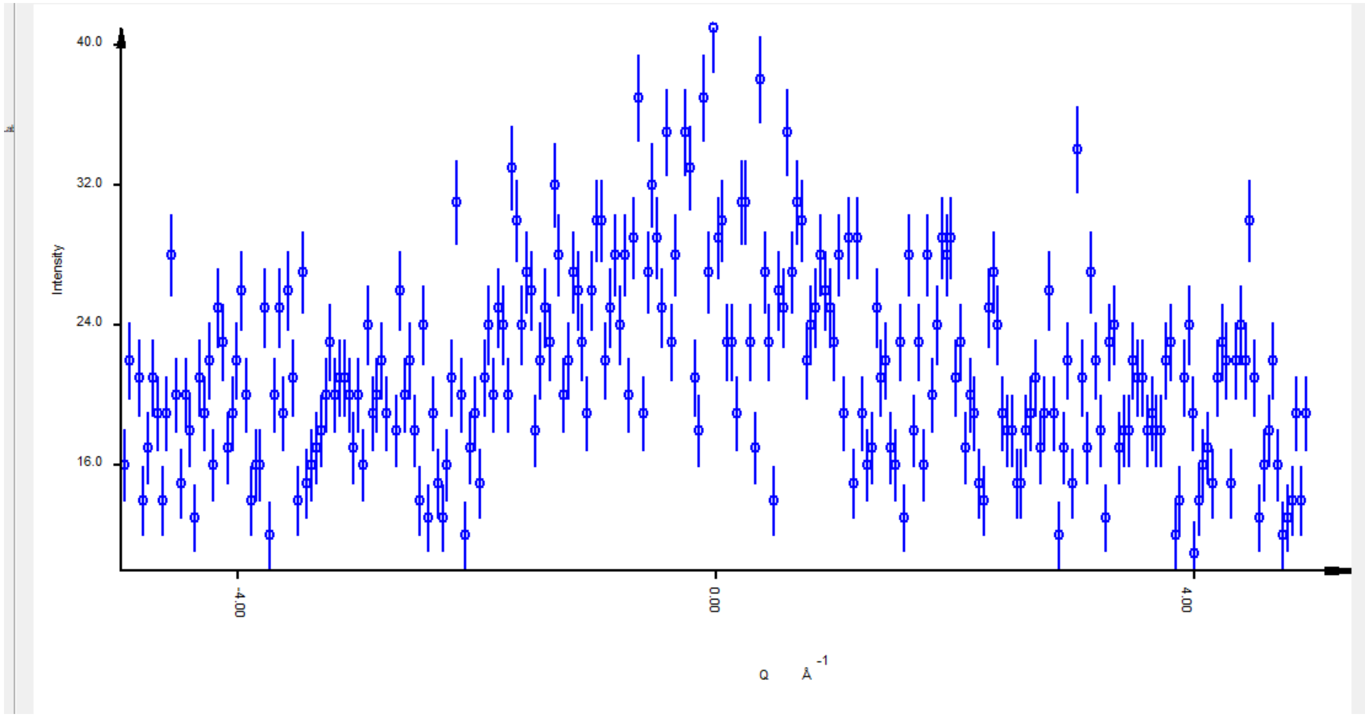
2023 CETS school

Data treatment

G. Pépy
gpepy@laposte.net
Budapest Neutron Center, P.O. Box 49, H-1525 Budapest, Hungary

After hard work you have got nice data...

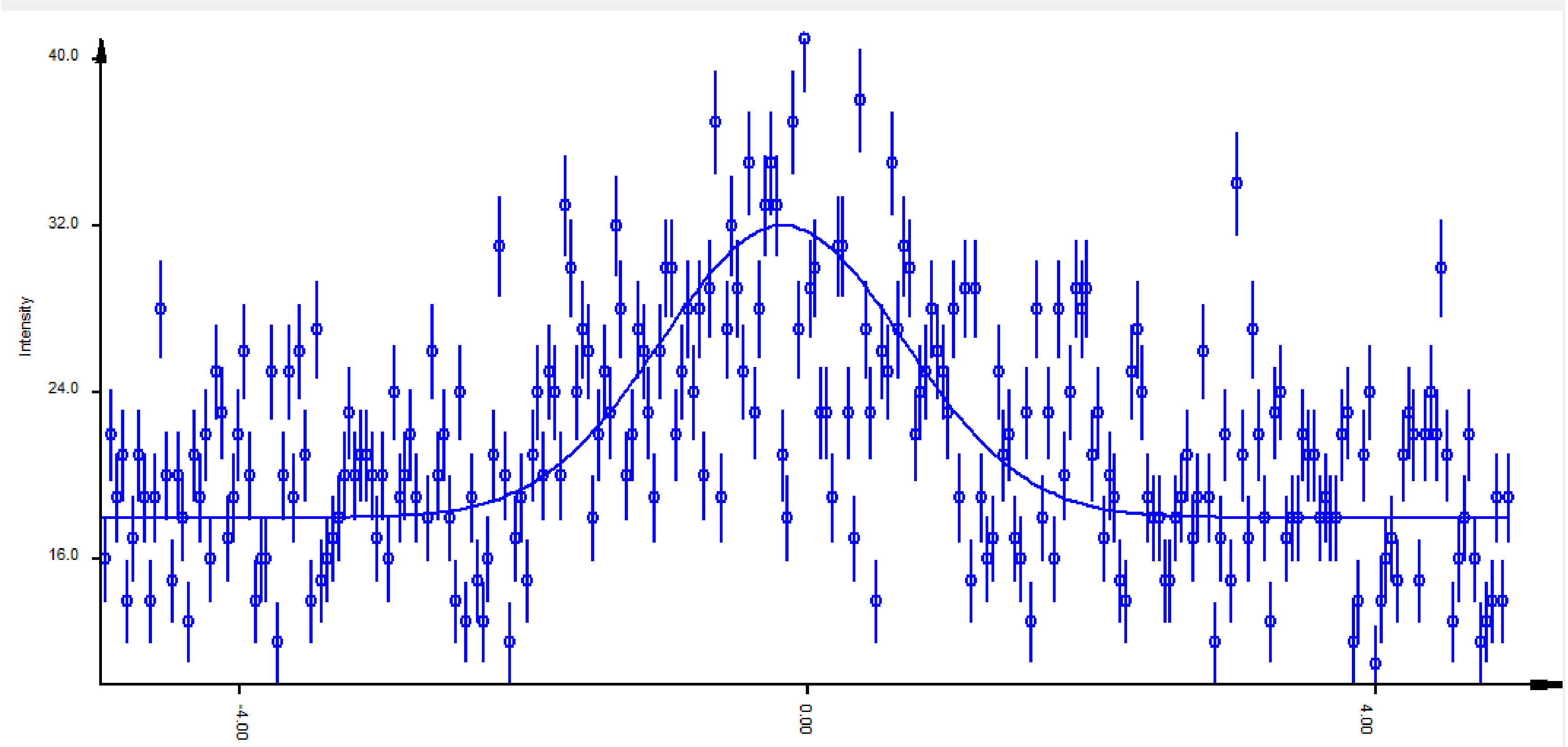


File example from Kuklin A., FLNP, JINR, Dubna

You just want to click on a button...

Run the fit

In order to get a nice fit !



However there is a high probability that you get into trouble...

Hopefully this lecture will help you understand the fit process and avoid traps.

Summary:

1 - probabilities

2 - data treatments

3 - examples

Most texts and pictures were found in Wikipedia articles, unless otherwise quoted.

$$Var(X) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \lambda)^2$$

Some definitions

- X a discrete random variable

- λ its average

$$\lambda = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- var(X) its variance

It is a measure of its dispersion

$$Var(X) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \lambda)^2$$

- σ its standard deviation

$$\sigma = \sqrt{Var(X)}$$

-The variance of the sum of two uncorrelated random variables is the sum of their variances:

$$Var(X+Y) = Var(X) + Var(Y)$$

which is not true for the standard deviation !

Binomial distribution

A binomial distribution with parameters ***n*** and ***p***

- is the discrete probability distribution of success

- in a sequence of ***n*** independent experiments,

- each asking a yes–no question,

- *success* (with probability ***p***) or *failure* (with probability ***q* = 1 − *p***)

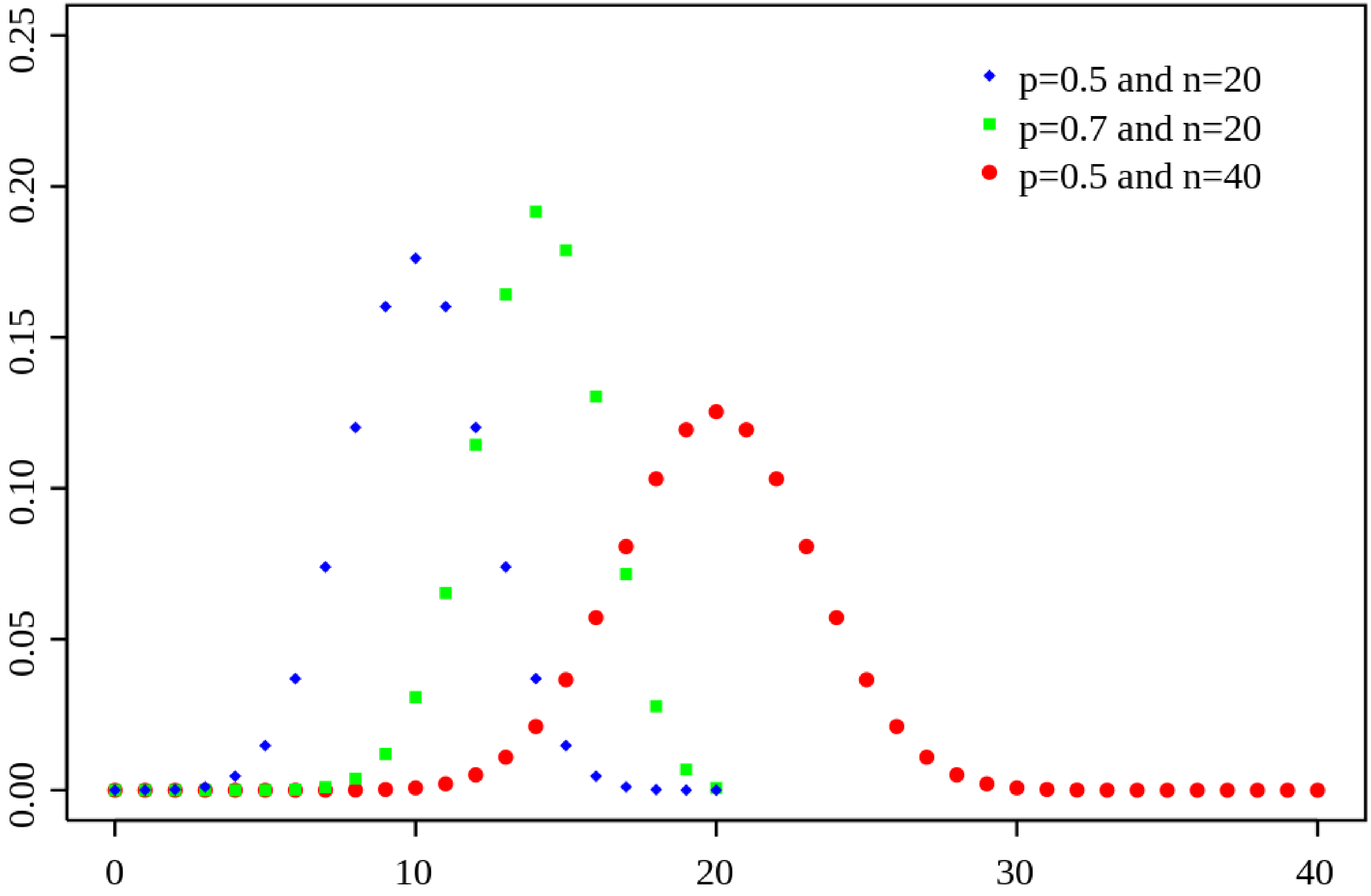
Probability of having k successes in n trials:

$$\Pr(X = k) = \frac{n!}{k!\,(n-k)!}p^k(1-p)^k$$

Average $\lambda = np$ (if np is an integer)

Variance Var(X) = np.(1-p)

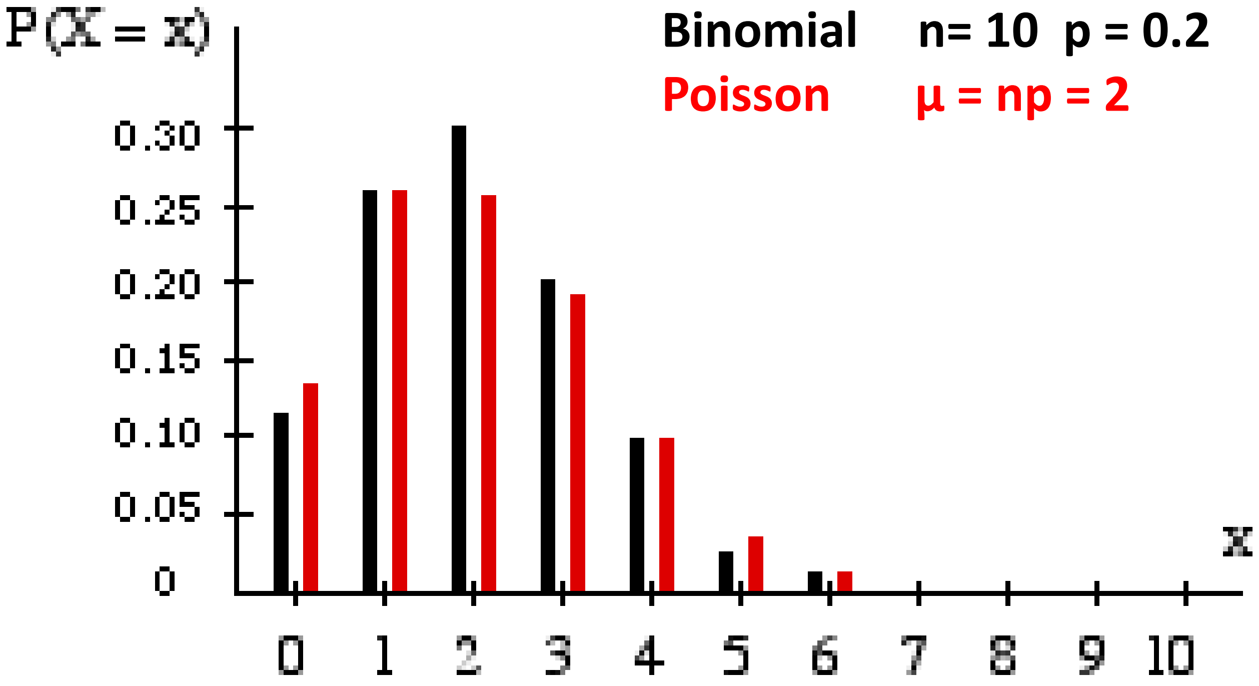
Binomial examples



Poisson approximation

In a binomial distribution if n is big and p is small a Poisson distribution is a good approximation.

Typically this approximation is good if n>20 and p<0.05



Binomial n= 10 p = 0.2

Poisson $\mu = np = 2$

Poisson distribution

A discrete probability distribution for the probability of
- a given number of events
- occurring in a fixed interval of time or space
- if these events occur with a known constant mean rate
- and independently of the time since the last event.

This is what happens in a gas neutron detector !

Poisson distribution	$\Pr(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$
Average	λ
Variance	**Var(X) =** λ

Central limit theorem

CLT states that, in many situations,
when independent random variables are summed up,
their properly normalized sum tends toward a normal distribution even if the original variables themselves are not normally distributed.

Therefore for large values, say λ>1000, the normal distribution is an excellent approximation to the Poisson distribution.
If λ > 10, then the normal distribution is a good approximation if correction is performed, i.e., if P($X \le x$) is replaced by P($X \le x + 0.5$).

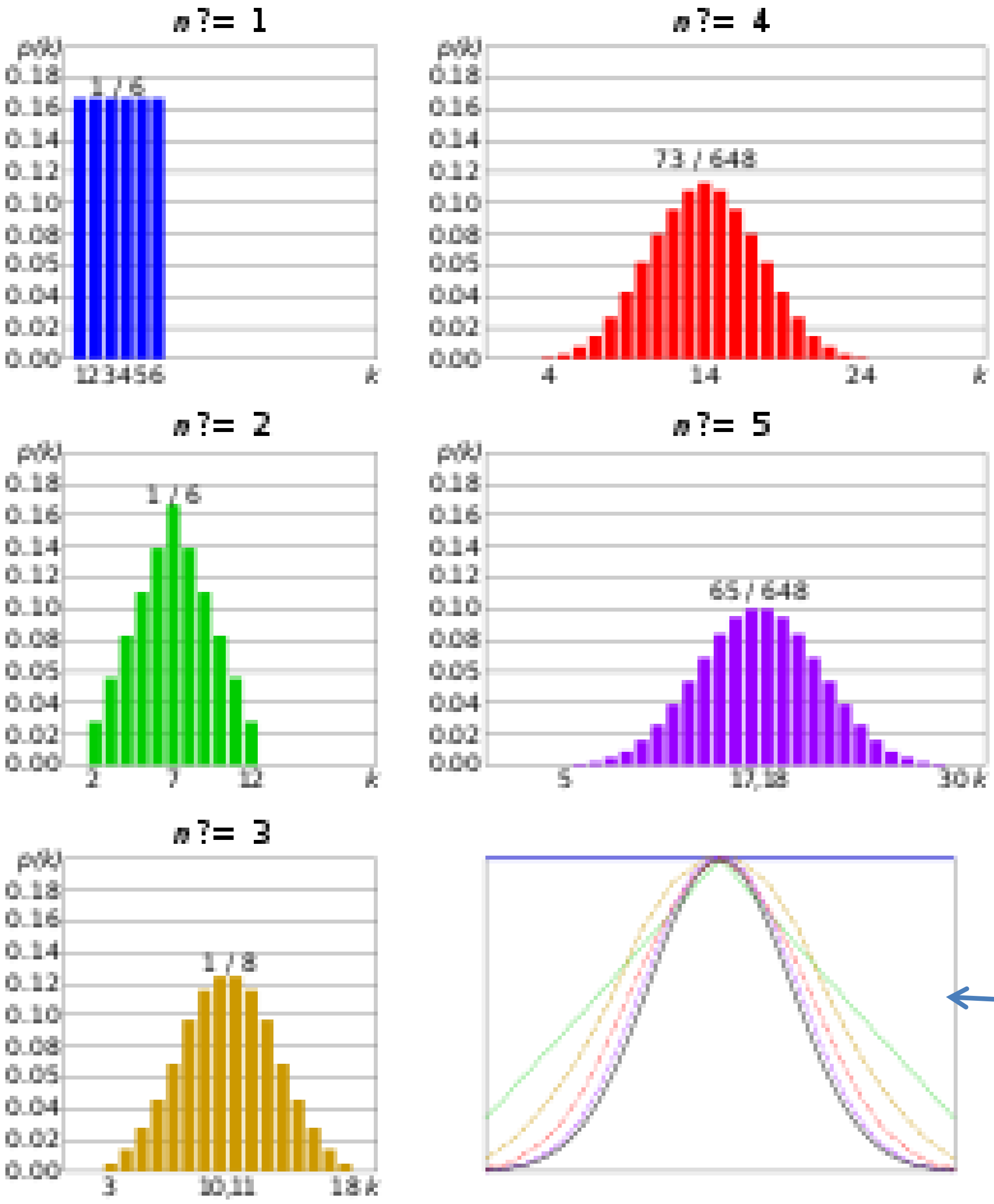
The normal distribution

In many detectors (notably gas detectors, but not CCDs !) counting follows a Poisson distribution law. As soon as the counting rate is high the « normal law » becomes a good approximation.
It is much easier to handle.

Normal distribution:

$$\Pr(X = x) = \frac{exp\left(-\frac{1}{2}\left(\frac{x-\lambda}{\sigma}\right)^2\right)}{\sigma\sqrt{\pi}}$$

Average	λ
Variance	**σ^2**
Mean square deviation	**σ**



Probability functions, p (k)
for the sum of n fair
6-sided dice

Smoothed profiles of the previous graphs are rescaled, superimposed and compared with a normal distribution (black curve).

Problem with the Poisson -> normal approximation:
the normal distribution is symetric
which means that for small numbers
exists a significant non-zero probability for negative occurences !

Impossible in real life !

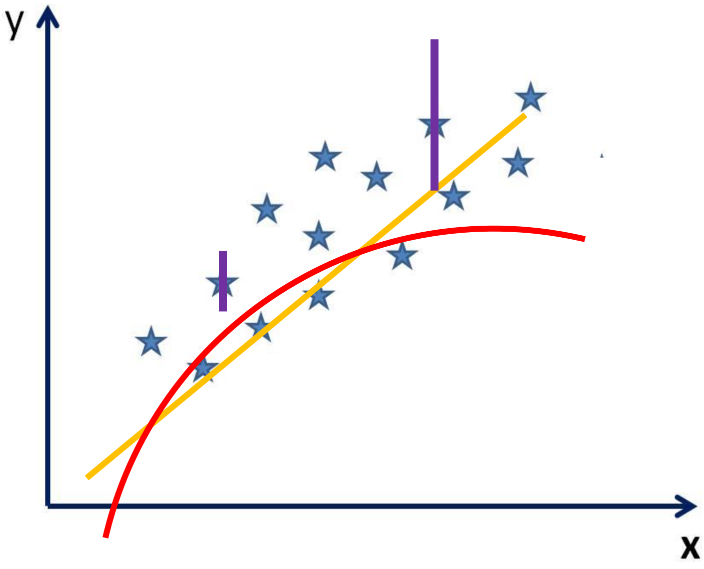
Fitting a model

Whatever the model appropriate for the observed scattering, you need a test function to appreciate the quality of the fit.
Such a function is a distance.
Several distances are known,
for 2 points with coordinates {x_i}, {y_i}:

$$1 - norm\ distance \quad d = \sum_{i=1}^{n} |x_i - y_i|$$

$$Euclidian\ distance \quad d = \left(\sum_{i=1}^{n} |x_i - y_i|^2\right)^{\frac{1}{2}}$$

$$p - norm\ distance \quad d = \left(\sum_{i=1}^{n} |x_i - y_i|^p\right)^{\frac{1}{p}}$$



Main property of a distance: scalar positive function

Discussion about distances in fitting by W.I.F. David in J. Res. Natl. Inst. Stand. Technol. 109, 107-123 (2004)

Distance for SANS data

$$\chi^2 = \frac{1}{N-p}\sum_{i=1}^{i=N}\left(\frac{\mathrm{I_i} - \mathrm{Y_i}(\{\mathrm{P}\})}{\Delta \mathrm{I_i}}\right)^2$$

Least squares

I_i	intensity in pixel i
N	number of data points
p	number of free parameters
{P}	set of parameters
Y_i	calculated intensity for pixel i
ΔI_i	uncertainty of I_i

If the random variables I_i are independent and follow a normal distribution, assuming that the normal law approximation is valid:

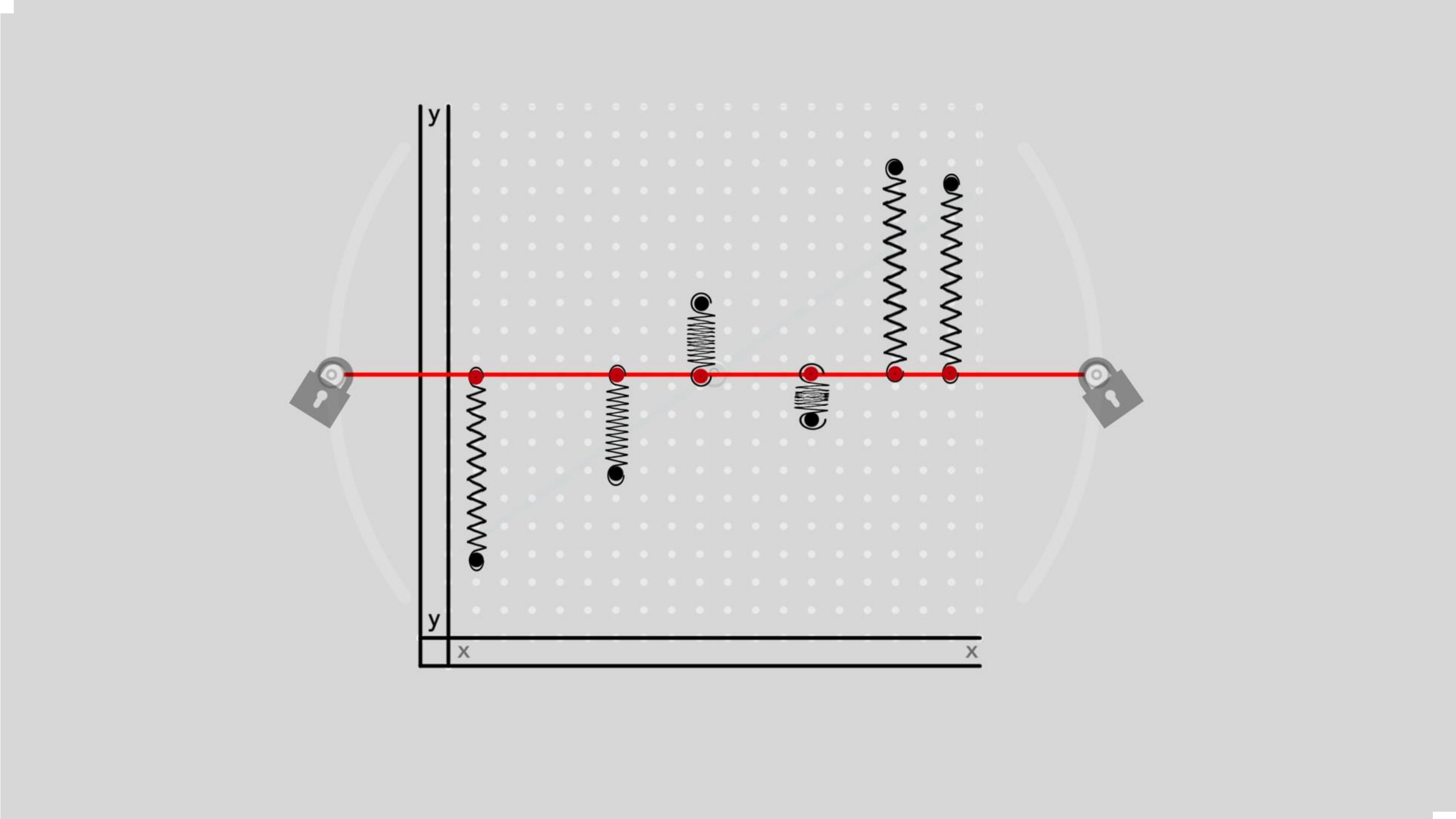
$$\Delta \mathrm{I_i} = \sqrt{\mathrm{I_i}}$$

then

$$\chi^2 \rightarrow 1$$

This is a good test for the fit.

Least squares : how does it work?



https://www.jmp.com/en_in/statistics-knowledge-portal/what-is-regression/the-method-of-least-squares.html

Fitting processes

1 – steepest descent (gradient)

Obvious method:

Calculate the gradient of the test function and make parameter increments in the opposite direction.

The main problem is to determine what are « good » increments. Usually one calculates the test function at a set of increments along the gradient opposite direction, in order to evaluate a good step.

Not efficient. No garantee to avoid a local minimum of the test function.

2 – least squares

Somewhat improved method.

- calculate a developpment
  of the χ² test function up to 2nd order

$$\chi^2 = \chi^2(0) + \sum_{i,j} \delta p_i \frac{\partial \chi^2}{\partial p_i} + \frac{1}{2} \delta p_i \delta p_j \frac{\partial^2 \chi^2}{\partial p_i \partial p_j}$$

- calculate the 1st order derivatives
  versus the parameters

$$\frac{d\chi^2}{dp_i} = \frac{\partial \chi^2}{\partial p_i} + \sum_{j} \delta p_j \frac{\partial^2 \chi^2}{\partial p_i \partial p_j}$$

- hypothesis : close to the χ² minimum
  all derivatives are 0

$$\frac{d\chi^2}{dp_i} = 0$$

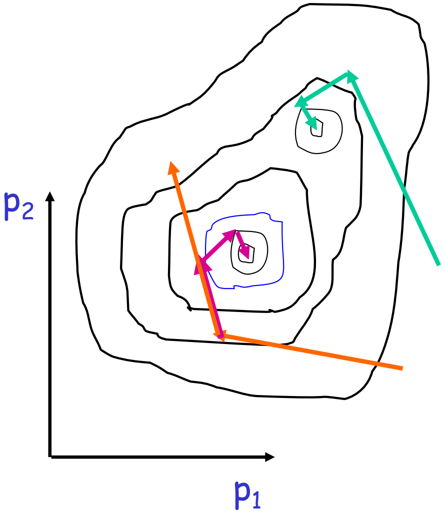
- this hypothesis provides a
linear equation system
[α] is the *curvature matrix*

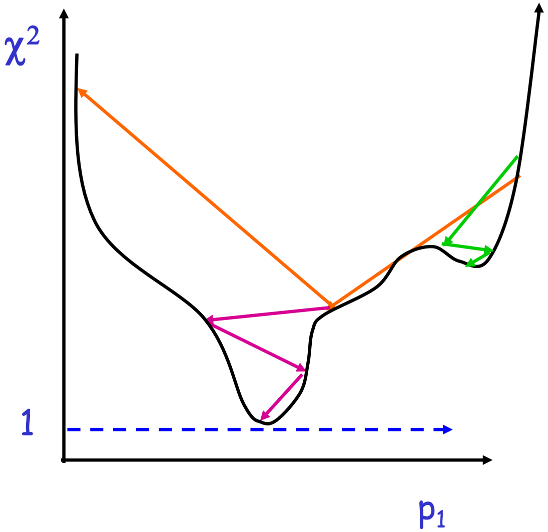
$$\sum_{j} \delta p_j \frac{\partial^2 \chi^2}{\partial p_i \partial p_j} = - \frac{\partial \chi^2}{\partial p_i} \quad or \quad [\alpha]\overrightarrow{\delta p} = \vec{\beta}$$

- its solution (matrix inversion)
provides a vector of parameters increments.
[C] is the *covariance matrix*

$$\overrightarrow{\delta p} = [\alpha]^{-1}\vec{\beta} \quad or \quad \overrightarrow{\delta p} = [C]\vec{\beta}$$

With some luck this vector leads to a point in the parameter hyperspace where χ^2 is smaller. It may not be the case as all this is highly non linear; one may try to shorten the vector keeping the same direction…
Then one repeats (iterates) the process.

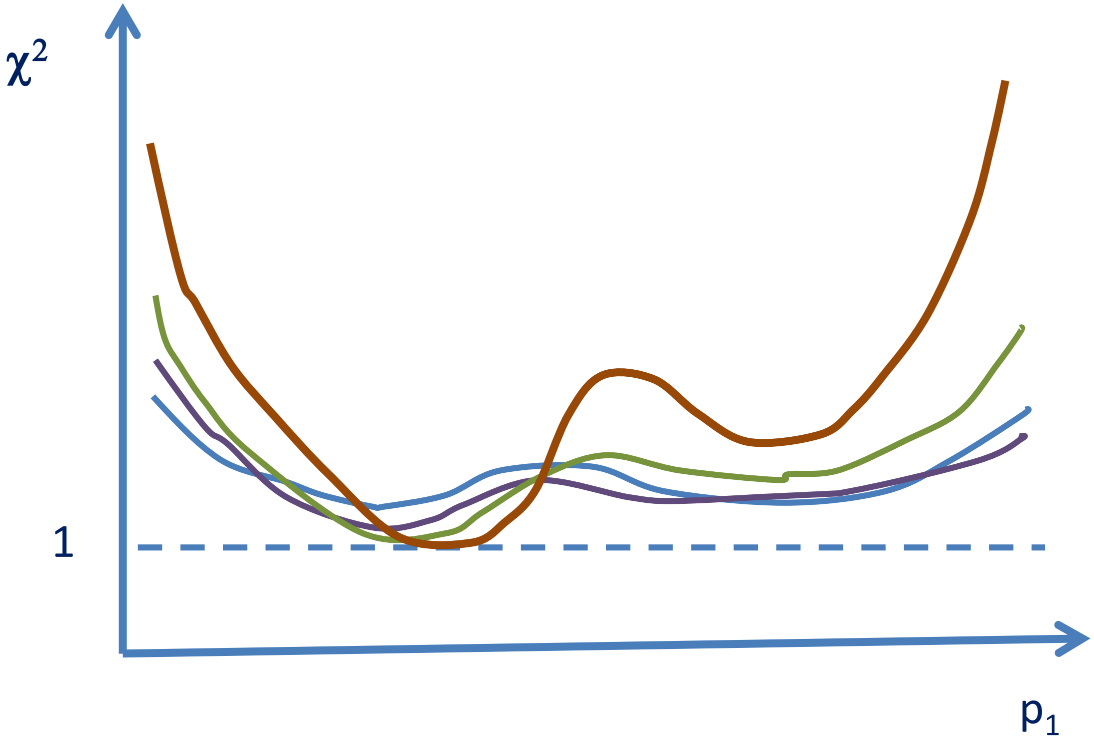




Some problems:

- non linearity
- the minima depths differentiate best if the pixels uncertainties are small
- systematic errors in the data or the model -> $\chi^2 \neq 1$
- nature of the probability law: Poisson or… worse
- bad evaluation of the pixel uncertainties

χ^2 in parameters hyperspace, statistics, phase transition



Notably, if the χ^2 is not a quadratic form this process is not efficient
(the hypothethis and calculation on derivatives are bad)

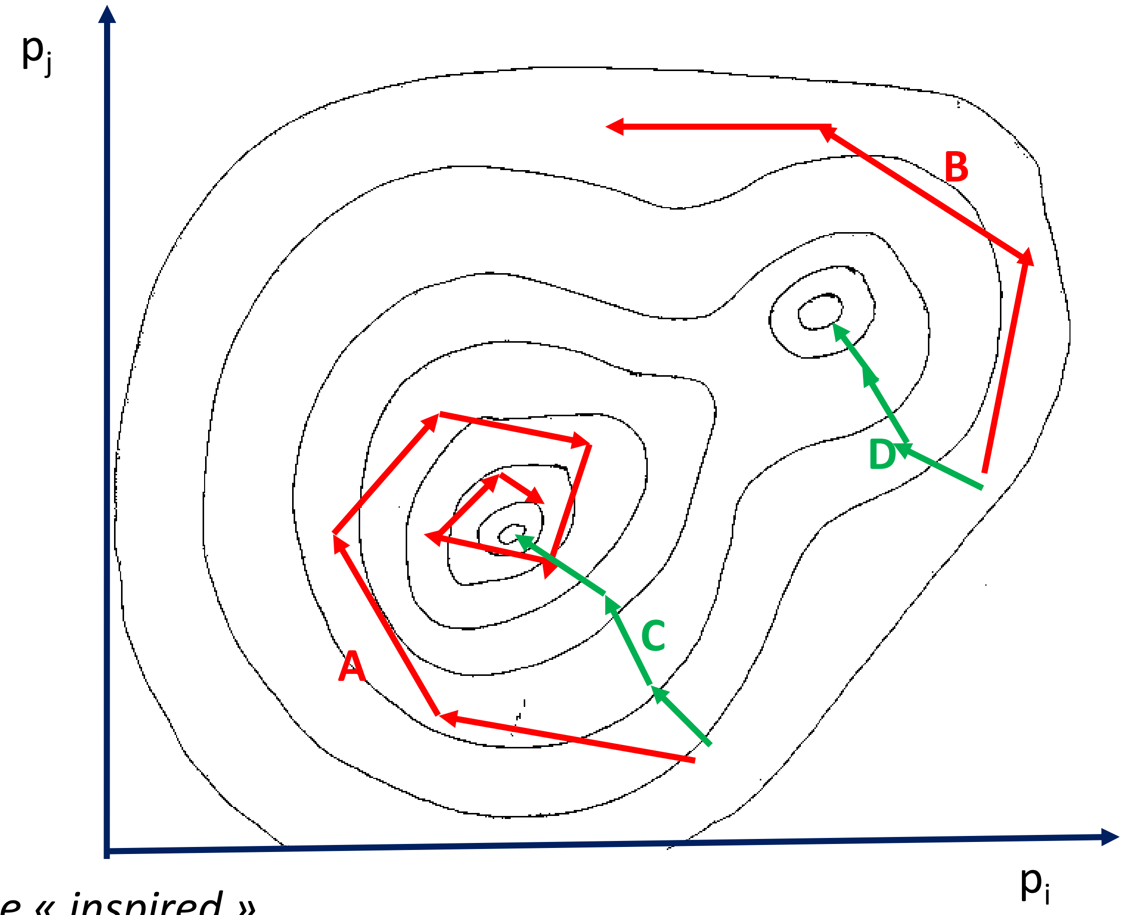
One prefers the **steepest descent method**,
taking a step C along the 1st derivative vector
C is now simply a scalar.

Levenberg-Marquardt have proposed a clever method
to pass from **“steepest descent”** to **“least square” :**

Replace **[α]** by **[α]’ = [α] + λ [I]**
(i.e. multiply the diagonal elements of the curvature matrix by 1 + λ)
1) start with a modest $\lambda \sim 1$,
2) compute **[α]** , **β** (and save it) and χ^2 and save them
3) calculate the parameter increments with **[α]’** diagonal χ^2 elements
4) compute new parameters and the corresponding χ^2
5) if the fit has converged, or too many iterations, stop !
6) if the fit improves, keep new parameters, divide λ by 10 and return to 2)
7) if the fit worsens, multiply λ by 10, return to 3)
(no new computation of **[α]** needed, so it is efficient)
NOTE - to obtain the proper error estimates on parameters
set $\lambda = 0$ for a final calculation.

least squares steepest descent Levenberg- Marquardt

Schematic Least Square & Steepest Descent, on a 2 parameters χ^2 surface



Least squares follows A
(well behaved, else might work
using partial shifts)
or B (blows up),
steepest descent follows C
or D (local minimum),
Marquardt steers
between B&D or A&C but
might fall into the local minimum.

λ is small for least squares or
λ is large for steepest descent

This slide and the previous one were « inspired »
from a talk at FLNP, Dubna, by Dr. Richard Heenan, ISIS Facility,

SANS experiments

smoothing

$$\frac{d\Sigma}{d\Omega} = Fw \frac{I_s/C_s - I_r/C_r}{I_c/C_c - I_b/C_b}$$

Pseudo calibration :
averaging lines and columns

$$C_i = \frac{(L_{SDi})^2}{A_i d_i M_i T_i}$$

The C coefficients index refer to:

s	sample
r	reference
c	calibration
b	background

with

A	beam area
d	sample thickness
M	monitor
T	transmission
L_{SD}	sample to detector distance
F_w	normalization prefactor

Before fitting SANS scattering, one has to make basic data corrections :
the most important is calibration (thanks to a flat scatterer like H_2O…).
Others consist in sample background or calibration background substractions.
Smoothing data may help avoid Poisson law.
Pseudo-calibration decreases the effect of calibration file uncertainties.
High quality experiments exhibit small uncertainties.

Data error bars (uncertainties)

$$\left(\frac{d\Sigma}{d\Omega}\right) = F_w \frac{I_s/C_s - I_r/C_r}{I_w/C_w - I_c/C_c}$$

Let

A area

d thickness

M monitor

T transmission

L

$$C_S = \frac{\boldsymbol{L^2_{SDs}}}{\boldsymbol{A_s d_s M_s T_s}}$$

Let the index hold for

s sample

r reference

w calibration (water)

c cuvette (sample container)

$$\Delta\left(\frac{d\Sigma}{d\Omega}\right) = F_w \left\{ \frac{\Delta I_s/C_s}{I_w/C_w - I_c/C_c} + \frac{\Delta I_r/C_r}{I_w/C_w - I_c/C_c} + \frac{\Delta I_w}{C_w}\frac{I_s/C_s - I_r/C_r}{(I_w/C_w - I_c/C_c)^2} + \frac{\Delta I_c}{C_c}\frac{I_s/C_s - I_r/C_r}{(I_w/C_w - I_c/C_c)^2} \right\}$$

Everything involved in the correction process will increase the final uncertainties. In his book about "Neutron diffraction", G.E Bacon (Oxford: Clarendon Press, 1955) already warned that ***«the uncertainty about background measurement must be as good as the uncertainty about sample data»***

Parameters error bars

The parameter error bars are best obtained after a least square fitting converges

Let us consider the χ^2 expression:
and the χ^2 change when a single parameter is changed

$$\chi^2 = \chi^2_{\min} + \overrightarrow{\delta p}^t [\alpha] \overrightarrow{\delta p}$$

$$\Delta\chi^2_1 = \chi^2_{M-1} - \chi^2_{\min}$$

It is possible to show that $\Delta\chi^2_1$ probability distribution is a normal law

Therefore one considers $\Delta\chi^2_1 = 1$
(which corresponds to a doubling of the χ^2 if χ^2_{min}=1 !)

and from the χ^2 development $\Delta\chi^2_1 = (\delta p_1)^2 / C_{11}$
with C11 the diagonal element of the covariance matrix for parameter 1

And as for $\Delta\chi^2 \cong 1$ $\delta p1 \cong 1\sigma$ (68% of cases, 1st standard deviation)

$$\delta p_i = \sqrt{C_{ii}}$$

Limitations of Least-Squares:

- data span many orders of magnitude -> another distance than the χ^2

For instance :absolute value instead of squares of the residuals

Robust fitting (FULLPROF for diffraction, now back to χ^2)

- significant outliers are present

Robust fitting proposes to modify the minimisation function by w_i weights.

$$\chi^2 = \frac{1}{N-p}\sum_{i=1}^{i=N}(w_i e_i)^2 \qquad e_i = \frac{I_i - Y_i\{P\}}{\Delta I_i}$$

Ii intensity in pixel i
N number of data points
P number of free parameters
{P} set of parameters
Yi calculated intensity in pixel i
ΔIi uncertainty of Ii
w_i weight
e_i estimator / residual

model	parameter k	$e_i < k$	$e_i > k$
Huber	1.345	$w_i = 1$	$w_i = k/e_i$
bisquare	4.685	$w_i = (1-(e_i/k)^2)^2$	$w_i = 0$

Huber P 1981 *Robust statistics* (New York: Wiley)

Fox J 2002 *An R and S-PLUS Appendix Companion to Applied Regression*

« Parabolic » smoothing

Table 1

1/16	1/8	1/16
1/8	1/4	1/8
1/16	1/8	1/16

Each pixel intensity is replaced by a weighted average over the neighbouring pixels

This modifies the uncertainty of pixel i which becomes:

$$\Delta \mathrm{Ii} = \sqrt{\sum_j \left(w_j . \Delta_j\right)^2}$$

Bevington P.R., Data reduction and error analysis, McGraw-Hill, (1969)

Pseudo-calibration file

Assume that the detector cells efficiency is mostly due to the pre-amplifiers efficiency. For most gaz detector there is one preamplifier for each line and column. Thence the trick: Calculate 2 vectors as the average of each line $\overrightarrow{A_i}$ and each column $\overrightarrow{B_j}$

and build a new calibration file C* as a normalised external product of these 2 vectors

$$[C^*] = \frac{1}{\hat{C}} \overrightarrow{A_i}^t \otimes \overrightarrow{B_j}$$

$\hat{C}$ is the average of the calibration file

The uncertainty is reduced by a factor equal to the number of cells in a row. This works very well for calibrators as vanadium or Plexiglas, not so well for H_2O as it exhibits a « cuvette » effect at larger Q.

Lindner P., Leclercq F., ILL

Simulated annealing

Fighting local minima:
Accept worse solutions to allow for a more extensive search for the optimal solution.
The name and inspiration come from **annealing in metallurgy**, a technique involving heating and controlled cooling of a material to improve its properties.

Initialization of the method:

- determine a parameter **neighbourhood** where the **parameters will be picked randomly**
- chose a start **« temperature » T**
- the system **energy E** will be the function to minimize: **χ^2** in our case
- define an **acceptance function P**
- if P **<** random number [0,1] go to the new parameter set

$$P = \exp\left(-\frac{\Delta E}{T}\right)$$

For a given T define a maximum number of trials, nTrials (it may be one)

- define how to decrease T and a maximum number of steps

Problem: costly in computing time

Important

During experiment:

- pay attention to the statistics, especially for the calibration file and the sample holder (cuvette) file .
- if this latter is very small versus the calibration forget it !

During preliminary data treatment (corrections)

- if some pixels exhibit small counting rate perform a smoothing of the data, in order to avoid Poisson distribution, which may spoil the fit.
- if you feel the calibration file statistics are poor, try the pseudo-calibration.

During the fit

- if the test function is a χ^2 its final value must be close to 1.

  This is an excellent test of the quality of the model

  and shows that a false minimum has been avoided,
- if it keeps far from 1, it means either
  - that the set of parameters is traped in a local minimum
  - or that there is a systematic error.

Especially when statistics are poor a fit often goes amok if there are many parameters. In a first phase keep some parameters static. Run the fit. Finally free all parameters and run the fit again.

Origin of systematic errors

- **bad centre definition.** When the scattering function changes very fast at small q it is very sensitive to the beam centre position (for instance power law).

A 2D fit including the centre coordinates usually solves this problem.

- **bad model.** Gaussian and Lorentzian are very similar at small scattering vector Q, but much different at larger Q. Impurities or surface scattering may generate power law scattering, which must be added to the main model function.
- **detector electronics.** May happen while uncommon. Erroneous pixels are found mostly on the detector border; it is easy to remove them during the corrections stage.

Advantage of 2D

- single I(0)
- use of all pixels of non isotropic data
- higher reliability for parameter comparison in various directions
- centre fit

Ethics

In your model use as little functions and parameters as possible.

It is always possible and meaningless to fit data with a lot of functions and parameters.

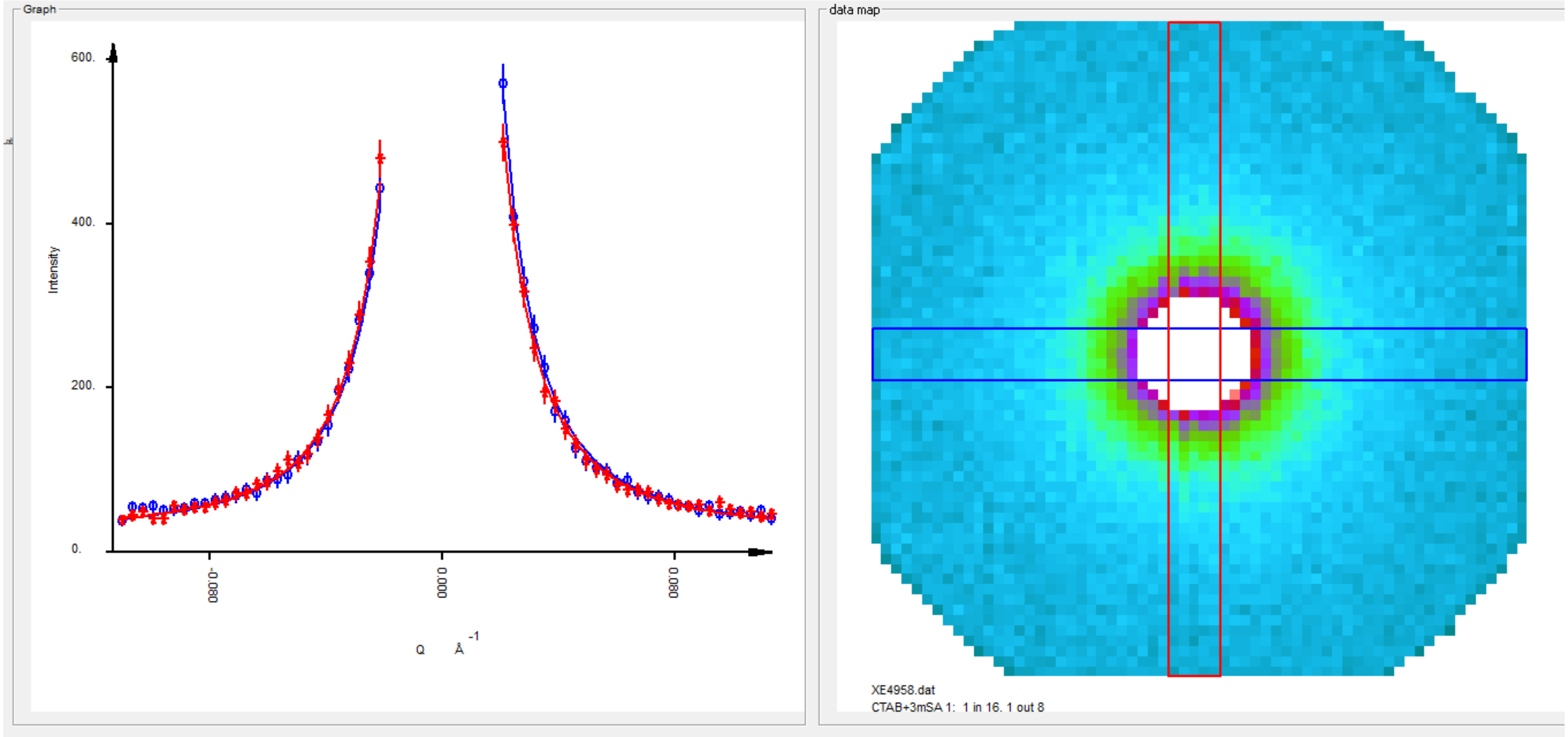
If you have 2 models compare them to the data. If both fit inside the data error bars, discard the most sophisticated.

If you are not happy with that make again your experiment with better statistics.

Examples

a) Is there any anisotropy in my data?
b) A problem of systematic error (Porod contamination) and statistics
c) Again a systematic error (centre position). Why it is interesting to substract the model to the data.
d) The anisotropic scattering by a liquid crystal polymer chain is spoiled by a smectic Bragg peak
e) 2 data files fitted with a single model
f) Représentation and model function in polar coordinates
g) « Embedded » data files, fitted together
h) SAXS by perfect nanochannels (track etched polymer membranes)
i) SANS by nanochannels (mica) , 2 data files together
g) Is the main chain of polymer nanotubes anisotropic ?
h) Again a systematic error (wrong model). Influence of the statistics of the calibration and sample holder file

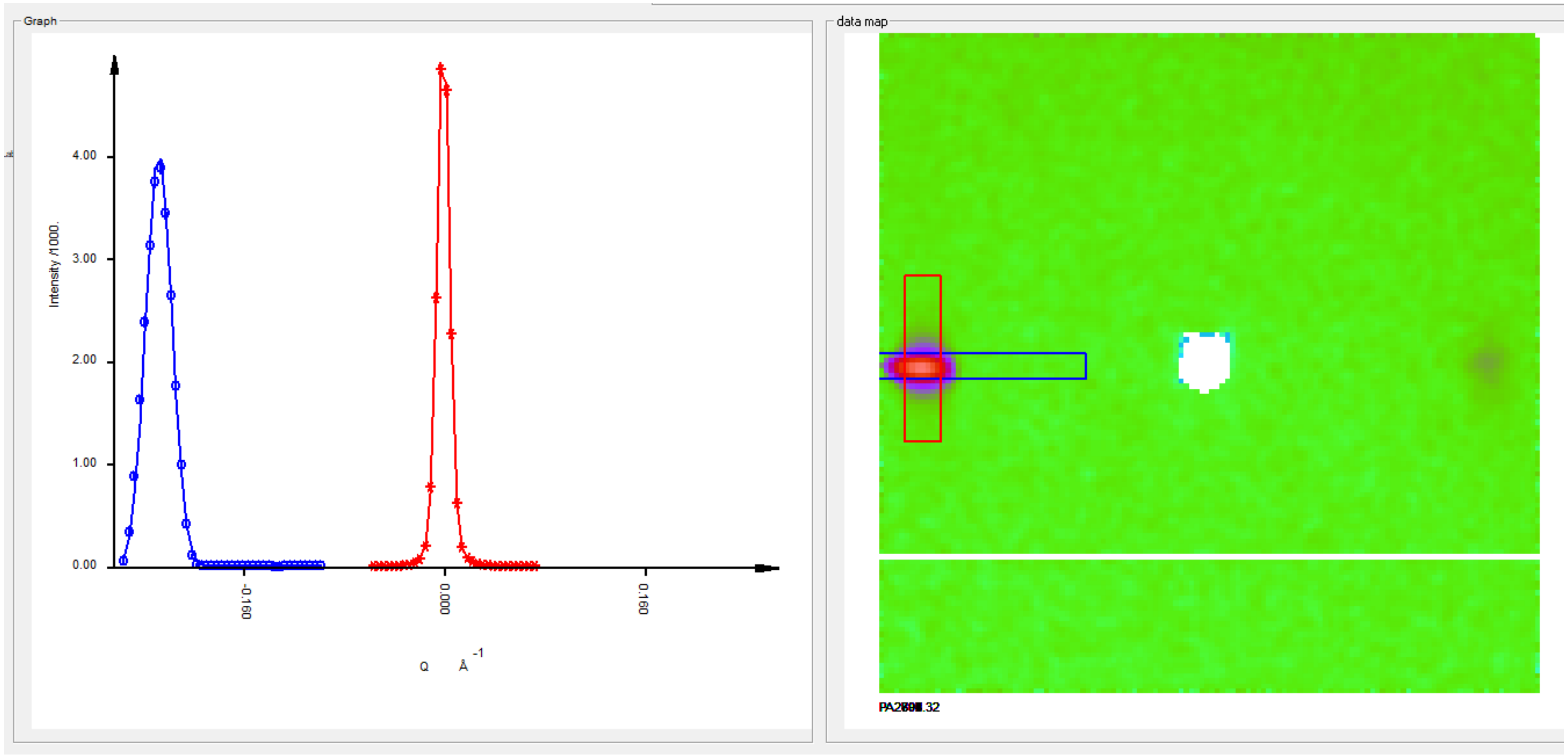
Is there any anisotropy ?



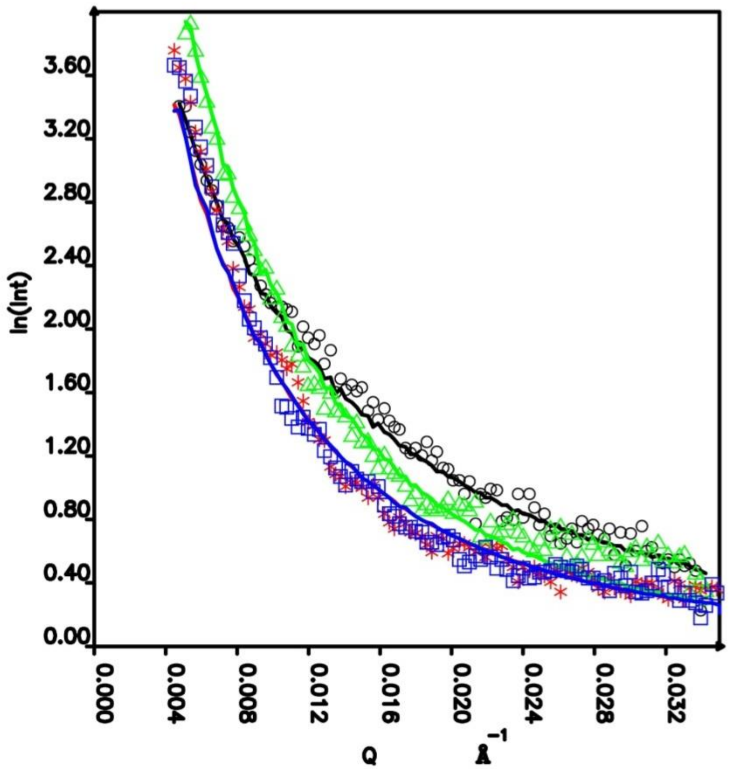
XE4958.dat
CTAB+3mSA 1: 1 in 16. 1 out 8

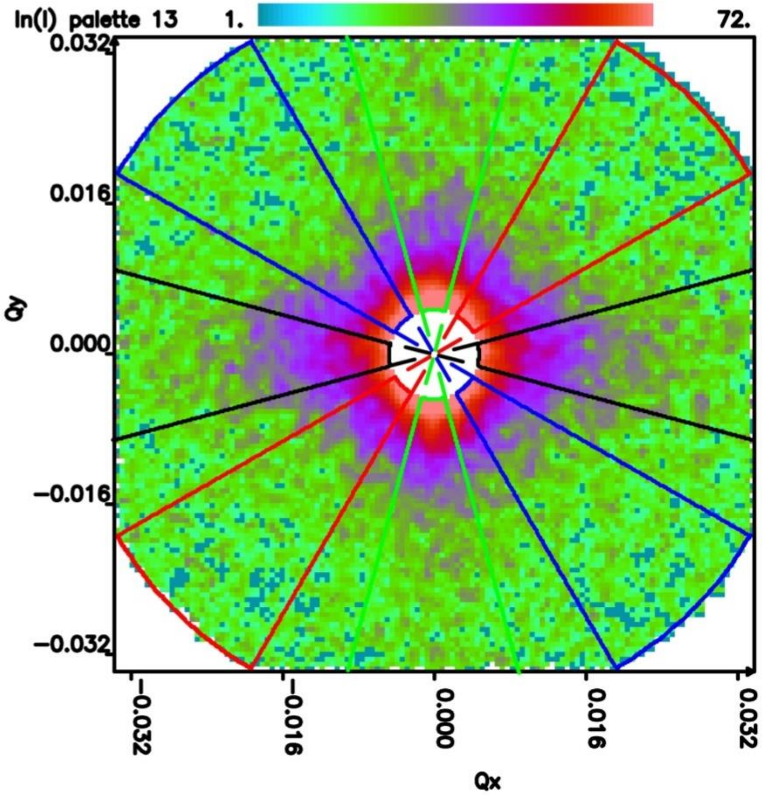
Pépy G. BNC

Systematic error (wrong model). Influence of correction file statistics.



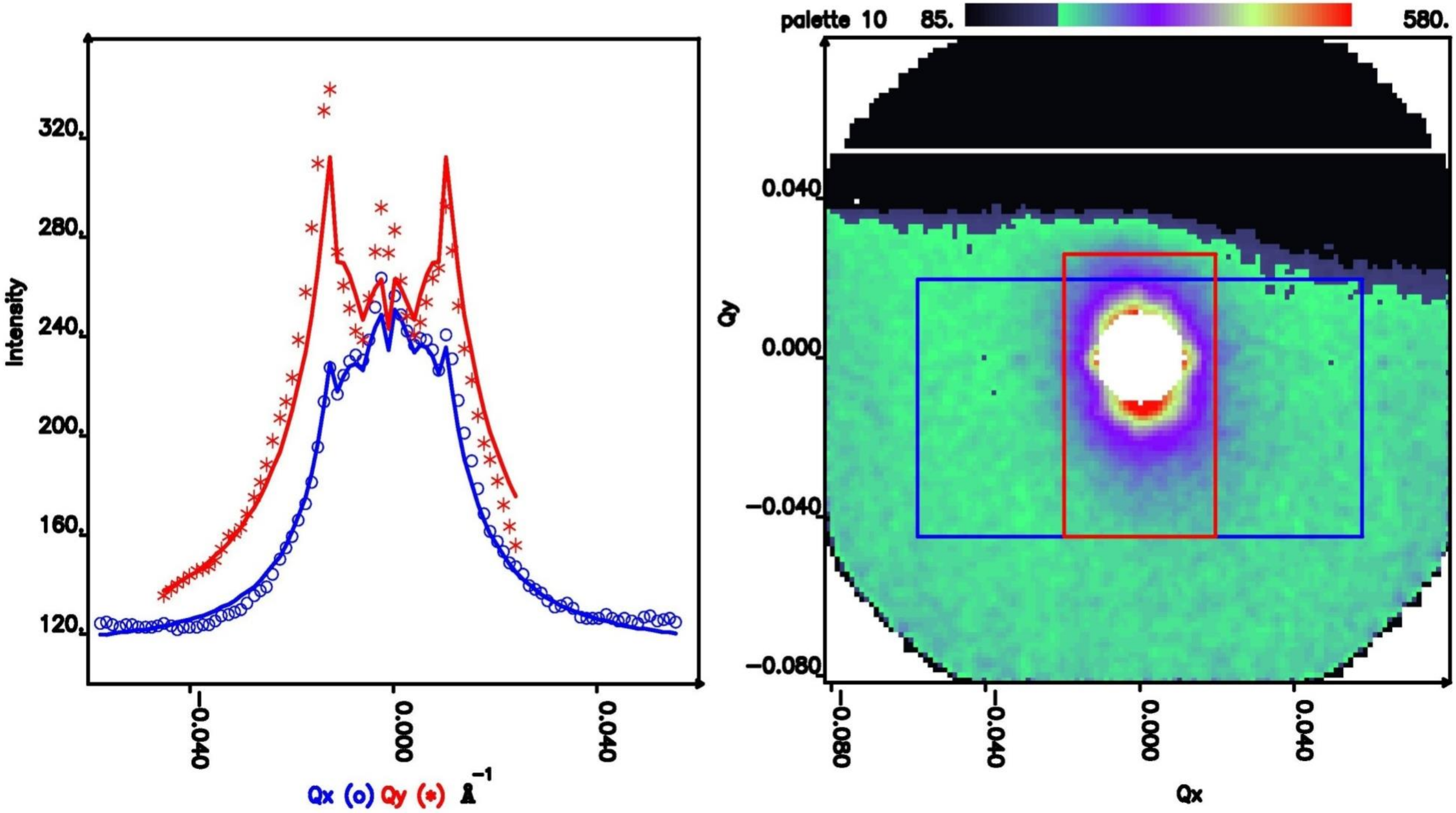
Poor statistics, Porod contamination





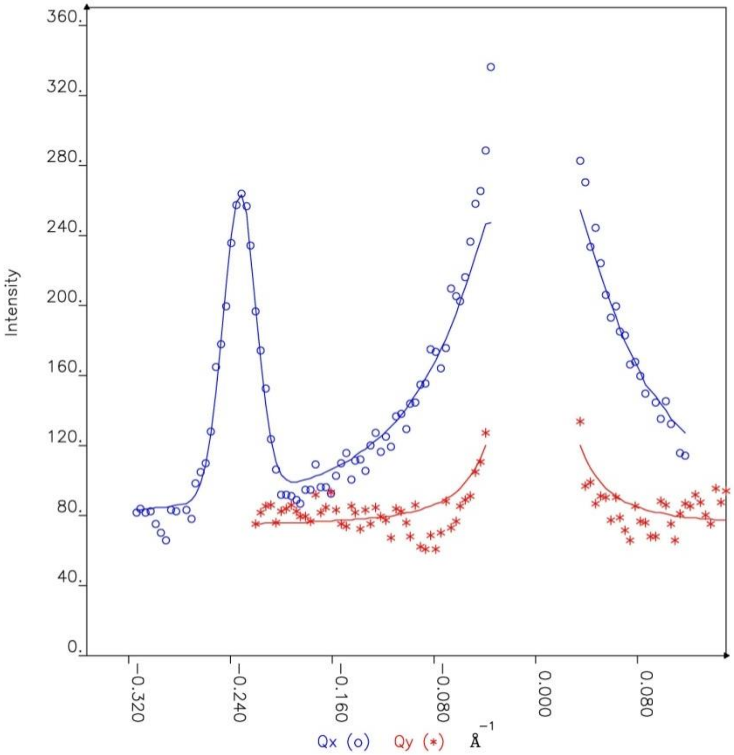
Mauzac M. CNRS

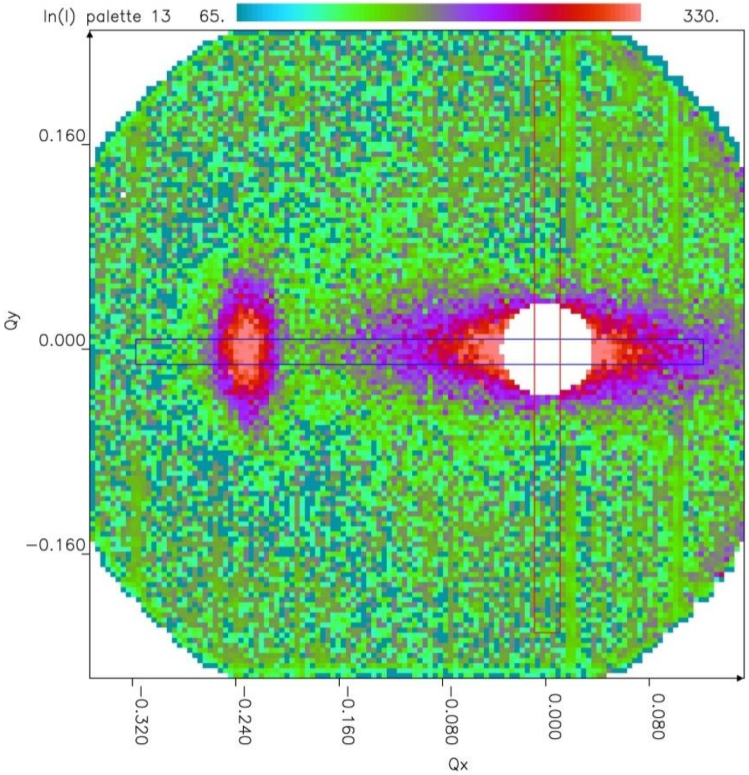
Find the origin of a systematic error…



Noirez L. LLB

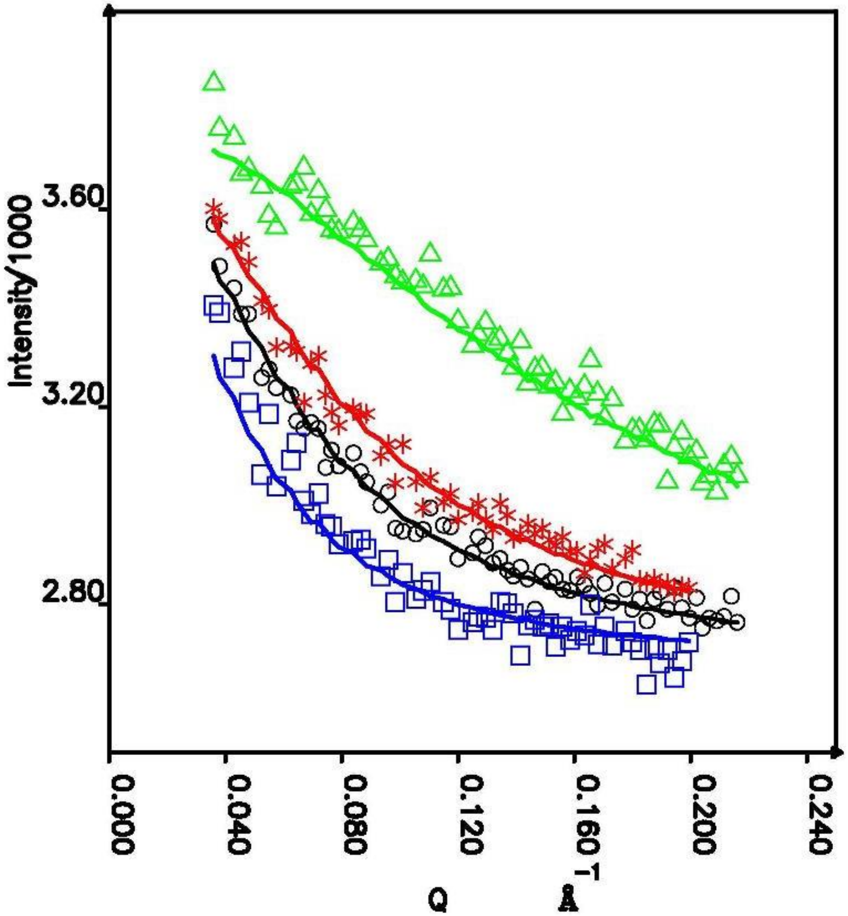
Nice polymer chain central scattering, with a smectic Bragg peak near-by

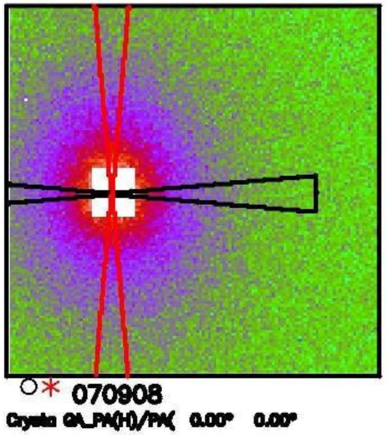


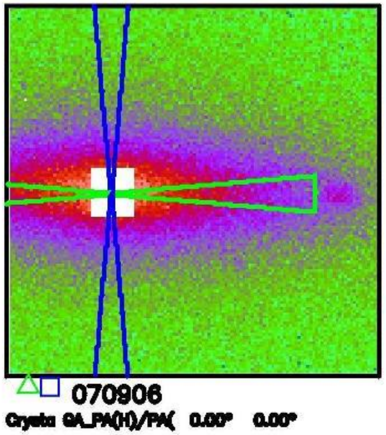


Cotton J.P. LLB

How the shape of a liquid crystal polymer changes with temperature

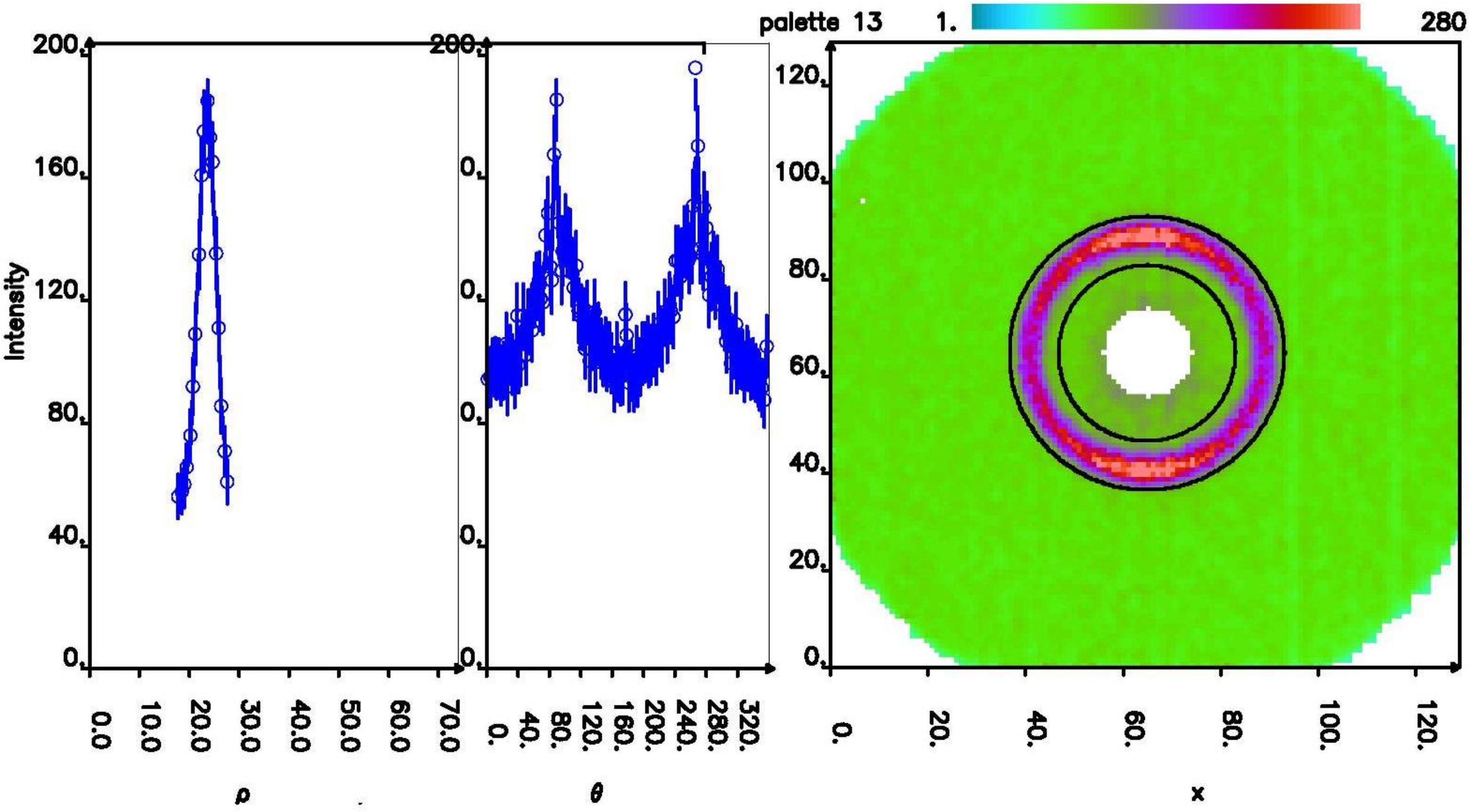






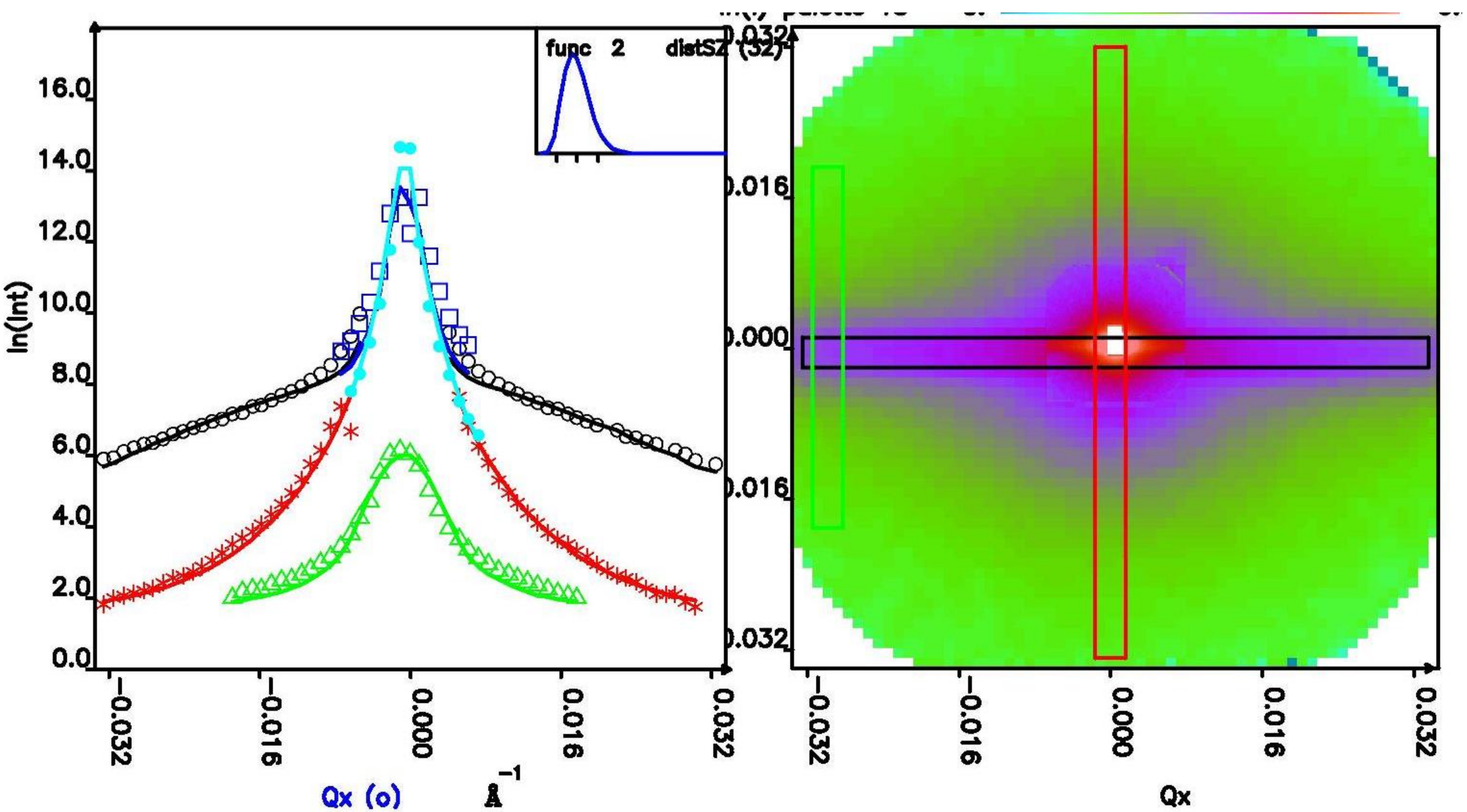
Noirez L. LLB

Polar representation may be helpful

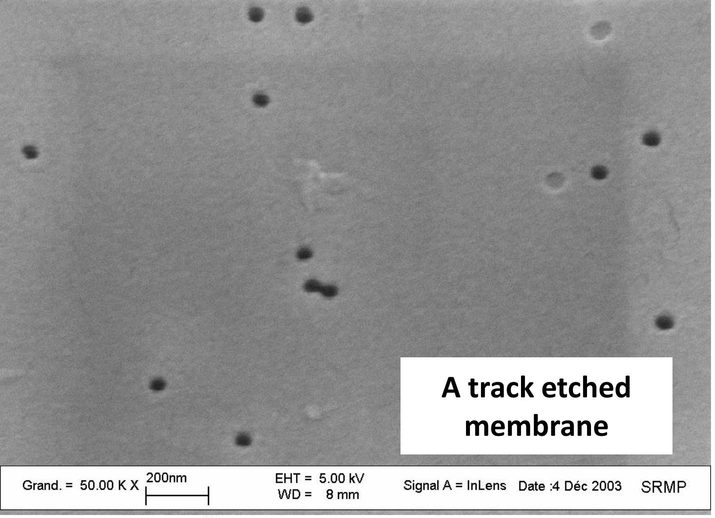


Kiselev, JINR

Bubbles in W wires. Large Q scale.



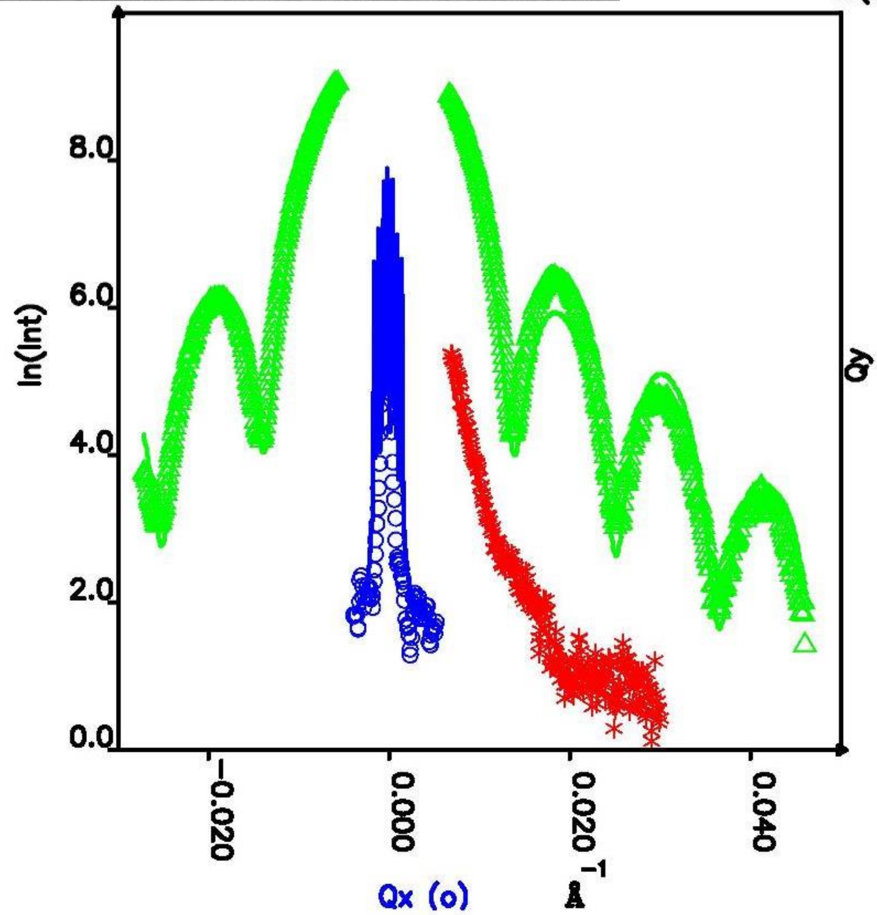
Len A. BNC

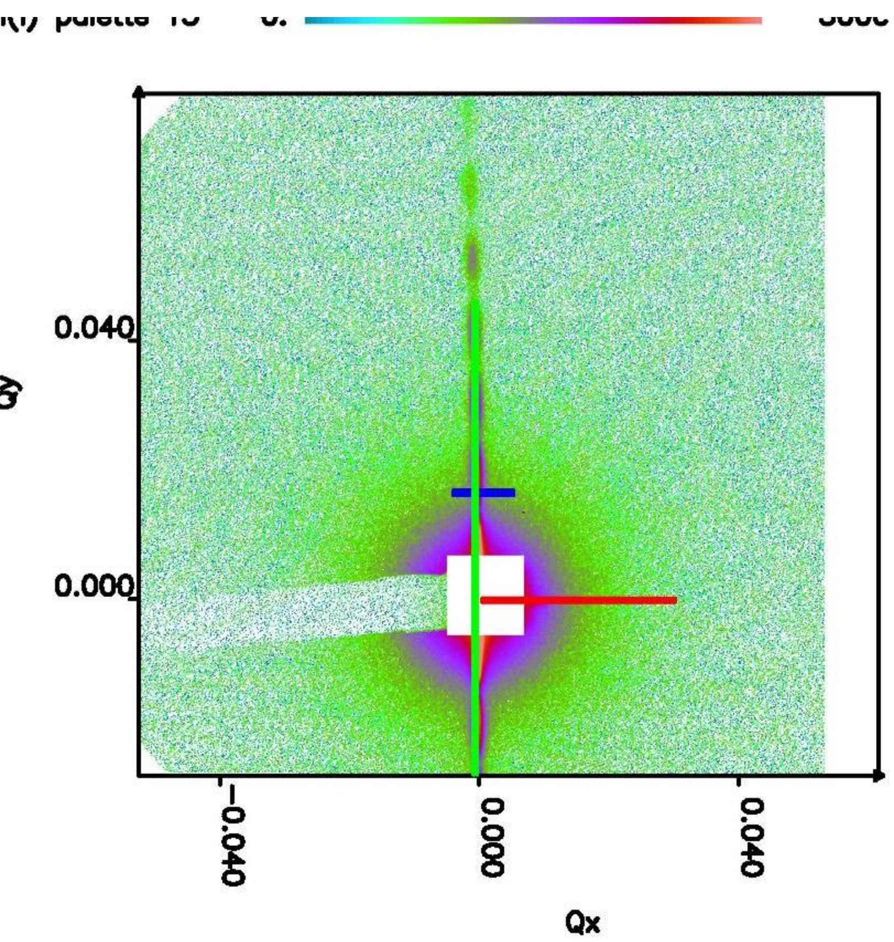


A track etched membrane

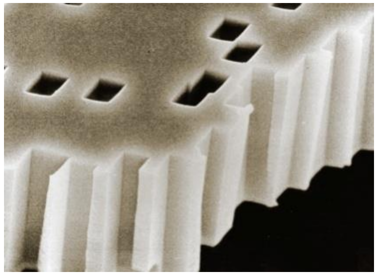
SAXS by nanochannels

Apel, Kuklin JINR





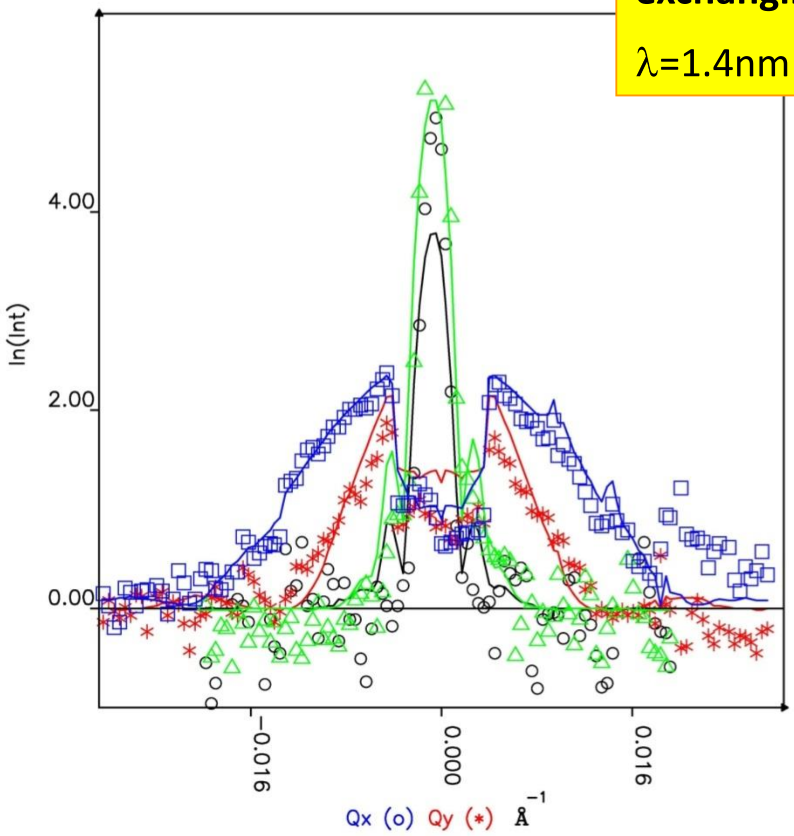
SANS by rhombohedric nanochannels in mica

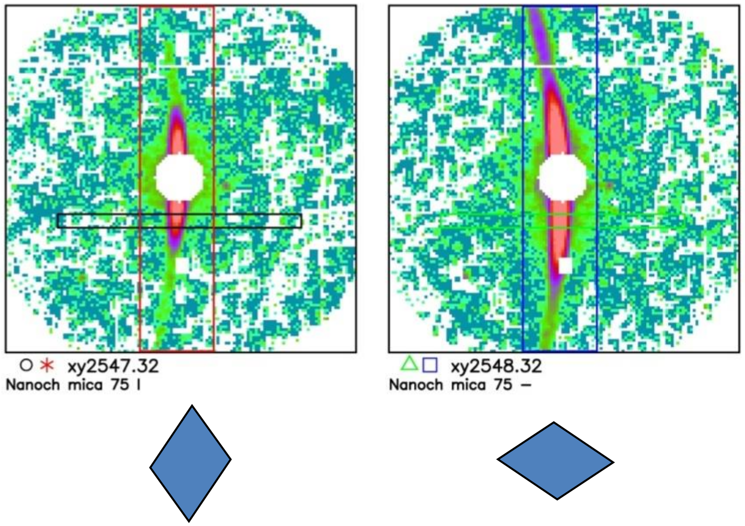


The data files were obtained at the LLB, Saclay, for two orientations of the channels at 90°.

Both files were fitted together, with the same function, exchanging Ra, Rb

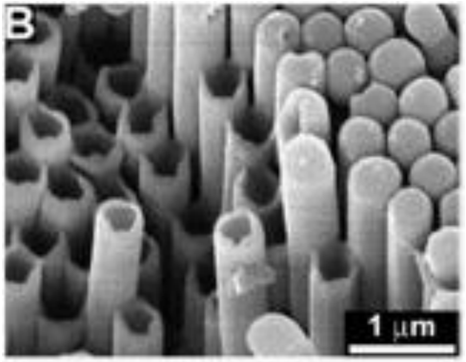
λ=1.4nm   D=14m   Ra=45nm   Rb=25nm





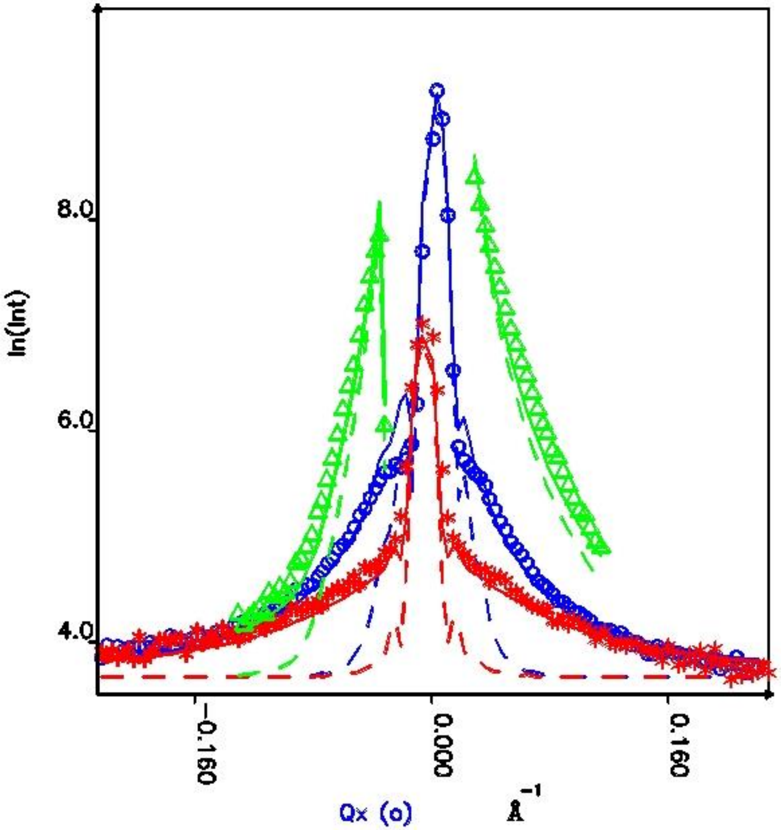
rhomboedric channel orientation

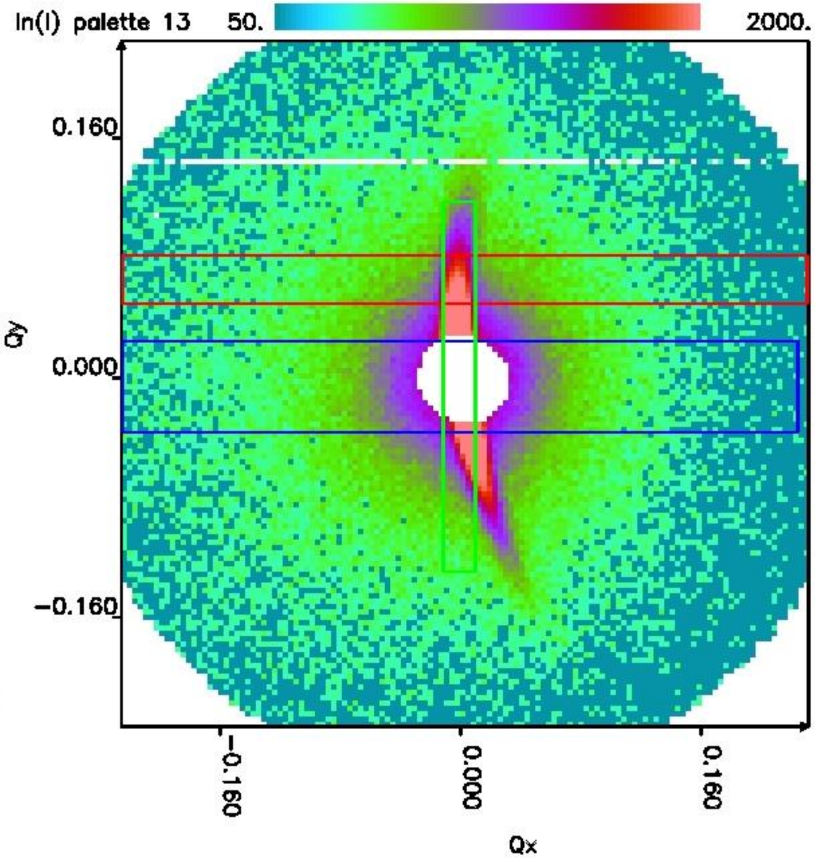
C. Trautmann, GSI, Germany



Is the main chain of polymer nanotubes anisotropic ?

Stillings, Germany





Thanks to contributors

Noirez L	LLB, Saclay
Martin N	LLB, Saclay
Rosta L.	Energia, Budapest
Len A.	Energia, Budapest
Baroni P.	LLB, Saclay
Papoular R.	LLB, Saclay
Islamov A	FLNP, JINR, Dubna
Kiselev M	FLNP, JINR, Dubna
Kuklin A	FLNP, JINR, Dubna
Apel P	FLNR, JINR, Dubna
Stillings C	Germany
Trautmann C	GSI Darmstadt

Dear students, it would be nice to tell me wether

- this lecture **was interesting**
- this lecture **was not what you expected**
- what was **most** interesting
- what I may **remove**

- **improvements ?**

gpepy@laposte.net

Thanks to the audience for the attention !