A Crash Course in Calibration Model Development

Barry M. Wise and Robert T. Roginski
Eigenvector Research, Inc.

EIGENVECTOR RESEARCH INCORPORATED

Contact Information

Eigenvector Research, Inc.
196 Hyacinth Road
Manson, WA 98831 USA
web: www.eigenvector.com

Barry M. Wise
e-mail: bmw@eigenvector.com
phone: 509-662-9213

Robert T. "Bob" Roginksi
e-mail: roginski@eigenvector.com

EIGENVECTOR RESEARCH INCORPORATED

Table of Contents

3

EIGENVECTOR RESEARCH INCORPORATED

Definition of Chemometrics

Chemometrics is the chemical discipline that uses mathematical and statistical methods to
1) relate *measurements* made on a *chemical* system to the *state* of the system
2) design or select optimal *measurement* procedures and experiments.

4

EIGENVECTOR RESEARCH INCORPORATED

What's in a Name?

- Chemometrics
  - Chemo — chemistry, metrics — measurements, good word!
- Artificial Intelligence
  - Theory and development of computer systems able to perform tasks that normally require human intelligence
- Machine Learning
  - Systems able to learn and adapt without following explicit instructions, by using algorithms and statistical models to analyze and draw inferences from patterns in data
- ~~Cheminformatics~~
  - Use of physical chemistry to predict molecular properties
- Data Science
  - Field that uses scientific methods, processes, algorithms and systems to extract knowledge and insight from data
- Chemical Data Science
  - Sounds right to me!

That which we call a rose by any other name would smell as sweet.

-- Romeo and Juliet

But it might not be found with a Google search!

5

https://eigenvector.com/we-used-to-call-it-chemometrics/

EIGENVECTOR RESEARCH INCORPORATED

Multivariate Analysis

Multivariate Statistical Analysis is concerned with data that consists of *multiple measurements* on a number of individuals, objects, or data samples. The measurement and analysis of *dependence between variables* is fundamental to multivariate analysis.

6

EIGENVECTOR RESEARCH INCORPORATED

Developing a Calibration Data Set

EIGENVECTOR RESEARCH INCORPORATED

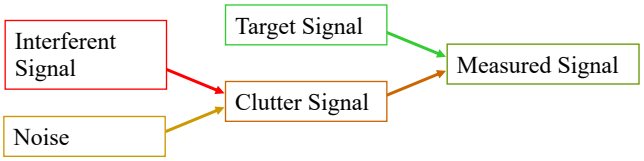
Data Collection: What Samples?

- Machine learning models (including Partial Least Squares regression) require calibration data, aka training or learning data
- The structure of the linear, factor based models like PLS is well suited to modeling spectroscopic data
- Even so
  - Models should be developed using data that covers the range expected during use of the model
  - This includes the analyte of interest but also clutter: any interferents or systematic non-idealities.
  - Models are only reliable within the range of the calibration data!

EIGENVECTOR RESEARCH INCORPORATED

Clutter

- *Clutter* is present in all measurements
  - X-block, Y-block



- The confounding effects of interfering chemical species, physical effects and instrument non-idealities

9

EIGENVECTOR RESEARCH INCORPORATED

Sources of Clutter

- Systematic background variability
  - Variation in chemical *interferents*
    - *Any substance whose presence interferes with an analytical procedure*
  - Physical effects such as scattering due to particles
- Other changes in the system being observed
  - T, P changes, variable sample matrix, "dark current"
- Variance due to physics of instrument
  - e.g., drift, instrument changes, variable baseline or gain
  - Non-linearity, saturation
- Non-systematic random noise
  - homoscedastic, heteroscedastic

10

EIGENVECTOR RESEARCH INCORPORATED

Advantage of Inverse Least Squares Methods

- PLS and other Inverse Least Squares (ILS) regression methods (MLR, PCR, etc.) don't require the concentration of all analytes, including interferents, be known, however…
- *Interferents and other clutter components must vary in the calibration data set for the regression model to be robust against them*
- This has important implications for calibration data collection
- Experimental design?

11

EIGENVECTOR RESEARCH INCORPORATED

How Much Calibration Data?

- 2-4 times the number of factors contributing to variation in the data. This includes
  - Analyte of interest
  - Interferent species
  - Physical effects like scatter, temperature, pressure
- Think hard about this before collecting data!
- Ignore potential variations at your own peril
- Experimental design is great
  - But not always possible to create variations you'd like to include
  - Get samples from "natural" variation

There are known knowns. These are things we know that we know. There are known unknowns. That is to say, there are things that we know we don't know. But there are also unknown unknowns. There are things we don't know we don't know.

--Donald Rumsfeld

12

EIGENVECTOR RESEARCH INCORPORATED

Finally…

- Be sure data is collected under the same conditions you expect when the method is in use
  - Instrument settings
  - Sample presentation
  - Extraneous effects
    - Lighting, temperature, humidity, etc. etc.
  - And what if it isn't?
    - May need to use calibration transfer/model updating methods
    - https://eigenvector.com/events/calibration-model-maintenance/
- Useful to review data as it is being collected to identify problems early

13

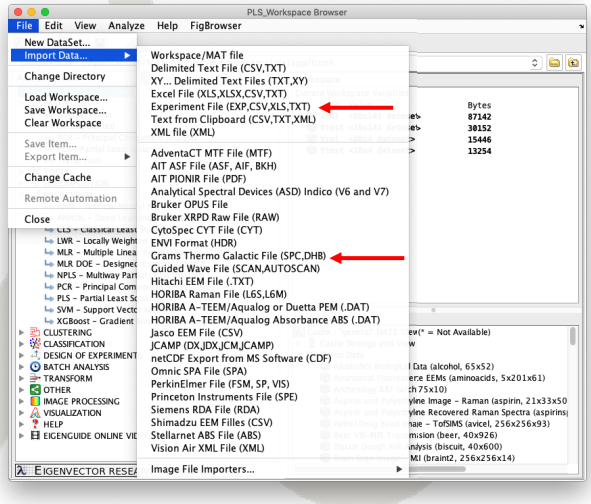
EIGENVECTOR RESEARCH INCORPORATED

Importing Data Into PLS_Toolbox & Solo

14

EIGENVECTOR RESEARCH INCORPORATED

Data Formats Supported



- PLS_Toolbox and Solo can import many types of data directly
- Starting from the Workspace Browser select File/Import Data…
- Experiment File
- Grams Thermo Galactic SPC

15

EIGENVECTOR RESEARCH INCORPORATED

Built-in Importers: Proprietary File Formats

- File format specific to an instrument manufacturer
- Examples:
  - Bruker OPUS files (which use a numeric extensions, .#, .####, etc).
  - Thermo Fisher Omnic files (.SPA)
  - Perkin Elmer (.SP) files
- Always happy to add more if we can get the file spec!

16

EIGENVECTOR RESEARCH INCORPORATED

Built-in Importers: Standardized File Formats

- Commonly used file formats
- Examples:
  - JCAMP (.JCAMP, .JDX, .DX, .JCM)
  - Galactic Thermo (.SPC)
  - netCDF (ANDI-MS)
    - network common data form
    - Common file format for chromatographic data (LC-MS and GC-MS)
    - .cdf file extension

17

EIGENVECTOR RESEARCH INCORPORATED

DataSet Object (DSO)

- Imported data files are DSOs in PLS_Toolbox/Solo
  - Simultaneously imported files in one DSO
- Data and all additional information in single variable
  - Labels, axis scales, class, author, modification dates, description
- Provides consistency tests for axis scales, labels
- Allows easy “classing” and “exclusion” (soft-delete) of samples (or variables)
- Smart concatenation
- User data for any other associated metadata
- History field
- DSOs can be edited directly or graphicly

18

EIGENVECTOR RESEARCH INCORPORATED

Experiment File Reader

- Building a regression model
- Allows reading in a file (.csv, .xlsx, .txt, .exp) that contains:
  - file names (or path)
  - reference values
  - calibration (Cal, C) or validation (Val, V)

```
Filename, Reference Value, Cal/Val
File1.xyz, 1.0, Cal
File2.xyz, 2.0, Cal
File3.xyz, 3.0, Cal
File4.xyz, 4.0, Val
```


EIGENVECTOR RESEARCH INCORPORATED

Experiment File Advantages

- Files are in order specified in Experiment File
  - No worrying about how system alphabetizes names
  - Specify reference (Y) values in file
- Creates matched DSOs for spectra (X) and reference values (Y)
- Nice record of what was done, easy to modify later

EIGENVECTOR RESEARCH INCORPORATED

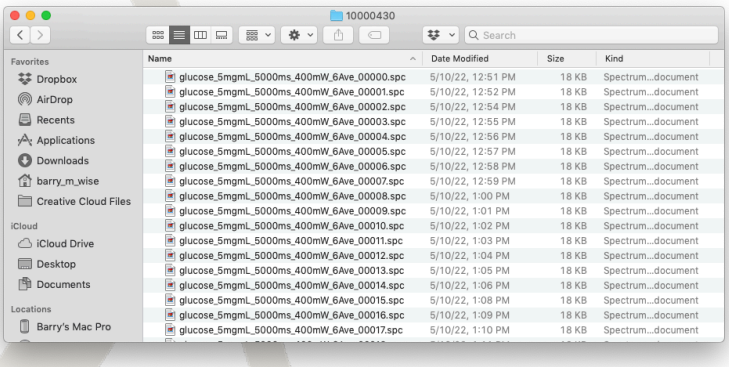
Example Data

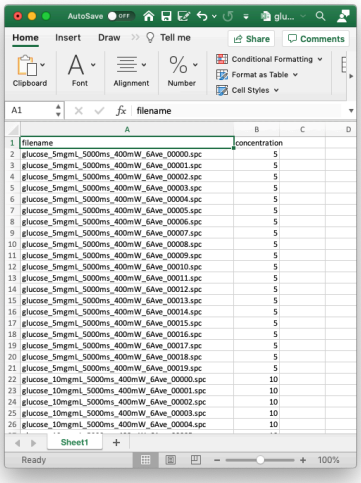
- Raman spectra of glucose in water
  - Feed to a bioreactor
- 120 Samples
  - Glucose range 5-40 mg/ml
- 2048 Variables (channels)
  - Raman shift 26 to 3283 cm^{-1}
- Collected on Thermofisher Ramina
- Thanks to Thermo for the data!

21

EIGENVECTOR RESEARCH INCORPORATED

Glucose Files Example



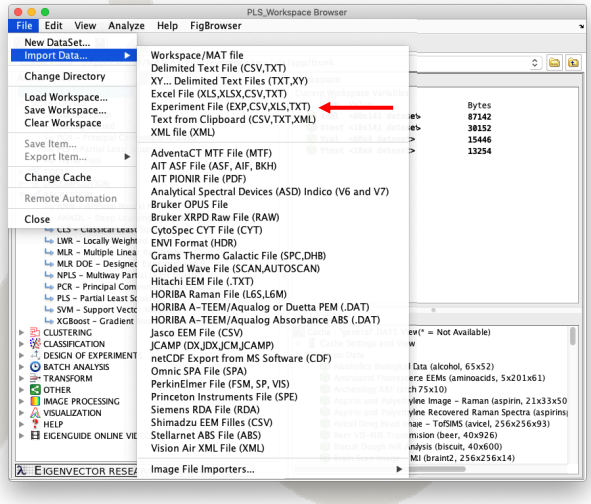


- Copy filenames from folder into Excel spreadsheet
- Add filename and concentration header
- Add concentration values
- Put .csv file into same folder

22

EIGENVECTOR RESEARCH INCORPORATED

Import from Browse Interface



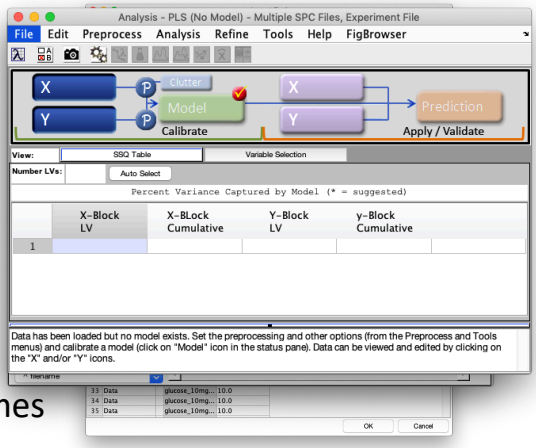
- Starting from the Workspace Browser select File/Import Data…
- Experiment File

23

EIGENVECTOR RESEARCH INCORPORATED

Import Experiment File

- In Workspace Browser select File/Import Data/Experiment Reader
- Navigate to .csv file and select
- Accept defaults in Text Import Settings
- Click OK in Import Tool
- In Experiment File DataSet Editor click ✓
- Data is pushed into Analysis interface
- Save Data into Workspace w/ desired names
- Save Workspace



24

EIGENVECTOR RESEARCH INCORPORATED

File Import Conclusions

- Many file types, many readers!
- If you can get it into Excel, you can get it into PLS_Toolbox/Solo
- Experiment Reader keeps things organized
- For more info see “EVRI-thing You Need to Know About Importing Data into PLS_Toolbox and Solo” on the webinar page at
  - https://eigenvector.com/aiovg_videos/evri-thing-you-need-to-know-about-importing-data-into-pls_toolbox-and-solo/
- And “EVRI-thing You Need to Know About Getting Started with PLS_Toolbox and Solo” at
  - https://eigenvector.com/aiovg_videos/evri-thing-you-need-to-know-to-get-started-with-pls_toolbox-and-solo/

EIGENVECTOR RESEARCH INCORPORATED