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

Laurent Duval[1,2], Marion Lacoue-Nègre[1], and Aurélie Pirayre[1]

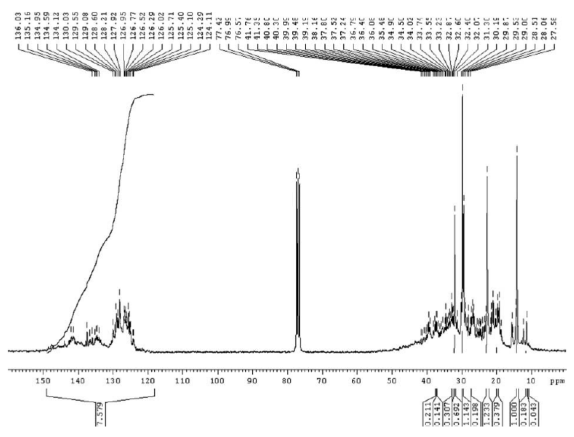
[1] IFP Energies nouvelles, [2] Université Paris-Est

Context

- Study of complex mixtures (petroleum fractions/biomass products):
  - Property (Y) analysis & quality assessment: fundamental needs
  - Standardized methods: "sufficient" quantities, time-consuming
- To increase experimental process efficiency:
  - High-throughput experiments (HTE) are developed
  - Smaller sample volumes: not compatible with standards
- Alternative: predict property Y from representative samples:
  - With analytical techniques (requiring small volumes)
  - Combined with processing workflow on analytical signals X

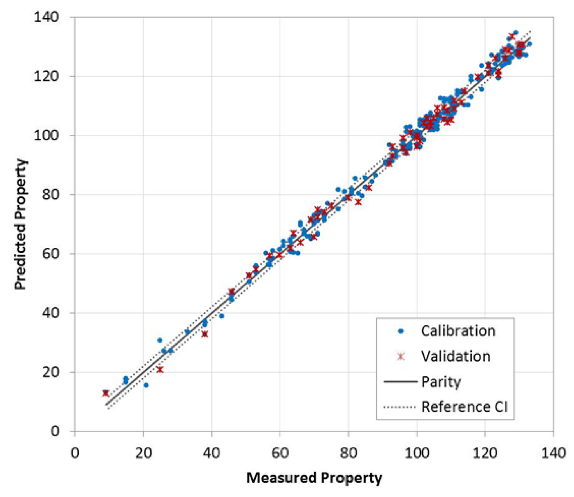
Background on prediction

"Ideal-style" model [4]



$$\mathrm{VI} = 203.0 - 41.32\left(\frac{A_2}{A_3}+\frac{A_7}{A_3}+\frac{A_{10}}{A_3}\right) - 2.206\cdot 10^{-3}\left(\frac{A_{\mathrm{aro}}}{A_3}\right)^5$$

Learned model



Measured/predicted parity plot

Issues

- Predictability of properties
- Predict models f: $Y \approx f(X)$ (PLS)
  - Model simplicity (sparsity)
- Sample base homogeneity
- Samples already "separated"
- Parasite effect resistance
  - Sample preparation, batch effect
  - Instrumental variation
  - Artifacts: abnormal, outlier data

Experimental background & challenge in VI prediction from NMR spectra

Crude oil —*hydrotreatment* / *hydrocracking*→ **Total Liquid Effluent (TLE)** → Distillation:

- Gas (LNG, LPG)
- Gasoline → RON, MON…
- Jet fuel → Smoke point, cetane…
- Diesel → Cloud point, pour point…
- Fuel oil… → **Viscosity Index (VI)**…

- "Ideal" [4]: **VI** prediction from fuel oil only

$$\text{Fuel oil} \longrightarrow \mathbf{VI}$$
$$\text{Fuel oil} \to \text{NMR} \xrightarrow{f} \mathbf{VI}$$

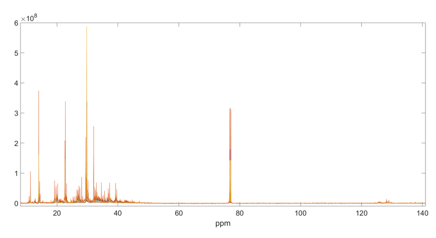
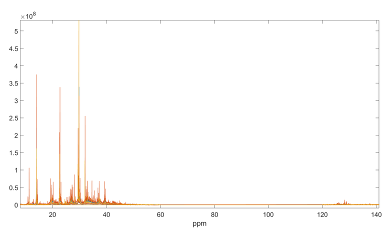
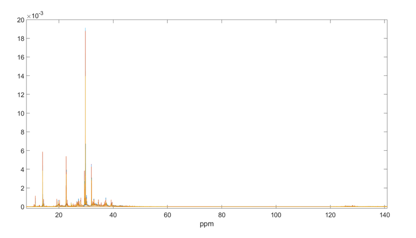
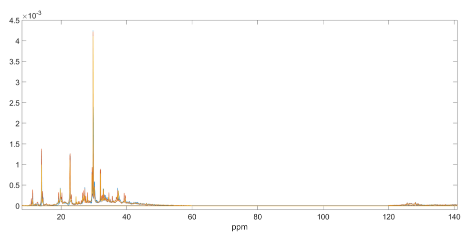
- Challenge: **VI** prediction from **TLE**

$$\mathbf{TLE} \xrightarrow{\textit{distillation}} \text{Fuel oil} \longrightarrow \mathbf{VI}$$
$$\mathbf{TLE} \to \text{NMR} \xrightarrow{f} \mathbf{VI}$$

Methodology

- Purpose (309 NMR spectra database)
  1. Ease tuning by non-statistians: shrink workflow complexity
  2. Improve model precision: robustify data/model performance
  3. Reduce over-parametrization: focus on important processing
- Lever identification
  - Align (*icoshift* [3]), normalize & filter (Savitzky-Golay [2])
  - Predict: from standard to sparse *snipls* [1]

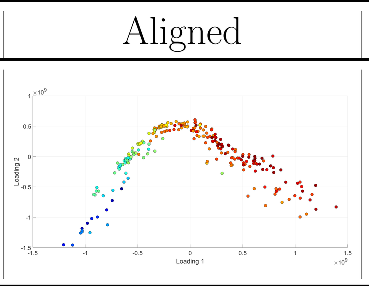
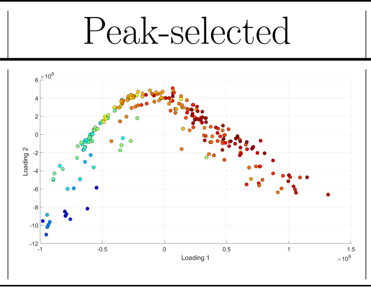
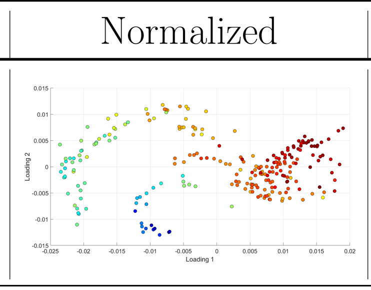
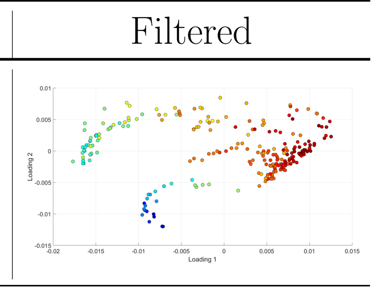
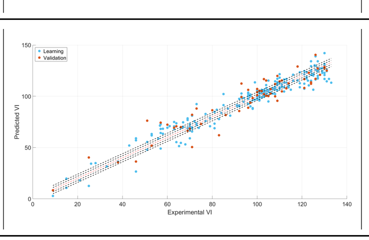
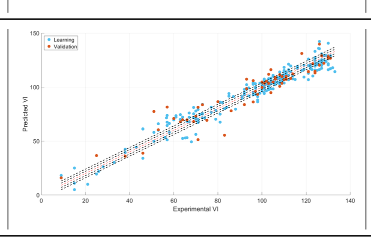
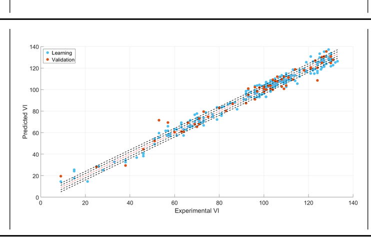
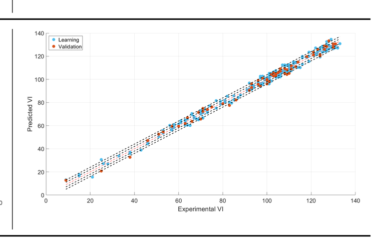
NMR processing workflow for VI

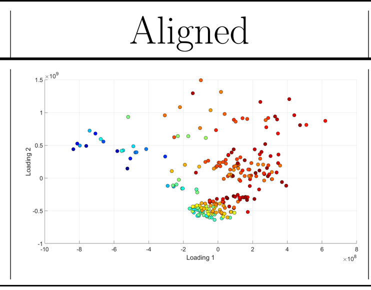
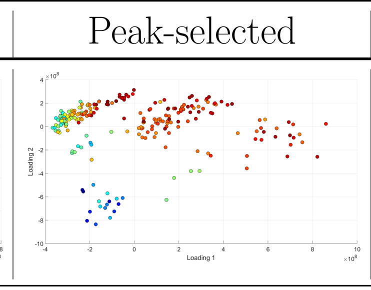
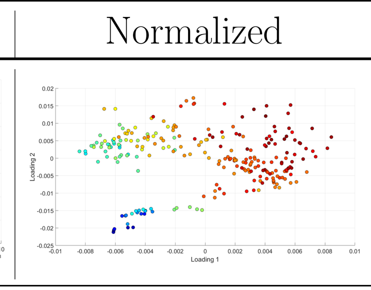
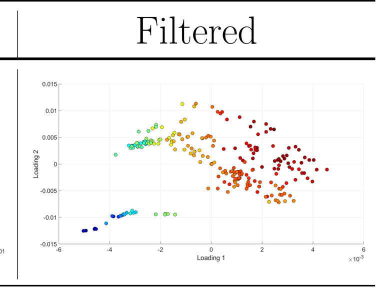
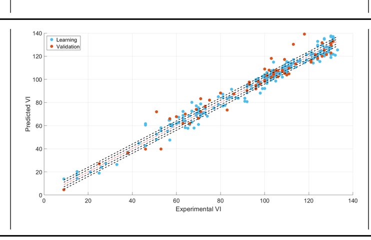
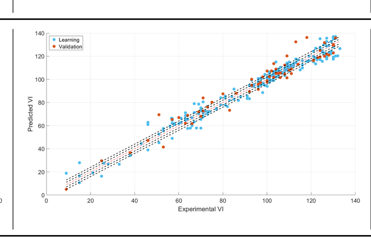
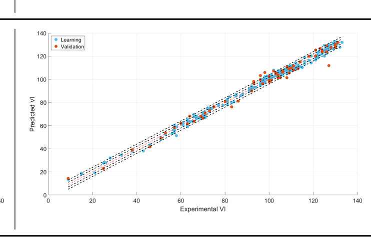
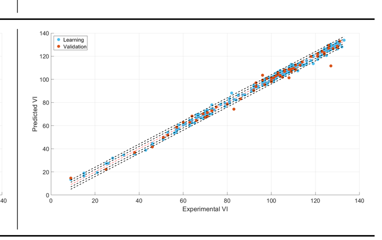
NMR spectra $\xrightarrow{\textit{alignment}}$ Aligned spectra $\xrightarrow{\textit{band selection}}$ Peak-selected spectra $\xrightarrow{\textit{area normalization}}$ Normalized spectra $\xrightarrow{\textit{filtering}}$ Filtered spectra $\xrightarrow{\textit{PLS}}$ Predicted VI

Pipeline improvement results

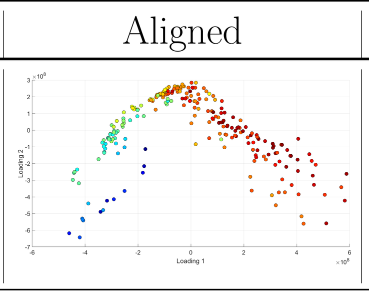
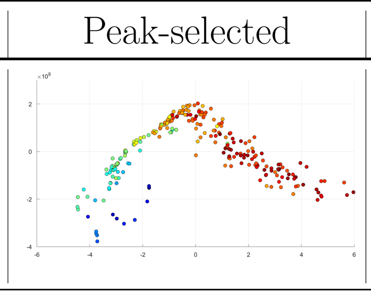
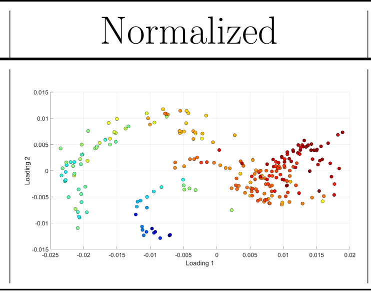
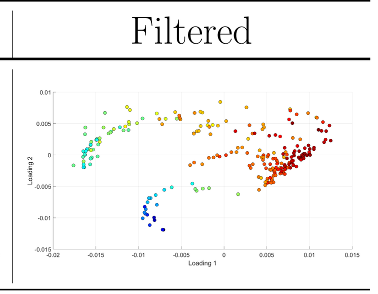
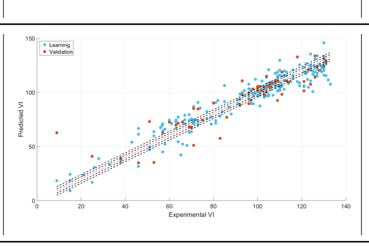
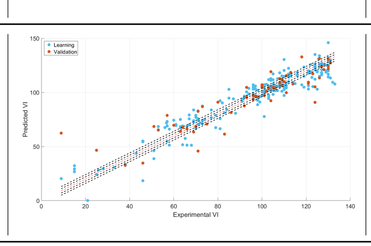
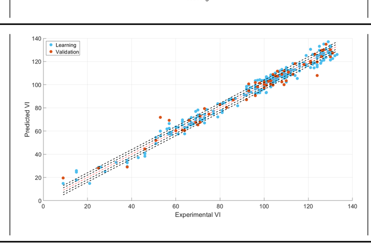
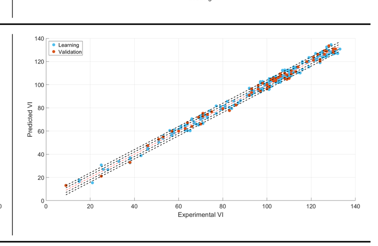
- MAE and Gain accross the workflow

	Aligned	Peak-selected	Normalized	Filtered
PLS				
Parity plot				
MAE	6.09	5.60	3.10	2.05
Gain (%)	–	**8.0**	**44.6**	**33.9**

- MAE and Gain using *icoshift*

	Aligned	Peak-selected	Normalized	Filtered
PLS				
Parity plot				
MAE	4.0	4.36	2.0	1.53
Gain (%)	**34.3**	**22.1**	**35.5**	**25.4**

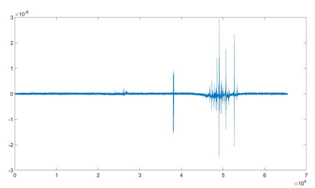
- MAE and Gain using the sparse PLS *snipls*

	Aligned	Peak-selected	Normalized	Filtered
PLS				
Parity plot				
MAE	5.89	5.41	2.93	1.95
Gain (%)	**3.1**	**3.4**	**5.5**	**4.9**

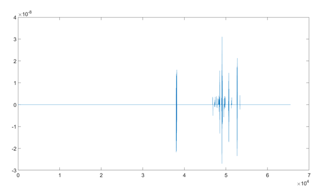
- MAE and Gain using SG filtering

	Raw	*icoshift*	*snipls*
MAE	1.60	1.23	1.54
Gain (%)	**21.9**	**18.9**	**21.0**

- Impact of sparse penalty on regression

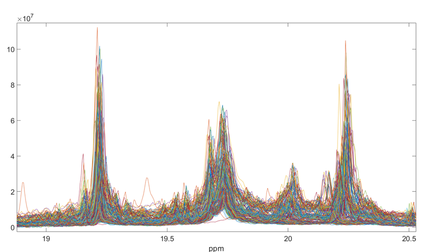


PLS coefficients

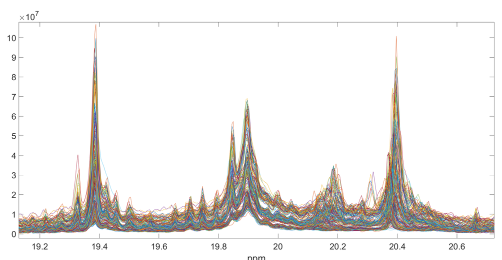


sparse PLS coefficients

- Impact of *icoshift* alignment



Aligned



Re-aligned with *icoshift*

Conclusions

- Leaving property predictability aside…
- Each processing step matters
- Various methods for each processing steps
- Some could be combined throughout modeling:
  - *eg.* band selection vs (group) sparsity
- Toward better integrated model $\mathcal{F}$/penalty $\mathcal{P}$ optimization?

$$\overline{y} = \arg\min_{\xi} \mathcal{F}(\mathcal{M}(x, y; p_\lambda(\xi))) + \mathcal{P}(x, y; \xi, p_\lambda(\xi))\,.$$

References

[1] Irene Hoffmann, Sven Serneels, Peter Filzmoser, and Christophe Croux. Sparse partial robust M regression. *Chemometr. Intell. Lab. Syst.*, 149:50–59, Dec. 2015.

[2] A. Savitzky and M. J. E. Golay. Smoothing and differentiation of data by simplified least squares procedures. *Anal. Chem.*, 36(8):1627—1639, July 1964.

[3] F. Savorani, F. Tomasi, and S. B. Engelsen. *icoshift*: A versatile tool for the rapid alignment of 1D NMR spectra. *J. Magn. Reson.*, 202(2):190–202, Feb. 2010.

[4] Sylvain Verdier, Joao A. P. Coutinho, Artur M. S. Silva, Ole F. Alkilde, and Jens A. Hansen. A critical approach to viscosity index. *Fuel*, 88(11):2199–2206, Nov. 2009.

LaTeX TikZposter