Chemometrics and Intelligent Laboratory Systems 146 (2015) 128–135

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemolab

Software Description

simrel — A versatile tool for linear model data simulation based on the concept of a relevant subspace and relevant predictors

Solve Sæbø [a,*], Trygve Almøy [a], Inge S. Helland [b]

[a] Dep. Chemistry, Biotechnology and Food Science, Norwegian University of Life Sciences, P.O. Box 5003, NO-1432 Ås, Norway
[b] Department of Mathematics, University of Oslo, P.O. Box 1053, NO-0316 Oslo, Norway

ARTICLE INFO

Article history:
Received 27 March 2015
Received in revised form 7 May 2015
Accepted 12 May 2015
Available online 19 May 2015

Keywords:
Data simulation
Linear model
R-package
Experimental design

ABSTRACT

In the field of chemometrics and other areas of data analysis the development of new methods for statistical inference and prediction is the focus of many studies. The requirement to document the properties of new methods is inevitable, and often simulated data are used for this purpose. However, when it comes to simulating data there are few standard approaches. In this paper we propose a very transparent and versatile method for simulating response and predictor data from a multiple linear regression model which hopefully may serve as a standard tool simulating linear model data. The approach uses the principle of a relevant subspace for prediction, which is known both from Partial Least Squares and envelope models, and is essentially based on a re-parametrization of the random *x* regression model. The approach also allows for defining a subset of relevant observable predictor variables spanning the relevant latent subspace, which is handy for exploring methods for variable selection. The data properties are defined by a small set of input-parameters defined by the analyst. The versatile approach can be used to simulate a great variety of data with varying properties in order to compare statistical methods. The method has been implemented in an *R*-package and its use is illustrated by examples.

© 2015 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

1. Introduction

In the process of developing new statistical methods for multiple linear regression, prediction and variable selection it is convenient to have a simple approach and accessible software for data simulation where the properties of the data can be controlled by a few parameters. Then it is easy to test the methodology on data with known properties, such as the number of predictor variables, the number of observations, the number of truly relevant predictor variables, the information content among other things, and even test out what predictor to use. Here we present a new *R*-package, *simrel* [1], making this readily available for all developers of statistical methodology. The simulations are based on a multivariate normal distribution giving rise to a best linear predictor for a response variable *y* given a set of predictor variables comprising a predictor matrix ***X***. The user defines the data properties by a set of input parameters, and the output is training data, test data (optional) and the vector of true regression coefficients.

There is a vast literature on simulation. The topic is among others exhaustively discussed in [2]. Also the performance of more advanced prediction methods is investigated by aims of simulations. Among earlier papers we can mention [3] on ridge regression, [4] on shrinkage estimators, [5] and [6] on subset selection methods, [7] comparing Ridge regression and PLS, and [8] conducted a study of the performance of PLS and PCR using the same concept of relevant components as is used in this paper. Although the literature on data simulation for method comparisons is vast, a systematic tool for doing such comparisons has in our knowledge not been available up until now.

The model parametrization is based on the concept of relevant components [9–11] where it is assumed that there exists a *y*-relevant subspace of the full variable space which is spanned by a subset of the eigenvectors of the covariance matrix of the *x*-variables. All relevant information for the prediction of *y* is contained in this sub-space and consequently, the orthogonal space is irrelevant. Here we also assume that the relevant sub-space is spanned by a subset of the predictor variables. In this way we may construct a set of relevant predictor variables with truly non-zero regression coefficients, which for instance should be recognized by variable selection methods. The user can control the signal to noise content in the predictor data by setting the true coefficient of determination, ρ^2, for the data. Other input parameters are the degree of collinearity in the predictor matrix (by controlling the decline in the eigenvalues of the *x*-covariance matrix) and the position of the relevant components (in the list of ordered eigenvectors).

[11] showed that prediction is relatively easy if the directions in the predictor space with large variability (large eigenvalues) are also the most relevant for prediction (given that ρ^2 is not very small), whereas

* Corresponding author.
E-mail addresses: solve.sabo@nmbu.no (S. Sæbø), trygve.almoy@nmbu.no (T. Almøy), ingeh@math.uio.no (I.S. Helland).

http://dx.doi.org/10.1016/j.chemolab.2015.05.012
0169-7439/© 2015 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

the opposite is true if the y-relevant information is associated with directions in the *x*-space with low variability (small eigenvalues).

The package also provides a tool for designing computer experiments based on the Multilevel Binary Replacement (MBR) design approach of [12]. The MBR-design provides a way of setting up a fractional design for large scale computer experiments in order to explore the effects of potentially many multi-level design factors. The design factors to be specified by the user, are:

- p: The number of predictors.
- n: The number of observations.
- q: The number of relevant predictors.
- m: The number of relevant components.
- $\mathcal{P}$: The set of indices for the relevant components.
- ρ^2: The population coefficient of determination.
- γ: A parameter defining the degree of collinearity in x.

The meaning of most of these design factors should be clear, but some need a closer explanation. We base our discussion on the random x regression model given by Eq. (2) in Section 2.1. We assume that there are p x-variables in total, and that q is the number of these x-variables that have non-trivial coefficient $\beta_j \neq 0$. The number m is related to the expansion of the regression vector $\boldsymbol{\beta}$ in terms of eigenvectors, $\boldsymbol{e}_j$ (for $j = 1,\ldots,p$), of the x-covariance matrix $\boldsymbol{\Sigma}_{xx}$:

$$\boldsymbol{\beta} = \sum_{j=1}^{p} \eta_j \boldsymbol{e}_j. \tag{1}$$

The number of terms in Eq. (1) may be reduced by two mechanisms: 1) Some of the η_j's may be 0; and 2) there are coinciding eigenvalues of $\boldsymbol{\Sigma}_{xx}$. Then it is enough to have one eigenvector for each space (stratum) corresponding to one value of the eigenvalue in the sum (Eq. (1)). Let m be the number of terms in Eq. (1) when this number is reduced as much as possible.

By this mechanism there are m eigenvectors/components that are relevant, and the positions of the relevant components is contained in the set of indices $\mathcal{P}$. Here it is assumed that the order of the components is defined by the declining set of eigenvalues $\boldsymbol{\Sigma}_{xx}$ such that $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p > 0$, hence, the eigenvalues of the m relevant components are given by $\lambda_{\mathcal{P}_1} > \lambda_{\mathcal{P}_2} > \ldots > \lambda_{\mathcal{P}_m}$. In the following, and in the *simrel*-package we will also refer to the set of positions for the relevant components as *relpos*. If for example $m = 3$ and $\mathcal{P} = \{2, 3, 10\}$, then the eigenvectors corresponding to λ_2, λ_3 and λ_{10} are relevant for the prediction of y. In the *R*-package *simrel* we correspondingly define the vector $relpos = c(2,3,10)$.

In *simrel* we have made the simplifying assumption that all p eigenvalues of $\boldsymbol{\Sigma}_{xx}$ are different and that they are decreasing exponentially as $e^{-\gamma \cdot (j-1)}$ for $j = 1,\ldots,p$ and some positive constant γ. When γ is large, we have very collinear x-variables.

In these specifications we may very well have $p > n$, but we must have $m < n$. Otherwise, the only restriction is that $m \leq q \leq p$, and that $\mathcal{P}$ is contained in a set $\mathcal{P}_x$ of indices of the relevant x-variables. For example, if the relevant components are defined by the set $\mathcal{P} = \{2, 3, 10\}$, then all sets of length q of the type $\mathcal{P}_x = \{2, 3, 10, \ldots\}$ of indices of relevant predictors are allowed, where "…" denotes any other set of variable(s) between 1 and p. In other words, this means that both the m relevant eigenvectors and the q relevant predictor variables are basis for the relevant space of dimension m.

In this paper and in [11] it is assumed to be known which components are relevant. This is of course rarely the case, but in the comparison of prediction methods it can serve to illustrate interesting cases. In the PLS-model of [9] and in the corresponding envelope model [13] only the dimension m of the relevant space is assumed known.

The purpose of data simulation is to investigate some measure of performance of one or several proposed methods and how this depends on parameters as those given above. Typical measures of performance are prediction error and success rate in variable selection. It goes without saying that if the performance is to be investigated under many settings of the above given input parameters, the computational burden will be quite large even for just a couple of levels of the design parameters. If two levels of each of the seven parameters are chosen, a single replicate of the design would require $2^7 = 128$ data sets to be analyzed. Typically several simulations are also required to better estimate the expected performance under each parameter setting. A more extensive, but reasonable, investigation could require four levels of each parameter. The number of runs in a single replicate would then be $4^7 = 16{,}384$. Obviously this is beyond what is convenient even on today's powerful computers. The MBR-design method provides an elegant way of choosing a fractional design for multi-factor and multi-level experiments which reduces the total number of runs dramatically, but still provides the possibility to estimate main effects and low-degree interaction effects of the design parameters on the performance measure used. The *simrel* package can provide both the MBR-design as well as a list of simulated data sets based on the chosen design.

The *simrel* package is freely available from CRAN (http://cran.r-project.org).

2. Statistical model

2.1. Model definition

The simulation model is the general linear model:

$$y = \mu_y + \boldsymbol{\beta}^t(\boldsymbol{x} - \boldsymbol{\mu}_x) + \epsilon \tag{2}$$

where y is the response variable, $\boldsymbol{x}$ is a vector of p predictor variables, $\boldsymbol{\beta}$ is the vector of regression coefficients and ϵ is the random error term assumed to be distributed as $N(0, \sigma^2)$. We here adopt a random regression framework as point of departure where $\boldsymbol{x} \sim N(\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_{xx})$ independent of ϵ. This is equivalent to

$$\begin{bmatrix} y \\ \boldsymbol{x} \end{bmatrix} \sim \mathcal{N}\left(\boldsymbol{\mu}_{yx}, \boldsymbol{\Sigma}_{yx}\right) = \mathcal{N}\left(\begin{bmatrix} \mu_y \\ \boldsymbol{\mu}_x \end{bmatrix}, \begin{bmatrix} \sigma_y^2 & \boldsymbol{\sigma}_{xy}^t \\ \boldsymbol{\sigma}_{xy} & \boldsymbol{\Sigma}_{xx} \end{bmatrix}\right) \tag{3}$$

where $\boldsymbol{\sigma}_{xy}$ is the vector of covariances between the predictors and y, and $\boldsymbol{\Sigma}_{xx}$ is the $(p \times p)$ covariance matrix of $\boldsymbol{x}$. According to the general theory on the multivariate normal distribution some of the properties of this model are:

- The noise variance and the minimum prediction error under expected quadratic loss is:

$$\sigma^2 = \sigma_y^2 - \boldsymbol{\sigma}_{xy}^t \boldsymbol{\Sigma}_{xx}^{-1} \boldsymbol{\sigma}_{xy}$$

- The true value of the regression coefficient vector is

$$\boldsymbol{\beta} = \boldsymbol{\Sigma}_{xx}^{-1} \boldsymbol{\sigma}_{xy}$$

- The population coefficient of determination is

$$\rho^2 = \boldsymbol{\sigma}_{xy}^t \boldsymbol{\Sigma}_{xx}^{-1} \boldsymbol{\sigma}_{xy} / \sigma_y^2 = 1 - \frac{\sigma^2}{\sigma_y^2}.$$

In order to simulate $(y, \boldsymbol{x})$ data from the model in Eq. (3) we will make use of the fact that any set of variables spanning the same p-dimensional predictor space as $\boldsymbol{x}$ will yield the same prediction of y

Download English Version:

https://daneshyari.com/en/article/7562970

Download Persian Version:

https://daneshyari.com/article/7562970

Daneshyari.com