simrel-m: A versatile tool for simulating multi-response linear model data

2017-04-04

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 Inge S Helland and Almøy ([1994](#ref-helland1994comparison)), which is equivalent to the envelope model R. Cook, Helland, and Su ([2013](#ref-cook2013envelopes)). It is a multi-response extension of simrel Sæbø, Almøy, and Helland ([2015](#ref-saebo2015simrel)), and as simrel the new approach is essentially based on random rotations of latent relevant components to obtain a predictor matrix , but in addition we introduce random rotations of latent components spanning a response space in order to obtain a multivariate response matrix . The properties of the linear relation between and 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, Almøy, and Helland ([2015](#ref-saebo2015simrel)).

# 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 presenting here is based on the principal of relevant space for prediction (Inge S Helland and Almøy [1994](#ref-helland1994comparison)) 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 are relevant for a latent subspace of the responses and collinearity structure of . This enables the possibility to construct data for evaluating estimation methods and methods developed for variable selection.

Among several publications on simulation (which publications), Ripley ([2009](#ref-ripley2009stochastic)) has exhaustively discussed the topic. In addition, many publications (which publications) are available on studies which has implemented simulated data in order to investigate new estimation methods and prediction strategy (see: R Dennis Cook and Zhang [2015](#ref-cook2015simultaneous); R. Cook, Helland, and Su [2013](#ref-cook2013envelopes); Inge S Helland et al. [2012](#ref-helland2012near)). 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, Almøy, and Helland ([2015](#ref-saebo2015simrel)) 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.

where, is a response vector with response variables with mean vector of and is vector of predictor variables with mean vector . Further,

|  |  |
| --- | --- |
|  | is variance-covariance matrix of |
|  | is variance-covariance matrix of variables |
|  | is matrix of covariance between and |

For model~(1), standard theory in multivariate statistics may be used to show that conditioned on corresponds to the linear model,

where, is a matrix of regression coefficient and is error term such that . The properties of the linear model in equation~(2) can be expressed in terms of covariance matrices from equation~(1).

Regression Coefficients

Coefficient of Determination

The diagonal elements of coefficient of determination matrix gives the amount of variation that has explained about in equation~(2).

Error variance

The conditional variance of given is,

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

Let us define a transformation of and as, and . Here, and are rotation matrices which rotates and giving and respectively. The model in equation~(1) can be expressed with these transformed variables as,

In addition, a linear model relating and can be written as,

where, is regression coefficient for the transformed model and . Further, if both and are orthonormal matrix such that and , the inverse transformation can be defined as,

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

Regression Coefficients

Error Variance

Further, the noise variance of transformed model~(4) is,

Population Coefficient of Determination

The population coefficient of determination for model~(4) is,

From eigenvalue decomposition principle, if and then and can be interpreted as principle components of and respectively. Here, and are diagonal matrix of eigenvalues corresponding to and respectively.

# Relevant Components

Let us consider a single response linear model with predictors.

where, and and are random and independent. Following the principle of relevant space and irrelevant space which are discussed extensively in Inge S Helland and Almøy ([1994](#ref-helland1994comparison)), Inge S. Helland ([2000](#ref-Helland_2000)), Inge S Helland et al. ([2012](#ref-helland2012near)), R. Cook, Helland, and Su ([2013](#ref-cook2013envelopes)), Sæbø, Almøy, and Helland ([2015](#ref-saebo2015simrel)) and Inge S. Helland et al. ([2017](#ref-helland2017)), we can assume that there exists a subspace of the full predictor space which is relevant for . An orthogonal space to this space does not contain any information about and is considered as irrelevant. Here, the relevant subspace of is spanned by a subset of eigenvectors of covariance matrix of , i.e. .

This concept can be extended to response so that the subspace of is relevant for a subspace of . This corresponds to the concept of simultaneous envelope (R. Dennis Cook and Zhang [2014](#ref-Cook_2014)) where relevant (material) and irrelevant (immaterial) space were discussed for both response and predictors.

# Model Parameterization

In order to construct a covariance matrix of and for model in equation~(3), we need to identify unknowns. For the purpose of this simulation, we implement some assumption to re-parameterize and simplify the model parameters. This enables us to construct diverse nature of model from few key parameters.

**Parameterization of**

If we let the rotation matrix be equal to the eigenvectors of , then is the set of principle components of . In that case is a diagonal matrix with eigenvalues . Further, we adopt the following parametric representation of these eigenvalues,

Here as increases, the decline of eigenvalues becomes steeper and hence a single parameter can be used for .

**Parameterization of**

Here, we assume that 's are independent and thus their covariance matrix is considered to be Identity .

**Parameterization of**

After parameterization of and , we are left with number of unknowns corresponding to . The elements in this covariance matrix depends on position of x-component that are relevant for . In order to re-parameterize this covariance matrix, it is necessary to discuss about the position of relevant components in details.

## Position of relevant components

Let only components are relevant for , components are relevant for and so on. Let the position of these components are given by the set respectively. Further, the covariance between and is non-zero only if is relevant for . If be the covariance between and then if where and and otherwise.

In addition, the corresponding regression coefficient for is,

where, for , is a vector with 1's and 0's such that if the position of relevant components for is in set and 0 otherwise.

The position of relevant components have heavy impact on prediction. Inge S Helland and Almøy ([1994](#ref-helland1994comparison)) have shown that if relevant components have large variance, prediction of from is relatively easy and if the variance of relevant components is small, the prediction becomes difficult given that coefficient of determination and other model parameters held constant. For example, if first and second components of are relevant for and fifth and sixth componets are relevant for , it is relatively easy to predict than . Since, the first and second principle components have larger variance than fifth and sixth components.

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

* The covariance matrix must be positive definite
* The covariance must satisfy user defined coefficient of determination

We have the relation,

Applying our assumption for simulation, and , we obtain,

Furthermore, we assume that there are no overlapping relevant components for any two , i.e, or for . The additional unknown parameters in diagonal should agree with user specified coefficient of determination for . i.e, is,

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

For some user defined , determined as follows,

1. Sample values from uniform distribution distribution. Let them be, .
2. Define,
3. for and

## Data Simulation

After the construction of covariance matrix,

observations are sampled from standard normal distribution of considering their mean to be zero, i.e. and . Let us define , such that . Since is positive definite, obtained from its Cholesky decomposition can serve as one of its square root and the matrix is sampled from standard normal distribution so that its covariance matrix . In addition the covariance matrix of is which satisfies all user defined population properties.

Here the first columns of will serve as and remaining columns will serve as . Further, each row of will be a vector sampled independently from joint normal distribution of . Finally, these simulated matrices and are orthogonally rotated in order to obtain and respectively. Following section discuss about these rotation matrices in details.

## Rotation of predictor space

In order to generate predictor variables and response variables , we construct matrices and and use them for orthogonal rotation of their respective principle components and . This defines a new basis for the same space as is spanned by the principle components. In principle, there are many possible candidate of rotation matrix and . Among them, the eigenvector matrix of and can be an option. However, in this reverse engineering both rotation matrix and the covariance matrices and are unknown. So, we are free to choose any and that satisfied the proportion of a real valued rotation matrix, i.e and so that and are orthonormal. Further there determinant becomes .

Among several methods (Anderson, Olkin, and Underhill [1987](#ref-anderson1987generation); Heiberger [1978](#ref-heiberger1978algorithm)) of generating random orthogonal matrix, in this paper we are using the matrix obtained from QR-decomposition of a matrix filled with standard normal variates. The rotation here can be a) restricted and b) unrestricted. The later one rotates all principle components together and makes all predictor variables somewhat relevant for all response variables. However, the former one performs a block-wise rotation so that it rotates certain selected principle components together. This gives control for specifying certain predictors as relevant for chosen response. This also lets us to simulate irrelevant predictors which can be detected during variables selection procedure.

## Rotation of response space

Simrel-M has considered an exclusive relevant predictor space for each response components, i.e. a set of predictor variables only influence one response component. However, it allows user to simulate more response variable than response components. In this case, noise are added during the orthogonal rotation of response components. For example, if user wants to simulation 5 response variation from 3 response components. Two standard normal vectors are combined with response components and rotated simultaneously. The rotation can be both restricted and unrestricted as discussed in previous section. The restricted rotation is carried out combining response vectors along with noise vector in a block-wise manner according to the users choice. Illustration in fig-...

Suppose, in our previous example, if response components are combined as -- , and . Here, any predictor variable is only relevant for and while and are noise. The resulting response variables are where, the first and fourth response variable spans the same space as by the first response components and noise component and so on. Thus, the predictors and predictor space relevant for response component is also relevant for response and .

# References

Anderson, Theodore W, Ingram Olkin, and Les G Underhill. 1987. “Generation of Random Orthogonal Matrices.” *SIAM Journal on Scientific and Statistical Computing* 8 (4). SIAM: 625–29.

Cook, R Dennis, and Xin Zhang. 2015. “Simultaneous Envelopes for Multivariate Linear Regression.” *Technometrics* 57 (1). Taylor & Francis: 11–25.

Cook, R. Dennis, and Xin Zhang. 2014. “Simultaneous Envelopes for Multivariate Linear Regression.” *Technometrics* 57 (1). Informa UK Limited: 11–25. doi:[10.1080/00401706.2013.872700](https://doi.org/10.1080/00401706.2013.872700).

Cook, RD, IS Helland, and Z Su. 2013. “Envelopes and Partial Least Squares Regression.” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 75 (5). Wiley Online Library: 851–77.

Heiberger, Richard M. 1978. “Algorithm as 127: Generation of Random Orthogonal Matrices.” *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 27 (2). JSTOR: 199–206.

Helland, Inge S, and Trygve Almøy. 1994. “Comparison of Prediction Methods When Only a Few Components Are Relevant.” *Journal of the American Statistical Association* 89 (426). Taylor &amp; Francis Group: 583–91.

Helland, Inge S, Solve Saebø, Ha Tjelmeland, and others. 2012. “Near Optimal Prediction from Relevant Components.” *Scandinavian Journal of Statistics* 39 (4). Wiley Online Library: 695–713.

Helland, Inge S. 2000. “Model Reduction for Prediction in Regression Models.” *Scandinavian Journal of Statistics* 27 (1). Wiley-Blackwell: 1–20. doi:[10.1111/1467-9469.00174](https://doi.org/10.1111/1467-9469.00174).

Helland, Inge S., S. Sæbø, T. Almøy, and R. Rimal. 2017. “Model and Estimators for Partial Least Squares.”

Ripley, Brian D. 2009. *Stochastic Simulation*. Vol. 316. John Wiley & Sons.

Sæbø, Solve, Trygve Almøy, and Inge S Helland. 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*. Elsevier.