A tool for simulating multi-response linear model data

Raju Rimal[a,*], Trygve Almøy[a], Solve Sæbø[b]

[a] *Faculty of Chemistry and Bioinformatics, Norwegian University of Life Sciences, Ås, Norway*
[b] *Prorector, Norwegian University of Life Sciences, Ås, Norway*

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 an extension to the R-package `simrel`, which is a versatile and transparent tool for simulating linear model data with an extensive range of adjustable properties. The method is based on the concept of relevant components, and is equivalent to the newly developed envelope model. It is a multi-response extension of R-package `simrel` which is available in R-package repository CRAN, and as `simrel` the new approach is essentially based on random rotations of latent relevant components to obtain a predictor matrix $\mathbf{X}$, but in addition we introduce random rotations of latent components spanning a response space in order to obtain a multivariate response matrix $\mathbf{Y}$. The properties of the linear relation between $\mathbf{X}$ and $\mathbf{Y}$ are defined by a small set of input parameters which allow versatile and adjustable simulations. Sub-space rotations also allow for generating data suitable for testing variable selection methods in multi-response settings. The method is implemented as an update to the R-package `simrel`.

Keywords: `simrel` package in r,data simulation,linear model

*Corresponding Author
Email addresses: `raju.rimal@nmbu.no` (Raju Rimal), `trygve.almoy@nmbu.no` (Trygve Almøy), `solve.sabo@nmbu.no` (Solve Sæbø)

1. Introduction

Technological advancement has opened a door for complex and sophisticated scientific experiments that were not possible before. Due to this change, enormous amounts of raw data are generated which contain massive information but is 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 and explore its properties. 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 extension to the r-package called `simrel`, that is versatile in nature and yet simple to use.

The simulation method we are presenting here is based on the principle of relevant space for prediction [13] which assumes that there exists a y-relevant subspace in the complete space of predictor variables that is spanned by a subset of eigenvectors of these predictor variables. Our extension to this principle is to introduce a subspace in $\mathbf{y}$ (material space) which contains the information that predictor space is relevant for. The concept of response reduction to the material space in response variable was introduced by Cook et al. [6]. Our r-package based on this principle lets the user specify various population properties such as; which latent components in $\mathbf{x}$ are relevant for a latent subspace of the responses $\mathbf{y}$ and the 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, Johnson [16]; Ripley [17] and Gamerman and Lopes [9] have exhaustively discussed the topic. In particular, methods based on covariance structure has been discussed by Arteaga and Ferrer [2]; Arteaga and Ferrer [3] and Camacho [4], following approaches to find simulated data satisfying the desired correlation structure. In addition, many publications have implemented simulated data in order to investigate new

estimation methods and prediction strategies [see: 8, 5, 14]. 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. [19] and as the r-package `simrel`. This paper extends `simrel` in order to simulate linear model data with multivariate response. The github repository of the package at http://github.com/simulatr/simrel has rich documentation with many examples and cases along with detailed descriptions of simulation parameters. In the following two sections, the discussion encircle the mathematical framework behind. In addition, in section 4 and 5 we have also discussed the input parameters needed for simrel function in brief. In section 4, an implementation is presented as a case example and the final section introduces the shiny web application for this tool.

2. Statistical Model

In this section we describe the model and the model parameterization which is assumed throughout this paper. We assume:

$$\begin{bmatrix}\mathbf{y}\\ \mathbf{x}\end{bmatrix} \sim N\left(\begin{bmatrix}\boldsymbol{\mu}_y\\ \boldsymbol{\mu}_x\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Sigma}_{yy} & \boldsymbol{\Sigma}_{yx}\\ \boldsymbol{\Sigma}_{yx}^t & \boldsymbol{\Sigma}_{xx}\end{bmatrix}\right) \tag{1}$$

where, $\mathbf{y}$ is a response vector with m response variables $y_1, y_2, \ldots y_m$ with mean vector $\boldsymbol{\mu}_y$, and $\mathbf{x}$ is vector of p predictor variables with mean vector $\boldsymbol{\mu}_x$. Further,

$\boldsymbol{\Sigma}_{yy}(m \times m)$	is the variance-covariance matrix of $\mathbf{y}$
$\boldsymbol{\Sigma}_{xx}(p \times p)$	is the variance-covariance matrix of variables $\mathbf{x}$
$\boldsymbol{\Sigma}_{yx}(m \times p)$	is the matrix of covariance between $\mathbf{x}$ and $\mathbf{y}$

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 $(m \times p)$ matrix of regression coefficients, and $\boldsymbol{\varepsilon}$ is an error term such that $\boldsymbol{\varepsilon} \sim N\left(0, \boldsymbol{\Sigma}_{y|x}\right)$. The properties of the linear model (2) can be expressed in terms of covariance matrices in (1).

Regression Coefficients The matrix of regression coefficients is given by

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

Coefficient of Determination Since, a matrix of coefficient-of-determination represents the proportion of variation explained by the predictors, we can write this matrix by its elements as,

$$\left(\rho_y^2\right)_{jj'} = \frac{\sigma_{xy_j}^t \Sigma_{xx}^{-1} \sigma_{xy_{j'}}}{\sqrt{\sigma_{y_j}^2 \sigma_{y_{j'}}^2}} \forall j, j' = 1 \ldots m$$

where, σ_{xy_j}, $\sigma_{xy_{j'}}$ are covariances between $\mathbf{x}$ and y_j, $y_{j'}$ respectively. Also, $\sigma_{y_j}^2$ and $\sigma_{y_{j'}}^2$ are unconditional variances of y_j and $y_{j'}$.

Here the numerator is equivalent to the covariance of fitted $\mathbf{y}$ in sample space. if $j = j'$, it corresponds to a population version of the mean sum of squares of regression. The denominator gives the total unconditional variation in $\mathbf{y}$. The diagonal elements of this matrix is the proportion of variation in a response $y_j, j = 1, \ldots m$ explained by the predictors.

Conditional variance The conditional variance-covariance matrix 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 minimum least squared error of prediction $\left[\mathrm{E}(y - \hat{y})^2\right]$ 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 that rotate $\mathbf{x}$ and $\mathbf{y}$ to yield $\mathbf{z}$ and $\mathbf{w}$,

respectively. The model (1) can be re-expressed in terms of these transformed variables as:

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

In addition, a linear model relating $\mathbf{w}$ conditioned on $\mathbf{z}$ 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 the regression coefficient vector 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 matrices, i.e., $\mathbf{Q}^t\mathbf{Q} = \mathbf{I}_m$ and $\mathbf{R}^t\mathbf{R} = \mathbf{I}_p$, the inverse transformation can be defined as,

$$\begin{aligned}\boldsymbol{\Sigma}_{yy} = \mathbf{Q}^t\boldsymbol{\Sigma}_{ww}\mathbf{Q} \quad \boldsymbol{\Sigma}_{yx} = \mathbf{Q}^t\boldsymbol{\Sigma}_{wz}\mathbf{R}\\ \boldsymbol{\Sigma}_{xy} = \mathbf{R}^t\boldsymbol{\Sigma}_{zw}\mathbf{Q} \quad \boldsymbol{\Sigma}_{xx} = \mathbf{R}^t\boldsymbol{\Sigma}_{zz}\mathbf{R}\end{aligned} \tag{5}$$

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

Regression Coefficients :

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

Conditional Variance Further, the conditional variance-covariance matrix of $\mathbf{w}$ given $\mathbf{z}$ is,

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

Coefficient of Determination The coefficient-of-determination matrix corresponding to $\mathbf{w}$ can be written as,

$$\begin{aligned}\left(\rho_w^2\right)_{jj'} &= \mathbf{\Sigma}_{ww}^{-1/2}\mathbf{\Sigma}_{wz}\mathbf{\Sigma}_{zz}^{-1}\mathbf{\Sigma}_{zw}\mathbf{\Sigma}_{ww}^{-1/2} \\ &= \frac{\sigma_{zw_i}^t \Sigma_{zz}^{-1}\sigma_{zw_{j'}}}{\sqrt{\sigma_{w_i}^2\sigma_{w_{j'}}^2}}\forall j, j' = 1\ldots m\end{aligned}$$

where, σ_{zw_j} and $\sigma_{zw_{j'}}$ are covariances of $\mathbf{z}$ with w_j and $w_{j'}$, respectively. Also, $\sigma_{w_j}^2$ and $\sigma_{w_{j'}}^2$ are unconditional variances of w_j and $w_{j'}$. For simplicity, we will denote $\sigma_{z_i w_j}$ by σ_{ij}.

Since the rotation matrices give a direct connection between the covariance of (1) and (3), a straight forward relationship can be worked out between the terms in the above given matrix and their counterpart covariance matrices of the $\mathbf{xy}$-space.

From the eigenvalue decomposition principle, if $\mathbf{\Sigma}_{xx} = \mathbf{R\Lambda R}^t$ and $\mathbf{\Sigma}_{yy} = \mathbf{Q\Omega Q}^t$ 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, $\mathbf{\Lambda}$ and $\mathbf{\Omega}$ are diagonal matrices of eigenvalues of $\mathbf{\Sigma}_{xx}$ and $\mathbf{\Sigma}_{yy}$, respectively.

3. Relevant Components

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 $\mathbf{x}$ is a vector of random predictors. Following the concept of relevant space and irrelevant space which is discussed extensively in Helland and Almøy [13], Helland [12], Helland et al. [14], Cook et al. [5], and Sæbø et al. [19], we can assume that there exists a subspace of the full predictor space which is relevant for y. An orthogonal space to this space does not contain any information about y and is considered as irrelevant. Here, the $y-$relevant

subspace of $\mathbf{x}$ is spanned by a subset of the principal components defined by the eigenvectors of the 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 envelopes [8] where relevant (material) and irrelevant (immaterial) space were discussed for both response and predictor variables.

3.1. Model Parameterization

In order to construct a fully specified and unrestricted covariance matrix of $\mathbf{z}$ and $\mathbf{w}$ for the model in equation (3), we need to identify $1/2(p+m)(p+m+1)$ unknown parameters. For the purpose of simulation, we implement some assumptions to re-parameterize and simplify the model. This enables us to construct a wide range of model properties from only few key parameters.

Parameterization of $\mathbf{\Sigma}_{zz}$ If we let the rotation matrix $\mathbf{R}$ correspond 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 same parametric representation as Sæbø et al. [19] for these eigenvalues:

$$\lambda_i = e^{-\gamma(i-1)}, \gamma > 0 \text{ and } i = 1, 2, \ldots, p \tag{6}$$

Here, as γ increases, the decline of eigenvalues becomes steeper, hence the parameter γ controls the level of multicollinearity in $\mathbf{x}$. We can write $\Sigma_{zz} = \Lambda = \text{diag}(\lambda_1, \ldots, \lambda_p)$.

Parameterization of $\mathbf{\Sigma}_{ww}$ In similar manner, a parametric representation of eigenvalues corresponding to Σ_{ww} is adopted as,

$$\kappa_j = e^{-\eta(j-1)}, \eta > 0 \text{ and } j = 1, 2, \ldots, m \tag{7}$$

Here, the decline of eigenvalues becomes steeper as η increases from zero. At $\eta = 0$, all w will have equal variance 1. Hence we can write $\Sigma_{ww} = \text{diag}(\kappa_1, \ldots, \kappa_m)$.

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 z is irrelevant for the given variable w. The non-zero elements define which of the z that are relevant for $\mathbf{w}$. We typically refer to the indices of these z variables as the positions of relevant components. In order to re-parameterize this covariance matrix, it is necessary to discuss the position of relevant components in detail.

3.1.1. Position of relevant components

Let k_1 components be relevant for w_1, k_2 components be relevant for 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 w_j and z_i is non-zero only if z_i is relevant for w_j. If σ_{ij} is the covariance between w_j and 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 α for w_j (4) 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 positions of the relevant components have heavy impact on prediction. Helland and Almøy [13] have shown that if the 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, z_1 and z_2, are relevant for w_1 and fifth and sixth components, z_5 and z_6, are relevant for w_2, it is relatively easier to predict w_1 than w_2, other properties being similar. This might be so, because the first and second principal components have larger variances than the fifth and sixth components.

Although the covariance matrix may depend 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,

$$\begin{aligned}\boldsymbol{\rho}_w^2 &= \boldsymbol{\Sigma}_{ww}^{-1/2}\boldsymbol{\Sigma}_{zw}^t\boldsymbol{\Sigma}_{zz}^{-1}\boldsymbol{\Sigma}_{zw}\boldsymbol{\Sigma}_{ww}^{-1/2} \\ &= \frac{\sigma_{ij}^t\Lambda^{-1}\sigma_{ij'}}{\sqrt{\sigma_j^2\sigma_{j'}^2}}\forall j, j' = 1\dots m\end{aligned}$$

Applying our assumptions that, $\boldsymbol{\Sigma_{ww}} = \text{diag}(\kappa_1, \dots, \kappa_m)$ (7) and $\boldsymbol{\Sigma}_{zz} = \Lambda = \text{diag}(\lambda_1, \dots, \lambda_p)$ (6), we obtain,

$$\rho_w^2 = \Sigma_{ww}^{-1/2}\Sigma_{zw}^t\Lambda^{-1}\Sigma_{zw}\Sigma_{ww}^{-1/2} = \begin{bmatrix} \sum_{i=1}^{p}\frac{\sigma_{i1}^2}{\lambda_i\kappa_1} & \dots & \sum_{i=1}^{p}\frac{\sigma_{i1}\sigma_{im}}{\lambda_i\sqrt{\kappa_1\kappa_m}} \\ \vdots & \ddots & \vdots \\ \sum_{i=1}^{p}\frac{\sigma_{i1}\sigma_{im}}{\lambda_i\sqrt{\kappa_1\kappa_m}} & \dots & \sum_{i=1}^{p}\frac{\sigma_{im}^2}{\lambda_i\kappa_m} \end{bmatrix}$$

Furthermore, we assume that there are no overlapping relevant components for any two w, i.e, $\mathcal{P}_j \cap \mathcal{P}_{j*} = \varnothing$ or $\sigma_{ij}\sigma_{ij*} = 0$ for $j \neq j*$. The additional unknown parameters in the diagonal of $\boldsymbol{\rho}_w^2$ should agree with user specified coefficients of determination for $\mathbf{w}$. i.e, $\rho_{w_j}^2$ is,

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

Here, only the relevant components have non-zero covariances with w_j, so,

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

For some user defined $\rho_{w_j}^2$ the σ_{ij}^2 is determined as follows,

1. Sample k_j values from a uniform distribution $\mathcal{U}(-1, 1)$ distribution. Let them be denoted $\mathcal{S}_{\mathcal{P}_1}, \dots, \mathcal{S}_{\mathcal{P}_{k_j}}$.

2. Define,

$$\sigma_{ij} = \text{Sign}(\mathcal{S}_i)\sqrt{\frac{\rho^2_{w_j}\,|\mathcal{S}_i|}{\sum_{k\in\mathcal{P}_j}|\mathcal{S}_k|}\lambda_i\kappa_j}$$

for $i \in \mathcal{P}_j$ and $j = 1, \ldots m$

This means that the covariances between the predictor components and the response components are sampled randomly, but with restriction that the requested $\rho^2_{w_j}$ values are satisfied. This also implies that the regression coefficients $\boldsymbol{\alpha}$ in (4) and $\boldsymbol{\beta}$ in (2) are sampled randomly under the same restriction.

3.1.2. Data Simulation

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

Similarly, in order to simulate n observations, we define $\underset{n\times(m+p)}{\mathbf{G}} = \mathbf{U}\boldsymbol{\Sigma}^{-1/2}$. 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 the 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. In the following section we discuss these rotation matrices in more detail.

3.2. Rotation of predictor space

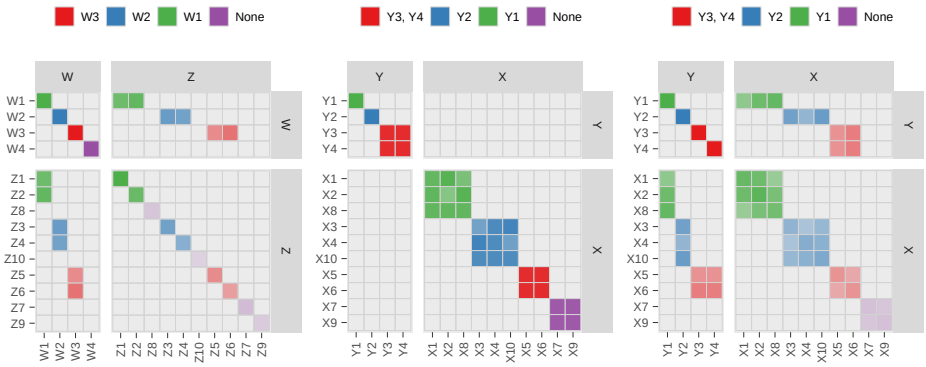
Initially, let us consider an example where a regression model with $p = 10$ predictors $(\mathbf{x})$ and $m = 4$ responses $(\mathbf{y})$. Let's assume that only three response components $(w_1, w_2$ and $w_3)$ are needed to describe all four response variables. Further, let the index sets $\mathcal{P}_1 = \{1, 2\}, \mathcal{P}_2 = \{3, 4\}$ and $\mathcal{P}_3 = \{5, 6\}$ define the positions of the predictor 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 predictor 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. [5].

Moreover, let $q_1 = 3, q_2 = 3$ and $q_3 = 2$ be the number of predictor variables we want to have relevant for w_1, w_2 and w_3 respectively. Then $q_1 = 3$ predictors may be obtained by rotating the predictor components in $\mathcal{P}_1$ along with one more irrelevant component. Similarly, $q_2 = 3$ predictors, relevant for w_2, can be obtained by rotating predictor components in $\mathcal{P}_2$ along with one more irrelevant component and finally, $q_3 = 2$ predictors, relevant for w_3, can be obtained by rotating the 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 $\mathcal{S}_k$ since in the process two irrelevant components were included in the rotations. 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 the complete variable space $\mathcal{S}_p$ into two orthogonal spaces – $\mathcal{S}_k$ relevant for $\mathbf{w}$ and $\mathcal{S}_{p-k}$ irrelevant for $\mathbf{w}$.

In the previous section, we discussed about the construction of a covariance matrix for the latent structure. Figure 1(a) shows a similar structure resembling the example here. The three colors represent the 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 predictor 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 the predictor components $\mathbf{z}$. This defines a new basis for the same space as is spanned by the predictor components. In principle, there are many possible options for defining a rotation matrix. Among them, the eigenvector matrix of $\mathbf{\Sigma}_{xx}$ can be a candidate. However, in this reverse engineering approach 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 satisfies the properties of a real valued rotation matrix, i.e $\mathbf{R}^{-1} = \mathbf{R}^t$ and $\det(\mathbf{R}) = \pm 1$ so that $\mathbf{R}$ is orthonormal. Here the rotation matrix $\mathbf{R}$ should be block diagonal as

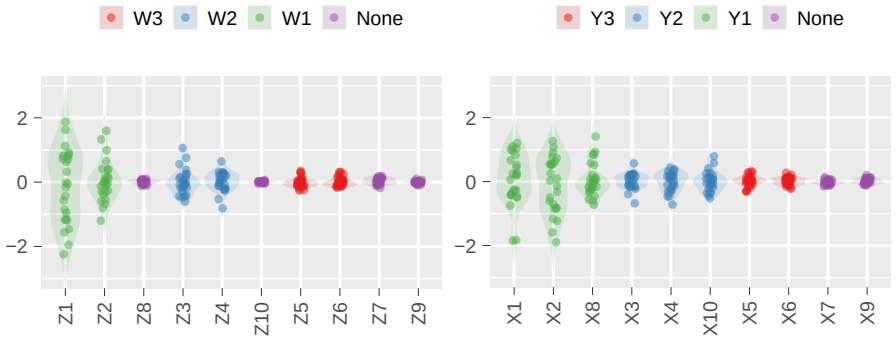


(a) Relevant Position (b) Rotation Matrix (c) Relevant Predictors

Figure 1: Simulation of predictor and response variables after orthogonal transformation of predictor and response components by rotation matrices Q and R shown as the upper left and the lower right block matrices in (b).

in Figure 1(b) in order to rotate spaces $\mathcal{S}_1, \mathcal{S}_2 \ldots$ separately. Figure 2(a) shows the simulated predictor components $\mathbf{z}$ that we are following in our example where we can see that the components z_1 and z_2 (relevant for w_1) is getting rotated together with an irrelevant component z_8. The resultant predictors (Figure 2(b)) x_1, x_2 and x_8 will hence also be relevant for w_1. In the figure, we can see that components z_7, z_8, z_9 and z_{10} are not relevant for any responses before rotation, however, the x_8, x_{10} predictors become relevant after rotation keeping x_7 and x_9 still irrelevant.

Among several methods [1, 11] 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 variates. The rotation here can be a) restricted and b) unrestricted. The latter rotates all components $\mathbf{z}$ together and makes all predictor variables somewhat relevant for all response components. However, the former performs a block-wise rotation so that it rotates certain selected predictor 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



(a) Before Rotation (b) After Rotation

Figure 2: Simulated data before and after rotation

which can be detected during variable selection procedures.

3.3. Rotation of response space

The previous example has four response variables with only three informative components w_1, w_2 and w_3. During the rotation procedure, the response space is also rotated along with the predictor space. Figure 1 shows that the informative response component w_3 is rotated together with the 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 only be relevant for y_1 and y_2 respectively, which we can see in Figure~2. Although the response components have exclusive set of relevant predictors, the rotation of the response space has the potential of creating several response variables that depend on the same relevant predictor space. In the r-package *simrel*, the combining of the response components is specified by a parameter `ypos`.

4. Implementation

This section demonstrates an application of multi-response extension of `simrel` with two examples in order to compare different estimation methods on

the basis of prediction error. These example are simply a demonstration of the use of `simrel` package rather than an extensive comparison of methods.

4.1. Example 1

For the comparison, we have considered four well established estimation methods.

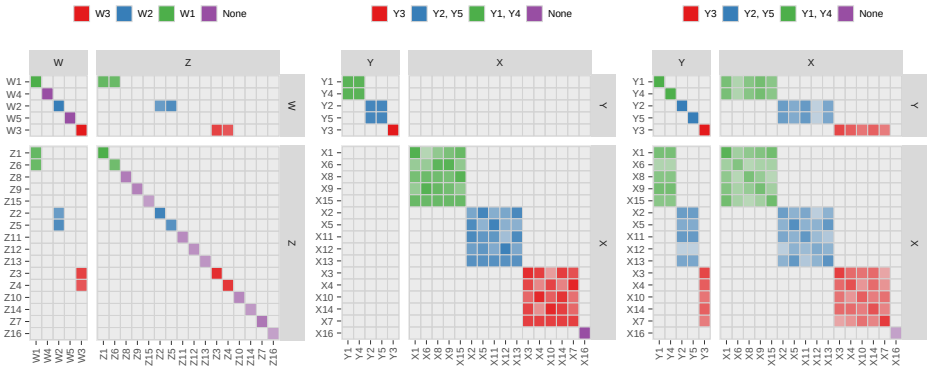
a) Ordinary Least Squares (OLS),
b) Principal Component Regression (PCR),
c) Partial Least Squares predicting individual response variable separately (PLS1) and
d) Partial Least Squares predicting all response variables together (PLS2).

We have also considered four relatively new estimation methods in multi-response regression:

a) Canonically Powered Partial Least Squares regression (CPPLS) [15],
b) Canonical Partial Least Squares regression (CPLS) [15],
c) Envelope estimation in predictor space (Xenv) [6],
d) Envelope estimation in response space (Yenv) [7] and
e) Simultaneous estimation of x- and y-envelope (Senv) [8]

From the possible combinations of two levels of coefficient of determination (ρ^2) and two levels of γ (6) (the factor that controls the multicollinearity in predictor variables), four simulation designs (design 1-4) were prepared. Replicating each design 20 times, 80 datasets with five response variables ($m = 5$) and 16 predictor variables ($p = 16$) were simulated using the method discussed in this paper. It was also assumed that three response components (w_1, w_2 and w_3) completely describe the variation present in five response variables ($y_1 \dots y_5$). Here, in this example we have assumed that all w's have equal variance, i.e. $\Sigma_{ww} = \mathbf{I}_m$, that is, $\eta = 0$ in (7). The four designs are presented in Table~1. All datasets contained 100 sampled observations and out of 16 predictor variables, three disjoint sets of five predictor variables each are relevant for response components

w_1, w_2 and w_3. Although the simulation method is well equipped to simulate data with $p \gg n$, for incorporating envelope estimation methods, which are based on maximization of likelihood, we have chosen a $n > p$ situation in the example. Further, predictor components z_1 and z_6 were relevant for response component w_1, predictor components z_2 and z_5 were relevant for response component w_2 and predictor component z_3 and z_4 were relevant for response component w_3. In addition, following the discussion about rotation of response space (section 3.3), w_1 was rotated together with w_4 and w_2 was rotated together with w_5. Figure 3 visualizes the covariance structure and relationship between the response and predictor variables for the first design.



(a) Relevant Position (b) Rotation Matrix (c) Relevant Predictors

Figure 3: Simulation of predictor and response variables for design one after orthogonal transformation of predictor and response components by rotation matrices Q and R shown as the upper left and the lower right block matrices in (b). Here (a) is the covariance structure of the latent space, which is rotated by the block diagonal rotation matrix in (b) resulting the covariance structure of simulated data in (c).

For each method, we can write an expected squared prediction error as,

$$\underset{m\times m}{\boldsymbol{\vartheta}} = \mathrm{E}\left[\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)^{t} \boldsymbol{\Sigma}_{xx} \left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)\right] + \boldsymbol{\Sigma}_{y|x} \tag{8}$$

where, $\hat{\boldsymbol{\beta}}$ is an estimate of the true regression coefficient $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}_{xx}$ is the true covariance structure of the predictor variables obtained from `simrel`. Also,

Table 1: Parameter setting of simulated data for comparison of estimation methods

	Decay of eigenvalues (γ)	Coef. of Determination ($\rho^2_{w_j}$)
Design1	0.2	0.8, 0.8, 0.4
Design2	0.8	0.8, 0.8, 0.4
Design3	0.2	0.4, 0.4, 0.4
Design4	0.8	0.4, 0.4, 0.4

$\mathbf{\Sigma}_{y|x}$ is the true minimum error of the model. Here $\hat{\boldsymbol{\beta}}$ varies across different estimation methods while the remaining terms are the same for each dataset design. The expression in (8) is estimated from 20 replicated calibration sets. Further, an overall prediction error of all responses is measured by the trace of $\boldsymbol{\vartheta}$ (8).

The minimum prediction error (measured as discussed above) for nine estimation methods averaged over 20 replications of four designs are shown in Table 2. The table also gives the number of predictor components (response components in case of `Yenv`), a method has used in order to obtain the minimum of average prediction error.

Table 2 shows that the simultaneous 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. The PCR model has the smallest prediction error (4.08) from 6 components in design 3 and Canonically 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 across different replicated dataset, but the component corresponding to minimum prediction error is discussed here. A detailed picture of prediction error for each estimation method obtained for each additional component is shown in Figure 4. Although designs 2 and 4 have higher levels of

Table 2: Minimum average prediction error (number of components corresponding to minimum prediction error, minimum prediction error) (For Yenv, the number of response components is given)

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, 3.60)	(1, 3.58)	(1, 4.57)	(1, 4.50)
PCR	(7, 3.28)	(6, 3.19)	(6, 4.08)	(6, 4.04)
PLS1	(2, 3.32)	(5, 3.20)	(1, 4.16)	(5, 4.07)
PLS2	(5, 3.29)	(6, 3.19)	(3, 4.11)	(6, 4.06)
Senv	(4, 3.17)	(5, 3.14)	(3, 4.35)	(5, 4.28)
Xenv	(5, 3.23)	(6, 3.20)	(5, 4.10)	(6, 4.11)
Yenv	(3, 3.24)	(3, 3.23)	(3, 4.29)	(3, 4.24)

multicollinearity, the performance of the estimation methods is indifferent to its effect. Since all methods, except OLS, are based on shrinking of estimates, they are less influenced by the multicollinearity problem.

The analysis presented in Figure 4 has addressed some questions such as how methods work when there exist a true reduced dimension in response space, but also raised other questions like why they perform differently. For example, what is the reason for the decreasing relative performance of the simultaneous envelope method as the ρ^2 values are reduced? Does this depend on the dimensions and shape of the $\mathbf{y}$ envelopes? Since the example is merely intended as a demonstration of how `simrel` can be used in scientific study, a more elaborative studies would be necessary to answer such questions, but for this purpose `simrel` would be a powerful tool.

4.2. Example 2

In this second example, wide matrices with 100 observations and 1000 predictor variables were simulated. Since wide matrices are common in various fields such as genomics, spectroscopy and chemometrics, we set up this second

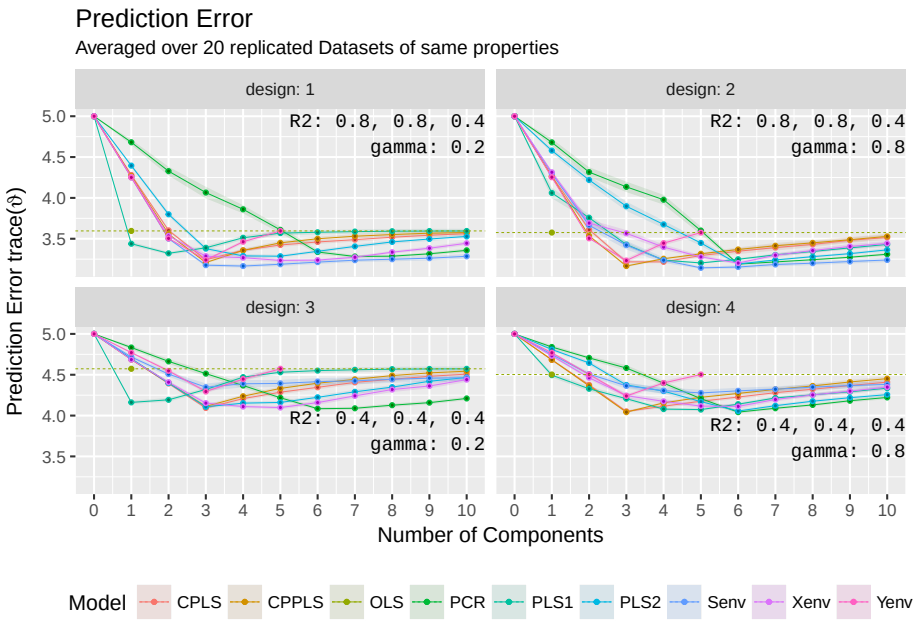


Figure 4: Minimum of Average Prediction Error

example to compare two variants of partial least square regression – PLS1 and PLS2. While estimating regression coefficients PLS1 uses each response variable separately, while PLS2 uses them all simultaneously. A simulation design was constructed as in Table 3. With each design, 20 replicated datasets were simulated having five response variables and a moderate level of multicollinearity within the predictor variables ($\gamma = 0.5$).

The comparison were based on the prediction error measured by root mean squares of prediction (RMSEP). In order to approximate the error to theoretically computed error, 1000 extra test samples were drawn from the same distribution as the training samples during simulation.

With the simulated data models with one to ten components were fitted and the prediction error was recorded for each response variable and each additional component.

The first and second design in Table 3 has one informative response component for which four predictor components are relevant at positions 2, 3, 5 and 7, and

Table 3: Simulation Design of second example

$\eta : 0.1$	$\eta : 0.8$	Parameter	Value
Single Informative Response Component			
Design 1	*Design 2*	relpos	2, 3, 5, 7
		q	1000
		R2	0.8
Two Informative Response Components			
Design 3	*Design 4*	relpos	2; 3
		q	500; 500
		R2	0.6; 0.6

the coefficient of determination is 0.8. Since the informative response component is rotated together with four uninformative response components, the information is shared among all five response variables after rotation.

The third and fourth design has two informative response components. The first response component has one relevant predictor component at position 2 and a coefficient of determination of 0.6. Similarly, the second response component has one relevant predictor component at position 3 and also here the coefficient of determination is 0.6.

In addition to having one and two response component models, two levels of variance structure of the response components is considered and defined by η parameters with values 0.1 and 0.8 respectively. In the first and third design, all response components vary in similar manner ($\eta = 0.1$), while in the second and fourth design the informative response components have higher variance ($\eta = 0.8$) than the uninformative ones as the eigenvalues of Σ_{ww} drop faster in this case.

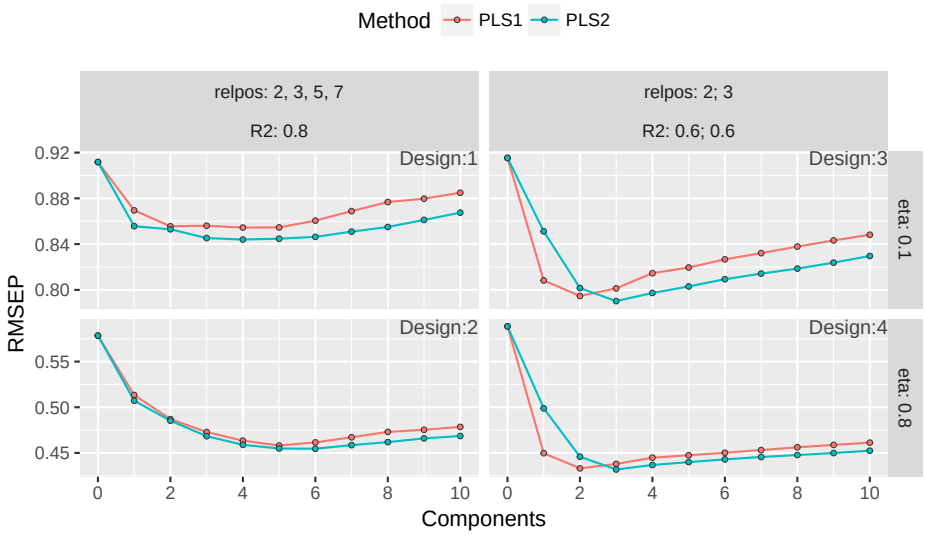


Figure 5: Root mean square of error of prediction of test observation averaged over all response variables.

Figure 5 shows the average prediction error of test observations modelled by PLS1 and PLS2 for all four designs. The prediction errors are averaged over all 20 replicated datasets.

In general, PLS2 dominates PLS1 with regard to minimum error achieved for these simulated designs. The difference is largest for the designs with $\eta = 0.1$ in which case the response are moderately correlated and prediction appears to be more difficult than for $\eta = 0.8$. The effect of number of relevant response and predictor components appears to have less influence on the results than the covariance structure of Σ_{yy}. This small example of the use of simrel indicates that a more elaborate comparison study should be done on PLS1 and PLS2 in this respect.

5. Web Interface

In order to give an alternative interface for `simrel`, we have created a shiny app which allows users to provide the simulation parameters through different input fields. Figure 6 shows a screenshot of 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 Figure 6 (a) so that a particular set of data can be re-simulated if needed. Figure 6 (b) has all the input panels where the user-dependent parameters for simulation can be entered. Here the user also has the option to simulate univariate, bivariate or multivariate response data. In addition, a simulated R-object comprising the simulated data can be downloaded in `Rdata` format (section (e) in Figure 6). The object holds the simulated data along with other properties such as coefficient of determination for each response, true regression coefficients and rotation matrices. Users can also download simulated data in JSON and CSV format.

All `simrel` parameters can be entered using a simple user interface where vector elements are separated with comma(,) and list elements are separated with semicolon(;). For instance, the relevant position discussed in the implementation (section 4) 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))`. An R expression equivalent to the input parameters as shown in Figure - 5(b) can be written as,

```
simrel(
  n      = 200, # Number of training observations
  ntest  = 50, # Number of test observations
  p      = 15, # Number of predictor variables
  q      = c(5, 4), # Number of relevant predictors
  relpos = list(c(1, 2), c(3, 4, 6)),
         # Relevant predictor components
  R2     = c(0.8, 0.7), # Rsq for each response component
  m      = 4, # Number of response variables
  gamma  = 0.6, # Decay factor of eigenvalues of predictors
  eta    = 0, # Decay factor of eigenvalues of responses
  ypos   = list(c(1, 3), c(2, 4)),
         # Combination of response components on rotation
```

```
  type   = "multivariate"
)
```

With the parameters for simulation in the screenshot (Figure-6) 200 training sets (`n`) and 50 test sets (`ntest`) will be simulated with 15 predictor variables (`p`) and 4 response variables (`m`). The 4 response variables will have a true latent dimension of two, which is spanned by the relevant *response components*. The first response component is rotated together with the third (irrelevant) response component and the second response component is rotated together with the fourth (irrelevant) response component as set in `ypos`. Out of 15 predictors, 5 will be relevant for the first response component and 4 will be relevant for the second response component, as set by `q`. The 5 predictor variables, that are relevant for the first response component, span the same space as the predictor components at position 1 and 2. Similarly, the 4 predictor variables that are relevant for the second response component, span the same space as the predictor components at position 3, 4 and 6 (`relpos`). The coefficient of determination for the first and second response components are 0.8 and 0.7, respectively (`R2`). The eigenvalues of the predictor components decay exponentially by the factor of 0.6 (`gamma`), whereas the eigenvalues of response components are constant (but can be set to exponential decay) (`eta`).

The application not only allows users to simulate data, but also gives some insight into simulated data properties. Section (c) in Figure 6 contains three plots – a) true regression coefficients b) relevant components and c) estimated relevant components. In the first plot (Figure-6(c) top) we can see that predictor variables (1, 2, 8, 9 and 13) are relevant for the first and third response variables (red and blue line) by their non-zero coefficients, whereas predictor variables (3, 4, 6 and 15) are relevant for the second and fourth response variables (purple and green line). The second plot (Figure 6(c) middle) shows the covariances between the response components and the predictor components along with the corresponding eigenvalues in the background (bar plot). In the plot the absolute value of the covariances after scaling with the largest covariance are shown. As

in our parameter setting, the plot shows that the first (red line) and second (green line) predictor components have non-zero covariance with the first and third response components, and the fourth and sixth predictor components have non-zero covariance with the second response component. The third plot (Figure-6(c) bottom) is the estimated covariances between the predictor components and the response variables, for the simulated data. Since the first and third response components are rotated together, in the plot, the covariance between the predictor components and the first and third response variables (red and blue line) are following similar patterns as the theoretical (6(c) middle). This also suggests that the predictor components which were relevant for the first response component, becomes relevant for the first and third response variables after rotation.

Along with these main sections, section (d) in the same figure contains additional analysis performed on the simulated data such as its estimation with different methods. This section is intended for educational purposes to show how changing the data properties influences the performances of different estimation and prediction methods. Beside this application, for Rstudio users, a gadget will be available after installing the r-package. This gadget provides an interface enabling users to input simulation parameters and access some of the properties.

Many scientific studies [14, 18, 8] 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 versatile simulation tool like `simrel` can play an important role. Gangsei et al. [10] and Sæbø et al. [19] are some examples where the univariate and bivariate form of `simrel` have been used for such purposes.

6. Conclusion

Whether comparing methods or assessing and understanding the properties of any method, tool or procedure; simulated data allows for controlled tests for researchers. However, researchers spend enormous amount of time creating such simulation tools so that they can obtain a particular nature of data. We believe

that this tool along with the R-package and the easy-to-use shiny web interface will become an assistive tool for researchers in this respect.

7. References

[1] Anderson, T. W., Olkin, I., Underhill, L. G., 1987. Generation of random orthogonal matrices. SIAM Journal on Scientific and Statistical Computing 8 (4), 625–629.

[2] Arteaga, F., Ferrer, A., 2010. How to simulate normal data sets with the desired correlation structure. Chemometrics and Intelligent Laboratory Systems 101 (1), 38–42.

[3] Arteaga, F., Ferrer, A., 2013. Building covariance matrices with the desired structure. Chemometrics and Intelligent Laboratory Systems 127, 80–88.

[4] Camacho, J., 2017. On the generation of random multivariate data. Chemometrics and Intelligent Laboratory Systems 160, 40–51.

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

[6] Cook, R. D., Li, B., Chiaromonte, F., 2010. Envelope models for parsimonious and efficient multivariate linear regression. Statistica Sinica, 927–960.

[7] Cook, R. D., Zhang, X., 2015. Foundations for envelope models and methods. Journal of the American Statistical Association 110 (510), 599–611.

[8] Cook, R. D., Zhang, X., 2015. Simultaneous envelopes for multivariate linear regression. Technometrics 57 (1), 11–25.

[9] Gamerman, D., Lopes, H. F., 2006. Markov chain Monte Carlo: stochastic simulation for Bayesian inference. CRC Press.

[10] Gangsei, L. E., Almøy, T., Sæbø, S., 2016. Theoretical evaluation of prediction error in linear regression with a bivariate response variable containing missing data. Communications in Statistics - Theory and Methods 0 (0), 1–9.
URL http://dx.doi.org/10.1080/03610926.2016.1222434

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

[12] Helland, I. S., Mar 2000. Model reduction for prediction in regression models. Scandinavian Journal of Statistics 27 (1), 1–20.
URL http://dx.doi.org/10.1111/1467-9469.00174

[13] Helland, I. S., Almøy, T., 1994. Comparison of prediction methods when only a few components are relevant. Journal of the American Statistical Association 89 (426), 583–591.

[14] Helland, I. S., Saebø, S., Tjelmeland, H., et al., 2012. Near optimal prediction from relevant components. Scandinavian Journal of Statistics 39 (4), 695–713.

[15] Indahl, U. G., Liland, K. H., Næs, T., 2009. Canonical partial least squares–a unified pls approach to classification and regression problems. Journal of Chemometrics 23 (9), 495–504.

[16] Johnson, M. E., 2013. Multivariate statistical simulation: A guide to selecting and generating continuous multivariate distributions. John Wiley & Sons.

[17] Ripley, B. D., 2009. Stochastic simulation. Vol. 316. John Wiley & Sons.

[18] Sæbø, S., Almøy, T., Flatberg, A., Aastveit, A. H., Martens, H., 2008. 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.

[19] Sæbø, S., Almøy, T., Helland, I. S., 2015. 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.

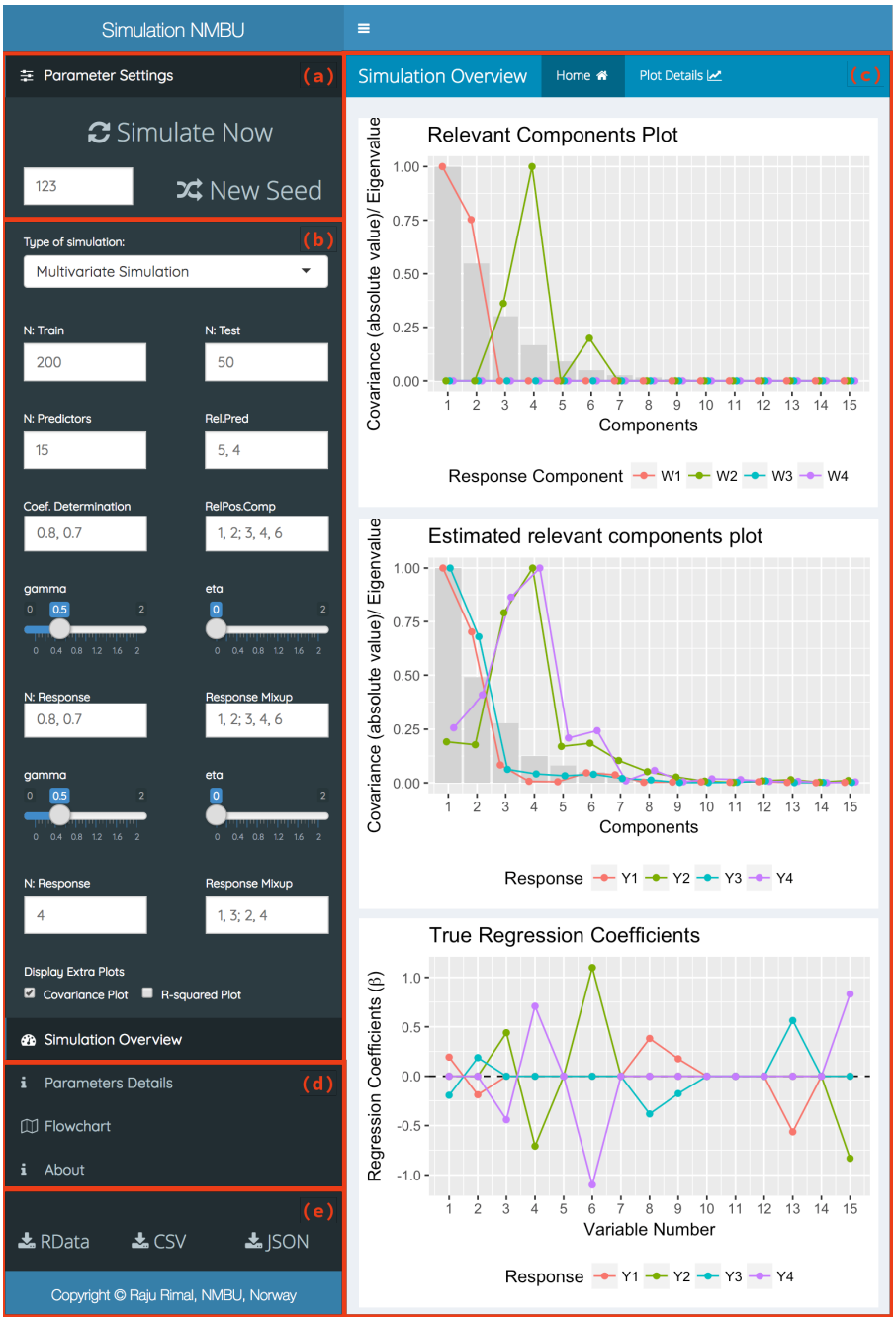


Figure 6: Web interface of shiny application of `simrel`: (**a**) Buttons to trigger simulation, (**b**) Parameters for simulation, (**c**) Visualization of the true properties of simulated data (regression coefficients, true and estimated covariance between response and predictors components) (**d**) Additional analysis (**e**) Download option of simulated data.