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

2017-05-24

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](#ref-helland1994comparison)), which is equivalent to the envelope model 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 (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. D. Cook and Zhang [2015b](#ref-cook2015simultaneous); Cook, Helland, and Su [2013](#ref-cook2013envelopes); 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 each .

Conditional 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,

Coefficient of Determination

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

From eigenvalue decomposition principal, if and then and can be interpreted as principal components of and respectively. In this paper, these principal components will be termed as *predictor components* and *response components* respectively. Here, and are diagonal matrices of eigenvalues of and respectively.

# Relevant Components

Let us consider a single response linear model with predictors.

where, and are random and independent. Following the principal of relevant space and irrelevant space which are discussed extensively in Helland and Almøy ([1994](#ref-helland1994comparison)), Helland ([2000](#ref-Helland_2000)), Helland et al. ([2012](#ref-helland2012near)), Cook, Helland, and Su ([2013](#ref-cook2013envelopes)), Sæbø, Almøy, and Helland ([2015](#ref-saebo2015simrel)) and 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 responses so that the subspace of is relevant for a subspace of . This corresponds to the concept of simultaneous envelope (R. D. Cook and Zhang [2015b](#ref-cook2015simultaneous)) 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 and for model in equation~(3), we need to identify unknown parametersSome Motivation for simplification of parameters(See comment from Solve). 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 few key parameters.

**Parameterization of**

If we consider the rotation matrix corresponds to the eigenvectors of , then becomes the set of principal 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, hence the parameter controls the level of multicollinearity in .

**Parameterization of**

Here, we assume that 's are independent unconditionally equally mutinormal distributed with variance 1, hence .

**Parameterization of**

After parameterization of and , we are left with number of unknowns corresponding to . Some of the elements of may be equal to zero, which implies that the given is irrelevant for the given variable . The non-zero elements define which of the are relevant for the . We typically refer to the indices of these 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 components be relevant for , components be relevant for and so on. Let the positions of these components be given by the index sets respectively. Further, the covariance between and is non-zero only if is relevant for . If is the covariance between and then if where and and otherwise.

In addition, the true regression coefficients for (equation~(4)) is given by:

The position of relevant components have heavy impact on prediction. Helland and Almøy ([1994](#ref-helland1994comparison)) have shown that if relevant components have large eigenvalues (variances), which here implies small index values in , prediction of from 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, and , are relevant for and fifth and sixth components, and , are relevant for , it is relatively easier to predict than , 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 , and must be positive definite
* The covariance must satisfy user defined coefficient of determination

We have the relation,

Applying our above given assumptions that, 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 of 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 , is determined as follows,

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

### Data Simulation

From the above given parameterizations and the user defined choices of model parameters, a fully defined and known covariance matrix of is given. For the simulation of a single observation of let us define such that . Here is obtained from Choleskey decomposition and serves as one of the square root of positive definite matrix and is simulated from standard normal distribution and has covariance .

Similarly, in order to simulate observations, we define such that . 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 make comments on predictor space, let us consider an example where a regression model with predictors and responses . Let's assume that only three principal components and are needed to describe all four response variables. Further, let the index sets and define the position of the principal components of that are relevant for and respectively. Let , and be the orthogonal spaces spanned by each set of principal components. These spaces together span which is the minimum relevant space and equivalent to the x-envelope as discussed by Cook, Helland, and Su ([2013](#ref-cook2013envelopes)).

Moreover, let and be the number of predictor variables we want to be relevant for and respectively. Then predictors may be obtained by rotating the principal components in along with one more irrelevant principal component. Similarly, predictors, relevant for , can be obtained by rotating principal components in along with one more irrelevant component and predictors, relevant for , can be obtained by rotating principal components in without any additional irrelevant component. Let the space spanned by the and number of predictors be , and . Together they span a space . This space is bigger than . Here, is orthogonal to and is orthogonal to . Generally speaking, here we are splitting complete variable space into two orthogonal space -- relevant for and irrelevant for .

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 and . Here we can see that and (first and second principal components of ) have non-zero covariance with (first latent component of response ). In the similar manner other non-zero covariances are self-explanatory.

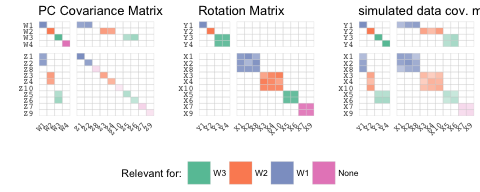


Figure 1 Simulation of predictor and response variables after orthogonal transformation of principal components by a rotation matrix

In order to simulate predictor variables , we construct matrix which then is used for orthogonal rotation of principal components . 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 can be a candidate. However, in this reverse engineering both rotation matrices and along with the covariance matrices are unknown. So, we are free to choose any that satisfied the properties of a real valued rotation matrix, i.e so that is orthonormal and its determinant becomes . Here the rotation matrix should be block diagonal as in figure~1 (middle) in order to rotate spaces separately. Figure~2 (left) shows the simulated principal components that we are following in our example where we can see that the principal component and relevant for is getting rotated together with an irrelevant component . The resultant predictors (Figure~2, right) and will also be relevant for . In the figure, we can see that principal components and are not relevant for any responses before rotation however predictors becomes relevant after rotation keeping and still irrelevant.

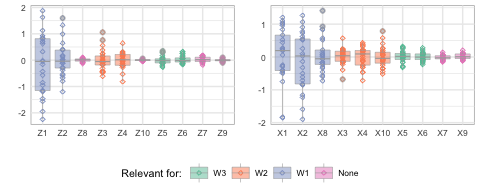


Figure 2 Simulated Data before (left) and after (right) rotation

Among several methods (Anderson, Olkin, and Underhill [1987](#ref-anderson1987generation); Heiberger [1978](#ref-heiberger1978algorithm)) for generating random orthogonal matrix, in this paper we are using orthogonal matrix 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 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 and 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 and . During the rotation of covariance matrix , the response space is also rotated separately along with the predictor space. Figure~1 shows that the informative response component is rotated together with uninformative response component so that the predictors which were relevant for will be relevant for response variables and . Similarly, response components and are rotated separately so that predictors relevant for and will also be relevant for and 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.

1. Ordinary Least Squares (OLS),
2. Principal Component Regression (PCR),
3. Partial Least Squares predicting individual response variable separately (PLS1) and
4. Partial Least Squares predicting all response variables together (PLS).

We have also considered four relatively new estimation methods

1. Canonically Powered Partial Least Squares regression (CPPLS) (Indahl, Liland, and Næs [2009](#ref-indahl2009canonical)),
2. Canonical Partial Least Squares regression (CPLS) (Indahl, Liland, and Næs [2009](#ref-indahl2009canonical)),
3. Envelope estimation in predictor space (xenv) (Cook, Li, and Chiaromonte [2010](#ref-cook2010envelope)),
4. Envelope estimation in response space (yenv) (R. D. Cook and Zhang [2015a](#ref-cook2015foundations)) and
5. Simultaneous estimation of x- and y-envelope (senv) (R. D. Cook and Zhang [2015b](#ref-cook2015simultaneous))

From the possible combinations of two levels of coefficient of determination 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 and 16 predictor variables are simulated using the method discussed in this paper. It is also assumed that three principle components of response variables ( and ) completely describes the variation present in five response variables (). 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 and . Further, predictor components and are relevant for response component , predictor components and are relevant for response component and predictor component is relevant for response component . In addition, following the discussion about [rotation of response space](#rotation-of-response-space), is rotated together with and is rotated together with .

Table 1 Parameter setting of simulated data for model comparison

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Design1 | Design2 | Design3 | Design4 |
| **Decay of eigenvalue** | 0.2 | 0.8 | 0.2 | 0.8 |
| **Coef. of Determination** | 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,

where, is an estimate of true regression coefficient and is the true covariance structure of predictor variable obtained from simrel-m. Also, is the true minimum error of the model. Here 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 Forbenius norm defined as (Golub and Van Loan [2012](#ref-golub2012matrix)),

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, prediction error)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Design: 1 | Design: 2 | Design: 3 | Design: 4 |
| *CPLS* | (3, 1.65) | (3, 1.64) | ***(3, 1.88)*** | (3, 1.86) |
| *CPPLS* | (3, 1.64) | (3, 1.63) | (3, 1.89) | ***(3, 1.86)*** |
| *OLS* | (1, 1.88) | (1, 1.87) | (1, 2.12) | (1, 2.09) |
| *PCR* | (7, 1.69) | (6, 1.66) | (6, 1.88) | (6, 1.87) |
| *PLS* | (4, 1.68) | (6, 1.66) | (3, 1.88) | (6, 1.88) |
| *PLS1* | (1, 1.71) | (5, 1.67) | (1, 1.91) | (5, 1.88) |
| *Senv* | ***(4, 1.63)*** | ***(5, 1.62)*** | (3, 1.99) | (5, 1.96) |
| *Xenv* | (5, 1.67) | (6, 1.67) | (5, 1.89) | (5, 1.9) |
| *Yenv* | (3, 1.65) | (3, 1.65) | (3, 1.96) | (3, 1.93) |

Table~2 shows that simulteneous envelope has prediction error of 1.63 and 1.62 in design 1 and design 2 respectively which is smaller than other methods. However the model was not able to show the same performance in design 3 and design 4. Cannonical PLS and Cannonically Powered PLS has out performed other methods in these designs. Here, the difference between CPLS and CPPLS is minimal. These methods has also shown a fair performance in the first two designs with only three components. 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.

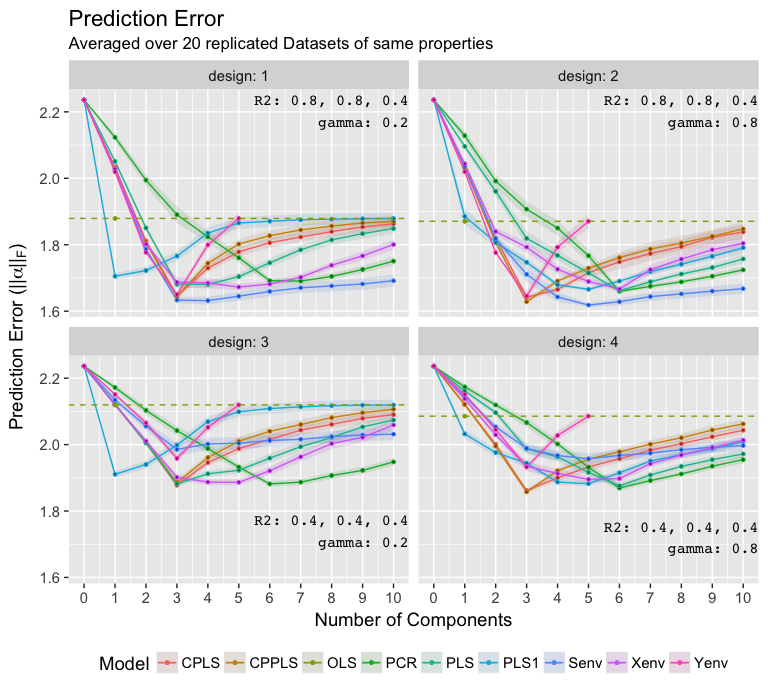


Figure 3 Minimum of Average Prediction Error

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, the reduced performance of simulteneous envelope going from design with to design with has arised question such as -- Does the performance of the method depends on or it is a random situation? Since, this paper is intended for a demonstration of how simrel-m can be used in scientific study, a more elaborative study is needed in order to answer such question which simrel-m can help as an useful instrument.

# Web Interface

In order to give an 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.

**Points to include in this section:**

1. User are able to select seed for simulation so that they can get exact same data with each individual seed.
2. The app also allows user simulate univariate (Sæbø, Almøy, and Helland [2015](#ref-saebo2015simrel)), bivariate (not yet published) and multivariate linear model data.
3. 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 considered [implementation](#implementation) section 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)).

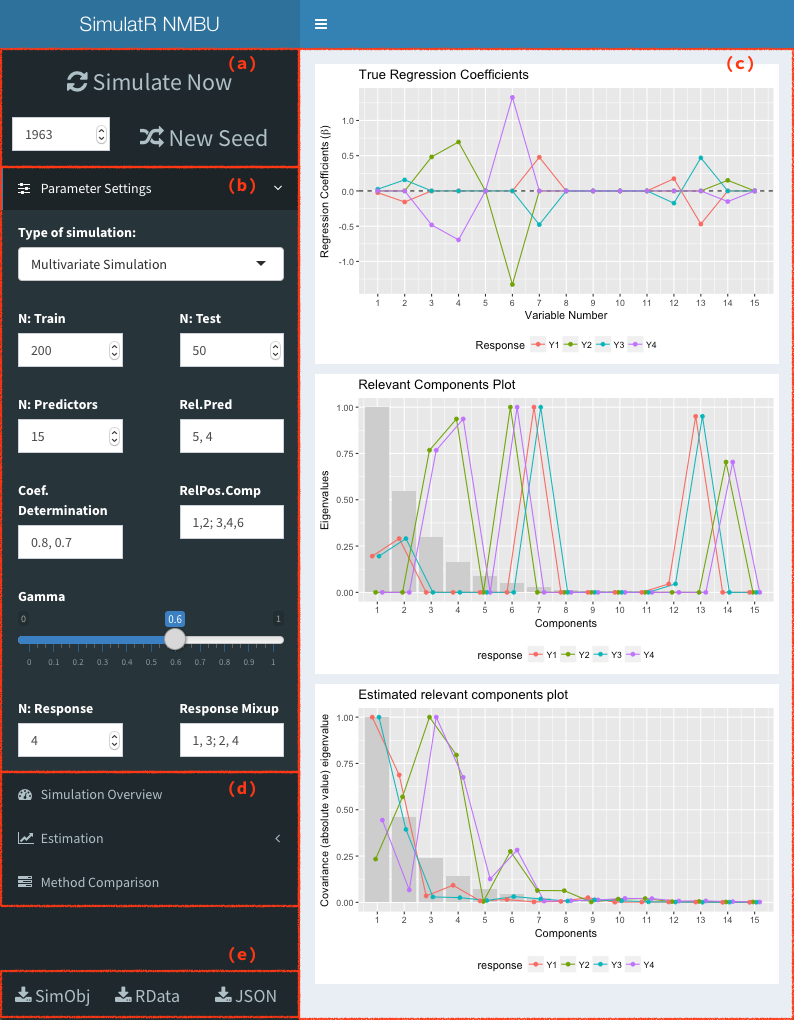


Figure 4 Application interface of simulatr. (a) Seed and simulation button (b) Parameter control panel (c) Properties of simulated data (d) Additional analysis (e) Download option of simulated data

# 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. 2015a. “Foundations for Envelope Models and Methods.” *Journal of the American Statistical Association* 110 (510). Taylor & Francis: 599–611.

———. 2015b. “Simultaneous Envelopes for Multivariate Linear Regression.” *Technometrics* 57 (1). Taylor & Francis: 11–25.

Cook, R Dennis, Bing Li, and Francesca Chiaromonte. 2010. “Envelope Models for Parsimonious and Efficient Multivariate Linear Regression.” *Statistica Sinica*. JSTOR, 927–60.

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.

Golub, Gene H, and Charles F Van Loan. 2012. *Matrix Computations*. Vol. 3. JHU Press.

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.”

Indahl, Ulf G, Kristian Hovde Liland, and Tormod Næs. 2009. “Canonical Partial Least Squares–a Unified Pls Approach to Classification and Regression Problems.” *Journal of Chemometrics* 23 (9). Wiley Online Library: 495–504.

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.