widehat{\mathbf{S}} - \mathbf{Y}\widehat{\mathbf{V}}_s)\|_F = \lambda. \tag{C.17}$$

This implies

$$\begin{aligned}
|\widehat{J}_s|\lambda^2 &= \sum_{k\in\widehat{J}_s}\|\mathbf{\Psi}_k^\top(\mathbf{\Psi}\widehat{\mathbf{S}} - \mathbf{Y}\widehat{\mathbf{V}}_s)\|_F^2 \\
&= \sum_{k\in\widehat{J}_s}\|\mathbf{\Psi}_k^\top(\mathbf{\Psi}\widehat{\mathbf{S}} - \mathbf{f}(\mathbf{X})\widehat{\mathbf{V}}_s - \mathbf{P}_{\widehat{J}_s}\mathbf{E}\widehat{\mathbf{V}}_s)\|_F^2 \\
&\leq 2\sum_{k\in\widehat{J}_s}\|\mathbf{\Psi}_k^\top(\mathbf{\Psi}\widehat{\mathbf{S}} - \mathbf{f}(\mathbf{X})\widehat{\mathbf{V}}_s)\|_F^2 + 2\sum_{k\in\widehat{J}_s}\|\mathbf{\Psi}_k^\top\mathbf{P}_{\widehat{J}_s}\mathbf{E}\widehat{\mathbf{V}}_s\|_F^2 \\
&\leq 2n\lambda_1(\Sigma_\Psi)(\widehat{\Delta}_s^2 + sd_1^2(\mathbf{P}_{\widehat{J}_s}\mathbf{E})).
\end{aligned} \tag{C.18}$$

Taking expectations on both sides and using the fact that $\widehat{J}_s \subseteq \tilde{J}$

$$\mathbb{E}(|\widehat{J}_s|) \leq \frac{2n\lambda_1(\Sigma_\Psi)}{\lambda^2}\left\{\mathbb{E}(\widehat{\Delta}_s^2) + \mathbb{E}(\mathbf{P}_{\tilde{J}}\mathbf{E})\right\} \tag{C.19}$$

Choose $\lambda^2 = C_2\sigma^2 \log(ep)nL_n s\lambda_1(\Sigma_P si)$ with $C_2 = 1584$ then we can combine (C.15), (C.16) and (C.19) to get

$$\begin{aligned} \mathbb{E}(\widehat{\Delta}_s^2) &\leq 10\Delta^2 + 130s\left\{6\sigma^2 q + 12\sigma^2 e^{-q/2} + 12\sigma^2\log(ep)|J|L_n\right\} \\ &\leq 20\|\mathbf{\Psi B} - \mathbf{\Psi A}\|_F^2 + 20\|\mathcal{E}(\mathbf{X})\|_F^2 + 130s\left\{6\sigma^2 q + 12\sigma^2 e^{-q/2} + 12\sigma^2\log(ep)|J|L_n\right\} \end{aligned} \tag{C.20}$$

Combining Cases I and II and ignoring numeric constants and lower order terms we finally have

$$\mathbb{E}(\widehat{\Delta}_s^2) \lesssim \|\mathbf{\Psi B} - \mathbf{\Psi A}\|_F^2 + \|\mathcal{E}(\mathbf{X})\|_F^2 + sq + s|J|L_n\log(ep). \tag{C.21}$$

Recall that here $\mathbf{A} \in \mathbb{R}^{n\times pL_n}$ with $rank(\mathbf{A}) \leq s$. Next we will simplify the bound in (C.21) for the case when the selected rank, $\hat{r} = r$ which is the true underlying rank. The other scenario, that is, $\hat{r} \neq r$ will be handled later. Let $\mathbf{\Psi B}$ has a SVD of the form

$$\mathbf{\Psi B} = \sum_{i=1}^{q} d_i u_i v_i^\top. \tag{C.22}$$

Choose $\mathbf{A} = \mathbf{B}\sum_{i=1}^{r} d_i u_i v_i^\top$, which implies

$$\|\mathbf{\Psi B} - \mathbf{\Psi A}\|_F^2 = \sum_{i=r+1}^{q} d_i^2(\mathbf{\Psi B}). \tag{C.23}$$

Now we apply Mirsky's theorem (*Mirsky*, 1960) to bound the right hand side of the above equation. Mirsky's theorem states that for generic matrices $\mathbf{M}_1 = \mathbf{M}_2 + \mathbf{M}_3$ we have

$$\sum_j (d_j(\mathbf{M}_1) - d_j(\mathbf{M}_2))^2 \leq \|\mathbf{M}_3\|_F^2. \tag{C.24}$$

Recall the equation for spline approximation given by, $\mathbf{f}(\mathbf{X}) = \mathbf{\Psi B} + \mathcal{E}(\mathbf{X})$. We apply the above result with $\mathbf{M}_1 = \mathbf{f}(\mathbf{X})$, $\mathbf{M}_2 = \boldsymbol{\Psi}\mathbf{B}$ and $\mathbf{M}_3 = \mathcal{E}(\mathbf{X})$ to get

$$\|\boldsymbol{\Psi}\mathbf{B} - \boldsymbol{\Psi}\mathbf{A}\|_F^2 = \sum_{i=r+1}^{q} d_i^2(\boldsymbol{\Psi}\mathbf{B}) \leq \|\mathcal{E}(\mathbf{X})\|_F^2 \leq C|J|qL_n. \tag{C.25}$$

Therefore under selection of true rank $\hat{r} = r$, plugging (C.25) back into (C.21) we get

$$\mathbb{E}(\|\mathbf{f}(\mathbf{X}) - \boldsymbol{\Psi}\widehat{\mathbf{W}}^{(r)}\|_F^2) \lesssim |J|qL_n + rq + r|J|L_n \log(ep). \tag{C.26}$$

To bound the expected risk in the general case we start by splitting it up as follows

$$\mathbb{E}(\|\mathbf{f}(\mathbf{X}) - \boldsymbol{\Psi}\widehat{\mathbf{W}}^{(\hat{r})}\|_F^2) = \mathbb{E}(\|\mathbf{f}(\mathbf{X}) - \boldsymbol{\Psi}\widehat{\mathbf{W}}^{(r)}\|_F^2 \mathbb{1}\{\hat{r} = r\}) + \mathbb{E}(\|\mathbf{f}(\mathbf{X}) - \boldsymbol{\Psi}\widehat{\mathbf{W}}^{(\hat{r})}\|_F^2 \mathbb{1}\{\hat{r} \neq r\}) \tag{C.27}$$

The preceding derivation bounds the first term on the right of (C.27) for $\hat{r} = r$. It remains to show that the second term is bounded by some constant. From the definition of $\widehat{\mathbf{W}}^{(\hat{r})}$ we have

$$\begin{aligned} \|\mathbf{Y} - \boldsymbol{\Psi}\widehat{\mathbf{W}}^{(\hat{r})}\|_F^2 + 2\lambda\|\widehat{\mathbf{W}}^{(\hat{r})}\|_{F,1} &\leq \|\mathbf{Y} - \boldsymbol{\Psi}\mathbf{0}\|_F^2 + 2\lambda\|\mathbf{0}\|_{F,1} = \|\mathbf{Y}\|_F^2 \\ \Rightarrow \|\mathbf{Y} - \boldsymbol{\Psi}\widehat{\mathbf{W}}^{(\hat{r})}\|_F^2 &\leq \|\mathbf{Y}\|_F^2. \end{aligned} \tag{C.28}$$

Using the above fact and norm inequalities we get,

$$\begin{aligned} \mathbb{E}\left[\|\mathbf{f}(\mathbf{X}) - \boldsymbol{\Psi}\widehat{\mathbf{W}}^{(\hat{r})}\|_F^2 \mathbb{1}\{\hat{r} \neq r\}\right] &\leq \mathbb{E}\left[\left\{2\|\mathbf{Y} - \mathbf{f}(\mathbf{X})\|_F^2 + 2\|\mathbf{Y} - \boldsymbol{\Psi}\widehat{\mathbf{W}}^{(\hat{r})}\|_F^2\right\} \mathbb{1}\{\hat{r} \neq r\}\right] \\ &\leq \mathbb{E}\left[\left\{2\|\mathbf{E}\|_F^2 + 2\|\mathbf{f}(\mathbf{X}) + \mathbf{E}\|_F^2\right\} \mathbb{1}\{\hat{r} \neq r\}\right] \\ &\leq \mathbb{E}\left[\left\{6\|\mathbf{E}\|_F^2 + 4\|\mathbf{f}(\mathbf{X})\|_F^2\right\} \mathbb{1}\{\hat{r} \neq r\}\right] \\ &\leq 6\mathbb{E}\left[\|\mathbf{E}\|_F^2 \mathbb{1}\{\hat{r} \neq r\}\right] + 4\|\mathbf{f}(\mathbf{X})\|_F^2 \mathbb{P}(\hat{r} \neq r) \end{aligned} \tag{C.29}$$

Using Cauchy-Schwartz inequality and the fact that $\|\mathbf{E}\|_F^2/\sigma^2$ follows a Chi-square distribution with nq degrees of freedom we have

$$\mathbb{E}\left[\|\mathbf{E}\|_F^2 \mathbb{1}\{\hat{r} \neq r\}\right] \leq \sqrt{\mathbb{E}[\|\mathbf{E}\|_F^4]\mathbb{P}(\hat{r} \neq r)} \leq \sqrt{2}nq\sigma^2 exp\left\{\frac{-C_1^2|J|qL_n}{4\sigma^2}\right\} \leq C_4. \tag{C.30}$$

Here C_3 is a numerical constant. The last inequality follows from the fact that $L_n = \mathcal{O}(n^\nu)$ for some $0 < \nu < 0.5$ and therefore the exponential term dominates the product. The boundedness of the second term on the right side of (C.29) follows from the condition **C4**. This completes the proof.

BIBLIOGRAPHY

BIBLIOGRAPHY

Akaike, H. (1974), A new look at the statistical model identification, *IEEE Transactions on Automatic Control*, *19*(6), 716–723.

Allman, E., C. Matias, and J. Rhodes (2009), Identiability of parameters in latent structure models with many observed variables, *Annals of Statistics*, *37*, 3099–3132.

Anderson, T. (1951), Estimating Linear Restrictions on Regression Coefficients for Multivariate Normal Distributions, *Annals of Mathematical Statistics*, *22*(3), 327–351.

Anderson, T. (1999), Asymptotic Distribution of the Reduced Rank Regression Estimator under General Conditions, *Annals of Statistics*, *27*(4), 1141–1154.

Anderson, T. (2002a), Specication and misspecication in reduced rank regression, *Sankhyā, Series A*, *64*, 193–205.

Anderson, T. (2002b), Reduced rank regression in cointegrated models, *Journal of Econometrics*, *106*, 203–216.

Arlot, S., and A. Celisse (2010), A survey of cross-validation procedures for model selection, *Statistics Surveys*, *4*, 40–79.

Bickel, P., Y. Ritov, and A. Tsybakov (2008), Simultaneous analysis of lasso and dantzig selector, *Annals of Statistics*, *37*, 1705–1732.

Breiman, L., and J. Friedman (1997), Predicting multivariate responses in multiple linear regression, *Journal of the Royal Statistical Society, Series B*, *59*, 3–37.

Bühlmann, P., and S. van de Geer (2011), *Statistics for High-Dimensional Data: Methods, Theory and Applications*, Springer Series in Statistics.

Bunea, F., Y. She, and M. Wegkamp (2011), Optimal Selection of Reduced Rank Estimators of High-Dimensional Matrices, *Annals of Statistics*, *39*(2), 1282–1309.

Bunea, F., Y. She, and M. Wegkamp (2012), Joint variable and rank selection for parsimonious estimation of high-dimensional matrices, *Annals of Statistics*, *40*(5), 2359–2388.

Chen, K., K. Chan, and N. Stenseth (2012), Reduced-rank stochastic regression with a sparse singular value decomposition, *Journal of the Royal Statistical Society, Series B*, *74*(2), 203–221.

Chen, L., and J. Huang (2012), Sparse reduced-rank regression for simultaneous dimension reduction and variable selection, *Journal of the American Statistical Association*, *107*(500), 1533–1545.

Cox, D., J. Little, and D. O'Shea (2007), *Ideals, Varieties, and Algorithms. An introduction to computational algebraic geometry and commutative algebra*, 3 ed., Springer, New York.

Davies, P., and M. Tso (1982), Procedures for Reduced-Rank Regression, *Applied Statistics*, *31*(3), 244–255.

de Leeuw, J. (2007), Derivatives of Generalized Eigen Systems with Applications, *Department of Statistics, UCLA*, pp. 1–28.

Donoho, D., and I. Johnstone (1995), Adapting to Unknown Smoothness via Wavelet Shrinkage, *Journal of the American Statistical Association*, *90*, 1200–1224.

Eckart, C., and G. Young (1936), The Approximation of One Matrix by Another of Lower Rank, *Psychometrika*, *1*(3), 211–218.

Efron, B. (2004), The Estimation of Prediction Error: Covariance Penalties and Cross-Validation (with discussion), *Journal of the American Statistical Association*, *99*, 619–642.

Golub, G., M. Heath, and G. Wahba (1979), Generalized Cross-Validation as a Method for Choosing a Good Ridge Parameter, *Technometrics*, *21*(2), 215–223.

Hartshorne, R. (1977), *Algebraic Geometry*, Springer-Verlag, New York.

Hastie, T., and R. Tibshirani (1990), *Generalized Additive Models*, Chapman and Hall, London.

Hastie, T., R. Tibshirani, and J. Friedman (2009), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer, New York.

Hocking, R., and R. Leslie (1967), Selection of the Best Subset in Regression Analysis, *Technometrics*, *9*, 531–540.

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

Huang, J., J. Horowitz, and F. Wei (2010), Variable selection in nonparametric additive models, *Annals of Statistics*, *38*(4), 2282–2313.

Izenman, A. (1975), Reduced-Rank Regression for the Multivariate Linear Model, *Journal of Multivariate Statistics*, *5*(2), 248–264.

Izenman, A. (2008), *Modern Multivariate. Statistical Techniques: Regression, Classification and Manifold Learning*, Springer, New York.

Laub, A. (2004), *Matrix Analysis for Scientists and Engineers*, Society for Industrial and Applied Mathematics, Philadelphia.

Li, M.-C., and K.-S. Chan (2007), Multivaraite reduced-rank nonlinear time series modeling, *Statistica Sinica*, *17*, 139–159.

Li, Y., and J. Zhu (2008), l_1-norm quantile regression, *Journal of Computational and Graphical Statistics*, *17*, 163–185.

Lin, Y., and H. Zhang (2006), Component selection and smoothing in multivariate non-parametric regression, *Annals of Statistics*, *34*, 2272–2297.

Lu, Z., R. Monteiro, and M. Yuan (2012), Convex optimization methods for dimension reduction and coefficient estimation in multivariate linear regression, *Mathematical Programming*, *131*, 163–194.

Magnus, J., and H. Neudecker (1998), *Matrix Differential Calculus with Applications in Statistics and Economatrix*, Wiley, New York.

Mallow, C. (1973), Some Comments on c_p, *Technometrics*, *15*, 661–675.

Massy, W. (1965), Principal Component Regression with Exploratory Statistical Research, *Journal of the American Statistical Association*, *60*, 234–246.

Meier, L., S. van de Geer, and P. Bühlmann (2009), High-dimensional additive modeling, *Annals of Statistics*, *37*, 3779–3821.

Meyer, M., and M. Woodroofe (2000), On the Degrees of Freedom in Shape-Restricted Regression, *Annals of Statistics*, *28*(4), 1083–1104.

Mirsky, L. (1960), Symmetric gage functions and unitarily invariant norms, *Quarterly Journal of Mathematics*, *11*, 50–59.

Moore, E. (1920), On the reciprocal of the general algebraic matrix, *Bulletin of the American Mathematical Society*, *26*(9), 394–395.

Mukherjee, A., and J. Zhu (2011), Reduced Rank Ridge Regression and its Kernel Extensions, *Statistical Analysis and Data Mining*, *4*(6), 612–622.

Neghaban, S., and M. Wainwright (2011), Estimation of (near) low-rank matrices with noise and high-dimensional scaling, *Annals of Statistics*, *39*, 1069–1097.

Obozinski, G., M. Wainwright, and M. Jordan (2011), Support union recovery in high-dimensional multivariate regression, *Annals of Statistics*, *39*(1), 1–17.

O'Neil, K. (2005), Critical Points of the Singular Value Decomposition, *SIAM Journal of Matrix Analysis and Applications*, *27*(2), 459–473.

Peng, J., A. Zhu, J. Bergamaschi, W. Han, D. Noh, J. Pollack, and P. Wang (2009), Regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer, *Annals of Applied Statistics*, *4*(1), 53–77.

Penrose, R. (1955), A generalized inverse for matrices, *Proceedings of the Cambridge Philosophical Society*, *51*, 406–413.

Rao, C. (1978), Matrix Approximations and Reduction of Dimensionality in Multivariate Statistical Analysis, in *In Proceedings of the Fifth International Symposium of Multivariate Analysis*, edited by P. R. Krishnaiah, pp. 3–22, North-Holland Publishing.

Ravikumar, P., H. Liu, J. Lafferty, and L. Wasserman (2009), Sparse additive models, *Journal of the Royal Statistical Society*, *71*, 1009–1030.

Reinsel, G., and R. Velu (1998), *Multivariate Reduced-Rank Regression: Theory and Applications*, Springer, New York.

Rohde, A., and A. Tsybakov (2011), Estimation of high-dimensional low-rank matrices, *Annals of Statistics*, *39*, 887–930.

Schumaker, L. (1981), *Spline Functions: Basic Theory*, Wiley, New York.

Schwarz, G. (1978), Estimating the Dimension of a Model, *Annals of Statistics*, *6*, 461–464.

Shao, J. (1993), Linear model selection by cross-validation, *Journal of the American Statistical Association*, *88*(422), 486–494.

She, Y. (2009), Thresholding-based iterative selection procedures for model selection and shrinkage, *Electronic Journal of Statistics*, *3*, 384–415.

She, Y. (2012), Reduced rank vector generalized linear models for feature extraction, *Statistics and Its Inference*, *To appear*, arXiv:1007.3098[stat.ML].

Shen, X., and J. Ye (2002), Adaptive Model Selection, *Journal of the American Statistical Association*, *97*, 210–221.

Skagerberg, B., J. MacGregor, and C. Kiparissdes (1992), Multivariate data analysis applied to low density polyethylene reactors, *Chemometrics and Intelligent Laboratory Systems*, *14*, 341–356.

Stein, C. (1973), Estimation of the Mean of a Multivariate Normal Distribution, *Tech. Rep. 48*, Stanford University.

Stein, C. (1981), Estimation of the Mean of a Multivariate Normal Distribution, *Annals of Statistics*, *9*, 1135–1151.

Stone, C. (1985), Additive regression and other nonparametric models, *Annals of Statistics*, *13*, 689–705.

Tibshirani, R. (1996), Regression Shrinkage and Selection via the Lasso, *Journal of the Royal Statistical Society*, *58*, 267–288.

Tibshirani, R., and J. Taylor (2011), Degrees of freedomin lasso problems, *Annals of Statistics*, *40*, 1198–1232.

Tsukuma, H. (2008), Admissibility and minimaxity of bayes estimators for a normal mean matrix, *Journal of Multivariate Analysis*, *99*, 2251 – 2264.

Turlach, B., W. Venables, and S. Wright (2005), Simultaneous variable selection, *Technometrics*, *47*(3), 349–363.

Tutuncu, R., K. Toh, and M. Todd (2003), Solving semidefinite-quadratic-linear programs using sdpt3, *Mathematical Programming*, *95*, 189–217.

Vaiter, S., C. Deledalle, G. Peyre, J. Fadili, and C. Dossal (2012), The degrees of freedom of the group lasso for a general design, *Tech. rep.*, Preprint Hal-00768896.

Velu, R. (1991), Reduced rank models with two sets of regressors, *Journal of the Royal Statistical Society, Series C*, *40*, 159–170.

Wahba, G. (), Spline models for observation data, *Society for Industrial and Applied Mathematics.*

Wille, A., et al. (2004), Sparse Graphical Gaussian Modeling of the Isoprenoid Gene Network in Arabidopsis Thaliana, *Genome Biology*, *5*, 1–13.

Wold, H. (1975), Soft modeling by latent variables: the non-linear iterative partial least squares approach, *Perspectives in Probability and Statistics, papers in honour of M. S. Bartlett.*

Ye, J. (1998), On Measuring and Correcting the Effects of Data Mining and Model Selection, *Journal of American Statistical Association*, *93*, 120–131.

Yee, T., and T. Hastie (2003), Reduced rank vector generalized linear models, *Statistical Modeling*, pp. 367–378.

Yuan, M., L. Ekici, Z. Lu, and R. Monteiro (2007), Dimension Reduction and Coefficient Estimation in Multivariate Linear Regression, *Journal of the royal Statistical Society*, *69*(3), 329–346.

Zou, H., T. Hastie, and R. Tibshirani (2007), On the "Degrees of Freedom" of the Lasso, *Annals of Statistics*, *35*(5), 2173–2192.