Description of the Semi-artificial Urine database

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

This article presents the Semi-artificial Urine database originally used in Rousseau (2011) and Martin & Govaerts (2019). This database was designed with spectroscopists from Eli Lilly and the University of Liege (ULg).

Incentive

The *semi-artificial urine database* was built to assess the quality of potential biomarkers' identification with known in advance biomarkers' location. This exact localisation enables to precisely quantify the statistical methods performances to retrieve the true spectral descriptors that were altered. To build this database, random artificial alterations were added to "control" rats urine spectra.

Sample acquisition and pre-processing

The primary dataset represents the status of physiological stability in rat urine. It is composed of FID signals issued from the COMET database (Lindon *et al.* , 2003) where all samples come from "control" rats that did not receive any treatment. These data were obtained with a flow injection process and a 600 MHz ^{1}H NMR spectroscope at the Imperial College in London.

To obtain interpretable spectra, pre-treatments (suppression of the residual water signal, apodization, baseline correction, and warping, window selection and bucketing to $m = 500$) of the raw data were handled in an automatic fashion by BUBBLE (Vanwinsberghe, 2005), the PepsNMR's precursor. A typical pre-processed spectrum can be seen in Figure 1. Note that non-informative areas with water and urine peaks (4.5 - 6 ppm) that are displayed in red in this graph were also removed from the spectra. Finally, some outliers were removed based on PCA and Euclidian/Mahalanobis distances.

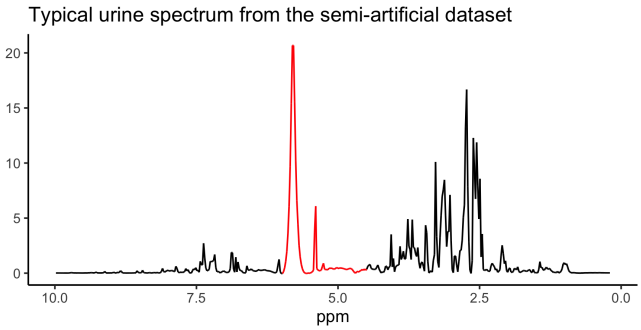

Figure 1: *Typical rat urine spectrum with the water and urine area to be deleted represented in red.*

Inclusion of artificial alterations

Half of the spectra are normal and half are altered. For the altered profiles, 46 descriptors are artificially positively altered with random heights from Gamma distributions. They are scattered across 10 different spectral regions, half localised in a noise-free area (6.48 - 7.26 ppm) and half in the noisy part (2.56 - 3.37 ppm) of the spectrum. Furthermore, some pairs of altered regions are correlated by using the same Gamma distribution to generate peaks, leaving only 6 independent biomarkers in the dataset (only six independent Gamma distributions have been used instead of ten). Note also that four peaks have a width of seven descriptors and the six other peaks have a smaller width of three descriptors, for a total of 46 descriptors or biomarkers. Finally, 2 out of the 6 independent biomarkers are randomly chosen to

alter each spectrum in order to mimic the natural variety of body response to a same stimulus. Figure 2 from Rousseau (2011) presents the location and the mean amplitude of all possibly added alterations. The index and ppm values of those biomarkers are given in Table 1.

Table 1: *Column indexes for artificial alterations peaks in the ^1H NMR spectra.*

B1	139 140 141 159 160 161
B2	147 148 149 150 151 152 153 167 168 169 170 171 172 173
B3	179 180 181
B4	339 340 341 359 360 361
B5	347 348 349 350 351 352 353 367 368 369 370 371 372 373
B6	379 380 381

Note that after the alterations were added, a constant sum normalisation was applied as well.

The final semi-artificial urine database

The total size of the final database is (800×500), were half of the samples (460) are artificially altered spectra. Finally, random draws of sizes 200 (2×100) and 60 (2×30) are extracted to create the sub-samples used in Chapter 3. The design matrix is represented by a binary variable coded $[1, 0]$ that indicates the presence/absence of the alterations. This dataset is available in the *DataSimul.RData* file that contains the x matrix with the NMR spectra and the binary vector y indicating if the spectra are altered $(y = 1)$ or not $(y = 0)$.

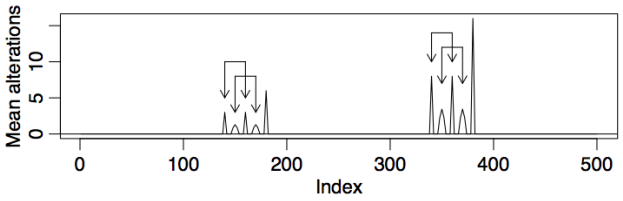

Figure 2: *Representation of the artificial alterations in the spectrum. Figure from Rousseau (2011).*

References

Lindon, J. C., Nicholson, J. K., Holmes, E., Antti, H., Bollard, M. E., Keun, H., Beckonert, O., Ebbels, T. M., Reily, M. D., Robertson, D., *et al.* . 2003. Contemporary issues in toxicology the role of metabonomics in toxicology and its evaluation by the COMET project. *Toxicology and Applied Pharmacology*, **187**(3), 137–146.

Martin, M., & Govaerts, B. 2019. *Feature Selection in metabolomics with PLS-derived methods.* Discussion Paper DP 2019/20. Institute of Statistics, Biostatistics and Actuarial Sciences, UCLouvain, Belgium.

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

Vanwinsberghe, J. 2005. *Bubble: development of a matlab tool for automated ^1H-NMR data processing in metabonomics.* Traineeship report (unpublished results), Strasbourg University.