Title: Robustifying a PLS property prediction workflow on NMR spectra with optimized processing

Author list : DUVAL Laurent (1,2), LACOUE-NEGRE Marion (3), PIRAYRE Aurélie (1)

(1) IFP Energies nouvelles, 92852 Rueil-Malmaison Cedex, France

(2) University Paris-Est, LIGM, Noisy-le-Grand, France

(3) IFP Energies nouvelles, 69360 Solaize, France

Keywords: data mining; NMR; precision; predictive models; property prediction

Property analysis and quality assessment are fundamental needs in the study of complex mixtures, for instance petroleum fractions or biomass products. For the purpose of experimental efficiency in process development, finding cheaper alternatives is a rising trend with the development of high-throughput experiments (HTE). These units produce smaller sample volumes that are not compatible with the standard analytical process applied to determine petroleum cuts properties (Y) such as density, viscosity, etc. It is time-consuming (up to two days for some properties).

An alternative for Y prediction is to combine analytical techniques (requiring a small volume of sample) with data mining. Given a subset of representative data (X), one strives to develop a predictive model P, such that P(X)~Y with sufficient precision compared to the standardized reference. Principal component regression (PCR) or Projection onto Latent Structures (PLS) prediction tools.

Such chemometric models do not resist straightforwardly to artifacts (X, Y, model limits). Exploratory data analysis (grouping patterns, infrequent data, outliers) and analytical signal preprocessing (normalization, alignment, denoising, detrending) may attenuate parasite effects (in sample preparation or instrumental variations). Preprocessing and prediction are classically applied in compartmentalized workflows, each with specific statistical assumptions and parameters [2]. They are at risk of producing over-parametrized solutions, with more tedious interpretation and generalization. Their control crucially depends on gray-box preprocessing, experimental diversity, model-maker skills. Not all of the former benefit from advanced optimization. Pressing needs in faster turnaround acquisition and modeling time impose building more precise and less fragile (to artifacts) data-driven processing and prediction workflows. The purpose is to reduce the over-parameterization burden, and help model-makers focus on their field-of-expertise, such as the conception of calibration/prediction databases, the determination of the numbers of latent variables, the evaluation of prediction accuracy.

In this contribution, we analyze a previously validated NMR predictive workflow. It relies on a curated dataset of 309 NMR spectra and its associated measured property (the Viscosity Index) to predict. The standardized method for the latter property raises questions regarding its repeatability and reproducibility [3]. However, in [4], a predictive PLS model was obtained with 12 latent variables (Figure 1). To increase the accuracy of the current prediction, we revisit this workflow to identify key leverage data processing and property prediction techniques throughout a detailed sensitivity analysis. We examine the relative importance of specific processing techniques, among which spectral selection, outlier rejection, trend [5] and noise reduction, with respect to novel alternatives aiming at introducing so-called “sparsity” or “robustness” to the workflow. Improvements in prediction accuracy will be evaluated on several accounts, including precision and resistance to outliers and model parameters, toward better integrated optimization algorithms.

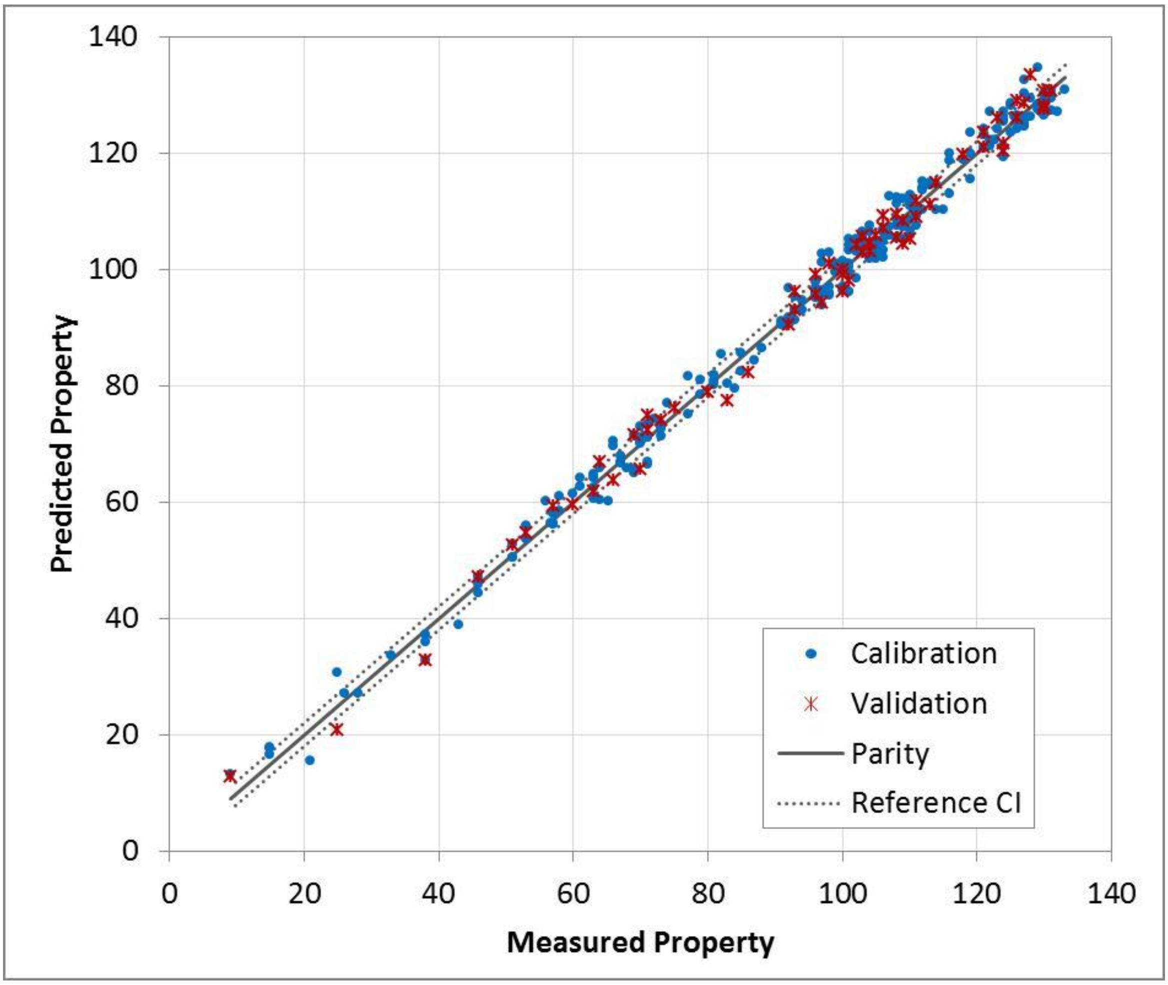

Figure 1: Parity plot for measured/predicted property.

1. S. Wold, M. Sjöström and L. Eriksson, PLS-regression: a basic tool of chemometrics, 2001, Chemometrics and Intelligent Laboratory Systems, DOI:10.1016/s0169-7439(01)00155-1

2. K. H. Liland, T. Almøy and B.-H. Mevik, Optimal Choice of Baseline Correction for Multivariate Calibration of Spectra, 2010, Applied Spectroscopy, DOI:10.1366/000370210792434350

3. S. Verdier *et al.*, A critical approach to viscosity index, 2009, Fuel, DOI:10.1016/j.fuel.2009.05.016

4. M. Lacoue-Nègre *et al.*, PANIC 2015

5. X. Ning, I. W. Selesnick and L. Duval, Chromatogram baseline estimation and denoising using sparsity (BEADS), 2014, Chemometrics and Intelligent Laboratory Systems, DOI:10.1016/j.chemolab.2014.09.014

The authors acknowledge the support from the Excellence Initiative of Aix-Marseille University — A*Midex — a French ”Investissements d’Avenir” program (BIFROST[1] project).

[1] https://www.i2m.univ-amu.fr/project/bifrost