SimSel - a new simulation feature selection method I

Martin Eklund and Silvelyn Zwanzig

Abstract

In pharmaceutical research there are data sets describing the interactions between proteins and molecules. The data sets include a huge number of independent variables (features) and the response variable is typically the binding strength. Thus, one of the most challenging problems is to find the features that have a real influence on the binding strength.

Here we present a feature selection method. The principle of the algorithm is to disturb each single feature by adding pseudo errors and to study the influence on the quality of the model fit. The main idea is that the change of unimportant features has no effect on the binding strength.

1 Introduction

Suppose a $n \times (p+1)$ data matrix $(\mathbf{y}, \mathbf{X})$, where the first column $\mathbf{y} = (y_i)$ is considered as response variable, $\mathbf{X} = (x_{ij})_{i=1..n,j=1..p}$ is a $n \times p$ matrix. The columns of $\mathbf{X}$ are related to different features denoted by $\mathbf{x}_{(j)}, j = 1, .., p$. The problem is to select the features which are really relevant for prediction of y.

Here we propose a procedure which indicate for every single feature the influence on the response.

The main idea arises from SIMEX methods in measurement errors models. There the measurement error is increased by adding pseudo errors and the influence of the pseudo errors is modelized in order to extrapolate backwards to the case with no measurement errors, compare [1]. Nowadays SIMEX is also applied to model selections procedures. There SIMEX is used for the choice of adaptive parameters, see [2].

Our method use only the simulation step, where pseudo errors are added. The extrapolation step is not done. Thus we introduce the name SimSel for simulation and selection. The goal of the method is to find out, whether a specific feature has some influence on the response. The selection part of the procedure is up to now not implemented.

The residual sum of squares is used as criterion for measuring the influence. It is not used as a criterion for fitting a model. It is enough to have an approximative model. If the underlying relationship is a nonlinear errors-in-variables model the procedure is still working. In this case the residual sum of squares corresponds to the naive method in an approximative quadratic model.

The main items of SimSel can be described as follows.

- The residual sum of squares is used to measure the influence of the features on the response.
- The x variable under consideration is disturbed by pseudo errors.
- The influence of the pseudo errors on the residual sum of squares is studied.
- If the residual sum of squares is unchanged, then this feature does not matter.

The report is organized as follows. First we present a linear and a quadratic algorithm and the related theoretical background. In Section 4 simulation results are summarized. Section 5 contains an application in pharmacy. The discussion of the simulations and of further research is given in Section 6. All proofs can be found in the Appendix. The beta version of the R-packages is available on the home page http://www.math.uu.se/ ˜zwanzig.

2 Linear SimSel

With out loss of generality we assume that the feature of interest is related to $\mathbf{x}_{(1)}$ the first column of $\mathbf{X}$.

The linear ordinary least squares method is defined by minimizing the sum of squares

$$\|\mathbf{y}-\mathbf{X}\beta\|^2=\sum_{i=1}^{n}\left(y_i-\sum_{j=1}^{p}x_{ij}\beta_j\right)^2.$$

Denote the residual sum of squares by

$$RSS=\min_{\beta\in\mathbb{R}^p}\|\mathbf{y}-\mathbf{X}\beta\|^2=\left\|\mathbf{y}-\mathbf{X}\widehat{\beta}\right\|^2.$$

Then the linear SimSel method consists of the following steps.

The linear SimSel Method

1. Choose a sequence of positive numbers $0\leq\lambda_1\leq\lambda_2\leq...\leq\lambda_K$.
2. For each $\lambda\in\{\lambda_1,..,\lambda_K\}$ pseudo errors $\varepsilon_1^*,...,\varepsilon_n^*$ are generated , *i.i.* P^* distributed with $E\ \varepsilon_i^*=0,\ Var(\varepsilon_i^*)=1$. The pseudo errors $\varepsilon^*=(\varepsilon_1^*,...,\varepsilon_n^*)^T$ are added to $\mathbf{x}_{(1)}$

   $$\mathbf{x}_{(1)}(\lambda)=\mathbf{x}_{(1)}+\sqrt{\lambda}\varepsilon^*.$$

   Note, the other columns are unchanged!

   We get a new data matrix $(\mathbf{y},\mathbf{X}(\lambda))$ with

   $$\mathbf{X}(\lambda)=\mathbf{X}+\sqrt{\lambda}\Delta,$$

   where Δ is the $(n\times p)-$ matrix

   $$\Delta=\begin{pmatrix}\varepsilon_1^* & 0 & \cdots & 0\\ \varepsilon_2^* & 0 & \cdots & 0\\ \vdots & \vdots & 0 & \vdots\\ \varepsilon_{n-1}^* & 0 & \cdots & \vdots\\ \varepsilon_n^* & 0 & \cdots & 0\end{pmatrix}.$$

3. Compute for each $\lambda \in \{\lambda_1, .., \lambda_K\}$

$$RSS(\lambda) = \min_{\beta \in \mathbb{R}^p} \|\mathbf{y} - \mathbf{X}(\lambda)\beta\|^2 .$$

4. Under the assumption that $\left(\mathbf{X}^T\mathbf{X}\right)^{-1}$ exists a simple linear regression with $RSS(\lambda_k), k = 1, ..., K$ as response variable and $(\lambda_1, .., \lambda_K)$ as design variable. The F statistic is saved. The R function is called *single.simsel.linear* .

5. The steps 2-4 are repeated several times and the density of the F statistic is estimated. The result is visualized by a violin plot. The R function called is *simsel.linear.f* .

6. If the F statistics are large, then we conclude that the feature $\mathbf{x}_{(1)}$ matters.

Let us now consider the theoretical background. The term (vector, matrix) $r_n = o_{P^*}(1)$ is defined by

$$\lim_{n\to\infty} P^*(\|r_n\| > \varepsilon) = 0 \text{ for all } \varepsilon > 0.$$

Theorem:

Under the assumption that $\left(\mathbf{X}^T\mathbf{X}\right)^{-1}$ exists, it holds

$$\frac{1}{n}RSS(\lambda) = \frac{1}{n}RSS + \frac{\lambda_1}{1 + h_{11}\lambda_1}\left(\widehat{\beta}_1\right)^2 + o_{P^*}(1)$$

where h_{11} is the $(1,1)-$element of $\left(\frac{1}{n}\mathbf{X}^T\mathbf{X}\right)^{-1}$and $\widehat{\beta}_1$ is the first component of the OLSE estimator $\widehat{\beta} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}$.

Proof: The proof is given in the Appendix.

Remark:

In the procedure we use a linear approximation, because for small $\lambda_1, .., \lambda_K$, it holds

$$\frac{\lambda_k}{1 + h_{11}\lambda_k} \approx \lambda_k; \quad k = 1, ..., K.$$

Remark:

We have not required any model assumption for this procedure. We compare only the ordinary least squares fit in an approximative models. In linear errors-in-variables model the OLSE is inconsistent. But if β_1 is zero, then OLSE converges also to zero. This gives the chance for a successful application of SimSel to errors-in-variables models.

3 Quadratic SimSel

Suppose a possible nonlinear relationship between the features and the response.

In this case we study the best quadratic approximation. For this method a center and scaling of all variables are essential.

$$\begin{aligned} y_{cent,i} &= \frac{y_i - \overline{y}}{\|y\|}, \;\; x_{cent,ji} = \frac{x_{ji} - \overline{x}_{(j)}}{\left\|x_{(j)}\right\|}, j = 1...p; \qquad (1) \\ \mathbf{y}_{cent} &= (y_{cent,i})_{i=1..n}, \;\; \mathbf{x}_{cent(j)} = (x_{cent,ji})_{i=1..n} \end{aligned}$$

We introduce the quadratic approximation

$$H\left(\mathbf{x}_{cent(1)}, ..., \mathbf{x}_{cent(p)}\right) = linear + mixture + quadratic$$

of length $m = \frac{1}{2}(p^2 + 3p)$. With out loss of generality we assume that the feature of interest is related to $\mathbf{x}_{(1)}$ the first column of $\mathbf{X}$. We organize the quadratic approximation, such that the first $p+1$ terms includes $\mathbf{x}_{(1)}$.

$$\begin{aligned} & H\left(\mathbf{x}_{cent(1)}, ..., \mathbf{x}_{cent(p)}\right) \\ = \; & \beta_1 \, \mathbf{x}_{cent(1)} + \beta_{p+1} \left(\mathbf{x}_{cent(1)}\mathbf{x}_{cent(2)}\right) + ...\beta_p \left(\mathbf{x}_{cent(1)}\mathbf{x}_{cent(p)}\right) + \beta_{p+1} \, \mathbf{x}^2_{cent(1)} \\ & + \beta_{p+2} \, \mathbf{x}_{cent(2)} + ... + \beta_m \, \mathbf{x}^2_{cent(p)} \\ = \; & \mathbf{H}\beta, \end{aligned}$$

where $\left(\mathbf{x}_{cent(1)}\mathbf{x}_{cent(2)}\right)$ is the vector produced by componentwisely multiplication. The columns $\mathbf{h}_{(1)}, ..., \mathbf{h}_{(m)}$ of $\mathbf{H}$ are centered again but without scaling:

$$h_{ce,ji} = h_{ji} - \frac{1}{n}\sum_{i=1}^{n} h_{ji}, \;\; \mathbf{h}_{ce(j)} = \left(h_{,ji}\left(\lambda\right)\right)_{i=1...n}, \; j = 1, ..., m \qquad (2)$$

Introduce the new matrix

$$\mathbf{H}_{ce} = \left(\mathbf{h}_{ce(1)}, ..., \mathbf{h}_{ce(m)}\right)$$

and the centered quadratic approximation by

$$\mathbf{H}_{ce}(\mathbf{x}_{cent(1)}, ..., \mathbf{x}_{cent(p)}) = \mathbf{H}_{ce}\beta.$$

The quadratic SimSel method is based on this centered quadratic approximation. We apply as criterion

$$RSS = \min_{\beta} \left\| \mathbf{y}_{cent} - \mathbf{H}_{ce}\beta \right\|^2 .$$

The quadratic SimSel Method

1. Center all variables by (1),
2. Choose a sequence of positive numbers $0 \leq \lambda_1 \leq \lambda_2 \leq ... \leq \lambda_K$.
3. For each $\lambda \in \{\lambda_1, .., \lambda_K\}$ generate pseudo errors $\varepsilon^* = (\varepsilon_1^*, ...; \varepsilon_n^*)^T$, *i.i.* P^* distributed with $E\ \varepsilon_i^* = 0,\ \ Var(\varepsilon_i^*) = 1,\ \ E(\varepsilon_i^*)^3 = 0$ and E $(\varepsilon_i^*)^4 = \mu + 1$.

   Add the pseudo errors to the first feature.

   $$\mathbf{x}_{cent(1)}\left(\lambda\right) = \mathbf{x}_{cent(1)} + \sqrt{\lambda}\varepsilon^*$$

   Note, the other variables are unchanged!
4. Calculate the quadratic expansion

   $$H_{ce}\left(\mathbf{x}_{cent(1)}\left(\lambda\right), \mathbf{x}_{cent(2)} ..., \mathbf{x}_{cent(p)}\right) = \mathbf{H}_{ce}\left(\lambda\right)\beta.$$

   with

   $$\mathbf{H}_{ce}\left(\lambda\right) = \left(\mathbf{h}_{ce(1)}\left(\lambda\right), ..., \mathbf{h}_{ce(m)}\left(\lambda\right)\right)$$
5. Compute for each $\lambda \in \{\lambda_1, .., \lambda_K\}$

   $$RSS\left(\lambda\right) = \min_{\beta \in \mathbb{R}^m} \left\| \mathbf{y} - \mathbf{H}_{ce}\left(\lambda\right)\beta \right\|^2 .$$

6. Calculate a simple linear regression with $RSS\left(\lambda_k\right), k = 1...K$ as response variable and $\left(\lambda_1, .., \lambda_K\right)$ as design variable. Save the F-statistic. The R function is called *single.simsel.quadratic* .
7. The steps 3 - 6 are repeated several times and the density of the F statistic is estimated. The result is visualized by a violin plot. The R function is called *simsel.quadratic.f* .
8. If the F statistics are large, then we conclude that the feature $\mathbf{x}_{(1)}$ matters.

The theoretical background is given by the following result.

Theorem: Under the assumption, that $(\frac{1}{n}\mathbf{H}_{ce}^T\mathbf{H}_{ce})^{-1}$ exists it holds

$$\frac{1}{n}RSS(\lambda) = \frac{1}{n}RSS + \widehat{\beta}^T\mathbf{D}(\lambda)\widehat{\beta} + o_{P^*}(1)$$

where $\widehat{\beta}^T\mathbf{D}(\lambda)\widehat{\beta}$ includes $\widehat{\beta}_1, ..., \widehat{\beta}_{p+1}$ only.

In the special case $p = 2$. It holds

$$\frac{1}{n}RSS\left(\lambda\right) = \frac{1}{n}RSS + \widehat{\beta}_{(1)}^T(\mathbf{D}\left(\lambda\right)^{-1} + \mathbf{H}_{(3,3)})^{-1}\widehat{\beta}_{(1)} + o_{P^*}(1)$$

with $\widehat{\beta}_{(1)} = (\widehat{\beta}_1, \widehat{\beta}_2, \widehat{\beta}_3)^T$ and

$$(\frac{1}{n}\mathbf{H}_{ce}^T\mathbf{H}_{ce})^{-1} = \left(\begin{array}{cc} \mathbf{H}_{(3,3)} & \mathbf{H}_{(4)} \\ \mathbf{H}_{(4)}^T & \mathbf{H}_{(5)} \end{array}\right)$$

and

$$D\left(\lambda\right) = \left(\begin{array}{ccc} \lambda & 0 & 0 \\ 0 & \lambda & 2\lambda\rho \\ 0 & 2\lambda\rho & 4\lambda + \lambda^2\mu \end{array}\right), \quad \rho = \frac{1}{n}\sum_{i=1}^{n} x_{cent,2i}x_{cent1,i}$$

Proof: The proof is given in the Appendix.

4 Simulation

Suppose an $n \times (p+1)$ data set $(\mathbf{y}, \mathbf{X})$. The simulation studies outlined below aim to show SimSel's performance in a:

- linear model setting where the columns in $\mathbf{X}$ are uncorrelated and correlated, respectively
- nonlinear model setting where the columns in $\mathbf{X}$ are uncorrelated and correlated, respectively
- linear errors-in-variable model setting where the columns in $\mathbf{X}$ are uncorrelated and correlated, respectively
- nonlinear errors-in-variable model setting where the columns in $\mathbf{X}$ are uncorrelated and correlated, respectively

Data sets for simulation studies were generated by assuming that $\mathbf{y}$ follow either a linear or a nonlinear model according to

$$y_i = \Sigma_{j=1}^{p} \beta_j x_{ij} + \epsilon_i \tag{3}$$

or

$$y_i = \sin(\Sigma_{j=1}^{p} \beta_j x_{ij}) + \epsilon_i, \tag{4}$$

respectively, where $\epsilon_i \sim N(0, \sigma^2)$, $i = 1, ..., 40$. σ^2 is chosen to adhere the y-variable with roughly ten percent error. We fix the number of independent variables to 4, i.e. $p = 4$, out of which two are relevant for the response variable $\mathbf{y}$, i.e. have $\beta_j \neq 0$. Arbitrarily we pick the relevant variables to be $j = 1$ and $j = 3$, both having $\beta = 1$, whereas the variables 2 and 4 have $\beta = 0$. The data matrix $\mathbf{X} = x_{ij}$ is sampled from a multivariate normal distribution. Thus $\mathbf{X} \sim N_4(\mathbf{0}, \Sigma_4)$, where $\mathbf{0}$ is a p-dimensional vector of zeros and Σ_4 is either the $\mathbf{I}_4$ identity matrix or the matrix $\mathbf{S}$ specified below, thus resulting in matrices $\mathbf{X}$ with uncorrelated or correlated variables.

$$S = \begin{pmatrix} 1 & 0.23 & -0.29 & 0.51 \\ 0.23 & 1 & -0.89 & 0.03 \\ -0.29 & -0.89 & 1 & -0.35 \\ 0.51 & 0.03 & -0.35 & 1 \end{pmatrix}$$

The result of the nine different simulation studies are shown below. The figures show the distribution of the F-statistic for the simple linear regression with $RSS\,(\lambda_1)\,, ..., RSS\,(\lambda_K)$ as response variable and $\lambda_1, ..., \lambda_K$ as design variable in the linear SimSel procedure and in the quadratic SimSel procedure description after 100 repetitions of the entire SimSel.

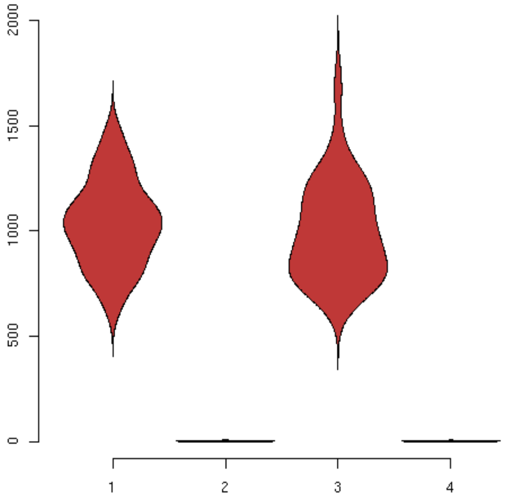

Figure 1: Linear data, no errors-in-variables and uncorrelated data.

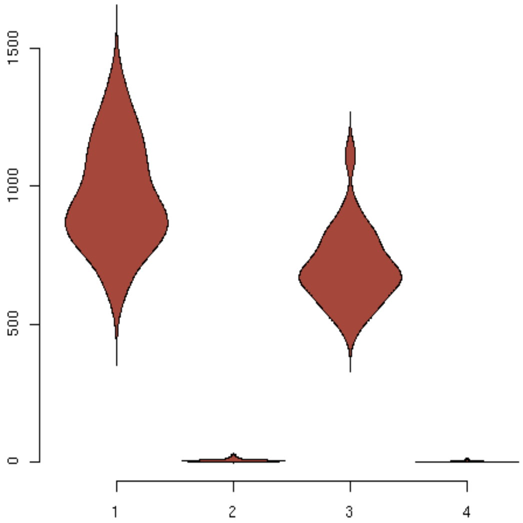

Figure 2: Linear data, no errors-in-variables and correlated data.

5 Application

In chemistry and pharmacology it is often desirable to characterize the interactions between molecules and proteins by for instance quantitative structure-

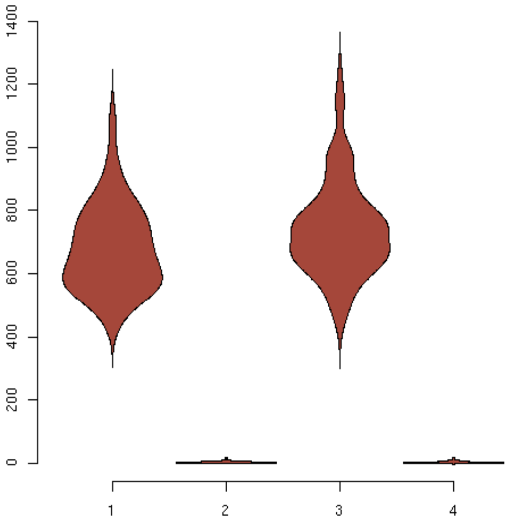

Figure 3: Linear data, errors-in-variables and uncorrelated data.

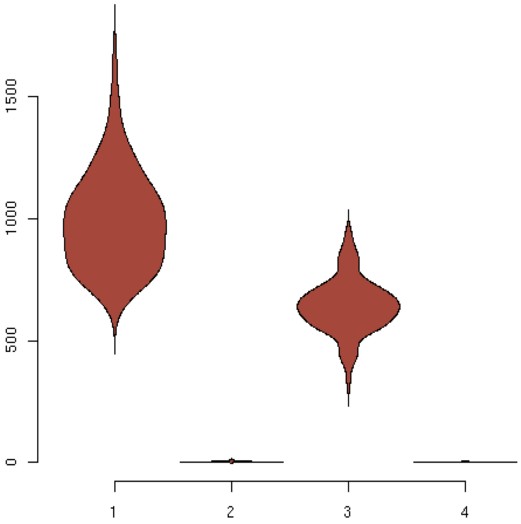

Figure 4: Linear data, errors-in-variables and correlated data.

activity relationship (QSAR) or Proteochemometrics (PCM) [5]. In these methods chemical structures are numerically described (by e.g. their physiochemical properties) and correlated with a well defined measurable process, such as biological activity or chemical reactivity, typically by the means of

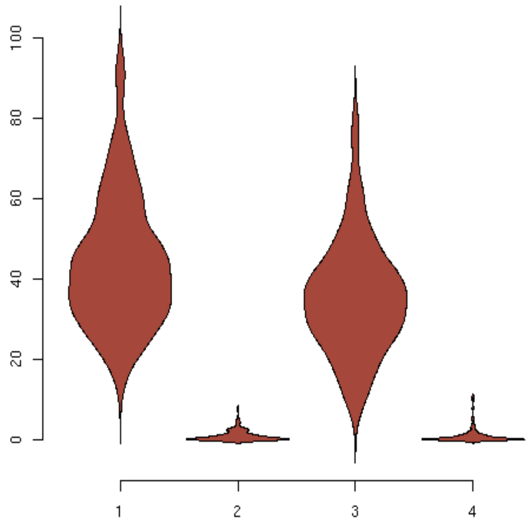

Figure 5: Nonlinear data, no errors-in-variables and uncorrelated data.

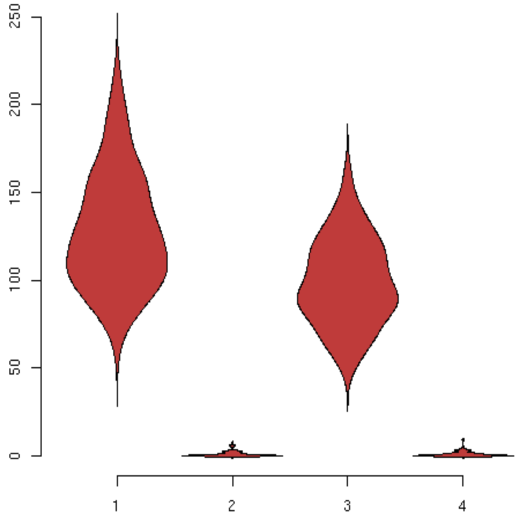

Figure 6: Nonlinear data, no errors-in-variables and correlated data.

linear or nonlinear regression methods. The data sets often suffer from having few observations in relation to the number of variables, often even $p > n$ (this problem is here referred to as 'underdeterminedness'). The underdeterminedness arises from the fact that it is very difficult to a priori say which

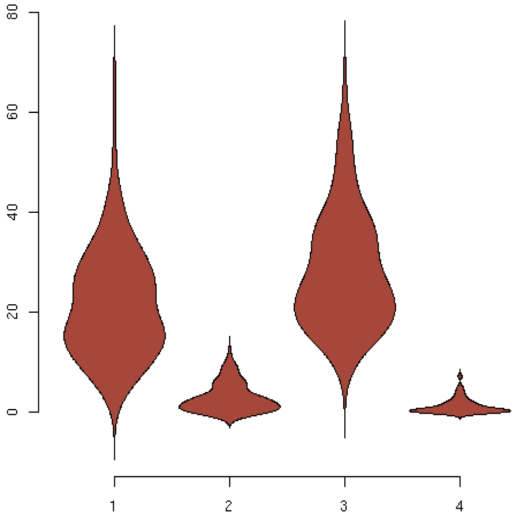

Figure 7: Nonlinear data, errors-in-variables and uncorrelated data.

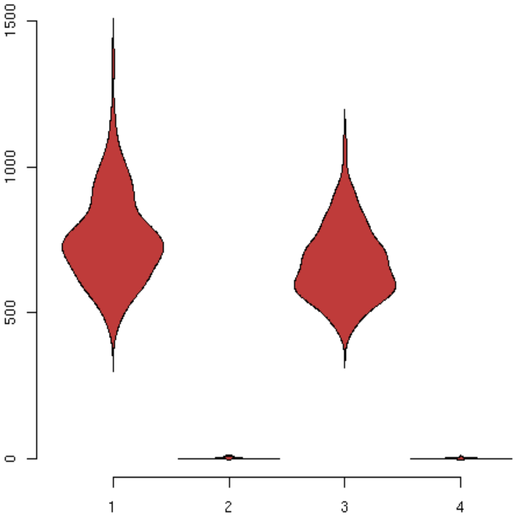

Figure 8: Linear data, errors-in-variables and correlated data.

molecular features that govern an interaction and the many available ways of numerically describing a chemical structure. Moreover, the variables in **X** are often strongly correlated and are impossible to observe without errors (occurring from either measurement errors or calculation errors), leading to errors-in-variables.

One way to deal with the problem of underdeterminedness is to employ feature-selection methods to reduce the number of variables in **X**. It is then important that the selection method works under the conditions stipulated by the problem domain, i.e. in chemistry and pharmacology that the method works for underdetermined, nonlinear errors-in-variable models with correlated variables. While SimSel not (yet) has been extended to work for underdetermined problems, we have demonstrated its usefulness on linear, nonlinear, linear EIV, and nonlinear EIV problems with uncorrelated and correlated design matrices on simulated data.

In order to make an evaluation of the methods performance on a small data set from a real application, we used the well-studied Selwood data set [4]. The Selwood data contains 31 observations of 53 variables and is thus grossly underdetermined. From the 53 variables we selected four out of which two are known from previous studies to be relevant for the response variable and the other two are known to not be relevant, see [3].

6 SimSel R-package

The SimSel procedure is implemented in the R-package SimSel 0.1 and is available at http://www.math.uu.se/∼zwanzig.

The two main functions are *simsel.linear* and *simsel.linear.f* for linear SimSel and *simsel.quadratic* and *simsel.quadratic.f* for the quadratic SimSel procedure. The *simsel.linear* and *simsel.quadratic* functions go through the linear and quadric SimSel procedures, respectively, and plot the points $\lambda_k, RSS\left(\lambda_k\right),\ i = 1, ..., p$. The *simsel.linear.f* and *simsel.quadratic.f* run the respective procedure a given, say B, number of times. Each time the F-statistic is collected. The distribution of the F-statistic for each variable over the B repetitions is plotted as a result.

The code for the simulations is also available in the SimSel package.

7 Discussion

The proposed method has some advantages:

- The method is model robust.
- It is easy to interpret and intuitive heuristic.
- No splitting of the data set in training and test subsets is required.

Open Problems:

- Relax the rank condition, that $(\frac{1}{n}\mathbf{H}_{ce}^T\mathbf{H}_{ce})^{-1}$ exists.
- Implement selection procedure steps.
- Generalize the procedure for testing more than one features simultaneously.
- Combine SimSel with an orthogonalization procedure.

8 Appendix

Proof for linear SimSel: It holds

$$\frac{1}{n}RSS\,(\lambda) = \frac{1}{n}\mathbf{y}^T\mathbf{y} - \frac{1}{n}\mathbf{y}^T P(\lambda)\mathbf{y} \tag{5}$$

with

$$P(\lambda) = \mathbf{X}(\lambda)\left(\mathbf{X}(\lambda)^T\mathbf{X}(\lambda)\right)^{-1}\mathbf{X}(\lambda)^T. \tag{6}$$

First we consider

$$\frac{1}{n}\mathbf{X}(\lambda)^T\mathbf{y} = \left(\frac{1}{n}\mathbf{X} + \frac{1}{n}\sqrt{\lambda}\Delta\right)^T\mathbf{y}.$$

Introduce the denotations:

$$\mathbf{e}_1 = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}; \varepsilon^* = \begin{pmatrix} \varepsilon_1^* \\ \varepsilon_2^* \\ \vdots \\ \varepsilon_n^* \end{pmatrix}$$

Then

$$\Delta = \varepsilon^{*}\mathbf{e}_1^T \text{ and } \Delta^T\mathbf{y} =\varepsilon^{*T}\mathbf{y}\mathbf{e}_1.$$

For arbitrary fixed $\mathbf{y}$ it holds by the law of large numbers

$$\frac{1}{n}\varepsilon^{*T}\mathbf{y} = o_{P^*}(1),$$

because

$$\frac{1}{n}\varepsilon^{*T}\mathbf{y} = \frac{1}{n}\sum_{i=1}^{n} y_i\varepsilon_i^*,\ Ey_i\varepsilon_i^* = 0,\ Var\left(y_i\varepsilon_i^*\right) < \infty.$$

Thus

$$\frac{1}{n}\mathbf{X}(\lambda)^T\mathbf{y} =\frac{1}{n}\mathbf{X}^T\mathbf{y}+o_{P^*}(1). \tag{7}$$

Consider now $\mathbf{X}(\lambda)^T\mathbf{X}(\lambda)$:

$$\begin{aligned}\frac{1}{n}\mathbf{X}(\lambda)^T\mathbf{X}(\lambda) &= \frac{1}{n}\left(\mathbf{X}+\sqrt{\lambda}\Delta\right)^T\left(\mathbf{X}+\sqrt{\lambda}\Delta\right) \\ &= \frac{1}{n}\mathbf{X}^T\mathbf{X}+\frac{1}{n}\sqrt{\lambda}\mathbf{X}^T\Delta+\frac{1}{n}\sqrt{\lambda}\Delta^T\mathbf{X}+\frac{1}{n}\lambda\Delta^T\Delta\end{aligned} \tag{8}$$

We have

$$\frac{1}{n}\mathbf{X}^T\Delta = \begin{pmatrix} \frac{1}{n}\varepsilon^{*T}\mathbf{x}_{(1)} & 0 & \cdots & 0 \\ \vdots & \vdots & 0 & \vdots \\ \frac{1}{n}\varepsilon^{*T}\mathbf{x}_{(p)} & 0 & \cdots & 0 \end{pmatrix}.$$

By the same argumentation as for $\frac{1}{n}\varepsilon^{*T}\mathbf{y}$ we obtain that

$$\frac{1}{n}\mathbf{X}^T\Delta = o_{P^*}(1);\ \frac{1}{n}\Delta^T\mathbf{X} =o_{P^*}(1). \tag{9}$$

Further

$$\frac{1}{n}\Delta^T\Delta = \begin{pmatrix} \frac{1}{n}\varepsilon^{*T}\varepsilon^* & \cdots & 0 \\ \vdots & 0 & \vdots \\ 0 & \cdots & 0 \end{pmatrix} = \frac{1}{n}\varepsilon^{*T}\varepsilon^*\mathbf{e}_1\mathbf{e}_1^T.$$

By the law of large numbers

$$\frac{1}{n}\varepsilon^{*T}\varepsilon^* = \frac{1}{n}\sum_{i=1}^{n}(\varepsilon_i^*)^2 = Var\left(\varepsilon_1^*\right) + o_{P^*}(1) = 1 + o_{P^*}(1).$$

Thus

$$\frac{1}{n}\Delta^T\Delta = \mathbf{e}_1\mathbf{e}_1^T + o_{P^*}(1). \tag{10}$$

Summarizing (8), (9) and (10) we have

$$\frac{1}{n}\mathbf{X}(\lambda)^T\mathbf{X}(\lambda) = \frac{1}{n}\mathbf{X}^T\mathbf{X}+\lambda\mathbf{e}_1\mathbf{e}_1^T + o_{P^*}(1).$$

Hence

$$\left(\frac{1}{n}\mathbf{X}(\lambda)^T\mathbf{X}(\lambda)\right)^{-1} = \left(\frac{1}{n}\mathbf{X}^T\mathbf{X} + \lambda\mathbf{e}_1\mathbf{e}_1^T\right)^{-1} + o_{P^*}(1).$$

Using the relation

$$(A+BCD)^{-1} = A^{-1} - A^{-1}B(C^{-1}+DA^{-1}B)^{-1}DA^{-1}$$

with $A = \frac{1}{n}\mathbf{X}^T\mathbf{X},\ \mathbf{B} = \mathbf{D}^T{=}\sqrt{\lambda}\mathbf{e}_1,\ C = 1$ and $h_{11} = \mathbf{e}_1^T\left(\frac{1}{n}\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{e}_1$ we get

$$\left(\frac{1}{n}\mathbf{X}^T\mathbf{X} + \lambda\mathbf{e}_1\mathbf{e}_1^T\right)^{-1} = \left(\frac{1}{n}\mathbf{X}^T\mathbf{X}\right)^{-1} - \frac{\lambda}{1+\lambda h_{11}}\left(\frac{1}{n}\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{e}_1\mathbf{e}_1^T\left(\frac{1}{n}\mathbf{X}^T\mathbf{X}\right)^{-1}. \tag{11}$$

Consider now the term $\frac{1}{n}\mathbf{y}^TP(\lambda)\mathbf{y}$ in (5). Using (7) and (11)

$$\begin{aligned}
\frac{1}{n}\mathbf{y}^TP(\lambda)\mathbf{y} &= \frac{1}{n}\mathbf{y}^T\mathbf{X}(\lambda)\left(\frac{1}{n}\mathbf{X}(\lambda)^T\mathbf{X}(\lambda)\right)^{-1}\frac{1}{n}\mathbf{X}(\lambda)^T\mathbf{y} \\
&= \left(\frac{1}{n}\mathbf{X}^T\mathbf{y}\right)^T\left(\frac{1}{n}\mathbf{X}^T\mathbf{X} + \lambda\mathbf{e}_1\mathbf{e}_1^T\right)^{-1}\left(\frac{1}{n}\mathbf{X}^T\mathbf{y}\right) + o_{P^*}(1) \\
&= \frac{1}{n}\mathbf{y}^TP\mathbf{y} - \frac{\lambda}{1+\lambda h_{11}}\widehat{\beta}^T\mathbf{e}_1\mathbf{e}_1^T\widehat{\beta} + o_{P^*}(1) \\
&= \frac{1}{n}\mathbf{y}^TP\mathbf{y} - \frac{\lambda}{1+\lambda h_{11}}(\widehat{\beta}_1)^2 + o_{P^*}(1).
\end{aligned}$$

Then the results follows from (5)

$$\begin{aligned}
\frac{1}{n}RSS\,(\lambda) &= \frac{1}{n}\mathbf{y}^T\mathbf{y} - \frac{1}{n}\mathbf{y}^TP\mathbf{y} + \frac{\lambda}{1+\lambda h_{11}}(\widehat{\beta}_1)^2 + o_{P^*}(1) \\
&= \frac{1}{n}RSS + \frac{\lambda}{1+\lambda h_{11}}(\widehat{\beta}_1)^2 + o_{P^*}(1).
\end{aligned}$$

Proof for quadratic SimSel: Consider two features only: $p = 2,\ m = 5$. The quadratic approximation is

$$H(\mathbf{x}_{(1)}, \mathbf{x}_{(2)}) = \beta_1 \mathbf{x}_{(1)} + \beta_2 \left(\mathbf{x}_{(1)}\mathbf{x}_{(2)}\right) + \beta_3 \mathbf{x}_{(1)}^2 + \beta_4 \mathbf{x}_{(1)} + \beta_5 \mathbf{x}_{(2)}^2$$

Step 4 of procedure gives

$$\mathbf{H}_{ce}\left(\lambda\right) = \left(\mathbf{h}_{ce(1)}\left(\lambda\right), ..., \mathbf{h}_{ce(5)}\left(\lambda\right)\right),$$

with $\mathbf{h}_{ce(j)}\left(\lambda\right) = \left(h_{j,i}\left(\lambda\right)\right)_{i=1...n}$, and $\mathbf{h}_{ce(j)} = \mathbf{h}_{ce(j)}\left(0\right) = \left(h_{cej,i}\right)_{i=1...n},\ j = 1...5$

$$\begin{aligned}
h_{1,i}\left(\lambda\right) &= (x_{cent,1i} + \sqrt{\lambda}\varepsilon_i^*) - \frac{1}{n}\sum_{k=1}^{n}(x_{cent,1k} + \sqrt{\lambda}\varepsilon_k^*) \\
&= h_{ce1,i} + \sqrt{\lambda}(\varepsilon_i^* - \frac{1}{n}\sum_{k=1}^{n}\varepsilon_k^*) \\
h_{ce1,i} &= x_{cent,1i}
\end{aligned}$$

$$\begin{aligned}
h_{2,i}\left(\lambda\right) &= (x_{cent,1i} + \sqrt{\lambda}\varepsilon_i^*)x_{cent,2i} - \frac{1}{n}\sum_{k=1}^{n}(x_{cent,1k} + \sqrt{\lambda}\varepsilon_k^*)x_{cent,2k} \\
&= h_{ce2,i} + 2\sqrt{\lambda}(\varepsilon_i^* x_{cent,2i} - \frac{1}{n}\sum_{k=1}^{n}\varepsilon_k^* x_{cent,2k}) \\
h_{ce2,i} &= x_{cent,1i}x_{cent,2i} - \frac{1}{n}\sum_{k=1}^{n}x_{cent,1k}x_{cent,2k}
\end{aligned}$$

$$\begin{aligned}
h_{3,i}\left(\lambda\right) &= (x_{cent,1i} + \sqrt{\lambda}\varepsilon_i^*)^2 - \frac{1}{n}\sum_{k=1}^{n}(x_{cent,1k} + \sqrt{\lambda}\varepsilon_k^*)^2 \\
&= h_{ce3,i} + 2\sqrt{\lambda}(\varepsilon_i^* x_{cent,1i} - \frac{1}{n}\sum_{k=1}^{n}\varepsilon_k^* x_{cent,1k}) \\
&\quad +\lambda(\varepsilon_i^{*2} - \frac{1}{n}\sum_{k=1}^{n}\varepsilon_k^{*2}) \\
h_{ce3,i} &= x_{cent,1i}^2 - \frac{1}{n}\sum_{k=1}^{n}x_{cent,1k}^2
\end{aligned}$$

$$\mathbf{h}_{ce(4)}\left(\lambda\right) = \mathbf{h}_{ce(4)} = \mathbf{x}_{cent(2)}$$

$$\mathbf{h}_{ce(5)}\left(\lambda\right) = \mathbf{h}_{ce(5)}, \quad h_{ce5,i} = x_{cent,2i}^2 - \frac{1}{n}\sum_{i=1}^{n} x_{cent,2k}^2.$$

Summarizing we obtained:

$$\mathbf{H}_{ce}\left(\lambda\right) = \mathbf{H}_{ce}+\sqrt{\lambda}\mathbf{D}_1 + \lambda\mathbf{D}_2$$

with $\mathbf{H}_{ce} = \mathbf{H}_{ce}\left(0\right) = \left(\mathbf{h}_{ce(1)}, ..., \mathbf{h}_{ce(5)}\right)$

$$\mathbf{D}_1 = \left(\varepsilon_i^* - \frac{1}{n}\sum_{k=1}^{n}\varepsilon_k^*, , \varepsilon_i^* x_{cent,2i} - \frac{1}{n}\sum_{k=1}^{n}\varepsilon_k^* x_{cent,2k}, 2(\varepsilon_i^* x_{cent,1i} - \frac{1}{n}\sum_{k=1}^{n}\varepsilon_k^* x_{cent,1k}), 0, 0\right)_{i=1..n}$$

and

$$\mathbf{D}_2 = \left(0, 0, (\varepsilon_i^{*2} - \frac{1}{n}\sum_{k=1}^{n}\varepsilon_k^{*2}), 0, 0\right)_{i=1..n}.$$

Introducing notations for the columns of $\mathbf{D}_1$ and $\mathbf{D}_2$ we have

$$\mathbf{D}_1 = (a, b, 2c, 0, 0) \;\; \mathbf{D}_2 = (0, 0, d, 0, 0).$$

Compare the RSS

$$\frac{1}{n}RSS\left(\lambda\right) = \frac{1}{n}\mathbf{y}_{cent}^T\mathbf{y}_{cent} - \frac{1}{n}\mathbf{y}_{cent}^T P(\lambda)\mathbf{y}_{cent} \tag{12}$$

with

$$\frac{1}{n}\mathbf{y}_{cent}^T P(\lambda)\mathbf{y}_{cent} = \frac{1}{n}\mathbf{y}_{cent}^T\mathbf{H}_{ce}(\lambda)\left(\frac{1}{n}\mathbf{H}_{ce}(\lambda)^T\mathbf{H}_{ce}(\lambda)\right)^{-1}\frac{1}{n}\mathbf{H}_{ce}(\lambda)^T\mathbf{y}_{cent}.$$

We show first

$$\frac{1}{n}\mathbf{H}_{ce}(\lambda)^T\mathbf{y}_{cent} = \frac{1}{n}\mathbf{H}_{ce}^T\mathbf{y}_{cent} + o_{P^*}(1). \tag{13}$$

We have

$$\frac{1}{n}\mathbf{D}_1^T\mathbf{y}_{cent} = o_{P^*}(1) \quad and \quad \frac{1}{n}\mathbf{D}_2^T\mathbf{y}_{cent} = o_{P^*}(1).$$

Because of $\sum_{i=1}^n y_{cent,i} = 0$ and of the law of large number with $E\varepsilon_i^* = 0$ it holds

$$\frac{1}{n} a^T y_{cent} = \frac{1}{n} \sum_{i=1}^n \varepsilon_i^* y_{cent,i} = \frac{1}{n} \sum_{i=1}^n E\varepsilon_i^* y_{cent,i} + o_{P^*}(1) = o_{P^*}(1) \tag{14}$$

and

$$\frac{1}{n} b^T y_{cent} = \frac{1}{n} \sum_{i=1}^n \varepsilon_i^* x_{cent,ik} y_{cent,i} = \frac{1}{n} \sum_{i=1}^n E\varepsilon_i^* x_{cent,ik} y_{cent,i} + o_{P^*}(1) = o_{P^*}(1).$$

Analogously $\frac{1}{n} c^T y_{cent} = o_{P^*}(1)$. Using $E(\varepsilon_i^*)^2 = 1$ and $\sum_{i=1}^n y_{cent,i} = 0$ we get

$$\begin{aligned} \frac{1}{n} d^T y_{cent} &= \frac{1}{n} \sum_{i=1}^n (\varepsilon_i^*)^2 y_{cent,i} = \frac{1}{n} \sum_{i=1}^n E(\varepsilon_i^*)^2 y_{cent,i} + o_{P^*}(1) \\ &= \frac{1}{n} \sum_{i=1}^n y_{cent,i} + o_{P^*}(1) = o_{P^*}(1). \end{aligned}$$

Consider now $\frac{1}{n} \mathbf{H}_{ce}(\lambda)^T \mathbf{H}_{ce}(\lambda)$:

$$\begin{aligned} \frac{1}{n} \mathbf{H}_{ce}(\lambda)^T \mathbf{H}_{ce}(\lambda) &= \frac{1}{n} (\mathbf{H}_{ce} + \sqrt{\lambda} \mathbf{D}_1 + \lambda \mathbf{D}_2)^T (\mathbf{H}_{ce} + \sqrt{\lambda} \mathbf{D}_1 + \lambda \mathbf{D}_2) \\ &= \frac{1}{n} (\mathbf{H}_{ce}^T \mathbf{H}_{ce} + \sqrt{\lambda} \mathbf{D}_1^T \mathbf{H}_{ce} + \lambda \mathbf{D}_2^T \mathbf{H}_{ce} + \\ &\quad \sqrt{\lambda} \mathbf{H}_{ce}^T \mathbf{D}_1 + \lambda \mathbf{D}_1^T \mathbf{D}_1 + \lambda \sqrt{\lambda} \mathbf{D}_2^T \mathbf{D}_1 \\ &\quad + \lambda \mathbf{H}_{ce}^T \mathbf{D}_2 + \lambda \sqrt{\lambda} \mathbf{D}_1^T \mathbf{D}_2 + \lambda^2 \mathbf{D}_2^T \mathbf{D}_2) \end{aligned}$$

Note the columns of $\mathbf{H}_{ce}$ are centered. We get analogously to (14)

$$\mathbf{D}_1^T \mathbf{H}_{ce} = o_{P^*}(1), \quad \mathbf{D}_2^T \mathbf{H}_{ce} = o_{P^*}(1).$$

Especially

$$\frac{1}{n} d^T h_{cent(4)} = \frac{1}{n} \sum_{i=1}^n \varepsilon_i^{*2} h_{cent,4i} = \frac{1}{n} \sum_{i=1}^n E\varepsilon_i^{*2} h_{cent,4i} + o_{P^*}(1) = o_{P^*}(1).$$

Remains to study $\lambda\mathbf{D}_1^T\mathbf{D}_1 + \lambda\sqrt{\lambda}(\mathbf{D}_2^T\mathbf{D}_1 + \mathbf{D}_1^T\mathbf{D}_2) + \lambda^2\mathbf{D}_2^T\mathbf{D}_2$. Consider $\frac{1}{n}\mathbf{D}_1^T\mathbf{D}_1$.

$$\frac{1}{n}a^T a = \frac{n-1}{n}s^2(\varepsilon^*) = \frac{n-1}{n}E(s^2(\varepsilon^*)) + o_{P^*}(1) = 1 + o(1) + o_{P^*}(1),$$

where $s^2(\varepsilon^*)$ is the sample variance and

$$\begin{aligned}
\frac{1}{n}a^T b &= \frac{1}{n}\sum_{i=1}^n \varepsilon_i^{*2}x_{cent,2i} - (\frac{1}{n}\sum \varepsilon_i^*)(\frac{1}{n}\sum \varepsilon_i^* x_{cent,2i}) \\
&= \frac{1}{n}\sum_{i=1}^n E\varepsilon_i^{*2}x_{cent,2i} - (\frac{1}{n}\sum E\varepsilon_i^*)(\frac{1}{n}\sum E\varepsilon_i^* x_{cent,2i}) + o_{P^*}(1) \\
&= \frac{1}{n}\sum_{i=1}^n x_{cent,2i} + o_{P^*}(1) = o_{P^*}(1)
\end{aligned}$$

$$\begin{aligned}
\frac{1}{n}c^T b &= \frac{1}{n}\sum_{i=1}^n \varepsilon_i^{*2}x_{cent,2i}x_{cent,1i} - (\frac{1}{n}\sum \varepsilon_i^* x_{cent,1i})(\frac{1}{n}\sum \varepsilon_i^* x_{cent,2i}) \\
&= \frac{1}{n}\sum_{i=1}^n E\varepsilon_i^{*2}x_{cent,2i}x_{cent,1i} + o_{P^*}(1) \\
&= \frac{1}{n}\sum_{i=1}^n x_{cent,2i}x_{cent,1i} + o_{P^*}(1) = \rho + o_{P^*}(1)
\end{aligned}$$

with

$$\rho = \frac{1}{n}\sum_{i=1}^n x_{cent,2i}x_{cent1,i}.$$

Further

$$\begin{aligned}
\frac{1}{n}c^T c &= \frac{1}{n}\sum_{i=1}^n \varepsilon_i^{*2}x_{cent,2i}^2 - (\frac{1}{n}\sum \varepsilon_i^* x_{cent,2i})^2 \\
&= \frac{1}{n}\sum_{i=1}^n E\varepsilon_i^{*2}x_{cent,2i}^2 + o_{P^*}(1) \\
&= \frac{1}{n}\sum_{i=1}^n x_{cent,2i}^2 + o_{P^*}(1) = 1 + o_{P^*}(1).
\end{aligned}$$

Thus

$$\frac{1}{n}\mathbf{D}_1^T\mathbf{D}_1 = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 2\rho & 0 & 0 \\ 0 & 2\rho & 4 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} + o_{P^*}(1).$$

Study now $\frac{1}{n}\mathbf{D}_1^T\mathbf{D}_2$

$$\begin{aligned} \frac{1}{n}a^T d &= \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^{*3}x_{cent2i} - (\frac{1}{n}\sum\varepsilon_i^{*2})(\frac{1}{n}\sum\varepsilon_i^{*}x_{cent,2i}) \\ &= \frac{1}{n}\sum_{i=1}^{n}E\varepsilon_i^{*3}x_{cent,2i} - (\frac{1}{n}\sum E\varepsilon_i^{*2})(\frac{1}{n}\sum E\varepsilon_i^{*}x_{cent,2i}) + o_{P^*}(1) \\ &= o_{P^*}(1) \end{aligned}$$

The other elements can estimated analogously. Thus

$$\frac{1}{n}\mathbf{D}_1^T\mathbf{D}_2 = o_{P^*}(1).$$

Consider $\frac{1}{n}\mathbf{D}_2^T\mathbf{D}_2$

$$\begin{aligned} \frac{1}{n}d^T d &= \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^{*4} - (\frac{1}{n}\sum\varepsilon_i^{*2})^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}E\varepsilon_i^{*4} - (\frac{1}{n}\sum E\varepsilon_i^{*2})^2 + o_{P^*}(1) \\ &= \mu + o_{P^*}(1). \end{aligned}$$

Summarizing:

$$\frac{1}{n}\mathbf{H}_{ce}(\lambda)^T\mathbf{H}_{ce}(\lambda) = \frac{1}{n}\mathbf{H}_{ce}^T\mathbf{H}_{ce} + \boldsymbol{\Delta} + o_{P^*}(1)$$

where

$$\boldsymbol{\Delta} = \begin{pmatrix} \lambda & 0 & 0 & 0 & 0 \\ 0 & \lambda & 2\lambda\rho & 0 & 0 \\ 0 & 2\lambda\rho & 4\lambda + \lambda^2\mu & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} = \begin{pmatrix} I_3 \\ 0 \end{pmatrix} \mathbf{D}(\lambda) \begin{pmatrix} I_3 & 0 \end{pmatrix}$$

with

$$\mathbf{D}\left(\lambda\right)=\left(\begin{array}{ccc}\lambda & 0 & 0\\ 0 & \lambda & 2\lambda\rho\\ 0 & 2\lambda\rho & 4\lambda+\lambda^{2}\mu\end{array}\right).$$

Apply now the relation

$$(A+BCD)^{-1}=A^{-1}-A^{-1}B(C^{-1}+DA^{-1}B)^{-1}DA^{-1}$$

with $A=\frac{1}{n}\mathbf{H}_{ce}^{T}\mathbf{H}_{ce},\ \mathbf{B}=\mathbf{D}^{T}=\left(\begin{array}{c}I_{3}\\ 0\end{array}\right),\ C=\mathbf{D}\left(\lambda\right)$ we get

$$\begin{aligned}&(\frac{1}{n}\mathbf{H}_{ce}^{T}\mathbf{H}_{ce}+\boldsymbol{\Delta})^{-1}\\ =\ &(\frac{1}{n}\mathbf{H}_{ce}^{T}\mathbf{H}_{ce})^{-1}-(\frac{1}{n}\mathbf{H}_{ce}^{T}\mathbf{H}_{ce})^{-1}\left(\begin{array}{c}I_{3}\\ 0\end{array}\right)(\mathbf{D}\left(\lambda\right)^{-1}+\mathbf{H}_{(3,3)})^{-1}\left(\begin{array}{cc}I_{3} & 0\end{array}\right)(\frac{1}{n}\mathbf{H}_{ce}^{T}\mathbf{H}_{ce})^{-1}\end{aligned}$$

with

$$(\frac{1}{n}\mathbf{H}_{ce}^{T}\mathbf{H}_{ce})^{-1}=\left(\begin{array}{cc}\mathbf{H}_{(3,3)} & \mathbf{H}_{(4)}\\ \mathbf{H}_{(4)}^{T} & \mathbf{H}_{(5)}\end{array}\right)$$

and

$$\mathbf{D}\left(\lambda\right)^{-1}=\tfrac{1}{\lambda}\tfrac{1}{4+\lambda\mu-4\rho^{2}}\left(\begin{array}{ccc}4+\lambda\mu-4\rho^{2} & 0 & 0\\ 0 & 4+\lambda\mu & -2\rho\\ 0 & -2\rho & 1\end{array}\right)$$

Note

$$\left(\begin{array}{cc}I_{3} & 0\end{array}\right)(\frac{1}{n}\mathbf{H}_{ce}^{T}\mathbf{H}_{ce})^{-1}\mathbf{H}_{ce}^{T}\mathbf{y}_{cent}=\widehat{\beta}_{(1)}$$

where

$$\widehat{\beta}_{(1)}=(\widehat{\beta}_{1},\widehat{\beta}_{2},\widehat{\beta}_{2})^{T}$$

Summarizing (12) and (13) we get

$$\frac{1}{n}\mathbf{y}_{cent}^{T}P(\lambda)\mathbf{y}_{cent}=\frac{1}{n}\mathbf{y}_{cent}^{T}P\mathbf{y}_{cent}-\widehat{\beta}_{(1)}^{T}(\mathbf{D}\left(\lambda\right)^{-1}+\mathbf{H}_{(3,3)})^{-1}\widehat{\beta}_{(1)}+o_{P^{*}}(1)$$

Thus

$$\frac{1}{n}RSS\left(\lambda\right)=\frac{1}{n}RSS+\widehat{\beta}_{(1)}^{T}(\mathbf{D}\left(\lambda\right)^{-1}+\mathbf{H}_{(3,3)})^{-1}\widehat{\beta}_{(1)}+o_{P^{*}}(1).$$

References

[1] J.R. Cook and L.A. Stefanski (1994), Simulation-Extrapolation Estimation in parametric measurement error models. JASA Vol 89, 1314-1327.

[2] Xiaohui Lou, L.A. Stefanski and D.G. Boos (2006), Tuning Variable Selection procedures by Adding Noise, Technometrics, May 2006, Vol 48, 165-175

[3] O. Nicolotti and A. Caraotti (2006), QSAR and QSPR studies of a highly structured physisochemical domain. J. Chem. Inf. Model. 2006, 46 (1), 264-76.

[4] D.L. Selwood et al. (1990), Structure-activity relationships of antifilarial antimycin analogs: a multivariate pattern recognition study. J. Med. Chem. 1990, p.136-42

[5] J.E.S. Wikberg, M. Lapinsh and P. Prusis (2004), Proteochemometrics: A tool for modelling the molecular interaction space. In Chemogenomics in Drug Discovery - A Medicinal Chemistry Perspective. Editors: H. Kubinyi and G. Müller.