Description of the Repeatability/reproducibility database

From Rousseau (2011) and Martin & Govaerts (2019)

This article presents the Repeatability/reproducibility database originally used in Rousseau (2011) and Martin & Govaerts (2019). This database was designed with spectroscopists from Eli Lilly and the University of Liège.

Motivations for creating this database

This serum database is dedicated to the exploration of the impact of multi-source biological factors on the spectral variability (*e.g.* variability between and within healthy patients, etc.). The goal of this database is not to describe all the biological variance components in detail. The aim instead is to get a reasonable idea of the global amplitude of natural biological variability due to inter- and intra-individual sources. Another purpose of this database is to compare the analytical variability and the biological one. For various further use of the discovered metabolomic biomarkers, it is very crucial to have an analytical variability lower than the biological one. If a discovered metabolomic spectral biomarker corresponds to a spectral zone presenting a systematically higher intensity but also a huge variability due to the acquisition or the pre-treatments of data, detection tools built with this biomarker will have a low sensitivity and specificity.

This database was created from the collection of blood samples taken from volunteers. Samplings were organised following a procedure allowing to observe the biological factors of interest. The resulting blood samples were then prepared and measured in order to explore some analytical factor.

Statistical experimental design

Four factors of variability were considered in the experimental design:

1. The volunteer (from 1 to 12): the blood of twelve volunteers was collected. Since the alterations of the spectra due to diseases is outside the scope of the current analysis, all selected volunteers were healthy. Among the volunteers, seven were women and five were men. By controlling the sex balance, sex is not a confounding factor in the study.
2. The sampling for a given volunteer (from 1 to 3): the blood sampling was performed on three non consecutive days under similar conditions. Similar conditions means that the three blood samplings were all carried out the morning around 10 a.m.. The volunteers were encouraged to have fasted. Thanks to these mea- sures, the fluctuation and the variability between the individuals was controlled.
3. The tube for a given volunteer and a given sampling (from 1 to 2): each blood sampling day, two tubes were collected. Since two samples are available for a given volunteer the experimental variance can be estimated. By experimental variance, we mean the variability due to the laboratory operations.
4. The time of measurement: each tube was measured twice the same day with a couple of hours between the two measurements. The replication allows determining whether the sample can be con- served after defrozing. Actually this has been already considered in the previous serum database. The only difference is the use of hu-

man serum instead of commercial serum from rats.

The two first factors are biological sources of variability while the two lasts are analytical sources.

Sample preparation

The preparation of the samples and the spectral acquisition were realised at the Laboratoire de Chimie Pharmaceutique of ULg, Belgium. Once the blood sample had been collected, it was processed into serum. Then phosphate buffer was added to control the pH. The trimethylsilyl propanoate (TMSP) was finally put in as internal reference. The protein signals were eliminated thanks to the pulse sequence CPMG. Serum samples were conserved frozen.

The pre-treatments

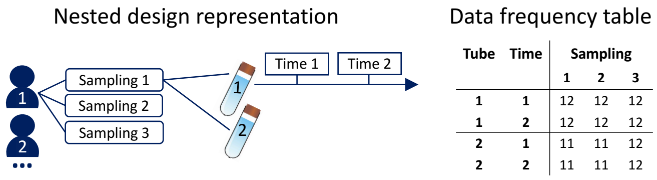
Each acquired spectrum was pre-treated according to the procedure proposed in Rousseau (2011). Each spectrum has 750 ppms and is normalised to have a sum equal to 750.

The final database

In total, $12 \times 3 \times 2 \times 2 = 144$ spectra were analysed. Each spectrum is available at a resolution of 750 buckets (*i.e.* variables) and observations are named with a series of 5 digits as: volunteer-sampling-tube-time ; e.g. 01312 corresponds to volunteer 01, sampling 3, tube 1 and time 2.

4 observations had an acquisition problem and were removed beforehand (04221, 08222, 11121,11122), rendering the design unbalanced with a total of 140 observations instead on 144, as represented in Figure 1.

The data are contained in the *HumanSerumData.Rdata* file with the *design* data.frame with the 4 variables from the experimental design and the *outcomes* matrix containing the spectra.



Tube	Time	Sampling		
		1	2	3
1	1	12	12	12
1	2	12	12	12
2	1	11	11	12
2	2	11	11	12

Figure 1: *Nested design representation and data frequency table for the Human Serum (Repeatability/reproducibility) dataset.*

A stacked representation of the spectra from this dataset is illustrated in Figure 2.

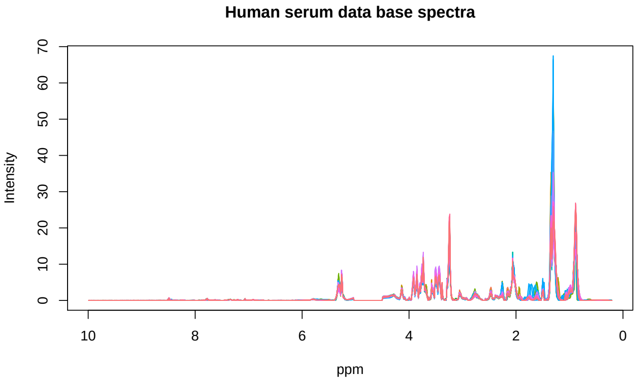


Figure 2: *Plot of the Repeatability/reproducibility database*

References

Martin, M., & Govaerts, B. 2019. *LiMM-PCA : combining ASCA$^+$ and linear mixed models to analyse high dimensional designed data.* Discussion Paper DP 2019/21. Institute of Statistics, Biostatistics and Actuarial Sciences, UCLouvain, Belgium.

Rousseau, R. 2011. *Statistical contribution to the analysis of metabonomics data in ^{1}H NMR spectroscopy.* Ph.D. thesis, Institute of Statistics, Biostatistics and Actuarial Sciences, Université catholique de Louvain, Belgium.