Statistical Treatment of Uncertainties

Dale E. Tronrud
Howard Hughes Medical Institute
Institute of Molecular Biology
University of Oregon, USA
http://www.uoxray.uoregon.edu/dale/

Introduction

- There has been much discussion of uncertainty in this workshop. All this talk simply means that the topic is critical to everything that we do.
- Unfortunately many people, and most of the users of our software, would prefer that everything is clear cut and certain.
- The world is filled with uncertainty and true understanding requires us to know the limits of our knowledge.

- My hope here is to clarify some of the terms and issues in this area. Terms tend to be used without clear definition and even the experts confuse them quite often.

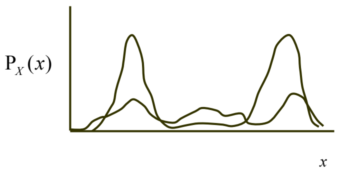
Major Topics in Uncertainty

- There are three (ok, maybe four) major topics in this area

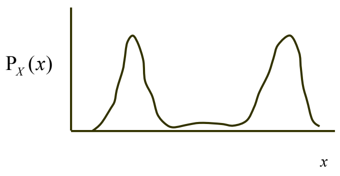
- Probability Distribution Functions -- PDF's
- Baysian Parameter Estimation
- Maximum Likelihood -- ML
- Least-Squares

Probability Distribution Functions (PDFs)

- Every quantity has an uncertainty. We like to say the uncertainty is plus or minus some amount but this is not sufficiently descriptive.
- With a PDF a probability is assigned to every conceivable value the quantity could "really" be.



Probability of the quantity *X* as a function of *x*



- People would rather deal with something simpler than this entire distribution.
- The basic characteristics are
  - The most probable value
  - The expectation value (also known as the "mean", "best", or "centroid")
  - The standard deviation (also known as "sigma")

- There are others that are less used
  - Skew, Kurtosis, and Entropy

- For the Normal distribution the most probable value is the mean and sigma is what we are used to.

Calculations in the presence of uncertainty

- If you have a quantity *X* what is the uncertainty of 2*X*?
- Actually you have to go to the PDF to find out. The rule is

$$P_{2X}(x) = P_X(x/2)$$

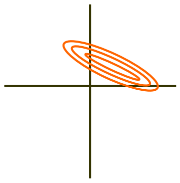
- Once you have created the PDF for 2*X* you can derive things like the "best" value.
- For linear functions the most probable value and "best" value transform with the variable. If you multiply *X* by 2 then the most probable and "best" values will also be multiplied by 2.
- This is not true for nonlinear functions.

$$P_{X^2}(x \mid x \geq 0) = P_X(\sqrt{x}) + P_X(-\sqrt{x})$$
$$P_{X^2}(x \mid x \text{ p } 0) = 0$$

- When doing math, the key is to work on the PDF and then recalculate the characterizing values and not to try to calculate the mean of the transformed variable by transforming the mean.
- If you have a single measurement, you must come up with some idea of its uncertainty. If you have no idea how confident you can be in a value, it is useless.
  - To generate a PDF for a single measurement you must understand how that value was acquired.
  - For most measurements the uncertainty is a Normal distribution and the sigma would be determined from your knowledge of the instrument.
- If you have multiple measurements, transform them all and then calculate the mean.
- A major limitation in practice is that often we are unsure of the PDF.
  - This leads to the realization that each point in the PDF must be represented by a PDF itself. This makes my head hurt.

Multiple Variables

- In any project there are many variables. While each has a PDF to represent it uncertainty, the uncertainties are often not independent of each other.
- For a reflection we usually consider the uncertainty in amplitude and phase to be independent, but that is not always the case. This leads to a two dimensional PDF



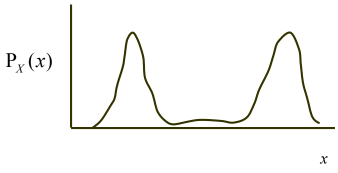
- Then again, the uncertainties of the amplitude and phase of a reflection are tied to its Fridel mate. This requires a four dimensional PDF.
- Then again, there is noncrystallographic symmetry. This requires an eight dimensional (or probably more) PDF.
- Let's face it, all reflections are tied to all others via crystal's contents. That requires a PDF that has a lot of dimensions.
- All interesting PDF's are multidimensional and usually they have an enormous number of dimensions.
- Finding some way to represent these PDF's is a problem yet to be solved.

Getting a Handle on the Matter

- If you have a big, multidimensional PDF it might be true that there is only one peak and that peak looks like a Normal distribution.
- In that special case, the most probable value is the "best" value and you can gage you uncertainty in that value.
  - For a 3 dimensional PDF this is not done with 3 sigmas.
  - You need a 3x3 covariance matrix.

$$\begin{pmatrix} \sigma_X^2 & \sigma_X\sigma_Y r_{XY} & \sigma_X\sigma_Z r_{XZ} \\ \sigma_X\sigma_Y r_{XY} & \sigma_Y^2 & \sigma_Y\sigma_Z r_{XZ} \\ \sigma_X\sigma_Z r_{XZ} & \sigma_Y\sigma_Z r_{YZ} & \sigma_Z^2 \end{pmatrix}$$

  - Macromolecular crystallographers keep asking for a sigma for each parameter, but that would be uninformative without the correlation coefficients.



- The danger is that there are more peaks hiding in your PDF.
- In that case you could find both and calculate a covariance matrix for each as though the other doesn't exist.

Baysian Parameter Estimation

- Kevin has talked quite a bit in the last session about this topic and I don't see a need to repeat what he said.

$$P(h \mid d) = \frac{P(d \mid h)\,P(h \mid knowledge)}{P(d \mid knowledge)}$$

- Each of these probabilities is a multidimensional PDF.
  - In refinement they have an incomprehensible number of dimensions.
- Even if you could determine the probabilities for all possible hypothesis's you would have to search though them all to find the most probable, or integrate over them all for the "best" hypothesis.
- This is equivalent to the Global Minimum problem that has plagued us in refinement.
- No one has come up with a solution for this, for macromolecular refinement.
- The solution is to find a guess for the most probable hypothesis some other way, assume there can be only one, and then search for the most probable nearby hypothesis in the neighborhood.

Maximum Likelihood

- Finding the most probable hypothesis in the Likelihood distribution, assuming there is only one peak, is the Maximum Likelihood method.
  - It finds the most probable, not necessarily the “best” hypothesis.
  - There may be whole worlds of stuff happening elsewhere in the likelihood distribution, but that is ignored.
- Baysian Parameter Estimation is a rigorous and robust procedure, but is very hard for our problems.
- Maximum Likelihood is a massive simplification that allows us to bring in many of the ideas of BPE but reduces the problem to something very similar to what our old programs did.

A Comparison of the Methods used in Macromolecular Refinement.

- There are three types of target functions that have been used in macromolecular refinement.
  - Energy Minimization
  - Least Squares
  - Maximum Likelihood

Energy Minimization

- The best model with the one with the lowest energy.
- How does one calculate the “energy” of a model?
- How does the diffraction data become “energy”?
- How does one reconcile the instantaneous nature of energy with the time averaged nature of the diffraction data?
- Why bother when a statistically based method has answers to all these questions?

The Least-squares Function

$Q_o(i)$	Observed quantity i
$\sigma_o^2(i)$	Observed variance of quantity i
$\mathbf{p}$	Parameters of a model
$Q_c(i,\mathbf{p})$	Corresponding quantity inferred from the current model

- The best $\mathbf{p}$ is that which minimizes

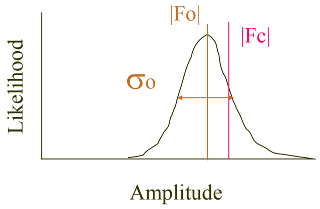
$$f = \sum_i \frac{(Q_o(i) - Q_c(i,\mathbf{p}))^2}{\sigma_i^2(i)}$$

Major Limitation of this Equation

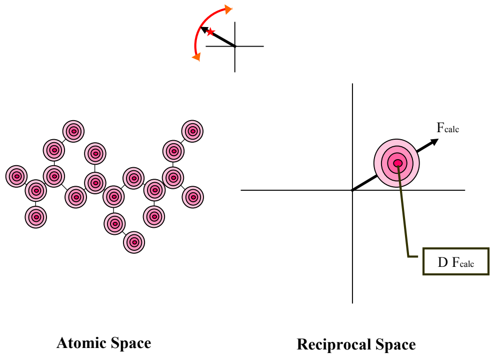
- The equation assumes that the observations are statistically independent. Often this is not the case.
  - Some programs use non-independent stereochemical restraint categories.
  - Many particular stereochemical targets are correlated.
  - The presence of noncrystallographic symmetry creates dependencies between some (many) reflections.
- This limitation also exists in all Maximum Likelihood implementations to date.

Maximum Likelihood

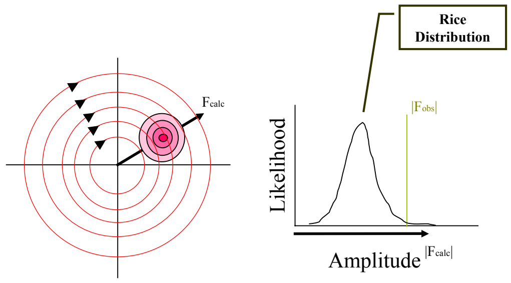
- Least Squares assumes that $\frac{|F_o| - |F_c|}{\sigma_o}$ for all reflections obeys a Normal distribution with a mean of zero and a standard deviation of one.
- Least Squares view of the world:



Is Structure Factor Calculation Hard?



Getting to Amplitudes



Difficulties in Maximum Likelihood

- What is the character of the uncorrected errors in the model?
  - Existing programs assume the errors behave like randomly, and with Normally distributed displacements of atomic positions and B factors.
  - Buster offers the option of non-uniform distribution of errors
    - It has a two state error model, where one part is treated in the usual way, but another part is identified only by a region of space and an elemental composition.
- How does one estimate the quantity of error in the model?
  - All ML programs use the agreement of the model to the test set to calibrate the error level.