

Lab2

Jakub Kuciński

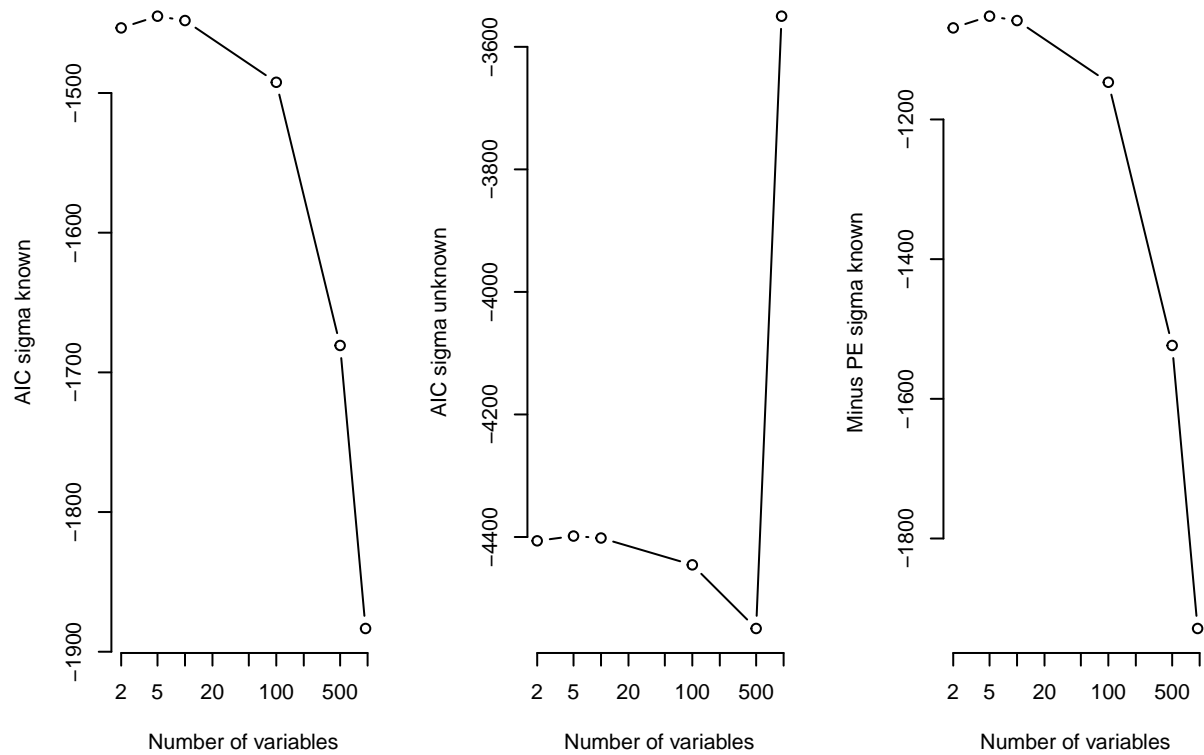
15 04 2022

Task 1

##	ncols	RSS	PE	PE_sigma_known	PE_sigma_unknown	PE_cross_valid
## 1	2	1064.91340	1002.289	1068.913	1069.182	1069.129
## 2	5	1042.12805	1005.389	1052.128	1052.602	1052.801
## 3	10	1038.38243	1009.135	1058.382	1059.360	1059.755
## 4	100	946.75916	1100.758	1146.759	1157.150	1168.589
## 5	500	523.60631	1523.911	1523.606	1570.819	2090.904
## 6	950	28.84945	2018.668	1928.849	1125.129	12078.323

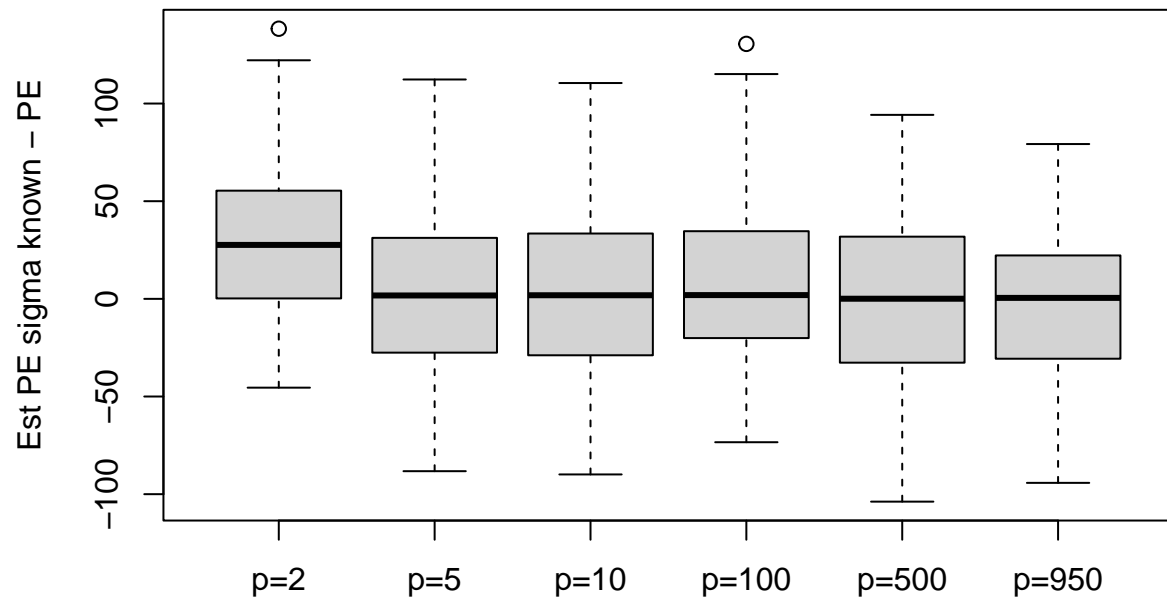
##	AIC_sigma_known	AIC_sigma_unknown
## 1	-1453.395	-4406.263
## 2	-1445.003	-4398.449
## 3	-1448.130	-4401.648
## 4	-1492.318	-4445.461
## 5	-1680.742	-4549.309
## 6	-1883.363	-3549.984

We can see that RSS drops with increasing number of variables, however it does not correspond to creating a better model as true expected value of prediction error increases. PE estimators are quite close to the true expected PE for up to 100 variables. PE estimation for known sigma is still quite close to the true expected PE for 500 variables. Estimation with unknown sigma vary more, especially for 950 variables. Estimation of PE using leave-one-out cross-validation gives much higher value for 500 variables and of order of magnitude bigger value for 950 variables.

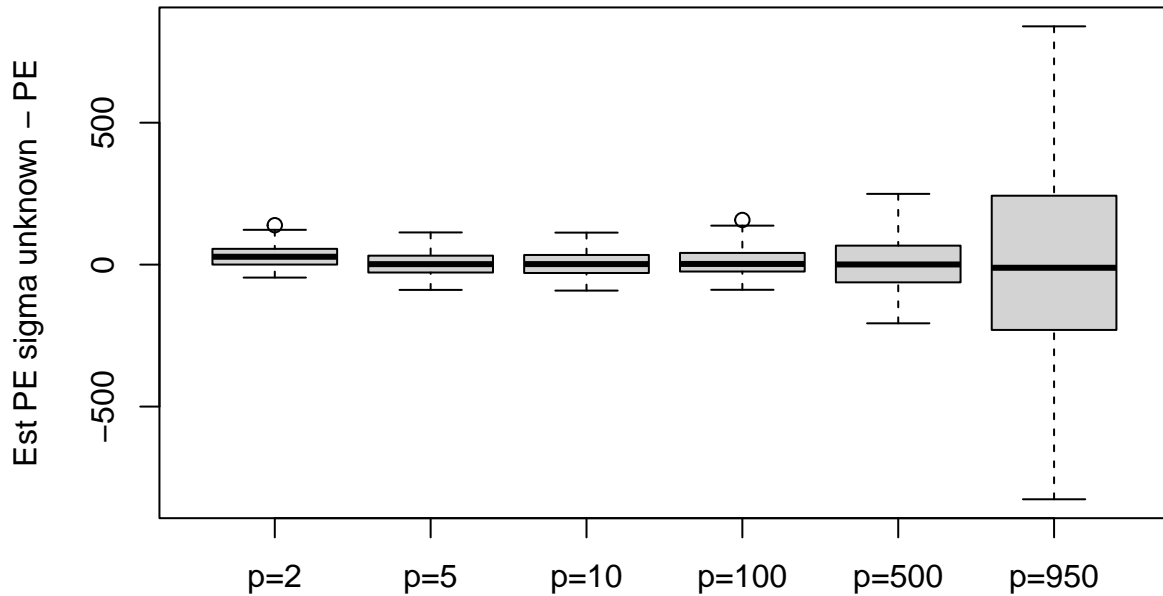


AIC with known sigma chooses model with 5 first variables, that is it chooses model with all and only true nonzero coefficients. AIC chooses models with very high number of parameters when number of variables is close to the number of samples. We can see that on plot of AIC with unknown sigma. The AIC value for model with 950 variables is significantly higher than for smaller models. We can also see, that plots of AIC for known sigma and minus PE for known sigma are similar, as maximizing AIC for known sigma is equivalent to minimizing $RSS + 2\sigma^2 p$, which is unbiased estimator of the prediction error in least squares regression.

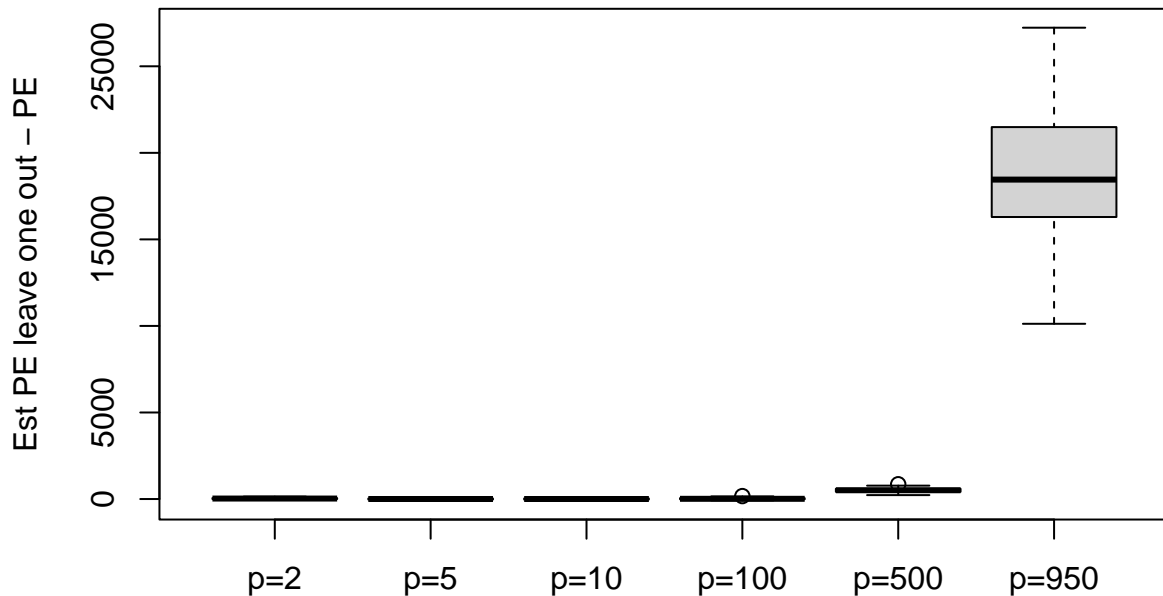
Boxplots of difference between PE and it's first estimator



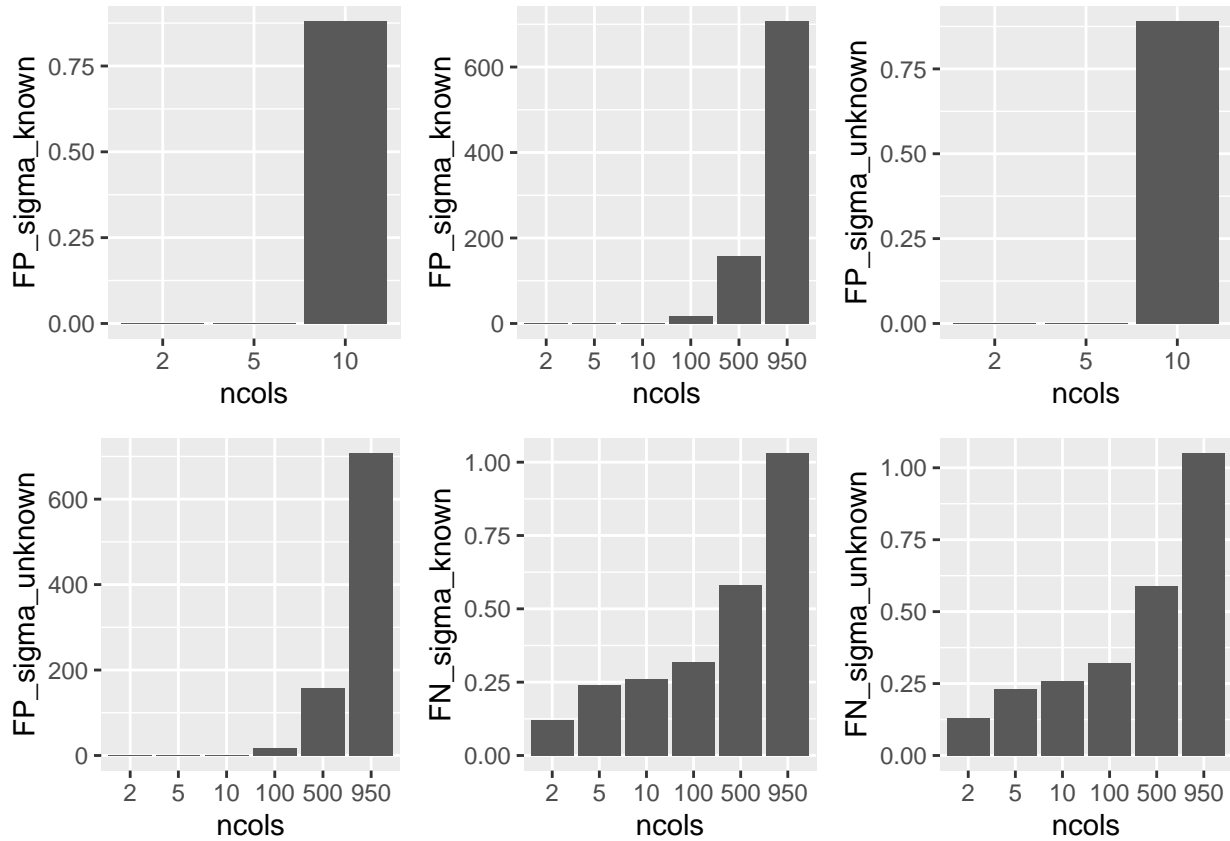
Boxplots of difference between PE and it's second estimator



Boxplots of difference between PE and it's third estimator

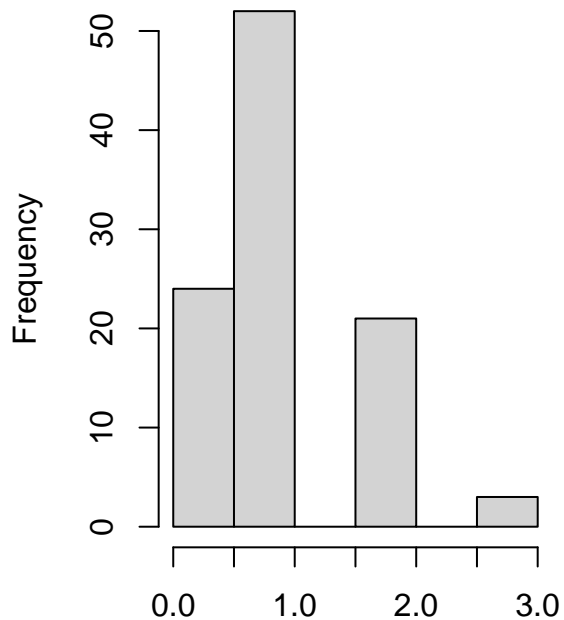


After 100 simulations we can look at the boxplots of differences of estimations of PE and its true expected value. For estimator with known sigma we can see, that differences are around zero and the standard deviation of differences are similar for all models. Differences of estimator with unknown sigma and true expected PE and estimator deviates more with increase of model size. For leave-one-out crossvalidation, the differences of the true expected PE and the estimator are getting big and deviates even more with increase of model size.

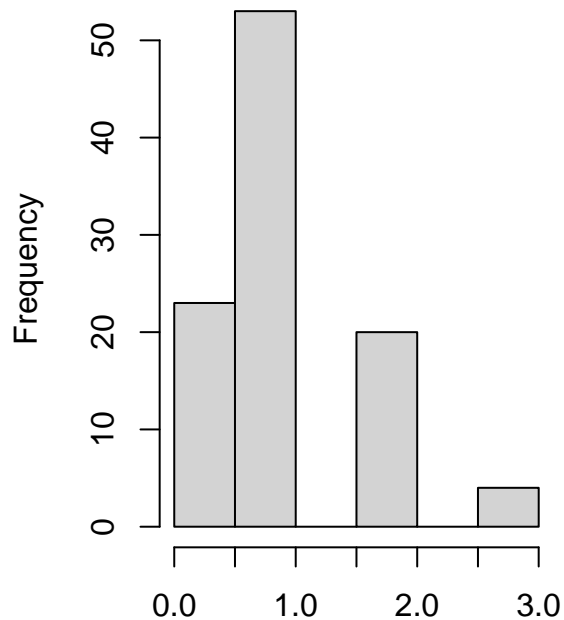


We can also look at the average number of FP and FN produced by AIC with known and unknown sigma for different model sizes. Of course for models with 2 and 5 first variables AIC can't produce FP (5 first variables have true non-zero coefficients). For model with 10 variables on average it is less than 1, but we can see, that number of FP for bigger models is getting very big. Average number of FN is below 1 for all models and grows with number of consider variables. We can conclude, that AIC is not a good selection criteria if we are interested in selecting small subset of relevant variables from big set containing many irrelevant variables.

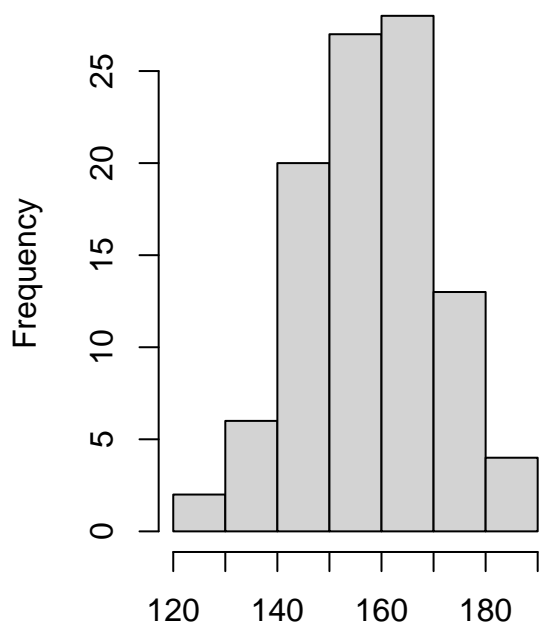
Histogram for 950 variables



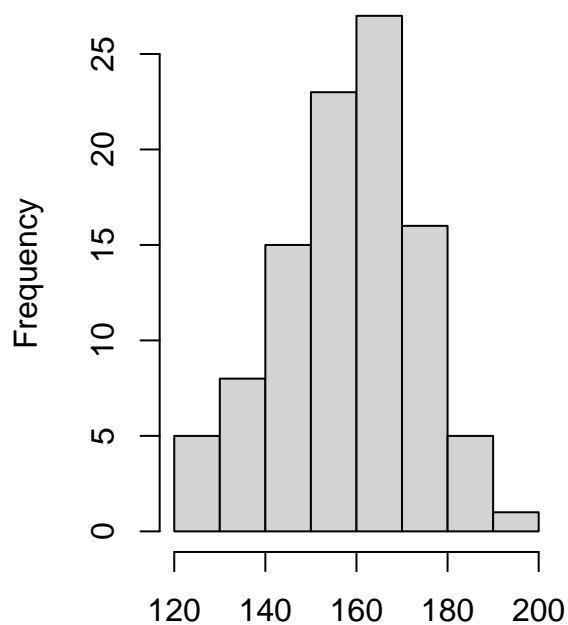
Histogram for 950 variables



FN_sigma_known
Histogram for 500 variables



FN_sigma_unknown
Histogram for 500 variables



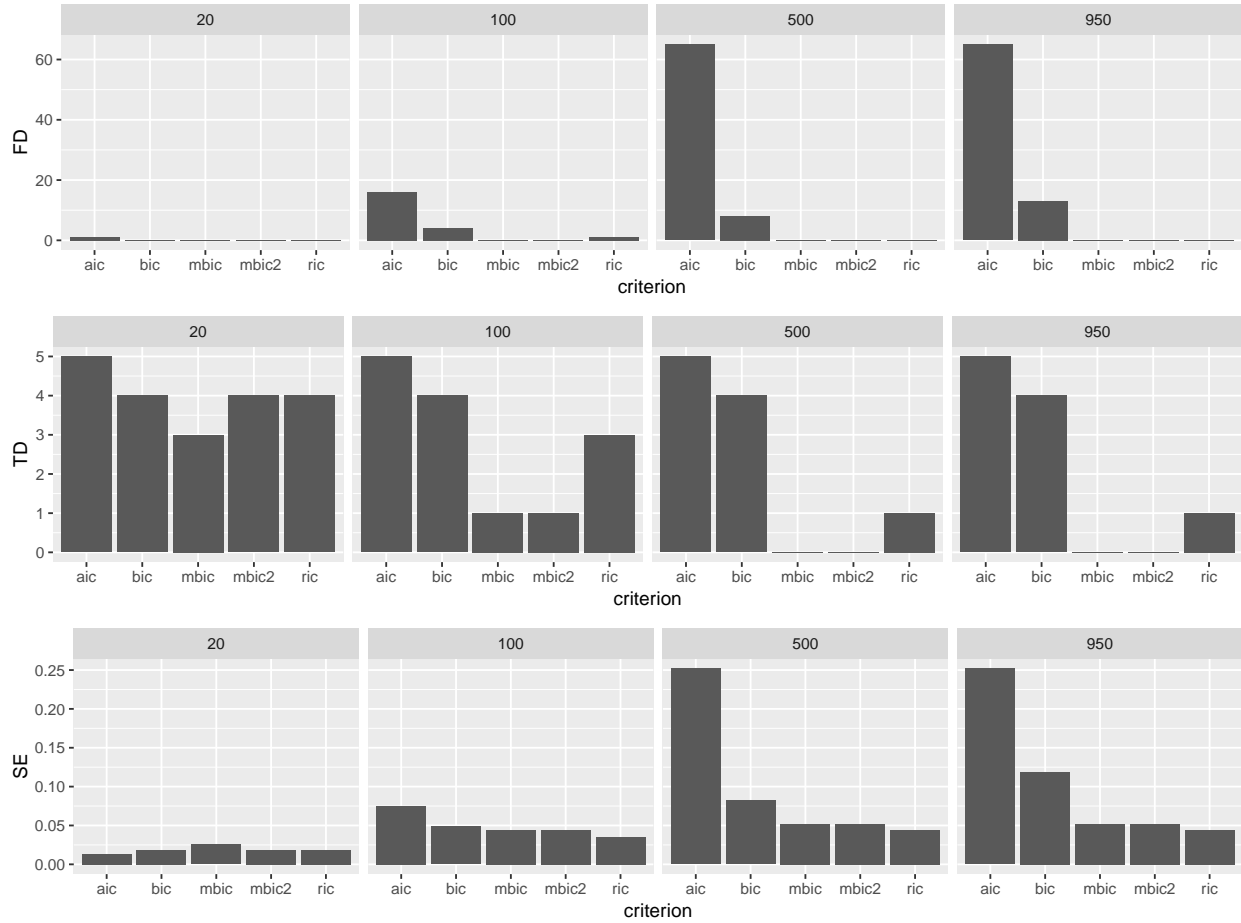
FP_sigma_known

FP_sigma_unknown

We can also have a look at selected histograms of the number of false negatives and false positives produced by both versions of AIC. We can see, that histograms with known and unknown sigmas of the number of FP and FN are very similar.

Task2

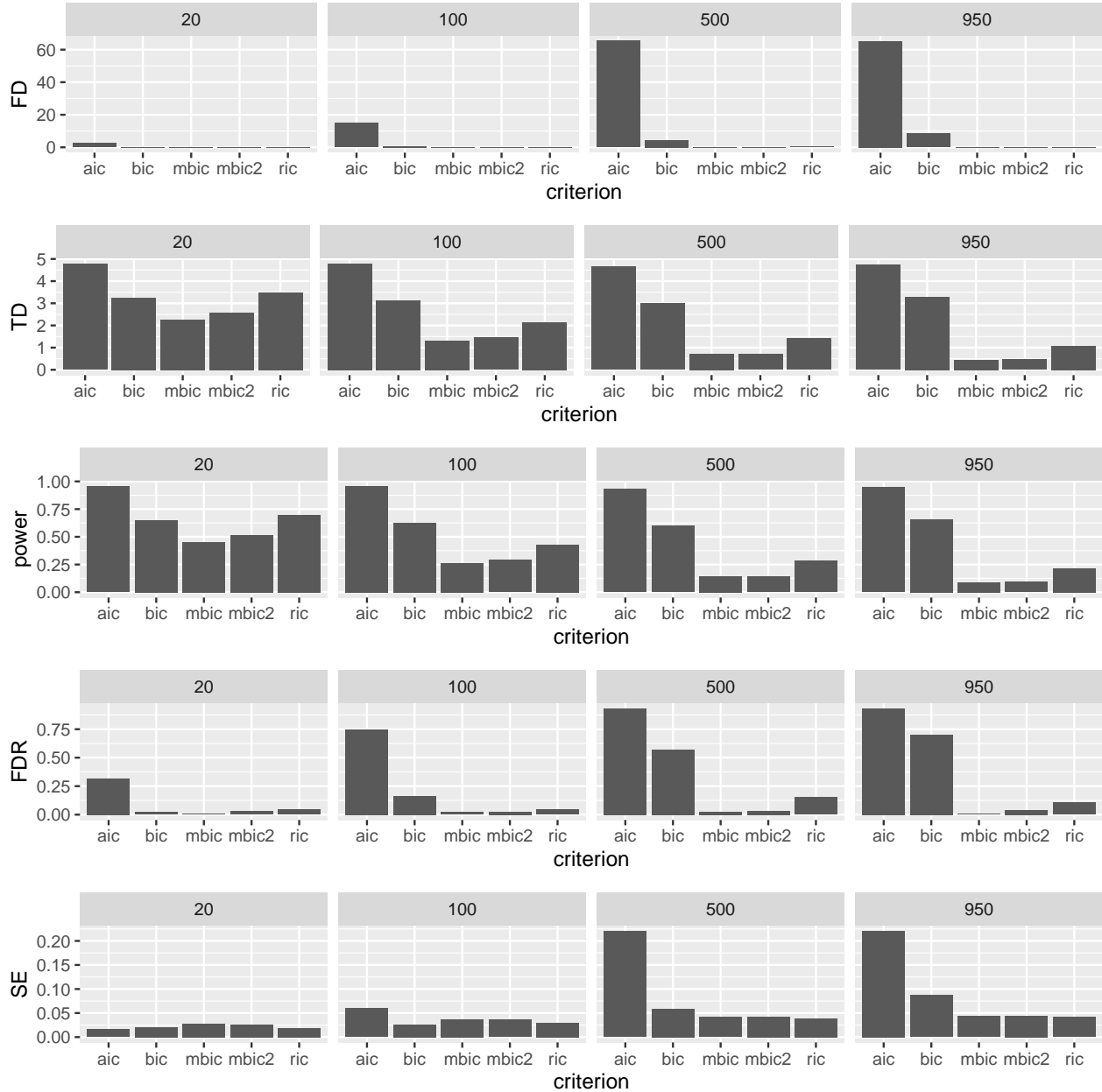
For all of the simulations I used fast forward, because step-wise was taking very long on my machine.



Looking at the plot we can see that AIC made the most TD, but it also made far the most FD and had very big square error. BIC had similar number of TD but also did some FD. Its SE was much lower than AIC. RIC had the smallest SE for bigger data, third best TD and no FD. mBIC and mBIC2 did less TD, no FD and they also had higher SE then RIC.

##	ncols	criterion	FD	TD	power	FDR	SE
## 1	20	ric	0.19	3.48	0.696	0.044	0.0188944844452284
## 2	20	bic	0.1	3.25	0.65	0.0208333333333333	0.0200958693899984
## 3	20	aic	2.47	4.77	0.954	0.318776862026862	0.0160252651686599
## 4	20	mbic	0.02	2.27	0.454	0.0075	0.0281972256598979
## 5	20	mbic2	0.13	2.58	0.516	0.03	0.0258957122229803
## 6	100	ric	0.14	2.15	0.43	0.046	0.0300749703471494
## 7	100	bic	0.72	3.14	0.628	0.162690476190476	0.02624390945103
## 8	100	aic	14.88	4.8	0.96	0.746200161708899	0.0598372674842826
## 9	100	mbic	0.05	1.32	0.264	0.0208333333333333	0.0371361059509894
## 10	100	mbic2	0.06	1.45	0.29	0.0258333333333333	0.0364124165977905
## 11	500	ric	0.32	1.41	0.282	0.151833333333333	0.0390720072428732
## 12	500	bic	4.47	2.99	0.598	0.57143623043623	0.0585629309234807
## 13	500	aic	65.34	4.66	0.932	0.933428571428571	0.219709011556112
## 14	500	mbic	0.04	0.72	0.144	0.025	0.0416433774730269
## 15	500	mbic2	0.06	0.71	0.142	0.0323333333333333	0.0423643237585623

```
## 16 950 ric 0.21 1.07 0.214 0.105 0.0415030084158327
## 17 950 bic 8.32 3.29 0.658 0.700310239253351 0.0871913312696208
## 18 950 aic 65.26 4.74 0.948 0.932285714285714 0.221165779677778
## 19 950 mbic 0.01 0.45 0.09 0.005 0.0431769576069547
## 20 950 mbic2 0.05 0.46 0.092 0.04 0.0437403869816306
```



By looking at the table and plot after 100 simulations (pasted above) we can see similar behaviors. AIC makes the most TD, but also does a lot of FD and thus has high power, but also very high FDR. It has also high SE for bigger models. BIC behaves similarly to AIC, but has smaller number of TD and FD and therefore smaller power, but also smaller FDR. mBIC, mBIC2 and RIC do very little FD (on average less than 0.32 with RIC making a bit more than mBIC and mBIC2). RIC do more TD than mBIC and mBIC2, especially for bigger data, however much less than AIC and BIC. RIC has bigger power and FDR than mBIC and mBIC2, but smaller than AIC and BIC. However RIC achieves the best SE for data base with very high number of variables (500 and 950). Therefore we can consider RIC as being the best criterion (among examined ones) for choosing important covariates in data sets with very high number of irrelevant variables,

as it does very little FD (just a little more than mBIC and mBIC2, so has low FDR), achieves the best SE and has reasonable power.

Task 3

##	Criterion	No. selected vars	Train MSE	Test MSE
## 1	ric	8	0.0219980271083248	0.0206446061719873
## 2	bic	70	0.00233448541044264	0.0399394408039285
## 3	aic	70	0.00856267976185865	0.0289034288279505
## 4	mbic	6	0.0233319961064578	0.0215663426226658
## 5	mbic2	10	0.0178267622123903	0.0372606827141793

From simulation we can see, that AIC and BIC chose 70 variables and achieve much smaller train MSE than RIC, mBIC and mBIC2, which chose around 8 variables. However we can see, that AIC and BIC did overfit to the training data, as test MSE is much higher than train MSE and actually higher than test MSE for the RIC, mBIC and mBIC2. On test dataset RIC achieved the best MSE so it was the best criterion for choosing variables, which is in line with the theory, as RIC was designed to identify important variables for datasets with very high number of variables (even higher than number of observations).