Review on Simrel Paper

Raju Rimal, Trygve Almøy and Solve Sæbø

First of all we are grateful to receive the constructive comments on our manuscript. Your comments helped us to improve our manuscript significantly. We have documented all the changes, as given below, based on your comments.

Changes in Assumptions

Your concern about the assumptions we have made, pointed us to a possibility of removing the assumption of covariance matrix of $\mathbf{w}$ to be identity. This triggered some changes in the calculation of the coefficient of determination. The changes concerning these issues are as follows,

Page 3:

Last Paragraph (Coefficient of Determination) :

~~**Coefficient of Determination** The diagonal elements of the coefficient-of-determination matrix $\boldsymbol{\rho}^2_y(m \times m)$ give the amount of variation in each response variable that is explained by $\mathbf{x}$.~~

$$\boldsymbol{\rho}^2_y = \boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}^{-1}_{xx}\boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}^{-1}_{yy}$$

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:

$$\left(\rho^2_y\right)_{jj'} = \frac{\sigma^t_{xy_j}\Sigma^{-1}_{xx}\sigma_{xy_{j'}}}{\sqrt{\sigma^2_{y_j}\sigma^2_{y_{j'}}}} \forall j, j' = 1 \dots m$$

where, σ_{xy_j}, $\sigma_{xy_{j'}}$ are covariances between $\mathbf{x}$ and y_j, $y_{j'}$ respectively. Also, $\sigma^2_{y_j}$ and $\sigma^2_{y_{j'}}$ 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 present in $\mathbf{y}$. The diagonal elements of this matrix is the proportion of variation in a response $y_j, j = 1, \dots m$ explained by the predictors.

Page 5:

~~**Coefficient of Determination** The coefficient-of-determination matrix for (4) is~~

$$\begin{aligned}\boldsymbol{\rho}_w^2 &= \boldsymbol{\Sigma}_{wz}\boldsymbol{\Sigma}_{zz}^{-1}\boldsymbol{\Sigma}_{zw}\boldsymbol{\Sigma}_{ww}^{-1} \\ &= \mathbf{Q}\boldsymbol{\Sigma}_{yx}\mathbf{R}^t\left(\mathbf{R}\boldsymbol{\Sigma}_{xx}\mathbf{R}^t\right)^{-1}\mathbf{R}\boldsymbol{\Sigma}_{xy}\mathbf{Q}^t\left(\mathbf{Q}\boldsymbol{\Sigma}_{yy}^{-1}\mathbf{Q}^t\right) \\ &= \mathbf{Q}\left[\boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}\boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yy}^{-1}\right]\mathbf{Q} = \mathbf{Q}\boldsymbol{\rho}_y^2\mathbf{Q}^t\end{aligned}$$

Coefficient of Determination The coefficient-of-determination matrix corresponding to **w** can be written as,

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

where, σ_{zw_j} and $\sigma_{zw_{j'}}$ are covariances of **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 xy-space.

Page 6

Equation number has been added to the expression of generating lambda from gamma. Subscript changed from j to i.

Added to equation of γ :

We can write $\Sigma_{zz} = \Lambda = \text{diag}(\lambda_1, \ldots, \lambda_p)$.

Parameterization of Σ_{ww} :

~~Here we assume that **w**'s are independent and has mutinormal distribution with variances 1, hence $\boldsymbol{\Sigma}_{ww} = \mathbf{I}_m$.~~ In similar manner, a parametric representation of eigenvalues corresponding to Σ_{ww} is adopted as,

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

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 $\Sigma_{ww} = \text{diag}(\kappa_1, \ldots, \kappa_m)$.

Page 8

Paragraph 2 :

$$\boldsymbol{\rho}_w^2 = \boldsymbol{\Sigma}_{zw}^t \boldsymbol{\Sigma}_{zz}^{-1} \boldsymbol{\Sigma}_{zw} \boldsymbol{\Sigma}_{ww}^{-1}$$

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

Paragraph 3 :

$$\mathrm{h}\Sigma_{ww} = \mathbf{I}_m$$

$$\Sigma_{ww} = \mathrm{diag}(\kappa_1, \dots, \kappa_m)$$

Paragraph 4 $\rho_w^2 = \Sigma_{zw}^t \Lambda^{-1} \Sigma_{zw} \Sigma_{ww}^{-1}$

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

Paragraph 6 :

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

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

Paragraph 8 :

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

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

Paragraph 12 (last display equation) :

$\sigma_{ij} = \mathrm{Sign}(\mathcal{S}_i) \sqrt{\frac{\rho_{wj}^2 |\mathcal{S}_i|}{\sum_{k \in \mathcal{P}_j} |\mathcal{S}_k|} \lambda_i}$

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

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

Before starting Data Simulation :

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 condition.

In addition, your suggestions about the papers are very much relevant in our context. It has helped us to understand alternative algorithmic approaches to simulating data based on correlation structures. Changes we have made are in,

Page 2:

Line 2 last paragraph :

In particular, methods based on covariance structure has been discussed by (Arteaga and Ferrer 2010, Arteaga and Ferrer (2013), Camacho (2017)), following approaches to find simulated data satisfying the desired correlation structure.

Further, in the following we have added a comment on the second assumption of non-overlapping relevant components.

Page 12 (Rotation of response space)

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.

Clarity on parameters and plot description

Your inquiry about the set of inputs the user needs to specify is reasonable. Since the first part (until implementation section) discusses mainly the theoretical background, we have added the following clarifying R-code in section 5.

Page 17

After first paragraph :

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"
)
```

In addition, we have mentioned the code repository at github where users can find rich documentation in detail.

Page 3:

After first Paragraph :

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 sections 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.

Changes in web interface

Page 16

Last Paragraph :

~~(uses `simrel` package in CRAN), (not yet available in CRAN), (`simrel-m`).~~

Second last line :

Users can also download simulated data in JSON and CSV format.

Figure-5 Caption: :

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.

In order to make the plots in screenshot more clear, we have changed the second and third paragraph of page 17 completely as follows,

With the parameters for simulation in the screenshot (Figure-5) 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 5) contains three plots – a) true regression coefficients b) relevant components and c) estimated relevant components. In the first plot (Figure-5)(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 5)(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-5)(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 (5(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.

Application of the simulation system

Since the paper is focused on describing the simulation method itself, we believe that a detailed case study is beyond the scope of this paper. However, we are currently working on a follow-up paper which will answer this concern. In addition, since we have illustrated the simulation approach for comparing a few estimation methods, among which one is based on maximum likelihood, we have not included an example with wide matrices here. But the simulation tool is capable of simulating such data and will be used in the follow up paper we mentioned earlier.

Conclusion added

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.

Additional Changes

Abstract

~~The method is based on the concept of relevant components [13], and is equivalent to the newly developed envelope model by Dennis Cook.~~

The method is based on the concept of relevant components, and is equivalent to the newly developed envelope model.

References

Arteaga, Francisco, and Alberto Ferrer. 2010. "How to Simulate Normal Data Sets with the Desired Correlation Structure." *Chemometrics and Intelligent Laboratory Systems* 101 (1).

Elsevier: 38–42.

———. 2013. "Building Covariance Matrices with the Desired Structure." *Chemometrics and Intelligent Laboratory Systems* 127. Elsevier: 80–88.

Camacho, José. 2017. "On the Generation of Random Multivariate Data." *Chemometrics and Intelligent Laboratory Systems* 160. Elsevier: 40–51.