# Chemical speciation by principal component analysis (and target transformation) in XAS

*K. Klementiev, MAX IV Laboratory*

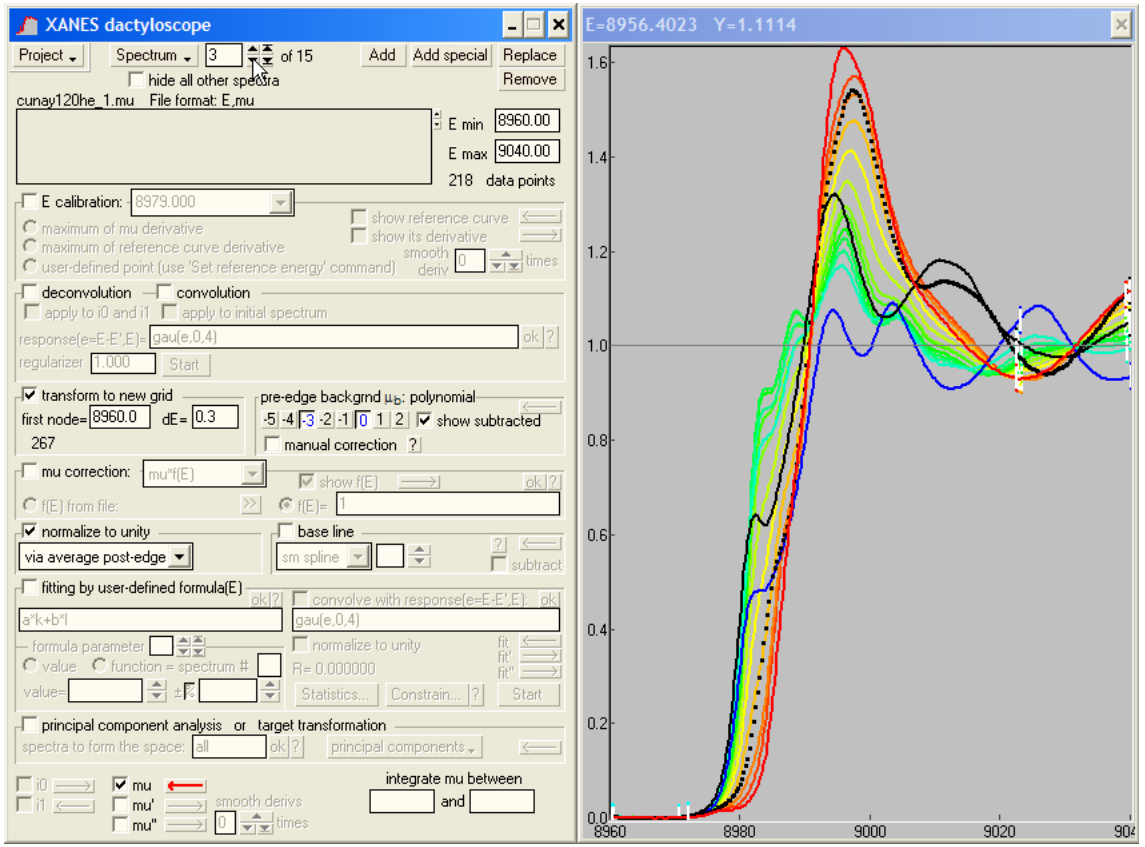
Recommended reading:

*XANES Dactyloscope Users Manual and Tutorial* (pdf)

# Conventional XANES Analysis Steps

- Instrumental corrections (energy calibration, deglitching etc.)
- Subtraction of pre-edge background
- Normalization data to edge step
- Principal component analysis (PCA), target transformation (TT)
- Fitting by a linear combination of reference spectra (if TT was successful)

*XANES dactyloscope*:



# Principal Component Analysis (PCA)

PCA answers the question “How many independent spectra?”

For the standard derivation based on SVD see

[1a] S. R. Wasserman, *J. Phys. IV France* **7** (1997) C2-203, which followed [1b]
[1b] E. R. Malinowski. Anal. Chem. **49** (1977) 606
[2] T. Ressler, et al., *Env. Sci. & Technol.* **34** (2000) 950

The standard derivations have two major drawbacks:

1) The PCA test spectra are compared with the data just 'by eye', without statistical grounds for the comparison. Therefore one cannot say how strongly the PCA mismatch (Malinowski's indicators) may differ from the experimental noise.

2) The experimental errors of different spectra may obviously differ. For instance, reference materials usually have much cleaner spectra than typical diluted samples. Therefore the comparison to noise (or, inversely, the estimation of noise) should be done individually for each spectrum, whereas the standard derivations concern only the global noise.

# Principal Component Analysis (PCA)

1) $N$ spectra $\rightarrow$ data matrix $D = \begin{pmatrix} \boldsymbol{d}_1 & \cdots & \boldsymbol{d}_N \\ \downarrow & \cdots & \downarrow \end{pmatrix}$

2) make $D^T D$ ($N \times N$ matrix) and find its eigen-pairs $\lambda_i$ and $\mathbf{e}_i$. Always: $\mathbf{1} \equiv \Sigma_i^N |\mathbf{e}_i\rangle\langle\mathbf{e}_i|$

3) for $M$ linearly independent data vectors, this sum can be truncated at $i=M$ and still $\mathbf{1} = \Sigma_i^M |\mathbf{e}_i\rangle\langle\mathbf{e}_i|$.

4) compare $D$ with $D \cdot \Sigma_i^M |\mathbf{e}_i\rangle\langle\mathbf{e}_i|$. If they coincide within noise then truncate further. Finally, $M$ gives the number of independent vectors.

5) alternatively, compare the cut-off part $D \cdot \Sigma_{i=M+1}^N |\mathbf{e}_i\rangle\langle\mathbf{e}_i|$ with noise. How strongly it may deviate from noise? Different spectra have different noise; which noise then to take? For the answers, see the manual of *XANES dactyloscope*.

# Questions answered by PCA

Two direct and two inverse questions:

PCA1) Given the global (average) noise level, how many spectra are linearly independent?

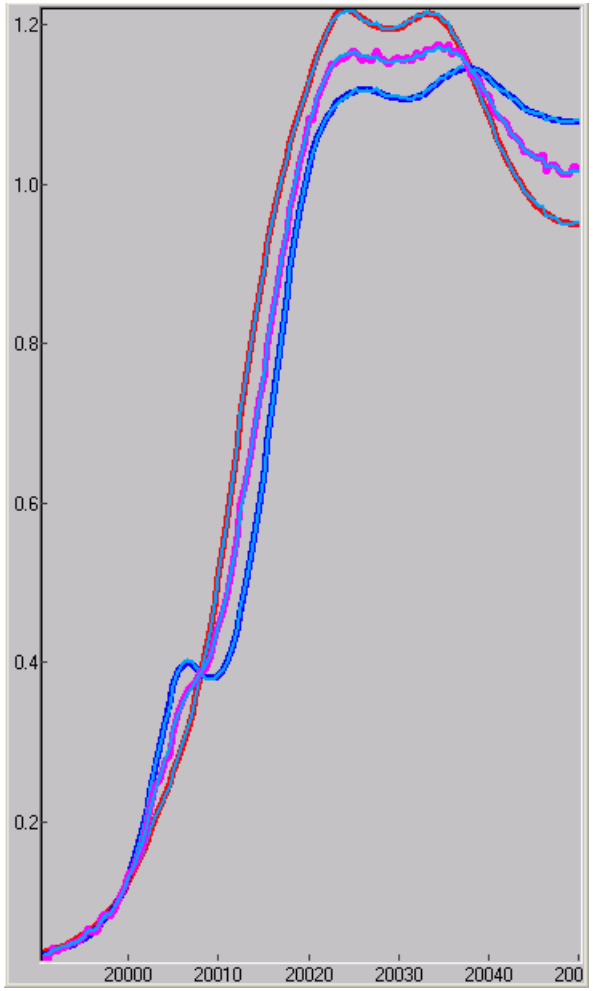
PCA1') How high must be the global noise level in order to have a given number of independent spectra?

PCA2) Given the noise level of a particular spectrum, how many principal components are needed to reproduce the spectrum?

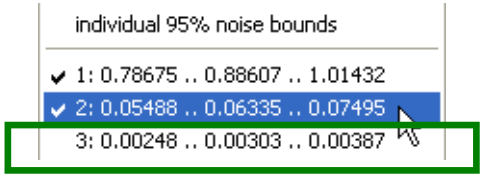
PCA2') How high must be the noise level of a particular spectrum in order to reproduce it by a specified number of principal components?

# PCA: test example 1

This example has two independent spectra (thick red and blue) and a spectrum constructed as the average of the two plus normal noise with nominal σ = 0.005 (thick magenta). The thin cyan curves are the PCA-transformed curves. **How many independent spectra we have within this set?**



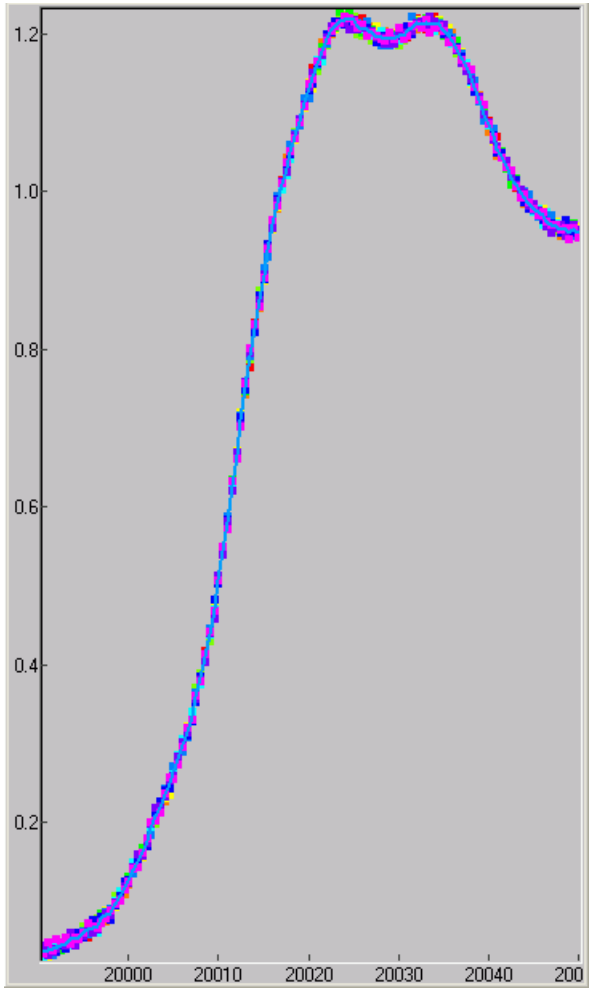

One needs 2 PCs to reproduce all 3 spectra




If we had all 3 spectra independent, the noise level had to be within these bounds with 95% probability.

# PCA: test example 2

This example has has 10 spectra which are all artificially constructed from a *single spectrum* with added 10 various realizations of normal noise of nominal σ = 0.005.




One needs 1 PC to reproduce all 10 spectra

individual 95% noise bounds

- ✔ 1: 0.78652 .. 0.88581 .. 1.01403
- 2: 0.00451 .. 0.00511 .. 0.00589
- 3: 0.00441 .. 0.00503 .. 0.00586
- 4: 0.00449 .. 0.00517 .. 0.00608
- 5: 0.00343 .. 0.00398 .. 0.00476
- 6: 0.00361 .. 0.00425 .. 0.00518
- 7: 0.00378 .. 0.00453 .. 0.00566
- 8: 0.00256 .. 0.00315 .. 0.00409
- 9: 0.00107 .. 0.00137 .. 0.00190
- 10: 0.00078 .. 0.00109 .. 0.00180

global 95% noise bounds

- ✔ 1: 0.78674 .. 0.88606 .. 1.01431
- 2: 0.00439 .. 0.00497 .. 0.00574
- 3: 0.00421 .. 0.00481 .. 0.00560
- 4: 0.00402 .. 0.00463 .. 0.00545
- 5: 0.00384 .. 0.00446 .. 0.00533
- 6: 0.00366 .. 0.00431 .. 0.00525
- 7: 0.00350 .. 0.00420 .. 0.00524
- 8: 0.00332 .. 0.00408 .. 0.00530
- 9: 0.00306 .. 0.00391 .. 0.00544
- 10: 0.00273 .. 0.00381 .. 0.00629

We have our noise within these bounds and thus we can say that with 95% probability there is 1 PC.

If we had our noise within these bounds (~1!) then we could tell that there is 0 PCs and all the data are just noise.

# Target Transformation (TT)

With TT one can answer the question whether a particular spectrum (material) represents a mixture of the basis spectra (materials). When answered positively, one can try fitting to a linear combination of the basis spectra.

1) Construct basis matrix $B$ from $N$ basis spectra. Make $B^TB$ ($N\times N$ matrix) and find its eigenvalues $\lambda_i$ and eigenvectors $\mathbf{e}_i$

2) **If the basis spectra are independent** then $(B^TB)^{-1}$ exists. Check this!

3) The matrix $$B(B^TB)^{-1}B^T \equiv \sum_i^N \frac{1}{\lambda_i} B|\mathbf{e}_i\rangle\langle\mathbf{e}_i|B^T$$

is an orthogonal projector to the basis space (it is equal to its square, check this). Hence, if a spectrum $S$ is a linear combination of the basis spectra then $B(B^TB)^{-1}B^TS = S$ and inversely.

4) In practice one checks if $B(B^TB)^{-1}B^TS$ coincides with $S$ within noise. How strongly the difference is allowed to deviate from noise? For the answer, see the manual of *XANES dactyloscope*.

# Questions answered by TT

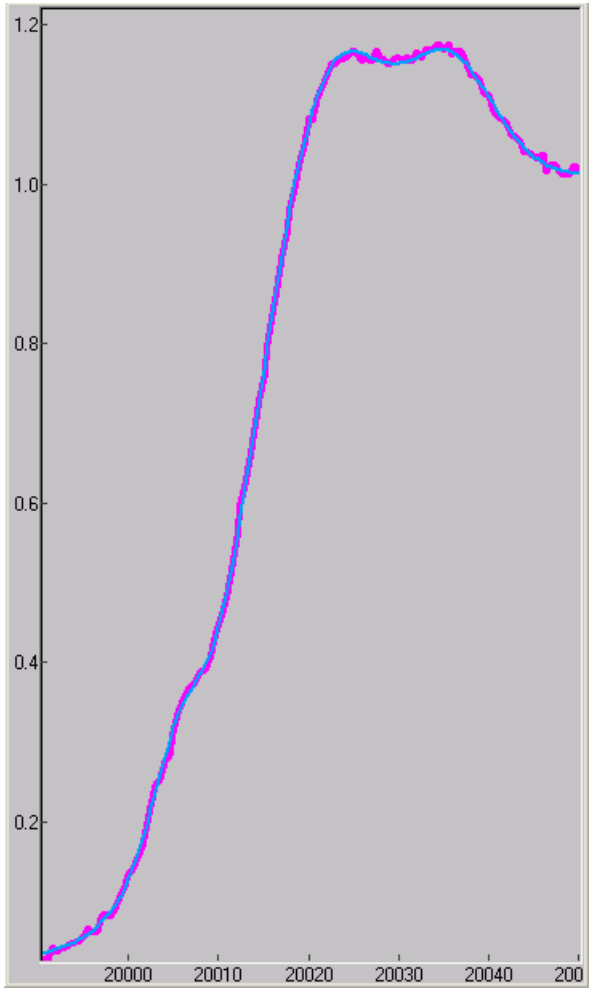
One direct and One inverse question:

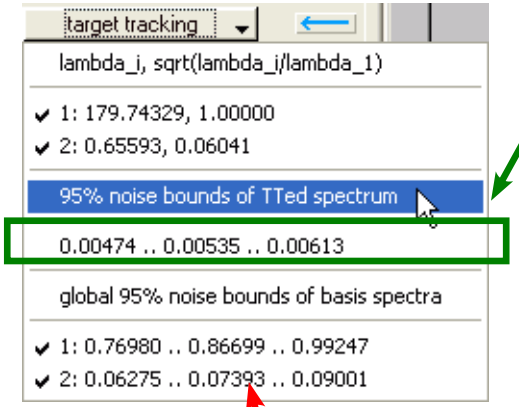
TT1) Given the noise level of a particular spectrum, can the spectrum be reproduced by a linear combination of the basis spectra?

TT1') How high must be the noise level of a particular spectrum in order to reproduce it by a linear combination of the basis spectra?

# TT: test example

The same example as on the slide 6 (two independent spectra and the average of them plus normal noise with σ = 0.005) but the question is different:
**Can the spectrum 3 be reproduced by a linear combination of the spectra 1 and 2?**



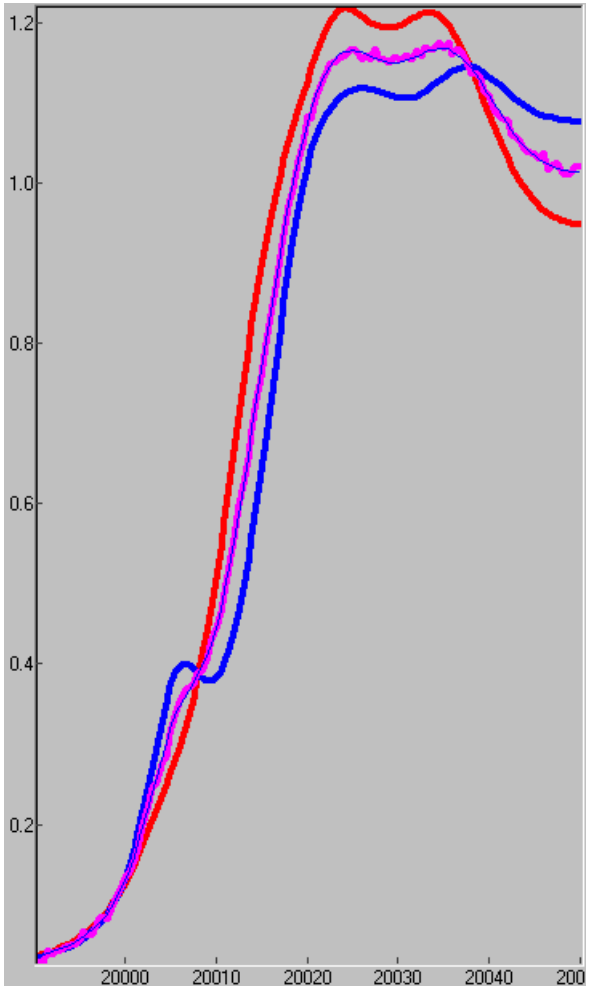



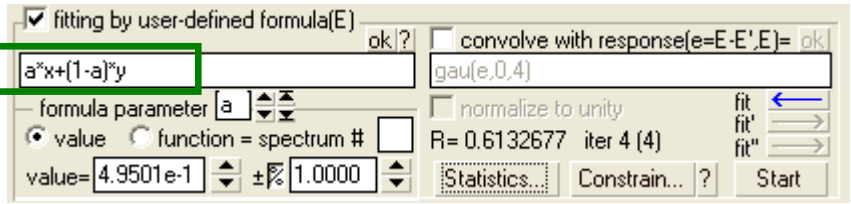

The answer is positive because our noise level is within these bounds:

Check also that the noise estimation of the last principle component of the basis set (here, 0.07393) is much higher than the experimental noise, otherwise the basis set is bad, i.e. internally linearly dependent.

# Linear combination fitting: test example

If the TT test is successful, we can perform *linear combination fitting*.
Again, as on slides 6 and 10, let us fit the average of two spectra plus some noise by a linear combination of the two spectra.

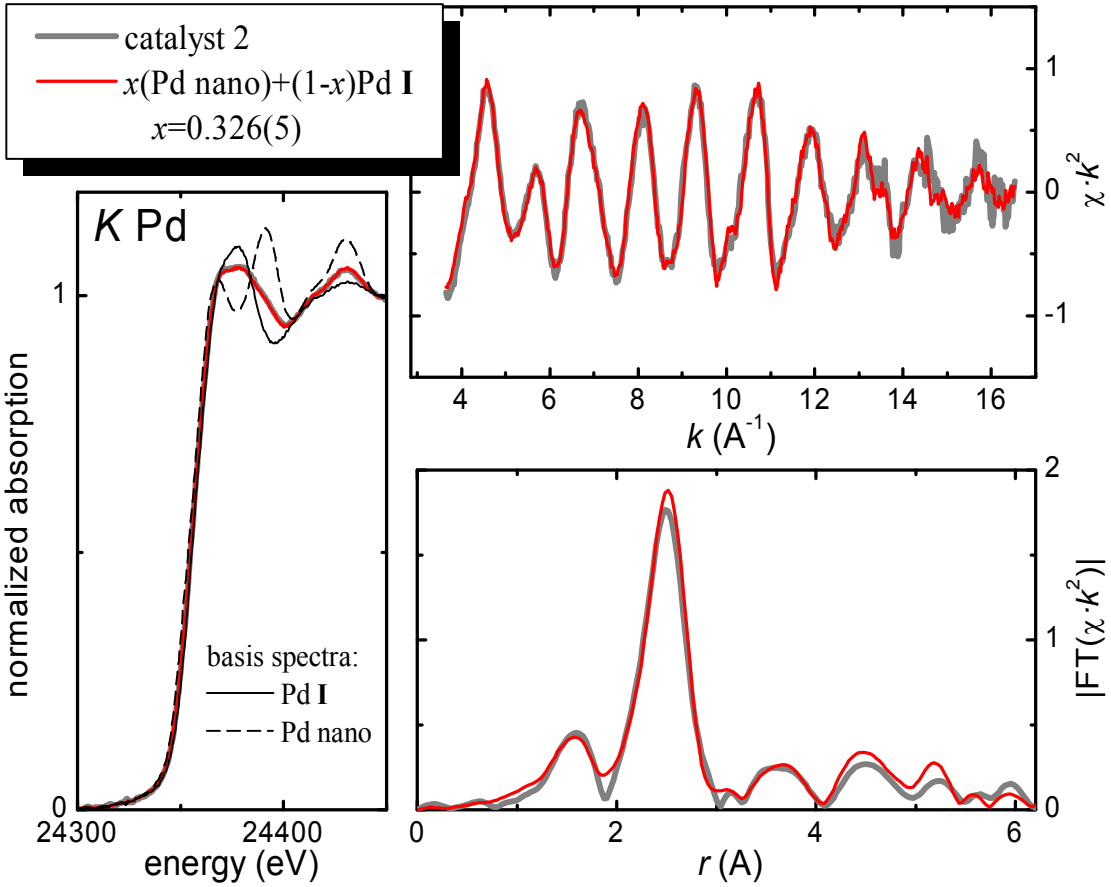



a = 0.495 ± 0.004
with expected a = 0.5

# Linear combination fitting: real example

1. *Basis spectra*: precursor (Pd **I**) and metallic Pd nano-particles.
2. The spectrum “catalyst 2” shows coincidence with its *target transformation*.
3. Linear combination fitting of **XANES** of “catalyst 2” by the two basis spectra.
4. The found linear combination is then successfully applied to **EXAFS**.




**Palladium Nanoparticles immobilized on Mesoporous Silica Support – New Efficient Catalysts for Aerobic Alcohol Oxidation in Supercritical Carbon Dioxide**

Z. Hou et al., *J. of Catalysis* **258** (2008) 315

# How to estimate the experimental noise?

It is almost impossible to use Fourier analysis for error estimations in XANES (and what is typical for EXAFS):

- the range is short
- the low-frequency signal (the absorption jump) dominates also in high-frequency FT.

Possible for XANES:

- propagate measurement errors (from detectors) through all the transformations towards the XANES spectrum.
- root mean square deviation of a series of XANES repeats.

# Summary

There are two main classes of applications of XANES:

Fingerprint analysis: presence/absence of pre-edge peaks, edge shift, white line height (see the introductory lecture). This analysis doesn't require sophisticated methods.

Factor analysis (PCA, TT) and linear combination fitting. This analysis requires some mathematics and statistics and careful attention to the estimation of noise. This part has been considered in the present lecture. The examples can be found among the example projects of *XANES dactyloscope* program and in its manual.

Assumptions:

- A mixture of chemical species leads to a linear combination of spectra.
- The basis set (of TT and linear fit) is complete.
- The basis set (of TT and linear fit) is 'good' = linearly independent.

You should check their validity.