A Crash Course in Calibration Model Development

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

©Copyright 1996-2022 Eigenvector Research, Inc. No part of this material may be photocopied or reproduced in any form without prior written consent from Eigenvector Research, Inc.



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



Table of Contents

- Introduction: Chemometrics and Machine Learning
- Developing a calibration data set
- Importing data into PLS_Toolbox/Solo
- Data review: plotting and Principal Components Analysis (PCA)
- Building a Partial Least Squares (PLS) regression model
- Optimizing Models: Preprocessing and Variable Selection
- Saving and exporting models
- Getting models online with Solo_Predictor and Model_Exporter
- Model Maintenance
- Overall Conclusions



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.



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

5

- Field that uses scientific methods, processes, algorithms and systems to extract knowledge and insight from data
- Chemical Data Science
 - Sounds right to me!

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

PLS PARAFAC MLSCA SW ASCA Artificial Intelligence PMF Machine Learning CHEMOMETRICS KNN SIMCA Data Science Process Analytics PCA LDA Chemical Data Science CLS UMAP & PLS-DA CLWR

> 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!



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.



Developing a Calibration Data Set



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!



Clutter

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



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



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



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?



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

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

EIGENVECTOR RESEARCH INCORPORATED

Importing Data Into PLS_Toolbox & Solo



Data Formats Supported

• •	PLS_Workspace Browser	
File Edit View Analyze	Help FigBrowser	
New DataSet		
Import Data >	Workspace/MAT file	
ennenne enner and ennenner a	Delimited Text File (CSV.TXT)	۵ 🖨
Change Directory	XY Delimited Text Files (TXT,XY)	
Load Workspace	Excel File (XLS,XLSX,CSV,TXT)	les T
Save Workspace	Experiment File (EXP.CSV.XLS.TXT)	Bytes
	Text from Clipboard (CSV,TXT,XML)	eb 87142
Clear Workspace	XML file (XML)	eb 30152
Save Item	AIME THE (AIME)	> 15446
Export Item >	AdventaCT MTF File (MTF)	> 13254
Export itemini oo	AIT ASF File (ASF, AIF, BKH)	
Change Cache	AIT PIONIR File (PDF)	
THE RECEIPTION	Analytical Spectral Devices (ASD) Indico (V6 and V7)	
Remote Automation	Bruker OPUS File	
Close ANDL - Deep Level	Bruker XRPD Raw File (RAW)	
Close	CytoSpec CYT File (CYT)	
LWR - Locally Weight	ENVI Format (HDR)	
he MLR - Multiple Linea	Grams Thermo Galactic File (SPC.DHB)	
he MLR DOE - Designed	Guided Wave File (SCAN.AUTOSCAN)	-
here NPLS - Multiway Part	Hitachi EEM File (.TXT)	
I→ PCR - Principal Com		
he PLS - Partial Least St	HORIBA Raman File (L6S,L6M)	
SVM − Support Vecto	HORIBA A-TEEM/Aqualog or Duetta PEM (.DAT)	0
Subsect - Gradient	HORIBA A-TEEM/Aqualog Absorbance ABS (.DAT)	
CLUSTERING	Jasco EEM File (CSV)	ew(* = Not Available)
CLASSIFICATION	JCAMP (DX JDX JCM JCAMP)	VI-W
DESIGN OF EXPERIMENT	netCDF Export from MS Software (CDF)	In the second second
BATCH ANALYSIS	Omnic SPA File (SPA)	Eata (alcohol, 65x52)
TRANSFORM GOTHER	PerkinElmer File (FSM, SP, VIS)	er.e EEMs (aminoacids, 5x201x61) th 75x10)
OTHER IMAGE PROCESSING	Princeton Instruments File (SPE)	vine image - Raman (aspirin, 21x33x)
MAGE PROCESSING VISUALIZATION	Siemens RDA File (RDA)	vine image - Raman (aspirin, 21x33x) vine Recovered Raman Spectra (aspirin
HELP	Shimadzu EEM Filles (CSV)	naje – TofSIMS (avicel, 256x256x93)
EIGENGUIDE ONLINE VIE	Stellarnet ABS File (ABS)	mision (beer, 40x926)
- LIGENGOIDE ONDIRE VIE	Vision Air XML File (XML)	indvsis (biscuit, 40x600)
		MI (braint2, 256x256x14)

• 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



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

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



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



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
```



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



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!

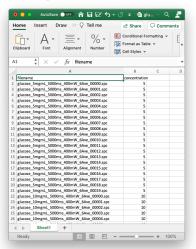




Glucose Files Example

• • • • 10000430				
< >		😻 🖌 Q. Search		
Favorites	Name	Date Modified	Size	Kind
😻 Dropbox	glucose_5mgmL_5000ms_400mW_6Ave_00000.spc	5/10/22, 12:51 PM	18 KB	Spectrumdocument
-	glucose_5mgmL_5000ms_400mW_6Ave_00001.spc	5/10/22, 12:52 PM	18 KB	Spectrumdocument
(initial and initial and initia and initial and initial and initial and initial and ini	glucose_5mgmL_5000ms_400mW_6Ave_00002.spc	5/10/22, 12:54 PM	18 KB	Spectrumdocument
Recents	glucose_5mgmL_5000ms_400mW_6Ave_00003.spc	5/10/22, 12:55 PM	18 KB	Spectrumdocument
Applications	glucose_5mgmL_5000ms_400mW_6Ave_00004.spc	5/10/22, 12:56 PM	18 KB	Spectrumdocument
-	glucose_5mgmL_5000ms_400mW_6Ave_00005.spc	5/10/22, 12:57 PM	18 KB	Spectrumdocument
🕑 Downloads	glucose_5mgmL_5000ms_400mW_6Ave_00006.spc	5/10/22, 12:58 PM	18 KB	Spectrumdocument
😭 barry_m_wise	glucose_5mgmL_5000ms_400mW_6Ave_00007.spc	5/10/22, 12:59 PM	18 KB	Spectrumdocument
Creative Cloud Files	glucose_5mgmL_5000ms_400mW_6Ave_00008.spc	5/10/22, 1:00 PM	18 KB	Spectrumdocument
	glucose_5mgmL_5000ms_400mW_6Ave_00009.spc	5/10/22, 1:01 PM	18 KB	Spectrumdocument
Cloud	glucose_5mgmL_5000ms_400mW_6Ave_00010.spc	5/10/22, 1:02 PM	18 KB	Spectrumdocument
iCloud Drive	glucose_5mgmL_5000ms_400mW_6Ave_00011.spc	5/10/22, 1:03 PM	18 KB	Spectrumdocument
— • •	glucose_5mgmL_5000ms_400mW_6Ave_00012.spc	5/10/22, 1:04 PM	18 KB	Spectrumdocument
Desktop	glucose_5mgmL_5000ms_400mW_6Ave_00013.spc	5/10/22, 1:05 PM	18 KB	Spectrumdocument
Documents	glucose_5mgmL_5000ms_400mW_6Ave_00014.spc	5/10/22, 1:06 PM	18 KB	Spectrumdocument
	glucose_5mgmL_5000ms_400mW_6Ave_00015.spc	5/10/22, 1:08 PM	18 KB	Spectrumdocument
Locations	glucose_5mgmL_5000ms_400mW_6Ave_00016.spc	5/10/22, 1:09 PM	18 KB	Spectrumdocument
Barry's Mac Pro	glucose_5mgmL_5000ms_400mW_6Ave_00017.spc	5/10/22, 1:10 PM	18 KB	Spectrumdocument

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

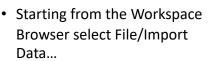




Import from Browse Interface

	PLS_Workspace Browser		
File Edit View Analyze	Help FigBrowser		
New DataSet			
Import Data >	Workspace/MAT file		
current Polder - proppincations	Delimited Text File (CSV.TXT)	۵ 🖴	
Change Directory	XY Delimited Text Files (TXT,XY)		
Load Workspace Save Workspace Clear Workspace	Excel File (XLS,XLSX,CSV,TXT) Experiment File (EXP,CSV,XLS,TXT) Text from Clipboard (CSV,TXT,XML) XML file (XML)	Bytes Bytes 87142 19:e5 30152	
Save Item ►	AdventaCT MTF File (MTF) AIT ASF File (ASF, AIF, BKH)	> 15446 > 13254	
Change Cache	AIT PIONIR File (PDF)		
Remote Automation	Analytical Spectral Devices (ASD) Indico (V6 and V7) Bruker OPUS File		
MIR - Multiple Linea MIR - Multiple Linea MIR DOE - Designer NPLS - Multitway Part PRS - Principal Com PRS - Principal Com Support Vector XOBOST - Gradient CLASSIFICATION SATCH ANALYSS MACE PROCESSING MACE PROCESSING MACE PROCESSING VSUAUZATION Y FILEP	Grams Thermo Galactic File (SPC,DHB) Guided Wave File (SCAN,4UTOSCAN) Hitachi EEM File (ISCAN,4UTOSCAN) HORBA Raman File (LSS,LSM) HORBA A-TEEM/Aqualog or Duetta PEM (.DAT) HORBA A-TEEM/Aqualog Absorbance ABS (.DAT) Jasco EEM File (CSV) JCAMP (OX,BJX,CM)CAMP) netCDF Export from MS software (CDF) Omnic SPA File (SPA) PerkinElmer File (SPA) Princeton Instruments File (SPE) Siemens RDA File (RDA) Simadzu EEM Files (CSV)	ev(* = Not Available) w Dia tackele, 65:52) err EBM (amound, 5:5212451) eh7:5310) yine Image - Raman (Saptirn, 21:33:55) yine Recovered Raman Spectra (25:525:65) are - ToSBM Schull, 25:525:65)	
► EI EIGENGUIDE ONLINE VIE	Stellarnet ABS File (ABS) Vision Air XML File (XML)	mision (beer, 40x926) miysis (biscuit, 40x600) MI (braint2, 256x256x14)	
λ EIGENVECTOR RESEA	Image File Importers		

23

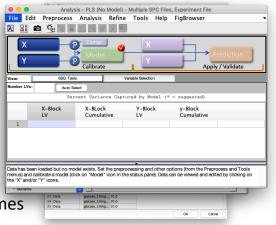


• Experiment File



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



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
 - <u>https://eigenvector.com/aiovg_videos/evri-thing-you-need-to-know-about-importing-data-into-pls_toolbox-and-solo/</u>
- And "EVRI-thing You Need to Know About Getting Started with PLS_Toolbox and Solo" at
 - <u>https://eigenvector.com/aiovg_videos/evri-thing-you-need-to-know-to-get-</u> started-with-pls_toolbox-and-solo/

