ADZD - list 3

Prediction error and information criteria

Introduction

We consider linear model $Y = \beta X + \epsilon$, where X is a $n \times p$ plan matrix, $\epsilon \sim N(0_n, \sigma^2 I_{n\times n})$ is a vector representing random noise and $\beta \in R^n$ is a vector of parameters. Let $\hat{\beta}$ be the estimate of β based on Y and some subset $X_{\tilde{p}}$ of X columns. This report discuss some criteria of $X_{\tilde{p}}$ selection.

Prediction error

Least squares estimator $\hat{\beta}_{LS}$ minimizes the error in the training sample (a.k.a. residual sum of squares) $RSS = ||Y - \hat{Y}||^2$, where $\hat{Y} = X\hat{\beta}$ and Y is the response used to fit the model. It might seem to be a good idea to choose $X_{\tilde{p}}$ resulting in smallest value of RSS, but it doesn't make sense when comparing models with different number of columns, because RSS never increases when we add more variables. The thing we want to minimize instead is the *prediction error* defined as

$$PE = E(\hat{Y} - Y^*)^2,$$

where $Y^* = X\beta + \epsilon^*$ and ϵ^* is a new noise independent of that in training sample. The expression can be rewritten as follows:

$$PE = E||X\hat{\beta} + \epsilon^* - X\beta||^2 = E\sum_{i=1}^{n}(X(\beta - \hat{\beta}) + \epsilon^*)^2 = E\sum_{i=1}^{n}[(X(\beta - \hat{\beta}))^2 + 2X(\beta - \hat{\beta})\epsilon^* + (\epsilon^*)^2] =$$

$$E\sum_{i=1}^{n}[(X(\beta-\hat{\beta}))^2+2\sum_{i=1}^{n}E[X(\beta-\hat{\beta})]E[\epsilon^*]+\sum_{i=1}^{n}E(\epsilon^*)^2] = E\sum_{i=1}^{n}(X(\beta-\hat{\beta}))^2+0+n\sigma^2 = E||X(\beta-\hat{\beta})||^2+n\sigma^2$$

If we use least squares estimator for parameters $\hat{\beta} = (X^TX)^{-1}X^TY$, then $\hat{Y} = X\hat{\beta} = X(X^TX)^{-1}X^TY = MY$ and by Stein's identity

$$E||X(\beta - \hat{\beta})||^2 = E[RSS] + 2tr(M)\sigma^2 - n\sigma^2 \implies PE = E[RSS] + 2tr(M)\sigma^2.$$

Trace of M is p if it's full rank and $PE = E[RSS] + 2p\sigma^2$ and if σ^2 is unknown it should be replaced with its unbiased estimator $s^2 = \frac{RSS}{n-p}$. This is how we can compute the true expected value of RSS:

$$E[RSS] = E||Y - \hat{Y}||^2 = E||X\beta + \epsilon - X\hat{\beta}||^2 = E||X\beta + \epsilon - X(X^TX)^{-1}X^TY||^2 =$$
$$E||X\beta + \epsilon - X(X^TX)^{-1}X^T(X\beta + \epsilon)||^2 = E||(I - M)\epsilon + X\beta - X(X^TX)^{-1}(X^TX)\beta||^2 =$$
$$E||(I - M)\epsilon + X\beta - X\beta||^2 = E||(I - M)\epsilon||^2 = E||z||^2 = \sum E[z_i^2]$$

It can be shown that $z = (I - M)\epsilon \sim N(0, \Sigma = \sigma^2(I - M))$, thus

$$\sum E[z_i^2] = \sum Var[z_i] = tr(\Sigma) = \sigma^2[tr(I - M)] = \sigma^2[tr(I) - tr(M)] = \sigma^2[n - p].$$

We have shown that $E[RSS] = \sigma^2(n-p)$. Finally we can write the true expected value of prediction error as

$$PE = n\sigma^2 - p\sigma^2 + 2p\sigma^2 = \sigma^2(n+p).$$

Instead of computing the true value we could use estimator $\hat{P}E = RSS + 2p\sigma^2$ or its equivalent with unknown σ. Another way of estimating the prediction error is by this formula which makes it easy to compute result of leave-one-out cross validation:

$$\hat{P}E = \sum_{i=1}^{n} \left(\frac{Y_i - \hat{Y}_i}{1 - M_{i,i}} \right)^2,$$

where $M = X(XX^T)^{-1}X^T$ is a matrix of projection onto $Lin(X)$ (X denotes here $X_{\tilde{p}}$, a subset of X).

Information Criteria

This section describe likelihood-based criteria of selection of best number of parameters for data model $M_k : f(x,\theta)$ with $\theta \in R^k$. We define likelihood function $L(X,\theta) = \Pi_{i=1}^{n} f(x,\theta)$ and log-likelihood as

$$l(X,\theta) = \log L(X,\theta).$$

Akaike Information Criterion (AIC) is the one that maximizes the value of

$$AIC(M_k) = l(x, \hat{\theta}_{MLE}) - k,$$

which in case of linear model $Y = X\beta + \epsilon$ can be rewritten as

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

Maximizing the expression above corresponds to minimizing $RSS + 2\sigma^2 k$, wchich is the same as SURE estimator (k denotes here the number of selected columns).

In case σ^2, the variance of random error in a sample is unknown it should be replaced with its **biased** estimator $\hat{\sigma}^2_{MLE} = \frac{RSS}{n}$ which leads to

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

and minimizing the value of $n\log(RSS) + 2k$.

AIC rewards goodness of fit assessed by the likelihood function, but at the same time imposes a penalty on increasing the number of parameters. Without that penalty, we would always choose a model with all avaiable variables which would lead to massive overfitting. Some remarks about practical use of AIC:

1. In most cases we cannot afford computing AIC for all possible models as it would require fitting a model 2^p times. Instead we use heuristic procedures which with large probability return model close to optimal, some of which are implemented by `bigstep` library and used to obtain results in this report.

2. It can be shown that in terms of multiple hypothesis testing, if X is orthogonal, AIC marks as significant variables $\hat{\beta}_i$ such that $|\hat{\beta}_i| > \sigma\sqrt{2}$, wchich in case of $\sigma^2 = 1$ gives probability of type I error equal to $2(1 - F_{N(0,1)}(\sqrt{(}2)) \approx 0.16$ and for large number of hypotheses the number of false discoveries is also large.

Bayesian Information Criterion (BIC) is similar to AIC, but it maximizes

$$BIC(M_k) = k\log(n) - 2\log L(X, \hat{\theta}_{MLE})$$

which is equivalent to minimizing the value of expression $RSS + \sigma^2 k \log(n)$. BIC selects variables such that $|\hat{\beta}_i| \geq \sigma\sqrt{\log(n)}$, which turns out to lead to much smaller chance of type I error (when $\sigma^2 = 1$ it is approximately 0.013.) Thus there are way less false discoveries.

Other criteria

There are also other, more complex information criteria used in this report: mBIC, mBIC2 (implemented in `bigstep` library). Both of them control probability of type I error better than AIC and BIC when p (all available columns) is large, because they depend on it in addition to k (selected columns).

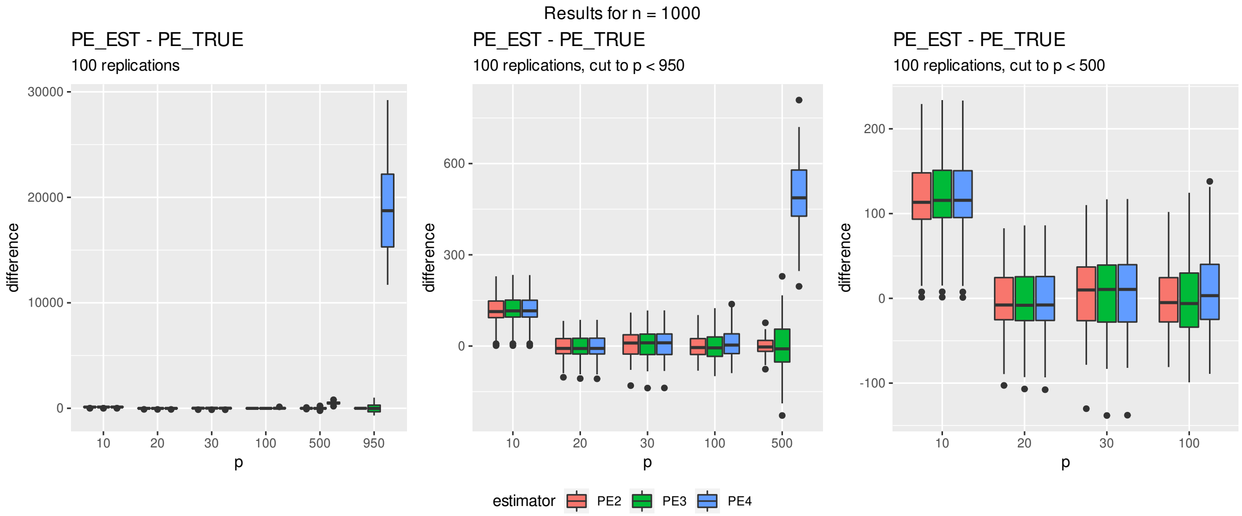
Task 1

The table below show numeric results of experiment replicated 100 times, where `PE1`, `PE2`, `PE3`, `PE4` are respectively: true prediction error, SURE with known σ^2, SURE with unknown σ^2 and LOO cross-validation estimator.

Table 1: Task 1 - numeric results averaged over 100 reps (seed = 600)

p	RSS	PE1	PE2	PE3	PE4	AIC_known	AIC_unknown
10	1110.10	1010	1130.10	1132.53	1132.57	-1483.99	-4434.60
20	977.69	1020	1017.69	1017.60	1018.05	-1427.78	-4381.13
30	976.66	1030	1036.66	1037.07	1037.90	-1437.27	-4390.43
100	899.04	1100	1099.04	1098.83	1110.31	-1468.46	-4419.07
500	497.96	1500	1497.96	1493.88	1994.20	-1667.92	-4523.48
950	49.78	1950	1949.78	1941.46	20711.25	-1893.83	-3811.85

Boxplots (Prediction Error)

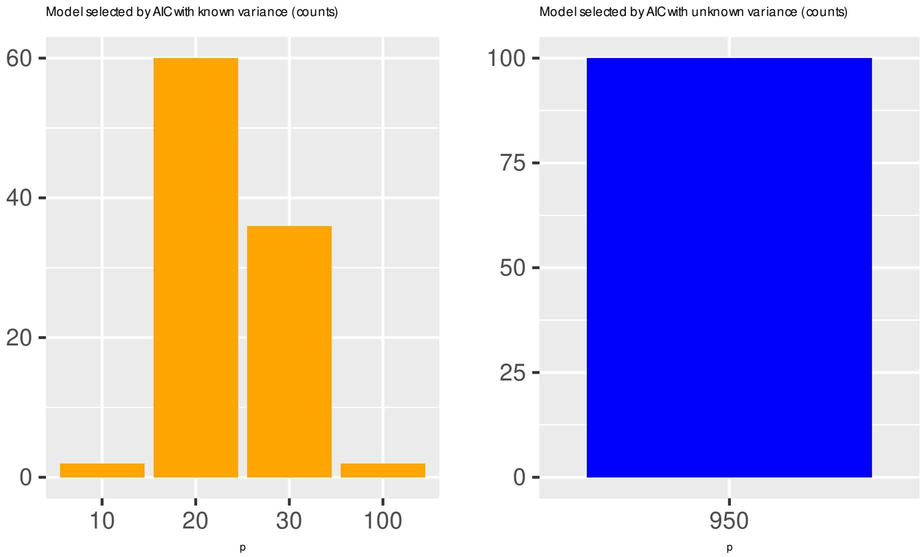


Comment:

- The median of difference $\hat{P}E - PE$ oscillates around 0 for first two estimators for all $p \geq 20$. It means that the estimators are unbiased. Results seem to be the best for $p = 20$ which is the true number of parameters in model. For $p = 10$ which is less than the true number of model parameters, estimated prediction error is greater by about 10% than the theoretical value of prediction error.

- For $p \geq 500$ interquartile range of estimator with unknown σ becomes obviously wider than for known σ.
- Cross-validation prediction error becomes strongly biased for $p \geq 500$.

Optimal model chosen by AIC



Comment:

- When σ is known AIC selects the true model about 60 out of 100 times, model with 30 variables instead of 20 with approximately 35 out of 100 times and 10 or 100 variables both about 2.5/100 out of times.
- When σ is unknown, then always model with all posible variables is selected. AIC is biased towards choosing model with large number of parameters.

Task 2

In this task we apply AIC, BIC, mBIC and mBIC2 to find significant variables. I used `fastforward` procedure which should give worse results that `stepwise`, but works much faster for some criteria.

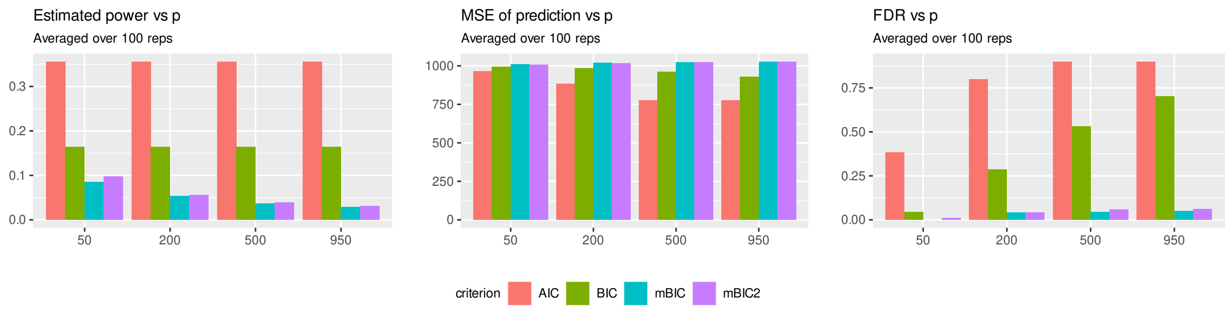


Table 2: Task 2 - estimated powers averaged over 100 reps (seed = 600)

p	AIC	BIC	mBIC	mBIC2
50	0.36	0.16	0.08	0.10
200	0.36	0.16	0.05	0.06
500	0.36	0.16	0.04	0.04
950	0.36	0.16	0.03	0.03

Table 3: Task 2 - MSE averaged over 100 reps (seed = 600)

p	AIC	BIC	mBIC	mBIC2
50	965.10	994.45	1010.93	1007.76
200	882.33	982.99	1018.41	1017.58
500	776.90	960.44	1023.74	1022.61
950	776.42	927.21	1026.46	1025.45

Table 4: Task 2 - FDR averaged over 100 reps (seed = 600)

p	AIC	BIC	mBIC	mBIC2
50	0.38	0.04	0.00	0.01
200	0.80	0.29	0.04	0.04
500	0.90	0.53	0.04	0.06
950	0.90	0.70	0.05	0.06

Comment:

- AIC has the best estimated power among criteria. It is independent of number of parameters, similarly to the power of BIC criterion, which is beacuse of the fact that both criteria don't depend on p. We can see that powers of other criteria keep decreasing as the number of parameters increases. The one with the worst power is mBIC.
- AIC has the best power, but at the same time its False Discovery Rate is the largest among all criteria. For small p it's much bigger than other FDRs and when p increases, FDR of BIC criterion goes up and starts to be more similar to FDR of AIC (but still smaller). Other criteria are really good in terms of FDR keeping it below value of 0.07 for all values of p (their FDR increases slightly with p).
- MSE of prediction is independent of p for all criteria except AIC and BIC. For AIC and BIC it decreases when p increases and the effect is more visible in AIC.

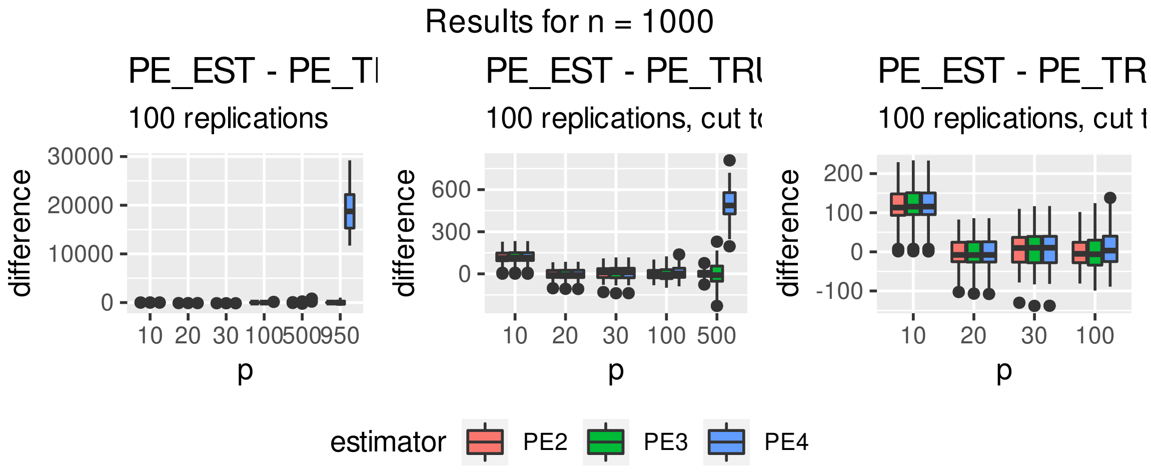
Task 3

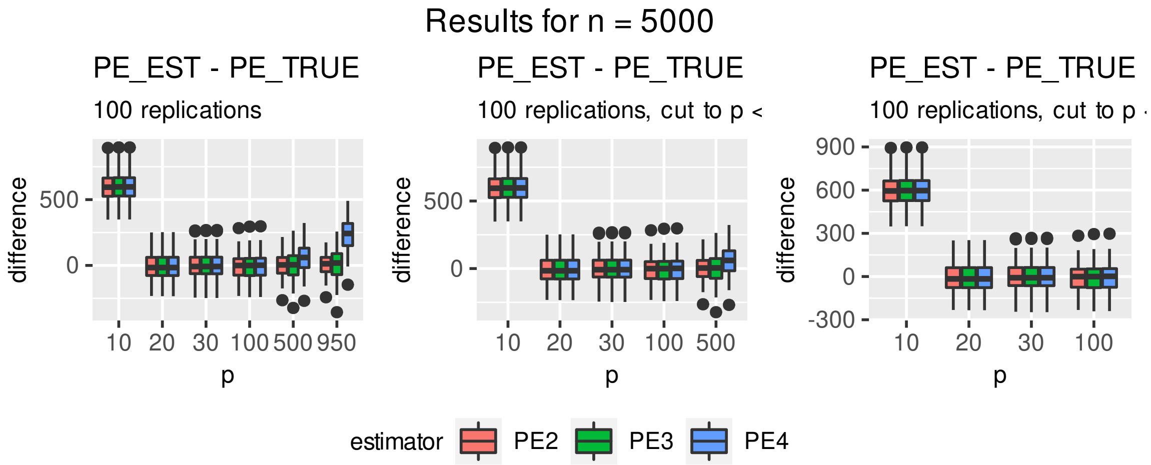
In this task we repeat experiments from task 1, but this time plan matrix has 5000 rows, which means that number of observations is much bigger than number of columns.

Table 5: Task 3 - numeric results averaged over 100 reps (seed = 600)

p	RSS	PE1	PE2	PE3	PE4	AIC_known	AIC_unknown
10	5585.87	5010	5605.87	5608.26	5608.28	-7397.63	-26174.28
20	4970.35	5020	5010.35	5010.27	5010.32	-7099.87	-25892.29
30	4968.74	5030	5028.74	5028.72	5028.91	-7109.06	-25901.53
100	4893.02	5100	5093.02	5092.73	5094.83	-7141.20	-25933.13
500	4496.98	5500	5496.98	5496.31	5552.33	-7343.18	-26122.10
950	4054.86	5950	5954.86	5957.14	6181.23	-7572.12	-26313.38

Boxplots (Prediction Error)

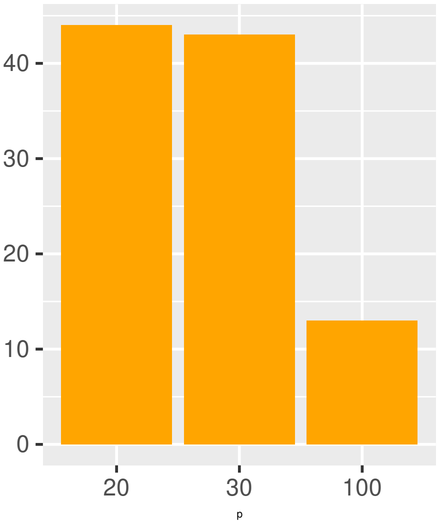




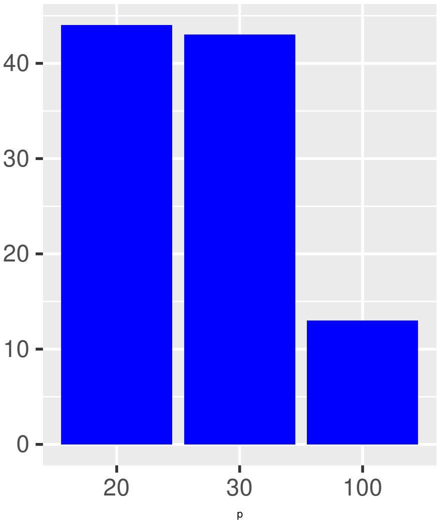
Comment:

Models selected by AIC

Model selected by AIC with known variance (counts)



Model selected by AIC with unknown variance (counts)


40
30
20
10
0
20
30
100
p