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

Raju Rimal, Trygve Almøy, Solve Sæbø[1,*]

Abstract

Data science is generating enormous amounts of data, and new and advanced analytical methods are constantly being developed to cope with the challenge of extracting information from such "big-data". Researchers often use simulated data to assess and document the properties of these new methods, and in this paper we present simrel-m, which is a versatile and transparent tool for simulating linear model data with extensive range of adjustable properties. The method is based on the concept of relevant components Helland and Almøy (1994), which is equivalent to the envelope model Cook et al. (2013). It is a multi-response extension of simrel Sæbø et al. (2015), and as simrel the new approach is essentially based on random rotations of latent relevant components to obtain a predictor matrix $\mathbf{X}$, but in addition we introduce random rotations of latent components spanning a response space in order to obtain a multivariate response matrix $\mathbf{Y}$. The properties of the linear relation between $\mathbf{X}$ and $\mathbf{Y}$ are defined by a small set of input parameters which allow versatile and adjustable simulations. Sub-space rotations also allow for generating data suitable for testing variable selection methods in multi-response settings. The method is implemented as an R-package which serves as an extension of the existing simrel packages Sæbø et al. (2015).

Keywords: simrel-2.0, simrel package in r, data simulation, linear model, simrel-m,

Introduction

General aspects

Technological advancement has opened a door for complex and sophisticated scientific experiments that was not possible before. Due to this change, enormous amounts of raw data are generated which contains massive information but difficult to excavate. Finding information and performing scientific research on these raw data is now becoming another problem. In order to tackle this situation new methods are being developed. However, before implementing any methods, it is essential to test its performance. Often, researchers use simulated data for the purpose which itself is a time-consuming process. The main focus of this paper is to present a simulation method, along with an r-package called

[*] Dep. of Chemistry and Food Science, NMBU, Ås (nmbu.no)

simrel-m, that is versatile in nature and yet simple to use.

The method is based on principal of relevant space for prediction which assumes that there exists a subspace in the complete space of responses that is spanned by a subset of eigenvectors of predictor variables. The method and the r-package based on this principle not only has ability to simulate wide range of multi-response linear model data but also let researcher to specify which components of predictors ($\mathbf{X}$) are relevant for a component of responses $\mathbf{Y}$. This enables the possibility to construct data for evaluating methods developed for variable selection.

A vast literature on simulation is present but most of them are developed to address the specific problems their study was dealing with. Sæbø et al. (2015) has presented a generic tool that is capable of simulating linear model data as an r-package simrel. This paper extends the methods to simulate multivariate response.

Application of simrel-m

The simple interface of simrel-m for sophisticated simulation opens for numerous applications in various disciplines among which some of them are discussed below.

Educational Purpose: Explaining multivariate statistics and edify people is a difficult and strategic task. Instructors spend lots of time just to find suitable datasets for explaining some issue. For instance –

Model and methods testing: Imagine a situation where a researcher is collecting enormous amount of data which takes both time and money. Before going into extensive sampling, a pilot project is started and samples were collected using various techniques. Each of them are modelled with different estimation methods. The researcher would like to compare the sampling methods or estimation procedures that will the suitable for the final project. A simulated dataset based on the pilot project may help to identify the most appropriate sampling methods or estimation procedures.

Understanding and developing multivariate statistics: New estimation methods are being developed to address mordern and complicated situations. A new method or technique could be difficult to understand. For example, the envelope model Cook and Zhang (2015), a recent estimation technique based on maximum likelihood, attempts to find a response envelope (relevant subspace) that contains all the information that the corresponding predictors can explain. Here, one can make use of simrel-m to simulate data with underlying informative response space. In addition, new methods such as CPLS based on PLS are steadily being developed.

Understanding population latent structure enables assessment of such models.

Model Specification

A multi-response multivariate general linear model in equation-(2) is conidered as a simulation model.

Being an extension of `simrel` package, a quick summary of the procedure used in that package helps to underestand the literature in this paper.

An overview of `simrel`

`Simrel` is based on uni-response linear model as in equation~(1).

$$\begin{bmatrix} y \\ \mathbf{X} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_y \\ \boldsymbol{\mu}_X \end{bmatrix}, \begin{bmatrix} \sigma_y^2 & {\boldsymbol{\sigma}_{Xy}}^t \\ \boldsymbol{\sigma}_{Xy} & \boldsymbol{\Sigma}_{XX} \end{bmatrix}\right) \tag{1}$$

$$\mathbf{Y} = \boldsymbol{\mu}_Y + \boldsymbol{B}^t(\mathbf{X} - \boldsymbol{\mu}_X) + \boldsymbol{\epsilon} \tag{2}$$

where $\mathbf{Y}$ is a response matrix with m response vector $y_1, y_2, \ldots y_m$, $\mathbf{X}$ is multivariate predictor matrix with p predictor variables and the random error term $\boldsymbol{\epsilon}$ is assumed to follow $N(\mathbf{0},\ \boldsymbol{\Sigma}_{Y|X})$. Equivalently,

$$\begin{bmatrix} \mathbf{Y} \\ \mathbf{X} \end{bmatrix} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = N\left(\begin{bmatrix} \boldsymbol{\mu}_Y \\ \boldsymbol{\mu}_X \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{YY} & \boldsymbol{\Sigma}_{XY}^t \\ \boldsymbol{\Sigma}_{XY} & \boldsymbol{\Sigma}_{XX} \end{bmatrix}\right) \tag{3}$$

Here,

$\boldsymbol{\Sigma}_{YY}$: Covariance Matrix of response $\mathbf{Y}$ without given $\mathbf{X}$
$\boldsymbol{\Sigma}_{XY}$: Covariance Matrix between $\mathbf{X}$ and $\mathbf{Y}$
$\boldsymbol{\Sigma}_{XX}$: Covariance matrix of predictor variables $\mathbf{X}$
$\boldsymbol{\mu}_X$ and $\boldsymbol{\mu}_Y$: Mean vectors of response $\mathbf{Y}$ and predictor $\mathbf{X}$ respective

According to the theory of Multivariate Normal Distribution, we can express different parameters interms of $\mathbf{X}$, $\mathbf{Y}$ and the covariance structure.

Model Parameterization

`Simrel-m` uses model parameterization which is based on the concept of relevant components Helland and Almøy (1994) where it is assumed that a subspace of response $\mathbf{Y}$ is spanned by a subset of eigenvectors corresponding to predictor space. A response space can be thought to have two mutually orthogonal space – relevant and irrelevant. Here the relevant space of response matrix is termed as response components, and we assume that each response component is spanned by an exclusive subset of predictor variables. In this way we can construct a set of predictor variables which has non-zero regression coefficients. This also enables user to have uninformative predictors which can be detected during variable selection procedure. In addition, user can control signal-to-noise ratio for each response components with a vector of population coefficient of determination $\rho_1, \ldots, \rho_q$. Further, the collinearity between predictor variables can also be controlled by a factor γ which guides the decay pattern of eigenvalue of $\mathbf{X}$ matrix. Helland and Almøy (1994) showed that if the direction of large variablity (i.e., component corresponding to large eigenvalues) are also relevant relevant predictor space, prediction is relatively easy. In contrast, if the relevant predictors are on the direction of low varibility, prediction becomes difficult.

Parameter Definition:

Before continuing any further, it is necessary to define the parameters used here,

Table 1: Parameters for simulation used in this study

Parameters	Description
n	number of observations
p	number of predictors
q	numbers of relevant predictors for each response components
l	number of response components
m	number of response
$\mathcal{P}$	set of position index of relevant components for each response components
γ	degree of collinearity, factor that control the decrease of eigenvalue of $\mathbf{X}$
$\mathcal{S}$	set of index of response components for simulation orthogonal rotation to simulation m responses
$\mathcal{R}^2$	population coefficient determination for each response components

In the following section, some of these parameters are discussed in detail. The discussion has considered random **X** regression model with m response as given in equation~(2).

Parameters Explanation and Notation Used:

Let m responses are spanned completely by l response components. These l response components are combined with $m - l$ standard normal vectors by user defined criteria to get m responses after successive orthogonal transformation. Out of q, let $q_j,\ j = 1, \ldots l$ be the number of predictors that is relevant for response j. Let c_j be the number of eigenvectors/ components that completely span j^{th} predictor space containing q_j number of predictors. Further, the position of these components for j^{th} response be in index set $\mathcal{P}_j$. Here it is also assumed that the eigenvalues corresponding to **X** declines successively such that $\lambda_i, i = 1, \ldots, p$ such that $\lambda_i \geq \lambda_k, i > k$ are the eigen values of **X**. The position index of eigenvalues corresponding to response j is in the set $\mathcal{P}_j$. We assume that the index are ordered within each sets so that, j^{th} index set contains c_j number of components. The eigenvalues corresponding to these components in $\mathcal{P}_j$ set is $\lambda_{\mathcal{P}_{jk}}, k = 1, \ldots c_j$ such that $\lambda_{\mathcal{P}_{jk}} > \lambda_{\mathcal{P}_{jk'}}$ for $k > k'$. In `Simrel-M` package, we refer this position by `relpos` argument. In addition, we suppose that the relevant components are exclusive for each response.

An Example:

Suppose we have a situation like,

Number of response	(m)	=	5
Number of response components	(l)	=	3
Position of relevant component for response 1	$(\mathcal{P}_1)$	=	$\{1,3\}$
Position of relevant component for response 2	$(\mathcal{P}_2)$	=	$\{2,4,5\}$
Position of relevant component for response 3	$(\mathcal{P}_3)$	=	$\{6\}$

such that, $\lambda_1 > \lambda_3$ in $\mathcal{P}_1$ and $\lambda_2 > \lambda_4 > \lambda_5$ in set $\mathcal{P}_2$. Here, the component (eigenvector) 1 and 3 are relevant for response component 1, component 2, 4 and 5 are relevant for response component 2 and component 6 is relevant for response component 3.

In `Simrel-M`, we have assumed that the eigenvalues are decreasing exponentially by factor γ and the largest eigenvalue is 1, i.e. for $\gamma > 0$, $\lambda_i = \mathrm{e}^{-\gamma(i-1)}$ for $i = 1, 2, \ldots p$.

Statistical Model

Regression Coefficients $(\boldsymbol{B})$ The measurement of effect of each predictor variables in the individual response are the regression coefficients. Predictors having regression coefficients closer to zero have lower impact on the response.

$$\mathbf{B} = \boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1} \tag{4}$$

Error Variance $\left(\boldsymbol{\Sigma}_{Y|X}\right)$ Error variance comprises the variation that is present in response but is not explained by the model. The true error variance can be written as,

$$\boldsymbol{\Sigma}_{Y|X} = \boldsymbol{\Sigma}_{YY} - \boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1}\boldsymbol{\Sigma}_{XY} \tag{5}$$

Population Coefficient of Determination $(\boldsymbol{\mathcal{R}}_{XY})$ Coefficient of determination measure how much the predictors explains the variation present in response matrix. The coefficient of determination in the case of **X** and **Y** relationship is,

$$\boldsymbol{\mathcal{R}}_{XY}^2 = \boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1}\boldsymbol{\Sigma}_{XY}\boldsymbol{\Sigma}_{YY}^{-1} \tag{6}$$

Simulation of $(\mathbf{Y}, \mathbf{X})$ for model~(3) requires the fact that – a set of latent variable spanning **X** and **Y** will contain same information in different structure. With two matrices $\mathbf{R}_{p\times p}$ and $\mathbf{Q}_{q\times q}$ with rank p and q respectively, lets define a transformation as $\mathbf{Z} = \mathbf{RX}$ and $\mathbf{W} = \mathbf{QY}$ so that,

$$\begin{aligned}
\begin{bmatrix} \mathbf{W} \\ \boldsymbol{Z} \end{bmatrix} &\sim N\left(\begin{bmatrix} \boldsymbol{\mu}_W \\ \boldsymbol{\mu}_Z \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{WW} & \boldsymbol{\Sigma}_{WZ}^t \\ \boldsymbol{\Sigma}_{ZW} & \boldsymbol{\Sigma}_{ZZ} \end{bmatrix}\right) \\
&= N\left(\begin{bmatrix} \boldsymbol{Q}\boldsymbol{\mu}_Y \\ \boldsymbol{R}\boldsymbol{\mu}_X \end{bmatrix}, \begin{bmatrix} \boldsymbol{Q}\boldsymbol{\Sigma}_{YY}\boldsymbol{Q}^t & \boldsymbol{Q}\boldsymbol{\Sigma}_{XY}^t\mathbf{R}^t \\ \boldsymbol{R}\boldsymbol{\Sigma}_{XY}\boldsymbol{Q}^t & \boldsymbol{R}\boldsymbol{\Sigma}_{XX}\mathbf{R}^t \end{bmatrix}\right)
\end{aligned} \tag{7}$$

Further, if both **Q** and **R** are orthogonal matrix their inverse equales to their transpose, i.e. $\mathbf{Q}^t\mathbf{Q} = \mathbf{I}_q$ and $\mathbf{R}^t\mathbf{R} = \mathbf{I}_p$. The inverse transformation can be defined as,

$$\begin{aligned}
\boldsymbol{\Sigma}_{XX} = \boldsymbol{R}^t\boldsymbol{\Sigma}_{ZZ}\boldsymbol{R}, \boldsymbol{\Sigma}_{XY} = \boldsymbol{R}^t\boldsymbol{\Sigma}_{ZW}\boldsymbol{Q} \\
\boldsymbol{\Sigma}_{YY} = \boldsymbol{Q}^t\boldsymbol{\Sigma}_{WW}\boldsymbol{Q}, \boldsymbol{\Sigma}_{YX} = \boldsymbol{Q}^t\boldsymbol{\Sigma}_{WZ}\boldsymbol{R}
\end{aligned} \tag{8}$$

The relationship above allows us to define the linear model relation in equation~(2) in terms of **W** and *athbfZ* as in equation~(9),

$$\mathbf{W} = \boldsymbol{\mu}_W + \mathbf{A}^t \left(\mathbf{Z} - \boldsymbol{\mu}_Z\right) + \boldsymbol{\tau}; \qquad \boldsymbol{\tau} \sim N\left(\mathbf{0}, \boldsymbol{\Sigma}_{W|Z}\right) \tag{9}$$

Here, in this setting the model parameters can be defined as follows where each of them can be related to the model parameter for model in equation~(2) using the ortogonal matrices **P** and **Q**.

Regression Coefficients $(\mathbf{A})$ Using the transformation matrix **P** and **Q**, we can obtain the regression coefficients corresponding to the latent structure of predictors.

$$\mathbf{A} = \boldsymbol{\Sigma}_{WZ}\boldsymbol{\Sigma}_{ZZ}^{-1} = \boldsymbol{Q}\boldsymbol{\Sigma}_{YZ}\mathbf{R}^t\left[\boldsymbol{R}\boldsymbol{\Sigma}_{XX}\mathbf{R}^t\right]^{-1} = \boldsymbol{Q}\boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1}R^t = \mathbf{QBR}^t \tag{10}$$

Error Variance $\left(\boldsymbol{\Sigma}_{W|Z}\right)$ The true unexplained variation present in population that the model between the latent structure of predictors and responses can also be expressed with the rotation matrix and the error variance $\Sigma_{Y|X}$.

$$\begin{aligned}
\boldsymbol{\Sigma}_{W|Z} &= \boldsymbol{Q}\boldsymbol{\Sigma}_{YY}\mathbf{Q}^t - \boldsymbol{Q}\boldsymbol{\Sigma}_{YX}\mathbf{R}^t\left[\boldsymbol{R}\boldsymbol{\Sigma}_{XX}\boldsymbol{R}^t\right]^{-1}\boldsymbol{R}\boldsymbol{\Sigma}_{XY}\mathbf{Q}^t \\
&= \boldsymbol{Q}\boldsymbol{\Sigma}_{YY}\mathbf{Q}^t - \boldsymbol{Q}\boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1}\boldsymbol{\Sigma}_{XY}\mathbf{Q}^t \\
&= \mathbf{Q}\left[\boldsymbol{\Sigma}_{YY} - \boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1}\boldsymbol{\Sigma}_{XY}\right]\mathbf{Q}^t \\
&= \mathbf{Q}\boldsymbol{\Sigma}_{Y|X}\mathbf{Q}^t
\end{aligned} \tag{11}$$

Since, **Y** is generated from orthogonal random rotation matrix **Q**, the error variance depends on **Q**. Further, the coefficient of determination is,

Population Coefficient of Determination $\left(\boldsymbol{\mathcal{R}}_{WZ}^2\right)$

$$\begin{aligned}
\boldsymbol{\mathcal{R}}_{XY}^2 &= \mathbf{Q}^t\boldsymbol{\Sigma}_{WZ}\boldsymbol{\Sigma}_{ZZ}^{-1}\boldsymbol{\Sigma}_{ZW}\boldsymbol{\Sigma}_{WW}^{-1}\boldsymbol{Q} \\
&= \mathbf{Q}^t\boldsymbol{\mathcal{R}}_{WZ}^2\mathbf{Q} \\
\text{i.e. } \boldsymbol{\mathcal{R}}_{WZ}^2 &= \mathbf{Q}\boldsymbol{\mathcal{R}}_{XY}^2\mathbf{Q}^t
\end{aligned} \tag{12}$$

Thus, on the basis of these mathematical backgrounds, the simulation strategy follows,

a) Construct covariance structure of $\mathbf{W}$ and $\mathbf{Z}$ satisfying given parameters

b) Simulate $\mathbf{W}$ and $\mathbf{Z}$ from random standard normal distribution

c) Rotation $\mathbf{Z}$ by orthonormal matrix $\mathbf{R}$ to yield $\mathbf{X} = \mathbf{R}^t\mathbf{Z}$

d) Rotation of $\mathbf{W}$ by orthogonal matrix $\mathbf{Q}$ to yield $\mathbf{Y} = \mathbf{Q}^t\mathbf{W}$

e) For simplification, we assume that no common components of $\mathbf{X}$, i.e. $\mathbf{Z}$, relevant for $\mathbf{W}$. For example, if component 1 and component 2 are relevant for $\mathbf{W}_1$, they are not relevant for other $\mathbf{W}$'s.

Model parameterization and relevant components

Eigenvalue decomposition principal states that a variance-covariance matrix $\mathbf{\Sigma}$ can be decomposed as,

$$\Sigma = \mathbf{E}\Lambda\mathbf{E}^t \tag{13}$$

where, $\mathbf{E} = (\mathbf{e}_1, \mathbf{e}_2, \ldots \mathbf{e}_p)$ is an orthogonal matrix of eigenvectors and Λ is a diagonal matrix of eigenvalues $\lambda_1 \leq \lambda_2 \leq \ldots \lambda_p$. From expression in equation~(8), $\mathbf{\Sigma}_{XX}$ and $\mathbf{\Sigma}_{WW}$ can have similar decomposition with some suitable choice of orthonormal matrix $\mathbf{R}$ and $\mathbf{Q}$ respectively.

In this study, all the components of $\mathbf{Y}$, i.e. $\mathbf{W}$ are considered to be uncorrelated. Since, the component structure also contains the irrelevant components, each of their correlation with others are considered to be zero. Hence, the unconditional covariance structure for the component matrix ($\mathbf{W}$) is $\mathbf{I}_m$. Furthermore, if $\mathbf{\Sigma}_{ZZ} = \Lambda = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_p)$, where $\lambda_i, i = 1, \ldots p$ are eigenvalues of $\mathbf{X}$, the expression in equationeqrefeq:cov-yx-wz helps to simulate $\mathbf{X}$ from $\mathbf{R}$, the orthonormal rotation matrix and its eigen structure $\mathbf{\Sigma}_{ZZ}$. Similarly from $\Sigma_{WW} = \mathbf{I}_m$ and rotaion matrix $\mathbf{Q}$, we can simulate $\mathbf{Y}$.

Let $\mathbf{W}_1, \ldots, \mathbf{W}_l$ are the components of Y that are relevant to $\mathbf{Z}$ and consequently $\mathbf{X}$, $\mathbf{W}_{l+1}, \ldots, \mathbf{W}_q$ are not the outcome of $\mathbf{Z}$, the principal components of $\mathbf{Z}$ that are relevant for $\mathbf{W}$ are applicable for $\mathbf{W}_1, \ldots, \mathbf{W}_l$ only. The covariance matrix of $\mathbf{W}$ and $\mathbf{Z}$ ($\mathbf{\Sigma}_{WZ}$) is constructed referring to the terminology in Helland and Almøy (1994) that the principal components are termed as relevant for which $\mathbf{\Sigma}_{WZ}$ are non-zero.

Assume $a_1, \ldots, a_l$ number of principal components of $\mathbf{X}$ are relevant to $\mathbf{W}_1, \ldots, \mathbf{W}_l$ respectively. Let $\mathcal{P}_1, \ldots, \mathcal{P}_l$ are the sets of positions of these components, then $(\Sigma_{WZ})_{ij} \neq 0$ if $j \in \mathcal{P}_i$, $i = 1, \ldots, l$ and zero otherwise. This follows us to the matrix of regression coefficients as,

$$\mathbf{A} = \begin{cases} \boldsymbol{\Sigma}_{WZ}\boldsymbol{\Sigma}_{ZZ}^{-1} = \sum_{j\in\mathcal{P}_i}\left(\frac{\sigma_{ij}}{\lambda_j}\mathbf{t}_j\right) & \text{for } i = 1,\ldots,l \\ 0 & \text{otherwise} \end{cases} \tag{14}$$

where, $\mathbf{t}_j$ is a p-vector with 1 at position j and zero otherwise. As in the previous version of simrel by Sæbø et al. (2015), eigenvalues of $\boldsymbol{\Sigma}_{XX}$ is assumed to be different and has adopted the parametric representation as $\lambda_j = e^{-\nu(j-1)}$ for $\nu > 0$ and $j = 1,\ldots p$. Here, the parameter ν regulates the decline of $\lambda_j, j = 1,\ldots p$. Without loss of generality, for further simplification, the first and largest eigenvalues are set to one.

For complete parametrization of the matrix Ξ_{WZ} in equation~(7), covariances between W and Z ($\boldsymbol{\Sigma}_{WZ}$) should be constructed such that it is positive definite and satisfy the relation,

$$\begin{aligned} \boldsymbol{\mathcal{R}}_{WZ}^2 &= \boldsymbol{\Sigma}_{WZ}\boldsymbol{\Sigma}_{ZZ}^{-1}\boldsymbol{\Sigma}_{ZW}\boldsymbol{\Sigma}_{WW}^{-1} \\ \text{i.e, } \boldsymbol{\mathcal{R}}_{WZ}^2\boldsymbol{\Sigma}_{WW} &= \boldsymbol{\Sigma}_{WZ}\boldsymbol{\Lambda}^{-1}\boldsymbol{\Sigma}_{ZW} \end{aligned} \tag{15}$$

For given $\boldsymbol{\mathcal{R}}_{WZ}^2$ and $\Sigma_{WW} = \mathbf{I}_m$, equation~(15) will be satisfied for some $\boldsymbol{\Sigma}_{WX}$ whose rows correspond to the relevant components for $\mathbf{W}$. As we have considered the situation that no relevant components are common, elements in $\boldsymbol{\Sigma}_{WZ}$ are sampled from a uniform distribution $\mathcal{U}(-1,1) = \{s_{\mathcal{P}_{1i}}, s_{\mathcal{P}_{2i}}, \ldots s_{\mathcal{P}_{pi}}\}$, for each $i = 1,\ldots q$ as in Sæbø et al. (2015) such that,

$$(\boldsymbol{\sigma}_{WZ})_{ij} = \text{sign}\left(s_{ij}\right)\sqrt{\frac{\boldsymbol{\mathcal{R}}_{WZ}^2 \cdot \left|s_{ij}\right|}{\sum_{k\in\mathcal{P}_i}\left|s_{ik}\right|}\lambda_j}$$

for $j \in \mathcal{P}_i$ and for each $i = 1,\ldots q$

Data Simulation

After the construction of Ξ_{WZ}, n samples are generated from standard normal distribution of$(\mathbf{W},\mathbf{Z})$ considering their mean to be zero, i.e.$\boldsymbol{\mu}_W = 0$ and$\boldsymbol{\mu}_Z = 0$. SinceΞ_{WZ} is positive definite,$\Xi_{WZ}^{1/2}$ obtained from its Cholesky decomposition, can serve as one of its square root. The simulation process constitute of following steps,

1) A matrix $\mathbf{U}_{n\times(p+q)}$ is sampled from standard normal distribution

2) Compute $\mathbf{G} = \boldsymbol{U}\Xi_{WZ}^{1/2}$

Here, first m columns of $\mathbf{G}$ will serve as $\mathbf{W}$ and remaining p columns will serve as $\mathbf{Z}$. Further, each row of $\mathbf{G}$ will be a vector sampled independently from joint normal distribution of $(\mathbf{W}, \mathbf{Z})$. The final step to generate $\mathbf{X}$ and $\mathbf{Y}$ from $\mathbf{Z}$ and $\mathbf{W}$ requires corresponding rotation matrices which is discusses on following section.

Rotation of predictor space

Simulation of predictor variables from principal components requires a construction of a rotation matrix $\mathbf{R}$ that defines a new basis for the same space as is spanned by the principle components. As any rotation matrix can be considered as $\mathbf{R}$, an eigenvalue matrix from eigenvalue decomposition of $\mathbf{\Sigma}_{XX}$ can be a candidate. Since simulation is a reverse engineering, the underlying covariance structure for the predictors are unknown. So, the method is free to construct a real valued orthogonal matrix that can serve for the purpose.

Among several methods (Anderson et al., 1987; Heiberger, 1978) to generate random orthogonal matrix the same method as is used in Sæbø et al. (2015) is implemented here. The $\mathcal{Q}$ matrix obtained from QR-decomposition of a matrix filled with standard normal variates can serve as the rotation matrix $\mathbf{R}$.

The rotation can be a) unrestricted and b) restricted. The former one rotates all p predictors making them some what relevant for the all response conponents and consequently all responses. However, only $q_i \leq p$ predictors are relevant for for i^{th} response component, the resticted rotation is implemented in `simrel-M`. This also ensure that $p - q_i$ predictors does not contribute anything on response component i and consequently the simulated data can also be used for testing variable selection methods.

Rotation of response space

`Simrel-M` has considered an exclusive relevant predictor space for each response components, i.e. a set of predictor variables only influence one response component. However, it allows user to simulate more response variable than response components. In this case, noise are added during the orthogonal rotation of response components. For example, if user wants to simulation 5 response variation from 3 response components. Two standard normal vectors are combined with response components and rotated simultaneously. The

Table 3: Parameter setting of simulated data for model comparison

Parameters	Design1	Design2
Number of observations	100	50
Number of predictors	20	40
Relevant number of predictors	5, 5, 7	15, 15, 9
Number of responses	6	5
Position of relevant components	1, 2, 3, 4, 6, 5	1, 2, 3, 4, 5, 7
Gamma (decay of eigenvalues)	0.9	0.4
Coefficient of Determination	0.8, 0.9, 0.7	0.8, 0.9, 0.5
Position of response components	1, 4, 2, 6, 3, 5	1, 4, 2, 5, 3
Number of test observations	1000	1000

rotation can be both restricted and unrestricted as discussed in previous section. The restricted rotation is carried out combining response vectors along with noise vector in a block-wise manner according to the users choice. Illustration in fig-...

Suppose, in our previous example, if response components are combined as – $\mathbf{W}_1, \mathbf{W}_4$, $\mathbf{W}_2$ and $\mathbf{W}_3, \mathbf{W}_5$. Here, any predictor variable is only relevant for $\mathbf{W}_1, \mathbf{W}_2$ and $\mathbf{W}_3$ while $\mathbf{W}_4$ and $\mathbf{W}_5$ are noise. The resulting response variables are $\mathbf{Y}_1 \ldots \mathbf{Y}_5$ where, the first and fourth response variable spans the same space as by the first response components $\mathbf{W}_1$ and noise component $\mathbf{W}_4$ and so on. Thus, the predictors and predictor space relevant for response component $\mathbf{W}_1$ is also relevant for response $\mathbf{Y}_1$ and $\mathbf{Y}_4$.

Implementation

Example of model comparison with simulated data from `simrel-m` package

In this section, `simrel-m` is implemented to simulate multi-response linear model dataset and use it to compare Principal Components Regression (PCR), Partial Least Squared Regression (PLS), Cannonically Powered Partial Least Squared Regression (CPPLS), Maximum Likelihood under Envelope Estimation and Ordinary Least Squared Regression (OLS) on the basis of their prediction ability of test observations. Here two design of parameters as in Table-3 are implemented for the simulation.

Here, the first design has large number of observations as compared to the number of variables while the number of variables in second design is nearly equals to its number of observations. Both the models have three informative response components from which first design generates 6 responses and the second design generates 5 responses. In addition, the eigenvalues of predictors decreases sharply in the first design than the second one. The prediction error are compared on the basis of 1000 test samples.

Prediction error is measured as mean squared error of prediction (MSEP) using the expression in equation~(16).

$$\text{MSEP}_{\text{train}} = \frac{\mathbf{Y}_{\text{train}} - \hat{\mathbf{Y}}_{\text{train}}}{n_{\text{train}}} \text{ and } \text{MSEP}_{\text{test}} = \frac{\mathbf{Y}_{\text{test}} - \hat{\mathbf{Y}}_{\text{test}}}{n_{\text{test}}}$$

where, $\hat{\mathbf{Y}}_{\text{train}} = \mathbf{X}_{\text{train}}\boldsymbol{\beta}$ and $\hat{\mathbf{Y}}_{\text{test}} = \mathbf{X}_{\text{test}}\boldsymbol{\beta}$

Comparison of Estimation Methods based on Prediction error

Figure-1 shows the performance of different models with components 1 to 10 . . .

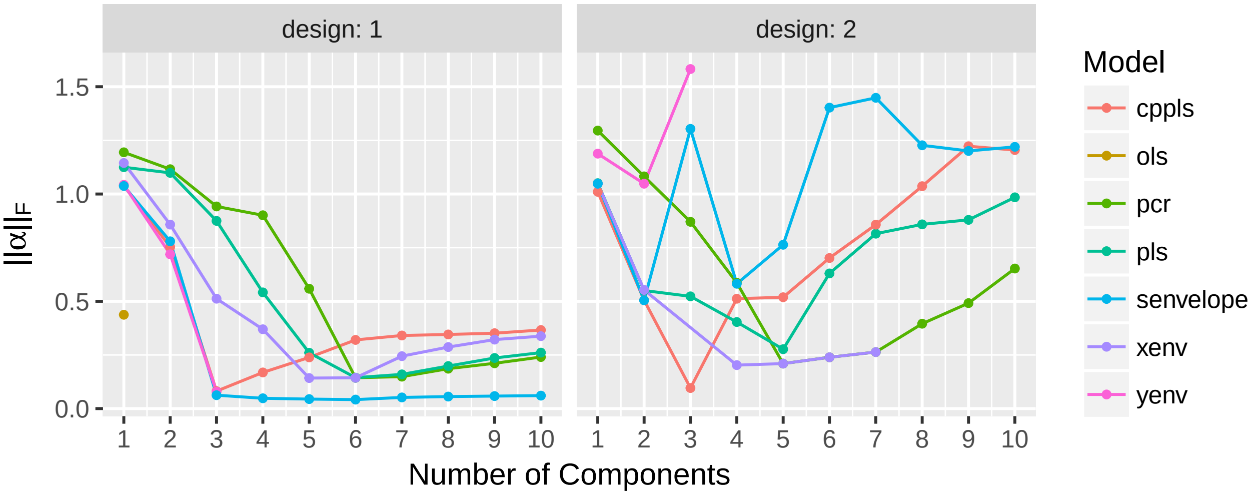

Figure 1: Model comparison based on test prediction error

Some more text and more and more

Appendix

Comprehensive Explanation of Notation

Following table present a comprehensive notation used in this paper.

Table 4: Notation used in this study

Symbol Definition	Symbols
Response	$y_1, y_2, \ldots, y_m$
Response Components	$W_1, W_2, \ldots, W_l, W_{l+1} \ldots, W_m$
Number of Predictors relevent for each response components	$q_1, q_2, \ldots, q_l, 0, \ldots, 0$
Number of **X** components relevant for each response components	$c_1, c_2, \ldots, c_l, 0, \ldots, 0$
Set of position index of relevant components	$\mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_l, \varnothing, \ldots, \varnothing$
Coefficient of determination for each response components	$\mathcal{R}_1^2, \mathcal{R}_2^2, \ldots, \mathcal{R}_l^2, 0, \ldots, 0$
Set of extra components index sampled randomly	$\mathcal{Q}_1, \mathcal{Q}_2, \ldots, \mathcal{Q}_l, \varnothing, \ldots, \varnothing$

Some properties and details of terms in Table~4 is explained below:

- Total number of relevant predictors ($\sum q_i$) is less than or equals to total number of predictor variables (p). Similarly, Total number of relevant predictor components ($\sum c_i$) is not greater than the total number of relevant predictor for that response component. In addition, the number of the relevant predictor components for each response never exceed the total number of relevant predictor for that component.

$$\sum q_i \leq p, \qquad \sum c_i \leq \sum q_i, \qquad c_i \leq q_i, \qquad \text{and ,} \qquad n(\mathcal{P}_i) = c_i$$

- In `simrel-m`, we have assumed that none of the predictor components are relevant for multiple response components, i.e.

$$\text{For } i \neq j, \; n(\mathcal{P}_i \cup \mathcal{P}_j) = 0$$

- Since, a predictor components can have lower dimension than predicotrs, i.e., a predictors can be spanned by lower dimensional subspace, the extra components required for orthogonal rotaion during simulation are sampled from the uninformative standard normal variates.

$$n(\mathcal{Q}_i) = q_i - c_i$$

- In simrel-m, the sampling is done without replacement.

$$n(\mathcal{P}_i \cup \mathcal{Q}_i) = q_i \text{ and } (\mathcal{P}_i \cup \mathcal{Q}_i) \cap (\mathcal{P}_j \cup \mathcal{Q}_j) = \varnothing$$

References

Theodore W Anderson, Ingram Olkin, and Les G Underhill. Generation of random orthogonal matrices. *SIAM Journal on Scientific and Statistical Computing*, 8(4):625–629, 1987.

R Dennis Cook and Xin Zhang. Foundations for envelope models and methods. *Journal of the American Statistical Association*, 110(510):599–611, 2015.

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

Richard M Heiberger. Algorithm as 127: Generation of random orthogonal matrices. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 27(2):199–206, 1978.

Inge S Helland and Trygve Almøy. Comparison of prediction methods when only a few components are relevant. *Journal of the American Statistical Association*, 89(426):583–591, 1994.

Solve Sæbø, Trygve Almøy, and Inge S Helland. 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*, 2015.