A STUDY OF PRE-VALIDATION

Holger Höfling* Robert Tibshirani†

July 3, 2007

Abstract

Pre-validation is a useful technique for the analysis of microarray and other high dimensional data. It allows one to derive a predictor for disease outcome and compare it to standard clinical predictors on the same dataset. An important step of the analysis is then to test if the microarray predictor has a significant effect for predicting the disease outcome. We show that the straightforward "one degree of freedom" analytical test is biased and we propose a permutation test to remedy this problem. In simulation studies, we show that the permutation test has the nominal level and achieves roughly the same power as the analytical test.

1 Introduction

An often encountered problem is to develop a prediction rule for an outcome based on a dataset. Since there are usually other competing predictors available for prediction of the same outcome, a comparison of the new prediction rule to the old rules is needed in order to determine if the new rule provides any additional benefit. Doing the comparison on the same dataset would favor the new rule as it was derived on this dataset and likely fits it very well. Another approach would be to split the data into separate training and test datasets, build the predictor on the training set and then fit it along with competing predictors on the test set. However with limited data, this may severely reduce the accuracy of the new prediction rule and/or the test set may be too small to have adequate power for the comparison.

*Dept. of Statistics, Stanford University, Stanford, CA, 94305, USA; hhoeflin@stanford.edu

†Departments of Health, Research & Policy, and Statistics, Stanford University, Stanford, CA, 94305, USA; tibs@stat.stanford.edu

Pre-validation (PV) (see Tibshirani and Efron (2002)) offers another approach to this problem. The new prediction rule is derived and compared to the old rules on the same dataset without biased results towards the new rule or big losses of power. Pre-validation is similar to cross-validation, except that the goal is to construct a "fairer" version of the prediction rule, rather than to directly estimate its prediction error. Before going into more detail, we explain how PV works on an example (see also Figure 1).

We have microarray data for n patients with breast cancer. On each array, measurements on p genes were taken. Also available are several non-microarray based predictors, which are commonly used in clinical practice (e.g. age, tumor size ...) to predict if the patient's prognosis is poor or good. We want to use the microarray data in order to predict the prognosis of a patient. In PV, the n patients are divided into K-folds. Leaving out one fold, a prediction rule using the microarray data for the remaining $K-1$ folds is fit (the internal model). Using this rule, the cancer types for the patients in the left out fold are predicted. This way, the data of the left-out fold is not used in building the rule and therefore no overfitting occurs. Repeating this procedure for every fold yields a vector of predictions for each patient. Each patient now has a prediction using a rule for which this patient's data was not used. We call this vector of predictions *pre-validated*. The pre-validated predictor can now be compared to the other non-microarray based predictions using a logistic regression model (the external model). If the coefficient of the pre-validated predictor in the logistic regression model is significant, we conclude that the new microarray-based prediction rule as an independent contribution over the existing rules. The effect of PV is to remove much of the bias that arises from using the same data to build the new prediction rule and compare it to the already established ones.

The goal of pre-validation is to construct a fairer version of our predictor that can be used on the same dataset and will act like a predictor applied to a new external dataset. That is, a one-dimensional pre-validated predictor should behave like a one degree of freedom predictor in a linear model. In this article, we will show that pre-validation is only partially successful in this goal: while the coefficient estimate for the pre-validated predictor is generally good, the one degree of freedom test can be biased, with a level differing from the target level. In this paper we propose a permutation test to solve this problem.

In section 3, we will show this bias analytically in the simple setting with a linear internal and a linear external model. Section 4 outlines the models that are used in the simulations, the amount of bias of the analytical test in these models and the permutation test. Section 5 presents the results of the simulations. In section 6, the method is applied to a microarray dataset of breast cancer patients.

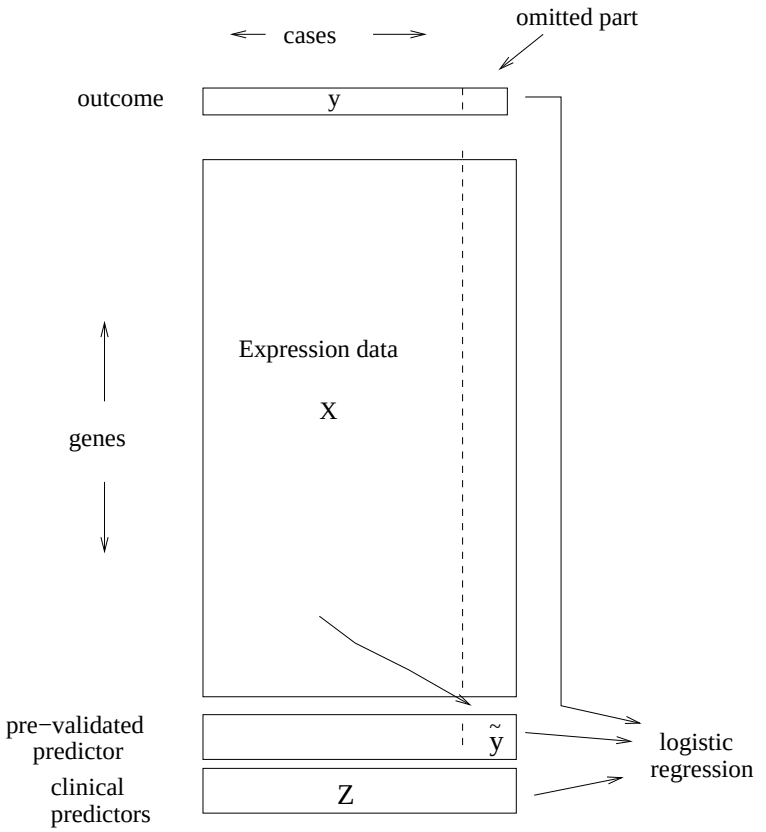

Figure 1: *A schematic of the Pre-Validation process. The cases are divided up into (say) 10 equal-sized groups. Leaving out one of the groups, a prediction rule is derived from the data of the remaining 9 groups. This prediction rule is then applied to the left out group, giving the pre-validated predictor $\tilde{y}$ for the cases in the left out group. Repeating this process for every group yields the pre-validated predictor $\tilde{y}$ for all cases. Finally, $\tilde{y}$ is included in a logistic regression model together with the clinical predictors to assess its relative strength in predicting the outcome.*

2 Pre-Validation

As mentioned above, deriving a prediction rule and comparing it to other rules on the same dataset can lead to a bias in favor of the new rule due to overfitting. This bias can be very large and an example of this effect will be shown later in section 6.

One way to avoid overfitting is to use cross-validation, which is just K applications of the training/test dataset approach mentioned above. For this type of problem, the procedure works as follows:

1. Divide the data in K separate groups.
2. Leave out one group and derive the prediction rule over the remaining $K-1$ groups.
3. Using the new prediction rule, predict the outcome for the left-out group.
4. Compare the strength of the prediction to the already existing predictors for the outcome (e.g. in a linear or logistic regression model, depending on the type of outcome) only in the left-out group. Test if the new predictor is significant.
5. Repeat steps 2-4 for every group and average the results.

However, depending on the choice of K, there are tradeoffs. If K is small, say 2 or 3, the prediction rule is derived on a smaller set of data, thus possibly losing accuracy. In situations as with microarray data, where the number of observations is usually small compared to the amount of available data, the reduction of prediction strength due to the lower number of observations can be substantial. On the other hand, if K is say 4 or larger, the comparison to the already existing prediction rules has to be done on a very small number of observations. If there are 5 (say) other predictors and a total of 50 observations, then with $K = 5$, the comparison of the new rule to the 5 old ones would have to be done using only 10 observations - it is very unlikely to find significant effects under these circumstances.

Pre-validation (see Figure 1) changes this procedure to avoid the for-mentioned problems:

1. Divide the data in K separate groups.
2. Leave out one group and derive the prediction rule over the remaining $K-1$ groups.
3. Using the new prediction rule, predict the outcome for the left-out group.

4. Repeat steps 2 and 3 for each group. Collect the predictions into a vector such that one prediction exists for every observation in every group (we call this predictor "pre-validated").

5. Compare the strength of the prediction to the already existing predictors for the outcome (e.g. in a linear or logistic regression model, depending on the type of outcome) using all observations and the predictor derived above. Test if the new predictor is significant.

For PV, the number of groups K is usually chosen to be 5 or 10. Leave-one-out PV ($K = n$) leads to high variance in estimates and lower values would decrease the size of the training set too much, as already discussed above. However, as in PV the predictions for all observations are collected before the comparison to the existing predictors, a high value of K does not compromise the power of this comparison.

When comparing the pre-validated predictor to the existing predictors, usually a linear or logistic regression model is fitted (depending on the outcome). The new prediction rule is judged to make a significant improvement over the old rules if the coefficient of the pre-validated predictor is significantly different from 0. As the new rule predicts the outcome, significant values for the coefficient would be positive. Therefore instead of a 2-sided test of $\beta_{PV} = 0$ vs $\beta_{PV} \neq 0$, we can get more power by doing a one sided test $\beta_{PV} = 0$ vs. $\beta_{PV} > 0$. For this, usually the standard analytical test for the model (i.e t-statistic or z-score) are used. In the next section, we will prove in a simple case that this analytical test is biased. In more complicated scenarios, simulations are used.

3 Analytical results on the bias of tests for pre-validated predictors

An analytical treatment of the distribution of test statistics in the external model is very difficult in the general case. However, the problem becomes tractable in a simplified setting. Consider PV with $K = n$, i.e. leave-one-out PV. Assume that $p < n$ and use a linear regression model for building the new prediction rule. Let there be e other external predictors for the same outcome y. Let X be the $n \times p$ matrix with the data used for the new prediction rule.

We assume that X and y have the following distributions

$$X_{ij} \sim N(0,1) \quad i.i.d \, \forall \, i = 1, \ldots, n; \; j = 1, \ldots, p$$

and

$$y_i \sim N(0,1) \quad i.i.d. \; \forall \; i = 1, \ldots, n$$

independent also of X. So here our data X is independent of the response y and we can therefore explore the distribution under the null in the external model ($\beta_{PV} = 0$).

3.1 No other predictors

For simplicity, let us first consider the case with $e = 0$, i.e. no other predictors. As a first step, we need an expression for the prediction using the internal linear model and leave-one-out pre-validation. Here let $H = X(X^T X)^{-1} X^T$ be the projection matrix used in linear regression. Let D be the matrix with the diagonal elements of H. Then the leave-one-out pre-validated predictor is

$$\tilde{y} = (I - D)^{-1}(H - D)y =: Py,$$

where I is the identity matrix.

Now use $\tilde{y}$ as the sole predictor in the external model, which is also linear. As there are no other predictors, this may not seem to make much sense, as the hypothesis that there is no relationship between X and y could be tested right away in the internal model. We apply the external model anyway, as it is very instructive as to what the problem is in more complicated settings.

So we now consider the model

$$y = \beta_{PV}\tilde{y} + \varepsilon$$

where $\varepsilon \sim N(0, \sigma^2 \cdot I)$. Then under these conditions, the following theorem holds

Theorem 1. *Under the assumptions described above, the t-statistic for testing the hypothesis $\beta_{PV} = 0$ has the asymptotic distribution*

$$t = \frac{\hat{\beta_{PV}}}{\hat{sd}(\hat{\beta_{PV}})} \rightarrow^d \frac{C - p}{\sqrt{C}} \quad as \; n \rightarrow \infty$$

where $C \sim \chi^2_p$.

Proof. See Appendix A.1. □

As it can be seen here, the statistic is not t-distributed as in a regular linear regression. This can lead to biases when the t-distribution is used for testing. The size of the bias will be explored numerically later in section 3.

3.2 Other predictors related to the response y

Now assume that we have several outside predictors for the response. As these are usually based on different data than X, we define the distribution of the outside predictors based on y and not on the internal model. So let Z be a $n \times e$ matrix with

$$Z_{ik} = y_i + \gamma_{ik}$$

where $\gamma_{ik} \sim N(0, \sigma_k^2)$ *i.i.d.* $\forall\ i = 1, \ldots, n;\ k = 1, \ldots, e$. Thus, the additional predictions are perturbed versions of the true response.

The internal model for the prediction of y using X is the same as before. The external linear model now becomes however

$$y = \tilde{y}\beta_{PV} + Z\beta + \varepsilon.$$

Again we want to test if $\beta_{PV} = 0$. In a linear model, this is usually done by calculating the t-statistic and calculating the quantile using the t-distribution with the right degrees of freedom. The following theorem gives the asymptotic distribution of the t-statistic under these assumptions.

Theorem 2. *Under the setup described above, the t-statistic for testing $\beta_{PV} = 0$ in the external linear model has the asymptotic distribution*

$$t = \frac{\hat{\beta}_{PV}}{\hat{sd}(\hat{\beta}_{PV})} \to^d \frac{(N^T N - p)}{\sqrt{N^T N}} - \frac{N^T A(\mathbf{1}\mathbf{1}^T + Cov(\gamma))^{-1}\mathbf{1}}{\sqrt{N^T N}(1 - \boldsymbol{1}^T(\boldsymbol{1}\boldsymbol{1}^T + Cov(\gamma))^{-1}\boldsymbol{1})} \quad as\ n \to \infty$$

where $N \sim N(0, I_p)$, $A = (A_1, \ldots, A_e)$ with $A_k \sim N(0, \sigma_k^2 \cdot I_p))$, $\boldsymbol{1} = (1, \ldots, 1)^T \in \mathbb{R}^e$ and $Cov(\gamma) = diag(\sigma_1^2, \ldots, \sigma_e^2)$.

Proof. See Appendix A.2 □

We can see that the asymptotic distribution of the t-statistic is not a t or normal distribution, as we already observed in the simple case above without external predictors.

In the next section, by using simulations, we will investigate the extent of the bias when the testing is done using a t-distribution.

4 Models, bias and permutation test

4.1 Models used in the simulations

In the section above we have seen that in the simple case where the internal and external models are linear regressions, the t-statistic does not have its usual distribution. We expect that same is true for more complicated scenarios, which are not tractable analytically. In order to investigate the amount of bias in more complex settings, we used the following 3 model combinations in our simulations.

4.1.1 Linear-Linear

This is the most simple model and was also used in the analytical analysis. Here, the internal and external models are standard linear regressions. Let n be the number of subjects and p be the number of predictors for the internal model. Let e be the number of external predictors. Then the internal predictors are a matrix X which is generated as

$$X_{ij} \sim N(0,1) \quad i.i.d. \quad i = 1, \ldots, n \quad j = 1, \ldots, p$$

Also $\beta_j \sim N(0, \sigma_b^2)$. Using this, the response is generated as

$$y \sim N(X\beta, I \cdot \sigma_I^2)$$

From this true response, the external predictors are derived as

$$Z_{ik} \sim N(y_i, \sigma_E^2) \quad i.i.d.\ i = 1, \ldots, n \quad k = 1, \ldots, e$$

The rationale for simulating the external predictors as a perturbation of the truth rather than the underlying model is that the external predictors would be derived using different models and maybe targeting other aspects of the phenomenon such that the underlying model here would not apply to them. From this perspective, modeling them as a noisy version of the truth seems more appropriate. For simplicity, we always choose $\sigma_b^2 = \sigma_I^2 = \sigma_E^2 = 1$ in the simulations.

4.1.2 Lasso-Linear

This model is an extension of the previous one. The predictor matrix X is generated in exactly the same way as before. However, instead of drawing all β_j from $N(0, \sigma_b^2)$, this is

done only for the first r. All others are set to 0 to ensure sparseness. The external predictors are then generated from y as described above.

For analyzing this artificial data, an internal lasso regression model will be used. The external model is linear regression as before. The internal model will be fit using the LARS algorithm, ensuring that the fitted model contains exactly a prespecified number l of non-zero coefficients.

4.1.3 Linear Discriminant Analysis (LDA) - Logistic

This model is intended to simulate something very to application on microarray data. Again, there are n observations, which are divided into 2 groups with n_1 and n_2 members ($n_1 + n_2 = n$). Also, p predictors (genes) will be generated for each observation independently. However, for the first p_1 out of the p genes, the means will be different. For the first group, $\mu_{ij} = 0 \; \forall i, j$, where i refers to the observation and j to the genes. For the second group of n_2 observations, the first p_1 genes will have $\mu_{ij} = \mu > 0$, a positive offset in the mean from the same genes in the first group. All others genes will also have mean 0 in the second group as well. Then we simulate the microarray data as

$$X_{ij} \sim N(\mu_{ij}, \sigma^2).$$

The external predictors are then generated by switching the label of the y_i independently with probability p_E.

In the internal model, first a number g of predictors is selected by choosing the predictor with the largest correlation with the response. Then an LDA model is fit to the chosen g predictors. In the external model, standard logistic regression is used.

4.2 Simulation of the type I error under the null

In each of the scenarios described above, we simulate artificial data and perform the PV algorithm $100,000$ times (without the permutation test). The analytical p-value of the pre-validated predictor is used to decide if the null hypothesis is rejected (t-statistic in linear regression model, z-score in logistic regression). Based on the simulations, the type I error of the analytical test is estimated (see Table 1).

The analytical tests in the external models show substantial upward and downward bias in the tested scenarios, depending on the choice of parameters. For the type I error level 0.01, this upward bias can double the size of the test and it is also substantial at level 0.05.

The remedy for this problem is a permutation test.

4.3 The permutation test

As we have just seen, the standard analytical test in the external models used (here linear and logistic) do not achieve their nominal level when they are being applied to pre-validated predictors. This can have serious consequences on the outcome of the test. A permutation test is a procedure that is very robust with respect to this problem.

The external predictors have usually been used and validated in this context before, so we were not concerned with evaluating their performance. In any case, extending the permutation test to cover them as well is straightforward. The variables that we have as input is the response y, the internal predictors X and the external predictors Z. As there is a relationship between y and Z, we do not permute y but instead the rows of X. Then, the pre-validation procedure is used and a test statistic in the external model collected (say β or t). This permutation is repeated often enough to get a sufficiently large sample of the test statistic (here usually 500 or 1000 permutations). The p-value is then estimated as the fraction of the permutation test statistic larger or equal to the observed test statistic (no randomization on the boundary). As the pre-validated predictor is a prediction for the response y, we expect its coefficient to be positive and therefore use a one-sided p-value (as we already did for the analytical test).

5 Simulation results

In this section we explore whether the permutation test achieves the intended level and what effect it has on the power of the test compared to the analytical solution. For this, artificial datasets according to the 3 scenarios described above are created and analyzed.

5.1 Level of the permutation test

For estimating the level of the test under the null hypothesis, the internal predictors X will be independent of the response and the external predictors Z. Several different parameter combinations will be used for this task. For each scenario and parameter choice, 1000 simulations were used where each test was based on 500 permutations.

Scenario	Parameters	CV-folds	Type I error		
			$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.1$
Linear-Linear	$n = 10, p = 5, k = 1$	5	0.022	0.079	0.137
		10	0.024	0.080	**0.139**
		n	**0.023**	**0.083**	0.140
	$n = 20, p = 5, k = 1$	5	0.018	0.069	0.123
		10	0.017	0.066	0.120
		n	0.018	0.067	0.119
	$n = 50, p = 5, k = 1$	5	0.016	0.064	0.115
		10	0.016	0.062	0.111
		n	0.015	0.060	0.109
Lasso-Linear	$n = 10, p = 100, k = 1, l = 5$	5	**0.008**	**0.033**	**0.062**
		10	0.011	0.040	0.072
	$n = 10, p = 100, k = 1, l = 10$	5	0.010	0.040	0.074
		10	0.016	0.053	0.091
	$n = 30, p = 100, k = 1, l = 5$	5	0.012	0.040	0.071
		10	0.014	0.046	0.076
	$n = 30, p = 100, k = 1, l = 10$	5	0.016	0.054	0.092
		10	0.021	0.065	0.105
	$n = 30, p = 100, k = 1, l = 20$	5	0.020	0.065	0.112
		10	**0.030**	**0.081**	**0.128**
LDA-Logistic	$n = 20, p = 1000, k = 1, l = 10$	5	0.003	0.025	0.076
		10	0.0096	0.047	0.100
	$n = 40, p = 1000, k = 1, l = 10$	5	0.018	**0.072**	0.122
		10	0.036	0.106	0.158
	$n = 80, p = 1000, k = 1, l = 10$	5	**0.019**	0.071	**0.122**
		10	0.053	0.126	0.179

Table 1: *Type I error in various scenarios. Each estimate is based on* 100000 *simulations, giving an SD of* $\leq .005$*. The most extreme values for each scenario are in bold.*

The levels of the permutation tests can be found in Tables 5, 6 and 7. The standard error for the $\alpha = .01$ estimate is 0.003, for $\alpha = .05$ it is 0.007 and for $\alpha = 0.1$, the standard error is 0.009. All the estimates in the tables are well within 2 standard deviations of their target value, so we see that the permutation tests are unbiased.

5.2 Power

The same scenarios that were used for estimating the level of the permutation tests will also be used to estimate the power under the alternative. As there is no distinct alternative hypothesis, several different choices will be used, depending on the specific scenario.

One of the most interesting aspects of this simulation is to compare the power of the permutation test to the power of the standard analytical test. However, as the analytical test is biased (usually upward), a straightforward comparison using the nominal test levels is inappropriate. In order to adjust for the bias, the simulations in the same scenario and parameters under the null hypothesis will be used. For each nominal level, a new cutoff for the p-values will be estimated such that the level of the analytical test is equal to its nominal level. This cutoff will then also be used to estimate its power.

The results can be seen in Tables 8, 9 and 10. As before, the estimates are based on 1000 simulations, each of which used 500 permutations for the tests. Here, the maximum standard deviation for the test is achieved for a power of 0.5, in which case the SD is 0.016. The power of the permutation test is in most cases very close to the power of the analytical test and sometimes even higher (although this may be a random occurrence). So, there does not seem to be a serious problem with loss of power when comparing the permutation tests to the analytical test.

However, the picture as to which choice of test statistic and number of folds to use for the permutation test is not very clear. For the Linear-Linear model, we used 5-fold PV, 10-fold PV, leave-on-out PV and permutation tests without PV ($K = 1$). For the other model, due to computation time constraints, we only used 5- and 10-fold PV as well as no PV. In the Linear-Linear scenario, leave-one-out PV performs slightly better than 5-fold and 10-fold PV. However, in all but the simplest models, performing leave-one-out PV comes with a serious increase in computation time so that just using 5- or 10-fold PV may be considered appropriate.

In some instances, the permutation test using no PV showed a lot more power than 5- or 10-fold PV permutation tests. However, especially in the LDA-Logistic model, the test without PV had power even below the nominal level of the test. This can be explained by overfitting

the data, leading to perfect separation of the classes even if there is no relationship between the class labels y and the internal predictors X. In these cases, the permutation test without PV does not give useable results.

Therefore, using 5-fold (or 10-fold) PV permutation test is the most reliable procedure, achieving the nominal level of the test without compromising power with respect to the analytical test. The choice of test statistic depends on the specific application, but all standard statistics we used had acceptable performance.

5.3 Performance of the estimator for the pre-validated coefficient

When the new prediction rule turns out to be a significant improvement over the performance of the old prediction rules, the value of the coefficient of the new predictor compared to the coefficients of the old predictors indicates how well the new predictor performs. Therefore it is important to know how well PV estimates the coefficient of the new prediction rule.

In order to have a comparison that is fair and relevant with respect to the amount of data available, we estimate the coefficient using PV over 1000 simulation runs in the scenarios presented above. As a benchmark method, we treat the dataset the PV was performed on as a training set to estimate the new prediction rule and do the comparison to the other prediction rules on an independently simulated test dataset of the same size as the training data. Our primary concern is that the coefficient estimated using PV is roughly unbiased w.r.t. the benchmark. The most straightforward approach would be to compare the mean over the simulations of the estimated coefficient using PV and using the benchmark. However, in the LDA-Logistic scenario, occasionally perfect separation occurs which makes the estimated coefficients extremely large. Mean-unbiasedness is not applicable in this case and we decided to use median-unbiasedness instead. As the difference between mean and median is quite small in all other scenarios and the median is more robust, we used the median in the remaining scenarios as well (Results see Table 14).

In general, PV tends to underestimate the coefficient compared to the Benchmark. The size of the underestimation depends on the scenario and the number of folds used in PV. The performance in the Linear-Linear model is very good with hardly any bias at all. For the Lasso-Linear and the LDA-Logistic scenario, the bias is bigger. The difference of the estimates for 5-fold and 10-fold PV show that at least part of the bias is due to the smaller training set used for deriving the prediction rule in PV. The bias also decreases with increasing number of observations, which can also be explained this way, as removing a certain percentage of observations has a smaller perturbing effect on the prediction rule when the total number of observations is large. Overall, PV does a good job of estimating the coefficient

of the new prediction rule.

6 Analysis of breast cancer data

Here we apply the permutation test to the dataset in van't Veer et al. (2002) and compare it to the analytical results. The data consists of microarray measurements on 4918 genes over 78 patients with breast cancer. 44 of these belong to the good prognosis group, 34 have a poor prognosis. Apart from the microarray data, a number of other clinical predictors exist:

- Tumor grade
- Estrogen receptor (ER) status
- Progestron receptor (PR) status
- Tumor size
- Patient age
- Angioinvasion

Based on the microarray data, a predictor for the cancer prognosis was constructed:

1. Select the 70 genes that have the highest correlation with the 78 class labels.
2. Find the centroid vector of the good prognosis group.
3. Compute the correlation of each case with the centroid of the good prognosis group. Find the cutoff such that only 3 cases in the poor prognosis group are misclassified.
4. Classify any new case as good prognosis if their correlation with the centroid is larger than the cutoff.

The result of the model fitting with and without using Pre-Validation can be found in Tables 2 and 3. We can immediately see how the significance of the microarray predictor is reduced when 10-fold PV is being used and thus the effect of fitting and testing the model on the same data removed. However, as PV chooses random folds, the results depend on the choice of folds. In order to get a clearer picture of the significance of the microarray predictor, we repeated the 10-fold PV 100 times and averaged the resulting p-values for the analytical and

Predictor	Coefficient	SD	Z-score	p-Value	Δ Deviance	p-Value (dev)
Microarray	4.0961	1.0921	3.751	0.000088[a]	25.016	$5.6 \cdot 10^{-7}$
Grade	-0.6974	1.0035	-0.695	0.487105	0.510	0.4750
ER	-0.5536	1.0444	-0.530	0.596041	0.282	0.5956
Angio	1.2085	0.8160	1.481	0.138613	2.290	0.1302
PR	1.2141	1.0569	1.149	0.250642	1.394	0.2378
Age	-1.5926	0.9113	-1.748	0.080549	3.478	0.0622
Size	1.4830	0.7322	2.026	0.042812	4.374	0.0365

[a]One sided test for pre-validated predictors and z-score

Table 2: *Summary of the coefficients in the external Logistic model without Pre-Validation. For each coefficient a test for $\beta = 0$ based on the z-score and the deviance is given. All p-values are for two-sided tests except for the z-score p-value of the Microarray predictor, which is a one-sided p-value for testing $\beta = 0$ versus $\beta > 0$.*

the permutation tests (see Table 4). The analytical test declares the microarray predictor to be significant, however all 3 permutation test statistics do not give significant results, though the difference of the analytical test to the z-score permutation test is quite small. A possible explanation for these different results is the bias of the analytical test.

7 Discussion

The problem often arises that, with a limited amount of data, one wants to find a prediction rule and verify its usefulness on the same dataset. Often, due to lack of power, doing full cross-validation is not feasible and pre-validation is a useful method to fill this gap, especially for finding reliable parameter estimates in the external model. However, we have found that using the standard analytical tests with the pre-validated predictor can yield a test with level above the nominal level.

The permutation test approach to the pre-validated predictor addresses the bias problem of the analytical test without compromising power and is therefore a more reliable way for assessing whether the new prediction rule is an improvement over previously established predictors. Its main drawback is that it is very computer-intensive, requiring us to refit the pre-validation model for every permutation. This can be a problem for especially large datasets. However, this will not often be a significant problem and the simple structure of the algorithm makes it easily accessible to parallelization to reduce computation time.

It might be possible to develop an analytical test that accounts for the special structure of

Predictor	Coefficient	SD	z-score	p-Value (z)	Δ Deviance	p-Value (dev)
Microarray	1.5449	0.7116	2.171	0.0150[a]	5.001	0.02533
Grade	0.5614	0.7473	0.751	0.4526	0.563	0.45299
ER	-0.6401	0.8967	-0.714	0.4754	0.517	0.47204
Angio	1.3466	0.6477	2.079	0.0376	4.568	0.03257
PR	0.4266	0.8336	0.512	0.6089	0.266	0.60603
Age	-1.4569	0.6925	-2.104	0.0354	4.817	0.02817
Size	0.8433	0.6026	1.400	0.1617	1.960	0.16156

[a]One sided test for pre-validated predictors and z-score

Table 3: *Summary of the coefficient in the external Logistic model using 10-Fold Pre-Validation. For each coefficient a test for* $\beta = 0$ *based on the z-score and the deviance is given. All p-values are for two-sided tests except for the z-score p-value of the Microarray predictor, which is a one-sided p-value for testing* $\beta = 0$ *versus* $\beta > 0$.

Statistic	Mean	$\% < 0.01$	$\% < 0.05$	$\% < 0.1$
Analytical z-score	0.046	15	66	91
Permutation with β	0.095	1	27	57
Permutation with z-score	0.050	17	62	86
Permutation with deviance	0.139	0	21	42

Table 4: *P-values for the microarray predictor over 100 runs of the Pre-Validation procedure. The mean values are reported as well as the percentage below the levels* 0.01, 0.05 *and* 0.1.

the pre-validated predictor. However, it is unclear if an analytical solution exists that holds for a large number of models. Since the internal models are usually tailored to the specific problem at hand, having to derive analytical solutions on a case by case basis would be very difficult. We believe that the permutation test is the best method currently available for the problem.

Acknowledgments Holger Höfling is supported by an Albion Walter Hewlett Stanford Graduate Fellowship.

Tibshirani was partially supported by National Science Foundation Grant DMS-9971405 and National Institutes of Health Contract N01-HV-28183.

A Proofs

A.1 Case of no outside predictors

For the proof, we first need a lemma:

Lemma 1. *Let X_{ij} be i.i.d. $N(0,1)$ for $i = 1, \ldots, n$ and $j = 1, \ldots, p$. Let $H = Proj(X) = X(X^T X)^{-1} X^T$ and $D = diag(H)$. Then $d_{ii} \sim O_P(n^{-1})$.*

Proof. By the strong law of large numbers,

$$\frac{1}{n} X^T X \to I_p \quad a.s.$$

and as taking the inverse of a matrix is a continuous operation

$$n(X^T X)^{-1} \to I_p \quad a.s.$$

Therefore

$$n d_{ii} = n x_i (X^T X)^{-1} x_i^T \to^d \chi_p^2$$

by continuous mapping, where x_i is the i-th row of X. □

Also note that as $trace(H) = \sum_i d_{ii} = p$ we have that $Cov(d_{ii}, d_{jj}) < 0 \quad \forall i \neq j$.

Now let us move on to the proof of Theorem 1.

Proof. Let the SVD of X be

$$X = UEV^T$$

with $U \in \mathbb{R}^{n\times p}$ orthogonal, $E \in \mathbb{R}^{p\times p}$ diagonal and $V \in \mathbb{R}^{p\times p}$ orthogonal. Then we can write $H = UU^T$, therefore the leave-one-out pre-validated predictor is

$$\tilde{y} = (I - D)^{-1}(UU^T - D)y$$

and

$$\hat{\beta}_{PV} = \frac{\tilde{y}^T y}{\tilde{y}^T \tilde{y}}.$$

Evaluating the numerator we get

$$\begin{aligned}\tilde{y}^T y &= y^T(UU^T - D)(I - D)^{-1}y = \\ &= y^T UU^T y + y^T(UU^T \left((I - D)^{-1} - I\right) y - y^T D((I - D)^{-1} - I)y - y^T Dy \to^d \\ &\to^d N^T N + 0 - p \quad \text{as} \quad n \to \infty\end{aligned}$$

where $N \sim N(0, I_p)$. This holds as $U^T y \sim N(0, I_p)$. The second term converges to 0 as $((I - D)^{-1} - I) \sim O_P(n^{-1})$ and $U^T y = N$ is bounded in probability. The third term converges to 0 in probability as $D((I - D)^{-1} - I) \sim O_P(n^{-2})$. For the fourth term observe that $E(y^T Dy) = E(E(y^T Dy|X)) = E(\sum d_{ii}) = p$. As $Cov(d_{ii}, d_{jj}) < 0$ for $i \neq j$, it is easy to show that $y^T Dy \to^P p$.

For the denominator we get

$$\begin{aligned}\tilde{y}^T \tilde{y} &= y^T(UU^T - D)(I - D)^{-2}(UU^T - D)y = \\ &= N^T N + N^T U^T((I - D)^{-2} - I)UN - 2y^T D(I - D)^{-2}UN + y^T D^2(I - D)^{-2}y.\end{aligned}$$

Here, the first term is $N^T N$ as above and the other terms converge to 0. The second and third summand converge to 0 as $(I - d)^{-2} - I \sim O_P(n^{-1})$ and $D(I - D)^{-2} \sim O_P(n^{-1})$ and for the fourth term we use that $D^2(I - D)^{-2} \sim O_P(n^{-2})$.

Now that we have the distribution of the numerator and denominator of $\hat{\beta}_{PV}$, consider $\hat{sd}(\hat{\beta}_{PV})$. This is estimated as

$$\hat{sd}(\hat{\beta}_{PV}0) = \hat{\sigma}\sqrt{\tilde{y}^T \tilde{y}}.$$

Only $\hat{\sigma}$ is left to treat, for which we can write

$$\begin{aligned}\hat{\sigma}^2 &= \frac{1}{n-1}(y - \hat{\beta}_{PV}\tilde{y})^T(y - \hat{\beta}_{PV}\tilde{y}) = \\ &= \frac{1}{n-1}\left(y^T y - 2\hat{\beta}_{PV}\tilde{Y}^t y + \hat{\beta}_{PV}^2 \tilde{y}^T \tilde{y}\right).\end{aligned}$$

We know that $\frac{1}{n-1}y^T y \to 1$ *a.s.*. The other terms go to 0 as it has been shown above that the second and third summand inside the bracket is bounded in probability.

So putting all this together yields the desired result. □

A.2 Case with outside predictors

The proof of Theorem 2 is along the lines of the proof for Theorem 1, but with more complicated algebra.

First recall a well known fact about the inverse of matrices. Assume we have a matrix with blocks of the form

$$M = \begin{pmatrix} A & B \\ C & D \end{pmatrix}$$

where A and D are non-singular square-matrices. Then we can write the inverse M^{-1} as

$$M^{-1} = \begin{pmatrix} (A - BD^{-1}C)^{-1} & -(A - BD^{-1}C)^{-1}BD^{-1} \\ -D^{-1}C(A - BD^{-1}C)^{-1} & D^{-1} + D^{-1}C(A - BD^{-1}C)^{-1}BD^{-1} \end{pmatrix}.$$

The proof of Theorem 2 is then:

Proof. Let $\beta = (\beta_{PV}, \beta_1^T)^T$ and $W = (\tilde{y}, Z)$. Then

$$\hat{\beta} = (W^T W)^{-1} W^T y$$

where

$$W^T W = \begin{pmatrix} \tilde{y}^T\tilde{y} & \tilde{y}^T Z \\ Z^T\tilde{y} & Z^T Z \end{pmatrix}$$

and as we are only interested in $\hat{\beta}_{PV}$, this can be written as

$$\hat{\beta}_{PV} = (\tilde{y}^T\tilde{y} - \tilde{y}^T Z(Z^T Z)^{-1}Z^T\tilde{y})^{-1}(\tilde{y}^T y - \tilde{y}^T Z(Z^T Z)^{-1}Z^T y),$$

using the formula for inverses of block matrices. Also define $\mathbf{1} = (1, \ldots, 1)^T \in \mathbb{R}^e$. Then

$$\begin{aligned} \frac{1}{n}Z^T Z &= \frac{1}{n}(y \cdot \mathbf{1}^T + \Gamma)^T(y \cdot \mathbf{1}^T + \Gamma) = \\ &= \frac{1}{n}(y^T y \mathbf{1}\mathbf{1}^T + 2 \cdot \mathbf{1} y^T \Gamma + \Gamma^T\Gamma) \to^P \\ &\to^P \mathbf{1}\mathbf{1}^T + 0 + Cov(\gamma) \end{aligned}$$

where $\Gamma_{ik} = \gamma_{ik}$ is the matrix of random errors of the external predictors and the convergence follows by the weak law of large numbers.

Also

$$\frac{1}{n} Z^T y = \frac{1}{n}(\mathbf{1} y^T y + \Gamma^T y) \to^P \mathbf{1} + 0$$

again using the weak law of large numbers and the independence of Γ and y. Furthermore

$$Z^T \tilde{y} = \mathbf{1} y^T \tilde{y} + \Gamma^T \tilde{y}.$$

As we already know that $y^T \tilde{y} \to^d N^T N - p$ where $N \sim N(0, I_p)$, we only have to determine the distribution of

$$\Gamma^T \tilde{y} = \Gamma^T (I - D)^{-1}(H - D)y = \Gamma^T (I - D)^{-1} U U^T y - \Gamma^T (I - D)^{-1} D y \to^d$$
$$\to^d A^T N - 0$$

where $N = U^T y \sim N(0, I_p)$ and $U^T (I - D)^{-1} \Gamma \to^d A = (A_1, \ldots, A_e)$ with $A_k \sim N(0, \sigma_k^2 \cdot I_p))$ i.i.d. So $Z^T \tilde{y}$ converges in distribution to

$$Z^T \tilde{y} \to^d N^T N - p + A^T N.$$

So combining the previous results we have

$$\tilde{y}^T Z (Z^T Z)^{-1} Z^T \tilde{y} = \frac{1}{n} \left(\tilde{y}^T Z \left(\frac{1}{n} Z^T Z \right)^{-1} Z^T \tilde{y} \right) \to^P 0$$

as the term inside the brackets is bounded in probability. Also

$$\tilde{y}^T Z (Z^T Z)^{-1} Z^T y = \tilde{y}^T Z \left(\frac{1}{n} Z^T Z \right)^{-1} \frac{1}{n} Z^T y \to^d$$
$$\to^d (\mathbf{1}^T (N^T N - p) + N^T A)(\mathbf{1}\mathbf{1}^T + Cov(\gamma))^{-1} \mathbf{1}.$$

Combining all this, we have that

$$\hat{\beta}_{PV} \to^d \frac{N^T N - p - (\mathbf{1}^T (N^T N - p) + N^T A)(\mathbf{1}\mathbf{1}^T + Cov(\gamma))^{-1} \mathbf{1}}{N^T N} =$$
$$= \frac{(N^T N - p)(1 - \mathbf{1}^T (\mathbf{1}\mathbf{1}^T + Cov(\gamma))^{-1} \mathbf{1}) - N^T A (\mathbf{1}\mathbf{1}^T + Cov(\gamma))^{-1} \mathbf{1}}{N^T N}.$$

In order to get the distribution of the t-statistic, the distribution of

$$\hat{sd}(\hat{\beta}_{PV}) = \sqrt{(W^T W)^{-1}_{11}} \hat{\sigma}$$

is needed. First, consider $(W^TW)^{-1}_{11}$:

$$(W^TW)^{-1}_{11} = (\tilde{y}^T\tilde{y} - \tilde{y}^TZ(Z^TZ)^{-1}Z^T\tilde{y})^{-1} \to^d (N^TN)^{-1}$$

as

$$\tilde{y}^TZ(Z^TZ)^{-1}Z^T\tilde{y} = \frac{1}{n}\tilde{y}^TZ\left(\frac{1}{n}Z^TZ\right)^{-1}Z^T\tilde{y} \to^P 0.$$

Next determine the asymptotic distribution of $\hat{\sigma}$:

$$\hat{\sigma} = \frac{1}{n-e-1}(y-\hat{y})^T(y-\hat{y}) = \frac{1}{n-e-1}\left(y^Ty - y^TW(W^TW)^{-1}W^Ty\right).$$

As before, $\frac{1}{n-e-1}y^Ty \to^P 1$. For the second term, first observe that

$$\frac{1}{n}W^Ty \to^P \begin{pmatrix} 0 \\ \mathbf{1} \end{pmatrix}$$

For $\frac{1}{n}(W^TW)^{-1}$ it is simple to show that all elements are asymptotically bounded in probability. For $\hat{\sigma}$, only the bottom right block is needed where

$$\frac{1}{n}(W^TW)^{-1}_{22} \to^P (\mathbf{1}\mathbf{1}^T + Cov(\gamma))^{-1} \quad \text{as} \quad n \to \infty.$$

Therefore

$$\hat{\sigma} \to^d 1 - \mathbf{1}^T(\mathbf{1}\mathbf{1}^T + Cov(\gamma))^{-1}\mathbf{1}.$$

Combining these results yields the claim. □

B Tables

Scenario	Stat.	α=.01				α=.05				α=.1			
		$K=1$	$K=5$	$K=10$	$K=n$	$K=1$	$K=5$	$K=10$	$K=n$	$K=1$	$K=5$	$K=10$	$K=n$
$n=10, p=5, k=1, \beta=0$	β perm.	0.009	0.004	0.011	0.006	0.060	0.044	0.051	0.041	0.114	0.104	0.099	0.103
	t perm.	0.010	0.008	0.013	0.008	0.057	0.044	0.047	0.049	0.113	0.090	0.104	0.098
$n=20, p=5, k=1, \beta=0$	β perm.	0.006	0.008	0.013	0.007	0.047	0.039	0.059	0.046	0.094	0.082	0.118	0.083
	t perm.	0.010	0.007	0.013	0.006	0.053	0.042	0.061	0.044	0.107	0.081	0.121	0.087
$n=50, p=5, k=1, \beta=0$	β perm.	0.011	0.008	0.010	0.011	0.047	0.055	0.044	0.059	0.108	0.100	0.089	0.112
	t perm.	0.013	0.009	0.010	0.011	0.048	0.056	0.042	0.061	0.105	0.100	0.091	0.114

Table 5: *Level of the Linear-Linear model under the null hypothesis for the permutation test based on β and t-statistic.*

Scenario	Stat.	α=.01			α=.05			α=.1		
		$K=1$	$K=5$	$K=10$	$K=1$	$K=5$	$K=10$	$K=1$	$K=5$	$K=10$
$n=10, p=100, k=1, \beta=0, s=5$	β perm.	0.008	0.016	0.004	0.054	0.053	0.028	0.099	0.098	0.083
	t perm.	0.012	0.014	0.016	0.053	0.058	0.062	0.104	0.098	0.124
$n=30, p=100, k=1, \beta=0, s=5$	β perm.	0.007	0.012	0.012	0.040	0.057	0.047	0.085	0.094	0.090
	t perm.	0.007	0.010	0.006	0.036	0.055	0.045	0.084	0.095	0.088
$n=30, p=100, k=1, \beta=0, s=10$	β perm.	0.009	0.009	0.008	0.062	0.053	0.050	0.114	0.105	0.103
	t perm.	0.006	0.009	0.011	0.061	0.053	0.059	0.115	0.105	0.103
$n=30, p=100, k=1, \beta=0, s=20$	β perm.	0.012	0.013	0.009	0.053	0.063	0.038	0.112	0.106	0.085
	t perm.	0.013	0.010	0.011	0.052	0.058	0.045	0.113	0.108	0.095

Table 6: *Level of the Lasso-Linear model under the null hypothesis for the permutation test based on β and t-statistic.*

Scenario	Stat.	α=.01			α=.05			α=.1		
		$K=1$	$K=5$	$K=10$	$K=1$	$K=5$	$K=10$	$K=1$	$K=5$	$K=10$
$n_1=n_2=10, p=1000, k=1, \mu=0, s=10$	β perm.	0.012	0.007	0.009	0.053	0.042	0.048	0.108	0.114	0.102
	z perm.	0.013	0.005	0.011	0.056	0.049	0.046	0.096	0.114	0.097
	dev. perm.	0.015	0.008	0.008	0.051	0.041	0.049	0.108	0.085	0.100
$n_1=n_2=20, p=1000, k=1, \mu=0, s=10$	β perm.	0.011	0.014	0.006	0.047	0.054	0.046	0.081	0.098	0.086
	z perm.	0.010	0.017	0.005	0.044	0.050	0.039	0.097	0.098	0.082
	dev. perm.	0.008	0.019	0.008	0.041	0.059	0.043	0.085	0.110	0.088
$n_1=n_2=40, p=1000, k=1, \mu=0, s=10$	β perm.	0.008	0.008	0.010	0.040	0.047	0.045	0.084	0.098	0.092
	z perm.	0.009	0.011	0.009	0.053	0.040	0.043	0.098	0.088	0.093
	dev. perm.	0.012	0.013	0.008	0.053	0.050	0.042	0.105	0.096	0.093

Table 7: *Level of the LDA-Logistic model under the null hypothesis for the permutation test based on β, t-statistic and deviance.*

Scenario	Stat.	α=.01				α=.05				α=.1			
		$K=1$	$K=5$	$K=10$	$K=n$	$K=1$	$K=5$	$K=10$	$K=n$	$K=1$	$K=5$	$K=10$	$K=n$
$n=10, p=5, k=1, \beta=.3$	β perm.	0.034	0.023	0.033	0.039	0.134	0.115	0.123	0.118	0.238	0.214	0.218	0.203
	t perm.	0.036	0.020	0.029	0.024	0.126	0.102	0.131	0.113	0.203	0.199	0.222	0.200
	analyt.	-	0.023	0.030	0.023	-	0.105	0.122	0.109	-	0.210	0.220	0.202
	test set	0.051	0.076	0.052	0.036	0.174	0.192	0.160	0.153	0.284	0.294	0.263	0.262
$n=20, p=10, k=1, \beta=.3$	β perm.	0.186	0.125	0.142	0.145	0.456	0.320	0.327	0.350	0.639	0.442	0.453	0.481
	t perm.	0.171	0.103	0.104	0.129	0.416	0.291	0.308	0.326	0.560	0.417	0.445	0.474
	analyt.	-	0.126	0.154	0.177	-	0.326	0.353	0.377	-	0.446	0.492	0.520
	test set	0.247	0.198	0.222	0.255	0.480	0.441	0.468	0.485	0.610	0.577	0.616	0.604
$n=50, p=10, k=1, \beta=.3$	β perm.	0.877	0.764	0.801	0.825	0.968	0.909	0.942	0.947	0.983	0.958	0.967	0.968
	t perm.	0.719	0.708	0.745	0.767	0.896	0.887	0.916	0.935	0.956	0.945	0.962	0.962
	analyt.	-	0.714	0.794	0.788	-	0.884	0.934	0.934	-	0.943	0.966	0.962
	test set	0.864	0.876	0.889	0.878	0.962	0.967	0.954	0.969	0.978	0.983	0.984	0.984
$n=20, p=10, k=5, \beta=.6$	β perm.	0.705	0.349	0.423	0.422	0.902	0.538	0.613	0.615	0.950	0.644	0.721	0.719
	t perm.	0.366	0.166	0.221	0.220	0.653	0.387	0.454	0.463	0.792	0.538	0.597	0.592
	analyt.	-	0.148	0.194	0.187	-	0.349	0.439	0.438	-	0.517	0.590	0.584
	test set	0.317	0.281	0.289	0.292	0.573	0.536	0.591	0.581	0.693	0.685	0.712	0.717
$n=50, p=10, k=5, \beta=.6$	β perm.	0.998	0.989	0.982	0.985	0.999	0.995	0.995	0.998	1.000	0.998	0.996	1.000
	t perm.	0.908	0.904	0.910	0.924	0.976	0.980	0.977	0.982	0.989	0.993	0.992	0.992
	analyt.	-	0.876	0.901	0.908	-	0.970	0.975	0.975	-	0.989	0.993	0.991
	test set	0.943	0.955	0.948	0.960	0.988	0.994	0.986	0.993	0.995	0.998	0.996	0.996

Table 8: *Power estimation for the Linear-Linear model for alternative scenarios. The permutation test was performed for the β and t-statistic. The power for the analytical test is corrected for the bias. The power when using an independent test set of size n is given as well.*

Scenario	Stat.	α=.01			α=.05			α=.1		
		$K=1$	$K=5$	$K=10$	$K=1$	$K=5$	$K=10$	$K=1$	$K=5$	$K=10$
$n=10, p=100, k=1, \beta=rep(1,5), s=5$	β perm.	0.008	0.007	0.024	0.041	0.063	0.081	0.103	0.129	0.147
	t perm.	0.015	0.007	0.023	0.071	0.070	0.086	0.122	0.127	0.141
	analyt.	-	0.008	0.030	-	0.072	0.084	-	0.133	0.143
	test set	0.047	0.031	0.039	0.123	0.120	0.123	0.195	0.199	0.195
$n=30, p=100, k=1, \beta=rep(1,5), s=5$	β perm.	0.336	0.231	0.248	0.554	0.428	0.447	0.669	0.560	0.561
	t perm.	0.244	0.207	0.226	0.417	0.421	0.434	0.551	0.555	0.552
	analyt.	-	0.209	0.252	-	0.415	0.453	-	0.551	0.566
	test set	0.501	0.529	0.531	0.705	0.726	0.714	0.785	0.807	0.800
$n=30, p=100, k=1, \beta=rep(5,5), s=5$	β perm.	0.655	0.433	0.503	0.802	0.611	0.693	0.859	0.714	0.773
	t perm.	0.542	0.396	0.458	0.704	0.599	0.674	0.791	0.706	0.763
	analyt.	-	0.399	0.495	-	0.604	0.690	-	0.702	0.774
	test set	0.773	0.745	0.724	0.877	0.873	0.858	0.918	0.920	0.897
$n=30, p=100, k=1, \beta=rep(1,10), s=10$	β perm.	0.097	0.145	0.158	0.330	0.315	0.346	0.487	0.455	0.487
	t perm.	0.115	0.128	0.140	0.332	0.312	0.324	0.462	0.447	0.477
	analyt.	-	0.139	0.198	-	0.305	0.417	-	0.432	0.548
	test set	0.352	0.351	0.330	0.566	0.569	0.550	0.704	0.694	0.664
$n=30, p=100, k=1, \beta=rep(5,10), s=10$	β perm.	0.152	0.195	0.217	0.414	0.389	0.418	0.562	0.508	0.538
	t perm.	0.186	0.167	0.189	0.402	0.363	0.410	0.549	0.505	0.530
	analyt.	-	0.168	0.276	-	0.374	0.488	-	0.504	0.601
	test set	0.464	0.469	0.473	0.672	0.670	0.708	0.773	0.759	0.802

Table 9: *Power estimation for the Lasso-Linear model for alternative scenarios. The permutation test was performed for the β and t-statistic. The power for the analytical test is corrected for the bias. The power when using an independent test set of size n is given as well. Here, s is the number of non-zero predictors.*

Scenario	Stat.	α=.01			α=.05			α=.1		
		$K=1$	$K=5$	$K=10$	$K=1$	$K=5$	$K=10$	$K=1$	$K=5$	$K=10$
$n_1=n_2=10, p=1000, k=1, \mu=rep(1,10), s=10$	β perm.	0.002	0.033	0.050	0.018	0.159	0.175	0.051	0.243	0.252
	z perm.	0.004	0.041	0.037	0.033	0.163	0.147	0.091	0.269	0.259
	dev. perm.	0.012	0.030	0.039	0.053	0.094	0.100	0.107	0.157	0.149
	analyt.	-	0.039	0.029	-	0.176	0.216	-	0.274	0.326
	test set	0.028	0.039	0.022	0.341	0.337	0.341	0.533	0.515	0.523
$n_1=n_2=10, p=1000, k=1, \mu=rep(2,10), s=10$	β perm.	0.000	0.629	0.653	0.000	0.790	0.872	0.001	0.891	0.923
	z perm.	0.000	0.169	0.073	0.031	0.326	0.189	0.124	0.397	0.255
	dev. perm.	0.003	0.722	0.726	0.022	0.886	0.882	0.038	0.930	0.936
	analyt.	-	0.556	0.565	-	0.735	0.721	-	0.802	0.794
	test set	0.717	0.735	0.720	0.759	0.766	0.766	0.793	0.799	0.803
$n_1=n_2=20, p=1000, k=1, \mu=rep(1,10), s=10$	β perm.	0.004	0.433	0.415	0.022	0.728	0.748	0.063	0.862	0.869
	z perm.	0.002	0.526	0.386	0.007	0.785	0.762	0.016	0.868	0.857
	dev. perm.	0.105	0.489	0.404	0.262	0.683	0.652	0.387	0.767	0.759
	analyt.	-	0.538	0.756	-	0.790	0.883	-	0.881	0.920
	test set	0.835	0.836	0.848	0.954	0.968	0.965	0.980	0.987	0.989
$n_1=n_2=40, p=1000, k=1, \mu=rep(1,10), s=10$	β perm.	0.351	0.999	0.996	0.436	1.000	1.000	0.495	1.000	1.000
	z perm.	0.001	0.927	0.841	0.002	0.983	0.956	0.002	0.992	0.979
	dev. perm.	0.884	0.999	0.998	0.977	0.999	1.000	0.991	0.999	1.000
	analyt.	-	0.968	0.972	-	0.992	0.991	-	0.998	0.997
	test set	0.954	0.949	0.932	0.988	0.987	0.980	0.995	0.995	0.992

Table 10: *Power estimation for the LDA-Logistic model for alternative scenarios. The permutation test was performed for the β, t, and deviance-statistic. The power for the analytical test is corrected for the bias. The power when using an independent test set of size n is given as well. Here s is the number of significant genes.*

Scenario	α=.01				α=.05				α=.1			
	$K=1$	$K=5$	$K=10$	$K=n$	$K=1$	$K=5$	$K=10$	$K=n$	$K=1$	$K=5$	$K=10$	$K=n$
$n=10, p=5, k=1, \beta=.3$	0.390	0.051	0.059	0.062	0.669	0.175	0.181	0.168	0.805	0.271	0.294	0.264
$n=20, p=10, k=1, \beta=.3$	0.954	0.186	0.205	0.227	0.990	0.380	0.399	0.443	0.999	0.497	0.537	0.550
$n=50, p=10, k=1, \beta=.3$	0.996	0.805	0.838	0.854	0.999	0.921	0.943	0.947	1.000	0.966	0.971	0.973
$n=20, p=10, k=5, \beta=.6$	0.844	0.206	0.257	0.269	0.944	0.427	0.491	0.500	0.972	0.570	0.630	0.626
$n=50, p=10, k=5, \beta=.6$	0.992	0.927	0.928	0.936	1.000	0.983	0.980	0.986	1.000	0.993	0.993	0.993

Table 11: *Unadjusted power for the analytical test for the Linear-Linear model for alternative scenarios.*

Scenario	α=.01			α=.05			α=.1		
	$K=1$	$K=5$	$K=10$	$K=1$	$K=5$	$K=10$	$K=1$	$K=5$	$K=10$
$n=10, p=100, k=1, \beta=rep(1,5), s=5$	0.925	0.006	0.032	0.981	0.042	0.070	0.992	0.081	0.110
$n=30, p=100, k=1, \beta=rep(1,5), s=5$	0.986	0.228	0.268	0.996	0.381	0.424	0.998	0.480	0.504
$n=30, p=100, k=1, \beta=rep(5,5), s=5$	0.993	0.424	0.503	0.998	0.575	0.661	1.000	0.652	0.736
$n=30, p=100, k=1, \beta=rep(1,10), s=10$	1.000	0.168	0.209	1.000	0.325	0.384	1.000	0.431	0.482
$n=30, p=100, k=1, \beta=rep(5,10), s=10$	1.000	0.226	0.289	1.000	0.389	0.443	1.000	0.485	0.537

Table 12: *Unadjusted power for the analytical test for the Lasso-Linear model for alternative scenarios.*

Scenario	α=.01			α=.05			α=.1		
	$K=1$	$K=5$	$K=10$	$K=1$	$K=5$	$K=10$	$K=1$	$K=5$	$K=10$
$n_1=n_2=10, p=1000, k=1, \mu=rep(1,10), s=10$	0.341	0.011	0.025	0.341	0.094	0.151	0.341	0.224	0.259
$n_1=n_2=10, p=1000, k=1, \mu=rep(2,10), s=10$	0.413	0.438	0.559	0.413	0.647	0.715	0.413	0.776	0.794
$n_1=n_2=20, p=1000, k=1, \mu=rep(1,10), s=10$	0.852	0.636	0.724	0.924	0.836	0.879	0.959	0.907	0.920
$n_1=n_2=40, p=1000, k=1, \mu=rep(1,10), s=10$	0.735	0.979	0.968	0.895	0.996	0.990	0.953	0.998	0.997

Table 13: *Unadjusted power for the analytical test for the LDA-Logistic model for alternative scenarios.*

Scenario	Parameters	Median of β using			% Difference of
		5-Fold PV	10-Fold PV	Benchmark	10-Fold PV - Benchmark
Linear-Linear	$n=10, p=5, k=1, \beta=.3$	0.153	0.182	0.188	3.4
	$n=20, p=10, k=1, \beta=.3$	0.219	0.237	0.278	14.6
	$n=50, p=10, k=1, \beta=.3$	0.356	0.362	0.382	5.2
	$n=20, p=10, k=5, \beta=.6$	0.201	0.240	0.275	12.7
	$n=50, p=10, k=5, \beta=.6$	0.342	0.349	0.368	5.1
Lasso-Linear	$n=10, p=100, k=1, \beta=rep(1,5), s=5$	-0.0910	-0.0617	0.0919	167.0
	$n=30, p=100, k=1, \beta=rep(1,5), s=5$	0.163	0.180	0.337	46.5
	$n=30, p=100, k=1, \beta=rep(5,5), s=5$	0.275	0.325	0.477	31.9
	$n=30, p=100, k=1, \beta=rep(1,10), s=10$	-	0.171	0.253	32.4
	$n=30, p=100, k=1, \beta=rep(5,10), s=10$	0.170	0.200	0.321	37.5
LDA-Logistic	$n_1=n_2=10, p=1000, k=1, \mu=rep(1,10), s=10$	0.00391	0.00887	0.05800	84.7
	$n_1=n_2=10, p=1000, k=1, \mu=rep(2,10), s=10$	0.211	0.663	1.250	46.8
	$n_1=n_2=20, p=1000, k=1, \mu=rep(1,10), s=10$	0.191	0.250	0.353	29.1
	$n_1=n_2=40, p=1000, k=1, \mu=rep(1,10), s=10$	0.590	0.648	0.774	16.3

Table 14: *Median value of the coefficient of the new prediction rule in the external model. The median value is given for 5-Fold PV, 10-Fold PV as well as the benchmark when the whole dataset is used and evaluation of the new rule w.r.t. the old rules is done on an independent test set of the same size.*

References

S. Dudoit, J. Fridlyand, and T. Speed. Comparison of discrimination methods for the classification of tumors using gene expression data. *J. Amer. Statist. Assoc.*, pages 1151–1160, 2001.

B. Efron. How biased is the apparent error rate of a prediction rule? *J. Amer. Statist. Assoc.*, 81:461–470, 1986.

B. Efron and R. Tibshirani. *An Introduction to the Bootstrap*. Chapman and Hall, London.

B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *Annals of Statistics*, 32(2):407–499, 2004.

T. Hastie and R. Tibshirani. *Generalized Additive Models*. Chapman and Hall, London, 1990.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer, 1st edition, 2001.

R. Tibshirani. Regression shrinkage and selection via the lasso. *J. Royal. Statist. Soc. B*, 58: 267–288, 1996.

Robert J. Tibshirani and Bradley Efron. Pre-validation and inference in microarrays. *Statistical Applications in Genetics and Molecular Biology*, 1:1–18, 2002.

L.J. van't Veer, H.D.M.J. van de Vijver, Y.D. He, A.A. Hart, M. Mao, H.L. Peterse, K. van der Kooy, M.J. Marton, G.J.S. Anke T. Witteveen, R.M. Kerkhoven, C. Roberts, P.S. Linsley, R. Bernards, and S.H. Friend. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415:530–536, 2002.

J. Ye. On measuring and correcting the effects of data mining and model selection. *J. Amer. Statist. Assoc.*, 93:120–131, 1998.