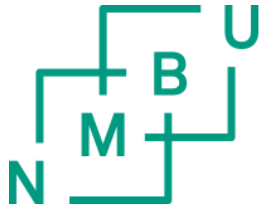
Analyzing the effect of multicollinearity and position of relevant components

STAT 360, 2019

Raju Rimal

april 10, 2023

Norges miljø- og biovitenskapelige universitet

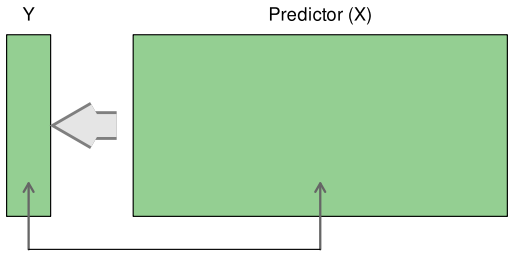


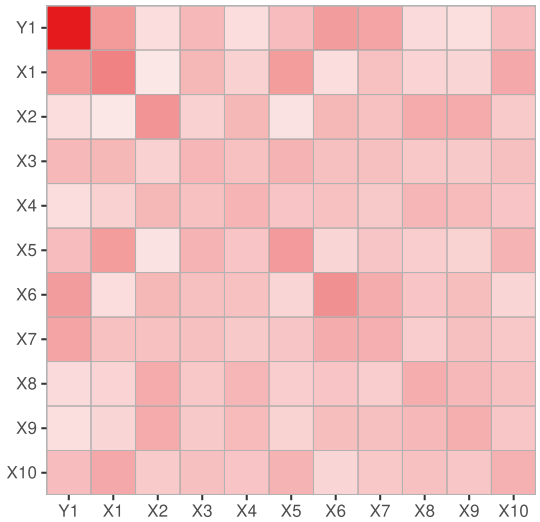
Norges miljø- og biovitenskapelige universitet

Linear Model

Relevant and irrelevant space in linear model

Linear Model





The Model:

$$\begin{bmatrix} y \\ \mathbf{x} \end{bmatrix} \sim \mathsf{N}\left(\begin{bmatrix} \mu_y \\ \boldsymbol{\mu}_x \end{bmatrix}, \begin{bmatrix} \sigma_y^2 & \boldsymbol{\sigma}_{yx} \\ \boldsymbol{\sigma}_{xy} & \boldsymbol{\Sigma}_{xx} \end{bmatrix} \right)$$

Linear Regression:

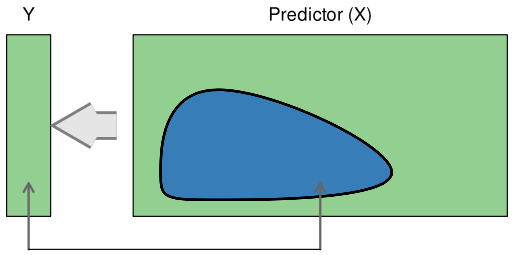
$$y = \mu_y + \boldsymbol{\beta}^t(\mathbf{x} - \boldsymbol{\mu}_x) + \varepsilon,\ \varepsilon \sim \mathsf{N}(0, \sigma^2)$$

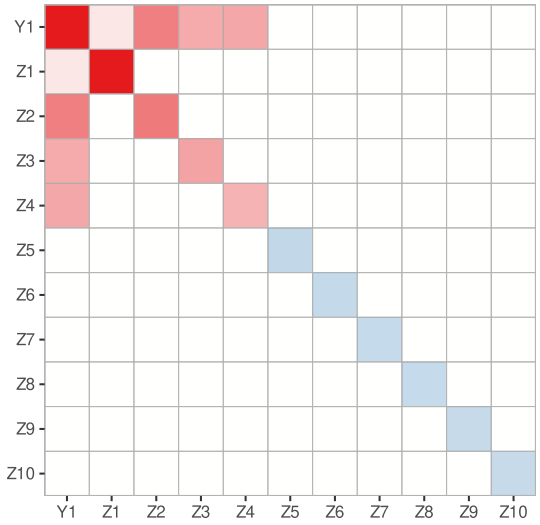
Regression Coefficients:

$$\boldsymbol{\beta} = \boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\sigma}_{xy}$$

Linear Model

Let us make a transformation as $\boldsymbol{z} = \mathbf{R}\boldsymbol{x}$ where $\mathbf{R}$ is an orthogonal matrix, i.e. $\mathbf{R}^t = \mathbf{R}^{-1}$.





New Model

$$\begin{bmatrix} y \\ \mathbf{z} \end{bmatrix} \sim \mathsf{N}\left(\begin{bmatrix} \mu_y \\ \boldsymbol{\mu}_z \end{bmatrix}, \begin{bmatrix} \sigma_y^2 & \boldsymbol{\sigma}_{yz} \\ \boldsymbol{\sigma}_{zy} & \boldsymbol{\Sigma}_{zz} \end{bmatrix} \right) = \mathsf{N}\left(\begin{bmatrix} \mu_y \\ \boldsymbol{\mu}_z \end{bmatrix}, \begin{bmatrix} \sigma_y^2 & \boldsymbol{\sigma}_{yx}\mathbf{R}^t \\ \mathbf{R}\boldsymbol{\sigma}_{xy} & \mathbf{R}\boldsymbol{\Sigma}_{xx}\mathbf{R}^t \end{bmatrix} \right)$$

Linear Regression

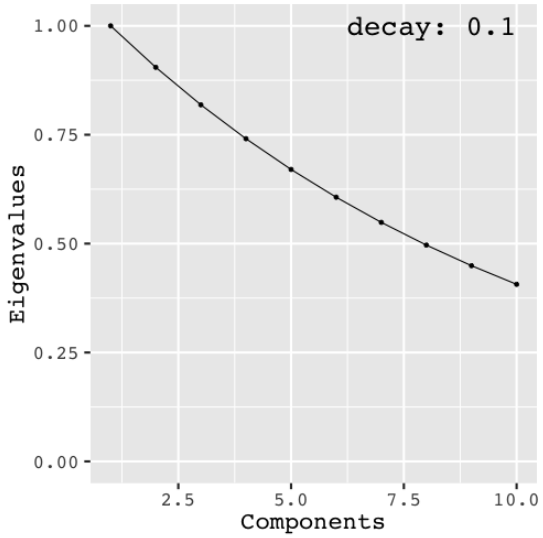
$$y = \mu_y + \boldsymbol{\alpha}^t(\mathbf{z} - \boldsymbol{\mu}_z) + \epsilon,\ \epsilon \sim \mathsf{N}(0, \tau^2)$$

Regression Coefficients

$$\boldsymbol{\alpha} = \mathbf{R}\boldsymbol{\beta} = \boldsymbol{\Sigma}_{zz}^{-1}\boldsymbol{\sigma}_{zy} = \Lambda^{-1}\boldsymbol{\sigma}_{zy} = \sum_{i=1}^{p} \frac{\sigma_{z_i y}}{\lambda_i}$$

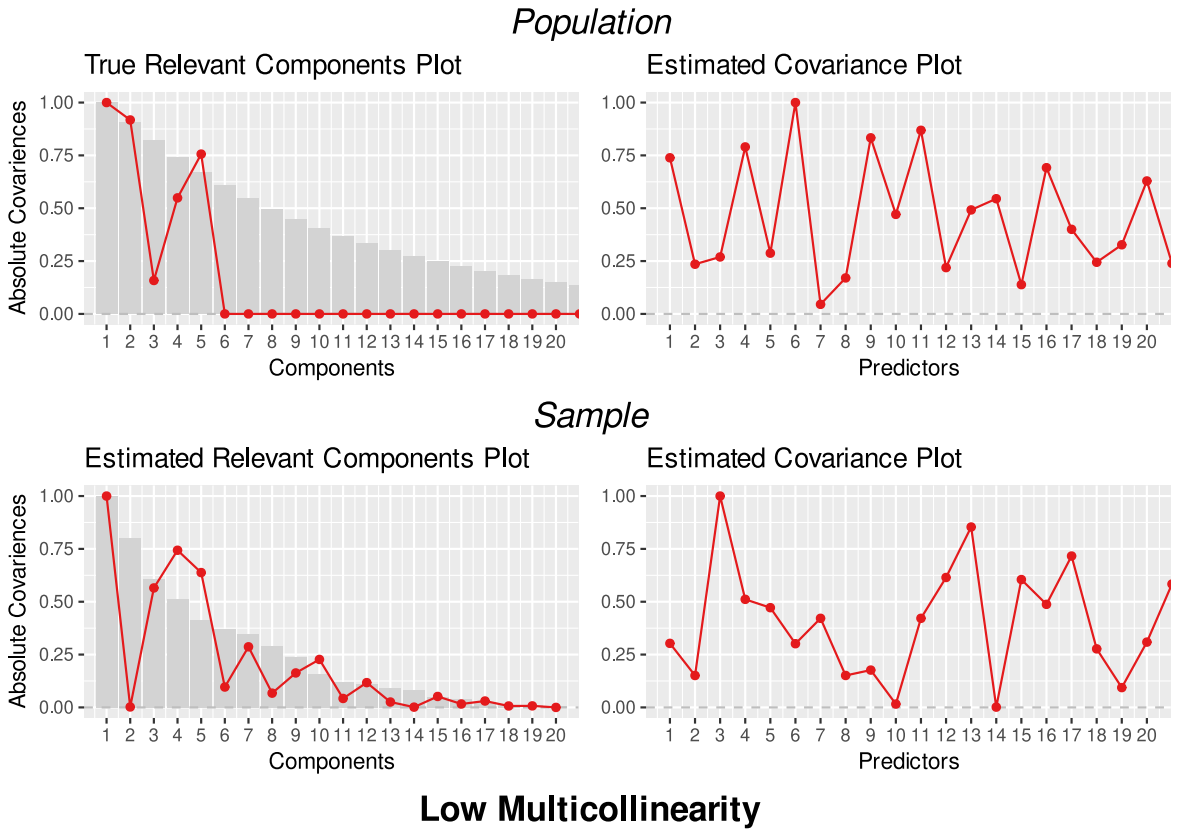
Simulation

```
    Design gamma relpos
1 Design 1   0.1    1:5
2 Design 2   0.1   5:10
3 Design 3   1.2    1:5
4 Design 4   1.2   5:10
```




```
# A tibble: 1 × 5
      n     p     q    R2 ntest
  <dbl> <dbl> <dbl> <dbl> <dbl>
1    20    30    30   0.8    50
```

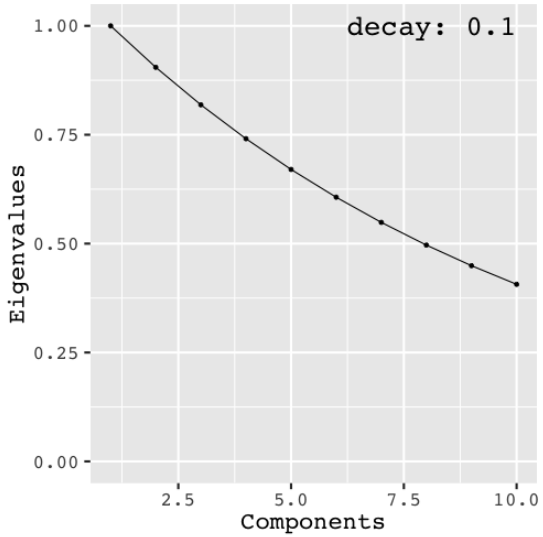
Relevant Components



Simulation

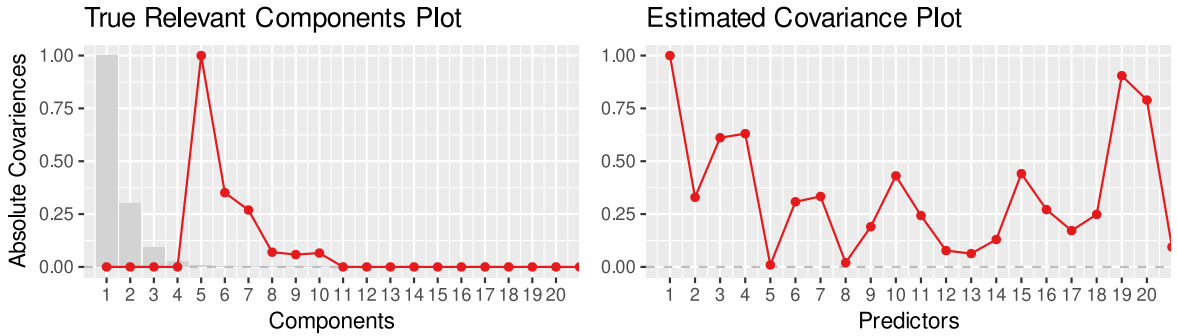
Relevant Components

```
    Design gamma relpos
1 Design 1   0.1    1:5
2 Design 2   0.1   5:10
3 Design 3   1.2    1:5
4 Design 4   1.2   5:10
```

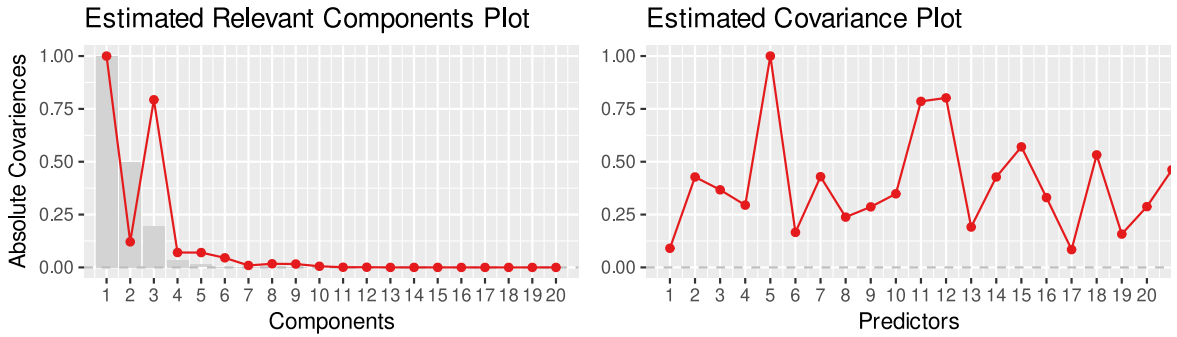



```
# A tibble: 1 × 5
      n     p     q    R2 ntest
  <dbl> <dbl> <dbl> <dbl> <dbl>
1    20    30    30   0.8    50
```

Population

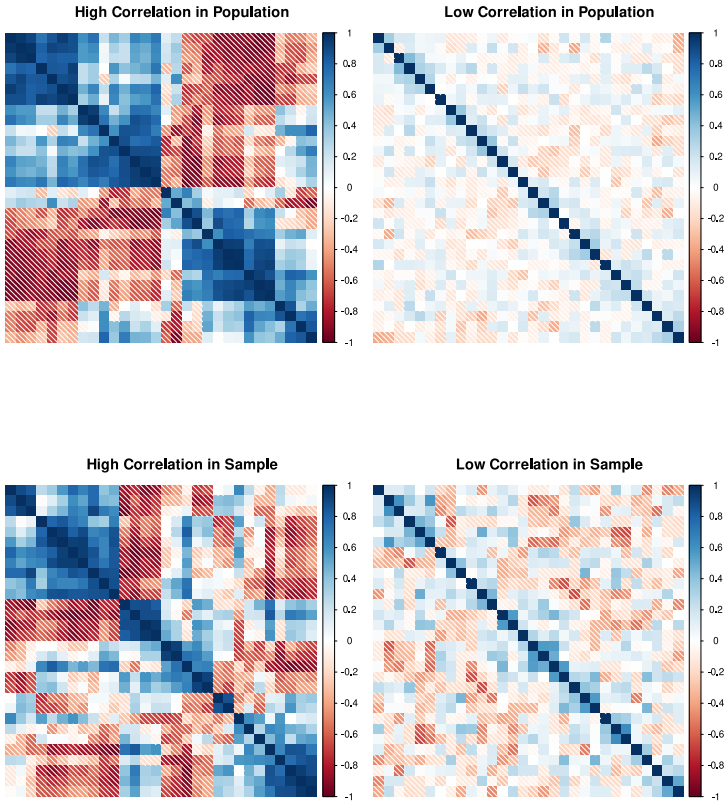


Sample



High Multicollinearity

Correlation Structure

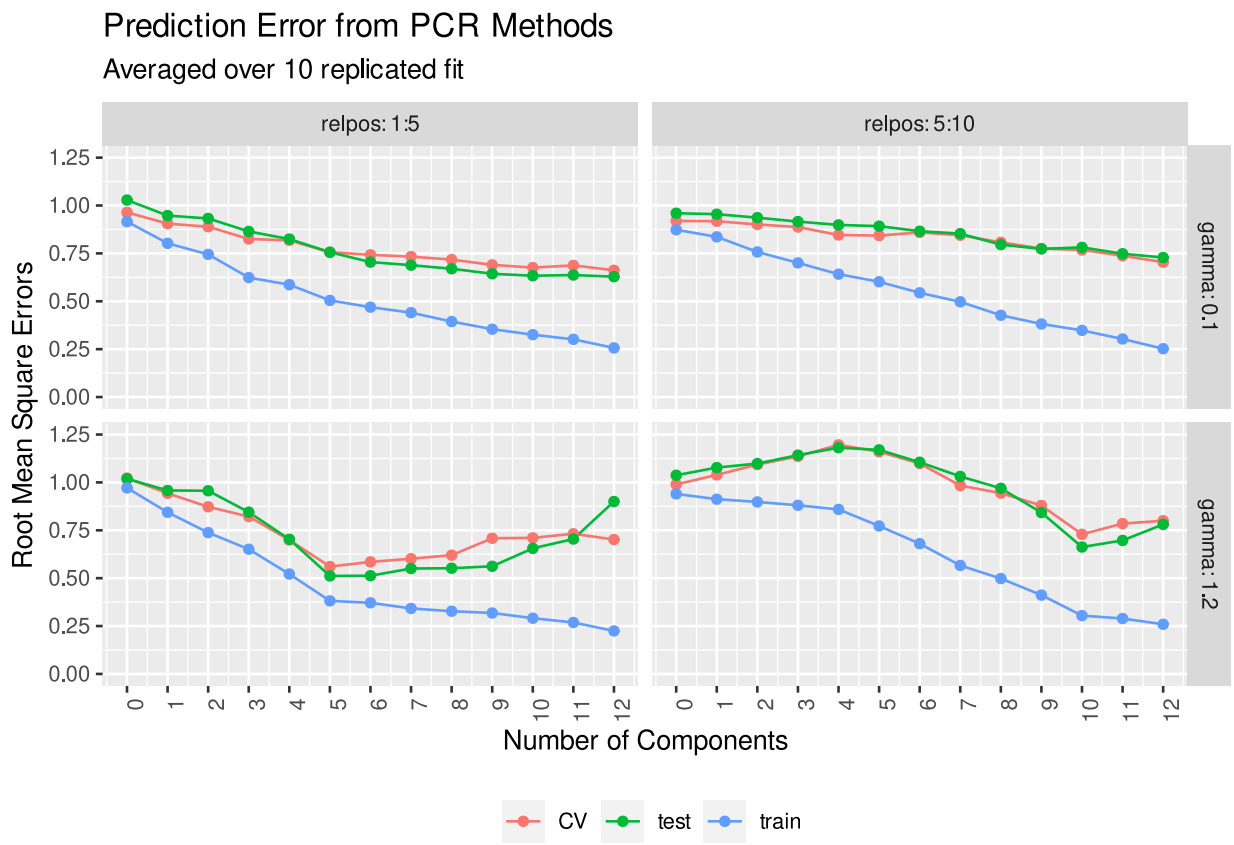


Structure of Simulated Data

y	x.1	x.2	x.3	x.4	...6	x.27	x.28	x.29	x.30
-1.162	-0.234	0.017	-0.242	0.033	...	0.054	0.013	-0.090	-0.015
0.395	-0.173	0.092	0.015	0.167	...	-0.072	0.306	0.124	-0.095
1.701	0.140	0.144	0.481	-0.021	...	-0.176	0.360	0.210	-0.089
0.849	0.117	0.011	0.229	-0.102	...	-0.061	0.030	0.020	0.005
-1.158	0.002	0.046	-0.282	0.002	...	0.082	-0.195	-0.084	-0.015
1.547	-0.219	-0.016	-0.543	-0.025	...	0.168	-0.259	-0.228	0.079
-0.782	0.223	-0.026	0.370	0.056	...	-0.073	0.115	0.190	-0.052
0.671	-0.052	0.120	0.097	0.064	...	-0.111	0.193	0.065	-0.049
-0.984	0.420	-0.091	0.355	-0.327	...	0.023	-0.182	-0.002	0.083
-0.179	0.002	-0.078	-0.153	-0.082	...	0.114	-0.135	-0.072	0.040

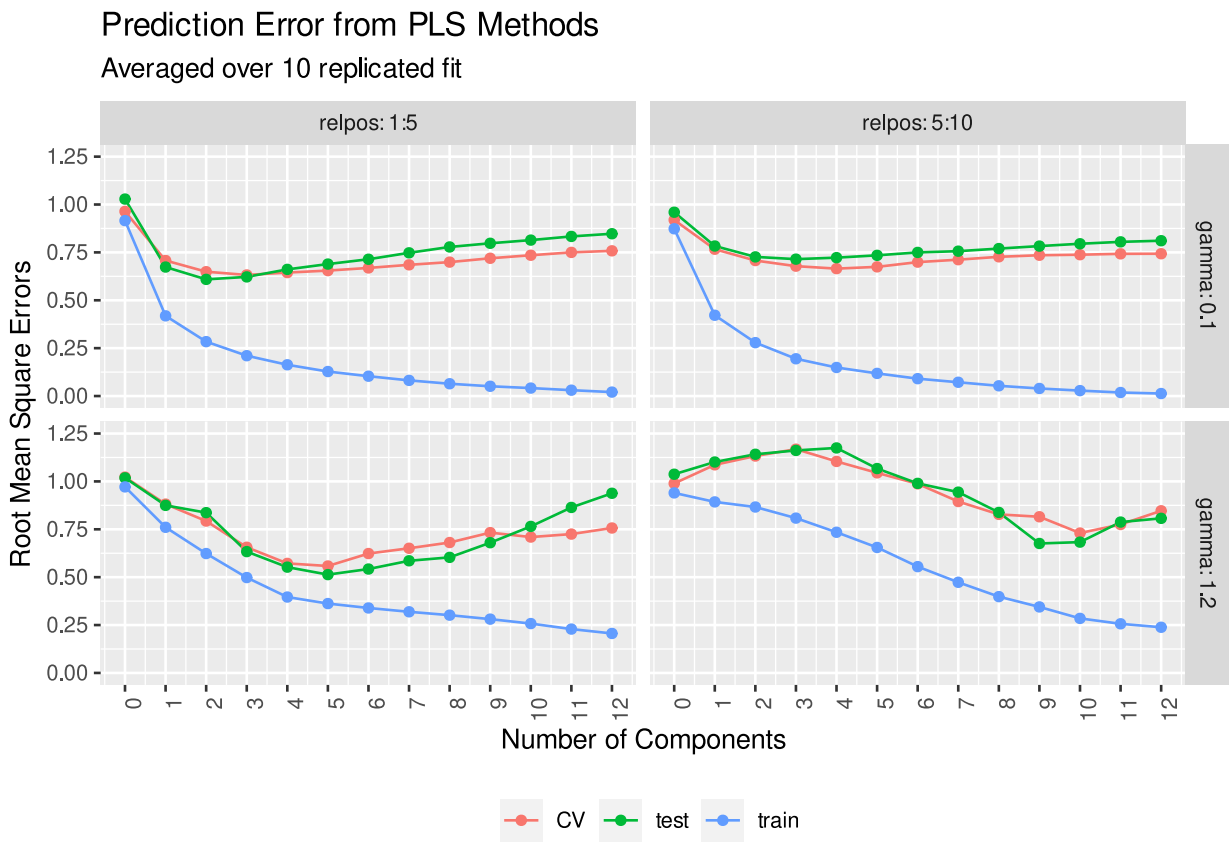
Prediction Performance

Principal Component Regression

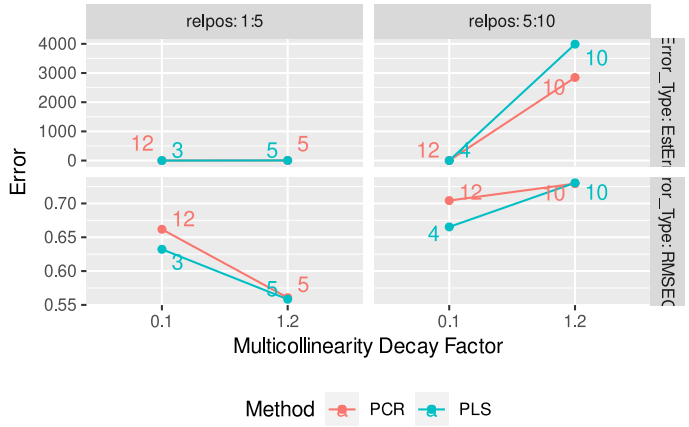


Prediction Performance

Partial Least Square Regression

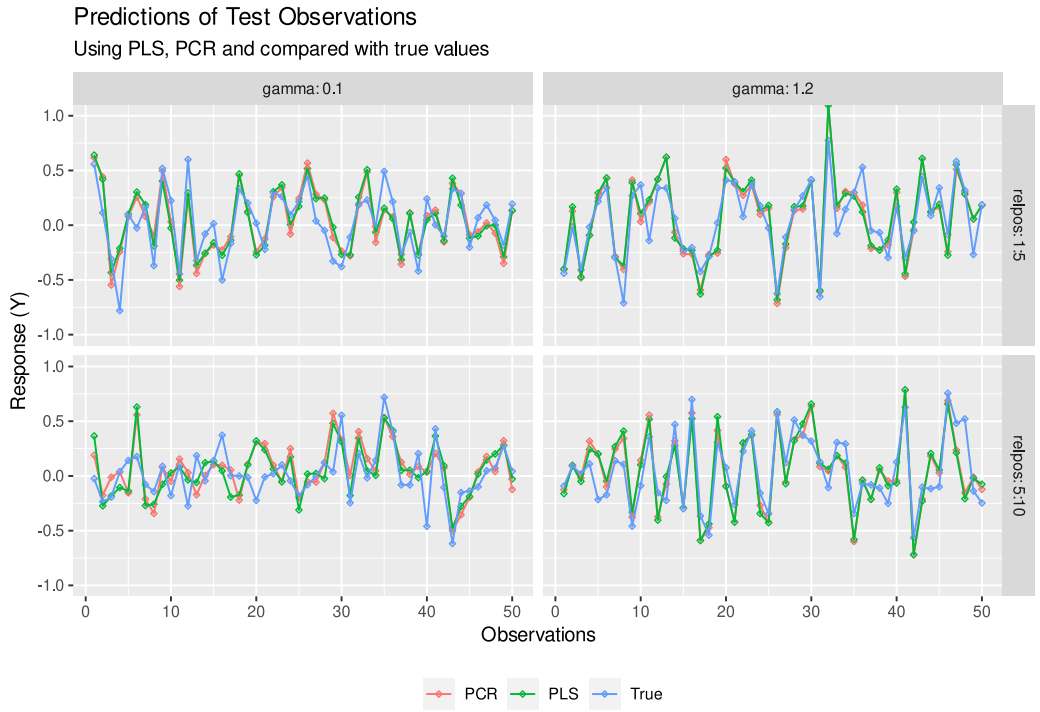


Error Comparison



Design	gamma	relpos	Method	Component	RMSEE	RMSEP
1	0.1	1:5	PCR	12	0.670	0.628
1	0.1	1:5	PLS	3	0.682	0.622
2	0.1	5:10	PCR	12	0.973	0.728
2	0.1	5:10	PLS	4	1.000	0.723
3	1.2	1:5	PCR	5	1.501	0.512
3	1.2	1:5	PLS	5	2.399	0.513
4	1.2	5:10	PCR	10	53.366	0.663
4	1.2	5:10	PLS	10	63.169	0.683

Estimation and Prediction Error



Prediction Error:

$$\mathsf{E}\left[(\beta-\hat{\beta})^t(X_\circ X_\circ^t)^{-1}(\beta-\hat{\beta})\right]$$

Estimation Error:

$$\mathsf{E}\left[(\beta-\hat{\beta})^t(\beta-\hat{\beta})\right]$$

References

Almøy, T. (1996). "A simulation study on comparison of prediction methods when only a few components are relevant". In: *Computational Statistics & Data Analysis* 21.1, pp. 87-107. DOI: 10.1016/0167-9473(95)00006-2. URL: https://doi.org/10.1016/0167-9473(95)00006-200006-2).

Helland, I. S. and T. Almøy (1994). "Comparison of prediction methods when only a few components are relevant". In: *Journal of the American Statistical Association* 89.426, pp. 583-591.

Helland, I. S., S. Sæbø, T. Almøy, et al. "Model and estimators for partial least squares regression". In: *Journal of Chemometrics*, p. e3044.

Rimal, R., T. Almøy, and S. Sæbø (2018). "A tool for simulating multi-response linear model data". In: *Chemometrics and Intelligent Laboratory Systems* 176, pp. 1-10.

Sæbø, S., T. Almøy, and I. S. Helland (2015). "simrel - A versatile tool for linear model data simulation based on the concept of a relevant subspace and relevant predictors". In: *Chemometrics and Intelligent Laboratory Systems*.

Installation

R-Package

```
install.packages("simrel")

if (!require(devtools)) install.packages("devtools")
devtools::install_github("simulatr/simrel")
```

Shiny Application

```
if (!require(simrel)) install.packages("simrel")
shiny::runGitHub("simulatr/AppSimulatr")
```