XIX CAC 2024
CHEMOMETRICS IN ANALYTICAL CHEMISTRY
Santa Fe, Argentina
9th - 12th september 2024

SPECTRAL MEASUREMENT ERRORS:
A BASIC TOOL FOR ANALYTICAL METHOD OPTIMISATION

UNIVERSITAT ROVIRA i VIRGILI

chemosens@urv.cat
www.chemosens.recerca.urv.cat

Jokin Ezenarro[1], Daniel Schorn-García[1,2], Olga Busto[1], Ricard Boqué[1]

[1] Universitat Rovira i Virgili. ChemoSens group, QAQO department, C/Marcel·lí Domingo 1, Tarragona, 43007, Spain
[2] University of Stellenbosch, SAGWRI, Department Viticulture and Oenology, Stellenbosch, 7602, South Africa

Introduction

Near-infrared (NIR) spectroscopy is increasingly utilized in industries such as cannabis production for its ability to perform rapid, non-destructive analysis. However, the accuracy of NIR data is often challenged by measurement errors, which can arise from instrument variability, environmental conditions, and sample preparation inconsistencies. These errors can significantly impact the reliability of the spectral data and the subsequent models used for predicting key properties [1,2].

Addressing these measurement errors is essential for improving the quality of regression models. Proper preprocessing techniques play a critical role in mitigating these errors, ensuring that the data are both accurate and reliable. This study focuses on the implications of measurement errors in NIR spectroscopy and underscores the importance of selecting appropriate preprocessing methods to enhance the performance of predictive models, particularly in the context of cannabis analysis [3].

PLS 'iid' assumption → Independent errors
'iid' assumption → Homoscedastic errors

Aim of study

To investigate the impact of measurement errors in NIR spectroscopy on the accuracy and reliability of predictive models, particularly in the analysis of cannabis samples. Also to develop and propose a metric to quantitatively assess and compare the effectiveness of different preprocessing techniques in mitigating these errors. By introducing this index, the study seeks to enhance the robustness and precision of NIR-based analytical methods, ensuring more reliable outcomes in cannabis analysis.

To propose a way to easily take into account the spectral measurement errors into the preprocessing and modelling.

Materials and Methods

15 NeoSpectra NIR spectrometers (Si-Ware Systems) and 135 virtual (simulated) instruments were used to acquire spectral data from **cannabis** samples, measuring in the 1351–2559 nm range. The instruments were calibrated with a Spectralon® blank standard, and measurements were taken with a 5-second scan, **6 replicates**. **31 samples**, originally measured for cannabinoid content prediction models, were selected based on their availability across multiple instruments but representing the whole cannabinoid concentration range.

- Data analysis was conducted using MATLAB R2022b and ProSpecTool v1.0 [4].
- Various spectral **preprocessing** techniques, including smoothing, scattering corrections, baseline corrections, derivatives, and their combinations, were tested: **680 datasets**.
- **Error covariance and correlation matrices (Σ)** were computed to assess measurement errors [1], and a new metric, the Integral Error Correlation Index (IECI), was introduced to evaluate preprocessing effectiveness.
- **Partial Least Squares Regression (PLSR)** models were built and validated using 5-fold cross-validation. Model dimensionality was selected based on the J-Score [5].

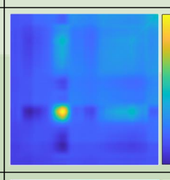
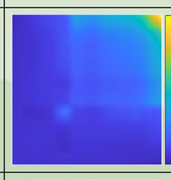
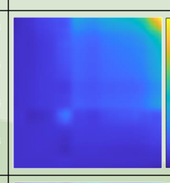
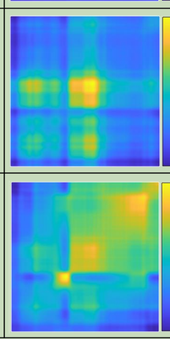
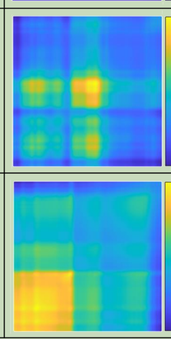
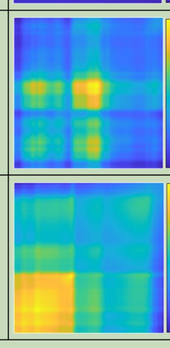
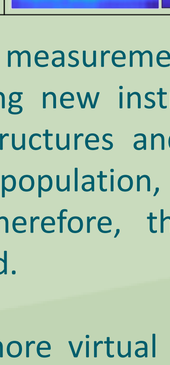
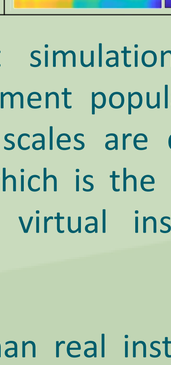
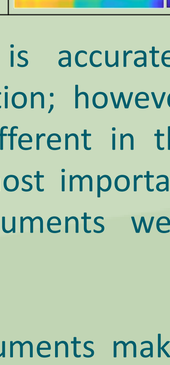
$$\boldsymbol{\Sigma}_{cov} = \frac{1}{(n-1)} \sum_{k=1}^{n} (\mathbf{X}_k - \overline{\mathbf{X}})^{\mathrm{T}} (\mathbf{X}_k - \overline{\mathbf{X}})$$

$$IECI = \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} \boldsymbol{\Sigma}_{\mathbf{corr}}(i,j) - \sum_{i=1}^{N} \boldsymbol{\Sigma}_{\mathbf{corr}}(i,i)}{N^2 - N}$$

$$J\text{-}Score = \left(\frac{RMSE_{CV}}{s_Y} + 1 - \frac{RMSE_{Cal}}{RMSE_{CV}} + NI_{RV}\right)/3$$

Results

Error structure exploration

Error covariance matrices	Real instruments	Virtual instruments	All instruments
Between instrumental replicates			
Between instruments			
Between samples			

Virtual measurement simulation is accurately modelling new instrument population; however, error structures and scales are different in the sample population, which is the most important part. Therefore, the virtual instruments were removed.

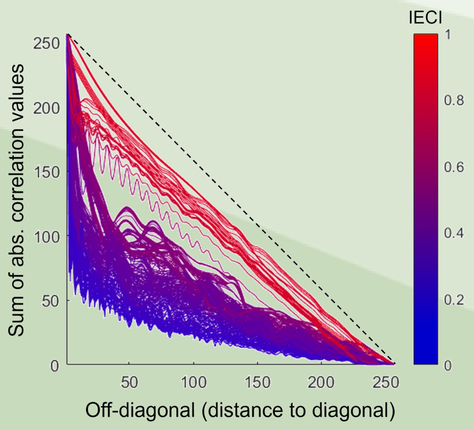
Using more virtual than real instruments makes the final error structures represent mostly the virtual ones, producing models that may not adequately adapt to new real instruments.

Error structure exploration may be primordial for assessing the adequacy of spectra simulation algorithms.

Independent errors

1) Convert Σ_{cov} to Σ_{corr}

2) Calculate their diagonal profile:

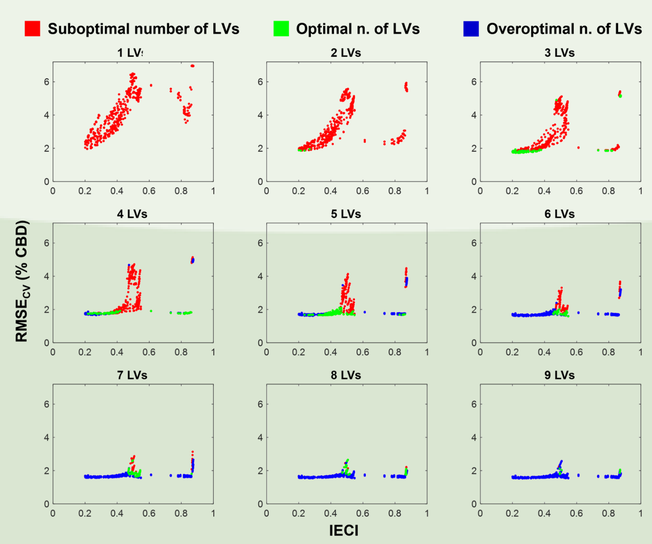


IECI
Integral Error Correlation Index

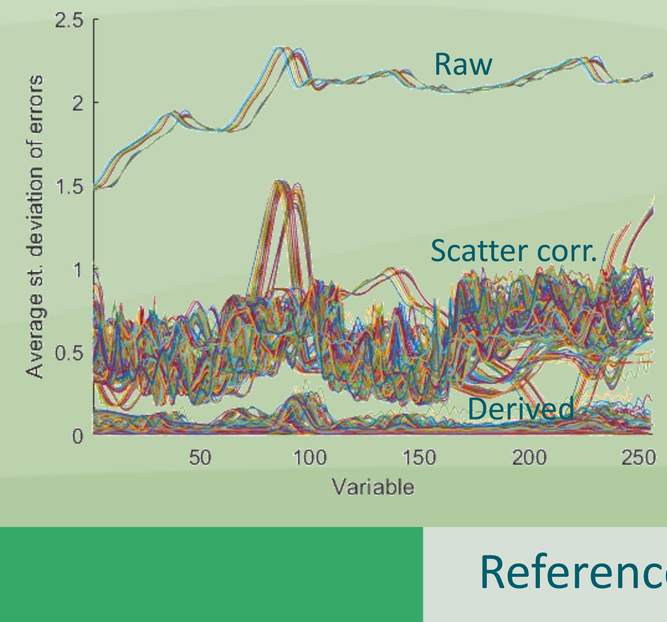
The ratio between the area under the curve and the area under the diagonal (full correlation).

Equivalent to the average absolute correlation coefficient.

The IECI is indicative of the subsequent PLS model complexity, allowing for the screening of different spectral preprocessing methods. The preprocessings that show lowest IECI should be tried in modelling.

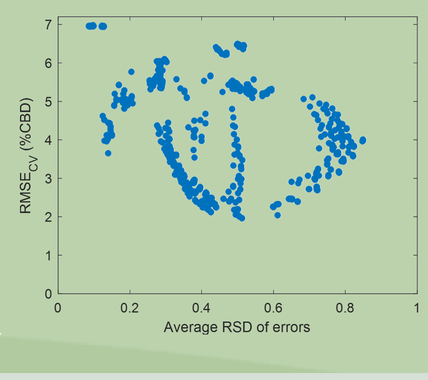


Homoscedastic errors



The errors seem to be somehow homoscedastic, as they have the same magnitude across all wavelengths. This is expected, as the measurements of the different channels in a portable NIR have an equivalent nature unlike other techniques or data sources.

Preprocessing does not seem to change this characteristic significantly, nor it seems to be directly related to the subsequent PLS model performance. It should be further studied.



Conclusions

- Virtual measurements must be carefully handled, explored and validated before including them in any model.
- Measurement errors in NIR spectroscopy can significantly impact the accuracy of analytical models.
- These errors play a major role when there are big sources of variability such as multiple instruments, even if they are of the same model.
- Spectral preprocessing methods play a crucial role in minimising errors and improving model performance; careful selection and evaluation of these methods are essential.
- The newly developed IECI is a valuable tool for quantifying error correlations and can be used to explore the effectiveness of preprocessing methods.

Error covariance and correlation matrices are a good tool to have in the model development kit.

References

[1] Wentzell (2014), Measurement errors in multivariate chemical data. *J Braz Chem Soc*. DOI: 10.5935/0103-5053.20130293

[2] Allegrini and Olivieri (2017), Recent advances in analytical figures of merit: Heteroscedasticity strikes back. *Anal Methods*. DOI: 10.1039/c6ay02916g

[3] Ezenarro *et al*. (2024), Measurement errors and implications for preprocessing in miniaturised near-infrared spectrometers: Classification of sweet and bitter almonds as a case of study. *Talanta*. DOI: j.talanta.2024.126271

[4] Ezenarro *et al*. (2024), ProSpecTool: A MATLAB toolbox for spectral preprocessing selection. ChemoLab. DOI: 10.1016/j.chemolab.2024.105096

[5] Ezenarro *et al*. (2023), J-Score: A new joint parameter for PLSR model performance evaluation of spectroscopic data. *ChemoLab*. DOI: 10.1016/j.chemolab.2023.104883

Acknowledgements

We thank Valenveras S.L. for providing the data.

Grant URV Martí i Franqués – Banco Santander (2021PMF-BS-12).

Chemometrics and Sensorics for Analytical Solutions (CHEMOSENS, ref.2021 SGR 00705, Departament de Recerca i Universitats, Generalitat de Catalunya).

Interesting links