MLNIRdata: 208 near-infrared spectroscopic spectra and densities of hydrocarbon mixtures for chemometrics, data science, machine learning or signal processing

Laurent Duval[*], Louna Alsouki, Jérémy Laxalde, Noémie Caillol

August 8, 2025

Abstract

This note describes the content of the MLNIRdata dataset. It has already been used in chemometrics for property prediction of chemical mixtures with tools like "Partial Least Squares" (PLS) or sparse PLS. Its publication as "open data" is meant for further analyses and benchmarks in chemometrics, data science, machine learning, signal processing or artificial intelligence applications (prediction, regression, clustering, training, etc.). Its formats (including "csv" files) can be imported into standard data processing frameworks (Matlab, Python, Julia, R). It is available at https://doi.org/10.5281/zenodo.16781223.

Keywords— analytical chemistry, artificial intelligence, chemoinformatic, chemometrics, data science, density, hydrocarbons, machine learning, near-infrared data, prediction, partial least squares, signal processing

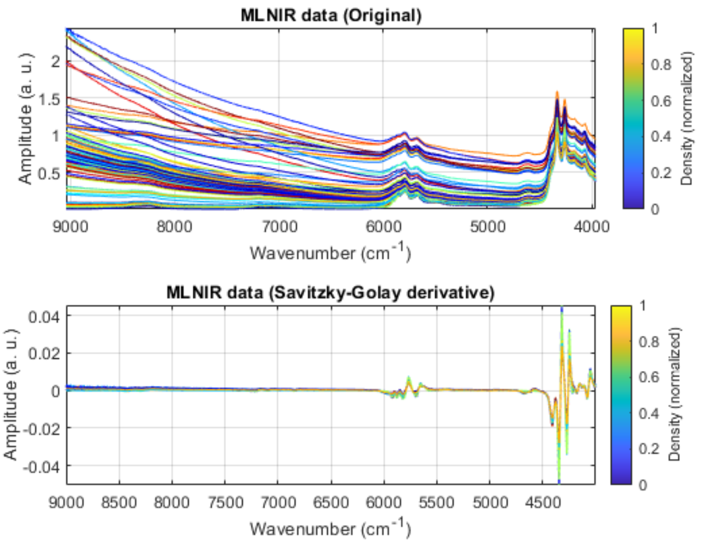

Figure 1: MLNIRdata 1D original (top) and (Savitzky-Golay) differentiated (bottom) of 208 near-infrared (NIR) spectra from hydrocarbons mixtures. Curve color corresponds to the normalized density of the hydrocarbon mixtures (displayed in the colorbar on the right).

1 MLNIRdata data description

The MLNIRdata dataset, depicted in Figure1, stems from the thesis of Jérémy Laxalde [Lax12]. The dataset proposed here results, after curation, from the analysis of $N = 208$ hydrocarbon samples. For each sample n, we provide a Near-Infrared (NIR or near-IR) spectrum (a 1D or vector signal, X_n) — corresponding to intensities measured at given wavenumbers — and a macroscopic property (a scalar, y_n) — density — obtained with the same apparatus and conditions for all samples.

[*] IFP Energies nouvelles, France. Corresponding author: laurent.duval@ifpen.fr

The 208 NIR data matrix and densities have been used to predict y for X, for instance using "Partial Least Squares" (PLS) or sparse PLS in [ADEH+23]. A more detailed data paper dedicated to MLNIRdata is being submitted [DALC25].

The original data consists in 208 1D signals, described by their NIR spectra X made of 2635 digital samples, whose amplitude is represented in homogeneous arbitrary units (abbreviated as a u). The x-axis is a discrete 1D vector of wavenumbers, uniformly sampled, whose units are given in cm^{-1}, and represented as reversed, see Figure 1-top. In infrared spectroscopy, low wavenumber corresponds to low energy bonds, while higher wavenumbers correspond to high energy bonds. For the same hydrocarbon samples, a global physico-chemical property named density y was acquired. A classical challenge in chemometrics is whether one may predict density from spectra using statistical models, and if one is able to identify which wavenumber ranges contribute the most to the predicted property. For MLNIRdata, the density property is a scalar, which has been normalized to the (unitless) interval 0 to 1. This normalization should not affect prediction performance.

Infrared spectroscopy signals are often preprocessed to reduce linear disturbances or non-linear artifacts. This step also often enhances the diversity between similar spectra. One common preprocessing consists in computing the first or second derivative of the data. To avoid noise amplification, it is sometimes coupled with smoothing. One such method is commonly known as Savitzky-Golay filtering **(S.-G.)** [SG64]. MLNIRdata contains a digital derivative of the original signals, characterized an order of 1 (derivative), a polynomial approximation of degree 2 and a length of 15. The resulting gradient signals have been trimmed to both ends, to yield another set of 2594-sample discrete signals. They are depicted in Figure 1-bottom, with the same color scheme as for the density property.

2 MLNIRdata files description

MLNIRdata is made available as ASCII CSV files, as described in Table 1. Spectra and their derivatives are given in `MLNIR_matrixX_NirSpectrumData.csv` and `MLNIR_matrixX_NirSpectrumDerivative.csv`, respectively. Their x-axis in wavenumber are in `MLNIR_matrixX_NirSpectrumDataAxis.csv` and `MLNIR_matrixX_NirSpectrumDerivativeAxis.csv`. The density is stored in `MLNIR_matrixY_NirPropertyDensityNormalized.csv`. Such files may easily be read from standard data analysis software, like R, Python or Julia. For Matlab users, the binary file `MLNIRdata_matrixXY_NirSpectrum_DensityNormalized.mat` contains all of the above five arrays. The Matlab script `MLNIRdata_Display.m` reproduces Figure 1.

Name	Type	Dimension
`MLNIR_matrixX_NirSpectrumData.csv`	CSV	2635×208
`MLNIR_matrixX_NirSpectrumDataAxis.csv`	CSV	2635×1
`MLNIR_matrixX_NirSpectrumDerivative.csv`	CSV	2594×208
`MLNIR_matrixX_NirSpectrumDerivativeAxis.csv`	CSV	2594×1
`MLNIR_matrixY_NirPropertyDensityNormalized.csv`	CSV	1×208
`MLNIRdata_matrixXY_NirSpectrum_DensityNormalized.mat`	Matlab .mat	N/A
`MLNIRdata_Display.m`	Matlab script	N/A

Table 1: File content description for MLNIRdata: Name, Type, Dimension.

3 MLNIRdata potential use

MLNIRdata may be used for any chemometrics, data science, machine learning, signal processing or artificial intelligence task: denoising, filtering, detrending or baseline removal, spectral enhancement, signal correction, wavelength selection, property prediction, clustering, regression, calibration/validation, learning...

References

[ADEH+23] Louna Alsouki, Laurent Duval, Rami El Haddad, Clément Marteau, and François Wahl. Dual-sPLS: a family of dual sparse partial least squares regressions for feature selection and prediction with tunable sparsity; evaluation on simulated and near-infrared (NIR) data. *Chemometr. Intell. Lab. Syst.*, 237:104813, Jun. 2023.

[DALC25] Laurent Duval, Louna Alsouki, Jérémy Laxalde, and Noémie Caillol. MLNIRdata: Machine learning near-infrared spectroscopy data for property prediction: 208 NIR hydrocarbon spectra with normalized density response. *PREPRINT*, 2025. Submitted, August 2025.

[Lax12] Jérémy Laxalde. *Analyse des produits lourds du pétrole par spectroscopie infrarouge*. PhD thesis, Université de Lille 1, 2012.

[SG64] 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.