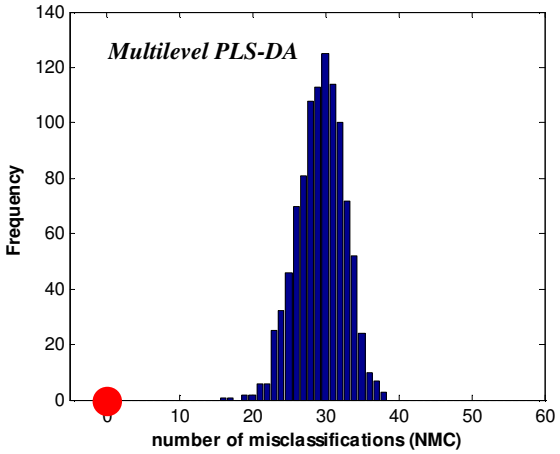
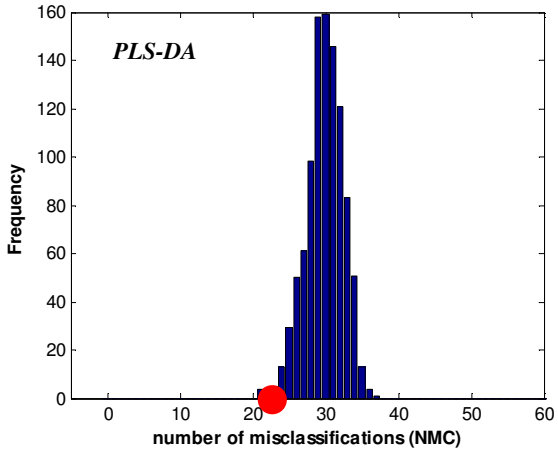
Multilevel data analysis

TUTORIAL





Installation

- Copy the Toolbox on your hard disk.
- Unzip the Toolbox in a directory, e.g. c:\matlab\work\CMV\Toolbox
- Start Matlab
- Add directory to search path by using the *addpath* command.

  Example

  addpath c:\matlab\work\CMV\Toolbox;

Preparation

- Load data sets

  Example

  load NMR_testset
  load class_labels
  load chemical_shifts
  load designmatrix

Simulated test dataset

The crossover designed NMR_testset contains simulated ^{1}H NMR urine spectra of 30 individuals (15 males and 15 females). These spectra were "acquired" after a long-term intervention of placebo (PLACEBO) and an antioxidant supplement (TREATMENT). Each individual in the dataset is therefore represented by 2 urine spectra.

The structure of the dataset is shown in **Figure 1**. The PLACEBO and TREATMENT spectra are grouped in two data blocks. The order of the individuals is exactly the same in both data blocks. The class labels that are associated with both interventions are -1 (PLACEBO) and +1 (TREATMENT).

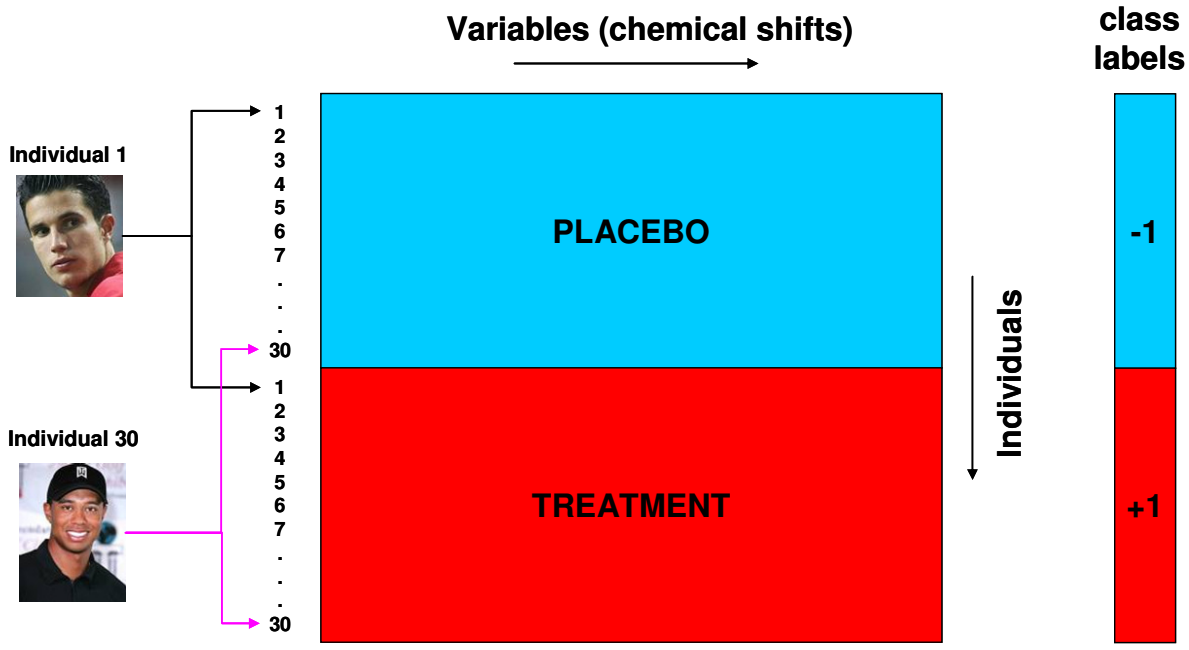


Figure 1 : Structure of dataset and the class labels.

- Plot the entire dataset. The urine profiles of the males are coloured red, whereas the urinary profiles of the females are coloured blue (**Figure 2**). Notice that the resonances in the region δ 2.0-2.2 ppm are elevated in the female-group as compared to the males. In the region δ 7.5-7.9 ppm no obvious differences can be observed.

  Example

  plotfigure1(chemical_shifts,NMR_testset);

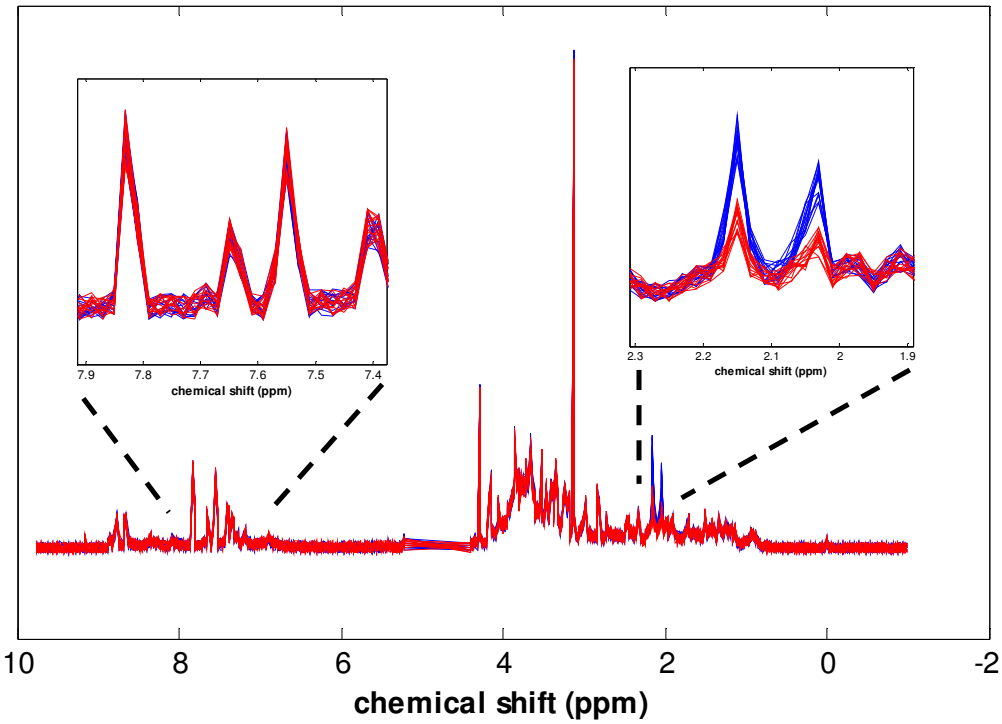


Figure 2: ^{1}H NMR urinary spectra in the NMR testset (urinary profiles). The spectra of the males are coloured red, and the spectra of the females are coloured blue. The resonance patterns in the region δ 2.0-2.2 ppm and δ 7.5-7.9 ppm are highlighted.

In the NMR testset two types of simulated variations are present. The first is related to gender differences (between-subject variation). The second is related to the antioxidant treatment (within-subject variation).

- Plot the NMR peaks of subject 1 (female) and subject 2 (male) to illustrate the gender-related differences between the subjects (**Figure 3a**).

  Example

  plotfigure2(chemical_shifts,between_testset);

- Plot the NMR peaks of subject 1 after placebo and antioxidant consumption to illustrate the treatment-related differences within the subjects (**Figure 3b)**.

  Example

  plotfigure3(chemical_shifts,within_testset);

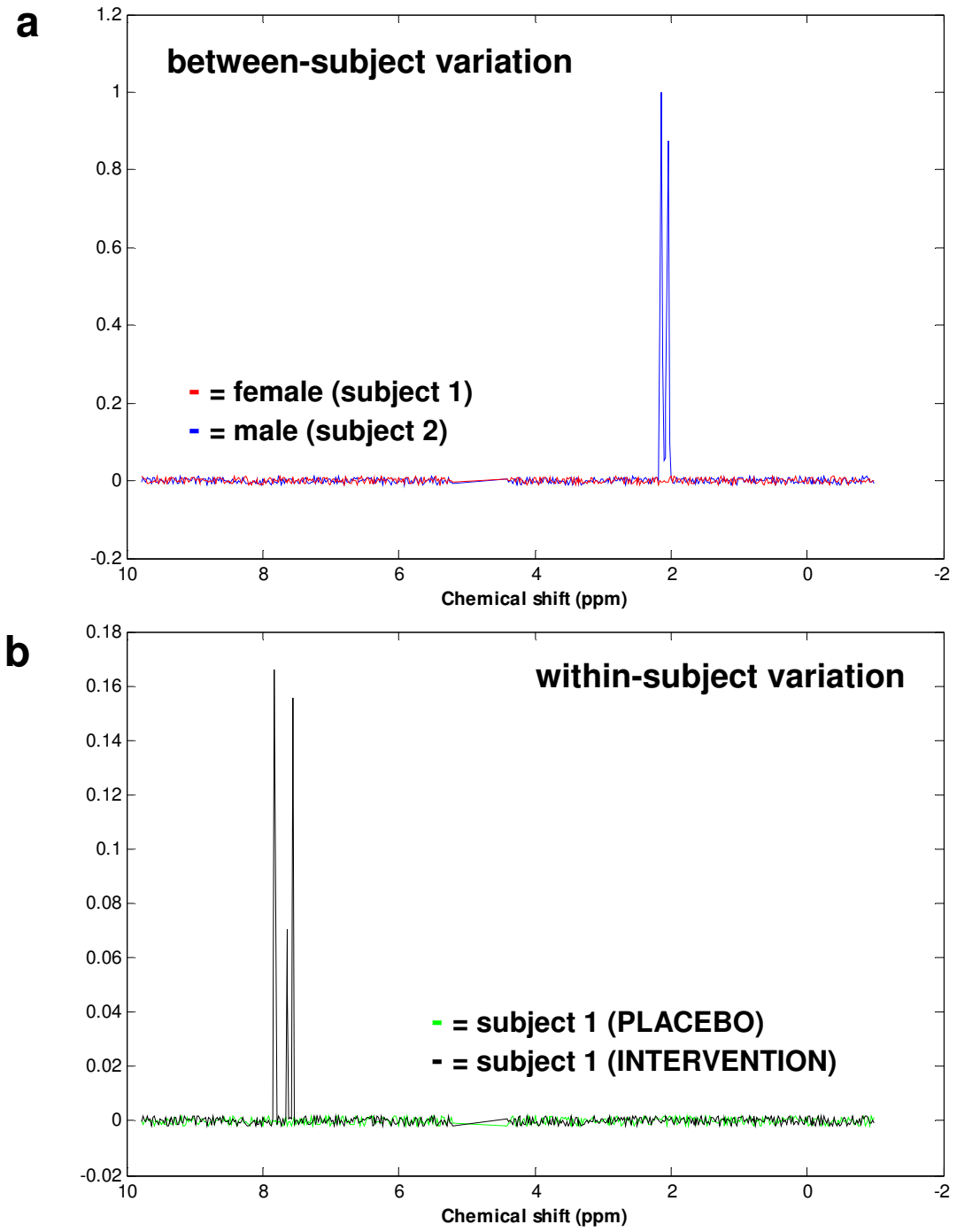
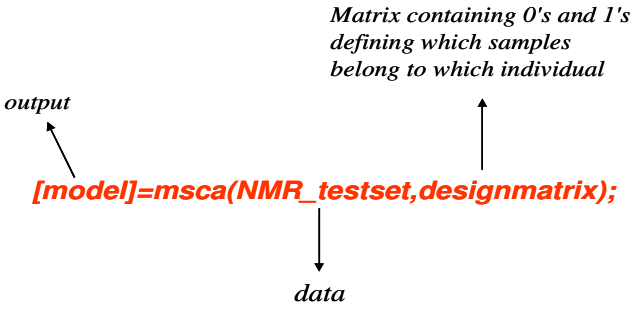


Figure 3: ^{1}H NMR signals that represent the simulated differences in the testset (a) between-subjects (δ 2.0-2.2 ppm) and (b) within-subjects (δ 7.5-7.9 ppm).

Variation splitting

- Split the between-subject variation and the within-subject variation in the NMR testset and perform PCA on both sets separately.
  Example



- Visualize the variation in the between-subject model by means of the PC1-PC2 scores (**Figure 4a**) and the PC1 loadings (**Figure 4b**). Notice that **Figure 4b** corresponds with **Figure 3a**.
  Example

```
plotfigure4(model.between.scores(:,1),model.between.scores(:,2));
plotfigure5(chemical_shifts,model.between.loadings(:,1));
```

- Visualize the variation in the within-subject model by means of the PC1-PC2 scores (**Figure 4c**) and the PC1 loadings (**Figure 4d**). Notice that **Figure 4d** corresponds with **Figure 3b**.
  Example

```
plotfigure6(model.within.scores(:,1),model.within.scores(:,2));
plotfigure5(chemical_shifts,model.within.loadings(:,1));
```

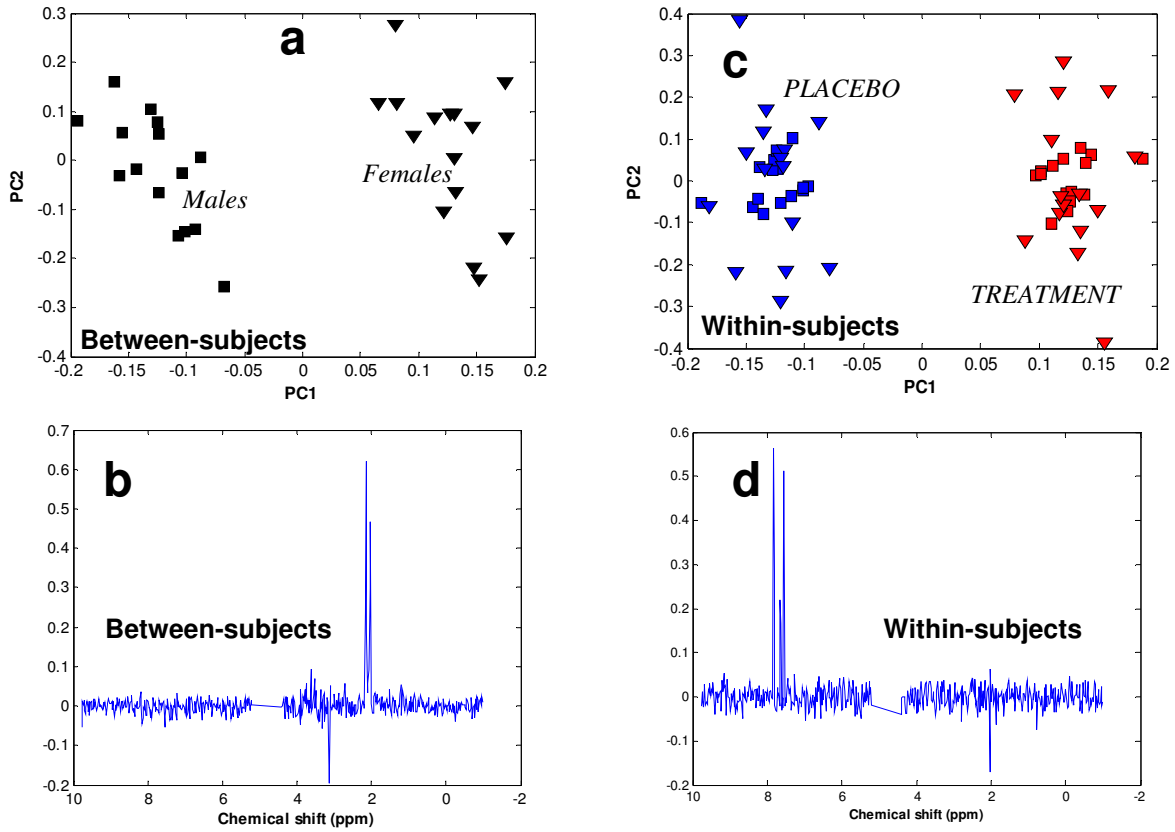
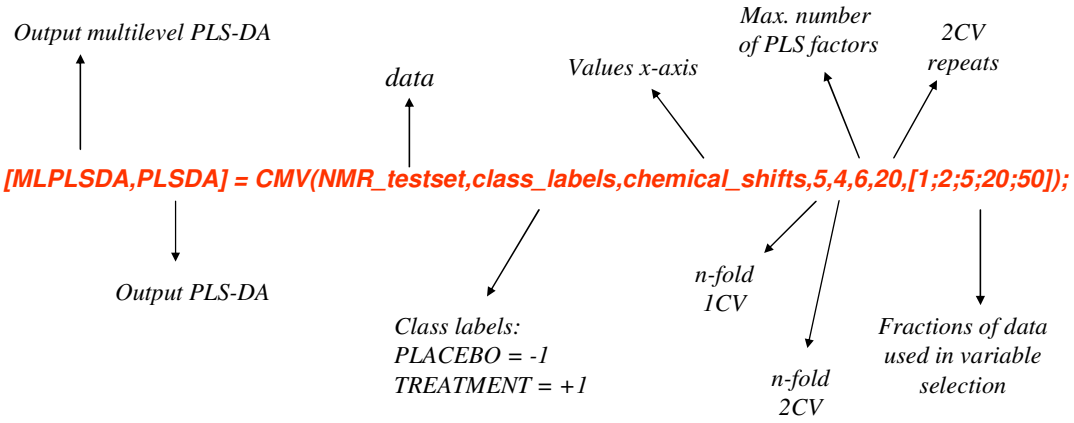


Figure 4: PC1-PC2 score plot of (a) the between-subject model and (c) the within-subject model. The PC1 loadings in the (b) between-subject model and (d) the within-subject model reveal the main sources of variation (gender, antioxidant treatment).

Multilevel PLS-DA and cross-model validation (CMV)

- Determine the treatment effect in NMR testset using multilevel PLS-DA and (ordinary) PLS-DA.
  Example



- Perform a permutation test (1000 CMV repeats) for both the multilevel PLS-DA model as well as for the (ordinary) PLS-DA model.
  Example

```
[MLPLSDAp,PLSDAp] = CMV_perm(NMR_testset,class_labels,chemical_shifts,5,4,6,20,1000,[1;2;5;20;50]);
```

CMV repeats

MLPLSDA, PLSDA, MLPLSDAp, PLSDAp are structured output-variables and contain all y-class predictions, Q2 values, AUROC values (area under the ROC curve) and NMC values (number of misclassifications) obtained in the performed CV1 and the CV2 rounds.

- The most discriminating variables between the PLACEBO and the TREATMENT group after multilevel PLS-DA are reflected in the Rank Product (RP). In **Figure 5** the lowest RP's represents the NMR signals that are strongly associated with the treatment effect.
  Example

```
plotfigure7(chemical_shifts,MLPLSDA.CV2.RP(1,:));
```

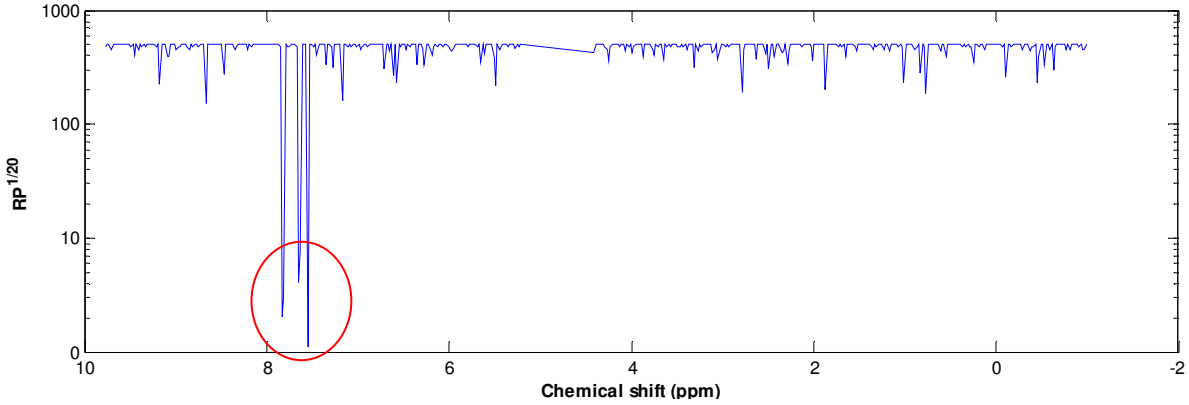


Figure 5: Rank product ($RP^{1/20}$) obtained after multilevel PLS-DA. The NMR signals with the lowest RP's contribute most to the observed treatment effect (within the subjects).

Permutation testing

- Compare the prediction error (NMC) of the multilevel PLS-DA model against the permutations (**Figure 6**).

  Example

  [frequency,nmc]=hist(mean(MLPLSDAp.CV2.NMC),[0:60]);
  plotfigure8(nmc,frequency,mean(MLPLSDA.CV2.NMC));

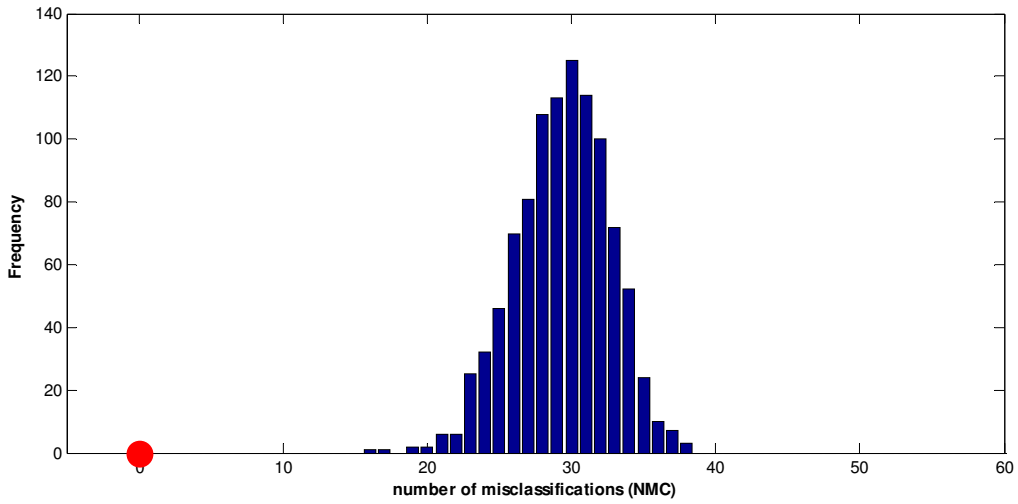


Figure 6: Comparison of the classification error (estimated in number of misclassifications) of the multilevel PLS-DA model (red) against the permutations (blue).

Since the prediction error (NMC=0) is much smaller as compared to the permutations, the statistically significance of this multilevel PLS-DA model is confirmed ($p<0.05$). Similar to the previous example, also the Q2 values and the AUROC values can be used as the prediction error.

- As shown in **Figure 7**, the classification error of the PLSDA model (NMC=25) is substantially larger.

  Example

  [frequency,nmc]=hist(mean(PLSDAp.CV2.NMC),[0:60]);
  plotfigure8(nmc,frequency,mean (PLSDA.CV2.NMC));

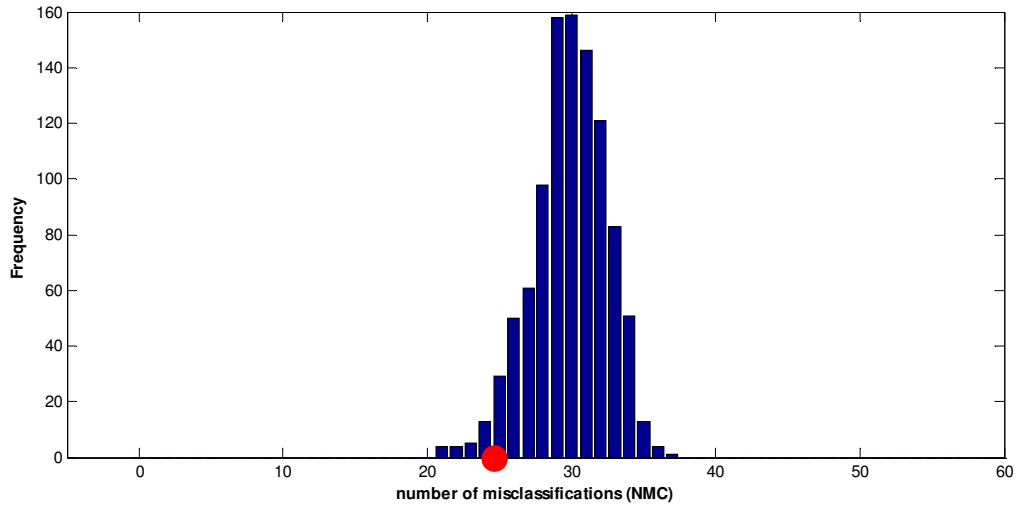


Figure 7: Comparison of the classification error (estimated in number of misclassifications) of the ordinary PLS-DA model (red) against the permutations (blue).

Using own data

When analyzing your own data set, the following considerations should be taken into account:

1- The new data should derive from a crossover designed experiment with a similar data structure as the NMR testset (**Figure 1**).
2- The designmatrix in the variation splitting procedure (msca) should be adjusted in such a way that the number of rows is equal to 2**I* (*I* = number of individuals) and the numbers of columns is equal to *I*. The structure of this new matrix should be similar to the designmatrix in the toolbox.
3- The number of CMV repeats (permutations), the number of 2CV repeats, number of PLS components, the "n-fold" of the 1CV and 2CV, and the number of PLS can be changed according the requirements of the user.