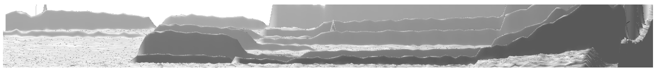


Profile refinement
Least-squares analysis and beyond

Bill David, ISIS Facility,
Rutherford Appleton Laboratory, UK

IUCr Computing School Siena 18-23 August 2005

Part I

When and why do we use least-squares analysis in crystallography?

Using least-squares analysis:
a basic part of crystallography

- It's worked all right up till now!
  - How accurate are your structural parameters?
- The fit looks pretty good!
  - How good are your data really?
  - How good, how complete is your model?
- If it ain't broke don't fix it!

Bert Lance 1977

> "Learn the fundamentals of the game and stick to them. Band-Aid remedies never last."
>
> Jack Nicklaus

- fundamental parameters
  - Pearson VIIs are good but fundamental parameters are better
  - "One of the intrinsic benefits of the fundamental parameters approach is that it is easily adapted to any laboratory diffractometer. Good fits can normally be obtained over the whole 2θ range without refinement using the known properties of the diffractometer (i.e. slit sizes, diffractometer radius and so on) and the emission profile."
  - you can understand and explain the peak shape function
- fundamental statistics
  - optimisation by plausible reasoning
  - probabilistic reasoning for least-squares analysis and beyond
  - minimise empiricism - no *deus ex machina*

What is least squares analysis?

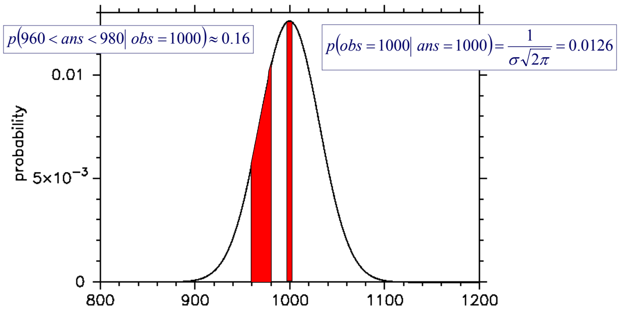
$$\chi^2 = \sum_i \left(\frac{\text{obs(i)} - \text{calc(i)}}{\text{esd(i)}}\right)^2$$

When do we use least squares analysis?

extremely specific extremely broad

Least squares analysis has its roots in the assumption <u>that the errors in the data follow a Gaussian (normal) probability distribution function.</u>

Least-squares analysis equates to the data following a Gaussian probability distribution



Finding the most probable solution = maximising the probability

likelihood

$$p(D_i|\,\mu,\sigma=\sqrt{\mu})=\prod_{i=1}^{N}\frac{1}{\sigma_i\sqrt{2\pi}}\exp\left(-\frac{1}{2\sigma^2}(D_i-\mu)^2\right)$$

Maximising the likelihood = minimising -(log-likelihood)

Minimise

$$-\ln\left(p(D_i|\,\mu,\sigma=\sqrt{\mu})\right)=\sum_{i=1}^{N}\left(\frac{1}{2}\ln(2\pi)+\ln\sigma_i+\frac{1}{2\sigma_i^2}(D_i-\mu)^2\right)$$
$$\cong C+\sum_{i=1}^{N}\left(\frac{1}{2\sigma_i^2}(D_i-\mu)^2\right)$$

c.f. least squares minimisation

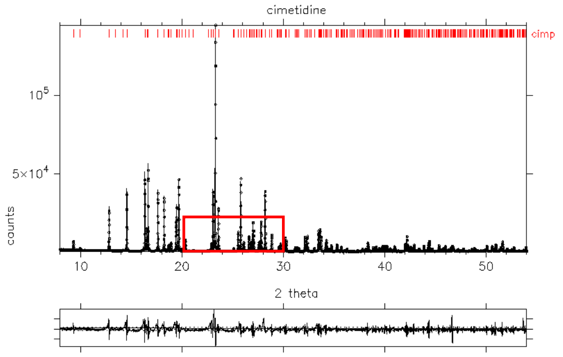
Minimise

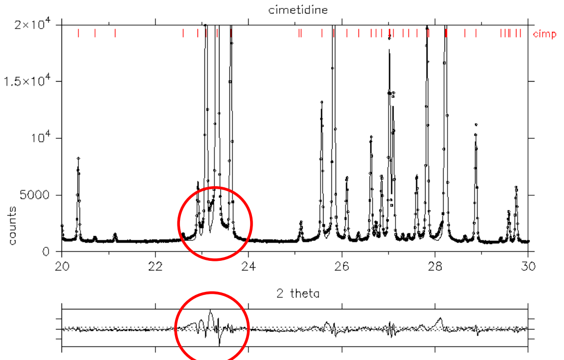
$$\chi^2=\sum_{i=1}^{N}\left(\frac{1}{\sigma_i^2}(D_i-\mu)^2\right)$$

Part II

What if the fit isn't that good?

What if the fit isn't that good?

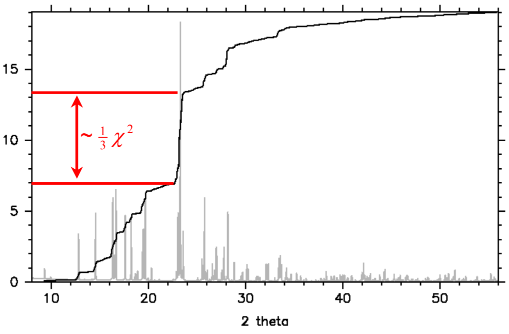




(obs-calc)/esd

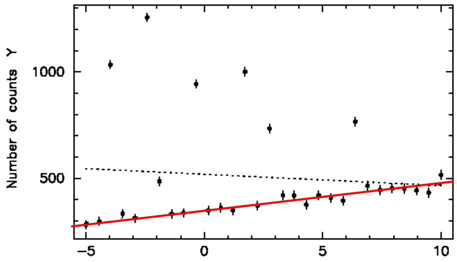
$$\chi_j^2=\sum_{i=1}^{j\le N_{obs}}\frac{1}{\sigma_i^2}(obs(i)-calc(i))^2$$

$$\chi_j^2=\sum_{i=1}^{j\le N_{obs}}\frac{1}{\sigma_i^2}(obs(i)-calc(i))^2$$



What if the fit isn't that good

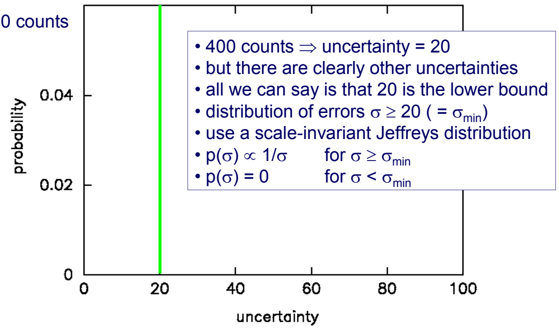
- Is it wise to have 150,000 counts in the biggest peak and 5000 counts in a very highly structured background?
  - No! Redo the experiment!
- Collect all Bragg peaks with similar fractional accuracy
  - variable counting time to give E/σ(E) constant
- If accuracy and precision are required be prepared to
  - comprehensively model structure and microstructure
  - perform fundamental line-shape analysis
  - undertake detailed "fundamental" background analysis
- If all else fails - use statistics / plausible reasoning!



What's gone wrong?

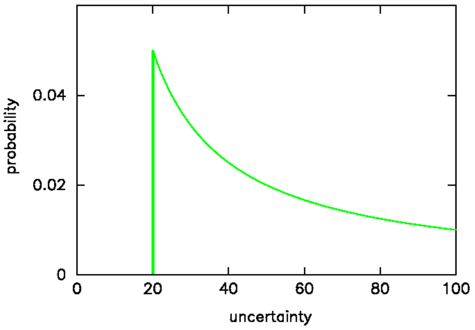
- We've performed a least-squares analysis and implicitly assumed that all errors follow a Gaussian PDF
- We've been <u>certain</u> about our <u>uncertainties</u>!

400 counts

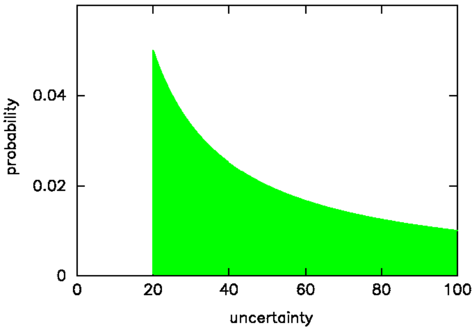


- 400 counts ⇒ uncertainty = 20
- but there are clearly other uncertainties
- all we can say is that 20 is the lower bound
- distribution of errors σ ≥ 20 (= σ_{min})
- use a scale-invariant Jeffreys distribution
- p(σ) ∝ 1/σ for $\sigma \ge \sigma_{min}$
- p(σ) = 0 for $\sigma < \sigma_{min}$

- Jeffreys prior
- p(σ) ∝ 1/σ for $\sigma \ge \sigma_{min}$
- p(σ) = 0 for $\sigma < \sigma_{min}$

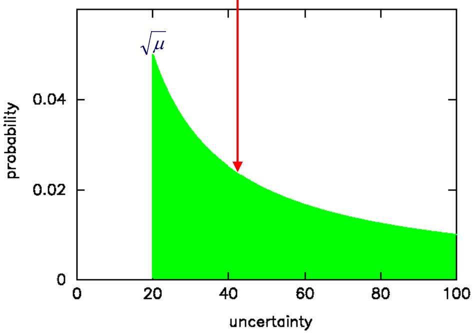


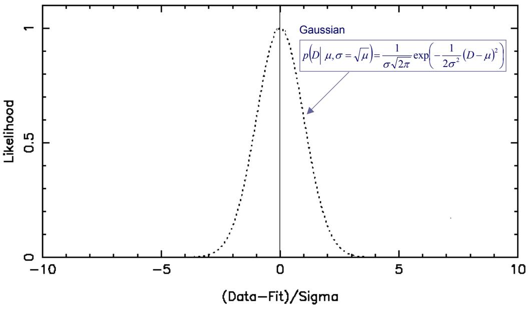
- Jeffreys prior
- p(σ) ∝ 1/σ for $\sigma \ge \sigma_{min}$
- p(σ) = 0 for $\sigma < \sigma_{min}$



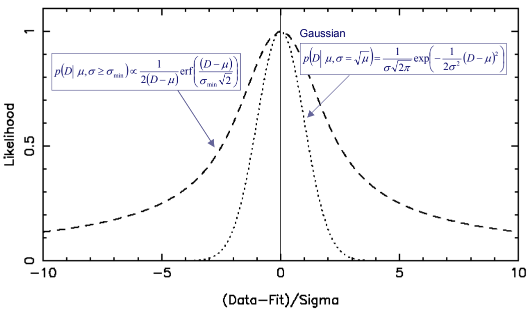
Gaussian

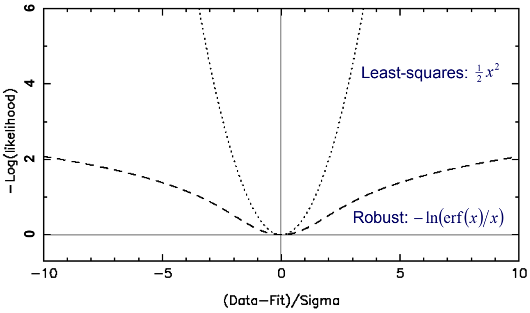
$$p(D|\mu,\sigma\ge\sqrt{\mu}) = \int_{\sigma_{\min}=\sqrt{\mu}}^{\infty} \boxed{prob(\sigma)}\, \boxed{\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{1}{2\sigma^2}(D-\mu)^2\right)}\, d\sigma$$

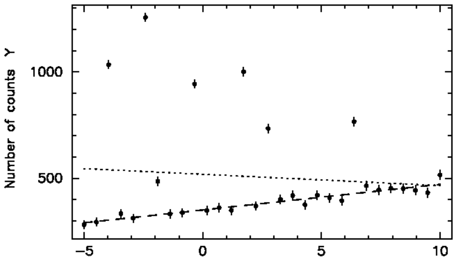


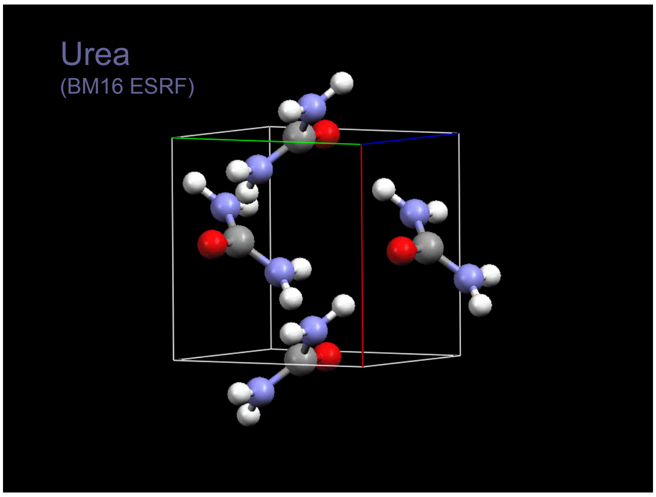


$$p(D|\mu,\sigma=\sqrt{\mu}) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{1}{2\sigma^2}(D-\mu)^2\right)$$





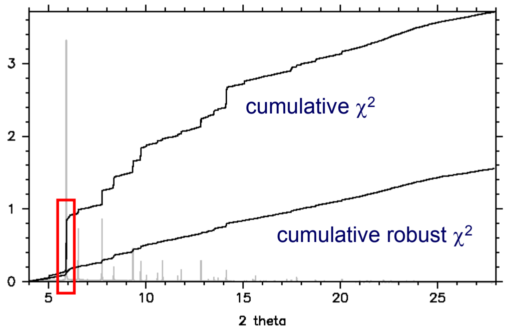




Urea

(BM16 ESRF)

Urea (BM16 ESRF)



Urea (BM16 ESRF)

	SXXD	Least Squares	LS-SXXD	Robust	R-SXXD
C1 z	0.3328(3)	0.3236(9)	-0.0092(10)	0.3319(13)	-0.0009(14)
O1 z	0.5976(4)	0.6013(5)	0.0037(6)	0.5984(7)	0.0008(8)
N1 x	0.1418(2)	0.1405(3)	-0.0013(4)	0.1423(7)	0.0005(7)
z	0.1830(2)	0.1807(5)	-0.0023(6)	0.1813(7)	-0.0017(7)
C1 U11	0.0353(6)	0.0348(20)	-0.0005(20)	0.0329(40)	0.0024(40)
U33	0.0155(5)	0.0396(30)	0.0241(30)	0.0413(40)	0.0258(40)
U12	0.0006(9)	0.0205(30)	0.0199(30)	0.0128(40)	0.0122(40)
O1 U11	0.0506(9)	0.0749(16)	0.0243(18)	0.0617(30)	0.0111(30)
U33	0.0160(6)	0.0080(14)	-0.0080(15)	0.0090(20)	-0.0070(20)
U12	0.0038(18)	0.0052(20)	0.0014(30)	-0.0011(35)	-0.0049(35)
N1 U11	0.0692(6)	0.0627(15)	-0.0065(18)	0.0697(25)	0.0005(25)
U33	0.0251(4)	0.0460(22)	0.0211(22)	0.0365(30)	0.0114(30)
U12	-0.0353(7)	-0.0252(18)	0.0101(20)	-0.0361(30)	-0.0008(30)
U13	-0.0003(3)	-0.0015(11)	-0.0012(12)	-0.0029(15)	-0.0026(15)

🟨 = diff > 4σ

9/14 > 4σ

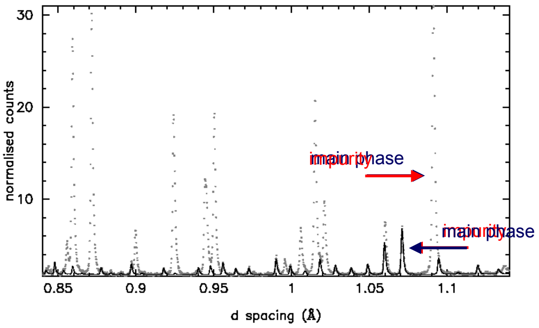
1/14 > 4σ

Part III

What if the model is incomplete?

- unknown impurity phase
- unknown fragment in crystal structure
  - e.g. disordered guest in zeolite
    unknown waters of hydration
    disordered oxygen in high T_c material
    incomplete models in structure solution
    F_{calc} has uncertainty as well as F_{obs}
    (c.f. maximum likelihood in structural biology)

What if there's an impurity phase?



Gaussian probability distribution

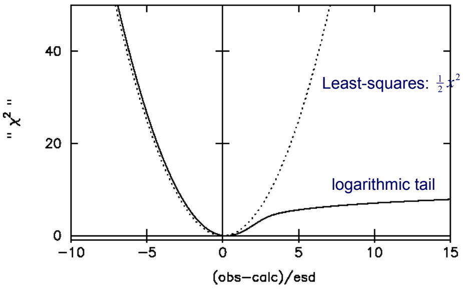
We're happy that the data errors follow a Gaussian distribution

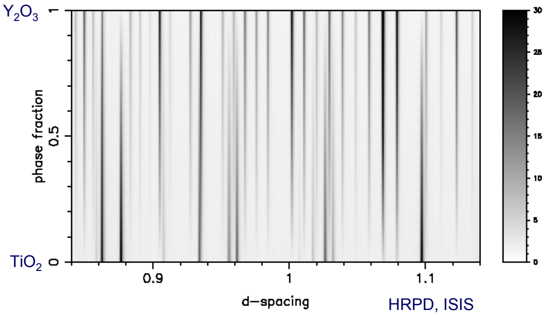
$$p(D|M,\sigma)=\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{1}{2\sigma^2}(D-M)^2\right)$$

Our problem is that the model is incomplete. We have a known phase with contribution, *P*, and an unknown (positive) component, *A*. i.e. *M=P+A*

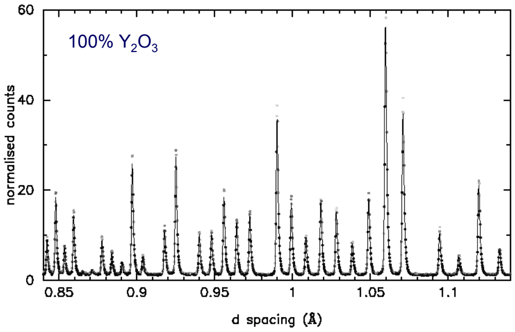
$$p(D_i|P,\sigma)=\int prob(A)\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{1}{2\sigma^2}(D-(P+A))^2\right)$$

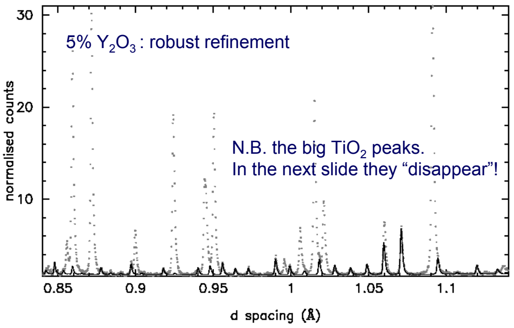
Use a scale invariant Jeffreys *1/A* prior for *prob(A)*





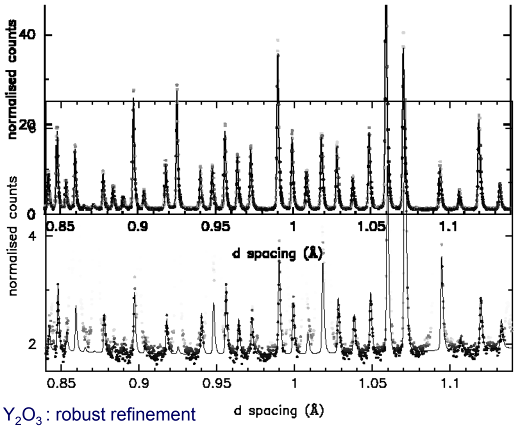
HRPD, ISIS





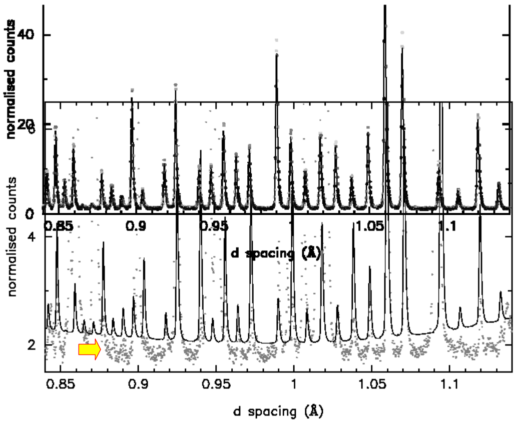
5% Y_2O_3 : robust refinement

N.B. the big TiO_2 peaks.
In the next slide they “disappear”!

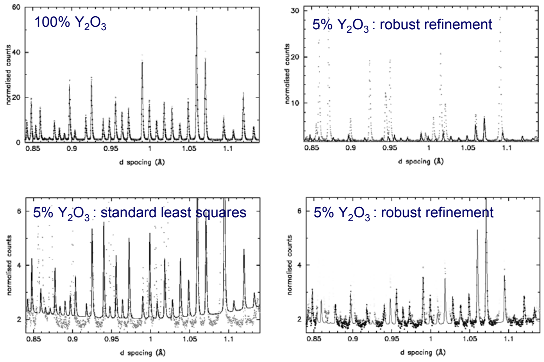


5% Y_2O_3 : robust refinement

The darkness of the dots indicates their relative impact in the robust analysis.
Large positive (obs-calc) points (i.e. mostly impurity peaks) are “invisible”.



5% Y_2O_3 : standard least-squares

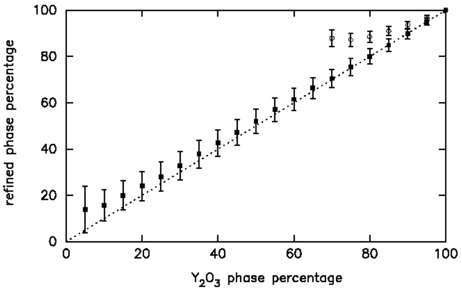


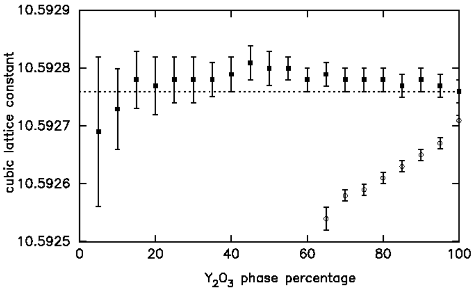
100% Y_2O_3

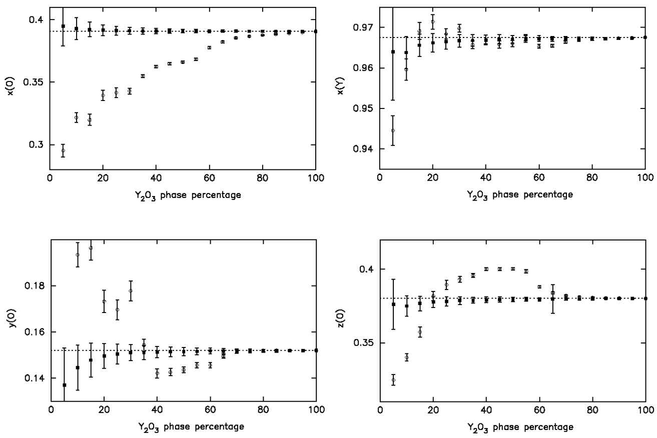
5% Y_2O_3 : robust refinement

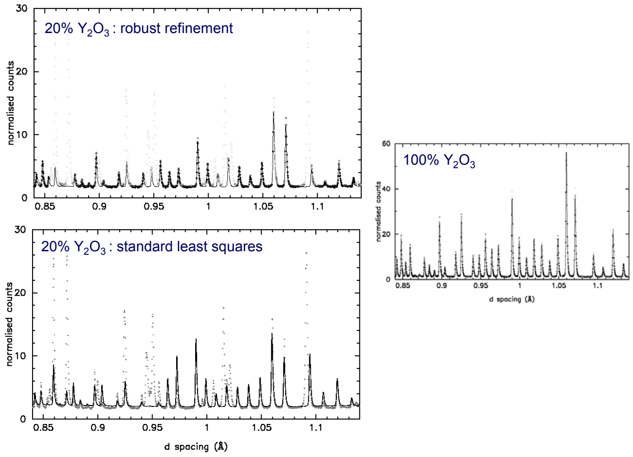
5% Y_2O_3 : standard least squares

5% Y_2O_3 : robust refinement









Summary of Part III

If the model is incomplete, careful construction of an appropriate probability distribution function can bring significant improvements over standard least-squares analysis.

J. Appl. Cryst. (2001). **34**, 691-698

Robust Rietveld refinement in the presence of impurity phases

W. I. F. David

Abstract: A modified least-squares analysis is presented that allows reliable structural parameters to be extracted from a powder diffraction pattern even in the presence of a substantial unmodelled impurity contribution. The algorithm is developed within the context of Bayesian probability theory. Experimental points that fall above those calculated, and are thus more probably from impurity peaks, are systematically down-weighted. This approach is illustrated with a two-phase example.

Acta Cryst. (2002). A**58**, 316-326

A maximum-likelihood method for global-optimization-based structure determination from powder diffraction data

A. J. Markvardsen, W. I. F. David and K. Shankland

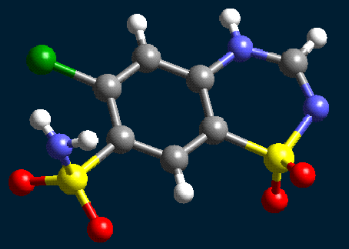
Abstract: A maximum-likelihood algorithm has been incorporated into a crystal structure determination from a powder diffraction data framework that uses an integrated-intensity-based global optimization technique. The algorithm is appropriate when the structural model being optimized is not a complete description of the crystal structure under study.

Conclusions

- least-squares is fairly ubiquitous but not always the most appropriate minimisation metric.
- try to keep things simple
  - do the best experiment possible
  - develop as complete a model as possible
- be as certain as possible about your uncertainties (probability distribution functions)
- be prepared to go beyond least-squares!

ISIS

2. Fourier recycling (MaxEnt)



D S Sivia and W I F David, *J. Phys. Chem. Solids* **62**, 2119-2127 (2001)

ISIS

Single crystal : minimise the integrated intensities χ_I^2

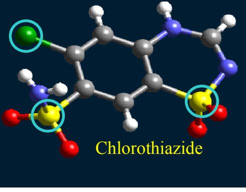
$$\chi_I^2 = \sum_{\mathbf{h}} \frac{1}{\sigma_h^2}\left(\left|F_{obs}(\mathbf{h})\right|^2 - \left|F_{calc}(\mathbf{h})\right|^2\right)^2$$

Powders : minimise the correlated integrated intensities χ_{CI}^2

$$\chi_{CI}^2 = \sum_{\mathbf{h}}\sum_{\mathbf{k}} w_{hk}\left(\left|F_{obs}(\mathbf{h})\right|^2 - \underline{\left|F_{calc}(\mathbf{h})\right|^2}\right)\left(\left|F_{obs}(\mathbf{k})\right|^2 - \left|F_{calc}(\mathbf{k})\right|^2\right)$$

Minimise the correlated integrated intensities χ_{CI}^2

$$\chi_{CI}^2 = \sum_{\mathbf{h}}\sum_{\mathbf{k}} w_{hk}\left(\left|F_{obs}(\mathbf{h})\right|^2 - \underline{\left|F_{model}(\mathbf{h}) + \Delta F(\mathbf{h})\right|^2}\right)\left(\left|F_{obs}(\mathbf{k})\right|^2 - \left|F_{model}(\mathbf{k}) + \Delta F(\mathbf{k})\right|^2\right)$$

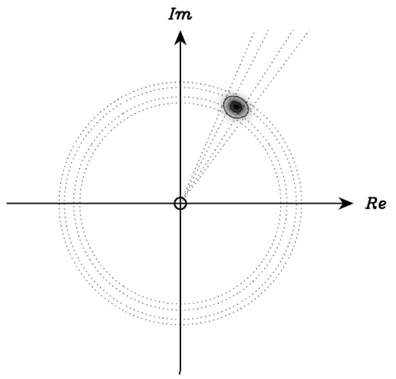


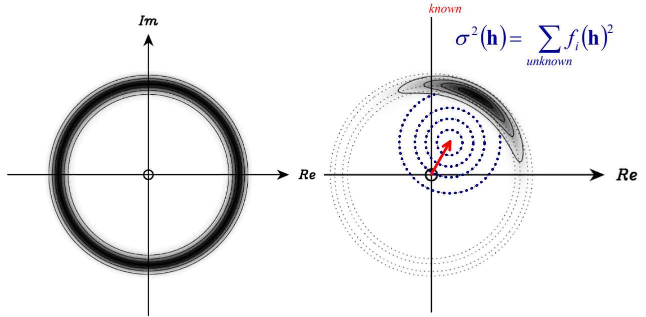
where $\Delta F(\mathbf{h}) = \sum_{\mathbf{r}} \Delta\rho(\mathbf{r})\exp(-2\pi i\mathbf{h}\cdot\mathbf{r})$

while maximising $\sum_{\mathbf{r}} \Delta\rho(\mathbf{r})\ln(\Delta\rho(\mathbf{r}))$

Chlorothiazide

The importance of phase information



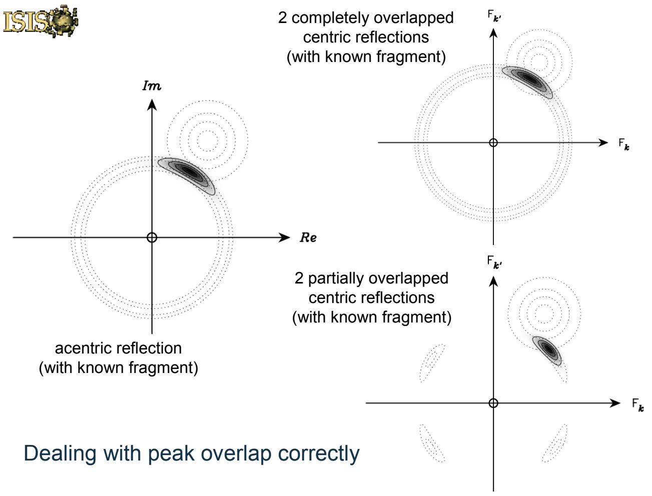


$$F_m(\mathbf{h}) = \sum_{known} f_i(\mathbf{h})\exp(2\pi i\mathbf{h}.\mathbf{r}_i)$$

$$\sigma^2(\mathbf{h}) = \sum_{unknown} f_i(\mathbf{h})^2$$

acentric reflection
(only magnitude known)

acentric reflection
(with known fragment)



2 completely overlapped
centric reflections
(with known fragment)

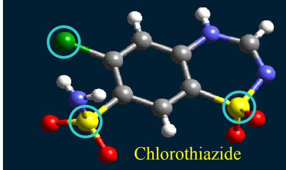
acentric reflection
(with known fragment)

2 partially overlapped
centric reflections
(with known fragment)

Dealing with peak overlap correctly

Minimise the correlated integrated intensities χ^2_{CI}

$$\chi^2_{CI} = \sum_{\mathbf{h}}\sum_{\mathbf{k}} w_{hk}\left(|F_{obs}(\mathbf{h})|^2 - |F_{model}(\mathbf{h}) + \Delta F(\mathbf{h})|^2\right)\left(|F_{obs}(\mathbf{k})|^2 - |F_{model}(\mathbf{k}) + \Delta F(\mathbf{k})|^2\right)$$



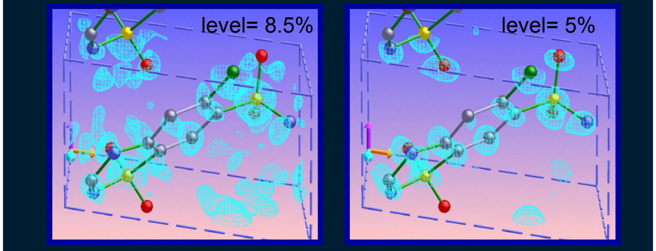
where $\Delta F(\mathbf{h}) = \sum_{\mathbf{r}} \Delta\rho(\mathbf{r})\exp(-2\pi i\mathbf{h}\cdot\mathbf{r})$

while maximising $\sum_{\mathbf{r}} \Delta\rho(\mathbf{r})\ln(\Delta\rho(\mathbf{r}))$

Chlorothiazide

D S Sivia and W I F David, *J. Phys. Chem. Solids* **62**, 2119-2127 (2001)

Bayesian map reconstruction from powder diffraction data

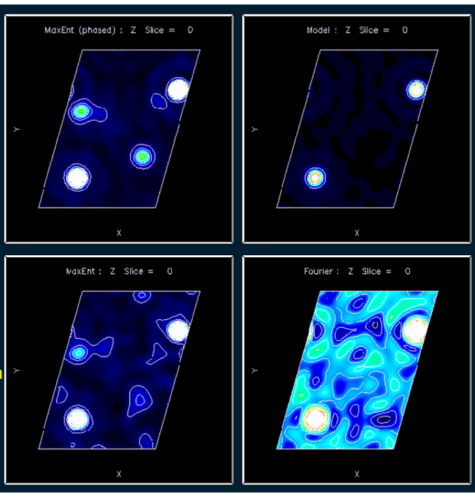


Ordinary Map

Bayesian Map

Chlorothiazide: $C_7H_5N_3O_4S_2Cl$

Bayesian approach: S_2Cl (32% scattering) to full structure

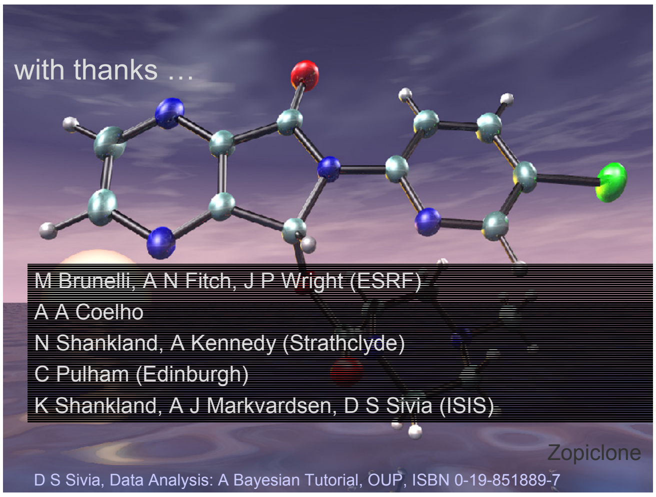


full structure

S_2Cl

Bayesian Fourier

Fourier



with thanks …

M Brunelli, A N Fitch, J P Wright (ESRF)
A A Coelho
N Shankland, A Kennedy (Strathclyde)
C Pulham (Edinburgh)
K Shankland, A J Markvardsen, D S Sivia (ISIS)

Zopiclone