simrel-m – A versatile tool for data simulation for multi-response linear model data based on the concept of relevant subspace of predictor space

Raju Rimal[1], Trygve Almøy[1], 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 & 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 $\mathbf{X}$, 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 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`,

1. Acknoledgement

2. Introduction

General aspects

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 is now becoming another problem. In order to tackle this situation new methods are being developed. However,

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

before implementing any methods, 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 method is based on principal of relevant space for prediction which assumes that there exists a subspace in the complete space of responses that is spanned by a subset of eigenvectors of predictor variables. The method and the r-package based on this principle not only has ability to simulate wide range of multi-response linear model data but also let researcher to specify which components of predictors ($\mathbf{X}$) are relevant for a component of responses $\mathbf{Y}$. This enables the possibility to construct data for evaluating methods developed for variable selection.

A vast literature on simulation is present but most of them are developed to address the specific problems their study was dealing with. Sæbø et al. (2015) has presented a generic tool that is capable of simulating linear model data as an r-package `simrel`. This paper extends the methods to simulate multivariate response.

Application of `simrel-m`

The simple interface of `simrel-m` for sophisticated simulation opens for numerous applications in various disciplines among which some of them are discussed below.

Educational Purpose: Explaining multivariate statistics and edify people is a difficult and strategic task. Instructors spend lots of time just to find suitable datasets for explaining some issue. For instance –

Model and methods testing: Imagine a situation where a researcher is collecting enormous amount of data which takes both time and money. Before going into extensive sampling, a pilot project is started and samples were collected using various techniques. Each of them are modelled with different estimation methods. The researcher would like to compare the sampling methods or estimation procedures that will the suitable for the final project. A simulated dataset based on the pilot project may help to identify the most appropriate sampling methods or estimation procedures.

Understanding and developing multivariate statistics: New estimation methods are being developed to address mordern and complicated situations. A new method or technique could be difficult to understand. For example, the envelope model Cook & Zhang (2015), a recent estimation technique based on maximum likelihood, attempts to find a response envelope (relevant subspace) that contains all the information that

the corresponding predictors can explain. Here, one can make use of `simrel-m` to simulate data with underlying informative response space. In addition, new methods such as CPLS based on PLS are steadily being developed. Understanding population latent structure enables assessment of such models.

2.1. Model Specification

A multi-response multivariate general linear model in equation-(2) 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.

2.1.1. An overview of `simrel`

`Simrel` is based on uni-response linear model as in equation~(1).

$$\begin{bmatrix} y \\ \mathbf{X} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_y \\ \boldsymbol{\mu}_X \end{bmatrix}, \begin{bmatrix} \sigma_y^2 & \boldsymbol{\sigma}_{X\boldsymbol{y}}{}^t \\ \boldsymbol{\sigma}_{X\boldsymbol{y}} & \boldsymbol{\Sigma}_{XX} \end{bmatrix} \right) \tag{1}$$

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

where $\mathbf{Y}$ is a response matrix with m response vector $y_1, y_2, \ldots y_m$, $\mathbf{X}$ is multivariate predictor matrix with p predictor variables and the random error term $\boldsymbol{\epsilon}$ is assumed to follow $N(\mathbf{0},\ \boldsymbol{\Sigma}_{Y|X})$. Equivalently,

$$\begin{bmatrix} \mathbf{Y} \\ \mathbf{X} \end{bmatrix} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = N\left(\begin{bmatrix} \boldsymbol{\mu}_Y \\ \boldsymbol{\mu}_X \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{YY} & \boldsymbol{\Sigma}_{XY}^t \\ \boldsymbol{\Sigma}_{XY} & \boldsymbol{\Sigma}_{XX} \end{bmatrix} \right) \tag{3}$$

Here,

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

According to the theory of Multivariate Normal Distribution, we can express different parameters interms of $\mathbf{X}$, $\mathbf{Y}$ and the covariance structure.

2.1.2. Model Parameterization

`Simrel-m` uses model parameterization which is based on the concept of relevant components Helland & 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 relevant space of response matrix 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 $\rho_1, \ldots, \rho_q$. Further, the collinearity between predictor variables can also be controlled by a factor γ which guides the decay pattern of eigenvalue of $\mathbf{X}$ matrix. Helland & Almøy (1994) showed that if the direction of large variablity (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 varibility, prediction becomes difficult.

Parameter Definition:

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

3. Stat Model

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

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

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

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*, .