PROOF COVER SHEET

Author(s): Riccardo De Bin

Article title: Handling co-dependence issues in resampling-based variable selection procedures: a simulation study

Article no: GSCS1378654

Enclosures: 1) Query sheet
2) Article proofs

Dear Author,

1. Please check these proofs carefully. It is the responsibility of the corresponding author to check these and approve or amend them. A second proof is not normally provided. Taylor & Francis cannot be held responsible for uncorrected errors, even if introduced during the production process. Once your corrections have been added to the article, it will be considered ready for publication.

Please limit changes at this stage to the correction of errors. You should not make trivial changes, improve prose style, add new material, or delete existing material at this stage. You may be charged if your corrections are excessive (we would not expect corrections to exceed 30 changes).

For detailed guidance on how to check your proofs, please paste this address into a new browser window: <http://journalauthors.tandf.co.uk/production/checkingproofs.asp>

Your PDF proof file has been enabled so that you can comment on the proof directly using Adobe Acrobat. If you wish to do this, please save the file to your hard disk first. For further information on marking corrections using Acrobat, please paste this address into a new browser window: <http://journalauthors.tandf.co.uk/production/acrobat.asp>

2. Please review the table of contributors below and confirm that the first and last names are structured correctly and that the authors are listed in the correct order of contribution. This check is to ensure that your name will appear correctly online and when the article is indexed.

Sequence	Prefix	Given name(s)	Surname	Suffix
1		Riccardo	De Bin	
2		Willi	Sauerbrei	

Queries are marked in the margins of the proofs, and you can also click the hyperlinks below. Content changes made during copy-editing are shown as tracked changes. Inserted text is in **red font** and revisions have a red indicator . Changes can also be viewed using the list comments function. To correct the proofs, you should insert or delete text following the instructions below, but **do not add comments to the existing tracked changes**.

AUTHOR QUERIES

General points:

- (1) **Permissions:** You have warranted that you have secured the necessary written permission from the appropriate copyright owner for the reproduction of any text, illustration, or other material in your article. Please see <http://journalauthors.tandf.co.uk/permissions/usingThirdPartyMaterial.asp>.
- (2) **Third-party content:** If there is third-party content in your article, please check that the rightsholder details for re-use are shown correctly.
- (3) **Affiliation:** The corresponding author is responsible for ensuring that address and email details are correct for all the co-authors. Affiliations given in the article should be the affiliation at the time the research was conducted. Please see <http://journalauthors.tandf.co.uk/preparation/writing.asp>.
- (4) **Funding:** Was your research for this article funded by a funding agency? If so, please insert ‘This work was supported by <insert the name of the funding agency in full>’, followed by the grant number in square brackets ‘[grant number xxxx]’.
- (5) **Supplemental data and underlying research materials:** Do you wish to include the location of the underlying research materials (e.g. data, samples or models) for your article? If so, please insert this sentence before the reference section: ‘The underlying research materials for this article can be accessed at <full link>/ description of location [author to complete]’. If your article includes supplemental data, the link will also be provided in this paragraph. See <<http://journalauthors.tandf.co.uk/preparation/multimedia.asp>> for further explanation of supplemental data and underlying research materials.
- (6) The **CrossRef database** (www.crossref.org/) has been used to validate the references. Mismatches will have resulted in a query.

QUERY NO.	QUERY DETAILS
AQ1	The disclosure statement has been inserted. Please correct if this is inaccurate.
AQ2	Please provide missing city for Ref. [13].

How to make corrections to your proofs using Adobe Acrobat/Reader

Taylor & Francis offers you a choice of options to help you make corrections to your proofs. Your PDF proof file has been enabled so that you can edit the proof directly using Adobe Acrobat/Reader. This is the simplest and best way for you to ensure that your corrections will be incorporated. If you wish to do this, please follow these instructions:

1. Save the file to your hard disk.
2. Check which version of Adobe Acrobat/Reader you have on your computer. You can do this by clicking on the “Help” tab, and then “About”.

If Adobe Reader is not installed, you can get the latest version free from <http://get.adobe.com/reader/>.

3. If you have Adobe Acrobat/Reader 10 or a later version, click on the “Comment” link at the right-hand side to view the Comments pane.

4. You can then select any text and mark it up for deletion or replacement, or insert new text as needed. Please note that these will clearly be displayed in the Comments pane and secondary annotation is not needed to draw attention to your corrections. If you need to include new sections of text, it is also possible to add a comment to the proofs. To do this, use the Sticky Note tool in the task bar. Please also see our FAQs here: <http://journalauthors.tandf.co.uk/production/index.asp>.

5. Make sure that you save the file when you close the document before uploading it to CATS using the “Upload File” button on the online correction form. If you have more than one file, please zip them together and then upload the zip file.

If you prefer, you can make your corrections using the CATS online correction form.

Troubleshooting

Acrobat help: <http://helpx.adobe.com/acrobat.html>

Reader help: <http://helpx.adobe.com/reader.html>

Please note that full user guides for earlier versions of these programs are available from the Adobe Help pages by clicking on the link “Previous versions” under the “Help and tutorials” heading from the relevant link above. Commenting functionality is available from Adobe Reader 8.0 onwards and from Adobe Acrobat 7.0 onwards.

Firefox users: Firefox’s inbuilt PDF Viewer is set to the default; please see the following for instructions on how to use this and download the PDF to your hard drive: http://support.mozilla.org/en-US/kb/view-pdf-files-firefox-without-downloading-them#w_using-a-pdf-reader-plugin

Handling co-dependence issues in resampling-based variable selection procedures: a simulation study

Riccardo De Bin^{a,b} and Willi Sauerbrei^c

^aDepartment of Medical Informatics, Biometry and Epidemiology, Ludwig-Maximilians-Universität of Munich, Germany; ^bDepartment of Mathematics, University of Oslo, Oslo, Norway; ^cFaculty of Medicine and Medical Center, Institute for Medical Biometry and Statistics, University of Freiburg, Freiburg, Germany

ABSTRACT

If a number of candidate variables are available, variable selection is a key task aiming to identify those candidates which influence the outcome of interest. Methods as backward elimination, forward selection, etc. are often implemented, despite their drawbacks. One of these drawbacks is the instability of their results with respect to small perturbations in the data. To handle this issue, resampling-based procedures have been introduced; using a resampling technique, e.g. bootstrap, these procedures generate several pseudo-samples that are used to compute the inclusion frequency of each variable, i.e. the proportion of pseudo-samples in which the variable is selected. Based on the inclusion frequencies, it is possible to discriminate between relevant and irrelevant variables. These procedures may fail in case of correlated variables. To deal with this issue, two procedures based on 2×2 tables of inclusion frequencies have been developed in the literature. In this paper we analyse the behaviours of these two procedures and the role of their tuning parameters in an extensive simulation study.

ARTICLE HISTORY

Received 2 May 2016
Accepted 8 September 2017

KEYWORDS

Correlation; inclusion frequencies; model building

AMS SUBJECT CLASSIFICATION

62F07; 62F40; 62H17

1. Introduction

Variable selection is a central topic in statistical practice. If the aim is the derivation of a model for explanation, it is important to identify those variables which are related to an outcome of interest, among the many candidate variables usually collected in a study [1]. If the goal is the prediction, the identification of the relevant (related to the outcome) variables is also important, because a reduced number of variables in the model reduces its variability and may make easier (and cheaper) the collection of data to predict the outcome of a future case. An important aspect often not adequately considered in the model-building process is the stability of the variable selection procedure. Given two or more experiments on the same problem, we consider stable a procedure that selects the same set of relevant variables in each case. In other words, a procedure that is not sensitive to small perturbations in the data. The problem of unstable results highly affects high-dimensional data studies (see, e.g. [2]), but it is also relevant in case of low-dimensional data (see, e.g. [3]). The traditional variable selection procedures (backward elimination, forward selection, all-subset

analysis, etc.), indeed, are affected by instability issues, and small modifications of the data may result in different sets of selected variables. Concerning model stability, the ‘strength’ of the effect of the variables is a key issue and the aim of an analysis has an influence whether the effect of a variable is considered as being strong or weak [4].

To investigate this issue and try to obtain more stable results, resampling-based approaches have been presented in the literature [5–7]. These methods use a resampling technique (traditionally bootstrap) to mimic data perturbations and use the concept of inclusion frequency, namely the proportion of ‘perturbed’ datasets in which a variable is selected, to discriminate between relevant and irrelevant variables. These methods have been investigated by Sauerbrei and Schumacher [8], who also suggested two strategies to handle issues related to co-dependence, i.e. situations in which the inclusion frequencies of two (or more) variables depend on each other. One strategy is concerned with the selection of variables with a strong effect only, whereas the other aims to also include variables with a weak effect. Besides the traditional studies, these strategies may be relevant for genetic data applications, in which resampling-based variable selection procedures are lately becoming popular (see, e.g. [2,9,10]). In gene expression data, for example, variables which refer to genes in the same biological pathway are highly correlated, and co-dependence issues are likely to occur. Before thinking to translate Sauerbrei and Schumacher’s strategies [8] to high-dimensional problems, however, it is important to have a clear and complete understanding of the results in the low-dimensional framework. This is the aim of this paper: we investigate in details the performances of the two strategies in an extensive simulation study. We also analyse the role and the sensitivity of the choice of their tuning parameters. We use low-dimensional data and, again for interpretability ease, we consider linear regression. Although some difficulties are known (see, e.g. [11,12]), results can be in general transferred to other regression models that include a linear predictor [13]. Popular models in biostatistics are logistic regression and Cox model for censored survival data.

The paper is structured as follows: in Section 2 we review the resampling-based variable selection procedure and the two strategies introduced by Sauerbrei and Schumacher [8]; Section 3 describes the simulation designs used in the paper; Section 4 contains the simulation results; we show a real example in Section 5; finally a brief discussion is presented in Section 6.

2. Methods

2.1. Variable selection based on the inclusion frequencies

The idea behind a variable selection approach based on variable inclusion frequencies is very simple. As a first step, several (let say, D) pseudo-samples are generated by applying a resampling technique to the original data. In this paper we implement bootstrap, mainly to be able to compare our results with those of Sauerbrei and Schumacher [8], but in this context alternatives such as subsampling may have noticeable advantages (for details, see [14]).

Once these D pseudo-samples are generated, a classical variable selection technique is applied to each of them, obtaining D sets of relevant variables $\mathcal{S}_d$, $d = 1, \dots, D$. In this paper we use backward elimination with a significance level of 0.05 as the variable selection technique, again for the matter of comparison with Sauerbrei and Schumacher [8], but

Table 1. Example of variable inclusion frequencies: for each pseudo-sample, 1 means that the variable is selected, 0 not.

Bootstrap pseudo-sample	variables					
	x_1	x_2	x_3	x_4	x_5	x_6
1	1	1	1	1	0	0
2	1	1	1	1	0	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
249	1	1	1	0	0	0
250	1	1	1	1	1	0
inclusion frequency	1.000	0.976	0.852	0.676	0.340	0.084

in principle any variable selection procedure or other significance levels can be applied. The inclusion frequency for a specific variable x_i is then defined as the proportion of $\mathcal{S}_d$ containing x_i . Ideally, the variables truly related to the outcome have inclusion frequency equal to 1, those with no effect (irrelevant variables) 0 or, by considering the type I error, equal to α , the significance level used as a selection criterion in the backward elimination procedure. Table 1 contains an example in which the inclusion frequencies of 6 variables are computed using $D = 250$ bootstrap pseudo-samples (in this case, the inclusion frequencies are also called ‘bootstrap inclusion frequencies’). We can then use this result to discriminate between relevant and irrelevant variables. The most straightforward procedure, that we denote ‘naive strategy’, consists in using a hard threshold k : the variables with inclusion frequency larger than k are considered relevant, those with inclusion frequency smaller than k are not. In this example, where $x_1 - x_4$ are generated as relevant, $x_5 - x_6$ without effect, by setting $k = 0.6$ we identify the correct variables. Finally, the relevant variables are considered when fitting the final model.

2.2. Co-dependence and effect on the inclusion frequencies

The situation described above, however, rarely happens in practical situations. First of all, some variables may have small effects, which may result in intermediate values for the inclusion frequencies; previous studies (see, e.g. [15]) showed inclusion frequencies between 20 % and 60% for these variables. The choice to include them (or part of them) is arbitrary, and related to k : the smaller its value, the larger the number of selected variables. Therefore, in situations in which only the really strong variables should be included, a large value of k is appropriate; conversely, if variables with weak effect are also of interest, a small value of k should be used (see also [8]).

A more serious issue is related to possible dependencies between the inclusion frequencies of two or more variables (co-dependence). Let us consider, as an example, the case of two highly correlated relevant variables: since the information that they bring is similar, it may happen that in the different pseudo-samples the two variables are alternatively selected. In the worst case, instead of inclusion frequencies equal to 1, they both have an inclusion frequency of 0.5. Obviously, this may cause a severe problem, especially if the goal is to select only variables with high effect (those with inclusion frequencies higher than, let say, 0.7). Despite their high effect, both variables are excluded from the final model, important information is ignored and the fit of the model selected by resampling may result worse than the fit of the individual models (i.e. those without resampling). A different issue may

Table 2. Extreme examples of 2×2 tables obtained in the case of co-dependent variables: two relevant variables (left); one relevant and one irrelevant variable (right).

x_i	x_j		
	in	out	Tot
	in	out	Tot
x_i	9 (0.036)	119 (0.476)	128 (0.512)
	122 (0.488)	0 (0.000)	122 (0.488)
Tot	131 (0.524)	119 (0.476)	250 (1.000)

x_i	x_j		
	in	out	Tot
	in	out	Tot
x_i	125 (0.500)	75 (0.300)	200 (0.800)
	0 (0.000)	50 (0.200)	50 (0.200)
Tot	125 (0.500)	125 (0.500)	250 (1.000)

Note: The relative frequencies are reported between brackets.

arise when one of the correlated variables is related to the outcome and one is not. In some replications, the irrelevant variable may be selected by chance. As a result, its inclusion frequency may be overestimated and it may be incorrectly considered in the final model.

Sauerbrei and Schumacher [8] investigated these issues and note that these kinds of co-dependence can be evaluated by considering 2×2 tables of variable inclusions. Table 2 reports two extreme examples of the aforementioned situations, obtained with $D = 250$ generated pseudo-samples. The left table shows the case of an alternate selection of two relevant variables, while the right table shows the case of a relevant and an irrelevant variable correlated. We can note, in the right table, that the variable x_j is included only when x_i is also included, never alone (conversely to x_i).

Thanks to these 2×2 tables we can detect the cases of co-dependence between two variables, and act accordingly. Obviously, the situations in real studies are much more complicated, and some guidelines are necessary to handle them. Sauerbrei and Schumacher [8] provided two strategies to deal with this issue and, as a consequence, to improve the variable selection based on the inclusion frequencies.

2.3. Strategies

2.3.1. Naive strategy: ignoring co-dependence

We have already described that the most immediate strategy for selecting the relevant variables for a model simply consists in considering separately their inclusion frequencies and retaining those with an inclusion frequency larger than a threshold value k . We called this approach ‘naive strategy’. Note that the co-dependence among variables is ignored. In this paper we mainly use the naive strategy for matter of comparison. Anyway, note that both strategies A and B start from the result of the naive strategy and then try to improve the result.

2.3.1.1. Practical example. Let us denote with h the vector of the inclusion frequencies. For the example reported in Table 1, $h = (1.000, 0.976, 0.852, 0.676, 0.340, 0.084)$. Depending on the threshold k that we want to use, we can select one ($0.976 < k \leq 1$), two ($0.852 < k \leq 0.976$), three ($0.676 < k \leq 0.852$), four ($0.340 < k \leq 0.676$), five ($0.084 < k \leq 0.340$) or all ($0 \leq k \leq 0.084$) variables. Using the values suggested in [8] for selecting only strong ($k = 0.7$) or strong and weak ($k = 0.3$) variables, we would select 3 and 5 variables, respectively. In neither of the two cases, the correct set of relevant variables is selected. For this specific example, the correct set would be selected for $0.340 < k \leq 0.676$.

2.3.2. Strategy A: considering strong and weak effect variables

The idea behind strategy A is to first include all variables which may have an effect on the outcome, and then check, through the 2×2 tables, if some of them have been included for co-dependence reasons. In this case, these variables are later removed. Let us call $h(x_i)$ the inclusion frequency for the variable x_i . Moreover, let $h(x_i|x_j+)$ and $h(x_i|x_j-)$ denote the proportion of times in which x_i is selected given that x_j is selected and x_j is not selected, respectively. Strategy A is defined as a four-step procedure:

- (1) compute the inclusion frequencies for all variables, i.e. $h(x_i)$ for $i = 1, \dots, p$;
- (2) remove the variables with very small inclusion frequencies, i.e. those for which $h(x_i) < k$;
- (3) for the remaining variables, compute the 2×2 tables like those shown in Table 2 and perform a test of independence (0.05 significance level as default). Those variables for which all tests do not reject the hypothesis of independence, are kept;
- (4) For each pair x_i and x_j judged dependent, consider the following conditions:
 - (i) $h(x_i) > c_1 h(x_i|x_j+)$;
 - (ii) $h(x_i|x_j-) > c_1 h(x_i)$ and $h(x_i) < c_2 k$;
 where c_1 and c_2 are tuning parameters, with default values $c_1 = c_2 = 1.4$ [8]. The variable x_i is removed if one of the two conditions holds. If both x_i and x_j satisfy one of the two conditions, only that with a smaller inclusion frequency is removed.

Please note that the authors recommend this strategy for studies in which weak effect variables should be selected; for this reason, in step 2 the parameter k should assume a small value (suggested 0.3).

2.3.2.1. Practical example. Using again the example reported in Table 1, we see how strategy A works in practice:

- (1) the vector of inclusion frequencies is $h = (1.000, 0.976, 0.852, 0.676, 0.340, 0.084)$;
- (2) using $k=0.3$, in the second step we remove x_6 , because $h(x_6) = 0.084 < 0.3$;
- (3) since x_1 is always included, it does not need to be considered in this step. For all pairs involving the other variables, we compute the 2×2 tables and perform a test on the independence (with 0.05 significance level). The inclusion frequencies of x_4 and x_5 (see below) are strongly dependent ($p\text{-value} = 1 \times 10^{-61}$), while for all other pairs there is no evidence against independence. In particular, the 2×2 table for the inclusion frequencies of x_4 and x_5 has form (relative frequencies between brackets)

		x_5		
		in	out	Tot
x_4	in	4 (0.016)	165 (0.660)	169 (0.676)
	out	81 (0.324)	0 (0.000)	81 (0.324)
Tot		85 (0.340)	165 (0.660)	250 (1.000)

- (4) both x_4 and x_5 satisfy condition (i): $0.676 = h(x_4) > c_1 h(x_4|x_5+) = 1.4 \times 0.047 = 0.066$ and $0.324 = h(x_5) > c_1 h(x_5|x_4+) = 1.4 \times 0.024 = 0.033$. In this case, x_4 has the higher inclusion frequency and therefore is kept in the model, x_5 is removed.

Please note that by using the naive strategy we would have incorrectly included the variable x_5 in the model, while applying strategy A we correctly identify $x_1 - x_4$ as the only relevant variables.

2.3.3. Strategy B: considering only strong effect variables

Strategy B follows a different idea to cope with co-dependent variables: in a first step, only variables with high inclusion frequency (i.e. $h(x_i) > k$, with, e.g. $k = 0.7$), i.e. those variables selected in almost all pseudo-samples, are included in the set of relevant variables. The following step looks for possible cases of co-dependence, with the aim of adding to the set of relevant variables those variables whose inclusion frequencies were smaller than the threshold k because co-dependence issues. The procedure is defined as

- (1) compute the inclusion frequencies for all variables, i.e. $h(x_i)$ for $i = 1, \dots, p$;
- (2) keep the variables with high inclusion frequencies, i.e. those for which $h(x_i) > k$;
- (3) for each pair of remaining variables, consider their pairwise joint inclusion frequency $h(x_i \vee x_j)$: if it is larger than a tuning parameter c_3 (default 0.9 [8]), include the variable with a higher inclusion frequency.

This strategy is recommended to identify only factors with high effect on the outcome [8].

2.3.3.1. Practical example. Let us see how strategy B works in practice in the same example considered before:

- (1) we have the same $h = (1.000, 0.976, 0.852, 0.676, 0.340, 0.084)$;
- (2) using the recommended $k = 0.7$, we note that the first three variables are directly defined as relevant $h(x_i) > k, i = 1, 2, 3$;
- (3) for the remaining variables, $h(x_4 \vee x_5) = 1.000, h(x_4 \vee x_6) = 0.712$ and $h(x_5 \vee x_6) = 0.388$; using the default value $c_3 = 0.9$, only the two former variables are recognized as co-dependent, and variable x_4 is included in the set of relevant variables due to its inclusion frequency higher than that of x_5 .

Again, the application of a strategy under consideration (in this case, B) improved the results of the naive strategy, that in this case would have missed to include the relevant variable x_4 .

2.4. Prediction ability

From the previous section, it is clear that strategy A is thought to derive a prediction model rather than a model for explanation. In contrast, the main aim of strategy B is to select the strong effect variables, which typically leads to a sparser model with a slightly inferior

prediction ability than the one selected by strategy A. This inferior prediction ability is due to the increase in the bias component of the prediction error. Due to the well-known bias/variance trade off, however, using a small number of variables may reduce the variance component of the prediction error, and therefore a sparser model may be preferred even if the bias is slightly increased. A sparser model also has severe benefits as far as generalizability, transportability and usefulness are concerned [16]. Note that both strategies, ‘play’ with this bias/variance trade-off in handling co-dependent variables. Often two variables are co-dependent because they provide very similar information. In this case, from a prediction point of view, one is redundant and strategies A and B select only the other. From a prediction point of view, indeed, it is more important to select a model with strong prediction ability rather than the ‘true’ one [1]. For a more detailed discussion about the differences between prediction and explanatory models, we refer the reader to Section 2.4 of Royston and Sauerbrei [4] and more specifically to the aforementioned paper by Shmueli [1]. The latter author nicely illustrates that these phrases are used differently among disciplines and illustrates relevant distinctions and practical implications for explanatory and predictive modelling. For example, in the social sciences the term explanatory model is used nearly exclusively for testing causal theory. Shmueli would probably prefer the term ‘descriptive modelling’ for an approach like strategy B.

Here we consider both variable selection and the prediction ability of the final model to contrast strategies A and B with the naïve approach. For this reason, in our simulation study we compute the ratios of true positive and of false negative, and the mean square prediction error of the models obtained with the different strategies. The mean square prediction error is computed on a large ‘test’ set, which is obtained by generating new data from the same design used for the training pseudo-data (see Section 3). We use this test set to compute the mean square prediction error of the final models identified in the training data by strategies A and B, as well as those obtained with the naïve strategy with $k = 0.3$ and $k = 0.7$.

3. Simulation designs

In order to evaluate the performance of the strategies introduced by Sauerbrei and Schumacher [8], we perform an extensive simulation study. We proceed as follows. We create 100,000 datasets, for each of which we generate D bootstrap samples. Consider the single dataset. To each bootstrap sample generated from it, we apply a backward elimination procedure, obtaining a table like Table 1, from which we compute the bootstrap inclusion frequencies for all variables. Based on the bootstrap inclusion frequencies, we select which variables are considered relevant using both the naïve strategy and strategies A and B. We repeat this procedure for each dataset and we summarize the results by computing the variable selection frequencies, i.e. we compute the proportion of times (out of the 100,000) in which the variables are identified as relevant by the different strategies. The variable selection frequency must not be confused with the bootstrap inclusion frequency: the former is the result of the simulation study, the latter is the key concept on which a resampling-based variable selection method is made. From these variable selection frequencies, we also compute the percentage of true positive (the proportion of datasets in which relevant variables are correctly identified – separately for strong and weak effect variables) and false negative (for irrelevant variables).

3.1. Main setting

We consider a multivariable linear regression example with 14 explanatory variables $x_1, \dots, x_{14}$ and a response variable with a Gaussian distribution with mean $\mu = 2.5 + x_1 + x_2 + x_3 + 0.5x_4 + 0.5x_5 + 0.5x_6 + 0.5x_7 + 0.5x_8$ and standard deviation 2: we have, therefore, 6 variables with no effect ($x_9 - x_{14}$) and 8 with effect, which can be separated in turn between 3 variables with strong effect ($x_1 - x_3$) and 5 variables with weak effect ($x_4 - x_8$). The explanatory variables have a multivariate Gaussian distribution with 0 mean and standard deviation 1 for all dimensions. To investigate the co-dependence issues, we consider a correlation $\rho = 0.85$ between some pairs of variables. Please note that the correlation value is relatively large but not uncommon in practice (see the real example of Section 5). This large value allows us to empathize the co-dependence effect. The pairs of correlated variables are the following:

- x_2 and x_9 , to investigate the co-dependence between strong and no effect variables;
- x_3 and x_8 , for the case of strong and weak effect correlated variables;
- x_6 and x_7 , two weak effect correlated variables;
- x_5 and x_{10} , in which a weak and a no effect variable are correlated;
- x_{11} and x_{12} , to see the effect of co-dependence between two no effect variables.

Please note that x_1, x_4, x_{13} and x_{14} are independent to all other variables. In particular, the former two have an effect on the outcome (strong and weak, respectively), while the latter two do not. This design adapts to our specific needs those of [van Houwelingen and Sauerbrei \[17\]](#) and [De Bin et al. \[14\]](#).

We perform the study using a sample size n equal to 150, and, in each iteration, computing the bootstrap inclusion frequencies based on $D = 250$ pseudo-samples. The explained variation R^2 computed for the full model is around 0.60 for all pseudo-datasets. For the implementation of the strategies, we use the recommended values of the parameters ($c_1 = 1.4, c_2 = 1.4, c_3 = 0.9$ [8]), as well as for the thresholds ($k = 0.3$ for strategy A, $k = 0.7$ for strategy B). In particular, these thresholds are also used for the naive strategy for the matter of comparison. Finally, we use $\alpha = 0.05$ as the significance level for the backward elimination performed in each bootstrap pseudo-sample.

As a further investigation, we use this setting to evaluate the effects of the parameters c_1, c_2, c_3, k . We change their values in turn, keeping the default values for all others ($c_1 = c_2 = 1.4, c_3 = 0.9, k = 0.3 \vee k = 0.7$), and contrast the resulting models in terms of mean square prediction error and average number of selected variables.

3.2. Alternative setting

3.2.1. Triplets of correlated variables

The main simulation setting allows us to easily understand the effect of strategies A and B. Nevertheless, its simple structure represents the most [favourable](#) case for the two strategies, because they have been explicitly designed to deal with pairs of co-dependent variables. To have a more challenging situation, we also considered a more complex correlation structure, in which the correlation involves triplets of variables. In this case, the correlation involves:

- strong/weak/weak effect variables, namely x_1 , x_6 and x_7 ;
- strong/no/no effect variables, namely x_2 , x_9 and x_{12} ;
- strong/weak/no effect variables, namely x_3 , x_8 and x_{11} ;
- weak/weak/no effect variables, namely x_4 , x_7 and x_{10} .

The explanatory and response data are generated as in the main setting design, using again for all marginal correlations (i.e. correlation between two variables of the triplets) the quite extreme but realistic value $\rho = 0.85$. Please note that the variables x_{13} and x_{14} are independent.

3.2.2. Smaller correlation value

Despite its plausibility, a correlation equal to 0.85 is not extremely common in practice, and, when it is present, rarely involves more than two variables. While this large value allows us to highlight the effect of strategies A and B in the previous simulation settings, one may ask whether the two strategies have any effect in situations with less correlation among variables. In order to show that strategies A and B affect the choice of the relevant variables in a larger number of cases, we replicate the analysis also using more moderate values for ρ , namely 0.7 and 0.5.

4. Results of the simulation study

4.1. Strategies' performances

4.1.1. Variable selection frequencies

Figure 1 shows the results of the simulation study using the main setting. Results of the naive strategy illustrate the effect of the choice of k , which is 0.3 (left) and 0.7 (right). Incorrect inclusion of variables without an effect ($x_9 - x_{14}$) hardly happens for $k = 0.7$, while it is much more likely for $k = 0.3$. Compared to the naive strategy with the corresponding choice of k , strategy B slightly increases the incorrect inclusions of these variables, whereas strategy A reduces them. In particular, this is true for the selection frequencies of x_9 and x_{10} , which are sometimes incorrectly selected by the naive strategy due to their correlation with x_2 (strong effect) and x_5 (weak effect), respectively. Both cases are examples of the situation described in the right side of Table 2. Similar relations can be seen for variables with an effect, which implies that strategy A incorrectly excludes variables more often than the naive strategy ($k = 0.3$) and strategy B includes relevant variables more often than the corresponding naive strategy ($k = 0.7$).

Please note the behaviours of strategies A and B for x_6 and x_7 . These two variables have the same (weak) effect and they are strongly correlated. The effect of strategies A and B is to move the selection frequencies from the incorrectly low (naive strategy with $k = 0.7$) or high (naive strategy with $k = 0.3$) value to a more reasonable 0.5. This means, in particular, that in the models selected by the two strategies at least one of the two variables is selected. Since x_6 and x_7 provide similar information, from a predictive point of view (see next section) it is not necessary that both are included in the model. Table 3 shows the 2×2 table obtained for the first iteration. Note that both inclusion frequencies (0.512 and 0.468) are larger than 0.3 and smaller than 0.7, therefore the naive strategy either include

Figure 1. Variable selection frequencies for the main setting design. The left graphic compares the strategy A and the naive strategy with $k = 0.3$, the right graphic the strategy B and the naive strategy with $k = 0.7$. The variables x_1, x_4, x_{13} and x_{14} are uncorrelated.

Table 3. The 2×2 tables for variables x_6 and x_7 (weak effect correlated variables) obtained in the first iteration.

		x_7		
		in	out	Tot
x_6	in	5 (0.020)	128 (0.512)	133 (0.532)
	out	117 (0.468)	0 (0.000)	117 (0.468)
	Tot	122 (0.488)	128 (0.512)	250 (1.000)

Note: The relative frequencies are reported between brackets.

($k = 0.3$) or remove ($k = 0.7$) both variables. Strategies A and B select the most relevant of the two, in this case x_6 .

4.1.2. True positive and false negative

Table 4 shows the results in terms of true positive and false negative for strategy A, strategy B and the naive strategy (with $k = 0.3$ and $k = 0.7$). Regarding the true positive, the results are given separately for relevant variables with strong and weak effects. The percentages are computed through the 100,000 iterations and are based on the selection frequencies of 3 (true positive, strong effect), 5 (true positive, weak effect) and 6 (false negative) variables. Note that the percentage of true positive can be interpreted as the power of the procedure, while 1 minus the percentage of true negative as the type I error.

For strategy A, we note an undesired decrease in terms of true positive for strategy A, especially for weak effect variables, which is hardly compensated by the small improve in terms of false negative. This is not unexpected, because strategy A focuses on prediction rather than explanation. On the contrary, we note a good performance for strategy B. There

Table 4. Percentage of true positive and false negative for models derived with strategy A, strategy B and naive strategy ($k = 0.3$ and $k = 0.7$).

	True positive		True negative
	Strong effect	Weak effect	
Strategy A	0.95	0.64	0.81
Naive strategy ($k = 0.3$)	0.98	0.76	0.76
Strategy B	0.92	0.49	0.96
Naive strategy ($k = 0.7$)	0.87	0.38	0.97

Notes: The percentage of true positive are computed separately for the variables with strong effect (3 variables, namely $x_1 - x_3$) and weak effect (5 variables, namely $x_4 - x_8$). The 6 variables without effect on which the percentage of false negative was computed are $x_9 - x_{14}$.

Table 5. Average results in terms of mean square prediction error (MSPE) and average number of variables for strategies A and B and for the naive strategy with the corresponding thresholds ($k = 0.3$ and $k = 0.7$, respectively).

	Strategy			
	A	Naive ($k = 0.3$)	B	Naive ($k = 0.7$)
Average MSPE	4.47	4.42	4.57	5.23
Average # variables	7.16	8.20	5.44	4.68

is a sensible improvement in terms of true positive for both strong and weak effect variables, while the loss in terms of false negative is negligible (around 0.01). Note that the true negative is computed based on 6 (out of which 2 are uncorrelated) variables while the true positive for weak effect variables is based on 5. Computations based only on the inclusions of the 4 irrelevant correlated variables ($x_9 - x_{12}$), however, do not show substantial differences with the results reported in Table 4.

4.1.3. Prediction ability

We also contrast strategies A and B with the naive approach in terms of prediction error. In order to mimic a test set, we generated 100,000 observations from the same design used to simulate our previous pseudo-data. On this test set, we compute the mean square prediction error of the final models identified by strategies A and B in the 100,000 iterations, as well as those identified by the naive strategy. The results are summarized in Table 5, where we also reported the average number of variables selected in the models.

In terms of mean square prediction error, we obtain an average value of 4.47 for strategy A, versus an average 4.42 using a naive strategy with the same threshold $k = 0.3$ (see Table 5). The exclusion of some relevant variables from the models, which we have seen (left graphic of Figure 1) leading to lower selection frequencies, seems to negatively influence the prediction performances of the models identified with strategy A. Nevertheless, it is worth noting that the average number of variables included in the models selected by strategy A is noticeably smaller than those selected with the naive approach (7.16 versus 8.20). In particular, this means that by applying strategy A, we obtain a model that is sparser (with the obvious advantage in interpretability and manageability) than that obtained with a naive strategy, at the cost of a small decrease in prediction ability. In other words, some variables that do not add too much information to the model due to their co-dependence with other selected variables are discarded, even if there is a small loss of information.

In some sense, it is the same compromise that we make when we remove a variable from the model in a classical backward elimination procedure.

Conversely, the prediction ability of the models obtained by implementing strategy B is in mean better than that of the models obtained with the naive strategy with the corresponding threshold ($k = 0.7$). The average mean square error for strategy B is 4.57, much smaller than the 5.23 obtained for the naive strategy. The average number of variables in the models selected by strategy B is obviously larger than that for the models obtained with the naive strategy (5.44 versus 4.68), but in this case it seems reasonable to include more than 4.68 variables in a model.

4.2. Influence of the parameters

4.2.1. Parameter c_1

We analyse now the effect of parameter c_1 , which influences the inclusion/exclusion of co-dependent variables in strategy A. In particular, we repeat the simulations performed above for different values of c_1 , namely for a grid of values between 0.2 and 4.1, with an interval of 0.3 between two points. All other parameters are kept fixed with their default values ($c_2 = 1.4$, $k = 0.3$, $n = 150$, $\alpha = 0.05$, $B = 250$). The results are reported in Figures 2 and 3. See also Table A1 in the supplementary material.

The former plot shows the selection frequencies for all variables. We note a general decrease of the effect of strategy A as c_1 increases, with the variable selection frequencies getting closer to those obtained with the naive strategy. Interestingly, the positive effect of strategy A on the selection frequencies of the irrelevant variables correlated with the relevant ones (i.e. of x_9 and x_{10}) is still evident for large values of c_1 , while the effect on the other irrelevant variables ($x_{10} - x_{14}$) disappears from $c_1 > 2$. Unfortunately, the negative effect of decreasing the selection frequencies of correlated relevant variables (see, e.g. x_5) also remains despite the increase of c_1 .

The decrease of the effect of the strategy as c_1 increases can also be seen in terms of mean square prediction error and average number of selected variables, which, in this case, means smaller prediction error and less sparse model (Figure 3). The improvement in terms of the prediction ability of the model, in particular, is very moderate, especially around the default value $c_1 = 1.4$ (see values between 1.1 and 2.3).

Single iteration for illustration. Let us see the effect of the parameter c_1 in an exemplary iteration. Table 6 reports the 2×2 tables obtained for the inclusions of x_3 and x_8 (left) and for x_5 and x_{10} (right). In particular, x_3 is a variable with strong effect, x_5 and x_8 with weak effect and x_{10} with no effect. In the first step of strategy A, all four variables are kept in the model, being their inclusion frequencies larger than 0.3. Both tables reveal the co-dependence between the variables when we test the independence with a chi-square test (p-values equal to 0.007 and 8×10^{-8} , respectively). Therefore, these four variables are considered by strategy A as possible redundant variables, because their high inclusion frequencies may be an effect of co-dependence.

In the first case, both x_3 and x_8 are relevant variables (their true regression coefficients are 1 and 0.5, respectively), and should be kept in the model. Let us see how they are treated by strategy A, in particular in step 4 of the algorithm (see Section 2.3.2). In this step, we are supposed to exclude a variable if one of conditions (i) and (ii) holds. In this case, the latter is never met because, using the standard values of c_2 and k , the inclusion frequencies of x_3 and

Figure 2. Selection frequencies of the 14 variables obtained by using strategy A with some selected values of c_1 . The results are compared with the selection frequencies for the naive strategy with $k = 0.3$ (black bars). The variables x_1 , x_4 , x_{13} and x_{14} are uncorrelated.

Table 6. Examples of 2×2 tables obtained in one selected iteration of our simulation study: in the left table, x_3 (strong effect) versus x_8 (weak effect); in the right table, x_5 (weak effect) versus x_{10} (no effect).

x_8				x_{10}					
x_3		in	out	Tot	x_5		in	out	Tot
	in	156	75	231		in	77	119	196
		(0.624)	(0.300)	(0.924)			(0.308)	(0.476)	(0.784)
	out	19	0	19		out	0	54	54
		(0.076)	(0.000)	(0.076)			(0.000)	(0.216)	(0.216)
Tot		175	75	250	Tot		77	173	250
		(0.700)	(0.300)	(1.000)			(0.308)	(0.692)	(1.000)

Notes: The two pairs of variables are highly correlated between each other ($\rho = 0.817$ and $\rho = 0.833$, respectively). The relative frequencies are reported between brackets.

x_8 are higher than 0.42. With regards to the former, instead, $1.036 = h(x_3)/h(x_3|x_8+) = h(x_8)/h(x_8|x_3+) > c_1$. This means that for small values of c_1 (in this case, smaller than 1.036) we would remove x_8 (the variable with smaller inclusion frequency). A large value for c_1 saves more variables from the elimination due to co-dependence.

Figure 3. Relative variations for strategy A with respect to the naive strategy in terms of mean square prediction error (left) and average number of variables (right) when varying c_1 . The dotted vertical line highlights the default value (1.4) in [8].

On the other hand, if we increase the value of c_1 too much, we risk including variables whose inclusion frequencies are high only as a result of the correlation with relevant variables. Let us consider x_5 and x_{10} . In this case the latter variable is irrelevant and it may be advantageous if it does not enter in the model. Again, condition (i) holds when $0.784 = h(x_5)/h(x_{10}|x_5+) = h(x_{10})/h(x_5|x_{10}+) > c_1$. If c_1 is too large, the condition for x_{10} is not met and x_{10} is not removed from the model.

As a conclusion, we note that c_1 regulates the effect of strategy A, which is stronger for smaller values of this parameter. In particular, smaller values mean the exclusion of more variables due to co-dependence. The default value 1.4 seems a good compromise between no effect (too many variables, scarce improvement in prediction ability) and overeffect (very sparse model with small prediction ability).

4.2.2. Parameter c_2

We repeated the analysis for the parameter c_2 , also involved in the inclusion/exclusion of co-dependent variables performed in step 4 of strategy A. Again, we consider values between 0.2 and 4.1, spaced by 0.3, and all the other parameters with default values. As we can see from Figure 4, the effect of this parameter is smaller than that of c_1 . In contrast to the previous case, however, larger values for the parameter c_2 lead to stronger effect of strategy A. Interestingly, in our simulation the effect involves only part of the variables; we note that x_2 , x_6 and x_7 have the same selection frequencies for all the considered values of c_2 . For the complete results, see Table A2 in the [supplementary material](#). Figure 5, moreover, shows the relatively small effect of c_2 on the prediction error and on the average number of variables in the model. Note that the suggested $c_2 = 1.4$ corresponds to the last value before the increase of the prediction error (left graphic), that, anyway, has a good effect on the sparsity of the model (right graphic).

Single iteration for illustration. Consider again the exemplary iteration seen before and the co-dependence between 2×2 tables of Table 6. We saw in Figure 5 that the parameter

Figure 4. Selection frequencies of the 14 variables obtained by using strategy A with some selected values of c_2 . The results are compared with the selection frequencies for the naive strategy with $k = 0.3$ (black bars). The variables x_1 , x_4 , x_{13} and x_{14} are uncorrelated.

c_2 does not have a strong effect. Nevertheless, for small values of this parameter, we have an increase in the inclusion of the variables. Condition (ii) of strategy A, indeed, holds when $h(x_i|x_j-) > c_1 h(x_i)$ (which does not depend on c_2) and $h(x_i) < c_2 k$. Let us consider x_8 . The condition $2.333 = h(x_8)/0.3 < c_2$ holds only for large values of c_2 , therefore, this variable is not a candidate to be removed. The smaller c_2 , the less variables excluded. Obviously, this is working for the noise variables as well. For example, $1.027 = h(x_{10})/0.3 < c_2$, and for a value of c_2 smaller than 1.027 the variable x_{10} would be incorrectly included.

It seems that c_2 acts in order to remove those variables whose inclusion frequencies are too small. In this sense, it seems to act like the parameter k . The smaller the value, the more variables included. Nevertheless, it is worth noting that the condition involving c_2 is only one term of the complete condition (ii) of strategy A. We can also see the role of c_2 as a control, in order to not exclude variables which may meet the other condition, $h(x_i|x_j-) > c_1 h(x_i)$, but have too high inclusion frequencies to be removed.

Figure 5. Relative variations for strategy A with respect to the naive strategy in terms of mean square prediction error (left) and average number of variables (right) when varying c_2 . The dotted vertical line highlights the default value (1.4) in Sauerbrei and Schumacher [8].

4.2.3. Parameter c_3

Strategy B, instead, relies on the parameter c_3 . We evaluate its possible values between 0.30 and 0.95 with intervals of 0.05. The results are reported in Figure 6 and show the importance of this parameter: the larger its value, the larger the effect of strategy B. Unfortunately, this effect involves both the positive (larger selection frequency for the relevant variable) and the negative (larger inclusion frequency for the variables without effect on the outcome) aspects. Table A3 in the Supplementary Material reports all results. Since the goal of strategy B is to select only the strong effect variables, a large value for c_3 , as the 0.9 recommended by Sauerbrei and Schumacher [8], seems to be adequate. This consideration is supported by Figure 7. We note a large effect on the sparsity of the model (right graphic) associated with a moderate increase of the prediction error, which is much smaller than that obtained with the naive strategy. This is likely an effect of the inability of the naive strategy to select the low-effect correlated variables x_6 and x_7 , as can be seen in Figure 6 by comparing the selection frequencies obtained with the naive strategy and with strategy B (also with the extreme value $c_3 = 0.95$).

Please note that, if we slightly modify strategy B in order to include only one variable in the case of multiple co-dependencies (i.e. if, let say, x_i , x_j and x_z are all co-dependent, instead of selecting the two with larger inclusion frequencies, as the strategy B does, we select only one of them), we would obtain a minimum for variable selection and average number of selected variables at $c_3 = 0.85$, very close to the suggested 0.9. We will go back to the issue of multiple co-dependencies in Section 6.

Single iteration for illustration. To show the effect of the parameter c_3 , let us consider two different pairs of variables in an exemplary iteration, namely x_6 versus x_7 and x_{12} versus x_{13} . The former two variables have both weak effect and they are correlated to each other. The latter two have no effect and are not correlated neither with each other nor with any relevant variable. Their inclusion frequencies are 0.592, 0.352, 0.476 and 0.452, respectively. No one

Figure 6. Selection frequencies of the 14 variables obtained by using strategy B with some selected values of c_3 . The results are compared with the selection frequencies for the naive strategy with $k = 0.3$ (black bars). The variables x_1 , x_4 , x_{13} and x_{14} are uncorrelated.

of them enters in the first selection operated by step 2 of strategy B (see Section 2.3.3). Therefore, step 3 considers them in order to find possible co-dependencies.

Table 7 shows the 2×2 tables. We can note in the left table that x_6 and x_7 are in a typical situation of alternate selection (co-dependence). In some iterations (146) x_6 is selected while x_7 is not, and in some (86) it is the other way around. In very few cases, 2 and 16, the variables are both selected and both excluded, respectively. Step 3 of strategy B considers the joint selection. In this case, one of the two variables is included if $0.936 = h(x_6 \vee x_7) > c_3$.

Therefore a small value for c_3 (in this case smaller than 0.936) is preferable, because it allows one of the co-dependent variables to enter in the model. On the other hand, we need a sufficiently large value for c_3 to avoid to include irrelevant variables by chance. Let us consider the second pair of variables. Both do not have any effect on the outcome, but one of them would enter in the model if $0.7 = h(x_{12} \vee x_{13}) > c_3$.

Once again, the value for the parameter should be chosen as a compromise between a too restrictive rule (c_3 too large) which reduces the selection of relevant variables affected

Figure 7. Relative variations for strategy B with respect to the naive strategy in terms of mean square prediction error (left) and average number of variables (right) when varying c_3 . The dotted vertical line highlights the default value (0.9) in Sauerbrei and Schumacher [8].

Table 7. Examples of 2×2 tables obtained in one iteration of our simulation study: in the left table, x_6 (weak effect) versus x_7 (weak effect); (b) x_{12} (no effect) versus x_{13} (no effect).

x_7				x_{13}					
x_6		in	out	Tot	x_{12}		in	out	Tot
	in	2	146	148		in	57	62	119
		(0.008)	(0.584)	(0.592)			(0.228)	(0.248)	(0.476)
	out	86	16	102		out	56	75	131
		(0.344)	(0.064)	(0.408)			(0.224)	(0.300)	(0.524)
Tot		88	162	250	Tot		113	137	250
		(0.352)	(0.648)	(1.000)			(0.452)	(0.548)	(1.000)

Note: The relative frequencies are reported between brackets.

by co-dependence issues, and a too permissive alternative, which allows irrelevant variables to enter into the model.

4.2.4. Parameter k in strategy A

Figure 8 shows the performance of strategy A with respect to the naive approach for three different values of the parameter k . The choice of this parameter is related to the specific problem, with smaller k leading to the selection of more (and weaker) explanatory variables. Here we want to show that strategy A is effective in removing those variables which do not provide much information due their co-dependence with other variables. The effect is larger for smaller thresholds k : in the top-left graphic of Figure 8 we note a larger decrease of the selection frequencies for all variables, which does not correspond to any substantial increase of the prediction error (see left graphic of Figure 9, in particular, the point 0.1). Before focusing on the prediction error, we want to mention that the effect of strategy A disappears for values of k larger than 0.5 (the selection frequencies for strategy A and the naive strategy are identical for $k = 0.6, 0.7, 0.8$ and 0.9 , graphics not shown). The absence of this effect can be seen in Figure 9: there, we compare strategy A and the naive strategy in

Figure 8. Selection frequencies of the 14 variables obtained by using strategy A (black) and the naive strategy (white) with three selected values of k . The variables x_1, x_4, x_{13} and x_{14} are uncorrelated.

terms of prediction error (left) and number of selected variables (right). The points corresponding to larger values of k lie on the bisecting line, i.e. there is no difference in terms of mean square prediction error and average number of selected variables between strategy A and the naive strategy. For smaller values of k , instead, there is a large save in number of variables (almost 3 variables, see the right graphic of Figure 9), without any noticeable loss in the predictive ability (left graphic).

It is not worth investigating the effect of k in a single iteration, since its effect is pretty obvious. Smaller values of k imply more variables to consider in the third and fourth steps of strategy A. More variables imply more cases in which strategy A has an effect and, therefore, the differences between applying and not applying the strategy is larger. When the threshold k is too high ($k > 0.5$), only variables without co-dependence are selected, and as a consequence the implementation of strategy A is pointless. Note that the case $k = 0.1$ is shown for illustration but does not make sense in practice.

Figure 9. Relative variations for strategy A with respect to the naive strategy in terms of mean square prediction error (left) and average number of variables (right) when varying k . The symbol 'X' highlights the default value (0.3) in Sauerbrei and Schumacher [8].

4.2.5. Parameter k in strategy B

Similarly to strategy A, strategy B has different effects for different values of k , with the difference that the latter strategy increases its effect for larger values of the parameter. For larger values of k , indeed, the naive strategy is not able to select those variables with inclusion frequencies smaller than the threshold as effect of co-dependence. In fact, we note the largest effect of strategy B when a threshold equal to 0.9 is implemented (bottom graphic of Figure 10). Also for strategy B, 0.5 is a sort of limit that separates the cases in which the strategy works (here, on the opposite of strategy A, for $k > 0.5$) and those cases in which there is no improvement ($k \leq 0.5$, graphics not shown).

Figure 11 shows the large improvement in terms of prediction ability of a model selected with strategy B with respect to one obtained with the naive strategy (left graphic). As noted above, strategy B has a strong effect for large values of k , and no effect for values smaller than 0.5 (values on the bisecting line). The same considerations work in terms of average number of variables in the model (Figure 11, right graphic). For large values of k (left tail) strategy B is able to select the relevant variables excluded by the naive strategy.

For strategy B as well we do not need to investigate the single case to evaluate how k affects the result. For larger values of the parameter, indeed, fewer variables are selected in the first two steps of strategy B (which correspond to the naive strategy), and more variables need to be checked in step 3, in order to recognize whether they have been excluded due to co-dependence. On the contrary, if we use a value of k ($k < 0.5$) too small, all the relevant variables are already selected in the first two steps (i.e. with the naive strategy), and the application of strategy B is pointless.

4.2.6. Influence of further quantities

For the matter of completeness, we also investigated the effects of the sample size n , the selection criterion α and the number of bootstrap replications D . The graphics are reported in the [supplementary material](#) (Figures A1–A4 for n , Figures A5–A8 for α and Figures

Figure 10. Selection frequencies of the 14 variables obtained by using strategy B (black) and the naive strategy (white) with three selected values of k . The variables x_1 , x_4 , x_{13} and x_{14} are uncorrelated.

A9–A.12 for D). The first four graphics show that strategies A and B are more effective for studies in which the sample size is moderate. Obviously, the larger the sample size, the more stable the results, and the issues related to co-dependence less influential. Conversely, the effects of the two strategies do not seem to change with respect to the choice of the selection criterion (Figures A5–A7). Only for strategy B, we have a moderate effect on the prediction error (Figure A8, left graphic) and a non-negligible effect on the average number of variables (Figure A8, right graphic) for all the investigated values ($\alpha = 0.01, 0.05, 0.10$ and 0.157 , the latter related to the Akaike information criterion [18]).

Neither strategy A nor strategy B seem to be highly influenced by the choice of the number of bootstrap replications. This result was expected for strategy B, but not for strategy A. The tests used to select the co-dependent variables implemented in step 3 of the latter strategy, indeed, depends on the number of iterations performed (more iterations, more statistical power, more variables maybe eliminated, see Sauerbrei and Schumacher [8],

Figure 11. Relative variations for strategy B with respect to the naive strategy in terms of mean square prediction error (left) and average number of variables (right) when varying k . The symbol 'X' highlights the default value (0.7) in [8].

[Remark 4 in Section 'The strategy']. From this simulation study, we note that it is not necessary to generate a very large number (already 200 may be sufficient) of pseudo-samples to obtain a satisfactory result.

4.3. Alternative settings

4.3.1. Triplets of correlated variables

Despite the fact that the correlation involves triplets of variables rather than pairs of them, in this alternative scenario strategies A and B perform similarly as in the main design simulation case, both in terms of selection frequencies (Figure A13 in the [supplementary material](#)) and mean square prediction error/average number of variables in the models (Table A4 in the [supplementary material](#)). The larger difficulties due to the more complex correlation structure lead to worse results for all strategies, but the relative performances of strategies A and B with respect to the naive strategy do not change. Strategy A decreases the selection frequencies of all variables (Figure A13, left graphic), leading to sparser models than the naive strategy (6.155 versus 7.370 variables in mean, see Table A4) which nevertheless have a very small decrease in mean square prediction error (4.498 versus 4.431).

As we also saw in the main design, strategy B increases the selection frequencies of all variables (Figure A13, right graphic), which results in a less sparse model (in mean, 4.060 versus 3.431 variables, see Table A4). As for the main design, the improvement in mean square prediction error seems to justify this small increase in model complexity. In this case, the mean square prediction error decreases from 5.693 (naive strategy) to 5.000 (strategy B). This means that, despite the fact that they are specific for pairwise co-dependencies, strategies A and B seem to be also able to deal with more complex situations.

4.3.2. Smaller correlation value

Figures A14–A15 and Tables A5–A6 report the results obtained in both settings (pairs and triplets of correlated variables) with correlation value $\rho = 0.7$. Although to a smaller extent, the results are very similar in both cases to those obtained with $\rho = 0.85$, which means that it is not necessary to have a really large correlation to appreciate the effect of strategies A and B. For the case $\rho = 0.5$, the specific effects of strategy A and B to remove or add variables is severely reduced (Figures A16–A17 and Tables A7–A8).

We note the same result in terms of true positive and false negative. In this case as well, the effects of strategies A and B are still evident, although to a smaller extent, for $\rho = 0.7$ (Table A9), while they are barely visible for $\rho = 0.5$ (Table A10). Same considerations are valid for the simulation setting with triplets of correlated variables (Tables A.11–A.13).

5. Body fat example

5.1. Data

As a real example, we analyse the body fat dataset introduced by Johnson in his 1996 paper [19]. It contains information on the percentage of body fat (*PBF*, continuous variable) of 252 men and on 13 variables with potential influence on *PBF*, namely age (x_1), weight (x_2), height (x_3), and 10 body circumference measurements (neck circumference (x_4), chest circumference (x_5), abdomen circumference ‘at the umbilicus and level with the iliac crest’ (x_6), hip circumference (x_7), thigh circumference (x_8), knee circumference (x_9), ankle circumference (x_{10}), extended biceps circumference (x_{11}), forearm circumference (x_{12}) and wrist circumference ‘distal to the styloid processes’ (x_{13})). The aim is to predict percentage body fat from the covariates. The study was first reported in the sport medicine literature and re-analysed several times. Many variables are strongly correlated (Table 8), which can cause variable selection instability. Royston and Sauerbrei [4] provide strong evidence that observation 39 is an influential point (see Section 2.7.2 and Table 8.3). Therefore, in this paper we decided to exclude this observation and our sample size is 251. The data are publicly available at http://portal.uni-freiburg.de/imbi/Royston-Sauerbrei-book/Multivariable_Model-building/downloads/datasets/edu_bodyfat_both.zip.

5.2. Results

5.2.1. Application of strategies A and B

If we apply the different strategies to the body fat data, we obtain the following results. With the naive strategy, we would select $x_1, x_3, x_4, x_6, x_{11}$ and x_{13} when using $k = 0.3$ as a threshold, and x_6 and x_{13} when using $k = 0.7$. In this specific example, strategy B does not affect the choice of the relevant variables, which remains x_6 and x_{13} . Strategy A, instead, identifies a co-dependence issue between x_4 and x_{13} (correlation 0.73, see Table 8). Let us analyse the 2×2 table for these two variables (Table 9). Condition (i) (see Section 2.3.2) does not hold neither for x_4 nor for x_{13} : $0.406 = h(x_4) < c_1 h(x_4|x_{13}+) = 1.4 \times 0.384 = 0.537$ and $0.928 = h(x_{13}) < c_1 h(x_{13}|x_4+) = 1.4 \times 0.872 = 1.221$. In contrast, condition (ii) holds for x_4 , $0.722 = h(x_4|x_{13}-) > c_1 h(x_4) = 1.4 \times 0.406 = 0.568$. Since condition

Table 8. Body fat example: Pearson correlation among the explanatory variables and their bootstrap inclusion frequencies computed with 250 iterations.

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}
x_1	1.00	−0.02	−0.25	0.12	0.18	0.24	−0.06	−0.21	0.02	−0.11	−0.04	−0.09	0.22
x_2	.	1.00	0.51	0.81	0.89	0.87	0.93	0.85	0.84	0.58	0.79	0.68	0.72
x_3	.	**	1.00	0.33	0.22	0.18	0.39	0.34	0.51	0.39	0.32	0.32	0.40
x_4	.	***	*	1.00	0.77	0.72	0.69	0.66	0.64	0.43	0.71	0.66	0.73
x_5	.	***	.	***	1.00	0.91	0.82	0.71	0.70	0.45	0.71	0.60	0.64
x_6	.	***	.	***	***	1.00	0.86	0.74	0.71	0.41	0.66	0.53	0.60
x_7	.	***	*	**	***	***	1.00	0.88	0.81	0.52	0.72	0.60	0.61
x_8	.	***	*	**	***	***	***	1.00	0.78	0.50	0.74	0.60	0.53
x_9	.	***	**	**	***	***	***	***	1.00	0.58	0.65	0.58	0.64
x_{10}	.	**	*	*	*	*	**	**	**	1.00	0.45	0.43	0.54
x_{11}	.	***	*	***	***	**	***	***	**	*	1.00	0.70	0.61
x_{12}	.	**	*	**	**	**	**	**	**	*	***	1.00	0.60
x_{13}	.	***	*	***	**	**	**	**	**	**	**	**	1.00
BIF	0.66	0.30	0.49	0.41	0.28	1.00	0.23	0.28	0.06	0.16	0.38	0.26	0.93

Note: The lower triangular summarizes the strength of the relations: ‘.’ if $|\rho| < 0.3$, ‘**’ if $0.3 \leq |\rho| < 0.5$, ‘***’ if $0.5 \leq |\rho| < 0.7$ and ‘****’ if $|\rho| \geq 0.7$.

Table 9. Body fat example: 2×2 table for the inclusion of the variables x_4 and x_{13} .

		x_{13}		Tot
		in	out	
x_4	in	89 (0.356)	13 (0.052)	102 (0.406)
	out	143 (0.572)	5 (0.020)	148 (0.592)
Tot		232 (0.928)	18 (0.072)	250 (1.000)

(ii) does not hold for x_{13} ($0.966 = h(x_{13}|x_4-) < c1 h(x_{13}) = 1.4 \times 0.928 = 1.299$ and $0.928 > 1.4 \times 0.3 = 0.420$), x_4 is removed and x_{13} kept.

5.2.2. Investigation of variable selection frequencies and prediction ability

In order to investigate the variable selection frequencies and the prediction ability of the models computed with the naive strategy, strategy A and strategy B, we repeated the analyses performed on the simulated data on 10,000 subsamples $\mathcal{S}_i$, $i = 1, \dots, 10,000$, of size $251 \times 2/3 \approx 167$, randomly drawn without replacement from the original dataset ($\mathcal{S}$). In this way, we mimic the presence of a learning ($\mathcal{S}_i$) and a test set ($\mathcal{S}_i^C = \mathcal{S} \setminus \mathcal{S}_i$, with sample size $251/3 \approx 84$). We compute the mean square prediction error on the data of the latter set, using the model derived from the former. As a consequence of the smaller sample size, in the subsamples we have a reduced power to include a variable.

The results are reported in Table 10 and Figure 12. We note again the behaviours of strategies A and B with respect to the naive strategy with the corresponding threshold: strategy A decreases the selection frequencies of the variables, while strategy B increases them. Please note that the effect of strategy A (variables removal) is quite strong in this dataset due to the presence of several strong relationships among variable inclusions and of several weak effect variables.

From a model point of view, we also experience a situation similar to that described in the simulation study, with a moderate increase in prediction error for a model constructed

Figure 12. Body fat example: selection frequencies of the 13 variables, obtained by applying strategy A, naive strategy with $k = 0.3$, strategy B and naive strategy with $k = 0.7$ to 10,000 subsample of size 167 (approximately 2/3 of 251) drawn without replacement from the original sample.

Table 10. Body fat example, subsampling: results based on 10,000 runs in terms of mean square prediction error (MSPE) and average number of variables for strategies A and B and for the naive strategy with the corresponding thresholds ($k = 0.3$ and $k = 0.7$, respectively).

	Strategy			
	A	Naive ($k = 0.3$)	B	Naive ($k = 0.7$)
Average MSPE	20.10	20.02	20.76	20.92
Average # variables	5.38	6.24	2.19	2.06

Note: The prediction error has been computed on the third ($n = 84$) of the original sample not used to derive the model.

using strategy A (rather than using the naive strategy with $k = 0.3$) and a noticeably saving in terms of average number of selected variables. Note that the prediction error has been measured through the average mean square prediction error computed only on the data included the test sets $\mathcal{S}_i^C$. For strategy B, instead, we have, with respect to the naive strategy with $k = 0.7$, a decrease in prediction error at the cost of a moderate increase in terms of variables included in the models (Table 10).

6. Discussion

In this paper we investigated strategies A and B introduced by Sauerbrei and Schumacher [8] and their ability of handling co-dependence issues. We compared their performances with those of a naive strategy which consists in simply evaluating the bootstrap inclusion frequencies with respect to a hard threshold. We also investigated in details the role of the tuning parameters c_1 , c_2 , c_3 and k . All in all, we saw that strategies A and B have a positive effect in balancing the prediction ability and the sparsity of a model. As a general result from our paper, we note the advantages in using the sparser model. Especially when applying strategy B, we obtain models with a smaller number of variables whose average mean square prediction error is very similar to that of models obtained with strategy A and the naive strategy with $k=0.3$, both selecting more variables (see Table 5). As stated in the introduction, the advantages of a sparser model also include better interpretability and useability, including easier and cheaper data collection for a future prediction. If we should suggest one procedure, therefore, we would choose the implementation of strategy B with a large value of c_3 (let us say between 0.85 and 0.95) and a k not far from 0.7.

Here we did not consider the comparison with a single application of a variable selection procedure (e.g. backward elimination) to the whole sample. We think that this point deserves a dedicated study, as done for a simple case (comparison to the naive strategy with $k=0.5$) by Austin [20]. Now that we have better understood the strategies' behaviours and the effect of different choices for the tuning parameters, a study comparing resampling-based model building with a usual variable selection represents a natural extension of the current one, and it is already in our plans. Together with variations of the significance level (here set to the classical 0.05), indeed, the choice of the tuning parameters (in particular, k) plays a relevant role in terms of prediction ability, sparsity and selection/exclusion of relevant/irrelevant variables. Obviously such a study needs to have a complex design and suitable analyses require systematic investigations of the role of various parameters in a regression model and the investigation of more complex relationships concerning variable inclusions. Assuming that more complex interactions are not relevant, here we concentrated on the investigation of two-dimensional co-dependencies. However, multicollinearity between covariates is an issue in many datasets and, in general, it is important to investigate dependencies of inclusion frequencies with more than two-dimensional approaches. Such issues were correctly raised by one of the reviewers and need to be considered carefully in a following study. Investigations of higher dimensional dependencies may be conducted similarly to [21,22], in which log-linear models were used for bootstrap result matrices. Obviously, such studies need a large 'sample size', an issue nowadays easy to solve by using a large number of simulation replications.

As an example, here we used the body fat data, which one of us (WS) had already large experience with. Earlier analyses had provided sufficient evidence that one observation has substantial influence in the model selection process and we decided to exclude it. Our investigation should not be harmed by any influential point. However, in real (new) data there is always the risk that single points have a strong influence on model building and such issues require more consideration. Just a single run of a classical variable selection procedures may provide misleading results in the presence of this kind of points [16]. Resampling-based procedures are probably more robust against the presence of influential

points, and such observations may be identified by investigating the resampling result matrix [23].

Another interesting aspect not considered in our study is the influence of the resampling technique. For a matter of comparison with Sauerbrei and Schumacher [8], here we use bootstrap, but alternative approaches can be implemented. In particular, De Bin et al. [14] showed that the use of subsampling instead of bootstrapping in a resampling-based variable selection procedure reduces the over-selection of irrelevant variables. This may be advantageous especially in the implementation of strategy A and, in any case, it is in line with our previous recommendation of favoring sparser models.

Finally, strategies A and B focus on pairwise co-dependencies. Although they seemed to perform relatively well for three co-dependent variable inclusion frequencies (Figure A13 and Table A4 in the supplementary material), the two strategies may need to be suitably extended to face cases of inclusion co-dependencies that involve more variables. While this may hardly be relevant in low-dimensional examples, it may be important for application to genetic studies, where several genes may belong to the same pathway [24].

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by Deutsche Forschungsgemeinschaft [BO3139/4-2, SA580/8-2].

References

- [1] Shmueli G. To explain or to predict? *Stat Sci.* 2010;25:289–310.
- [2] Meinshausen N, Bühlmann P. Stability selection. *J R Stat Soc Ser B.* 2010;72:417–473.
- [3] Sauerbrei W, Buchholz A, Boulesteix A-L, et al. On stability issues in deriving multivariable regression models. *Biom J.* 2015;57:531–555.
- [4] Royston P, Sauerbrei W. Multivariable Model-building: a pragmatic approach to regression analysis based on fractional polynomials for modelling continuous variables. Chichester: Wiley; 2008.
- [5] Gong G. Some ideas on using the bootstrap in assessing model variability. *Computer Science and Statistics: Proceedings of the 14th Symposium on the Interface*; New York: Springer; 1982. p. 169–173.
- [6] Chen C-H, George SL. The bootstrap and identification of prognostic factors via Cox's proportional hazards regression model. *Stat Med.* 1985;4:39–46.
- [7] Altman DG, Andersen PK. Bootstrap investigation of the stability of a Cox regression model. *Stat Med.* 1989;8:771–783.
- [8] Sauerbrei W, Schumacher M. A bootstrap resampling procedure for model building: application to the Cox regression model. *Stat Med.* 1992;11:2093–2109.
- [9] Herold T, Jurinovic V, Metzeler K, et al. An eight-gene expression signature for the prediction of survival and time to treatment in chronic lymphocytic leukemia. *Leukemia.* 2011;25:1639–1645.
- [10] Sill M, Saadati M, Benner A. Applying stability selection to consistently estimate sparse principal components in high-dimensional molecular data. *Bioinformatics.* 2015;btv197.
- [11] Robinson LD, Jewell NP. Some surprising results about covariate adjustment in logistic regression models. *Int Statist Rev./Rev. Int. Statist.* 1991;227–240.
- [12] Demidenko E. Sample size determination for logistic regression revisited. *Stat Med.* 2007;26:3385–3397.

- [13] Andersen PK, Skovgaard LT. Regression with linear predictors. Springer Science & Business Media; 2010.
- [14] De Bin R, Janitza S, Sauerbrei W, et al. Subsampling versus bootstrapping in resampling-based model selection for multivariable regression. *Biometrics*. 2016;72:272–280.
- [15] Buchholz A, Holländer N, Sauerbrei W. On properties of predictors derived with a two-step bootstrap model averaging approach: a simulation study in the linear regression model. *Comput Stat Data Anal*. 2008;52:2778–2793.
- [16] Sauerbrei W, Royston P, Binder H. Selection of important variables and determination of functional form for continuous predictors in multivariable model building. *Stat Med*. 2007;26:5512–5528.
- [17] van Houwelingen HC, Sauerbrei W. Cross-validation, shrinkage and variable selection in linear regression revisited. *Open J Stat*. 2013;3:79–102.
- [18] Sauerbrei W. The use of resampling methods to simplify regression models in medical statistics. *J R Stat Soc Ser C*. 1999;48:313–329.
- [19] Johnson RW. Fitting percentage of body fat to simple body measurements. *J Stat Educ*. 1996;4:265–266.
- [20] Austin PC. Bootstrap model selection had similar performance for selecting authentic and noise variables compared to backward variable elimination: a simulation study. *J Clin Epidemiol*. 2008;61:1009–1017.
- [21] Royston P, Sauerbrei W. Stability of multivariable fractional polynomial models with selection of variables and transformations: a bootstrap investigation. *Stat Med*. 2003;22:639–659.
- [22] Binder H, Sauerbrei W. Stability analysis of an additive spline model for respiratory health data by using knot removal. *J R Stat Soc Ser C*. 2009;58:577–600.
- [23] DeBin R, Sauerbrei W, Boulesteix A-L. Detection of influential points as a byproduct of resampling-based model-building procedures. *Comput Stat Data Anal*. 2017. Under second revision.
- [24] Costanzo M, Baryshnikova A, Bellay J, et al. The genetic landscape of a cell. *Science*. 2010;327:425–431.