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 were 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`.

Statistical Model

Let us consider a random regression model in equation~(1) as our point of departure.

$$\begin{bmatrix} \mathbf{Y} \\ \mathbf{X} \end{bmatrix} \sim N\left(\begin{bmatrix} \boldsymbol{\mu}_Y \\ \boldsymbol{\mu}_X \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{YY} & \boldsymbol{\Sigma}_{YX} \\ \boldsymbol{\Sigma}_{XY} & \boldsymbol{\Sigma}_{XX} \end{bmatrix} \right) \tag{1}$$

where, $\mathbf{Y}$ is a response matrix with m response variables $y_1, y_2, \ldots y_m$ with mean vector of $\boldsymbol{\mu}_Y$ and $\mathbf{X}$ is vector of p predictor variables. Further,

$\boldsymbol{\Sigma}_{YY}$	is variance-covariance matrix of response $\mathbf{Y}$
$\boldsymbol{\Sigma}_{XX}$	is variance-covariance matrix of predictor variables $\mathbf{X}$
$\boldsymbol{\Sigma}_{XY}$	is matrix of covariance between $\mathbf{X}$ and $\mathbf{Y}$
$\boldsymbol{\mu}_Y$ and $\boldsymbol{\mu}_X$	are mean vectors of response $\mathbf{Y}$ and predictor $\mathbf{X}$ respective.

A linear relationship between $\mathbf{X}$ and $\mathbf{Y}$ for model~(1) can be imagined as,

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

where, $\boldsymbol{\beta}^t$ is regression coefficient and $\boldsymbol{\varepsilon}$ is error term such that $\boldsymbol{\varepsilon} \sim N\left(0, \boldsymbol{\Sigma}_{Y|X}\right)$. The properties of the linear model in equation~(2) can be expressed in terms of covariance matrices from equation~(1).

Regression Coefficients

$$\boldsymbol{\beta} = \boldsymbol{\Sigma}_{XY}\boldsymbol{\Sigma}_{XX}^{-1}$$

Coefficient of Determination $\boldsymbol{\rho}_Y^2$ The diagonal elements of coefficient of determination matrix $\boldsymbol{\rho}_Y^2$ gives the amount of variation that $\boldsymbol{X}$ has explained about $\boldsymbol{Y}$ in equation~(2).

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

Error variance The minimum error $\boldsymbol{\Sigma}_{Y|X}$ of the model is,

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

Let us define a transformation of $\boldsymbol{X}$ and $\boldsymbol{Y}$ as, $\mathbf{Z} = \mathbf{R}\mathbf{X}$ and $\mathbf{W} = \mathbf{Q}\mathbf{Y}$. Here, $\mathbf{R}_{p\times p}$ and $\mathbf{Q}_{m\times m}$ are rotation matrices which rotates $\mathbf{X}$ and $\mathbf{Y}$ giving $\mathbf{Z}$ and $\mathbf{W}$ respectively. The random regression model in equation~(1) can be expressed with these transformed variables as,

$$\begin{aligned}\begin{bmatrix}\mathbf{W}\\ \mathbf{Z}\end{bmatrix} &\sim N\left(\begin{bmatrix}\boldsymbol{\mu}_W\\ \boldsymbol{\mu}_Z\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Sigma}_{WW} & \boldsymbol{\Sigma}_{WZ}\\ \boldsymbol{\Sigma}_{ZW} & \boldsymbol{\Sigma}_{ZZ}\end{bmatrix}\right)\\ &= N\left(\begin{bmatrix}\boldsymbol{Q}\boldsymbol{\mu}_Y\\ \boldsymbol{R}\boldsymbol{\mu}_X\end{bmatrix}, \begin{bmatrix}\boldsymbol{Q}\boldsymbol{\Sigma}_{YY}\boldsymbol{Q}^t & \boldsymbol{Q}\boldsymbol{\Sigma}_{YX}\mathbf{R}^t\\ \boldsymbol{R}\boldsymbol{\Sigma}_{XY}\boldsymbol{Q}^t & \boldsymbol{R}\boldsymbol{\Sigma}_{XX}\mathbf{R}^t\end{bmatrix}\right)\end{aligned} \tag{3}$$

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

$$\mathbf{W} = \boldsymbol{\mu}_W + \boldsymbol{\alpha}^t\left(\mathbf{Z} - \boldsymbol{\mu}_Z\right) + \boldsymbol{\tau} \tag{4}$$

where, $\boldsymbol{\alpha}$ is regression coefficient for the transformed model and $\boldsymbol{\tau} \sim N\left(\mathbf{0}, \boldsymbol{\Sigma}_{W|Z}\right)$. 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,

$$
\begin{aligned}
\boldsymbol{\Sigma}_{XX} = \mathbf{R}^t\boldsymbol{\Sigma}_{ZZ}\mathbf{R} \quad &\Rightarrow \quad \boldsymbol{\Sigma}_{ZZ} = \mathbf{R}\boldsymbol{\Sigma}_{XX}\mathbf{R}^t \\
\boldsymbol{\Sigma}_{XY} = \mathbf{R}^t\boldsymbol{\Sigma}_{ZW}\mathbf{Q} \quad &\Rightarrow \quad \boldsymbol{\Sigma}_{ZW} = \mathbf{R}\boldsymbol{\Sigma}_{XY}\mathbf{Q}^t \\
\boldsymbol{\Sigma}_{YX} = \mathbf{Q}^t\boldsymbol{\Sigma}_{WZ}\mathbf{R} \quad &\Rightarrow \quad \boldsymbol{\Sigma}_{WZ} = \mathbf{Q}\boldsymbol{\Sigma}_{YX}\mathbf{R}^t \\
\boldsymbol{\Sigma}_{YY} = \mathbf{Q}^t\boldsymbol{\Sigma}_{WW}\mathbf{Q} \quad &\Rightarrow \quad \boldsymbol{\Sigma}_{WW} = \mathbf{Q}\boldsymbol{\Sigma}_{YY}\mathbf{Q}^t
\end{aligned}
\tag{5}
$$

Here, we can find a direct connection between different population properties between (2) and (4).

Regression Coefficients

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

Error Variance Further, the noise variance of transformed model~(4) 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~(4) is,

$$
\begin{aligned}
\boldsymbol{\rho}_W^2 &= \boldsymbol{\Sigma}_{WZ}\boldsymbol{\Sigma}_{ZZ}^{-1}\boldsymbol{\Sigma}_{ZW}\boldsymbol{\Sigma}_{WW}^{-1} \\
&= \mathbf{Q}^t\boldsymbol{\Sigma}_{YX}\mathbf{R}^t\left(\mathbf{R}\boldsymbol{\Sigma}_{XX}\mathbf{R}^t\right)^{-1}\mathbf{R}\boldsymbol{\Sigma}_{XY}\mathbf{Q}^t\left(\mathbf{Q}\boldsymbol{\Sigma}_{YY}^{-1}\mathbf{Q}^t\right) \\
&= \mathbf{Q}^t\left[\boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}\boldsymbol{\Sigma}_{XY}\boldsymbol{\Sigma}_{YY}^{-1}\right]\mathbf{Q} \\
&= \mathbf{Q}\boldsymbol{\rho}_Y^2\mathbf{Q}^t
\end{aligned}
$$

From eigenvalue decomposition principle, if $\boldsymbol{\Sigma}_{XX} = \mathbf{R}\boldsymbol{\Lambda}\mathbf{R}^t$ and $\boldsymbol{\Sigma}_{YY} = \mathbf{Q}\boldsymbol{\Omega}\mathbf{Q}^t$ then $\mathbf{Z}$ and $\mathbf{W}$ can be principal components of $\mathbf{X}$ and $\mathbf{Y}$ respectively. Here, $\boldsymbol{\Sigma}$ and $\boldsymbol{\Omega}$ are diagonal matrix of eigenvalues corresponding to $\mathbf{X}$ and $\mathbf{Y}$ respectively.

Relevant Components

Let us consider a single response linear model with p predictors.

$$\mathbf{y} = \mu_y + \beta^t (\mathbf{X} - \mu_x) + \epsilon$$

where, $\epsilon \sim N(0, \sigma^2)$ and $\mathbf{X}$ and ϵ are random and independent. Following the principal of relevant space and irrelevant space which are discussed extensively in Helland and Almøy (1994), Helland (2000), Helland et al. (2012), Cook et al. (2013), Sæbø et al. (2015) and Helland et al. (2017), we can assume that there exists a subspace of complete variable space which is relevant for $\mathbf{y}$. An orthogonal space to this space does not contain any information about $\mathbf{y}$ and are irrelevant. Here, the $y-$relevant subspace of $\mathbf{X}$ is spanned by a subset of eigenvectors $(\mathbf{e})$ of covariance matrix of $\mathbf{X}$, i.e. $\mathbf{\Sigma}_{XX}$).

This concept can be extended for m response so that the subspace of $\mathbf{X}$ is relevant for a subspace of $\mathbf{Y}$. This corresponds to the concept of simultaneous envelope (Cook and Zhang, 2014) where relevant (material) and irrelevant (immaterial) space were discussed for both response and predictors.

Model Parameterization

In order to construct a covariance matrix of $\mathbf{Z}$ and $\mathbf{W}$ for model in equation~(3), we need to identify $1/2(p+m)(p+m+1)$ unknowns. For the purpose of this simulation, we implement some assumption to re-parameterize and simplify the model parameters. This enables us to construct diverse nature of model from few key parameters.

Parameterization of $\mathbf{\Sigma}_{ZZ}$: Since $\mathbf{X}$'s are principal components of $\mathbf{X}$, the $\mathbf{\Sigma}_{ZZ}$ can be a diagonal matrix with eigenvalues $\lambda_1, \ldots, \lambda_p$ of predictors $\mathbf{X}$. Further, we adopt following approximate parametric representation of these eigenvalues,

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

Here as γ increases, the decline of eigenvalues becomes steeper and hence a single parameter γ can be used for $\mathbf{\Sigma}_{ZZ}$.

Parameterization of $\mathbf{\Sigma}_{WW}$: Here, we assume that $\mathbf{W}$'s are independent and thus their covariance matrix is considered to be Identity $\mathbf{I}_m$.

Parameterization of $\mathbf{\Sigma}_{ZW}$: After parameterization of $\mathbf{\Sigma}_{ZZ}$ and $\mathbf{\Sigma}_{WW}$, we are left with $m \times p$ number of unknowns corresponding to $\mathbf{\Sigma}_{ZW}$. The elements in this covariance matrix depends on position of x-component that are relevant for $\mathbf{Y}$. In order to

re-parameterize this covariance matrix, it is necessary to discuss about the position of relevant components in details.

Position of relevant components

Let only k_1 components are relevant for $\mathbf{w}_1$, k_2 components are relevant for $\mathbf{w}_2$ and so on. Let the position of these components are given by the set $\mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_m$ respectively. Further, the covariance between $\mathbf{w}_j$ and $\mathbf{z}_i$ is non-zero only if $\mathbf{z}_i$ is relevant for $\mathbf{w}_j$. If σ_{ij} be the covariance between $\mathbf{w}_j$ and $\mathbf{z_i}$ then $\sigma_{ij} \neq 0$ if $i \in \mathcal{P}_j$ where $i = 1, \ldots, p$ and $j = 1, \ldots, m$ and $\sigma_{ij} = 0$ otherwise.

In addition, the corresponding regression coefficient for $\mathbf{w}_j$ is,

$$\boldsymbol{\alpha}_j = \Lambda^{-1}\sigma_{ij} = \sum_{i \in \mathcal{P}_j} \frac{\sigma_{ij}}{\lambda_i}\mathbf{t}_{ij}, \qquad j = 1, 2, \ldots m$$

where, $\mathbf{t}_{ij}$ is a matrix with column vectors of 1's and 0's such that $\mathbf{t}_{ij} = 1$ if the position relevant components for $\mathbf{w}_j$ in set $\mathcal{P}_j$ and 0 otherwise.

The position of relevant components have heavy impact on prediction. Helland and Almøy (1994) have shown that if relevant components have large variance, prediction of $\mathbf{Y}$ from $\mathbf{X}$ is relatively easy and if the variance of relevant components is small, the prediction becomes difficult given that coefficient of determination and other model parameters held constant. For example, if first and second components of $\mathbf{X}$ are relevant for $\mathbf{Y}_1$ and fifth and sixth componets are relevant for $\mathbf{Y}_2$, it is relatively easy to predict $\mathbf{Y}_1$ than $\mathbf{Y}_2$. Since, the first and second principal components have larger variance than fifth and sixth components.

Although the covariance matrix depends only on few relevant components, we can not choose these covariances freely since we also need to satisfy following two conditions:

- The covariance matrix must be positive definite
- The covariance σ_{ij} must satisfy user defined coefficient of determination

We have the relation,

$$\boldsymbol{\rho}^2_W = \boldsymbol{\Sigma}^t_{ZW}\boldsymbol{\Sigma}^{-1}_{ZZ}\boldsymbol{\Sigma}_{ZW}\boldsymbol{\Sigma}^I_{WW}$$

Applying our assumption for simulation, $\boldsymbol{\Sigma_{WW}} = \mathbf{I}_m$ and $\boldsymbol{\Sigma}_{ZZ} = \Lambda$, we obtain,

$$\begin{aligned}\boldsymbol{\rho}_W^2 &= \boldsymbol{\Sigma}_{ZW}^t \Lambda^{-1} \boldsymbol{\Sigma}_{ZW} \mathbf{I}_m \\ &= \begin{bmatrix} \sum_{i=1}^p \sigma_{i1}^2/\lambda_i & \dots & \sum_{i=1}^p \sigma_{i1}\sigma_{im}/\lambda_i \\ \vdots & \ddots & \vdots \\ \sum_{i=1}^p \sigma_{i1}\sigma_{im}/\lambda_i & \dots & \sum_{i=1}^p \sigma_{im}^2/\lambda_i \end{bmatrix}\end{aligned}$$

Furthermore, we assume that there are no overlapping relevant components for any two $\mathbf{W}$, i.e, $n\left(\mathcal{P}_j \cap \mathcal{P}_{j*}\right) = 0$ or $\sigma_{ij}\sigma_{ij*} = 0$ for $j \neq j*$. The additional unknown parameters in diagonal should agree with user specified coefficient of determination for $\mathbf{W}_j$. i.e, ρ_{wj}^2 is,

$$\rho_{wj}^2 = \sum_{i=1}^{p} \frac{\sigma_{ij}^2}{\lambda_i}$$

Here, only the relevant components have non-zero covariances with $\mathbf{w}_j$, so,

$$\rho_{wj}^2 = \sum_{i \in \mathcal{P}_j} \frac{\sigma_{ij}^2}{\lambda_i}$$

For some user defined ρ_{jw}^2, σ_{ij}^2 determined as follows,

1) Sample k_j values from uniform distribution $\mathcal{U}(-1, 1)$ distribution. Let them be, $\mathcal{S}_{\mathcal{P}_1}, \dots, \mathcal{S}_{\mathcal{P}_{k_j}}$.
2) Define,
$$\sigma_{ij} = \text{Sign}\left(\mathcal{S}_i\right) \sqrt{\frac{\rho_{wj}^2 \left|\mathcal{S}_i\right|}{\sum_{k \in \mathcal{P}_j} \left|\mathcal{S}_k\right|} \lambda_i}$$
for $i \in \mathcal{P}_j$ and $j = 1, \dots, m$

Data Simulation

After the construction of $\boldsymbol{\Xi}_{WZ}$, 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$\boldsymbol{\Xi}_{WZ}$ is positive definite,$\boldsymbol{\Xi}_{WZ}^{1/2}$ 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, Jan 2014. ISSN 1537-2723. doi: 10.1080/00401706.2013.872700. URL http://dx.doi.org/10.1080/00401706.2013.872700.

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. Model reduction for prediction in regression models. *Scandinavian Journal of Statistics*, 27(1):1–20, Mar 2000. ISSN 1467-9469. doi: 10.1111/1467-9469.00174. URL http://dx.doi.org/10.1111/1467-9469.00174.

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.

Inge S. Helland, S. Sæ bø, T. Almø y, and R. Rimal. Model and estimators for partial least squares. 2017.

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.