Interpretation of Regression Vectors:

Past, Present Views

Christopher D. Brown, Robert L. Green

chrisbrown@chemist.com
Ahura Corporation
46 Jonspin Road
Wilmington, MA USA 01887
www.ahuracorp.com

10th International Conference on Chemometrics in Analytical Chemistry (CAC)
Aguas de Lindola, Brazil, September 2006

Divining the Model – Why?

In general, because people want to believe that their math/software is really doing chemistry. More specifically:

- to support/disprove an apparent relationship by an assertion of **sensitivity**, and/or **selectivity**
- for chemical **discovery**

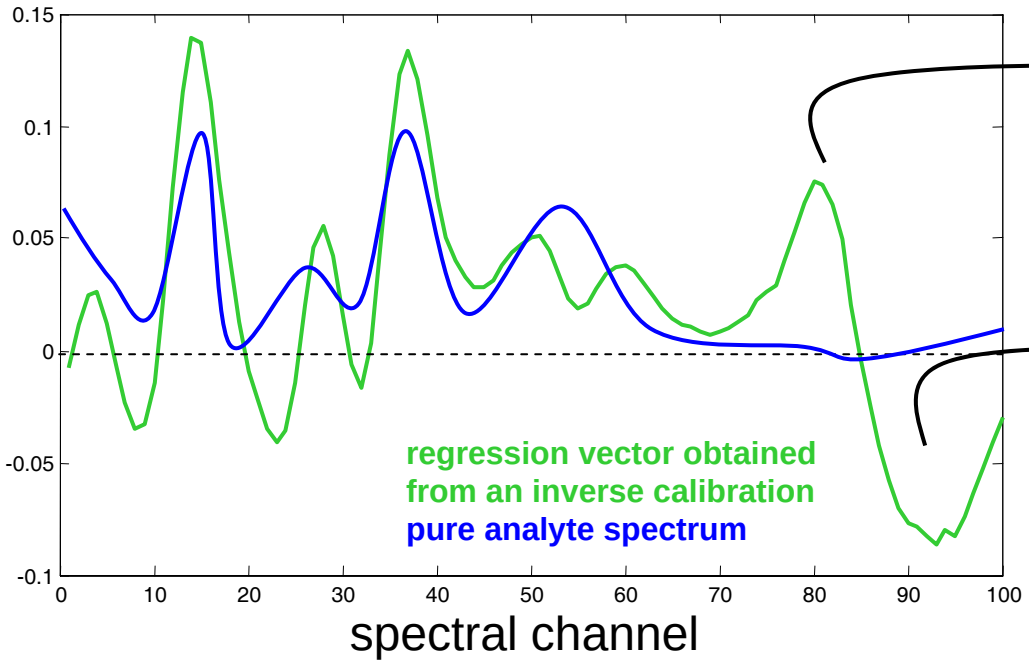
People also do it:

- for wavelength selection (not this talk)
- because people are told to do it (not this talk)

AHURA

Reading the Tea Leaves (a virtual example)*

large positive features should correspond to bands from the analyte, **THEREFORE**
If you don't see them, the model is not sensitive



large positive features not related to the analyte, **THEREFORE** the model is not selective

large negative bands not corresponding to the analyte, **THEREFORE**, the analyte is anti-correlated with an interferent

AHURA

* There are many such examples in the literature, and several guiding, granting or regulatory agencies have encouraged the practice. See Enejder *et al.*, *Journal of Biomedical Optics* **10(3)** (2005) for a recent literature example.

Interpretation of Regression Coefficients

Some Definitions:

"The expected change in the response, per unit change in the variable."

True for univariate regression, but only true for multivariate regression if the variable is orthogonal to all other variables.

"The expected change in the response, per unit change in the variable, if all other variables in the regression are held constant."

More true, if it is indeed sensible to change a variable and keep all others constant. In instrumental data it is usually impossible to consider changing one wavelength while holding all others constant

"The expected change in the response, per unit change in the variable, if the variable and response are linearly adjusted for all other variables in the regression."

True. This is an awful lot of conditionality to keep track of for subjective interpretation.

See Mosteller & Tukey, *Data Analysis and Regression*, Ch. 13 for a broader discussion of the difficulties in interpreting regression coefficients in multiple linear regression.

AHURA

Seasholtz et al.

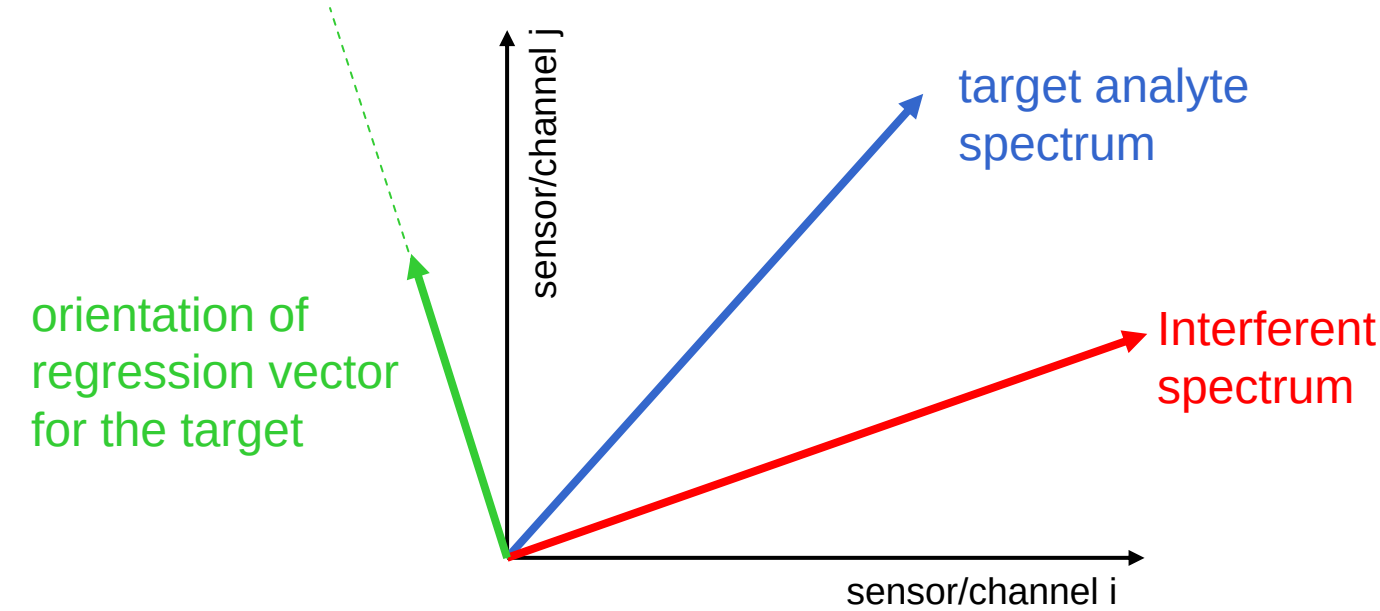
Applied Spectroscopy **44**:1337-1348 (1990)

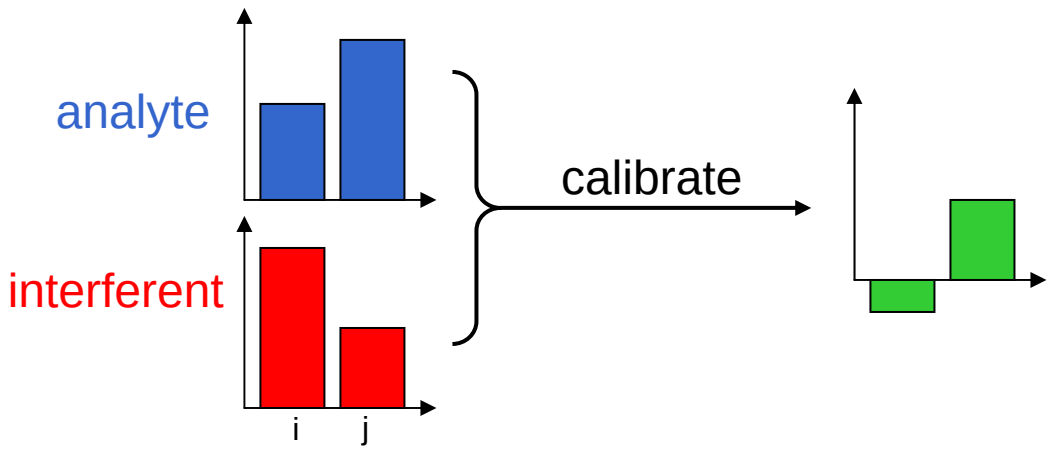
"[Implicit modeling techniques] should be both quantitatively (*e.g.*, cross-validation) and qualitatively validated. Qualitative validation is performed in two steps:
(1) identification of sources of variation that are included in the model and, with that information,
(2) confirmation that the model is including only variance that is chemically meaningful."

(paraphrasing) There are two reasons why care must be taken for the interpretation of regression vectors:
(1) the contravariance constraint (net analyte signal)
(2) they depend on the samples in the calibration

AHURA

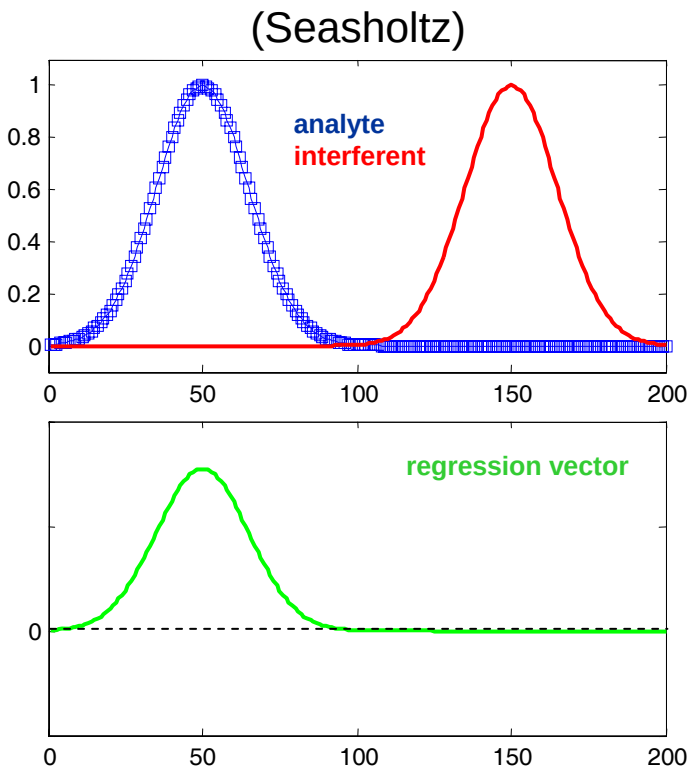
Contravariance Constraint

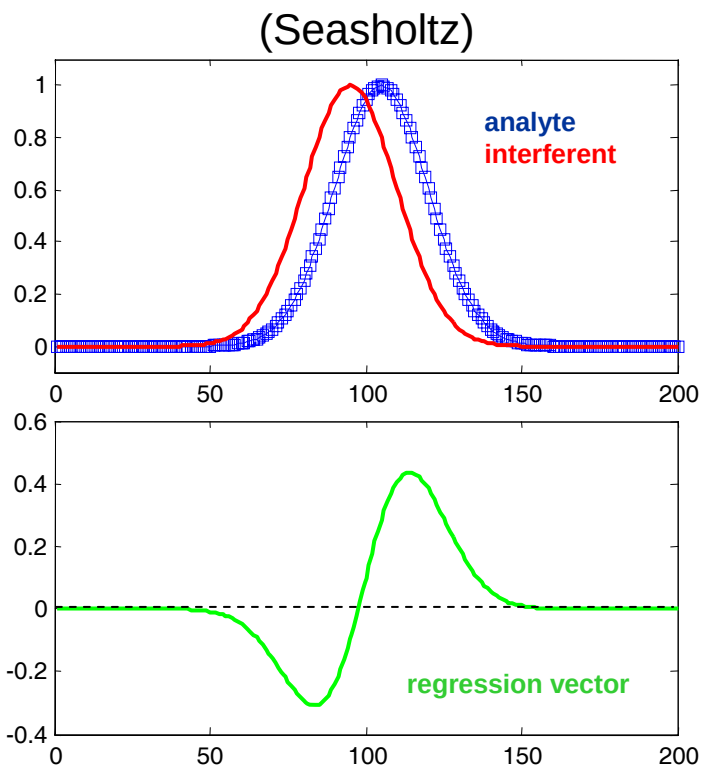


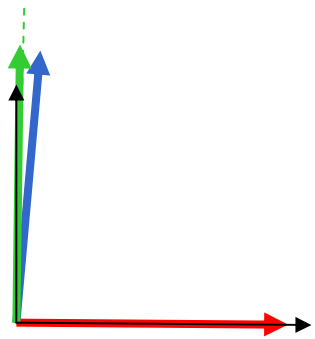


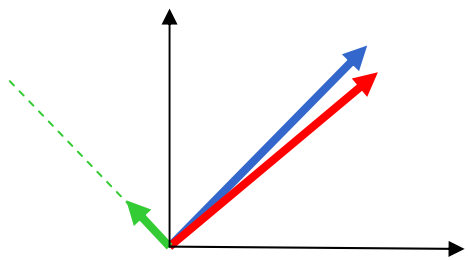
Commentary: A cartoon illustrating the contravariance constraint for a two component system (a vector representation above and the channel graph below). Per the illustration, negative elements in the regression vector arise naturally in the system.

Contravariance (continued)

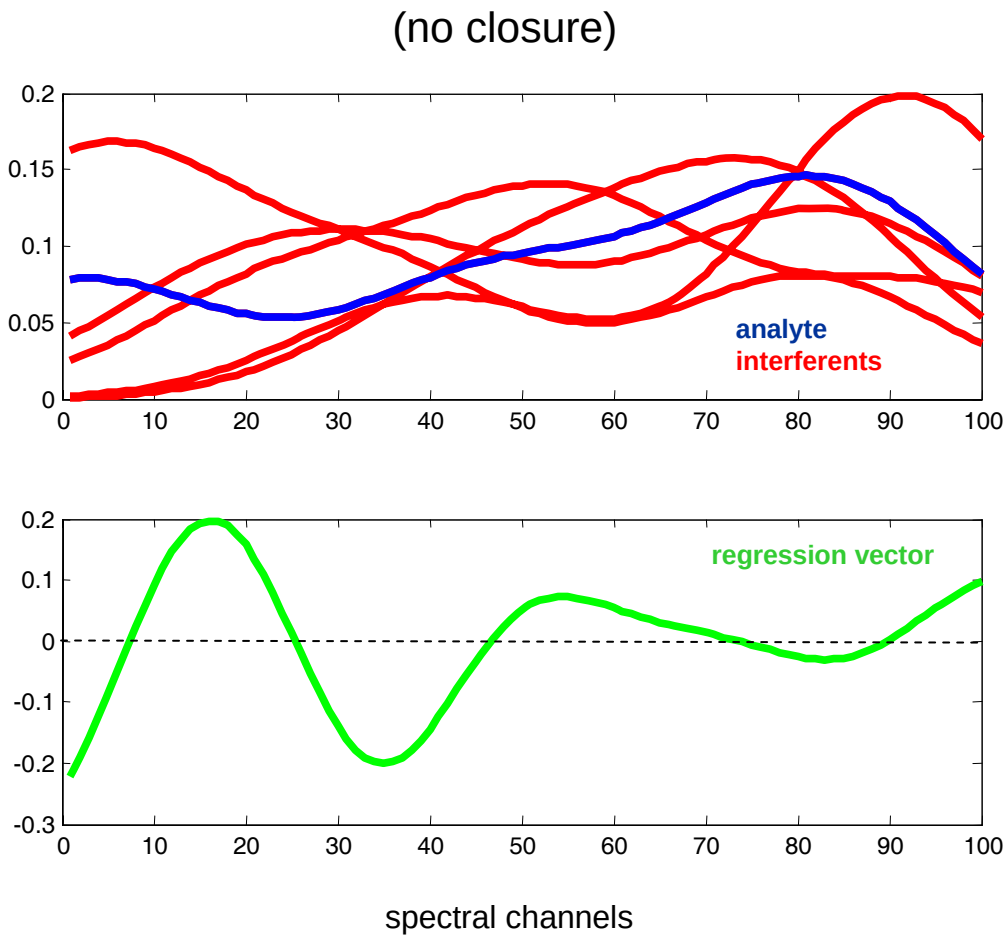








Contravariance (continued)



Commentary: The cases shown by Seasholtz and Kowalski very clearly illustrate the effects in play due to the contravariance constraint. But they are also comparatively simple systems. On this slide we show a (still comparatively simple) six component system. Note that

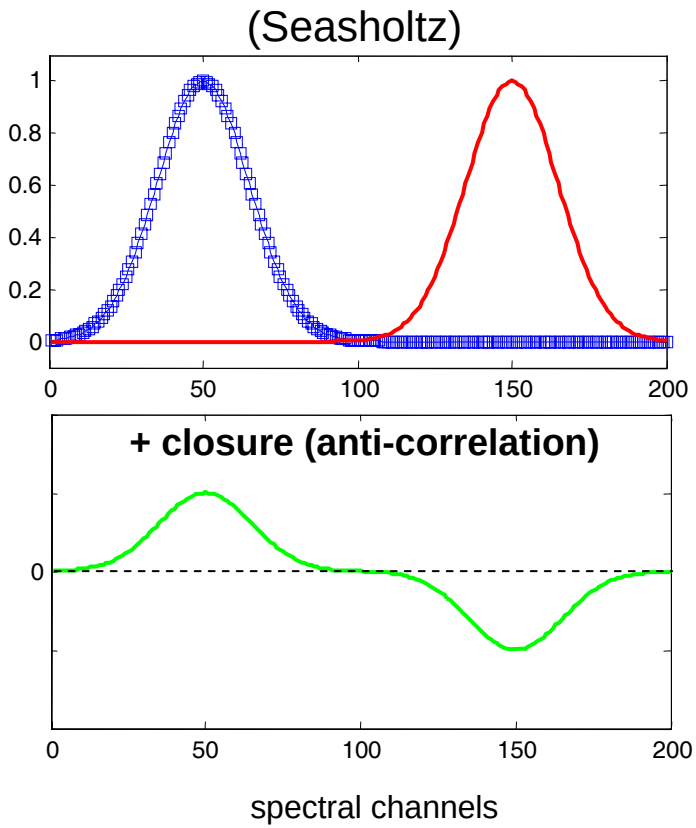
(1) the area of maximum absorption for the analye (around channel 80) carries essentially no weight in the regression vector. In fact the coefficients are slightly negative around channel 85.

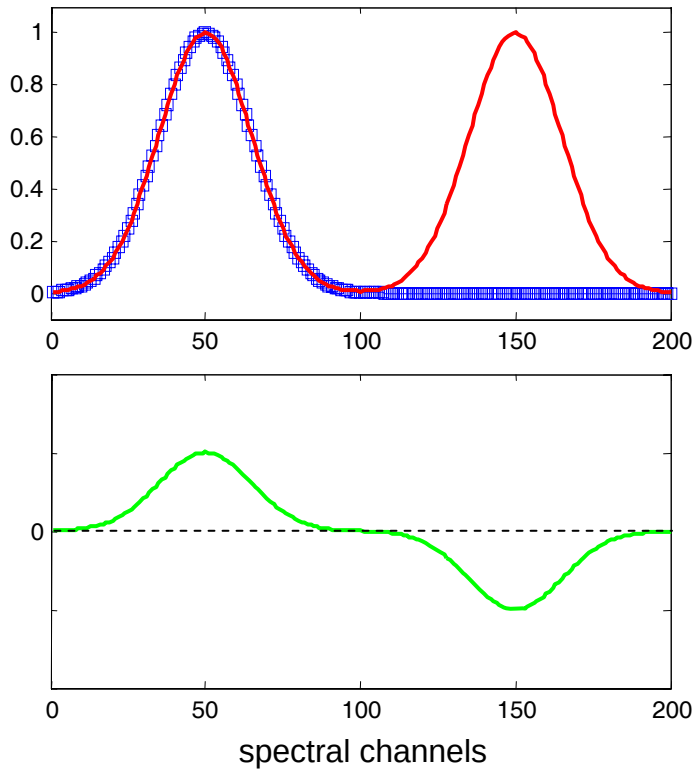
(2) the largest coefficients in the regression vector are aligned with a region of minimum analyte absorption

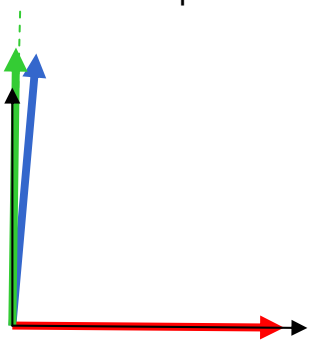
In short, given only the regression vector, and the analyte spectrum, it is impossible to know if things are "proper" even if they are indeed "proper."

AHURA

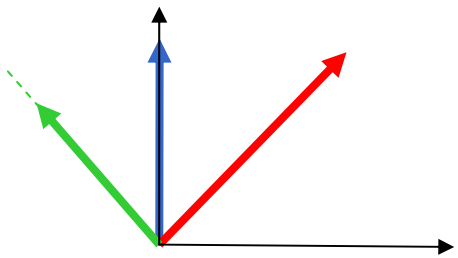
Contravariance (continued)



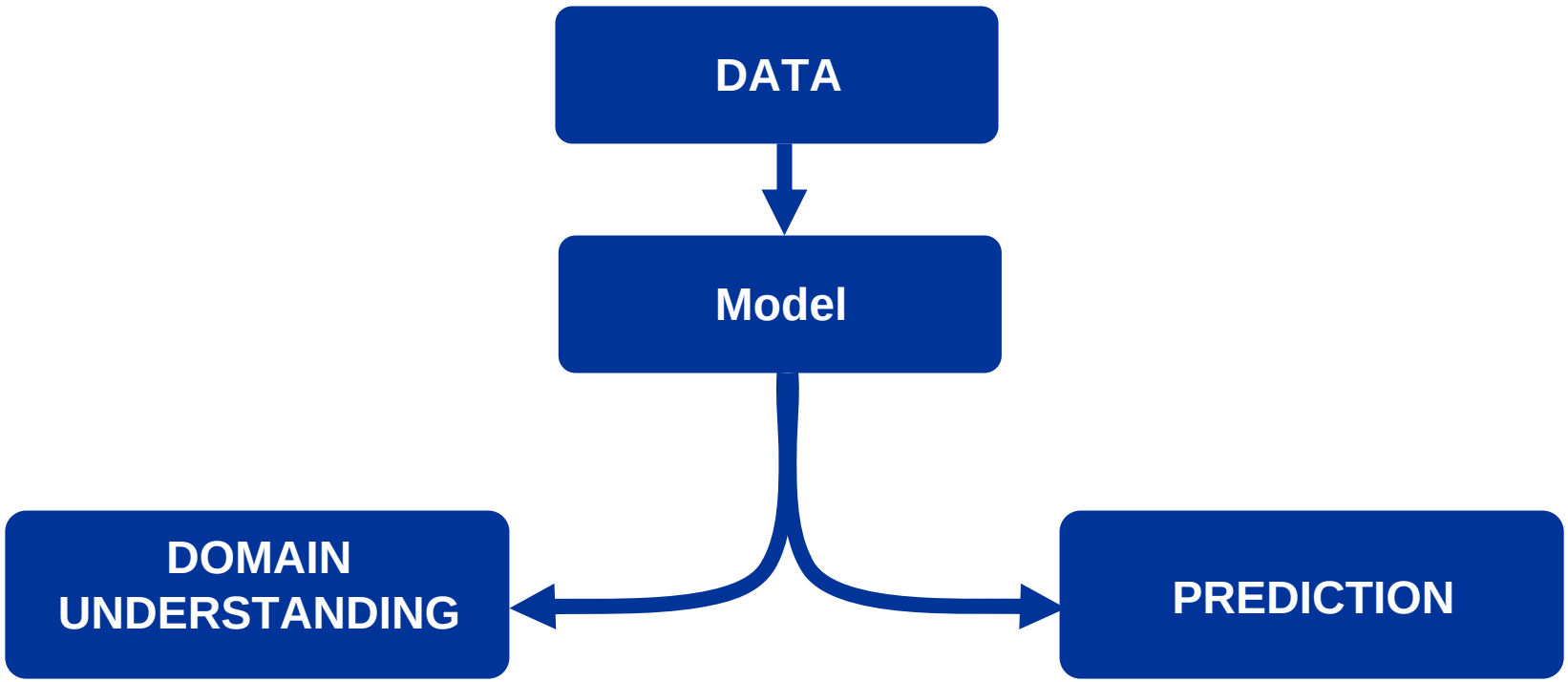




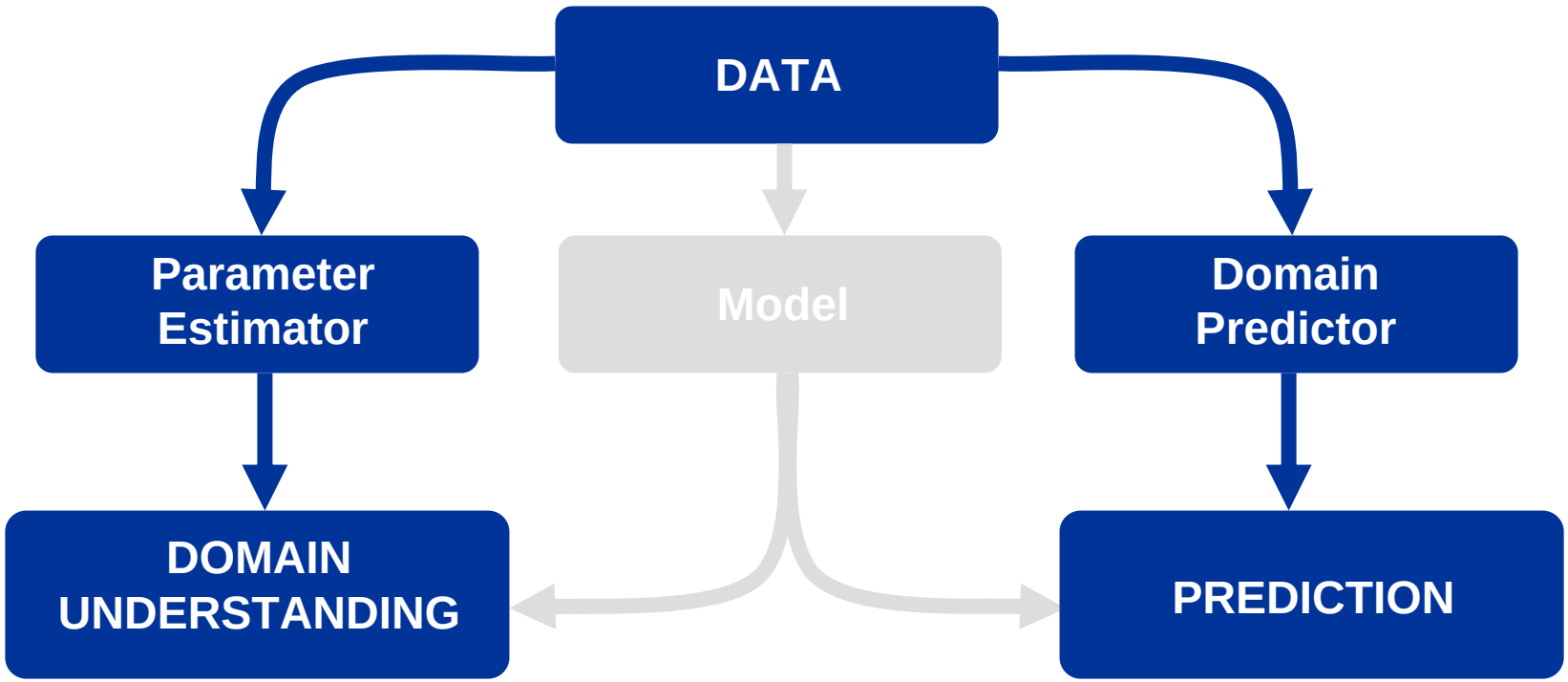
Commentary: In this case we can generate two identical regression vectors with completely different pure-component spectra. The case noted by Seasholtz exhibits closure (concentration covariance) while the classical net analyte signal model can not adequately describe this complication.



Observational Modeling



Observational Modeling

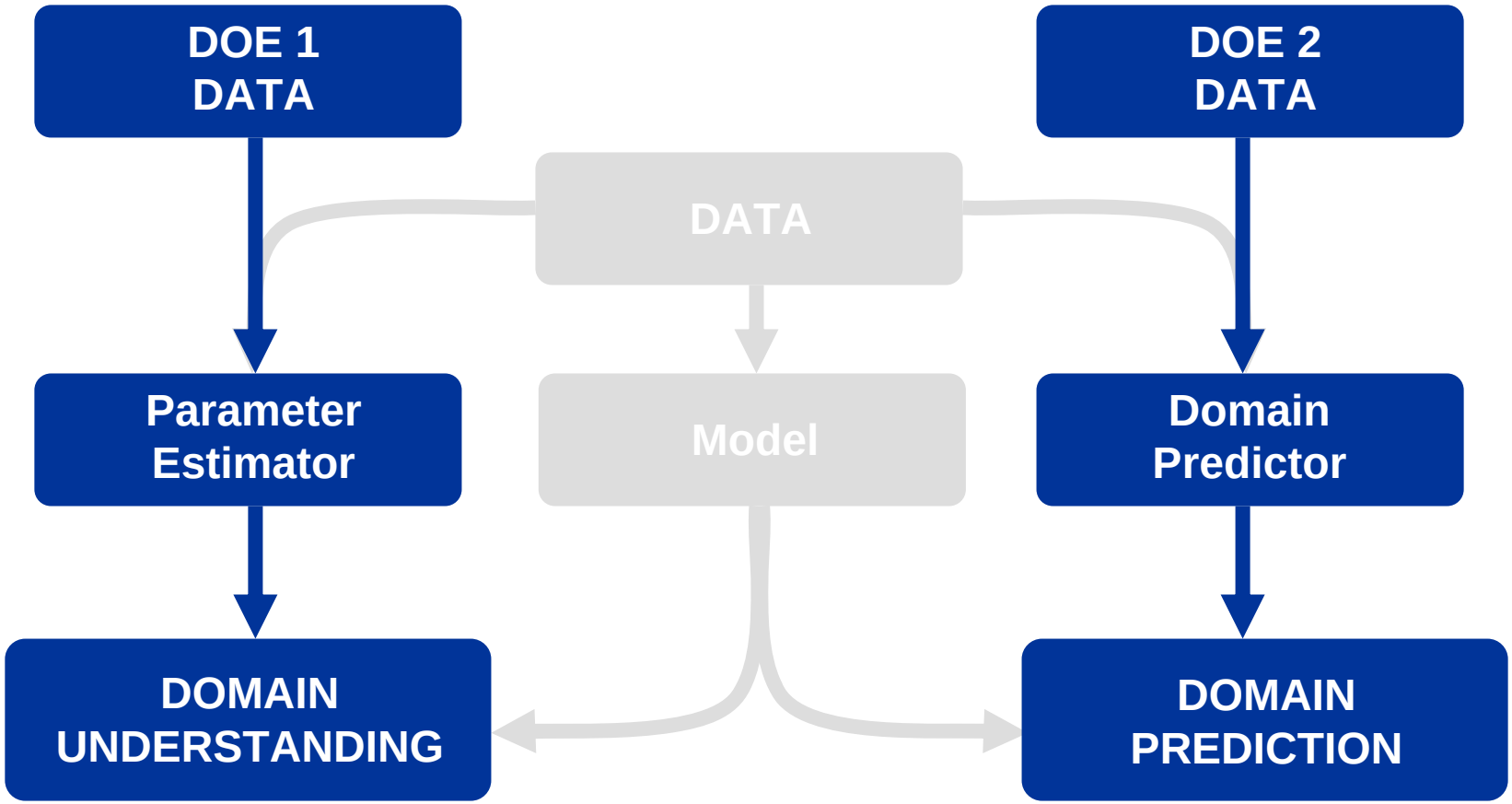


classical calibration (K-matrix)
NAS Theory (Lorber), PCA
curve resolution, EIV

inverse calibration
(P-matrix)

AHURA

Observational Modeling



classical calibration (K-matrix)
NAS Theory (Lorber), PCA
curve resolution, EIV

inverse calibration
(P-matrix)

AHURA

see C.D. Brown, *Anal. Chem.* **76**:4364 (2004) for a univariate example illustrating this fissure between parameter estimators and predictors, and DOE

Classical v. Inverse (BLP) Calibration

$$\mathbf{X} = \mathbf{CK} + \mathbf{E}$$

$$\hat{\mathbf{C}} \Leftarrow \mathbf{XB}$$

'Classical'

Unbiased estimates of $\mathbf{C}$

$$\mathbf{B} = \mathbf{K}^T \left(\mathbf{K}\mathbf{K}^T \right)^{-1}$$

'Inverse' **

minimum MSE estimates of $\mathbf{C}$

$$\mathbf{B}_{BLP} = \left(\mathbf{K}^T \mathbf{\Psi} \mathbf{K} + \mathbf{\Sigma} \right)^{-1} \mathbf{K}^T \mathbf{\Psi}$$

concentration covariance

$\mathbf{\Psi} = E(\mathbf{c}^T\mathbf{c})$

instrument measurement error covariance

$\mathbf{\Sigma} = E(\mathbf{e}^T\mathbf{e})$

AHURA

** C.D. Brown, *Anal. Chem.* **76**:4364 (2004) This paper discusses the divergence between classical net analyte signal theory and inverse methods of multivariate calibration (e.g., MLR, PCR, PLSR). *c.f.* Krutchkoff, *Technometrics* **9**:425 (1967) and the numerous papers in the 10 years that followed for a lively discussion of inverse versus classical prediction; Boaz Nadler has also recently published "small-*n*" results for PLS in *J. Chemometrics* **19**:107 (2005)

Seasholtz et al.

Applied Spectroscopy **44**:1337-1348 (1990)

"[Implicit modeling techniques] should be both quantitatively (*e.g.*, cross-validation) and qualitatively validated. Qualitative validation is performed in two steps:
(1) identification of sources of variation that are included in the model and, with that information,
(2) confirmation that the model is including only variance that is chemically meaningful."

(paraphrasing) **There are two reasons why care must be taken for the interpretation of regression vectors:**
(1) the contravariance constraint (net analyte signal)
(2) they depend on the samples in the calibration and/or the implicit covariance of the components
(3) they depend on the SNR of the data

AHURA

Illustration/Simulation

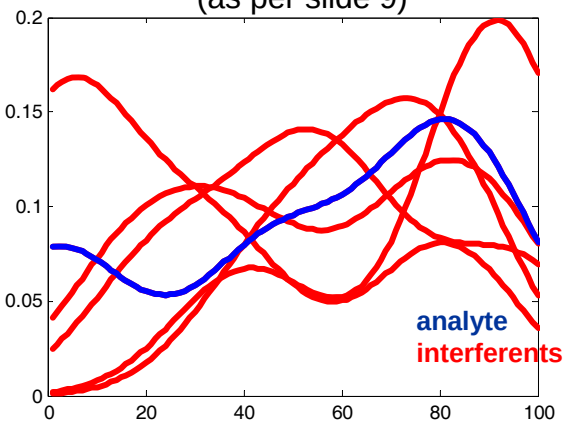
$$\mathbf{B}_{BLP} = \left(\mathbf{K}^T \boldsymbol{\Psi} \mathbf{K} + \boldsymbol{\Sigma}\right)^{-1} \mathbf{K}^T \boldsymbol{\Psi}$$

Simulation: change only σ^2, with all other parameters fixed, and look at the regression vectors that result.

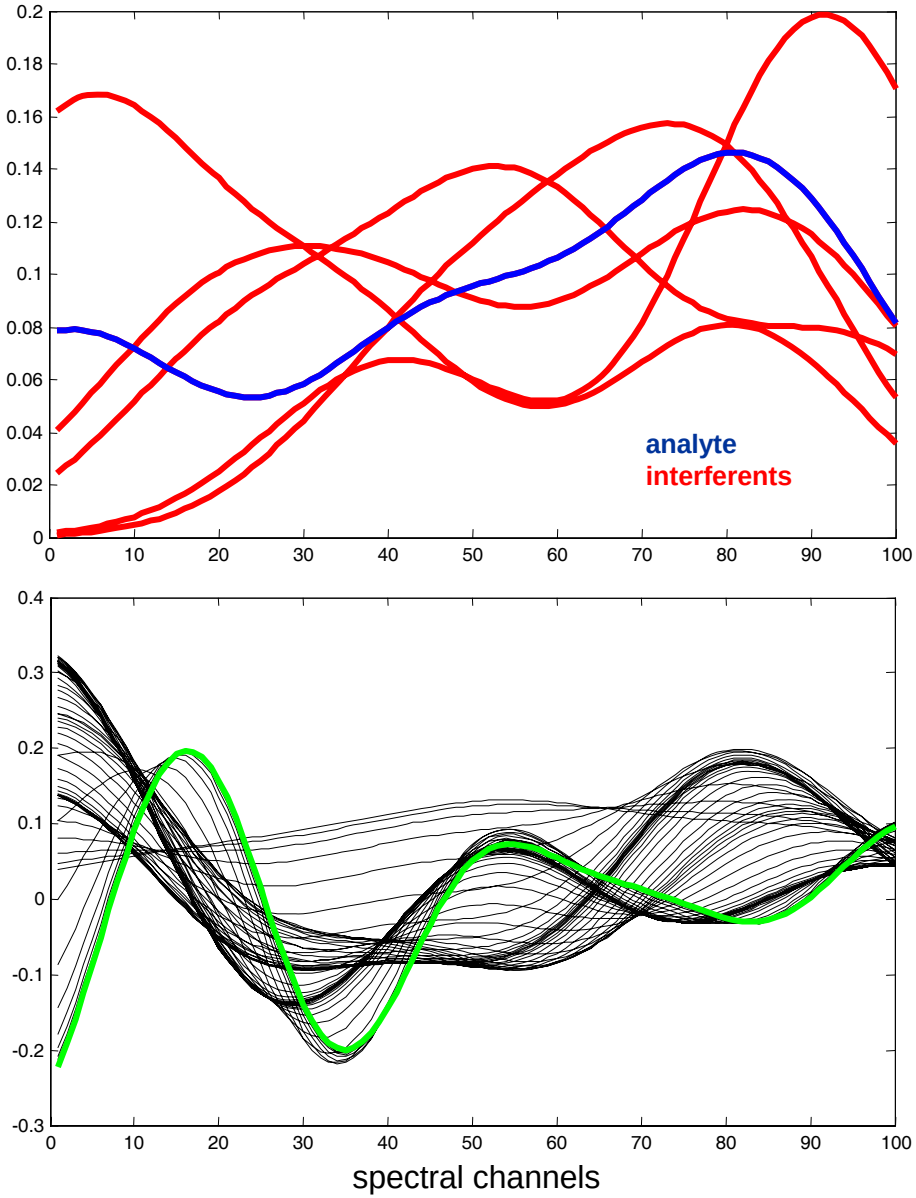
Assume *iid* noise: $\boldsymbol{\Sigma} = \sigma^2 \mathbf{I}$

Designate $\boldsymbol{\Psi}$ as being diagonal and fixed (no concentration covariates).

(as per slide 9)



AHURA



Commentary: Family of regression vectors plotted in black with variable *k*, pure-components, concentrations are fixed in the system. The *NAS* (contravariant/classical solution) is shown in green for reference.

Three major points of relevance are:

(1) the inverse regression vector has no rigid contravariance constraint, so it meanders about to minimize *MSEP* in a given situation.

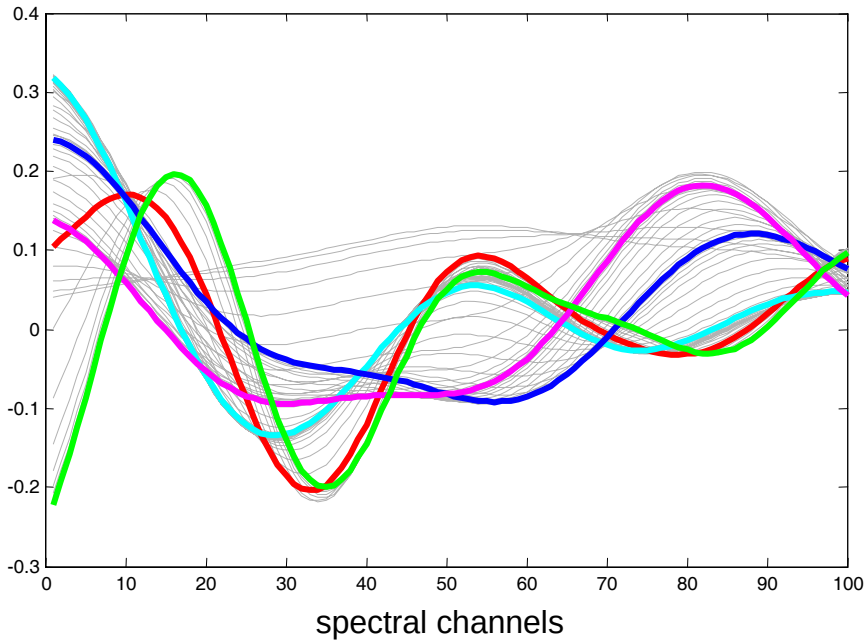
(2) note that the sign of the regression coefficients can easily change, and for some cases a selected coefficient is large, while for other cases the coefficient is near zero.

(3) in many of these regression vectors the largest coefficients in the regression have no correspondence to the largest bands in the absorption spectra of the analyte.

AHURA

Different, but Useful

"Models, of course, are never true, but fortunately it is only necessary that they be useful." *George E. P. Box*, in *JASA* **74**:1-4 (1979)



$R^2(\hat{\mathbf{c}}_{pred}, \mathbf{c}_{ref})$

1.0000
1.0000
0.9988
0.7629

Commentary: These selected regression vectors have widely varying directionality in the multivariate domain, but yet they are all 'useful' in the sense that there is very little difference in their predictive ability (pink being perhaps an exception) .

AHURA

The Alternatives

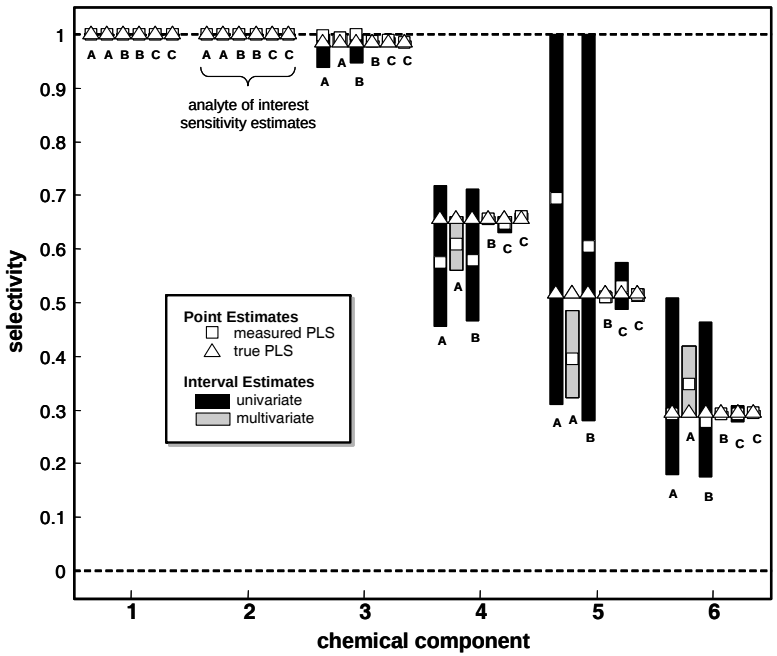
- Use tools that are designed for domain understanding (parameter estimators), and

Figures of Merit

- Sensitivity, selectivity can be assessed quantitatively with analytical figures of merit
  - see A. Olivieri *et al.*, IUPAC Guidelines for Multivariate Calibration Part 3, *Pure Appl. Chem.* and references therein for a current review of the state of the art
  - *c.f.* Vessman *et al.*, *Pure Appl. Chem* **73**:1381-1386 (2001); K. Danzer, *Fres. J. Anal. Chem.* **369**:397-402 (2001)
- On the next slide we show an example of how sensitivity and selectivity can be evaluated quantitatively using figures of merit
  - C.D. Brown *et al.*, *Appl. Spec.* **59**:787-803 (2005)

AHURA

Component-Specific FOM Estimates



component	PLS-25 SEL	PLS-25 var (%)	PLS-500 SEL	PLS-500 var (%)
1	0.999	0.319	0.999	0.041
2 (sensitivity)	0.999	1.148	0.999	0.667
3	0.985	3.796	0.987	6.029
4	0.655	33.890	0.999	0.088
5	0.517	15.019	0.748	10.852
6	0.295	38.480	0.454	38.682
$b^T\Sigma b$		7.349		43.640
<chemical MSEP>		0.586 (92.7%)		0.249 (56.4%)
<total MSEP>		0.632		0.441

Commentary: This analysis was conducted on the regression vector colored cyan in slide 18. This definition of selectivity is component-specific so information is provided separately for each analyte / interferent. It is also purely a *predictive* figure of merit in that it simply characterizes the outcome of the calibration experiment, treating the resultant regression vector as a constant. The Var (%) columns in the table give the variance components of the *MSEP* for the analyte of interest (component 2), which directly quantify how much of the *MSEP* is attributable to each chemical component in the system, as well as the disturbance term ($b^T\Sigma b$). The direct *predictive* interpretation means that if one could eliminate a component from the system (or hold it constant) in prediction, the *MSEP* for the analyte of interest would drop by the corresponding % while using the exact same regression vector.

With a 25 sample calibration set (PLS-25), we can see the non-selectivity of the predictor contributes almost 93% of the total *MSEP*, and we can determine whether the model is selective against specific interferents, as well as sensitive to the analyte. The PLS-500 model (500 calibration samples) is still quite non-selective against interferences – about 56% of the *MSEP* is due to interfering components, a result which is very close to theoretical values from the optimal inverse predictor (equation on slide 13).

AHURA

This example is from C.D. Brown *et al.*, *Appl. Spec.* **59**:787-803 (2005); simulated data and theory given in Part 1, experimental examples from NIR spectroscopy applications in Part 2

Summary

- Regression vectors are highly compressed representations of the data, honed for a specific task
- It is extremely difficult to infer properties such as the matrix composition, selectivity or sensitivity of predictors from the regression vector (and analyte spectrum) alone
- If one is modeling for predictive purposes using inverse methods, the situation can be greatly complicated
- Many tools have matured, or are maturing for fully characterizing the properties of interest
  - Data discovery tools and DOE
  - Figures of merit

AHURA