Exploring the capability of generative AI in chemometrics. Towards a natural language programming?

XIX CAC 2024
CHEMOMETRICS IN ANALYTICAL CHEMISTRY

Rocha-de-Oliveira R[1], Mir-Cerdà A[1,2], Gómez-Sánchez A[3]

[1]Chemometrics Group, Universitat de Barcelona, Barcelona, Spain
[2]INSA, Universitat de Barcelona, Santa Coloma de Gramenet, Spain
[3]LASIRE (UMR 8516), CNRS, Université de Lille, Lille, France

Hey, chat! Can you help me?

The integration of large language models (LLMs), specifically Generative Pre-trained Transformer (GPT) models, has the potential to **revolutionize the field of chemometrics**. This transformation extends to **how chemometrics are taught**, enables **faster generation of scripts**, and allowing users to independently **create and interact with chemometric tools**, significantly changing the landscape of chemometric education and practice.

Originally developed by OpenAI, GPT models have showcased remarkable abilities in various **natural language processing** (NLP) tasks, including text generation and translation. NLP refers to the ability of GPTs to **interpret, manipulate, and comprehend human language**.

We preliminarily examined how GPT-4o, the latest GPT version released by OpenAI in May 2024, can enhance different stages of chemometric analysis, including data visualization, coding, and complex data interpretation, by leveraging its advanced NLP capabilities. Typical univariate problems were posed to GPT-4o, including the identification of appropriate statistical tests, creating user-friendly interfaces for data preprocessing, such as asymmetric least squares (AsLS), Savitzky-Golay derivative, and a combination of binning and AsLS, and the analysis of hyperspectral image datasets using principal component analysis. **You can see the full conversation by scanning the QR codes of each example.**

Solving univariate problems

t-test

User: Are these two samples the same?

sampl1repl1	818.1478
sampl1repl2	622.0344
sampl1repl3	873.8661
samp2repl1	727.73
samp2repl2	635.6827
samp2repl3	757.3943

GPT: Here are the results comparing the two samples:
Sample 1: Mean: 771.35 Standard Deviation: 132.28
Sample 2: Mean: 706.94 Standard Deviation: 63.46
The t-test results show: t-statistic: 0.76 p-value: 0.489

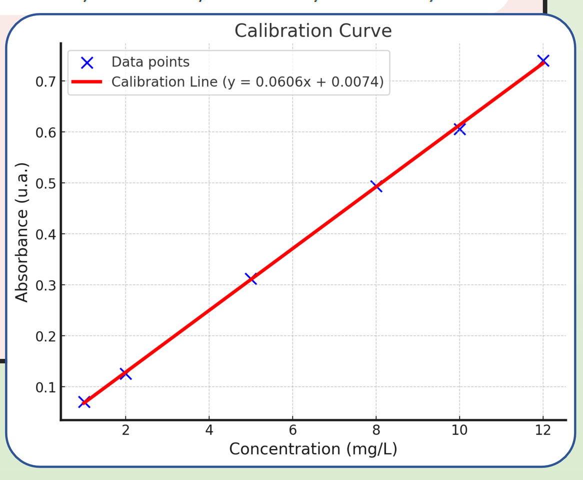
"Since the p-value is quite high (greater than 0.05), this suggests that there is no statistically significant difference between the two samples, meaning they are likely not different in a statistically meaningful way."

Univariate regression

User: Hey, give me the calibration function and the scatter plot

Concentració (mg/L)	1,0	2,0	5,0	8,0	10,0	12,0
Absorbància (u.a.)	0,0702	0,1259	0,3125	0,4936	0,6058	0,7401

GPT: The calibration function derived from the data is: $y = 0.0606x + 0.0074$
The scatter plot with the calibration line is also displayed. Let me know if you need further analysis!



- The GPT model **accurately performed significance tests and univariate calibration**.
- A statistical comparison is not always done with a single prompt if not clearly stated.
- New prompts can be added to redesign the plot, add new data or numeric results.
- However, **be careful, too many prompts may lead to unwanted results** due to mixing up different steps.

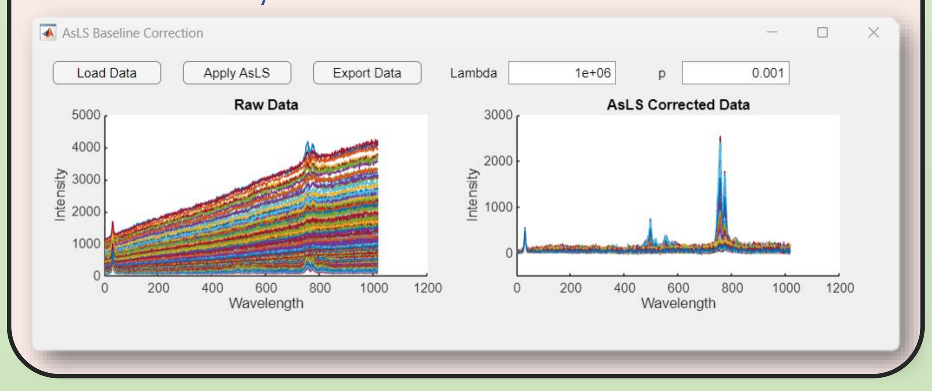
Creating user-friendly interfaces

GPT MATLAB

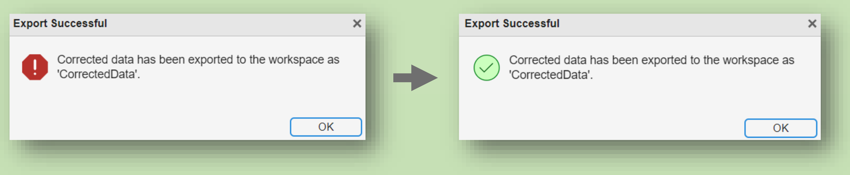
AsLS interface

User: Create an interface for AsLS baseline correction of spectroscopic data with dual-pane display, adjustable parameters, and user-friendly controls.

GPT: Certainly! Here it is:



- The GPT model **successfully provided a MATLAB code with a functional interface for AsLS in one shot**.
- The export button was initially missing, and posteriorly successfully implemented.
- The redesign of the exported data confirmation dialogue was requested and successfully implemented.

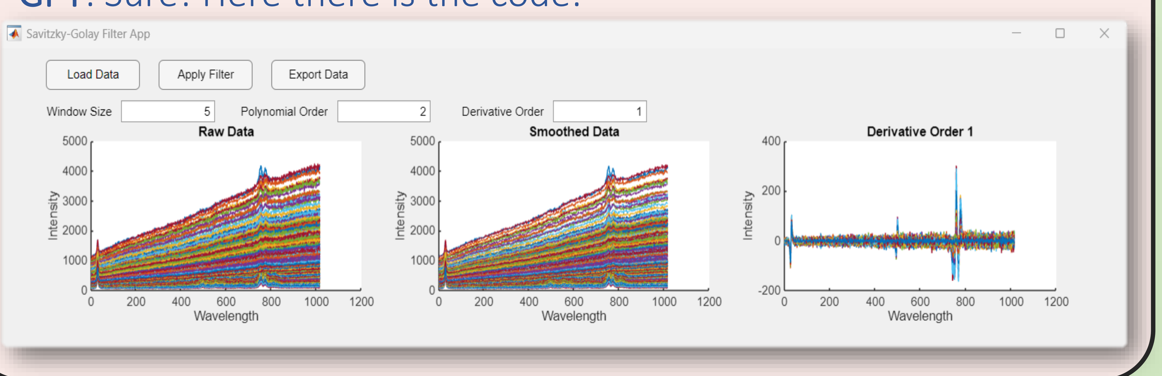


Savitsky-Golay filter interface

GPT MATLAB

User: Create an interface for Savitzky-Golay filtering with triple-plane display, adjustable parameters, and user-friendly controls.

GPT: Sure! Here there is the code:



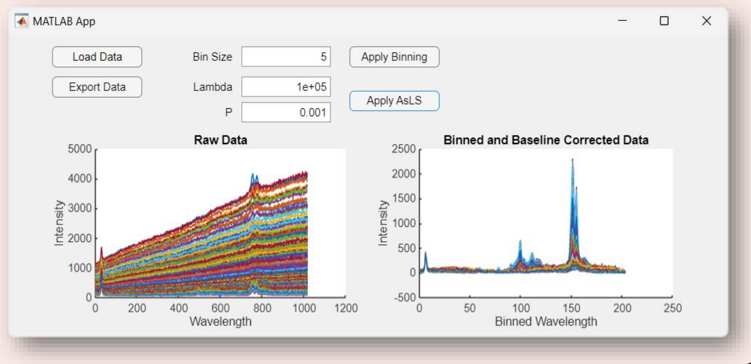
- The GPT model **successfully provided MATLAB code with a functional interface for Savitzky-Golay filtering in one shot**, even considering the necessity for the window size to be odd without an explicit requirement.
- An additional request was made to add a derivative option to the interface.
- The initial modification included errors in handling the dimensions of the data after the derivative and a misunderstanding in applying the convolution coefficients.
- After several interactions **under supervision**, the GPT model identified and fixed the errors, corrected the tail effects, and provided a fully functional and interactive interface.

Binning + AsLS

GPT MATLAB

User: Create an interface with binning and AsLS baseline correction, featuring parameter adjustment and dual-pane visualization.

GPT: Here is the interface, including both preprocessing!



- The GPT model **successfully provided MATLAB code with a functional interface for applying both spectral binning and AsLS in one shot**, though the initial code was very slow.
- When asked to improve efficiency, an error related to AsLS was introduced.
- **Under supervision**, GPT focused on and fixed the error, resulting in a fully functional and interactive interface that includes both preprocessing steps.

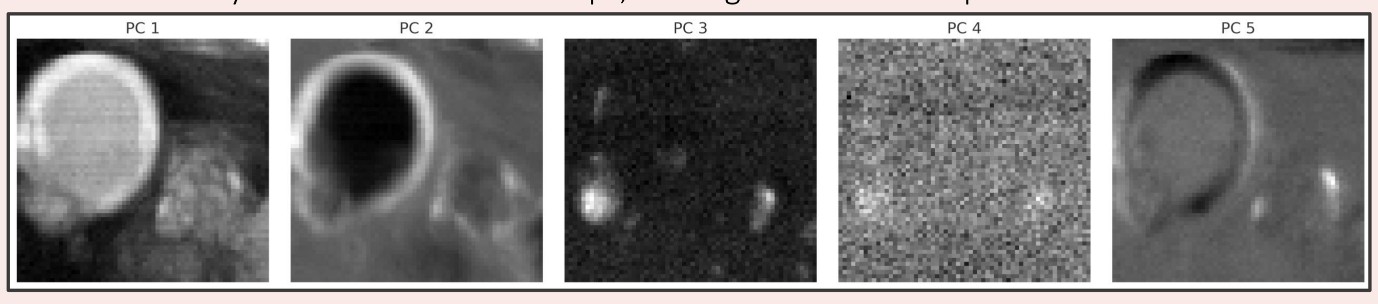
Testing the GPT ability to analyze multivariate data

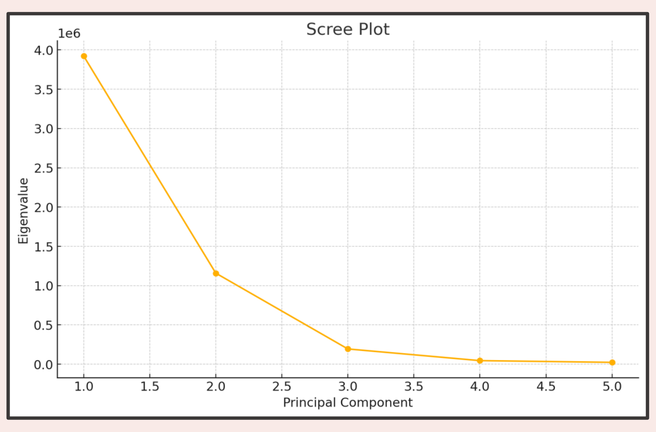
GPT PCA HSI

The GPT model can analyze data using Python and Jupyter Notebooks, leveraging libraries such as Pandas, NumPy, and Matplotlib for data manipulation, statistical analysis, and visualization.
We tested this capability by asking GPT to perform a PCA analysis on the classical Raman hyperspectral image of the oil-water emulsion.

User: Hey! Could you analyze this data set using PCA?

GPT: Yes! Here you can see the score maps, loadings and the scree plot:





The GPT model was able to:

- PCA calculation:

Correct calculation of the PCA components by executing in their internal server a self-generated code in Python.

- Scree Plot Analysis:

Correct identification of the elbow point and significant PCs.

- Loadings Analysis:

Correct identification of key Raman shifts but only after supervision.

Correct identification of vibrational modes of each Raman peak.

The GPT model failed in:

- Scores Analysis:

General answers for the interpretation of the scores. The GPT did not analyze the scores, but gave general indications.

- Peak Detection Loading Analysis:

Failure to consider negative peaks, missing significant loadings features. Needed for explicitly point the problem.

Access the open Github repository with the MATLAB interfaces developed in this work. Try yourself and let us know ;)

Conclusions

The integration of GPT models into chemometrics **has shown transformative potential**. ChatGPT demonstrates an impressive ability to interact naturally with users, effectively solving univariate problems while providing correct and comprehensive explanations. This natural language processing capability makes it **a valuable tool for education and practical applications in chemometrics**. However, they clearly **still require professional supervision** for accurate interpretation to ensure comprehensive and reliable results.

Despite this, there is **huge potential for GPT models to enhance productivity and capacity in chemometrics**. As competition in the field grows with new models like **Anthropic's Claude 3.5 Sonnet**, we can expect continuous advancements in natural language model technologies, which will likely transform how chemometric tools are taught, learned, applied, and developed.

Université de Lille
Universitat de Barcelona
INSA UB
Excelencia María de Maeztu 2023-2026

A.G.S. and R.R.O. acknowledge financial support from the Spanish government Project PID2023-146465NB-I00. A.M.C. acknowledge financial support from Unidad de Excelencia INSA-UB María de Maeztu (Subvención CEX2021-001234-M) funded by MICIU/AEI and EU ERDF.