MA8701 Advanced methods in statistical inference and learning

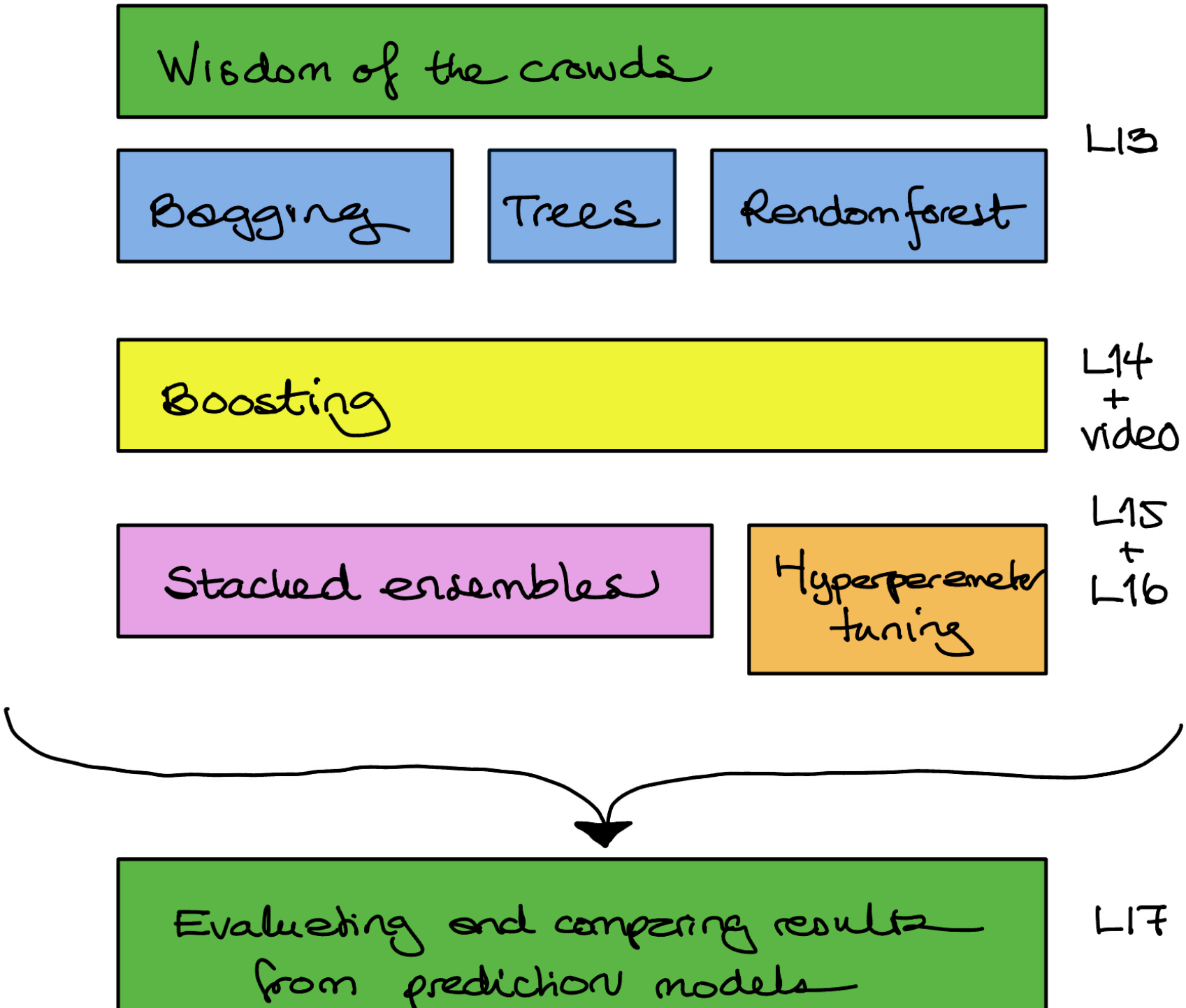
Part 3: Ensembles. L17: Evaluating and comparing results from prediction models

Mette Langaas

Lectured 13.03.2023

~~3/12/23~~

Before we start



Literature

There is a long list of references in the end of this document, but for our reading list this document will suffice.

Evaluating and comparing results from prediction models

We will only consider using *one data set*. For comparing methods across many data sets see Boulesteix et al (2015).

We are not interesting in general "unconditional" results (for all possible training sets from some distribution) - and not to know if method A *in general* is better than method B in situations similar to ours.

We also have the "No free lunch theorem" of Wolpert (1996) stating that there is no such thing as the "best" learning algorithm.

We consider two different set-ups:

Data rich situation:

- We have used our *training set* to tune our model (choosing hyperparameters) - possibly by using cross-validation or some other technique.
- Then we have fitted the finally chosen model to the full training set, and used this final model to make predictions on the independent *test set*.
- If we want to compare results from two or more prediction models (A and B), when the same test set is used for all the models.

Data poor situation:

- We don´t have enough data to set aside observations for a test set.
- We need to use some type of resampling to evaluate and compare prediction models - the "common" choice is k-fold cross-validation.
- This is more difficult than for the data rich situation, because now *independence* of observations for testing cannot be assumed (more below).

What do we want to report?

Classification

We will only look at binary classification, but parts of the results may be used for each of the categories (vs the rest) for more than two classes.

- Estimate and confidence interval for misclassification rate or ROC-AUC (or other) on test observations for one prediction model.
- Is the misclassification rate (or ROC-AUC, or other) for prediction method A better than for prediction method B?
- Can this be extended to more than two methods?

This is by far the most popular situation in the literature.

Regression

Relate to ESL Ch7.1 with Err and Err_T.

- Estimate and confidence interval for evaluation criterion (mean square error of predictions) on test observations for one prediction model.
- Is prediction model A better than prediction model B?
- Can this be extended to more than two methods?

Much more difficult to "find" literature with methods here than for classification - seems to be far less popular.

?

Keep in mind that not only error rates govern which prediction models to use, also aspects like

- training time and
- interpretability plays an important role.

There might be

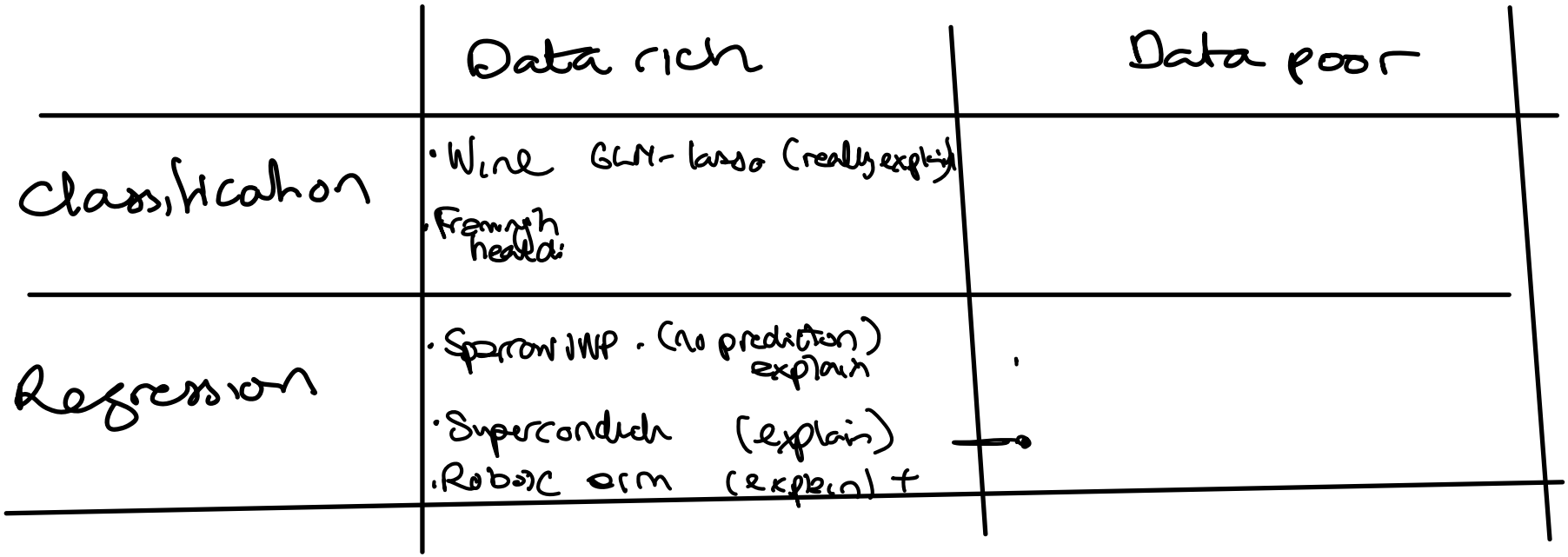
- controllable and
- uncontrollable factors

that influence the model fit and add variability to our model predictions.

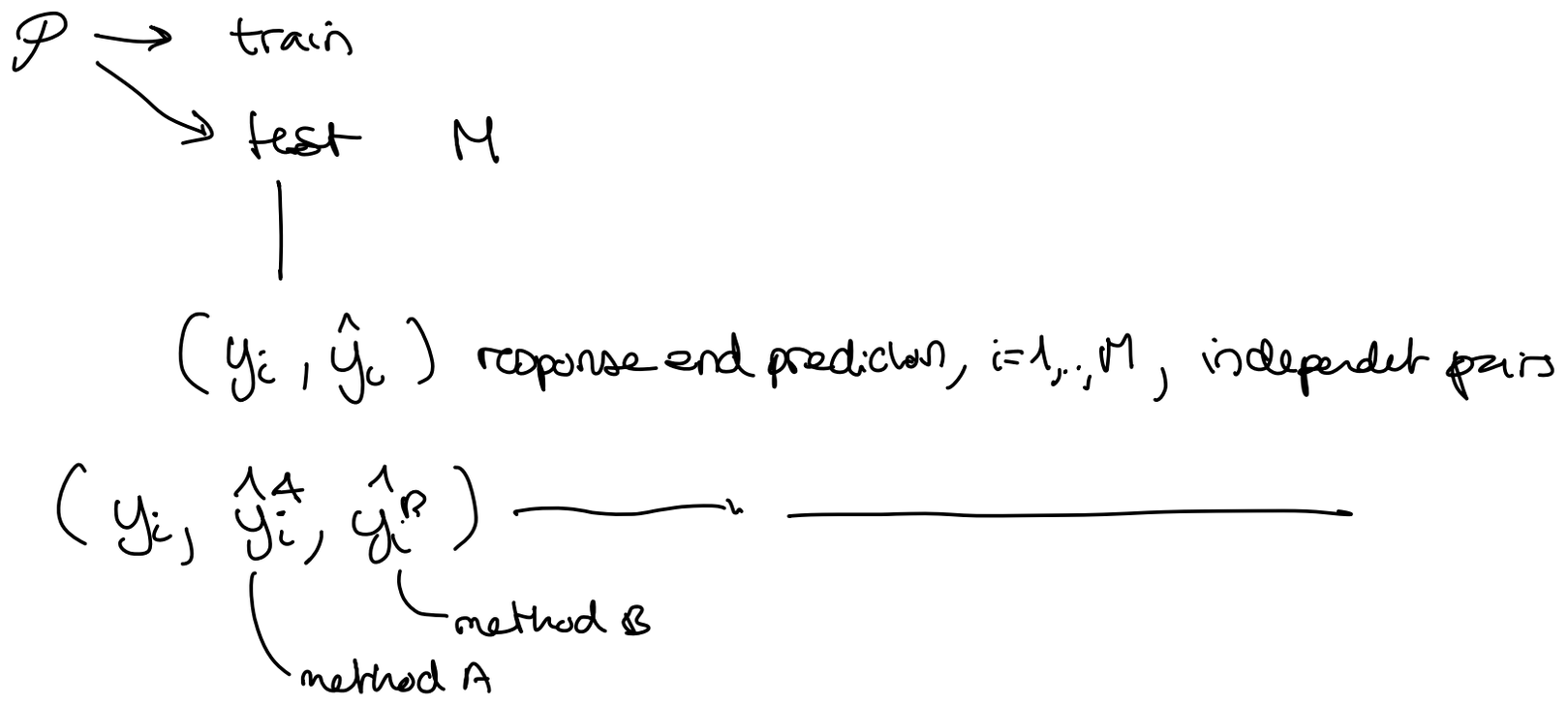
It is always wise (helpful) to present results in graphical displays.

Group discussion

For your data analysis project, which of the above is relevant? Explain!



DATA RICH SITUATION



Example

We will use the classical data set of *diabetes* from a population of women of Pima Indian heritage in the US, available in the R `MASS` package. The following information is available for each woman:

- diabetes: 0= not present, 1= present
- npreg: number of pregnancies
- glu: plasma glucose concentration in an oral glucose tolerance test
- bp: diastolic blood pressure (mmHg)
- skin: triceps skin fold thickness (mm)
- bmi: body mass index (weight in kg/(height in m)^2)
- ped: diabetes pedigree function.
- age: age in years

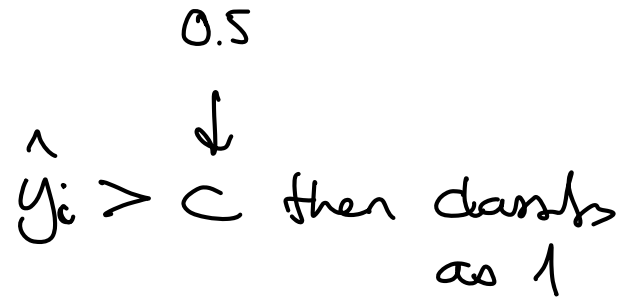
We will use the default division into training and test in the MASS library, with 200 observations for training and 332 for testing.

```
Pima.tr$diabetes=as.numeric(Pima.tr$type)-1
Pima.te$diabetes=as.numeric(Pima.te$type)-1
```

Test set classification

223 non-diabetes and 109 diabetes cases

```
[1] "Lasso"
   classlasso
      0   1
0   213  10
1    61  48
[1] "Random forest"
   classrf
      0   1
0   191  32
1    46  63
```



Let M = number of independent trials ←

X = number of correct classifications

$X \sim \text{binomial}(M, p)$ $\quad E(X) = Mp$, $\text{Var}(X) = Mp(1-p)$

↑

true "classic classification rate"

$\hat{p} = \frac{X}{M}$ → how to get a CI for p?

A common way to construct a confidence interval for the success probability is to use the normal approximation

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{M}}} \sim N(0,1)$$

which gives the $(1-\alpha)100\%$ confidence interval

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{M}}$$

The Agresti-Coull interval adds 4 trials and 2 sucesses for a better performance (asymptotic method).

$$\left(\tilde{p} = \frac{X+2}{M+4} \right)$$

Exact versions (not using asymptotic normality) are the

- Clopper-Pearson intervals
- Blaker intervals by Blaker (2000) as implemented in Klaschka (2010).

Clopper - Pearson :

$(1-\alpha) \cdot 100\%$ CI

$$H_0 : p = p_0 \quad \text{vs} \quad H_1 : p \neq p_0$$

Tail method: solve for p_0 to get an upper and lower limit of CI for p

$$\left.\begin{aligned} &\sum_{k=0}^{x} \binom{n}{k} p_0^k (1-p_0)^{n-k} \leq \frac{\alpha}{2} \\ &\sum_{k=x}^{n} \binom{n}{k} p_0^k (1-p_0)^{n-k} \leq \frac{\alpha}{2} \end{aligned}\right\} \text{for which } p_0 \text{ does this hold}$$

Since x is discrete it is often not possible to get exactly $\frac{\alpha}{2}$

$\Rightarrow$ method said to be conservative $\Rightarrow$ too wide interval

Fig 1.3 from Agresti shown in class.

Blaker's CI: found from inverting a two-sided p-value

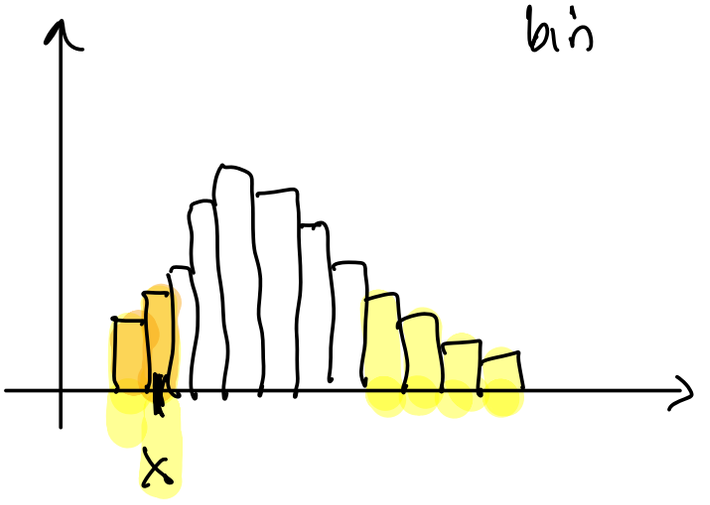
$H_0: p = p_0$ vs $H_1: p \neq p_0$

Test statistic

$$T = \min(P(X \leq x), P(X \geq x))$$

P-value: $\sum_{x: T \leq t} P(X = x)$

- $x: T \leq t$: sum over all possible outcomes x
- t: observed



Then invert into CI = all p_0 when

p-value $\leq \alpha$ $(1-\alpha)$

R-package blakeCI

Figure 1 from Blaker (2000) shown in class.

```
[1] "lasso"

[1] "Normal approx CI"

[1] 0.7420393 0.8302498

[1] "Clopper Pearson CI"


    Exact binomial test

data:  X and M
number of successes = 261, number of trials = 332, p-value < 2.2e-16
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.7380713 0.8290302
sample estimates:
probability of success
             0.7861446

[1] "Blaker CI"

[1] 0.7386136 0.8276581
```

Normal

CP

coverage big enough

Blaker ← valid and most powerful (compared to CP)

not biased

```
[1] "randomforest"

[1] "Normal approx CI"

[1] 0.7194560 0.8106645

[1] "Clopper Pearson CI"


    Exact binomial test

data:  X and M
number of successes = 254, number of trials = 332, p-value < 2.2e-16
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.7156949 0.8096267
sample estimates:
probability of success
             0.7650602

[1] "Blaker CI"

[1] 0.7159697 0.8096206
```

Next: Is method A better then method B?

lasso

0.786

$\hat{P}_{lasso}$

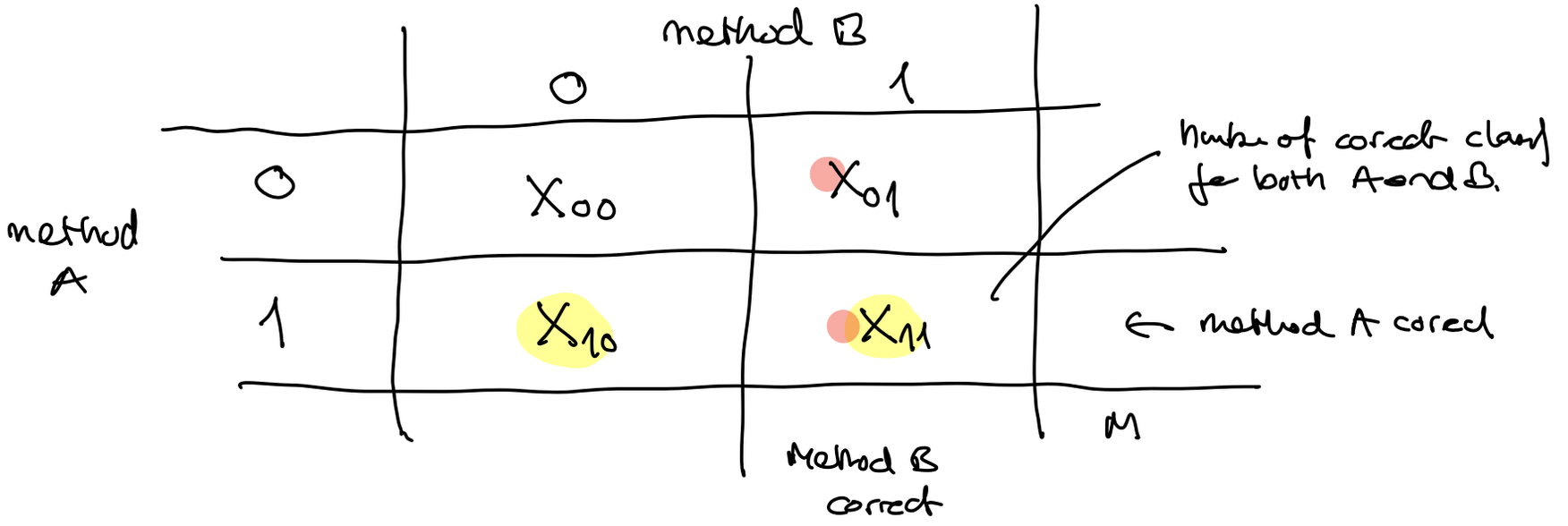
RF

0.765

$\hat{P}_{RF}$

- interpretability?
- classif error - might not be the to consider ROC-AUC

For each observation $i = 1, .., M$
count the nr of
0 = fail
1 = correct classification for both A, B



$$(X_{01}, X_{10}, X_{00}, X_{11}) \sim \text{multinomal}(M, q_{01}, q_{10}, q_{0\cdot}, q_{11})$$

$$H_0: q_{10} + q_{11} = q_{01} + q_{11} \iff q_{10} = q_{01}$$

↑ equally good

Test statistic:

$$\frac{(X_{10} - X_{01})^2}{(X_{01} + X_{10})} \approx \chi^2_1$$

↑

McNemar's test

The sum $X_{01} + X_{10}$ need to be large (rule of thumb at least 25), unless a two-sided binomial version of the test is recommended (with $n = X_{01} + X_{10}$ and $p = 0.5$ and number of successes equal X_{01}). This is a conditional test (conditional tests are valid).

An exact conditional p-value can also be calculated by enumeration.

```
tab=table(classlasso==test$diabetes,classrf==test$diabetes)
tab
```

```
       FALSE TRUE
  FALSE    52   19
  TRUE     26  235
```

CONCLUSION:
Lasso RF

```
mcnemar.test(tab,correct=FALSE)
```

H_0: lasso eq RF H_1: not so

⇒

```
    McNemar's Chi-squared test

data:  tab
McNemar's chi-squared = 1.0889, df = 1, p-value = 0.2967
```

```
binom.test(tab[1,2],n=tab[1,2]+tab[2,1],p=0.5)
```

```
    Exact binomial test

data:  tab[1, 2] and tab[1, 2] + tab[2, 1]
number of successes = 19, number of trials = 45, p-value = 0.3713
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.2765670 0.5784967
```

Confidence intervals for paired proportions

Confidence interval for the difference between success-proportions can be calculated using for example an asymptotic Wald interval or by inverting hypotheses tests p-values.
See Fagerland et al (2014) for this and other choices, not R package but see references for R-scripts.
The package `ExactCIdiff` is presented in the R Journal

ROC-AUC

In a two class problem - assume the classes are labelled "-" (non disease,0) and "+" (disease,1). In a population setting we define the following event and associated number of observations.

	Predicted -	Predicted +	Total
True -	True Negative TN	False Positive FP	N
True +	False Negative FN	True Positive TP	P
Total	N*	P*	

(N in this context not to be confused with our sample size...which we have called M)

Remind yourself — what is this and how to proceed to make ROC and calculate ROC-AUC?

Let now X = covariates and $f(x)$ the prediction model $f(x) \in [0,1]$ and classify to class 1 if $f(x) > c$.

The true class is $Y \in \{0,1\}$

Probabilistic confusion matrix

	Truth $P(Y=1)=\pi$	Truth $P(Y=0)=1-\pi$
Predicted $P(f(x)>c)$	$P(f(x)>c \cap Y=1)$	$P(f(x)>c \cap Y=0)$
$P(f(x)\leq c)$	$P(f(x)\leq c \cap Y=1)$	$P(f(x)\leq c \cap Y=0)$

Realized confusion

	$Y=1$ P	$Y=0$ N	
Predicted $P(f(x)>c)$	TP	FP	P^*
$P(f(x)\leq c)$	FN	TN	N

Sensitivity (recall)

$$P(f(X) > c \mid Y = 1) \qquad TPR = \frac{TP}{P} \leftarrow TP + FN$$

1 − Specificity

$$P(f(X) > c \mid Y = 0) \qquad FPR = \frac{FP}{N}$$

Positive predictive value PPV (precision)

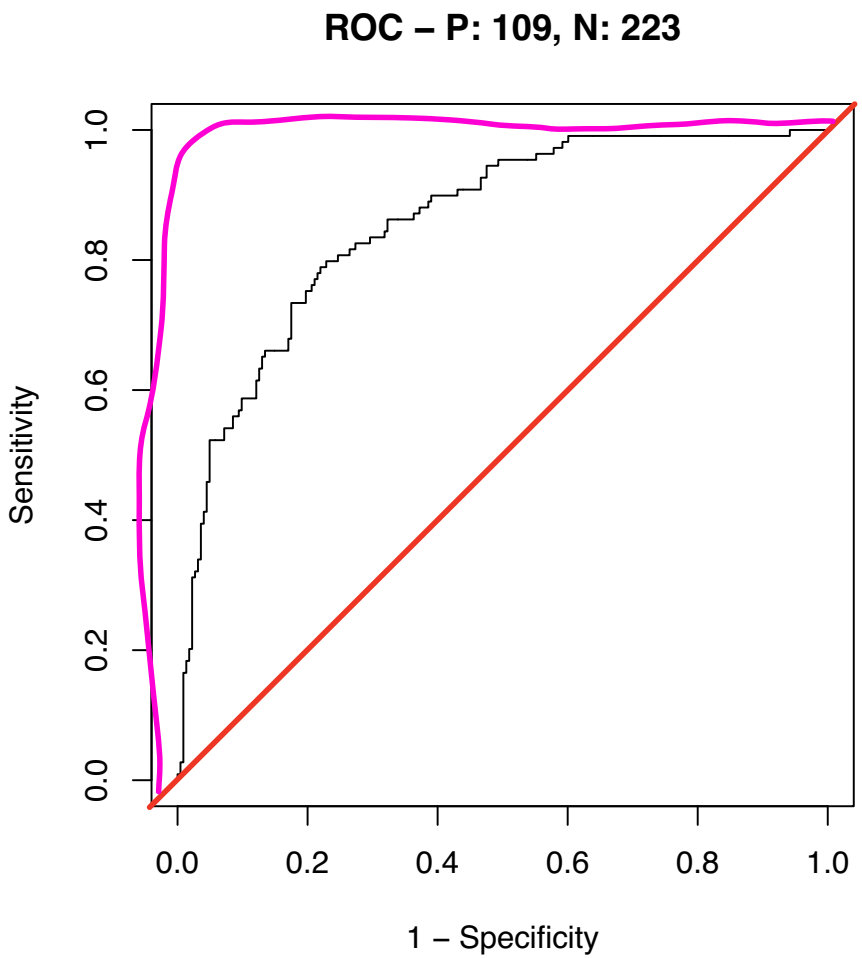
$$P(Y = 1)\ f(X > c) \qquad PPV = \frac{TP}{P^*}$$

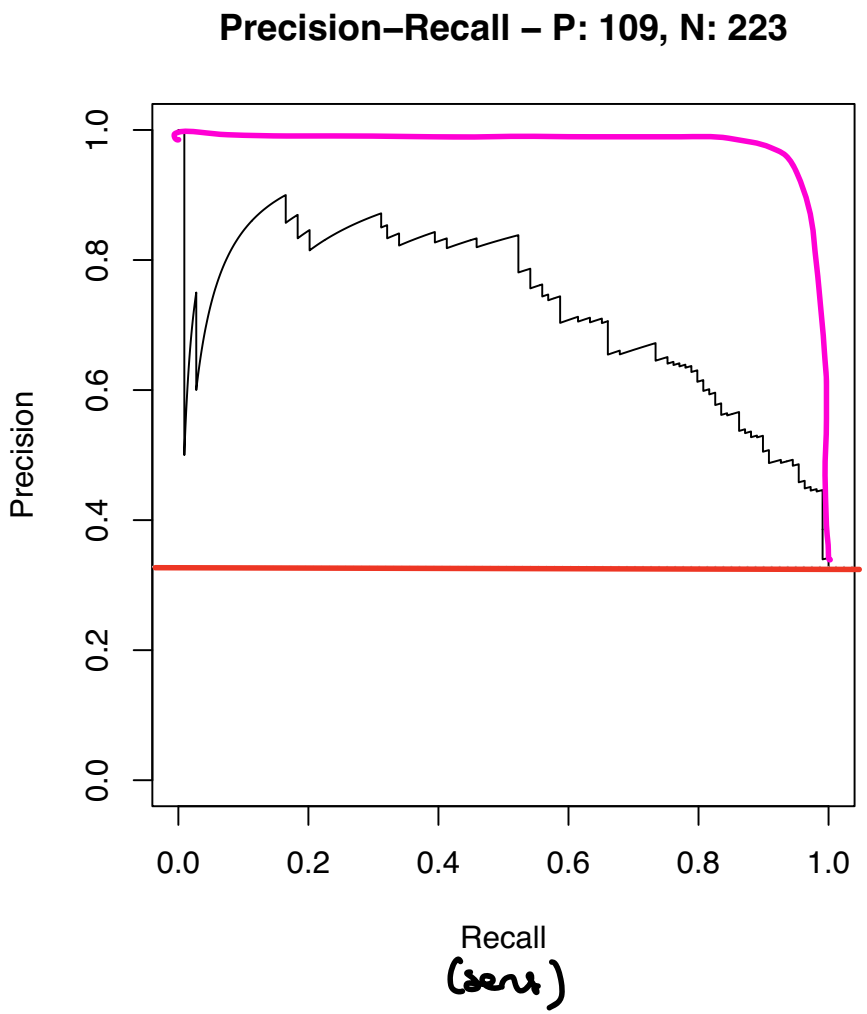
ROC: sensitivity (y) vs 1-specificity (x) for all c

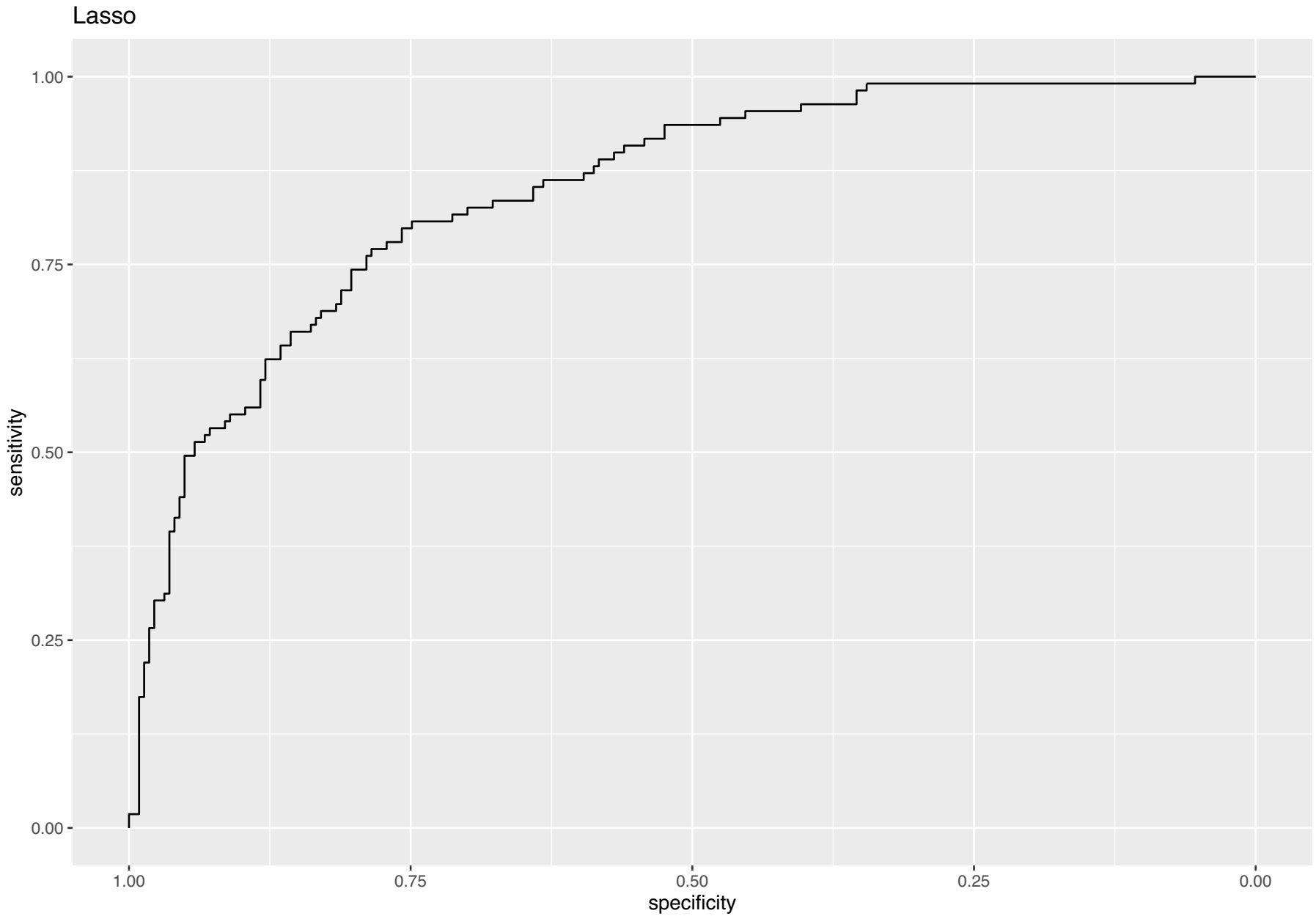
PR : PPV (y) vs sensitivity (x) —"—

LASSO

noskill






Lasso
1.00
0.75
0.50
0.25
0.00
sensitivity
1.00
0.75
0.50
0.25
0.00
specificity

ROC-AUC: the probability that a randomly selected positive sample (obs) will rank higher than a randomly selected negative sample.

$$P_0\left(f(X_1) > f(X_2) \mid Y_1 = 1, Y_2 = 0\right) \Rightarrow \text{Wilcoxon MW}$$

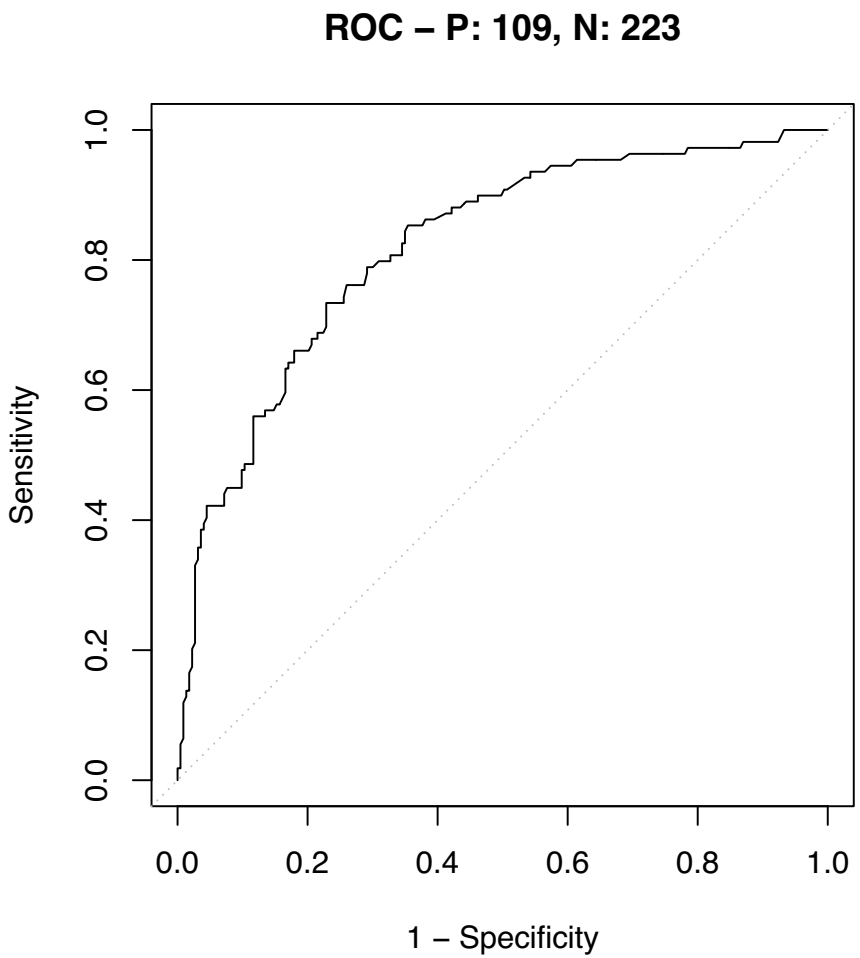
Below we use:

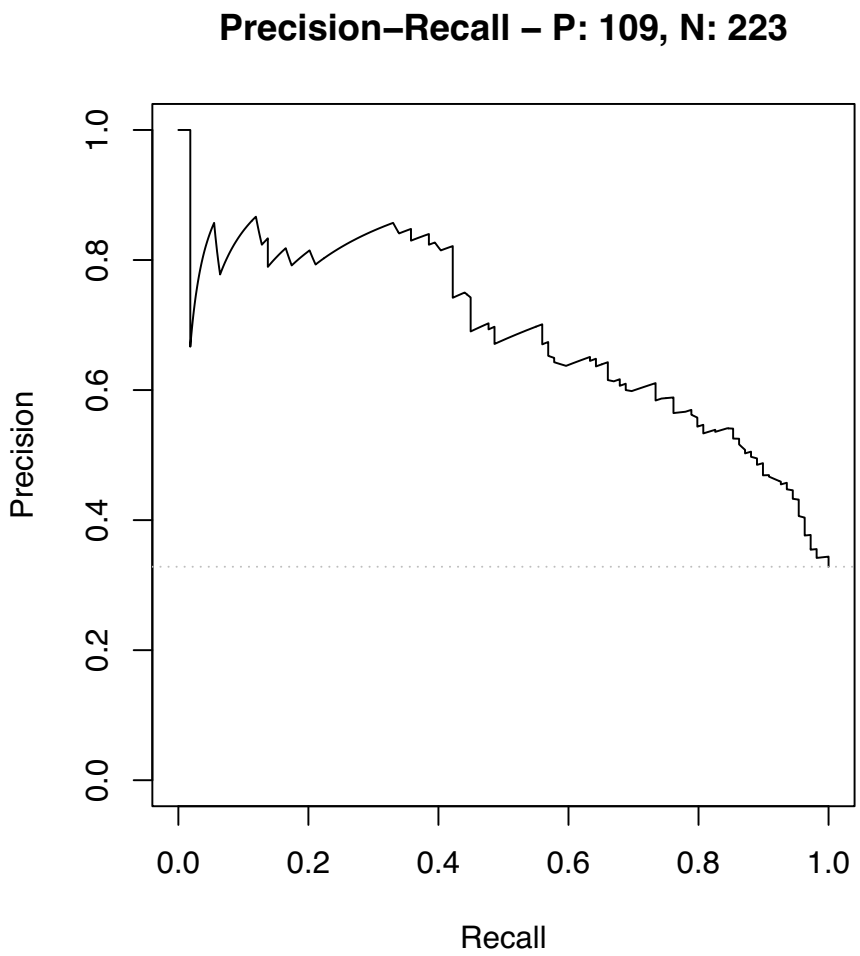
- DeLong et al confidence intervals for the ROC and the ROC-AUC for each prediction method.
- DeLong et al test for two paired (correlated) ROC curves. This test is based on asymptotic normal theory for the U-statistic.

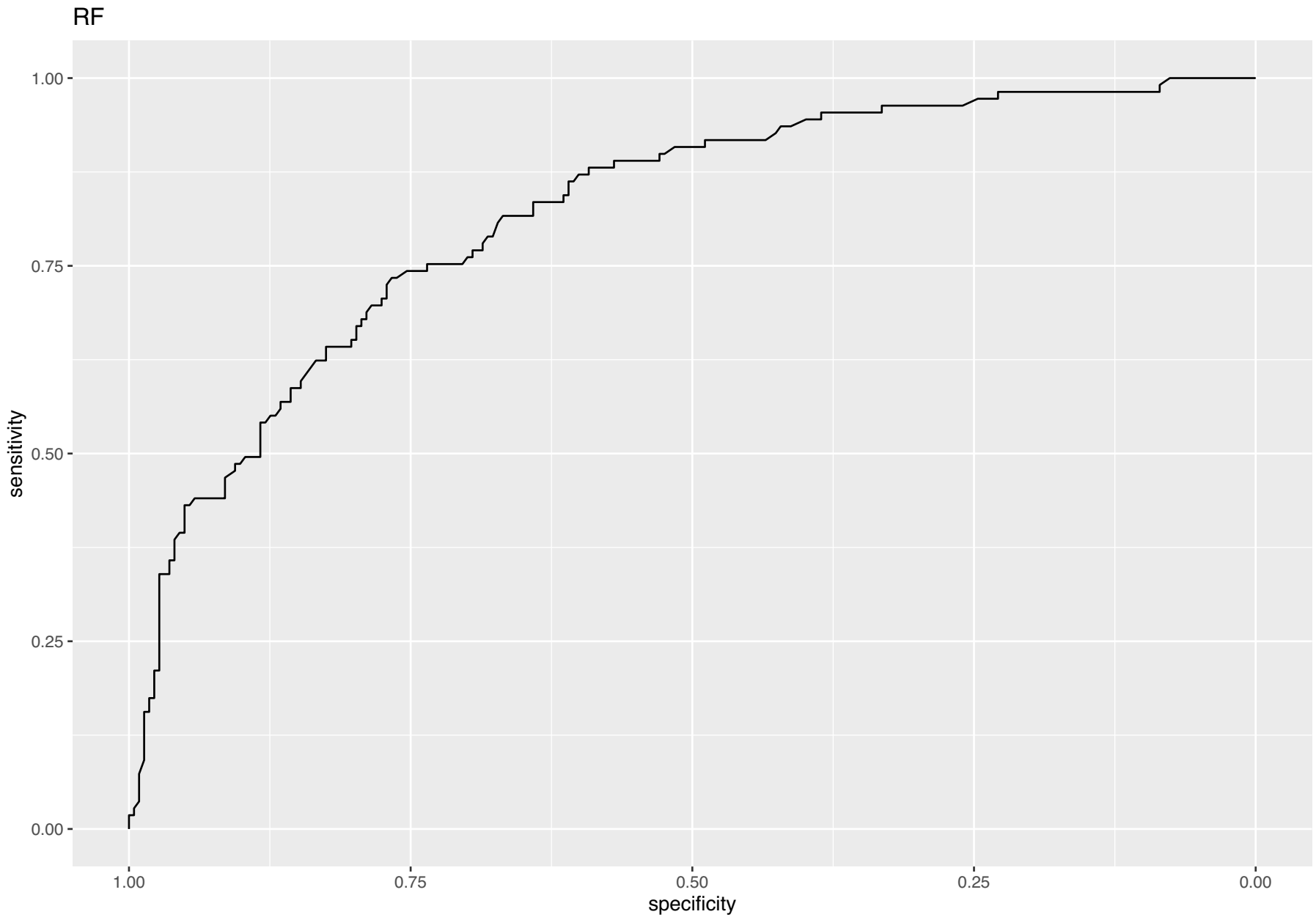
```
[1] "Lasso ROC-AUC with CI"

Area under the curve: 0.8486

95% CI: 0.8054-0.8918 (DeLong)
```


ROC – P: 109, N: 223
Sensitivity
1 – Specificity
0.0
0.2
0.4
0.6
0.8
1.0


Precision–Recall – P: 109, N: 223
Precision
Recall
0.0
0.2
0.4
0.6
0.8
1.0


RF
1.00
0.75
0.50
0.25
0.00
sensitivity
1.00
0.75
0.50
0.25
0.00
specificity

```
[1] "RF ROC-AUC with CI"

95% CI: 0.7709-0.8667 (DeLong)

[1] "Comparing AUC for lasso and RF"


    DeLong's test for two correlated ROC curves

data:  lassoroc and rfroc
Z = 1.972, p-value = 0.04861
alternative hypothesis: true difference in AUC is not equal to 0
95 percent confidence interval:
 0.000181508 0.059472090
sample estimates:
AUC of roc1 AUC of roc2
  0.8486033   0.8187765
```

⇒ Lasso better than RF for ROC-AUC

Regression

For regression we would like to focus on providing an estimate for the Err_T for a squared error rate.

$$\text{Err}_T = \text{E}[L(Y, \hat{f}(X)) \mid T]$$

Here the expected value is with respect to (X, Y), but the training set is fixed - so that this is the test set error is for this specific training set T.

In ELS Ch7.1 we saw that the *mean squared error on the test set* was a natural estimator.

In the unconditional version, we take expected value over ALL that is random - including the training set

$$\mathrm{Err} = \mathrm{E}(\mathrm{E}[L(Y, \widehat{f}(X)) \mid T]) = \mathrm{E}_T[\mathrm{Err}_T]$$

However, we did not work to provide an estimate of the *variability* of this estimate - or how to provide a confidence interval for Err_T.

Let the mean *squared* error on the test set be denoted $\widehat{\text{MSEP}}$.

If we can assume that the "residuals" on the test set $y_i - \hat{y}_i$ follow a normal distribution with some mean μ_i and some variance σ_i^2, then there is a relationship between the $\widehat{\text{MSEP}}$ and a sum of non-central χ^2 distributions, see Faber (1999). However, it is not clear how to turn that into a confidence interval for Err_T.

Not seen in literature: Another possibility is to use bootstrapping on the "test set residuals". This can provide a bootstrap confidence interval for the Err_T. With bootstrapping it would also be possible to look at randomly flipping the A and B method to get the distribution of the $\widehat{\text{MSEP}}$ under the null hypothesis that the two methods are equal, and use tha percentage of times the bootstrap samples are larger than the observed $\widehat{\text{MSEP}}$ to be the p-value.

-DATA POOR-

Cross-validation

Remember from ELS Ch 7.10 that with cross-validation the Err estimate:

- The allocation of observation $\{1, \ldots, N\}$ to folds $\{1, \ldots, K\}$ is done using an indexing function $\kappa : \{1, \ldots, N\} \to \{1, \ldots, K\}$, that for each observation allocate the observation to one of K folds.
- Further, $\hat{f}^{-k}(x)$ is the fitted function, computed on the observations except the kth fold (the observations from the kth fold is removed).
- The CV estimate of the expected prediction error $\text{Err} = \text{E}_T\text{E}_{X^0,Y^0}[L(Y^0, \hat{f}(X^0)) \mid T]$ is then

$$\text{CV}(\hat{f}) = \frac{1}{N} \sum_{i=1}^{N} L(y_i, \hat{f}^{-\kappa(i)}(x_i))$$

Can the validation fold results be handled like the test set?

Question:

Can we handle the predictions in the hold-out folds $\widehat{y}_i$ as *independent predictions* at the observations x_i - as we did in the data rich situation above (when we had a separate test set and used the "same" full training set for fitting the model)?

To address this a simulation study is conducted. Here

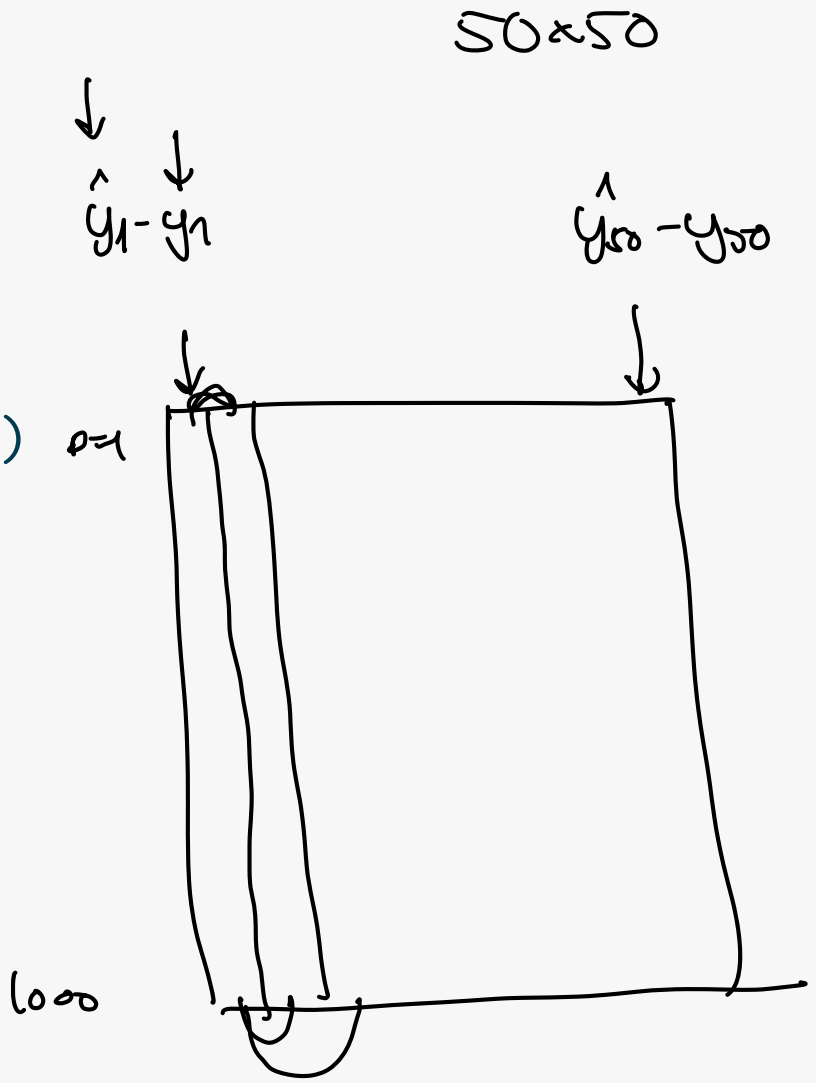
- data are simulated to follow a simple linear regression.
- $N = 50$.
- The observations are divided into 5 fold of 10 observations.
- Then a 5-fold CV is performed where a simple linear regression is fitted on the training folds and predictions are performed in the test fold.
- Residuals are then formed for the test fold.

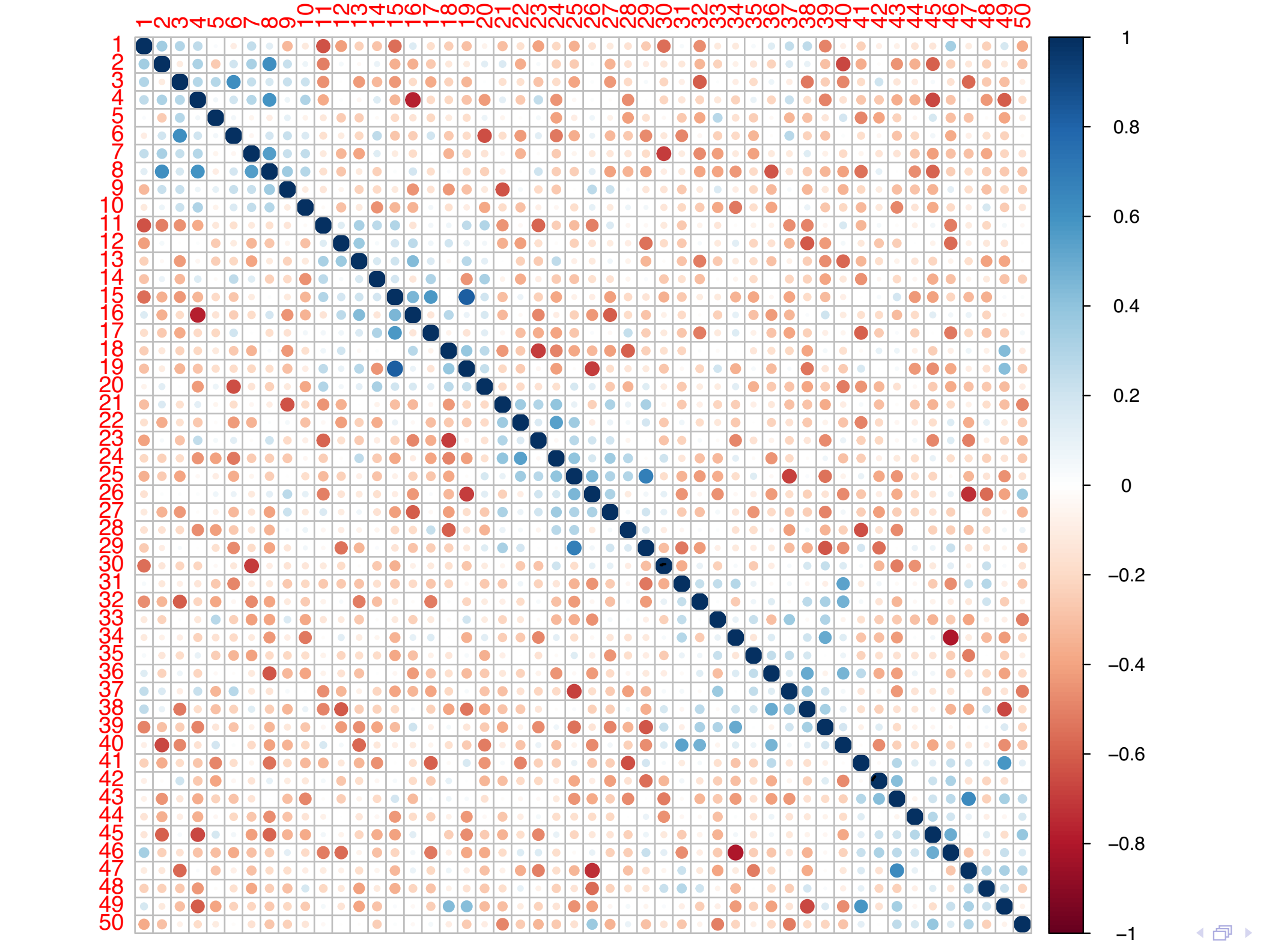
The simulations are repeated B=1000 times, and correlation between the N residuals for the test folds are calculated.

The question to be checked is if the residuals for observations in the same fold are correlated in a different way than residuals in different folds. If that is the case, then the residuals can not be seen to be independent, and standard methods to construct CI and perform a test is not valid.

```r
K=5
B=1000
N=50
b0=0
b1=2
sigma=0.2
k=rep(1:K,each=N/K)
predmat=matrix(ncol=N,nrow=B)
resmat=matrix(ncol=N,nrow=B)

set.seed(123)
for (b in 1:B)
{
  x=runif(N,0,1)
  eps=rnorm(N,0,sigma)
  y=b0+b1*x+eps
  for (i in 1:K)
  {
    fit=lm(y~x,subset=(k!=i))
    predmat[b,k==i]=predict(fit,newdata=data.frame(x=x[k==
```

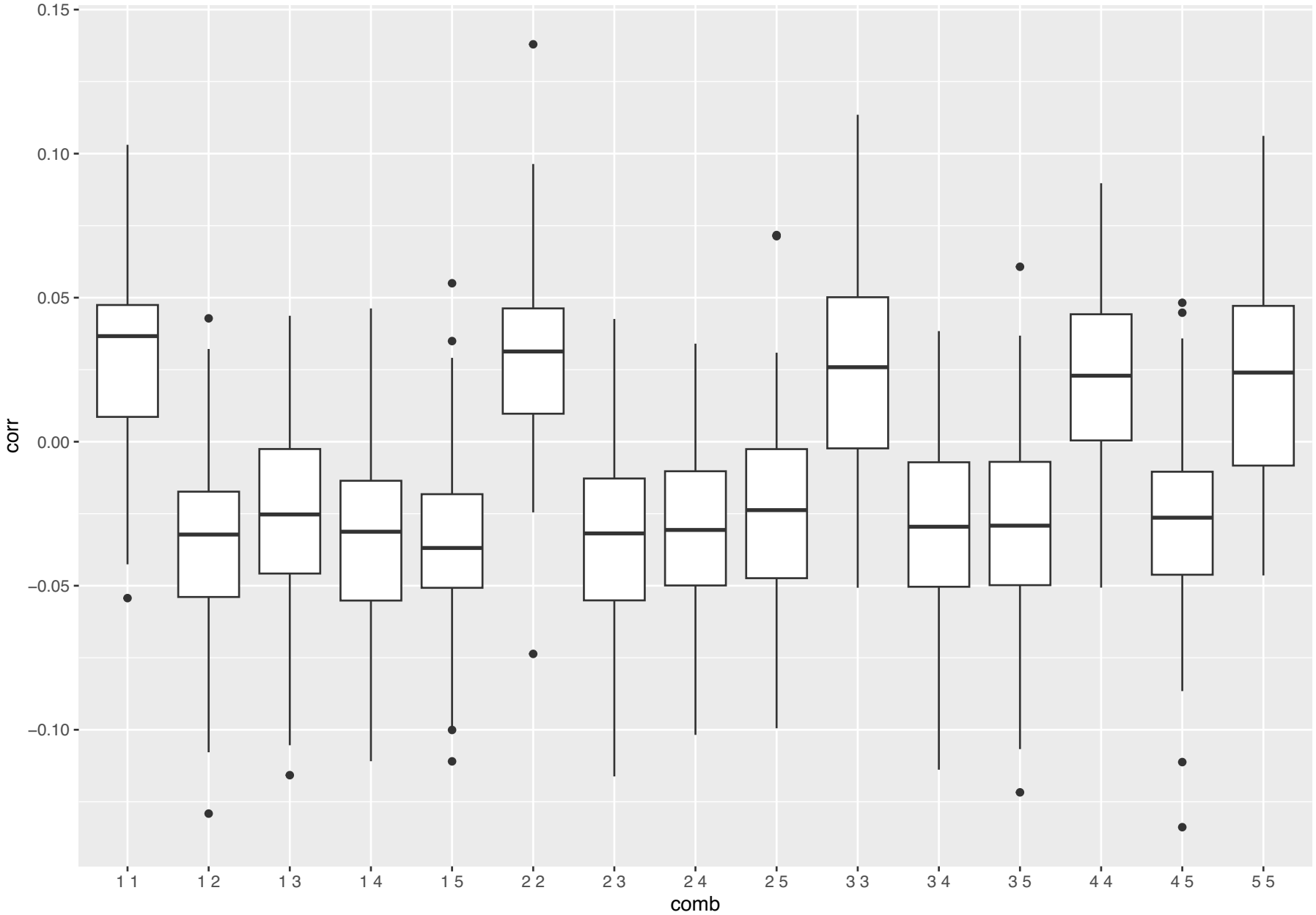



1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1

There are $50 * 49/2 = 1225$ unique pairs of observations (residuals) for the simulated example. There are 5 folds and the average correlation for the 5 times $10 * 9/2 = 45$ pairs $=$ 225 pairs within each fold is 0.0253342.

The average correlation for the 1000 pairs between folds is -0.0304969.

However - testing if the correlation is different from null for all possible pairs of observation of the residuals (with 50 observation we have $50 * 49/2$ pairs), only gave a significant result for 12 using FDR cut-off 0.05.


0.15
0.10
0.05
0.00
−0.05
−0.10
corr
1 1
1 2
1 3
1 4
1 5
2 2
2 3
2 4
2 5
3 3
3 4
3 5
4 4
4 5
5 5
comb

The boxplot of the correlation between residuals are taken between two folds, labelled on the horizonal axes.

Most articles state that this is a substantial problem, mainly because for constructing tests the variance of the test statistics is underestimated with positively correlated tests. However, other articles like Wong and Yang (2017) do not consider this a problem.

What can we present from the CV?

We have now focus on some loss function, like squared loss, binomial deviance, cross-entropy loss.

When we performed model selection with CV for the lasso we plotted some mean and standard error. How did we then calculate the standard error and the mean? Can we use this standard error to calculate a confidence interval?

We had N observations in the training set and choose K-fold CV:

$$\text{CV}(\hat{f}) = \frac{1}{N} \sum_{i=1}^{N} L(y_i, \hat{f}^{-k(i)}(x_i))$$

Assuming that $N = K \cdot N_K$ so that the number of observations in each fold N_j is the same and equal to N_K.

$$\mathsf{CV}(\hat{f}) = \frac{1}{K}\sum_{j=1}^{K}\frac{1}{N_K}\sum_{i\in k(i)} L(y_i, \hat{f}^{-k(i)}(x_i)) = \frac{1}{K}\sum_{j=1}^{K}\widehat{\mathsf{CV}}_j$$

What we plotted was the $\frac{1}{K}\sum_{j=1}^{K}\widehat{\mathsf{CV}}_j$ as the estimator for the evaluation criterion, and then ± 1 standard error of this mean.

The variance of the mean was estimated as

$$\mathsf{SE}^2(\hat{f}) = \frac{1}{K}\left(\frac{1}{K-1}\sum_{j=1}^{K}(\widehat{\mathsf{CV}}_j - \mathsf{CV}(\hat{f}))^2\right)$$

Since the residuals within a fold are positively correlated and between folds are negatively correlated, we only present plots of

$$\mathsf{CV}(\hat{f}) \pm \mathsf{SE}(\hat{f})$$

and are happy with that.

ROC-AUC on CV data

For the ROC-AUC two different strategies are possible:

- For each CV fold separately calculate the ROC-AUC, and then report average and standard error (as above) over the fold. This is called *average approach*.
- Use all predictions (across all folds) to calculate ROC_AUC. This is called *pooled approach*. Then results from the DeLongi method might not be completely correct due to the observations being positively correlated within folds and negatively correlated between folds.

Airola et al (2010) suggest an hybrid combination of the two methods.

None of these approaches provides CIs or hypothesis tests.

LeDell (2015, Section 5, page 53,55) develop CIs for cross-validated AUC.

The starting point is that the ROC-AUC theoretically can be interpreted as: "the probability that a randomly selected positive sample will rank higher than a randomly selected negative sample".

$$AUC(P_0, f) = P_0(f(X_1) > f(X_2)|Y_1 = 1, Y_2 = 0)$$

where (X_1, Y_1) and (X_2, Y_2) are samples from P_0.

The empirical AUC can be written

$$AUC(P_n, f) = \frac{1}{n_0 n_1} \sum_{i=1}^{n_0} \sum_{j=1}^{n_1} I(f(X_j) > f(X_i))$$

where n_0 is the number of observations with $Y = 0$ and n_1 with $Y = 1$.

To arrive at an estimator based on V-fold CV the empirical formula above is used for each fold and then the V-fold CV ROC-AUC is the average of this over the folds.

The influence function (a core idea of the phd of LeDell) is used to find the variance of the cross-validated ROC-AUC (taking into account the correlation between folds) and to establish a CI. This is implemented in the R-package `cvAUC`.

```
$cvAUC
[1] 0.8973285

$se
[1] 0.01651137

$ci
[1] 0.8649668 0.9296902

$confidence
[1] 0.95
```

TAKE HOME MESSAGE

Data rich + classification: Bleier
McNemar
ROC-AUC Delong

regression: Help!

k-fold CV & classif: fine: Model ROC-AUC

& regression: happy with $CV(\hat{f}) \pm SE(\hat{f})$?