Reduced Rank Ridge Regression and Its Kernel Extensions

Ashin Mukherjee* and Ji Zhu

Department of Statistics, University of Michigan, Ann Arbor, MI 48109, USA

Received 28 January 2011; revised 11 August 2011; accepted 21 August 2011
DOI:10.1002/sam.10138
Published online in Wiley Online Library (wileyonlinelibrary.com).

Abstract: In multivariate linear regression, it is often assumed that the response matrix is intrinsically of lower rank. This could be because of the correlation structure among the prediction variables or the coefficient matrix being lower rank. To accommodate both, we propose a reduced rank ridge regression for multivariate linear regression. Specifically, we combine the ridge penalty with the reduced rank constraint on the coefficient matrix to come up with a computationally straightforward algorithm. Numerical studies indicate that the proposed method consistently outperforms relevant competitors. A novel extension of the proposed method to the reproducing kernel Hilbert space (RKHS) set-up is also developed. © 2011 Wiley Periodicals, Inc. Statistical Analysis and Data Mining 4: 612–622, 2011

Keywords: reduced rank regression; ridge regression; RKHS

1. INTRODUCTION

Multivariate linear regression is the simple extension of the classical univariate regression model to the case where we have $Q > 1$ responses and P predictors. It is commonly used in chemometrics, econometrics, and other similar quantitative fields where one is interested in predicting several responses generated by a single production process.

We can express the multivariate linear regression model in matrix notation. Let $\mathbf{X}$ denote the $N \times P$ predictor or design matrix, with ith row $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iP})$. Similarly the response matrix is denoted by $\mathbf{Y}$, $N \times Q$ where the ith row is $\mathbf{y}_i = (y_{i1}, y_{i2}, \ldots, y_{iQ})$. The regression parameter is given by the coefficient matrix $\mathbf{B}$ which is $P \times Q$. Note that the qth column of $\mathbf{B}$, $\beta_q = (\beta_{1q}, \beta_{2q}, \ldots, \beta_{Pq})$ is the regression coefficient vector for regressing the kth response on the predictors. Let $\mathbf{E}$ denote the $N \times Q$ random error matrix, then the model is,

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E}. \tag{1}$$

Note that this will reduce to the classical univariate regression model if $Q = 1$. For notational simplicity we will assume that the columns of the response and the predictors are centered and scaled so that the intercept terms can be omitted. The most standard approach to estimating the coefficient matrix $\mathbf{B}$ is by Ordinary Least Squares approach. The estimator is,

$$\hat{\mathbf{B}}_{\mathrm{OLS}} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y}. \tag{2}$$

This amounts to performing Q separate univariate regression of Y_q's on the predictors. The OLS approach fails to take advantage of any relationship or dependence between the responses, thus performs suboptimally when the true response dimension is $< Q$. In addition it is well known that this type of estimators perform poorly when the predictor variables are highly correlated.

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* [1], *Partial Least Squares* [2], *Canonical Correlation Analysis* [3]. 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.

Correspondence to: Ashin Mukherjee (ashinm@umich.edu)

Another dimensionality reduction approach called *Reduced Rank Regression* [4–7] minimizes the least squares criterion subject to the constraint $\text{rank}(\mathbf{B}) \leq r$ for some $r \leq \min\{P, Q\}$. This problem can also be motivated from latent variable regression, where we assume that the Q responses are functions of r underlying latent variables. The solution to the reduced rank regression problem involves projection of the usual OLS estimator to a r-dimensional space that explains the maximum variation in terms of the Frobenius norm. As OLS estimator performs poorly when the predictor variables are highly correlated the performance of the reduced rank estimator is also affected when the predictors are collinear.

Yuan et al. [8] proposed a novel dimension reduction method called *Factor Estimation and Selection* (FES). They try to minimize the constrained least squares criterion,

$$\hat{\mathbf{B}} = \arg\min_{\mathbf{B}} \|\mathbf{Y} - \mathbf{XB}\|_F^2 + \lambda \sum_{j=1}^{\min\{P,Q\}} \sigma_j(\mathbf{B}), \qquad (3)$$

where $\sigma_j(\mathbf{B})$ denotes the jth 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 Eq. (3) is shown to be equivalent to a second 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 [9]. Unlike reduced rank regression solution this provides a continuous regularization path. But as with the reduced rank regression this method also fails to account for the correlation among the predictor variables. The situations where the singular values of $\hat{\mathbf{B}}_{\text{OLS}}$ is a poor approximation to σ_j the FES method may suffer heavily.

To directly exploit the correlation structure between the response variables [10] proposed a method they call the *Curds and Whey* (CW) procedure. The main idea is to borrow strength from the separate OLS regressions by performing a second round of regression of the responses on the OLS estimates. Intuitively if some responses are heavily correlated then we will be able to obtain a better, more stable predictor by averaging over the corresponding OLS estimates. Notationally, CW predictor takes the form $\tilde{\mathbf{Y}} = \hat{\mathbf{Y}}_{\text{OLS}}\mathbf{M}$, where $\mathbf{M}$ is a $Q \times Q$ shrinkage matrix obtained by the second round of regression. The authors show that the CW procedure has some close connections to the canonical covariate analysis, they also develop an easy to implement GCV type criterion to efficiently perform cross-validatory shrinkage.

Several other penalization approaches have been proposed to improve the performance of least square estimates. Most commonly studied are Ridge regression [11] and LASSO regression [12] in the univariate situation, that is, $Q = 1$. LASSO is used as a tool for variables selection, specifically suited to the case where the number of predictors p is large but only a few of them actually have some effect on the response variables, more commonly known as the sparse set-up. Ridge regression introduces an ℓ_2 penalty, thus it performs shrinkage to handle the issues caused by collinearity in the predictor variables, rather than dimension reduction. Zou and Hastie [13] proposed the *Elastic Net* another variable selection method which combines the ℓ_1 and ℓ_2 penalties in an effort to utilize the favorable properties of the LASSO and the Ridge at the same time. Elastic Net achieves dimension reduction while controlling for the correlated predictors thus enjoys a grouping property which is useful in many real life scenario. Turlach et al. [14] proposed the ℓ_∞ penalty on the rows of $\mathbf{B}$ to enhance simultaneous variable selection. The method is recommended for model identification rather than prediction because of the bias induced due to the ℓ_∞ penalty. Peng et al. [15] proposed a joint constraint function of the form $C(\mathbf{B}) = \lambda_1 \sum_{p=1}^{P} \sum_{q=1}^{Q} |\beta_{pq}| + \lambda_2 \sum_{p=1}^{P} \|\mathbf{B}_p\|_2$ for the identification of *Master Predictors*. The first penalty encourages sparsity in $\mathbf{B}$ whereas the second penalty shrinks some of the entire rows of $\mathbf{B}$ to 0 thus enhancing the selection of the *Master Predictors*. The model is shown to outperform separate LASSO regressions and leads to highly interpretable estimated models in cancer studies. But this model is not exactly designed for the situation where our underlying assumption is that the Q responses actually live in a lower dimensional space.

In this paper we propose a procedure that combines some of the strengths of the estimators discussed above. The underlying assumption is that the true model is rank deficient, that is, $\text{rank}(\mathbf{B}) \leq \min\{P, Q\}$. Thus the response matrix would approximately be of low rank. Here it is important to note that the response matrix can have approximately low rank when the predictor matrix $\mathbf{X}$ is highly collinear even if the true coefficient matrix $\mathbf{B}$ is of full rank. 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 paper is organized as follows: In Section 2 we formally introduce the reduced rank ridge regression model and discuss some of the finer details. Section 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 setting in Section 4, and show a real data application. Section 5 concludes with a summary and brief discussion.

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 ensures 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,

$$\hat{\mathbf{B}}(\lambda, r) = \underset{\{\mathbf{B}:\text{rank}(\mathbf{B})\le r\}}{\arg\min} \|\mathbf{Y} - \mathbf{X}\mathbf{B}\|_F^2 + \lambda\|\mathbf{B}\|_F^2, \qquad (4)$$

where $r \le \min\{P, Q\}$. $\|\cdot\|_F^2$ 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,

$$\mathbf{X}^*_{(N+P)\times P} = \begin{pmatrix} \mathbf{X} \\ \sqrt{\lambda}\mathbf{I} \end{pmatrix}, \qquad \mathbf{Y}^*_{(N+P)\times Q} = \begin{pmatrix} \mathbf{Y} \\ \mathbf{0} \end{pmatrix}. \qquad (5)$$

Then it is a matter of simple algebra to notice that the minimization problem in Eq. (4) is equivalent to the following reduced rank regression problem:

$$\hat{\mathbf{B}}(\lambda, r) = \underset{\{\mathbf{B}:\text{rank}(\mathbf{B})\le r\}}{\arg\min} \|\mathbf{Y}^* - \mathbf{X}^*\mathbf{B}\|_F^2. \qquad (6)$$

Now we can use the orthogonal projection property of the OLS estimate 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 the same as the OLS estimate obtained from the linear model $\mathbf{Y}^* = \mathbf{X}^*\mathbf{B} + \mathbf{E}^*$. Note that the first term does not involve $\mathbf{B}$ hence we get the following equivalent form for the minimization problem (6) as,

$$\hat{\mathbf{B}}(\lambda, r) = \underset{\{\mathbf{B}:\text{rank}(\mathbf{B})\le r\}}{\arg\min} \|\hat{\mathbf{Y}}^*_R - \mathbf{X}^*\mathbf{B}\|_F^2. \qquad (7)$$

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

$$\hat{\mathbf{Y}}^*_r = \sum_{i=1}^{r} \sigma_i u_i v_i^{\mathrm{T}}. \qquad (8)$$

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

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

Hence we are able to show that the proposed solution $\hat{\mathbf{B}}(\lambda, r) = \hat{\mathbf{B}}^*_R\mathbf{P}_r$ is the minimizer of the optimization problem (4), 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:

$$\hat{\mathbf{B}}(\lambda, r) = \hat{\mathbf{B}}^*_R\mathbf{P}_r = \left(\mathbf{X}^{*\mathrm{T}}\mathbf{X}^*\right)^{-1}\mathbf{X}^{*\mathrm{T}}\mathbf{Y}\mathbf{P}_r$$
$$= \left(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y}\mathbf{P}_r, \qquad (9)$$

$$\hat{\mathbf{Y}}(\lambda, r) = \mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y}\mathbf{P}_r = \hat{\mathbf{Y}}_\lambda\mathbf{P}_r. \qquad (10)$$

$\hat{\mathbf{Y}}_\lambda$ 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.

2.1. 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,

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

The first two 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 $\mathbf{Y}^{\mathrm{T}}\mathbf{Y}$ 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 correlation among predictors and that is precisely where Reduced Rank Ridge regression improves by adding ridge penalty.

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 $\{1, 2, \ldots \min\{P, Q\}\}$. 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,

$$(\hat{\lambda}, \hat{r}) = \arg\min_{\lambda, r} \sum_{k=1}^{K} \|\mathbf{Y}^{(k)} - \mathbf{X}^{(k)}\hat{\mathbf{B}}^{(-k)}(\lambda, r)\|_F^2, \qquad (11)$$

where $\mathbf{X}^{(k)}$ and $\mathbf{Y}^{(k)}$ denote the predictor and response matrix for the kth fold, and $\hat{\mathbf{B}}^{(-k)}(\lambda, r)$ denotes the estimated regression coefficient matrix computed leaving out the observations in the kth 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.

3. NUMERICAL EXAMPLES

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 square* (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 Square* (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 [10]. The model error of an estimate $\hat{\mathbf{B}}$ is given by,

$$\mathrm{ME}(\hat{\mathbf{B}}) = \mathrm{trace}\left[(\mathbf{B} - \hat{\mathbf{B}})^{\mathrm{T}}\mathbf{\Sigma}_X(\mathbf{B} - \hat{\mathbf{B}})\right], \qquad (12)$$

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

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 three 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 (Fig. 1).

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 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

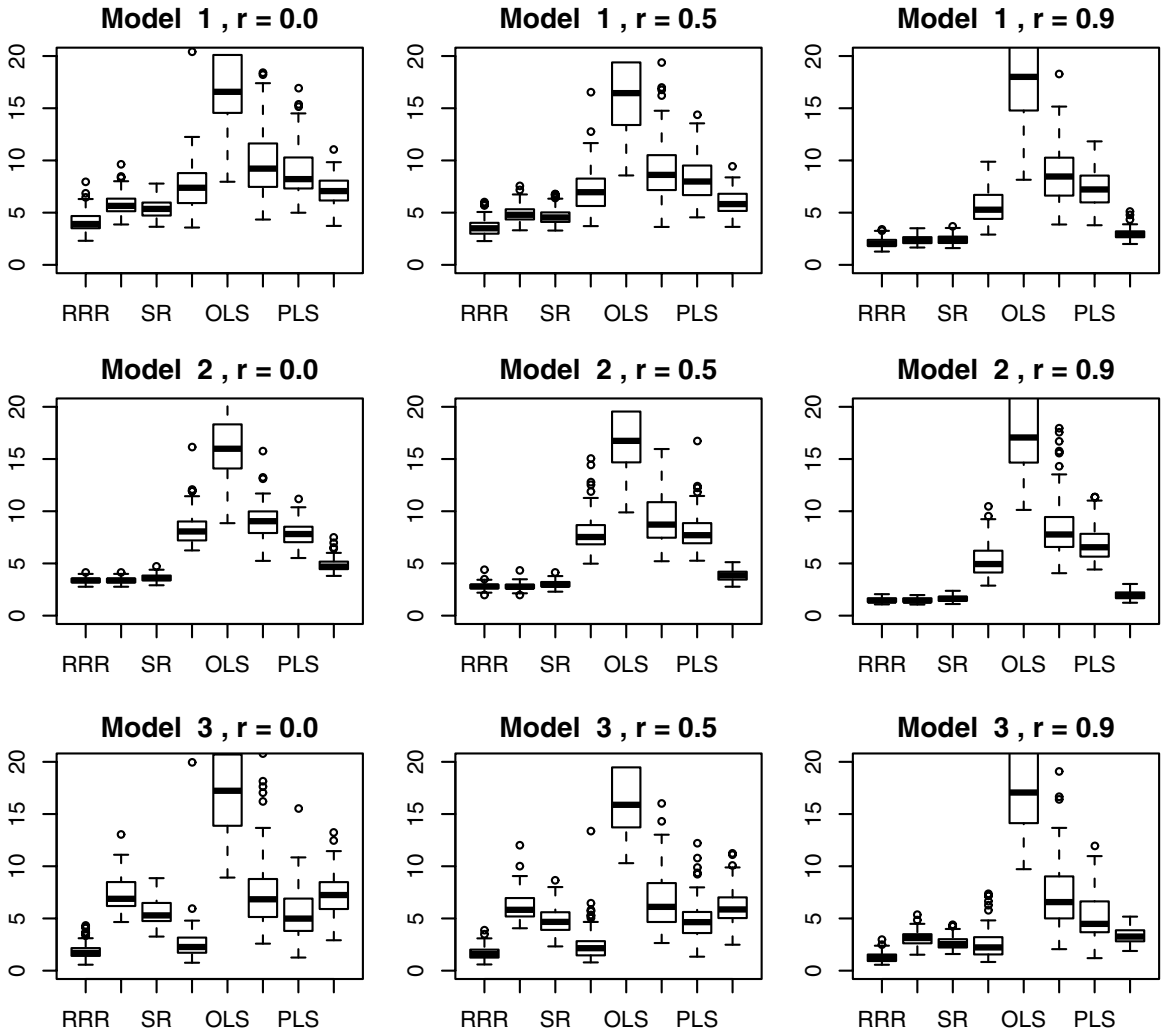

Fig. 1 Boxplot of ME of each method over 100 replicates under each combination of settings.

the singular values nonzero. 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 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 (Figs 2 and 3).

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.

3.2. Application in Chemometrics Example

It is originally from ref. [16]. 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 consist 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

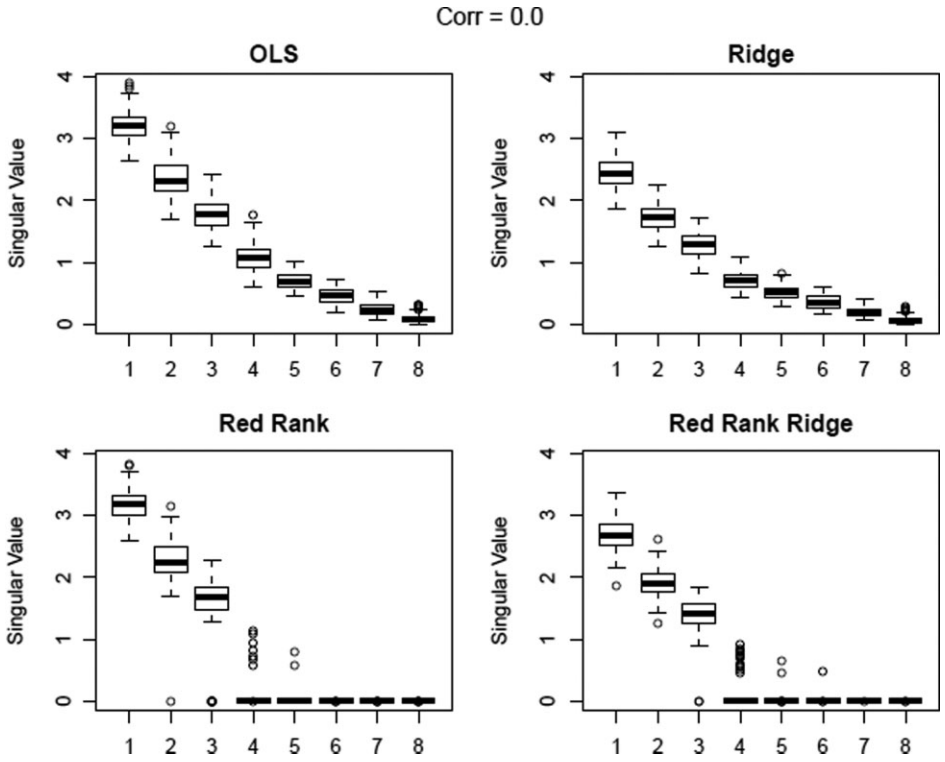

Fig. 2 Singular values of $\hat{\boldsymbol{B}}$, $\rho = 0.0$.

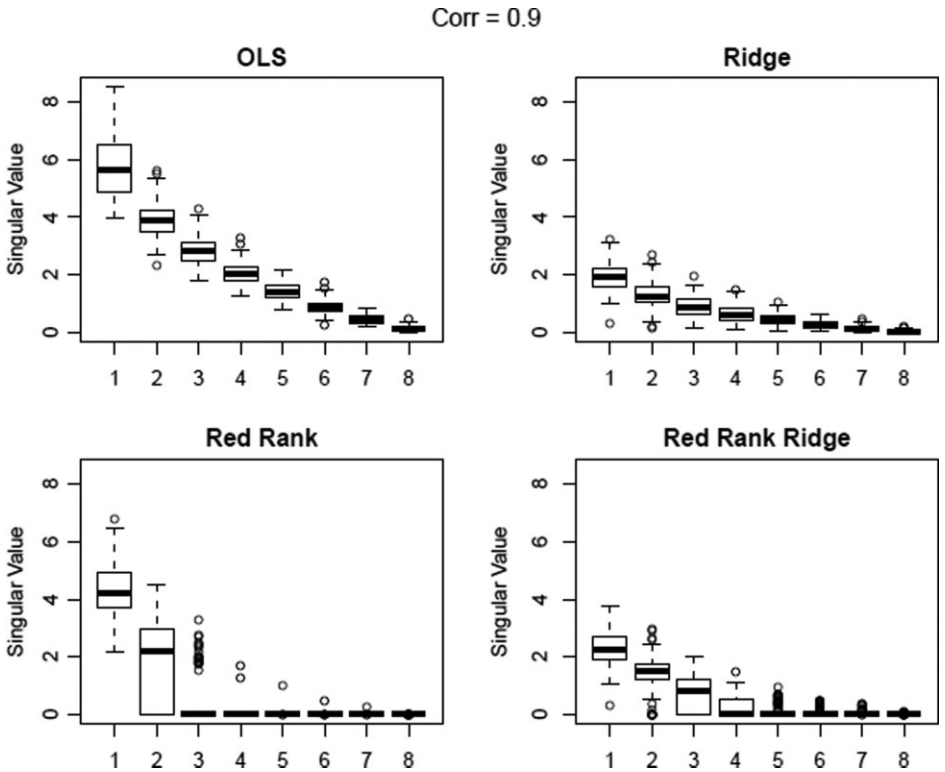

Fig. 3 Singular values of $\hat{\boldsymbol{B}}$, $\rho = 0.9$.

standardized them. The response correlation matrix is as follows:

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

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 (Table 1).

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 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.

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.

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 $\langle x, y\rangle = x^{\mathrm{T}}y$, $\mathbb{L}^2$-space of all square functions that can integrate on $\mathbb{R}$ with $\langle f, g\rangle = \int_{\mathbb{R}} f(x)g(x)\mathrm{d}x$. 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 nonparametric way. $\mathbb{L}_2$ is *too big* for our purpose as it contains too many nonsmooth functions. One way to obtain such spaces of smooth functions which allows us to fit a nonparametric 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 well known [17] 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.

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,

$$J_\lambda(f) = \sum_{i=1}^N (y_i - f(x_i))^2 + \lambda \|f\|_{\mathcal{H}}^2, \tag{13}$$

where $\|\cdot\|_{\mathcal{H}}^2$ 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 (13) can be written as,

$$f(x) = \sum_{i=1}^N \alpha_i K(x, x_i), \quad \text{for } (\alpha_1, \alpha_2, \ldots, \alpha_N) \in \mathbb{R}^N. \tag{14}$$

Table 1. Performance comparison for the Chemometrics data.

	OLS	CW-gcv	PLS	RR	MVR	RRR
Y1	0.49	0.49	0.68	0.44	0.15	0.15
Y2	1.12	0.74	0.90	0.46	0.22	0.22
Y3	0.53	0.49	0.45	0.65	0.39	0.39
Y4	0.24	0.18	0.18	0.14	0.26	0.24
Y5	0.30	0.22	0.26	0.18	0.28	0.27
Y6	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

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

$$J_\lambda(f_1, f_2, \ldots, f_Q) = \sum_{q=1}^{Q}\sum_{i=1}^{N}\|y_{iq} - f_q(x_i)\|^2 + \lambda\sum_{q=1}^{Q}\|f_q\|^2_{\mathcal{H}}. \tag{15}$$

Like in the linear case it is fairly easy to see that in absence of any constraint on the functions $\left(f_1, f_2, \ldots, f_Q\right)$ 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 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,

$$\min_{f_1, f_2, \ldots, f_Q \in \mathcal{H}} J_\lambda\left(f_1, f_2, \ldots, f_Q\right) \quad \text{subject to}$$
$$\dim\left(\text{span}\{f_1, f_2, \ldots, f_Q\}\right) \le r, \tag{16}$$

where $1 \le r \le Q$ and $J_\lambda\left(f_1, f_2, \ldots, f_Q\right)$ is defined as in Eq. (15). The solution has the following representation,

$$f_q(x) = \sum_{i=1}^{N}\alpha_{iq}K(x, x_i), \quad \text{for } q = 1, 2, \ldots, Q, \quad \alpha_{iq} \in \mathbb{R}. \tag{17}$$

The constraint $\dim\left(\text{span}\{f_1, \ldots, f_Q\}\right) \le r$ 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 of Eq. (16) becomes equivalent to a rank constraint on the coefficient matrix $\mathbf{A} = \left[\alpha_{iq}\right]_{N\times Q}$, because that would allow us to extend the reduced rank ridge regression solution developed in Section 2 in a natural way to the kernel setting.

PROPOSITION 2: If $K(\cdot,\cdot)$ is strictly positive definite and $\{x_1, x_2, \ldots x_N\}$ are distinct then

$$\dim\left(\text{span}\{f_1, f_2, \ldots f_Q\}\right) \le r \Rightarrow \text{rank}(\mathbf{A}) \le r,$$
$$\text{where, } \left[f_1, \ldots f_Q\right] = \left[K(\cdot, x_1), \ldots K(\cdot, x_N)\right]\underset{N\times Q}{\mathbf{A}}.$$

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 multiquadratic 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.

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. For a given point $x \in \mathbb{R}^Q$ (row vector) prediction had the form,

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

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

$$\mathbf{Y}^\text{T}\mathbf{X}\left(\mathbf{X}^\text{T}\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\mathbf{X}^\text{T}\mathbf{Y} = \mathbf{Y}^\text{T}\mathbf{K}\left(\mathbf{K} + \lambda\mathbf{I}\right)^{-1}\mathbf{Y} \tag{18}$$

$$x\left(\mathbf{X}^\text{T}\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\mathbf{X}^\text{T}\mathbf{Y} = K(x)\left(\mathbf{K} + \lambda\mathbf{I}\right)^{-1}\mathbf{Y}. \tag{19}$$

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 $\mathbf{Y}^\text{T}\mathbf{K}\left(\mathbf{K} + \lambda\mathbf{I}\right)^{-1}\mathbf{Y}$ by $\mathbf{P}_r^K$ then the final prediction for the point $x \in \mathbb{R}^P$ is given by,

$$\hat{Y}_x(\lambda, r) = K(x)\left(\mathbf{K} + \lambda\mathbf{I}\right)^{-1}\mathbf{Y}\mathbf{P}_r^K, \tag{20}$$

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

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

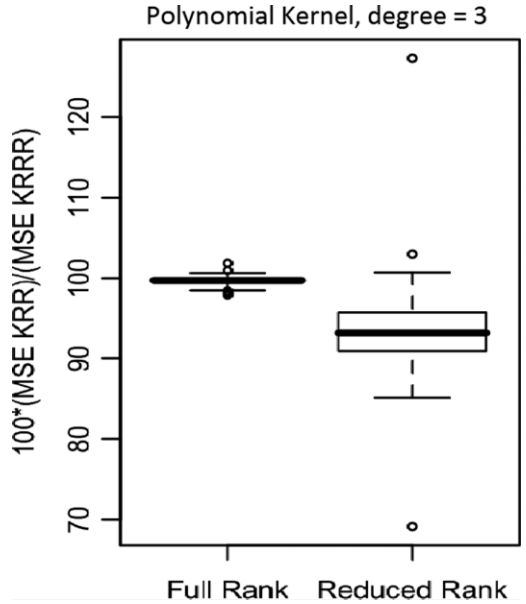

Fig. 4 Polynomial kernel, % of MSE compared to kernel ridge regression.

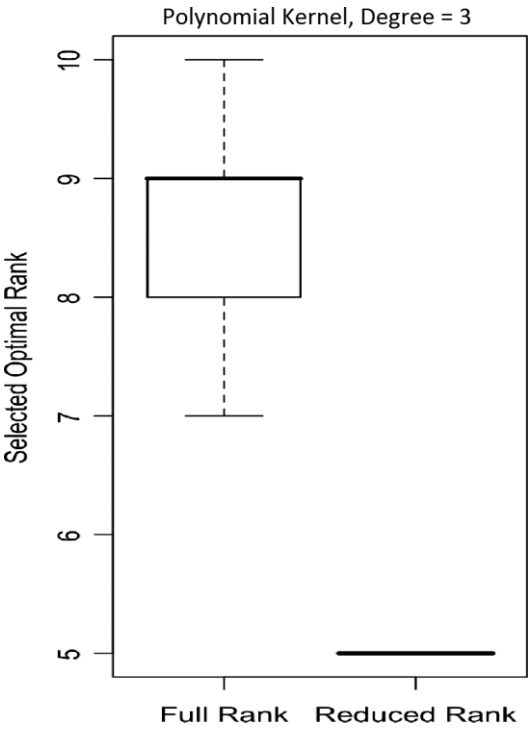

Fig. 5 Polynomial kernel, Box plot of the estimated rank over 100 replications.

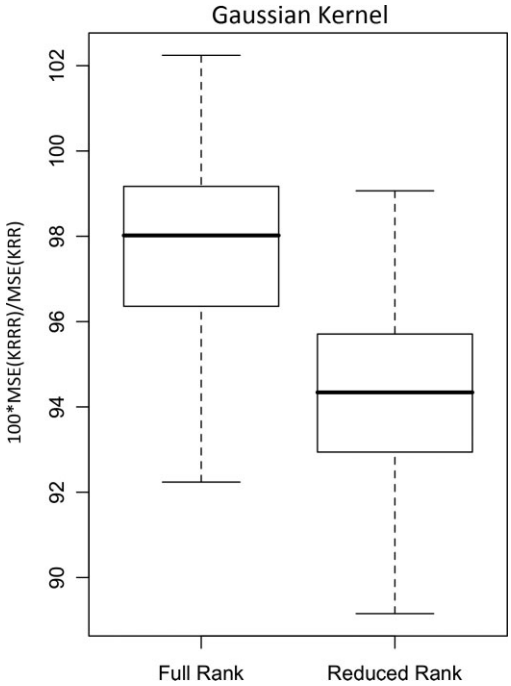

Fig. 6 Gaussian kernel, % of MSE compared to kernel ridge regression.

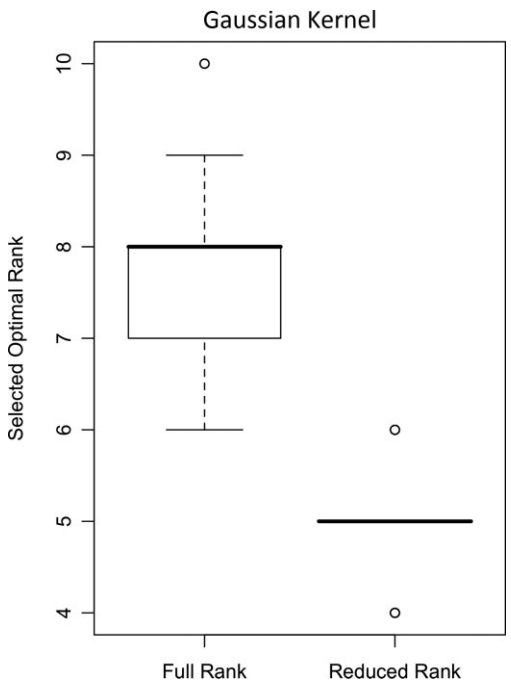

Fig. 7 Gaussian kernel, Box plot of the estimated rank over 100 replications.

Proposition 2 and the polynomial kernel which is clearly finite-dimensional and hence does not satisfy the sufficient condition provided in Proposition 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 $K(\cdot, b_j)$ where $\{b_j : j = 1, 2, \ldots, 10\}$ were generated

independently from a multivariate Gaussian distribution. We consider two cases,

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

The tuning parameters, that is, (λ, r, σ) were chosen using independently generated validation data sets of same size. In Figs 4–7 we present the box-plots of the percentage ratio of MSE of kernel RRR and kernel Ridge Regression over 100 replications of the experiment.

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 2. Also the estimated optimal rank seems 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.

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 multiquadratic kernel $K(x, x') = \frac{1}{\sqrt{\|x-y\|^2+c}}$. Both predictors and responses were standardized for this analysis. An eight 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 multiquadratic kernel (Table 2).

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 three 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.

5. SUMMARY AND DISCUSSION

We propose Reduced Rank Ridge Regression to produce a low-rank estimator of the regression coefficient matrix **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 **X** and low-rank of the true coefficient matrix **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

Table 2. Performance comparison for Kernel RRR and Kernel RR for the Chemometrics data with Gaussian and Inverse multi-quadratic kernels.

		Gaussian Kernel		Inverse multiquadratic Kernel	
	Linear RRR	Kernel Ridge	Kernel RRR	Kernel Ridge	Kernel RRR
Y1	0.153	0.088	0.087	0.111	0.120
Y2	0.250	0.148	0.129	0.224	0.210
Y3	0.230	0.113	0.111	0.160	0.161
Y4	0.188	0.054	0.044	0.098	0.094
Y5	0.205	0.107	0.071	0.125	0.101
Y6	0.211	0.070	0.064	0.092	0.097
Avg.	0.206	0.097	0.084	0.135	0.131

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 help us understand some statistical insights into the working of the proposed Reduced Rank Ridge Regression method.

APPENDIX

Proof of Proposition 1: Let $\left(f_1, f_2, \ldots, f_Q\right)$ be the minimizer to Eq. (16). Define,

$$\mathcal{F}_K = \text{span}\{K(\cdot, x_i) : i = 1, 2, \ldots N\}. \tag{21}$$

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

$$f_q(x_i) = \langle f_q^* + f_q^0, K(\cdot, x_i)\rangle = f_q^*(x_i),$$
$$\|f_q\|_{\mathcal{H}}^2 = \|f_q^*\|_{\mathcal{H}}^2 + \|f_q^0\|_{\mathcal{H}}^2.$$

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

$$f_j(x) = \sum_{i=1}^{N} \alpha_{iq} K(x, x_i), \quad \text{for } q = 1, 2, \ldots Q, \quad \alpha_{iq} \in \mathbb{R}. \tag{22}$$

Proof of Proposition 2: If $r = Q$ then the result holds vacuously. If $r < Q$ then $\exists$ nontrivial linear combinations $\sum_{q=1}^{Q} c_q f_q(\cdot) \equiv 0$. Equivalently, we have, $\|\sum_{q=1}^{Q} c_q f_q(\cdot)\|_{\mathcal{H}}^2 = 0$:

$$\left\| \sum_{q=1}^{Q} c_q f_q(\cdot) \right\|_{\mathcal{H}}^2 = 0 \Leftrightarrow c_{Q\times 1}^{\mathrm{T}} \mathbf{A}^{\mathrm{T}} \left[(K(x_i, x_{i'})\right]_{i,i'=1}^{N} \mathbf{A} c_{Q\times 1} = 0.$$

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, $T(c) = \sum_{q=1}^{Q} c_q f_q(\cdot)$ Then using the *Rank-Nullity Theorem* and the previous part,

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

REFERENCES

[1] W. Massy, Principal components regression with exploratory statistical research, J Am Stat Assoc 60 (1965), 234–246.

[2] H. Wold, Soft modeling by latent variables: the non-linear iterative partial least squares approach, In Perspect Prob Stat, papers in Honor of M.S. Bartlett, S. Gani, ed. New York Academic Press, 1975.

[3] H. Hotelling, The most predictable criterion, J Ed Psychol 26 (1935), 139–142.

[4] T. Anderson, Estimating linear restrictions on regression coefficients for multivariate normal distributions, Ann Math Stat 22 (1951), 327–351.

[5] A. Izenman, Reduced-rank regression for the multivariate linear model, J Multivariate Anal 5(2) (1975), 248–264.

[6] P. Davies, and M. Tso, Procedures for reduced-rank regression, Appl Stat 31(3) (1982), 244–255.

[7] G. Reinsel, and R. Velu, Multivariate Reduced-Rank Regression: Theory and Applications, New York, Springer, 1998.

[8] M. Yuan, A. Ekici, Z. Lu, and R. Monteiro, Dimension reduction and coefficient estimation in multivariate linear regression, J R Stat Soc B 69(3) (2007), 329–346.

[9] R. Tutuncu, K. Toh, and M. Todd, Solving semidefinite-quadratic-linear programs using SDPT3, Math Program 95 (2003), 189–217.

[10] L. Breiman, and J. Friedman, Predicting multivariate responses in multiple linear regression, J R Stat Soc B 59 (1997), 3–37.

[11] A. Hoerl, and R. Kennard, Ridge regression: biased estimation for non-orthogonal problems, Technometrics 8 (1970), 27–51.

[12] R. Tibshirani, Regression shrinkage and selection via the Lasso, J R Stat Soc B 58 (1996), 267–288.

[13] H. Zou, and T. Hastie, Regularization and variable selection via the elastic Neta, J R Stat Soc B 67 (2005), 301–320.

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

[15] J. Peng, J. Zhu, A. Bergamaschi, W. Han, D. Noh, J. Pollack, and P. Wang, Regularized multivariate regression for identifying master prediction with application to intergrative genomics study of breast cancer, Ann Appl Stat 4(1) (2009), 53–77.

[16] B. Skagerberg, J. MacGregor, and C. Kiparissdes, Multivariate data analysis applied to low density polyethylene reactors, Chem Intell Lab Syst 14 (1992), 341–356.

[17] G. Wahba, Spline models for observation data, Soc Ind Appl Math 59 (1990), 1–171.