Open Journal of Statistics, **2013, 3, 79-102**
http://dx.doi.org/10.4236/ojs.2013.32011 Published Online April 2013 (http://www.scirp.org/journal/ojs)

Cross-Validation, Shrinkage and Variable Selection in Linear Regression Revisited

Hans C. van Houwelingen[1], Willi Sauerbrei[2]
[1]Department of Medical Statistics and Bioinformatics, Leiden University Medical Center, Leiden, The Netherlands
[2]Institut fuer Medizinische Biometrie und Medizinische Informatik, Universitaetsklinikum Freiburg, Freiburg, Germany
Email: wfs@imbi.uni-freiburg.de

Received December 8, 2012; revised January 10, 2013; accepted January 26, 2013

ABSTRACT

In deriving a regression model analysts often have to use variable selection, despite of problems introduced by data-dependent model building. Resampling approaches are proposed to handle some of the critical issues. In order to assess and compare several strategies, we will conduct a simulation study with 15 predictors and a complex correlation structure in the linear regression model. Using sample sizes of 100 and 400 and estimates of the residual variance corresponding to R^2 of 0.50 and 0.71, we consider 4 scenarios with varying amount of information. We also consider two examples with 24 and 13 predictors, respectively. We will discuss the value of cross-validation, shrinkage and backward elimination (BE) with varying significance level. We will assess whether 2-step approaches using global or parameterwise shrinkage (PWSF) can improve selected models and will compare results to models derived with the LASSO procedure. Beside of MSE we will use model sparsity and further criteria for model assessment. The amount of information in the data has an influence on the selected models and the comparison of the procedures. None of the approaches was best in all scenarios. The performance of backward elimination with a suitably chosen significance level was not worse compared to the LASSO and BE models selected were much sparser, an important advantage for interpretation and transportability. Compared to global shrinkage, PWSF had better performance. Provided that the amount of information is not too small, we conclude that BE followed by PWSF is a suitable approach when variable selection is a key part of data analysis.

Keywords: Cross-Validation; LASSO; Shrinkage; Simulation Study; Variable Selection

1. Introduction

In deriving a suitable regression model analysts are often faced with many predictors which may have an influence on the outcome. We will consider the low-dimensional situation with about 10 to 30 variables, the much more difficult task of analyzing ‘omics’ data with thousands of measured variables will be ignored. Even for 10+ variables selection of a more relevant subset of these variables may have advantages as it results in simpler models which are easier to interpret and which are often more useful in practice. However, variable selection can introduce severe problems such as biases in estimates of regression parameters and corresponding standard errors, instability of selected variables or an overoptimistic estimate of the predictive value [1-4].

To overcome some of theses difficulties several proposals were made during the last decades. To assess the predictive value of regression model cross-validation is often recommended [2]. For models with a main interest in a good predictor the LASSO by [5] has gained some popularity. By minimizing residuals under a constraint it combines variable selection with shrinkage. It can be regarded, in a wider sense, as a generalization of an approach by [2], who propose to improve predictors with respect to the average prediction error by multiplying the estimated effect of each covariate with a constant, an estimated shrinkage factor. As the bias caused by variable selection is usually different for individual covariates, [4] extends their idea by proposing parameterwise shrinkage factors. The latter approach is intended as a post-estimation shrinkage procedure after selection of variables. To estimate shrinkage factors the latter two approaches use cross-validation calibration and can also be used for GLMs and regression models for survival data.

When building regression models it has to be distinguished whether the only interest is a model for prediction or whether an explanatory model, in which it is also important to assess the effect of each individual covariate on the outcome, is required. Whereas the mean square error of prediction (MSE) is the main criterion for the earlier situation, it is important to consider further quality criteria for a selected model in the latter case. At least interpretability, model complexity and practical usefulness are relevant [6]. For the low-dimensional situation we consider backward elimination (BE) as the most suitable variable selection procedure. Advantages compared to other stepwise procedure were given by [7]. For a more general discussion of issue in variable selection and arguments to favor BE to other stepwise procedures and to subset selection procedures using various penalties (e.g. AIC and BIC) see [4] and [8]. To handle the important issue of model complexity we will use different nominal significance levels of BE. The two post-estimation shrinkage approaches mentioned above will be used to correct parameter estimates of models selected by BE. There are many other approaches for model building. Despite of its enormous practical importance hardly any properties are known and the number of informative simulation studies is limited. As a result many issues are hardly understood, guidance to built multivariable regression models is limited and a large variety of approaches is used in practice.

We will focus on a simple regression model

$$Y = \beta_0 + X^{\mathrm{T}}\beta_1 + \varepsilon$$

with X a p-dimensional covariate. Let there be n observations $(y_1, x_1), \cdots, (y_n, x_n)$ used to obtain estimates b_1 and $b_0 = \bar{y} - \bar{x}^{\mathrm{T}} b_1$ of the regression parameters.

The standard approach without variable selection is classic ordinary least squares (OLS). In a simulation study we will investigate how much model building can be improved by variable selection and cross-validated based shrinkage. The paper reviews and extends early work by the authors [2,4,9]. Elements added are a thorough reflection on the value of cross-validation and a comparison with Tibshirani's LASSO [5]. With an interest in deriving explanatory models we will not only use the MSE as criteria, but will also consider model complexity and the effects of individual variables. Two larger studies analyzed several times in the literature will also be used to illustrate some issues and to compare results of the procedures considered.

The paper is structured in the following way. Section 2 describes the design of the simulation study. Section 3 reviews the role of cross-validation in assessing the prediction error of a regression model and studies its behavior in the simulation study. Section 4 reviews global and parameterwise shrinkage and assesses the performance of cross-validation based shrinkage in the simulation data. The next Sections 5 and 6 discuss the effect of model selection by BE and the usefulness of cross-validation and shrinkage after selection. Section 7 compares the performance of post-selection shrinkage with the LASSO. Two real-life examples are given in Section 8. Finally, the findings of the paper are summarized and discussed in Section 9.

2. Simulation Design

The properties of the different procedures are investigated by simulation using the same design as in [10]. In that design the number of covariates $p = 15$, the covariates have a multivariate normal distribution with mean $\mu_j = 0$, standard deviation $\sigma_j = 1$ for all covariates. Most correlations are zero, except $R_{1,5} = 0.7$, $R_{1,10} = 0.5$, $R_{2,6} = 0.5$, $R_{4,8} = -0.7$, $R_{7,8} = 0.3$, $R_{7,14} = 0.5$, $R_{9,13} = 0.5$ and $R_{11,12} = 0.7$. The covariates X_3, X_8 and X_{15} are uncorrelated with all other ones. The regression coefficients are taken to be $\beta_0 = 0$ (intercept), $\beta_1 = \beta_2 = \beta_3 = 0$, $\beta_4 = -0.5$, $\beta_5 = \beta_6 = \beta_7 = 0.5$, $\beta_8 = \beta_9 = 1$ and $\beta_{10} = 1.5$, $\beta_{11} = \beta_{12} = \beta_{13} = \beta_{14} = \beta_{15} = 0$.

The variance of the linear predictor $X^{\mathrm{T}}\beta = X_1\beta_1 + \cdots + X_{15}\beta_{15}$ in the model equals $\operatorname{var}(X^{\mathrm{T}}\beta) = \beta^{\mathrm{T}} C_X \beta = 6.25$, where C_X is the covariance matrix of the X's. The residual variances are taken to be $\sigma^2 = 6.25$ or $\sigma^2 = 2.5$. The corresponding values of the multiple correlation coefficient $R^2 = \operatorname{var}(X^{\mathrm{T}}\beta) / \operatorname{var}(X^{\mathrm{T}}\beta + \sigma^2)$ are $R^2 = 0.50$ and $R^2 = 5/7 = 0.714$, respectively. Sample sizes are $n = 100$ or $n = 400$. For each of the four (σ^2, n) combinations, called scenarios, $N = 10{,}000$ samples are generated and analyzed. The scenarios are ordered on the amount of information they carry on the regression coeffients. Scenario 1 is the combination $(n = 100, \sigma^2 = 6.25)$, scenario 2 is $(n = 100, \sigma^2 = 2.50)$, scenario 3 is $(n = 400, \sigma^2 = 6.25)$ and scenario 4 is $(n = 400, \sigma^2 = 2.50)$.

Since the covariates are not independent, the contribution of X_j to the variance of the linear predictor $\operatorname{var}(X^{\mathrm{T}}\beta)$ is not simply equal to $\beta_j^2 \operatorname{var}(X_j) = \beta_j^2$. Moreover, the regression coefficients have no absolute meaning, but depend on which other covariates are in the model. To demonstrate this, it is studied how dropping one of the covariates influences the optimal regression coefficients of the other covariates, the variance of the linear predictor $\operatorname{var}(X^{\mathrm{T}}\beta)$ and the increase of the residual variance σ^2, which is equal to the decrease of $\operatorname{var}(X^{\mathrm{T}}\beta)$. This is only done for $X_4, \cdots, X_{10}$ which have non-zero coefficients in the full model. The results are shown in **Table 1**.

Table 1. Effects of dropping one covariate with non-zero β's. The other 14 covariates remain in the model. The main body of the table gives the regression coefficients. The last rows show the resulting values of $\mathrm{var}\left(X^{\mathrm{T}}\beta\right)$, the increase in the residual variance σ^2 and the multiple correlation R^2. The latter is computed for the case of $\sigma^2 = 6.25$. Covariates X_1, X_2, X_3, X_{11}, ···, X_{15} have $\beta = 0$. Dropping them will not affect the β's of the model under "none".

	cov	covariate dropped						
		none	4	5	6	7	8	9	10
coeff.	1	0	0	0.47	0	0	0	0	1.47
	2	0	0	0	0.25	0	0	0	0
	3	0	0	0	0	0	0	0	0
	4	−0.5	-	−0.50	−0.50	−0.29	−1.20	−0.50	−0.50
	5	0.5	0.50	-	0.50	0.50	0.50	0.50	0.53
	6	0.5	0.50	0.50	-	0.50	0.50	0.50	0.50
	7	0.5	0.34	0.50	0.50	-	0.90	0.50	0.50
	8	1.0	1.40	1.00	1.00	1.29	-	1.00	1.00
	9	1.0	1.00	1.00	1.00	1.00	1.00	-	1.00
	10	1.5	1.50	1.27	1.50	1.50	1.50	1.50	-
	11	0	0	0	0	0	0	0	0
	12	0	0	0	0	0	0	0	0
	13	0	0	0	0	0	0	0.50	0
	14	0	0	0	0	0.25	−0.20	0	0
	15	0	0	0	0	0	0	0	0
$\mathrm{var}\left(X^{\mathrm{T}}\beta\right)$		6.25	6.139	6.163	6.063	6.107	5.860	5.500	5.103
increase σ^2		0	0.111	0.087	0.187	0.143	0.390	0.750	1.147
R^2		0.50	0.491	0.493	0.485	0.488	0.469	0.440	0.408

The table also shows the resulting R^2 for the case that $\sigma^2 = 6.25$. Apparently, the effect of each covariate is partly "*inherited*" by some of the other covariates. A simple pattern of inheritance is seen for X_6. It only correlates with X_2 and the pair $\left(X_2, X_6\right)$ is independent of the rest. If X_6 is dropped, X_2 gets the regression coefficient $\beta_{2,\text{inher}} = R_{2,6}\beta_6 = 0.25$. This saves a little bit of the variance of the linear predictor. It drops from 6.250 to 6.063, while it would have dropped to 6.000 if X_6 were independent of the other predictors. A more complicated pattern is seen for X_7. If that one is dropped, X_{14}, X_8 and X_4 inherite the effects. The covariates X_{14} and X_8 show up because they are directly correlated with X_7. Covariate X_4 shows up because it is correlated with X_8. The variance of the linear predictor drops from 6.250 to 6.107.

Since $\left(X_3, X_{11}, X_{12}, X_{15}\right)$ are independent of the other covariates, they cannot inherit effects. However, $\left(X_1, X_2, X_{13}, X_{14}\right)$ can partly substitute $X_4, \cdots, X_{10}$, although they have coefficients $\beta_i = 0$ in the full model.

3. The Value of Cross-Validation

Cross-validation is often recommended as a robust way of assessing the predictive value of a statistical model. The simplest approach is leave-one-out cross-validation in which each observation is predicted from a model using all the other observations. The generalization is k-fold cross-validation in which the observations are randomly divided into k "folds" of approximately equal size and observation in one fold are predicted using the observations in the other folds. In the paper leave-one-out cross-validation will be used $\left(k = n\right)$, but the formulas presented apply more generally. Let $\bar{y}_{(-i)}, \bar{x}_{(-i)}, b_{1,(-i)}$ be obtained in the cross-validation subset, in which observation i is not included. The cross-validation based estimate of the prediction error is defined as

$$\hat{Err}_{CV} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \left(\bar{y}_{(-i)} + \left(x_i - \bar{x}_{(-i)}\right)^{\mathrm{T}} b_{1,(-i)}\right)\right)^2$$

The true prediction error of the model with estimates b_0 and b_1 from the "original" model using all n observations is defined as

$$Err = E\left[\left(Y_{\text{new}} - \left(b_0 + X_{\text{new}}^{\mathrm{T}} b_1\right)\right)^2\right].$$

In the simulation study it is given by

$$Err = \sigma^2 + b_0^2 + (b_1 - \beta)^{\mathrm{T}} C_X (b_1 - \beta).$$

The results in the simulation study using all covariates without any selection are given in **Table 2**.

The results show that $\hat{Err}_{CV}$ does a good job in estimating the mean value of Err over all simulations. However, since the correlation between $\hat{Err}_{CV}$ and Err over all simulation runs is virtually equal to zero, it must be concluded that it does a very poor job in estimating the prediction error of the individual models.

Notice that the standard deviation of $\hat{Err}_{CV}$ is much larger than that of Err. The explanation is that a lot of the variation in $\hat{Err}_{CV}$ is due to the estimation of the unknown σ^2. Cross-validation might be do a better job in picking up the systematic prediction part of the prediction error caused by the error in the estimated β's. That can be checked by studying the behavior of $\hat{Err}_{CV} - s^2$ which is an estimate of the systematic part $b_0^2 + (b_1 - \beta)^{\mathrm{T}} R (b_1 - \beta)$. Here s^2 is the usual unbiased estimator of σ^2. The results are shown in **Table 3**. It nicely shows that the systematic error decreases monotonically from scenario 1 to scenario 4.

Means are very similar but standard deviations from the CV estimates are much smaller. CV somehow shrinks the estimate of the systematic error towards the mean. The table shows that the correlations between the estimate $\hat{Err}_{CV} - s^2$ and the true value $Err - \sigma^2$ are still very low. The warning issued in Section 4 of [2] still holds. It is nearly impossible to estimate the prediction error of a particular regression model. Cross-validation is of very little help in estimating the actual error. It can only estimate the mean error, averaged over all potential "training sets". However, it might be helpful in selecting procedures that reduce the prediction error.

Finally, it should be pointed out that the cross-validation results are in close agreement with the model based estimates of the prediction error as discussed in the same section of [2].

4. Cross-Validation Based Shrinkage without Selection

4.1. Global Shrinkage

As argued by [2,11], the predictive performance of the resulting model can be improved by shrinkage of the model towards the mean. This gives the predictor

$$\hat{Y} = \bar{y} + c \cdot (X - \bar{x})^{\mathrm{T}} b_1$$

with shrinkage factor c, $0 \le c \le 1$. In the following c will be called global shrinkage factor. Under the assumption of homo-skedasticity, the optimal value for c can be estimated as

$$\hat{c}_{\text{heur}} = 1 - \frac{p \cdot s^2}{SS_{\exp}}$$

with $SS_{\exp}$ the explained sum of squares, s^2 the estimate of the residual variance and p the number of predictors.

A model free estimator can be obtained by means of cross-validation. Let $\bar{y}_{(-i)}, \bar{x}_{(-i)}, b_{1,(-i)}$ be obtained in the cross-validation subset, in which observation i is not included, then c can be estimated by minimizing

$$\sum_{i=1}^{n} \left(y_i - \bar{y}_{(-i)} - c \cdot \left(x_i - \bar{x}_{(-i)} \right)^{\mathrm{T}} b_{1,(-i)} \right)^2$$

Table 2. Simulation results for $\hat{Err}_{CV}$ and Err and their correlation (corr.) in models without selection.

			$\hat{Err}_{CV}$		Err		corr.
scenario	n	σ^2	mean	st.dev	mean	st.dev.	
1	100	6.25	7.470	1.160	7.461	0.473	0.024
2	100	2.50	3.029	0.474	2.980	0.188	0.029
3	400	6.25	6.505	0.470	6.511	0.096	-0.009
4	400	2.50	2.611	0.189	2.604	0.038	0.002

Table 3. Simulation results for $\hat{Err}_{CV} - s^2$ and $Err - \sigma^2$ and their correlation (corr.) in models without selection.

			$\hat{Err}_{CV} - s^2$		$Err - \sigma^2$		corr.
scenario	n	σ^2	mean	st.dev	mean	st.dev.	
1	100	6.25	1.216	0.275	1.211	0.473	0.063
2	100	2.50	0.526	0.150	0.480	0.188	0.045
3	400	6.25	0.261	0.048	0.261	0.096	0.095
4	400	2.50	0.114	0.031	0.104	0.038	0.044

resulting in

$$\hat{c}_{cal} = \frac{\sum_{i=1}^{n}\left(y_i-\bar{y}_{(-i)}\right)\cdot\left(x_i-\bar{x}_{(-i)}\right)^{\mathrm{T}} b_{1,(-i)}}{\sum_{i=1}^{n}\left(\left(x_i-\bar{x}_{(-i)}\right)^{\mathrm{T}} b_{1,(-i)}\right)^2}.$$

This estimate can be obtained by regressing $y_i-\bar{y}_{(-i)}$ on $\left(x_i-\bar{x}_{(-i)}\right)^{\mathrm{T}} b_{1,(-i)}$ in a model without an intercept. It differs slightly from the one obtained by regressing y_i on $\hat{y}_{(-i)}$ as proposed in [2]. The definition allows *k*-fold cross-validation and is not restricted to leave-one-out cross-validation. The results of application of global shrinkage in the simulation data, ignoring the restriction $0 \le c \le 1$, are shown in **Table 4**. Actually, $c<0$ was never observed and $c>1$ only occasionally.

The table shows that global shrinkage can help to reduce the prediction error if the amount of information in the data is low. For scenario 1 the mean of the shrinkage factor is 0.84 and the mean reduction of prediction error is 0.14. Corresponding values for scenario 4 are 0.98 and 0.001. For the latter all shrinkage factors are close to one and predictors with and without shrinkage are nearly identical. However, the positive correlation between the shrinkage factor *c* and the reduction in prediction error is counter-intuitive. To get more insight the data for scenario 1 with a small amount of information $\left(n=100, \sigma^2=6.25\right)$ is shown in **Figure 1**.

The relation between reduction in prediction error due to shrinkage and the prediction error of the OLS models are shown for three categories of the shrinkage factor c, namely $\left(c<0.8\right),\left(0.8<c<0.9\right)$ and $\left(c>0.9\right)$. The frequencies of these categories among the 10,000 simulations are 1754, 7740 and 506, respectively. The upper panel shows the apparent (estimated) prediction errors based on cross-validation and the apparent reduction achi- eved by global shrinkage. The differences between the three categories are small, but they are in line with the intuition that the largest reduction is achieved when the shrinkage factor is small. The quartiles (25%, 50%, 75%) of the apparent reduction are 0.09, 0.15, 0.27 for $c<0.8, -0.01, 0.04, 0.15$ for $0.8<c<0.9$ and −0.07, −0.02, 010 for $c>0.9$. The middle panel shows the actual (true) prediction error based on our knowledge of the true model. Here, the picture is completely different. Reduction of the prediction error only occurs when the shrinkage factor is close to one and the OLS prediction error is large. Substantial shrinkage with $c<0.8$ tends to increase the prediction error. The quartiles of the true reduction are −0.29, −0.13, 0.04 for *c* < 0.8, 0.05, 0.18, 0.31 for $0.8<c<0.9$ and 0.19, 0.28, 0.38 for $c>0.9$. The lower panel shows the relation between the apparent and the actual reduction. At first sight the results our counter-intuitive. This phenomenon is extensively discussed in [9]. What happens could be understood from the heuristic shrinkage factor $\hat{c}_{\text{heur}} = 1-\left(p\cdot s^2/SS_{\text{exp}}\right)$. If b is "large" by random fluctuation, the observed explained sum of squares SS_{exp} is large and $\hat{c}_{\text{heur}}$ stays close to 1 and does not "push" b in the direction of the true β. If b is "small" by random fluctuation, SS_{exp} is small and $\hat{c}_{\text{heur}}$ will be closer to 0 and might "push" in the wrong direction. This explains the overall negative correlation $r=-0.253$ between apparent and actual reduction of the prediction error. It must be concluded that it is impossible to predict from the data whether shrinkage will be helpful for a particular data set or not. The chances are given under "frac. pos." in **Table 4**. They are quite high in noisy data, but that gives no guarantee for a particular data set.

4.2. Parameterwise Shrinkage

[4] suggested a covariate specific shrinkage factor, coined parameterwise shrinkage factor (PWSF), to be defined as

$$\hat{Y} = \bar{y} + \left(X-\bar{x}\right)^{\mathrm{T}}\left(c \star b_1\right).$$

Here, c is a vector of shrinkage factors with $0 \le c \le 1$ for $j=1,\cdots,p$ and "$\star$" stands for coordinate-wise multiplication: $\left(c \star b_1\right)_j = c_j \cdot b_{1,j}$. This way of regulation is in the spirit of Breiman's Garrote [12]. See also [9,13]. Sauerbrei suggests to use the parameterwise shrinkage after model selection and to estimate the vector c by cross-validation. As for the global shrinkage

Table 4. Simulation results for the reduction in prediction error (compared with OLS) achieved by global shrinkage in models without selection; "frac. pos." stands for the fraction with positive reduction, "corr." stand for the correlation between the shrinkage factor and the reduction.

			shrinkage factor c		reduction of pred. error		frac. pos.	corr.
scenario	n	σ^2	mean	st.dev	mean	st.dev		
1	100	6.25	0.839	0.045	0.139	0.245	0.746	0.617
2	100	2.50	0.929	0.018	0.022	0.067	0.657	0.487
3	400	6.25	0.962	0.006	0.007	0.025	0.636	0.679
4	400	2.50	0.984	0.002	0.001	0.006	0.573	0.495

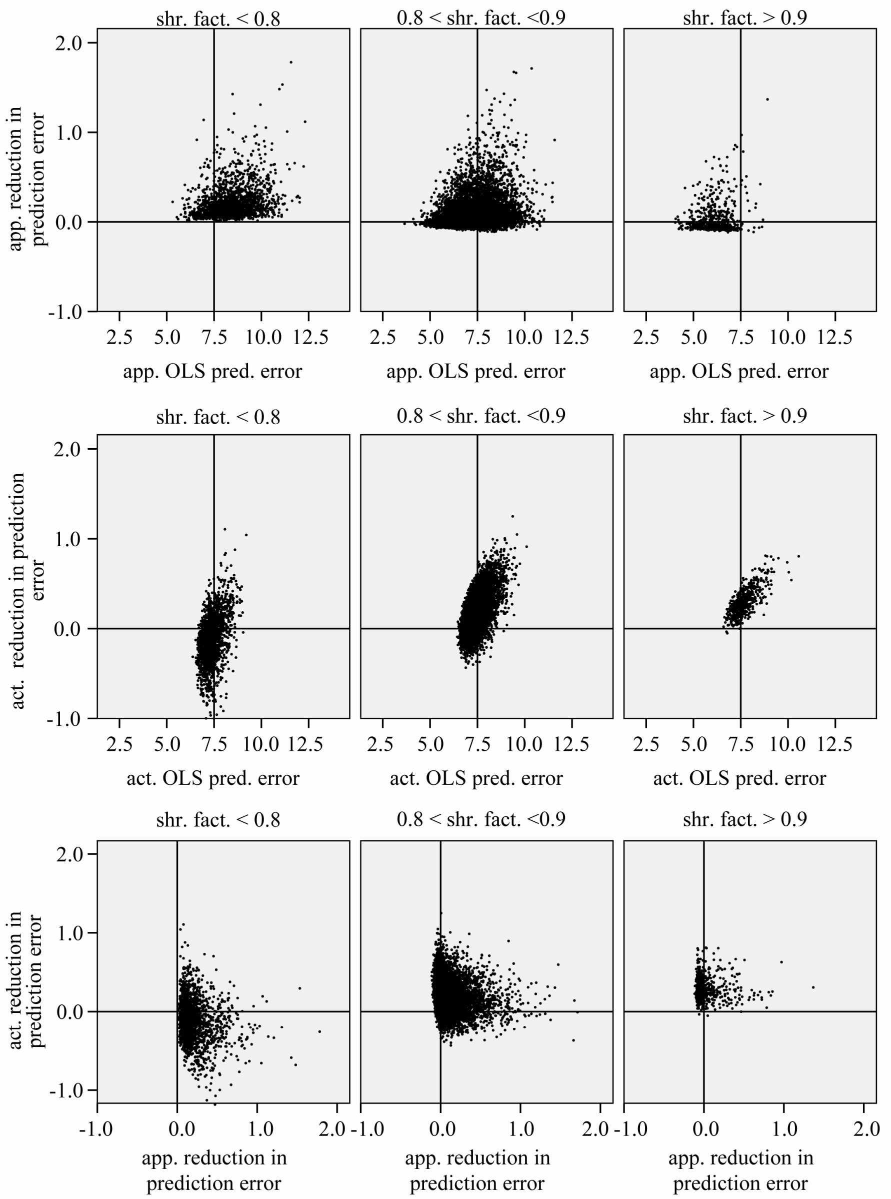

Figure 1. Reduction of prediction error achieved by global shrinkage for different categories of the shrinkage factor c; data from scenario 1 $(n = 100, \sigma^2 = 6.25)$. The upper panel shows the apparent prediction errors obtained through cross-validation, the middle panel shows the actual (true) prediction errors and the lower panel shows the relation between the apparent and the actual reduction.

this could be obtained by regression without intercept of $y_i - \bar{y}_{(-i)}$ on $\left(x_i - \bar{x}_{(-i)}\right) \star \left(b_{1,(-i)}\right)$.

Although this is against the advice of [4] parameterwise shrinkage was applied in the simulation data in models without selection, ignoring the restrictions $0 \le c_j \le 1$ for $j = 1, \cdots, p$. A summary of the results is given in **Table 5**.

Using PWSF the average prediction error increases when compared with the OLS predictor. The increase is large (about 10%) in scenario 1. In scenario 4 the increase is moderate, but still present. Moreover, the estimated prediction error obtained from the cross-validation fit is far too optimistic. (Data not shown). The explanation is that parameterwise shrinkage is not able to handle the redundant covariates with no effect at all. This can be seen from the box plots in **Figure 2**.

For the redundant covariates the shrinkage factors are all over the place. Even variables with a weak effect have sometimes negative PWSF values. For the strong covariates they are quiet well-behaved despite the erratic behavior for the other ones. The conclusion must be that

Table 5. Comparison of the prediction errors of OLS, global shrinkage and parameterwise shrinkage in models without selection.

			mean prediction error		
scenario	n	σ^2	OLS	global shr.	parameterwise shr.
1	100	6.25	7.46	7.32	8.10
2	100	2.50	2.98	2.96	3.24
3	400	6.25	6.51	6.50	6.94
4	400	2.50	2.60	2.60	2.77

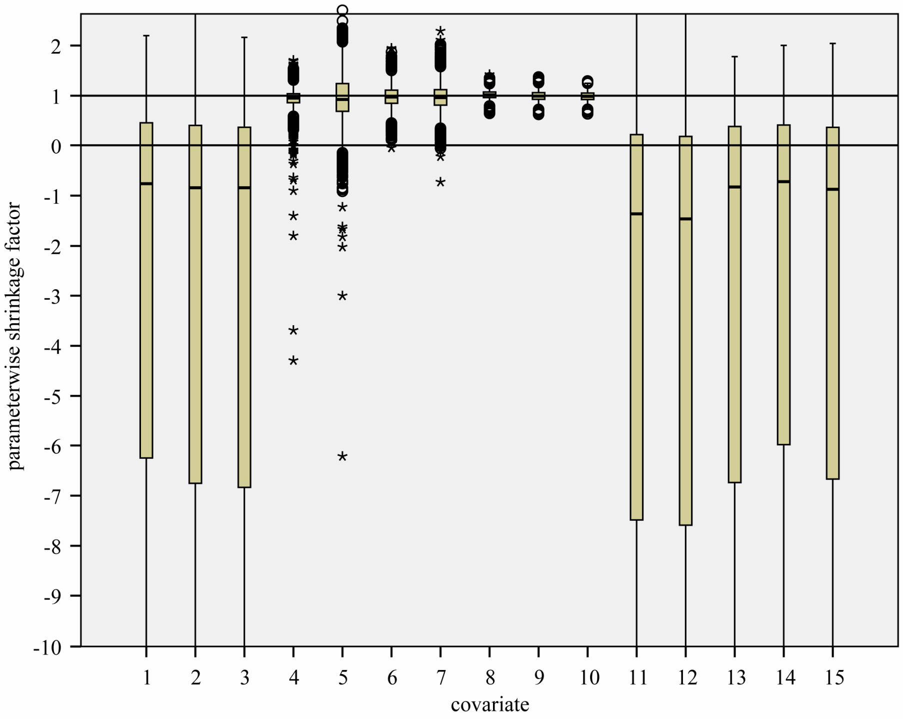

Figure 2. Box plot of parameterwise shrinkage factors in models without selection. Results from scenario 4.

in models with many predictors without selection on stronger predictors parameterwise shrinkage cannot be recommended. The behavior is a bit better if negative shrinkage values will be set to zero, but altogether such a constraint is not sufficient. This is completely in line with Sauerbrei's original suggestion.

5. Model Selection

Following [10] models are selected by backward elimination at significance levels of $\alpha = 0.1573, \alpha = 0.05$ and $\alpha = 0.01$. An impression of which covariates are selected is, shown in **Figure 3**.

5.1. Number of Variables Selected, Type I and Type II Error

The "softest" definition of model selection is the selection of covariates to be used in further research. The "optimal" model contains only the important covariates, the ones that have an influence on the outcome in the full model. In the simulation those are the covariates $X_4, \cdots, X_{10}$. If these are selected the other ones are redundant. However if one of the important covariates is not selected, other non-important covariates can come to the rescue if they are correlated with the non-selected important covariate(s). In the simulation data non-important covariates that can play such a role are covariates X_1, X_2, X_{13}, X_{14} as can be seen from **Table 1**. The effect of omitting important covariates is the loss of explained variation in the optimal model after selection or, equivalently, the introduction of additional random error. If no selection takes place there is no loss of explained variation, but there is a large number of non-important covariates, leading to larger estimation errors. Even more important is a severe loss in general usability of such predictors [4,6]. Consider, for example, a prognostic model comprising many variables. All constituent variables would have to be measured in an identical or at least in a

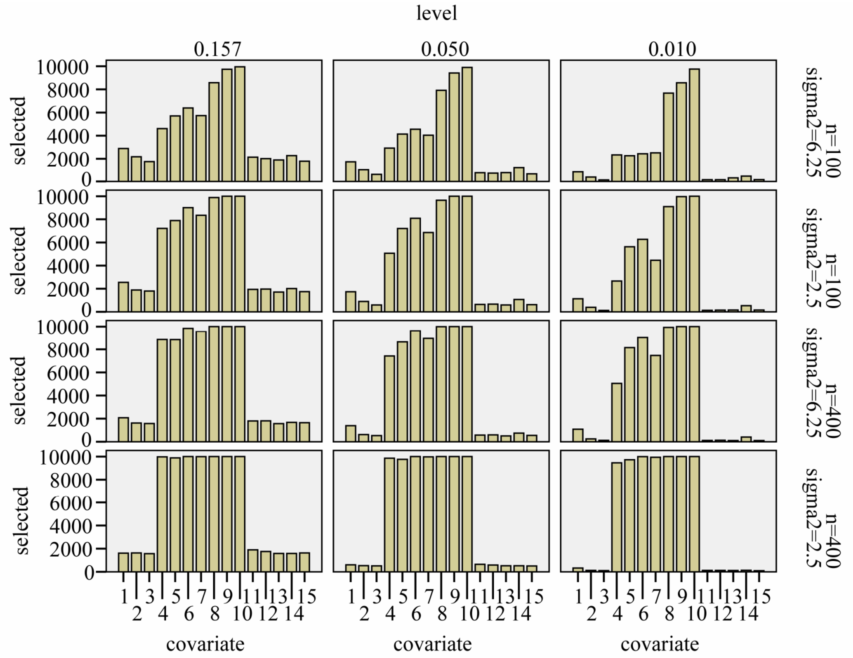

Figure 3. Frequency of selection per covariate.

similar way, even when their effects were very small. Such a model is impractical, therefore "not clinically useful" and likely to be "quickly forgotten" [14].

From **Figure 3** we learn that in the easy situation with a lot of information (scenario 4) all important variables are selected in nearly all replications. For the 8 variables without an influence selection frequencies agree closely to the nominal significance level. For $\alpha = 15.73\%$ selection frequencies are between 15.7% and 19.0%. The corresponding relative frequencies are between 5.1% and 6.55% for $\alpha = 5\%$ and between 0.9% and 3.3% for $\alpha = 1\%$. Results are much worse for situations with less information. For the most extreme scenario 1 it is obvious that all selected models deviate substantially from the true model. For $\alpha = 1\%$ selection frequencies are only between 23.4% and 25.1% for the 4 relevant variables $(X_4 - X_7)$ with a weak effect. Even for $\alpha = 15.73\%$ these frequencies are only between 46.1% and 57.3%. Of these X_4 has the lowest frequency, which is probably caused by the strong correlation with X_8. With 28.9% X_1 has the largest frequency of selection among the non-important covariates. That is explained by its strong correlation with X_5. In 19.4% of the simulations X_1 is selected while the important variable X_5 is not selected.

The effect of selection at the three levels is shown in **Figures 4-6**.

Figure 4 summarizes the number of included variables for the different scenarios. In the simple scenario 4 all seven relevant variables were nearly always selected. In addition irrelevant variables were selected in close agreement to the significance level. In 87.7% the correct model was selected for $\alpha = 0.01$ whereas redundant variables without influence were added when using $\alpha = 0.157$ For scenarios with less information the number of selected variables is usually smaller and the significance level has a stronger influence on the inclusion frequencies. Strong predictors are still selected in most of the replications, but the rest of the selected model does hardly ever agree to the true model. For each of the variables with a weaker effect the power for variable inclusion is low.

The comparison of **Figures 5** and **6** nicely shows the balance between allowing redundant covariates (covariates without effect in the selected model) and allowing loss of explained variation. The number of redundant covariates shown in **Figure 5**, corresponds directly to the type I error. It hardly depends on the residual variance σ^2 and the sample size n. Despite of some stronger correlations in our design the distribution is very close to binomial $(8, \alpha)$. The type II error is reflected by the loss of the variance of the optimal linear predictor $\text{var}(X^{\text{T}}\beta)$, or ,equivalently, the increase in residual variance σ^2, caused by not selecting all important covariates. It could be translated into a loss of R^2 by dividing it by the marginal variance $\text{var}(Y) = \text{var}(X^{\text{T}}\beta) + \sigma^2$, which is equal to 12.5 in scenarios 1 and 3 ($\sigma^2 = 6.25$) and 8.75 in scenarios 2 and 4 ($\sigma^2 = 2.5$). This is shown in **Figure 6**.

Generally speaking, loss of R^2 depends on both the significance level used in the selection process and the

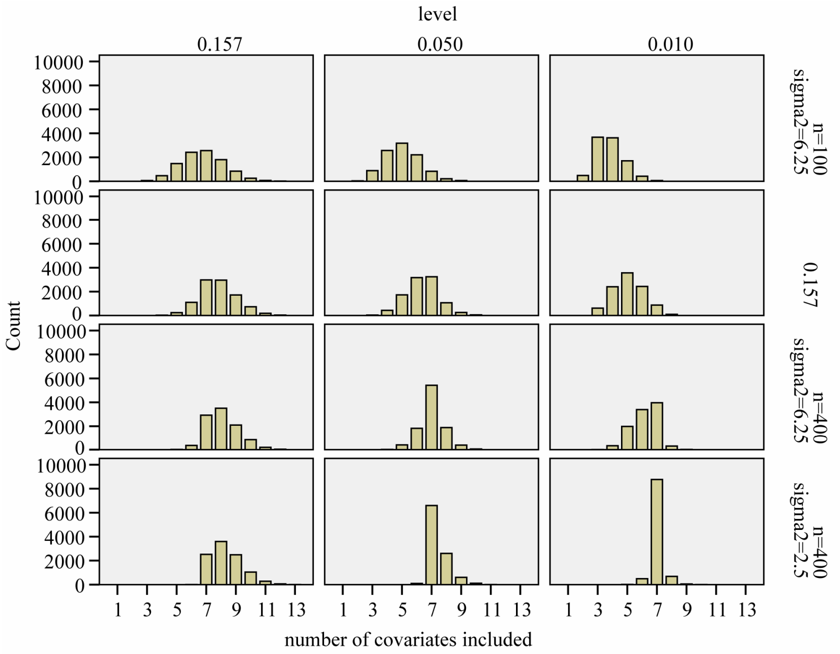

Figure 4. Number of included covariates.

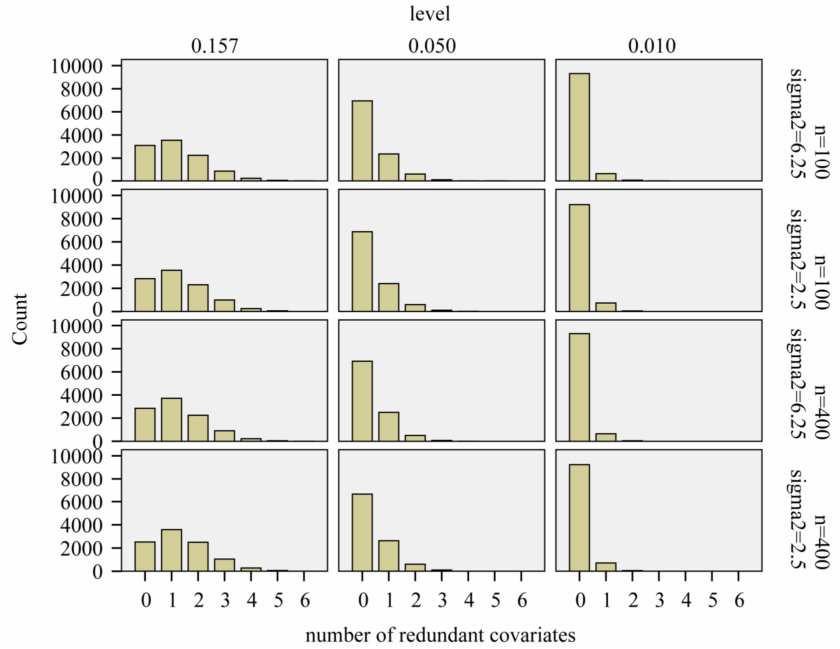

Figure 5. Number of redundant covariates.

amount of information in the data reflected by residual variance σ^2 and the sample size n. In scenario 4 with a high amount of information, the loss of R^2 is negligible and the number of redundant covariates can be controlled by taking $\alpha = 0.01$. In scenario 1 there is a substantial loss of R^2 if the selection is too strict and $\alpha = 0.1573$ might be more appropriate.

As mentioned above and seen in **Table 1** the variable X_1 can partly take over the role of X_5 (correlation coefficient is 0.7) if the latter is not selected. That is shown in **Table 6** for scenario 2 and $\alpha = 0.05$. It has to be kept in mind that other variables are also deleted making a direct comparison difficult. Concerning these two variables the correct model includes X_5 and excludes X_1, which is the case in 69% of the replications. Here the mean loss is smallest. X_5 is erroneously excluded in about 28% of the models. In about half of them the correlated variable X_1 is included, reducing the R^2 loss

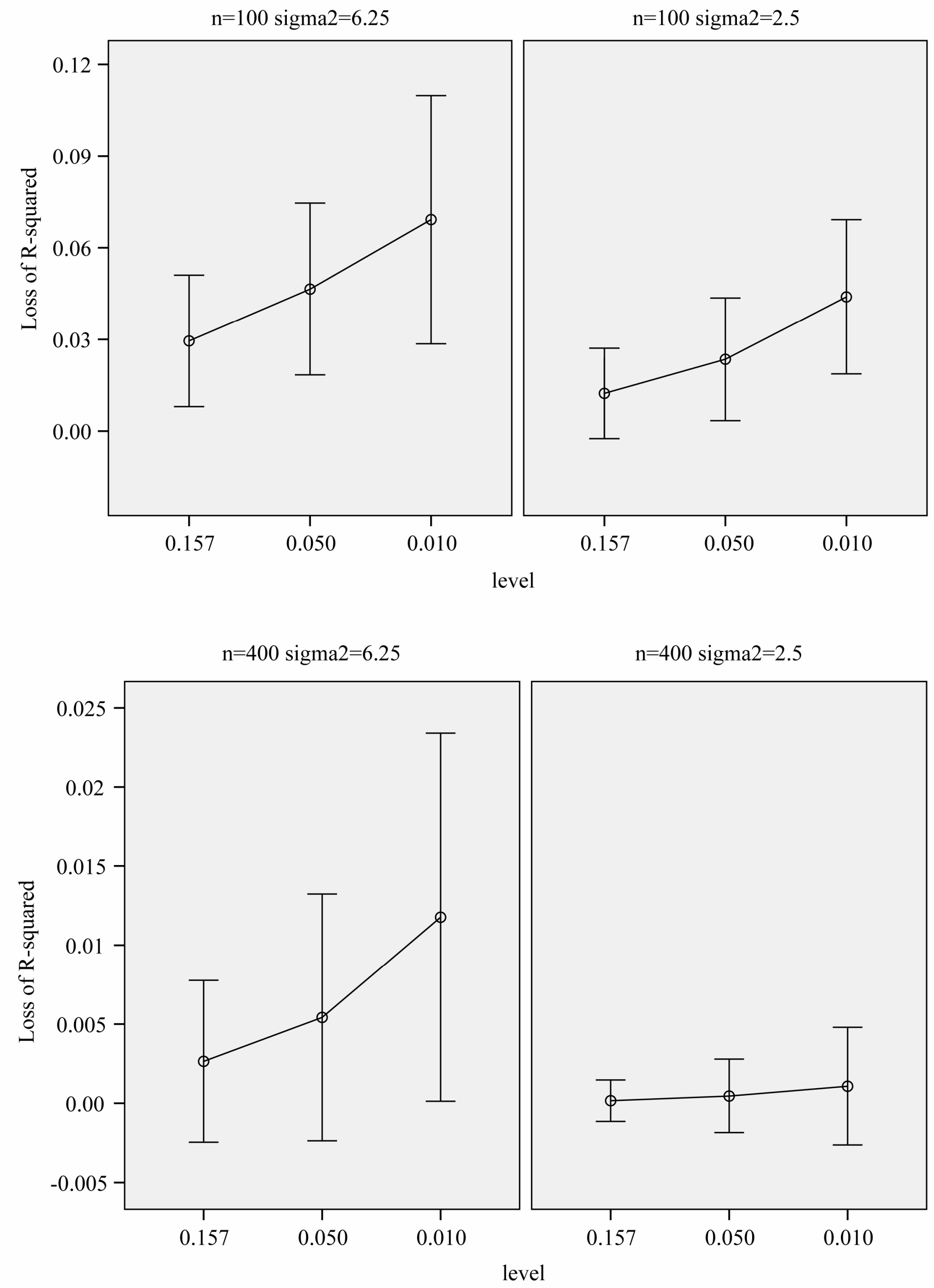

Figure 6. Loss of R^2 in comparison to the full model. The bars show mean ± 1 st.dev. in the population of replications.

Table 6. Inclusion frequencies and average loss of R^2 for combinations of X_1 and X_5 for scenario 2 and $\alpha = 0.05$.

X_1 included	X_5 included	Frequency	mean loss of R^2
no	no	1365	0.0485
no	yes	6900	0.0179
yes	no	1433	0.0274
yes	yes	302	0.0206

caused by exclusion of X_5 substantially.

5.2. Assessment of Prediction Error

The prediction error of a particular model depends on the number of redundant covariates and the loss of explained variation. The average reduction of prediction error when compared with no selection (significance level $\alpha = 1$) is shown in **Figure 7**. Those numbers could be best compared with the systematic component of the prediction error, $Err - \sigma^2$, as reported in **Table 3**.

It nicely shows that the optimal level depends on the amount of information in the data. It also shows that moderate selection at $\alpha = 0.1573$, in a univariate situation equivalent with AIC or Mallows' CP, can do very little harm. Even $\alpha = 0.05$ gives always better results than no selection. For a small sample size the relative reduction in prediction error is small. For the large sample size elimination of several variables reduces the relative prediction error substantially for $\alpha = 0.05$. The average number of selected variables is 7.02 for scenario 3 and 7.40 for scenario 4.

Figure 8 shows the prediction error ranking of the different levels in the individual simulation data sets. Mean ranks were used in replications where the same model was selected. As observed before, the variation is very big and there is no outspoken winner, but there are some outspoken losers. In all scenarios it is bad not to select at all. However, for a small sample size it is even worse to use a very small selection level.

The prediction errors of the selected models can be estimated by cross-validation in such a way that in each cross-validation data set the whole selection procedure is carried out. As in Section 3 this will yield a correct estimate of the average prediction. Thus, it could be used to select the "optimal" significance level in general, but it will not necessarily yield the best procedure for the actual data set at hand.

6. Post-Selection Cross-Validation

A common error in model selection is to use cross-validation after selection to estimate the prediction error and to select the best procedure. As pointed out by [15], this is a bad thing to do. That is exemplified by **Figures 9** and **10**.

Comparing **Figures 7** and **9** shows that cross-validation after selection is far too optimistic about the reduction of the prediction error and is not able to notice the poor performance of selection at $\alpha = 0.01$ for the scenarios 1 and 2. Moreover, as can be seen from **Figure 10**, post-selection cross-validation tends to favor selection at $\alpha = 0.1573$ for all scenarios. This is no surprise because in univariate selection $\alpha = 0.1573$ is equivalent with using AIC, which is very close to using cross-validated prediction error if the normal model holds true.

6.1. Post-Selection Shrinkage

While cross-validation after selection is not able to select

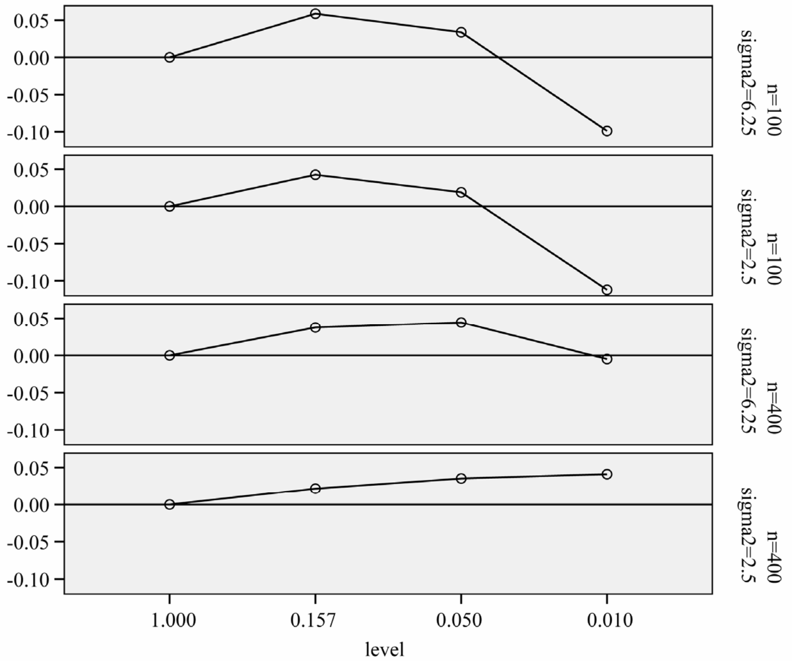

Figure 7. Mean reduction of the true prediction error as reported in Tables 2 and 3 for different α-levels. Values of $Err - s^2$ are 1.21, 0.48, 0.26 and 0.104 in scenarios 1-4, respectively (see Table 3).

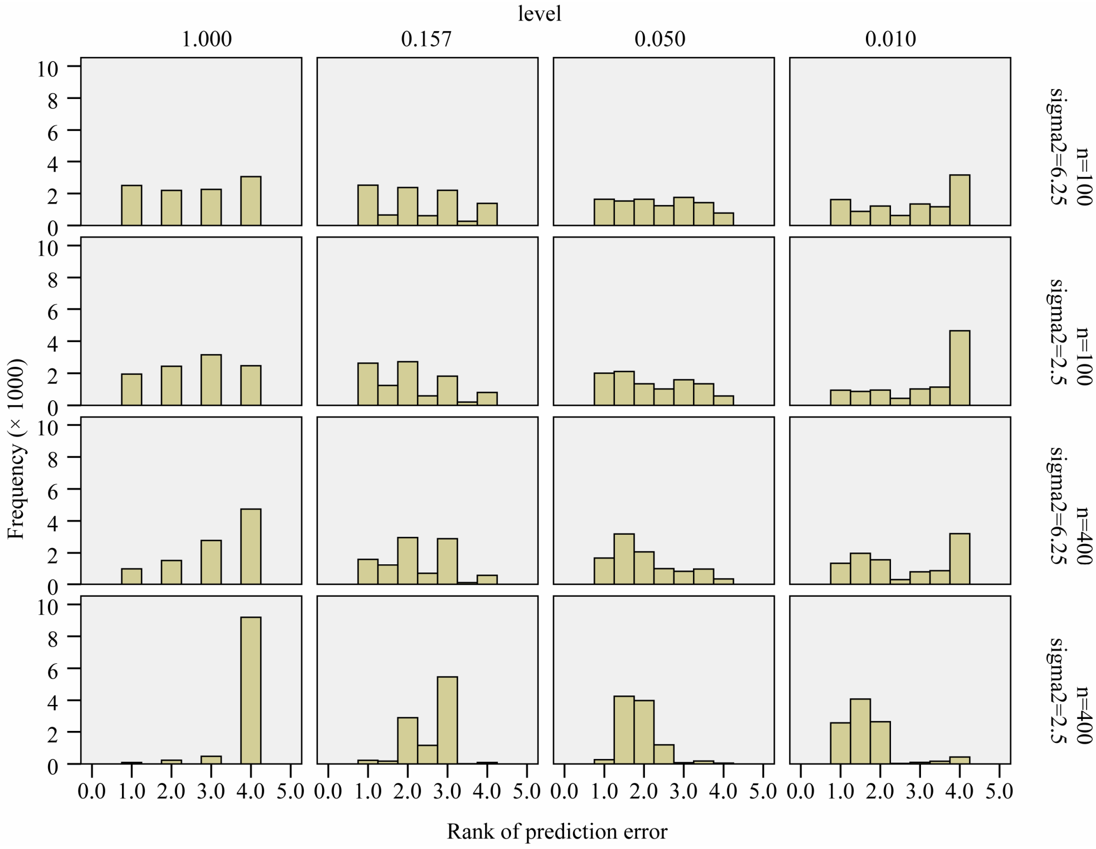

Figure 8. Ranks of prediction error for different levels; rank = 1 is best, rank = 4 is worst.

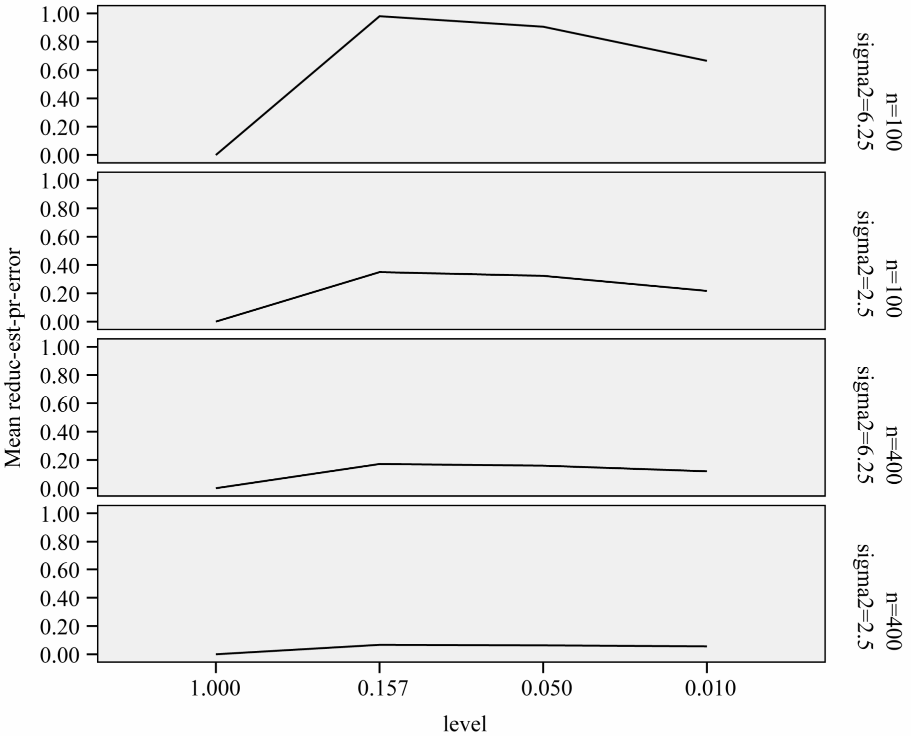

Figure 9. Reduction of estimated prediction error, obtained through post-selection cross-validation, after backward elimination with three selection levels.

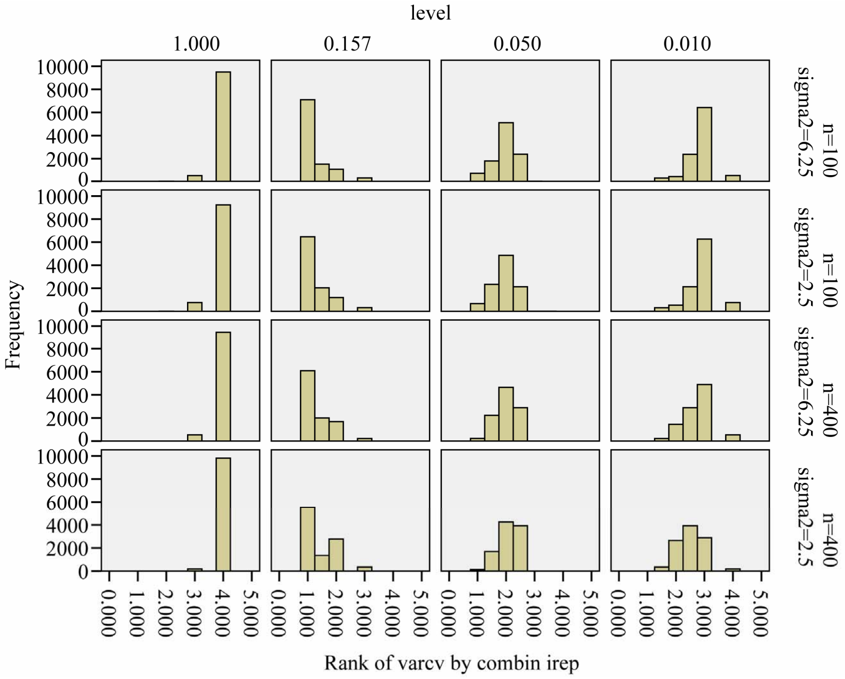

Figure 10. Ranks of estimated prediction error obtained from post-selection cross-validation for different levels; rank = 1 is best, rank = 4 is worst. The ranks of the true prediction error are shown in Figure 8.

the best model, it might be of interest to see whether cross-validation based shrinkage after selection can help to improve the model. The results are shown in **Figure 11**.

On the average, parameterwise shrinkage gives better predictions than global shrinkage, when applied after selection. An intuitive explanation is that small effects that just survive the selection, are more prone to selection bias and, therefore, can profit from shrinkage. In contrast, selection bias plays no role for large effects and shrinkage is not needed, see [16] and chapter 2 of [8]. Whereas global shrinkage "averages" over all effects, parameterwise shrinkage aims to shrink according to these individual needs.

This can also be investigated by looking into the mean squared estimation errors $\overline{\left(\hat{\beta}_j - \beta_{opt,j}\right)^2}$ of the regression coefficients conditional on the selection of covariates in the model. The coefficients $\beta_{opt,j}$ are the optimal regression coefficients in the selected model. As discussed in Section 2 the $\beta_{opt,j}$ coefficients can differ from the β_j coefficients if there is correlation between the covariates. **Figure 12** shows the mean squared errors of the shrinkage based estimators relative to the mean squared errors of the OLS estimators for sample size $n = 100$, scenarios 1 and 2. Sample size $n = 400$ is not shown, because post-selection shrinkage has hardly any effect.

It is clear that the parameterwise shrinkage helps to reduce the effect of redundant covariates that only enter by chance, while the global shrinkage is not able to make that distinction. The precise mechanism is not quite clear yet. To get some better feeling what is going on, the observed scatterplot of parameterwise shrinkage factors versus the OLS estimator are shown in **Figure 13** for covariate X_3 (no effect), X_6 (weak effect) and X_9 (strong effect). These covariates are selected because the optimal parameter value when selected does not depend on which other covariates are selected as well. X_3 is independent of all other variables and the other two variables are only correlated with one variable without influence. Therefore parameter estimates are theoretically equal to the true value in the full model. If the optimal value varies with the selection the graphs are a bit harder to interpret.

For X_3, the covariate without effect and no correlation, the inclusion frequency is close to the type I error and in about half of these cases parameter estimates are positive and negative. The variable is selected in replications in which the estimated regression coefficient is by chance most heavily overestimated (in absolute terms) compared to the true (null) effect. One would hope that PWFS would correct these chance inclusions by a rule like "*the larger the absolute effect* $|b|$, *the smaller the shrinkage factor* c". Although most shrinkage factors are much lower than one, **Figure 13** shows a different cloud: "*the larger* $|b|$, *the larger* c".

A similar observation transfers to the plot for X_9, which is selected in all replications. Therefore, selection bias is no issue for this covariate. The hope is that PWSF would move the estimate close to the true value $\beta = 1.0$.

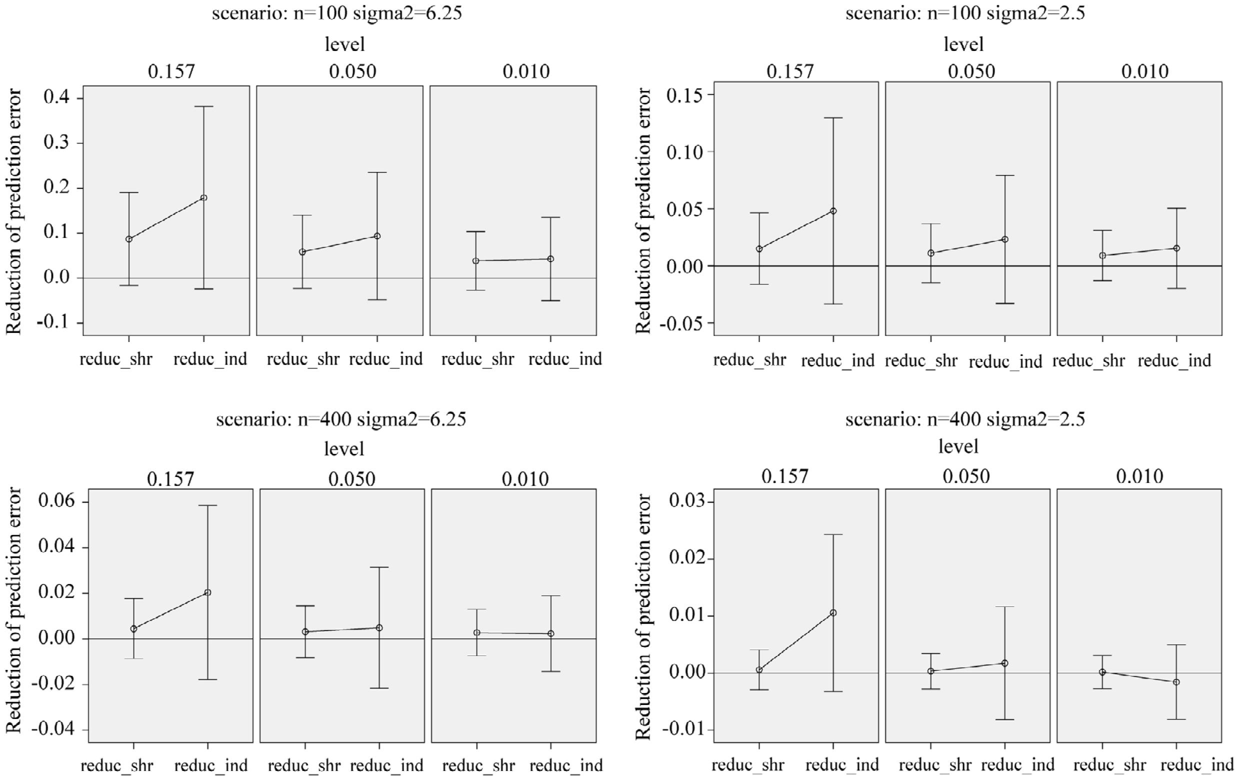

Figure 11. Reduction in prediction error obtained through shrinkage after selection. Error bars show mean ± 1 standard deviation.

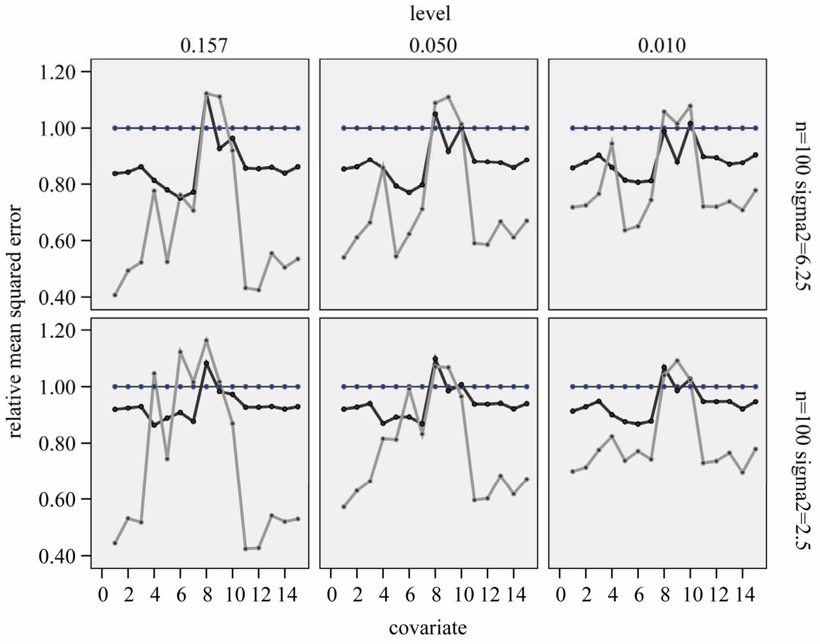

Figure 12. Relative mean squared estimation errors of the partial regression coefficient per covariate (compared to OLS in the selected model) for the parameter estimates obtained by global (black) or parameterwise post-selection shrinkage (grey). Covariates 1-3 and 11-15 have no effect in the full model.

Generally speaking that does not happen: *c* increases slowly with *b*. Most values are close to 1, indicating that shrinkage is not required. The only "hoped for" observation can be made for the cases where the correlated

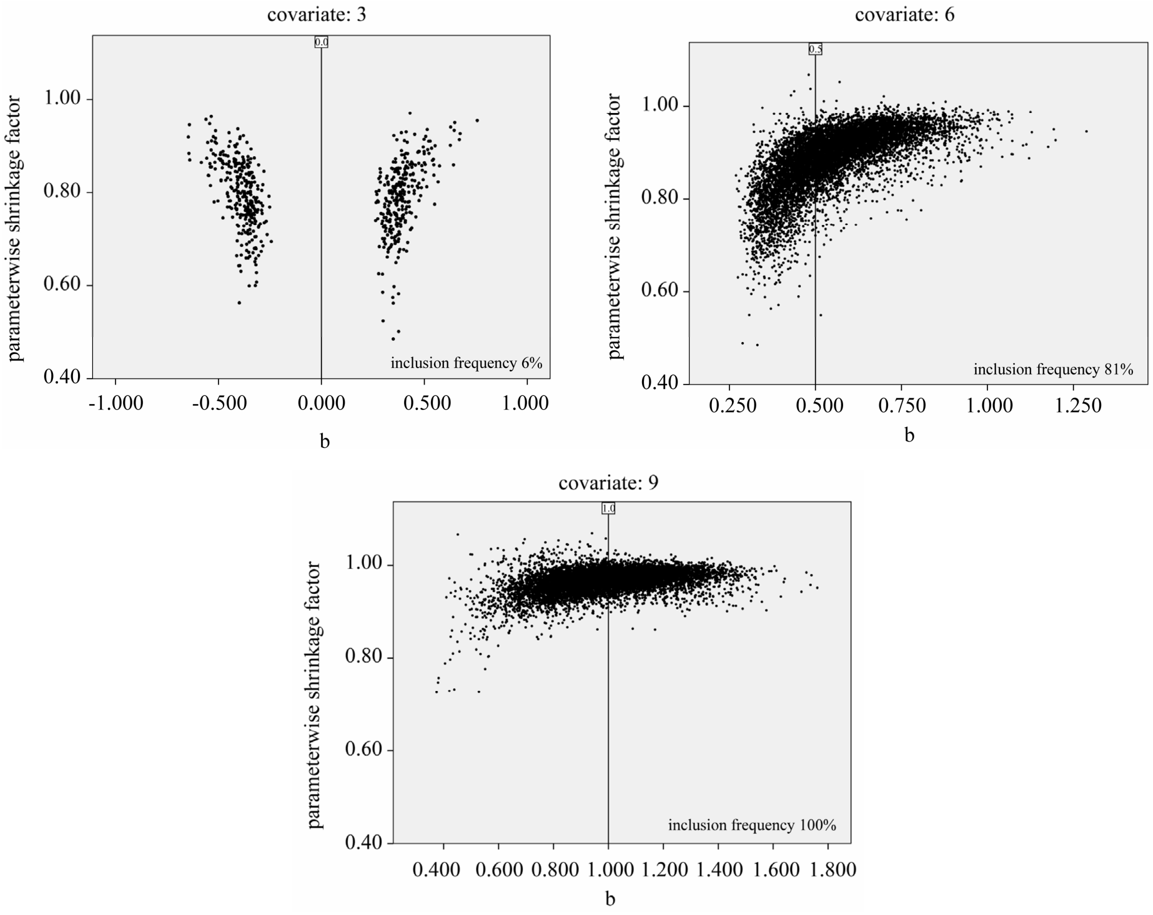

Figure 13. Parameterwise shrinkage factors versus OLS estimates from selected models for covariates 3, 6 and 9; α = 0:05 and scenario 2; reference lines refer to the true value of the parameter.

variable X_{13} is included (plot not shown). X_{13} has no effect and the selection frequency (5.9%) agrees well to the type I error. If X_{13} is included, the shrinkage factor for X_9 show a decreasing trend with c-values clearly below 1 if the estimate b overestimates the true value $\beta = 1$, and values around $c = 1$ if $b < 1$. One might say that parameter shrinkages helps to correct for chance inclusions of X_{13} but not for estimation errors.

X_6 is a covariate with a weak effect. It is not included in 19% of the replications, certainly cases in which the true effect was underestimated by chance. The overall picture for the case where it is included shows a stronger increasing trend (compared with X_9) tending to a value of about $c = 0.96$, if b is large. Here, X_2 plays the role of a confounding redundant covariate. In cases where it is included (plot not shown), the shrinkage factor for X_6 is rather stable with median value about $c = 0.88$.

Some understanding can be obtained by the observation in [2] that in univariate models the optimal shrinkage factor is given by $c_{uni} = 1 - \mathrm{var}\left(\hat{\beta}\right) / \beta^2$. If $|\beta|$ is small, this quantity is very hard to estimate. If $|\beta|$ is large it could be estimated by $\hat{c}_{uni} = 1 - 1/t^2$. The parameterwise shrinkage factor behaves very similarly. This could be seen from plotting PWSF against $|t|$ (Graphs not shown). Such a plot clearly shows that $|t| \approx 2$ (for $\alpha = 0.05$) is the cut-point for an inclusion and that PWSF tends to increase with $|t|$ for included variables. For large absolute t-values, PWSF's are close to one. whereas they drop to about 0.8 for $|t|$ close to 2. The relation between PWSF and $|t|$ is similar for all three covariates X_9. The difference between the covariates is the size of the effect and correspondingly the range of $|t|$ after selection.

The conclusion so far is that coordinate-wise shrinkage is helpful after selection. However it is not clear how to select the significance level. In a real analysis, the level should be determined subjectively by the aim of the study [4]. In the following we will compare backwards elimination with a procedure like the LASSO, that combines selection, shrinkage and fine-tuning.

7. Comparison with LASSO

The simulations discussed above were compared with the results of the LASSO with cross-validation based selection of the penalty parameter λ. Because LASSO is quite time-consuming it was only applied on the first 2000 data sets for each combination of n and σ^2. **Figure 14** shows the distribution of the cross-validation based λ for each combination. The variation in the penalty parameter λ, even in the simple situation of scenario 4 is surprisingly large. There is some correlation with the estimated variance in the full model, but that does not explain the huge variation.

The next **Figure 15** shows the inclusion frequencies

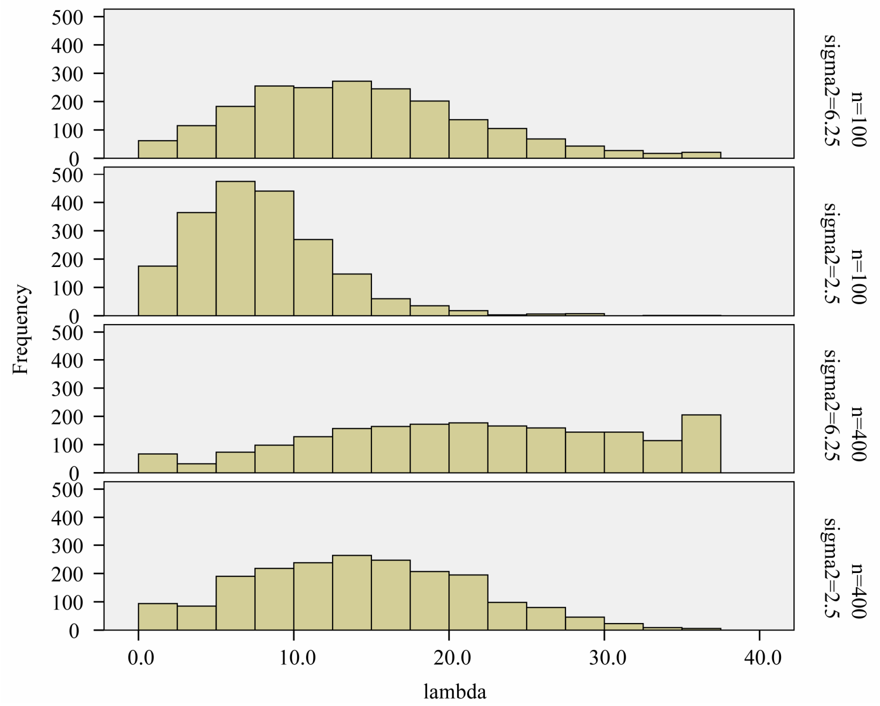

Figure 14. LASSO: histogram of the cross-validation based λ's for the different scenarios.

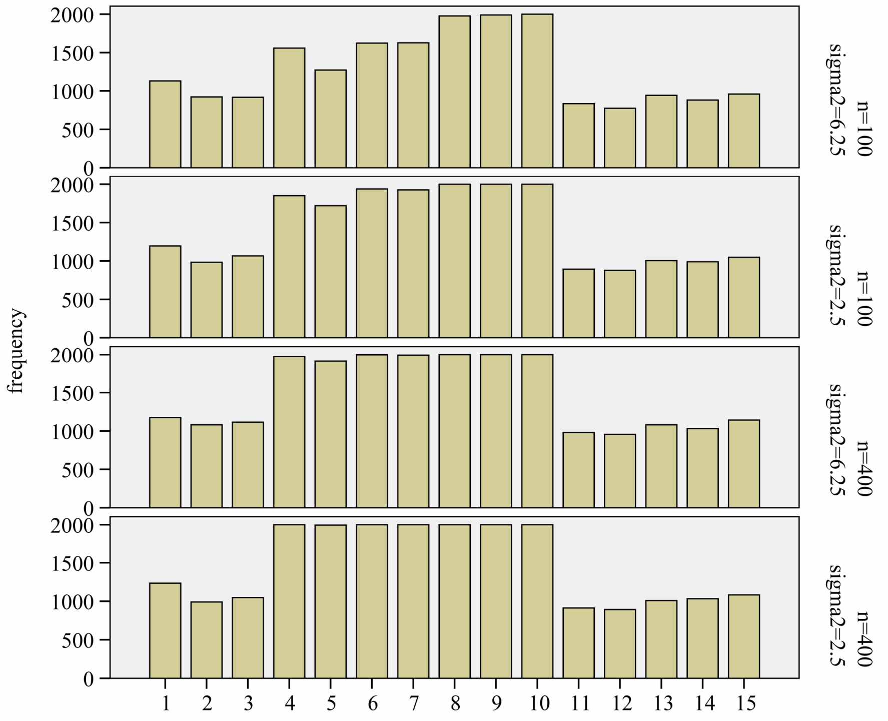

Figure 15. LASSO: inclusion frequencies of the covariates.

for the different covariates. Relevant variables are nearly always included, but LASSO is not able to exclude redundant covariates if there is much signal in the other ones. For example, in scenario 4 inclusion frequencies are 52% for X_3 and 54% for X_{15}, the two uncorrelated variables without influence. The probable reason is that selection and shrinkage are controlled by the same penalty term λ. The phenomenon is also nicely illustrated by **Figure 16**.

Finally, the question how the prediction error of LASSO compares with the models based on selection and shrinkage is answered by **Figure 17**.

The conclusion must be that LASSO is no panacea. Concerning prediction error, it seems to be OK for noisy data (scenarios 1 and 2), but it is beaten by variable selection followed by some form of shrinkage if the data are less noisy (scenario 4). Most likely, that is caused by the inclusion of two many variables without effect. Variable selection combined with parameterwise shrinkage performs quit well. The choice of a suitable significance level seems to depend on the amount of information in the data. Whereas $\alpha = 0.01$ has the best performance in scenario 4, this level seems to be too low in the other scenarios. In these cases selections with $\alpha = 0.157$ or $\alpha = 0.05$ have better prediction performance. Using post selection shrinkage slightly reduces the prediction errors with an advantage for parameterwise shrinkage.

8. Examples

8.1. Ozone Data

For illustration, we consider one specific aspect of a study on ozone effects on school children's lung growth. The study was carried out from February 1996 to October 1999 in Germany on 1101 school children in first and second primary school classes (6 - 8 years). For more details see [17]. As in [18] we use a subpopulation of 496 children and 24 variables. None of the continuous variables exhibited a strong non-linear effect, allowing to assume a linear effect for continuous variables in our analyses.

First, the whole data set is analyzed using backward elimination in combination with global and parameterwise shrinkage and LASSO. Selected variables with corresponding parameter estimates are given in **Table 7** and mean squared prediction errors are shown in **Table 8**. The *t*-values are only given for the full model to illustrate the relation between the *t*-value and the parameterwise shrinkage factor. For variables with very large $|t|$-values, PWSF are close to 1. In contrast, PWSFs are all over the place if $|t|$ is small, a good indication that variables should be eliminated.

Mean squared prediction errors for the full model and the BE models were obtained through double cross-validation in the sense that for each cross-validated prediction, the shrinkage factors were determined by cross-validation within the cross-validation training set. Prediction error for the LASSO is based on single cross-validation because double cross-validation turned out to be too time-consuming. Therefore, the LASSO prediction error might be too optimistic.

MSE is very similar for all models, irrespective of applying shrinkage or not (range 0.449 - 0.475; the full model with PWSF is the only exception), but the number

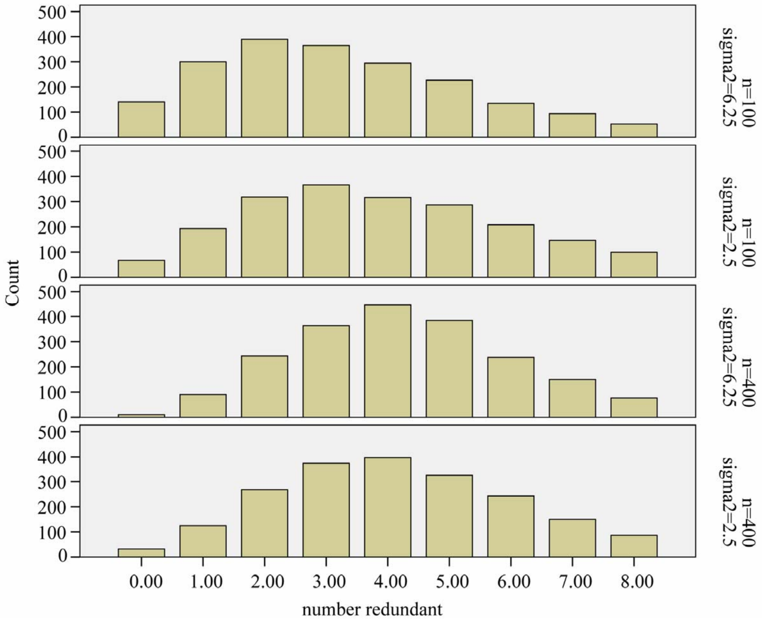

Figure 16. LASSO: distribution of the number of redundant covariates. There are eight redundant covariates in the design.

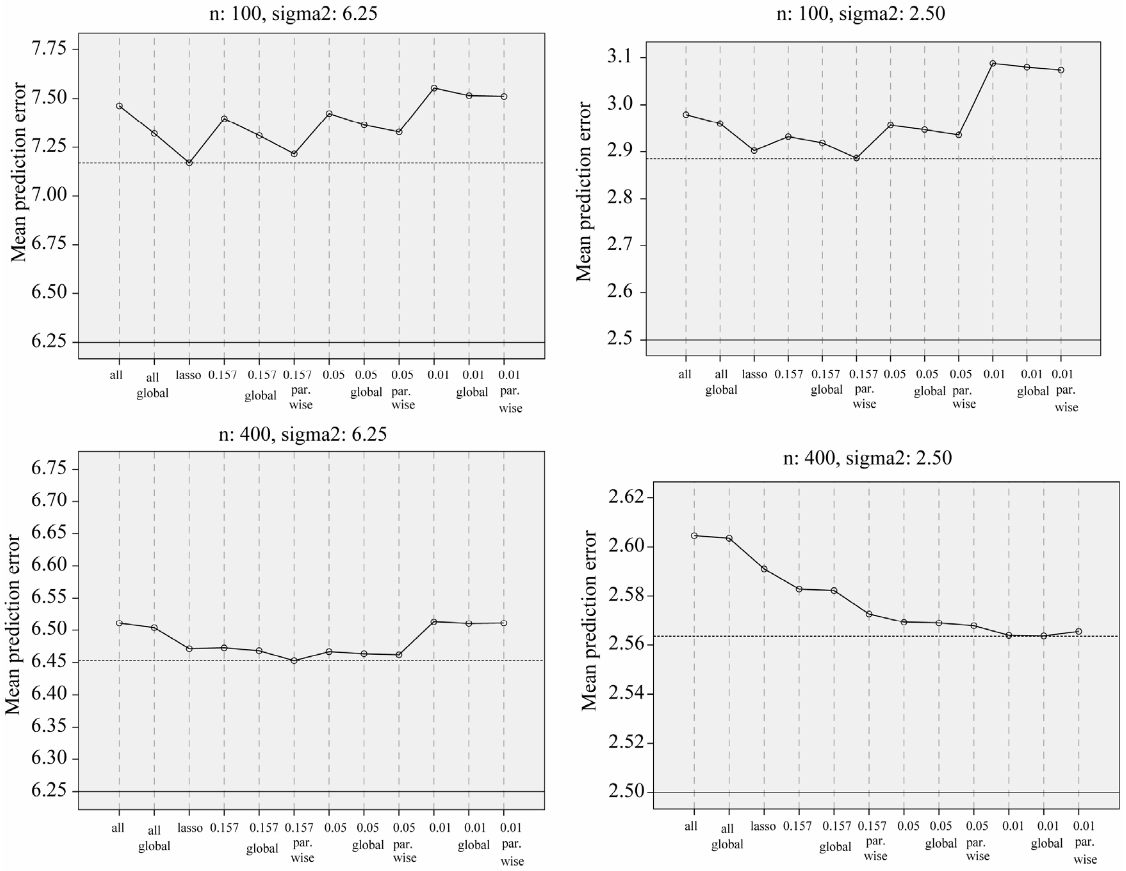

Figure 17. Average prediction errors for different strategies.

Table 7. Analysis of full ozone data set using standardized covariates. The unadjusted R^2 equals 0.67 in the full model and drops only slightly to 0.64 for the BE(0.01) model. For the LASSO model it is 0.66.

method	**full model**			**BE(0.1573)**		**BE(0.05)**		**BE(0.01)**		**LASSO**
c_{global}	0.973			0.9876		0.9927		0.9958		−
variables	b	$\|t\|$	c_{par}	b	c_{par}	b	c_{par}	b	c_{par}	b
ALTER	0.016	1.42	1.09							0.015
ADHEU	−0.010	0.90	1.03							−0.005
SEX	−0.099	10.04	1.00	−0.101	0.98	−0.098	0.99	−0.096	0.99	−0.094
HOCHOZON	−0.033	2.52	0.78	−0.036	0.64	−0.026	0.72			−0.014
AMATOP	−0.002	0.15	−33.9							
AVATOP	−0.007	0.70	−0.20							−0.003
ADEKZ	0.004	0.38	−8.93							
ARAUCH	0.003	0.31	−9.15							
AGEBGEW	0.010	0.97	−0.00							0.007
FSNIGHT	0.008	0.77	−1.09							0.004
FLGROSS	0.173	11.42	0.97	0.181	1.01	0.181	1.01	0.184	1.01	0.172
FMILB	−0.021	1.56	0.45	−0.018	0.64					−0.011
FNOH24	−0.036	2.85	0.72	−0.038	0.72	−0.032	0.79			−0.020
FTIER	−0.004	0.37	−5.60							−0.002
FPOLL	−0.026	1.32	−1.13	−0.020	0.79	−0.025	0.80			−0.011
FLTOTMED	−0.019	1.93	0.86	−0.020	0.63					−0.012
FO3H24	0.033	1.58	0.46	0.038	0.22					
FSPT	0.015	0.65	−3.61							
FTEH24	−0.030	1.53	0.36	−0.032	0.25					−0.002
FSATEM	0.023	1.88	1.08	0.023	0.76					0.019
FSAUGE	0.003	0.30	−6.61							
FLGEW	0.086	6.04	1.16	0.090	0.98	0.090	0.98	0.090	0.97	0.086
FSPFEI	0.027	2.20	0.68	0.027	0.87	0.032	0.89	0.026	0.90	0.019
FSHLAUF	−0.008	0.75	−1.16							

Table 8. Mean squared prediction errors.

model	no shrinkage	global shrinkage	parameterwise shrinkage
full	0.0461	0.0461	0.0629
BE(0.1573)	0.0449	0.0449	0.0456
BE(0.05)	0.0475	0.0475	0.0470
BE(0.01)	0.0465	0.0465	0.0464
LASSO			0.0458

of variables in the model is very different. BE(0.01) selects a model with 4 variables, corresponding PWSF are all close to 1. Three variables are added if 0.05 is used as significance level. Using 0.157 selects a model with 12 variables. Two of them have a very low (below 0.3) PWSF, indicating that these variables may better be excluded. LASSO selects a complex model with 17 variables.

Although they carry relevant information the double cross-validation results for the full data set lack the intuitive appeal of the split-sample approach. To get closer to that intuition the following "dynamic" analysis scheme is applied. First the data are sorted randomly, next the first n_{train} observations are used to derive a prediction model which is used to predict the remaining $n - n_{\text{train}}$ observations. This is done for n_{train} = 100, 150, 200, 250, 300, 350 and repeated 100 times. In that way an impression is obtained how the different approaches behave with growing information. The results are shown in graphs. **Figure 18** shows the mean number of covariates included. More variables are included with increasing sample size (larger power) and differences between the procedures are substantial. For n_{train} = 350 LASSO selects on average 14.4 variables, whereas BE(0.01) selects only 4.0 variables. **Figure 19** shows the evolution of the global shrinkage factor. For models selected with BE(0.01) it is always around 0.98 and for BE(0.157) it varies between 0.96 and 0.98, but for the full model the global shrinkage factor is much lower, starting from 0.83 with an increase to 0.95. **Figure 20** shows the mean squared prediction errors for the different strategies. Without selection PWSF has a very bad performance, whereas the global shrinkage factor can slightly reduce MSE of prediction. PWSF improves the predictor if variable selection is performed and has smaller MSE than using a global shrinkage factor. The most important conclusion is that "BE(0.01) followed by PWSF" and LASSO have very similar prediction MSE's, but the LASSO has to include many more covariates to achieve that.

8.2. Body Fat Data

In a second example we will illustrate some issues in a study with one dominating variable. The data were first analysed in [19] and later used in many papers. 13 continuous covariates (age, weight, height and 10 body circumference measurements) are available as predictors for percentage of body fat. As in the book of [8] we excluded case 39 from the analysis. The data are available on the website of the book. In **Table 9** we give mean squared prediction errors for several models and shrinkage approaches. Furthermore we give these estimates for variables excluding X_6, the dominating predictor. This analysis assumes that X_6 would not have been measured or that it would not have been publicly available. A related analysis is presented in chapter 2.7 of [8] with the aim to illustrate the influence of a dominating predictor and to raise the issue about the term "full model" and whether a full model approach has advantageous properties compared with variable selection procedures. Using all variable MSEs of the models are very similar with a range from 18.76 - 20.80. Excluding X_6 leads to a severe increase, but differences between models are still negligible (25.87 - 27.47); with the full model followed by PWSF as an exception. This agrees well with the results of the ozone data. For BE(0.157), BE(0.01) and LASSO we give parameter estimates in **Table 10**.

Excluding X_6 results in the inclusion of other variables for all approaches. As in the ozone data LASSO hardly eliminates any variable, but the MSE is not better than from BE(0.01) followed by PWSF. PWSF of all variables selected by BE(0.01) are close to 1, whereas variables selected additionally by BE(0.157) all have PWSF values below 0.9 and sometimes around 0.6. This example confirms that BE(0.01) followed by PWSF gives similar prediction MSE's, but includes a much smaller number of variables.

9. Discussion and Conclusions

Building a suitable regression model is a challenging task if a larger number of candidate predictors is available. Having a situation with about 10 to 30 variables in mind the full model is often unsuitable and some type of variable selection is required. Obviously subject matter knowledge has to play a key role in model building, but often it is limited [20] and data-driven model-building is required. Despite of some critique in the literature [3,20,

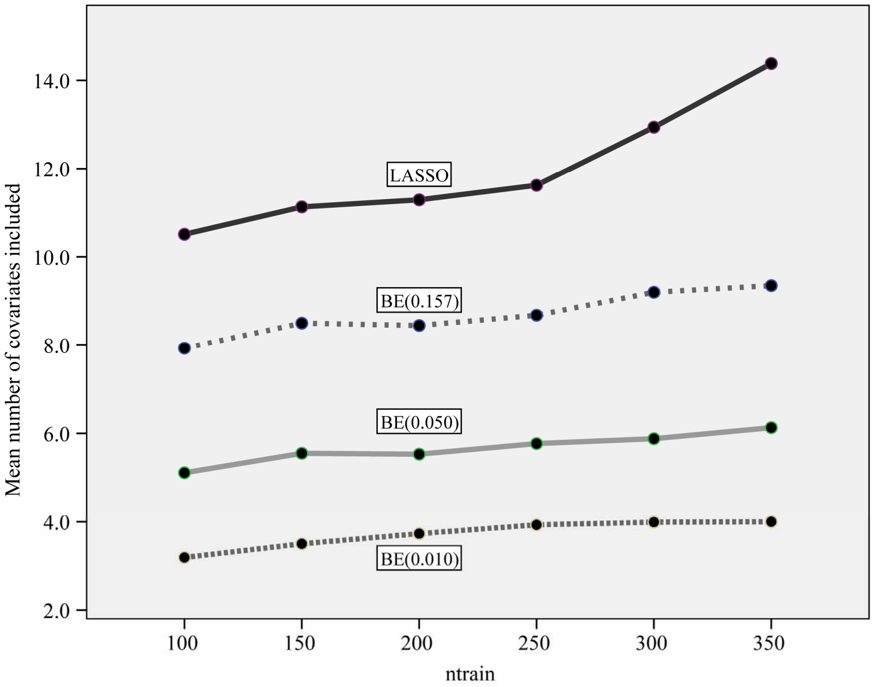

Figure 18. Mean number of covariates included under different strategies.

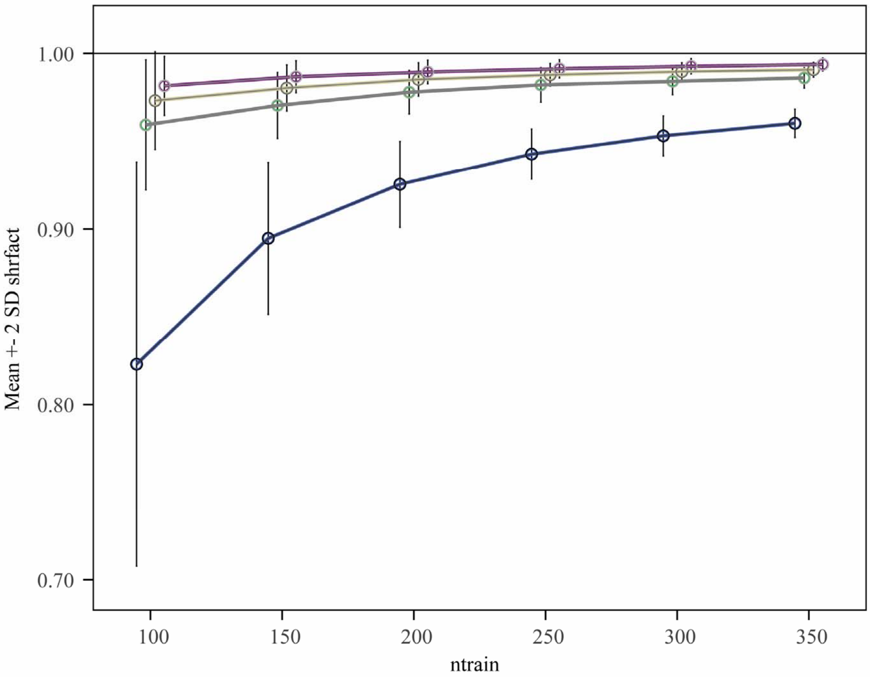

Figure 19. Evolution of the global shrinkage factor for different selection levels. From top to bottom BE(0.010), BE(0.050), BE(0.157) and BE(1.000) = "No selection".

21] we consider backward elimination as a suitable approach, provided the sample size is not too small and the significance level is sensibly chosen according to the aim of a study. For a more detailed discussion see chapter 2 of [8]. Under- and overfitting, model instability and bias of parameter estimates are well known issues of selected regression models.

In a simulation study and two examples we discuss the value of cross-validation, assess a global and a parameterwise cross-validation shrinkage approach, both without and with variable selection, and compare results with the LASSO procedure which combines variable

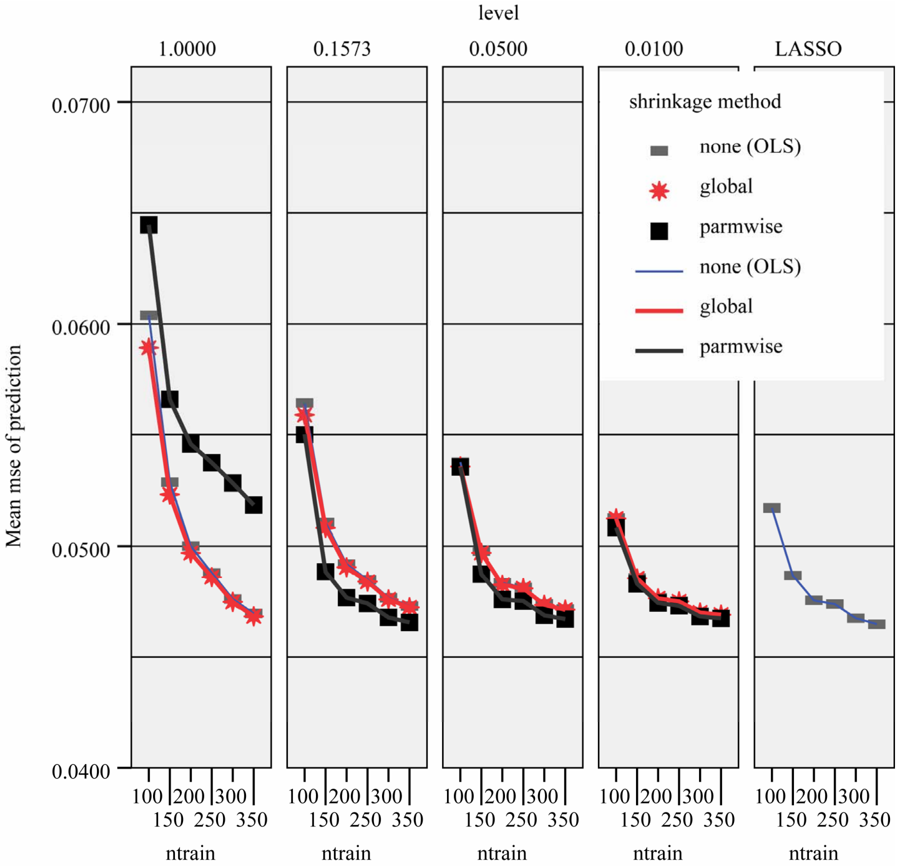

Figure 20. Mean squared prediction error for the different strategies.

Table 9. Mean squared prediction errors.

model	no shrinkage		global shrinkage		param.wise shrinkage	
	all	no X_6	all	no X_6	all	no X_6
full	19.05	26.74	19.04	26.71	20.59	31.47
BE(0.1573)	20.80	27.47	20.78	27.45	20.22	27.22
BE(0.05)	18.76	26.10	18.76	26.10	18.83	26.85
BE(0.01)	19.54	25.87	19.54	25.87	19.54	25.90
LASSO					19.08	26.22

selection with shrinkage [2,4,5]. As discussed in the introduction it is often necessary to derive a suitable explanatory model which means that the effects of individual variables are important. In this respect a sparse model has advantages, both from a statistical point of view and from a clinical point of view. In a related context [22] refer to parameter sparsity and practical sparsity.

9.1. Design of Simulation Study

Our simulation design was used before for investigations of different issues [10]. We consider 15 covariates with seven of them having an effect on the outcome. In addition, multicollinearity between variables introduces indirect effects of irrelevant variables if a relevant variable is not included. It seems consensus that stepwise and other variable selection procedures have limited value as tools for model building in small studies ([8,21] Section 2.3) and 10 observations per variable is often considered as a lower boundary to derive an explanatory model [23]. Based on this knowledge it is obvious that a sample size of $n = 100$ (6.7 per variable) is low. As a more realistic scenario we also consider $n = 400$. Concerning the residual variance we have chosen two scenarios resulting in $R^2 = 0.5$ and $R^2 = 0.71$. The 4 scenarios reflect the situation with a low to a large(r) amount of information in the data. As expected from related studies the amount of information in the data has an influence on the results

Table 10. Parameter estimates for the standardized variables.

	BE(0.1573)				BE(0.01)				LASSO	
	all		no X_6		all		no X_6		all	no X_6
var.	b	c_{par}	b	c_{par}	b	c_{par}	b	c_{par}	b	b
X_1	0.77	0.85	2.76	0.95			2.49	0.98	0.89	2.49
X_2			8.46	0.99			8.86	0.99		4.74
X_3	−0.84	0.91	−2.90	0.99	−1.11	0.93	−2.99	0.99	−0.81	−2.17
X_4	−0.76	0.69	−0.93	0.63					−0.80	−0.47
X_5	−1.09	0.71							−0.76	1.56
X_6	8.93	0.96			8.01	0.99			8.54	
X_7									−0.57	1.41
X_8			1.01	0.71					0.54	0.99
X_9										
X_{10}									0.18	
X_{11}									0.40	
X_{12}	0.73	0.59							0.43	
X_{13}	−1.58	0.98	−2.64	1.00	−1.57	0.97	−2.97	0.96	−1.64	−2.32

and therefore on the comparison between different procedures.

9.2. Cross-Validation and Shrinkage without Selection

The findings of Section 3 confirm that cross-validation does not estimate the performance of the model at hand but the average performance over all possible “training sets”. The results of Section 4 confirm that global shrinkage can help to improve prediction performance in data with little information [2,11] like in the first scenario with $n = 100$ and $\sigma^2 = 6.25$. However, the results show that the actual value of the global shrinkage factor is hard to interpret [9]. Shrinkage is a bit counter-intuitive. Considerable shrinkage is a sign that something is wrong and application of shrinkage might even increase the prediction error. That is evident from the negative correlation $\rho = -0.253$ between apparent and actual reduction in prediction error in the simulations from scenario 1. For the more informative scenarios 2-4 all shrinkage factors are close to one and predictors with and without shrinkage are nearly identical. It must be concluded that it is impossible to predict from the data whether shrinkage will be helpful for a particular data set or not.

Our results confirm that it does not make any sense to use parameterwise shrinkage in the full model [4]. Estimated shrinkage factors are not able to handle redundant covariates with no effect at all. Therefore they can have positive and negative signs and some of them have values far away from the intended range between 0 and 1.

9.3. Variable Selection and Post Selection Shrinkage

Most often prediction error is the main criteria to compare results from variable selection procedures. This implies that a suitable predictor with favorable statistical criteria are the main (or only) interest of an analysis. In contrast to such a statistically guided analysis researchers have often the aim to derive a suitable explanatory model and are willing to accept a model with slightly inferior prediction performance [6]. Excluding relevant variables results in a loss of R^2 and the inclusion of variables without effect complicates the model unnecessarily and usually increases the variance of a predictor. We used several criteria to compare the full model and models derived by using backward elimination with several values of the nominal significance level, the key criterion to determine complexity of a selected model, and the LASSO procedure. The number of variables is a key criterion for the interpretability and the practical usefulness of a predictor. BE(0.01) always selects the sparsest model, but such a low significance level may be dangerous in studies with a low amount of information. Our results confirm that BE(0.01) is very well suited if a lot of information is available. All stronger predictors are always included and only a small number of irrelevant variables is selected. Altogether BE with significance levels of 0.05 or 0.157 selected reasonable models. For studies with a sample size of 400 the loss in R^2 is acceptable and more than compensated when using the prediction error as criterion. Obviously, considering several statistical criteria the full model does not have advantages and the simulation study illustrates that some selection is always sensible. The results of Section 5 show that parameterwise shrinkage after selection can help to improve predictive performance by only correcting the regression coefficients that are borderline significant.

9.4. Comparison with LASSO and Similar Procedures

With the hype for high-dimensional data the LASSO approach [5] became popular. However, results in our simulation study and in the two examples are disappointing. As originally proposed we used cross-validation to select the penalty parameter lambda. However, even in the simplest situation of scenario 4 the variation is surprisingly large. In all scenarios a large number of redundant variables were selected which means that this approach is less suitable for variable selection. This is confirmed in the examples. From 24 candidate variables 17 were included in the model derived for the ozone data. In contrast, BE(0.01) selected a small model with only 4 variables, but the MSE was similar. The simultaneous combination of variable selection with shrinkage is often considered as an important advantage of LASSO and some of its followers, such as SCAD [24] or the elastic net [25]. We compare the LASSO results with post selection shrinkage procedures, in principle two-step procedures combining variable selection and shrinkage. In contrast to LASSO and related procedures post selection shrinkage is not based on optimizing any criteria under a given constraint and the approaches are somehow ad-hoc. Using cross-validation a global shrinkage approach was proposed [2] and later extended to parameterwise shrinkage [4]. These post selection shrinkage approaches can be easily used in all types of GLMs or regression models for survival data after the selection of a model. Whereas global shrinkage “averages” over all effects, parameterwise shrinkage aims to shrink according to individual needs caused by the selection bias. Our results confirm that parameterwise shrinkage helps to reduce the effect of redundant covariates that only enter by chance, while the global shrinkage is not able to make that distinction. The better performance concerning the individual factors results also in better predictions for parameterwise shrinkage compared to global shrinkage. The PWSF results confirm observations from another type of study on the use of cross-validation to reduce bias caused by model building [16]. Small effects that just survived selection are prone to selection bias and can profit from shrinkage. In contrast, large effects are always selected and do not need any shrinkage.

From the large number of newer proposals combining variable selection and shrinkage simultaneously, we considered only the LASSO in this work. In a comparison of approaches in low-dimensional survival settings [26] also compared results from the elastic net, SCAD and some boosting approaches. Using stability, sparseness, bias and prediction performance as criteria they conclude “*overall results did not differ much. Especially in prediction performance*, ⋯, *the variation of them was not too large*”. They did not consider backward elimination followed by post-selection shrinkage, the approach which gave better results than the LASSO in our investigation, both in simulation and the examples. With a suitably chosen significance level for BE prediction performance was not worse compared to the LASSO and models selected were much sparser, an important advantage for interpretation and transportability of models. The investigation also shows that the PWSF approach has advantages compared to the global shrinkage factor. As our two step approach, BE followed by PWSF, can generally be used in regression models we consider it as a suitable approach when variable selection is a key part of data analysis.

9.5. Directions for Future Research

Like the choice of λ in LASSO, the choice of the significance level α in the variable selection is crucial. Double cross-validation might be helpful in selecting α, but reflection is needed about the criterion to be used. Prediction error is the obvious choice but does not reflect the need for a sparse model [27]. In order to improve research on selection procedures for high-dimensional data, several approaches to determine a more suitable λ or use two penalty parameters were proposed during the last years. It would be important to investigate whether they can improve model building in the easier low-dimensional situations

As mentioned above, the approach of this paper can be easily implied for generalized linear models like model logistic regression and survival analysis. It would be interesting to see such applications.

10. Acknowledgements

We thank Gabi Ihorst for providing the ozone study data.

REFERENCES

[1] C. Chen and S. L. George, “The Bootstrap and Identification of Prognostic Factors via Cox’s Proportional Hazards Regression Model,” *Statistics in Medicine*, Vol. 4, No. 1, 1985, pp. 39-46. doi:10.1002/sim.4780040107

[2] J. C. van Houwelingen and S. le Cessie, “Predictive

Value of Statistical Models," *Statistics in Medicine*, Vol. 9, No. 11, 1990, pp. 1303-1325. doi:10.1002/sim.4780091109

[3] F. E. Harrell, K. L. Lee and D. B. Mark, "Multivariable Prognostic Models: Issues in Developing Models, Evaluating Assumptions and Adequacy, and Measuring and Reducing Errors," *Statistics in Medicine*, Vol. 15, No. 4, 1996, pp. 361-387. doi:10.1002/(SICI)1097-0258(19960229)15:4<361::AID-SIM168>3.0.CO;2-4

[4] W. Sauerbrei, "The Use of Resampling Methods to Simplify Regression Models in Medical Statistics," *Journal of the Royal Statistical Society Series C—Applied Statistics*, Vol. 48, No. 3, 1999, pp. 313-329. doi:10.1111/1467-9876.00155

[5] R. Tibshirani, "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society*, *Series B*, Vol. 58, No. 1, 1996, pp. 267-288.

[6] W. Sauerbrei, P. Royston and H. Binder, "Selection of Important Variables and Determination of Functional Form for Continuous Predictors in Multivariable Model Building," *Statistics in Medicine*, Vol. 26, No. 30, 2007, pp. 5512-5528. doi:10.1002/sim.3148

[7] N. Mantel, "Why Stepdown Procedures in Variable Selection?" *Technometrics*, Vol. 12, No. 3, 1970, pp. 621-625. doi:10.1080/00401706.1970.10488701

[8] P. Royston and W. Sauerbrei, "Multivariable Model-Building: A Pragmatic Approach to Regression Analysis Based on Fractional Polynomials for Modelling Continuous Variables," Wiley, Chichester, 2008. doi:10.1002/9780470770771

[9] J. C. van Houwelingen, "Shrinkage and Penalized Likelihood as Methods to Improve Predictive Accuracy," *Statistica Neerlandica*, Vol. 55, No. 1, 2001, pp. 17-34. doi:10.1111/1467-9574.00154

[10] W. Sauerbrei, N. Holländer and A. Buchholz, "Investigation about a Screening Step in Model Selection," *Statistics and Computing*, Vol. 18, No. 2, 2008, pp. 195-208. doi:10.1007/s11222-007-9048-5

[11] J. B. Copas, "Regression, Prediction and Shrinkage (with Discussion)," *Journal of the Royal Statistical Society Series B-Methodological*, Vol. 45, No. 3, 1983, pp. 311-354.

[12] L. Breiman, "Better Subset Regression Using the Nonnegative Garrote," *Technometrics*, Vol. 37, No. 4, 1995, pp. 373-384. doi:10.1080/00401706.1995.10484371

[13] K. Vach, W. Sauerbrei and M. Schumacher, "Variable Selection and Shrinkage: Comparison of Some Approaches," *Statistica Neerlandica*, Vol. 55, No. 1, 2001, pp. 53-75. doi:10.1111/1467-9574.00156

[14] J. C. Wyatt and D. G. Altman, "Prognostic Models: Clinically Useful or Quickly Forgotten?" *British Medical Journal*, Vol. 311, No. 7019, 1995, pp. 1539-1541. doi:10.1136/bmj.311.7019.1539

[15] S. Varma and R. Simon, "Bias in Error Estimation When Using Cross-Validation for Model Selection," *BMC Bioinformatics*, Vol. 7, No. 91, 2006. doi:10.1186/1471-2105-7-91

[16] M. Schumacher, N. Holländer and W. Sauerbrei, "Resampling and Cross-Validation Techniques: A Tool to Reduce Bias Caused by Model Building?" *Statistics in Medicine*, Vol. 16, No. 24, 1997, pp. 2813-2827. doi:10.1002/(SICI)1097-0258(19971230)16:24<2813::AID-SIM701>3.0.CO;2-Z

[17] G. Ihorst, T. Frischer, F. Horak, M. Schumacher, M. Kopp, J. Forster, J. Mattes and J. Kuehr, "Long- and Medium-Term Ozone Effects on Lung Growth Including a Broad Spectrum of Exposure," *European Respiratory Journal*, Vol. 23, No. 2, 2004, pp. 292-299. doi:10.1183/09031936.04.00021704

[18] A. Buchholz, N. Holländer and W. Sauerbrei, "On Properties of Predictors Derived with a Two-Step Bootstrap Model Averaging Approach—A Simulation Study in the Linear Regression Model," *Computational Statistics and Data Analysis*, Vol. 52, No. 5, 2008, pp. 2778-2793. doi:10.1016/j.csda.2007.10.007

[19] R. W. Johnson, "Fitting Percentage of Body Fat to Simple Body Measurements," *Journal of Statistics Education*, Vol. 4, No. 1, 1996.

[20] F. E. Harrell, "Regression Modeling Strategies, with Applications to Linear Models, Logistic Regression and Survival Analysis," Springer, New York, 2001.

[21] E. Steyerberg, R. Eijkemans, F. Harrell and J. Habbema, "Prognostic Modelling with Logistic Regression Analysis: A Comparison of Selection and Estimation Methods in Small Data Sets," *Statistics in Medicine*, Vol. 19, No. 8, 2000, pp. 1059-1079. doi:10.1002/(SICI)1097-0258(20000430)19:8<1059::AID-SIM412>3.0.CO;2-0

[22] J. Bien, J. Taylor and R. Tibshirani, "A Lasso for Hierarchical Interactions," Submitted 2012.

[23] F. E. Harrell, K. L. Lee, R. M. Califf, D. B. Pryor and R. A. Rosati, "Regression Modeling Strategies for Improved Prognostic Prediction," *Statistics in Medicine*, Vol. 3, No. 2, 1984, pp. 143-152. doi:10.1002/sim.4780030207

[24] J. Q. Fan and R. Z. Li, "Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties," *Journal of the American Statistical Association*, Vol. 96, No. 456, 2001, pp. 1348-1360. doi:10.1198/016214501753382273

[25] H. Zou and T. Hastie, "Regularization and Variable Selection via the Elastic Net," *Journal of the Royal Statistical Society Series B*, Vol. 67, No. 2, 2005, pp. 301-320. doi:10.1111/j.1467-9868.2005.00503.x

[26] C. Porzelius, M. Schumacher and H. Binder, "Sparse Regression Techniques in Low-Dimensional Survival Data Settings," *Statistics and Computing*, Vol. 20, No. 2, 2010, pp. 151-163. doi:10.1007/s11222-009-9155-6

[27] C. L. Leng, Y. Lin and G. Wahba, "A Note on the Lasso and Related Procedures in Model Selection," *Statistica Sinica*, Vol. 16, 2006, pp. 1273-1284.