Statistical learning with Sparsity – The LASSO and generalisations.

Chapter 6: Inference. Plus some other considerations.

Menelaos Pavlou

Department of Statistical Science, UCL, UK

Overview

- Development and validation of risk models for predicting continuous, binary or survival outcomes. Some simulations regarding the practical application of Lasso.
- Early approaches to inference
  - Bayesian Lasso
  - Bootstrap
- Recent approaches to inference
  - The covariance test.
  - Extensions (The Spacing Test, briefly).
- Discussion

Prediction models

Stage 1: Model Development (Training)

For *independent* binary outcomes, developing a prediction model can be as simple as fitting a multivariable logistic regression model

$$\mathrm{logit}(P(Y_i|\boldsymbol{X}_i) = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \cdots + \beta_p X_{i,p} = \boldsymbol{\beta}\, \boldsymbol{X_i}$$

regression coefficients: $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \dots, \beta_p)^T$

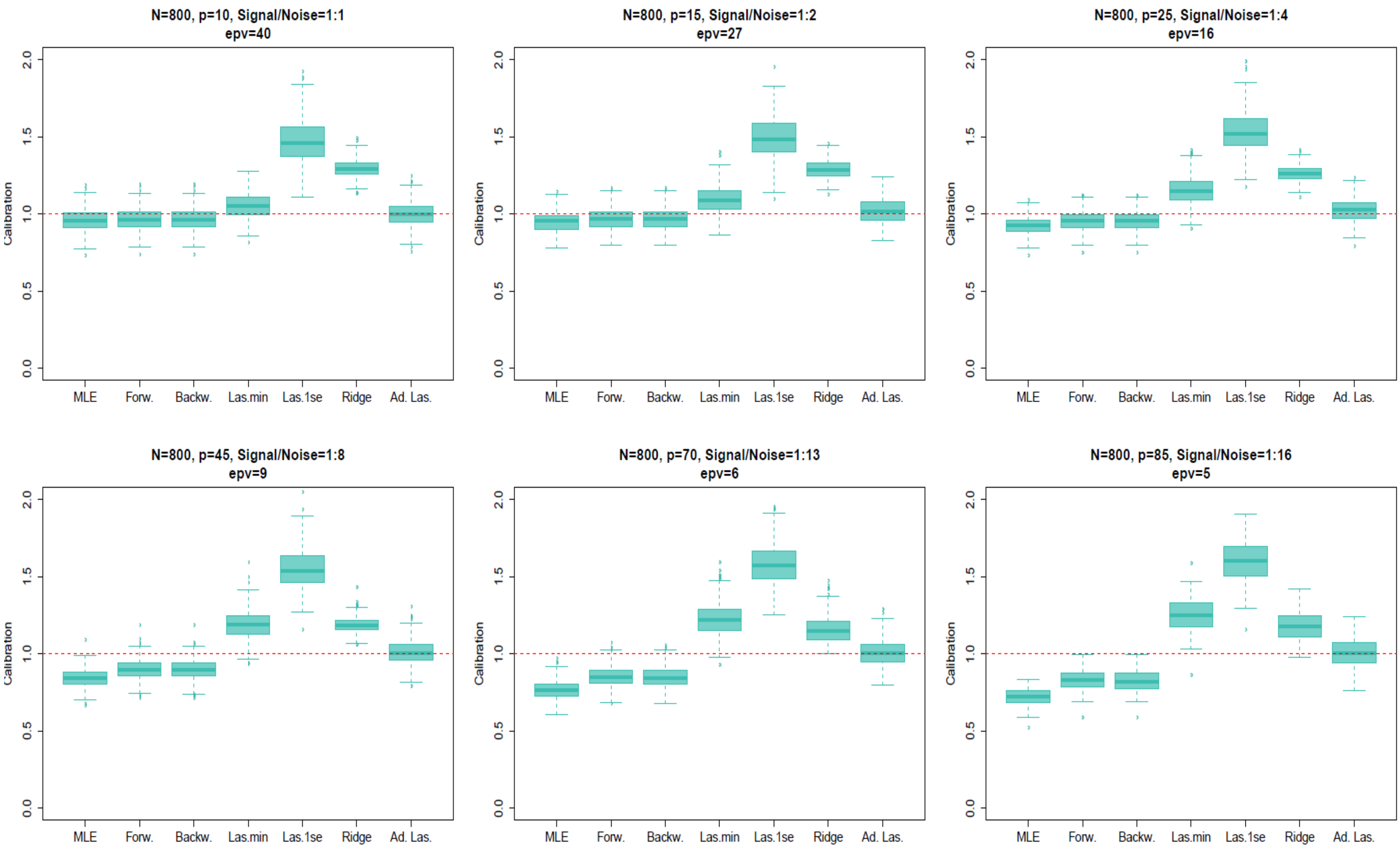
predictors: $\boldsymbol{X_i} = (X_{i,1}, X_{i,2}, \dots, X_{i,p})^T, i = 1, \dots, N$

Stage 2: Model Validation (Testing)

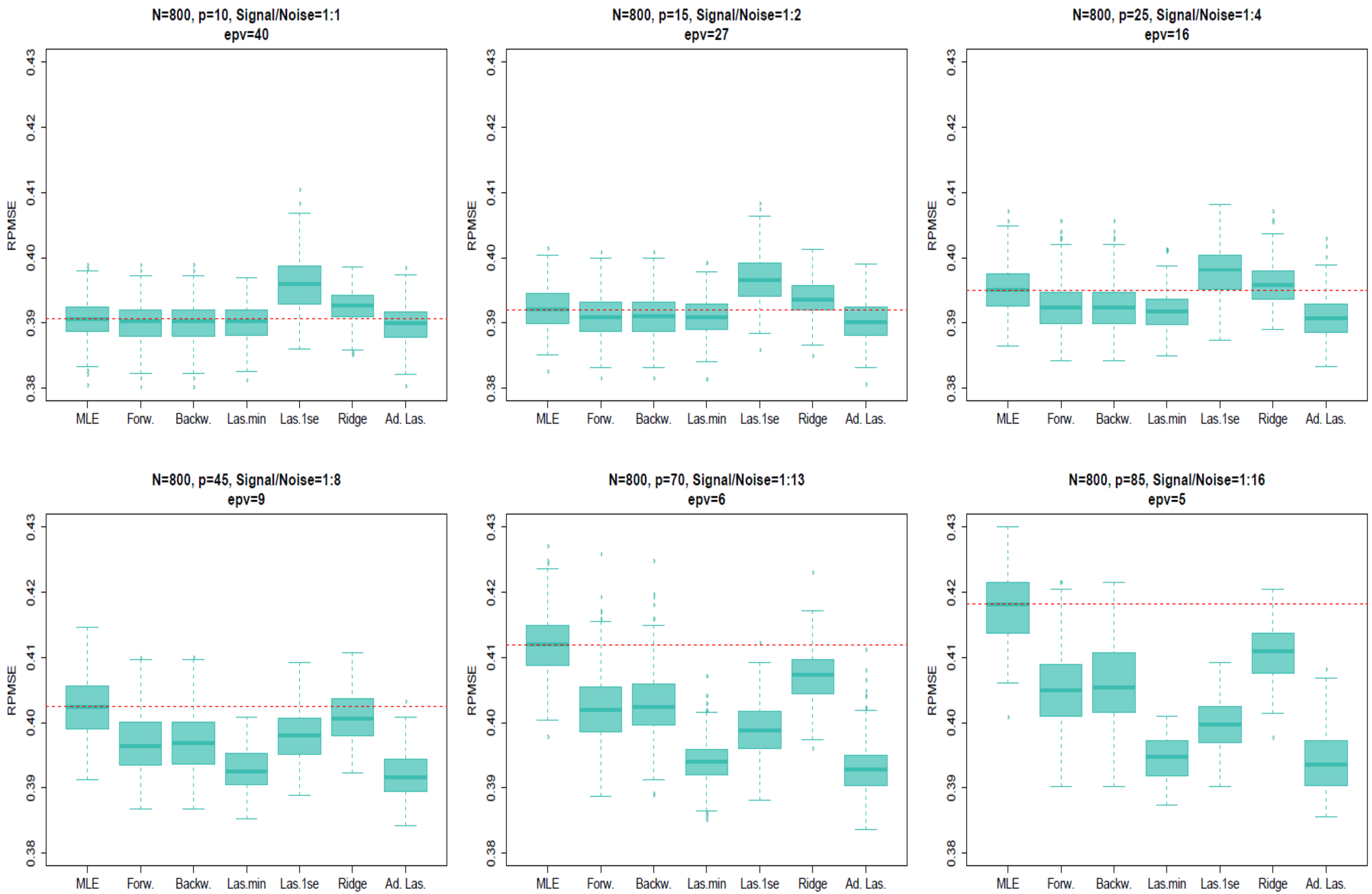
Predictive performance needs assessing before the model is used in practice.

Common measures are: Predictive accuracy (Brier Score or Predictive Mean Squared Error in Simulations), Calibration (calibration slope as a measure of overfitting), Discrimination (C-statistic, AUC for binary outcomes).

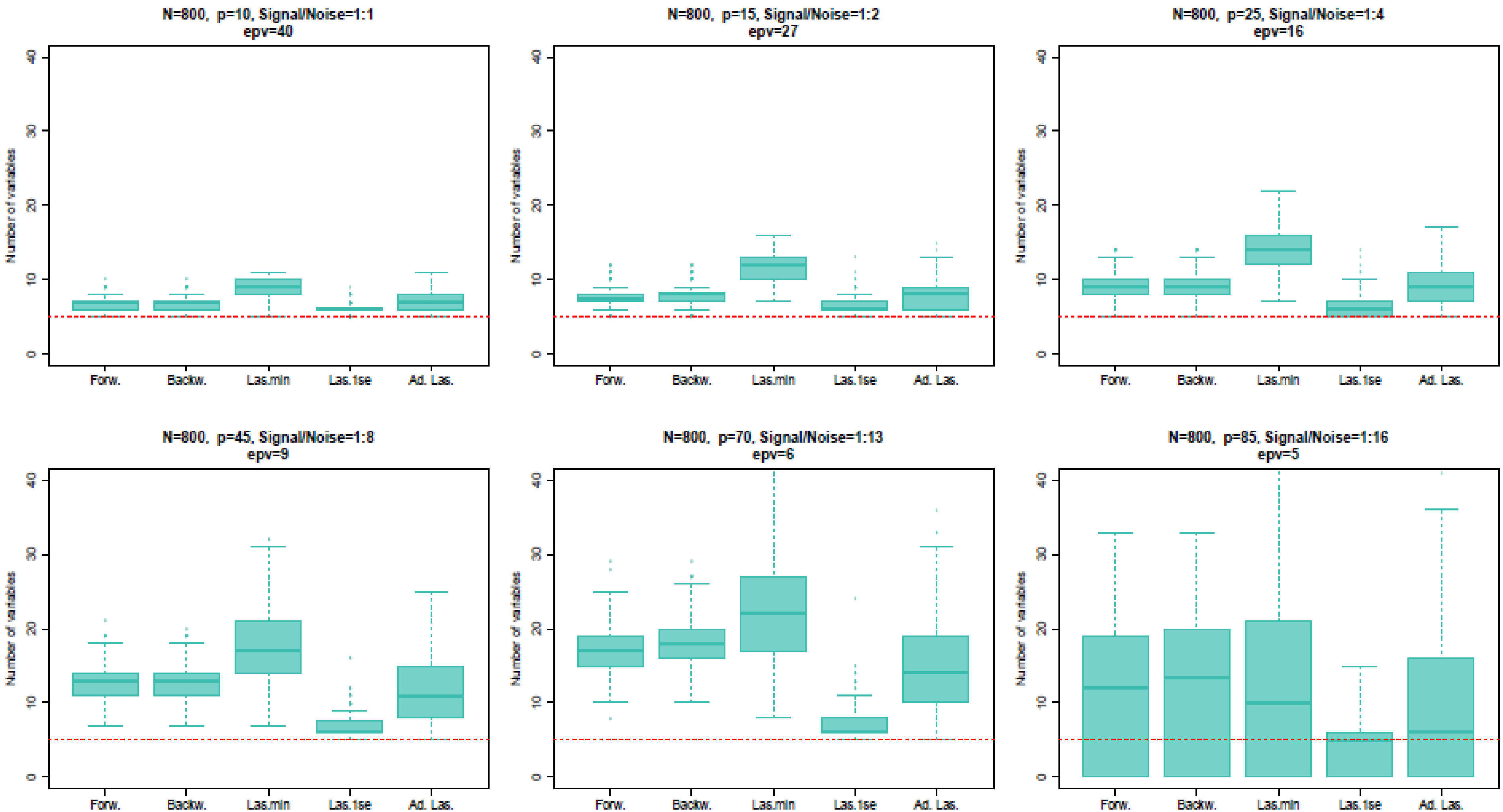
Simulations Binary Outcome - Calibration



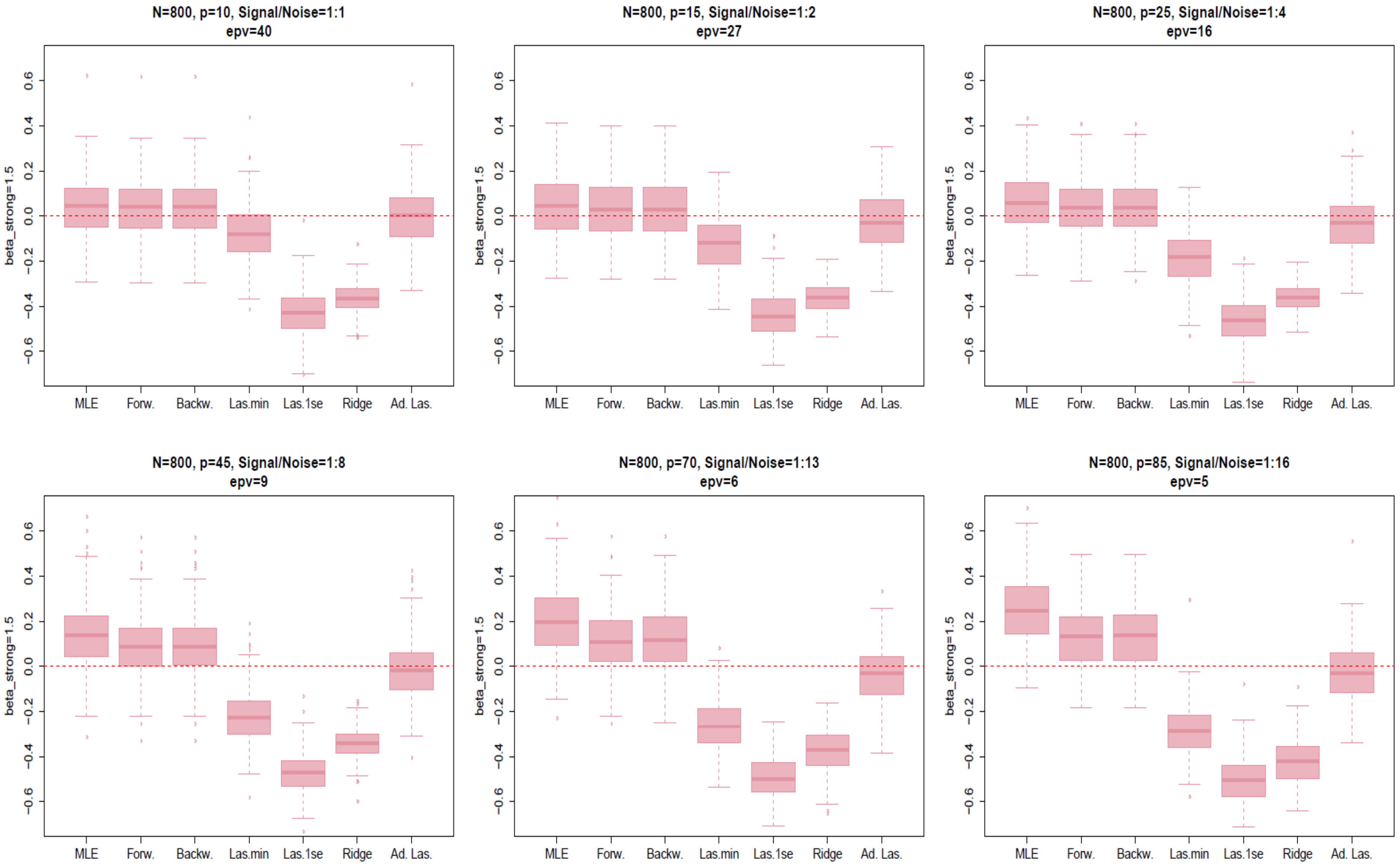
Simulations Binary Outcome – Predictive MSE



Simulations Binary – Number of selected variables



Simulations Binary – Bias on a large coefficient



The bootstrap

- In fitting a model using lasso we obtain we typically obtain *point* estimates of the regression coefficients, $\widehat{\boldsymbol{\beta}}(\hat{\lambda}_{CV})$, some of which may be zero. For a different partition in the crossvalidation it is likely that different coefficients are set to zero.
- It is desirable to get a feeling about the sampling distribution of $\widehat{\boldsymbol{\beta}}(\hat{\lambda}_{CV})$
- One way of doing this is using the bootstrap.

- **Algorithm: (Non parametric) Bootstrap**

1. Create a 'new' dataset, *i*, by sampling with replacement from the original dataset.
2. Fit Lasso to the bootstrapped dataset to obtain $\widehat{\boldsymbol{\beta}_{\boldsymbol{i}}}(\hat{\lambda}_{CV})$.
3. Go to 1 and repeat B times (e.g. 1000 times).
4. Assess the sampling distribution of $\widehat{\boldsymbol{\beta}}(\hat{\lambda}_{CV})$.

- **Parametric bootstrap**

Bayesian Lasso

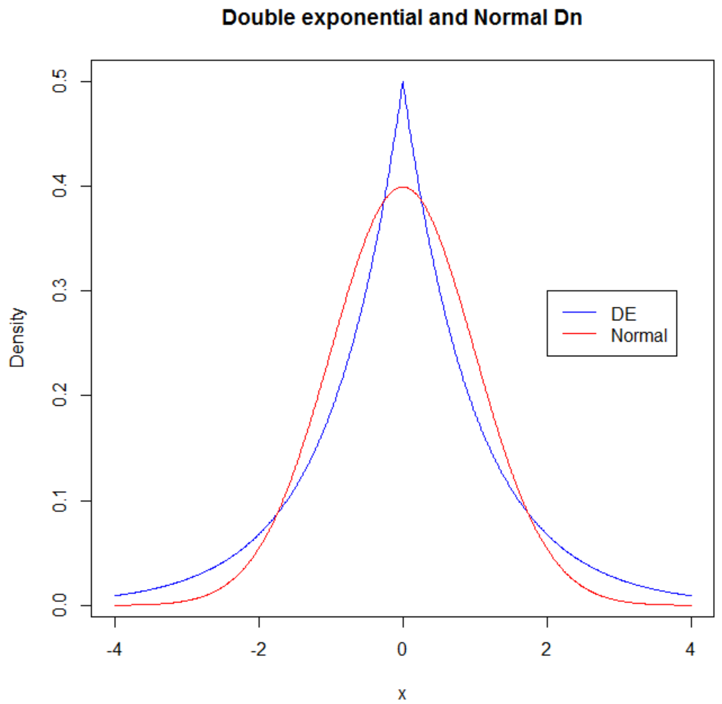
- ***For regression coefficients:***

Laplace (double exponential prior)

$$\beta_k|\sigma^2 \sim \mathrm{N}(0, \sigma^2)$$

$$\sigma^2|\lambda \sim \mathrm{Exp}(0.5\,\lambda)$$

$$\lambda \sim \mathrm{Gamma}(\alpha, \beta)$$



- Assess variability of estimates using the entire posterior distribution.
- Relation to frequentist Lasso (for fixed λ).
- Variable selection using hard-shrinkage of the regression coefficients.
- Sensitivity to the selection of hyper-parameters?

Bootstrap vs Bayesian Lasso

- Both provide a way to assess variability of Lasso estimates.
- Bootstrap is faster for larger problems seems to scale closer to $\mathcal{O}(p)$ while Bayesian lasso to $\mathcal{O}(p^2)$.

Table 6.1 *Timings for Bayesian lasso and bootstrapped lasso, for four different problem sizes. The sample size is $N = 400$.*

p	Bayesian Lasso	Lasso/Bootstrap
10	3.3 secs	163.8 secs
50	184.8 secs	374.6 secs
100	28.6 mins	14.7 mins
200	4.5 hours	18.1 mins

- For GLMs the computational complexities for Bayesian Lasso grow.
- Bayesian approach leans more heavily on parametric assumptions?

! For bootstrap, the sampling distribution for coefficients close to zero maybe non-normal posing issues to valid inference (e.g.Kyung et al. (2010)).

Example: The crimes data (I)

- Crimes data

Table 2.1 *Crime data: Crime rate and five predictors, for $N = 50$ U.S. cities.*

city	funding	hs	not-hs	college	college4	crime rate
1	40	74	11	31	20	478
2	32	72	11	43	18	494
3	57	70	18	16	16	643
4	31	71	11	25	19	341
5	67	72	9	29	24	773
⋮	⋮	⋮	⋮	⋮		
50	66	67	26	18	16	940

- Continuous outcome: Crime. 5 Continuous predictors.

	Least squares	LASSO lambda.min	LASSO lambda.1se	Bayesian Lasso
funding	10.98	9.80	1.93	4.56
Hs	-6.09	-2.72	0.00	-1.01
not.hs	5.48	3.25	0.00	2.74
college	0.38	0.00	0.00	-0.23
college4	5.50	0.00	0.00	0.19

Example: The crimes data - Bootstrap

```
#Bootstrap
p<-5
B<-100000
n<-nrow(x)
beta.1se.bs<-matrix(NA,B,p)
beta.min.bs<-matrix(NA,B,p)

for (i in 1:B){
  ind<-sample(n,replace=TRUE)
  data.bs<-data[ind,]
  x.bs<-x[ind,]
  y.bs<-y[ind]
  fit<-glmnet(x.bs,y.bs,alpha=1,lambda=lambda.se)
  #fit<-cv.glmnet(x.bs,y.bs,alpha=1)
  beta.1se.bs[i,]<-as.vector(coef(fit,s="lambda.1se"))[-1]
  fit<-glmnet(x.bs,y.bs,alpha=1,lambda=lambda.min)
  #fit<-cv.glmnet(x.bs,y.bs,alpha=1)
  beta.min.bs[i,]<-as.vector(coef(fit,s="lambda.min"))[-1]
  print(i)
}
```

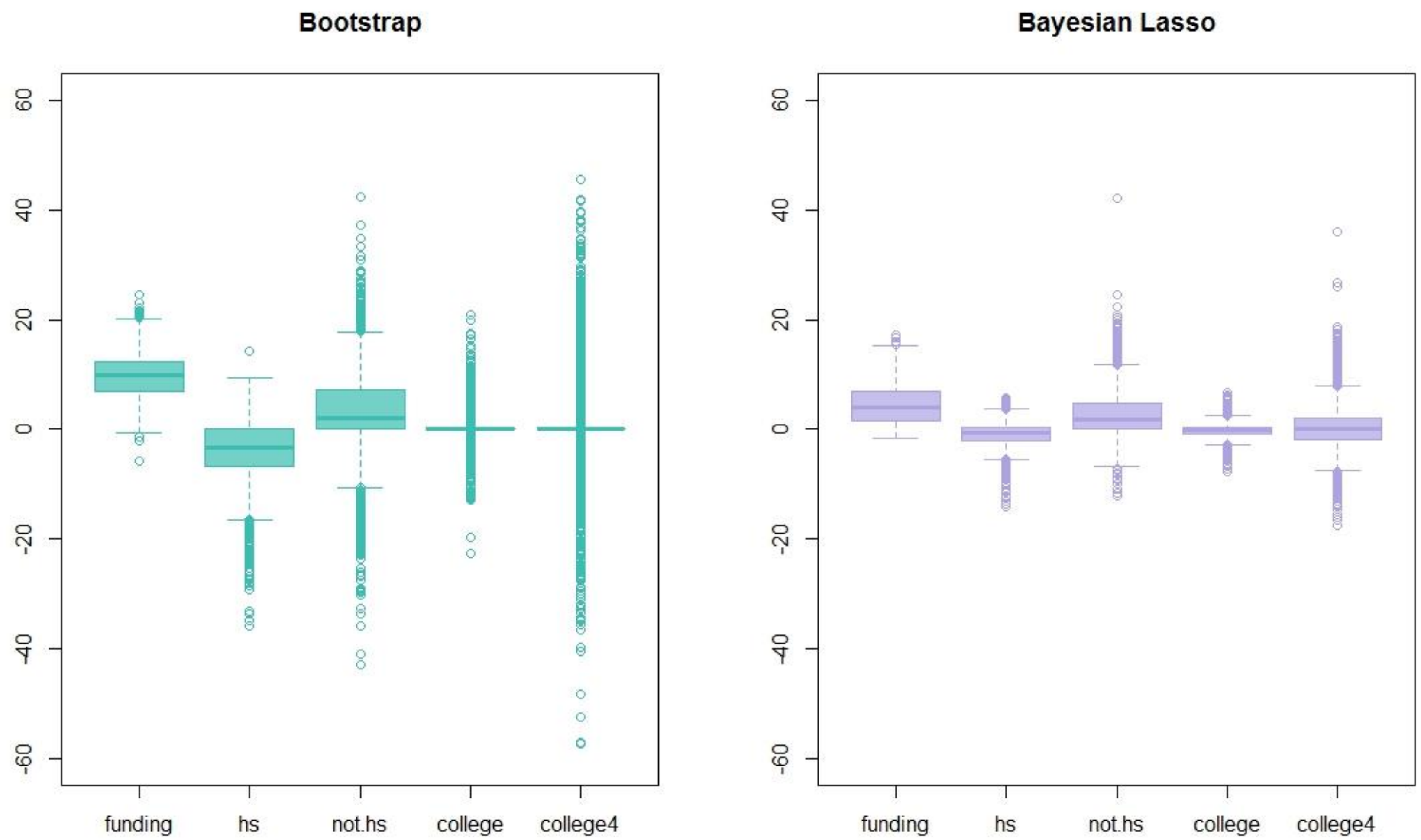
Example: The crimes data – Bayesian Lasso

```
#Bayesian lasso

#require(monomvn)

fit<- blasso(x,y,T=3000,RJ=FALSE)
plot(fit,ylim=c(-40,40))
beta.blasso<-fit$beta
beta.blasso.mean<-colMeans(fit$beta)
beta.blasso.se<-sqrt(apply(fit$beta,2,var))
beta.blasso.ci=cbind(beta.blasso.mean-1*beta.blasso.se,
beta.blasso.mean+1*beta.blasso.se)
cbind(beta.blasso.mean,beta.blasso.ci)
```

Example: The crimes data (II)



Post-selection inference for Lasso and other methods

- Significance testing from Linear Modelling
- What’s wrong with Forward Stepwise regression
- The covariance Test
  - The test
  - Example
  - Contrast with forward stepwise selection.
- Extensions (briefly).

Significance testing for Linear Modelling

- Linear regression setting:

  - *N* observations, *N*-dimensional outcome vector $\boldsymbol{Y}$.

  - $\boldsymbol{Y} = \boldsymbol{X}\beta + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N(0, \sigma^2 \boldsymbol{I}_{N\times N})$

  - *p*-dimensional vector of regression coefficients.

- Let M and $M \cup j$ be fixed subsets of *{1, . . . p}*
- Let $RSS_{M\cup j}$ and RSS_M be the residual sum of squares *(RSS)* from regression on $M \cup j$ and M.

- For nested models such as the ones above one uses the chi-square test:

  Test Statistic: $R_j = (RSS_{M\cup j} - RSS_M)/\sigma^2$. (*)

  and compare it to the ${\chi_1}^2$ distribution.

* If σ^2 is unknown estimate it from the data and use *F*-test.

What's wrong with forward stepwise regression (I)?

- Standard forward stepwise selection: repeated comparison of nested models.

  → Enters predictors one at a time, in a stepwise manner, comparing nested models as described earlier.

  → Chooses the predictor that reduces most the residual sum of squares, *after checking all available predictors*

- RSS_k: residual sum of squares for the model with k predictors

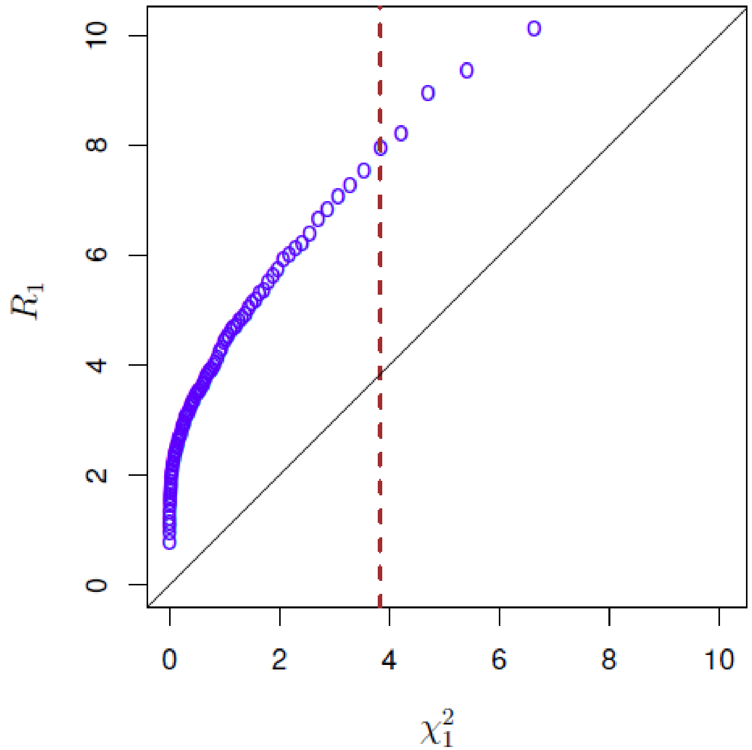
- Similarly as before , form the test-statistic:

  $R_k = 1/\sigma^2(RSS_{k-1} - RSS_k)$

  and compare it to the ${\chi_1}^2$ distribution.

What’s wrong with forward stepwise regression (II)?

- This significance testing is alright when comparing nested models that are fixed. In stepwise regression, the sets we are comparing *are not fixed*, but a result of an adaptive procedure.
- This adaptivity invalidates the use of the ${\chi_1}^2$ null distribution for the statistic.
- R_1, refers to the addition of the *first* predictor in a model (Figure 6.8 example with 10 predictors).
- Using simulation, we see that the theoretical quantiles of the assumed distribution *do not agree* with the quantiles of R_1.
- Connection to degrees of freedom (later).



What's wrong with forward stepwise regression (III)?

- Reason: The chi-square statistic assumes that the models are pre-specified and *not* data-driven.
- The fact that the strongest predictor is chosen among all available predictors given the data, yields a larger drop (improvement) in RSS that it is actually expected. As a results the p-values are biased downward. The method is too 'liberal'.
- One solution would be to perform this form of variable selection on a different sample (split sample), so assess the drop in RSS in *new data*. That would entail a loss of information.

- Target of recent work: Derive a test that accounts for adaptivity in model selection.

The Covariance Test (I)

- Based on the LAR algorithm of Efron et al. (2004) which traces the solutions as λ decreases from ∞ to 0.
- Consider the knots returned by the LAR algorithm
$$\lambda_1 > \lambda_2 > \cdots > \lambda_K.$$
'knots'= the values of the regularization parameter where there is a change in the set of active predictors.
- **Proposal**: A test statistic for the significance of predictor added at the k^{th} step, at λ_κ.
- **Notation**

- $\mathbf{A}$: set of predictors with non-zero coefficients just before step k (active set).
- Predictor *j* enters at λ_κ.
- $\widehat{\boldsymbol{\beta}}(\lambda_{\kappa+1})$: the solution at the *next knot* with active set $\boldsymbol{A} \cup \{j\}$.
- Refit Lasso, with $\lambda = \lambda_{\kappa+1}$ just using predictors in $\boldsymbol{A}$. Gives $\widetilde{\boldsymbol{\beta}}_{\mathbf{A}}(\lambda_{\kappa+1})$.

The Covariance Test (II)

- Covariance Test statistic:

$$T_\kappa = \frac{1}{\sigma^2}\left(< \boldsymbol{y}, \boldsymbol{X}\widehat{\boldsymbol{\beta}}(\lambda_{\kappa+1}) > - < \boldsymbol{y}, \boldsymbol{X}_A\, \widetilde{\boldsymbol{\beta}}_{\mathbf{A}}(\lambda_{\kappa+1}) >\right)$$

(Lockhart et al. (2014). As significance test for the Lasso), The Annals of Statistics, 42(2), 413-468.

- Notes:

→ Covariance between outcome and fitted model can be attributed to the predictor just added.

→ Intuitively: The larger the covariance of $\boldsymbol{y}$ with $\boldsymbol{X}\widehat{\boldsymbol{\beta}}$ compared to that with $\boldsymbol{X}_A\, \widetilde{\boldsymbol{\beta}}_{\mathbf{A}}$, the more important the role of predictor *j* in the model $\boldsymbol{A} \cup \{j\}$.

→ Why evaluate solution at $\lambda = \lambda_{\kappa+1}$?

The Covariance Test (III)

- Covariance Test statistic:

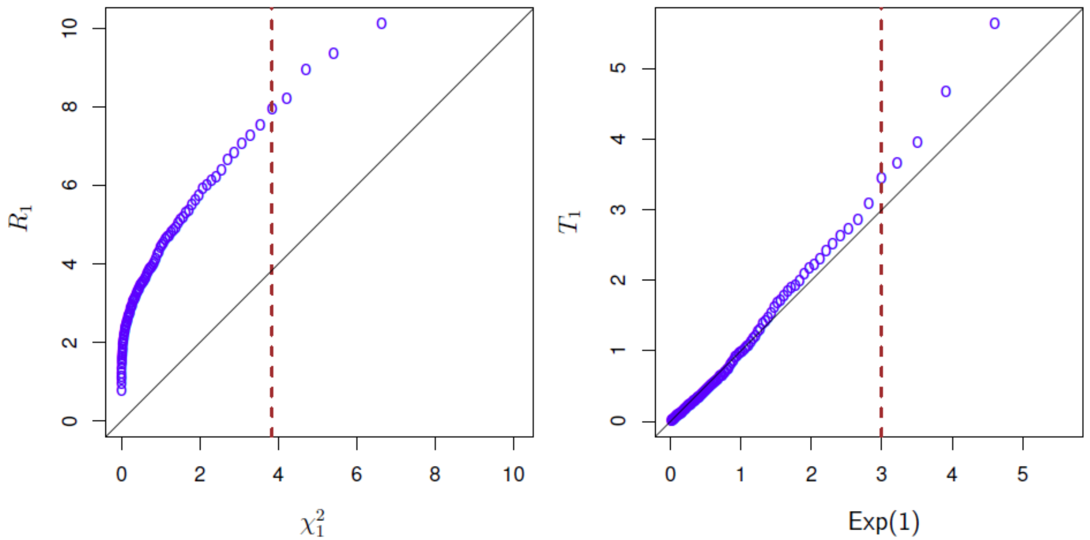
$$T_\kappa = \frac{1}{\sigma^2}\left(< \boldsymbol{y}, \boldsymbol{X}\widehat{\boldsymbol{\beta}}(\lambda_{\kappa+1}) > - < \boldsymbol{y}, \boldsymbol{X}_A\, \widetilde{\boldsymbol{\beta}}_{\mathbf{A}}(\lambda_{\kappa+1}) >\right)$$

- Choice of this form? Other choices?

- $T_\kappa \xrightarrow{d} Exp(1)$.

- $Exp(1)$ is the analogue of ${\chi_1}^2$ distribution for adaptive fitting.

- Quantile-quantile plots we saw earlier.



(a) Forward stepwise (b) Lasso

The Covariance Test (IV)

- Covariance Test statistic:

$$T_\kappa = \frac{1}{\sigma^2}\left(< \boldsymbol{y}, \boldsymbol{X}\widehat{\boldsymbol{\beta}}(\lambda_{\kappa+1}) > - < \boldsymbol{y}, \boldsymbol{X}_A \widetilde{\boldsymbol{\beta}}_{\mathbf{A}}(\lambda_{\kappa+1}) >\right)$$

- Conditions:

  - On data matrix $\boldsymbol{X}$: signal variables (with non-zero coefficients) are not too highly correlated with the noise variables.
  - On the outcome: it is normally distributed.
  - $N, p \to \infty$
  - non-zero coefficients are large enough in magnitude.

- Interpretation – important!

  The p-values at each row are *conditional on the variables in the active set* at a given knot!

Covariance Test – Diabetes example

Table 6.2 *Results of forward stepwise regression and LAR/lasso applied to the diabetes data introduced in Chapter 2. Only the first ten steps are shown in each case. The p-values are based on (6.4), (6.5), and (6.11), respectively. Values marked as 0 are* < 0.01.

	Forward Stepwise		LAR/lasso		
Step	Term	p-value	Term	p-value Covariance	p-value Spacing
1	bmi	0	bmi	0	0
2	ltg	0	ltg	0	0
3	map	0	map	0	0.01
4	age:sex	0	hdl	0.02	0.02
5	bmi:map	0	bmi:map	0.27	0.26
6	hdl	0	age:sex	0.72	0.67
7	sex	0	glu^2	0.48	0.13
8	glu^2	0.02	bmi^2	0.97	0.86
9	age^2	0.11	age:map	0.88	0.27
10	tc:tch	0.21	age:glu	0.95	0.44

Covariance Test – Example

Implementation in R

```
#Covariance test

require(covTest)

fit<-lars(x,y)

covTest(fit,x,y)
```

Covariance Test: Practical issues

1. Lasso entered a predictor into the active set at each step. Possibility of a predictor entering active set *more than once* along the lasso path (since it may leave it at some point). Each entry is treated as separate problem. Discuss.
2. By design the covariance test is applied in sequential manner, estimating p-values for every predictor as it enters the lasso path. A more difficult problem is to test the significance of any of the active predictors at some arbitrary value of the tuning parameter.

   → Proposed extensions

Connection to the degrees of freedom

- Connection between the covariance test and degrees of freedom of fitting procedure.
- Linear regression setting the definition of df is:

$$df(\hat{y}) = 1/\sigma^2 \sum_{i=1}^{n} Cov(y_i, \hat{y_i})$$

- The more adaptive a procedure is, the higher the covariance is, the more the degrees of freedom. With *k* predictors in the model, *a forward stepwise procedure* uses substantially *more* than k degrees of freedoms.

- What changes with *Lasso*?

→ For a model with *k* predictors, the degrees of freedom for a Lasso fit is equal *k* (in expectation or exactly).

→ Reason is *shrinkage*! Chooses adaptively *but also shrinks*. As a result the cost in extra degrees of freedom induced by adaptivity is balanced out by shrinking non-zero coefficients (by a right amount so the df is totally *k*).

The Spacing Test (briefly)

- More general scheme that can be applied for
  - exact p-values and confidence intervals in the Gaussian case.
  - for finite N and p.
  - works for any $\boldsymbol{X}$.

- It can be applied to any procedure where the selection event can be seen as a set of inequalities in $\boldsymbol{y}$.
  - to successive steps of the LAR algorithm
  - to Lasso at a fixed choice of the tuning parameter λ.
  - to forward stepwise regression. See:

  Loftus and Taylor (2014). *A significance test for forward stepwise model selection*, arXiv:1405.3920v1 .

- R-package `selectiveInference`

Conclusions

- Bootstrap and Bayesian lasso early approaches to inference for Lasso coefficients. No studies regarding the coverage properties of Confidence/ Credible Intervals?
- More recent approaches include the Covariance Test, Spacing Test.
- Ongoing line of research- not limiting to Gaussian Outcomes (Ref).
- Relevant to Inference and explanation. Thoughts about presentation of results for prediction modelling?