Maximising the value of your data – multivariate analysis

Consequences and opportunities of breakthrough in scientific understanding

Ralf Marbach
VTT Optical Instrument Center, www.vtt.fi/oic
ralf.marbach@vtt.fi , Ph. +358 40 770 8375

VTT Optical Instruments Center

- VTT = Finland's main R&D service provider
  - Approx. 2,800 staff ; multidisciplinary
- OIC = One of 45 knowledge centres; **optical measurement technology for industrial applications**
- 50+ staff of which about 50% on PAT
- 20+ years of R&D and Engineering in optical instruments; focus on *on-line* applications
- **Expertise**
  - Spectroscopy and machine vision
  - Optics, electronics, high precision mechanics, software, embedded software, system engineering, optoelectronic components, packaging, digital signal processing, wireless communication, low cost manufacturing techniques, **calibration**, various aspects of applied physics and chemistry, …
- Independent organization, serving both instrument suppliers and end-users

Two Messages …

The good news:

- *The advantages* of "science-based" calibration (SBC)
- Three insights (break-aways from traditional thinking) necessary for scientific understanding
  - Multivariate calibration as simple as intuitive as univariate

The other news:

- *The "ugly" side* of the scientific understanding provided by SBC
  - Exact definitions of SPECIFICITY ("Selectivity") and SENSITIVITY now available in multivariate case
  - How the *existing* methods of calibration are affected
  - **Purpose:** Initiate discussion in the user community

Spectrometry – Background

Measured spectrum [AU]

all interfering spectra

$$\mathbf{x}(t) = y(t)\cdot\mathbf{g} + c_1(t)\cdot\mathbf{k}_1 + c_2(t)\cdot\mathbf{k}_2 + \mathrm{K} + \mathrm{K} + \mathbf{i}_{baseline}(t) + \mathrm{K} + \mathbf{i}_{noise}(t)$$

all instrumental effects

true concentration of analyte of interest [mol/L]

"response spectrum" of analyte of interest [AU/(mol/L)]

History & Status of "Chemometrics"

- "Classical" (or "physical" or "K-matrix") calibration
- Simple cases only

- "Statistical" (or "inverse" or "P-matrix") calibration
- Widely applied (**PLS, PCR** …)

SBC method (1/2) − First mental leap (“Signal or else”)

Measured spectrum [AU]: $\mathbf{x}^T = y \cdot \mathbf{g}^T + \mathbf{x}_n^T$

	"Spectral signal"		"Spectral noise"	
Mean:	$\bar{y} \cdot \mathbf{g}^T$	[AU]	$\bar{\mathbf{x}}_n^T$	[AU]
Std:	$\sigma_y \cdot \mathbf{g}^T$	[AU]	$\boldsymbol{\Sigma}$	[AU^2]

- Interfering spectra **and** electronic noise, sampling noise, …
- $\boldsymbol{\Sigma}$ easy to determine in practice

SBC (2/2) – Optimal solution (for measuring **g** in **Σ**)

$$\mathbf{b}_{opt} = \frac{\boldsymbol{\Sigma}^{-}\mathbf{g}}{\mathbf{g}^{T}\boldsymbol{\Sigma}^{-}\mathbf{g}}$$

[(mol/L) / AU] cmp. Ref. 1

Prediction: $$\hat{y}_{pred}(t) = \left[y(t)\cdot\mathbf{g}^{T} + \mathbf{x}_{n}^{T}(t)\right]\cdot\frac{\boldsymbol{\Sigma}^{-}\mathbf{g}}{\mathbf{g}^{T}\boldsymbol{\Sigma}^{-}\mathbf{g}} = y(t) + \frac{\mathbf{x}_{n}^{T}(t)\cdot\boldsymbol{\Sigma}^{-}\mathbf{g}}{\mathbf{g}^{T}\boldsymbol{\Sigma}^{-}\mathbf{g}}$$

Best possible "prediction" error: $$SEP_{opt} = \sqrt{\frac{1}{\mathbf{g}^{T}\cdot\boldsymbol{\Sigma}^{-}\cdot\mathbf{g}}}$$

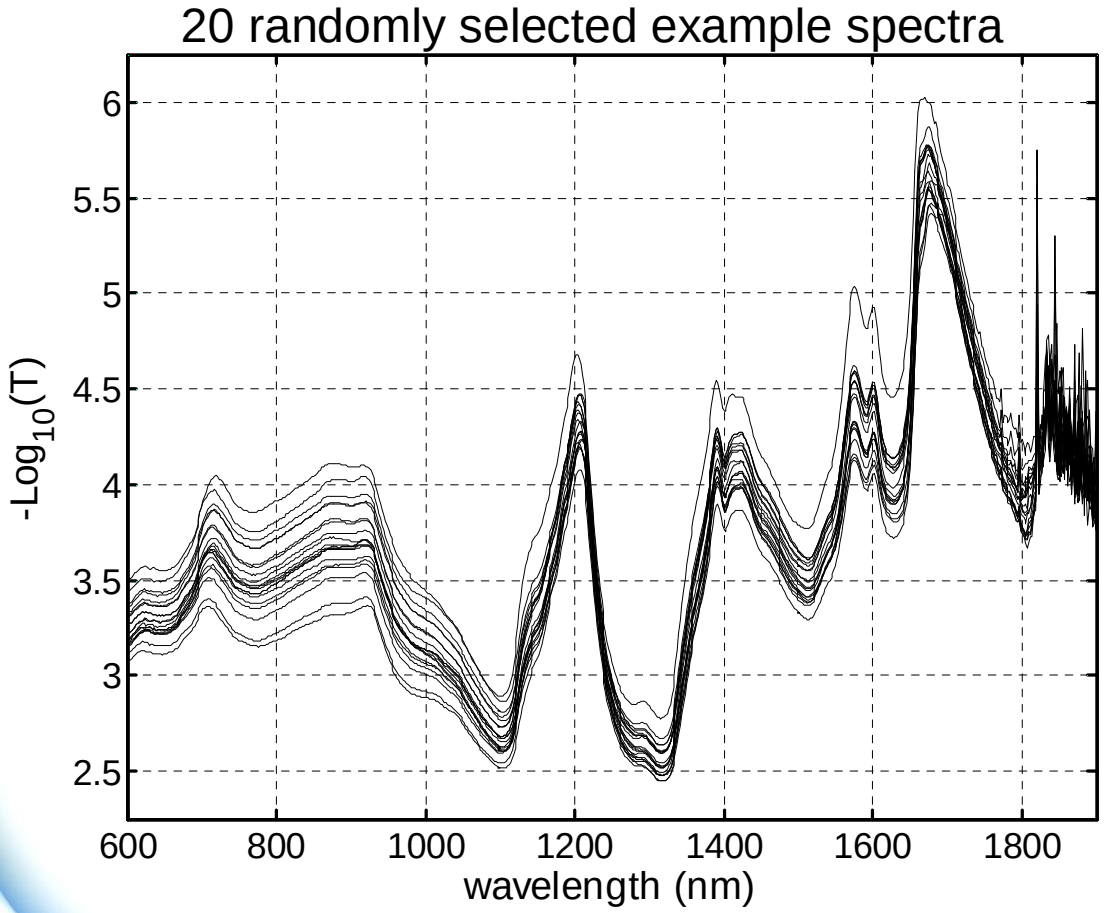
[mol/L] RMS

Multivariate Limit of Sensitivity

SBC Advantages

- Spectrometry becomes "primary" method
  - Need for lab-reference values virtually eliminated (**$$**)
  - Need for “designer samples” eliminated, i.e. stable industrial processes with minimal amount of analyte variation can also be calibrated (**$$**)
  - Development of new, application-specific analyzers much faster, less risky because hardware spec's can be derived in advance (**$$**)
  - Specificity of response can be proven **(!)**
  - Improved possibilities for calibration transfer
  - Improved possibilities for dealing with non-linearities, instationarities
  - Calibration process transparent and communicate-able ("science-based")
  - …

Chambersburg *Shoot-out 2002* Example – API in tablets

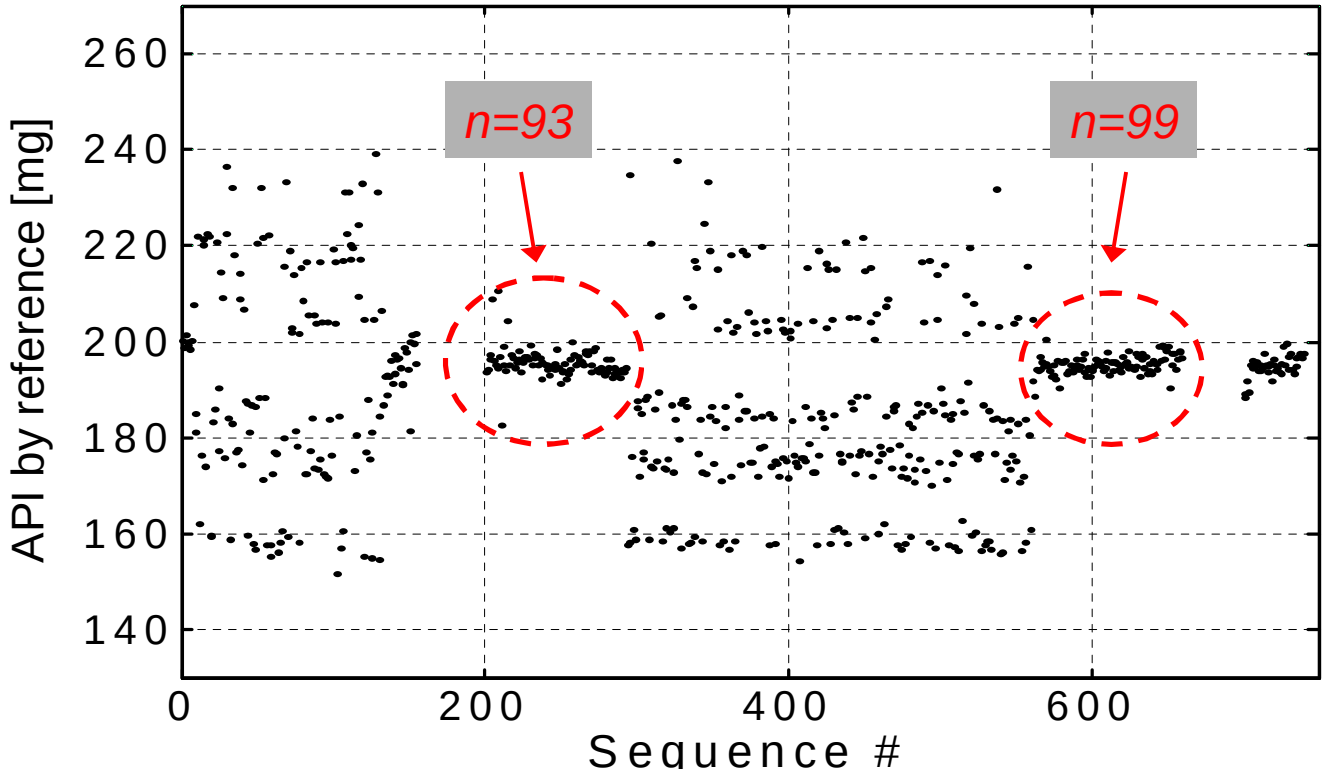


- 655 tablets (155 cal, 460 test, 40 val)
- NIR diffuse transmittance
- Foss NIRSystems Multitab Analyzer (2 units)
- data provided courtesy of Purdue Pharma L.P and Gary Ritchie
- data available at www.idrc-chambersburg.org/shootout_2002.htm

- G.E. Ritchie; R.W. Roller; E.W. Ciurczak; H. Mark; C. Tso; S.A. MacDonald, *Validation of a near-infrared transmission spectroscopic procedure - Part B: Application to alternate content uniformity and release assay methods for pharmaceutical solid dosage form*, J. Pharm. Biomed. Anal. 29, 159-171 (2002)

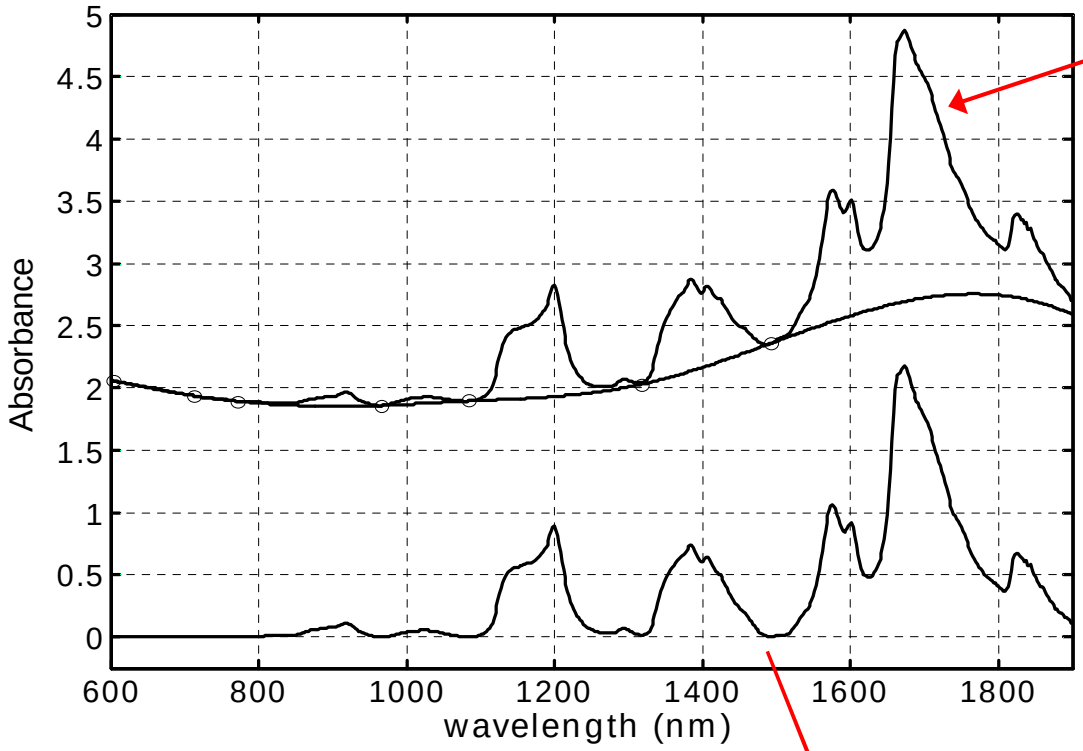
Shoot-out 2002 Example – Lab-reference values

Calibration first (n=155), then *Test* (n=460), then *Validation* (n=40)



- (two instruments:) *2x(93+99) = 384* "noise" spectra; $\boldsymbol{\Sigma} \cong \frac{\tilde{\mathbf{X}}_{384}^{T}\tilde{\mathbf{X}}_{384}}{384-1}$
- Inversion using first 20 "PCR factors"

Shoot-out 2002 Example – API response spectrum



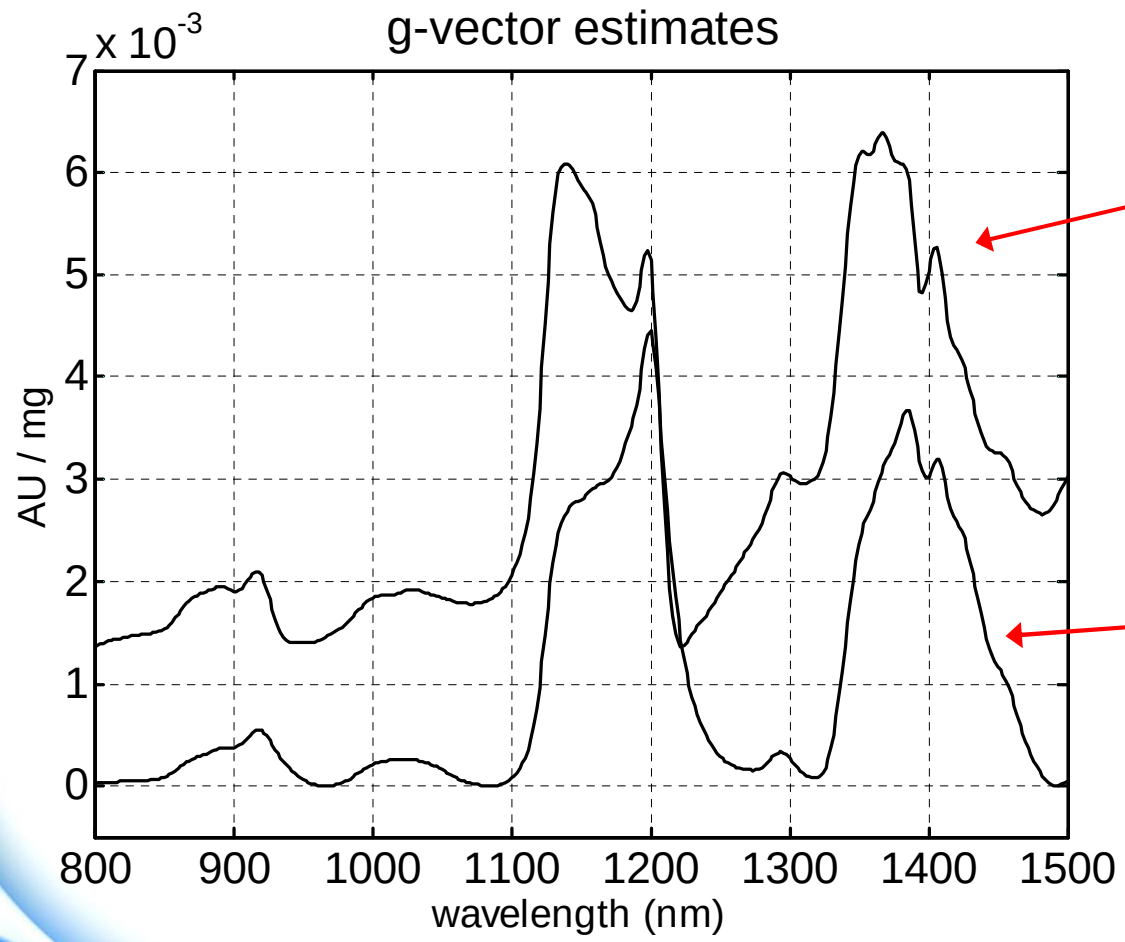
"Pure active" from: D.W. Hopkins, *Shout-out 2002: transfer of calibration for content of active in a pharmaceutical tablet*, NIR News 14 (5), 10 - 13 (2003)

baseline fit valid from about **800 - 1500 nm**

after baseline-subtraction

$$g = \frac{this\ spectrum\,[AU]}{200\ mg}$$

Shoot-out 2002 Example — API response spectrum (cont'd)

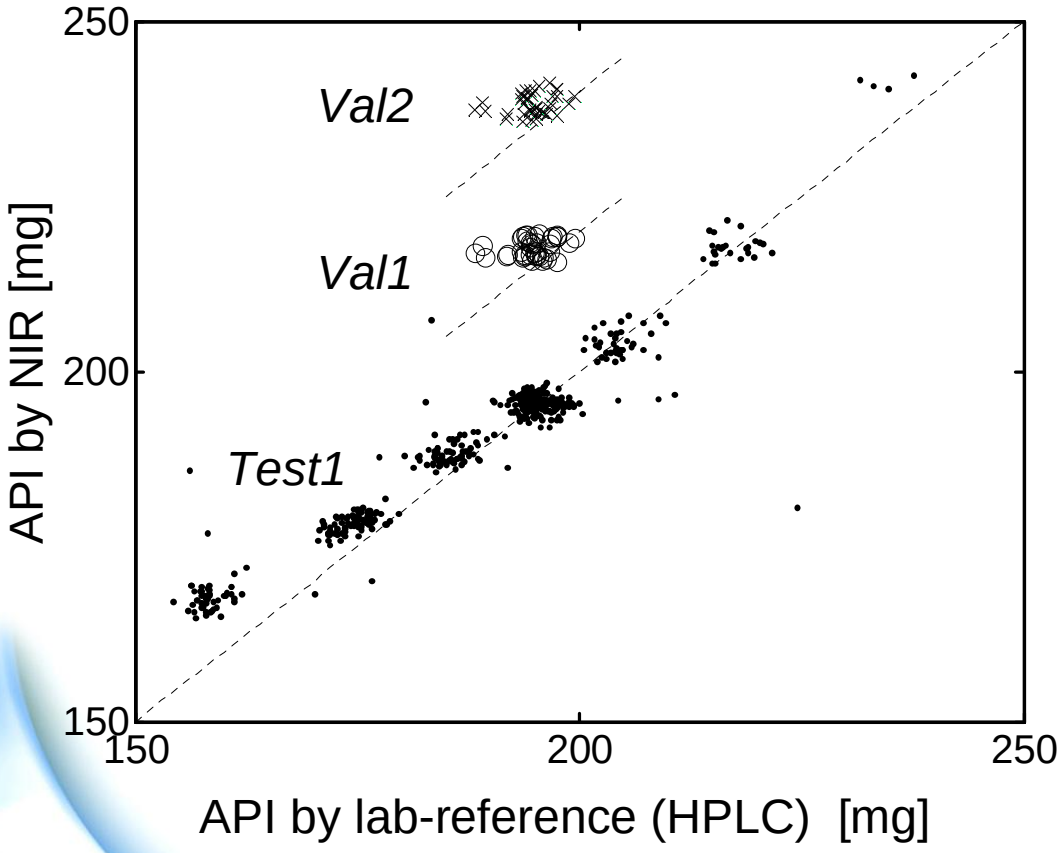


For comparison: PLS, PCR on *Calibration1* (155 lab-ref's; "designer tablets;" **no offset added**)

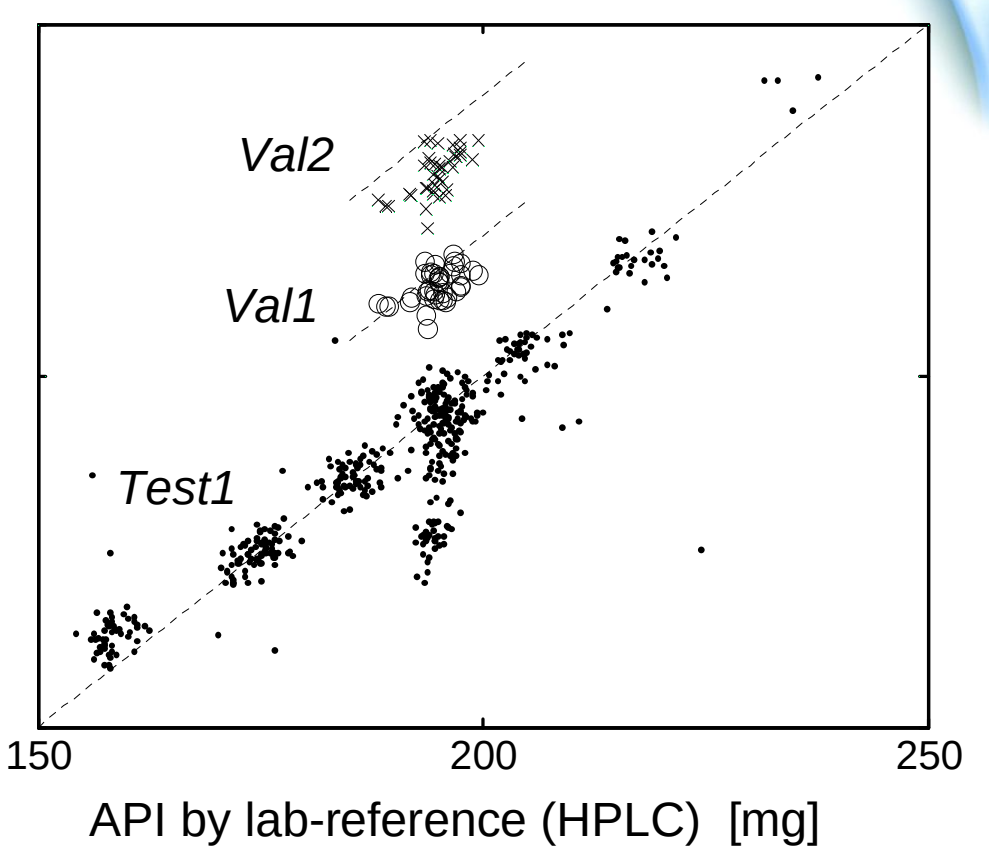
from literature "pure active" (1 lab-ref. needed)

Shoot-out 2002 Example – Prediction results

"**Science-based**" method



PLS on *Calibration1*, 3 factors



Second mental leap − ***All*** calibrations can be written in this form

$$\mathbf{b}_c = \frac{\boldsymbol{\Sigma}_c^{-}\,\mathbf{g}_c}{\mathbf{g}_c^{T}\,\boldsymbol{\Sigma}_c^{-}\,\mathbf{g}_c}$$

- *"Űber formula"* − multivariate or univariate, classical or inverse
- SBC brings common language to **all** methods
- ***Something*** is always used *as* signal
- ***Something*** is always used *as* noise

Example 1 – “Classical" calibration

Model ("K-matrix"):

$$\mathbf{x}_{pred} = \begin{bmatrix}\mathbf{g} & \mathbf{K}\end{bmatrix} \cdot \begin{bmatrix} y \\ \mathbf{c} \end{bmatrix} + \mathbf{r} \qquad \text{[AU]}$$

Equivalent b-vector:

$$y_{pred} = \left(\begin{pmatrix} 1 & 0 & \dots & 0 \end{pmatrix} \cdot \left\{ \begin{bmatrix} \mathbf{g}^T \\ \mathbf{K}^T \end{bmatrix} \cdot \begin{bmatrix} \mathbf{g} & \mathbf{K} \end{bmatrix} \right\}^{-1} \begin{bmatrix} \mathbf{g}^T \\ \mathbf{K}^T \end{bmatrix} \right) \cdot \mathbf{x}_{pred} = \mathbf{b}_{eq}^T \cdot \mathbf{x}_{pred}$$

$$\mathbf{b}_{eq} = \frac{\left(\mathbf{I} - \mathbf{K}(\mathbf{K}^T\mathbf{K})^{-1}\mathbf{K}^T\right)\cdot\mathbf{g}}{\mathbf{g}^T \cdot \left(\mathbf{I} - \mathbf{K}(\mathbf{K}^T\mathbf{K})^{-1}\mathbf{K}^T\right)\cdot\mathbf{g}} \qquad \text{[(mol/L) / AU]} \qquad \textbf{\textit{“Net analyte signal”}}$$

- Good estimate of the spectral signal, **g**
- Bad estimate of the spectral noise, **Σ**

Example 2 – "Statistical" calibration (PLS, PCR)

Inverse model: $\mathbf{y}_R = \mathbf{X} \cdot \mathbf{b} + \mathbf{e}$ [mg]

b-vector: $\mathbf{b} = \left(\tilde{\mathbf{X}}^T \tilde{\mathbf{X}}\right)^{-PLS\,/\,PCR} \tilde{\mathbf{X}}^T \tilde{\mathbf{y}}_R$

Everything that correlates is used ***as*** signal

$$= \frac{\left\{\tilde{\mathbf{X}}^T \left(\mathbf{I} - \frac{\tilde{\mathbf{y}}_R \cdot \tilde{\mathbf{y}}_R^T}{\tilde{\mathbf{y}}_R^T \tilde{\mathbf{y}}_R}\right) \tilde{\mathbf{X}}\right\}^{-PLS\,/\,PCR} \left(\frac{\tilde{\mathbf{X}}^T \tilde{\mathbf{y}}_R}{\tilde{\mathbf{y}}_R^T \tilde{\mathbf{y}}_R}\right) \cdot \tilde{\mathbf{y}}_R^T \tilde{\mathbf{y}}_R}{1 + \tilde{\mathbf{y}}_R^T \tilde{\mathbf{y}}_R \cdot \left(\frac{\tilde{\mathbf{X}}^T \tilde{\mathbf{y}}_R}{\tilde{\mathbf{y}}_R^T \tilde{\mathbf{y}}_R}\right)^T \left\{\tilde{\mathbf{X}}^T \left(\mathbf{I} - \frac{\tilde{\mathbf{y}}_R \cdot \tilde{\mathbf{y}}_R^T}{\tilde{\mathbf{y}}_R^T \tilde{\mathbf{y}}_R}\right) \tilde{\mathbf{X}}\right\}^{-PLS\,/\,PCR} \left(\frac{\tilde{\mathbf{X}}^T \tilde{\mathbf{y}}_R}{\tilde{\mathbf{y}}_R^T \tilde{\mathbf{y}}_R}\right)}$$

- Unreliable and expensive estimate of **g**; often haunted by spurious & unspecific correlations
- Good estimate of **Σ**, but expensive

Details see Ref. 2

Example 3 − Simple two-wavelength calibration

"Analytical" absorbance band:

Signal: $\mathbf{g}_c = \begin{pmatrix} g \\ 0 \end{pmatrix}$ [AU / (mol/L)]

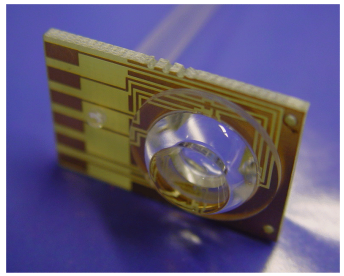
Noise: $\boldsymbol{\Sigma}_c = \sigma_x^2 \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$ [AU²]

$$\mathbf{b}_c = \frac{\boldsymbol{\Sigma}_c^- \mathbf{g}_c}{\mathbf{g}_c^T \boldsymbol{\Sigma}_c^- \mathbf{g}_c} = \frac{1}{g} \begin{pmatrix} +1 \\ -1 \end{pmatrix}$$ [(mol/L) / AU]

Example 4 – Univariate Case

Signal: $\mathbf{g}_c = g$ [AU / (mol/L)]

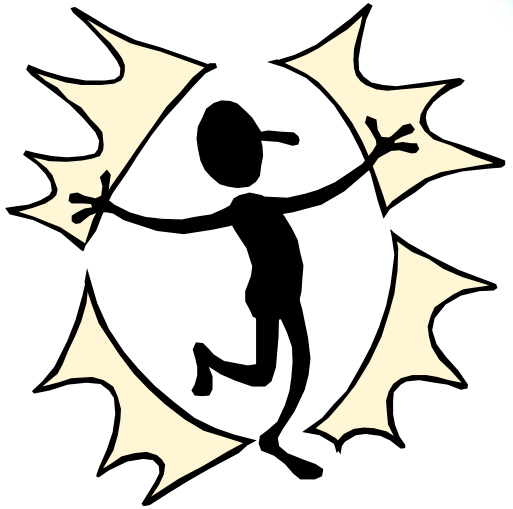
Noise: $\boldsymbol{\Sigma}_c = \sigma_x^2$ [AU²]



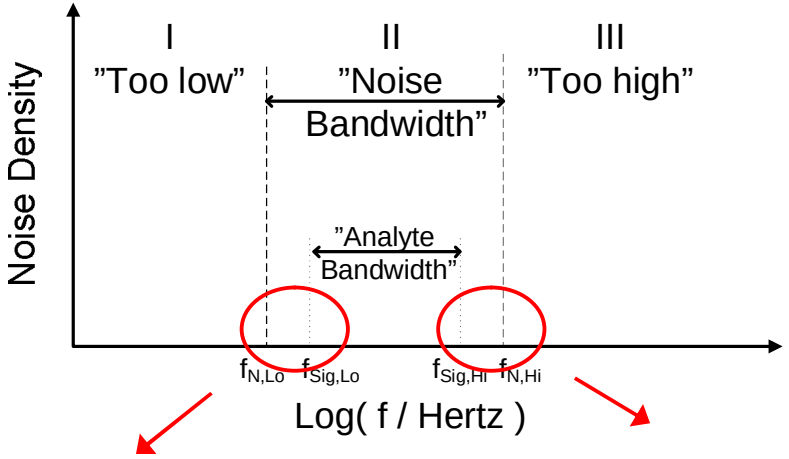
$$\mathbf{b}_c = \frac{\boldsymbol{\Sigma}_c^{-} \mathbf{g}_c}{\mathbf{g}_c^T \boldsymbol{\Sigma}_c^{-} \mathbf{g}_c} = \frac{1}{g}$$ [(mol/L) / AU]

Third mental leap − Time (a.k.a. frequency) axis is **important**

Measured spectrum: $$\mathbf{x}^T(t) = y(t)\cdot\mathbf{g}^T + \mathbf{x}_n^T(t)$$



"**Unspecific correlations**" vs. "noise"

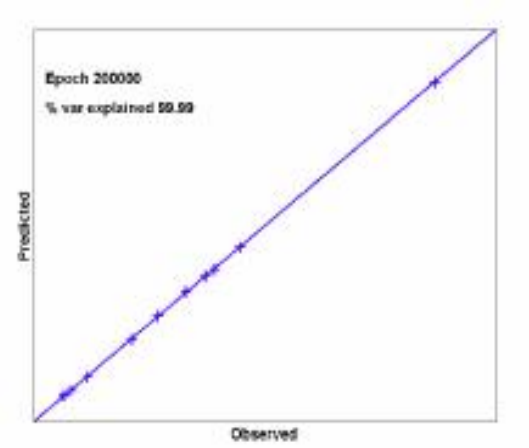


"Slow" noises
→ Stability, Accuracy

"Fast" noises
→ Repeatability, Precision

"Spurious" vs. *"unspecific" correlations*

- **Spurious correlations** appear only in "statistical" calibration (PLS, PCR)
  - Caused by random fluctuations in calibration data set. Disappear with increasing number of calibration standards. Can be bad …



T. Fearn, *Valid validation*, in: Proc. 12th Int'l Conf. on Near Infrared Spectroscopy, Auckland, New Zealand, 9 – 15 April 2005, G.R. Burling-Claridge, S.E. Holroyd and R.M.W. Sumner (eds.), 283 - 284

  - … but are only the tip of the iceberg.

    The REAL problem are ***unspecific* correlations -- chemically unspecific but statistically reproducible**. *All* methods of calibration can be affected ("statistical", "classical", and SBC).

Unspecific correlation – Example

- Measurement of **cholesterol in human blood** (e.g., IR spectrometry)
- Cholesterol (analyte) and triglycerides (interferent), $r \approx 0.5$

$$y_{trig}(t) = S \cdot r \cdot y_{chol}(t) + \left(y_{trig}(t) - S \cdot r \cdot y_{chol}(t) \right) \quad \text{where} \quad S = \sigma_{trig} / \sigma_{chol}$$

- Measured spectrum:

$$\mathbf{x}^T(t) = y_{chol}(t) \cdot \mathbf{g}_{chol}^T + y_{trig}(t) \cdot \mathbf{k}_{trig}^T + \mathbf{x}_{n,\,all\,but\,trig}^T(t)$$

$$= y_{chol}(t) \cdot \mathbf{g}_{chol}^T + \left\{ S \cdot r \cdot y_{chol}(t) + \left(y_{trig}(t) - S \cdot r \cdot y_{chol}(t) \right) \right\} \cdot \mathbf{k}_{trig}^T + \mathbf{x}_{n,all\,but\,trig}^T(t)$$

$$= y_{chol}(t) \cdot \mathbf{g}_{chol}^T + \underbrace{S \cdot r \cdot y_{chol}(t) \cdot \mathbf{k}_{trig}^T}_{\text{UC}} + \mathbf{x}_n^T(t)$$

Unspecific correlation − Example cont'd

- Prediction:

$$\hat{y}_{chol}(t) = \mathbf{x}^T(t) \cdot \mathbf{b}_c$$

$$= \left(y_{chol}(t) \cdot \mathbf{g}_{chol}^T + S \cdot r \cdot y_{chol}(t) \cdot \mathbf{k}_{trig}^T + \mathbf{x}_n^T(t) \right) \cdot \frac{\Sigma_c^- \mathbf{g}_{chol}}{\mathbf{g}_{chol}^T \Sigma_c^- \mathbf{g}_{chol}}$$

$$= y_{chol}(t) \left(1 + S \cdot r \cdot \frac{\mathbf{k}_{trig}^T \Sigma_c^- \mathbf{g}_{chol}}{\mathbf{g}_{chol}^T \Sigma_c^- \mathbf{g}_{chol}} \right) + \frac{\mathbf{x}_n^T(t) \Sigma_c^- \mathbf{g}_{chol}}{\mathbf{g}_{chol}^T \Sigma_c^- \mathbf{g}_{chol}}$$

Specificity ***(slope)*** *Sensitivity* ***(scatter)***

Unspecific correlation − General Case

- Several UC components:

$$S_1 \cdot r_1 \cdot y(t) \cdot \mathbf{k}_1^T + S_2 \cdot r_2 \cdot y(t) \cdot \mathbf{k}_2^T + \mathrm{K}$$

$$= \left(S_1 \cdot r_1 + S_2 \cdot r_2 + \mathrm{K}\right) \cdot y(t) \cdot \left(\frac{S_1 \cdot r_1}{S_1 \cdot r_1 + S_2 \cdot r_2 + \mathrm{K}} \mathbf{k}_1^T + \frac{S_2 \cdot r_2}{S_1 \cdot r_1 + S_2 \cdot r_2 + \mathrm{K}} \mathbf{k}_2^T + \mathrm{K} \right)$$

$$= \left(S \cdot r\right) \cdot y(t) \cdot \mathbf{u}^T$$

- Measured spectrum:

$$\mathbf{x}^T(t) = y(t) \cdot \mathbf{g}^T + \underbrace{\left(S \cdot r\right) \cdot y(t) \cdot \mathbf{u}^T}_{\text{UC from } \textbf{all} \text{ components}} + \mathbf{x}_n^T(t)$$

Figures of Merit − General Case

Situation: $\mathbf{x}^T(t) = y(t)\cdot\mathbf{g}^T + S\cdot r\cdot y(t)\cdot\mathbf{u}^T + \mathbf{x}_n^T(t)$ where $Cov[\mathbf{x}_n] = \boldsymbol{\Sigma}$

Calibration: $\mathbf{b}_c = \dfrac{\boldsymbol{\Sigma}_c^- \mathbf{g}_c}{\mathbf{g}_c^T \boldsymbol{\Sigma}_c^- \mathbf{g}_c}$ where $\mathbf{g}_c = \mathbf{g} + \Delta\mathbf{g}$

Prediction:

$$\hat{y}(t) = \mathbf{x}^T(t)\cdot\mathbf{b}_c = \left(y(t)\cdot\mathbf{g}^T + S\cdot r\cdot y(t)\cdot\mathbf{u}^T + \mathbf{x}_n^T(t)\right)\cdot\mathbf{b}_c$$

$$= y(t)\left\{1 + \frac{(S\cdot r\cdot\mathbf{u} - \Delta\mathbf{g})^T \boldsymbol{\Sigma}_c^- \mathbf{g}_c}{\mathbf{g}_c^T \boldsymbol{\Sigma}_c^- \mathbf{g}_c}\right\} + \frac{\mathbf{x}_n^T(t)\boldsymbol{\Sigma}_c^- \mathbf{g}_c}{\mathbf{g}_c^T \boldsymbol{\Sigma}_c^- \mathbf{g}_c}$$

Specificity *Sensitivity*

Proof of specificity =

1. $\mathbf{g}_c$ "looks right," **and**
2. $\mathbf{g}_c$ "stands tall" in $\boldsymbol{\Sigma}_c^-$

Scatter, $MSE = \dfrac{\mathbf{g}_c^T \boldsymbol{\Sigma}_c^- \cdot \boldsymbol{\Sigma} \cdot \boldsymbol{\Sigma}_c^- \mathbf{g}_c}{\left(\mathbf{g}_c^T \boldsymbol{\Sigma}_c^- \mathbf{g}_c\right)^2} \geq \dfrac{1}{\mathbf{g}_c^T \boldsymbol{\Sigma}^- \mathbf{g}_c}$

Proof of specificity − Praxis

Proof of specificity =

1. $\mathbf{g}_c$ "looks right," **and**
2. $\mathbf{g}_c$ "stands tall" in Σ_c^{-}

(1) Determine $\mathbf{g}$ from spectroscopic expertise & application knowledge

(2) Determine Σ_c -- estimate Σ and then start adding "extra" noises

$$\Sigma_c = \hat{\Sigma} + \sigma^2_{c,trig}\left(\hat{\mathbf{k}}_{trig}\hat{\mathbf{k}}^T_{trig}\right) + \sigma^2_{c,urea}\left(\hat{\mathbf{k}}_{urea}\hat{\mathbf{k}}^T_{urea}\right) + \mathrm{K}$$

→ Specificity of response
→ Long-term stability of calibration
→ Transferability instr.-to-instr.
→ …

Effective trade-off between specificity and sensitivity requires user-control over BOTH $\mathbf{g}_c$ and Σ_c (SBC)

Proof of specificity − Praxis (cont'd)

… **Two** ways to prove the **second** step

Proof of specificity =

1. $\mathbf{g}_c$ "looks right," **and**
2. $\mathbf{g}_c$ "stands tall" in $\sum_c^-$

(1) From spectroscopic expertise and application knowledge

(2) From spectroscopic first-principles;

OR

Empirically: By checking slope vs. a reference method (*)

(*) Slope check is only valid AFTER step (1) is passed

Current testing practice does not ensure Step (1) is passed !!

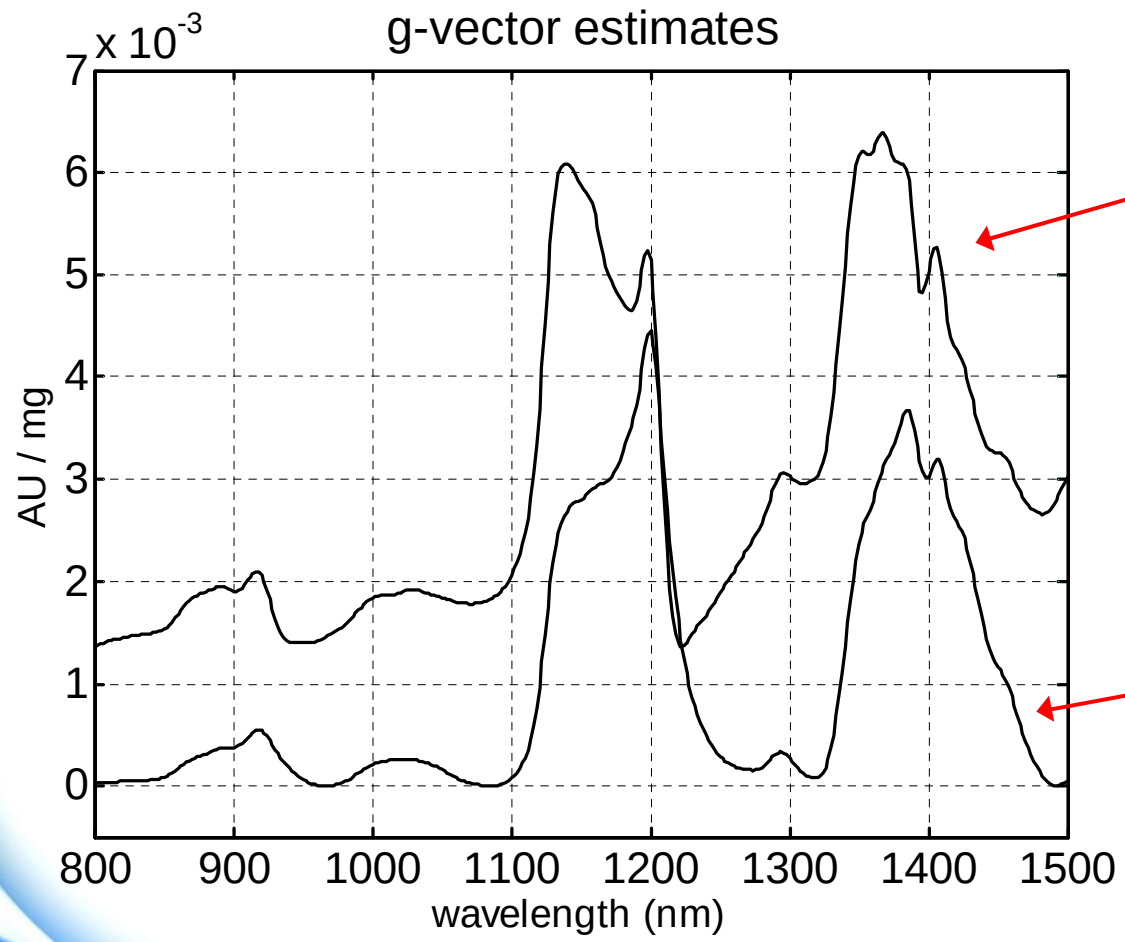
- If a “statistical” calibration (PLS, PCR) is affected by UC and then tested on so-called “independent” data that are affected by the *same* UC, seemingly good predictions (slope=1) can result even w/o any analyte signal
  - Current practice of predicting “independent” spectra guards only against *spurious* correlations, but not against *unspecific* correlations
- Current guidelines of checking for specificity (ICH Q2B; ASTM 1655; etc.) are incomplete / fail in the case of “statistical” calibration (PLS, PCR) and should be amended
  - Step (1) can only be passed by spectroscopic expertise & application knowledge, i.e., a ’responsible scientist’ must *define and approve*

Issue with “statistical” calibration (PLS, PCR)

- **“Statistical” calibrations are special** because they actively search for *any* correlations to use them *as* signal
  - *Spurious* correlations pose additional problem
  - The (implicitly used) estimate, $\mathbf{g}_c$, is virtually guaranteed to be affected whenever *unspecific* correlations are present

à Step (1) often not passed

Chambersburg *Shoot-out 2002* Example − API in tablets



Response spectrum implicitly used by PLS, PCR (from data set *Calibration1;* 155 standards); **no offset added**

Response spectrum from literature

Summary of older (non-SBC) calibration methods

1. **"Classical" and "simple" (few wavelengths) calibrations:**
   - Sensitivity − likely sub-optimal
   - Specificity − if slope OK, then likely OK
2. **"Statistical" calibrations** (PLS, PCR):
   - Sensitivity − often "better" than optimal
   - Specificity − often not measuring "the right thing" (fail Step 1)
     - If sensitivity and specificity are to be defined, then "a-posteriori sanctioning" of $\mathbf{g}_c$ is necessary
       - With good application knowledge, "specificity in design space" may be argued
       - When testing prediction slope, **"really" independent prediction samples** are vital
     - OR, user can try to retreat to statistical point of view

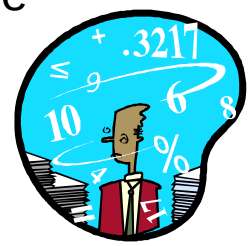
"Measurement Science" vs. "Statistics"

Dividing line:

Knowledge of the "response spectrum" $\mathbf{g}_c$ [AU/mg] of the analyte of interest

Multivariate Statistics

- Specificity & Sensitivity can *not* be defined
- Only *one* performance metric, viz. correlation ("PRESS$^{1/2}$", …)
- Proper word: "Prediction"
- Historically, chemometrics has focused here



Multivariate Measurement Science

- Specificity & Sensitivity can be defined (*two* performance metrics)
- Proper word: "Measurement"
- Historically done in "classical" and "simple" (few wavelengths) calibrations
- Now also possible in the general case where "noise matching" is needed

Example: “BTX” in NIR

- Benzene, Toluene, *o*-Xylene, *m*-Xylene, and *p*-Xylene
- “How well can the three Xylenes be predicted?”
- **Typical PLS results:**
  - FT-NIR, 1 mm cuvette
  - 6300 – 4150 cm^{-1} (1587 – 2410 nm)

Table 1. Standard errors of predictions obtained for the quantification of xylene isomers with NIR

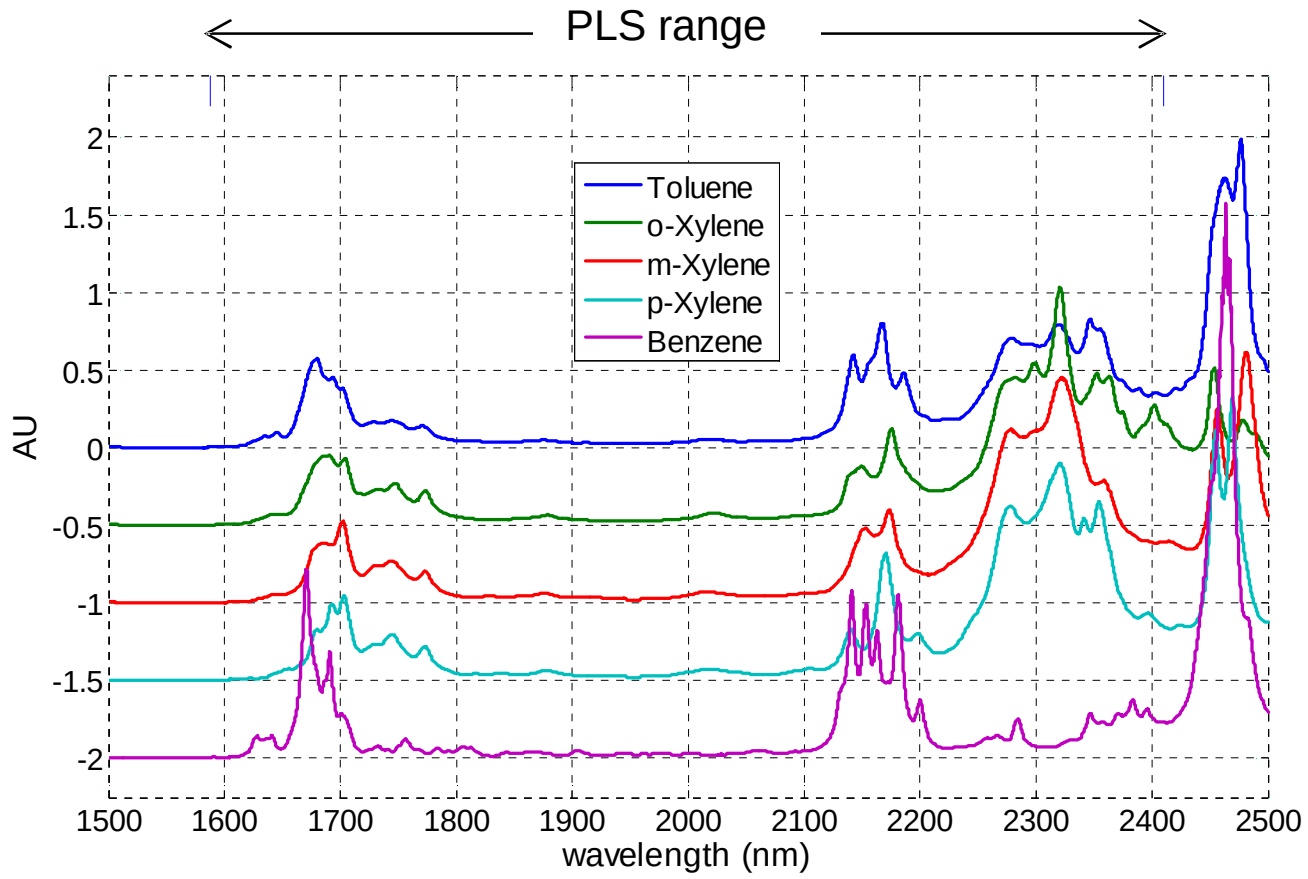
Resolution (cm^{-1})	SEP for *o*-xylene (%)	SEP for *m*-xylene (%)	SEP for *p*-xylene (%)
16	0.023	0.020	0.017
8	0.021	0.019	0.013
4	0.038	0.035	0.036
2	0.027	0.029	0.028

All concentrations in [%] v-v

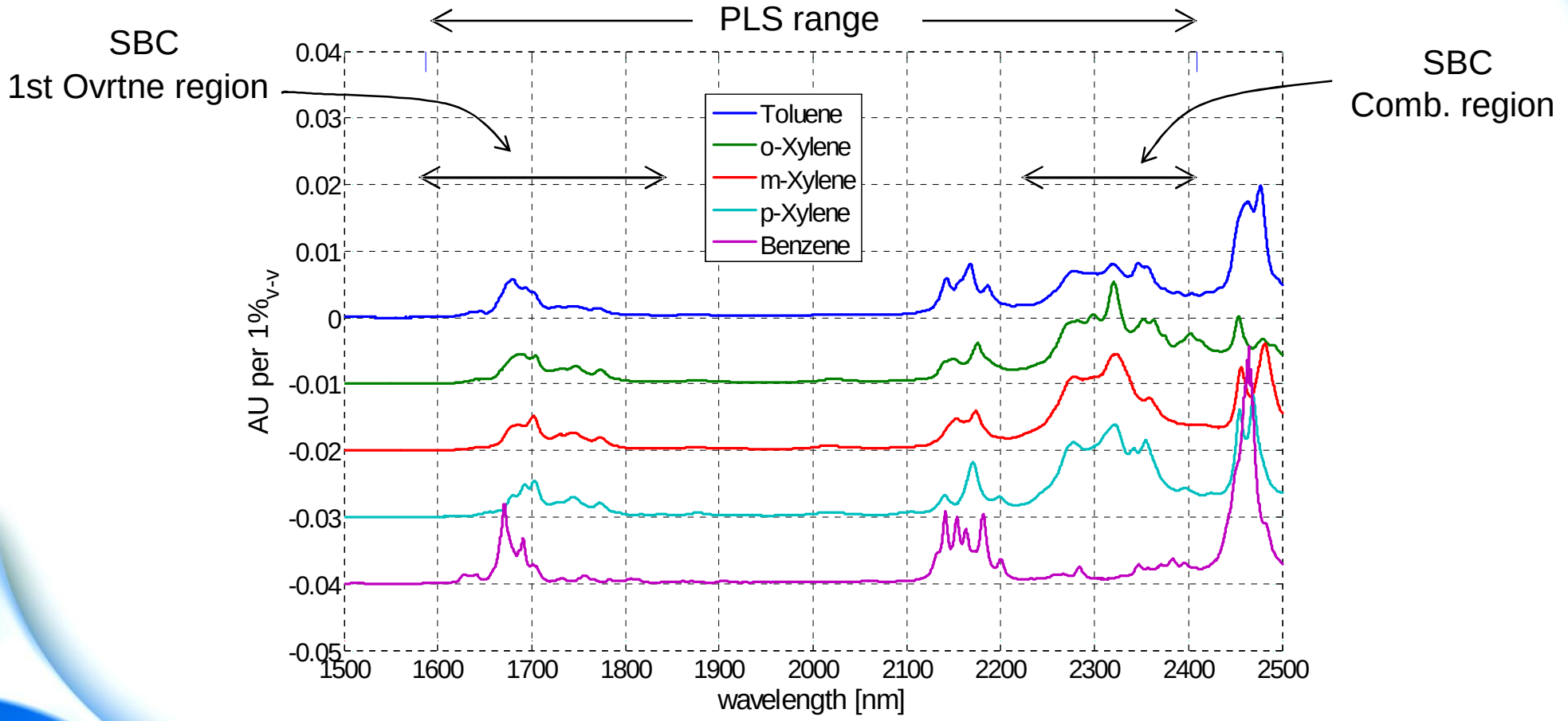
From: T. Meyer, J. Oelichmann, H. Kellerhals, *Resolution and suppression of mechanical noise in FT-NIR Spectroscopy*, Trends in Analytical Chemistry 25(1), 19-23, 2006

Response spectra for SBC

- Cary-5000 double-beam spectrometer, SBW=0.5 nm, point spacing 0.2 nm
- 1mm cuvette, “pure” component spectra (p.a.)



Response spectra (cont'd)



"BTX" example: Data set used in PLS analysis

- 50 mixtures ("designer samples")
- 6 repeats each (300 spectra)

  - Overview:

	Toluene	o-Xylene	m-Xylene	p-Xylene	Benzene
Mean	10.37	2.19	1.29	1.52	84.63
Std	5.61	1.58	1.23	1.42	5.61

From: T. Meyer, J. Oelichmann, H. Kellerhals, *Resolution and suppression of mechanical noise in FT-NIR Spectroscopy*, Trends in Analytical Chemistry 25(1), 19-23, 2006

SBC Method, e.g., for Toluene

- Signal, $\mathbf{g} = \mathbf{g}_{Tol} \quad [AU/\%_{v-v}]$
- Noise,

$$\boldsymbol{\Sigma}_{c,Tol} = (1.58\,\%)^2 \left(\mathbf{g}_{oXy} \cdot \mathbf{g}_{oXy}^T\right) + (1.23\,\%)^2 \left(\mathbf{g}_{mXy} \cdot \mathbf{g}_{mXy}^T\right) + (1.42\,\%)^2 \left(\mathbf{g}_{pXy} \cdot \mathbf{g}_{pXy}^T\right) + (5.61\,\%)^2 \left(\mathbf{g}_{Ben} \cdot \mathbf{g}_{Ben}^T\right) + \mathrm{K}$$

$$\mathrm{K} + \begin{pmatrix} n_{\lambda_1}^2 & 0 & 0 \\ 0 & \mathrm{O} & 0 \\ 0 & 0 & n_{\lambda_k}^2 \end{pmatrix} + (20\,mAU)^2 \begin{pmatrix} 1 \\ \mathrm{M} \\ 1 \end{pmatrix} \cdot \begin{pmatrix} 1 & \mathrm{L} & 1 \end{pmatrix} + (0.02\,nm)^2 \left(\frac{\partial \bar{\mathbf{x}}}{\partial \lambda}\right) \cdot \left(\frac{\partial \bar{\mathbf{x}}^T}{\partial \lambda}\right) \qquad [AU^2]$$

Hardware noise floor
- 50 µAU RMS at 1900 nm
- single-beam intensity scaled at other λ's

Baseline offset noise
- 20 mAU RMS

Wavelength Shift Noise
- 0.02 nm RMS shift
- computed on average absorbance spectrum

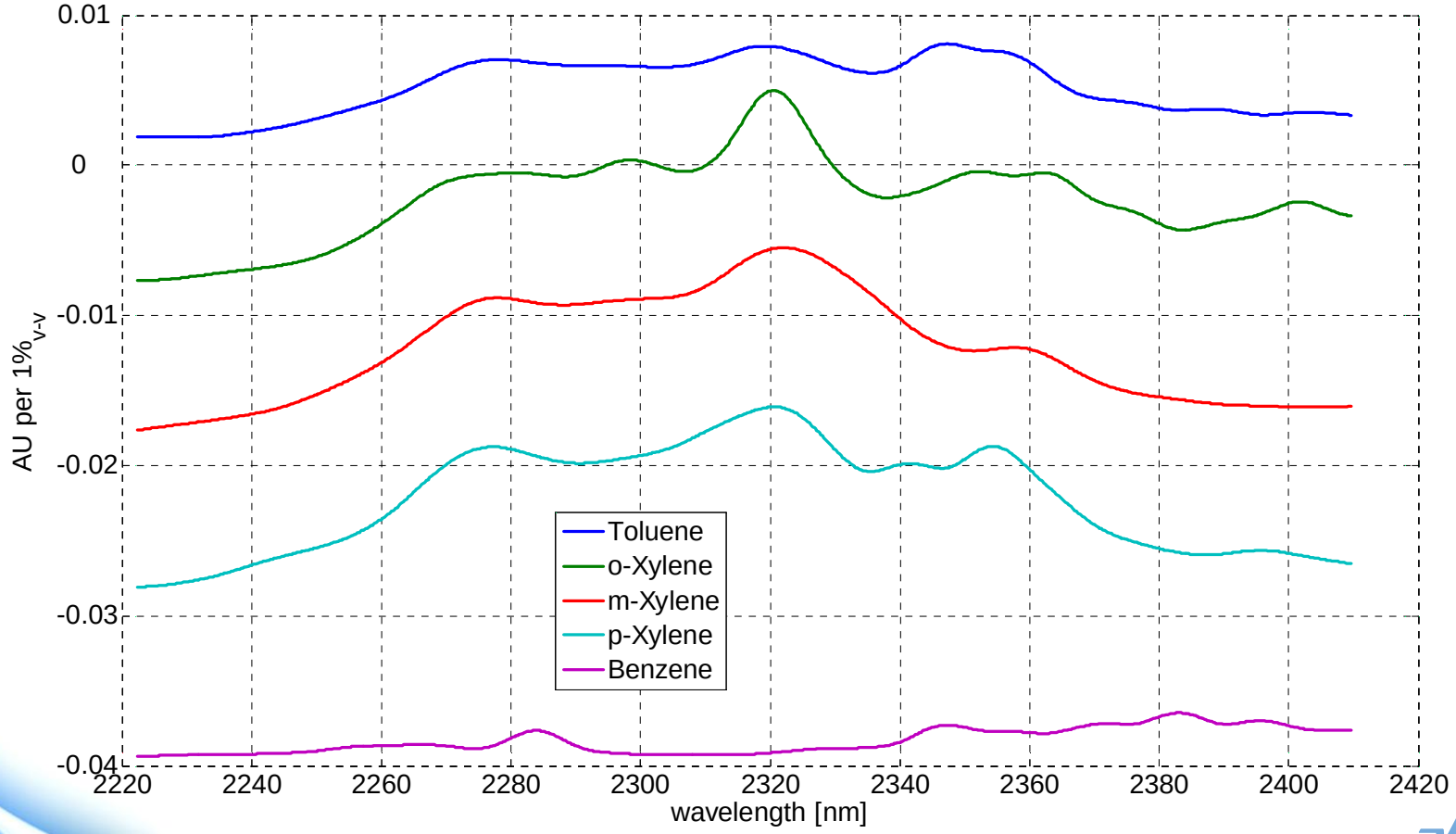
SBC Results, Combination region, 4500 – 4150 cm^{-1}

$$\left(\Delta\nu = 16\ cm^{-1}\right) \cong \left(\Delta\lambda = 8\ nm\right)$$

MATLAB Results created 09-Jun-2007 21:32:40									
2222.3999 - 2409.6001 nm, NoRanges=1, NoPixels=937									
					Res = 7.9964 nm FWHM,				
Results in "absolute" units [% v-v]									
							Interferents ...		
Analytes	*NominalCo*	*Sensitivity*	*Repeatability*	*Sen/Rep*	***Toluene***	***o-Xylene***	***m-Xylene***	***p-Xylene***	***Benzene***
	[%v-v]	*[%v-v] rms*	*[%v-v] rms*		*Ratio-of-slopes ... [%v-v] / [+1%v-v from interferent]*				
Toluene	10.37	**0.0095**	0.0095	1.0000107	1	0.0000	0.0000	0.0000	0.0000
o-Xylene	2.19	**0.0028**	0.0028	1.0000018	0.0000	1	0.0000	0.0000	0.0000
m-Xylene	1.29	**0.0034**	0.0034	1.0000042	0.0000	0.0000	1	0.0000	0.0000
p-Xylene	1.52	**0.0072**	0.0072	1.0000027	0.0000	0.0000	0.0000	1	0.0000
Benzene	84.63	**0.0062**	0.0062	1.0000057	0.0000	0.0000	0.0000	0.0000	1
Results in "relative" units [%] of nom. concentration									
							Interferents ...		
Analytes	*NominalCo*	*Sensitivity(*)*	*Repeatability(*)*	*Sen/Rep*	***Toluene***	***o-Xylene***	***m-Xylene***	***p-Xylene***	***Benzene***
	[%]	[%] rms	[%] rms		Ratio-of-slopes(**) ... [%] / [+1% from interferent]				
Toluene	100	0.0917	0.0917	1.0000107	1	0.0000	0.0000	0.0000	0.0000
o-Xylene	100	0.1263	0.1263	1.0000018	0.0000	1	0.0000	0.0000	0.0000
m-Xylene	100	0.2613	0.2613	1.0000042	0.0000	0.0000	1	0.0000	0.0000
p-Xylene	100	0.4751	0.4751	1.0000027	0.0000	0.0000	0.0000	1	0.0000
Benzene	100	0.0073	0.0073	1.0000057	0.0000	0.0000	0.0000	0.0000	1
		(*) Highlighted if > 1% RMS			*(**) Highlighted if absolute value >0.05*				

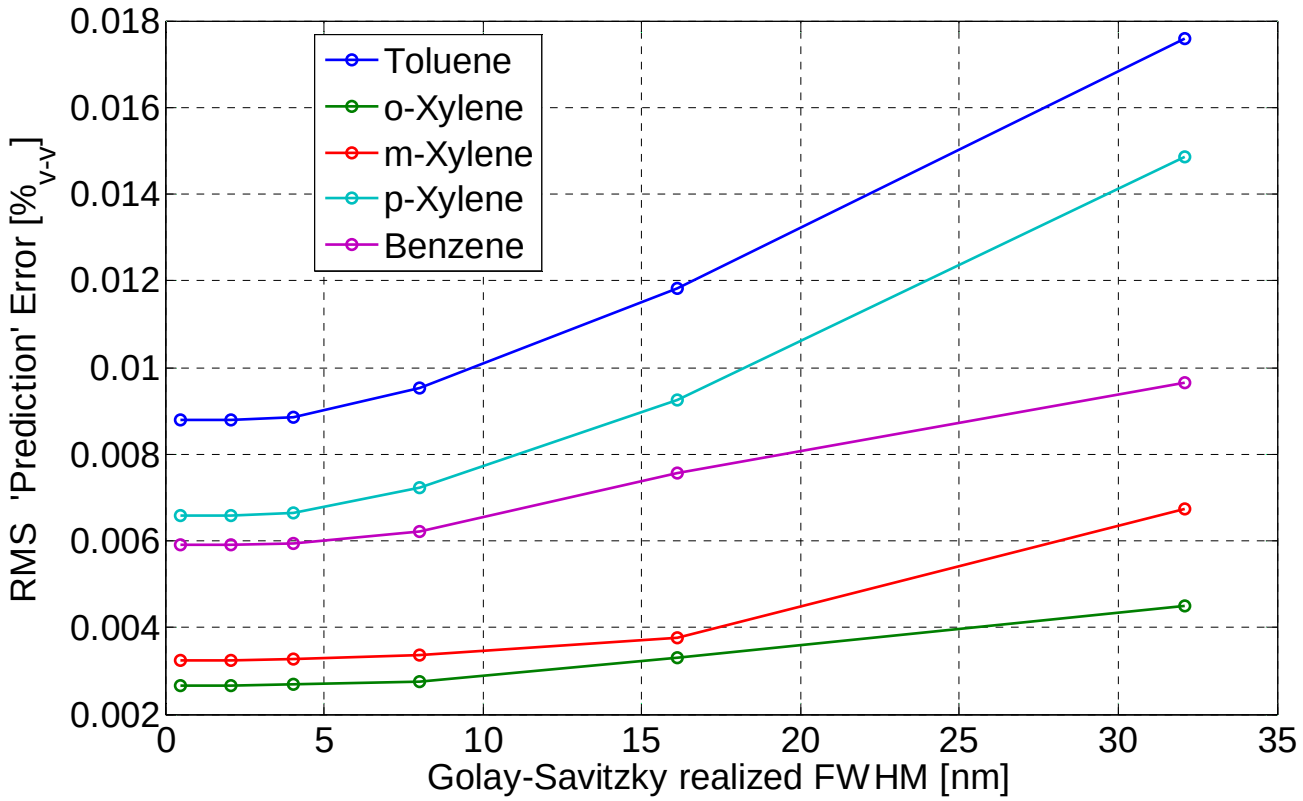
Response Spectra at ~16 cm^{-1} in combination region

- Response spectra measured on Cary-5000 (offset by -0.01 [AU/%])
- Golay-Savitzki FWHM = 7.9964 nm



Sensitivity as function of spectral resolution, Combination region

- 4500 – 4150 cm-1
- SBC calibration



~ 16 cm^{-1}

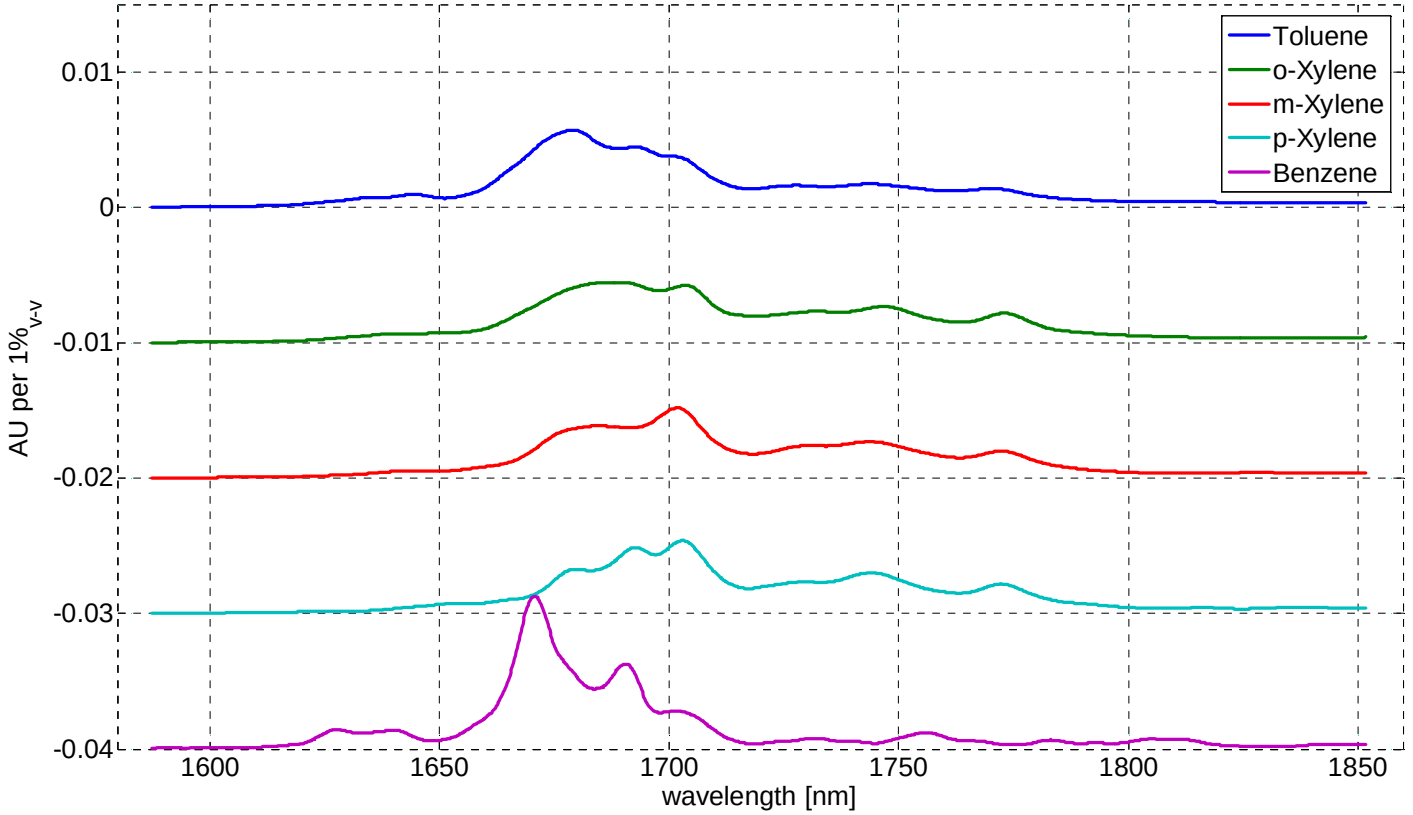
SBC results, 1st Overtone region, 6300 – 5400 cm^{-1}

$$\left(\Delta\nu = 16\ cm^{-1}\right) \cong \left(\Delta\lambda = 5\ nm\right)$$

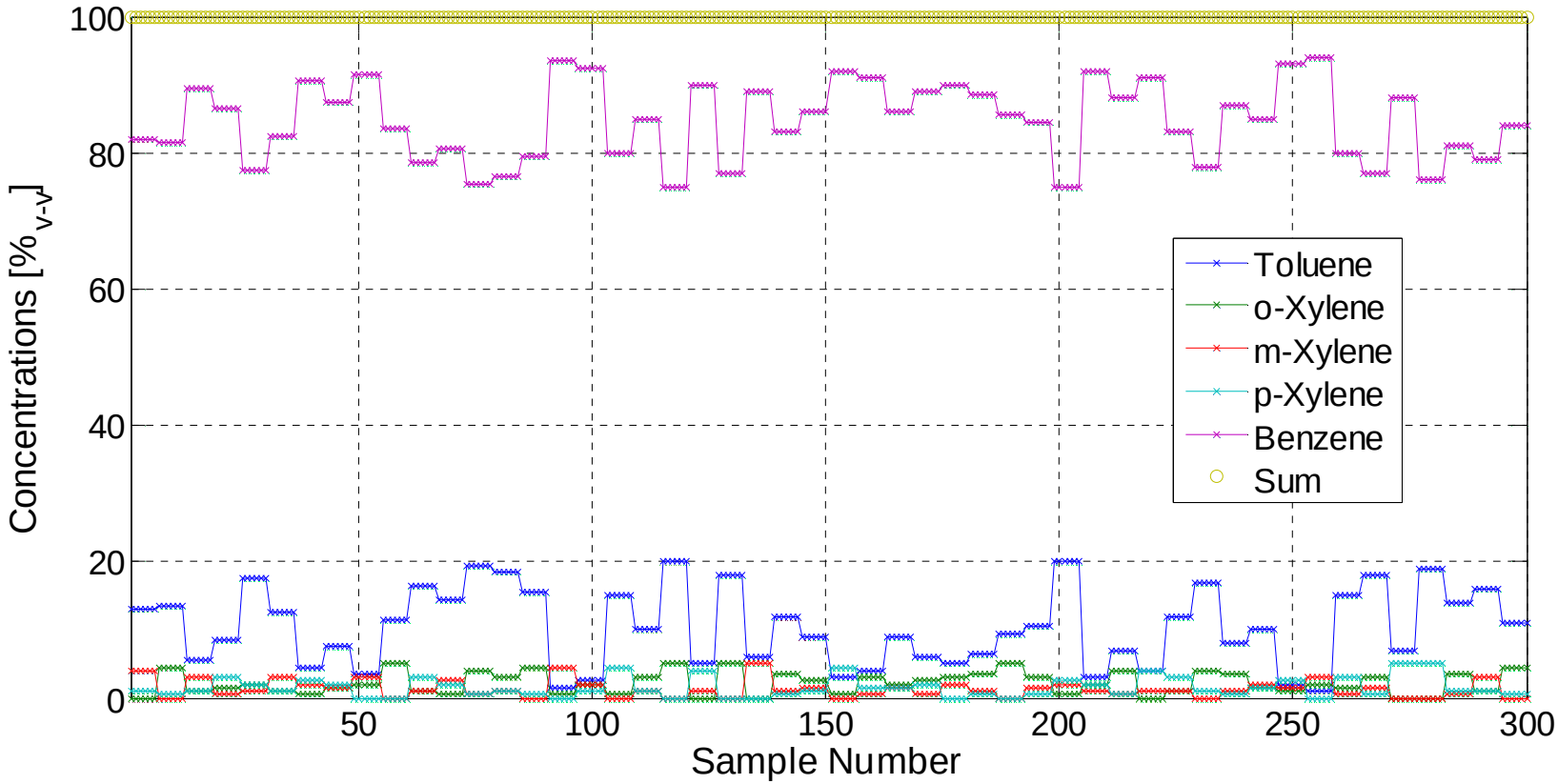
MATLAB Results created 10-Jun-2007 13:53:49									
1587.4 - 1851.8 nm, NoRanges=1, NoPixels=1323									
					Res = 5.1484 nm FWHM,				
Results in "absolute" units [% v-v]									
							Interferents ...		
Analytes	*NominalCc*	*Sensitivity*	*Repeatability*	*Sen/Rep*	***Toluene***	***o-Xylene***	***m-Xylene***	***p-Xylene***	***Benzene***
	[%v-v]	*[%v-v] rms*	*[%v-v] rms*		*Ratio-of-slopes ... [%v-v] / [+1%v-v from interferent]*				
Toluene	10.37	**0.0127**	0.0127	1.0000241	1	0.0000	0.0001	0.0000	0.0000
o-Xylene	2.19	**0.0132**	0.0132	1.000014	0.0000	1	0.0000	0.0000	0.0000
m-Xylene	1.29	**0.0190**	0.0190	1.0000379	0.0000	0.0000	1	0.0001	0.0000
p-Xylene	1.52	**0.0149**	0.0149	1.0000767	0.0000	0.0000	0.0001	1	0.0000
Benzene	84.63	**0.0055**	0.0055	1.000012	0.0000	0.0000	0.0000	0.0000	1
Results in "relative" units [%] of nom. concentration									
							Interferents ...		
Analytes	*NominalCc*	*Sensitivity(*)*	*Repeatability(*)*	*Sen/Rep*	***Toluene***	***o-Xylene***	***m-Xylene***	***p-Xylene***	***Benzene***
	[%]	[%] rms	[%] rms		Ratio-of-slopes(**) ... [%] / [+1% from interferent]				
Toluene	100	0.1228	0.1228	1.0000241	1	0.0000	0.0000	0.0000	0.0000
o-Xylene	100	0.6013	0.6013	1.000014	0.0000	1	0.0000	0.0000	0.0000
m-Xylene	100	**1.4755**	**1.4755**	1.0000379	0.0000	0.0001	1	0.0001	-0.0001
p-Xylene	100	0.9818	0.9817	1.0000767	0.0000	0.0000	0.0001	1	0.0000
Benzene	100	0.0065	0.0065	1.000012	0.0000	0.0000	0.0000	0.0000	1
		(*) Highlighted if > 1% RMS			*(**) Highlighted if absolute value >0.05*				

Response spectra at ~16 cm^{-1} in 1st Overtone region

- Response spectra from Cary-5000 (offset by -0.01 [AU/%])
- Golay-Savitzki FWHM = 5.1484 nm

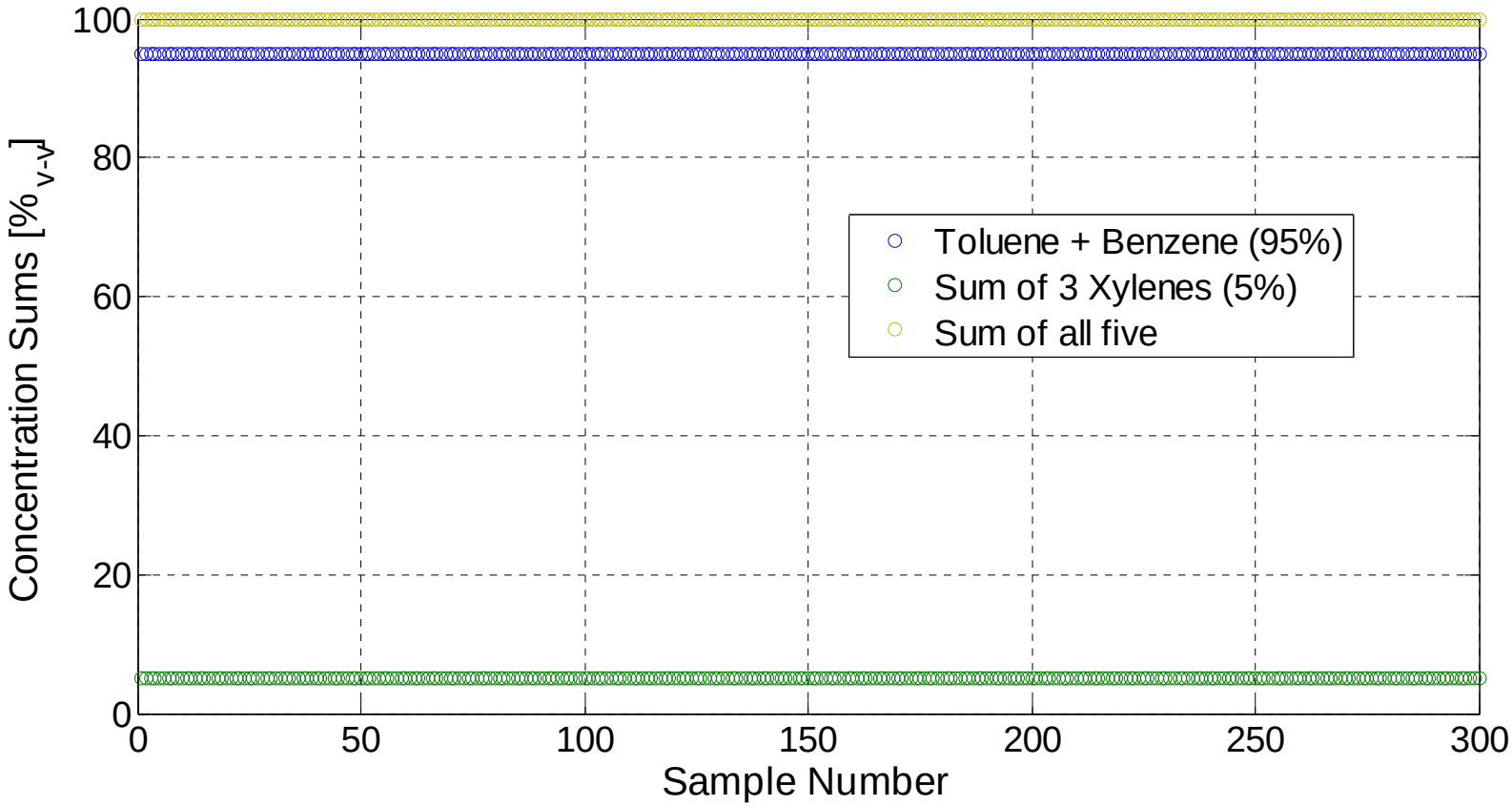


PLS reference concentrations



- 50 standards (“designer samples”), 6 repeats each

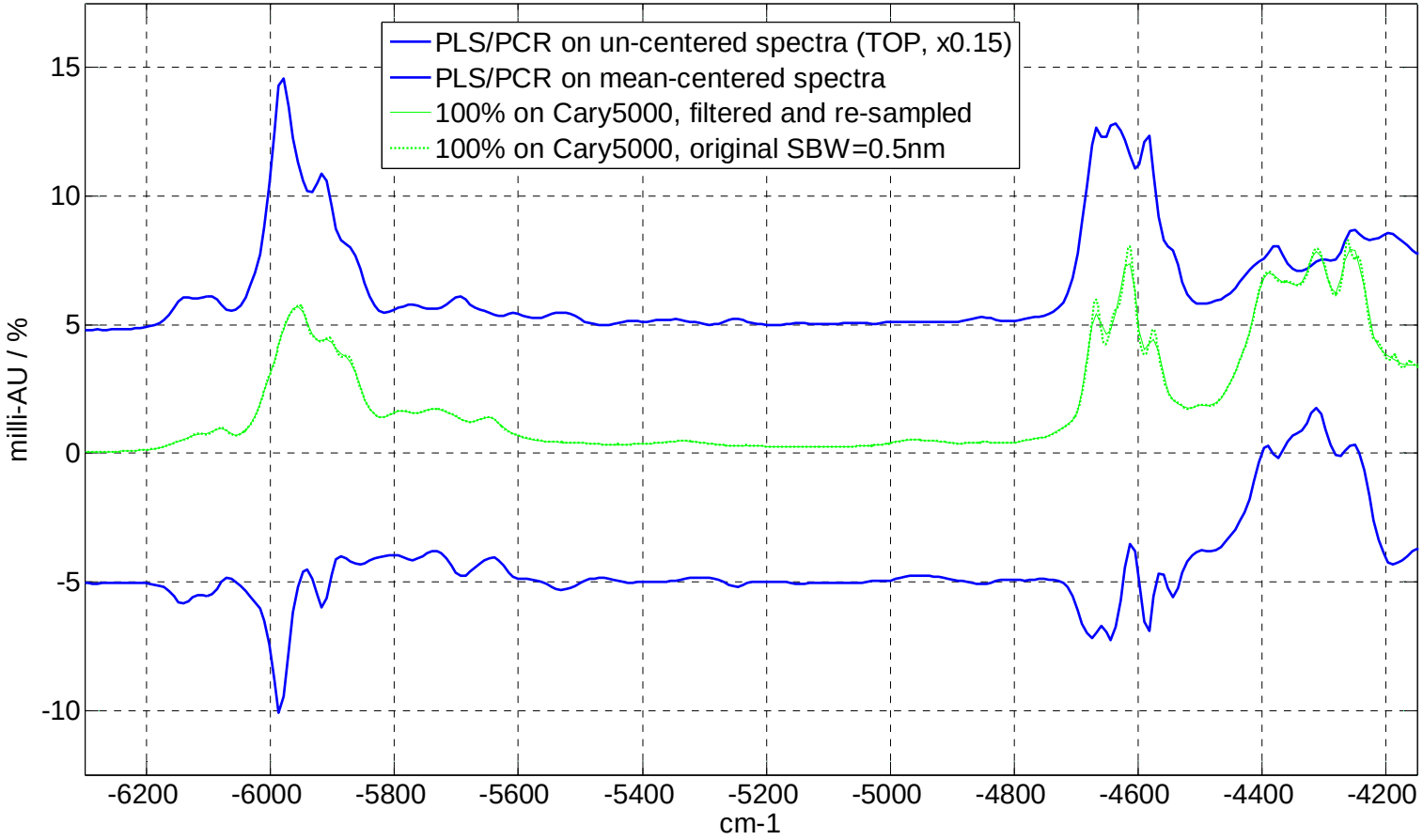
Standards affected by closure (2x)



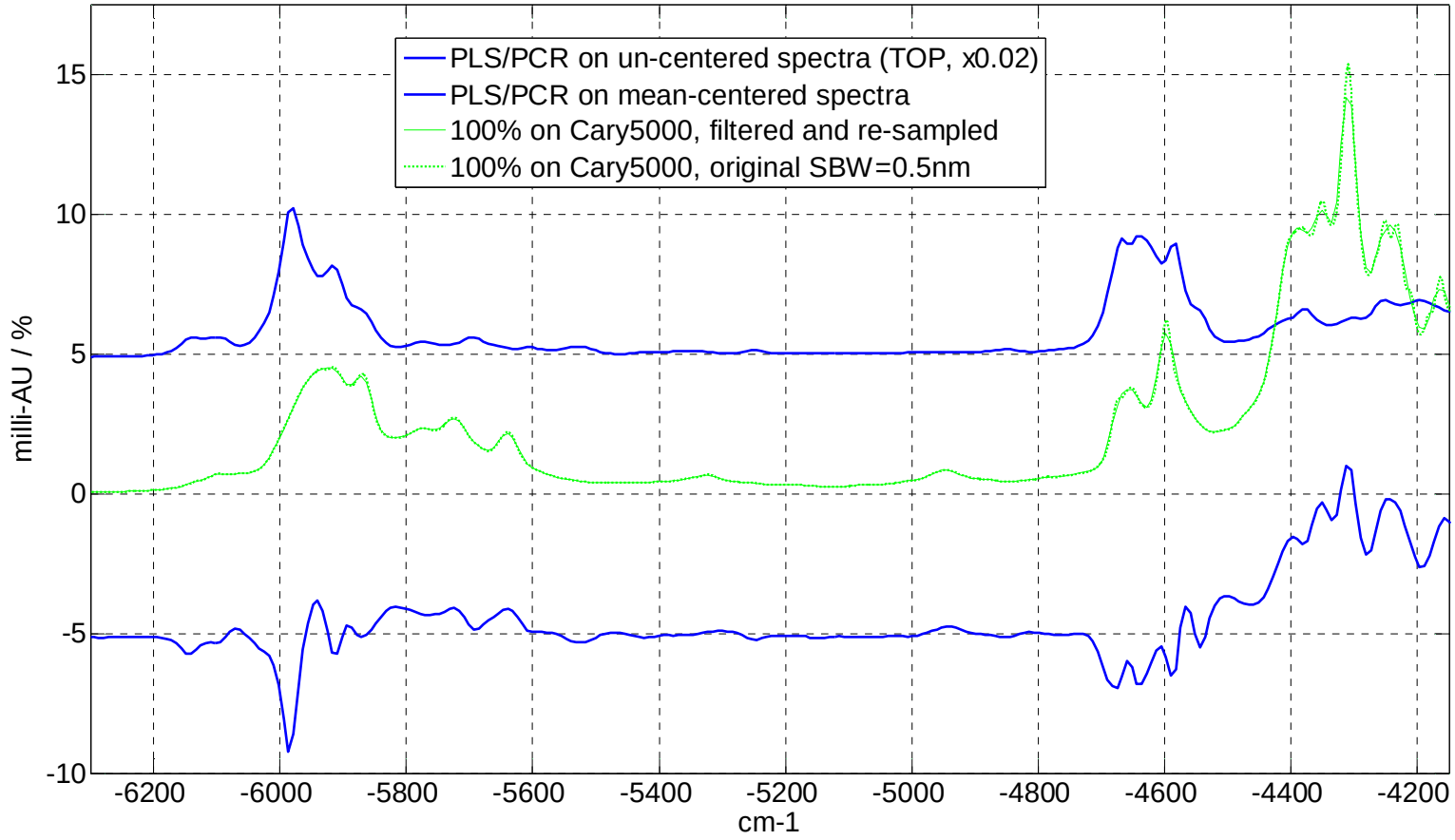
Closure causes non-zero correlation coefficients

	Toluene	o-Xylene	m-Xylene	p-Xylene	Benzene	oX+mX	oX+pX	mX+pX
Toluene	1	0.2689	-0.3208	-0.0217	**-1.0000**	0.0217	0.3208	0.3208
o-Xylene	0.2689	1	-0.5121	-0.6694	-0.2689	0.6694	0.5121	0.5121
m-Xylene	-0.3208	-0.5121	1	-0.2953	0.3208	0.2953	**-1.0000**	**-1.0000**
p-Xylene	-0.0217	-0.6694	-0.2953	1	0.0217	-1.0000	0.2953	0.2953
Benzene	**-1.0000**	-0.2689	0.3208	0.0217	1	-0.0217	-0.3208	-0.3208
oX + mX	0.0217	0.6694	0.2953	-1.0000	-0.0217	1	-0.2953	-0.2953
oX + pX	0.3208	0.5121	**-1.0000**	0.2953	-0.3208	-0.2953	1	**1.0000**
mX + pX	0.3208	0.5121	**-1.0000**	0.2953	-0.3208	-0.2953	**1.0000**	1

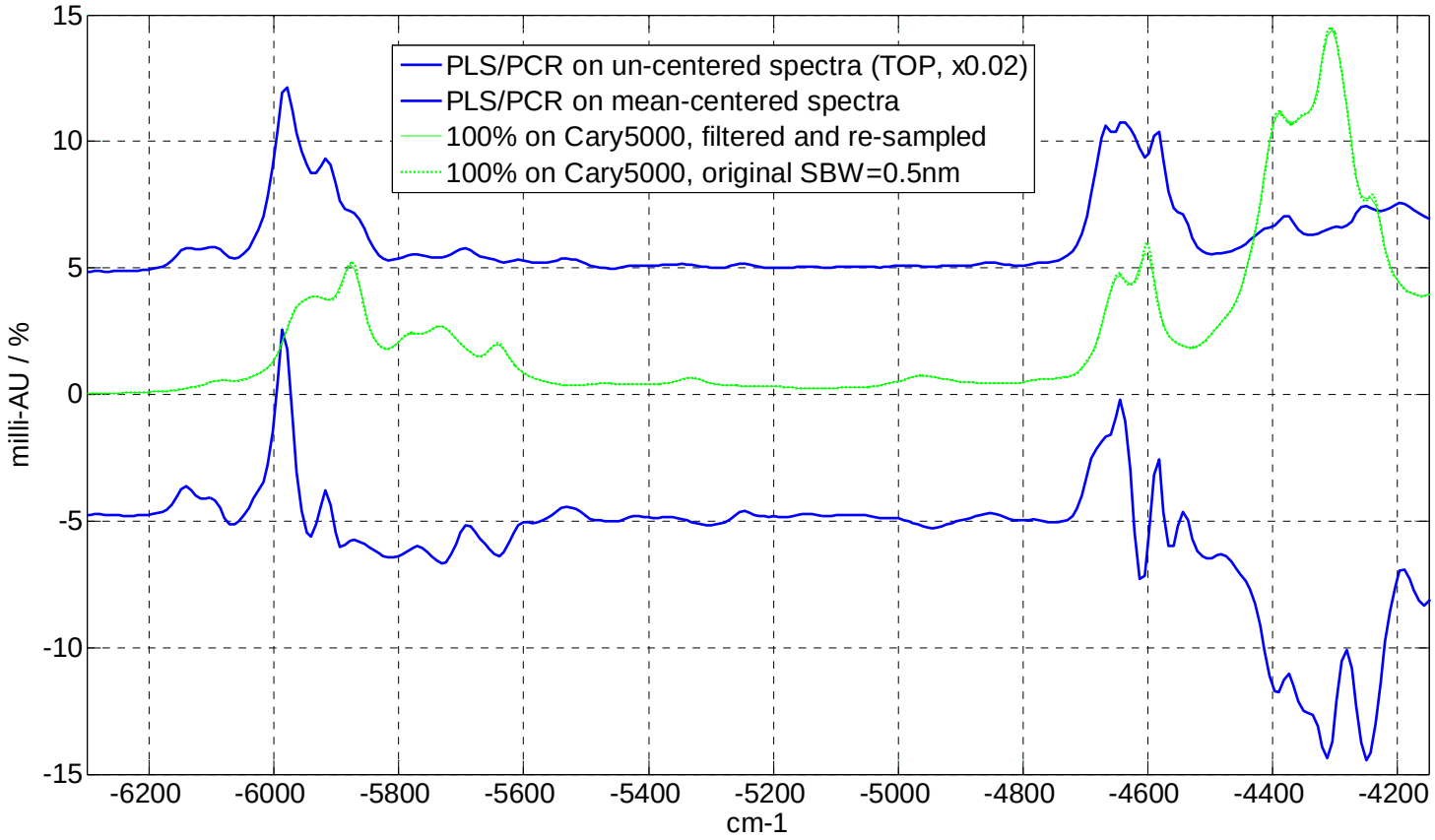
PLS “response” spectrum used for Toluene



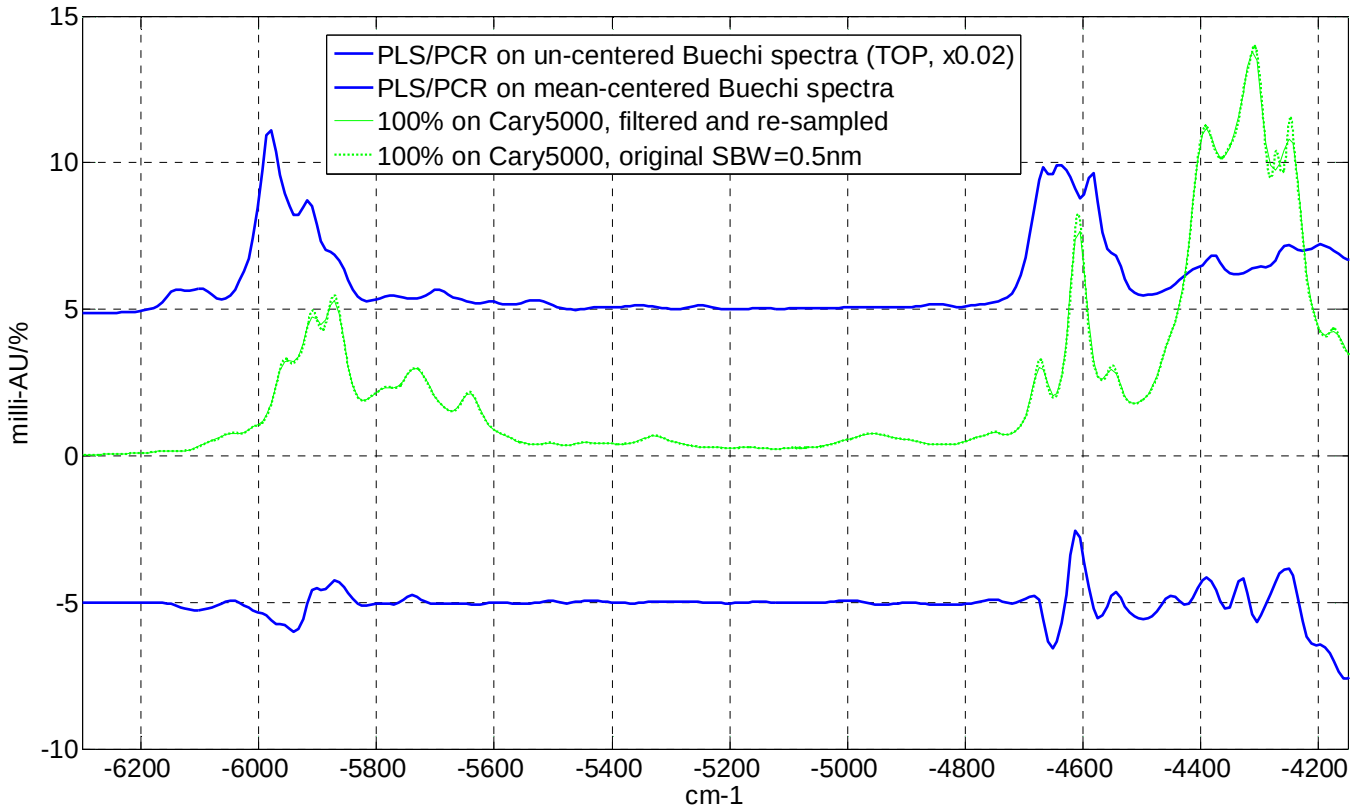
PLS “response” spectrum used for o-Xylene



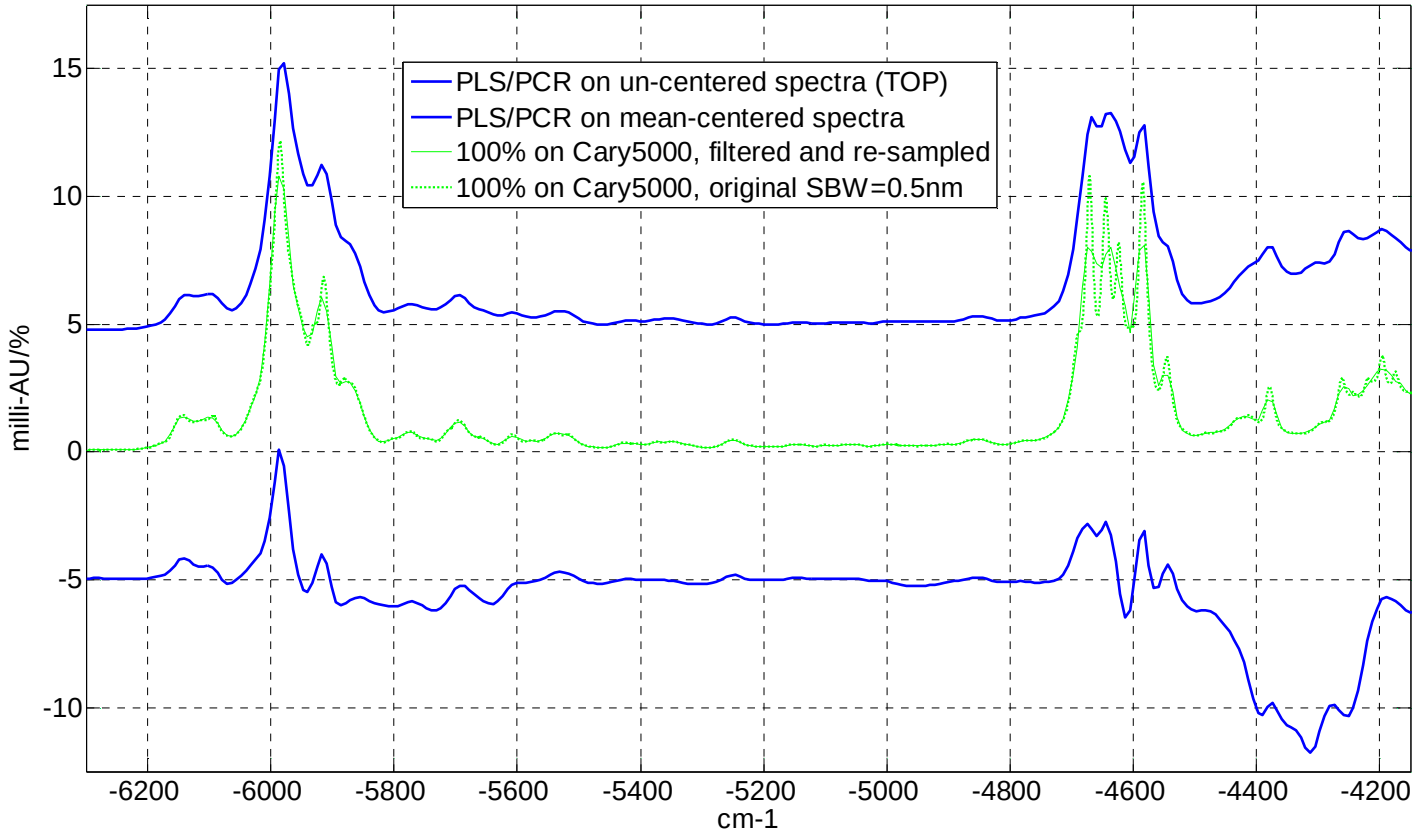
PLS “response” spectrum used for m-Xylene



PLS “response” spectrum used for p-Xylene



Response spectra used for Benzene



Conclusions

- Multivariate *measurement science* is different from multivariate *statistics*
- Multivariate calibration is NOT complicated
  - Three breaks from traditional thinking, and everything becomes clear
  - Multivariate calibration consists of only two parts, $\mathbf{g}_c$ and $\boldsymbol{\Sigma}_c^-$
- Taking control over **both** inputs is the only way to (a) prove specificity from spectroscopic first-principles and (b) trade-off specificity vs. sensitivity in an effective and user-controlled way
  - "Enabling" technology; used in Finland since 2001
- SBC makes multivariate calibration as simple & intuitive as univariate
  - Only difference: Analyst can influence the trade-off between specificity/robustness vs. short-term noise/repeatability
- Existing statistical calibrations (PLS, PCR) should be re-evaluated in light of the scientific understanding generated by SBC
  - As much as several 10% of NIR applications may be affected by unspecific correlations

…

Conclusions (cont'd)

- Current methods for testing specificity (ASTM 1655; ICH Q2B; etc.) should be amended. **Spectrometric community should start discussion.**
- Role of chemometricians will **grow** in future. Focus will shift back to spectroscopy & chemistry à "responsible application scientist"
- Community & Scientific Journals should start to enforce a rule:

  "Every manuscript must plot the (implicitly used) response spectrum"
- The best days of spectrometry are still ahead !!!

References:

SBC method

1. R. Marbach, *On Wiener filtering and the physics behind statistical modeling*, J. Biomed. Optics 7, 130-147 (2002)
2. R. Marbach, *Methods to significantly reduce the calibration cost of multichannel measurement instruments*, US Pat. No. *6,629, 041*, 30 Sep. 2003
3. R. Marbach, *A New Method for Multivariate Calibration*, J. Near Infrared Spectroscopy 13, 241 – 254 (2005)
4. R.P. Cogdill and C.A. Anderson, *Efficient spectroscopic calibration using net analyte signal and pure component projection methods,* J. Near Infrared Spectrosc. 13, 119-132 (2005)

Figures of merit

5. A.C. Olivieri, N.M. Faaber, J. Ferre, R. Boque, J.H. Kalivas, H. Mark, *Uncertainty estimation and figures of merit for multivariate calibration,* (IUPAC Technical Report), Pure Appl. Chem. 78, 633-661 (2006)
6. R. Marbach, *Figures of merit for multivariate measurement systems: selectivity and sensitivity*, in preparation

Thank you for your attention!

Ralf Marbach
VTT Optical Instrument Center, www.vtt.fi/oic
ralf.marbach@vtt.fi , Ph. +358 40 770 8375