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.

The simulation method we are discussing here is based on principal of relevant space

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

March 9, 2017

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

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

The r-package `simrel-m` uses model parameterization which is based on the concept of relevant components Helland and Almøy (1994) where it is assumed that a subspace of response $\mathbf{Y}$ is spanned by a subset of eigenvectors corresponding to predictor space. A response space can be thought to have two mutually orthogonal space – relevant and irrelevant. Here the space of response matrix for which the predictors are relevant is termed as response components, and we assume that each response component is spanned by an exclusive subset of predictor variables. In this way we can construct a set of predictor variables which has non-zero regression coefficients. This also enables user to have uninformative predictors which can be detected during variable selection procedure. In addition, user can control signal-to-noise ratio for each response components with a vector of population coefficient of determination $\boldsymbol{\rho}^2$. Further, the collinearity between predictor variables can also be adjusted by a factor γ which controls decay factor of eigenvalue of $\mathbf{X}$ matrix. Helland and Almøy (1994) showed that if the direction of large variability (i.e., component corresponding to large eigenvalues) are also relevant relevant predictor space, prediction is relatively easy. In contrast, if the relevant predictors are on the direction of low variability, prediction becomes difficult. Table~1 shows all the parameters that a user can specify in `simrel-m`.

Table 1: Parameters for simulation used in this study

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

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

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

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

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

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

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

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

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

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

Here we have assumed that the relevant components are know, as in Helland and Almøy (1994), which is rare in practice. But in comparative studies of prediction methods, this can help to explain interesting cases.

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

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

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

Model Specification

A multi-response multivariate general linear model in equation-(3) is conidered as a simulation model.

Being an extension of `simrel` package, a quick summary of the procedure used in that package helps to underestand the literature in this paper.

An overview of `simrel`. `Simrel` is based on uni-response linear model as in equation~(4).

$$\begin{bmatrix} y \\ \mathbf{X} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_y \\ \boldsymbol{\mu}_X \end{bmatrix}, \begin{bmatrix} \sigma_y^2 & {\boldsymbol{\sigma}_{Xy}}^t \\ \boldsymbol{\sigma}_{Xy} & \boldsymbol{\Sigma}_{XX} \end{bmatrix}\right) \tag{4}$$

Equivalently,

$$\begin{bmatrix} \mathbf{Y} \\ \mathbf{X} \end{bmatrix} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = N\left(\begin{bmatrix} \boldsymbol{\mu}_Y \\ \boldsymbol{\mu}_X \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{YY} & \boldsymbol{\Sigma}_{XY}^t \\ \boldsymbol{\Sigma}_{XY} & \boldsymbol{\Sigma}_{XX} \end{bmatrix}\right) \tag{5}$$

Here,

$\boldsymbol{\Sigma}_{YY}$: Covariance Matrix of response **Y** without given **X**
$\boldsymbol{\Sigma}_{XY}$: Covariance Matrix between **X** and **Y**
$\boldsymbol{\Sigma}_{XX}$: Covariance matrix of predictor variables **X**
$\boldsymbol{\mu}_X$ and $\boldsymbol{\mu}_Y$: Mean vectors of response **Y** and predictor **X** respective

According to the theory of Multivariate Normal Distribution, we can express different parameters interms of **X**, **Y** and the covariance structure.

Model Parameterization. Parameter Definition:

Before continuing any further, it is necessary to define the parameters used here,

In the following section, some of these parameters are discussed in detail. The discussion has considered random $\mathbf{X}$ regression model with m response as given in equation~(3).

Parameters Explanation and Notation Used:

Let m responses are spanned completely by l response components. These l response components are combined with $m - l$ standard normal vectors by user defined criteria to get m responses after successive orthogonal transformation. Out of q, let $q_j,\ j = 1, \ldots l$ be the number of predictors that is relevant for response j. Let c_j be the number of eigenvectors/ components that completely span j^{th} predictor space containing q_j number of predictors. Further, the position of these components for j^{th} response be in index set $\mathcal{P}_j$. Here it is also assumed that the eigenvalues corresponding to $\mathbf{X}$ declines successively such that $\lambda_i, i = 1, \ldots, p$ such that $\lambda_i \geq \lambda_k, i > k$ are the eigen values of $\mathbf{X}$. The position index of eigenvalues corresponding to response j is in the set $\mathcal{P}_j$. We assume that the index are ordered within each sets so that, j^{th} index set contains c_j number of components. The eigenvalues corresponding to these components in $\mathcal{P}_j$ set is $\lambda_{\mathcal{P}_{jk}}, k = 1, \ldots c_j$ such that $\lambda_{\mathcal{P}_{jk}} > \lambda_{\mathcal{P}_{jk'}}$ for $k > k'$. In `Simrel-M` package, we refer this position by `relpos` argument. In addition, we suppose that the relevant components are exclusive for each response.

An Example:

Suppose we have a situation like,

Number of response	(m)	=	5
Number of response components	(l)	=	3
Position of relevant component for response 1	$(\mathcal{P}_1)$	=	$\{1,3\}$
Position of relevant component for response 2	$(\mathcal{P}_2)$	=	$\{2,4,5\}$
Position of relevant component for response 3	$(\mathcal{P}_3)$	=	$\{6\}$

such that, $\lambda_1 > \lambda_3$ in $\mathcal{P}_1$ and $\lambda_2 > \lambda_4 > \lambda_5$ in set $\mathcal{P}_2$. Here, the component (eigenvector) 1 and 3 are relevant for response component 1, component 2, 4 and 5 are relevant for response component 2 and component 6 is relevant for response component 3.

In `Simrel-M`, we have assumed that the eigenvalues are decreasing exponentially by factor γ and the largest eigenvalue is 1, i.e. for $\gamma > 0$, $\lambda_i = \mathrm{e}^{-\gamma(i-1)}$ for $i = 1, 2, \ldots p$.

References

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.

Inge S Helland and Trygve Almøy. Comparison of prediction methods when only a few components are relevant. *Journal of the American Statistical Association*, 89(426):583–591, 1994.

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

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

Solve Sæbø, Trygve Almøy, and Inge S Helland. simrel-a versatile tool for linear model data simulation based on the concept of a relevant subspace and relevant predictors. *Chemometrics and Intelligent Laboratory Systems*, 2015.