Spectra Preprocessing

Chemometrics for Spectroscopists

Intensive Course Kraków
2021-11-29 – 12-03

Claudia Beleites
Chemometric Consulting Claudia Beleites

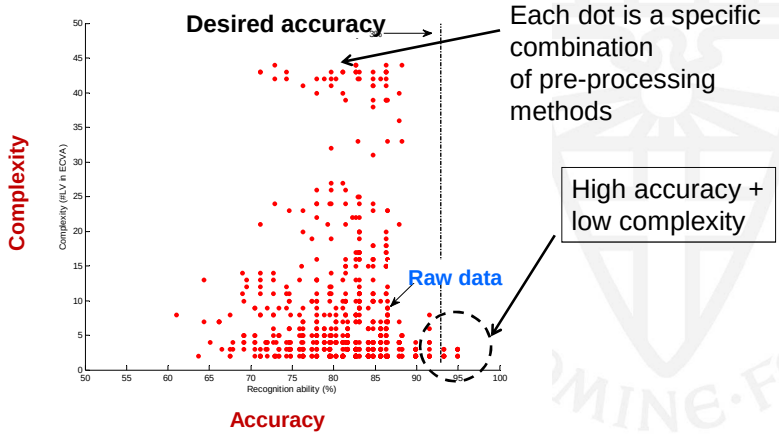
Preprocessing

is part of model training

- Pre-processing needs to take into account both
  - the raw data and
  - the modeling to follow
- Pre-processing enhances signal-to-noise-ratio (SNR)
  - signal is wrt. the task at hand
  - noise is wrt. the task at hand
- Pre-processing removes confounding effects on top of the signal

〰 think about instrumental, physical, chemical, biological reason and how to model/remove that

Preprocessing: A Case Study



Courtesy Prof. Buydens – J. Engel et al., Trends in Analytical chemistry 50 (2013) 96–106.

Baseline correction

- Absorption (transmission): losses due to stray light
  - “white loss” ⇝ offset
  - e.g. NIR λ-dependency ⇝ 2nd grade polynomial
  - Mie scattering: more sophisticated modelling, see e.g. Bassan *et al.*: Resonant Mie Scattering (RMieS) correction of infrared spectra from highly scattering biological samples, Analyst, 2010, 135, 268–277.
- Absorption or emission due to solvent, cuvette material, optical materials, matrix composition
  ⇝ background spectrum
- Raman: stray Rayleigh-scattered photons in the spectrograph
- Raman: stray light from outside
  ⇝ background spectrum

Baseline correction

- Derivatives
  - enhance position information of band
  - inflate noise: need very good SNR to start with
  - ⇝ smoothing derivatives, e.g. Savitzky-Golay filter
- ✔ Polynomials
  - variety of methods to get support points
  - piecewise polynomials
- ✔ Extended Multiplicative Signal Correction (EMSC)
  - chemometric model of background and signal
- lots of other heuristics
- Chemometric model: may not need baseline correction

Normalization

- correct unwanted intensity changes:
- transflecion in MIR: standing waves with $\lambda \approx d$
- Raman: excitation intensity changes
- optical path length in sample changes
- optical properties of sample chamber change (microfluidic chip)
- sample thickness changes
- micro-spectroscopy: focus changes
  chromatic aberration: can be wavelength dependent
- Raman ν^4-dependency

Normalization

- min-max-normalization:
  - $I_c(\lambda) = I(\lambda) \cdot \frac{1}{max(I(\lambda)) - min(I(\lambda))}$
  - chemically meaningful: choose specific band
  - needs very good SNR to start with

✔ area normalization
  - $I_c(\lambda) = I(\lambda) \cdot \frac{1}{\sum I(\lambda)}$
  - chemically meaningful: choose specific band
  - OK with low SNR
  - baseline correction prerequisite

✔ internal standard

✔ EMSC

Normalization

- SNV
  - subtract mean: $I_{c1}(\lambda) = I(\lambda) - \bar{I}(\lambda)$
    then divide by std.: $I_c(\lambda) = I_{c1}(\lambda)/s(I_{c1}(\lambda))$
  - careful: can *cause* non-linearities in the data
  - if at all sensible, then only if all spectra should be very similar

- vector normalization
  - $I_c(\lambda) = I(\lambda) \cdot \frac{1}{\sqrt{\sum I^2(\lambda_i)}}$
  - projects onto spherical surface

  ⇝ use only with models that work with angle between spectra

Raman: Cosmic Ray Spikes

- take very different shapes depending on spectrograph configuration, e.g.
  - Renishaw moving grating: sawtooth
  - Witec moving stage: always 2 spectra affected
  - Kaiser close to textbook, but typically several pixels affected
  - Instrument-side processing of spectra influences shape

✔ repeated spectra
  - detect differences
  - exclude spiky region or whole repetition with spike

- detection within single spectrum difficult, needs strong assumptions about system
- detecting that a spike occured is simple
- difficult: detecting which spectral range exactly is affected

✔ pragmatic: detect spike ⇝ exclude spectrum

Outliers

Outlier Removal Rules

- Do what you like in **training**,
- but **validate** with excluded spectra.
  Can be done as separate subtask of validation.

Centering and Scaling

- . . . are typical “statistical” preprocessing methods
- work on the columns (wavelengths), not on the spectra

- **centering**: subtract mean spectrum
  - poor man’s substitute for baseline correction
  - can help numerically
  - part of some methods
- **scaling**: divide each column by factor
  - typically divide by standard deviation
  - can help numerically
  - part of some methods
  - ✘ rarely adequate for spectra: scales up noise of wavelengths with baseline intensity

Smoothing Interpolation

- Smoothing in itself rarely sensible
- ✔ better combined with downsampling
  ⇝ dimensionality reduction
- correct for shift in wavelength axis of instrument
- interferogram: smoothing downsamplig during FT
- smoothing derivative to extract band position
  e.g. Savitzky-Golay filter
- cut spectral range
  ⇝ dimensionality reduction

Dimensionality reduction & Projecton methods

- …are really chemometric models
- EMSC
- PCA, PLS

Summary

Do what is physically, chemically, spectroscopically sensible