Cross Validation (CV)

- After circulation, the selected number of parameters is that, for which the average modeling error is smallest.
- For the optimal model, the parameter values are calculated from the whole data set.

Cross Validation (CV)

- Selection is based on average goodness of ML estimates of the groups.
- Good estimate $\rightarrow$ small modeling error.
- Groups are used to reflect 'a part of the problem'. On the other words, part of the data is removed to avoid over fitting
- Two-part criteria: all the data is used, parameter number is penalized, CV: part of the data is used, parameter number is not penalized.

Cross Validation (CV)

- Popular method for the selection of the number of the parameters
- In this method the available data is divided to groups. One group is used for testing, others for training. Modeling error is calculated for the testing set
- This is repeated until each group has been a testing set. Average modeling error is calculated over the errors of the testing sets.
- Circulation is repeated over the parameter range.

Cross Validation (CV)

- Data is divided to $N$ groups, where $N$ is between 2 - 10.
- The groups do not have to have equal size
- An exception is Leave-One-Out (LOO) -method, where only one sample is used for testing.
- Experiences from using LOO have shown that it should used only when the data set is small. Otherwise it favors large parameter number
- In theory LOO approaches AIC, when we have infinite data set.
Cross Validation (CV)

- It should be noticed that there is no method for selecting the number of groups.
- The allocation of the samples to groups affects the performance of the CV procedure, also does the number of groups. Therefore CV method is not purely data driven method.
- There are various implementations of CV. Usually CV varied according the properties of the modeling problem at hand (how to divide samples to groups, model goodness criteria).

Cross Validation (CV)

- CV procedure: Divide the data for $l$ groups, and define the parameter space $K$
- for $k = 1:K$
  1. for $i = 1:l$
  2. Assign group $i$ for testing
  3. With the training groups solve parameters $\hat{a}$
  4. Calculate modeling error ($E_i = y - X\hat{a}$) using the test set.
- Calculate average error $E_k = 1/l\sum^l_i E_i$

Cross Validation (CV)

- Dividing the samples to groups must be done carefully.
- Usual approaches are random selection without substitution or a time interval.
- Whatever method is used, the internal dependences of the samples must not be forgotten

Cross Validation (CV)

- Example: Use CV procedure to solve the optimal number of parameters for a linear regression problem.
- We have set of $n$ observations $y_i$, and model class
  $y_i = a_0 + \sum^m a_x x_i + \epsilon$. We assume that error is gaussian, eg. $\epsilon \sim N(0,\sigma^2)$. In matrix form this is $y = Xa + \epsilon$ ($X$ is $n \times m + 1$, augmented with column of ones).
- Optimal parameters $\hat{a}$ can be solved with pseudoinverse solution
  $$\hat{a} = (X^T X)^{-1} X^T y$$
**Bootstrap**

- In this method the generalization capability is studied through different combinations of variations.
- So called bootstrap sample is formed by sampling the observations. The sampling is done with equal probability for each sample and without substitution.
- The size of sample is constant.

**Bootstrapping**

- For a model class $M$ with $k$ parameters $\theta$, the model goodness can be estimated with $N$ bootstrap replicates $\theta[B]_k$

\[
S(K) = \frac{1}{N} \sum_{b=1}^{N} p(x|\theta[B]_k)
\]

- Remember ML, choose parameters which make the data set most probable.
- The best number of parameters $k^*$ is that which maximizes $S(k)$.

**Cross Validation (CV)**

- Select $k$ for which $E_k$ is smallest.
- Calculate the optimal model $\hat{y} = f(x|\hat{\theta})$ from the whole data set.

**Bootstrap**

- Several (100, 1000, 10000) bootstrap samples are generated.
- From the bootstrap samples are solved optimal (ML) parameters, modeling error and average modeling error.
- This is repeated over parameter space and that number of parameters is chosen, for which the average modeling error is smallest (ML estimate is greatest).
- Finally the optimal parameter values for the optimal number of parameters are solved using whole data set.
Bootstrap

- There are various implementations of bootstrap methods.
- Also in bootstrap method, the number and size of the bootstrap samples has to be decided before the implementation, making bootstrap also non data driven method.