EXPO&*more* International Workshop
Crystallographic Software for Powder Diffraction Data
30 September – 3 October 2019
Organized by the Institute of Crystallography (IC) – CNR Bari, Italy

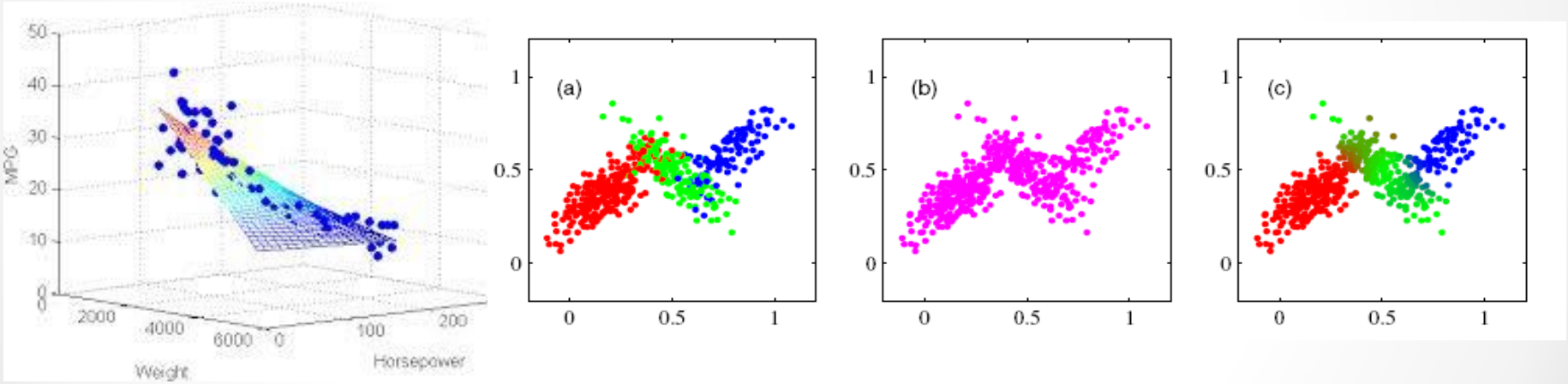
New algorithms for fast extraction of information from *in situ* powder diffraction data

DEI - DIPARTIMENTO DI INGEGNERIA ELETTRICA E DELL'INFORMAZIONE
POLITECNICO DI BARI

Pietro Guccione – Assistant Professor in Signal Processing
(pietro.guccione@poliba.it, http://dei.poliba.it/il-dipartimento/il-personale/guccione.html)

Motivation /1

Multivariate Analysis is a powerful and well-established set of methods (regression, clustering, classification, dimensionality reduction, density estimation, …) for retrieving information from large datasets and combining data from different sources. It consists in a statistical, mathematical and graphical set of techniques that consider multiple variables simultaneously.



Chemometrics has born while applying these methods to Chemistry and it has the aim of extracting information from chemical systems by data-driven means.

Motivation /2

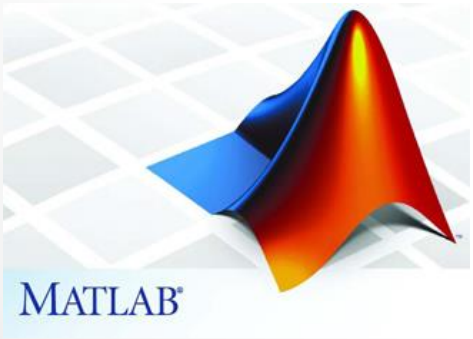
Usual methods for such analysis are well supported in calculus software environment such as Matlab® .

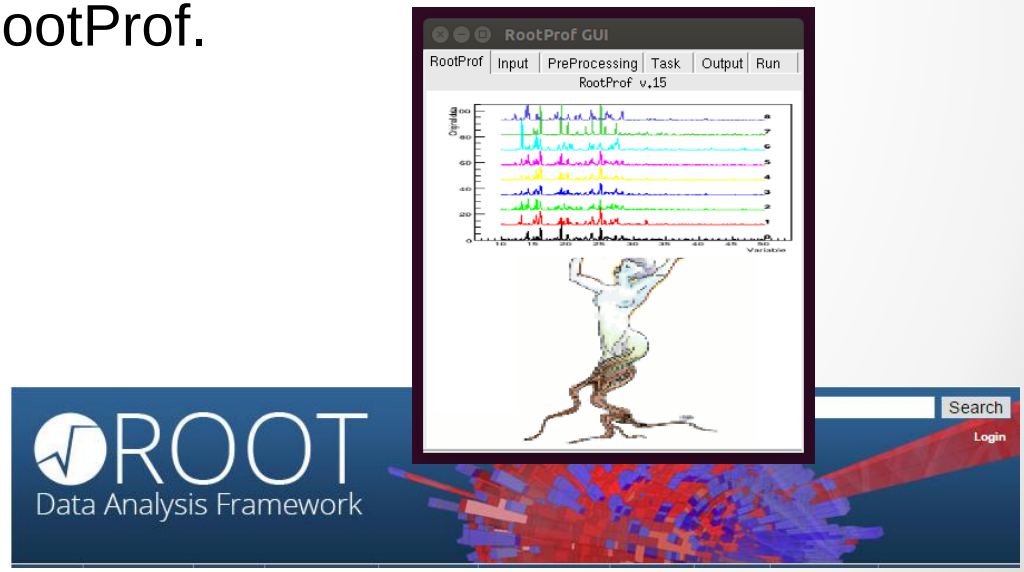
Recent software, more specific for Powder Diffraction Data (PDD) as **Rootprof**, are starting to develop such tools.

The purpose of the talk is to provide:

(i) Some view and basis of the methods;

(ii) Some study case faced with multivariate analysis and supported (now and in future) by RootProf.





Summary

- Introduction to the dimensionality problem: meaning and need of reduction

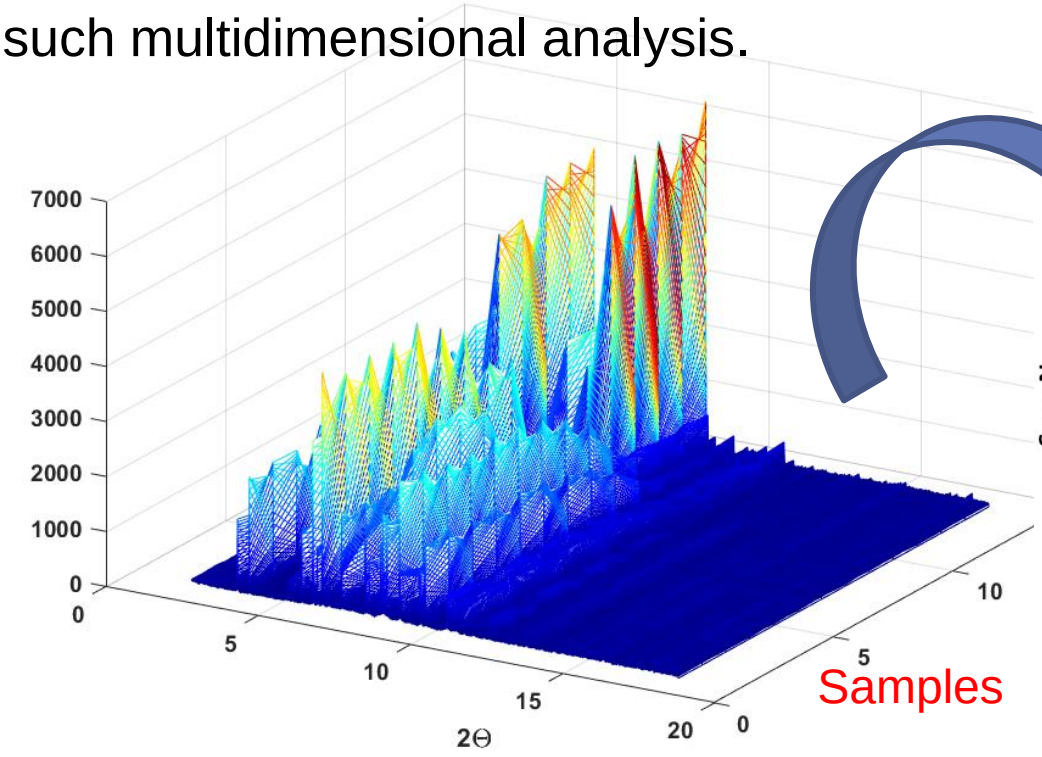
- Principal Component Analysis: meaning and related tools
- Extension of PCA and relaxation of orthogonality: OCCR.
- Case study: analysis of XRPD dataset

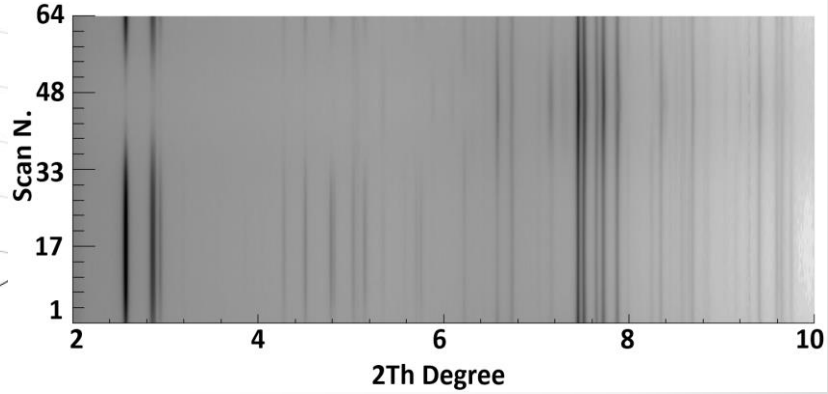
- Kinetics of Solid-state reaction: optimized-PCA analysis.
- Case study: evaluation of kinetics triplet from XRPD

Thinking at many dimensions

Powder Diffraction Data are a set of spectra acquired with slight different conditions along time.
Change of structural crystalline characteristics (occupancy, lattice, etc) provides different spectra.

Retrieve the “basic components” of such spectra and the “causes of modification” with little or null information about the dataset is the aim of such multidimensional analysis.



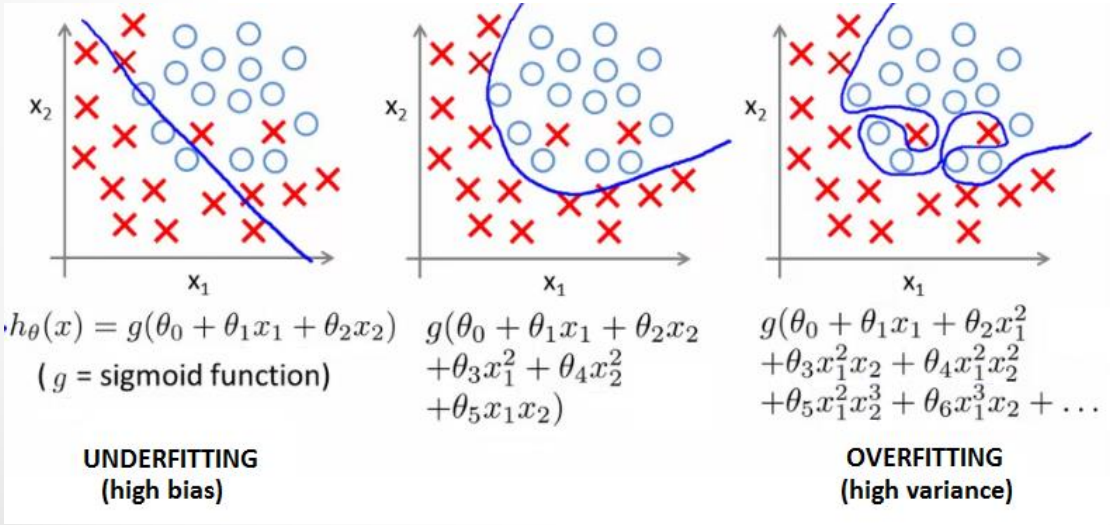


Dimensions!
(>> samples)

The high dimensionality problem

More variables than observations (Hughes phenomenon):
When the number of variables is too high compared to the number of the samples, the analysis algorithm is unable to find a proper structure within data that can be generalized to other dataset of the same experiment.

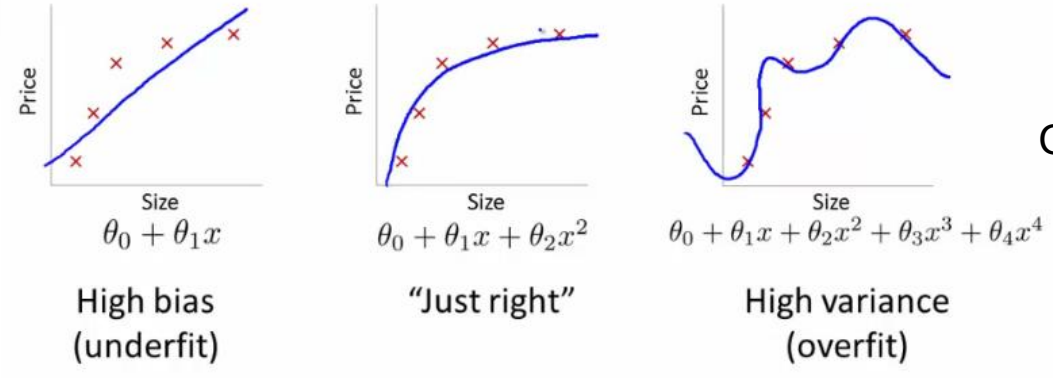
This is known as the *curse of dimensionality* or *Hughes phenomenon*.
It may commonly occur in **PDD:** diffraction angles may be thousands, as well, compared to few dozens of measured spectra



Visual example:
Overfitting in classification

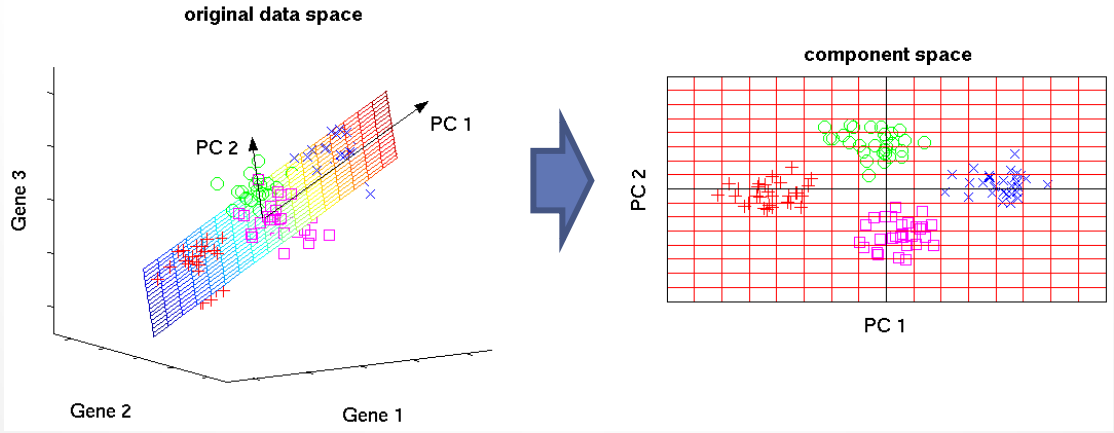
Dimensionality reduction

- The problem of high dimensionality involves also the estimation of parameters in hidden models (e.g.: the number of coefficient in a regression problem) or of latent variables (e.g.: number of mixtures in a density estimation problem).



Overfitting in regression

- The problem of dimensionality depends on both the data and the algorithm. Possible solutions are: trying to change algorithm or trying to **reduce the dimensionality** of the problem

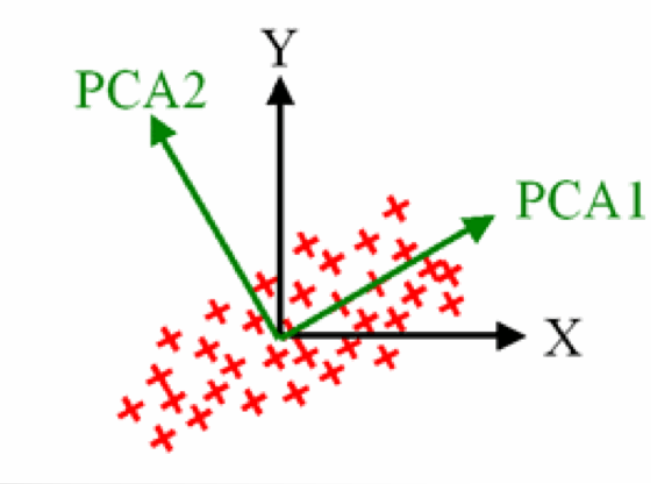


Dimensionality Reduction: the PCA

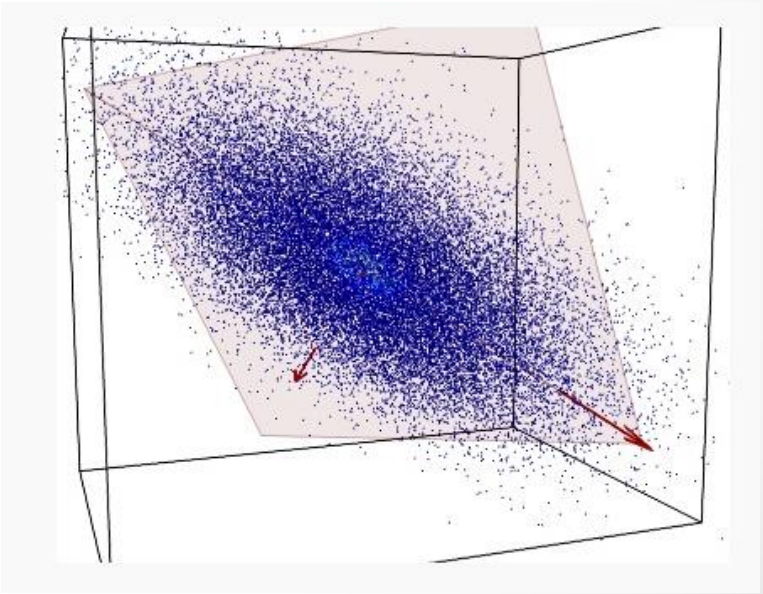
Principal Component Analysis is a standard technique for visualizing high dimensional data and for data pre-processing. PCA may reduce the dimensionality (the number of variables) of a data set by maintaining as much variance (i.e. energy) as possible.

PCA:

- finds the directions of maximum variation of the data
- decorrelates the original variables by using orthogonal transformation
- The set of uncorrelated variables are said ***principal components***



Retain all the dimensions



Reduce the dimensions

PCA: mathematical details

Principal Component Analysis is an **orthogonal linear transformation** that transforms the data to a new coordinate system such that the greatest variance lies on the first coordinate, the second greatest on the second coordinate, and so on.

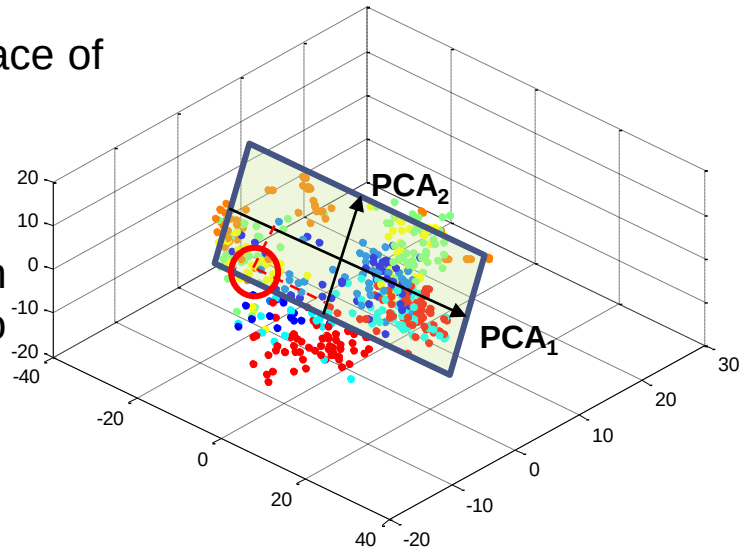
Organize data in a matrix, **X** [N x P], N samples (repetition of the experiment), P variates (the features of the experiment). The full principal components decomposition of **X** can be given as:

$$X = TW'$$

- The principal components **T** (called **scores**) are achieved as linear combination of data and a set of weights (called ***loadings***)
- The (column) weights **W** (that are the loadings) are the eigenvectors of the sample covariance matrix of data

PCA: meaning

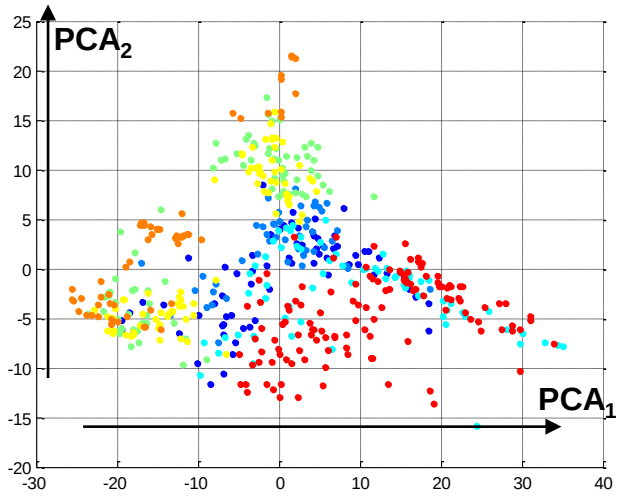
In PCA data are decomposed by projecting in a new space of the same dimension.
Samples are described in a multi-dimensional space.



- The loadings are the weights by which each standardized original variable should be multiplied to get the component score

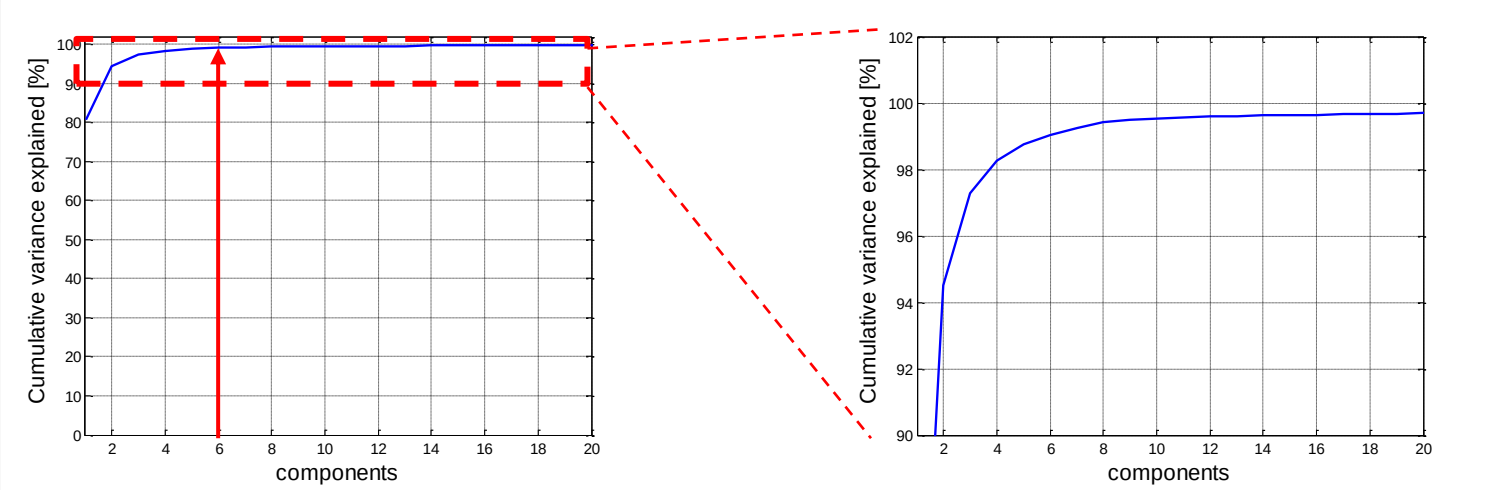
- The scores are the transformed variable values corresponding to each sample $T = XW$

Decomposition is done to maximize the variance (the energy) of the data in the first (few) dimensions

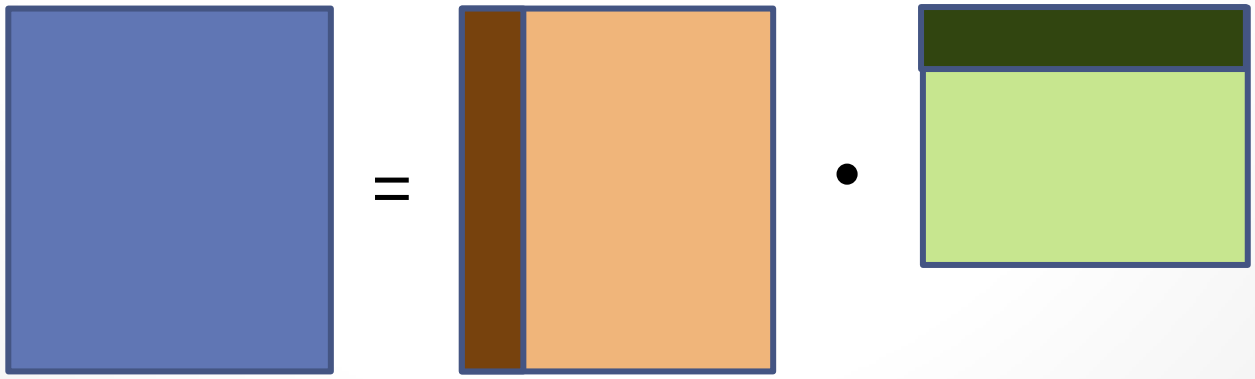


PCA: dimensionality reduction

Not all the principal components are equally important. Their relative importance is given by the explained variance. A typical plot of the variance is given by:



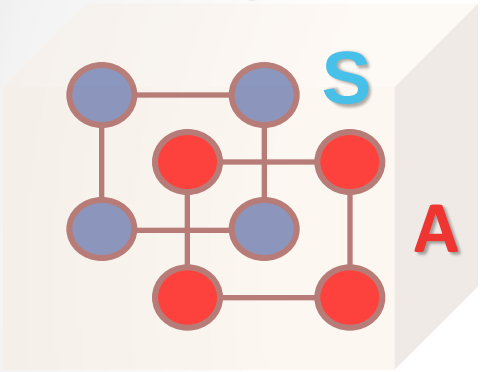
We want 99% of variance explained → n_c=6 components are enough



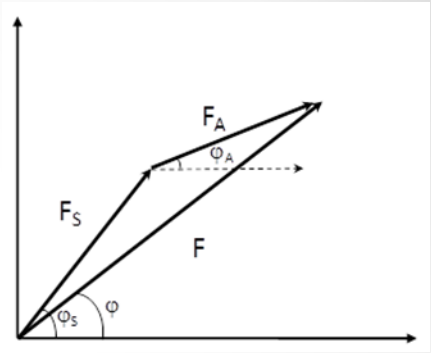
Modulated Enhanced Diffraction XPD data

External periodic stimulus

Crystal

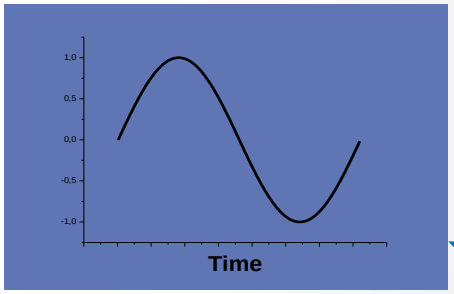


Atoms A responding elastically



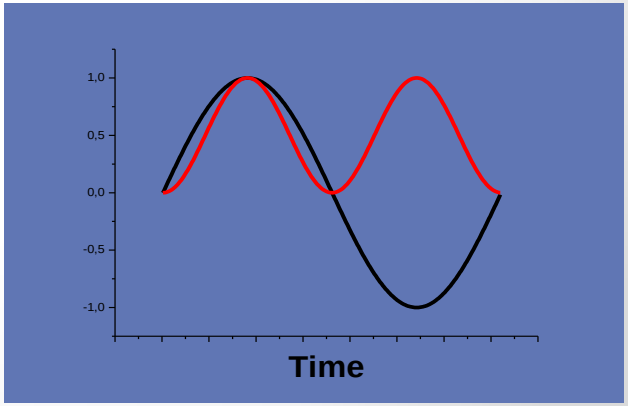
Detector

$$\left| F_S + F_A \right|^2 \rightarrow F_S^2 + 2 F_A F_S \cos(\Phi_A - \Phi_S) + F_A^2$$



Same frequency of stimulus

Doubled frequency



Acta Crystallographica Section A
Foundations of Crystallography
ISSN 0108-7673

Received 11 January 2011
Accepted 22 March 2011

Kinematic diffraction on a structure with periodically varying scattering function

Dmitry Chernyshov,[a]* Wouter van Beek,[a,b] Hermann Emerich,[a] Marco Milanesio,[b] Atsushi Urakawa,[c] Davide Viterbo,[b] Luca Palin[b] and Rocco Caliandro[d]

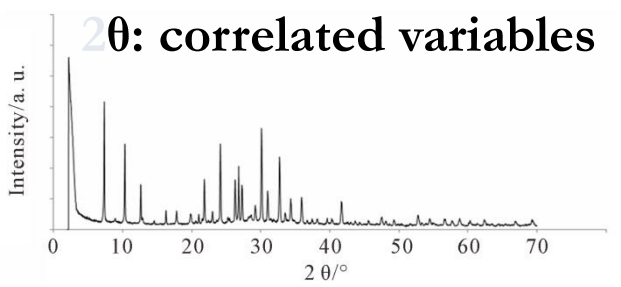
PCA applied to XRPD MED data

Loadings

c_i

$$\mathbf{u} = \sum_{i=1}^{N} c_i \mathbf{e}_i$$

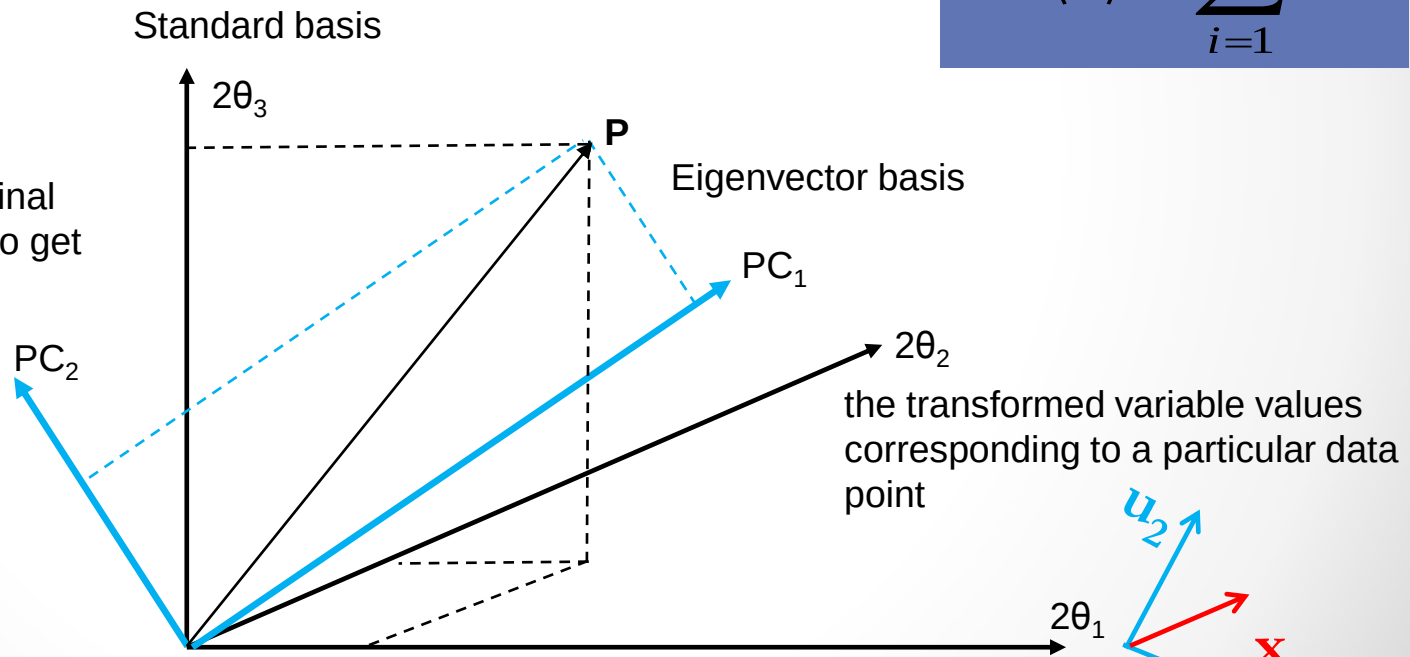
the weight by which each original variable should be multiplied to get the component score



Scores

b_i

$$x - \langle x \rangle = \sum_{i=1}^{N} b_i \mathbf{u}_i$$

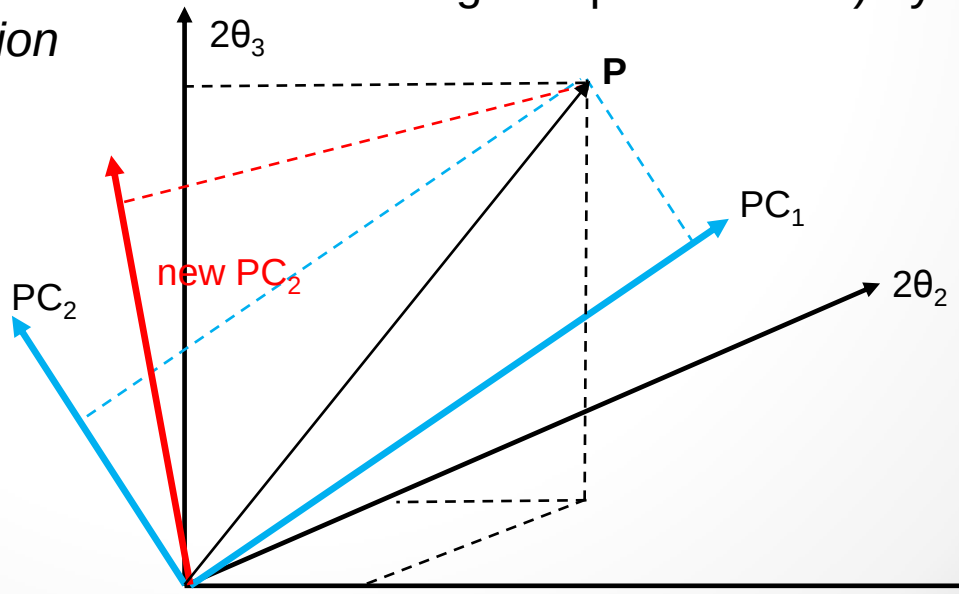


Principal components: uncorrelated variables

Extension: component rotation

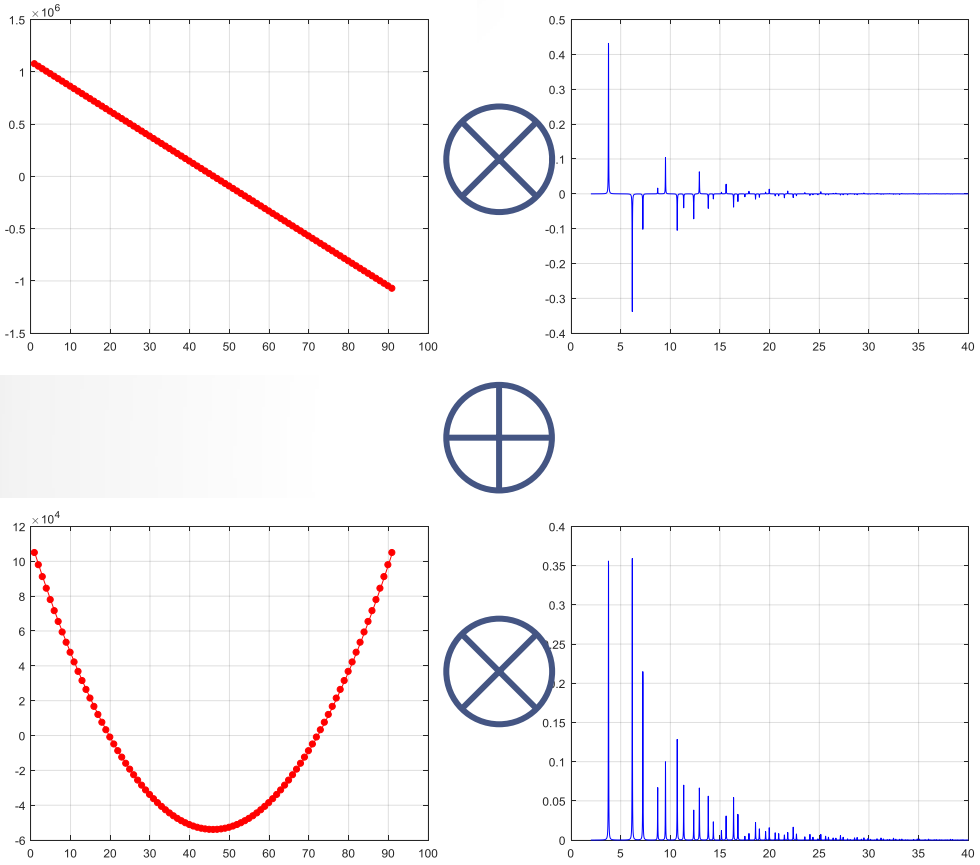
OCCR
Orthogonal Constrainted Component Rotation
In order to supply to futher condition in the problem, components may be changed and **no** longer constrained to be **orthogonal** each other (they may be *partially correlated*), so to allow the constraints to be applied.

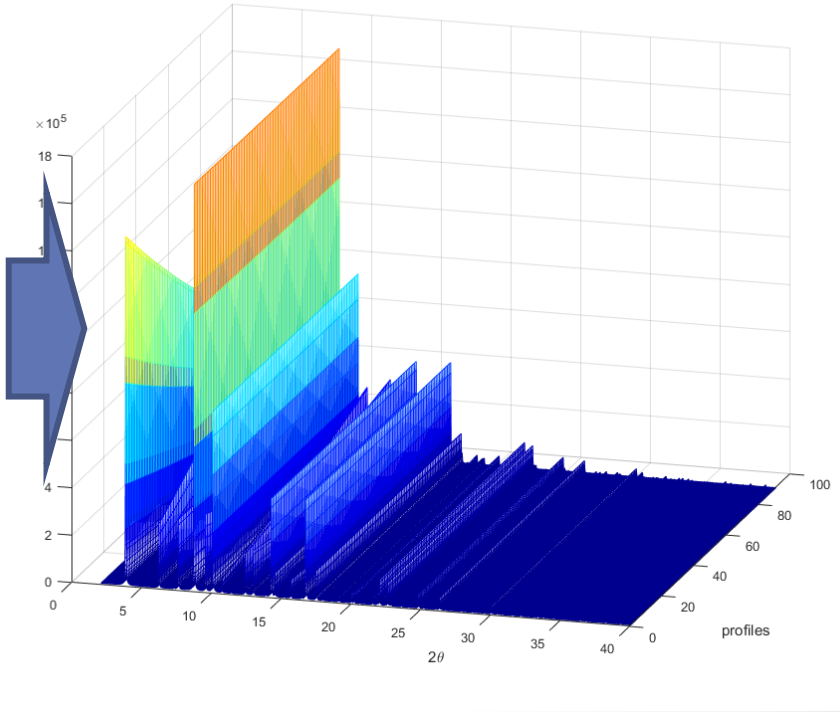
The score axes change their directions, by exploring the *k*-dimensional space (already reduced to the principal components) *driven by a properly defined cost function*.
The idea is that we are able to detect the optimal rotated axes of a low-dimensional space (where data still have a meaningful representation) by *minimizing an objective function*



PCA/OCCR decomposition

Visual scheme





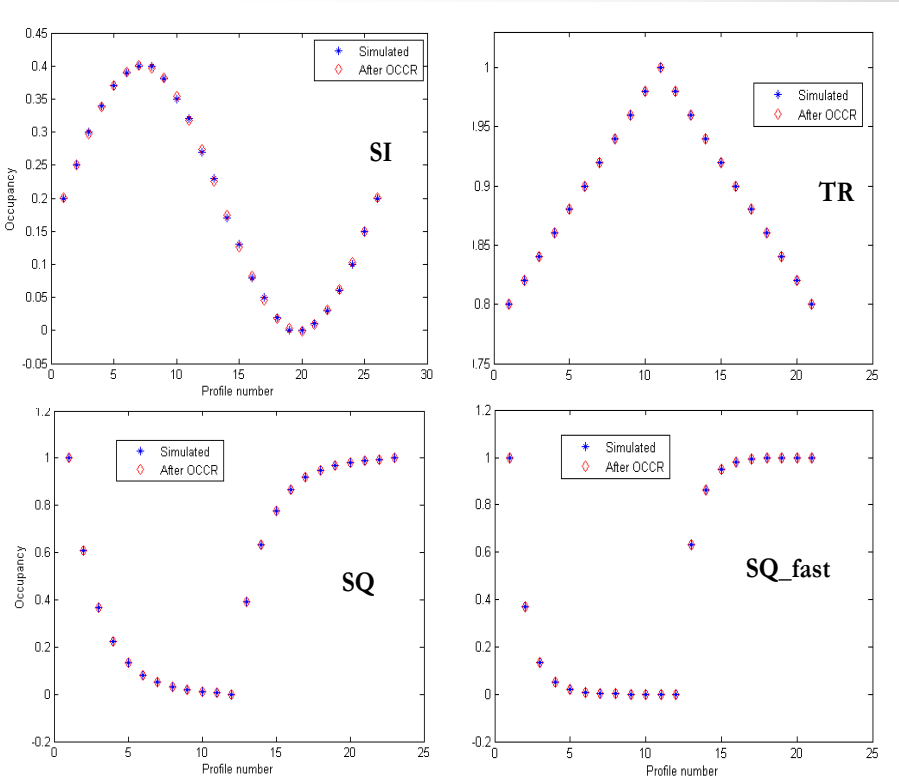
OCCR: $PC_2=PC_1^2$ Imposed
(in PCA dictated by data)

ρ = 0.0687
(in PCA ρ = 0)

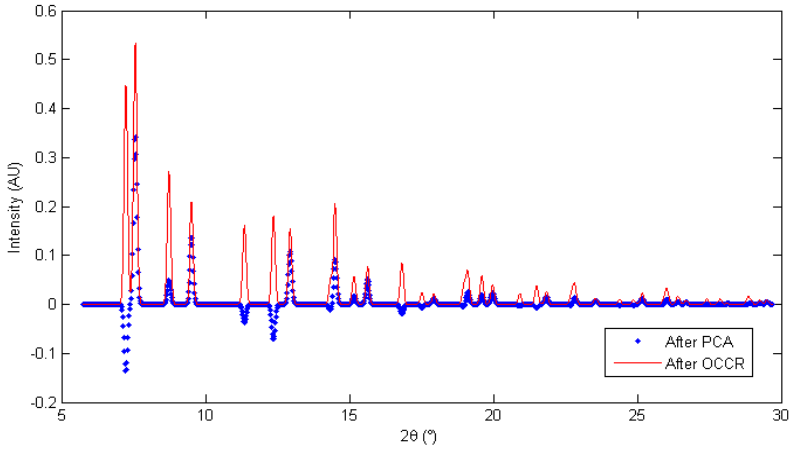
Results on Simulations

System response	Acronym	PCA	OCCR	PSD*
Sinusoidal [0 0.4]	SI	0.940	1.000	1.000
Triangular [0.8 1]	TR	0.478	1.000	1.000
Square, slow decay [0 1]	SQ	0.695	1.000	1.000
Square, fast decay [0 1]	SQ_fast	0.704	1.000	0.860
Square, asymmetric decay [0 1]	SQ_asym	0.684	1.000	0.316
Sinusoidal [0.8 1]	--	0.521	1.000	1.000
Ramp [0.8 1]	--	0.609	1.000	0.788
Ramp [0 1]	--	0.919	1.000	0.215

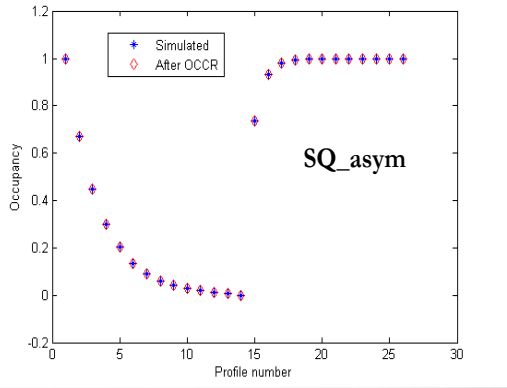
Correlation coefficient



The occupancy of the Cu atom is varied according to various functions. **Correlation coefficient** between the calculated XRD profile of the Cu atom and those obtained by PCA or OCCR decomposition, or by Phase Sensitive Detection demodulation (a traditional method). The intervals spanned by the occupancy values are in brackets



SI



Case study

Modulated Enhanced Diffraction

XPD data

Problem

A set of X-Ray Powder Diffraction data (XRPD), have been simulated by applying on the sample a known stimulus profile along time.

We want to retrieve, separately, the crystalline phases and the trend in time of the phases evolution. No prior knowledge of the model is supposed, although the data may behave accordingly to two models:

- Case 1: Two crystalline phases, without active atom [$CuFe_2O_4$+Cu]
- Case 2: A single crystalline phase [$CuFe_2O_4$] and one active atom species [Cu]

Recall PCA contribution

It has been already observed that **Principal Component Analysis** is able to separate the contributions forming the dataset supposing the different components **uncorrelated**.

In detail,

PCA *scores* explain the time trend of the crystalline phases,

PCA *loadings* express the pure spectra, if uncorrelation among components is a reasonable hypothesis.

If the crystalline model is simple, the components are expected to be well separated and PCA working well.

Case 1: Mathematical Model

The specific case study analyzed in simulation concerns:
$CuFe_2O_4$+Cu, a case in which there are two crystalline phases and no active atoms.
The mathematical model underlying the change of spectra evolution with time is the following:

$$X(2\vartheta,t) = m(t)\cdot\left|F_1(2\vartheta)\right|^2 + n(t)\cdot\left|F_2(2\vartheta)\right|^2$$

$$n(t) = 1 - m(t)$$

where X(2ϑ,t) are the data, $F_1(2\vartheta)$ the first phase and $F_2(2\vartheta)$ the second crystalline phase.
The two phases have been simulated so that at any time they complement each other, i.e.

$$n(t) + m(t) = 1$$

- *Dataset name: 1_frazioni_in_peso_CuFe2O4_scala_lineare*

Conditions

$$X(2\vartheta,t) = m(t)\cdot\left[\left|F_1(2\vartheta)\right|^2 - \left|F_2(2\vartheta)\right|^2\right] + \left|F_2(2\vartheta)\right|^2$$

In **PCA**: PC1: it should follow the external stimulus

In **PCA**: loading1: it should have positive (related to $|F_1|^2$) and negative (related to $|F_2|^2$) parts

To analyze the results, the figures of merit used have been:

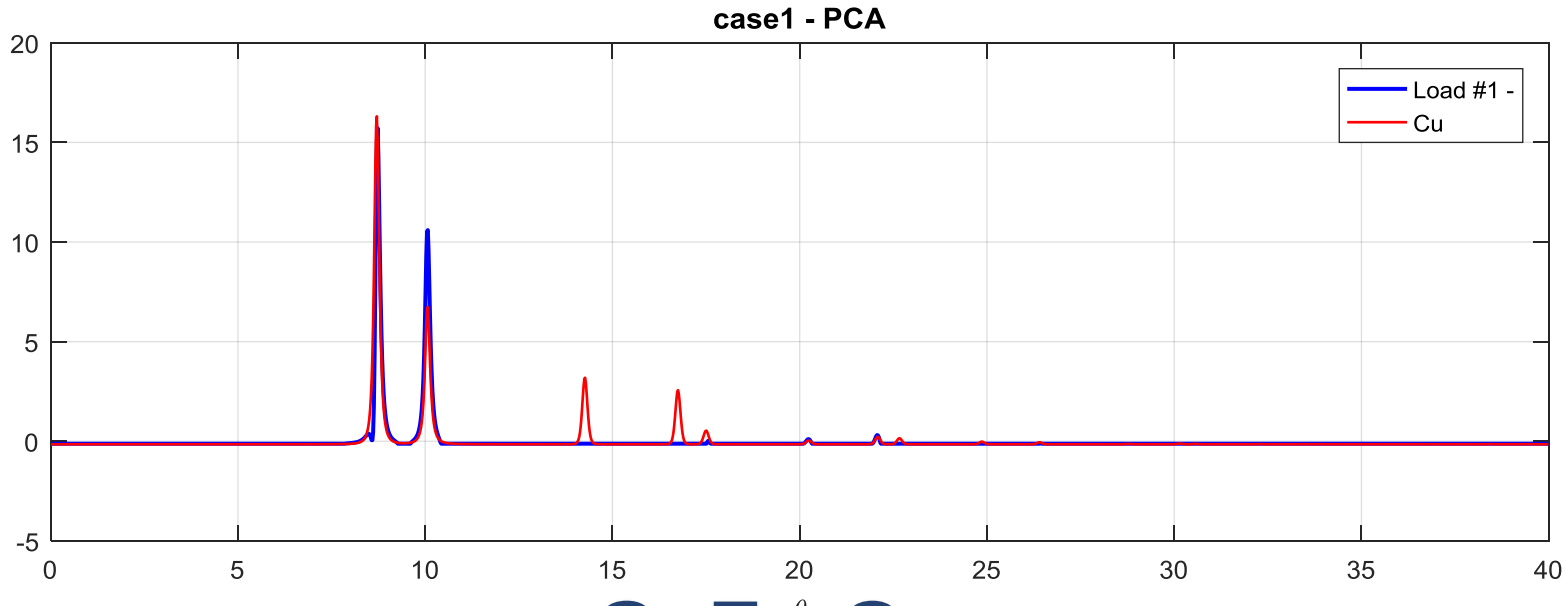
- Correlation between the linear stimulus with PC1 [only the knowledge of stimulus is supposed].
- Correlation of positive part of loading 1 with pure reference spectrum
- Correlation of negative part of loading 1 with pure reference spectrum

[although in practical situation the pure spectra are not known].

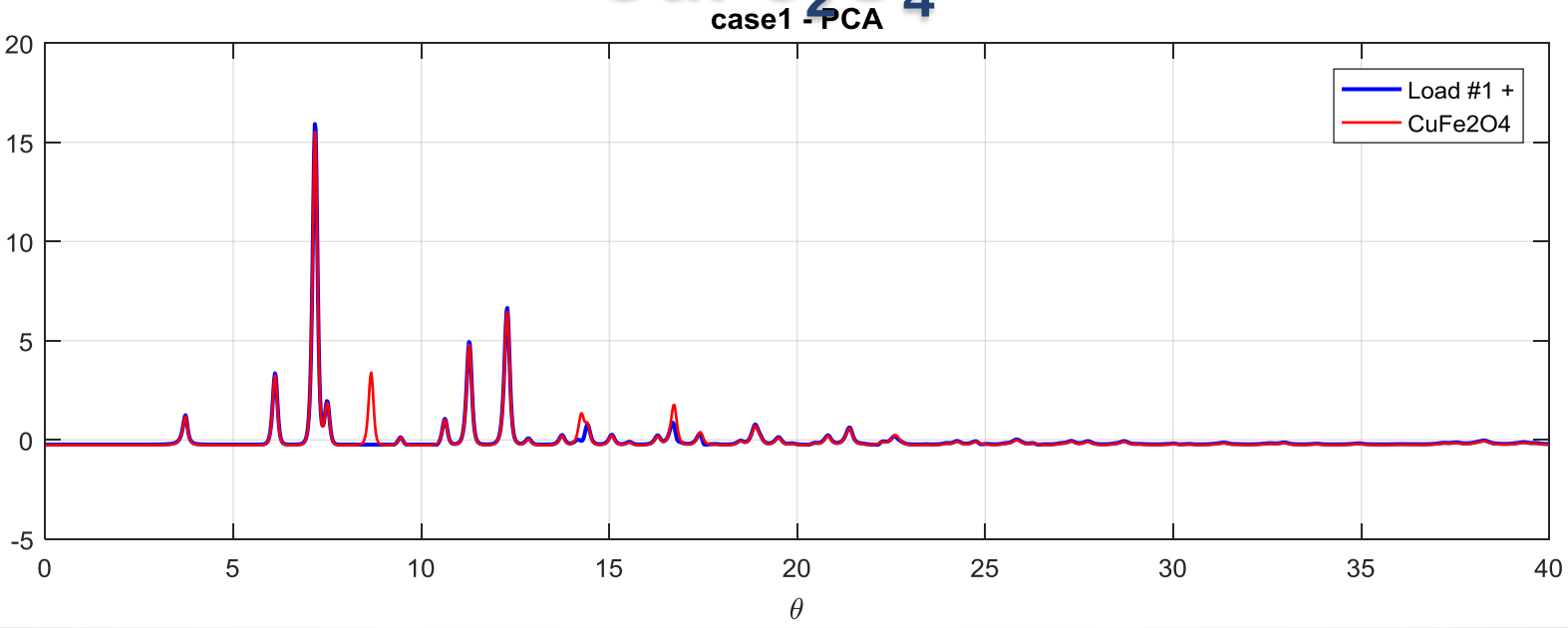
Correlation Results

Method	FoM type	FoM description	Value
PCA	INTRINSIC	Correlation coefficient of the first stimulus with PC1	-1.0000
	EXTERNAL	Correlation coefficient of loading 1+ with CuFe2O4	0.9998
		Correlation coefficient of loading 1- with Cu	0.9999

Cu


case1 - PCA
Load #1 -
Cu
20
15
10
5
0
-5
0
5
10
15
20
25
30
35
40
θ

$CuFe_2O_4$


case1 - PCA
Load #1 +
CuFe2O4
20
15
10
5
0
-5
0
5
10
15
20
25
30
35
40
θ

Case 2: Mathematical Model

The specific case study analyzed in simulation concerns a single crystalline phase with one active atom (Cu) species, the $CuFe_2O_4$+Cu.

The mathematical model underlying the change of spectra evolution with time is the following:

$$X(2\vartheta,t) = \left|m(t)F_a(2\vartheta) + F_s(2\vartheta)\right|^2 =$$

$$m^2(t)\cdot\left|F_a(2\vartheta)\right|^2 + 2m(t)\cdot\left|F_a(2\vartheta)\right|\left|F_s(2\vartheta)\right|\cdot\cos\delta + \left|F_s(2\vartheta)\right|^2$$

where X(2ϑ,t) are the data, $F_a(2\vartheta)$ is the spectrum of the active atoms (i.e. the ones responding to the external stimulus) and $F_s(2\vartheta)$ the spectrum of the silent atoms. It is expected that the behavior of the trend in the active atom *is somewhat related to the external stimulus* but in general it is unknown.

In the simulation of Case 2, the external stimulus is linear.

$$m(iT) = \frac{i}{N}, i =, 0..., N$$

Dataset name: 2_atomo_Cu_spinello_step_001_scale_ok_Cu_occ_0.87_scala

Conditions

$$X(2\vartheta,t) = m^2(t)\cdot\left|F_a(2\vartheta)\right|^2 + 2m(t)\cdot\left|F_a(2\vartheta)\right|\left|F_s(2\vartheta)\right|\cdot\cos\delta + \left|F_s(2\vartheta)\right|^2$$

In **PCA**: PC2: it should follow the square of the external stimulus

In **PCA**: loading2: it should be only positive

In **PCA**: PC1: it should follow the external stimulus

In **PCA**: loading1: it should have positive and negative parts

To analyze the results, the figures of merit used have been:

- Correlation between the linear stimulus with PC1; quadratic with PC2; positivity of loading 2 [only the knowledge of stimulus is supposed].
- Correlation of loading 2 with pure reference spectrum of active atoms

[although in practical situation the pure spectra are not known].

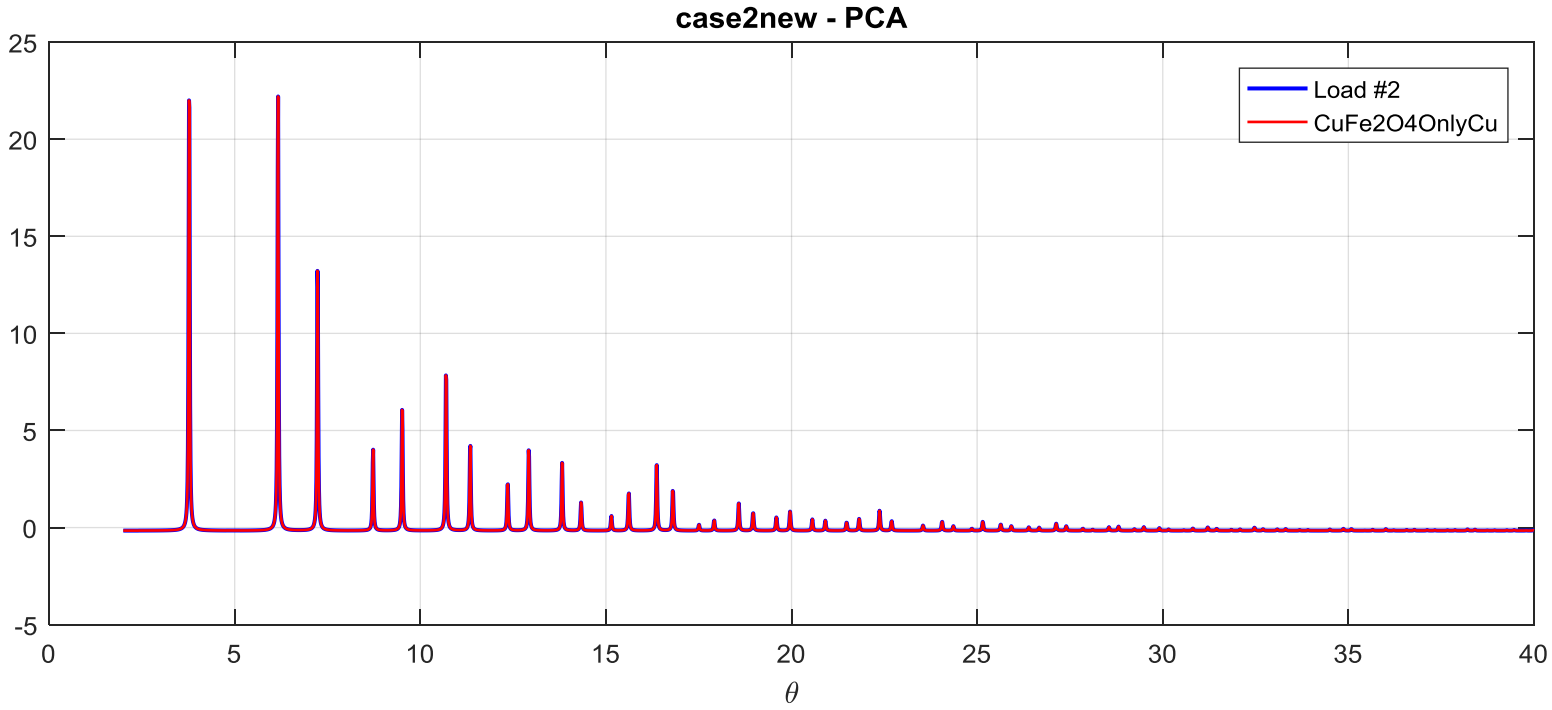
Correlation Results

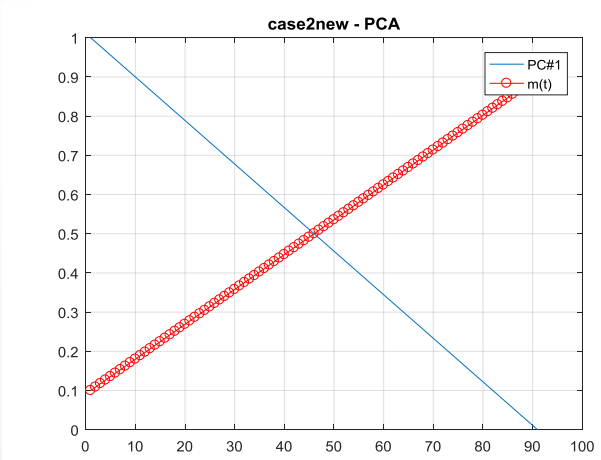
FoM type	FoM description	PCA	OCCR load2	OCCR corr coef	OCCR comb
INTRINSIC	Positivity degree of loading 2	1.0000	1.0000	1.0000	1.0000
	Correlation coefficient of PC2 with $PC1^2$	0.9998	0.9998	1.0000	0.9998
	Correlation coefficient of PC1 with m(t)	-1.0000	-1.0000	-1.0000	-1.0000
	Correlation coefficient of PC2 with $m(t)^2$	1.0000	1.0000	1.0000	1.0000
INTRINSIC	**Geometric mean of the previous figures**	**1.0000**	**1.0000**	**1.0000**	**1.0000**
EXTERNAL	Correlation coefficient of loading 2 with CuFe2O4-OnlyCu	0.9978	0.9978	1.0000	0.9978

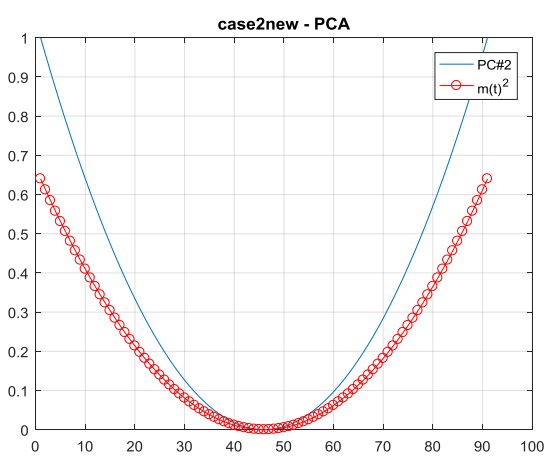
PERFECT!

Different running conditions for OCCR (i.e. different optimality criterion applied):
Load2: highest positivity of second loading
Corrcoef: highest correlation coefficient of PC_1^2 and PC_2
Combined: geometric mean of the previous figures.

$CuFe_2O_4$ only Cu







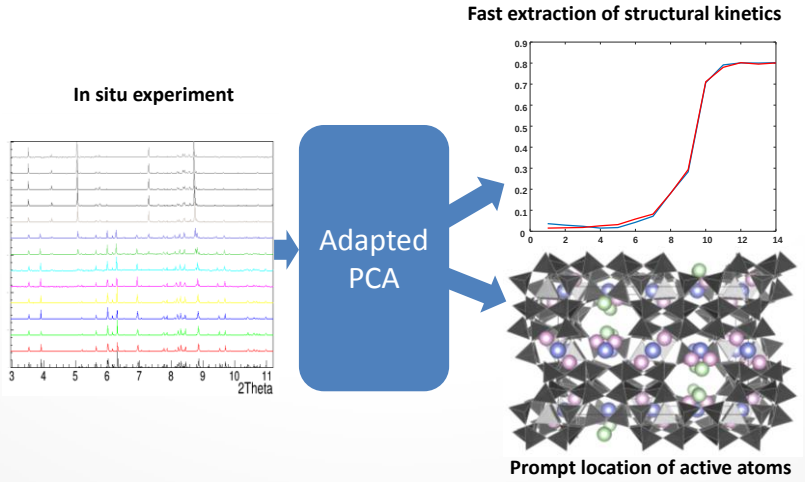
Case study:

Kinetics of Solid-state reaction

X-ray Diffraction profiles during 2 solid phases changes

General purpose of the study:

- Analysis of a two solid state transformation and **estimation of the kinetic triplet parameters**.
- The kinetic has been investigated through X-ray Powder Diffraction method, collecting a set of spectra as a function of temperature (in case of non-isothermal experiment) or as a function of time (in case of iso-thermal experiment).
- The general idea is that the spectra may capture information about the kinetic of transformation and then that it is possible to infer equation parameters observing the transformation of the spectra with time or temperature.



Solid-state transformation basis /1

Solids transformation from one crystalline phase (state of matter) into another has been observed.
Said α the extent of conversion, the following dynamic equation holds:

$$\frac{d\alpha}{dt} = K(T) \cdot f(\alpha)$$

where K(T) is a temperature-dependent reaction rate and f(α) a kinetic-dependent model function.

The Arrhenius equation links explicitly K to temperature:

$$K(T) = A \cdot \exp\left(-\frac{E_a}{RT}\right)$$

with E_a the activation energy of the reaction, R the universal gas constant and T the temperature (*A* is called frequency factor and it is an unknown, together with E_a).

Solid-state transformation basis /2

The triplet {A,E_a,f(α)} is called kinetic triplet and characterizes a unique decomposition reaction.
Some models for f(α) are reported in literature and, highlighted in green, the ones used in the experiments of our interest.

No.	Symbol	Reaction model	$f(\alpha)$
1	P_1	Power law	$4\alpha^{3/4}$
2	P_2	Power law	$3\alpha^{2/3}$
3	P_3	Power law	$2\alpha^{1/2}$
4	P_4	Power law	$2/3\alpha^{-1/2}$
5	R_2	Phase-boundary controlled reaction(contracting area, *i.e.* bidimensional shape)	$2(1-\alpha)^{1/2}$
6	R_3	Phase-boundary controlled reaction(contracting volume, *i.e.* tridimensional shape)	$3(1-\alpha)^{2/3}$
7	F_1	First-order (Mampel)	$(1-\alpha)$
8	A_2	Avrami-Erofeev($n=2$)	$2(1-\alpha)[-\ln(1-\alpha)]^{1/2}$
9	A_3	Avrami-Erofeev($n=3$)	$3(1-\alpha)[-\ln(1-\alpha)]^{2/3}$
10	D_1	One-dimensional diffusion	$1/2\alpha$
11	D_2	Two-dimensional diffusion (bidimentional particle shape) Valensi equation	$1/[-\ln(1-\alpha)]$
12	D_3	Three-dimensional diffusion (tridimentional particle shape) Jander equation	$3(1-\alpha)^{1/3}/2[(1-\alpha)^{-1/3}-1]$

Z.A. Alothman, R. Mahfouz, 'Kinetic Studies of the Non-Isothermal Decomposition of Unirradiated and gamma-Irradiated Gallium Acetylacetonate', Progress in Reaction Kinetics and Mechanism - May 2010

Optimization: general strategy

XPD data have been taken during transformation between two phases with the purpose of estimate the kinetic parameters:

$$\{A, E_a, n\}$$

Principal Component Analysis has been used on the dataset. In detail, the first score has been supposed to follow the general trend of the implied transformation

$$\mathbf{t}_1 \propto \alpha$$

The idea is to relate the first score with the explicit expression of α derived from the models, which is function of the three unknowns {A,E_a,n}.
For a given set of the triplet, it is possible to infer the expression of α, that is used to force the decomposition so that α is just the first score.

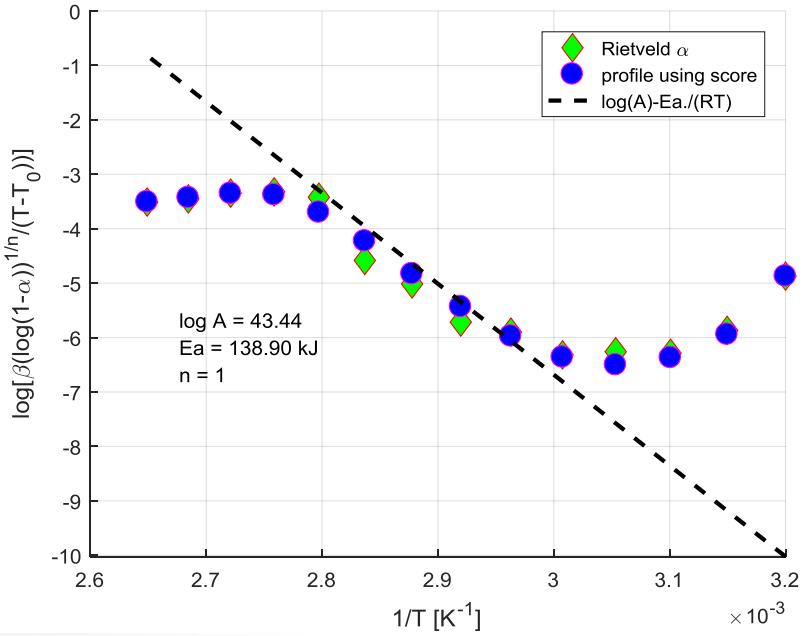
CR model

$$\alpha = 1 - \exp\left[-T^{2n} \left[A \cdot \exp\left(-\frac{E_a}{RT} \right) \right]^n \right]$$

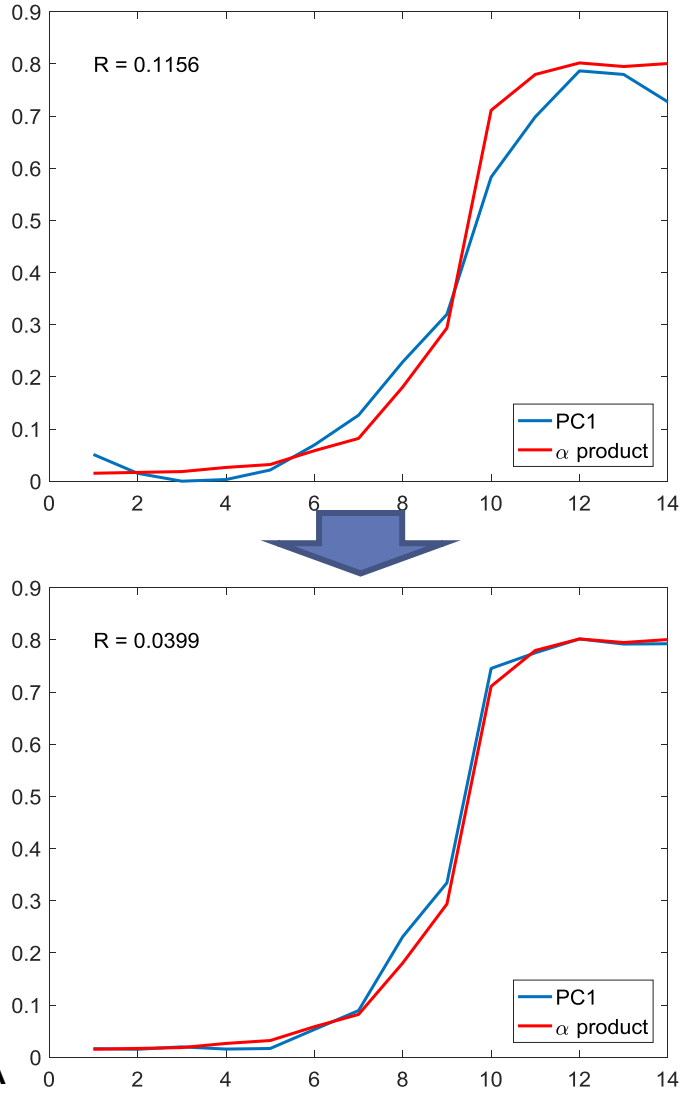
KC model

$$\alpha = 1 - \exp\left[-\left(\frac{(T - T_0)}{\beta} A \right)^n \cdot \exp\left(-\frac{nE_a}{RT} \right) \right]$$

Diffractograms: Naphtalene dataset



PCA



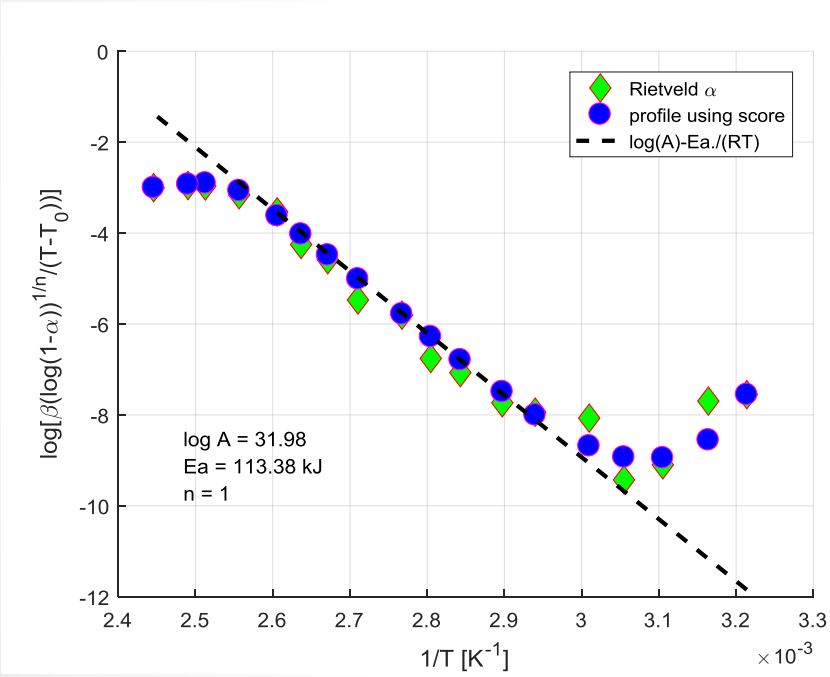
Constr PCA

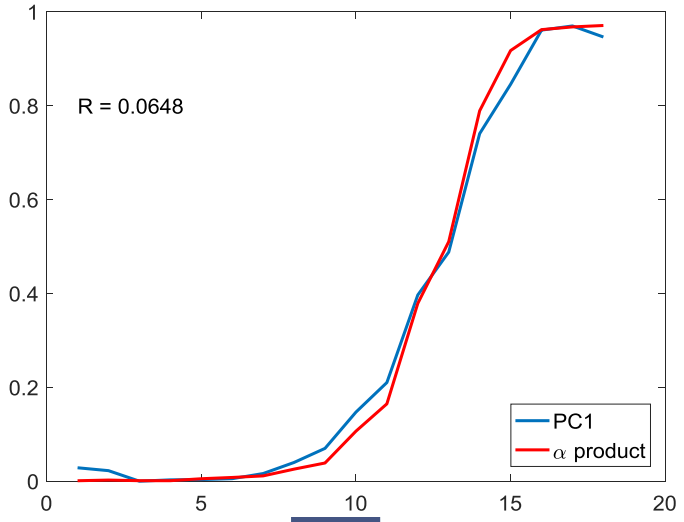
Comparison with Rietveld refinement method (based on least square approach, H. M. Rietveld, J. Appl. Crystallogr., 1969, 2, 65) that is more computationally intensive is in red)

fluorene (FL), naphthalene (NA) and anthracene (AN) as donor moieties and 7,7,8,8-tetracyanoquinodimethane (TCNQ) as an acceptor moiety

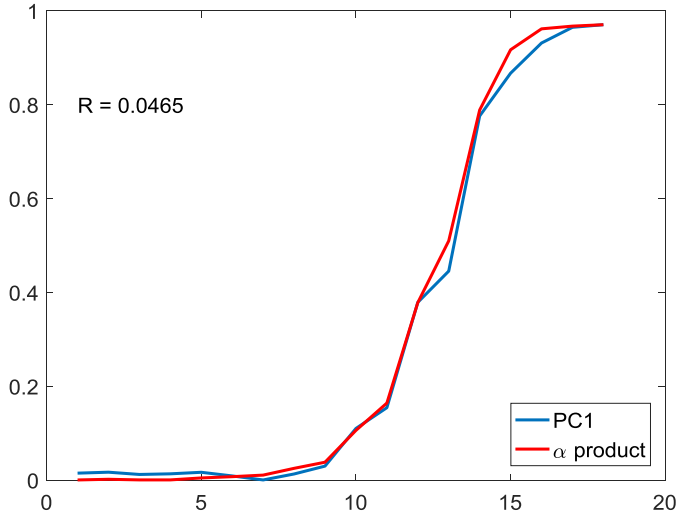
Diffractograms: Fluorene dataset

PCA

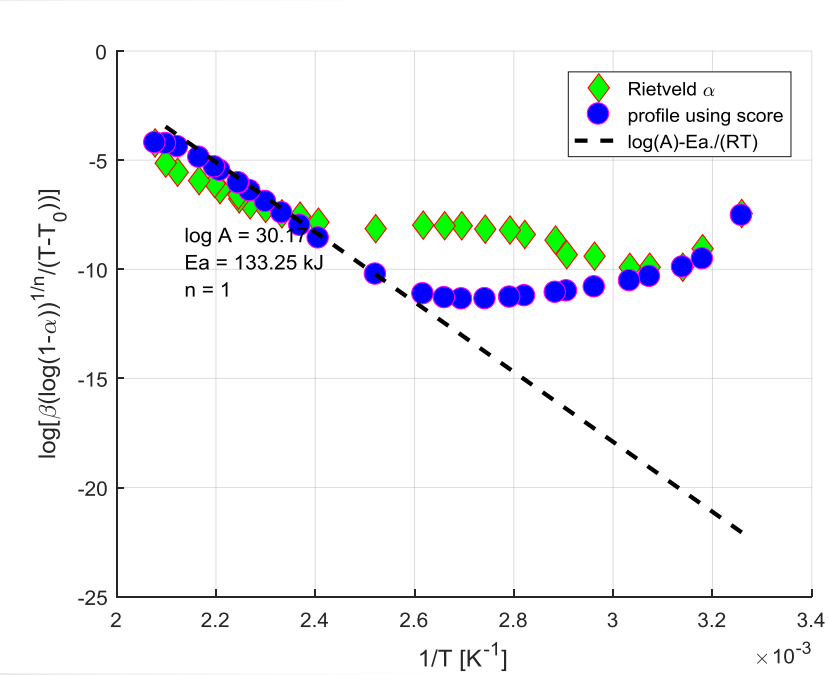




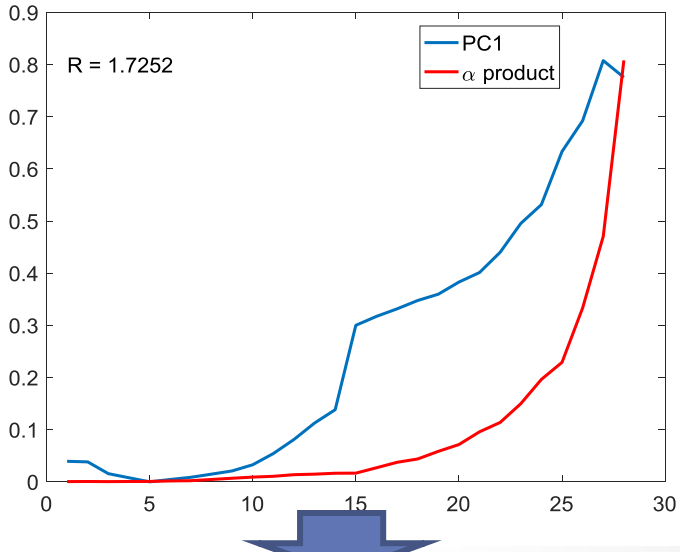
Constr PCA

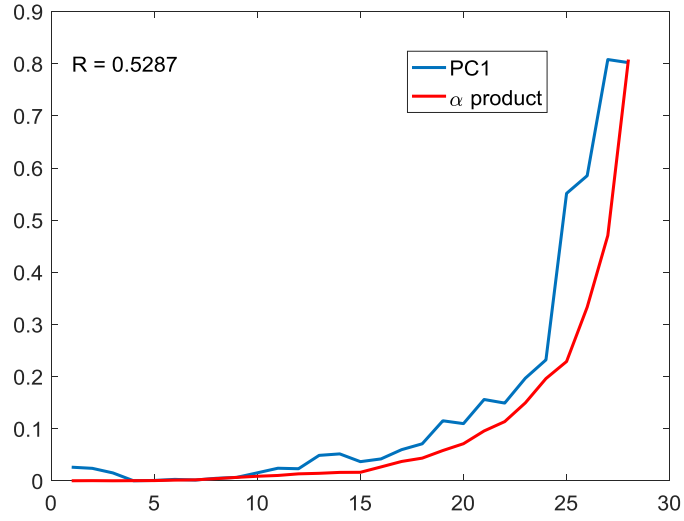


Diffractograms: Anthracene dataset



PCA





Constr PCA

Discussion & Conclusions

Multivariate Analysis performs decomposition of pure spectrum and stimulus in Modulated Enhanced Diffraction of X-Ray Powder Diffracted Data. It has been used also to infer the kinetic reaction parameters

Novelty w.r.t. traditional methods:

- No need to know the underlined model, at least in principle,
- Very accurate decomposition for simple models, good accuracy for more complicated models,
- Fast and completely automated method. In RootProf PCA is implemented; OCCR and constrained-PCA (for triplet estim.) in future versions

Limits:

- Some problem with the sign of the loadings (positive/negative)
- Decomposition supposes uncorrelated spectrum, which is not exactly the truth
- First score could not contain all the ‘trend’ of the dataset.