Comparison of Multi-response Prediction Methods

March 5, 2019

While data science is battling to extract information from the enormous explosion of data, many estimators and algorithms are being developed for better prediction. Researchers and data scientists often introduce new methods and evaluate them based on various aspects of data. However, studies on the impact of/on model with multiple response variables are limited. This study compares some newly-developed (envelope) and well-established (PLS, PCR) prediction methods based on real data and simulated data specifically designed by varying properties such as multicollinearity, correlation between multiple responses and position of relevant principal components of predictors. This study aims to give some insight into these methods and help researchers to understand and use them in further studies.

# Introduction

Prediction has been an essential component of modern data science, whether in the discipline of statistical analysis or machine learning. Modern technology has facilitated a massive explosion of data; however, such data often contain irrelevant information that consequently makes prediction difficult. Researchers are devising new methods and algorithms in order to extract information to create robust predictive models. Such models mostly contain predictor variables that are directly or indirectly correlated with other predictor variables. In addition, studies often consist of many response variables correlated with each other. These interlinked relationships influence any study, whether it is predictive modelling or inference.

Modern inter-disciplinary research fields such as chemometrics, econometrics and bioinformatics handle multi-response models extensively. This paper attempts to compare some multivariate prediction methods based on their prediction performance on linear model data with specific properties. The properties include correlation between response variables, correlation between predictor variables, number of predictor variables and the position of relevant predictor components. These properties are discussed more fully in the [Experimental Design](#experimental-design) section. Among others, Sæbø, Almøy and Helland ([2015](#ref-saebo2015simrel)) and Almøy ([1996](#ref-Alm_y_1996)) have conducted a similar comparison in the single response setting. In addition, Rimal, Almøy and Sæbø ([2018](#ref-Rimal2018)) have also conducted a basic comparison of some prediction methods and their interaction with the data properties of a multi-response model. The main aim of this paper is to present a comprehensive comparison of contemporary prediction methods such as simultaneous envelope estimation (Senv) (R. Dennis Cook and Zhang [2015](#ref-cook2015simultaneous)) and envelope estimation in predictor space (Xenv) (R Dennis Cook, Li, and Chiaromonte [2010](#ref-cook2010envelope)) with customary prediction methods such as Principal Component Regression (PCR), Partial Least Squares Regression (PLS) using simulated dataset with controlled properties. In the case of PLS, we have used PLS1 which fits individual response separately and PLS2 which fits all the responses together. An experimental design and the methods under comparison are discussed further, followed by a brief discussion of the strategy behind the data simulation.

# Simulation Model

Consider a model where the response vector with elements and predictor vector with elements follow a multivariate normal distribution as follows,

where, and are the variance-covariance matrices of and , respectively, is the covariance between and and and are mean vectors of and , respectively. A linear model based on (1) is,

where, is a matrix of regression coefficients and is an error term such that . Here, and

In a model like (2), we assume that the variation in response is partly explained by the predictor . However, in many situations, only a subspace of the predictor space is relevant for the variation in the response . This space can be referred to as the relevant space of and the rest as irrelevant space. In the similar way, for a certain model, we can assume that a subspace in the response space exists and contains the information that the relevant space in predictor can explain (Figure-1). R. Dennis Cook, Li, and Chiaromonte ([2010](#ref-cook2010envelope)) and R. Dennis Cook and Zhang ([2015](#ref-cook2015simultaneous)) have referred to the relevant space as material space, and the irrelevant space as immaterial space.

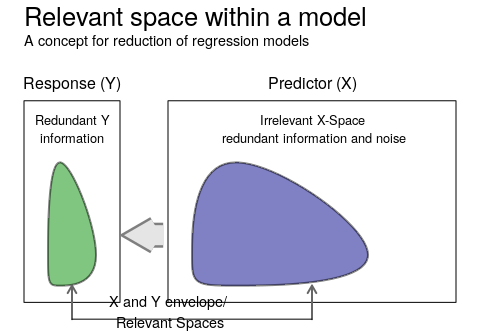


Figure 1 Relevant space in a regression model

With an orthogonal transformation of and to latent variables and , respectively, by and , where and are orthogonal rotation matrices, an equivalent model to (1) in terms of the latent variables can be written as,

where, and are the variance-covariance matrices of and , respectively. is the covariance between and . and are the mean vector of and respectively.

Here, the elements of and are the principal components of responses and predictors, which will be referred to respectively as “response components” and “predictor components”. The column vectors of respective rotation matrices and are the eigenvectors corresponding to these principal components. We can write a linear model based on (3) as,

where, is a matrix of regression coefficients and is an error term such that .

Following the concept of relevant space, a subset of predictor components can be imagined to span the predictor space. These components can be regarded as relevant predictor components. Naes and Martens ([1985](#ref-Naes1985)) introduced the concept of relevant components which was explored further by Helland ([1990](#ref-helland1990partial)), Næs and Helland ([1993](#ref-naes1993relevant)), Helland and Almøy ([1994](#ref-Helland1994b)) and Helland ([2000](#ref-Helland2000)). The corresponding eigenvectors were referred to as relevant eigenvectors. A similar logic is introduced by R. Dennis Cook, Li, and Chiaromonte ([2010](#ref-cook2010envelope)) and later by R. D. Cook, Helland and Su ([2013](#ref-cook2013envelopes)) as an envelope which is the space spanned by the relevant eigenvectors (R. Dennis Cook [2018](#ref-cook2018envelope), 101).

In addition, various simulation studies have been performed with the model based on the concept of relevant subspace. A simulation study by Almøy ([1996](#ref-Alm_y_1996)) has used a single response simulation model based on reduced regression and has compared some contemporary multivariate estimators. In recent years, Helland, Saebø and Tjelmeland ([2012](#ref-helland2012near)), Sæbø, Almøy and Helland ([2015](#ref-saebo2015simrel)), I. S. Helland et al. ([2018](#ref-helland2016algorithms)) and Rimal, Almøy and Sæbø ([2018](#ref-Rimal2018)) implemented simulation examples similar to those we are discussing in this study. This paper, however, presents an elaborate comparison of the prediction using multi-response simulated linear model data. The properties of the simulated data are varied through different levels of simulation parameters based on an experimental design. Rimal, Almøy and Sæbø ([2018](#ref-Rimal2018)) provide a detailed discussion of the simulation model that we have adopted here. The following section presents the estimators being compared in more detail.

# Prediction Methods

Partial least squares regression (PLS) and Principal component regression (PCR) have been used in many disciplines such as chemometrics, econometrics, bioinformatics and machine learning, where wide predictor matrices, i.e. (number or predictors) > (number of observation) are common. These methods are popular in multivariate analysis, especially for exploratory studies and predictions. In recent years, a concept of envelope introduced by R. Dennis Cook, Li and Chiaromonte ([2007](#ref-Cook2007a)) based on reduction in regression model was implemented for the development of different estimators. This study compares these prediction methods based on their prediction performance on data simulated with different controlled properties.

*Principal Components Regression (PCR):*

Principal components are the linear combinations of predictor variables such that the transformation makes the new variables uncorrelated. In addition the variation of the original dataset captured by the new variables are sorted in descending order. In other words, each successive component captures maximum variation left by the preceding components in predictor variables (Jolliffe [2002](#ref-Jolliffe2002)). Principal components regression uses these principal components as a new predictor to explain the variation in the response.

*Partial Least Squares (PLS):*

Two variants of PLS: PLS1 and PLS2 are used for comparison. The first one considers individual response variables separately, i.e. each response is predicted with a single response model, while the latter considers all response variables together. In PLS regression, the components are determined so as to maximize a covariance between response and predictors (Jong [1993](#ref-DeJong1993)). R-package pls (Mevik, Wehrens and Liland [2018](#ref-pls2018)) is used for both PCR and PLS methods.

*Envelopes:*

The envelope, introduced by R. Dennis Cook, Li and Chiaromonte ([2007](#ref-Cook2007a)), was first used to define response envelope (R Dennis Cook, Li and Chiaromonte [2010](#ref-cook2010envelope)) as the smallest subspace in the response space so that the span of regression coefficients lies in that space. Since a multivariate linear regression model contains relevant (material) and irrelevant (immaterial) variation in both response and predictor, the relevant part provides information, while the irrelevant part increases the estimative variation. The concept of the envelope uses the relevant part for estimation while excluding the irrelevant part, consequently increasing the efficiency of the model (R. Dennis Cook and Zhang [2016](#ref-cook2016algorithms)).

The concept was later extended to the predictor space, where the predictor envelope was defined (R. D. Cook, Helland and Su [2013](#ref-cook2013envelopes)). Further R. Dennis Cook and Zhang ([2015](#ref-cook2015simultaneous)) used envelopes for joint reduction of the responses and predictors and argued to produce efficiency gains greater than using individual envelops either of the response and predictors. All the variants of envelope estimations are based on maximum likelihood estimation. Here we have used predictor envelope (Xenv) and simultaneous envelope (Senv) for the comparison. R-package Renvlp (Lee and Su [2018](#ref-env2018)) is used for both Xenv and Senv methods.

## Modification in envelope estimation

Since envelope estimators (Xenv and Senv) are based on maximum likelihood estimation (MLE), it fails to estimate in the case of wide matrices, i.e. . To incorporate these methods in our comparison, we have used the principal components of the predictor variables as predictors, using the required number of components for capturing 97.5% of the variation in for the designs where . The new set of variables were used for envelope estimation. The regression coefficients corresponding to these new variables were transformed back to obtain coefficients for each predictor variable as,

where is a matrix of eigenvectors with first number of components. Since, only simultaneous envelope allows us to specify the dimension of response envelope which we have fixed at two in the simulation study. In the case of Senv, when the envelope dimension for response is the same as the number of responses, it degenerates to the Xenv method and if the envelope dimension for predictor is the same as the number of predictors, it degenerates to the standard multivariate linear regression (Lee and Su [2018](#ref-env2018)).

# Experimental Design

This study compares prediction methods based on their predictive ability. Data with specific properties are simulated, some of which are easier to predict than others. These data are simulated using the R-package simrel, which is discussed in Sæbø, Almøy and Helland ([2015](#ref-saebo2015simrel)) and Rimal, Almøy and Sæbø ([2018](#ref-Rimal2018)). Here we have used four different factors to vary the property of the data: a) number of predictors (p), b) multicollinearity in predictor variables (gamma), c) correlation in response variables (eta) and d) position of predictor components relevant for the response (relpos). Using two levels of p, gamma and relpos and four levels of eta, 32 set of distinct properties are designed for the simulation.

Number of predictors:

To observe the performance of the methods on tall and wide predictor matrices, 20 and 250 predictor variables are simulated with the number of observations fixed at 100. Parameter p controls these properties in the simrel function.

Multicollinearity in predictor variables:

Highly collinear predictors can be explained completely by few components. The parameter gamma () in simrel controls decline in the eigenvalues of the predictor variables as (5).

Here, are eigenvalues of the predictor variables. We have used 0.2 and 0.9 as different levels of gamma. The higher the value of gamma, the higher the multicollinearity will be, and vice versa.

Correlation in response variables:

Correlation among response variables has been explored to a lesser extent. Here we have tried to explore that part with four levels of correlation in the response variables. We have used the eta () parameter of simrel for controlling the decline in eigenvalues corresponding to the response variables as (6).

Here, are the eigenvalues of the response variables and m is the number of response variables. We have used 0, 0.4, 0.8 and 1.2 as different levels of eta. The larger the value of eta, the larger the correlation will be between response variables, and vice versa.

Position of predictor components relevant to the response:

The principal components of the predictors are ordered. The first principal component captures most of the variation in the predictors. The second captures the most in the rest that is left by the first principal components and so on. In highly collinear predictors, the variation captured by the first few components is relatively high. However, if those components are not relevant for the response, prediction becomes difficult (Helland and Almøy [1994](#ref-Helland1994b)). Here, two levels of the positions of these relevant components are used as 1, 2, 3, 4 and 5, 6, 7, 8.

Moreover, a complete factorial design from the levels of the above parameters gave us 32 designs. Each design is associated with a dataset having unique properties. Figure~2, shows all the designs. For each design and prediction method, 50 datasets were simulated as replicates. In total, there were , i.e. 8000 simulated datasets.

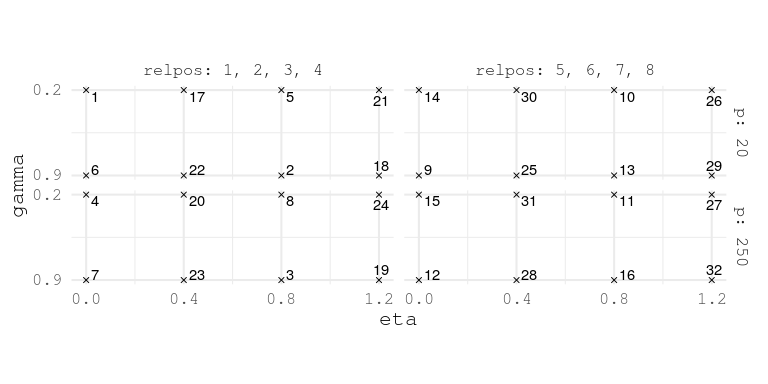


Figure 2 Experimental Design of simulation parameters. Each point represents a unique data property.

Common parameters:

Each dataset was simulated with number of observation and response variables. Furthermore, the coefficient of determination corresponding to each response component in all the designs is set to and 0.8. In addition, we have assumed that there is only one informative response component. Hence, the informative response component is rotated orthogonally together with three uninformative response components to generate four response variables. This spreads out the information in all simulated response variables. For further details on the simulation tool, see (Rimal, Almøy and Sæbø [2018](#ref-Rimal2018)).

An example of simulation parameters for the first design is as follows:

|  |  |  |
| --- | --- | --- |
| simrel( |  |  |
| n | = 100, | * Training samples |
| p | = 20, | * Predictors |
| m | = 4, | * Responses |
| q | = 20, | * Relevant predictors |
| relpos | = list(c(1, 2, 3, 4)), | * Relevant predictor components index |
| eta | = 0, | * Decay factor of response eigenvalues |
| gamma | = 0.2, | * Decay factor of predictor eigenvalues |
| R2 | = 0.8, | * Coefficient of determination |
| ypos | = list(c(1, 2, 3, 4)), |  |
| type | = "multivariate" |  |
| ) |  |  |

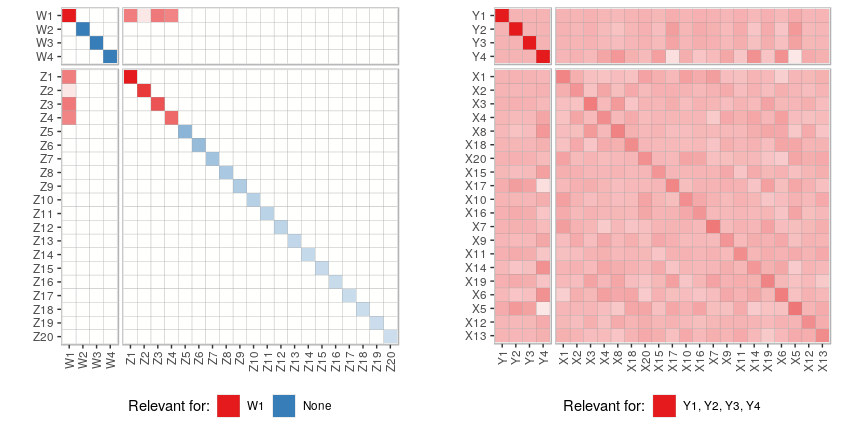


Figure 3 (left). Covariance structure of latent components. (right) Covariance structure of predictor and response

The covariance structure of the data simulated with this design in Figure 3 shows that the predictor components at positions 1, 2, 3 and 4 are relevant for the first response component. After the rotation with orthogonal rotation matrix, all predictor variables are somewhat relevant for all response variables, satisfying other desired properties such as multicollinearity and coefficient of determination. For the same design, Figure 4 (top left) shows that the predictor components 1, 2, 3 and 4 are relevant for the first response component. All other predictor components are irrelevant and all other response components are uninformative. However, due to orthogonal rotation of the informative response component together with uninformative response components, all response variables in the population have similar covariance with the relevant predictor components (Figure 4 (top right)). The sample covariances between the predictor components and predictor variables with response variables are shown in Figure 4 (bottom left) and (bottom right) respectively.

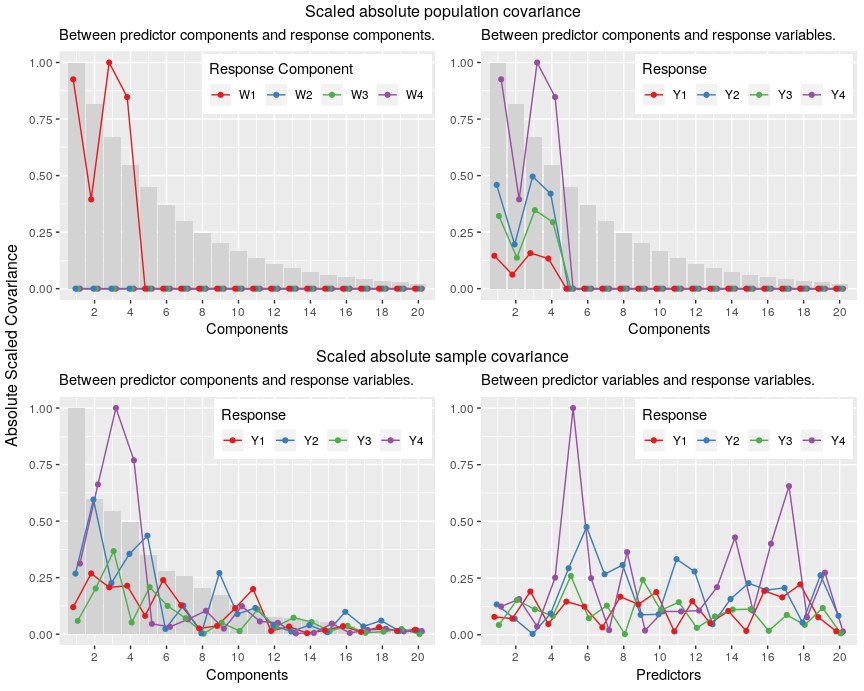


Figure 4 Expected Scaled absolute covariance between predictor components and response components (top left). Expected Scaled absolute covariance between predictor components and response variables (top right). Sample scaled absolute covariance between predictor components and response variables (bottom left). Sample scaled absolute covariance between predictor variables and response variables (bottom right). The bar graph in the background represents eigenvalues corresponding to each component in population (top plots) and in sample (bottom plots). One can compare the top-right plot (true covariance of the population) with bottom-left (covariance in the simulated data) which shows a similar pattern for different components.

A similar description can be made for all 32 designs, where each of the designs holds the properties of the data they simulate. These data are used by the prediction methods discussed in the previous section. Each prediction method is given independently simulated datasets in order to give them an equal opportunity to capture the dynamics in the data.

# Basis of comparison

This study focuses mainly on the prediction performance of the methods with emphasis specifically on the interaction between the properties of the data controlled by the simulation parameters, and the prediction methods. The prediction performance is measured based on the following:

1. The average prediction error that a method can give using arbitrary number of components and
2. The average number of components used by the method to give the minimum prediction error

Let us define,

as a prediction error of response for a given design and method using number of components. Here, is the true covariance matrix of the predictors, unique for a particular design and for response is the true model error. Here prediction error is scaled by the true model error to remove the effects of influencing residual variances. Since both the expectation and the variance of are unknown, the prediction error is estimated using data from 50 replications as follows,

where is the estimated prediction error averaged over replicates.

The following section focuses on the data for the estimation of these prediction errors that are used for the two models discussed above in a) and b) of this section.

# Data Preparation

A dataset for estimating (7) is obtained from simulation which contains a) five factors corresponding to simulation parameters, b) prediction methods, c) number of components, d) replications and e) prediction error for four responses. The prediction error is computed using predictor components ranging from 0 to 10 for each 50 replicates as,

Thus there are 32 (designs) 5 (methods) 11 (number of components) 50 (replications), i.e. 88000 observations corresponding to the response variables from Y1 to Y4.

Since our discussions focus on the average minimum prediction error that a method can obtain and the average number of components they use to get the minimum prediction error in each replicate, the dataset discussed above is summarized as constructing the following two smaller datasets. Let us call them *Error Dataset* and *Component Dataset*.

*Error Dataset*:

For each prediction method, design and response, an average prediction error is computed over all replicates for each component. Next, a component that gives the minimum of this average prediction error is selected, i.e.,

Using the component , a dataset of is used as the *Error Dataset*. Let for be the outcome variables measuring the prediction error corresponding to the response number in the context of this dataset.

*Component Dataset*:

The number of components that gives the minimum prediction error in each replication is referred to as the *Component Dataset*, i.e.,

Here is the number of components that gives minimum prediction error for design , response , method and replicate . Let for be the outcome variables measuring the number of components used for minimum prediction error corresponding to the response in the context of this dataset.

# Exploration

This section explores the variation in the *error dataset* and the *component dataset* for which we have used Principal Component Analysis (PCA). Let and be the principal component score sets corresponding to PCA run on the and matrices respectively. The scores density in Figure-5 corresponds to the first principal component of , i.e. the first column of .

Since higher prediction errors correspond to high scores, the plot shows that the PCR, PLS1 and PLS2 methods are influenced by the two levels of position of relevant predictor components. When the relevant predictors are at positions 5, 6, 7, 8, the eigenvalues corresponding to them are relatively smaller. This also suggests that PCR, PLS1 and PLS2 depend greatly on the position of the relevant components, and the variation of these components affects their prediction performance. However, the envelope methods appeared to be less influenced by relpos in this regard.

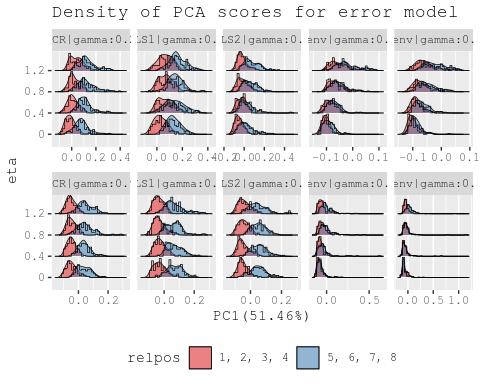


Figure 5 Scores density corresponding to first principal component of error dataset () subdivided by methods, gamma and eta and grouped by relpos.

In addition, the plot also shows that the effect of gamma, i.e., the level of multicollinearity, has a lesser effect when the relevant predictors are at positions 1, 2, 3, 4. This indicates that the methods are somewhat robust for handling collinear predictors. Nevertheless, when the relevant predictors are at positions 5, 6, 7, 8, high multicollinearity results in small variance of these relevant components and consequently yields poor prediction. This is in accordance with the findings of Helland and Almøy ([1994](#ref-Helland1994b)).

Furthermore, the density curves for PCR, PLS1 and PLS2 are similar for different levels of eta, i.e., the factor controlling the correlation between responses. However the envelope models have been shown to have distinct interactions between the positions of relevant components (relpos) and eta. Here higher levels of eta have yielded higher scores and clear separation between two levels of relpos.

In the case of high multicollinearity, envelope methods have resulted in some large outliers indicating in some cases that the methods can result in unexpected prediction.

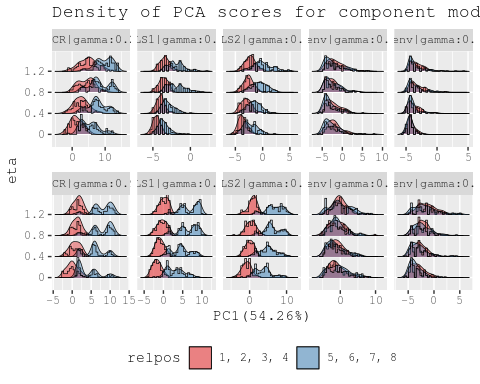


Figure 6 Score density corresponding to first principal component of component dataset () subdivided by methods, gamma and eta and grouped by relpos.

In Figure 6, the higher scores suggest that the methods have used a larger number of components to give minimum prediction error. The plot also shows that the relevant predictor components at 5, 6, 7, 8 give larger prediction errors than those in positions 1, 2, 3, 4. The pattern is more distinct in large multicollinearity cases and PCR and PLS methods. Both the envelope methods have shown equally enhanced performance at both levels of relpos and gamma. However, for data with low multicollinearity (), the envelope methods have used a lesser number of components on average than in the high multicollinearity cases to achieve minimum prediction error.

# Statistical Analysis

This section has modelled the *error data* and the *component data* as a function of the simulation parameters to better understand the connection between data properties and prediction methods using multivariate analysis of variation (MANOVA).

Let us consider a model with third order interaction of the simulation parameters (p, gamma, eta and relpos) and Methods as in (11) and (12) using datasets and , respectively. Let us refer them as the *error model* and the *component model*.

Error Model:

Component Model:

where, is a vector of prediction errors in the *error model* and is a vector of number of components used by a method to obtain minimum prediction error in the *component model*.

Although there are several test-statistics for MANOVA, all are essentially equivalent for large samples (Johnson and Wichern [2018](#ref-johnson2018applied)). Here we will use Pillai’s trace statistic which is defined as,

Here the matrix holds between-sum-of-squares and sum-of-products for each of the predictors. The matrix has a within sum of squares and sum of products for each of the predictors. represents the eigenvalues corresponding to (Rencher [2003](#ref-rencher2003methods)).

For both the models (11) and (12), Pillai’s trace statistic is used for accessing the effect of each factor and returns an F-value for the strength of their significance. Figure 7 plots the Pillai’s trace statistics as bars with corresponding F-values as text labels for both models.

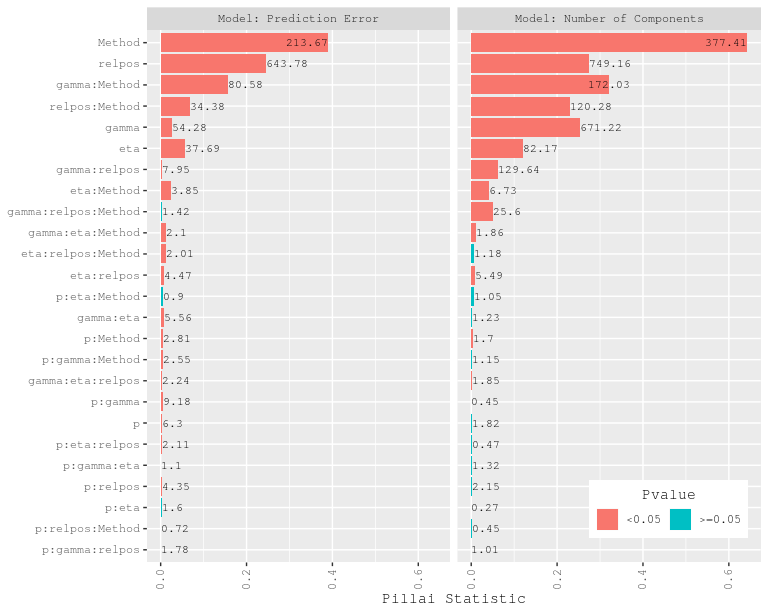


Figure 7 Pillai Statistic and F-value for the MANOVA model. The bar represents the Pillai Statistic and the text labels are F-value for corresponding factor.

Error Model:

Figure 7 (left) shows the Pillai’s trace statistic for factors of the *error model*. The main effect of Method followed by relpos, eta and gamma have largest influence on the model. A highly significant two-factor interaction of Method with eta followed by relpos and gamma clearly shows that methods perform differently for different levels of these data properties. The significant third order interaction between Method, eta and gamma suggests that the performance of a method differs for a given level of multicollinearity and the correlation between the responses. Since only some methods consider modelling predictor and response together, the prediction is affected by the level of correlation between the responses (eta) for a given method.

Component Model:

Figure 7 (right) shows the Pillai’s trace statistic for factors of the *component model*. As in the *error model*, the main effects of the Method, relpos, gamma and eta have significantly large effect on the number of components that a method has used to obtain minimum prediction error. The two factor interactions of Method with simulation parameters are larger in this case. This shows that the Methods and these interactions have larger effect on the use of number of component than the prediction error itself. In addition, a similar significant high third-order interaction as found in the *error model* is also observed in this model.

The following section will continue to explore the effects of different levels of the factors in the case of these interactions.

## Effect Analysis of Error Model

The large difference in the prediction error for the envelope models in Figure 8 (left) are intensified when the position of relevant predictor is at 5, 6, 7, 8. The results also show that the envelope methods are more sensitive to the levels of eta than the rest of the methods. In the case of PCR and PLS, the difference in the effect of levels of eta is small.

In Figure 8 (right), we can see that the multicollinearity (controlled by gamma) has affected all the methods. However, envelope methods have better performance on low multicollinearity, as opposed to high multicollinearity, and PCR, PLS1 and PLS2 are robust for high multicollinearity. Despite handling high multicollinearity, these methods have higher prediction error in both cases of multicollinearity than the envelope methods.

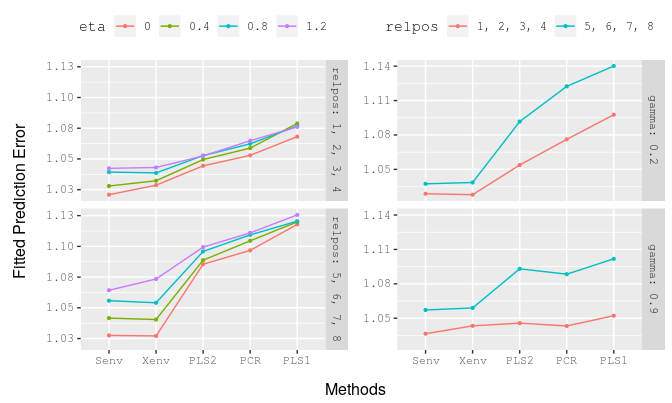


Figure 8 Effect plot of some interactions of the multivariate linear model of prediction error

## Effect Analysis of Component Model

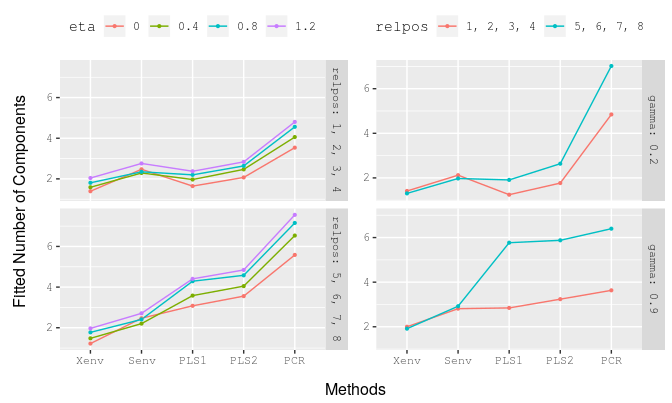


Figure 9 Effect plot of some interactions of the multivariate linear model of number of components to get minimum prediction error

Figure 9 (left) shows that the number of components used by the methods to obtain minimum prediction error is less affected by the levels of eta. All methods appear to use on average more components when eta increases. Envelope methods are able to obtain minimum prediction error by using components ranging from 1 to 3 in both the cases of relpos. This value is much higher in the case of PCR as its prediction is based only on the principal components of the predictor matrix. The number of components used by this method ranges from 3 to 5 when relevant components are at positions 1, 2, 3, 4 and 5 to 8 when relevant components are at positions 5, 6, 7, 8.

When relevant components are at position 5, 6, 7, 8, the eigenvalues of relevant predictors becomes smaller and responses are relatively difficult to predict. This becomes more critical for high multicollinearity cases. Figure 9 (right) shows that the envelope methods are less influenced by the level of relpos and are particularly better in achieving minimum prediction error using a fewer number of components than other methods.

# Examples

In addition to the analysis with the simulated data, the following two examples explore the prediction performance of the methods using real datasets. Since both examples have wide predictor matrices, principal components explaining 97.5% of the variation in them are used for envelope methods. The coefficients were transformed back after the estimation.

## Raman spectra analysis of contents of polyunsaturated fatty acids (PUFA)

This dataset contains 44 training samples and 25 test samples of fatty acid information expressed as: a) percentage of total sample weight and b) percentage of total fat content. The dataset is borrowed from Næs et al. ([2013](#ref-naes2013multi)) where more information can be found. The samples were analysed using Raman spectroscopy from which 1096 wavelength variables were obtained as predictors. Raman spectroscopy provides detailed chemical information from minor components in food. The aim of this example is to compare how well the prediction methods that we have considered are able to predict the contents of PUFA using these Raman spectra.

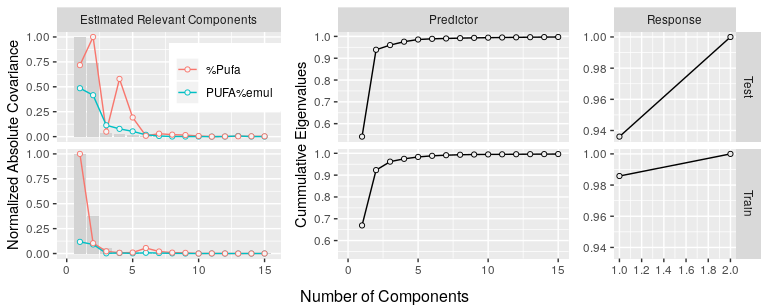


Figure 10 (Left) Bar represents the eigenvalues corresponding to Raman Spectra. The points and line are the covariance between response and the principal components of Raman Spectra. All the values are normalized to scale from 0 to 1. (Middle) Cumulative sum of eigenvalues corresponding to predictors. (Right) Cumulative sum of eigenvalues corresponding to responses. The top and bottom row corresponds to test and training datasets respectively.

Figure 10 (left) shows that the first few predictor components are somewhat correlated with response variables. In addition, the most variation in predictors is explained by less than five components (middle). Further, the response variables are highly correlated, suggesting that a single latent dimension explains most of the variation (right). We may therefore also believe that the relevant latent space in the response matrix is of dimension one. This resembles Design 19 (Figure 2) from our simulation.

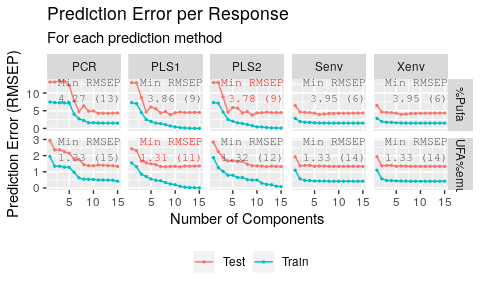


Figure 11 Prediction Error of different prediction methods using different number of components.

Using a range of components from 1 to 15, regression models were fitted using each of the methods. The fitted models were used to predict the test observation, and the root mean squared error of prediction (RMSEP) was calculated. Figure 11 shows that PLS2 obtained a minimum prediction error of 3.783 using 9 components in the case of response %Pufa, while PLS1 obtained a minimum prediction error of 1.308 using 11 components in the case of response PUFA%emul. However, the figure also shows that both envelope methods have reached almost minimum prediction error in fewer number of components. This pattern is also visible in the simulation results (Figure 9).

## Example-2: NIR spectra of biscuit dough

The dataset consists of 700 wavelengths of NIR spectra (1100–2498 nm in steps of 2 nm) that were used as predictor variables. There are four response variables corresponding to the yield percentages of (a) fat, (b) sucrose, (c) flour and (d) water. The measurements were taken from 40 training observations of biscuit dough. A separate set of 32 samples created and measured on different occasions were used as test observations. The dataset is borrowed from Indahl ([2005](#ref-indahl2005twist)) where further information can be obtained.

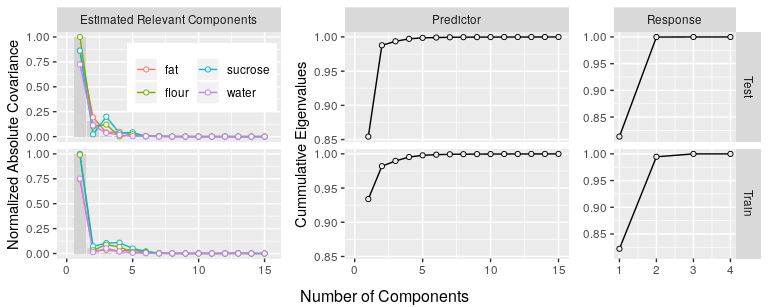


Figure 12 (Left) Bar represents the eigenvalues corresponding to NIR Spectra. The points and line are the covariance between response and the principal components of NIR Spectra. All the values are normalized to scale from 0 to 1. (Middle) Cumulative sum of eigenvalues corresponding to predictors. (Right) Cumulative sum of eigenvalues corresponding to responses.

Figure 12 (left) shows that the first predictor component has largest variance and also has large covariance with all response variables. The second component, however, has larger variance (middle) than the succeeding components but has a small covaration with all the responses, which indicates that the component is less relevant for any of the responses. In addition, two response components have explained most of the variation in response variables (right). This structure is also somewhat similar to Design 19, although it is uncertain whether the dimension of the relevant space in the response matrix is larger than one.

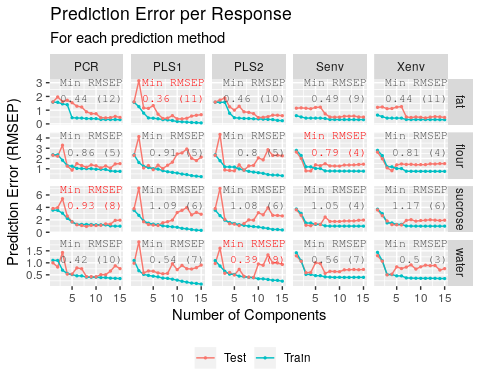


Figure 13 Prediction Error of different prediction methods using different number of components.

Figure 13 (corresponding to Figure 11) shows the root mean squared error for both test and train prediction of the biscuit dough data. Here four different methods have minimum test prediction error for the four responses. As the structure of the data is similar to that of the first example, the pattern in the prediction is also similar for all methods.

The prediction performance on the test data of the envelope methods appears to be more stable compared to the PCR and PLS methods. Furthermore, the envelope methods achieve good performance generally using fewer components, which is in accordance with Figure 6.

# Discussion and Conclusion

Analysis using both simulated data and real data has shown that the envelope methods are more stable, less influenced by relpos and gamma and in general perform better than PCR and PLS methods. These methods are also found to be less dependent on the number of components.

Since the facets in Figures 5 and 6 have their own scales, despite having some large prediction errors seen at the right tail, envelope methods still have a smaller prediction error and have used fewer number of components than the other methods.

Particularly in the case of simultaneous envelope, since users can specify the number of dimension for the response envelope, the method can leverage the relevant space of response while PCR, PLS and Xenv are constrained to play only on predictor space.

Furthermore, we have fixed the coefficient of determination () as a constant throughout all the designs. Initial simulations (not shown) indicated that low affects all methods in a similar manner and that the MANOVA is highly dominated by . Keeping the value of fixed has allowed us to analyse other factors properly.

Two clear comments can be made about the effect of correlation of response on the prediction methods. Highly correlated response has shown highest prediction error in general and the effect is most distinct in envelope methods. Since the envelope methods identifies the relevant space as the span of relevant eigenvectors, the methods are able to obtain the minimum average prediction error by using a lesser number of components for all levels of eta.

To our knowledge, the effect of correlation in the response on PCR and PLS methods has been explored only to a limited extent. In this regard, it is interesting to see that these methods have applied a large number of components and returned a larger prediction error than envelope methods in the case of highly correlated responses. To fully understand the effect of eta, it is necessary to study the estimation performance of these methods with different numbers of components.

In addition, since using principal components or actual variables as predictors in envelope methods has shown similar results, we have used principal components that have explained 97.5% of the variation as mentioned previously in the cases of envelope methods for the designs where . As the envelope methods are based on MLE, this can be an alternative way of using the methods in data with wide predictors. The results from this study will help researchers understand these methods for different nature of data. We encourage researchers to use newly developed methods such as the envelope with the nature of data they are working on.

Since, this study has focused entirely on prediction performance, a further analysis of the estimative properties of these methods is required. A study of estimation error and the behaviour of methods on non-optimal number of components may result in a deeper understanding of these methods.

A shiny application (Chang et al. [2018](#ref-shiny)) is available at <http://therimalaya.shinyapps.io/Comparison> where all the results related to this study can be visualized. In addition, a github repository at <https://github.com/therimalaya/03-prediction-comparison> can be used to reproduce this study.

# Acknowledgment

We are grateful to Inge Helland for his input on this paper throughout the period. His guidance on the envelope models and his review of the paper helped us greatly. Our gratitude also goes to Kristian Lillan, Ulf Indahl, Tormod Næs, Ingrid Måge and the team for providing the data for analysis.

# References

Almøy, Trygve. 1996. “A simulation study on comparison of prediction methods when only a few components are relevant.” *Computational Statistics & Data Analysis* 21 (1). Elsevier {BV}: 87–107. doi:[10.1016/0167-9473(95)00006-2](https://doi.org/10.1016/0167-9473(95)00006-2).

Chang, Winston, Joe Cheng, JJ Allaire, Yihui Xie, and Jonathan McPherson. 2018. *Shiny: Web Application Framework for R*. <https://CRAN.R-project.org/package=shiny>.

Cook, R Dennis, Bing Li, and Francesca Chiaromonte. 2010. “Envelope Models for Parsimonious and Efficient Multivariate Linear Regression.” *Statistica Sinica* 20 (3). JSTOR: 927–1010.

Cook, R. D., I. S. Helland, and Z. Su. 2013. “Envelopes and partial least squares regression.” *Journal of the Royal Statistical Society. Series B: Statistical Methodology* 75 (5). Wiley Online Library: 851–77. doi:[10.1111/rssb.12018](https://doi.org/10.1111/rssb.12018).

Cook, R. Dennis. 2018. *An introduction to envelopes : dimension reduction for efficient estimation in multivariate statistics*. 1st ed. Hoboken, NJ : John Wiley & Sons, 2018.

Cook, R. Dennis, and Xin Zhang. 2015. “Simultaneous envelopes for multivariate linear regression.” *Technometrics* 57 (1). Taylor & Francis: 11–25. doi:[10.1080/00401706.2013.872700](https://doi.org/10.1080/00401706.2013.872700).

———. 2016. “Algorithms for Envelope Estimation.” *Journal of Computational and Graphical Statistics* 25 (1). Taylor & Francis: 284–300. doi:[10.1080/10618600.2015.1029577](https://doi.org/10.1080/10618600.2015.1029577).

Cook, R. Dennis, Bing Li, and Francesca Chiaromonte. 2007. “Dimension reduction in regression without matrix inversion.” *Biometrika* 94 (3): 569–84. doi:[10.1093/biomet/asm038](https://doi.org/10.1093/biomet/asm038).

Helland, Inge S. 1990. “Partial least squares regression and statistical models.” *Scandinavian Journal of Statistics* 17 (2). JSTOR: 97–114. doi:[10.2307/4616159](https://doi.org/10.2307/4616159).

———. 2000. “Model Reduction for Prediction in Regression Models.” *Scandinavian Journal of Statistics* 27 (1). Wiley-Blackwell: 1–20. doi:[10.1111/1467-9469.00174](https://doi.org/10.1111/1467-9469.00174).

Helland, Inge S., and Trygve Almøy. 1994. “Comparison of prediction methods when only a few components are relevant.” *Journal of the American Statistical Association* 89 (426): 583–91. doi:[10.1080/01621459.1994.10476783](https://doi.org/10.1080/01621459.1994.10476783).

Helland, Inge S., Solve Saebø, and Ha Kon Tjelmeland. 2012. “Near Optimal Prediction from Relevant Components.” *Scandinavian Journal of Statistics* 39 (4). Wiley Online Library: 695–713. doi:[10.1111/j.1467-9469.2011.00770.x](https://doi.org/10.1111/j.1467-9469.2011.00770.x).

Helland, Inge Svein, Solve Saebø, Trygve Almøy, Raju Rimal, Solve Sæbø, Trygve Almøy, and Raju Rimal. 2018. “Model and estimators for partial least squares regression.” *Journal of Chemometrics* 32 (9). Wiley Online Library: e3044. doi:[10.1002/cem.3044](https://doi.org/10.1002/cem.3044).

Indahl, Ulf. 2005. “A twist to partial least squares regression.” *Journal of Chemometrics* 19 (1). Wiley Online Library: 32–44. doi:[10.1002/cem.904](https://doi.org/10.1002/cem.904).

Johnson, R.A., and D.W. Wichern. 2018. *Applied Multivariate Statistical Analysis (Classic Version)*. Pearson Modern Classics for Advanced Statistics Series. Pearson Education Canada. <https://books.google.no/books?id=QBqlswEACAAJ>.

Jolliffe, I T. 2002. *Principal Component Analysis, Second Edition*. doi:[10.2307/1270093](https://doi.org/10.2307/1270093).

Jong, Sijmen de. 1993. “SIMPLS: An alternative approach to partial least squares regression.” *Chemometrics and Intelligent Laboratory Systems* 18 (3): 251–63. doi:[10.1016/0169-7439(93)85002-X](https://doi.org/10.1016/0169-7439(93)85002-X).

Lee, Minji, and Zhihua Su. 2018. *Renvlp: Computing Envelope Estimators*. <https://CRAN.R-project.org/package=Renvlp>.

Mevik, Bjørn-Helge, Ron Wehrens, and Kristian Hovde Liland. 2018. *Pls: Partial Least Squares and Principal Component Regression*. <https://CRAN.R-project.org/package=pls>.

Naes, Tormod, and Harald Martens. 1985. “Comparison of prediction methods for multicollinear data.” *Communications in Statistics - Simulation and Computation* 14 (3): 545–76. doi:[10.1080/03610918508812458](https://doi.org/10.1080/03610918508812458).

Næs, Tormod, and Inge S Helland. 1993. “Relevant components in regression.” *Scandinavian Journal of Statistics* 20 (3). JSTOR: 239–50.

Næs, Tormod, Oliver Tomic, Nils Kristian Afseth, Vegard Segtnan, and Ingrid Måge. 2013. “Multi-Block Regression Based on Combinations of Orthogonalisation, Pls-Regression and Canonical Correlation Analysis.” *Chemometrics and Intelligent Laboratory Systems* 124. Elsevier: 32–42.

Rencher, Alvin C. 2003. *Methods of Multivariate Analysis*. Vol. 492. John Wiley & Sons.

Rimal, Raju, Trygve Almøy, and Solve Sæbø. 2018. “A tool for simulating multi-response linear model data.” *Chemometrics and Intelligent Laboratory Systems* 176 (May). Elsevier: 1–10. doi:[10.1016/j.chemolab.2018.02.009](https://doi.org/10.1016/j.chemolab.2018.02.009).

Sæbø, Solve, Trygve Almøy, and Inge S. Helland. 2015. “Simrel - A versatile tool for linear model data simulation based on the concept of a relevant subspace and relevant predictors.” *Chemometrics and Intelligent Laboratory Systems* 146. Elsevier B.V.: 128–35. doi:[10.1016/j.chemolab.2015.05.012](https://doi.org/10.1016/j.chemolab.2015.05.012).