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 presenting here is based on the 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 specify various population properties such as which components of predictors $(\mathbf{x})$ are relevant for a latent subspace of the responses $\mathbf{y}$ and collinearity structure of $\mathbf{x}$. This enables the possibility to construct data for evaluating estimation methods and methods developed for variable selection.

Among several publications on simulation (`which publications`), Ripley (2009) has exhaustively discussed the topic. In addition, many publications (`which publications`) 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 the 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 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 vector 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 with mean vector $\boldsymbol{\mu}_y$. Further,

$\boldsymbol{\Sigma}_{yy}$	is variance-covariance matrix of $\mathbf{y}$
$\boldsymbol{\Sigma}_{xx}$	is variance-covariance matrix of variables $\mathbf{x}$
$\boldsymbol{\Sigma}_{xy}$	is matrix of covariance between $\mathbf{x}$ and $\mathbf{y}$

For model~(1), standard theory in multivariate statistics may be used to show that $\mathbf{y}$ conditioned on $\mathbf{x}$ corresponds to the linear model,

$$\mathbf{y} = \boldsymbol{\mu}_y + \boldsymbol{\beta}^t(\mathbf{x} - \boldsymbol{\mu}_x) + \boldsymbol{\varepsilon} \tag{2}$$

where, $\boldsymbol{\beta}^t$ is a matrix of 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

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

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 $\mathbf{X}$ has explained about $\mathbf{Y}$ in equation~(2).

$$\boldsymbol{\rho}_y^2 = \boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yy}^{-1}$$

Error variance The conditional variance of $\mathbf{y}$ given $\mathbf{x}$ is,

$$\boldsymbol{\Sigma}_{y|x} = \boldsymbol{\Sigma}_{yy} - \boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xy}.$$

The diagonal elements of this matrix equals the theoretical minimum errors of prediction for each of the response variables.

Let us define a transformation of $\mathbf{X}$ and $\mathbf{Y}$ 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 model in equation~(1) can be expressed with these transformed variables as,

$$\begin{bmatrix}\mathbf{w}\\ \mathbf{z}\end{bmatrix} \sim N\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right) \tag{3}$$

$$= N\left(\begin{bmatrix}\boldsymbol{\mu}_w\\ \boldsymbol{\mu}_z\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Sigma}_{ww} & \boldsymbol{\Sigma}_{wz}\\ \boldsymbol{\Sigma}_{zw} & \boldsymbol{\Sigma}_{zz}\end{bmatrix}\right)$$

$$= N\left(\begin{bmatrix}\boldsymbol{Q}\boldsymbol{\mu}_y\\ \boldsymbol{R}\boldsymbol{\mu}_x\end{bmatrix}, \begin{bmatrix}\boldsymbol{Q}\boldsymbol{\Sigma}_{yy}\boldsymbol{Q}^t & \boldsymbol{Q}\boldsymbol{\Sigma}_{yx}\mathbf{R}^t\\ \boldsymbol{R}\boldsymbol{\Sigma}_{xy}\boldsymbol{Q}^t & \boldsymbol{R}\boldsymbol{\Sigma}_{xx}\mathbf{R}^t\end{bmatrix}\right) \tag{4}$$

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

$$\mathbf{w} = \boldsymbol{\mu}_w + \boldsymbol{\alpha}^t \left(\mathbf{z} - \boldsymbol{\mu}_z\right) + \boldsymbol{\tau} \tag{5}$$

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,

$$\begin{aligned} \boldsymbol{\Sigma}_{yy} = \mathbf{Q}^t\boldsymbol{\Sigma}_{ww}\mathbf{Q} \quad \boldsymbol{\Sigma}_{yx} = \mathbf{Q}^t\boldsymbol{\Sigma}_{wz}\mathbf{R} \\ \boldsymbol{\Sigma}_{xy} = \mathbf{R}^t\boldsymbol{\Sigma}_{zw}\mathbf{Q} \quad \boldsymbol{\Sigma}_{xx} = \mathbf{R}^t\boldsymbol{\Sigma}_{zz}\mathbf{R} \end{aligned} \tag{6}$$

Here, we can find a direct connection between different population properties between (2) and (5).

Regression Coefficients

$$\begin{aligned} \boldsymbol{\alpha} = \boldsymbol{\Sigma}_{wz}\boldsymbol{\Sigma}_{zz}^{-1} &= Q\boldsymbol{\Sigma}_{YZ}\mathbf{R}^t\left[R\boldsymbol{\Sigma}_{xx}\mathbf{R}^t\right]^{-1} \\ = \mathbf{Q}\left[\boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}\right]\mathbf{R}^t &= \mathbf{Q}\boldsymbol{\beta}\mathbf{R}^t \end{aligned}$$

Error Variance Further, the noise variance of transformed model~(5) is,

$$\begin{aligned} \boldsymbol{\Sigma}_{w|z} &= Q\boldsymbol{\Sigma}_{yy}\mathbf{Q}^t - Q\boldsymbol{\Sigma}_{yx}\mathbf{R}^t\left[R\boldsymbol{\Sigma}_{xx}R^t\right]^{-1}R\boldsymbol{\Sigma}_{xy}\mathbf{Q}^t \\ &= Q\boldsymbol{\Sigma}_{yy}\mathbf{Q}^t - Q\boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xy}\mathbf{Q}^t \\ &= \mathbf{Q}\left[\boldsymbol{\Sigma}_{yy} - \boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xy}\right]\mathbf{Q}^t \\ &= \mathbf{Q}\boldsymbol{\Sigma}_{y|x}\mathbf{Q}^t \end{aligned}$$

Coefficient of Determination The coefficient of determination for model~(5) 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 interpreted as principle components of $\mathbf{x}$ and $\mathbf{y}$ respectively. Here, $\boldsymbol{\Sigma}$ and $\boldsymbol{\Omega}$ are diagonal matrix of eigenvalues of $\boldsymbol{\Sigma}_{xx}$ and $\boldsymbol{\Sigma}_{yy}$ respectively.

Relevant Components

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

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

where, $\epsilon \sim N(0, \sigma^2)$ and $\mathbf{x}$ are random and independent. Following the principle 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 the full predictor space which is relevant for $\mathbf{y}$. An orthogonal space to this space does not contain any information about $\mathbf{y}$ and is considered as irrelevant. Here, the $y-$relevant subspace of $\mathbf{x}$ is spanned by a subset of eigenvectors of covariance matrix of $\mathbf{x}$, i.e. $\boldsymbol{\Sigma}_{xx}$.

This concept can be extended to 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~(4), we need to identify $1/2(p+m)(p+m+1)$ unknown parameters. 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 $\boldsymbol{\Sigma}_{zz}$ If we consider the rotation matrix $\mathbf{R}$ equals to the eigenvectors of $\boldsymbol{\Sigma}_{xx}$, then $\mathbf{z}$ becomes the set of principle components of $\mathbf{x}$. In that case $\boldsymbol{\Sigma}_{zz}$ is a diagonal matrix with eigenvalues $\lambda_1, \ldots, \lambda_p$. Further, we adopt the following 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 $\boldsymbol{\Sigma}_{zz}$.

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 k_1 components be relevant for $\mathbf{w}_1$, k_2 components be relevant for $\mathbf{w}_2$ and so on. Let the position of these components be 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} is 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, for $j = 1, \ldots m$, $\mathbf{t}_{ij}$ is a vector with 1's and 0's such that $\mathbf{t}_{ij} = 1$ if the position of relevant components for $\mathbf{w}_j$ is 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 eigenvalues (variance), prediction of $\mathbf{y}$ from $\mathbf{x}$ is relatively easy and if the eigenvalues (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 principle 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 Σ_{zz}, Σ_{ww} and Σ must be positive definite
- The covariance σ_{ij} must satisfy user defined coefficient of determination

We have the relation,

$$\boldsymbol{\rho}_w^2 = \boldsymbol{\Sigma}_{zw}^t \boldsymbol{\Sigma}_{zz}^{-1} \boldsymbol{\Sigma}_{zw} \boldsymbol{\Sigma}_{ww}^I$$

Applying our assumption for simulation, $\boldsymbol{\Sigma}_{\boldsymbol{ww}} = \mathbf{I}_m$ 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 of $\boldsymbol{\rho}_w^2$ 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 covariance matrix,

$$\mathbf{\Sigma} = \begin{pmatrix} \mathbf{\Sigma}_{ww} & \mathbf{\Sigma}_{wz} \\ \mathbf{\Sigma}_{zw} & \mathbf{\Sigma}_{zz} \end{pmatrix}$$

n observations are sampled from standard normal distribution of $(\mathbf{w}, \mathbf{z})$ considering their mean to be zero, i.e. $\boldsymbol{\mu}_w = 0$ and $\boldsymbol{\mu}_z = 0$. Let us define $\mathbf{G} = \mathbf{U}\mathbf{\Sigma}^{1/2}$, such that $\mathbf{G}^t\mathbf{G} = \mathbf{\Sigma}$. Since $\mathbf{\Sigma}$ is positive definite, $\mathbf{\Sigma}^{1/2}$ obtained from its Cholesky decomposition can serve as one of its square root and the matrix $\mathbf{U}_{n\times(p+q)}$ is sampled from standard normal distribution so that its covariance matrix $\mathbf{U}^t\mathbf{U} = \mathbf{I}$. In addition the covariance matrix of $\mathbf{G}$ is $\mathbf{\Sigma}$ which satisfies all user defined population properties.

Here the 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 $(\mathbf{w}, \mathbf{z})$. Finally, these simulated matrices $\mathbf{w}$ and $\mathbf{z}$ are orthogonally rotated in order to obtain $\mathbf{y}$ and $\mathbf{x}$ respectively. Following section discuss about these rotation matrices in details.

Rotation of predictor space

In order to make comments on predictor space, let us consider an example where a regression model with $p = 10$ predictors $(\mathbf{x})$ and $m = 4$ responses $(\mathbf{y})$. Let only 3 principle components $(w_1, w_2$ and $w_3)$ are needed to describe all 4 response variables. Further, let $\mathcal{P}_1 = \{1, 2\}, \mathcal{P}_2 = \{3, 4\}$ and $\mathcal{P}_3 = \{5, 6\}$ principle components of $\mathbf{x}$ are relevant for w_1, w_2 and w_3 respectively. Let $\mathcal{S}_1$, $\mathcal{S}_2$ and $\mathcal{S}_3$ be the space spanned by them. These space together $\mathcal{S}_k = \mathcal{S}_1 \oplus \mathcal{S}_2 \oplus \mathcal{S}_3$ is the minimum relevant space similar to the x-envelope as discussed by Cook et al. (2013).

Moreover, let $q_1 = 3, q_2 = 3$ and $q_3 = 2$ number of predictor variables we want to be relevant for w_1, w_2 and w_3 respectively. So that $q_1 = 3$ predictor can be obtained by rotating the principle components in $\mathcal{P}_1$ along with one more irrelevant principle components. Similarly, $q_2 = 3$ predictors, relevant for w_2, can be obtained by rotating principle components in $\mathcal{P}_2$ along with one more irrelevant components and $q_3 = 2$ predictors, relevant for w_3, can be obtained by rotating principle components in $\mathcal{P}_3$ without any additional irrelevant components. Let the space spanned by q_1, q_2 and q_3 number of predictors be $\mathcal{S}_{q_1}$, $\mathcal{S}_{q_2}$ and $\mathcal{S}_{q_3}$. Together they form a space $\mathcal{S}_q = \mathcal{S}_{q_1} \oplus \mathcal{S}_{q_2} \oplus \mathcal{S}_{q_3}$. This space

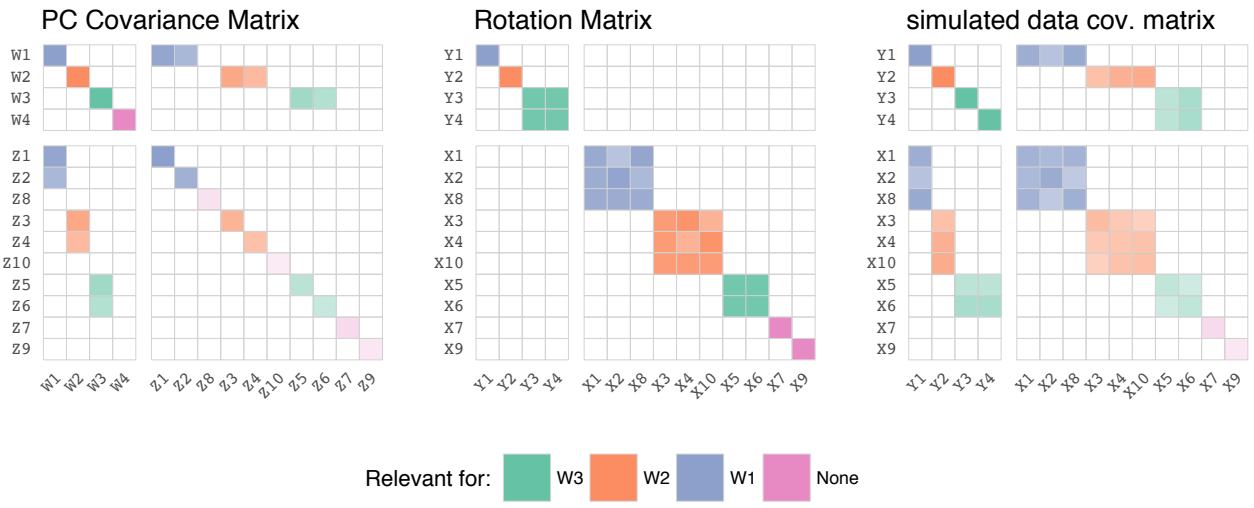

Figure 1: Simulation of predictor and response variables after orthogonal transformation of principle components by a rotation matrix

is bigger than $\mathcal{S}_k$. Here, $\mathcal{S}_k$ is orthogonal to $\mathcal{S}_{p-k}$ and $\mathcal{S}_q$ is orthogonal to $\mathcal{S}_{p-q}$. Generally speaking, here we are splitting complete variable space $\mathcal{S}_p$ into two orthogonal space – $\mathcal{S}_k$ relevant for $\mathbf{y}$ and $\mathcal{S}_{p-k}$ irrelevant for $\mathbf{y}$.

In the previous section, we discussed about constructing covariance matrix of latent structure. Figure~1 (left) shows a similar structure resembling the example here. The three colors represents their relevance with three latent structure of response $(w_1, w_2$ and $w_3)$. Here we can see that z_1 and z_2 (first and second principle components of $\mathbf{x}$) have non-zero covariance with w_1 (first latent component of response $\mathbf{y}$). In the similar manner other non-zero covariances are self-explanatory.

In order to simulate predictor variables $(\mathbf{x})$, we construct matrix $\mathbf{R}$ which then is used for orthogonal rotation of principle components $\mathbf{z}$. This defines a new basis for the same space as is spanned by the principle components. In principle, there are many possible options for a rotation matrix. Among them, the eigenvector matrix of $\mathbf{\Sigma}_{xx}$ can be a candidate. However, in this reverse engineering both rotation matrices $\mathbf{R}$ and $\mathbf{Q}$ along with the covariance matrices $\mathbf{\Sigma}_{xx}$ are unknown. So, we are free to choose any $\mathbf{R}$ that satisfied the properties of a real valued rotation matrix, i.e $\mathbf{R}^{-1} = \mathbf{R}^t$ so that $\mathbf{R}$ is orthonormal and its determinant becomes ± 1. Here the rotation matrix $\mathbf{R}$ should be block diagonal as in figure~1 (middle) in order to rotate spaces $\mathcal{S}_1, \mathcal{S}_2 \ldots$ separately. Figure~2 (left) shows the simulated principle components $\mathbf{z}$ that we are following in our example where we can see that the principle component z_1 and z_2 relevant for w_1 is getting rotated together with an irrelevant z_10. The resultant predictors (Figure~2, right) x_1, x_2 and x_10 will also be relevant

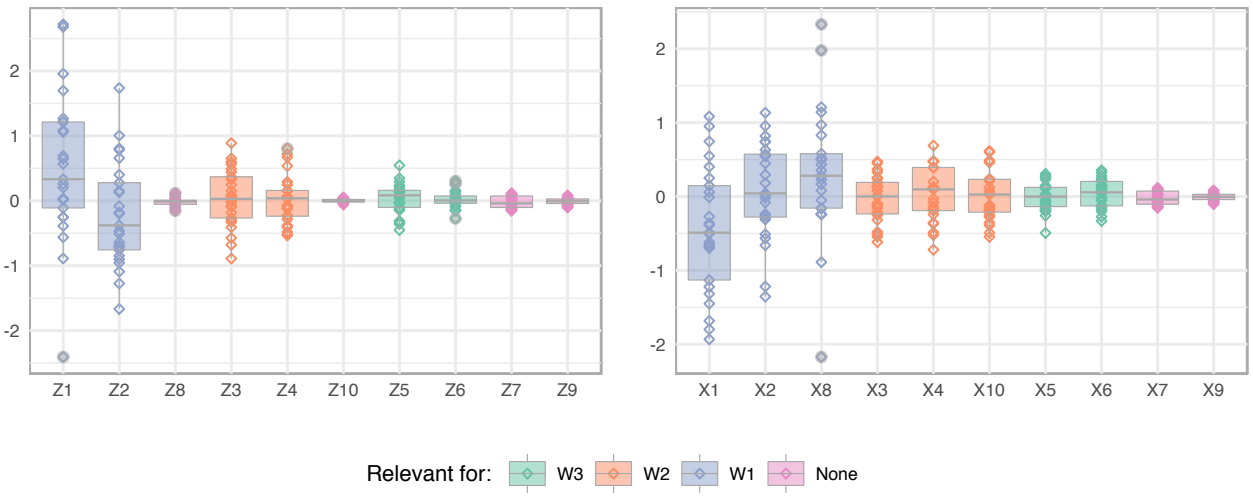

Figure 2: Simulated Data before (left) and after (right) rotation

for w_1. In the figure, we can see that principle components z_8 and z_9 and corresponding predictors are not relevant for any of the response.

Among several methods (Anderson et al., 1987; Heiberger, 1978) for generating random orthogonal matrix, in this paper we are using orthogonal matrix $\mathcal{Q}$ obtained from QR-decomposition of a matrix filled with standard normal variate. The rotation here can be a) restricted and b) unrestricted. The later one rotates all principle components $\mathbf{z}$ together and makes all predictor variables somewhat relevant for all response variables. However, the former one performs a block-wise rotation so that it rotates certain selected principle components together. This gives control for specifying certain predictors as relevant for selected response, which was discussed in our example above. This also lets us to simulate irrelevant predictors such as x_7 and x_8 which can be detected during variables selection procedure.

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.