Discrimination of olive oils origin based on FTIR Spectroscopy data

Evgeniia Martynova (s1038931)

Lili Mészáros (s1015790)

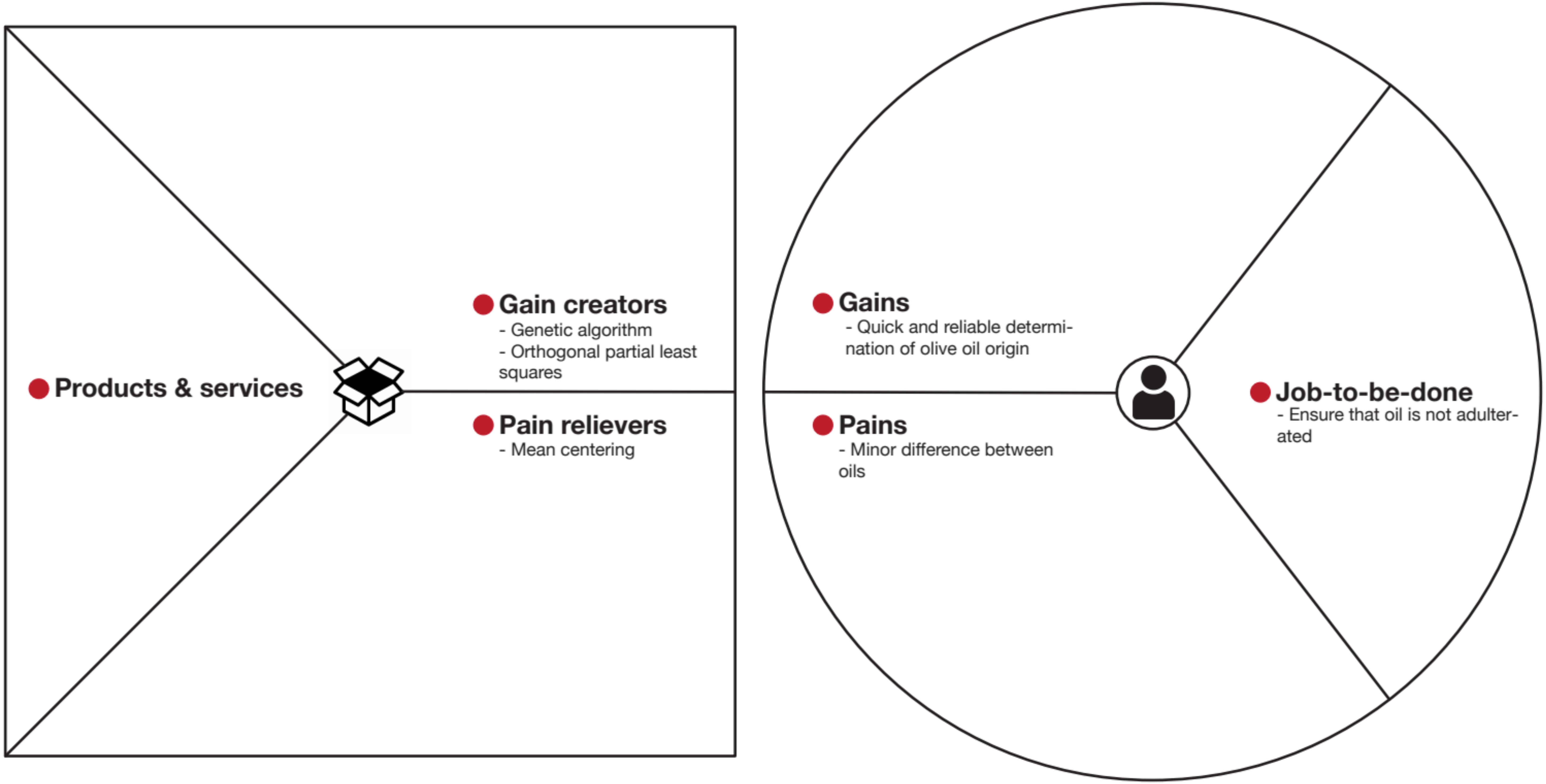
Radboud University

Data. Spectrum is obtained by Fourier transform infrared spectroscopy

- FTIR spectroscopy is fast and no complex samples pre-processing is needed
- 2 measurements of each sample made within 1-24 days interval
- 60 samples of olive oils from 4 countries obtained for the original study [1]

Group designation	Country of origin	No. of samples
1	Greece	10
2	Italy	17
3	Portugal	8
4	Spain	25
	total:	60

[1] Henri S. Tapp, Marianne Defernez, E. Katherine Kemsley, FTIR Spectroscopy and Multivariate Analysis Can Distinguish
the Geographic Origin of Extra Virgin Olive Oils
J. Agric. Food Chem. 2003



● **Products & services**

● **Gain creators**

- Genetic algorithm
- Orthogonal partial least squares

● **Pain relievers**

- Mean centering

● **Gains**

- Quick and reliable determination of olive oil origin

● **Pains**

- Minor difference between oils

● **Job-to-be-done**

- Ensure that oil is not adulterated

Experiment design and motivation

Original study:

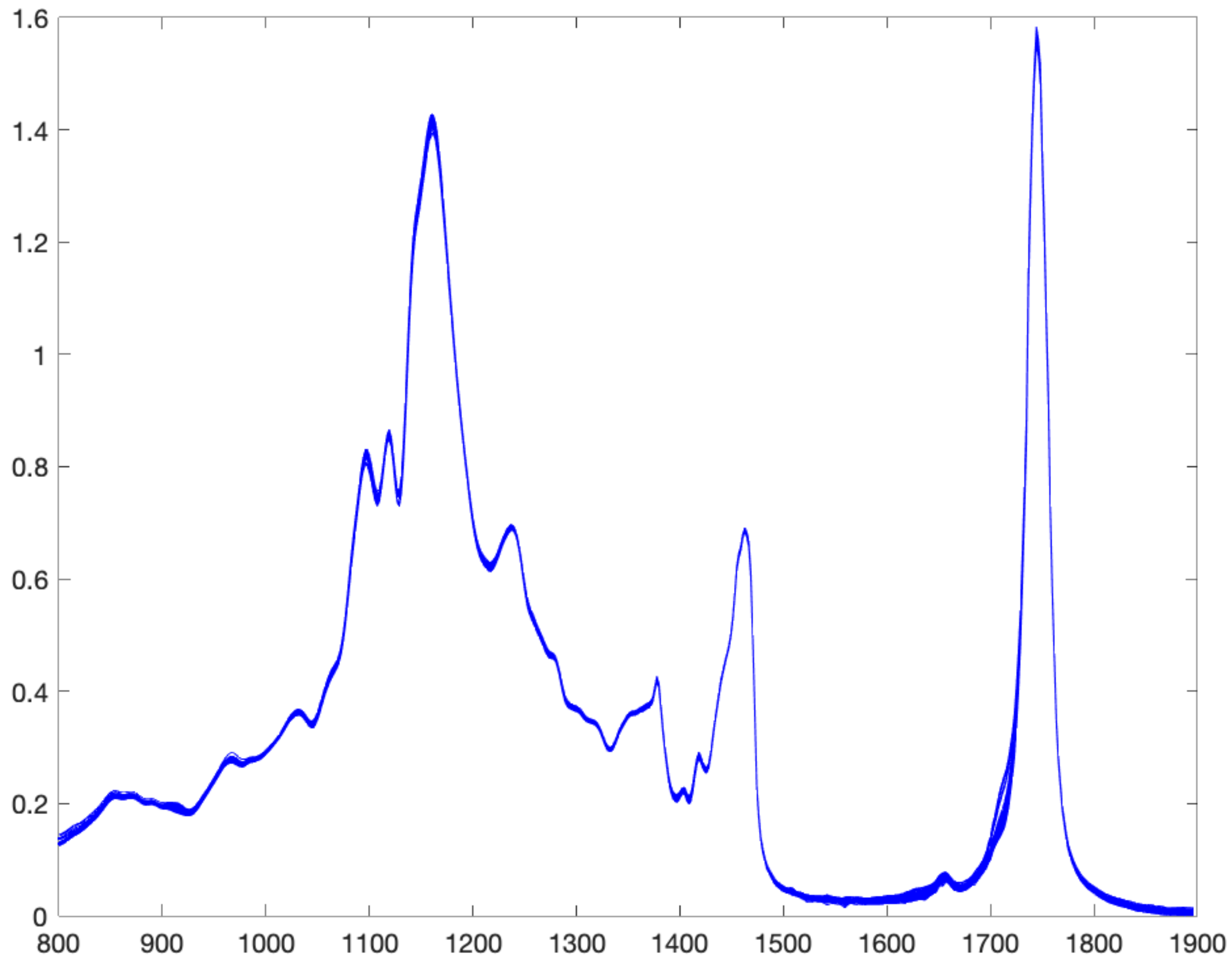
- Used internal cross-validation
- PLS
- Variable selection with genetic algorithm

Our study:

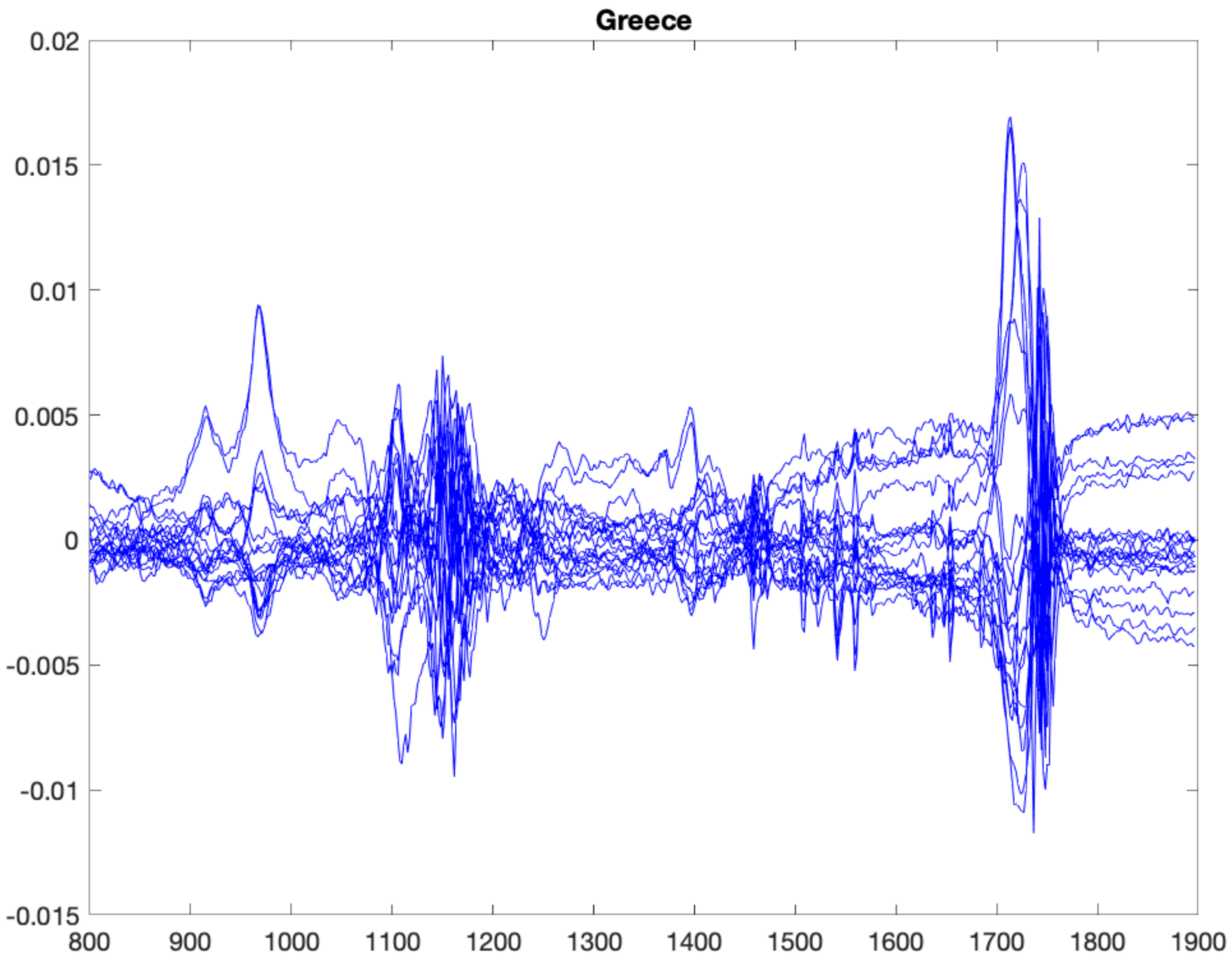
- Used Double cross-validation (LOO as for internal cross-validation)
- OPLS
- Further exploration of solutions with GA

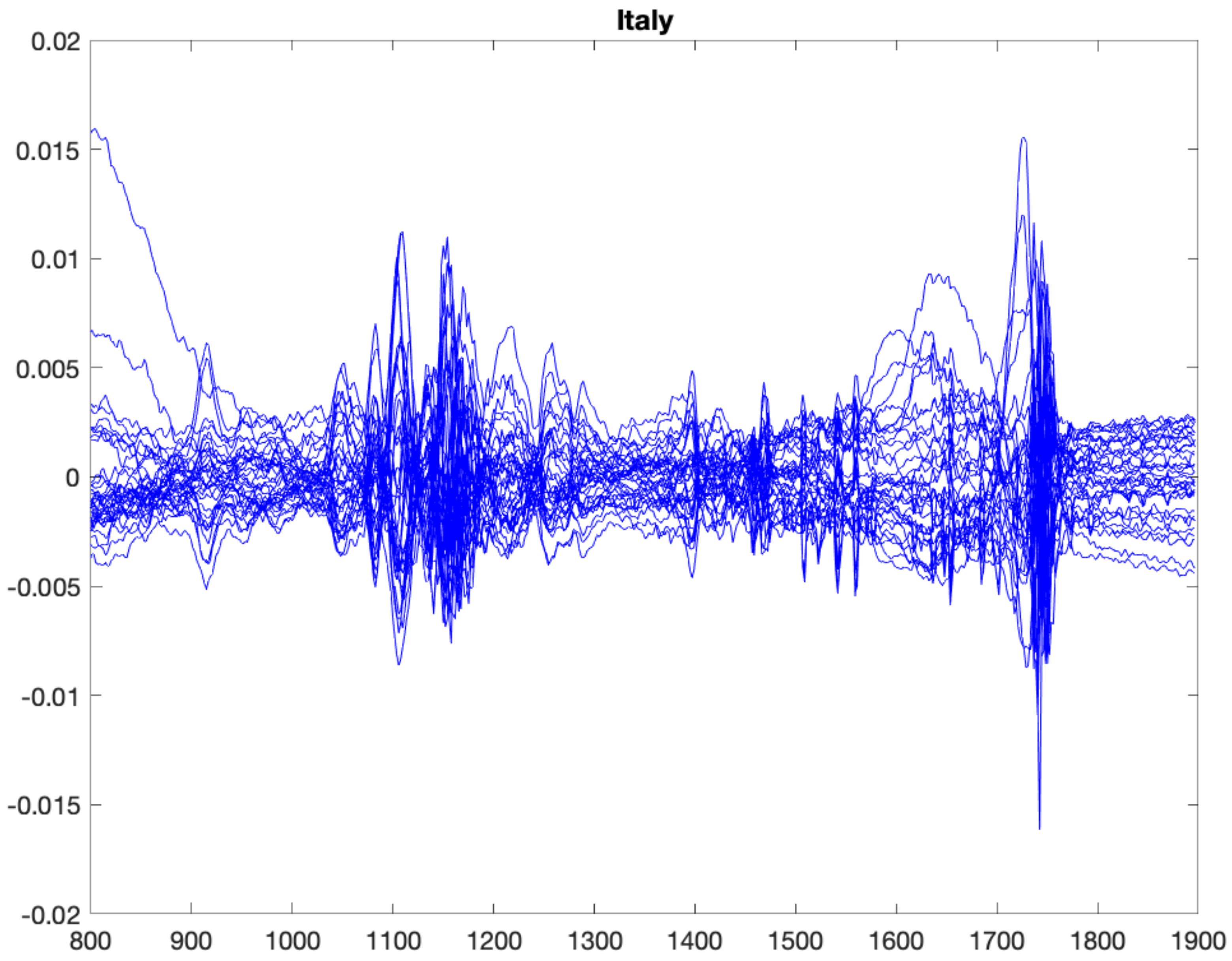
Preprocessing. Raw data

- **Mean-centering** – is used to make wavelength values more comparable and spectrum of samples from different countries more distinctive
- **Autoscaling** – should not be used for the FTIR data, because we would loose important information about peaks in spectrum

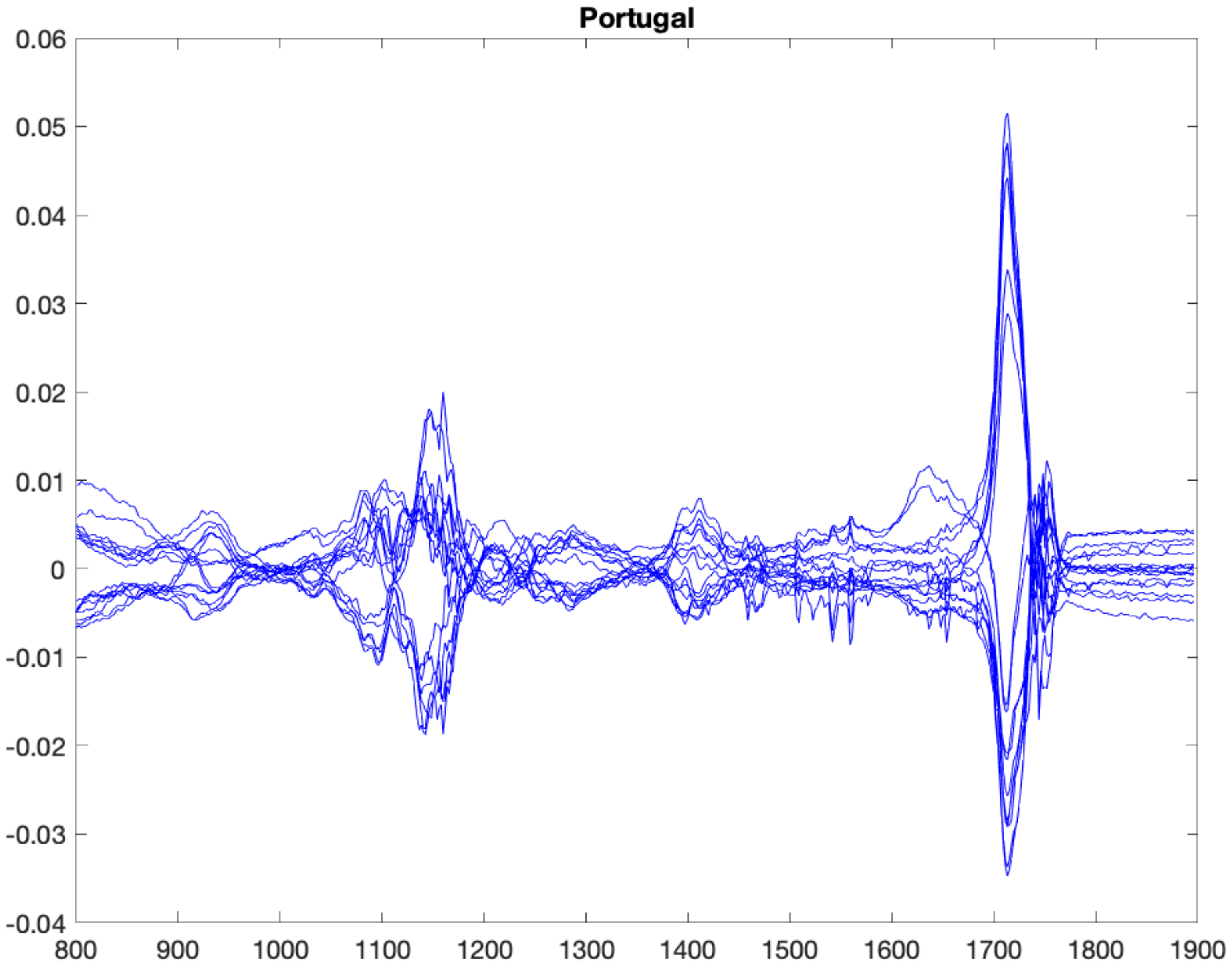


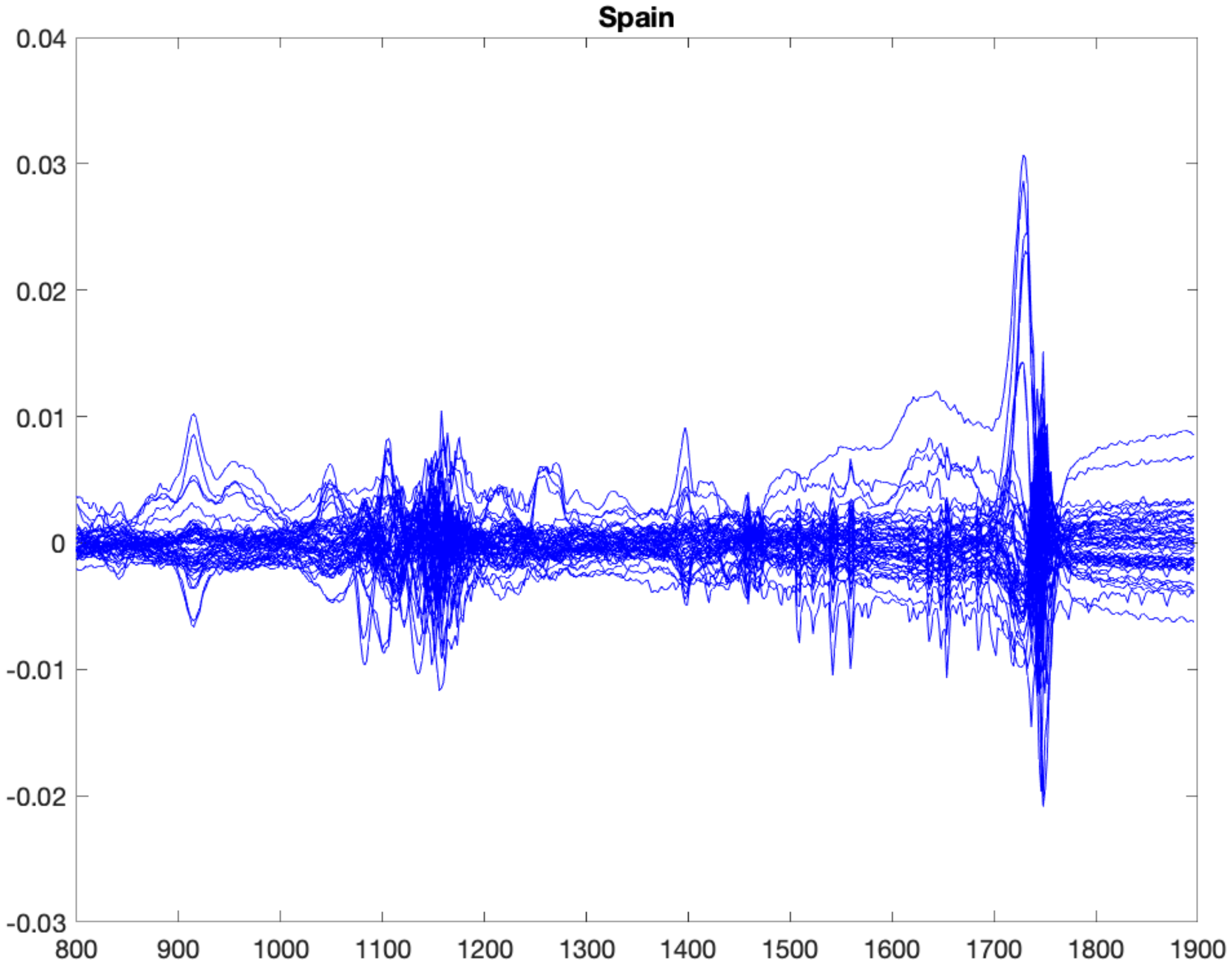
Images of raw and mean centered spectrum. Greece and Italy





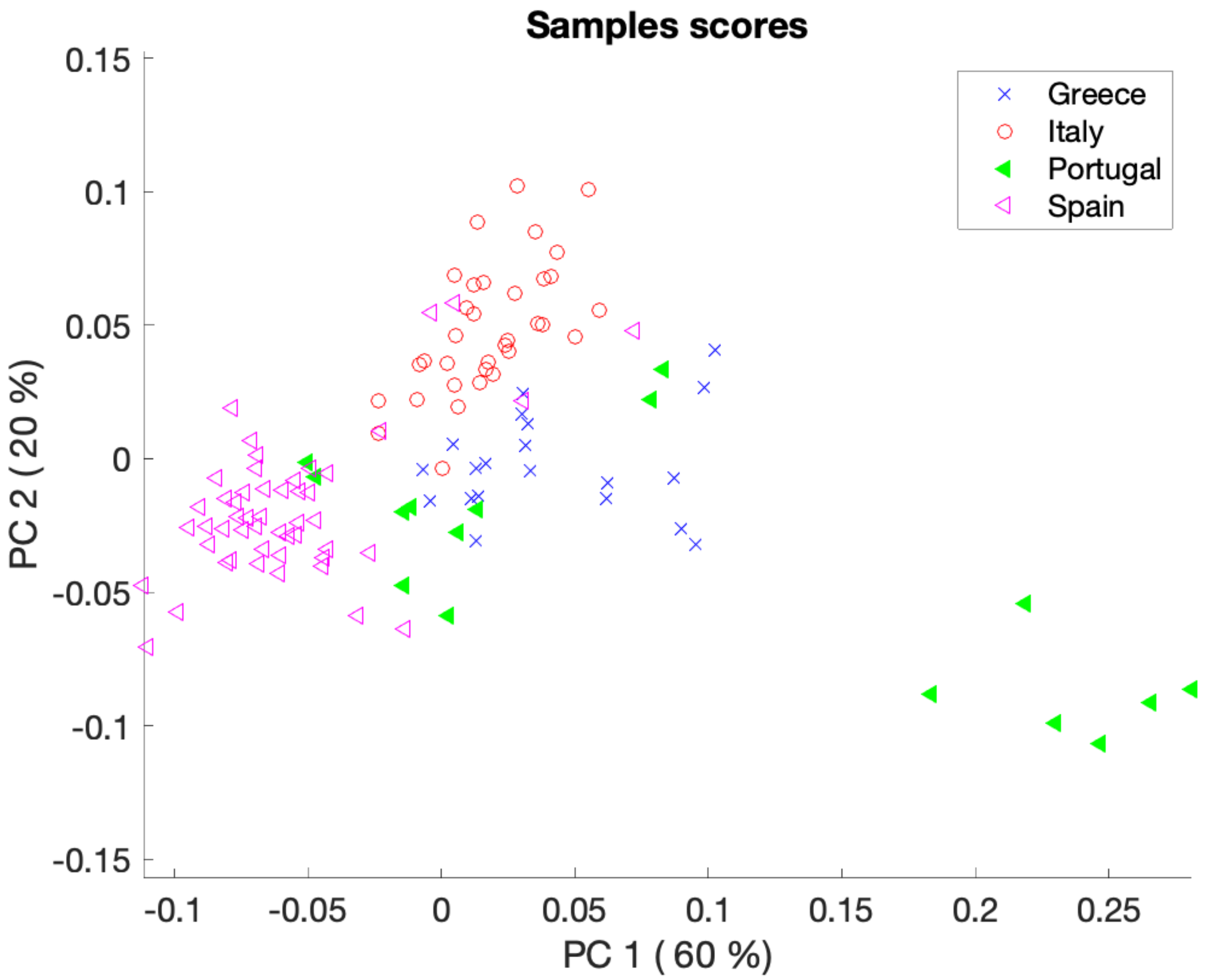
Images of raw and mean centered spectrum. Portugal and Spain



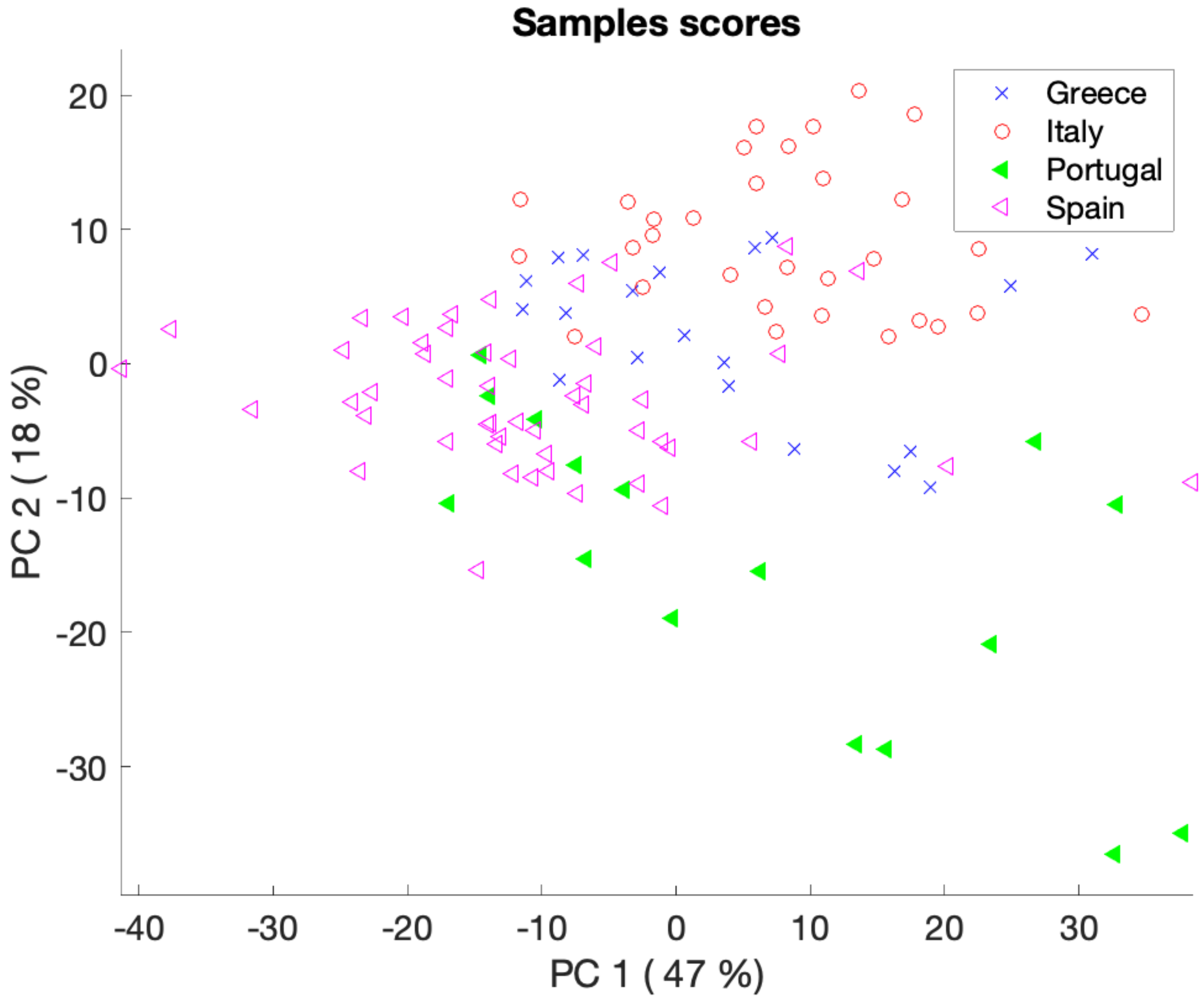


PCA analysis

Mean-centered



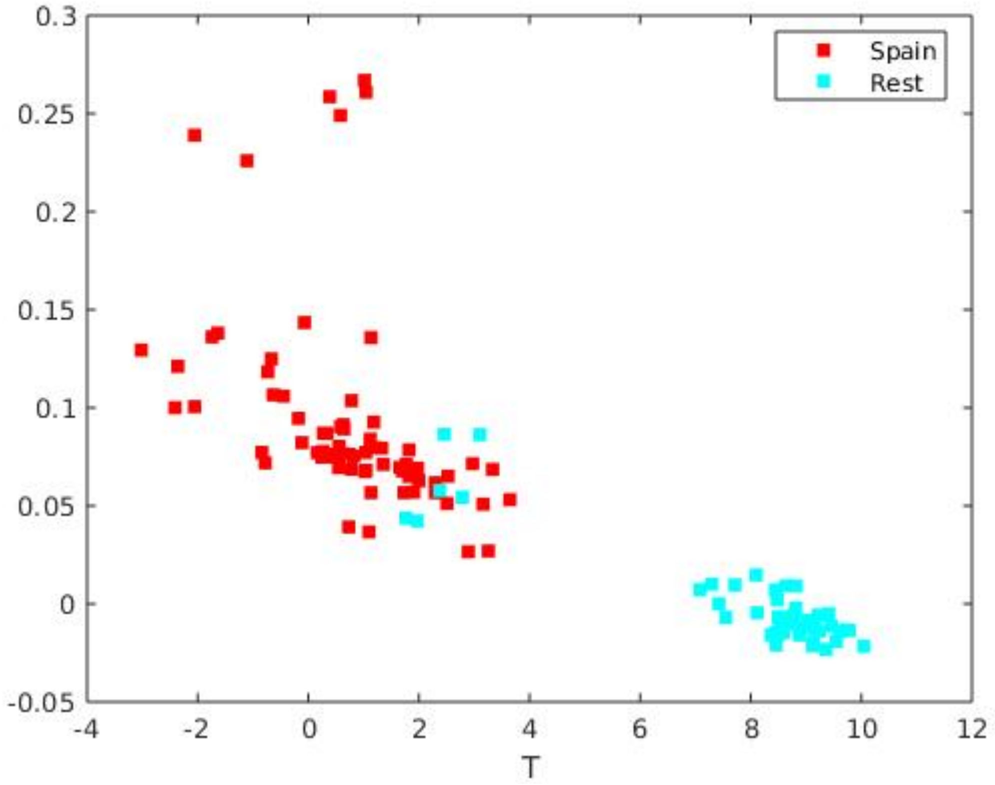
Autoscaled



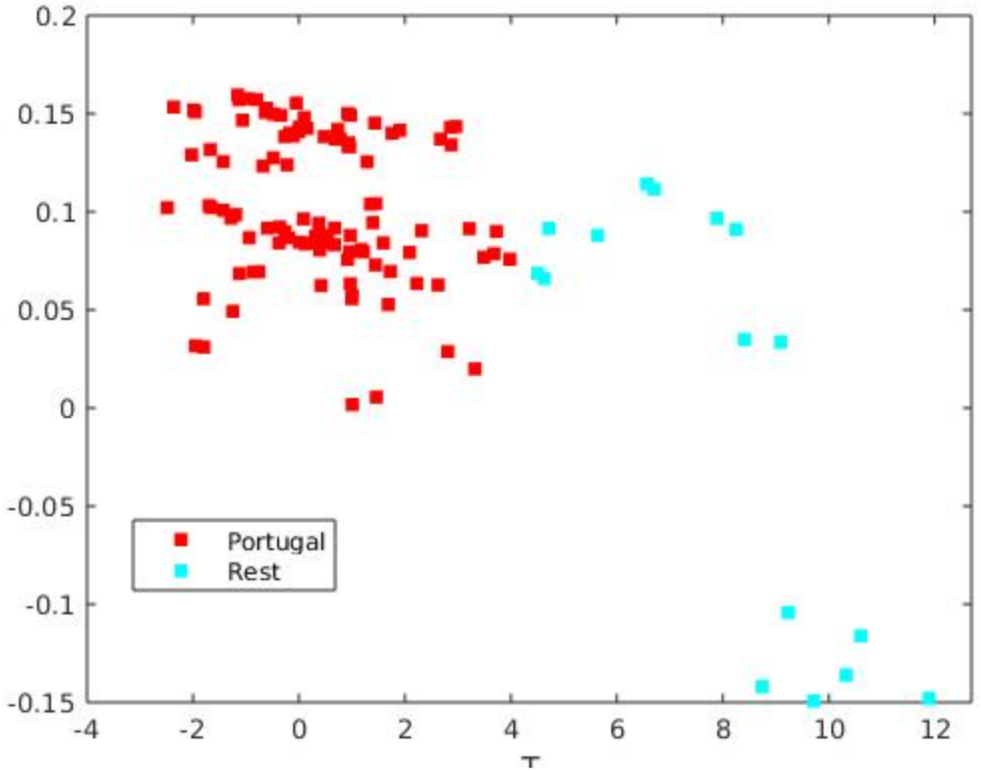
Double cross validation

- **25 %** of the data set will be used as independent **test set**
- The data is splitted into validation and test set **randomly, but**
- Since we have 2 measurements for each sample we ensure that both go to the same set to avoid sharing information between validation and test set
- We ensure that samples from different countries are evenly distributed in validation and test set
- We use the analogue of LOO cross-validation (with the correction of duplicates) for model tuning

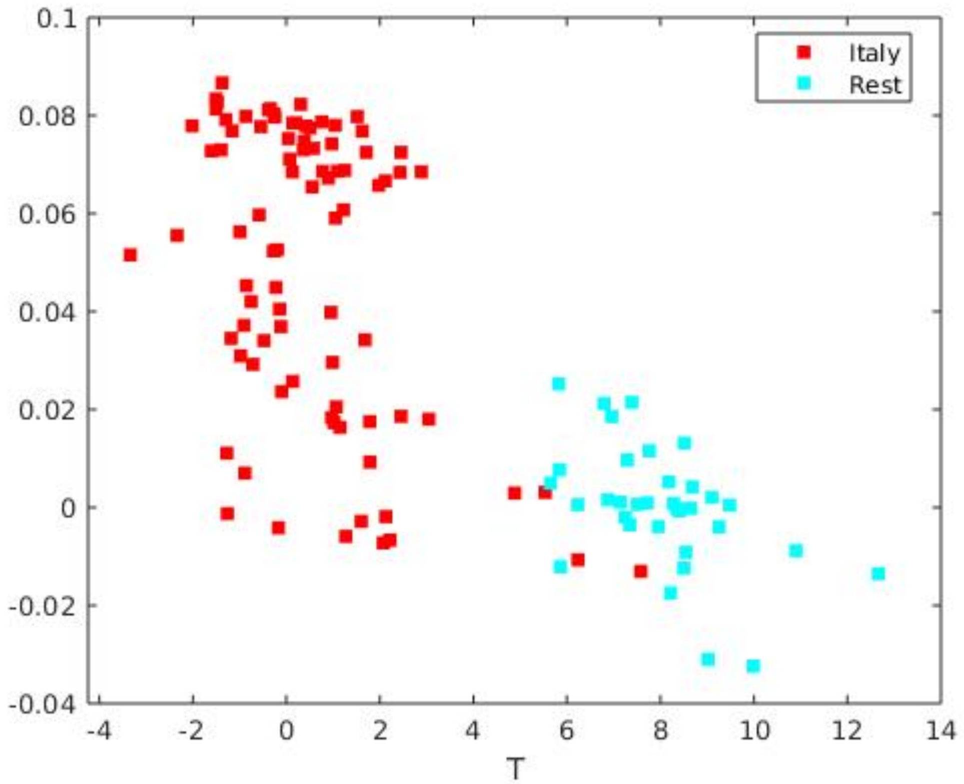
OPLS



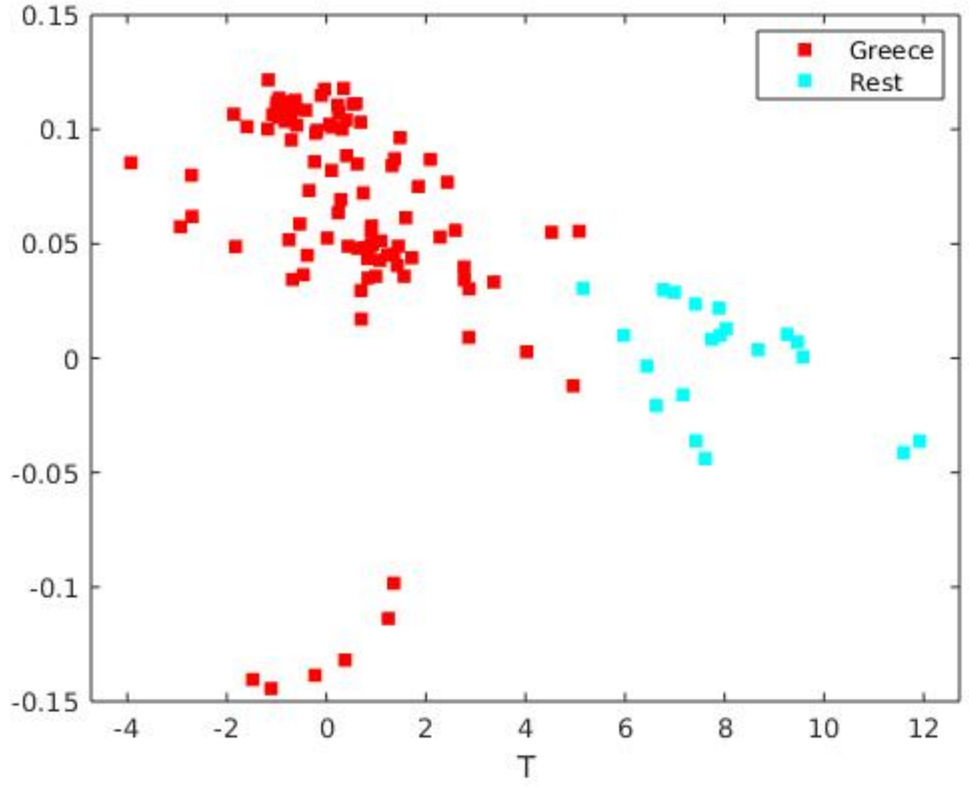
🟥 Spain
🟦 Rest
0.3
0.25
0.2
0.15
0.1
0.05
0
-0.05
-4
-2
0
2
4
6
8
10
12
T



🟥 Portugal
🟦 Rest
0.2
0.15
0.1
0.05
0
-0.05
-0.1
-0.15
-4
-2
0
2
4
6
8
10
12
T



🟥 Italy
🟦 Rest
0.1
0.08
0.06
0.04
0.02
0
-0.02
-0.04
-4
-2
0
2
4
6
8
10
12
14
T

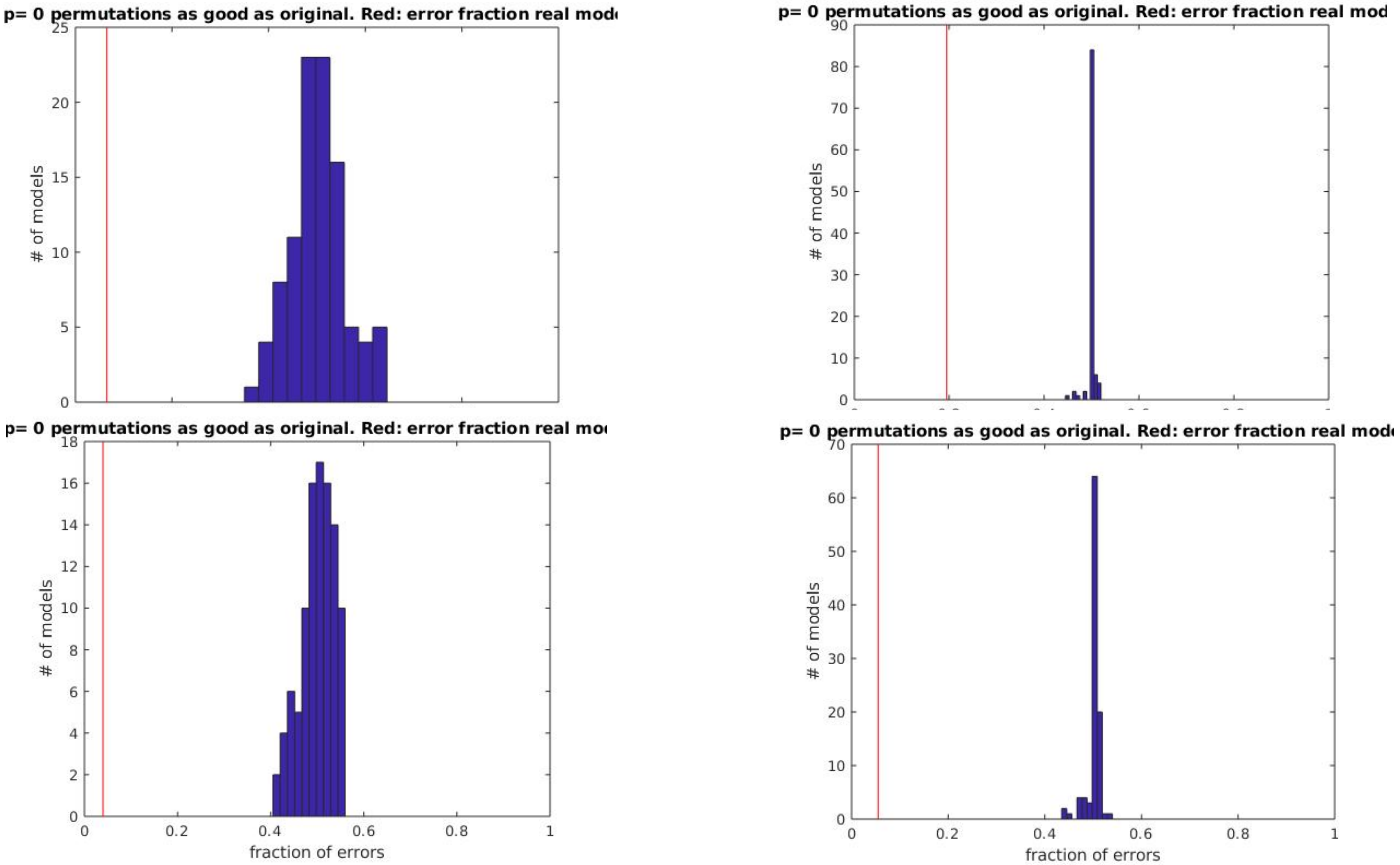


🟥 Greece
🟦 Rest
0.15
0.1
0.05
0
-0.05
-0.1
-0.15
-4
-2
0
2
4
6
8
10
12
T

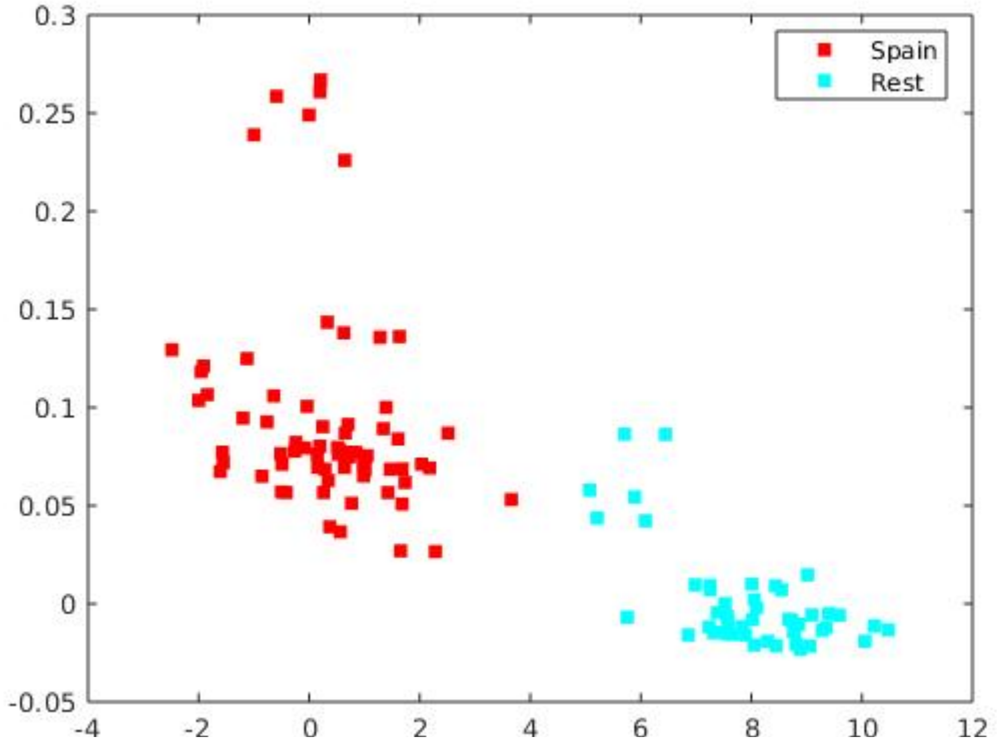
OPLS Cross Validation

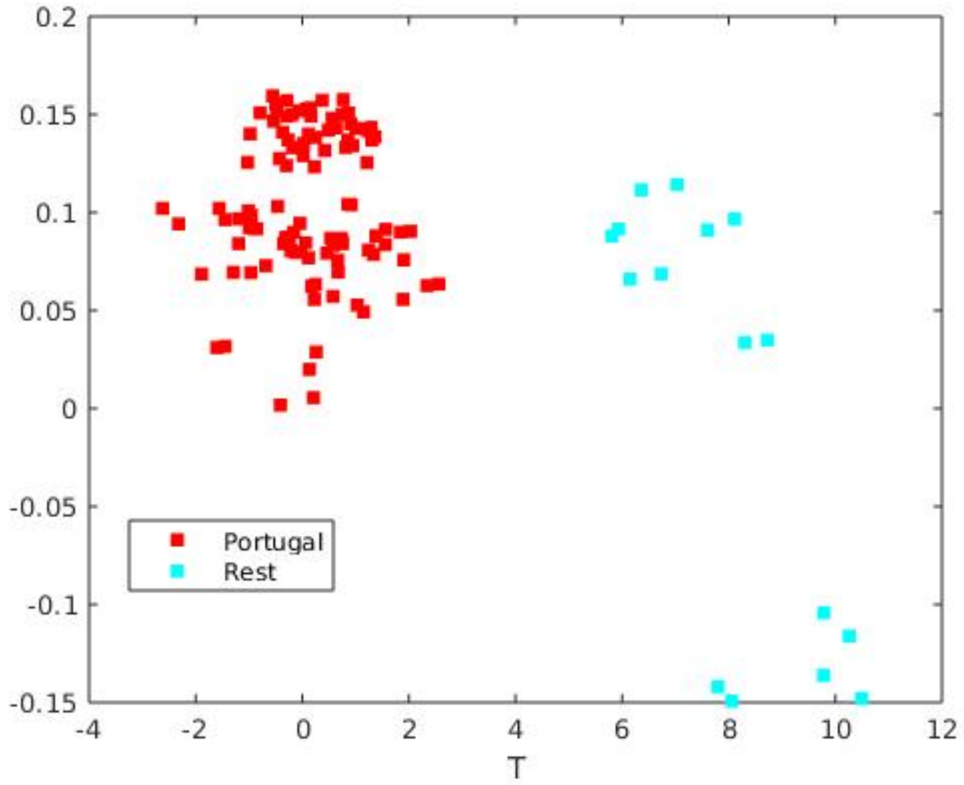
- 5 random testsets were used, the number of latent variables was 4
- Number of missclassified samples varies, but is generally low, no higher than 4%
- Q2 value qualifies as significant, generally around 0.55-0.7

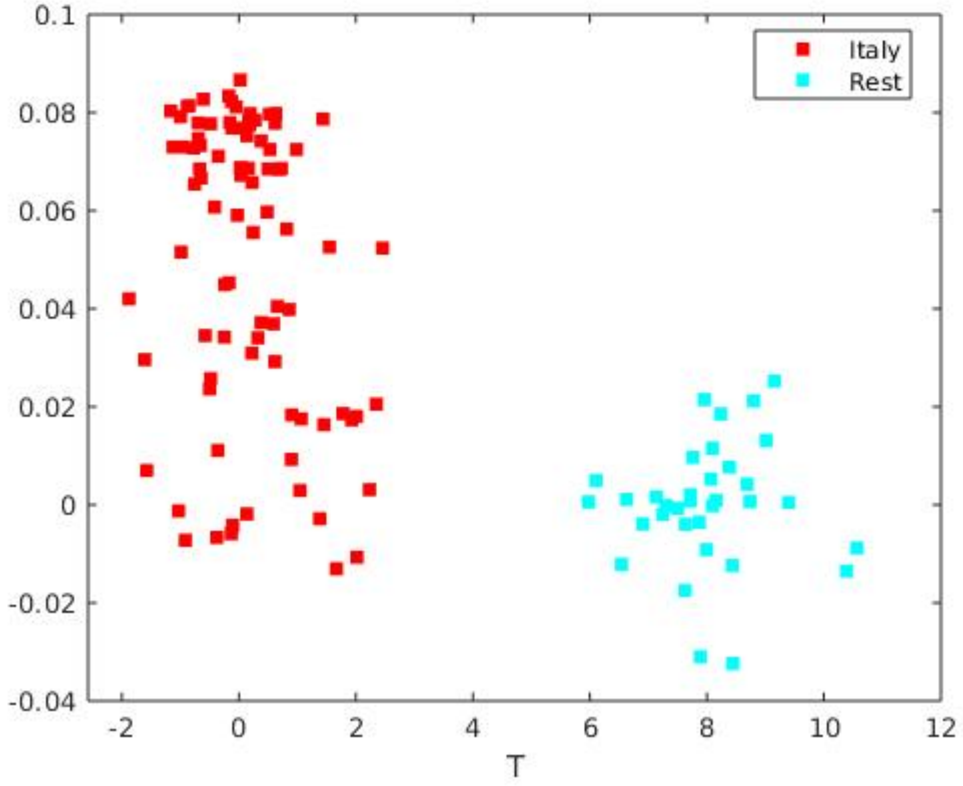
Significance test

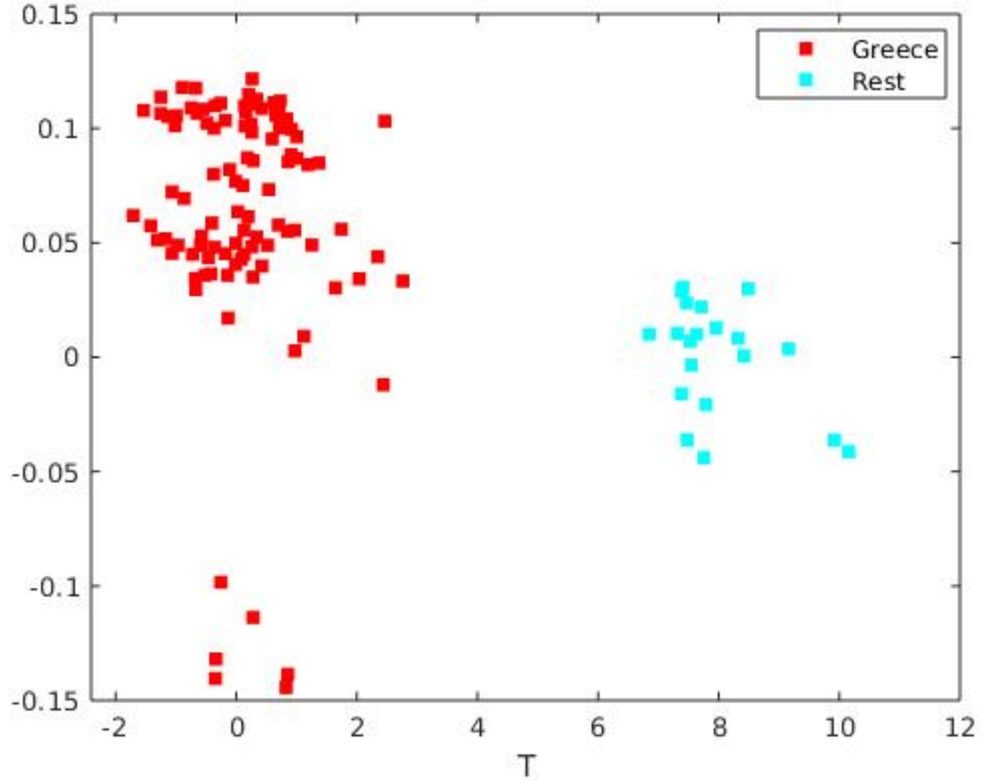


OPLS with 8 latent variables









OPLS Cross Validation

- 5 random testsets were used, the number of latent variables increased to 8
- Number of missclassified samples decreases, usually 0%, occasionally 1%
- Q2 value increases, generally around 0.7-0.8

GA implementation

- Based on the results of OPLS we aim to find a subset of 8 wavelengths to classify oils with LDA
- Initial population is generated **randomly**
- Encoding: **real values** (more precisely integers)
- For new generations: **2-points crossover** with 0.8 rate and mutation by **shifting a wavelength number** in (-10, 10) interval with 0.05 rate
- Generational reproduction with ranked-based selection strategy is used
- Fitness function: LOO internal validation
- Final evaluation with independent test set (25%)
- Stop criterion: 0 error rate or max number of generations is reached

GA – tried parameters settings

1. GA1 - population size: 200, max number of iterations: 30. 100 runs
2. GA2 - population size: 50, max number of iterations: 50. 100 runs

Baseline – average error rate of 1000 classifications based on randomly selected wavelengths subset

Method	Average error rate	Convergence	Zero error rate runs
Random baseline 1	22.11	-	-
GA1	18	100 of 100 (min 1, max 12)	0
Random baseline 2	10.44	-	-
GA2	4.85	77 of 100 (min 4)	24

GA2 zero error rate solutions evaluation

We evaluate zero error rate the solutions obtained by GA2 with 4-fold cross-validation (to have 25% test set).

CV is performed 10 times with different training and test sets splits and the results are averaged for each solution

- Min – 0.33
- Max – 3.42

Top ten solutions error rates
0.33
1.24
1.25
1.25
1.48
1.73
1.83
1.83
1.83
1.88

Top-10 selected wavelengths statistics

Wavelengths range: 799 -1897

Wawelength	Occurences
1007.3	24
1134.6	19
1132.7	18
1202.2	17
1617.0	15
1128.8	13
1130.8	13
1005.3	11
1013.1	11
1620.9	11

GA issues and possible improvements

Main issues:

- 100 runs of both GA versions took about 3 hours
- The results tend to be affected by a separation into training and tests set
- LOO validation tend to overfit

Possible improvements:

- Generate initial population based on obtained knowledge
- More GA runs
- More detailed analysis of top selected wavelenghts
- GA implementation