Model Assessment and Selection

Florian Markowetz & Rainer Spang

Courses in Practical DNA Microarray Analysis

NGFN
Nationales Genomforschungsnetz

A short test on what you have learned so far…

1. What is overfitting and what is the overfitting disaster ?

2. What is the difference between prediction and separation ?

3. How does Regularization work?

4. How is Regularization implemented PAM and in SVM ?

Open Problems:

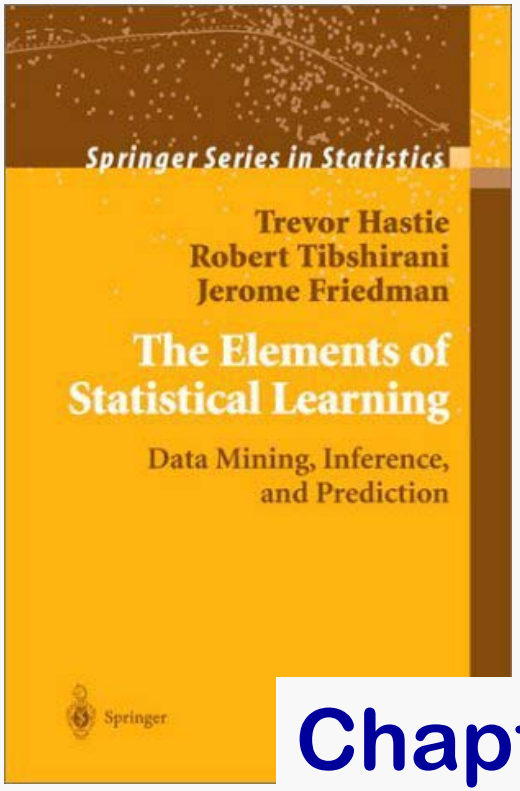
1. **How much regularization is good?**

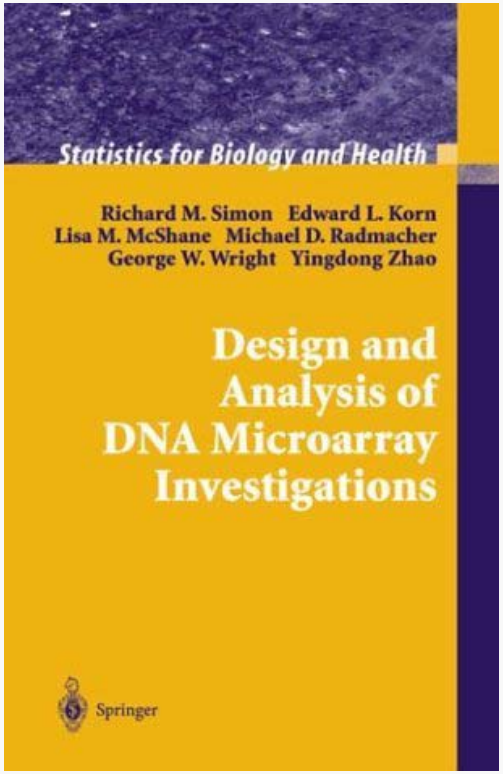
   - adaptive model selection -

2. **If I have found a signature, how do I know whether it is meaningful and predictive or not?**

   - validation -

Model Selection & Validation

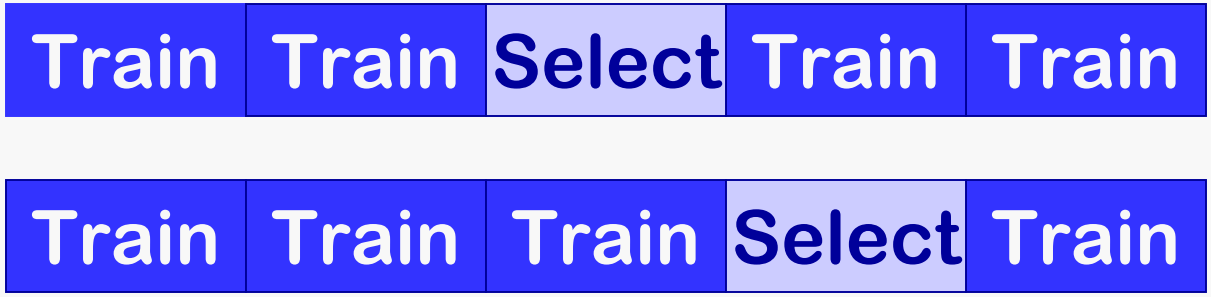




Chapter 7

We only discuss Cross-Validation

Cross-Validation



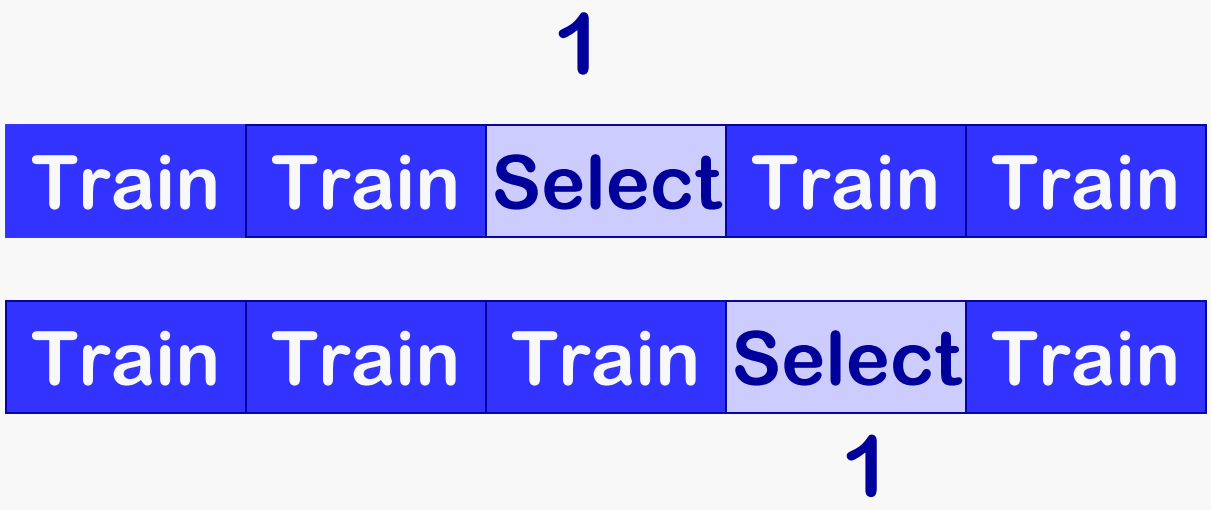
Chop up the training data (don‘t touch the test data) into 10 sets

Train on 9 of them and predict the other

Iterate, leave every set out once

- 10-Fold Cross Validation -

Leave one out Cross-Validation

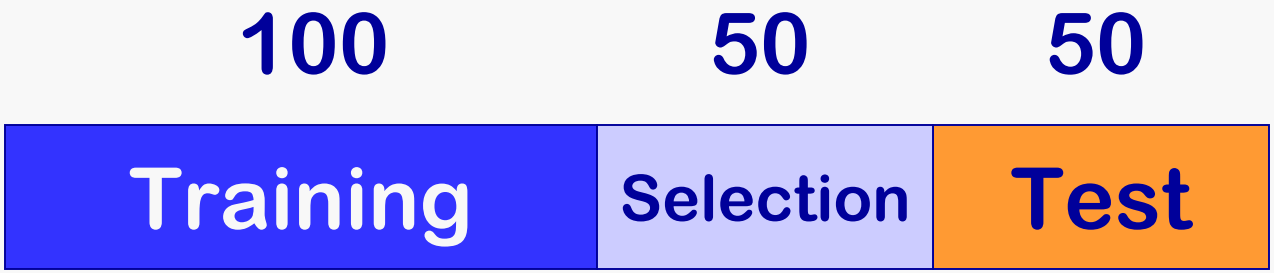


Essentially the same

But you only leave one sample out at a time and predict it using the others

Good for small training sets

Model Selection with separate data



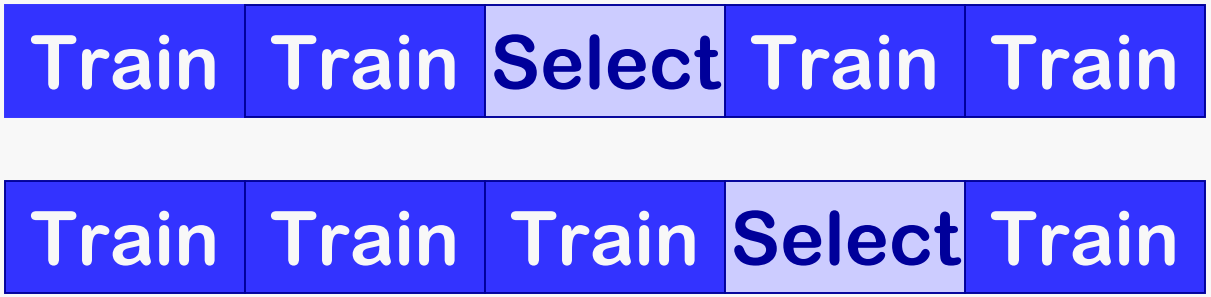
Split of some samples for Model Selection

Train the model on the training data with different choices for the regularization parameter

Apply it to the selection data and optimize this parameter ***- Adaptive Model Selection -***

Test how good you are doing on the test data
- Validation -

How much shrinkage is good in PAM ?



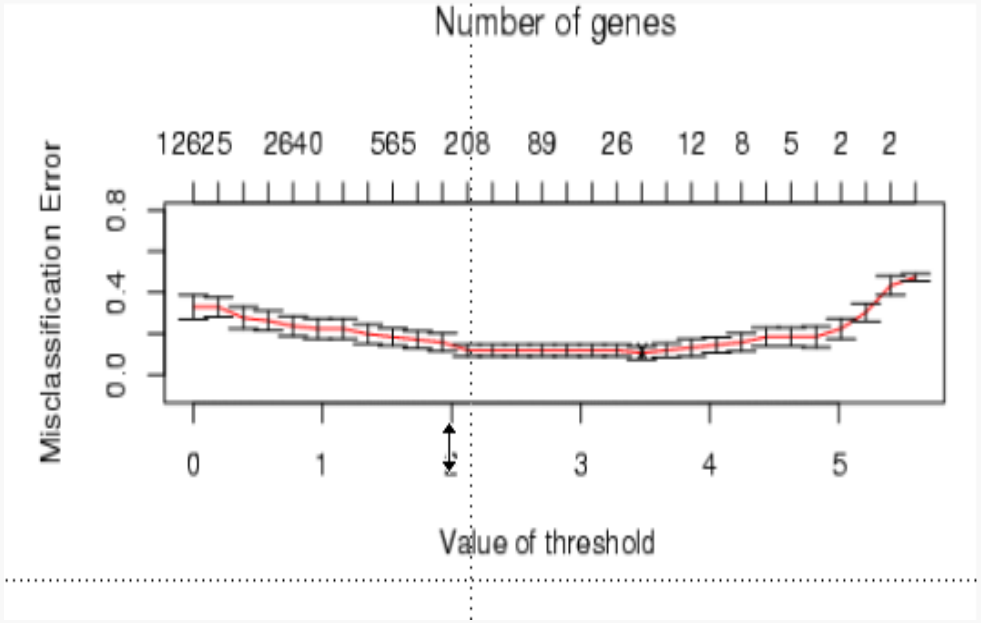
Compute the CV-Performance for several values of Δ

Pick the Δ that gives you the smallest number of CV-Misclassifications

Adaptive Model Selection

PAM does this routinely

Model Selection Output of PAM



Small Δ, many genes poor performance due to overfitting

High Δ, few genes, poor performance due to lack of information – *underfitting* -

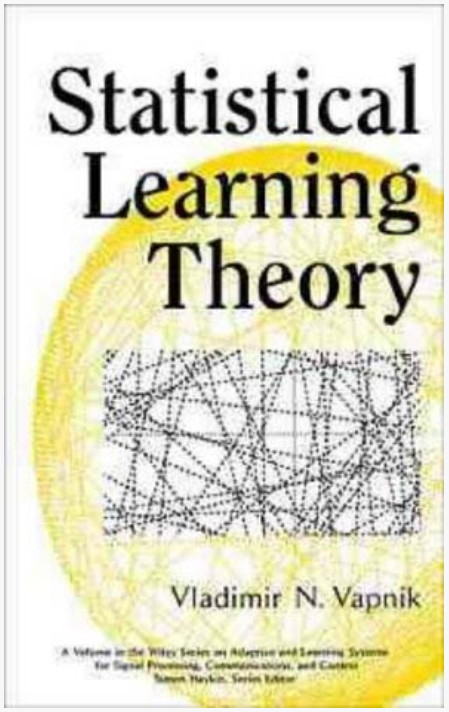
The optimal Δ is somewhere in the middle

Adaptive Model Selection of SVM

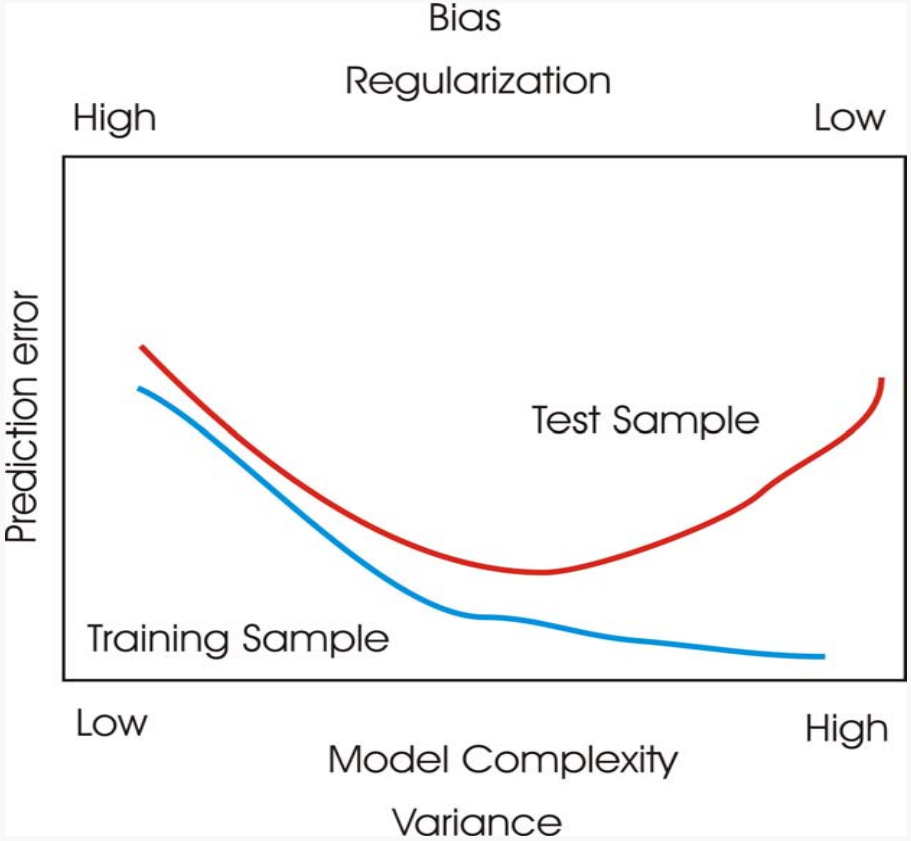
SVM optimize the margin of separation

There are theoretical results connecting the margin to an upper bound of the test error (V. Vapnik)

- structural risk minimization -



The overfitting underfitting trade off



Model Complexity:

-max number of genes

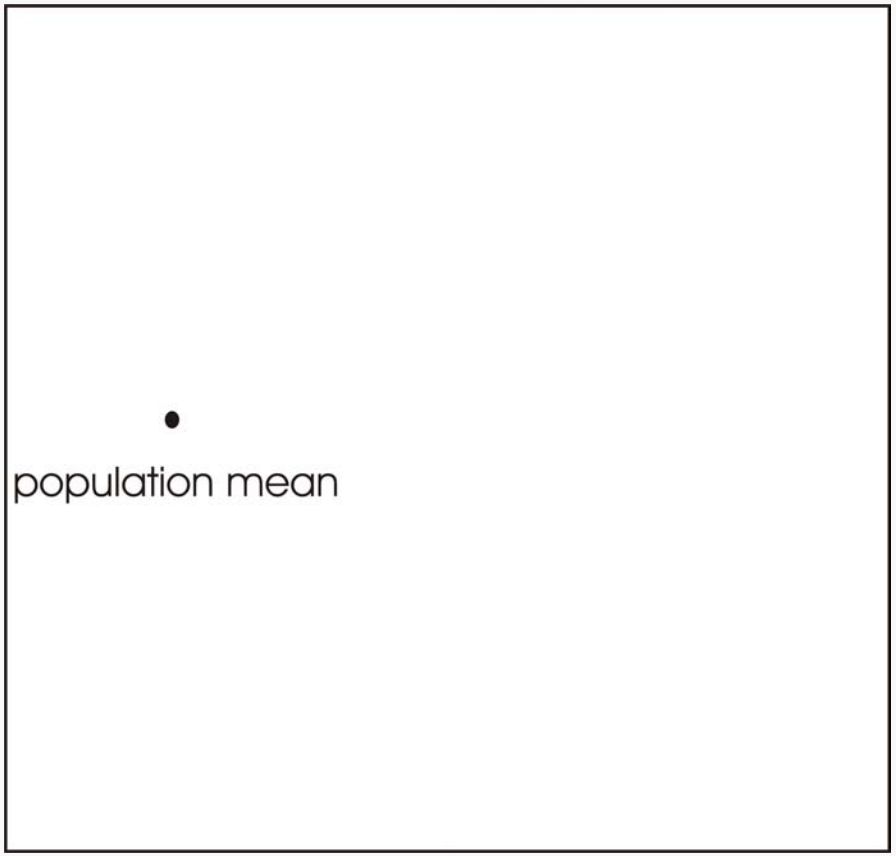
-shrinkage parameter

-minimal margin

-etc

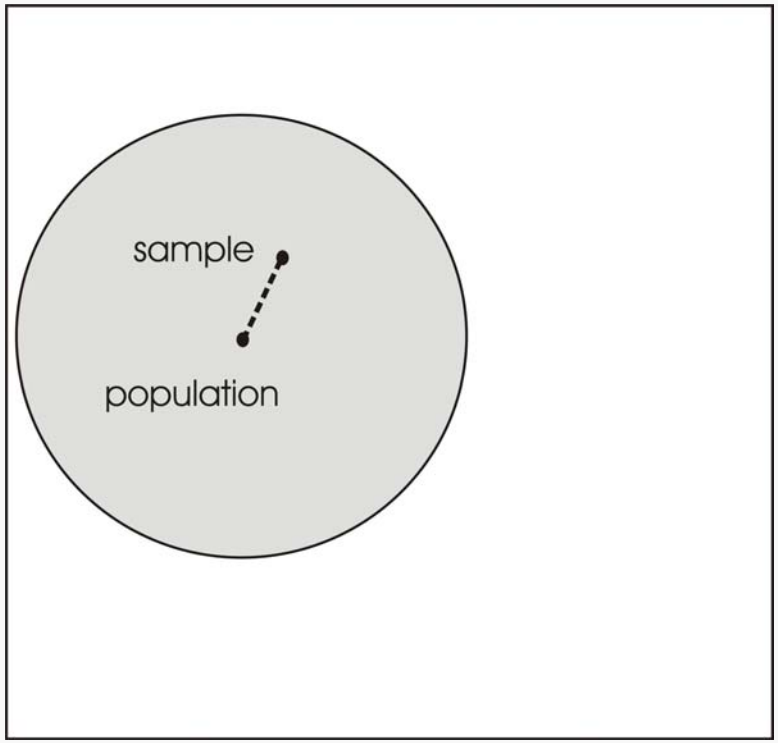
Population mean:

Genes have a certain mean expression and correlation in the population

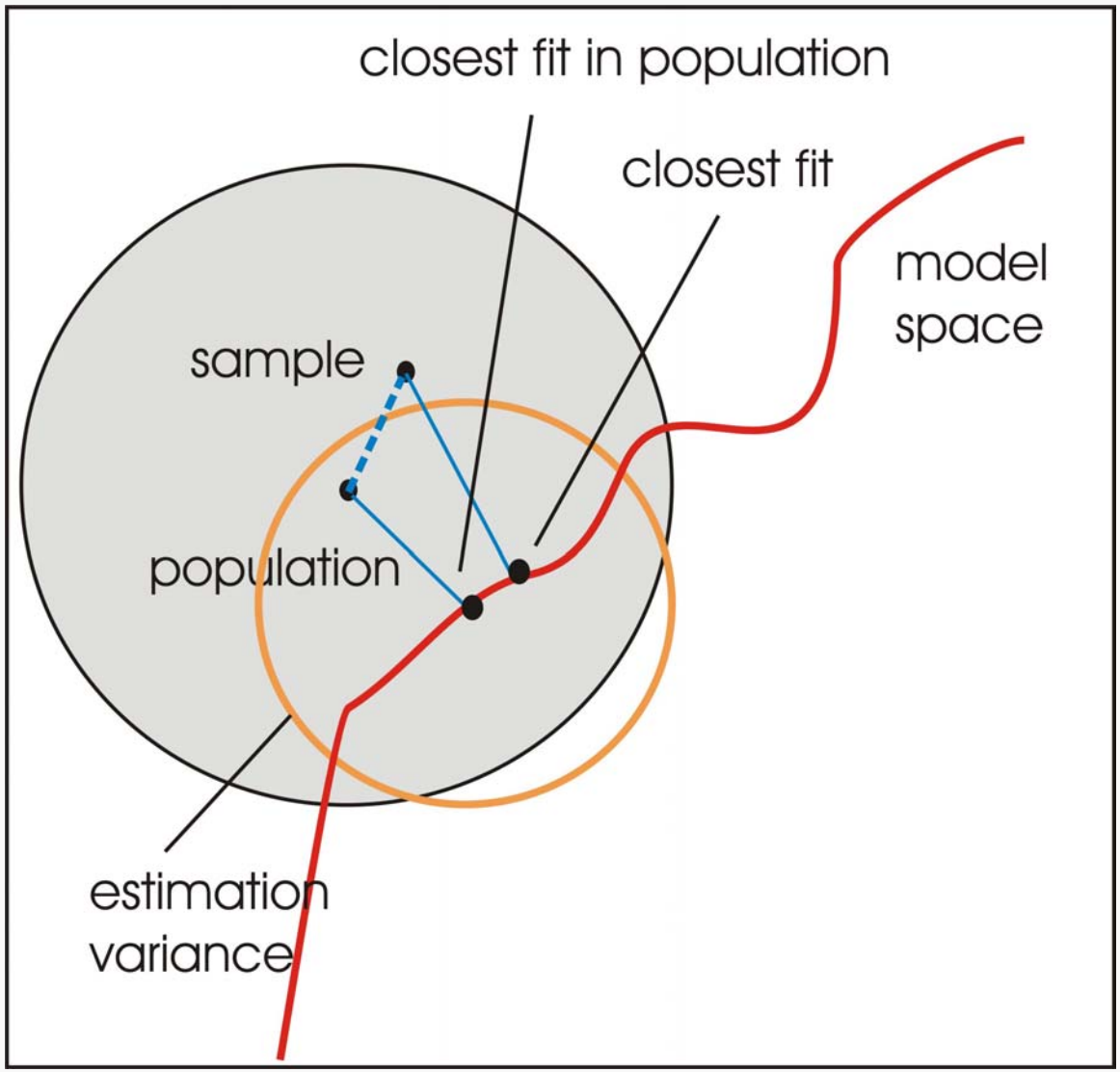


Sample mean:

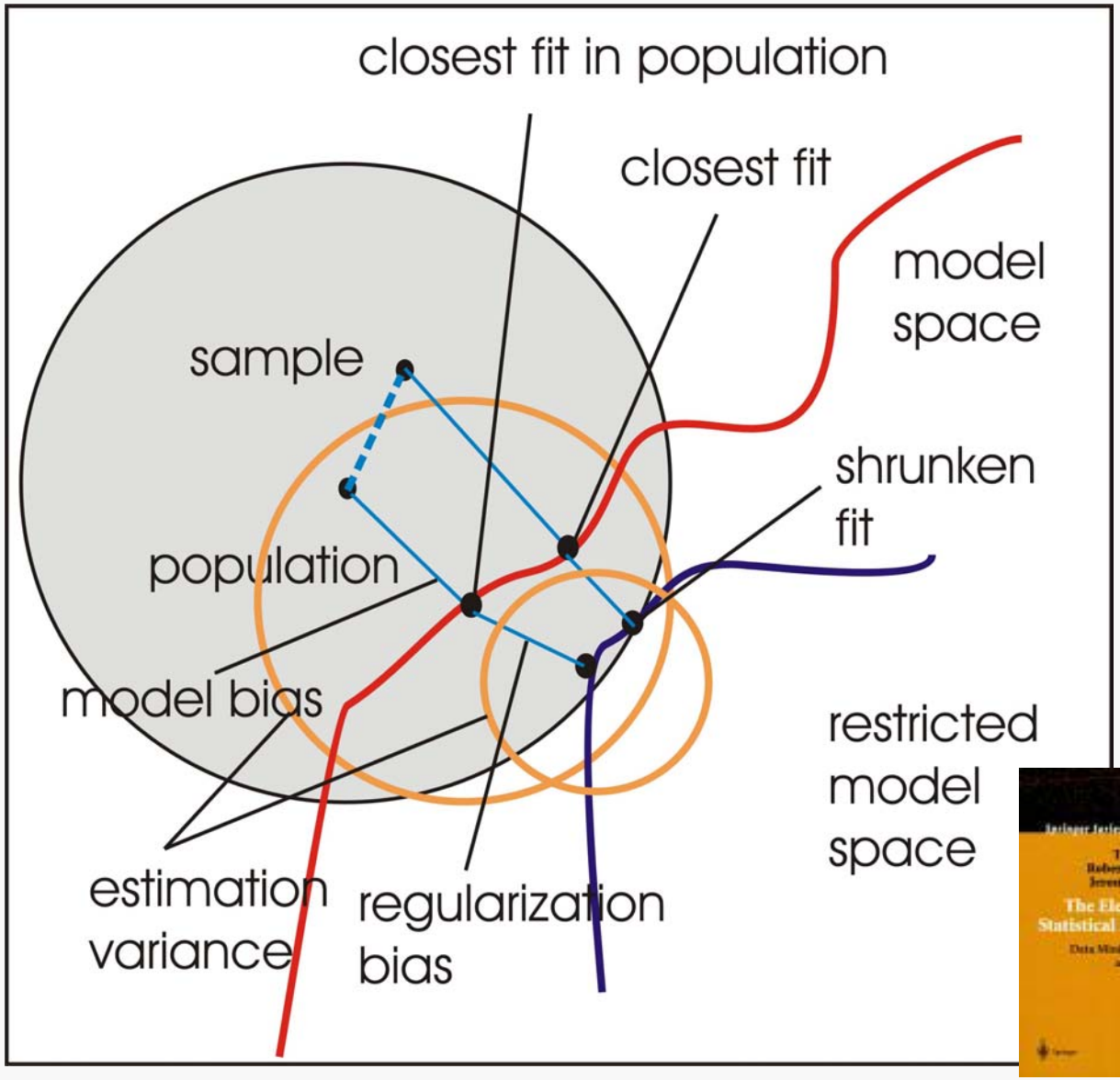
We observe average expression and empirical correlation

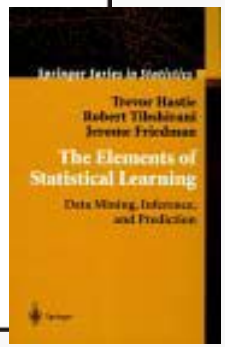


Fitted model:



Regularization





Validation

How well did I do?

Can I use my signature for clinical diagnosis?

How well will it perform?

How does it compare to traditional methods?

Validation

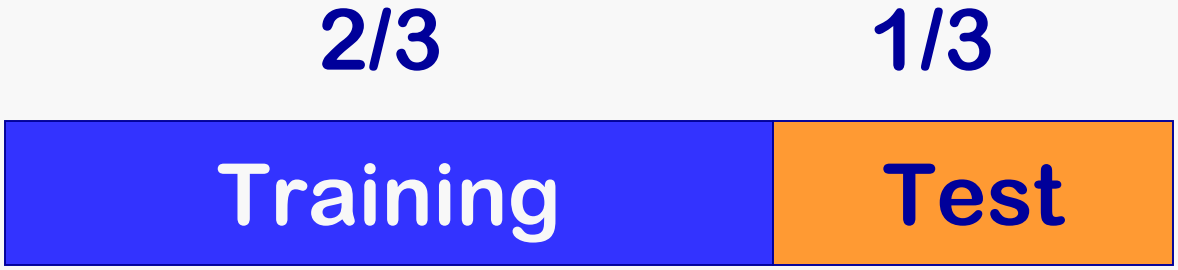
A Internal

1. Independent Test Set
2. Nested CV

B External (FDR)

1. Completely new prospective study

Test and Training Data



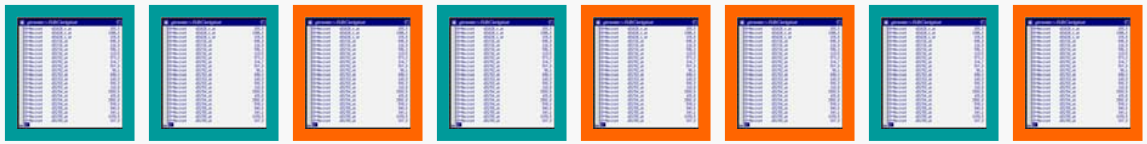
Split your profiles randomly into a training set and a test set

Train your model only using the data in the training set

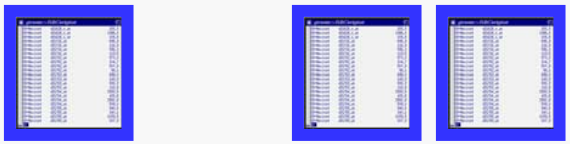
(define centroids, calculate normal vectors for large margin separators, ...)

Apply the model to the test data ...

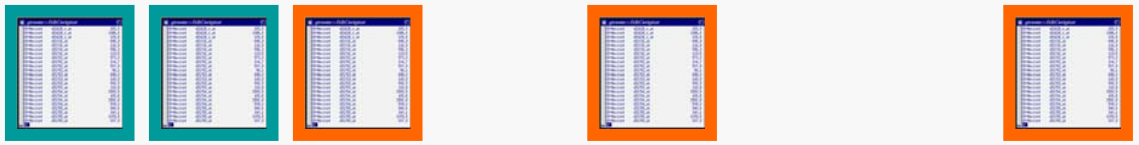
Recall the idea of a test set?



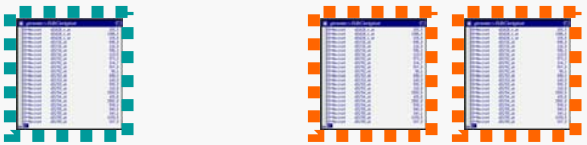
Take some patients from the original training samples and blind the outcome



These are now called test samples



Only the remaining samples are still training samples. Use them to learn how to predict



ok ok mistake

Predict the test samples and compare the predicted outcome to the true outcome

Validation Ideas

You want to validate the predictive performance of your signature

Validation is usually done using an independent test set

This mimics the prediction of new patients

Information of the outcome of the patients must not be used at any time before the final prediction

Scenario 1

1. You train a SVM on using all genes and all patients and you observe not a single misclassification

2. You conclude that your signature does not make any (or only very little) mistakes

What is wrong ?

The most important consequence of understanding the overfitting disaster:

If you find a separating signature, it does not mean (yet) that you have a top publication …

… in most cases it means nothing.

Scenario 2

1. You find the 500 genes with the highest average fold change between all type A patients and all type B patients

2. You split the patients into a test and a training set. Using only the training set you fit a SVM and applying it to both the test and trainings data, you observe 5% errors.

3. You conclude that your signature will be wrong in only 5% of all future cases

What is wrong ?

*Gene selection is part of training and **must not** be separated from it*

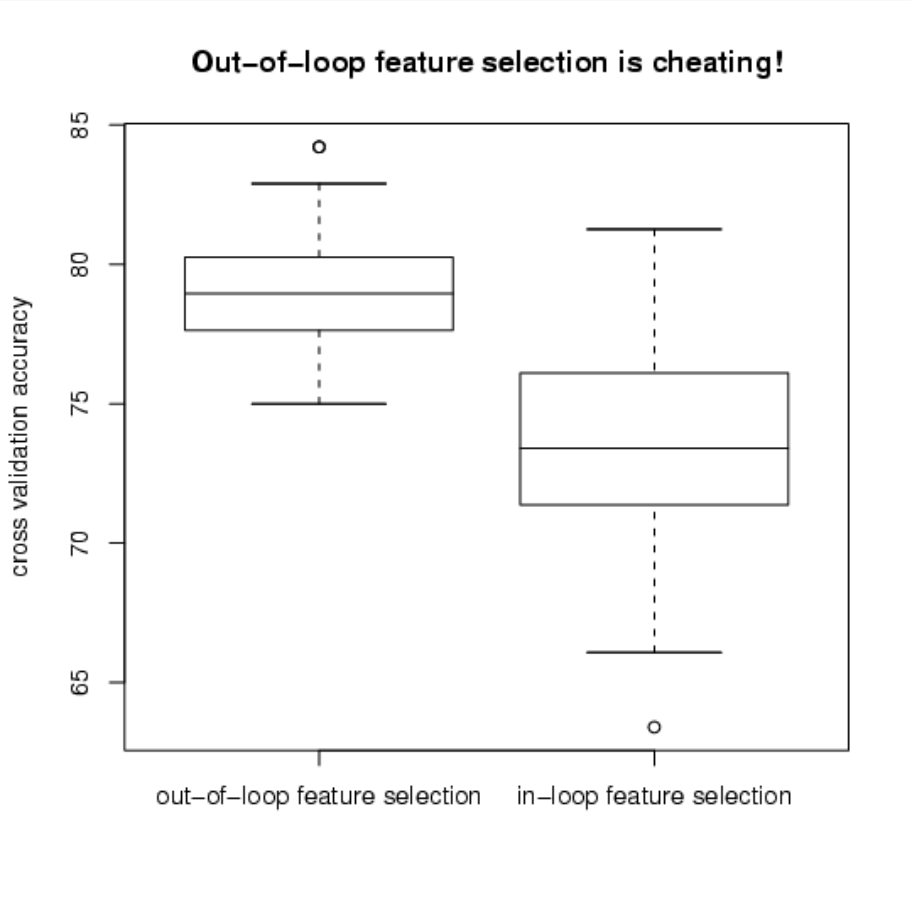
You can not select 20 genes using all your data and then with this 20 genes split test and training data and evaluate your method.

There is a difference between a model that restricts signatures to depend on only 20 genes and a data set that only contains 20 genes

Your validation result will look much better than it should

- selection bias -

The selection bias



Out-of-loop and in-loop gene selection

Scenario 3

1. You run PAM using adaptive model selection. CV Performance varies between 5% -10%
2. You choose the optimal Δ which yields 5% misclassifications
3. You conclude that your signature will be wrong in only 5% of all future cases

What is wrong ?

*Adaptive model selection is part of the training **and not** part of the validation*

Choosing the optimal Δ always means choosing the optimal Δ for your training data

The performance on new patients is in general a little worse

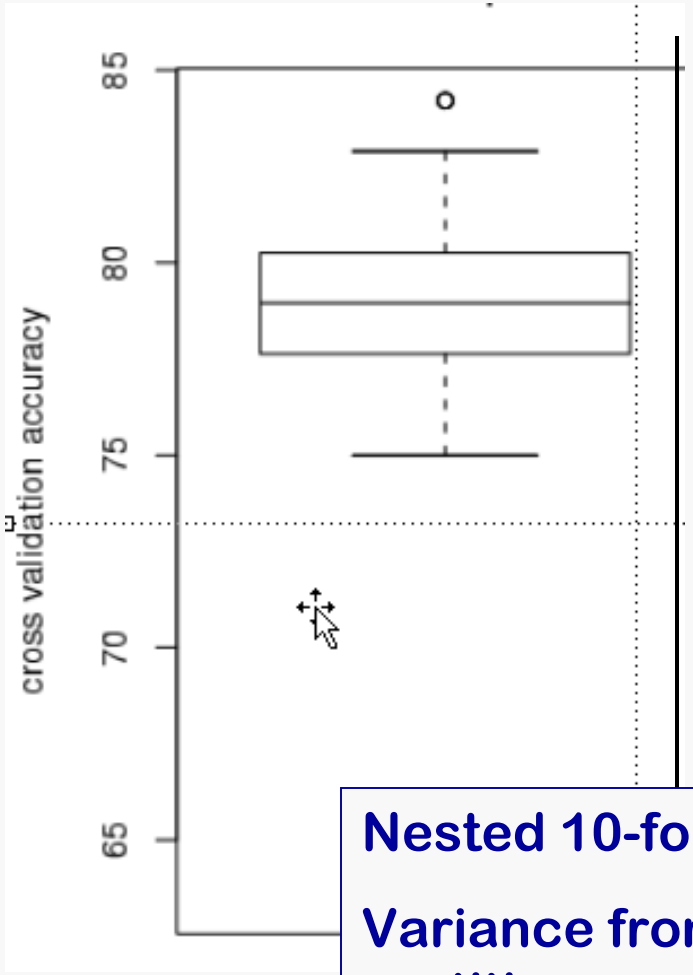
You can see this using test data

Scenario 4

1. You split your data in test and training data
2. Using only the training data you rum PAM including adaptive model selection. The optimal CV-Error is achieved for Δ=3
3. You apply the Δ=3 signature to the test data and observe an error of 7%
4. You conclude that your signature will be wrong in not more than 7% of all future cases

What is wrong ?

What you get is an estimation of performance ...



... and estimators have variance.

If the test set is mall this variance can be big.

Nested 10-fold- CV

Variance from 100 random partitions.

DOs AND DONTs :

1. Decide on your diagnosis model (PAM,SVM,etc...) and don‘t change your mind later on

2. Split your profiles randomly into a training set and a test set

3. Put the data in the test set away ... far away

4. Train your model only using the data in the training set

(select genes, define centroids, calculate normal vectors for large margin separators, perform adaptive model selection ...)

don‘t even think of touching the test data at this time

5. Apply the model to the test data ...

don‘t even think of changing the model at this time

6. Do steps 1-5 only once and accept the result ...

don‘t even think of optimizing this procedure

Thank you