Principal Component Analysis: Getting an Edge on EXAFS

Stephen R. Wasserman

MediChem Life Sciences

Acknowledgments

- INE-FZK
  - Melissa Denecke
  - Kathy Dardenne
  - Jorg Röthe
- Anatoly Frenkel
  - Irit Sagi
  - Oded Kleifeld
- Spectra obtained at HASYLAB, NSLS, and APS
- BP
  - Gerry Zajac
  - David Sinclair
- FZ-Rossendorf
  - Andre Rossberg
- State of Illinois Department of Commerce and Community Affairs

Mixtures

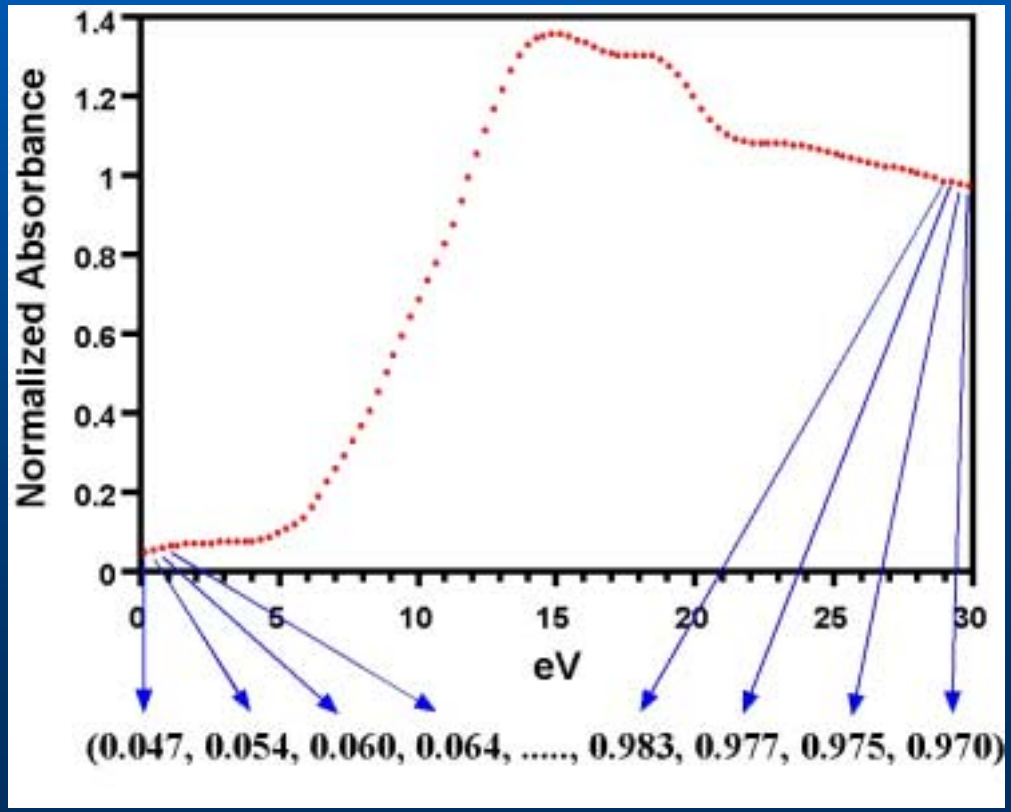
- **Series of XAS spectra: may or may not be pure**
- **Traditional approach to analysis:**
  - **1) Choose pure model compounds or simulate with *ab initio* calculations**
  - **2) Fit to these standards**

The Problem

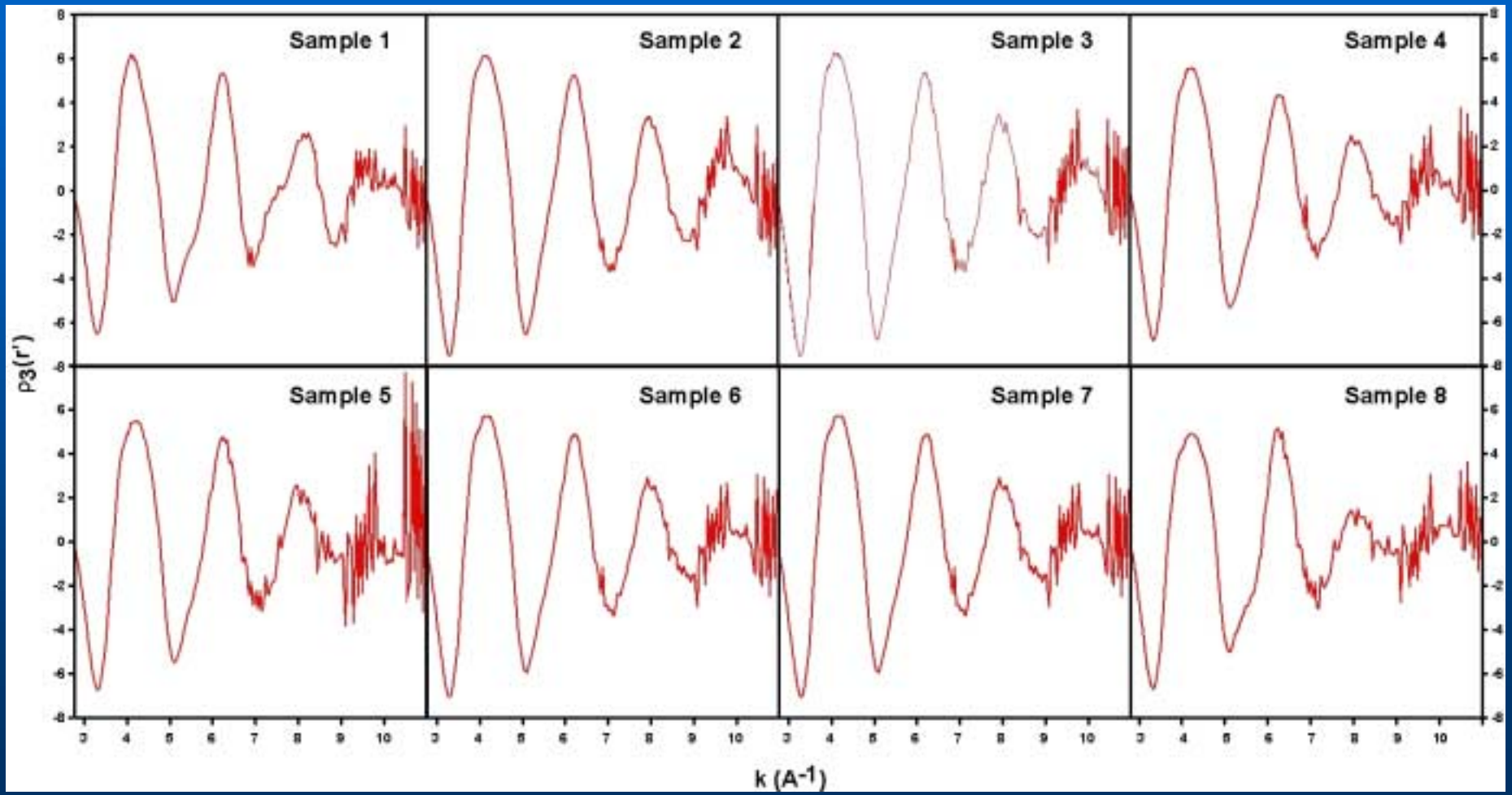
- How many standards are needed?
- How do we know the models are reasonable?
- "If you have the wrong group of standards..., there is no way you can get the right answer"

Vectors

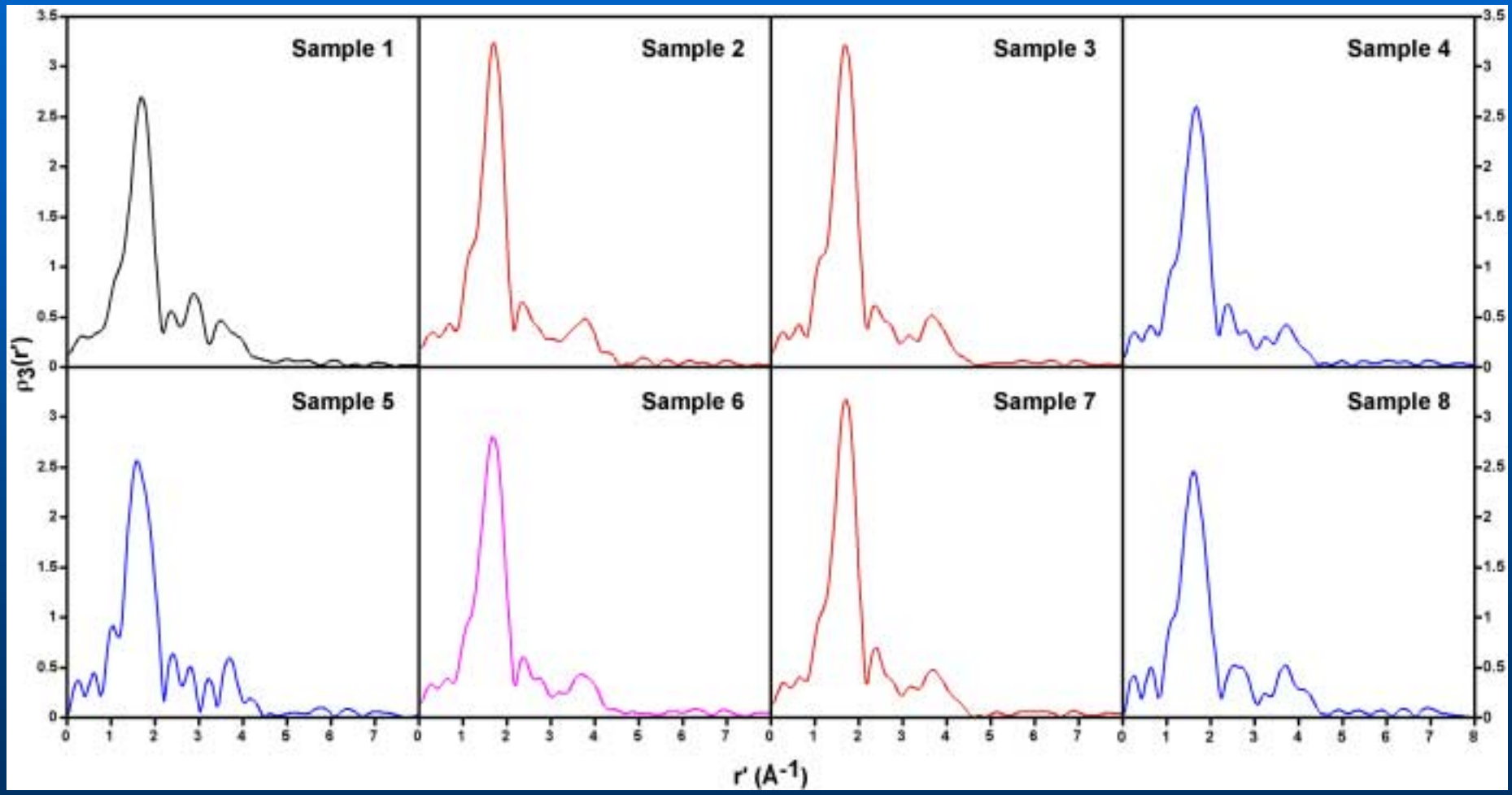
- Each spectrum can be represented as a vector
- Interpolated to same abscissa
  - Energy
  - Wave vector



Catalyst


Sample 1
Sample 2
Sample 3
Sample 4
Sample 5
Sample 6
Sample 7
Sample 8
ρ3(r')
k (A-1)

Catalyst


Sample 1
Sample 2
Sample 3
Sample 4
Sample 5
Sample 6
Sample 7
Sample 8
ρ3(r')
r' (A-1)

Artificial Standards

- **Mathematical constructs**
- **Derive from the original spectra (not separate standards)**
- **No two standards reproduce the same features**
- **Weighting factor to measure how important each artificial standard is in reproducing the entire series**
- **Unique**

Linear Algebra

- **Vector space**
- **Orthogonal**
- **Eigenvalue**
- **Normalized eigenvector**

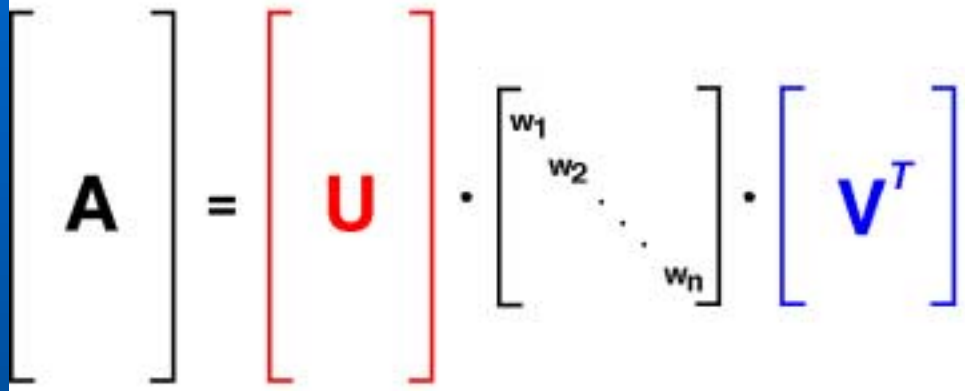
- **PCA terminology: components**

Components

- Number of derived components equals number of original spectra
  - Not equally important (weighting factor)
- Use of all the components will reproduce the original spectra EXACTLY (including experimental noise)
- Most important components (large weighting factor) contain real spectral features
- Least important components (small weighting factor) represent noise and other errors
- Each experimental spectrum contains both physically meaningful data and noise

Singular Value Decomposition

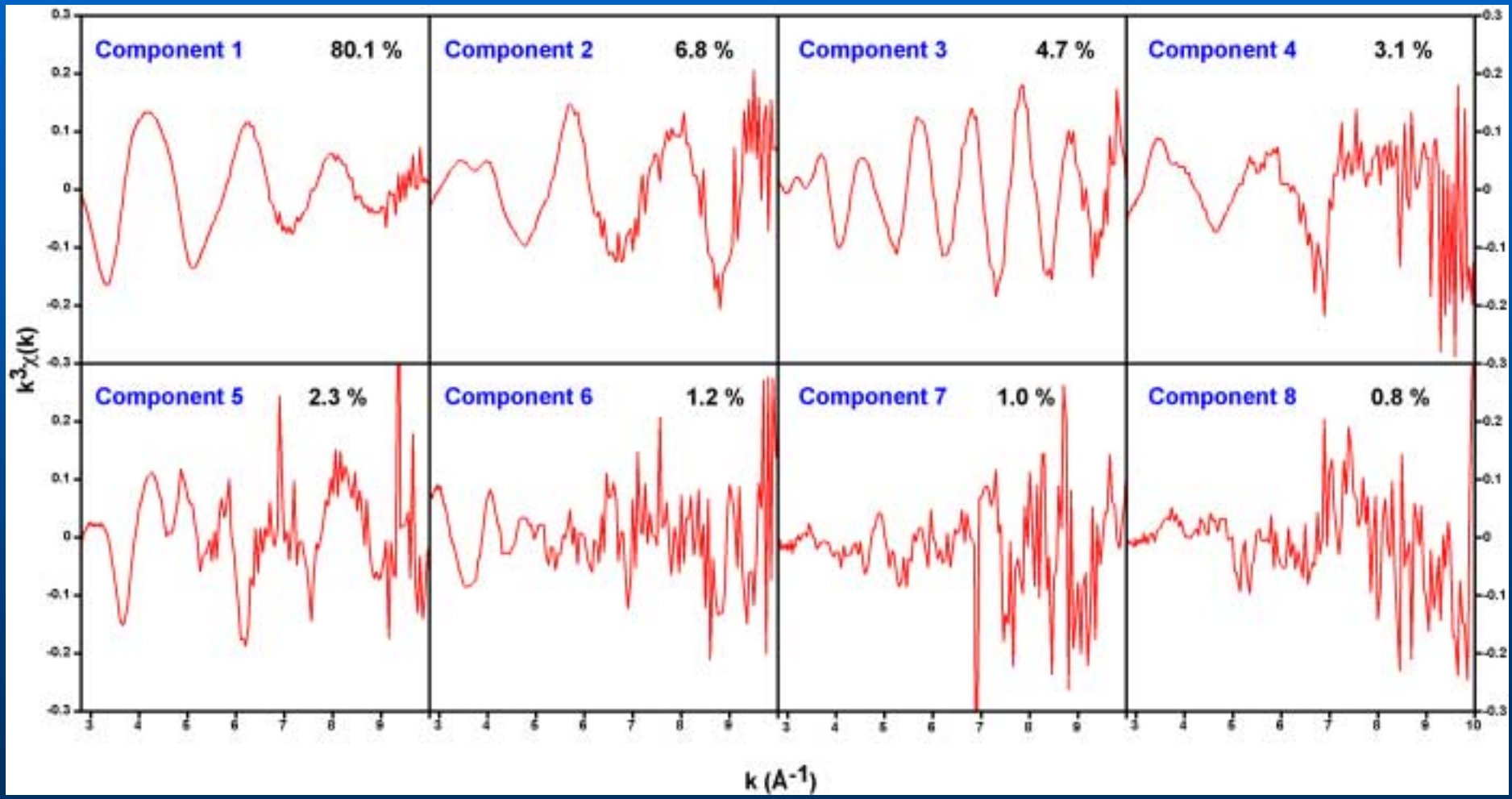
Matrix with M rows and N columns



M x N	M x N	N x N	N x N
	Columns orthogonal	Diagonal (positive or zero)	Orthogonal (square matrix)
Data	Components	Eigenvalues	Weights
		(importance for entire series)	(importance for an individual spectrum)

- PCA assumes data matrix is "effectively" singular

Components: Catalyst



Heart of PCA

- **Goal: Find number of components that can reproduce the spectra to within experimental error**
- **Number of used components equals the number of pure species in the spectra**
- **Theorem: If a vector space has one basis with a finite number of elements, then all other bases are finite and have the same number of elements.**

Primary Components

- Correspond to meaningful data
- Determination:

  1) Error analysis

  2) Reproduce the experimental data with components

  3) For EXAFS, examine the Fourier transform

  4) Amplitude of component times weighting factor

Secondary Components

- Not used in reproduction of data
- Weighting factors describe the contribution of a component to ENTIRE series
- Sum of weighting factors from secondary components represents errors
- Examine this sum as a function of number of primary components

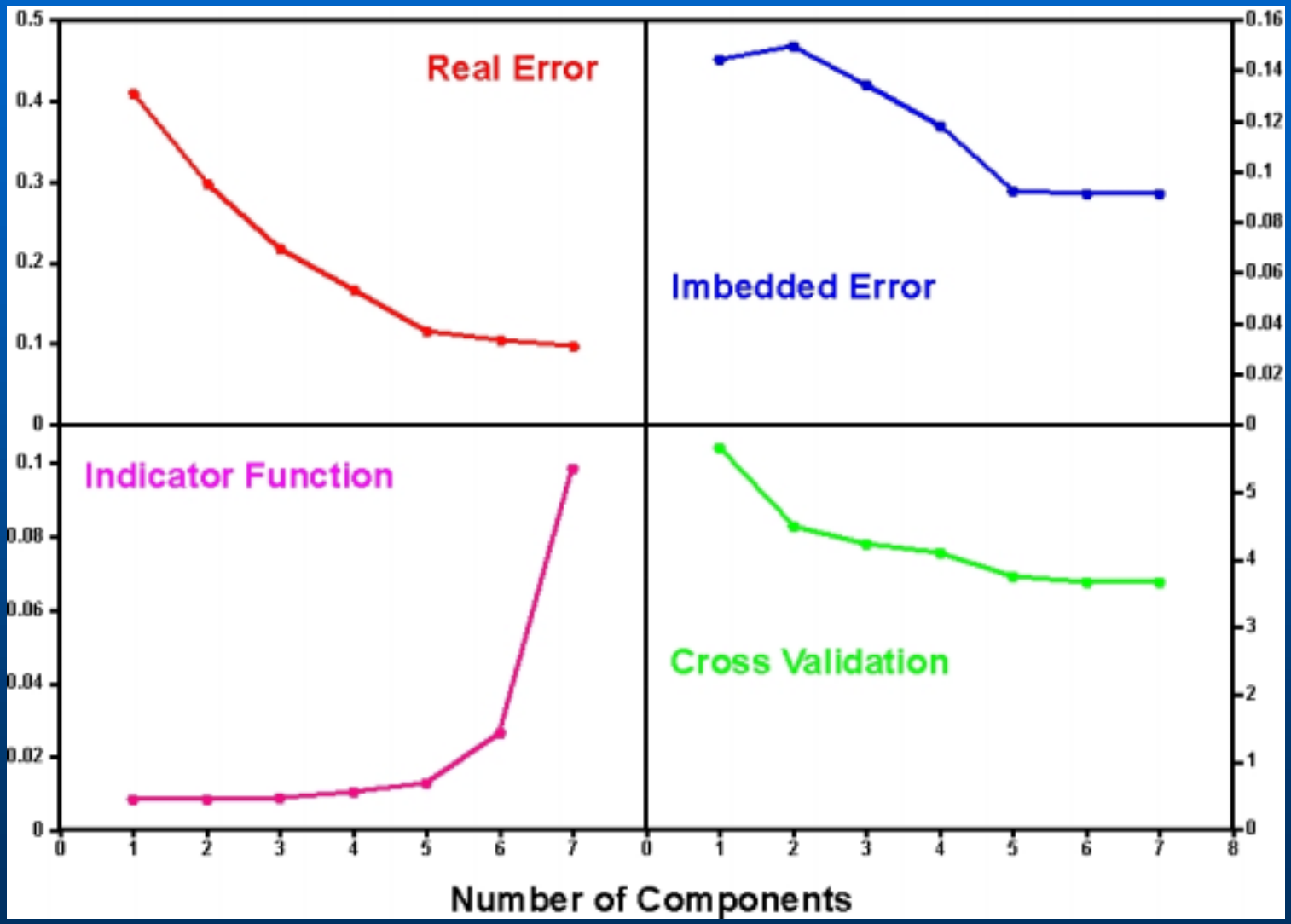
Error Analysis

- **Real Error (RE)**
  - Difference between pure data and experimental data
  - Discontinuity as a function of number of primary components: change in type of information in components from data to noise
- **Extracted error (XE)**
  - Error separated from data by PCA
- **Imbedded Error (IE)**
  - Error present in primary components
  - Discontinuity

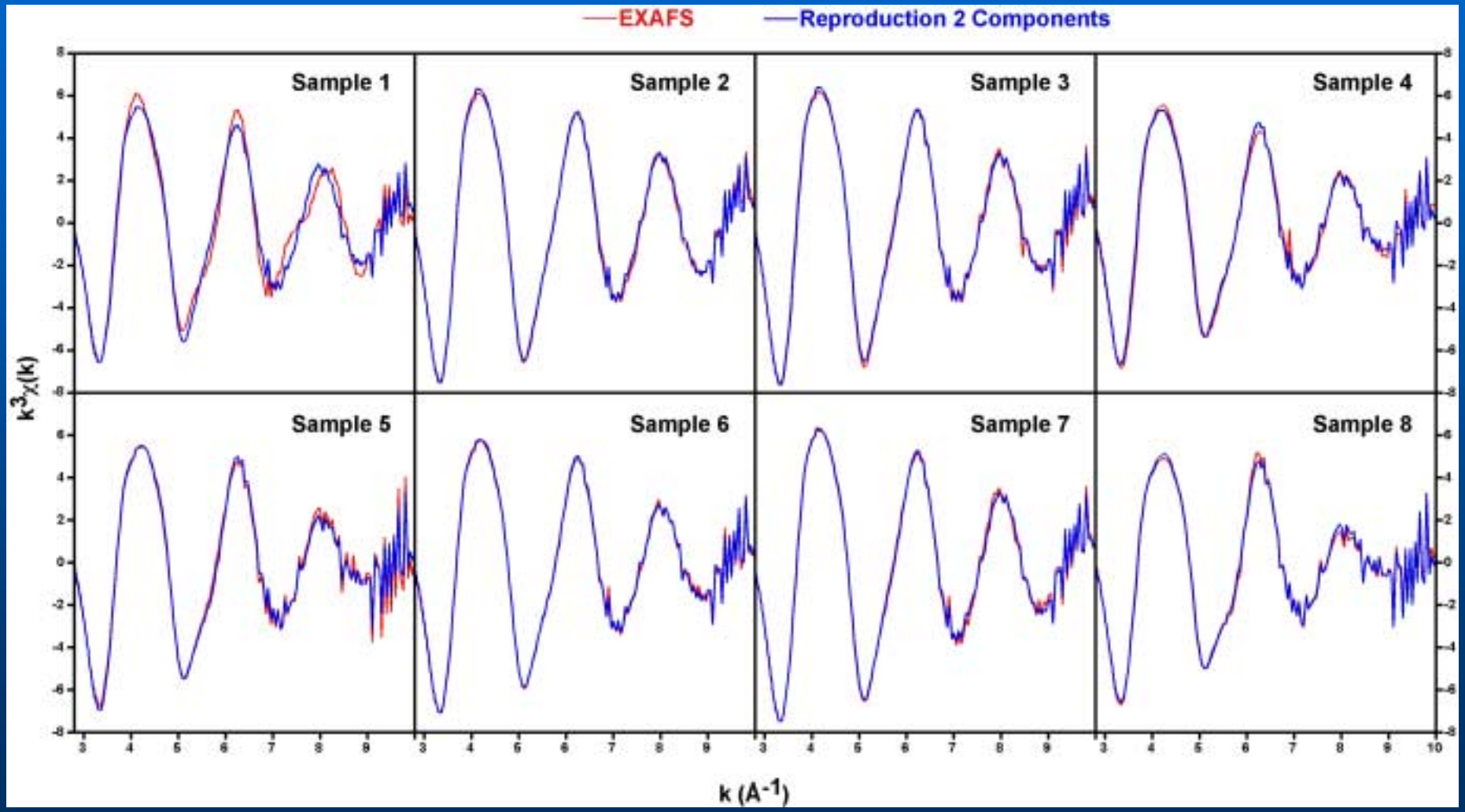
Error Analysis

- **Indicator Function**
  - **Penalty for inclusion of more primary components**
  - **Minimum**
- **Cross-Validation**
  - **Leave one out**
  - **Less sensitive to noise**
  - **Discontinuity**

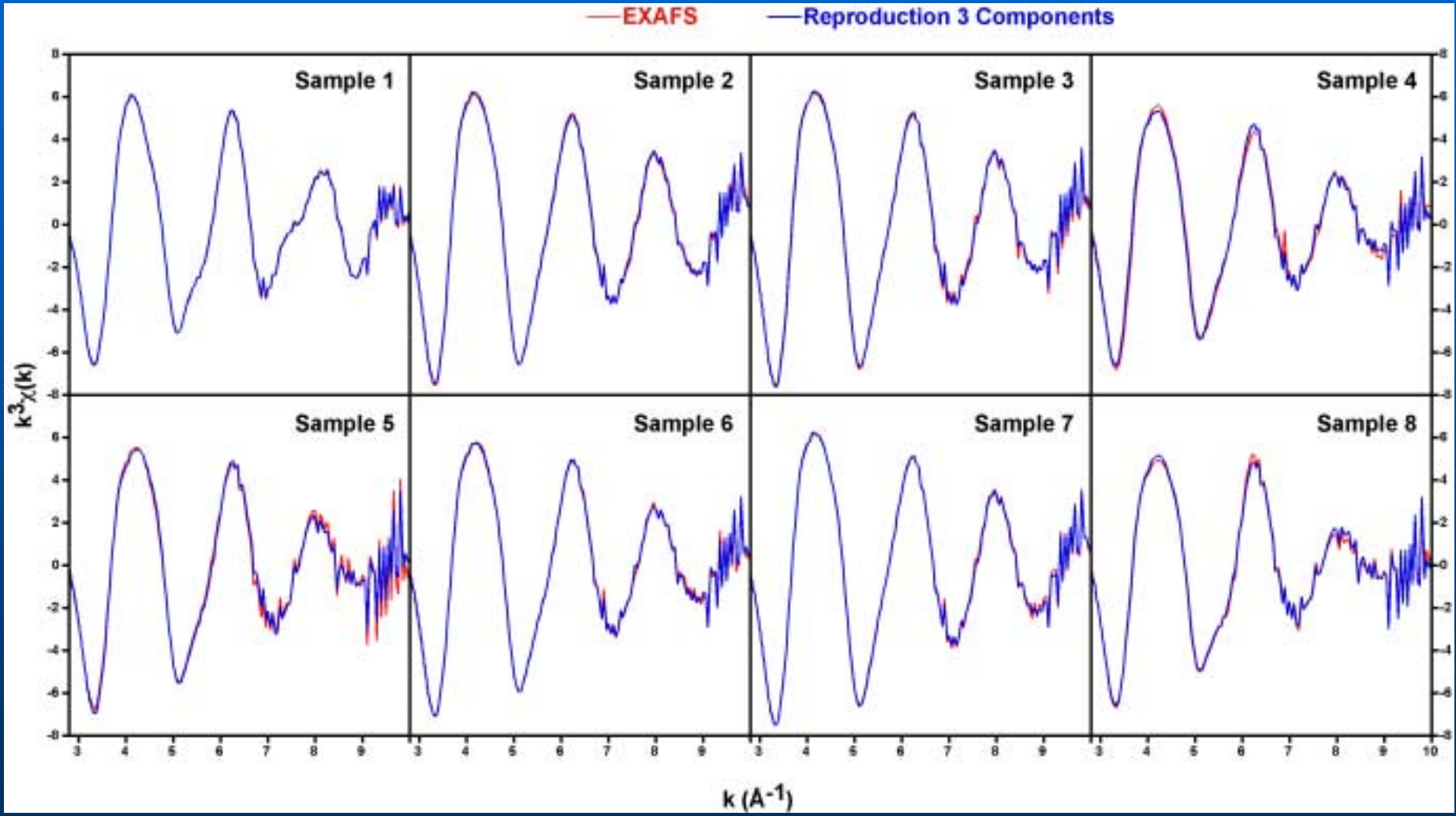
Errors


Real Error
Imbedded Error
Indicator Function
Cross Validation
Number of Components

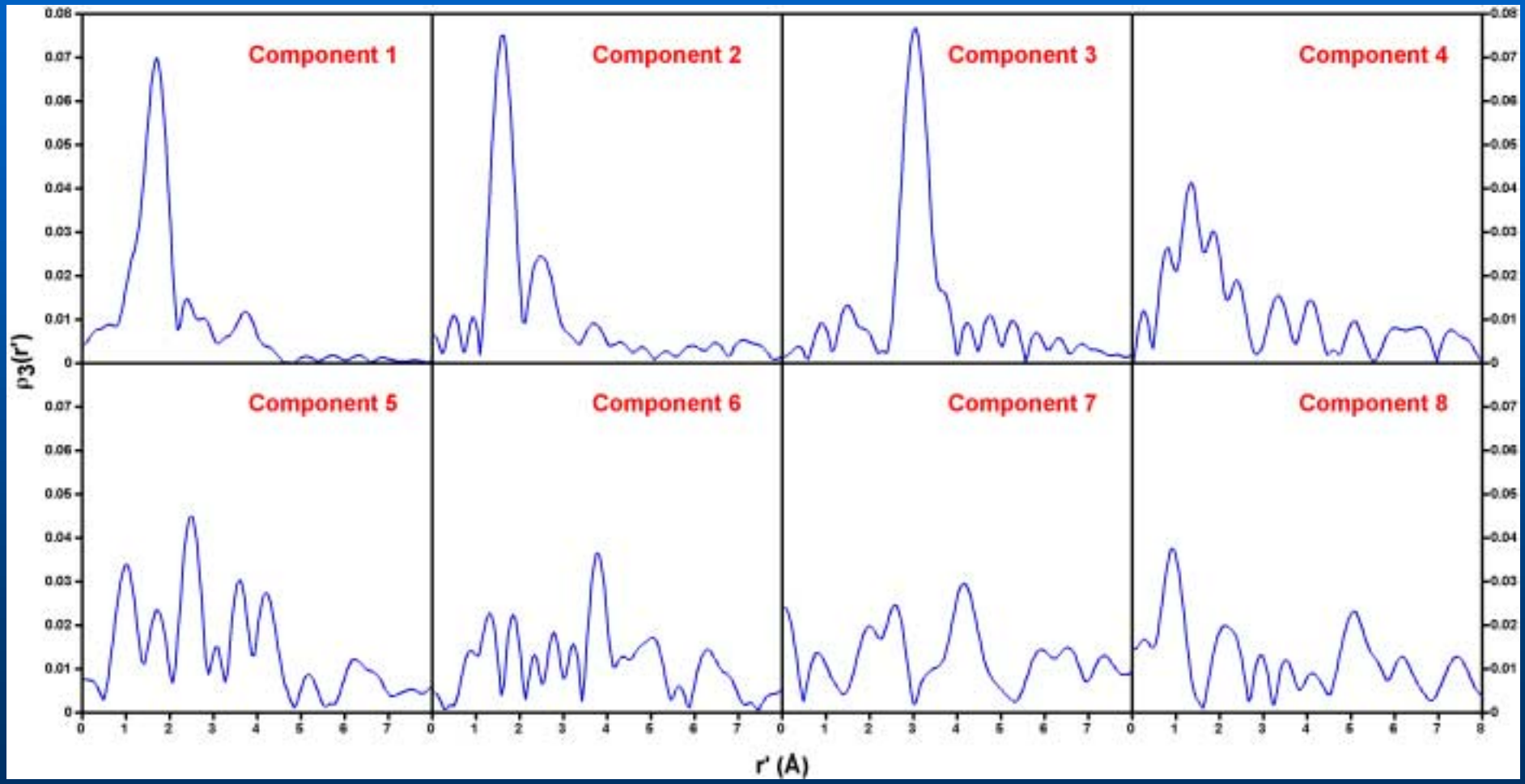
Reproduction: 2 Components



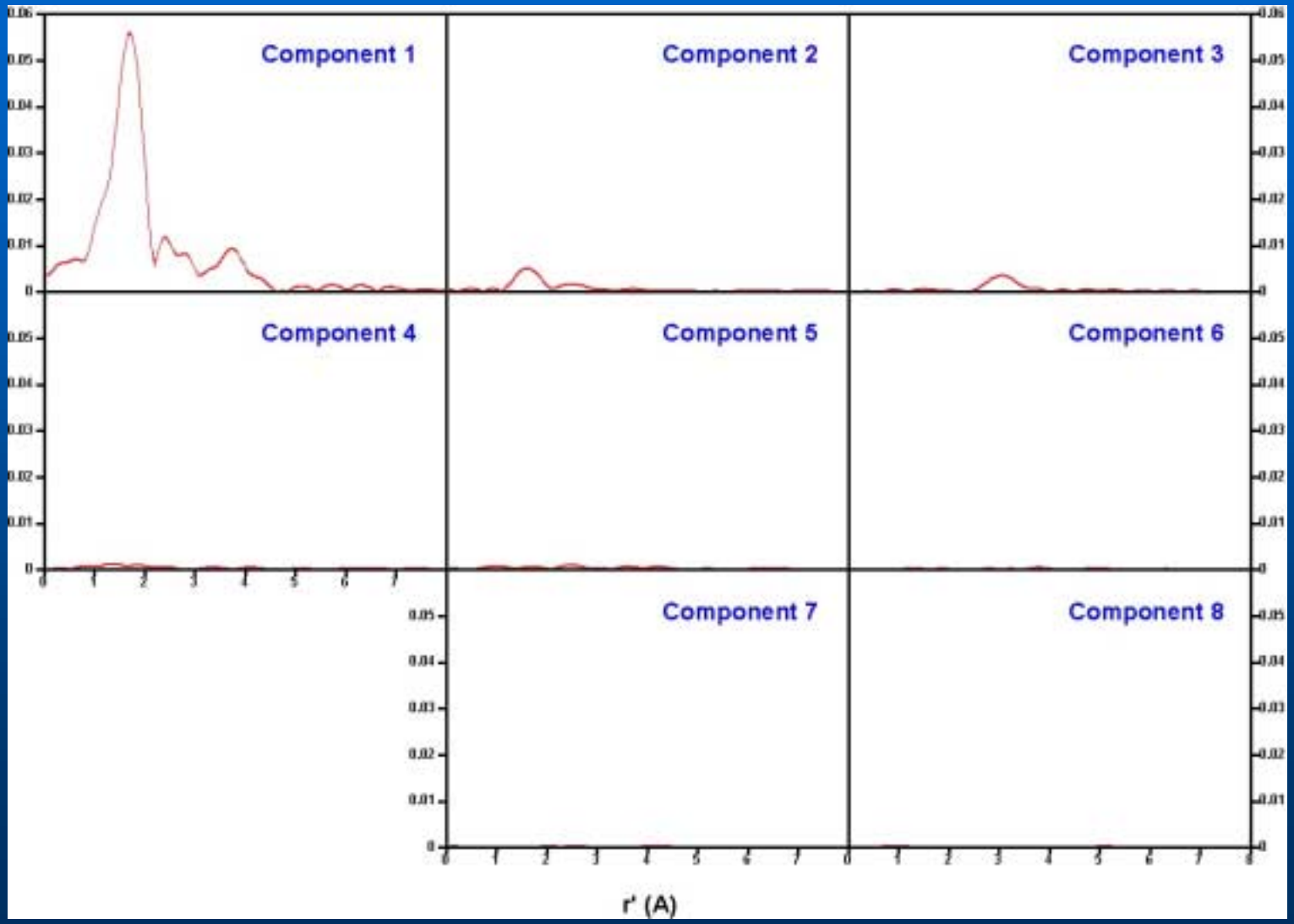
Reproduction: 3 Components



Components: Fourier Transform

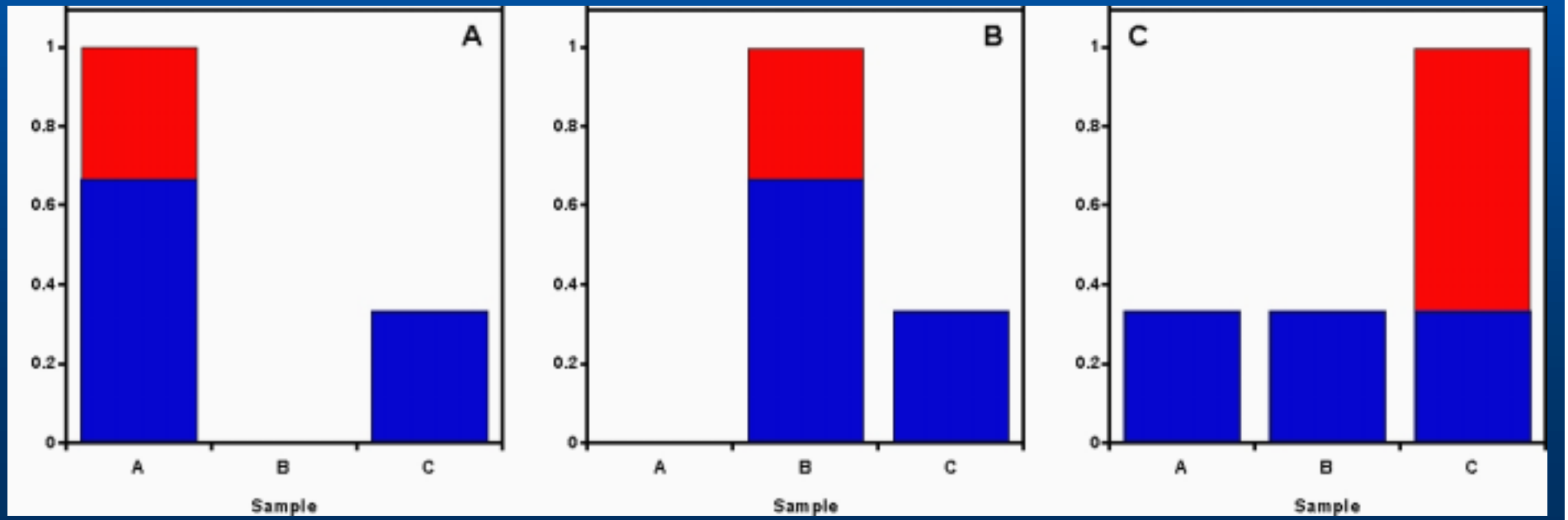


Weighted Components

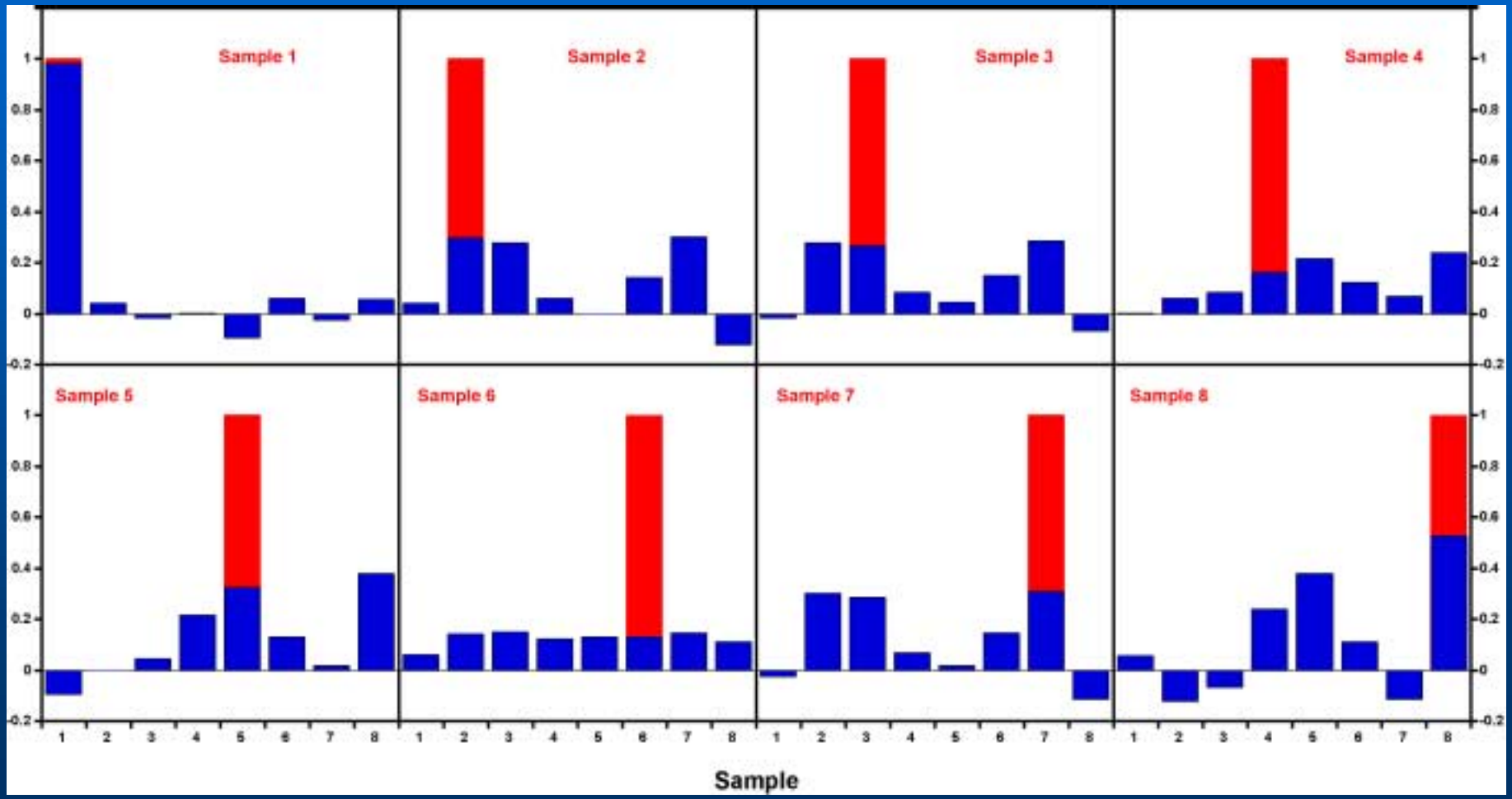


Common Species: Uniqueness Tests

- 3 Samples
  - A: pure
  - B: pure
  - C: Equal mixture of A and B



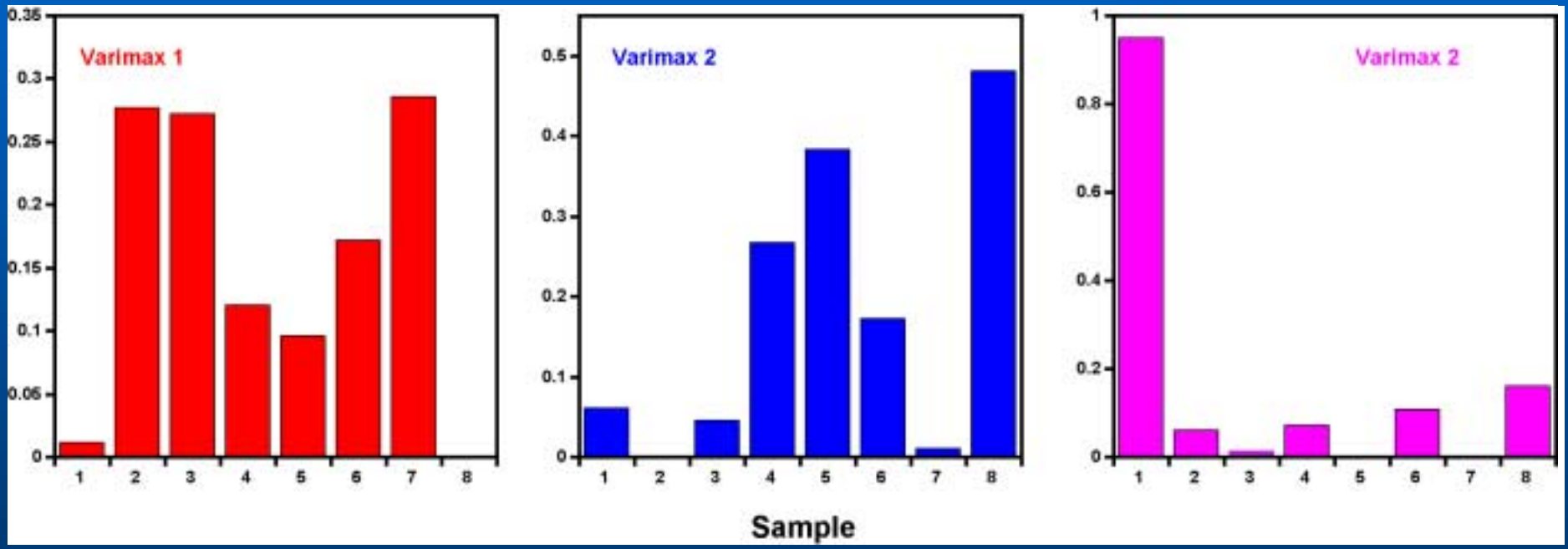
Catalyst: Uniqueness Tests



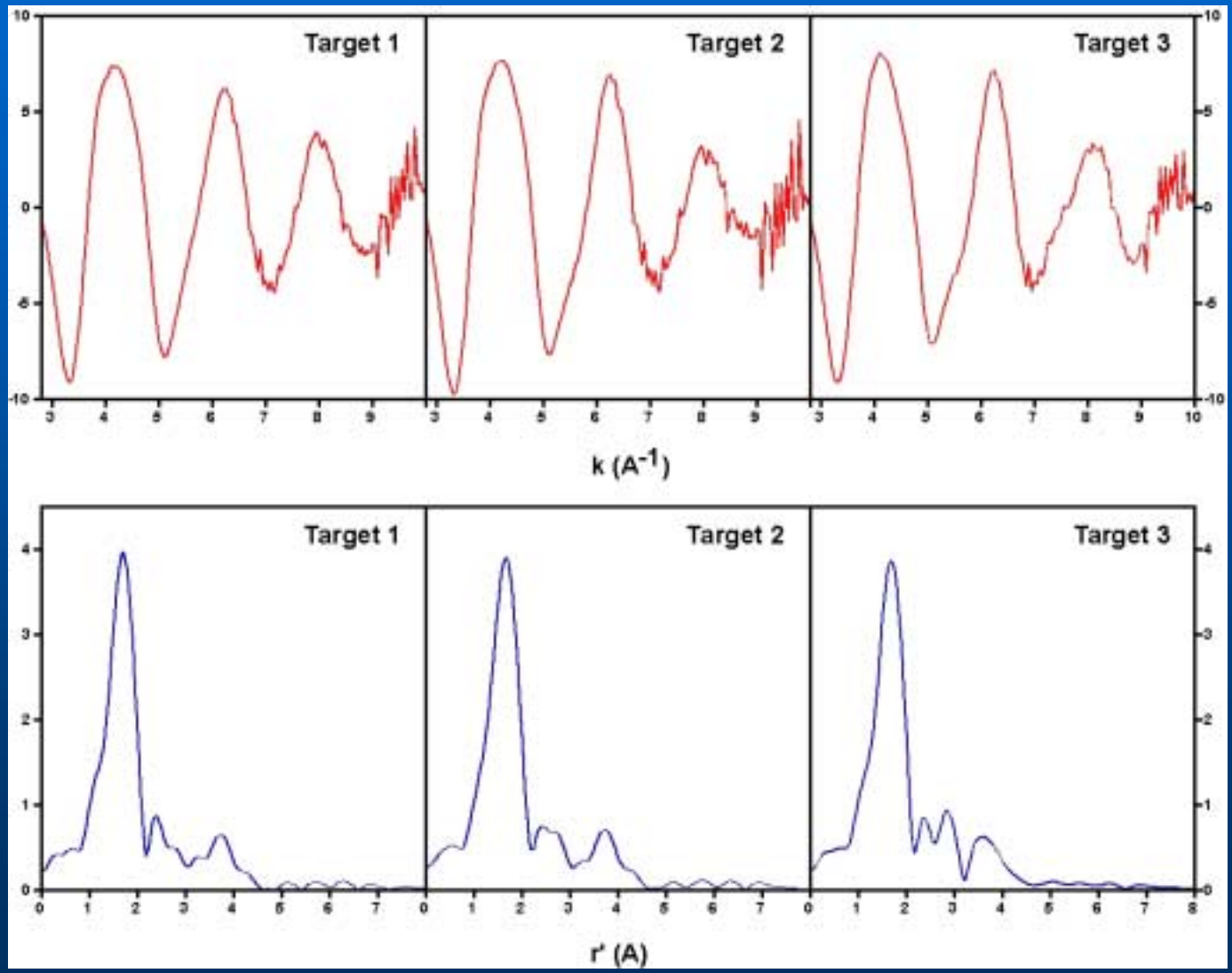
Finding the Underlying Spectra

- **Rotation**
  - **Orthogonal**
    - **Varimax**
  - **Oblique**
    - **Target Transformation**

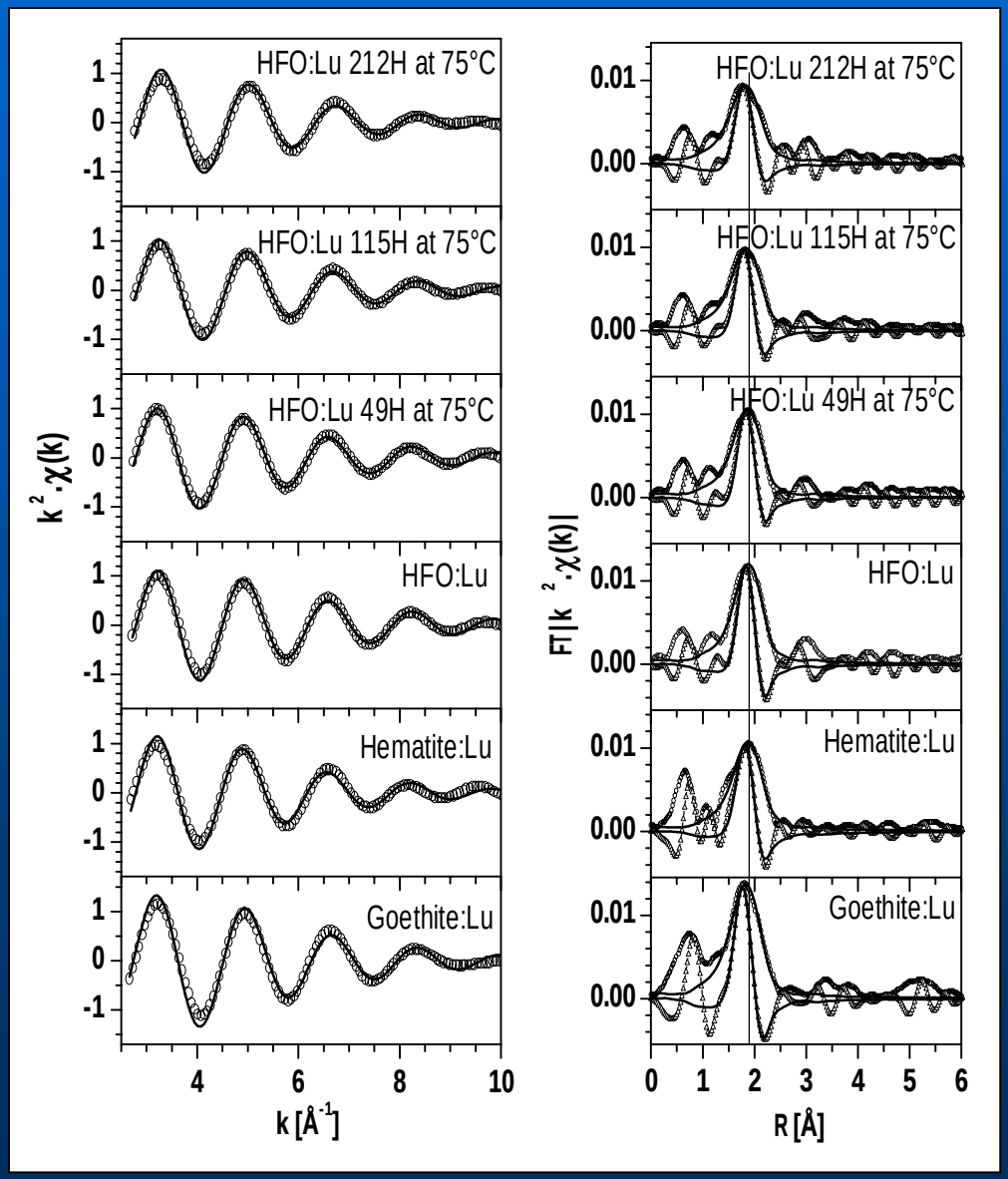
Varimax Rotation



Iterative Target Factor Analysis (ITFA)



Sorption of Lu(III) on Hydrous Ferric Oxide



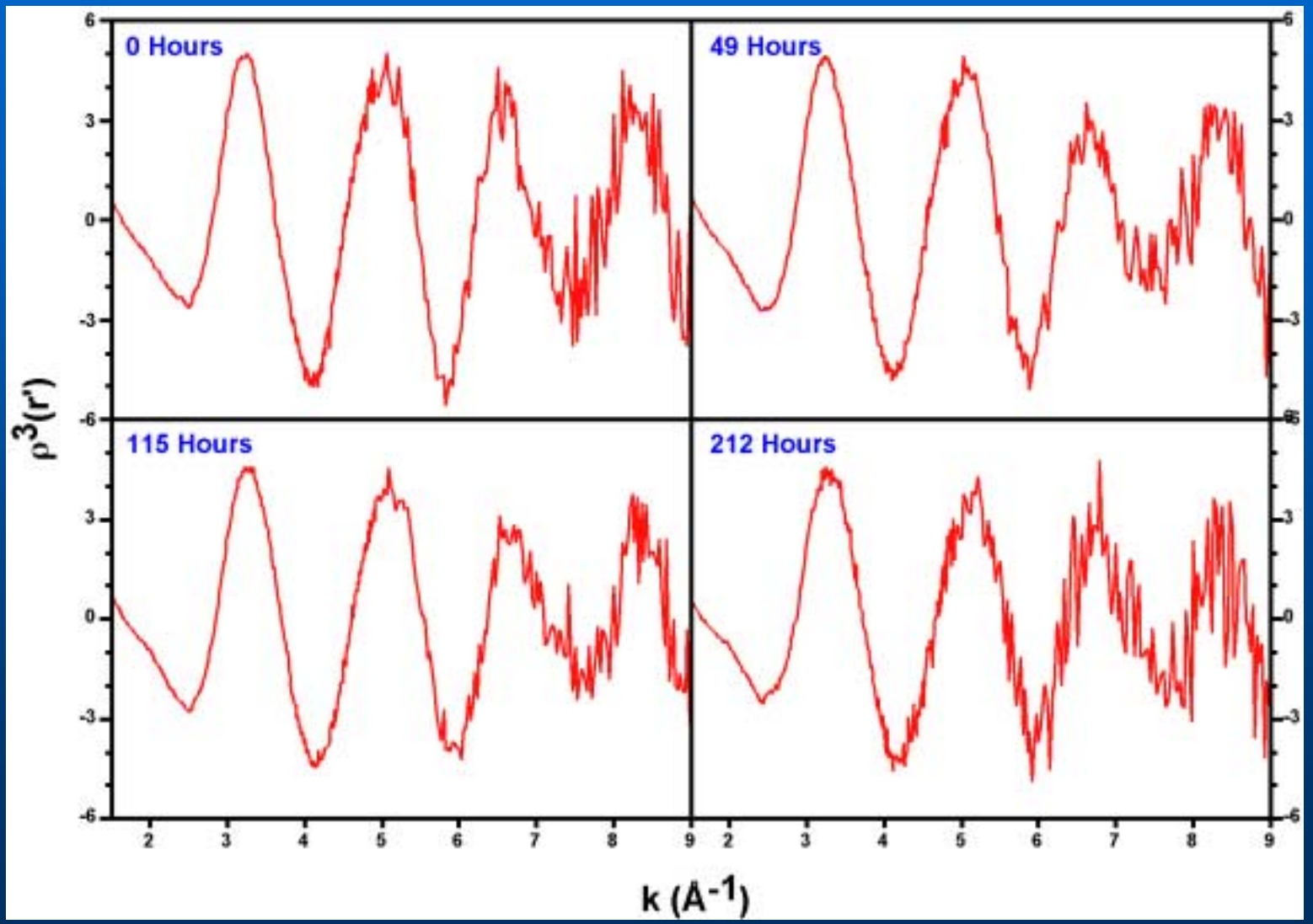
2.25 A

2.26 A

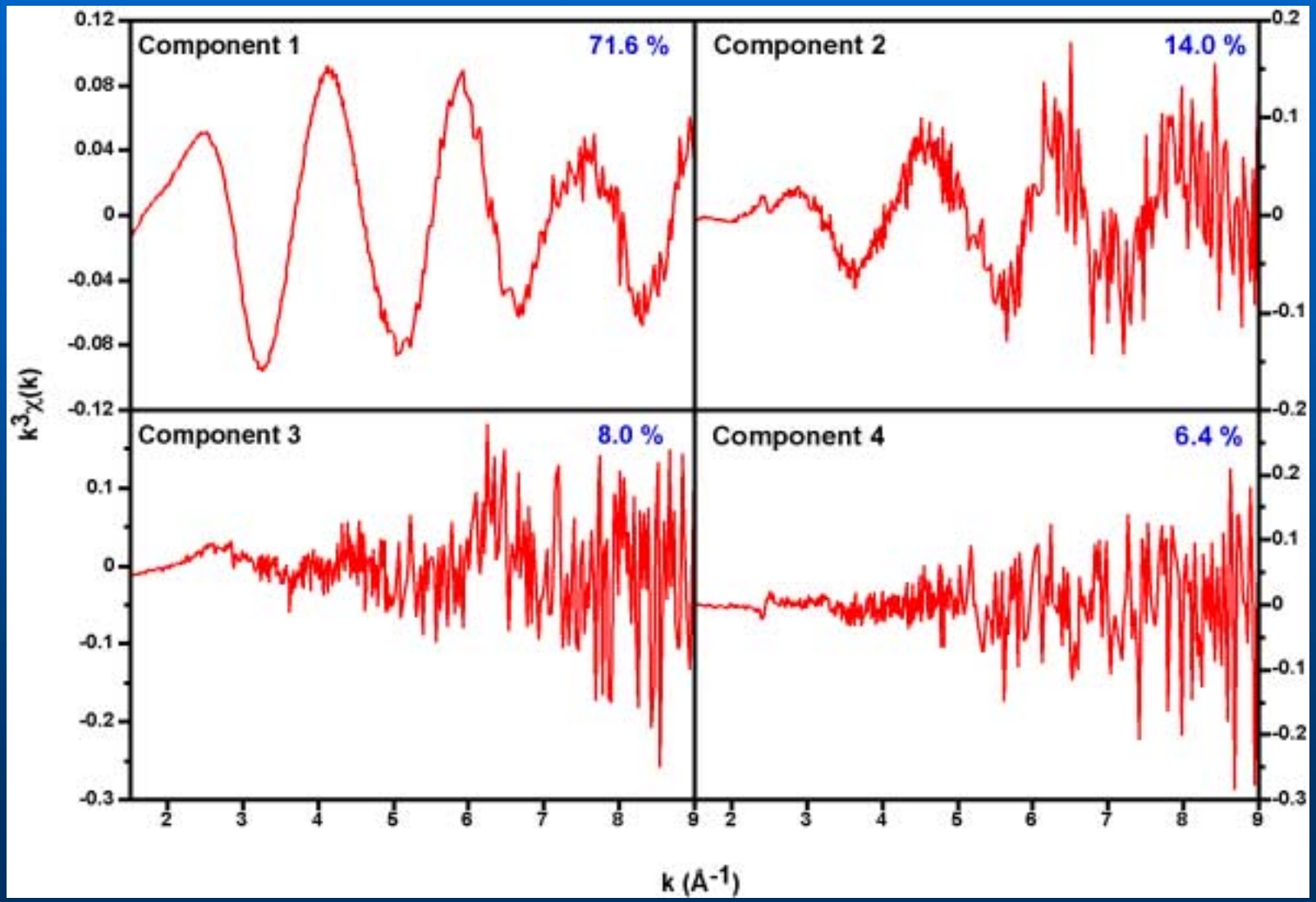
2.29 A

2.30 A

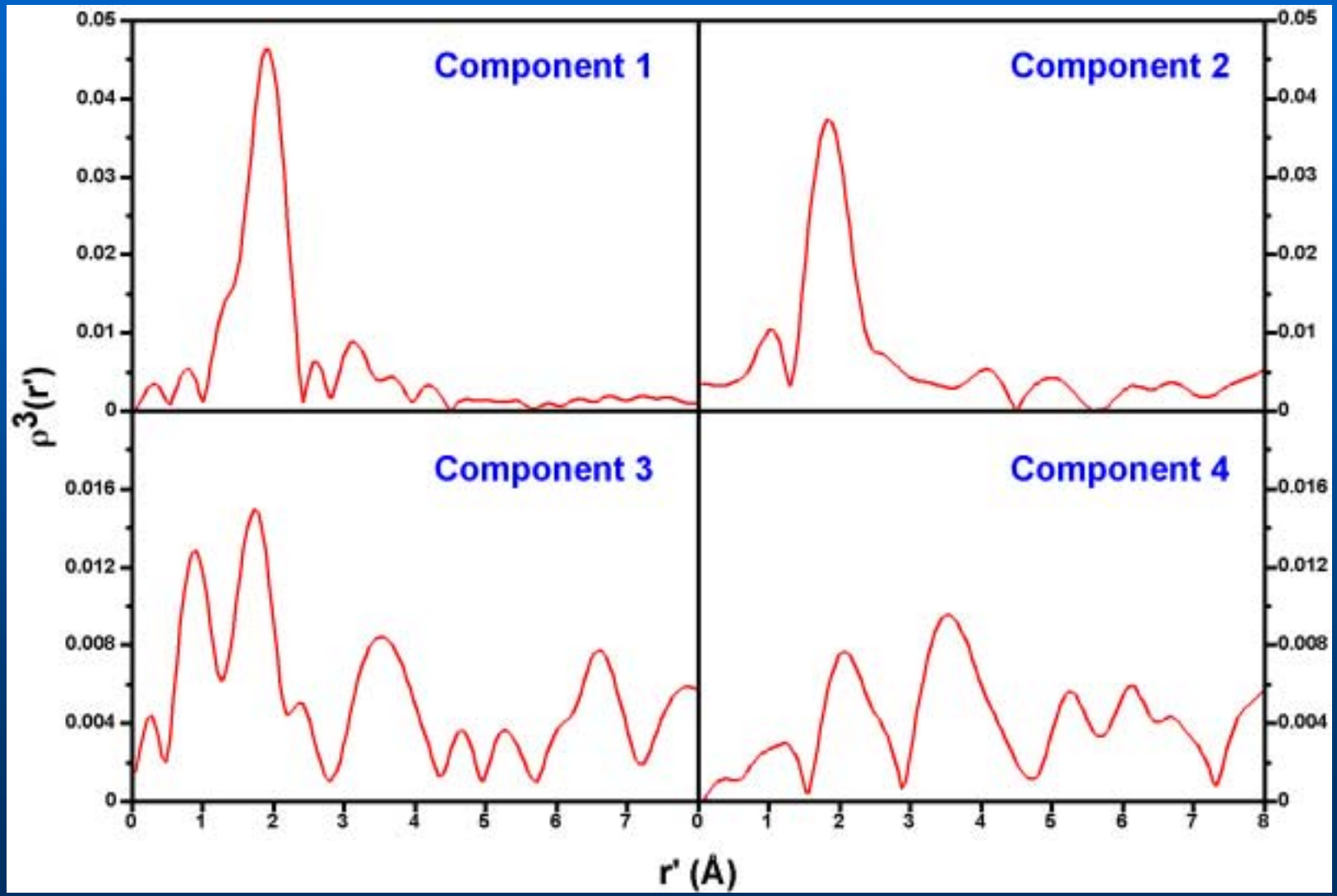
Sorption of Lu(III) on HFO



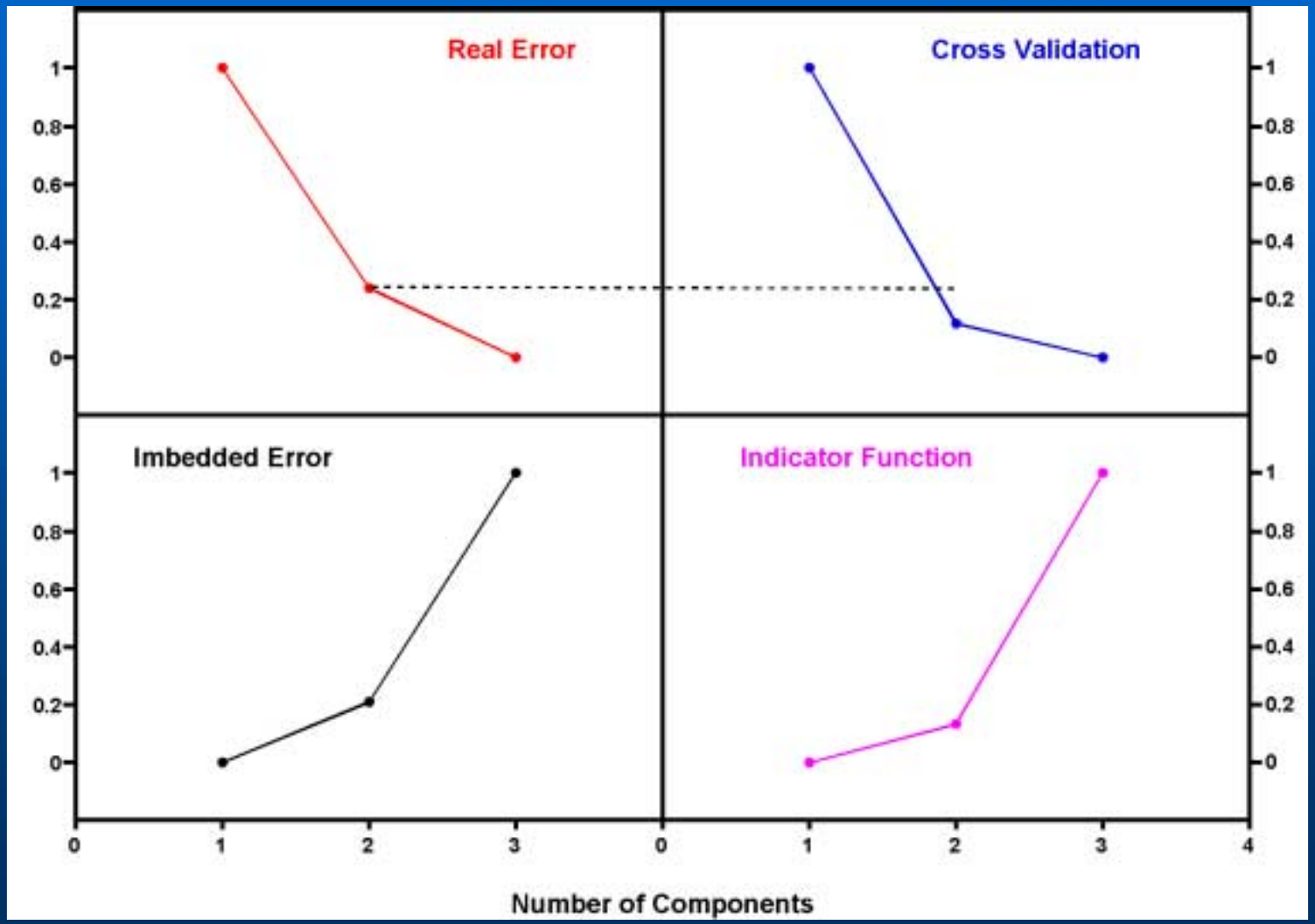
Lu(III) on HFO: Components



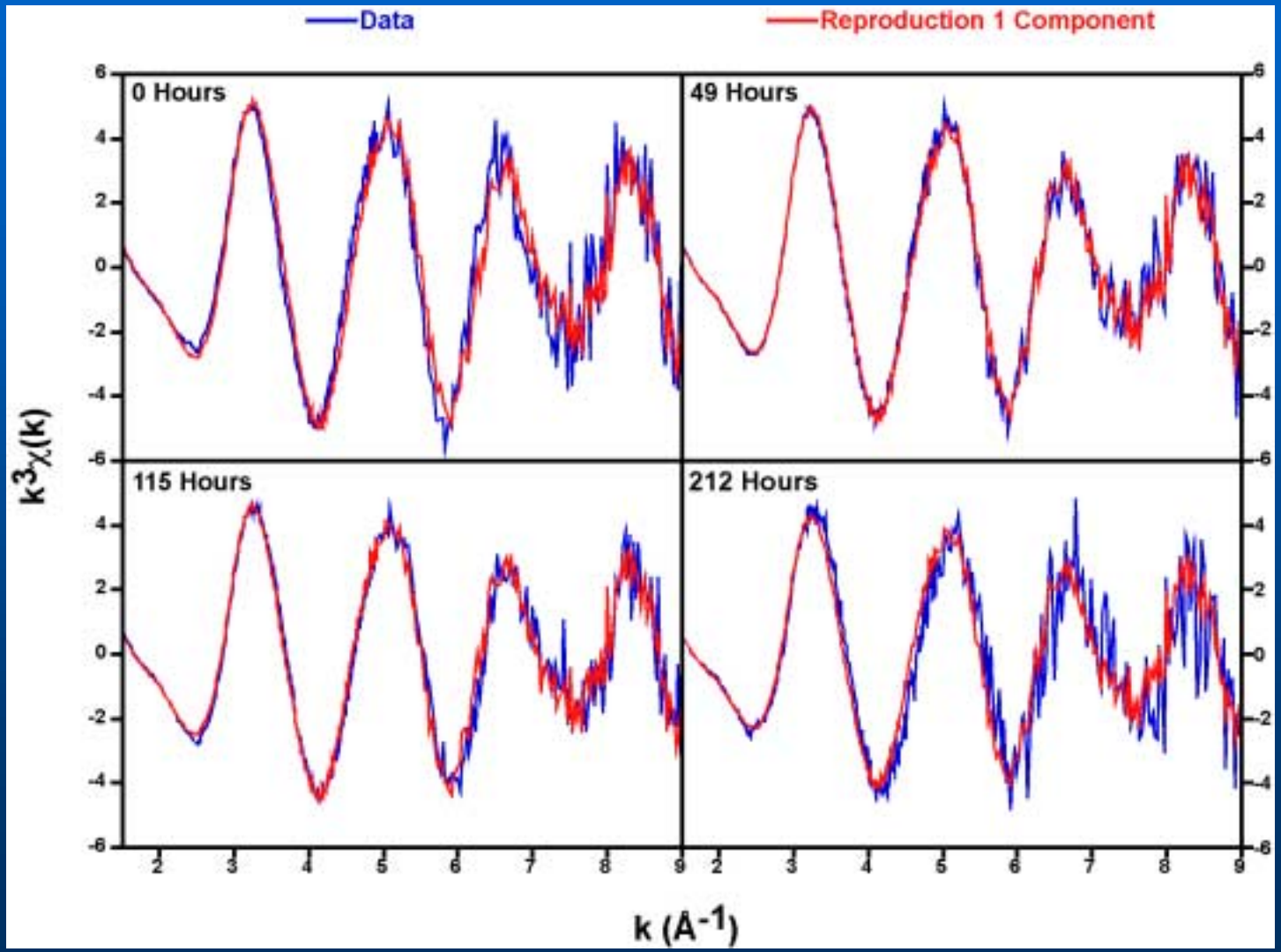
Lu(III) on HFO: Components



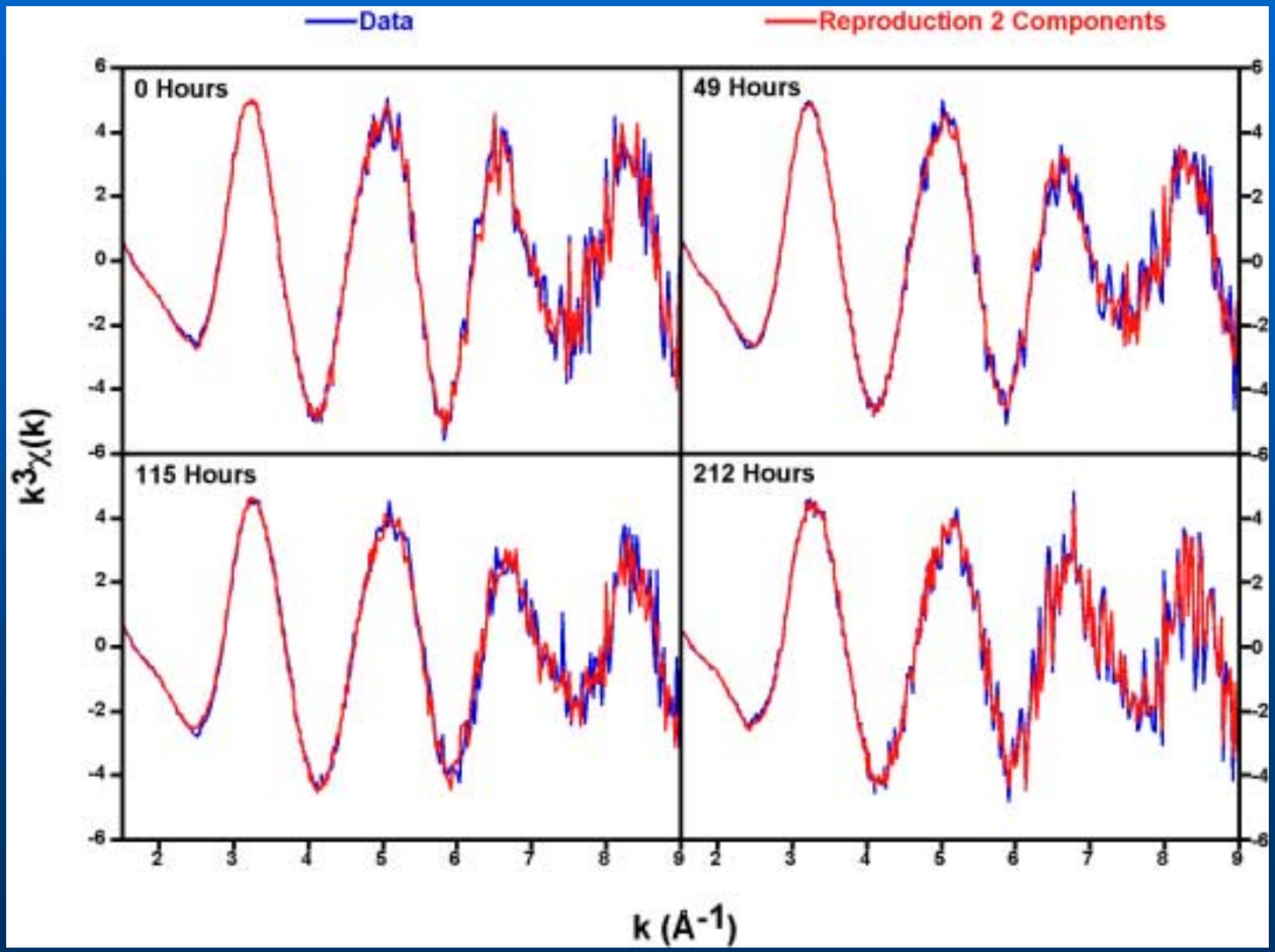
Errors: Lu(III) on HFO



Reproduction: 1 Component



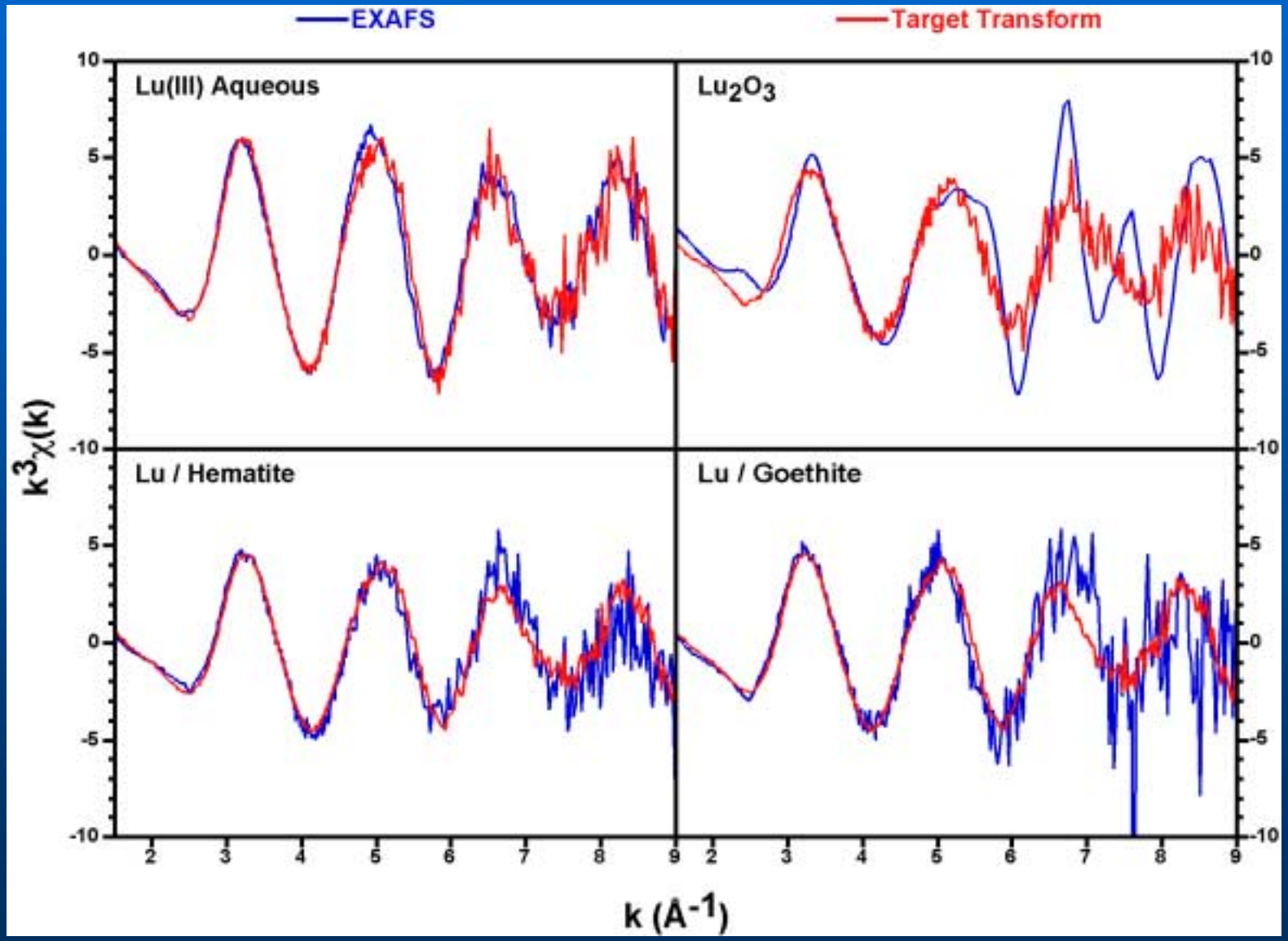
Reproduction: 2 Components



Real Standards

- **A real standard must be fit by the primary components**
- **Examine a standard compound WITHOUT knowing which other compounds are present**
  - **Invert traditional method of fitting edges to standards**
  - **Mathematically, primary components are equivalent to the original spectra (except for error)**
- **PCA can eliminate potential standards**
- **Only determine if a standard is reasonable**

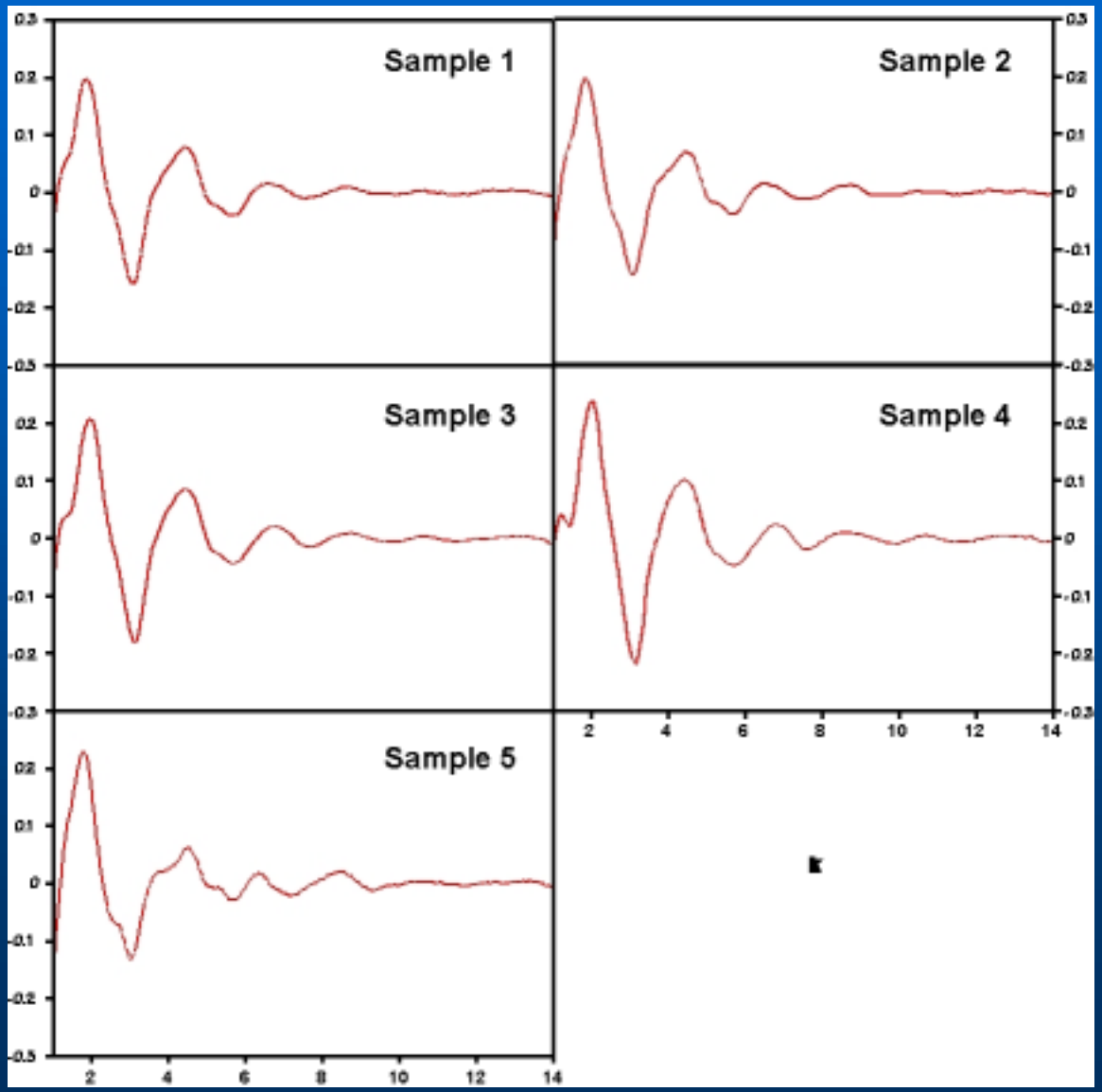
Target Transforms



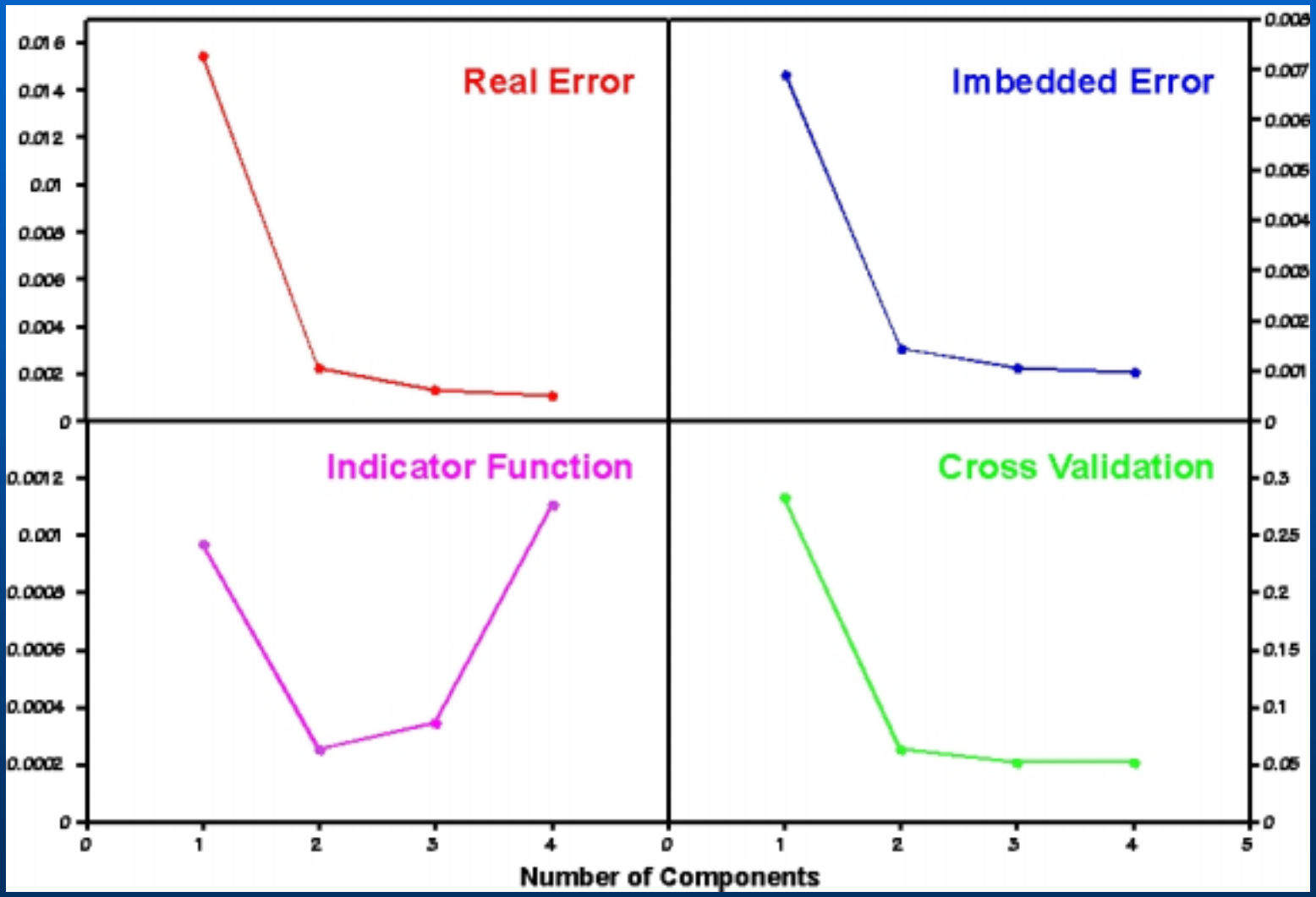
Degrees of Freedom: Lu(III) on HFO

- $\delta k = 7.5$
- $\delta r \sim 1.5$
- d.o.f ~ 7
- Restricted to 1 shell (unless fix parameters)
- Increase d.o.f. using series of samples

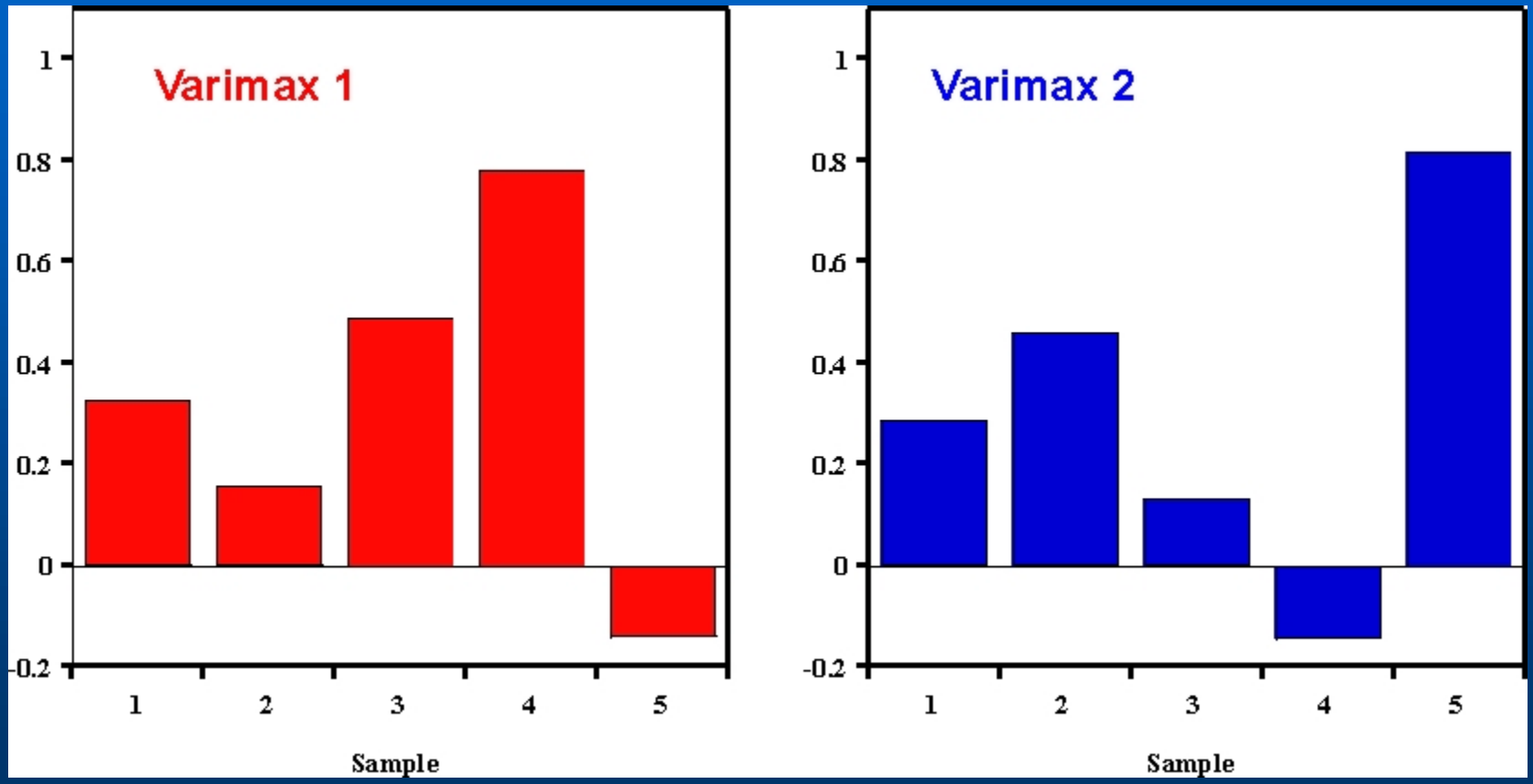
Co-AcAc and Co-TPP



Co-AcAc and Co-TPP: Errors



Co-AcAc and Co-TPP: Varimax



PCA vs. FT

- PCA creates an alternative representation in vector space
- FT creates a representation in conjugate space
- For both, underlying data are unchanged

Limitations

- **Experimental artifacts can be interpreted as components**
  - **Energy calibration**
  - **Mode of detection**
    - **Transmission**
    - **Fluorescence**
    - **Electron yield**
  - **Resolution**
    - **Monochromator crystals**
    - **Divergence**
  - **Thickness effects**
  - **Harmonic rejection**

PCA

- The use of PCA requires a change in perspective on XAS analysis.
- Advantages
  - Few spectra must be analyzed
  - The shells to fit are generally intense
- Difference from traditional methods
  - A single coordination shell will often appear in two or more components
  - Fit parameters apply to several spectra
- Disadvantages
  - Negative coordination numbers
  - Simultaneous fitting of two or more components may be necessary