PhD Midway Seminar

Simulation Tool and its application

Raju Rimal

Supervisors:

Solve Sæbø Tryge Almøy

08 March, 2017

Introduction

My PhD Plan

My PhD Plan

PhD Program	
Phase 1	Make a simulation Tool
Phase 2	Apply it for comparing different estimation Methods
Phase 3	Extend the simulation tool for model with background information
Phase 4	Apply it to test multi-matrix extension of PLS models such as LPLS and UPLS

My PhD Plan

PhD Program	
Phase 1	Make a simulation Tool
Phase 2	Apply it for comparing different estimation Methods
Phase 3	Extend the simulation tool for model with background information
Phase 4	Apply it to test multi-matrix extension of PLS models such as LPLS and UPLS

Why I am doing this

Important for:

My PhD Plan

PhD Program	
Phase 1	Make a simulation Tool
Phase 2	Apply it for comparing different estimation Methods
Phase 3	Extend the simulation tool for model with background information
Phase 4	Apply it to test multi-matrix extension of PLS models such as LPLS and UPLS

Why I am doing this

Important for:

- Research

My PhD Plan

PhD Program	
Phase 1	Make a simulation Tool
Phase 2	Apply it for comparing different estimation Methods
Phase 3	Extend the simulation tool for model with background information
Phase 4	Apply it to test multi-matrix extension of PLS models such as LPLS and UPLS

Why I am doing this

Important for:

- Research
- Education and

My PhD Plan

PhD Program	
Phase 1	Make a simulation Tool
Phase 2	Apply it for comparing different estimation Methods
Phase 3	Extend the simulation tool for model with background information
Phase 4	Apply it to test multi-matrix extension of PLS models such as LPLS and UPLS

Why I am doing this

Important for:

- Research
- Education and
- Method Evaluation

What I learn

What I learn

- Advanced Multivariate methods and their properties

What I learn

- Advanced Multivariate methods and their properties
- Programming concept for developing statistical packages and applications for various statistical methods

What I learn

- Advanced Multivariate methods and their properties
- Programming concept for developing statistical packages and applications for various statistical methods
- Extending and improving existing methods in statistics

What I learn

- Advanced Multivariate methods and their properties
- Programming concept for developing statistical packages and applications for various statistical methods
- Extending and improving existing methods in statistics
- And, obviously, to properly document what I have done

Today’s Special

Today's Special

Today I will talk about:

Today's Special

Today I will talk about:

- A comparative study of various estimation techniques by simulating linear model data using `simulatr` in single response situation Demonstration

Today's Special

Today I will talk about:

- A comparative study of various estimation techniques by simulating linear model data using `simulatr` in single response situation Demonstration

Today's Special

Today I will talk about:

- A comparative study of various estimation techniques by simulating linear model data using `simulatr` in single response situation Demonstration

Today's Special

Today I will talk about:

- A comparative study of various estimation techniques by simulating linear model data using `simulatr` in single response situation Demonstration
- Simulation tool (`simulatr`) we are building

Today's Special

Today I will talk about:

- A comparative study of various estimation techniques by simulating linear model data using `simulatr` in single response situation Demonstration
- Simulation tool (`simulatr`) we are building

A comparative study of different estimation methods using simulated data

Overview

Four estimtion methods were considered

Overview

Four estimtion methods were considered

Ordinary Least Squares (OLS)

- Although **unbiased**, suffer highly from **multicollinearity**
- Widely used and can be used as **reference for comparison**

Partial Least Squares (PLS)

- **Well established** and widely used method
- Based on Latent Structure and **free of multicollinearity problem**

Overview

Four estimation methods were considered

Envelope

- Relatively **new method** (Cook, Helland, & Su, 2013) and is also based on reduction of regression model
- Based on **Maximum Likelihood** but works better than OLS in p approaches n

Bayes PLS

- **Bayesian Estimation** of regression coefficient
- **Promising performance** was shown in previous studies (I. S. Helland, Sæbø, & Tjelmeland, 2012)

Simulation Design

Simulation Design

Population Parameters were set as follows:

Simulation Design

Population Parameters were set as follows:

- **Number of sample observations**: *50*

Simulation Design

Population Parameters were set as follows:

- **Number of sample observations**: *50*
- **Number of predictor variables**: *15* and *40*

Simulation Design

Population Parameters were set as follows:

- **Number of sample observations**: *50*
- **Number of predictor variables**: *15* and *40*
- **Coefficient of determination** (R^2): *0.5* and *0.9*

Simulation Design

Population Parameters were set as follows:

- **Number of sample observations**: *50*
- **Number of predictor variables**: *15* and *40*
- **Coefficient of determination** (R^2): *0.5* and *0.9*
- **Level of multicollinearity**: *0.5* and *0.9*

Simulation Design

Population Parameters were set as follows:

- **Number of sample observations**: *50*
- **Number of predictor variables**: *15* and *40*
- **Coefficient of determination** (R^2): *0.5* and *0.9*
- **Level of multicollinearity**: *0.5* and *0.9*
- **Position of relevant components**: *1* and *2*; *1* and *3*; *2* and *3*; *1*, *2* and *3*

Simulation Design

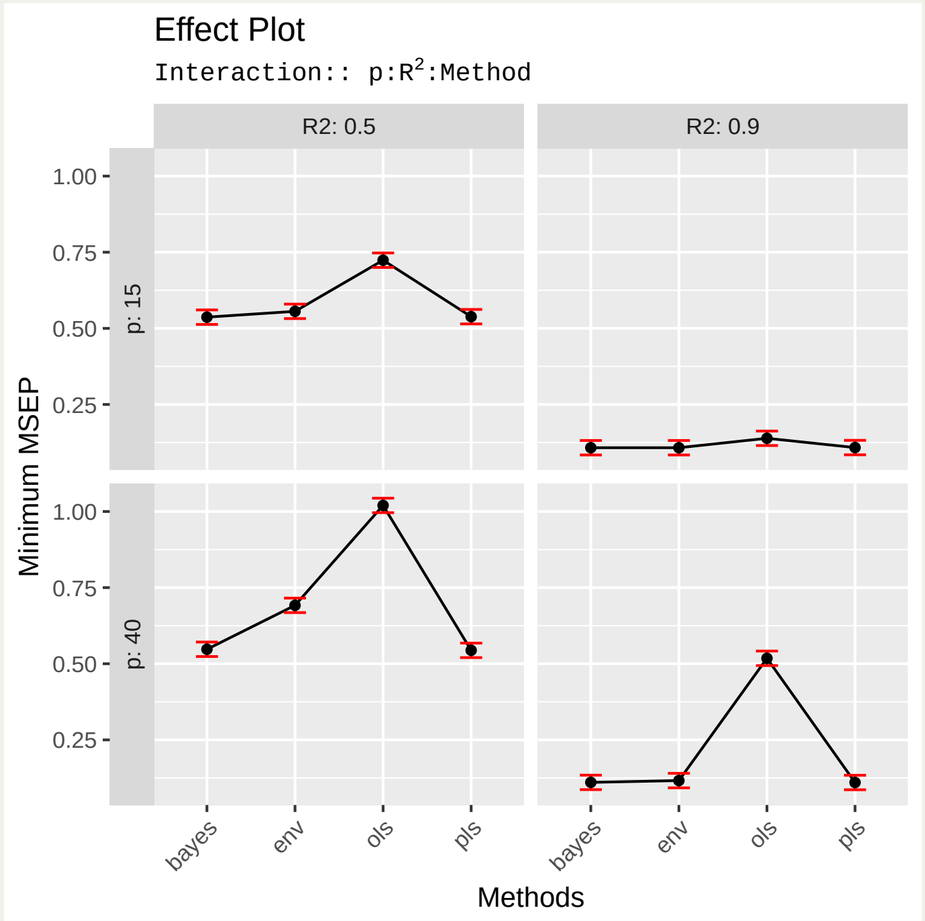
Population Parameters were set as follows:

- **Number of sample observations**: *50*
- **Number of predictor variables**: *15* and *40*
- **Coefficient of determination (R^2)**: *0.5* and *0.9*
- **Level of multicollinearity**: *0.5* and *0.9*
- **Position of relevant components**: *1* and *2*; *1* and *3*; *2* and *3*; *1*, *2* and *3*

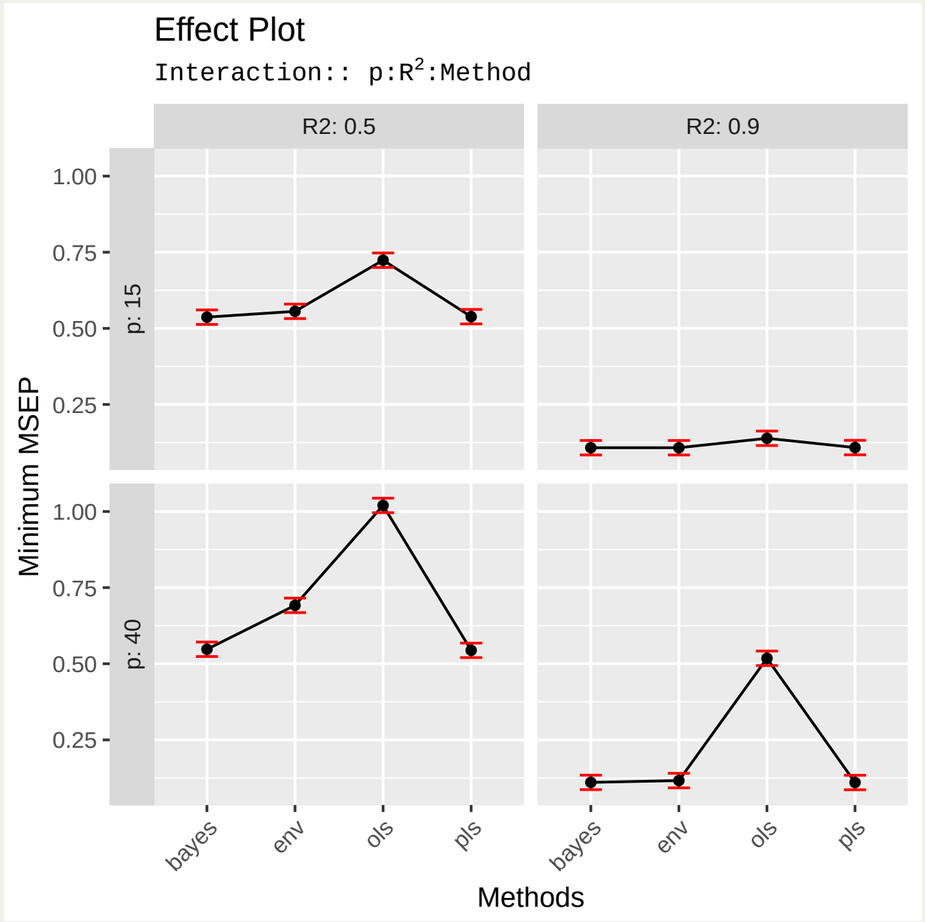
From the combination of above parameters, **32** datasets were simulated with **5** replication of each, i.e. **160 datasets** with 5 of them having similar population properties.

A Systematic Comparison

A Systematic Comparison

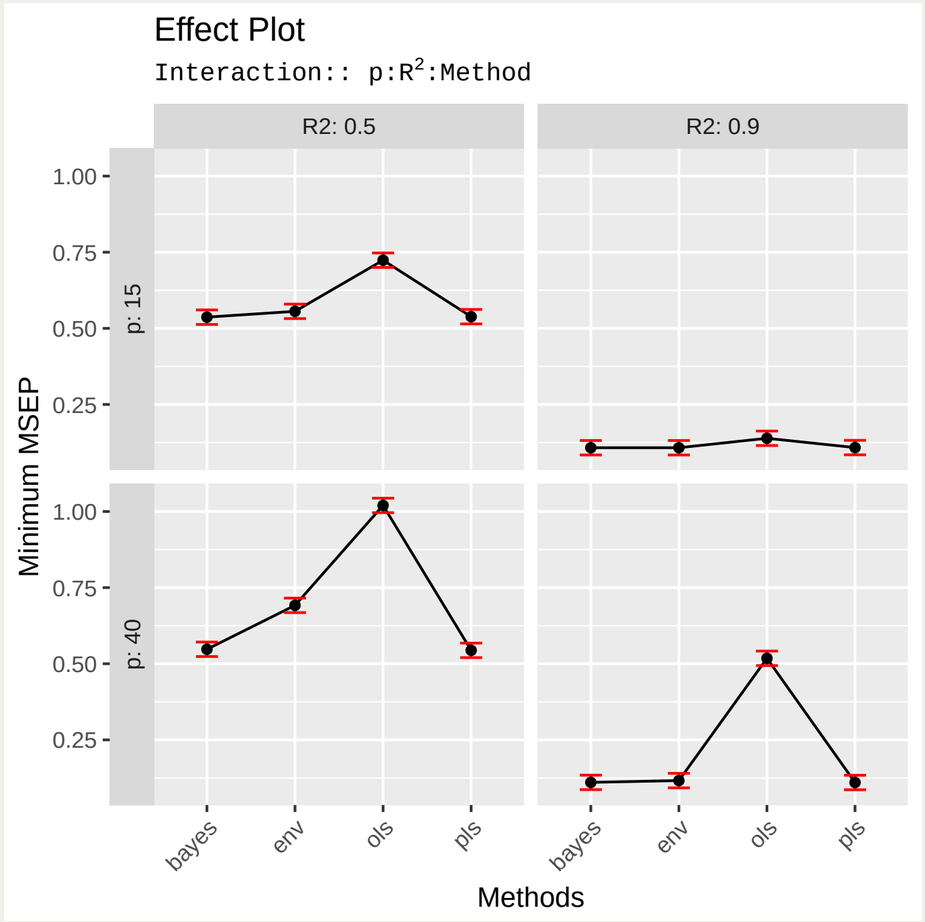


A Systematic Comparison



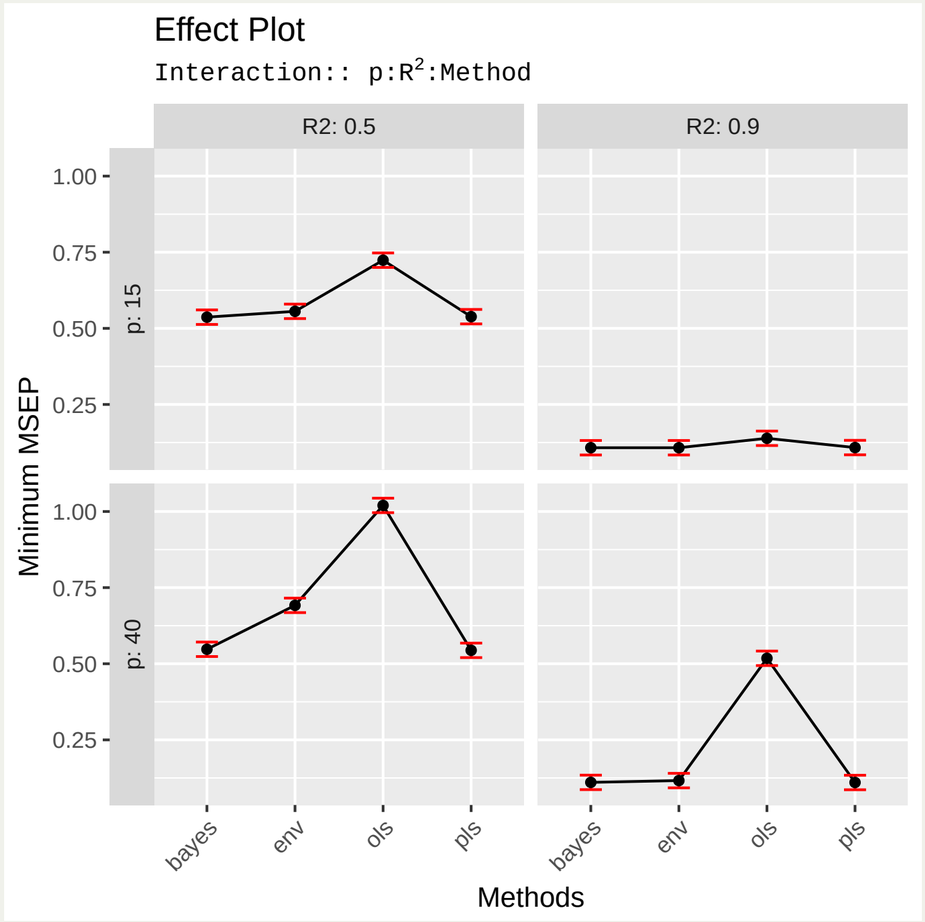
- Bayes PLS has out-performed others methods

A Systematic Comparison



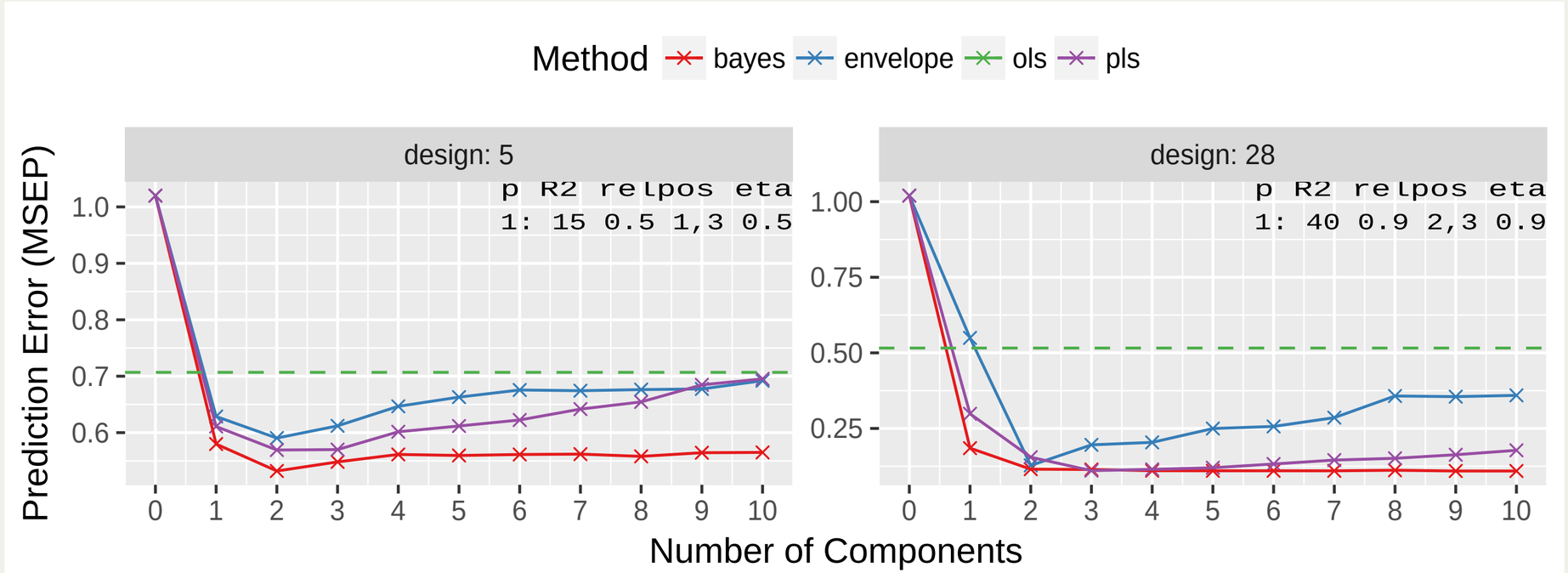
- Bayes PLS has out-performed others methods
- Envelope performed better than OLS

A Systematic Comparison

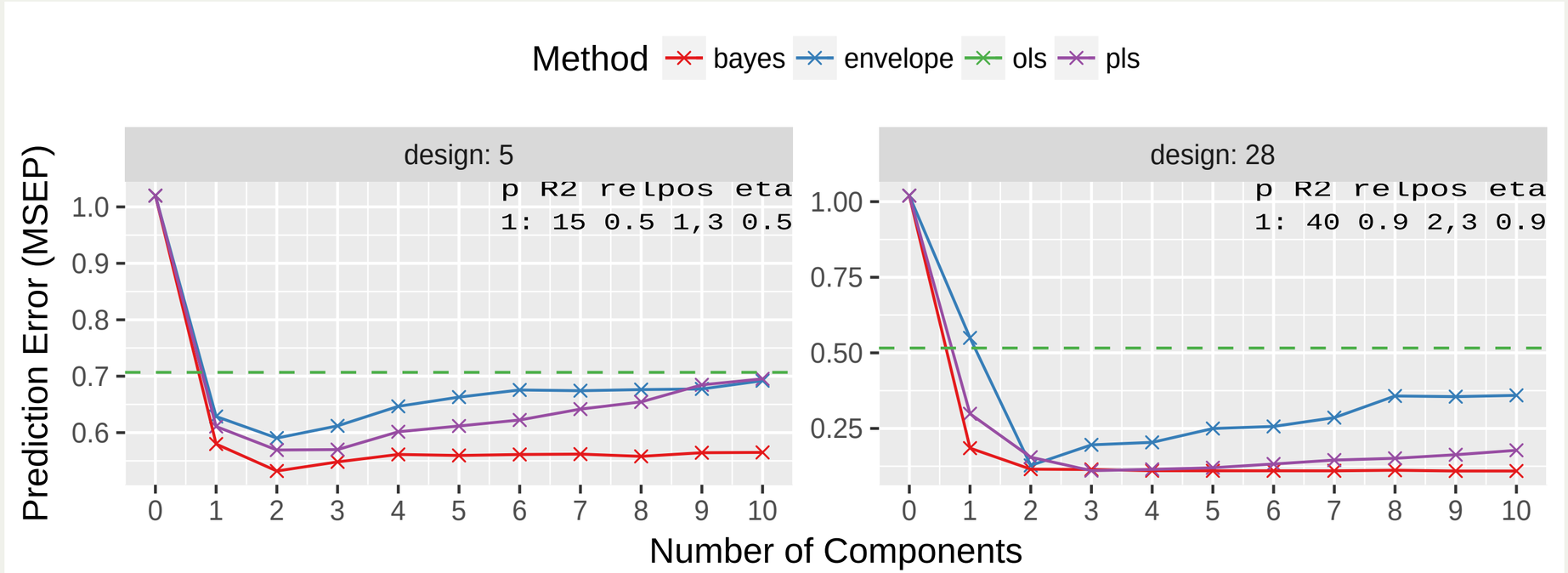


- Bayes PLS has out-performed others methods
- Envelope performed better than OLS
- OLS prediction: very poor in noisy data

A Systematic Comparison

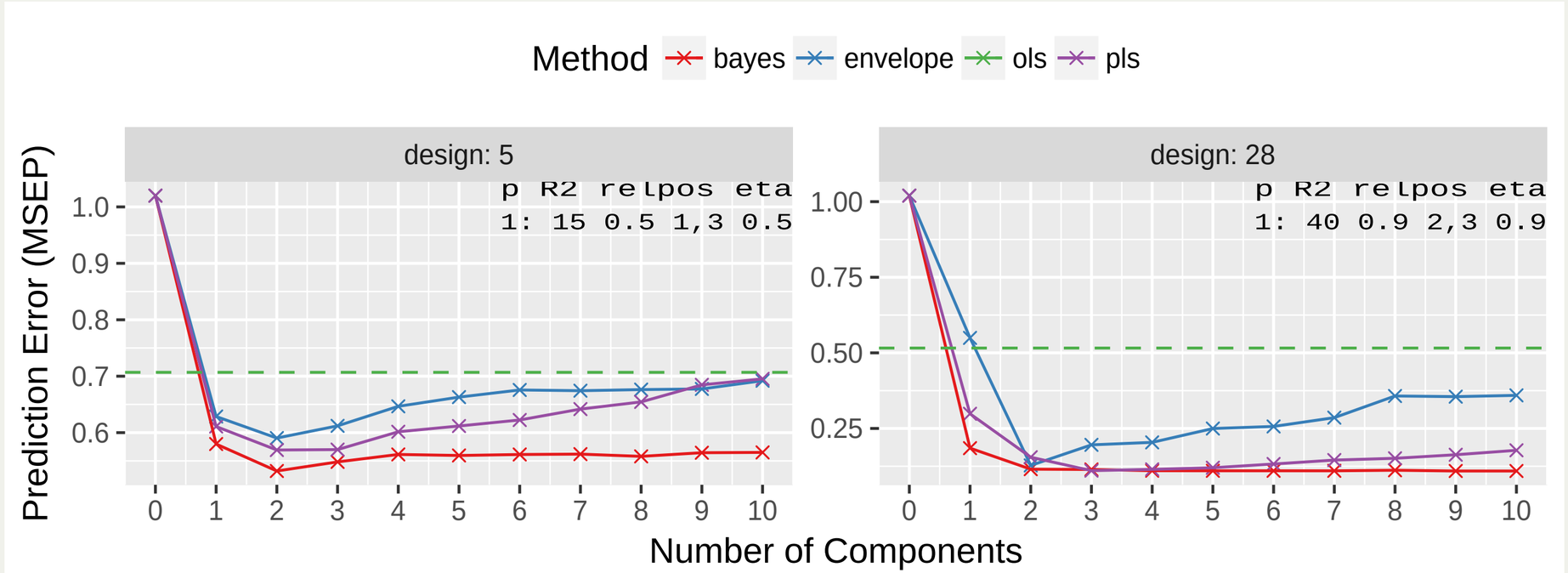


A Systematic Comparison



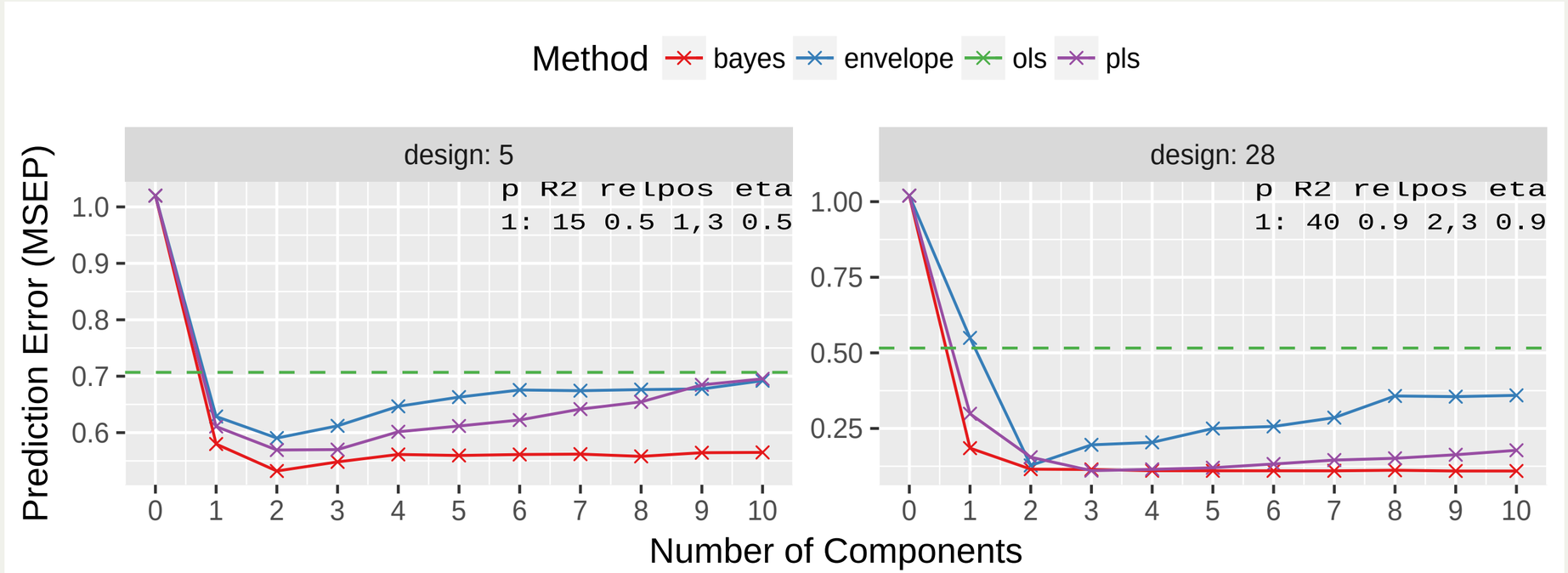
- Bayes PLS has approached to its minimum error with very few component and remained low for additional component

A Systematic Comparison



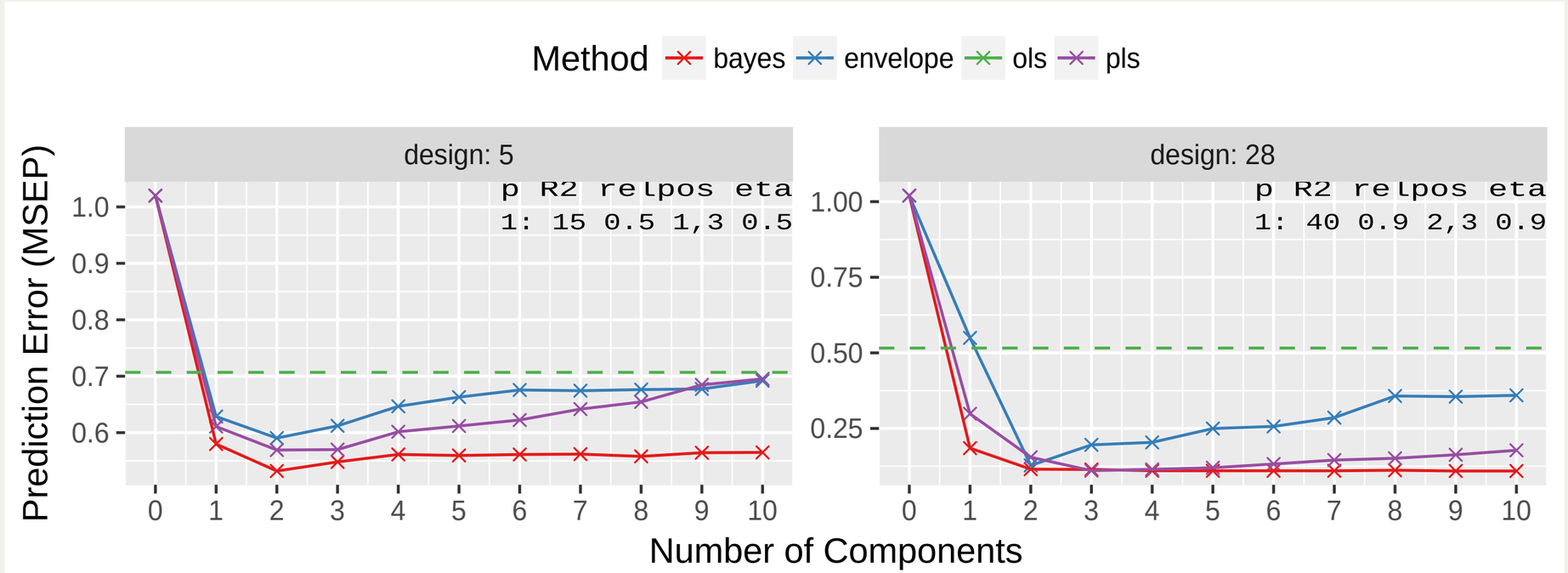
- PLS has moderate performance but better than envelope in many situations.

A Systematic Comparison



- OLS prediction is poor especially with large number of predictor

A Systematic Comparison



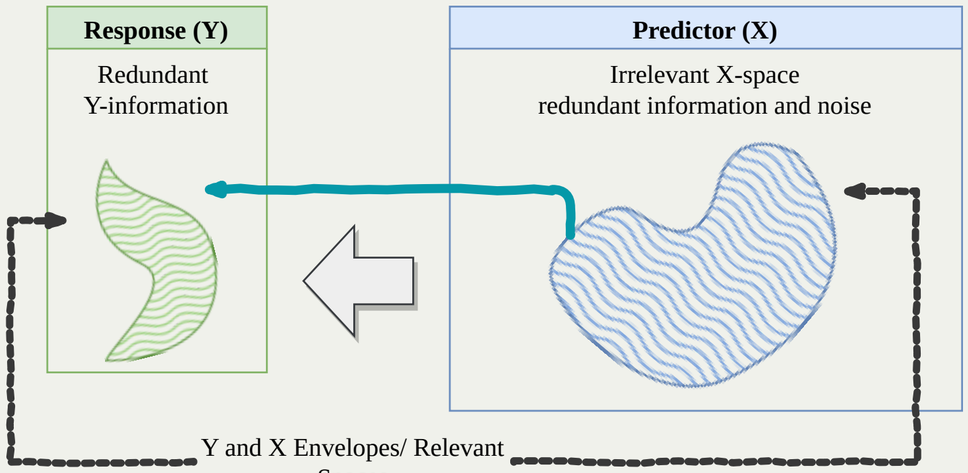
- Envelope method captured its minimum error and the error increased with additional components

`simrel-m`: A versatile tool for simulating multi-response linear model data

simrel-m

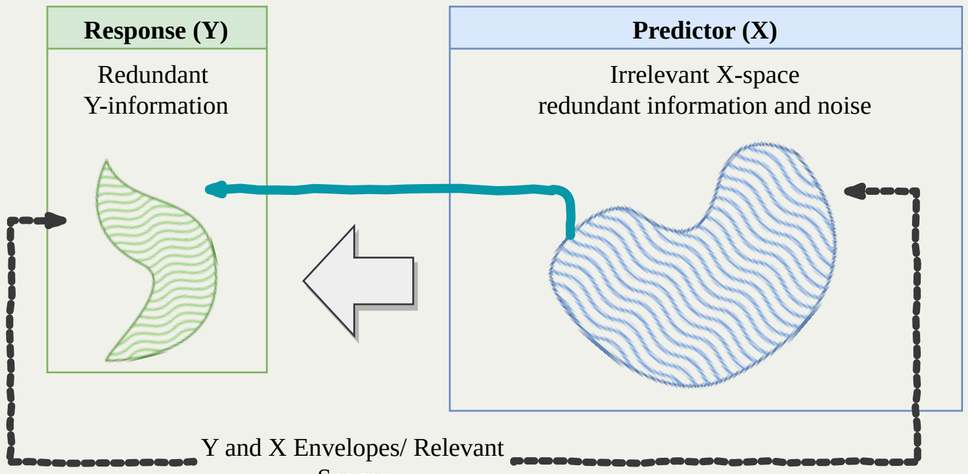
simrel-m

It is an extension of `simrel` *(Sæbø, Almøy, & Helland, 2015) r-package for simulating* ***multi-response data***



simrel-m

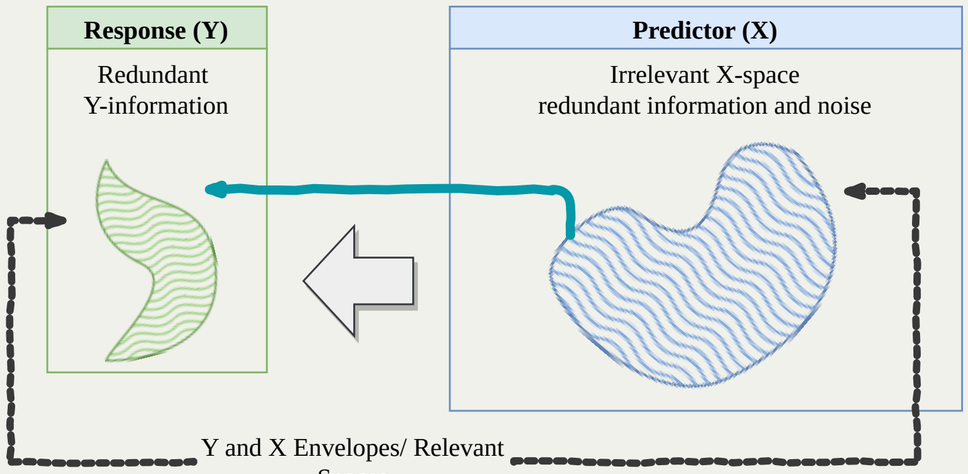
It is an extension of `simrel` *(Sæbø, Almøy, & Helland, 2015) r-package for simulating* ***multi-response data***



- Based on idea of **reduction of random regression model**

simrel-m

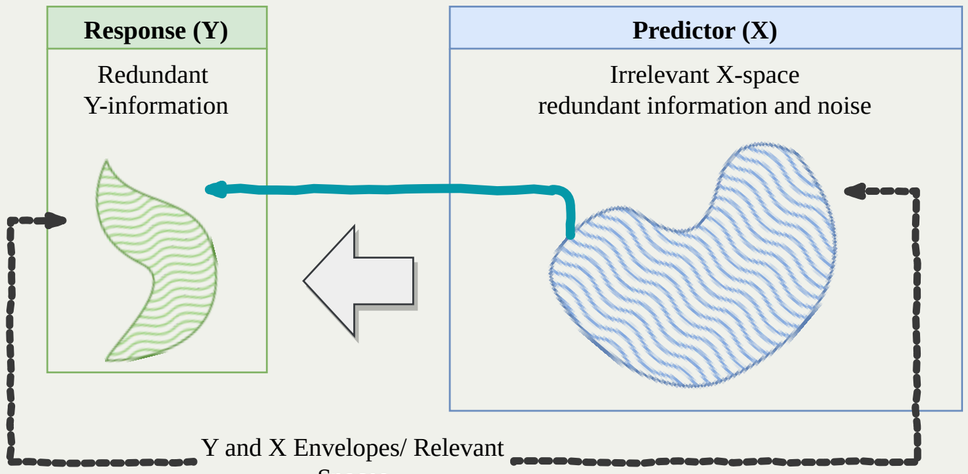
It is an extension of `simrel` *(Sæbø, Almøy, & Helland, 2015) r-package for simulating* ***multi-response data***



- It separates X into subspaces that is relevant and irrelevant for predicting each response

simrel-m

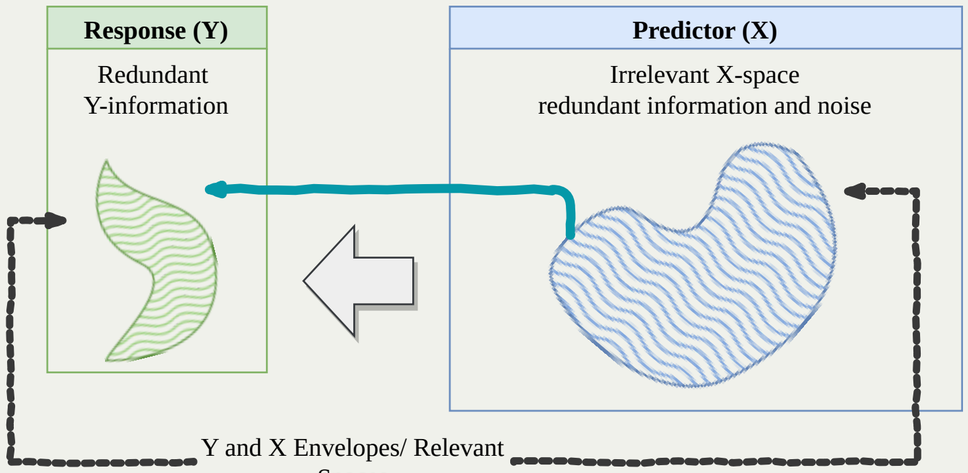
It is an extension of `simrel` *(Sæbø, Almøy, & Helland, 2015) r-package for simulating* ***multi-response data***



- It re-parameterize the population model,
  $\mathbf{Y} = \boldsymbol{\mu}_Y + \mathbf{B}^t\left(\mathbf{X} - \boldsymbol{\mu}_X\right) + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim N(0, \boldsymbol{\Sigma}_{Y|X})$

simrel-m

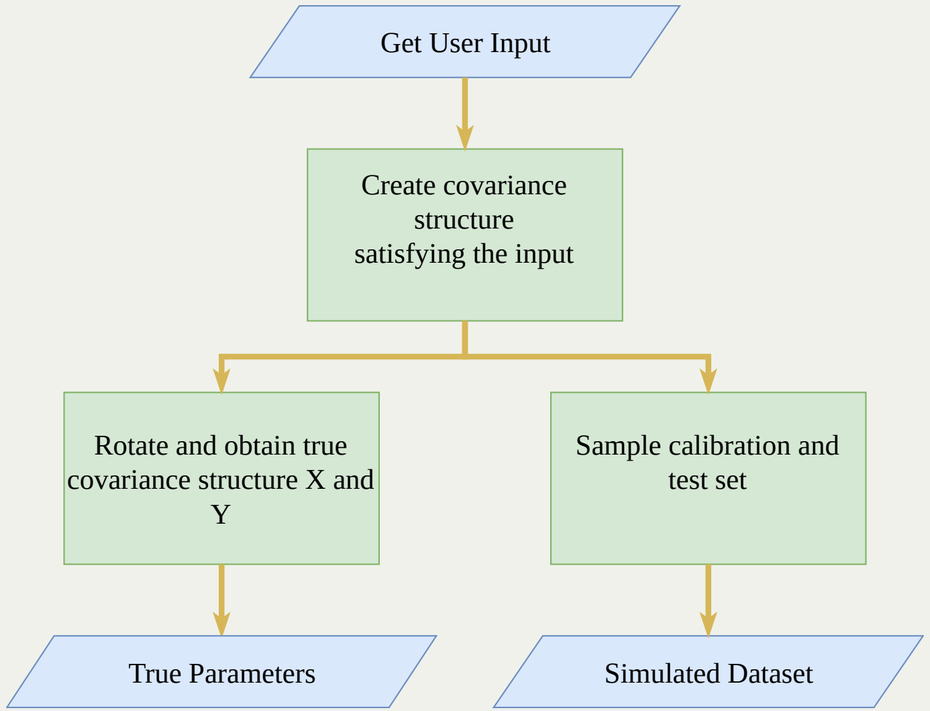
It is an extension of `simrel` *(Sæbø, Almøy, & Helland, 2015) r-package for simulating* ***multi-response data***



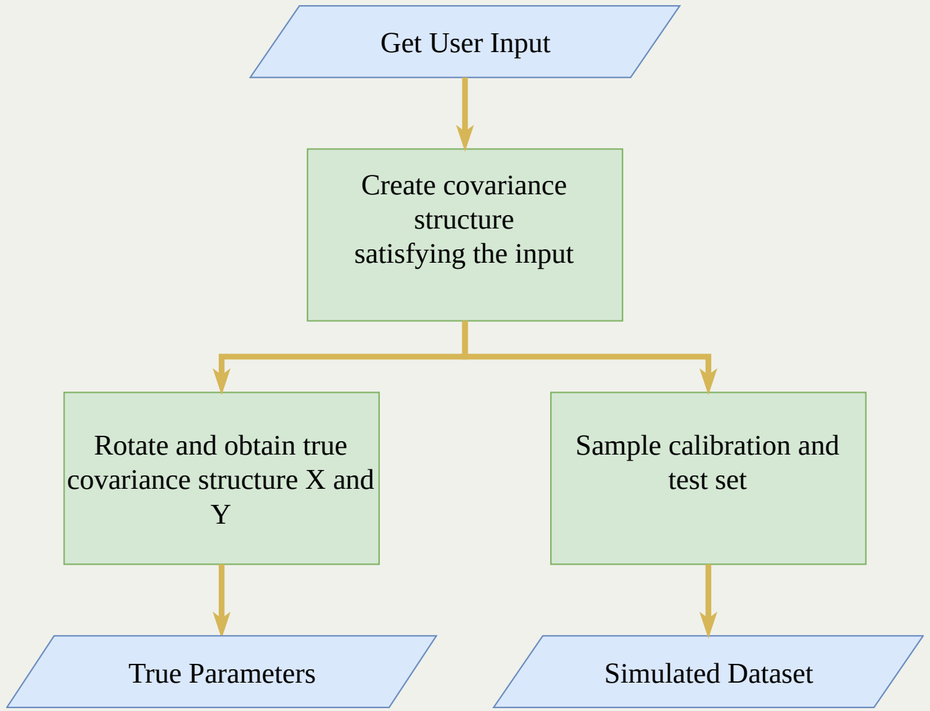
- It can simulate diverse nature of data with very few parameters

How it works

How it works

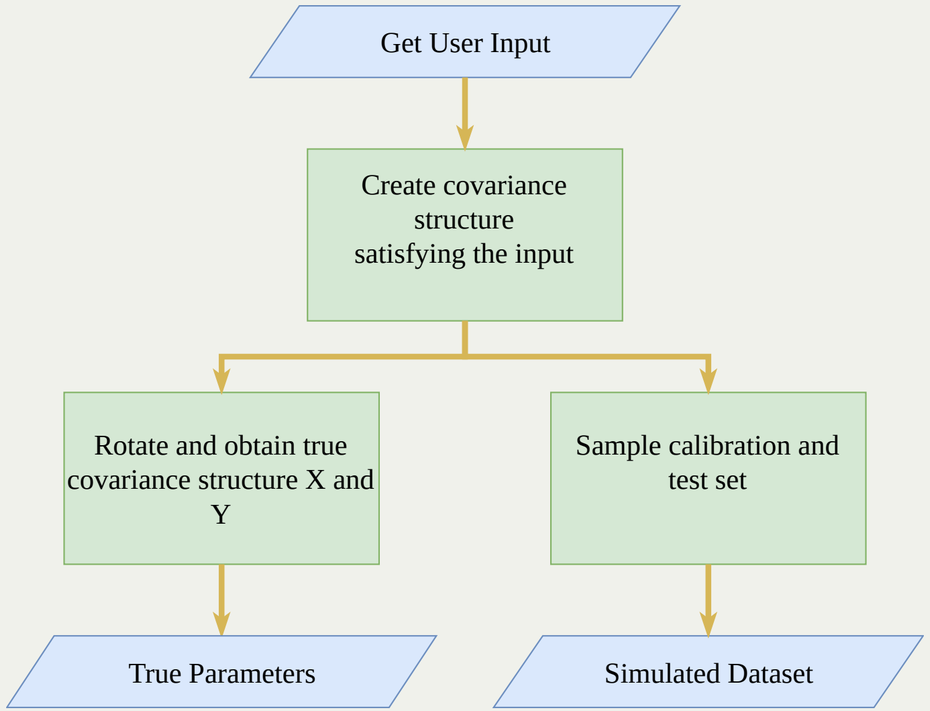

Get User Input
Create covariance structure satisfying the input
Rotate and obtain true covariance structure X and Y
Sample calibration and test set
True Parameters
Simulated Dataset

How it works



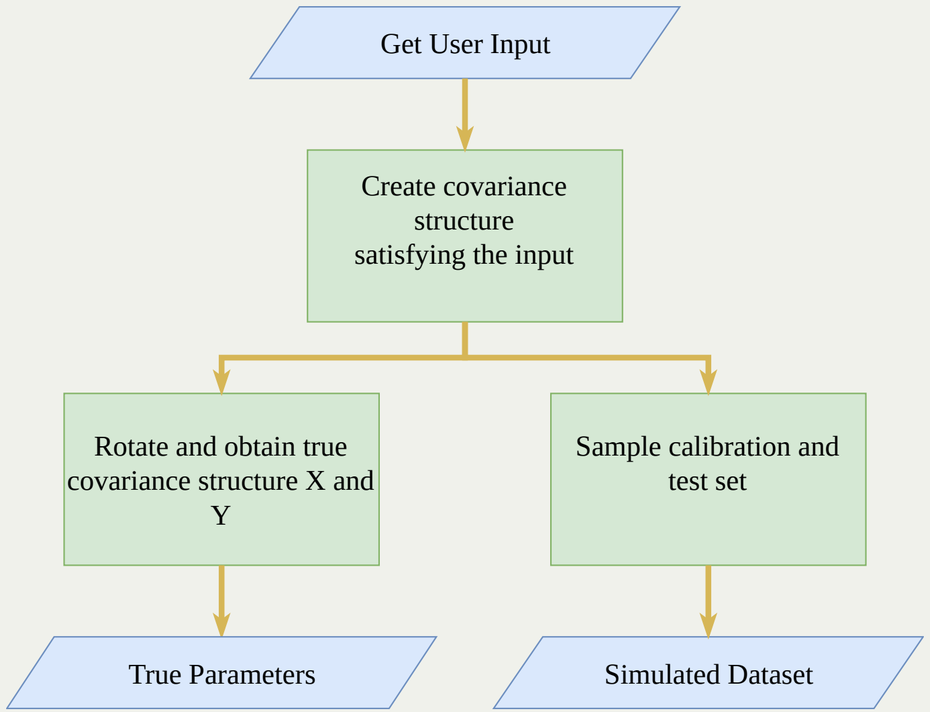
- Collect input parameters from user

How it works



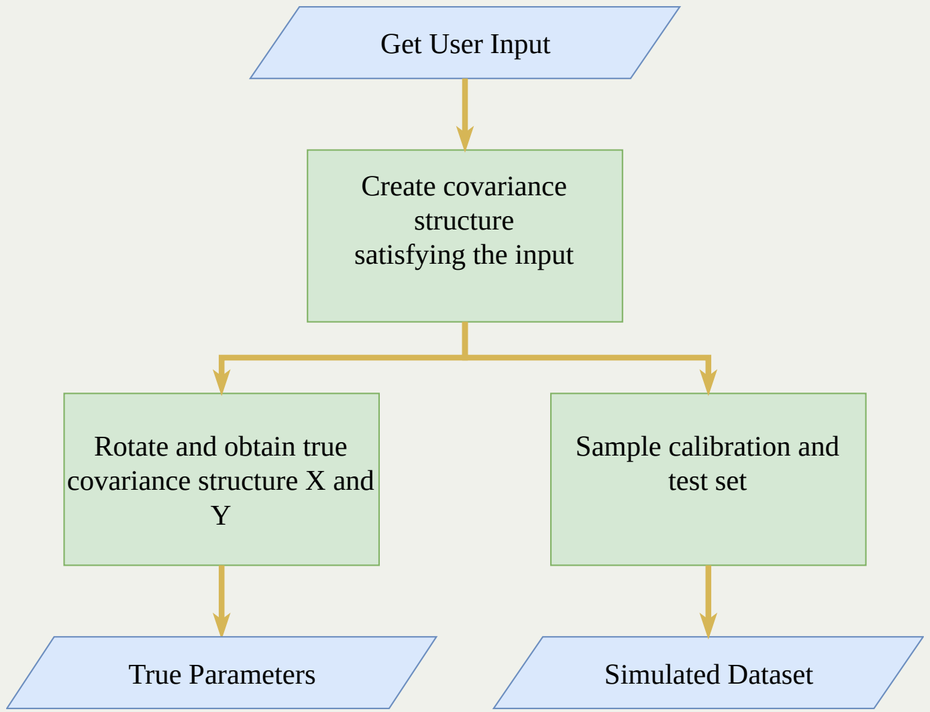
- Collect input parameters from user
- Make a covariance matrix satisfying those input parameters

How it works



- Collect input parameters from user
- Make a covariance matrix satisfying those input parameters
- Computes true population properties such as regression coefficients

How it works



- Collect input parameters from user
- Make a covariance matrix satisfying those input parameters
- Computes true population properties such as regression coefficients
- Sample calibration and validation sets

Demonstration

SimulatR NMBU

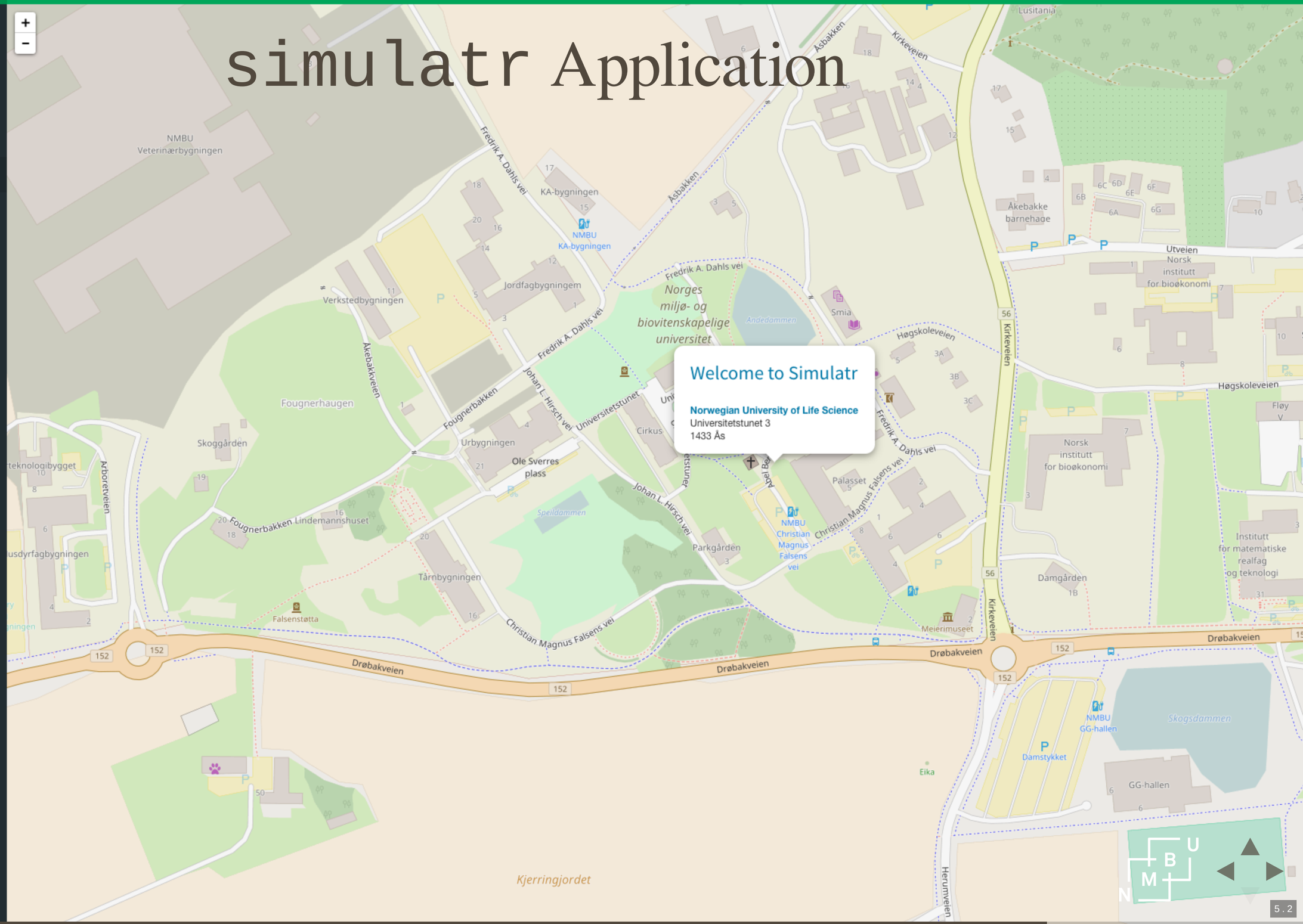
New Seed

Type of simulation:

Bivariate Simulation

- Parameter Settings
- Simulation Overview
- Estimation
- Model Comparison

simulatr Application



5.2

Salamat
Dakujem
GRACiAS
ধন্যবাদ
TAKK
ASANTE
TAK
Teşekkür ederim
SUWUN
hvala
Salamat
HVALA
mersi
شكرا جزيل
Ευχαριστώ
감사합니다
GRAZZii
GRAZAS
kiitos
merci
DANKE
Salamat
TAKK
MAHALO
Рахмет
kiitos
ARiGATO
Thank You
arigato
takk
suwun
धन्यवाद
HVALA
GRAZiE
DAKUJEM
MERCi
Teşekkür ederim
kiitos
hvala
mahalo
GRACiAS
GRAZAS
TAKK
ありがとう
DANKE
Salamat
SUWUN
Благодарам
TAKK
ASANTE
多謝
SALAMAT
gracias
спасибо

References

References

Cook, R., Helland, I., & Su, Z. (2013). Envelopes and partial least squares regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *75*(5), 851–877.

Helland, I. S., Sæbø, S., & Tjelmeland. (2012). Near optimal prediction from relevant components. *Scandinavian Journal of Statistics*, *39*(4), 695–713.

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*, *146*, 128–135.