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)

June 22, 2017

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, Ripley (2009) and Gamerman and Lopes (2006) has exhaustively discussed the topic. In addition, many publications have implemented simulated data in order to investigate new estimation methods and prediction strategy (see: Cook and Zhang, 2015b; 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 $\boldsymbol{\mu}_x$. 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 each $\mathbf{Y}$.

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

Conditional 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 least square error 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,

$$\begin{aligned}
\begin{bmatrix}\mathbf{w}\\ \mathbf{z}\end{bmatrix} &\sim N\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right) \qquad (3)\\
&= N\left(\begin{bmatrix}\boldsymbol{\mu}_w\\ \boldsymbol{\mu}_z\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Sigma}_{ww} & \boldsymbol{\Sigma}_{wz}\\ \boldsymbol{\Sigma}_{zw} & \boldsymbol{\Sigma}_{zz}\end{bmatrix}\right)\\
&= N\left(\begin{bmatrix}\boldsymbol{Q}\boldsymbol{\mu}_y\\ \boldsymbol{R}\boldsymbol{\mu}_x\end{bmatrix}, \begin{bmatrix}\boldsymbol{Q}\boldsymbol{\Sigma}_{yy}\boldsymbol{Q}^t & \boldsymbol{Q}\boldsymbol{\Sigma}_{yx}\mathbf{R}^t\\ \boldsymbol{R}\boldsymbol{\Sigma}_{xy}\boldsymbol{Q}^t & \boldsymbol{R}\boldsymbol{\Sigma}_{xx}\mathbf{R}^t\end{bmatrix}\right) \qquad (4)
\end{aligned}$$

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,

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

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

Regression Coefficients

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

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

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

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 principal, 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 principal components of $\mathbf{x}$ and $\mathbf{y}$ respectively. In this paper, these principal components will be termed as *predictor components* and *response components* respectively. Here, $\mathbf{\Sigma}$ and $\mathbf{\Omega}$ are diagonal matrices of eigenvalues of $\mathbf{\Sigma}_{xx}$ and $\mathbf{\Sigma}_{yy}$ respectively.

Relevant Components

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

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

where, $\epsilon \sim N(0, \sigma^2)$ and $\mathbf{x}$ 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 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. $\mathbf{\Sigma}_{xx}$.

This concept can be extended to m responses 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, 2015b) where relevant (material) and irrelevant (immaterial) space were discussed for both response and predictors.

Model Parameterization

In order to construct a fully specified 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. However, Σ_{zz} is a diagonal matrix of eigenvalues of Σ_{xx}, the number of parameter reduced by $1/2p(p-1)$ parameters. In addition, the covariances of $\mathbf{z}$ that are not relevant for $\mathbf{w}$ will be zero. This further reduces the 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 a wide range of model properties from only few key parameters.

Parameterization of $\mathbf{\Sigma}_{zz}$ If we consider the rotation matrix $\mathbf{R}$ corresponds to the eigenvectors of $\mathbf{\Sigma}_{xx}$, then $\mathbf{z}$ becomes the set of principal components of $\mathbf{x}$. In that case $\mathbf{\Sigma}_{zz}$ is a diagonal matrix with eigenvalues $\lambda_1, \ldots, \lambda_p$. Further, we adopt the following parametric representation of these eigenvalues,

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

Here as γ increases, the decline of eigenvalues becomes steeper, hence the parameter γ controls the level of multicollinearity in $\mathbf{x}$.

Parameterization of $\mathbf{\Sigma}_{ww}$ Here, we assume that $\mathbf{w}$'s are independent unconditionally equally mutinormal distributed with variance 1, hence $\mathbf{\Sigma}_{ww} = \mathbf{I}_m$.

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 $\mathbf{\Sigma}_{zw}$. Some of the elements of $\mathbf{\Sigma}_{zw}$ may be equal to zero, which implies that the given $\mathbf{z}$ is irrelevant for the given variable $\mathbf{w}$. The non-zero elements define which of the $\mathbf{z}$ are relevant for the $\mathbf{w}$. We typically refer to the indices of these $\mathbf{z}$ variables as the position of relevant components. In order to re-parameterize this covariance matrix, it is necessary to discuss 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 positions of these components be given by the index sets $\mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_m$ 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 true regression coefficients for w_j (equation~(5)) is given by:

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

The position of relevant components have heavy impact on prediction. Helland and Almøy (1994) have shown that if relevant components have large eigenvalues (variances), which here implies small index values in $\mathcal{P}_j$, prediction of $\mathbf{y}$ from $\mathbf{x}$ is relatively easy and if the eigenvalues (variances) of relevant components are small, the prediction becomes difficult given that the coefficient of determination and other model parameters are held constant. For example, if the first and second components, $\mathbf{z}_1$ and $\mathbf{z}_2$, are relevant for $\mathbf{w}_1$ and fifth and sixth components, $\mathbf{z}_5$ and $\mathbf{z}_6$, are relevant for $\mathbf{w}_2$, it is relatively easier to predict $\mathbf{w}_1$ than $\mathbf{w}_2$, other properties being similar. This is so, because the first and second principal components have larger variances than the fifth and sixth components.

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

- The covariance matrices 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,

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

Applying our above given assumptions that, $\boldsymbol{\Sigma}_{\boldsymbol{ww}} = \mathbf{I}_m$ 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 is determined as follows,

1. Sample k_j values from a 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

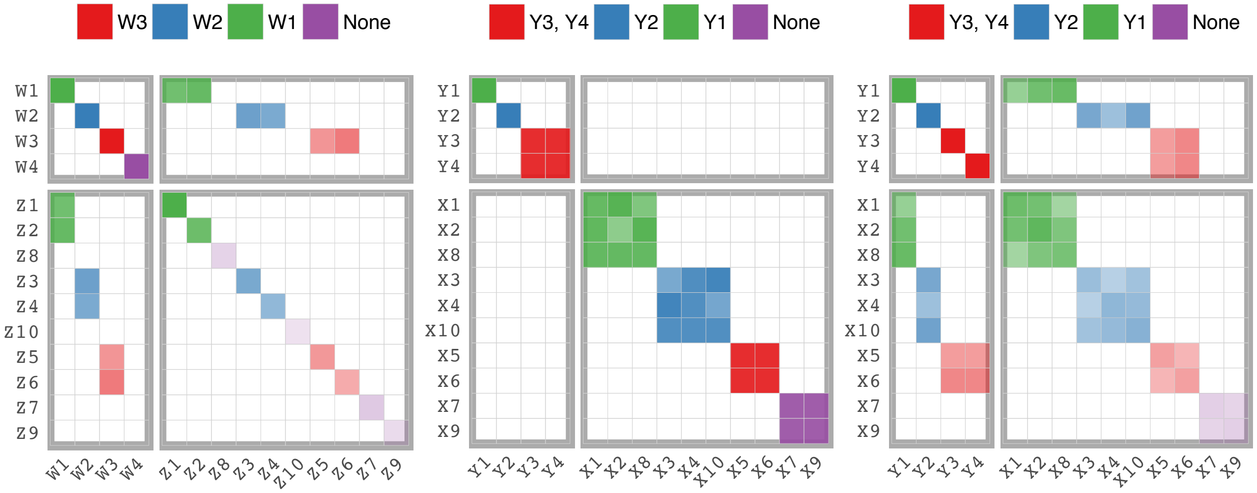
From the above given parameterizations and the user defined choices of model parameters, a fully defined and known covariance matrix $\mathbf{\Sigma}$ of $(\mathbf{w}, \mathbf{z})$ is given. For the simulation of a single observation of $(\mathbf{w}, \mathbf{z})$ let us define $\mathbf{g} = \mathbf{\Sigma}^{-1/2}\mathbf{u}$ such that $\text{cov}(\mathbf{g}) = \mathbf{\Sigma}$. Here $\mathbf{\Sigma}^{-1/2}$ is obtained from Choleskey decomposition and serves as one of the square root of positive definite matrix $\mathbf{\Sigma}$ and $\mathbf{u}$ is simulated from standard normal distribution and has covariance $\text{cov}(u) = \mathbf{I}$.

Similarly, in order to simulate n observations, we define $\underset{n\times(m+p)}{\mathbf{G}} = \mathbf{U}\mathbf{\Sigma}^{-1/2}$ such that $\text{cov}(\mathbf{G}) = \mathbf{\Sigma}$. 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's assume that only three principal components $(w_1, w_2$ and $w_3)$ are needed to describe all four response variables. Further, let the index sets $\mathcal{P}_1 = \{1, 2\}, \mathcal{P}_2 = \{3, 4\}$ and $\mathcal{P}_3 = \{5, 6\}$ define the position of the principal components of $\mathbf{x}$ that are relevant for w_1, w_2 and w_3 respectively. Let $\mathcal{S}_1$, $\mathcal{S}_2$ and $\mathcal{S}_3$ be the orthogonal spaces spanned by each set of principal components. These spaces together span $\mathcal{S}_k = \mathcal{S}_1 \oplus \mathcal{S}_2 \oplus \mathcal{S}_3$ which is the minimum relevant space and equivalent to the x-envelope as discussed by Cook et al. (2013).

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



(a) Relevant Position

(b) Rotation Matrix

(c) Relevant Predictors

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

$\mathcal{S}_{p-k}$ 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 the three latent response components (w_1, w_2 and w_3). Here we can see that z_1 and z_2 (first and second principal 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 principal components $\mathbf{z}$. This defines a new basis for the same space as is spanned by the principal components. In principal, there are many possible options for a rotation matrix. Among them, the eigenvector matrix of $\boldsymbol{\Sigma}_{xx}$ can be a candidate. However, in this reverse engineering both rotation matrices $\mathbf{R}$ and $\mathbf{Q}$ along with the covariance matrices $\boldsymbol{\Sigma}_{xx}$ 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 principal components $\mathbf{z}$ that we are following in our example where we can see that the principal component z_1 and z_2 relevant for w_1 is getting rotated together with an irrelevant component z_8. The resultant predictors (Figure~2, right) x_1, x_2 and x_8 will also be relevant for w_1. In the figure, we can see that principal components x_7, x_8, x_9 and

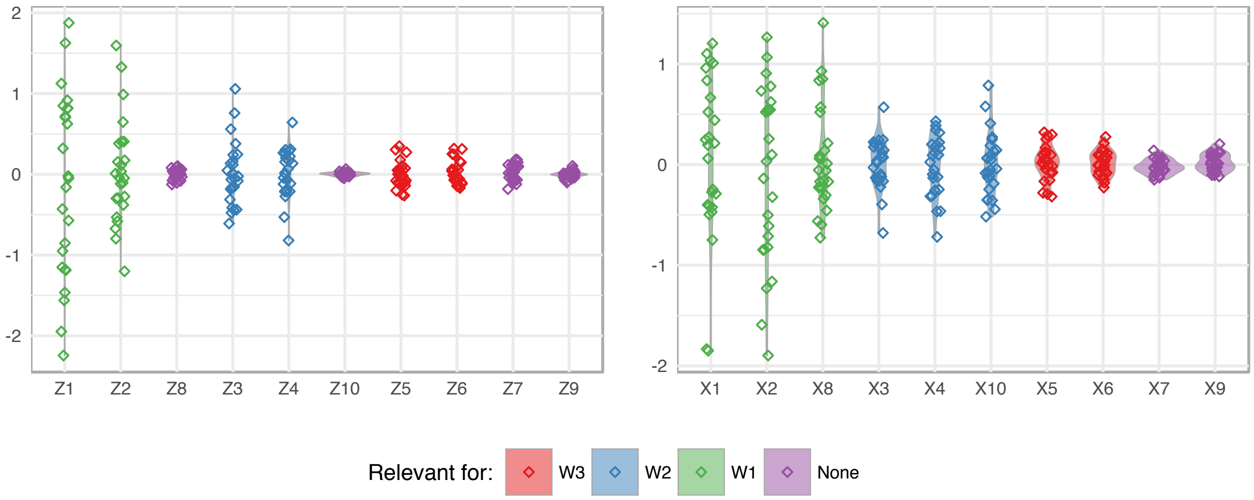


Figure 2: Simulated Data before (left) and after (right) rotation

x_{10} are not relevant for any responses before rotation however x_8, x_{10} predictors becomes relevant after rotation keeping x_7 and x_9 still irrelevant.

Among several methods (Anderson et al., 1987; Heiberger, 1978) for generating random orthogonal matrix, in this paper we are using orthogonal matrix Q obtained from QR-decomposition of a matrix filled with standard normal variates. The rotation here can be a) restricted and b) unrestricted. The latter rotates all principal 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 principal components together. This gives control for specifying certain predictors as relevant for selected responses, which was discussed in our example above. This also allows us to simulate irrelevant predictors such as x_7 and x_9 which can be detected during variables selection procedures.

Rotation of response space

The previous example has four response variables with only three informative principal components w_1, w_2 and w_3. During the rotation of covariance matrix $\mathbf{\Sigma}$, the response space is also rotated separately along with the predictor space. Figure~1 shows that the informative response component w_3 is rotated together with uninformative response component w_4 so that the predictors which were relevant for w_3 will be relevant for response variables y_3 and y_4. Similarly, response components w_1 and w_2 are rotated separately so that predictors relevant for w_1 and w_2 will also be relevant for y_1 and y_2 respectively which we can see in Figure~2. In the r-package *simrel-m*, the combining of the response components is specified by a parameter `ypos`.

Implementation

This section demonstrates an application of `simrel-m` in order to compare different estimation methods on the basis of prediction error. For the comparison, we have considered four well established estimation methods.

a) Ordinary Least Squares (OLS),
b) Principal Component Regression (PCR),
c) Partial Least Squares predicting individual response variable separately (PLS1) and
d) Partial Least Squares predicting all response variables together (PLS).

We have also considered four relatively new estimation methods

a) Canonically Powered Partial Least Squares regression (CPPLS) (Indahl et al., 2009),
b) Canonical Partial Least Squares regression (CPLS) (Indahl et al., 2009),
c) Envelope estimation in predictor space (xenv) (Cook et al., 2010),
d) Envelope estimation in response space (yenv) (Cook and Zhang, 2015a) and
e) Simultaneous estimation of x- and y-envelope (senv) (Cook and Zhang, 2015b)

From the possible combinations of two levels of coefficient of determination (R^2) and two levels of `gamma` (factor that controls the multicollinearity in predictor variable), four simulation designs (design 1, design 2, design 3, design 4) are prepared. Replicating each design 20 times, 80 datasets with five response variables $(m = 5)$ and 16 predictor variables $(p = 16)$ are simulated using the method discussed in this paper. It is also assumed that three principle components of response variables (w_1, w_2 and w_3) completely describes the variation present in five response variables ($y_1 \ldots y_5$). The four designs are presented in the Table~1. All datasets contains 100 sample observations and out of 16 predictor variables, three disjoint set of five predictor variables are relevant for response components w_1, w_2 and w_3. Further, predictor components z_1 and z_6 are relevant for response component w_1, predictor components z_2 and z_5 are relevant for response component w_2 and predictor component z_3 is relevant for response component w_3. In addition, following the discussion about rotation of response space, w_1 is rotated together with w_4 and w_2 is rotated together with w_5.

Table 1: Parameter setting of simulated data for model comparison

	Design1	Design2	Design3	Design4
Decay of eigenvalue (γ)	0.2	0.8	0.2	0.8
Coef. of Determination (ρ^2)	0.8, 0.8, 0.4	0.8, 0.8, 0.4	0.4, 0.4, 0.4	0.4, 0.4, 0.4

For each method, an estimate of test prediction error is computed as,

$$\underset{m\times m}{\boldsymbol{\alpha}} = \left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)^t \boldsymbol{\Sigma}_{xx} \left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) + \boldsymbol{\Sigma}_{Y|X}$$

where, $\hat{\boldsymbol{\beta}}$ is an estimate of true regression coefficient $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}_{xx}$ is the true covariance structure of predictor variable obtained from `simrel-m`. Also, $\boldsymbol{\sigma}^2$ is the true minimum error of the model. Here $\hat{\beta}$ vary accross different estimation methods while the remaining terms are same for each dataset design. Further, an overall prediction error of all responses is measured by the trace of $\boldsymbol{\alpha}$.

The minimimum prediction error (measured as discussed above) for nine estimation methods averaged over 20 replications of four designs are in Table~2. The table also shows that the number of components a method has used in order to obtain the minimum of average prediciton error.

Table 2: Minimum average prediction error (number of components corresponding to minimum prediction error, minimum prediction error)

Model	Design: 1	Design: 2	Design: 3	Design: 4
CPLS	(3, 3.24)	(4, 3.22)	(3, 4.09)	(3, 4.05)
CPPLS	(3, 3.21)	(3, 3.17)	(3, 4.11)	*(3, 4.04)*
OLS	(1, 3.6)	(1, 3.58)	(1, 4.57)	(1, 4.5)
PCR	(7, 3.28)	(6, 3.19)	*(6, 4.08)*	(6, 4.04)
PLS	(5, 3.29)	(6, 3.19)	(3, 4.11)	(6, 4.06)
PLS1	(2, 3.32)	(5, 3.2)	(1, 4.16)	(5, 4.07)
Senv	*(4, 3.17)*	*(5, 3.14)*	(3, 4.35)	(5, 4.28)
Xenv	(5, 3.23)	(6, 3.2)	(5, 4.1)	(6, 4.11)
Yenv	(3, 3.24)	(3, 3.23)	(3, 4.29)	(3, 4.24)

Table~2 shows that simulteneous envelope has prediction error of 3.17 and 3.14 in design~1 (with 4 components) and design~2 (with 5 components) respectively which is smaller than other methods. However the method was not able to show the same performance in design~3 and design~4. PCR model has least prediction error (4.08) from 6 components in design 3 and Cannonically Powered PLS has minimum prediction error (4.04) from 3 components in design 4. In design 3, we can also see that the Canonical PLS method has second best performance with only three components. The number of components vary accross different replicated dataset but the component corresponding to minimum prediction error is discussed here. A detail picture of prediction error for each estimation method obtained for each additional component is shown in Figure~3. Although design 2 and design 4 has higher level of multicollinearity, the performance of the estimation methods is indifferent to its effect. Since all the methods, except OLS, are based on shrinking of estimates, they are less influenced by multicollinearity problem.

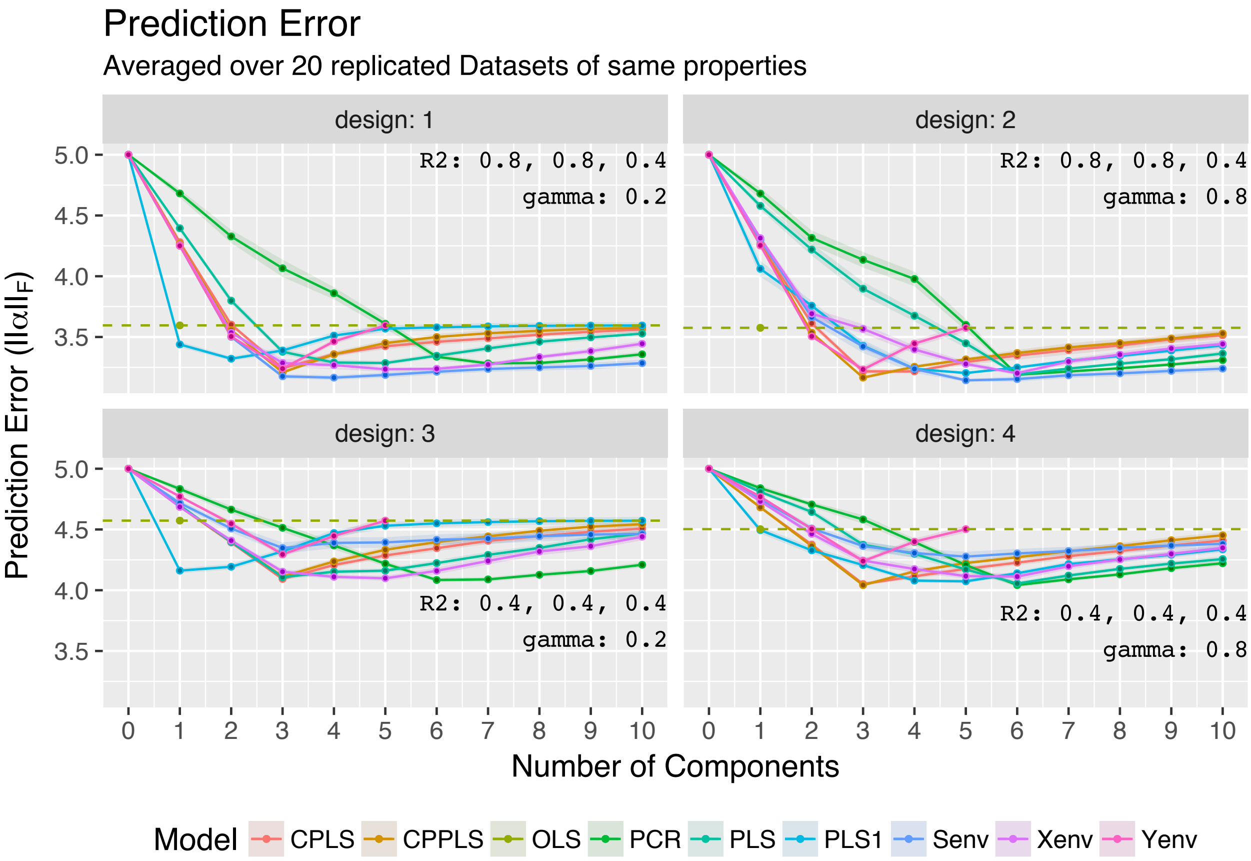


Figure 3: Minimum of Average Prediction Error

Above analysis has answered some questions such as how methods works when there exist a true reduced dimension in response space but also arised question like why they perform differently. For example, what is the reason for the decreasing performance of the simultaneous envelope method as the ρ^2 values are reduced? Does this depend on the dimensions and spheric shape of the **y** envelopes? Since, the example is intended as a demonstration of how `simrel-m` can be used in scientific study, and a more elaborative study would be necessary in order to answer such questions, but for this purpose `simrel-m` would be a powerful tool.

Web Interface

In order to give an alternative interface for `simrel-m`, we have created a shiny app which allows users to input the simulation parameters through different input fields. Figure~4 shows a screenshot for the application. The application contains three main sections through which the user can interact with this simulation approach. A random seed can be selected using section in Figure~4 (a) so that a particular set of data can be simulated. Figure~4 (b) has all the input panel where parameter for simulation can be entered. Here the user also has the option to simulate univariate (uses `simrel` package in CRAN), bivariate (not yet available in CRAN) and multivariate simulation (`simrel-m`). In addition, a simulated R-object is comprised by the simulated data can be download as `Rdata` format (section (e) in Figure~4). The object constitute of the simulated data along with other properties such as coefficient of determination for each response, true regression coefficients and rotation matrices.

All `simrel-m` parameters can be entered using simple user interface where a vectors are separated with comma(,) and lists are separated with semicolon(;). For instance, the relevant position discussed in implementation section of this paper can be entered as `1, 6; 2, 5; 3, 4` which is equivalent to R syntax `list(c(1, 6), c(2, 5), c(3, 4))`.

The application not only allows users to simulate data but also gives some insight on simulated data. The example used in the screenshot has simulated 200 training and 50 test samples with 15 predictor and 4 response variables. There are two latent variables (response component) that completely spans the informative response space. Five predictor variables are relevant for the first response component and explains 80 percent of the variation. In addition, the first and second predictor components spans the same space spanned by these five relevant predictors. Similarly, another set of four predictor variables are relevant for second response component and explains 70 percent of the variation. Here, third, fourth and sixth predictor components spans the same space spanned by these four relevant predictors. Further, the first response component is rotated together with a normally distributed random vector to obtain first and third response variable and second resposne components is rotated together with another normally distributed random vector to obtain second and fourth response variable.

Section (c) in Figure~4 contains three plots – a) regression coefficients b) relevant components and c) estimated relevant components. In the first plot we can see that predictor variables (1, 2, 8, 9 and 13) are relevant for first and third response variable and predictor variables (3, 4, 6 and 15) are relevant for second and fourth response variable. The second plot shows covariance between response components and the predictor components along with the corresponding eigenvalues in the background (bar plot). As in our parameter setting, the plot shows that first and second predictor components have non-zero covariance with first response component and third, fourth and sixth predictor components have non-zero covariance with second response component. The third plot is the estimated covariance between predictor components with response variable, for the simulated data. Since the first and third response components are rotated together, in the plot, the covariance between predictor components and first and third response variable is following similar pattern. This also suggests that the predictor components which were relevant for first response components gets relevant for first and third response variables after rotation.

Along with these main sections, section (d) in the figure contains additional analysis performed with the simulated data such as its estimation with different methods. This section, which is still under development, is intended for educational purposes to show how changing the data properties influences the performances of different estimation and prediction methods.

Many scientific studies (Cook and Zhang, 2015b; Helland et al., 2012; Sæbø et al., 2008) are using simulated data in order to compare their findings with others or assess its properties. In many of these situations, a user-friendly and versitle simulation tool like `simrel-m` can play an important role. Gangsei et al. (2016a); Gangsei et al. (2016b); and Sæbø et al. (2015) are some examples where the univariate and bivariate form of `simrel` have been used for their study.

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, 2015a.

R Dennis Cook and Xin Zhang. Simultaneous envelopes for multivariate linear regression. *Technometrics*, 57(1):11–25, 2015b.

R Dennis Cook, Bing Li, and Francesca Chiaromonte. Envelope models for parsimonious and efficient multivariate linear regression. *Statistica Sinica*, pages 927–960, 2010.

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.

Dani Gamerman and Hedibert F Lopes. *Markov chain Monte Carlo: stochastic simulation for Bayesian inference*. CRC Press, 2006.

Lars Erik Gangsei, Trygve Almøy, and Solve Sæbø. Theoretical evaluation of prediction error in linear regression with a bivariate response variable containing missing data. *Communications in Statistics-Theory and Methods*, (just-accepted), 2016a.

LE Gangsei, T Almøy, and S Sæbø. Linear regression with bivariate response variable containing missing data. *An empirical Bayes strategy to increase prediction precision. Submitted manuscript to Communications in Statistics–Simulation and Computation*, 2016b.

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.

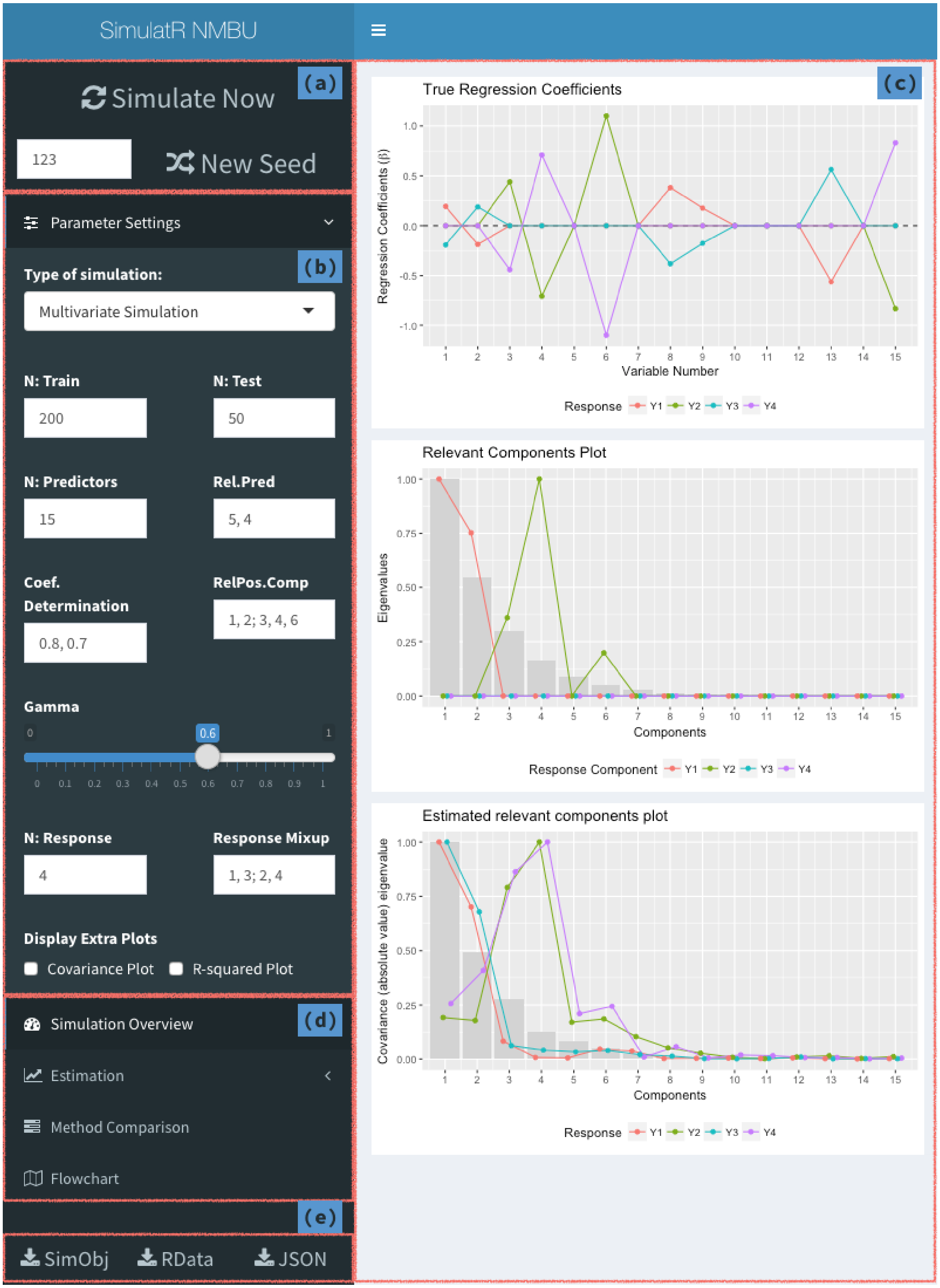
Inge S. Helland, S. Sæ bø, T. Almø y, and R. Rimal. Model and estimators for partial least squares. 2017.

Ulf G Indahl, Kristian Hovde Liland, and Tormod Næs. Canonical partial least squares–a unified pls approach to classification and regression problems. *Journal of Chemometrics*, 23(9):495–504, 2009.

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

Solve Sæbø, Trygve Almøy, Arnar Flatberg, Are H Aastveit, and Harald Martens. Lpls-regression: a method for prediction and classification under the influence of background information on predictor variables. *Chemometrics and Intelligent Laboratory Systems*, 91(2): 121–132, 2008.

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.



Seed: 123

Type of simulation: Multivariate Simulation

N: Train: 200

N: Test: 50

N: Predictors: 15

Rel.Pred: 5, 4

Coef. Determination: 0.8, 0.7

RelPos.Comp: 1, 2; 3, 4, 6

Gamma: 0.6

N: Response: 4

Response Mixup: 1, 3; 2, 4

Display Extra Plots

□ Covariance Plot □ R-squared Plot

Figure 4: Application interface of ‘simulatr‘. (a) Seed and simulation button (b) Parameter control panel (c) Properties of simulated data (d) Additional analysis (e) Download option of simulated data