Comparison of Multivariate Estimation Methods

Raju Rimal[a,*], Trygve Almøy[a], Solve Sæbø[b]

[a]*Faculty of Chemistry and Bioinformatics, Norwegian University of Life Sciences, Ås, Norway*
[b]*Prorector, Norwegian University of Life Sciences, Ås, Norway*

Abstract

While Data science is battling to extract information from 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, these studies seldom study the impact of/on multiple response model. This study compares some newly-developed (envelope estimation) and well-established estimators (PLS, PCR, Ridge and Lasso) based on the simulated data specifically designed by variying properties such as multicollinearity, correlation between multiple response and amount of information content in latent variables. This study aims to identify itself as an additional eye for researcher to look at these methods.

Keywords: model-comparison,multi-response,simrel

1. Introduction

"Big Data" is becoming a focal discussion in many discipline. Massive explosion of data, integrated with information in several variables and features, have made exploration and analysis more difficult. Researchers are devising new methods and algorithms in order to extract such information weather to study the relationship between variables or to create a predictive model. Mostly, the models used for study contains predictor variables that are directly or indirectly correlated with other predictor variables and many response

*Corresponding Author
Email addresses: raju.rimal@nmbu.no (Raju Rimal), trygve.almoy@nmbu.no (Trygve Almøy), solve.sabo@nmbu.no (Solve Sæbø)

variables. These interlinked relationship observed as a nature of data influences any study whether it is a predictive modeling or an inference.

Modern inter-disciplinary research fields such as chemometrics, echonometrics and bioinformatics are handling multi-response models extensively. This paper attempts to compare some of such methods and their performance on linear model data with specifically designed properties. The properties includes coefficient of determination, correlation between response variables, correlation between predictor variables and number of predictor variables. These properties are discussed more in Experimental Design section.

Sæbø et al. [2015] and Almøy [1996] have made similar comparison with single response setting. Rimal et al. [2018] has introduced a method for simulating multi-response data based on data properties as parameters. The study also has made a basic comparison on some estimation 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 estimation methods with customary estimation methods using dataset simulated with particular properties. An experimental design and the methods under comparison are discussed followed by a brief discussion of strategy behind the data simulation.

2. Simulation Model

Consider a model where response variable $(\mathbf{y})$ and predictor variable $(\mathbf{x})$ follows normal distribution as follows,

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{x} \end{bmatrix} \sim \mathbf{N} \left(\begin{bmatrix} \boldsymbol{\mu}_y \\ \boldsymbol{\mu}_x \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{yy} & \boldsymbol{\Sigma}_{yx} \\ \boldsymbol{\Sigma}_{xy} & \boldsymbol{\Sigma}_{xx} \end{bmatrix} \right) \tag{1}$$

where, $\boldsymbol{\Sigma}_{yy}$ is the variance-covariance matrix of $\mathbf{y}$ and $\boldsymbol{\Sigma}_{xx}$ is the variance-covariance matrix of $\mathbf{x}$. The covariance between $\mathbf{x}$ and $\mathbf{y}$ is $\boldsymbol{\Sigma}_{xy}$. $\boldsymbol{\mu}_y$ and $\boldsymbol{\mu}_x$ are mean vector of $\mathbf{x}$ and $\mathbf{y}$ respectively. An equivalent regression model of (1) can be defined as,

$$\mathbf{y} = \mu_y + \boldsymbol{\beta}^t(\mathbf{x} - \mu_x) + \epsilon \tag{2}$$

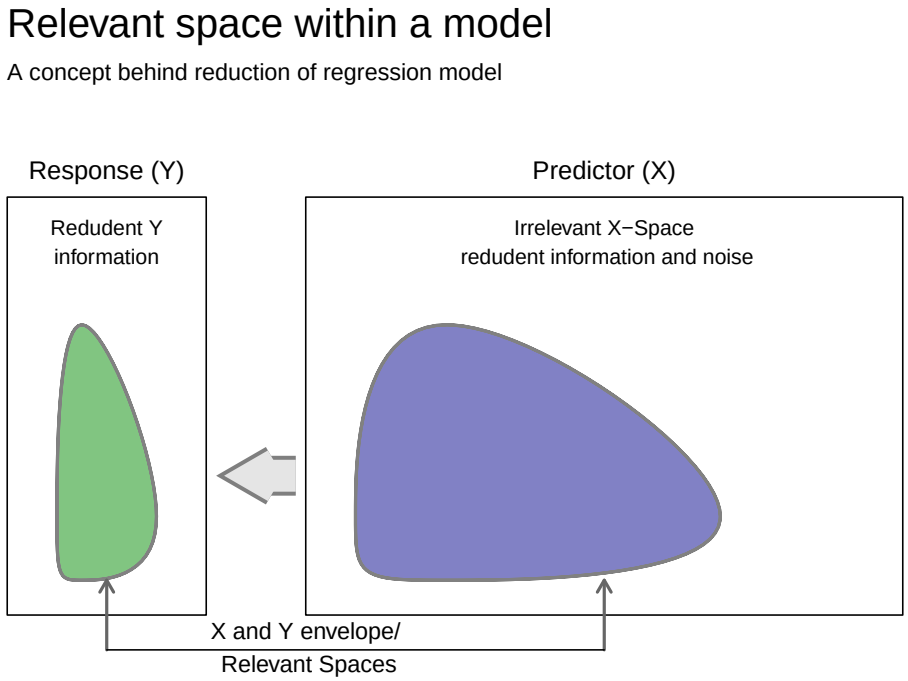

Figure 1: Relevant space in a regression model

where, $\boldsymbol{\beta}^t$ is a matrix of regression coefficients and ϵ is an error term such that $\epsilon \sim N(0, \Sigma_{y|x})$

In a casual relationship model like (2), we assume that the variation in response **y** is caused by the predictor **x**. However, in many situations, only small proportion of the variation in **x** have some connection with **y**. This space can be referred as relevant or material space of **x** and rest as irrelevant or immaterial space. In the similar manner, we can assume that only some part of variation in **y** is caused by the predictors or contains the information (Figure-1). Following the concept of relevant and irrelevant space **(Helland & Trygve, 1994)**, a set of latent variables (such as principal components) that spans these space can be imagined in both predictors and response space. Let us consider the components that spans the relevant predictor space as predictor components and the components that spans the relevant response space as response components. Further let us define the components that spans the irrelevant space as irrelevant components.

Estimation methods like Partial Least Square Regression and many of its derivetives have been using the idea of relevant components in predictor space for better estimation and

prediction (*perhaps need citation*). More recent methods such as Simultaneous Envelope [Cook and Zhang, 2015b] have extended the concept to formulate informative response space (envelope) to decrease error variance in both prediction and estimation.

Define a transformation $\mathbf{w} = \mathbf{Q}\mathbf{y}$ and $\mathbf{z} = \mathbf{R}\mathbf{x}$ such that $\mathbf{Q}$ and $\mathbf{R}$ are orthogonal rotation matrix which transform $\mathbf{x}$ and $\mathbf{y}$ into latent components $\mathbf{z}$ and $\mathbf{w}$ respectively. An equivalent model to (1) can be defined as,

$$\begin{bmatrix} \mathbf{w} \\ \mathbf{z} \end{bmatrix} \sim \mathbf{N}\left(\begin{bmatrix} \boldsymbol{\mu}_w \\ \boldsymbol{\mu}_z \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{ww} & \boldsymbol{\Sigma}_{wz} \\ \boldsymbol{\Sigma}_{zw} & \boldsymbol{\Sigma}_{zz} \end{bmatrix} \right) = \mathbf{N}\left(\begin{bmatrix} \mathbf{Q}\boldsymbol{\mu}_w \\ \mathbf{R}\boldsymbol{\mu}_z \end{bmatrix}, \begin{bmatrix} \mathbf{Q}\boldsymbol{\Sigma}_{ww}\mathbf{Q}^t & \mathbf{Q}\boldsymbol{\Sigma}_{wz}\mathbf{R}^t \\ \mathbf{R}\boldsymbol{\Sigma}_{zw}\mathbf{Q}^t & \mathbf{R}\boldsymbol{\Sigma}_{zz}\mathbf{R}^t \end{bmatrix} \right) \tag{3}$$

Here, $\mathbf{w}$ and $\mathbf{z}$ can be considered as principal components of response (response components) and predictors (predictor components) respectively. A regression model equivalent to (3) can be defined as,

$$\mathbf{w} = \mu_w + \boldsymbol{\alpha}^t(\mathbf{z} - \mu_z) + \tau \tag{4}$$

where, $\boldsymbol{\alpha}^t$ is a matrix of regression coefficients and τ is an error term such that $\tau \sim N(0, \Sigma_{w|z})$

If we should write this part perhaps we should describe how data are simulated from the model above

---

In this study, data is simulated based on model-(1) using an R-package "`simrel`". Rimal et al. [2018] has discussed the concept behind the simulation in detail that is implemented in the R-package. Properties in the simulated data such as its dimension, correlation structure and signal-to-noise ratio are controlled through parameters. The estimation methods under comparison, discussed in Methods section, are tested against these dataset and evaluated on the basis of their performance on both prediction and estimation.

3. Experimental Design

A complete factorial design with different levels of parameters are used as follows.

Number of predictors: In order to observe the performance of estimators on tall and wide predictor matrices, 20 and 250 predictor variables are simulated. Parameter `p` controls this properties in `simrel` function in the R Package.

Multicollinearity in predictor variables: Highly collinear predictors can be explained completely by few components. Parameter `gamma` (γ) in `simrel` controls the eigenvalues of the predictor variables as (5).

$$\lambda_i = e^{-\gamma(i-1), \gamma>0} \text{ and } i = 1, 2, \ldots, p \tag{5}$$

Here, $\lambda_i, i = 1, 2, \ldots p$ are eigenvalues of predictor variables. Here we have used 0.2 and 0.9 as different levels of `gamma`. Higher the value of gamma, higher will be the correlation between predictors and vice versa.

Correlation in response variables: Correlation in response variables is less explored area. Here we have tried to explore that part with 4 levels of correlation in response variables. We have used `eta` η parameter of `simrel` for controlling the eigenvalues corresponding to response variables as (6).

$$\kappa_i = e^{-\eta(i-1), \eta>0} \text{ and } j = 1, 2, \ldots, m \tag{6}$$

Here, $\kappa_i, i = 1, 2, \ldots m$ are eigenvalues of response variables and `m` is number of response variables. Here we have used 0, 0.4, 0.8 and 1.2 as different levels of `eta`. Larger the value of eta, larger will be the correlation between response variables and vice versa.

Coefficient of determination: Coefficient of determination controls signal-to-noise ratio in simulated data and influence the prediction heavily. `R2` parameters in `simrel` package is used to specify coefficient of determination correlation for a response components. Here we have used single response component which contains information

Table 1: Simulation parameters of all designs

Design	p	eta	gamma	R2	Design	p	eta	gamma	R2
1	20	0.0	0.2	0.8	**17**	20	0.4	0.2	0.8
2	20	0.8	0.9	0.8	**18**	20	1.2	0.9	0.8
3	250	0.8	0.9	0.8	**19**	250	1.2	0.9	0.8
4	250	0.0	0.2	0.8	**20**	250	0.4	0.2	0.8
5	20	0.8	0.2	0.8	**21**	20	1.2	0.2	0.8
6	20	0.0	0.9	0.8	**22**	20	0.4	0.9	0.8
7	250	0.0	0.9	0.8	**23**	250	0.4	0.9	0.8
8	250	0.8	0.2	0.8	**24**	250	1.2	0.2	0.8
9	20	0.0	0.9	0.8	**25**	20	0.4	0.9	0.8
10	20	0.8	0.2	0.8	**26**	20	1.2	0.2	0.8
11	250	0.8	0.2	0.8	**27**	250	1.2	0.2	0.8
12	250	0.0	0.9	0.8	**28**	250	0.4	0.9	0.8
13	20	0.8	0.9	0.8	**29**	20	1.2	0.9	0.8
14	20	0.0	0.2	0.8	**30**	20	0.4	0.2	0.8
15	250	0.0	0.2	0.8	**31**	250	0.4	0.2	0.8
16	250	0.8	0.9	0.8	**32**	250	1.2	0.9	0.8

for the relevant predictor components. The single informative response components is *blended* in `m` response variables. Here we have used 0.8 and 0.8 levels of coefficient of determination `R2` (ρ^2) corresponding to the single response component.

Further, a complete factorial design from the levels of above parameters gave us 32 designs. Each design is associated with a dataset having unique properties. Table~1, shows all the design obtained from above factors. For each design and estimation method, 50 datasets were simulated for replication. In total, there were $7 \times 32 \times 50$, i.e. 11200 dataset simulated.

Common parameters: Each dataset was simulated with $n = 100$ number of observation and $m = 4$ response variables. Further, the position of relevant predictor components are set to at position 1, 2, 3, 4 and 5, 6, 7, 8. In addition, we have assumed that there is only 1 number of informative response component. So, that the first informative response component is rotated orthogonally together with 3 uninformative response components. This spread out the information in all simulated response variables. For further details see: [Rimal et al., 2018].

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

```
simrel(
  n      = 100L,                  ## Observations
  p      = 20L,                   ## Predictors
  m      = 4L,                    ## Response
  q      = 20,                    ## Relevant predictors
  relpos = list(c(1, 2, 3, 4)),   ## Position of predictor components
  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 2 shows the covariance structure of this design. The left plot shows that only first four predictor components at position 1, 2, 3 and 4 have high covariance with first response components. Remaining predictor components are not relevant for any of the response components and since response space is explained by single components, response components except 1 has zero covariance with the predictors components. This is the population covariance structure of the simulated. The right plot, where as, shows that the all the predictor components have non-zero covariance with all the response components where first few components have large covariance than the later components. The `eta` parameters controls the eigenvalues corresponding to response variables, it should be relatively easier to predict fourth response than first since it has less variation present in the data and will be easier to explain by the predictors. This is explored in analysis section of this paper.

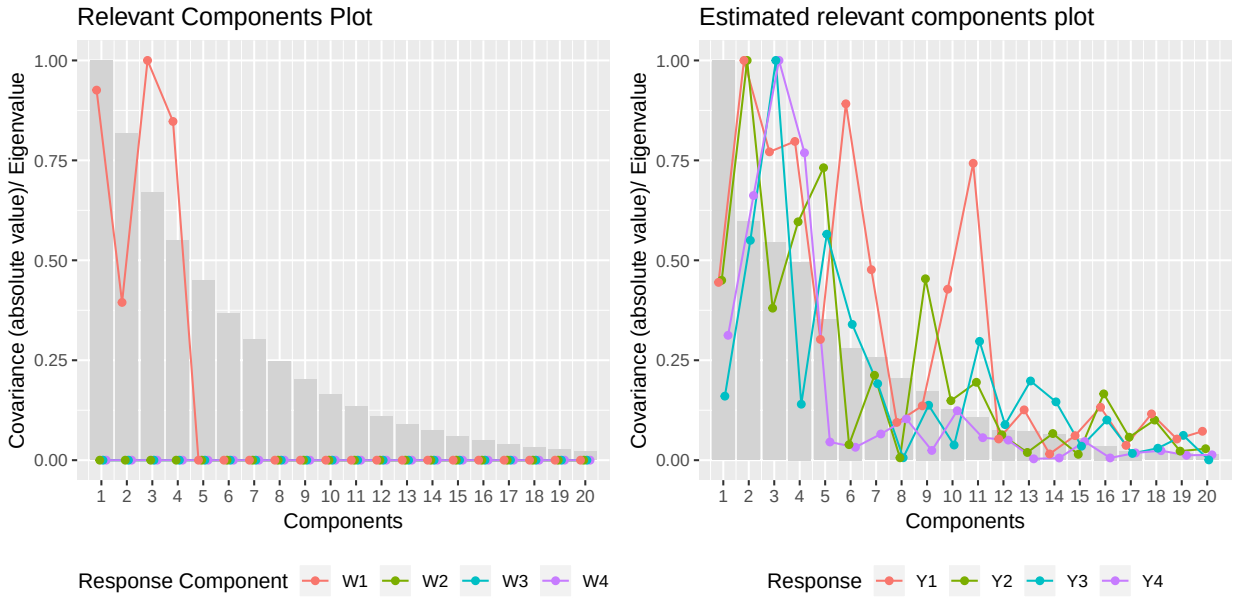

Figure 2: Scaled covariance between predictor components and response components (left) and scaled expected covariance between predictor components and response variables (the one we obtain after rotation). The bar in the background eigenvalues corresponding to each components.

4. Methods

The aim of this paper is to compare some customary methods such as Principal Component Regression (PCR), Partial Least Squares (PLS), Ridge and Lasso with some stat-of-the-art methods such as envelope estimation and simultaneous envelope estimation. This will also make comparison on the methods based on reduced regression such as PCR, PLS and envelope with that of Shrinkage methods such as Ridge and Lasso. The estimation methods used for comparison as follows.

a) Principal Component Regression (PCR)
b) Partial Least Squares Regression using single response per model (PLS1)
c) Partial Least Squares Regression using all response together (PLS2)
d) Ridge Regression
e) Lasso Regression
f) Envelope estimation in predictor space (Xenv)[Cook and Zhang, 2015a]
g) Simultaneous estimation of envelope in both predictor and response space

(Senv)[Cook and Zhang, 2015b]

4.1. Modification in envelope estimation

Since envelope estimators (Xenv and Senv) are based on maximum likelihood estimation (MLE), it fails to estimate on wide matrices, i.e. $p > n$. In order to incorporate these method in our comparison, we have used the principal components $(\mathbf{z})$ of predictor variables $(\mathbf{x})$ using required number of components for capturing 97.5% of the variation in it. The new set of variables $\mathbf{z}$ were used for envelope estimation. The regression coefficients $(\hat{\boldsymbol{\alpha}})$ corresponding to these new variables $\mathbf{z}$ were transformed back to obtain coefficients for each predictor variables $(\hat{\boldsymbol{\beta}})$ as,

$$\hat{\boldsymbol{\beta}} = \mathbf{e}_k \hat{\boldsymbol{\alpha}_k}$$

where, $\mathbf{e}_k$ is the eigenvectors with k number of components.

4.2. Comparison Criteria

The models are compared on following three basis.

Estimation Error Diagonal elements of (7) is used as estimation error of each response in every fitted model.

$$\text{estimation error} = \left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right)^t \left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right) \tag{7}$$

where, $\boldsymbol{\beta}$ and $\hat{\boldsymbol{\beta}}$ are true and estimated regression coefficient corresponding to (1).

Prediction Error Diagonal elements of (8) is used as estimation error of each response in every fitted model.

$$\text{prediction error} = \left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right)^t \Sigma_{xx} \left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right) + \Sigma_{y|x} \tag{8}$$

where, Σ_{xx} is the true covariance matrix of predictor and $\Sigma_{y|x}$ is the true model error both obtained from simulation.

Comparison of estimated coefficients with true coefficients A graphical comparison of estimated coefficient from different estimators is compared with the true coefficients. (*need some statistical way of doing so*) This shows how each additional components contributes on finding beta coefficients closer to the true value. This also helps us to see if the additional components contains noise how the estimates deviates from the true coefficients.

5. Analysis

In order to carry out a proper statistical comparison, a multivariate linear model is used with prediction and estimation error with respect to each response variables as dependent variables and the interaction of simulation parameters (`p`, `gamma`, `eta` and `R2`) and `Methods` as independent variables as (9).

$$\mathbf{y} = \boldsymbol{\mu} + \texttt{p} \times \texttt{gamma} \times \texttt{eta} \times \texttt{R2} \times \texttt{Methods} + \boldsymbol{\varepsilon} \tag{9}$$

where, y_j is the prediction error in one model and estimation error in another corresponding to response j. Let us discuss the results of Multivariate Analysis of Variance (MANOVA) for both of these models.

5.1. Prediction error model:

Using the prediction error corresponding to response j as y_j in (9) a 2nd interaction of methods and simulation parameters as predictors, a clear high F-statistic corresponding to Pillai statistic for coefficient of determination (`R2`) has suggested high influence of `R2` in the model. In order to look for effects of other interactions, we have used model (10) fitted for low and high coefficient of determination separately.

$$\mathbf{y} = \boldsymbol{\mu} + \texttt{p} \times \texttt{gamma} \times \texttt{eta} \times \texttt{Methods} + \boldsymbol{\varepsilon} \tag{10}$$

From the ANOVA tables **??** and **??**, we can see that the factor `eta` has high effect on prediction error in both the cases of low and high coefficient of determination. Also, Also,

the tables show that Methods have significantly large second and third order interaction with `gamma` and `eta`, however this is mostly dominated by the high main effect of `eta` which is same in both the models.

Although, methods have similar prediction performance for different levels of `eta`, `R2` and `p`, the difference is considerable for low and high multicollinearity in case of both low (Figure **??**) and high (Figure **??**) `R2`.

5.2. Estimation error

Some plots and descriptions on estimation error models

Based on the simulation design and ANOVA analysis, following analysis revolves abound the inter-connection between both prediction and estimation errors and the properties of data. The analysis is based on comparison of true and estimated regression coefficients, estimation error and the prediction error. Following four groups can be identified from the simulation design (see table 1):

a) Wide vs Tall prediction matrix (20 and 250)
b) High vs low multicollinearity (0.2 and 0.9)
c) High vs low coefficient of determination (0.8 and 0.8)
d) Different levels of correlation between responses (0, 0.4, 0.8 and 1.2)

5.3. Error (prediction and estimation) Analysis

5.4. Regression Coefficients

6. Discussion

References

References

Almøy, T., 1996. A simulation study on comparison of prediction methods when only a few components are relevant. Computational Statistics & Data Analysis 21, 87–107. URL: https://doi.org/10.1016%2F0167-9473%2895%2900006-2, doi:10.1016/0167-9473(95)00006-2.

Cook, R.D., Zhang, X., 2015a. Foundations for envelope models and methods. Journal of the American Statistical Association 110, 599–611.

Cook, R.D., Zhang, X., 2015b. Simultaneous envelopes for multivariate linear regression. Technometrics 57, 11–25.

Rimal, R., Almøy, T., Sæbø, S., 2018. A tool for simulating multi-response linear model data. Chemometrics and Intelligent Laboratory Systems 176, 1–10. URL: https://doi.org/10.1016%2Fj.chemolab.2018.02.009, doi:10.1016/j.chemolab.2018.02.009.

Sæbø, S., Almøy, T., Helland, I.S., 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 .