Report 2

Selecting Features with Information Criteria

Klaudia Balcer

12/17/2021

Contents

Introduction

When considering statistical regression models, our goal is to get good predictions for new variables and detect real relationships in data. As it is about linear regression, we need to select variables well (in order to find relationships) and estimate the coefficients (to get good predictions for new observations). One can use various methods to select the variables. In the report we will study the properties of several information criteria, both using real and simulated data.

The first two tasks are simulation studies. We will consider a model with five sufficient variables and a different (rather large) number of insufficient variables. First, we will see how the prediction error behaves when selecting insufficient variables. In the second task, we will compare the properties of different information criteria. In the last task, we will make a case study with data from the genetics domain.

Task 1

We will simulate the design matrix from the normal distribution:

$$X_{1000\times 950} \sim \mathcal{N}\Big(0, \sigma = \frac{1}{\sqrt{1}000}\Big)$$

i.i.d.

The real vector of coefficients have five signals of strength tree and different number $(q-5)$ of zero entries:

$$\beta = (3, 3, 3, 3, 3, 0, \ldots, 0)^T$$

.

Random noise will be added to the data:

$\epsilon \sim \mathcal{N}(\vec{0}, \mathbb{I})$.

The length of the vector of coefficients:

- $q = 2, 5, 10, 100, 500, 950$.

The number of repetitions of the experiment: 100.

Prediction Error As discussed in the introduction, looking only at the fit of the model may lead to overfitting. It is important to look at how our model behaves for new data. A good measurement is the prediction error. It is the expected norm of the difference between the estimated responses and values of the response variable generated by the same design matrix and a new error term. In a perfect situation, the MSE of the estimators should be the same for the data used for training and the new data (obtained by using a new error term). We will observe the prediction error and its estimators for different qs.

Prediction error:

$$PE = \mathbb{E}\|X(\beta - \hat{\beta}) + \epsilon^*\|^2,$$

where $\epsilon^* \sim \mathcal{N}(\vec{0}, \mathbb{I})$ is a new noise vector.

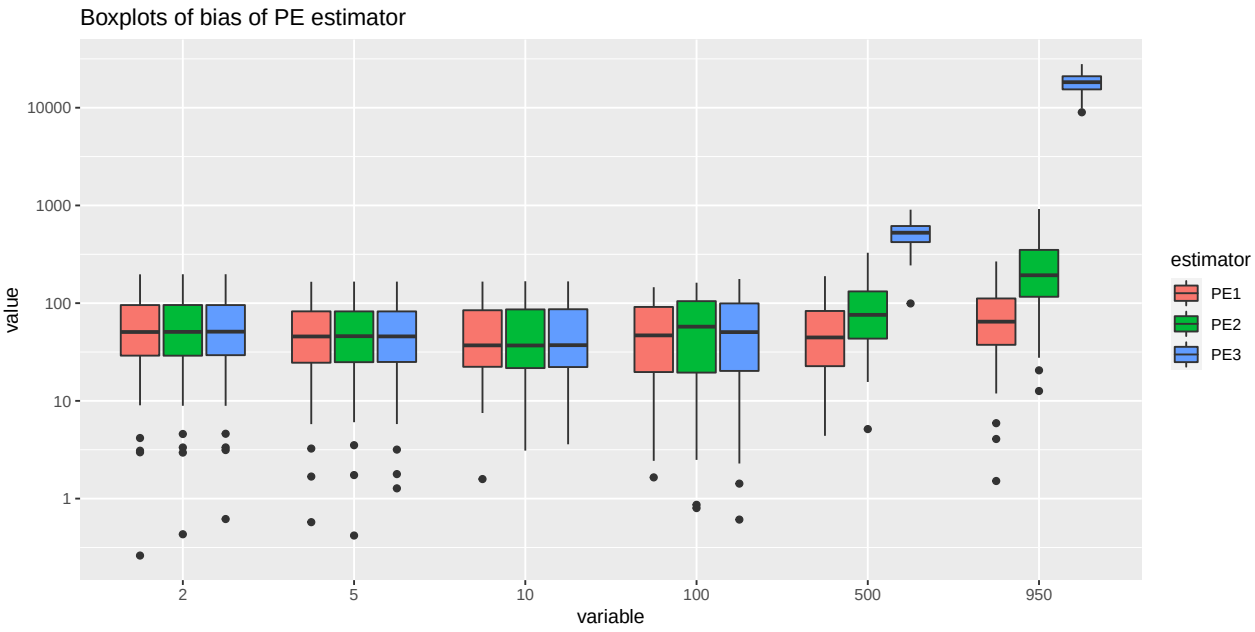
SURE estimator of the PE (in orthogonal design):

- σ known:
$$\widehat{PE}_1 = RSS + 2\sigma^2 p$$
- σ unknown:
$$\widehat{PE}_2 = RSS + 2\hat{\sigma}^2 p = \frac{(n+p)RSS}{n-p}$$

Leave-one-out CV PE estimator:

$$\widehat{PE}_3 = \sum_i^n \left(\frac{Y_i - \hat{Y}_i}{1 - H_{i,i}}\right)^2,$$

where $H = X(X'X)^{-1}X'$ is the projection matrix.



When increasing the dimension of the design matrix (the number of columns p), the fit is always better. However, adding insufficient variables may cause the prediction properties to worsen despite better fit. Such a situation can be called overfitting. We can observe it at the prediction error estimators' boxplots.

AIC Optimizing the Akkaike Infomation Criteria (AIC) corresponds with a decrease in the prediction error. The geenral formula of the criterion is:

$$AIC(M_p) = ln\mathcal{L}(X, \hat{\theta}_{MLE}) - p,$$

where p is the length of the parameter vector ($\theta \in \mathbb{R}^p$).

When sigma is known, the AIC can be calculated using the formula:

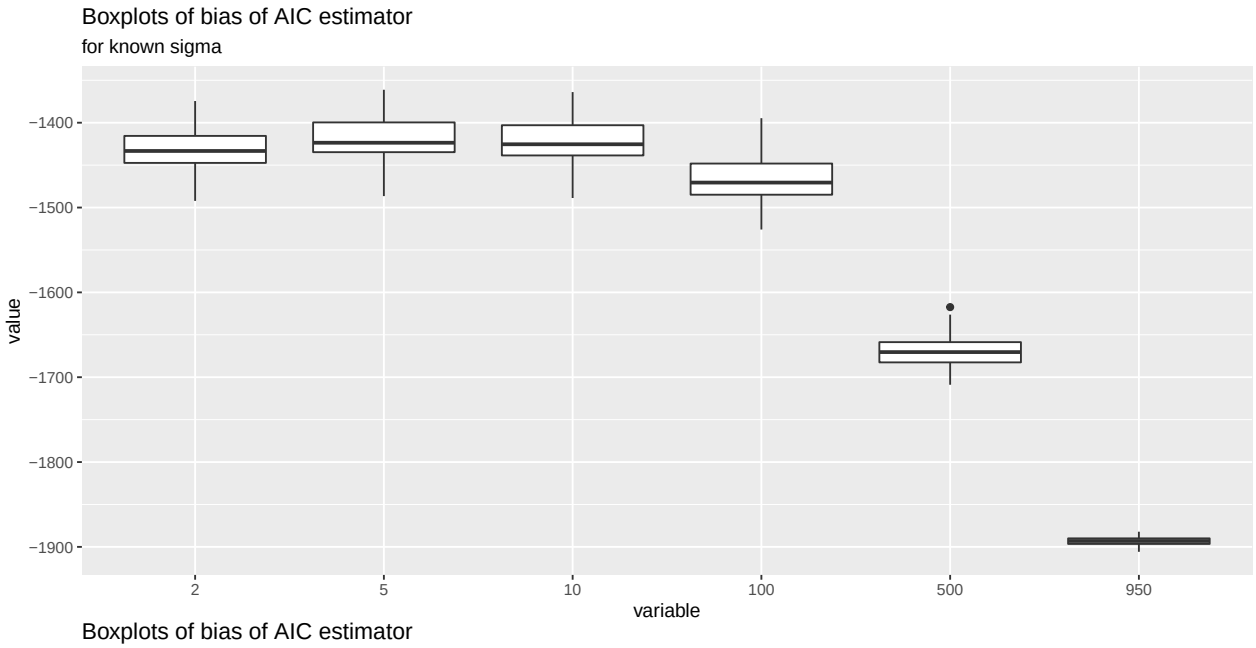
$$AIC(M_k) = C_{n,\sigma} - \frac{RSS}{2\sigma^2} - k$$

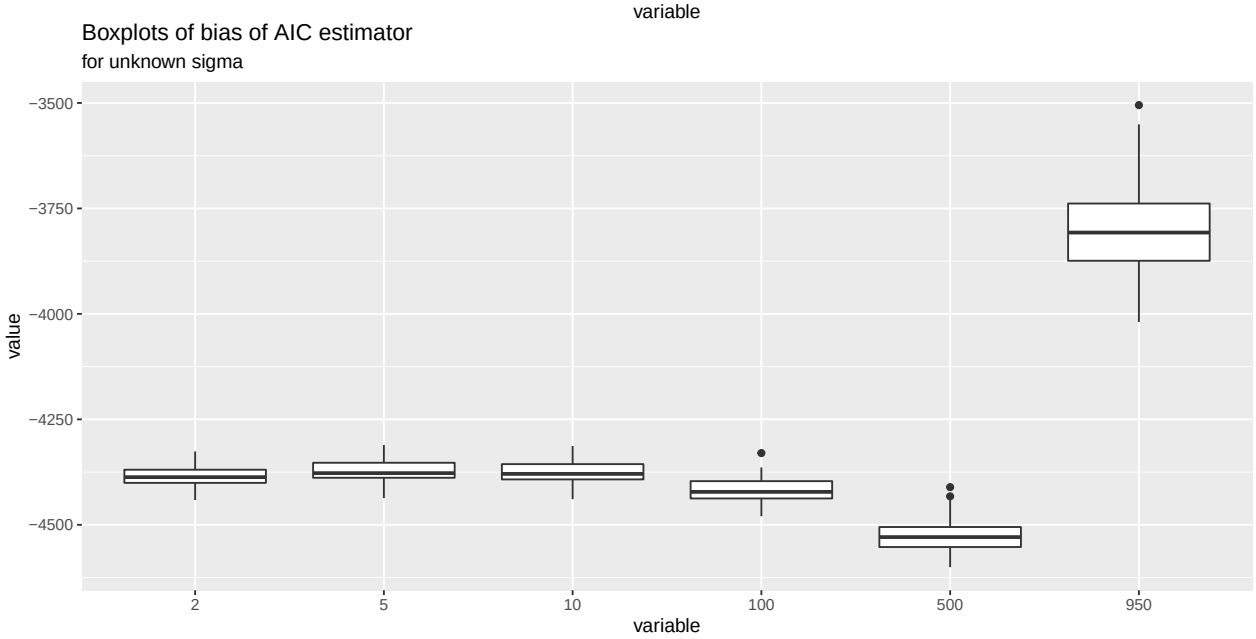
Maximizing AIC is equivalent to minimizing $\frac{RSS}{2\sigma^2} + k$.

When sigma is unknown, the AIC can be calculated using the formula:

$$AIC(M_k) = C_n - \frac{n}{2}log(RSS) - k$$

Maximizing AIC is equivalent to minimizing $\frac{n}{2}log(RSS) + k$.





When σ is known, the AIC criterion leads the proper model (we obtain the largest AIC for the model with only significant variables). For unknown σ the criterion has a strange aberration for $p = 950$ and leads to the model with the most insufficient variables.

Table 1: Results for BIC criterion

	20	100	500	950
FP	0.150	0.920	4.350	9.570
FN	1.690	1.740	1.720	1.780
power	0.662	0.652	0.656	0.644
SSE	20.163	26.734	55.760	97.839

Table 2: Results for AIC criterion

	20	100	500	950
FP	2.370	16.000	112.340	0.110
FN	0.280	0.300	0.420	4.110
power	0.944	0.940	0.916	0.178
SSE	15.797	63.617	343.940	41.097

Task 2

In this task we will consider the same setup as in the first one.

Presented criteria have different properties. Akaike Information Criterion (AIC) selects the model with the best predictive properties. Bayesian Information Criterion (BIC) is consistent. Namely, with n growing to infinity, the probability of detecting the proper model converges to one. The Risk Inflation Criterion (RIC) controls the FDR (false discovery rate). The probability of type I error is very low when using RIC (much lower than for BIC and AIC). Modified versions of BIC (mBIC and mBIC2) were developed for the case when the number of variables is greater than the number of observations, which is not our case in this task. mBIC is ABOS (asymptotically Bayes optimal under sparsity) for extremely sparse signals and corresponds to Bonferroni's correction. mBIC2 adapts well to unknown sparsity (is ABOS for unknown sparsity) and corresponds to Benjamini-Hochberg's procedure.

We used those methods to select the best model. However, checking 2^q models could be computationally very expensive. Thus we need to apply a procedure that leads to a good (even if not optimal) solution. Examples of such procedures are forward selection, backward elimination, and stepwise selection.

In the simulations, we have used the stepwise procedure. The only exception is the Akaike Criterion for $p = 950$ when the forward selection was used (due to the high time complexity).

In this task, we consider $n > p$. However, for $p = 950$ the difference between p and n is minimal. We can observe how the criteria good for a lower number of variables (AIC, BIC, RIC) lose their desired (expected) properties. The results are to find in tables 1-5.

The BIC criterion makes quite many false discoveries for a big number of features. However, it holds its power (around 0.65) when the number of observations increases. The AIC is much more powerful. It also makes many more false discoveries. The stepwise computations for AIC were very time-consuming, it was not possible to obtain results in a reasonable time for $n = 950$. Thus, the results are presented for the

Table 3: Results for RIC criterion

	20	100	500	950
FP	0.200	0.140	0.170	0.180
FN	1.420	2.460	3.380	3.710
power	0.716	0.508	0.324	0.258
SSE	18.087	27.161	36.096	39.336

Table 4: Results for mBIC criterion

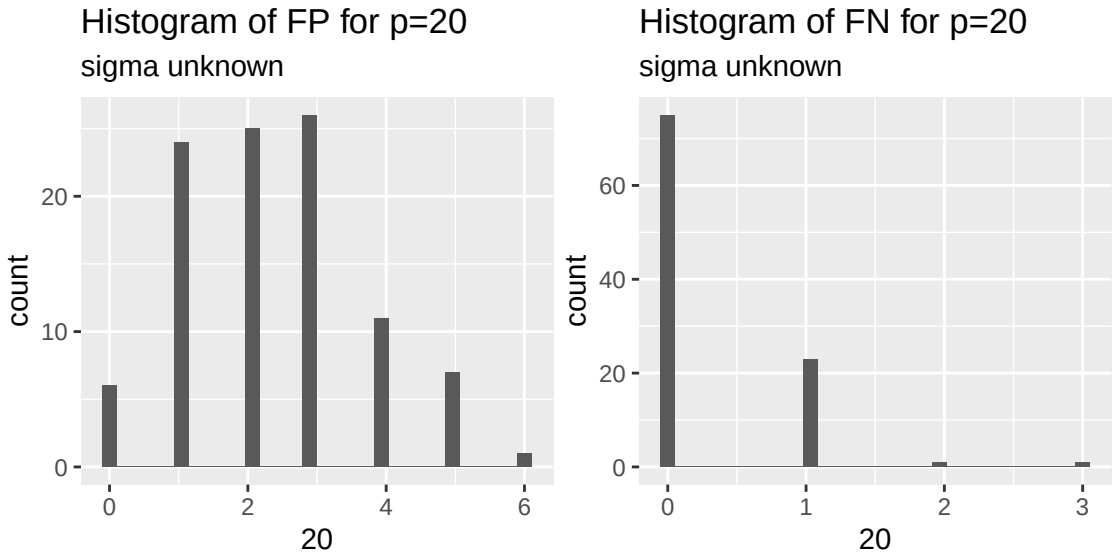
	20	100	500	950
FP	0.020	0.020	0.020	0.010
FN	2.660	3.600	4.190	4.400
power	0.468	0.280	0.162	0.120
SSE	27.771	35.905	40.880	42.319

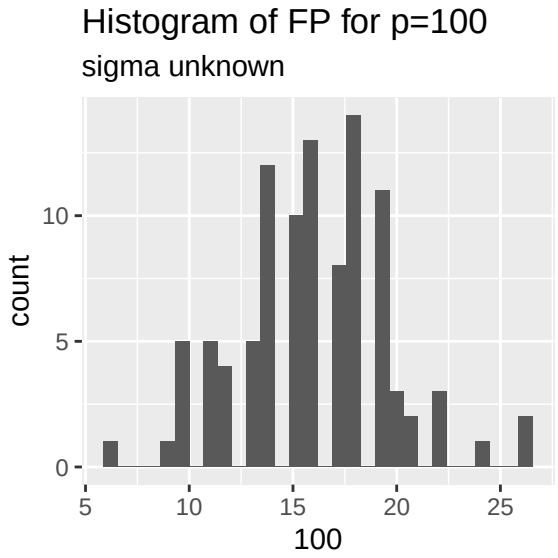
Table 5: Results for mBIC2 criterion

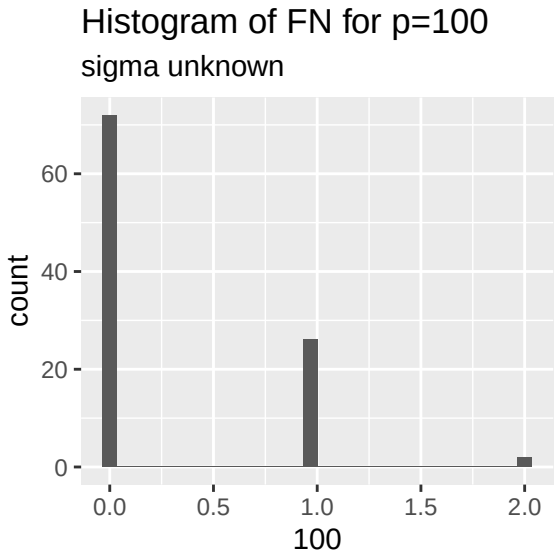
	20	100	500	950
FP	0.120	0.060	0.060	0.030
FN	1.980	3.170	4.090	4.290
power	0.604	0.366	0.182	0.142
SSE	22.474	32.692	40.661	41.750

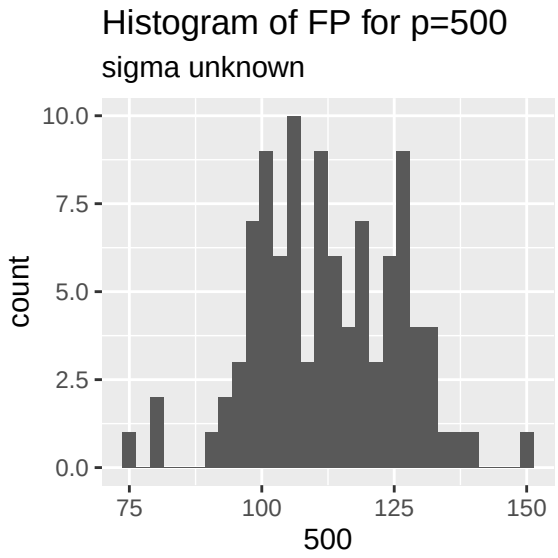
forward selection. This procedure did not work well. The RIC is very conservative. It controls the FDR independently of the value of n. The cost of this property is very low power for large p. The mBIC is also a very conservative criterion. On the one hand, its power is even lower than RIC's. On the other hand, mBIC makes almost no false discoveries. The second modification of the Bayesian information criterion worked better than the presented first. However, it is not dedicated to such a case. mBIC2 controls FDR.

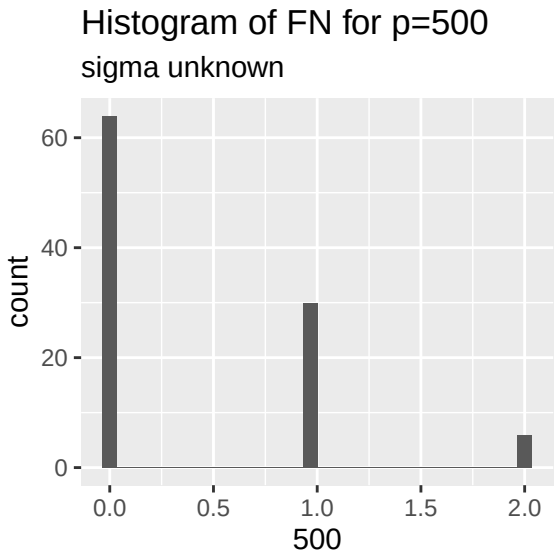
Additionally, below you can find the histograms for the number of false discoveries (FP) and the number of undiscovered signals (FN) for the Akaike criterion (in default setup with unknown sigma).

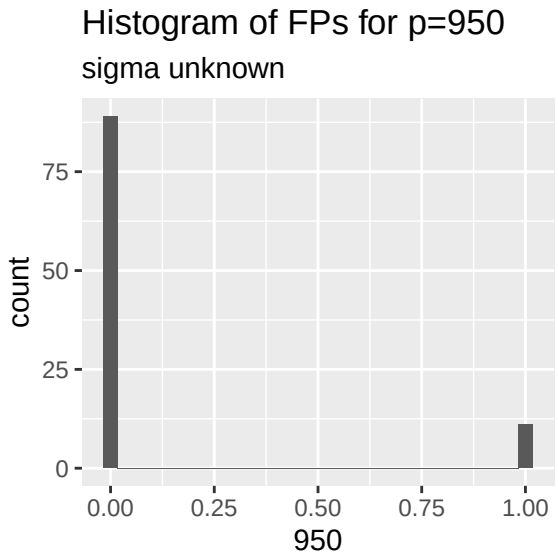


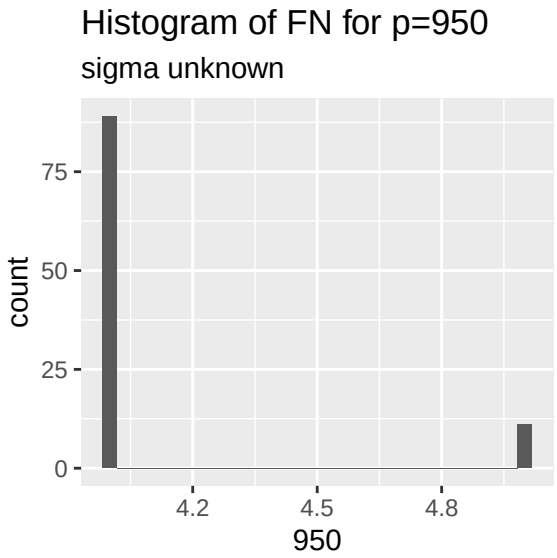

Histogram of FP for p=100
sigma unknown
count
100


Histogram of FN for p=100
sigma unknown
count
100


Histogram of FP for p=500
sigma unknown
count
500


Histogram of FN for p=500
sigma unknown
count
500


Histogram of FPs for p=950
sigma unknown
count
950


Histogram of FN for p=950
sigma unknown
count
950

Task 3

The third task is a case study. We work with data from the genetic domain. The number of features (3200) is much larger than the number of observations (210). Thus the traditional methods of selecting variables do not work well. Stepwise procedure was used with several criteria. The error of the predictions is significantly lower when using methods that control FDR (RIC, mBIC, mBIC2). The best results are obtained by RIC and mBIC2.

Table 6: Prediction Error for different criteria

BIC	AIC	RIC	mBIC	mBIC2
2.544	2.174	0.647	0.776	0.647