List 2 - Prediction error and information criteria

Statistical learning

Paulina Podgórska

Let us generate the design matrix $X_{n\times p} = X_{1000\times 950}$ such that its elements are iid random variables from $N(0, \sigma = \frac{1}{\sqrt{1000}})$. Then, we generate the vector of values of the response variable $Y = X\beta + \epsilon$, where $\beta = (3, 3, 3, 3, 3, 0, \ldots, 0)^T$ and $\epsilon \sim N(0, I)$.
We will perform analysis using 6 models containing only:

- the first 2 columns of X $(p = 2)$,
- the first 5 columns of X $(p = 5)$,
- the first 10 columns of X $(p = 10)$,
- the first 100 columns of X $(p = 100)$,
- the first 500 columns of X $(p = 500)$,
- all 950 columns $(p = 950)$.

We will start our data analysis with estimating β with the Least Squares method and calculating the residual sum of squares:

$$RSS = ||\hat{Y} - Y||^2 = \sum_{i=1}^{n} e_i^2$$

.

Table 1: RSS of our 6 models

	X_2	X_5	X_{10}	X_{100}	X_{500}	X_{950}
RSS	1079.735	1044.011	1038.788	923.858	509.082	61.919

The model with 2 variables has the highest RSS value. As the number of variables increases, the RSS decreases significantly. The RSS difference between the X_2 and the X_{950} model is: 1017.82.

Prediction error

The formula of the true expected value of the prediction error, conditional on the training sample is as follows:

$$PE = E_{\epsilon^*}||X(\beta - \hat{\beta}) + \epsilon^*||,$$

where $\epsilon^* \sim N(0, I)$ is a new noise vector, independent on the training sample.

Prediction error estimation

We can estimate the prediction error in three ways:

1. Using RSS with the true value of σ: $\widehat{PE}_1 = RSS + 2\sigma^2 p$, where p - number of variables in the model,

2. Using RSS with the regular unbiased estimator $\hat{\sigma}$: $\widehat{PE}_2 = RSS + 2\hat{\sigma}^2 p$, where $\hat{\sigma}^2 = \frac{RSS}{n-p}$,
3. Using leave-one-out cross-validation:
   $CV = \sum_{i=1}^{n} (\frac{Y_i - \hat{Y}_i}{1 - M_{ii}})$, where $M = X_k(X_k'X_k)^{-1}X_k', X_k = (X^{(1)} \cdots X^{(k)})$.

Table 2: Prediction error estimation

	X_2	X_5	X_{10}	X_{100}	X_{500}	X_{950}
PE	978.5455	969.3322	963.8175	1038.674	1422.538	1924.373
$\widehat{PE}_1$	1083.7347	1054.0115	1058.7876	1123.858	1509.082	1961.919
$\widehat{PE}_2$	1084.0623	1054.5041	1059.7732	1129.160	1527.246	2414.840
CV	1083.9293	1054.4208	1059.5118	1142.199	2045.779	26481.590

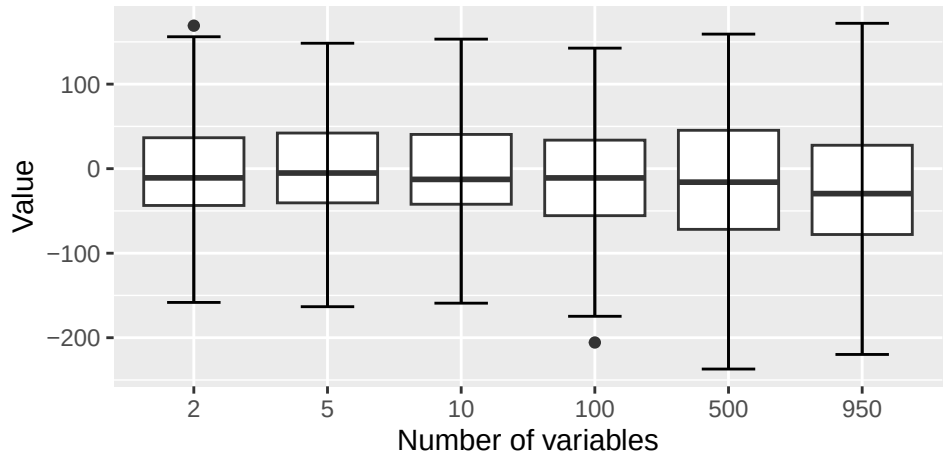
The estimator $\widehat{PE}_1$ determined using RSS with the true value of σ has the values closest to the true expected values of PE. For a number of variables greater than 100, the CV estimator achieves much higher values compared to the others - for X_{950} PE is over 10 times higher.

The model is optimal when its prediction error is small. Thus, we can see that X_5 and X_{10} are the best - they have the smallest prediction error (for each estimator). On the other hand, the model with 950 variables is significantly worse.

100 repetitions

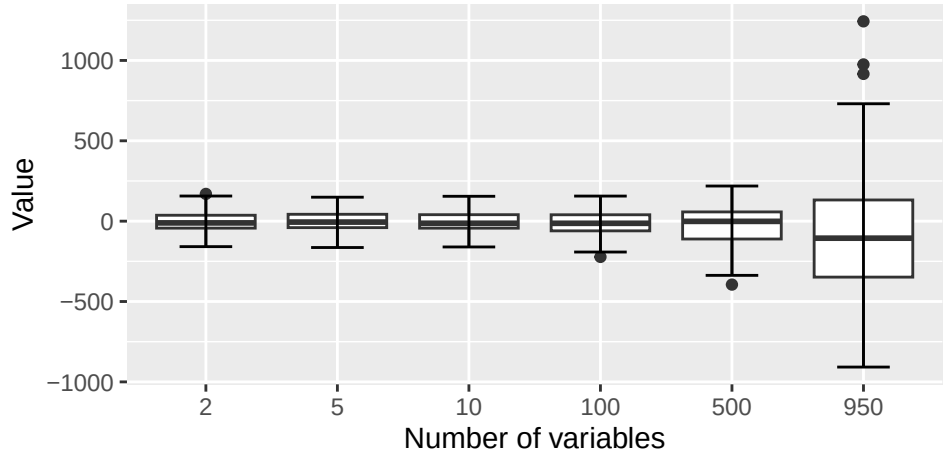
To better test the differences between the true expected value of PE and its estimators, we will repeat above analysis 100 times and create boxplots of $\widehat{PE} - PE$ for each model.

$\widehat{PE}_1 - PE$



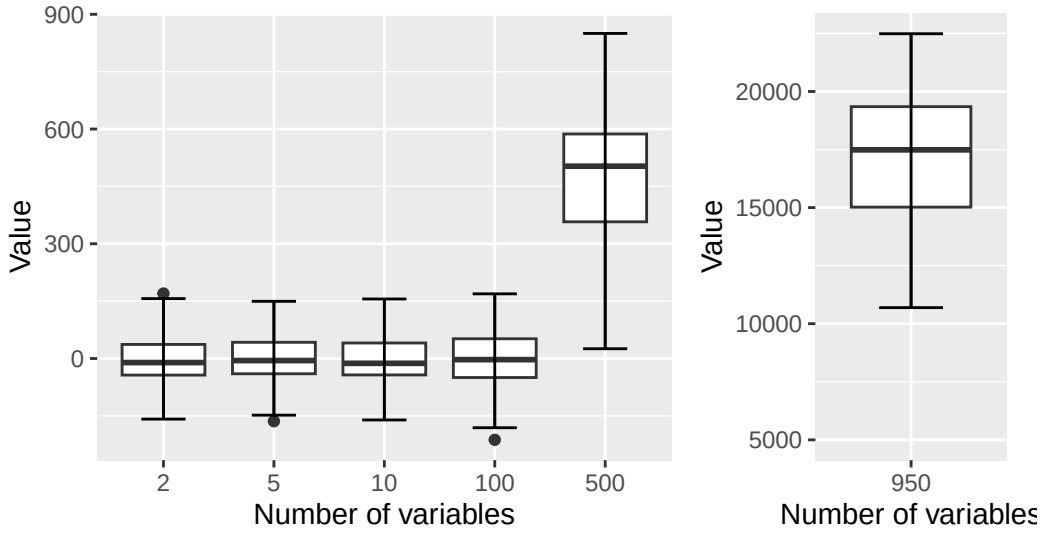
For $\widehat{PE}_1 - PE$ the interquartile range is similar for each of the models. The median value is close to 0 in each case. We can conclude from this that the difference between $\widehat{PE}_1$ and PE is at a similar level, regardless of the number of variables in the model.

$\widehat{PE_2} - PE$



Based on the box plots, we can infer that for models with a number of variables ≤ 500, the $\widehat{PE_2}$ estimator performs similarly to $\widehat{PE_1}$. This conclusion is supported by the similar median, Q1 and Q3 values. In all cases, the predicted value of PE is not significantly different from the expected value of PE. However, for X_{950}, the $\widehat{PE_2}$ estimator differs from PE, with a larger interquartile range and more extreme minimum and maximum values.

$CV - PE$



The CV estimator produces results that are significantly different from the expected values. In the case of 500 variables, the median of the difference is above 500, the IQR and the min and max values are large. With 950 variables, $CV - PE$ is very large, with a median close to 20 000. This estimation method performs the worst for models with a large number of variables. The boxplots confirm that the estimator that gives the results closest to the expected value is $\widehat{PE_1}$.

Model selection using AIC

The AIC criterion is used to choose between models with different numbers of predictors. It is also an indicator of the fit of the model to the data. We will consider three versions of AIC:

- AIC_1 – with known σ: we will choose a model that minimizes $RSS + 2\sigma^2 k$,
- AIC_2 – with σ estimated by the unbiased estimator $\hat{\sigma}$: we will choose a model that minimizes $RSS + 2k\frac{RSS}{n-p}$,
- AIC_3 – version for unknown σ: we will choose a model that minimizes $n \cdot log(RSS) + 2k$.

Table 3: The number of variables in the optimal model selected by three versions of AIC

	AIC_1	AIC_2	AIC_3
Number of variables	128	161	161

100 repetitions

Let us repeat above model selections 100 times in order to calculate the average number of false negatives and false positives produced by all three versions of AIC.

Table 4: Average number of false positives and false negatives for each method

	AIC_1	AIC_2	AIC_3
False positives	114.01	165.67	168.09
False negatives	0.46	0.45	0.45

BIC, AIC, RIC, mBIC, mBIC2

Next, let's consider additional information criteria for selecting significant variables to include in the model.

- BIC – we choose a model that minimizes $RSS + \sigma^2 klog(n), n > 8$,
- RIC – we choose a model that minimizes $RSS + 2k\sigma^2 log(p)$,
- mBIC – we choose a model that minimizes $log(RSS) + klog(n) + 2klog(\frac{p}{c})$, where c - average number of significant variables,
- mBIC2 – we choose a model that minimizes $log(RSS) + klog(n) + 2klog(\frac{p}{c}) - 2log(k!)$.

We will perform analysis using 4 models containing: 20, 100, 500 and 950 first variables.

False and true discoveries, square error

The square error of the estimation of the vector of expected values of Y:

$$SE = ||X\beta - \hat{Y}||^2.$$

Table 5: 20 and 100 variables

Criterion	TD	FD	SE	Criterion	TD	FD	SE
AIC	5	1	10.42	AIC	5	21	73.88
BIC	4	0	15.61	BIC	4	0	15.61
RIC	5	1	10.42	RIC	4	0	15.61
mBIC	4	0	15.61	mBIC	4	0	15.61
mBIC2	4	0	15.61	mBIC2	4	0	15.61

Table 6: 500 and 950 variables

Criterion	TD	FD	SE	Criterion	TD	FD	SE
AIC	5	65	231.44	AIC	5	65	231.44
BIC	4	10	100.72	BIC	4	15	136.59
RIC	4	1	29.01	RIC	4	0	15.61
mBIC	2	0	30.72	mBIC	1	0	36.90
mBIC2	2	0	30.72	mBIC2	1	0	36.90

AIC resulted in a relatively high number of true and false discoveries. In contrast, mBIC and mBIC2 criteria achieved a lower number of false discoveries, but in the presence of a large number of variables, the number of true discoveries was limited and the square error of estimation was higher. In comparison, RIC achieved the smallest squared error value, with a good trade-off between true and false discoveries. Thus, the RIC criterion performed best.

100 repetitions

Table 7: 20 and 100 variables

Criterion	TD	FD	FDR	Power	MSE	Criterion	TD	FD	FDR	Power	MSE
AIC	4.68	2.23	0.297	0.936	14.810	AIC	4.68	14.94	0.753	0.936	60.149
BIC	3.08	0.16	0.038	0.616	20.827	BIC	3.08	0.99	0.221	0.616	27.853
RIC	4.58	1.83	0.258	0.916	14.636	RIC	2.47	0.38	0.122	0.494	28.749
mBIC	1.97	0.03	0.018	0.394	29.562	mBIC	1.06	0.02	0.010	0.212	36.957
mBIC2	2.34	0.10	0.036	0.468	26.838	mBIC2	1.20	0.04	0.018	0.240	36.051

Table 8: 500 and 950 variables

Criterion	TD	FD	FDR	Power	MSE	Criterion	TD	FD	FDR	Power	MSE
AIC	4.68	65.25	0.933	0.936	219.260	AIC	4.68	65.32	0.933	0.936	219.434
BIC	3.08	4.48	0.575	0.616	56.470	BIC	3.08	8.43	0.725	0.616	88.601
RIC	1.26	0.17	0.078	0.252	37.281	RIC	0.93	0.21	0.109	0.186	40.920
mBIC	0.53	0.03	0.020	0.106	41.252	mBIC	0.41	0.04	0.030	0.082	42.434
mBIC2	0.56	0.04	0.025	0.112	41.132	mBIC2	0.41	0.05	0.033	0.082	42.602

In terms of power, the AIC criterion outperforms all other criteria in all cases. However, it also yields a high rate of false discoveries and a large mean squared error. The BIC criterion comes in second in terms of power, with a large number of true discoveries and significantly fewer false discoveries. Other criteria show similar performance, particularly when the number of variables is small.

To minimize false discoveries, the mBIC or mBIC2 criterion is the best choice, with the lowest false discovery rate at the cost of lower power. If the goal is to achieve the most optimal result, the BIC criterion performs well, with a high number of true discoveries and low standard error but relatively high false discovery rate. Lastly, if the goal is to identify as many true discoveries as possible, even at the cost of many false discoveries, the AIC criterion is the best choice.

realdata.Rdata

We will test the AIC, BIC, RIC, mBIC, and mBIC2 criteria on real data. Our data set contains the expression levels of 3221 genes for 210 individuals, which we have split into a test and a train set. The test set comprises 30 randomly selected individuals. We then construct a regression model to explain the expression level of gene 1 as a function of the expression levels of other genes and use the mentioned criteria to select explanatory variables. Let us see how many variables were selected by each criterion.

Table 9: Number of variables selected by criteria

	AIC	BIC	RIC	mBIC	mBIC2
Number of variables	193	194	7	6	8

Now let's test the accuracy of our models predictions on the test set using square error.

Table 10: SE

	AIC	BIC	RIC	mBIC	mBIC2
SE	31.7597	1.8156	0.0638	0.1101	0.0082

Comparing different criteria for selecting explanatory variables, we see that AIC performs poorly. The Bayesian Information Criterion also does not yield satisfactory results. In contrast, RIC and mBIC show promise with smaller errors. However, mBIC2 criterion emerges as the most effective, delivering the best prediction accuracy among all the criteria tested.