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 **X**, but in addition we introduce random rotations of latent components spanning a response space in order to obtain a multivariate response matrix **Y**. The properties of the linear relation between **X** and **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

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 has now become another problem. In order to tackle this situation new methods are being developed. However, before implementing any method, 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 simrel-m, that is versatile in nature and yet simple to use.

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

The simulation method we are discussing here is based on principal of relevant space for prediction (Helland and Almøy, 1994) which assumes that there exists a subspace in the complete space of response variables that is spanned by a subset of eigenvectors of predictor variables. The r-package based on this method lets user to specify various population properties such as which components of predictors ($\mathbf{X}$) are relevant for a component of responses $\mathbf{Y}$ and how the eigenvalues of $\mathbf{X}$ decreases. This enables the possibility to construct data for evaluating estimation methods and methods developed for variable selection.

Among several literatures in simulation (`which literatures`), Ripley (2009) has exhaustively discussed the topic. In addition, many literatures (`which literatures`) are available on studies which has implemented simulated data in order to investigate new estimation methods and prediction strategy (see: Cook and Zhang, 2015; Cook et al., 2013; Helland et al., 2012). However, most of the simulations in these studies is developed to address their specific problem. A systematic tool for simulating linear model data with single response, which could serve as a general tool for all such comparisons, was presented in Sæbø et al. (2015) and as r-package `simrel`. This paper extends `simrel` in order to simulate linear model data with multivariate response with an r-package `simrel-m`.

The r-package `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 space of response matrix for which the predictors are relevant 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 ρ^2. Further, the collinearity between predictor variables can also be adjusted by a factor γ which controls decay factor of eigenvalue of $\mathbf{X}$ matrix. Helland and Almøy (1994) showed that if the direction of large variability (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 variability, prediction becomes difficult. Table~1 shows all the parameters that a user can specify in `simrel-m`.

Table 1: Parameters for simulation used in this study

Parameters	Description
n	number of observations
p	number of predictor variables
q	numbers of relevant predictors for each latent component of response variables
l	number of informative latent component of response variables (response components)
m	number of response variables
γ	degree of collinearity (factor that control the decrease of eigenvalue of $\mathbf{X}$)
$\mathcal{P}$	position index of relevant predictor components for each response components
$\mathcal{S}$	position index of response components to combine relevant and irrelevant response variable
$\boldsymbol{\rho}^2$	population coefficient determination for each response components

Based on random regression model in equation~(3), we discuss some of these parameters in details.

Out of p predictor variables only q of them are relevant and has non-zero regression coefficients. If $\boldsymbol{e}_i,\ i = 1, 2, \ldots, p$ be the eigenvectors corresponding to $\mathbf{X}$, then, for some vector of $\boldsymbol{\eta}_j, j = 1, \ldots, m$ for m response variables,

$$\mathbf{B} = (\beta_{ij}) = \begin{bmatrix} \boldsymbol{\eta}_{11} & \cdots & \boldsymbol{\eta}_{1p} \\ \vdots & \ddots & \vdots \\ \boldsymbol{\eta}_{m1} & \cdots & \boldsymbol{\eta}_{mp} \end{bmatrix} \begin{bmatrix} \mathbf{e}_{11} & \vdots & \mathbf{e}_{1p} \\ \vdots & \ddots & \vdots \\ \mathbf{e}_{p1} & \vdots & \mathbf{e}_{pp} \end{bmatrix} \tag{1}$$

$$= \begin{bmatrix} \boldsymbol{\eta}_1 \\ \vdots \\ \boldsymbol{\eta}_m \end{bmatrix}_{m \times p} \begin{bmatrix} \mathbf{e}_1 & \ldots & \mathbf{e}_p \end{bmatrix}_{p \times p} \tag{2}$$

The number of terms in equation~(2) may be reduced by two mechanisms:

a) Some elements in $\boldsymbol{\eta}_j, j = 1, \ldots m$ are zero
b) There are coinciding eigenvalues of $\boldsymbol{\Sigma}_{xx}$ such that it is enough to have one eigenvector in equation~(2).

Let there are k number predictor components that are relevant for any of the response. The $\mathcal{P}$ contains the indices of these position for each response components. Here the order of components of predictor is defined by a decreasing set of eigenvalues such that $\lambda_1 \geq \ldots \geq \lambda_p > 0$. In `simrel-m` package, these set of position is referred as `relpos`. For example, if $\mathcal{P} = 1, 2, 3, 5, 4$ then we can say that there are 3 informative space in response such that components 1 and 2 of predictor variable is relevant for first response component; component 3 and 5 of predictor are relevant for second response component and fourth predictor component is relevant for third response component. In addition, $\lambda_1 \geq \ldots \geq \lambda_5$ are relevant for some response.

In `simrel-m`, these 3 response components are combined with non-informative vectors for desired number of response variables. This is referred as $\mathcal{S}$ and `ypos` in `simrel-m` package. If $\mathcal{S} = 1, 4, 2, 3, 5$ then we can say that there are 5 response variables which has 3 dimensional informative space. Since response components 1, 2 and 3 are informative, from the indices of $\mathcal{S}$, first response component is combined with non-informative fourth component and third informative component is combined with fifth non-informative component. In this way, we will obtain a set of 5 response variables for which predictor component 1 and 2 will be relevant for response 1 and 4; predictor component 3 and 5 will be relevant for response 2 and predictor component 4 will be relevant for response 3 and 5.

For simplification, an assumption is made that all p eigenvalues of $\boldsymbol{\Sigma}_{XX}$ decrease exponentially as $e^{-\gamma(i-1)}$ for $i = 1, \ldots p$ and some positive constant γ. This way, the p eigenvalues depends on single variables γ such that when γ is large, eigenvalues decreases sharply referring high degree of multi-collinearity in predictor variables.

Here we have assumed that the relevant components are know, as in Helland and Almøy (1994), which is rare in practice. But in comparative studies of prediction methods, this can help to explain interesting cases. For example, simultaneous envelope (Cook and Zhang, 2015) has discussed about the informative (material) and uninformative (immaterial) space on both $\mathbf{X}$ and $\mathbf{Y}$. In this case, `simrel-m` can provide a simulated data with various interesting cases for its assessment and comparison with other methods.

Statistical Model

Let us consider a random regression model in equation~(3) as our point of departure.

$$\mathbf{Y} = \boldsymbol{\mu}_Y + \mathbf{B}^t(\mathbf{X} - \boldsymbol{\mu}_X) + \boldsymbol{\epsilon} \tag{3}$$

where $\mathbf{Y}$ is a response matrix with m response variables $y_1, y_2, \ldots y_m$ with mean vector of $\boldsymbol{\mu}_Y$; $\mathbf{X}$ is vector of p predictor variables and the random error term $\boldsymbol{\epsilon}$ is assumed to follow $N(\mathbf{0},\ \boldsymbol{\Sigma}_{Y|X})$. In addition, we assume equation~(3) as a random regression model where $\mathbf{X} \sim N(\boldsymbol{\mu}_X, \boldsymbol{\Sigma}_{XX})$ independent of $\boldsymbol{\epsilon}$. Equivalently, this relationship can be written as,

$$\begin{bmatrix}\mathbf{Y}\\ \mathbf{X}\end{bmatrix} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = N\left(\begin{bmatrix}\boldsymbol{\mu}_Y\\ \boldsymbol{\mu}_X\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Sigma}_{YY} & \boldsymbol{\Sigma}_{XY}^t\\ \boldsymbol{\Sigma}_{XY} & \boldsymbol{\Sigma}_{XX}\end{bmatrix}\right) \tag{4}$$

Here, $\boldsymbol{\Sigma}_{YY}$ is Covariance Matrix of response $\mathbf{Y}$; $\boldsymbol{\Sigma}_{XY}$ is Covariance Matrix between $\mathbf{X}$ and $\mathbf{Y}$; $\boldsymbol{\Sigma}_{XX}$ is Covariance matrix of predictor variables $\mathbf{X}$; $\boldsymbol{\mu}_Y$ and $\boldsymbol{\mu}_X$ are Mean vectors of response $\mathbf{Y}$ and predictor $\mathbf{X}$ respective.

Simulation of $(\mathbf{Y}, \mathbf{X})$ for model~(4) requires the fact that – a set of latent variable spanning $\mathbf{X}$ and $\mathbf{Y}$ 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 $\mathbf{Z} = \mathbf{R}\mathbf{X}$ and $\mathbf{W} = \mathbf{Q}\mathbf{Y}$ so that,

$$\begin{aligned}\begin{bmatrix}\mathbf{W}\\ \boldsymbol{Z}\end{bmatrix} &\sim N\left(\begin{bmatrix}\boldsymbol{\mu}_W\\ \boldsymbol{\mu}_Z\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Sigma}_{WW} & \boldsymbol{\Sigma}_{WZ}^t\\ \boldsymbol{\Sigma}_{ZW} & \boldsymbol{\Sigma}_{ZZ}\end{bmatrix}\right)\\ &= N\left(\begin{bmatrix}\boldsymbol{Q}\boldsymbol{\mu}_Y\\ \boldsymbol{R}\boldsymbol{\mu}_X\end{bmatrix}, \begin{bmatrix}\boldsymbol{Q}\boldsymbol{\Sigma}_{YY}\boldsymbol{Q}^t & \boldsymbol{Q}\boldsymbol{\Sigma}_{XY}^t\mathbf{R}^t\\ \boldsymbol{R}\boldsymbol{\Sigma}_{XY}\boldsymbol{Q}^t & \boldsymbol{R}\boldsymbol{\Sigma}_{XX}\mathbf{R}^t\end{bmatrix}\right)\end{aligned} \tag{5}$$

Further, if both $\mathbf{Q}$ and $\mathbf{R}$ are orthonormal matrix such that $\mathbf{Q}^t\mathbf{Q} = \mathbf{I}_q$ and $\mathbf{R}^t\mathbf{R} = \mathbf{I}_p$, the inverse transformation can be defined as,

$$
\begin{aligned}
\boldsymbol{\Sigma}_{XX} = \mathbf{R}^t\boldsymbol{\Sigma}_{ZZ}\mathbf{R} \quad &\Rightarrow \quad \boldsymbol{\Sigma}_{ZZ} = \mathbf{R}\boldsymbol{\Sigma}_{XX}\mathbf{R}^t \\
\boldsymbol{\Sigma}_{XY} = \mathbf{R}^t\boldsymbol{\Sigma}_{ZW}\mathbf{Q} \quad &\Rightarrow \quad \boldsymbol{\Sigma}_{ZW} = \mathbf{R}\boldsymbol{\Sigma}_{XY}\mathbf{Q}^t \\
\boldsymbol{\Sigma}_{YX} = \mathbf{Q}^t\boldsymbol{\Sigma}_{WZ}\mathbf{R} \quad &\Rightarrow \quad \boldsymbol{\Sigma}_{WZ} = \mathbf{Q}\boldsymbol{\Sigma}_{YX}\mathbf{R}^t \\
\boldsymbol{\Sigma}_{YY} = \mathbf{Q}^t\boldsymbol{\Sigma}_{WW}\mathbf{Q} \quad &\Rightarrow \quad \boldsymbol{\Sigma}_{WW} = \mathbf{Q}\boldsymbol{\Sigma}_{YY}\mathbf{Q}^t
\end{aligned}
\tag{6}
$$

In addition, a linear model relating $\mathbf{W}$ and $\mathbf{Z}$ can be written as,

$$
\mathbf{W} = \boldsymbol{\mu}_W + \mathbf{A}^t\left(\mathbf{Z} - \boldsymbol{\mu}_Z\right) + \boldsymbol{\tau}; \qquad \boldsymbol{\tau} \sim N\left(\mathbf{0}, \boldsymbol{\Sigma}_{W|Z}\right) \tag{7}
$$

Here, we can find a direct connection between different population properties between (3) and (7). Some of them are:

Regression Coefficients Regression coefficients for model~(3) is,

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

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

$$
\begin{aligned}
\mathbf{A} = \boldsymbol{\Sigma}_{WZ}\boldsymbol{\Sigma}_{ZZ}^{-1} \qquad &= \boldsymbol{Q}\boldsymbol{\Sigma}_{YZ}\mathbf{R}^t\left[\boldsymbol{R}\boldsymbol{\Sigma}_{XX}\mathbf{R}^t\right]^{-1} \\
&= \mathbf{Q}\left[\boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1}\right]\mathbf{R}^t \\
&= \mathbf{Q}\mathbf{B}\mathbf{R}^t
\end{aligned}
$$

Error Variance The noise variance and the minimum prediction error for model~(3) is,

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

Further, the noise variance of transformed model~(7) is,

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

Population Coefficient of Determination The population coefficient of determination for

model~(3) is,

$$\boldsymbol{\rho}^2_{XY} = \boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}^{-1}_{XX}\boldsymbol{\Sigma}_{XY}\boldsymbol{\Sigma}^{-1}_{YY}$$

Further, the coefficient of determination corresponding to model~(7) is,

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

Model parameterization and relevant components

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

$$\Sigma = \mathbf{E}\Lambda\mathbf{E}^t \tag{8}$$

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~(6), $\boldsymbol{\Sigma}_{XX}$ and $\boldsymbol{\Sigma}_{WW}$ 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 $\boldsymbol{\Sigma}_{ZZ} = \Lambda = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_p)$, 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 $\boldsymbol{\Sigma}_{ZZ}$. 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 $\mathcal{Y}$ 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 $\mathbf{Z}$ ($\boldsymbol{\Sigma}_{WZ}$) is constructed referring to the terminology in Helland and Almøy (1994) that the principal components are termed as relevant for which $\boldsymbol{\Sigma}_{WZ}$ 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,

$$\mathbf{A} = \begin{cases} \mathbf{\Sigma}_{WZ}\mathbf{\Sigma}_{ZZ}^{-1} = \sum_{j \in \mathcal{P}_i} \left(\frac{\sigma_{ij}}{\lambda_j} \mathbf{t}_j \right) & \text{for } i = 1, \ldots, l \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

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 $\mathbf{\Sigma}_{XX}$ 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 $\mathbf{\Xi}_{WZ}$ in equation~(5), covariances between W and Z ($\mathbf{\Sigma}_{WZ}$) should be constructed such that it is positive definite and satisfy the relation,

$$\begin{aligned} \boldsymbol{\mathcal{R}}_{WZ}^2 &= \mathbf{\Sigma}_{WZ}\mathbf{\Sigma}_{ZZ}^{-1}\mathbf{\Sigma}_{ZW}\mathbf{\Sigma}_{WW}^{-1} \\ \text{i.e, } \boldsymbol{\mathcal{R}}_{WZ}^2\mathbf{\Sigma}_{WW} &= \mathbf{\Sigma}_{WZ}\mathbf{\Lambda}^{-1}\mathbf{\Sigma}_{ZW} \end{aligned} \tag{10}$$

For given $\boldsymbol{\mathcal{R}}_{WZ}^2$ and $\Sigma_{WW} = \mathbf{I}_m$, equation~(10) will be satisfied for some $\mathbf{\Sigma}_{WX}$ whose rows correspond to the relevant components for $\mathbf{W}$. As we have considered the situation that no relevant components are common, elements in $\mathbf{\Sigma}_{WZ}$ 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,

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

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

Data Simulation

After the construction of $\mathbf{\Xi}_{WZ}$, 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$\boldsymbol{\mu}_Z = 0$. Since$\mathbf{\Xi}_{WZ}$ is positive definite,$\mathbf{\Xi}_{WZ}^{1/2}$ 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 $\mathbf{G} = \boldsymbol{U}\boldsymbol{\Xi}_{WZ}^{1/2}$

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 $\boldsymbol{\Sigma}_{XX}$ 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 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 $\mathbf{Y}_4$.

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. Simultaneous envelopes for multivariate linear regression. *Technometrics*, 57(1):11–25, 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.

Inge S Helland, Solve Saebø, Ha Tjelmeland, et al. Near optimal prediction from relevant components. *Scandinavian Journal of Statistics*, 39(4):695–713, 2012.

Brian D Ripley. *Stochastic simulation*, volume 316. John Wiley & Sons, 2009.

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.