Bayesian analysis in nuclear physics

Ken Hanson

T-16, Nuclear Physics; Theoretical Division
Los Alamos National Laboratory

Tutorials presented at LANSCE
Los Alamos Neutron Scattering Center
July 25 – August 1, 2005

Los Alamos
NATIONAL LABORATORY

This presentation available at
http://www.lanl.gov/home/kmh/

LA-UR-05-5680

Goals of tutorials

My aim is to

- present overview of Bayesian and probabilistic modeling
- cover basic Bayesian methodology relevant to nuclear physics, especially cross section evaluation
- point way to how to do it
- convince you that
  - Bayesian analysis is a reasonable approach to coping with measurement uncertainty

- Many thanks to my T-16 colleagues
  - Gerry Hale, Toshihiko Kawano, Patrick Talou

Outline – four tutorials

1. **Bayesian approach**

   probability – quantifies our degree of uncertainty

   Bayes law and prior probabilities

2. **Bayesian modeling**

   Peelle's pertinent puzzle

   Monte Carlo techniques; quasi-Monte Carlo

   Bayesian update of cross sections using Jezebel criticality expt.

3. **Bayesian data analysis**

   linear fits to data with Bayesian interpretation

   uncertainty in experimental measurements; systematic errors

   treatment of outliers, discrepant data

4. **Bayesian calculations**

   Markov chain Monte Carlo technique

   analysis of Rossi traces; alpha curve

   background estimation in spectral data

Slides and bibliography

- These slides can be obtained by going to my public web page:
  http://public.lanl.gov/kmh/talks/
  - link to **tutorial slides**
  - short **bibliography** relevant to topics covered in tutorial
  - other presentations, which contain more detail about material presented here
- Noteworthy books:
  - D. Sivia, *Data Analysis: A Bayesian Tutorial* (1996); lucid pedagogical development of the Bayesian approach with an experimental physics slant
  - D. L. Smith, *Probability, Statistics, and Data Uncertainties in Nuclear Science and Technology* (1991); lots of good advice relevant to cross-section evaluation
  - G. D'Agostini, *Bayesian Reasoning in Data Analysis: A Critical Review*, (World Scientific, New Jersey, 2003); Bayesian philosophy
  - A. Gelman et al., *Bayesian Data Analysis* (1995); statisticians' view
  - W. R. Gilks et al., *Markov Chain Monte Carlo in Practice* (1996); basic MCMC text

Tutorial 1
Bayesian approach

Uncertainty quantification

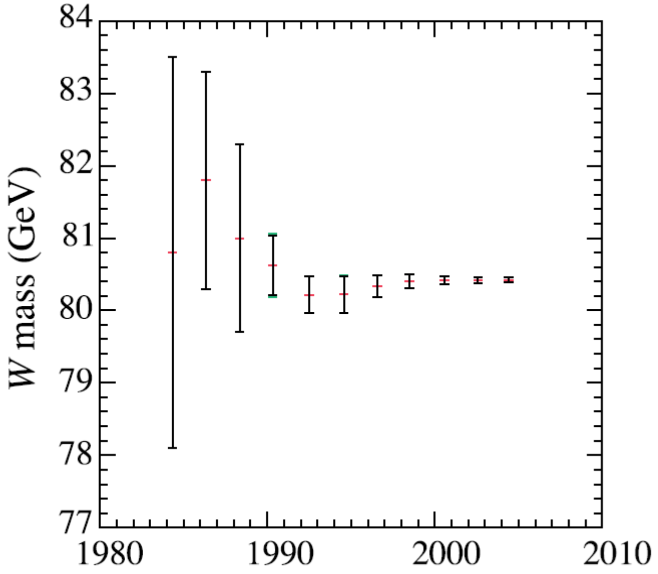
We need to know uncertainty in data:

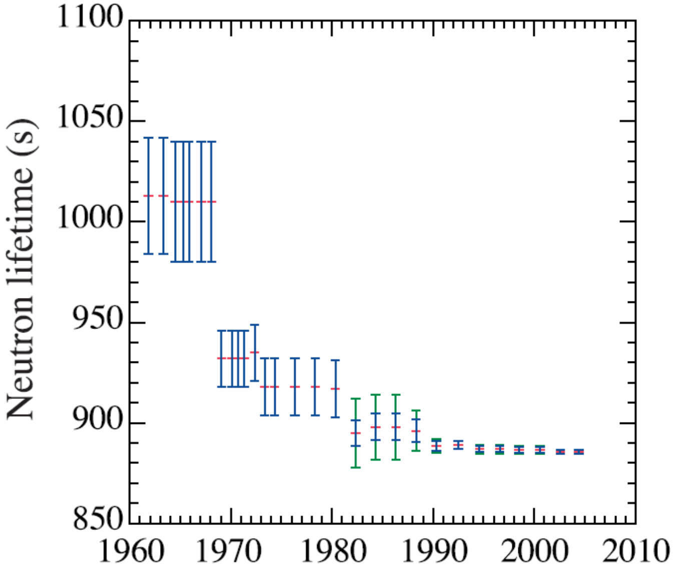
- To determine agreement among data, or between data and theory
- Inference about validity of models requires knowing degree of uncertainty
- We typically assume uncertainty described by a Gaussian pdf
  - often a good approximation
  - width of Gaussian characterized by its standard deviation σ
  - σ provides the metric for uncertainty about data
  - when combining measurements, weight by inverse variance σ^{-2}

- Nomenclature – uncertainty or error?
  - error – state of believing what is incorrect; wrong belief; mistake
  - uncertainty – lack of certainty, sureness; vagueness
  - **uncertainty analysis** seems to convey appropriate meaning

History of particle-properties measurements

- Plots show histories of two "constants" of fundamental particles
- Mass of W boson
  - logically ordered history
  - all within error bar wrt last (best?) measurement
- Neutron lifetime
  - disturbing history
  - periodic jumps with periods of extreme agreement
  - most earlier measurements disagree with latest ones
  - plot demonstrates possible sociological and psychological aspects of experimental physics

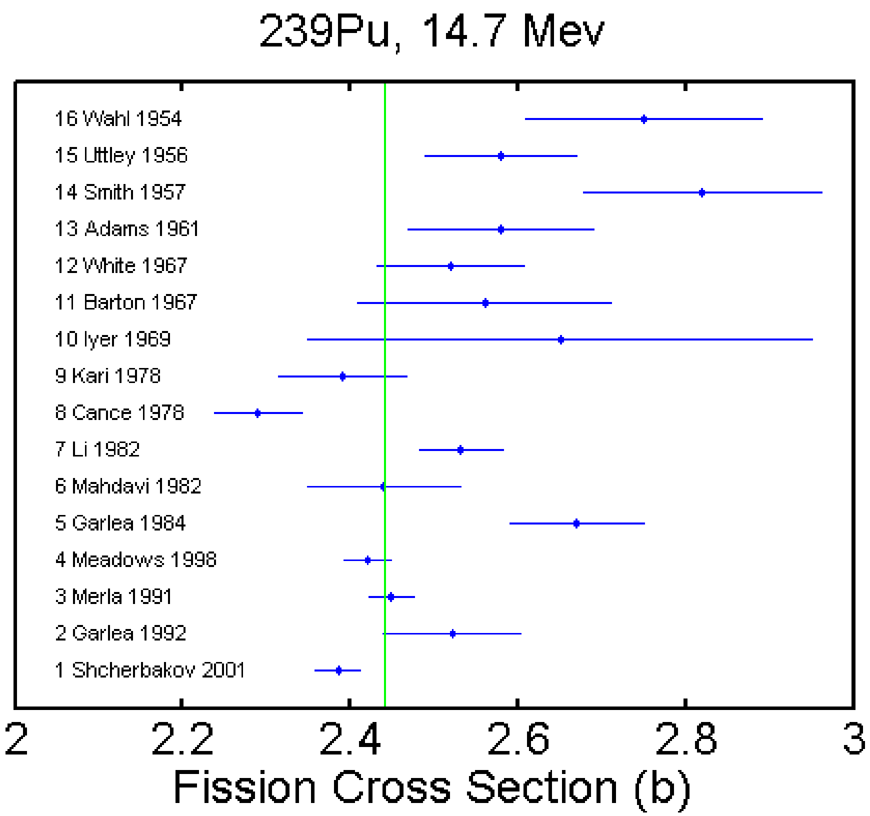




plots from Particle Data Group 2004

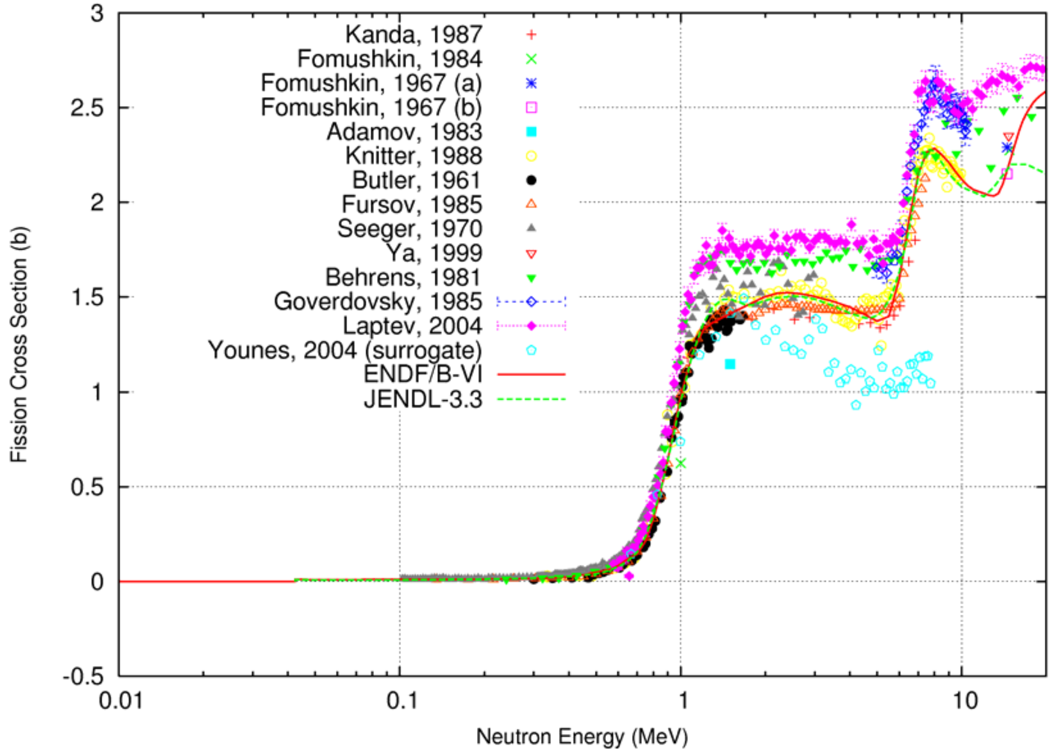
Neutron fission cross section data for ^{239}Pu

- Graph shows 16 measurements of fission cross-section for ^{239}Pu at 14.7 MeV
- Data exhibit fair amount of scatter
- Quoted error bars get smaller with time
- Minimum $\chi^2 = 44.6$, $p = 10^{-4}$ indicates a problem
  - dispersion of data larger than quoted error bars
  - outliers?; three data contribute 24 to χ^2, more than half



Neutron fission cross-section data

^{243}Am fission cross section



plot from P. Talou

- Neutron cross sections measured by many experimenters
  - sometimes data sets differ significantly
  - often little information about uncertainties, esp. systematic errors
  - many directly measure ratios of cross sections, e.g., ^{243}Am/ ^{235}U
  - a thorough analysis must go back to original data and consider all discrepancies

Bayesian analysis of experimental data

- Bayesian approach
  - focus is as much on uncertainties in parameters as on their best (estimated) value
  - provides means for coping with Uncertainty Quantification (UQ)
  - quantitative support of scientific method
  - use of prior knowledge, e.g., previous experiments, modeling expertise, subjective
  - experiments should provide decisive information
  - model-based analysis
  - model checking –
    does model agree with experimental evidence?
- Goal is to estimate **model parameters and their uncertainties**

Bayesian approach to model-based analysis

- Models
  - used to describe and analyze physical world
  - parameters inferred from data
- Bayesian analysis
  - uncertainties in parameters described by probability density functions (pdf)
  - prior knowledge about situation may be incorporated
  - quantitatively and logically consistent methodology for making inferences about models
  - open-ended approach
    - can incorporate new data
    - can extend models and choose between alternatives

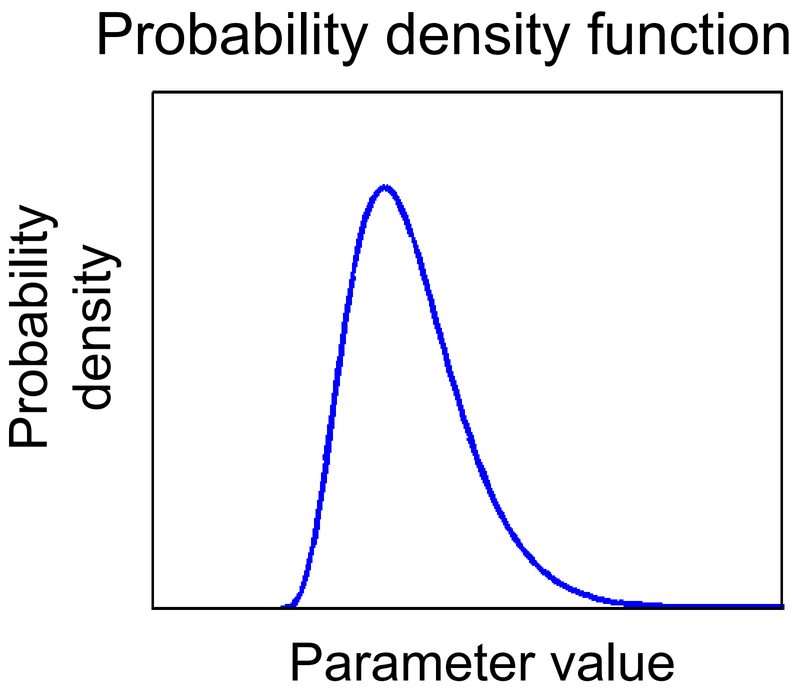
Bayesian approach to model-based analysis

- Bayesian formalism provides framework for modeling
  - choice of model is up to analyst (as in any analysis)
  - many ways to do it
  - calling an analysis Bayesian does not distinguish it

- Because it is a Bayesian analysis does not necessarily mean it is a good analysis; it can also be bad or inappropriate

Uncertainties and probabilities

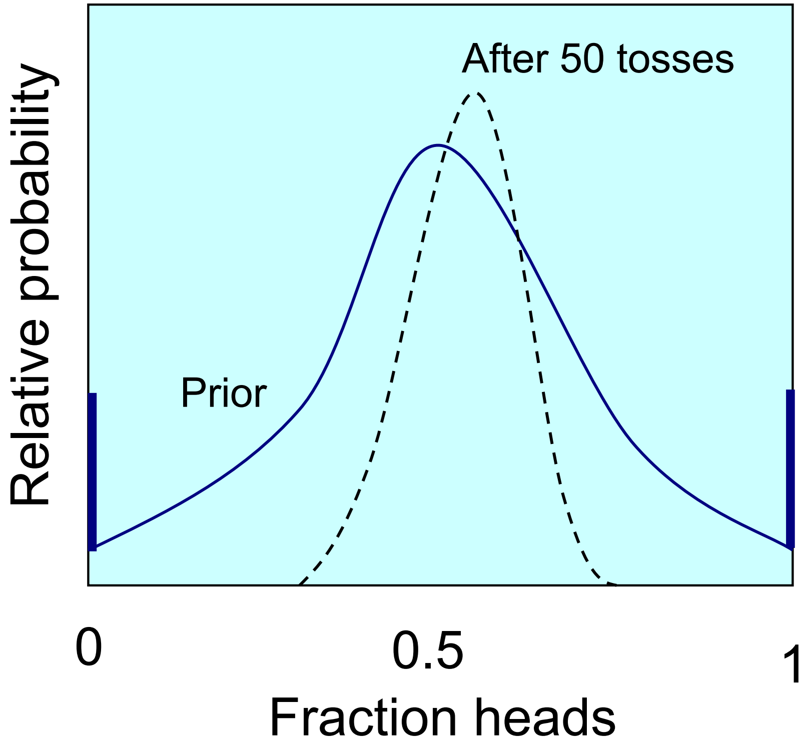
- Uncertainties in parameters are characterized by probability density functions (pdf)
- Probability interpreted as quantitative measure of "**degree of belief**"
- This interpretation is referred to as "subjective probability"
  - different for different people with different knowledge
  - changes with time
  - in science, we seek consensus, avoid bias
- Rules of classical probability theory apply
  - provides firm foundation with mathematical rigor and consistency



Subjective probability can be quantitative

Example – coin toss

- Hypothesis: for a specific coin, fraction of tosses that come up heads = 50%
- Hypothesis seems so reasonable that you might believe it is true
- On basis of this subjective probability, you might be willing to bet with 1:1 odds
- Before any tosses, you might have a prior as shown
- After 50 tosses, you would know better whether coin is fair



Coherent bet quantifies subjective probability

- A property of the Gaussian distribution is that random draws from it will fall inside the interval from -σ to +σ 68% of time
- Suppose, on basis of what you know, you specify the standard error σ of your measurement of a quantity, assuming Gaussian
- If you truly believe in the value of σ you have assigned, you should be willing to accept a bet, randomly chosen between two options:
  - 2:1 bet that a much more accurate measurement would differ from your measured value by **less** than one σ
  - OR 1:2 bet that a much more accurate measurement would differ from your measured value by **more** than one σ
- Your willingness to take bet either way makes this a **coherent bet**
- As physicists, we should make honest effort to assign uncertainties in this spirit, and communicate what we have done

Rules of probability

- Continuous variable *x*; *p*(*x*) is a probability density function (**pdf**)
- **Normalization**: $\int p(x)dx = 1$
- Decomposition of **joint distribution** into conditional distribution:

$$p(x, y) = p(x \mid y)\, p(y)$$

where $p(x \mid y)$ is **conditional** pdf (probability of *x* given *y)*

- ▸ if $p(x \mid y) = p(x),\ x$ is independent of *y*

- **Bayes law** follows:

$$p(y \mid x) = \frac{p(x \mid y)\, p(y)}{p(x)}$$

- **Marginalization**:

$$p(x) = \int p(x, y)\, dy = \int p(x \mid y)\, p(y)\, dy$$

is probability of *x*, without regard for *y* (nuisance parameter)

Rules of probability

- Change of variables: if **x** transformed into **z**, **z** = *f*(**x**), the pdf in terms of **z** is

$$p(\mathbf{z}) = \left|\mathbf{J}\right|^{-1} p(\mathbf{x})$$

where **J** is the Jacobian matrix for the transformation:

$$\mathbf{J} = \begin{pmatrix} \frac{\partial z_1}{\partial x_1} & \cdots & \frac{\partial z_3}{\partial x_1} \\ \vdots & \ddots & \vdots \\ \frac{\partial z_1}{\partial x_3} & \cdots & \frac{\partial z_3}{\partial x_3} \end{pmatrix}$$

Bayesian analysis of experimental data

- Bayes rule

$$\boldsymbol{p}(\boldsymbol{a} \mid \boldsymbol{d}, I) = \frac{\boldsymbol{p}(\boldsymbol{d} \mid \boldsymbol{a}, I)\, \boldsymbol{p}(\boldsymbol{a} \mid I)}{\boldsymbol{p}(\boldsymbol{d} \mid I)}$$

  - where
    d is the vector of measured data values
    a is the vector of parameters for model that predicts the data
  - $p(\boldsymbol{d} \mid \boldsymbol{a}, I)$ is called the **likelihood** (of the data given the true model and its parameters)
  - $p(\boldsymbol{a} \mid I)$ is called the **prior** (on the parameters ***a***)
  - $p(\boldsymbol{a} \mid \boldsymbol{d}, I)$ is called the **posterior** – fully describes final uncertainty in the parameters
  - *I* stands for whatever **background information** we have
    about the situation, results from previous experience,
    our expertise, and the model used
  - denominator provides normalization: $\boldsymbol{p}(\boldsymbol{d}) = \int \boldsymbol{p}(\boldsymbol{d} \mid \boldsymbol{a})\, \boldsymbol{p}(\boldsymbol{a})\, d\boldsymbol{a}$
    i.e., is integral of numerator

Auxiliary information – *I*

All relevant information about the situation may be brought to bear:

- Details of experiment
  - laboratory set up, experiment techniques, equipment used
  - potential for experimental technique to lead to mistakes
  - expertise of experimenters

  more subjective
- Relationship between measurements and theoretical model
- History of kind of experiment
- Appropriate statistical models for likelihood and prior
- Experience and expertise

- We usually leave *I* out of our formulas, but keep it in mind

Likelihood

- Form of the likelihood $p(\boldsymbol{d} \,|\, \boldsymbol{a}, I)$ depends on how we model the uncertainties in the measurements $\boldsymbol{d}$
- Choose pdf that appropriately describes uncertainties in data
  - Gaussian – good generic choice
  - Poisson – counting experiments
  - Binomial – binary measurements (coin toss …)
- Outliers exist
  - likelihood should have a long tail, i.e., there is some probability of large fluctuation
- Systematic errors
  - caused by effects common to many (all) measurements
  - model by introducing variable that affects many (all) measurements; then marginalize it out

Priors

- Noncommittal prior
  - uniform pdf; $p(\theta)$ = const. when θ is offset parameter
  - uniform in $\log(\theta)$; $p(\log\theta)$ = const. when θ is scale parameter
  - choose pdf with maximum entropy, subject to known constraints
- Physical principles
  - cross sections are nonnegative $\Rightarrow p(\theta) = 0$ when $\theta < 0$
  - invariance arguments, symmetries
- Previous experiments
  - use posterior from previous measurements for prior
  - Bayesian updating
- Expertise
  - elicit pdfs from experts in the field, avoiding common info sources
  - elicitation, an established discipline, may be useful in physical sciences

Priors

- Conjugate priors
  - for many forms of likelihood, there exist companion priors that make it easy to integrate over the variables
  - these priors facilitate analytic solutions for posterior
  - example: for the Poisson likelihood in n and λ, the conjugate prior is a Gamma distribution in λ with parameters α and β, which determine the position and width of the prior
  - conjugate priors can be useful and their parameters can often be chosen to create a prior close to what the analyst has in mind
  - however, in the context of numerical solution of complicated overall models, they loose their appeal

Posterior

- Posterior $p(\boldsymbol{a} \mid \boldsymbol{d}, I)$
  - net result of a Bayesian analysis
  - summarizes our state of knowledge
  - it provides fully quantitative description of uncertainties
  - usual practice is to characterize posterior in terms of an estimated value of the variables and their variance
- Visualization
  - difficult to visualize directly because it is a density distribution of many variables (dimensions)
  - Monte Carlo allows us to visualize the posterior through it effect on the model that has been used in the analysis

Visualization of uncertainties

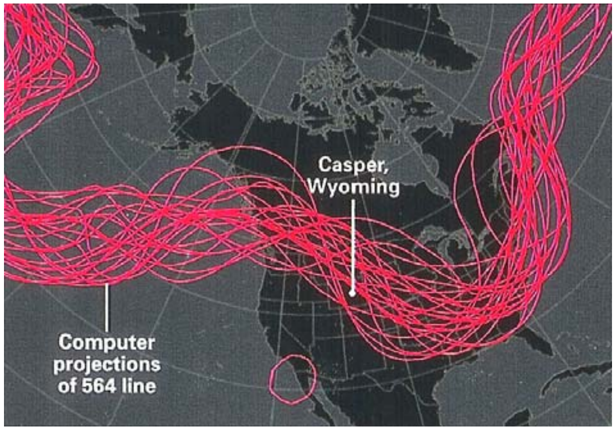
- Visualization plays an important role in developing an understanding of a model and communicating its consequences
- Monte Carlo is often a good choice – choose sets of parameters from their uncertainty distribution and visualize corresponding outputs from the model
- Random sampling from posterior is typically done
- Quasi-random sampling is noteworthy alternative; it provides more uniform sets of samples

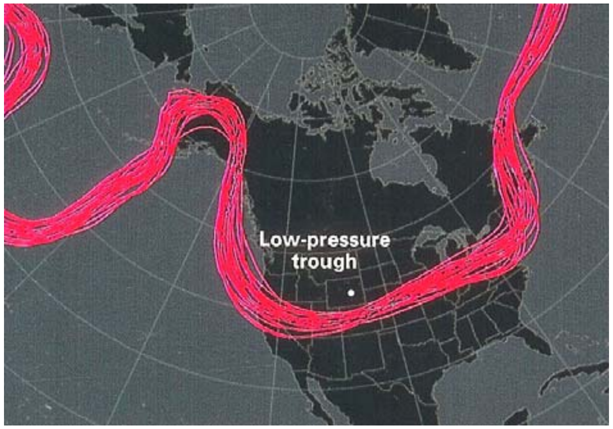
Probability in weather forecasting

- Metrological forecast for Oct. 30, 2003 for Casper, Wyoming
- 22 predictions of 564 line (500 mb) obtained by varying input conditions; indicate plausible outcomes
- Density of lines conveys certainty/probability of winter storms

7 days ahead

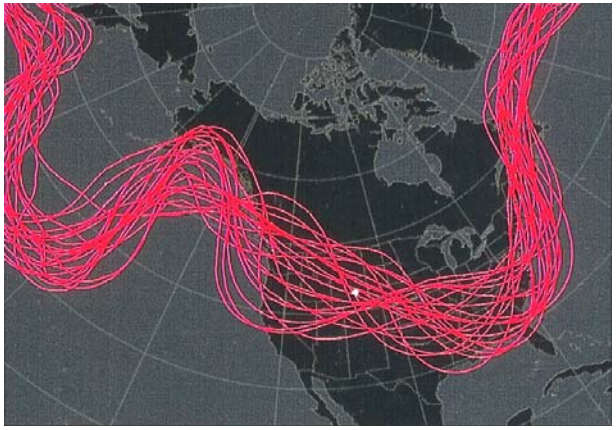
564 line; predictive of winter storms





1 day ahead

4 days ahead



what happened?
20-inches of snow!

National Geographic, June 2005

Posterior – quantitative results

- Quantitative results are obtained by characterizing the posterior:
  - ► mean (first moment): $\hat{x} = \langle x \rangle = \int x\, p(x)\, dx$
    - mean minimizes quadratic cost function
  - ► maximum (peak position); same as mean if pdf symmetric
  - ► standard deviation (second moment): $\sigma_x = \sqrt{\int \left(x - \langle x \rangle\right)^2 p(x)\, dx}$
    - **standard error**
  - ► covariance matrix: $\mathrm{cov}(x, y) = \mathbf{C}_{xy} = \int \left(x - \langle x \rangle\right)\left(y - \langle y \rangle\right) p(x, y)\, dxdy$
    - correlation matrix: $\mathrm{corr}(x, y) = \mathbf{R}_{xy} = \sigma_{xy}^2 / \sigma_x \sigma_y$
  - ► credible (confidence) interval, e.g., 95% credible interval
- Means for estimating these include:
  - ► can use calculus if posterior is in convenient analytic form
  - ► second-order approximation around peak (numerical)
  - ► Monte Carlo (numerical)

Higher-order inference

- One can make inferences about models, not just parameters
- The posterior for a model is

$$
\begin{aligned}
p(M \mid \boldsymbol{d}) &= \int p(\boldsymbol{a}, M \mid \boldsymbol{d})\, d\boldsymbol{a} = \int p(\boldsymbol{a}, M \mid \boldsymbol{d})\, d\boldsymbol{a} \\
&\propto \int p(\boldsymbol{d} \mid \boldsymbol{a}, M)\, p(\boldsymbol{a}, M)\, d\boldsymbol{a} \\
&= p(M) \int p(\boldsymbol{d} \mid \boldsymbol{a}, M)\, p(\boldsymbol{a} \mid M)\, d\boldsymbol{a}
\end{aligned}
$$

  - the final integral is the normalizing denominator in original Bayes law for $p(\boldsymbol{a}|\boldsymbol{d})$; it is called the **evidence**
  - while the evidence is not needed for parameter inference, it is required for model inference
- May be used for **model selection**, e.g., deciding between two or more models
  - e.g., how many terms to include in a functional analysis

Summary

In this tutorial:

- Need for uncertainty quantification
- Bayesian fundamentals
  - ► subjective probability, nevertheless quantifiable
  - ► Bayesian use of probability theory
  - ► posterior sampling
  - ► visualization of uncertainties – Monte Carlo
  - ► higher-order inference

Tutorial 2
Bayesian modeling

Peelle's Pertinent Puzzle (1987)

Overview:

- Paradoxical result produced by strong correlations in uncertainties
- Probabilistic view of PPP
- Specific probabilistic model for PPP elucidates how correlations in uncertainties arise
- Plausible experimental situation consistent with PPP result
- Bayesian approach to coping with uncertainty in model
- With probabilistic modeling, you can go beyond simple linear, additive models
- PPP underlines the need to specify **how** uncertainties contribute to reported data

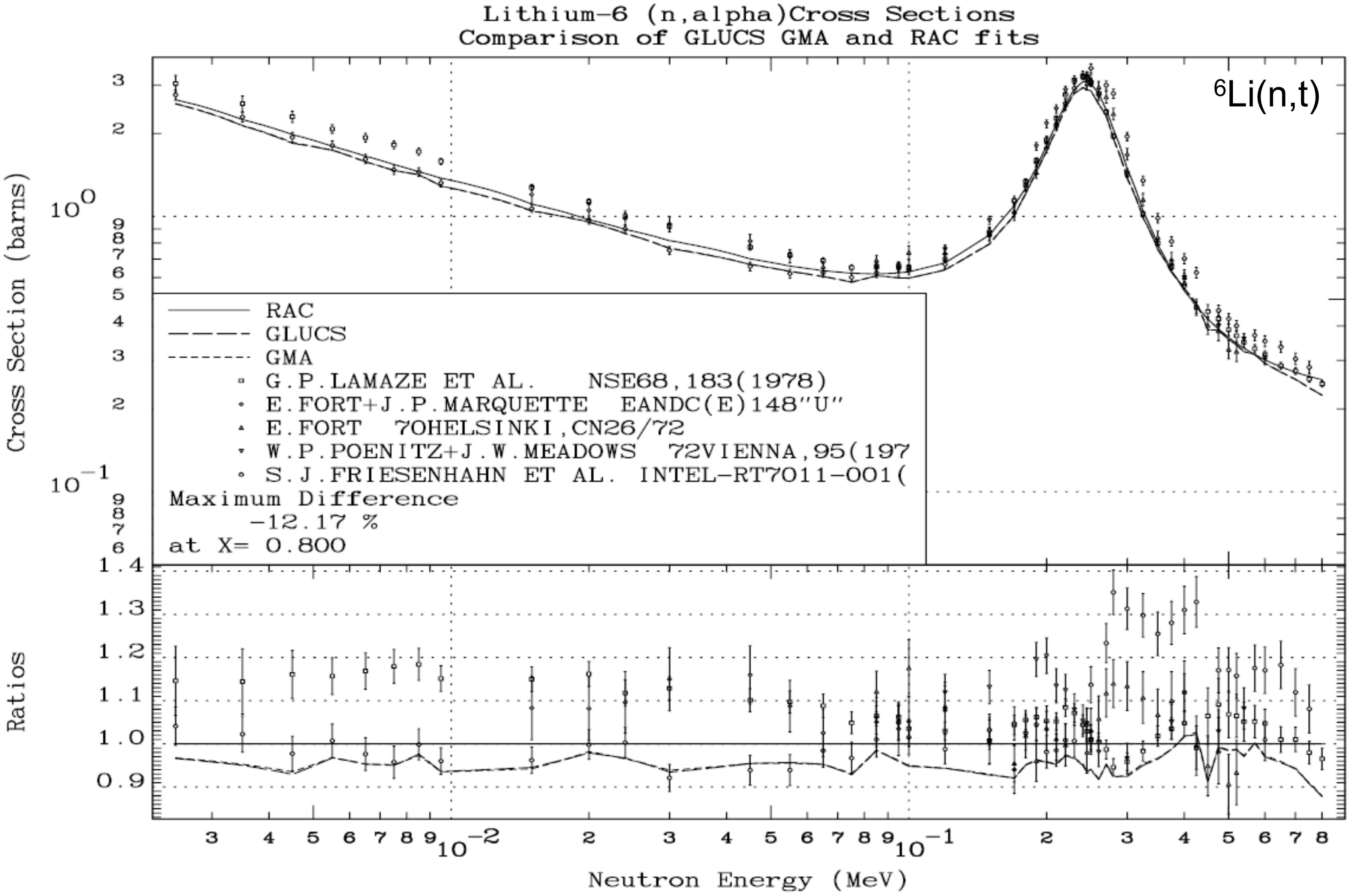
Peelle's pertinent puzzle

- Robert Peelle (ORNL) posed the PPP in 1987:
  Given two measurements of same quantity *x*:
  $m_1 = 1.5$; $m_2 = 1.0$,
  each with independent standard error of 10% ,
  and fully correlated standard error of 20% .
  Weighted average using least-squares is $x = 0.88 \pm 0.22$

- Peelle asks "under what conditions is this result reasonable?"
- By extension, if this not reasonable, what answer is appropriate?
- PPP is pertinent – its effect has been observed in nuclear data evaluation for decades
- Comment – PPP description of errors is ambiguous, which leads to numerous plausible interpretations

PPP in cross-section evaluation

- Although the PPP problem may seem academic, it has significant real-world consequences in cross-section evaluation
  - historically, fits to several data sets fall below lowest measurements



from Pronyaev, *INDC(NDS)*-438, p. 163 (2003)

note large data discrepancies

Standard solution to PPP

- The solution given in PPP is based on standard matrix equations for least-squares result:

  estimated value $\boldsymbol{x} = (\boldsymbol{G}^{\boldsymbol{T}}\boldsymbol{C}^{-1}\boldsymbol{G})^{-1}\boldsymbol{G}^{\boldsymbol{T}}\boldsymbol{C}^{-1}\boldsymbol{m}$

  covariance in estimate $\boldsymbol{V} = (\boldsymbol{G}^{\boldsymbol{T}}\boldsymbol{C}^{-1}\boldsymbol{G})^{-1}$

  where the sensitivity matrix is $\boldsymbol{G} = [1.0\ \ 1.0]$

  and the measurements are the vector $\boldsymbol{m} = [1.5\ \ 1.0]^{\boldsymbol{T}}$

  with covariance matrix $\boldsymbol{C} = \begin{pmatrix} 1.5^2 * (0.1^2 + 0.2^2) & 1.5 * 1.0 * 0.2^2 \\ 1.5 * 1.0 * 0.2^2 & 1.0^2 * (0.1^2 + 0.2^2) \end{pmatrix}$

- Result is $x = 0.88 \pm 0.22$
- This result is smaller than both measurements, which seems implausible

Probabilistic view of standard PPP solution

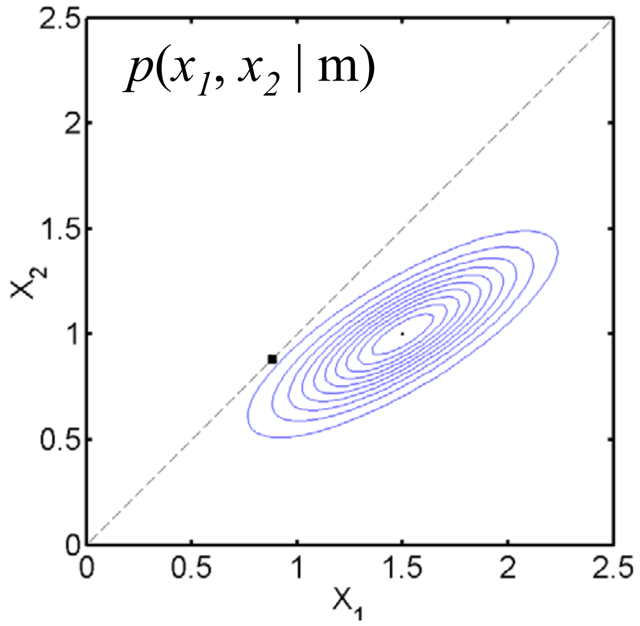
- Consider the probability density function (pdf) for the variables

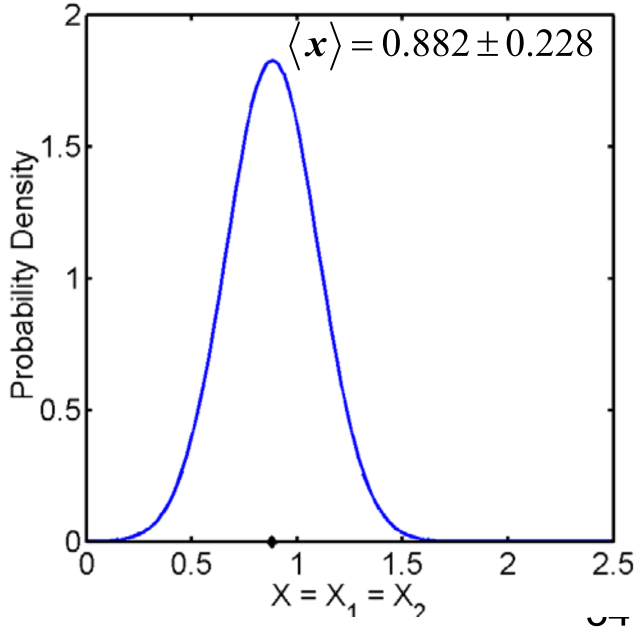
$$\boldsymbol{x} = [x_1 \quad x_2]^T$$

$$p(\boldsymbol{x} \,|\, \boldsymbol{m}) \propto \exp\left\{ -\frac{1}{2}^{T} (\boldsymbol{x}-\boldsymbol{m})^T \boldsymbol{C}^{-1} (\boldsymbol{x}-\boldsymbol{m}) \right\}$$

where measurements are $\boldsymbol{m} = [1.5 \;\; 1.0]^T$ and their covariance matrix is

$$\boldsymbol{C} = \begin{pmatrix} 1.5^2 * (0.1^2 + 0.2^2) & 1.5 * 1.0 * 0.2^2 \\ 1.5 * 1.0 * 0.2^2 & 1.0^2 * (0.1^2 + 0.2^2) \end{pmatrix}$$





- For $x = x_1 = x_2$ (diagonal of 2D pdf), $p(x|\boldsymbol{m})$ is normal distribution centered at 0.88

Probabilistic model for additive error

- Represent common uncertainty in measurements by systematic additive offset Δ: $x_1 = m_1 + \varepsilon_1 + \Delta; \quad x_2 = m_2 + \varepsilon_2 + \Delta$
  - where the ε_i represent the random fluctuations
- Bayes law gives joint pdf for *x* and Δ

$$p(x, \Delta \mid \boldsymbol{m}) = p(\boldsymbol{m} \mid x, \Delta)\, p(x)\, p(\Delta)$$

where priors *p*(*x*) is uniform and *p*(Δ) assumed normal (σ_Δ= 0.2)

- Writing $p(x, \Delta \mid \boldsymbol{m}) \propto \exp\{-\varphi\}$ and assuming normal distributions

$$2\varphi = \frac{\left(x_1 - m_1 - \Delta\right)^2}{\sigma_1^2} + \frac{\left(x_2 - m_2 - \Delta\right)^2}{\sigma_2^2} + \frac{\Delta^2}{\sigma_\Delta^2}$$

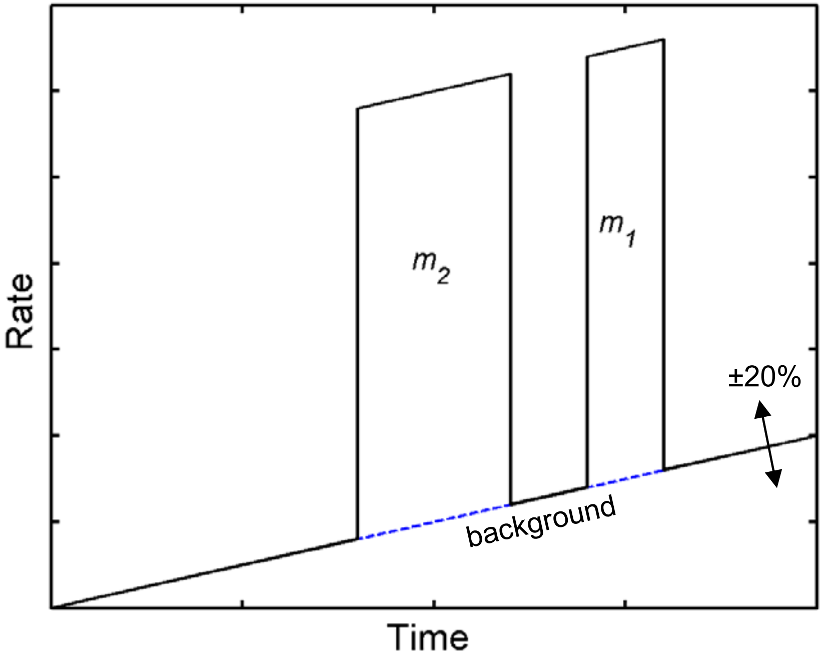
where $\sigma_1 = 0.1 * m_1; \quad \sigma_2 = 0.1 * m_2; \quad \sigma_\Delta = 0.2$

- Pdf for *x* obtained by integration: $p(x \mid \boldsymbol{m}) = \int p(x, \Delta \mid \boldsymbol{m})\, \mathrm{d}\Delta$

- This model equivalent to $p(\boldsymbol{x} \mid \boldsymbol{m}) \propto \exp\left\{ -\frac{1}{2}^{\boldsymbol{T}} \left(\boldsymbol{x} - \boldsymbol{m}\right)^{\boldsymbol{T}} \boldsymbol{C}^{-1} \left(\boldsymbol{x} - \boldsymbol{m}\right) \right\}$

Plausible experimental scenario

- Under what conditions is PPP result reasonable?
- Suppose that
  - measurements made in intervals shown
  - from experience with apparatus, we know background increases linearly in time
  - background subtraction for m_1 is 1.5 times larger than for m_2; leads to stated covariance matrix
- For this scenario, the additive model is appropriate, and the PPP solution, 0.88, is the correct answer



Probabilistic model for normalization error

- Represent common uncertainty in measurements by systematic error in normalization factor c: $cx = m_1 + \varepsilon_1; \quad cx = m_2 + \varepsilon_2$
  - where the ε_i represent the random fluctuations
- Following same development as before, where prior $p(c)$ assumed normal with expected value of 1 and $\sigma_c = 0.2$
- Writing $p(cx, c \mid \boldsymbol{m}) \propto \exp\{-\varphi\}$

$$2\varphi = \frac{(cx - m_1)^2}{\sigma_1^2} + \frac{(cx - m_2)^2}{\sigma_2^2} + \frac{(c-1)^2}{\sigma_c^2}$$

where $\sigma_1 = 0.1 * m_1; \quad \sigma_2 = 0.1 * m_2; \quad \sigma_c = 0.2$

- Divide $p(cx, c)$ by Jacobian $J = 1/c$ to get $p(x, c)$, which is a log-normal distribution
- $p(x)$ obtained by numerical integration: $p(x \mid \boldsymbol{m}) = \int p(x, c \mid \boldsymbol{m})\,\mathrm{d}c$
- This approach promoted by D. Smith (1991)

Probabilistic view of normalization error

- Consider the probability density function (pdf) for variables $\boldsymbol{x} = [x_1 \;\; x_2]^T$

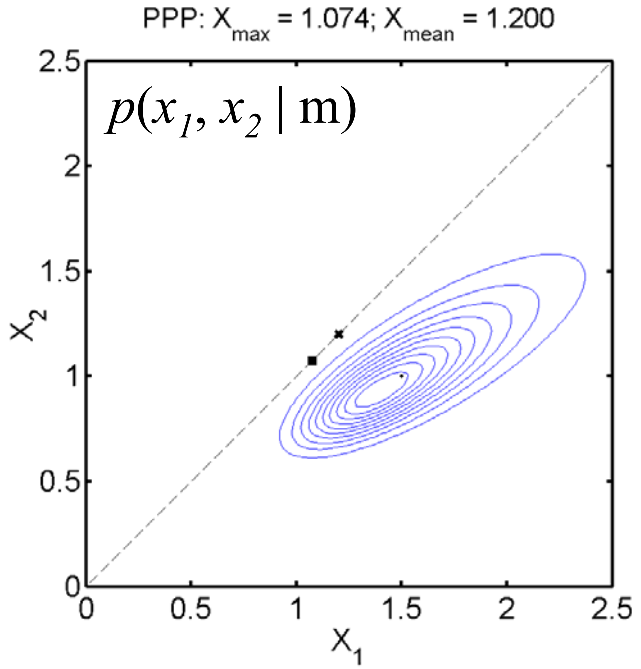
$$\chi^2 = \left(\frac{cx_1 - m_1}{m_1\rho_1}\right)^2 + \left(\frac{cx_2 - m_2}{m_2\rho_2}\right)^2 + \left(\frac{c-1}{\sigma_c}\right)^2;$$

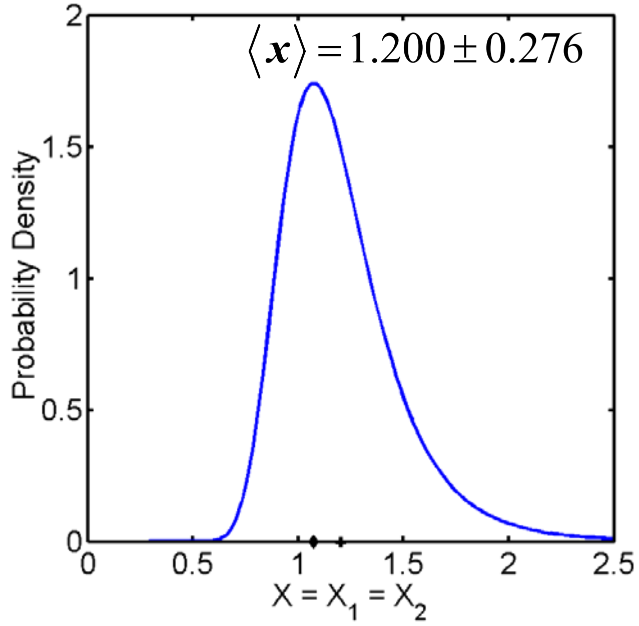
$$\sigma_c = \rho_c;$$

where measurements are $\boldsymbol{m} = [1.5 \;\; 1.0]^T$

- ▸ also, divide $p(cx, c)$ by Jacobian $J = 1/c$ to get $p(x, c)$,
- ▸ for $x = x_1 = x_2$ (diagonal of 2D pdf), $p(x|\boldsymbol{m})$ is not a simple normal distribution
- ▸ max at: $x_{max} = 1.074$
- ▸ posterior mean and rmsd:

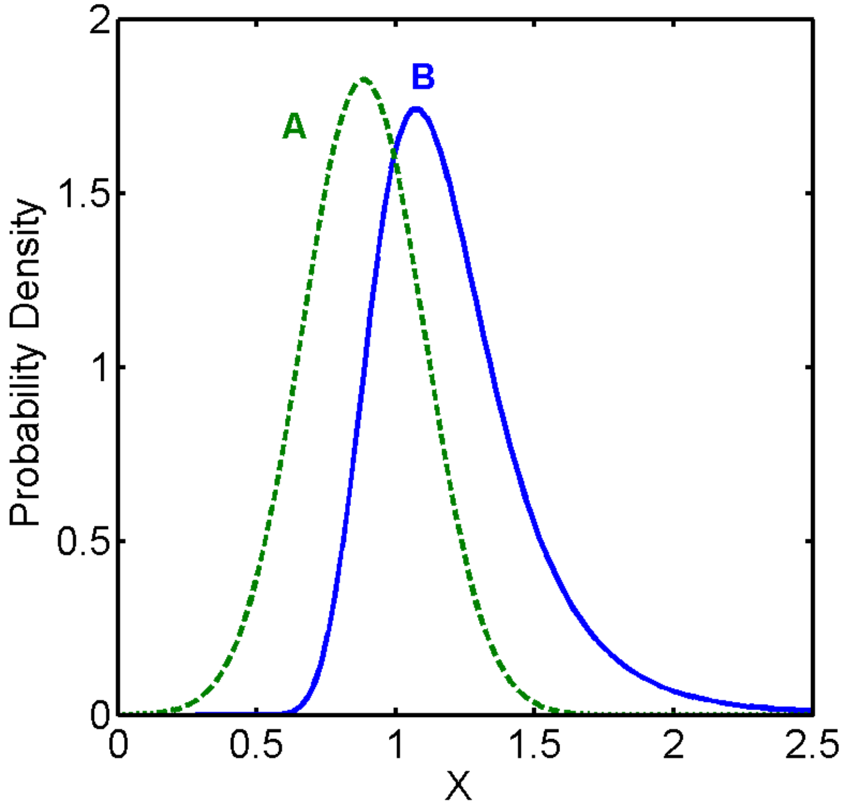
$$x = 1.200 \pm 0.276$$





Probabilistic model for normalization error

- Compare pdfs for two models for correlated effect:
  A – additive offset
  B – normalization factor
- Observe significant difference in two results
  - emphasizes need to know which kind of effect leads to correlation



- Probabilistic modeling is capable of handling a variety of known effects

But which model should we use?

- Ambiguity in specifying source of correlation leads to uncertainty about which model to use
- Bayesian approach can handle model uncertainty

$$p(x \mid \boldsymbol{m}) = \int p(x, M \mid \boldsymbol{m})\,\mathrm{d}M$$

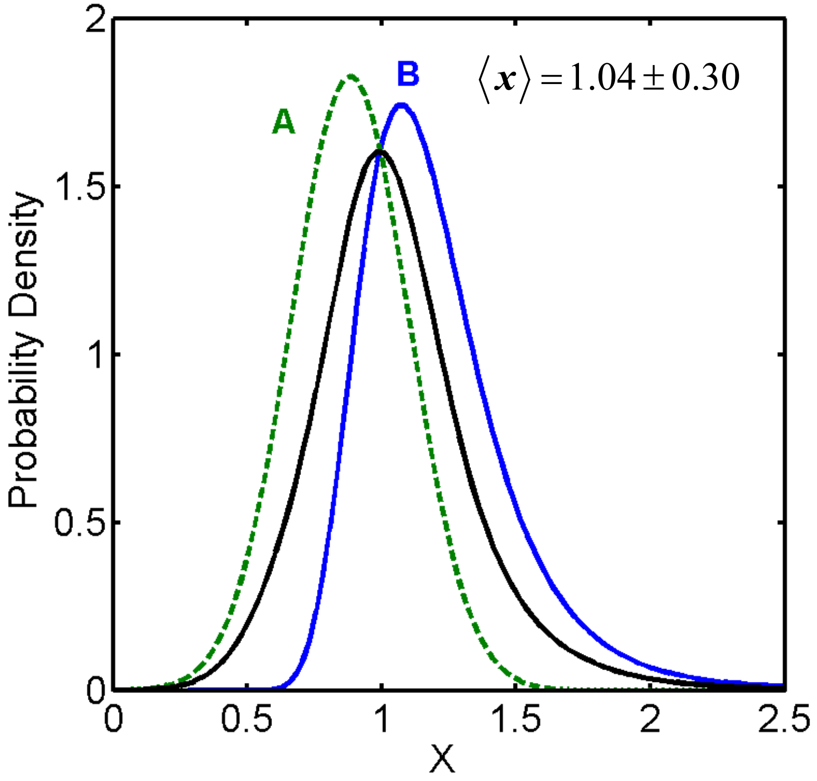
$$= \int p(x \mid \boldsymbol{m}, M)\, p(M)\,\mathrm{d}M$$

$$= \frac{1}{2}\, p(x \mid \boldsymbol{m}, M_1) + \frac{1}{2}\, p(x \mid \boldsymbol{m}, M_2)$$

  - for two equally likely models M_1 and M_2

- Answer is average **both** pdfs!!

$$x = 1.04 \pm 0.30$$



solid black line is average of A and B

An alternative approach

- Devinder Sivia offers an variation on this approach
- Use data to help decide which model to use

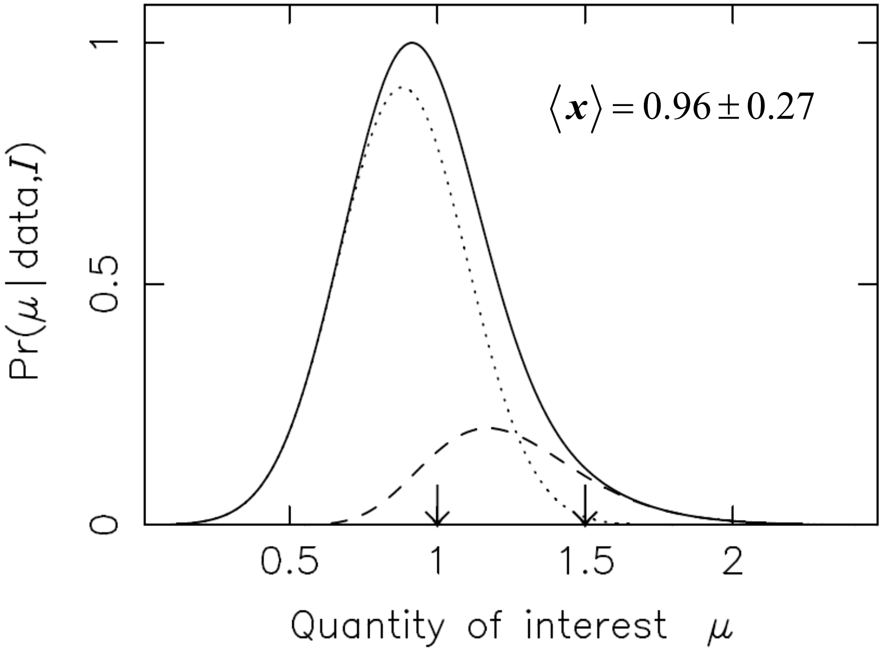
$$p(x \mid \boldsymbol{m}) = \sum_i p(x, M_i \mid \boldsymbol{m})$$

$$= \sum_i p(x \mid \boldsymbol{m}, M_i)\, p(M_i \mid \boldsymbol{m})$$

$$= w_1\, p(x \mid \boldsymbol{m}, M_1) + w_2\, p(x \mid \boldsymbol{m}, M_2)$$

  - where w_i is proportional to the evidence integral for $p(M_i \mid \boldsymbol{m})$

- Answer is: $x = 0.96 \pm 0.27$
- Comment: relative weights depend heavily on resp. priors; perhaps not a good situation



solid black line is weighted average of other two distributions

from D. Sivia, *Proc. AMCTM Conf.*, (World Scientific, 2005)

Conclusions

- PPP result is consistent with plausible experimental scenario
  - in which correlated (systematic) error contributes additively to result
- Ambiguous statement of the PPP leads to other interpretations
  - some of which yield more plausible answers
- Analysts need better information to analyze data without guessing

- Probabilistic modeling can cope with various known uncertainty effects

Conclusions

- **Experimenters – please provide measurement details**
- Some of the details needed:
  - ▸ specify standard errors as precisely as possible, indicating where uncertainties in their assessment lie
  - ▸ specify components in uncertainties and whether they are
    - independent, or correlated, e.g., systematic errors
    - given relative to measured quantities or inferred values
    - additive (background subtraction) or multiplicative (normalization)
- **Correlation matrix by itself is not enough**
- Another issue in PPP is inconsistency between two measurements: one can cope with this discrepancy by introducing notion that the true errors may differ from quoted errors, i.e., treatment of outliers

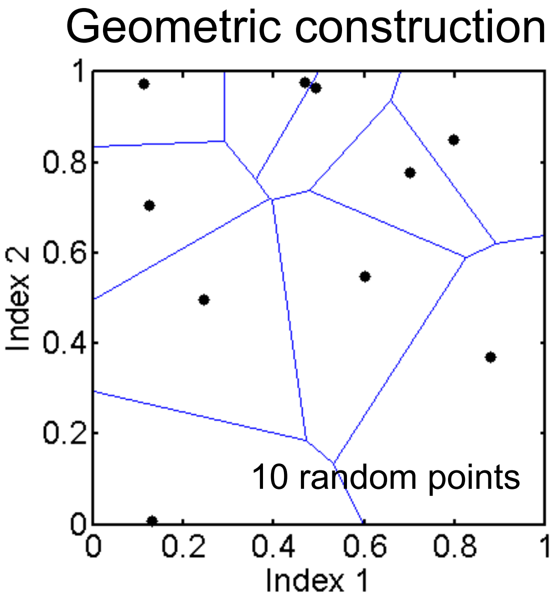
Monte Carlo techniques

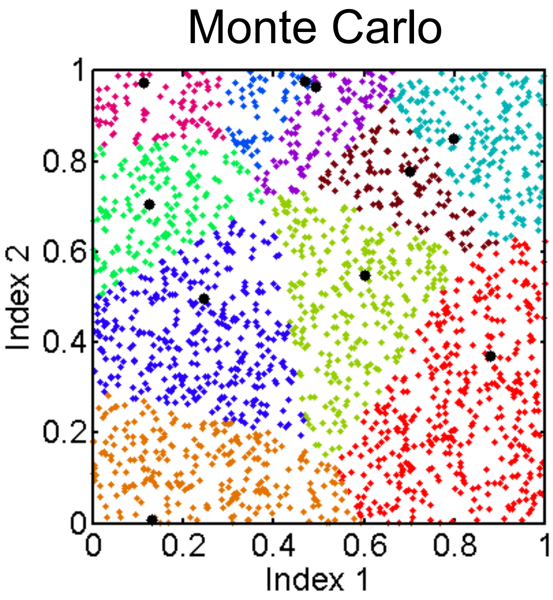
Monte Carlo – represent pdf by a set of point samples

- Typically use MC to draw samples from posterior for parameters, which are fed into model to get prediction; **predictive distribution**
- Visualization of pdf, uncertainty
- Numerical calculations
  - estimation of mean, standard deviation, correlations
  - integration, marginalization
- Quasi-Monte Carlo – select points with more uniform distribution
  - provide more accurate estimates for fixed number of samples
  - often deterministic point sets
- Markov chain Monte Carlo
  - draw random samples for numerically-defined pdf
  - facilitates inference through numerical calculations

Voronoi analysis

- Voronoi diagram
  - partitions domain into polygons
  - points in *i*th Voronoi region are closest to *i*th generating point, $\boldsymbol{x}_i$
  - boundaries often obtained by geometrical construction
- Monte Carlo technique for Voronoi analysis
  - randomly throw large number of points $\boldsymbol{z}_k$ into region
  - compute distance of each $\boldsymbol{z}_k$ to all generating points $\{\boldsymbol{x}_i\}$
  - z_k belongs to Voronoi region of closest $\boldsymbol{x}_j$
  - can compute volume, first moment , radial moments, identify neighbors, …
- Readily extensible to high dimensions





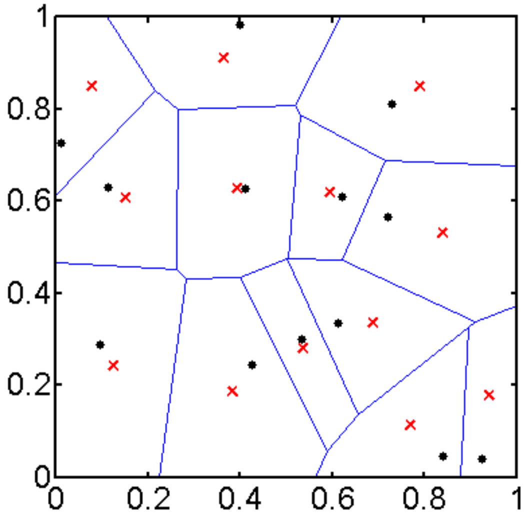
Centroidal Voronoi Tessellation

- Plot shows 13 random points (·) and the centroids of their Voronoi regions (×)
- A point set is called a Centroidal Voronoi Tessellation (CVT) when the generating points $\mathbf{z}^j$ coincide with the centroids their Voronoi regions; a CVT minimizes

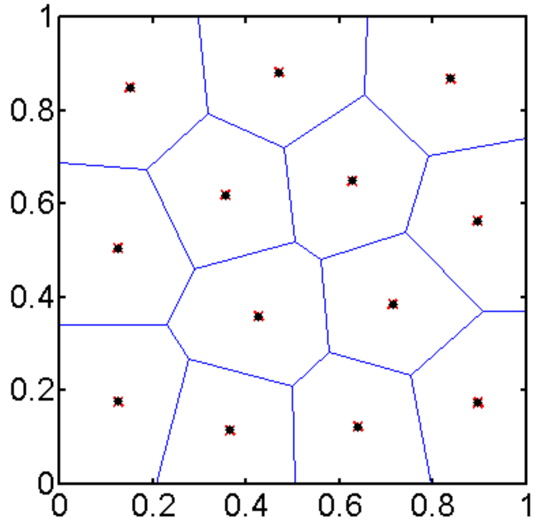
$$\sum_j \int_{V_j} \left|\mathbf{z}^j - \mathbf{x}\right|^2 \mathrm{d}\mathbf{x}$$

- Algorithm (McQueen)
  - start with arbitrary set of generating points
  - perform Voronoi analysis using Monte Carlo
  - move each generating point to its Voronoi centroid
  - iterate lasts two steps until convergence
- Final CVT points are uniformly distributed

Start with random points

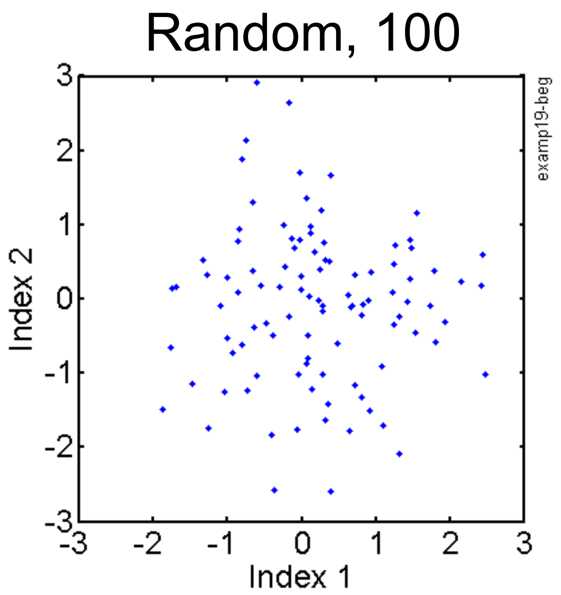


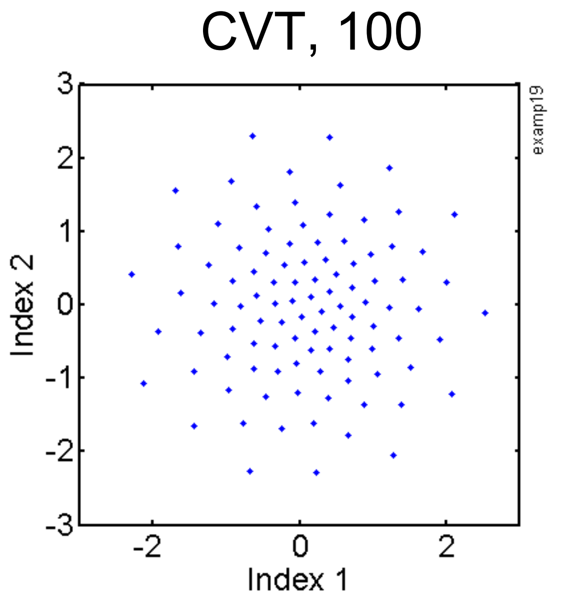
Final CVT point set



CVT for multi-variate normal distribution

- CVT algorithm works for an arbitrary density function, e.g., a normal distribution
- In above MC algorithm for Voronoi analysis, simply draw random numbers from desired distribution
- Plots show starting random point set and final CVT set
- Radii of points are rescaled to achieve desired average variance along axes
- CVT points appear uniformly distributed within constraint of adhering to unit-variance normal distribution
- This kind of distribution may have benefits for MC calculations and visualizations





Sampling from correlated normal distribution

- Want to draw samples from multi-variate normal distribution with known covariance $\mathbf{C_x}$
- Important to include correlations among uncertainties, i.e., off-diagonal elements
- Algorithm:
  - ► perform eigenanalysis of covariance matrix of *d* dimensions

    $$\mathbf{C_x} = \mathbf{U \Lambda U}^{\mathrm{T}}$$

    where $\mathbf{U}$ is orthogonal matrix of eigenvectors and
    $\mathbf{\Lambda}$ is the diagonal matrix of eigenvalues
  - ► draw *d* samples from uncorrelated unit-variance normal distr., ξ_i
  - ► scale this vector by $\lambda_i^{½}$
  - ► transform vector into parameter space using the eigenvector matrix
  - ► to summarize, fluctuations are given by: $\Delta\mathbf{x} = \mathbf{U \Lambda}^{1/2} \boldsymbol{\xi}$

Sampling from correlated normal distribution

Proof of algorithm:

- Want to draw samples from multi-variate normal distribution with specified covariance $\mathbf{C_x}$
- Algorithm:
  - fluctuations given by: $\Delta\mathbf{x} = \mathbf{U}\boldsymbol{\Lambda}^{1/2}\boldsymbol{\xi}$
    where ξ_i randomly drawn from uncorrelated normal pdf and
    $\mathbf{U}$ and $\boldsymbol{\Lambda}$ come from an eigenanalysis of $\mathbf{C_x}$: $\mathbf{C_x} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^{\mathrm{T}}$
    where $\mathbf{U}$ is orthogonal matrix of eigenvectors and
    $\boldsymbol{\Lambda}$ is the diagonal matrix of eigenvalues
- Proof:
  - Covariance of an ensemble of $\mathbf{x}$ vectors is

$$\mathbf{C} = \left\langle \Delta\mathbf{x}\,\Delta\mathbf{x}^{\mathrm{T}} \right\rangle = \left\langle \mathbf{U}\boldsymbol{\Lambda}^{1/2}\boldsymbol{\xi}\boldsymbol{\xi}^{\mathrm{T}}\boldsymbol{\Lambda}^{1/2}\mathbf{U}^{\mathrm{T}} \right\rangle$$

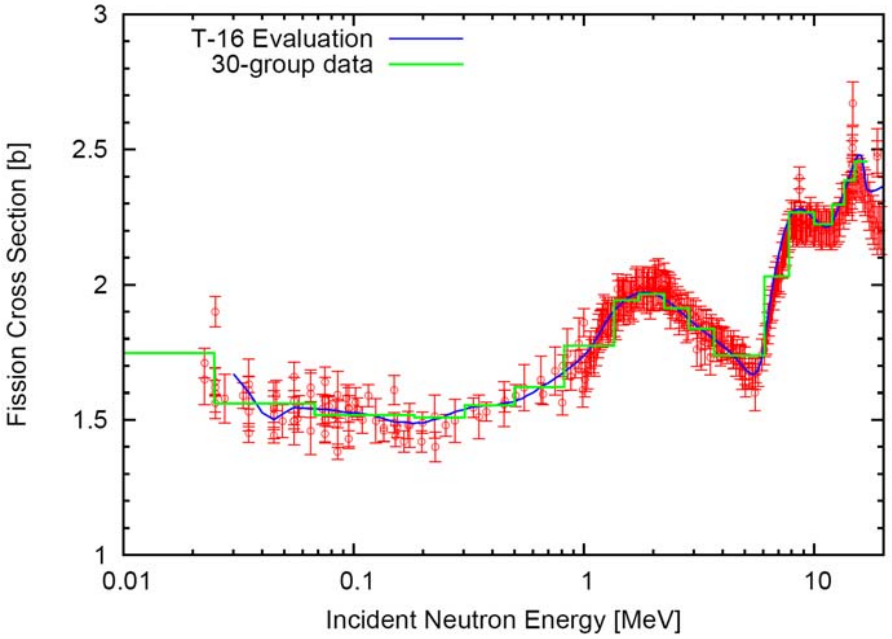
$$= \mathbf{U}\boldsymbol{\Lambda}^{1/2}\left\langle \boldsymbol{\xi}\boldsymbol{\xi}^{\mathrm{T}} \right\rangle \boldsymbol{\Lambda}^{1/2}\mathbf{U}^{\mathrm{T}} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^{\mathrm{T}} = \mathbf{C_x}$$

  - thus, the fluctuations $\Delta\mathbf{x}$ have the desired covariance

Neutron cross sections

- Plot shows
  - measured fission cross sections for neutrons on ^{239}Pu; red data points
  - inferred cross sections; blue line
  - weighted average in 30 energy bins (groups); green histogram
- PARITSN code simulates neutron transport based on multigroup, discrete-ordinates method
  - uses 30 energy bins (groups)
  - calculates criticality for specified configuration of fissile-material
  - establish dependence of criticality experiment to cross sections

^{239}Pu cross sections



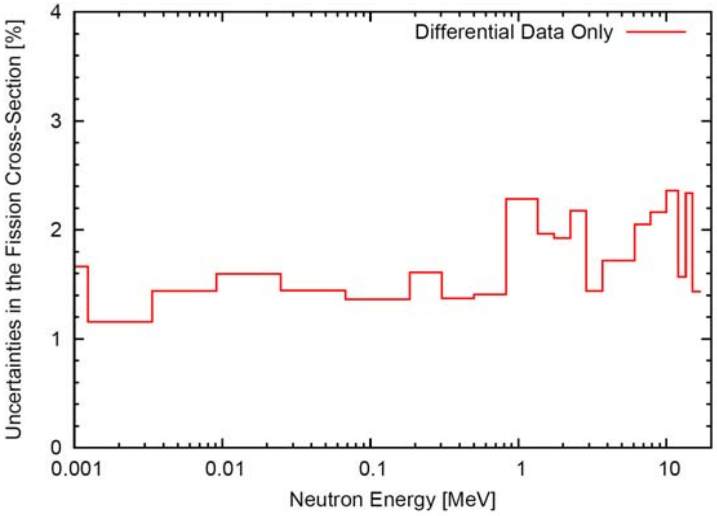
cross section evaluation, P. Young et al.

Neutron cross sections - uncertainties

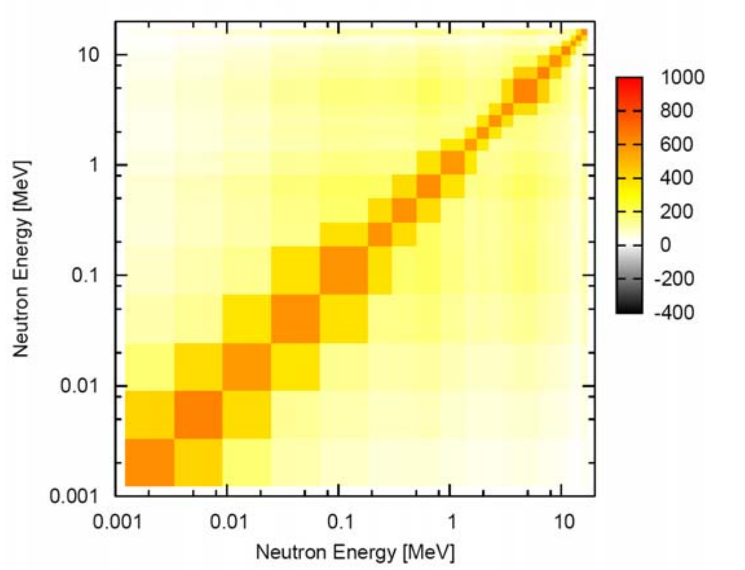
- Analysis of measured cross sections yields a set of evaluated cross sections

- Uncertainties in evaluated cross sections are ~ 1.4-2.4 %
- Covariance matrix important
- Strong positive correlations caused by normalization uncertainties in each experiment

standard error in cross sections



correlation matrix

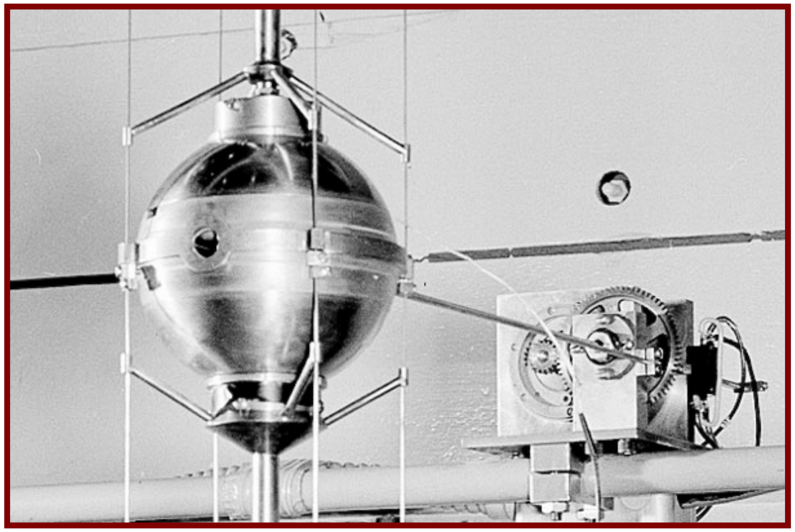


JEZEBEL – criticality experiment

- JEZEBEL experiment (1950-60)
  - fissile material ^{239}Pu
  - measure neutron multiplication as function of separation of two hemispheres of material
  - summarize criticality with neutron multiplication factor, $k_{eff} = 0.9980 \pm 0.0019$
  - very accurate measurement

- Our goal – use highly accurate JEZEBEL measurement to improve our knowledge of ^{239}Pu cross sections

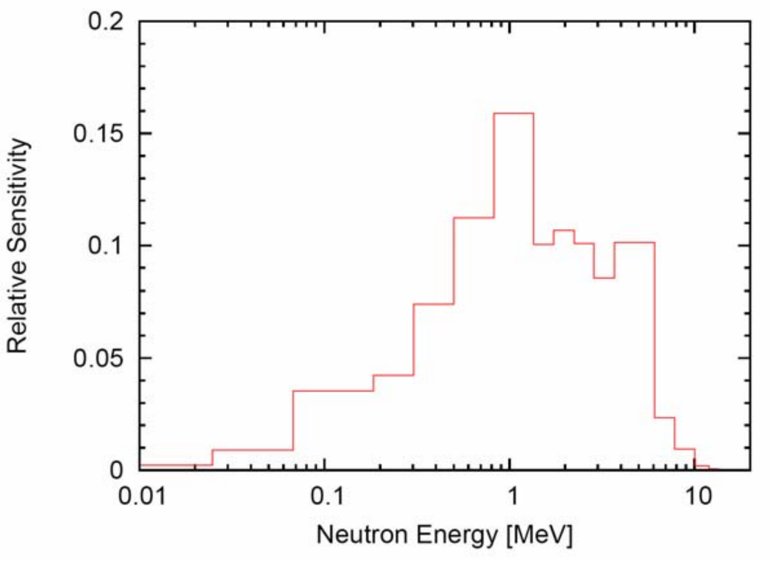
JEZEBEL set up



JEZEBEL – sensitivity analysis

- PARITSN code calculates k_{eff} on basis of neutron cross sections
- Sensitivity of k_{eff} to cross sections found by perturbing cross section in each energy bin by 1% and observing increase in k_{eff}
- Observe that 1% increase in all cross sections results in 1% increase in k_{eff} , as expected

k_{eff} sensitivity to cross sections



Bayesian update

- For data linearly related to the parameters, the Bayesian (aka Kalman) update for Gaussian distributions is

$$\mathbf{C}_1^{-1}\mathbf{x}_1 = \mathbf{C}_0^{-1}\mathbf{x}_0 + \mathbf{S}_\mathbf{y}^\mathrm{T}\mathbf{C}_\mathbf{y}^{-1}\mathbf{S}_\mathbf{y}(\mathbf{y} - \mathbf{y}_0)$$

$$\mathbf{C}_1^{-1} = \mathbf{C}_0^{-1} + \mathbf{S}_\mathbf{y}^\mathrm{T}\mathbf{C}_\mathbf{y}^{-1}\mathbf{S}_\mathbf{y}$$

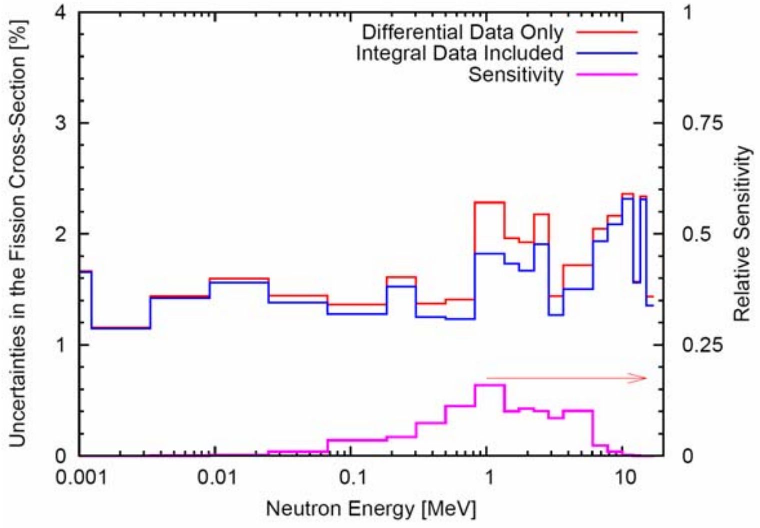
  - $\mathbf{x}_0$ and $\mathbf{x}_1$ are parameter vectors before and after update
  - $\mathbf{C}_0$ and $\mathbf{C}_1$ are their covariance matrices
  - $\mathbf{y}$ and $\mathbf{C}_\mathbf{y}$ are the measured data vector and its covariance
  - $\mathbf{y}_0$ is the value of $\mathbf{y}$ for $\mathbf{x}_0$
  - $\mathbf{S}_\mathbf{y}$ is the matrix of the sensitivity of $\mathbf{y}$ to $\mathbf{x}$; $\partial\mathbf{y}/\partial\mathbf{x}$

- For the JEZEBEL case, $\mathbf{y}$ is a scalar (k_{eff}), $\mathbf{C}_\mathbf{y}$ is a scalar (variance), and $\mathbf{S}_\mathbf{y}$ is a vector

Updated cross sections

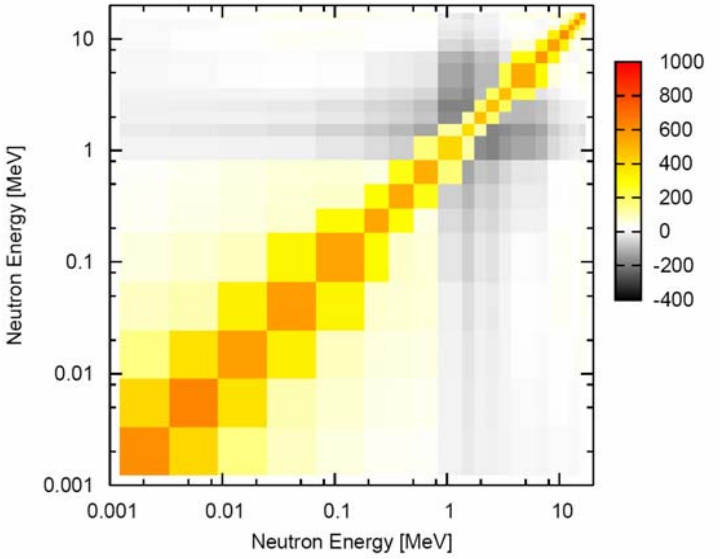
- Plot shows uncertainties in cross sections before and after using JEZEBEL measurement
- Modest reduction in uncertainties; follows energy dependence of sensitivity

standard error in cross sections



- Correlation matrix is significantly altered
- Strong negative correlations introduced by integral constraint of matching JEZEBEL's k_{eff}
  - reduction in uncertainties in future prediction depends on how closely its sensitivity matches JEZEBEL's

correlation matrix

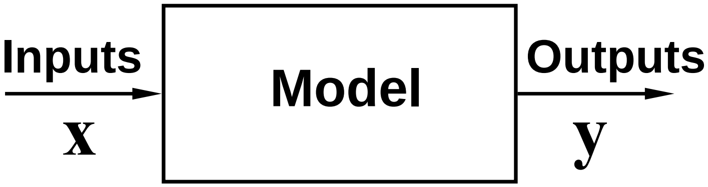


Linear-response model – output uncertainty

- Assume outputs of a model are linearly related to perturbations in the inputs,

$$\delta \boldsymbol{y} = \mathbf{S}_{\mathbf{y}}^{\mathrm{T}} \delta \mathbf{x}$$

  - where $\mathbf{S}_{\mathbf{y}}$ is sensitivity matrix $\partial \mathbf{y} / \partial \mathbf{x}$



- The covariance in the output **y** is

$$\mathbf{C}_{\mathbf{y}} = \mathbf{S}_{\mathbf{y}}^{\mathrm{T}} \mathbf{C}_{\mathbf{x}} \mathbf{S}_{\mathbf{y}}$$

  - when output y is a scalar,
    the covariance $\mathbf{C}_{\mathbf{y}}$ is a scalar (variance),
    and $\mathbf{S}_{\mathbf{y}}$ is a vector

- If linear model is sufficient and one knows $\mathbf{S}_{\mathbf{y}}$, then predictive distribution is easily characterized
- For complex simulations, $\mathbf{S}_{\mathbf{y}}$ is not usually known

Uncertainty in subsequent simulations

- Our goal is to use updated cross sections in new calculations
  - expect that integral constraint will reduce uncertainties
- Demonstrate usefulness of quasi-MC in form of CVT point sets by "predicting" k_{eff} measured in JEZEBEL
  - for this demo, assume linear model with known sensitivity vector
  - under this assumption, we can calculate exact answer and compare to MC-style sampling to obtain predictive distribution
- For a new physical scenario, we would not have sensitivity vector and would have to do full simulation calculation
  - thus, only a modest number of function evaluations can be done

Accuracy of predicted k_{eff} and its uncertainty

- Prediction based on liner model with know sensitivities
  - only 30 sample sets allowed for neutronics calc. because of time
  - check accuracy of predicted mean and standard deviation
- Conclude – CVT is more accurate than random sampling

Performance summary from 1000 runs, each with set of 30 sample vectors; 'rot' indicates single sample set randomly rotated to achieve each new one

	est. mean k_{eff}		est. std. dev. k_{eff}	
	avg.	rms dev.	avg.	rms dev.
random	0.99788	0.00037	0.00191	0.00028
random-rot	0.99824	0.00010	0.00218	0.00010
CVT-rot	0.99796	0.00001	0.00197	0.00002
exact-linear	0.99796	-	0.00195	-

Summary

In this tutorial:

- Peelles' pertinent puzzle
  - impact on cross-section evaluation
  - probabilistic modeling; additive and multiplicative systematic effects
  - experimenters need to provide more than correlation matrices
- Monte Carlo
  - generation of samples with specified covariance matrix
  - quasi-Monte Carlo – more uniformly spaced points than random
  - Centroidal Voronoi Tessellation (CVT) algorithm
- Bayesian updating of cross sections to include integral data
  - JEZEBEL criticality experiment
  - integral constraint results in negative correlations
  - CVT point set improves prediction accuracy

Tutorial 3
Bayesian data analysis

Types of measurement uncertainties

- Generally two major types of uncertainties
    - random uncertainty – different for each measurement of same quantity
        - in repeated measurements, get a different answer each time
        - often assumed to be statistically independent, but aren't always
    - systematic uncertainty – same for each measurement within a group
        - component of measurements that remains unchanged
        - for example, caused by error in calibration or zeroing
        - this kind of uncertainty needs more attention
- Nomenclature varies
    - physics – random uncertainty and systematic uncertainty
    - statistics – random and bias
        - metrology standards (NIST, ASME, ISO) –
          random and systematic uncertainties (now)
    - trend toward quoting **standard error**

Measurement uncertainties in cross sections

In cross-section experiments, sources of uncertainties include:

- Random uncertainties
  - counting statistics for primary process and monitoring process
  - background
- Systematic uncertainties
  - integrated beam intensity
  - target thickness, target impurities
  - detector efficiency
  - count rate corrections
  - geometry
  - corrections for contamination from other processes
- Try to reduce systematic uncertainties through calibration, design
- Random uncertainties usually easy to assess;
  systematic uncertainties require judgment

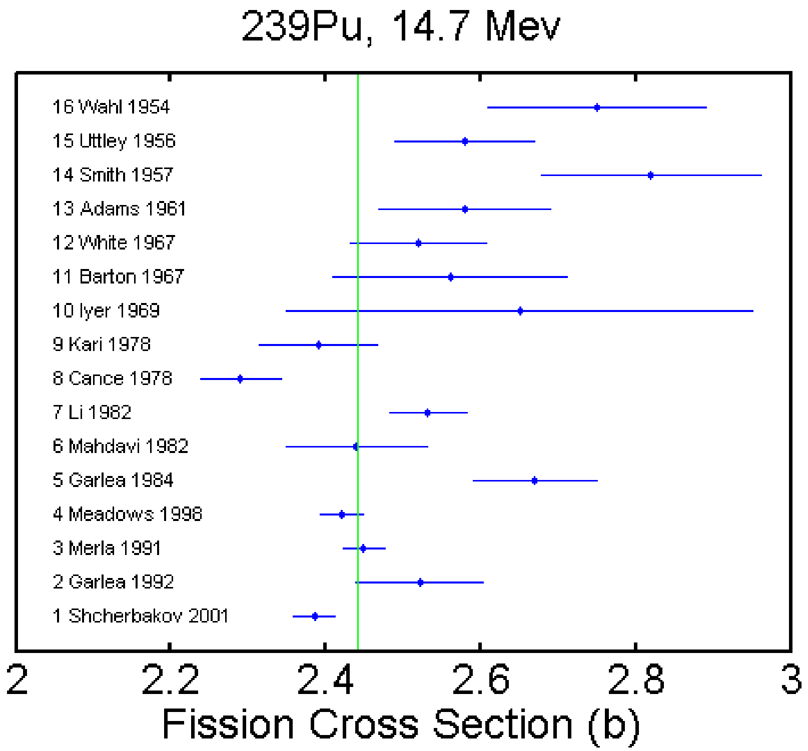
Characterization of measurement uncertainties

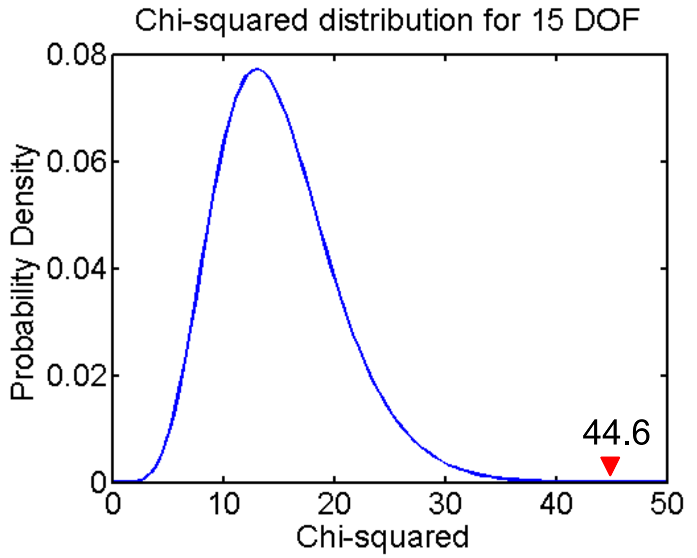
- The best analysis is based on a thorough understanding of probabilistic nature of the fluctuations in the data
- In nuclear physics we are fortunate to have control over measurements; we can calibrate and study apparatus
- Look closely at measurements to characterize random fluctuations
  - shape of pdf
  - standard deviation (variance) of fluctuations,
  - presence of outliers
  - covariance, correlation: $\mathrm{cov}(\boldsymbol{d}) \equiv \mathbf{C}_{\boldsymbol{d}} = \left\langle (\boldsymbol{d}-\hat{\boldsymbol{d}})(\boldsymbol{d}-\hat{\boldsymbol{d}})^{\mathrm{T}} \right\rangle$
  - usually need to assume stationarity, same characteristics everywhere
  - autocorrelation function useful for estimating correlations

$$\rho(l) = \frac{1}{N}\sum_{i=1}^{N} y(i)\,y(i-l)$$

Neutron fission cross section data for ^{239}Pu

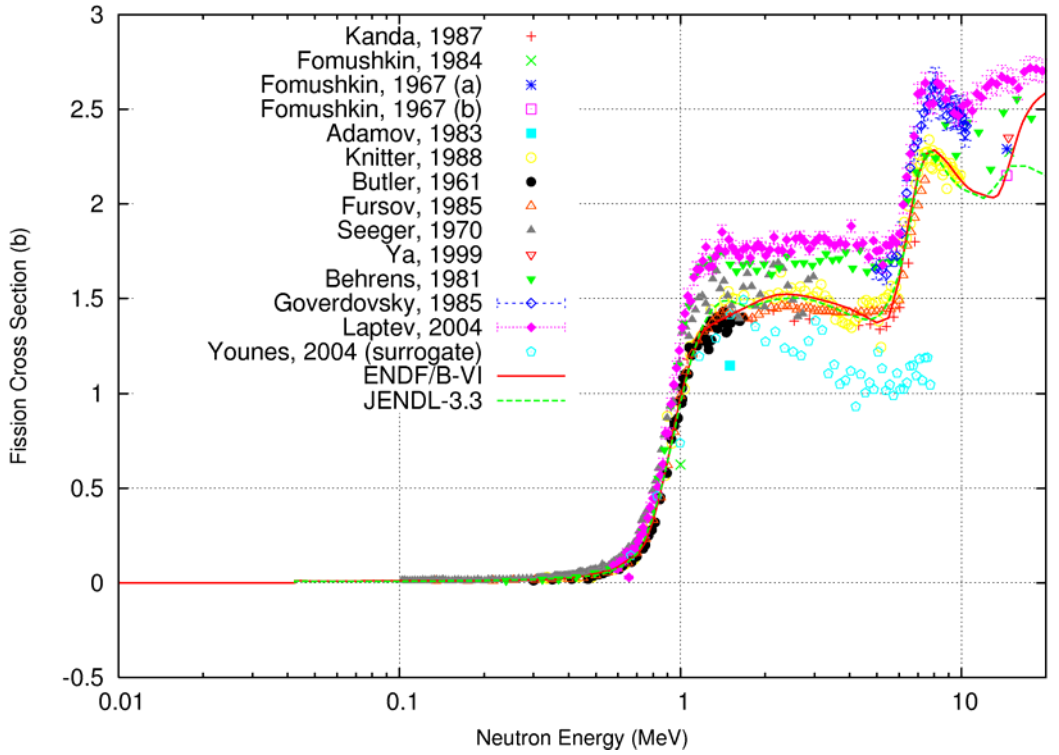
- Graph shows 16 measurements of fission cross-section for ^{239}Pu at 14.7 MeV
- Data exhibit fair amount of scatter
- Quoted error bars get smaller with time
- Minimum $\chi^2 = 44.6$ ($p = 10^{-4}$) indicates a problem
  - dispersion of data larger than quoted error bars by factor $\sqrt{3}$
  - outliers?; three data contribute 24 to χ^2, more than half





Neutron fission cross-section data

^{243}Am fission cross section



plot from P. Talou

- Neutron cross sections measured by many experimenters
  - sometimes data sets differ significantly
  - often little information about uncertainties, esp. systematic errors
  - many directly measure ratios of cross sections, e.g., ^{243}Am/ ^{235}U
  - thorough analysis must take into account all discrepancies

Inference using Bayes rule

- We wish to infer the parameters $\boldsymbol{a}$ of a model M, based on data $\boldsymbol{d}$
- Use Bayes rule, which gives the *posterior*:

$$p(\boldsymbol{a} \,|\, \boldsymbol{d}, M, I) \propto p(\boldsymbol{d} \,|\, \boldsymbol{a}, M, I)\; p(\boldsymbol{a} \,|\, M, I)$$

  - where I represents general information we have about the situation
  - $p(\boldsymbol{d} \,|\, \boldsymbol{a}, M, I)$ is the *likelihood*, the probability of the observed data, given the parameters, model, and general info
  - $p(\boldsymbol{a} \,|\, M, I)$ is the *prior*, which represents what we know about the parameters exclusive of the data
- Note that inference requires specification of the prior

Likelihood

- Form of the likelihood $p(\boldsymbol{d} \,|\, \boldsymbol{a}, I)$ based on how we model the uncertainties in the measurements $\boldsymbol{d}$
- Choose pdf that appropriately describes uncertainties in data
  - Gaussian – good generic choice
  - Poisson – counting experiments
  - Binomial – binary measurements (coin toss …)
- Outliers exist
  - likelihood should have a long tail, i.e., there is some probability of large fluctuation
- Systematic errors
  - caused by effects common to many (all) measurements
  - model by introducing variable that affects many (all) measurements; marginalize out

The model and parameter inference

- We write the model as

$$\boldsymbol{y} = \boldsymbol{y}(\boldsymbol{x}, \boldsymbol{a})$$

  - ► where $\boldsymbol{y}$ is a vector of physical quantities, which is modeled as a function of the independent variables vector $\boldsymbol{x}$ and $\boldsymbol{a}$ represents the parameter vector for the model

- In inference, the aim is to determine:
  - ► the parameters $\boldsymbol{a}$ from a set of n measurements d_i of $\boldsymbol{y}$ under specified conditions x_i
  - ► **and** the uncertainties in the parameter values

- This process is called parameter inference, model fitting (or regression); however, uncertainty analysis is often not done, only parameters estimated

The likelihood and chi-squared

- The form of the likelihood $p(\boldsymbol{d} \,|\, \boldsymbol{a}, I)$ depends on how we model the uncertainties in the measurements $\boldsymbol{d}$
- Assuming the error in each measurement d_i is normally (Gaussian) distributed with zero mean and variance σ_i^2, and that the errors are statistically independent,

$$p(\boldsymbol{d} \,|\, \boldsymbol{a}) \propto \prod_i \exp\left[-\frac{[d_i - y_i(a)]^2}{2\sigma_i^2} \right]$$

- where y_i is the value predicted for parameter set $\boldsymbol{a}$
- The above exponent is one-half chi squared

$$\chi^2 = -2\log\left[p(\boldsymbol{d} \,|\, \boldsymbol{a}) \right] = \sum_i \left[\frac{[d_i - y_i(a)]^2}{\sigma_i^2} \right]$$

- For this error model, likelihood is $p(\boldsymbol{d} \,|\, \boldsymbol{a}) \propto \exp(-\tfrac{1}{2}\chi^2)$

Likelihood analysis

- For a non-informative **uniform prior**, the posterior is proportional to the likelihood
- Given the relationship between chi-squared and the likelihood, the posterior is

$$p(\boldsymbol{a} \mid \boldsymbol{d}) \propto p(\boldsymbol{d} \mid \boldsymbol{a}) \propto \exp(-\tfrac{1}{2}\chi^2)$$

- Parameter estimation based on **maximum likelihood** is equivalent to that based on **minimum chi squared** (or **least squares**)

Likelihood analysis – chi squared

- When the errors in each measurement are Gaussian distributed and independent, likelihood is related to chi squared:

$$p(\boldsymbol{d}\,|\,\boldsymbol{a}) \propto \exp(-\tfrac{1}{2}\chi^2) = \exp\left\{-\tfrac{1}{2}\sum_i\left[\frac{[d_i - y_i(\boldsymbol{a})]^2}{\sigma_i^2}\right]\right\}$$

- near minimum, χ^2 is approximately quadratic in the parameters $\boldsymbol{a}$

$$\chi^2(\boldsymbol{a}) = \tfrac{1}{2}\left(\boldsymbol{a}-\hat{\boldsymbol{a}}\right)^{\mathrm{T}}\boldsymbol{K}\left(\boldsymbol{a}-\hat{\boldsymbol{a}}\right) + \chi^2(\hat{\boldsymbol{a}})$$

  - where $\hat{\boldsymbol{a}}$ is the parameter vector at minimum χ^2 and $\boldsymbol{K}$ is the χ^2 curvature matrix (aka the *Hessian*)

- The covariance matrix for the uncertainties in the estimated parameters is

$$\mathrm{cov}(\boldsymbol{a}) \equiv \left\langle(\boldsymbol{a}-\hat{\boldsymbol{a}})(\boldsymbol{a}-\hat{\boldsymbol{a}})^{\mathrm{T}}\right\rangle \equiv \boldsymbol{C} = 2\boldsymbol{K}^{-1}$$

Characterization of chi squared

- Expand vector $\boldsymbol{y}$ around $\boldsymbol{y}^0$, and approximate:

$$y_i = y_i(x_i, \boldsymbol{a}) = y_i^0 + \sum_j \left.\frac{\partial y_i}{\partial a_j}\right|_{a^0} (a_j - a_j^0) + \cdots$$

- The derivative matrix is called the *Jacobian*, $\boldsymbol{J}$
- Estimated parameters $\hat{\boldsymbol{a}}$ minimize χ^2 (MAP estimate)
- As a function of $\boldsymbol{a}$, χ^2 is approximately quadratic in $\boldsymbol{a} - \hat{\boldsymbol{a}}$

$$\chi^2(\boldsymbol{a}) = \tfrac{1}{2}\left(\boldsymbol{a} - \hat{\boldsymbol{a}}\right)^{\mathrm{T}} \boldsymbol{K} \left(\boldsymbol{a} - \hat{\boldsymbol{a}}\right) + \chi^2(\hat{\boldsymbol{a}})$$

  - where $\boldsymbol{K}$ is the χ^2 curvature matrix (aka the *Hessian*);

$$\left[\boldsymbol{K}\right]_{jk} = \left.\frac{\partial^2 \chi^2}{\partial a_j \partial a_k}\right|_{\hat{a}} \; ; \quad \boldsymbol{K} = 2\boldsymbol{J}\boldsymbol{\Lambda}\boldsymbol{J}^{\mathrm{T}} \; ; \quad \boldsymbol{\Lambda} = \mathrm{diag}(\sigma_1^{-2}, \sigma_2^{-2}, \sigma_3^{-2}, \ldots)$$

- Jacobian useful for finding min. χ^2 , i.e., optimization

Multiple data sets and Gaussian prior

- Analysis of multiple data sets
  - to combine the data from multiple, independent data sets into a single analysis, the combined chi squared is

$$\chi^2_{all} = \sum_k \chi^2_k$$

  - where $p(\boldsymbol{d}_k \mid \boldsymbol{a}, I)$ is the likelihood from kth data set
- Include Gaussian priors through Bayes theorem

$$p(\boldsymbol{a} \mid \boldsymbol{d}, I) \propto p(\boldsymbol{d} \mid \boldsymbol{a}, I)\; p(\boldsymbol{a} \mid I)$$

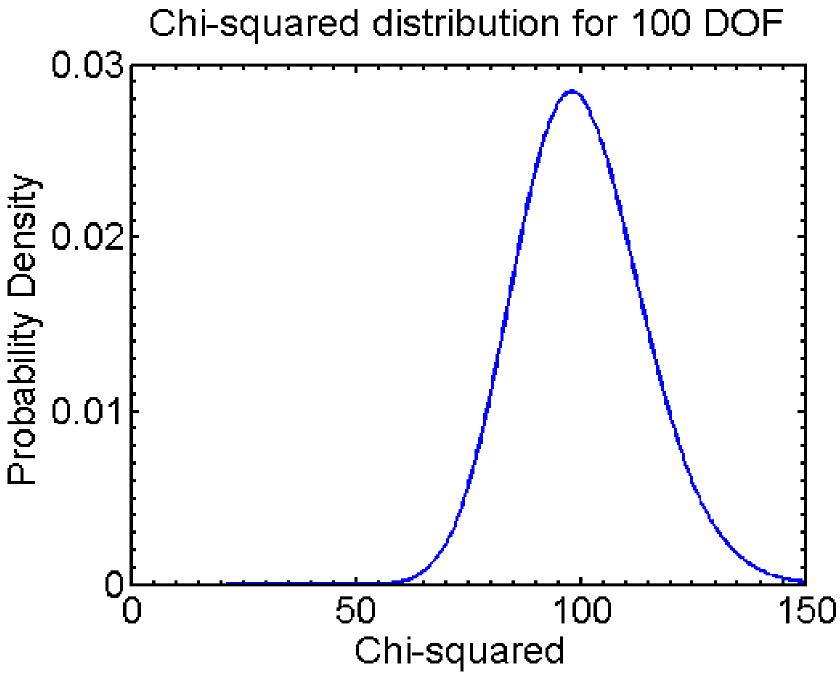
  - for a Gaussian prior on a parameter a_j

$$-\log p(\boldsymbol{a} \mid \boldsymbol{d}, I) = \varphi(\boldsymbol{a}) = \tfrac{1}{2}\chi^2 + \frac{\left(a_j - \tilde{a}_j\right)^2}{2\sigma_j^2}$$

  - where $\tilde{a}_j$ is the default value for a_j and σ_j^2 is assumed variance

Chi-squared distribution

- Plot shows χ^2 distribution for number of degrees of freedom, $\nu = 100$
- Generally,
  - mean = ν
  - rms dev = $\sqrt{2/\nu}$
- Cumulative distribution gives *p* value, probability of $\chi^2 \geq$ observed value
- *p* often used a measure of goodness of fit
- Checks self-consistency of models used to explain data (weakly)



Goodness of fit

- Check of minimum chi-squared value only weakly confirms validity of models used
- Chi-squared value depends on numerous factors:
  - assumption that errors follow Gaussian distribution and are statistically independent
  - proper assignment of standard deviation of errors
  - correctness of model used to calculate measured quantity
  - measurements correspond to calculated quantity (proper measurement model)
- Thus, a reasonable chi-squared *p* value does not necessarily mean everything is OK, because there may be compensating effects

Fit linear function to data – minimum χ^2

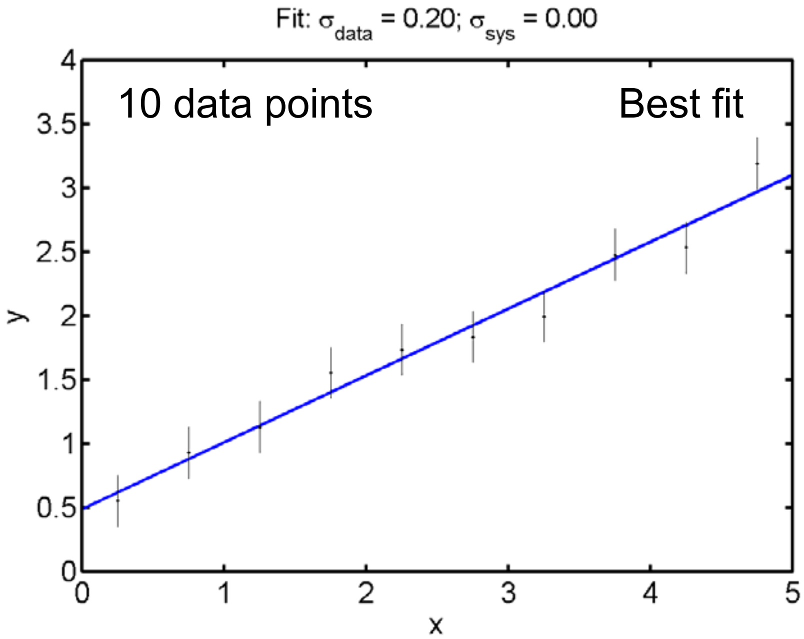
- Linear model: $y = a + bx$
- Simulate 10 data points, $\sigma_y = 0.2$ exact values: $a = 0.5 \quad b = 0.5$
- Determine parameters, intercept a and slope b, by minimizing chi-squared (standard least-squares analysis)
- Result: $\chi^2_{\min} = 4.04 \quad p = 0.775$

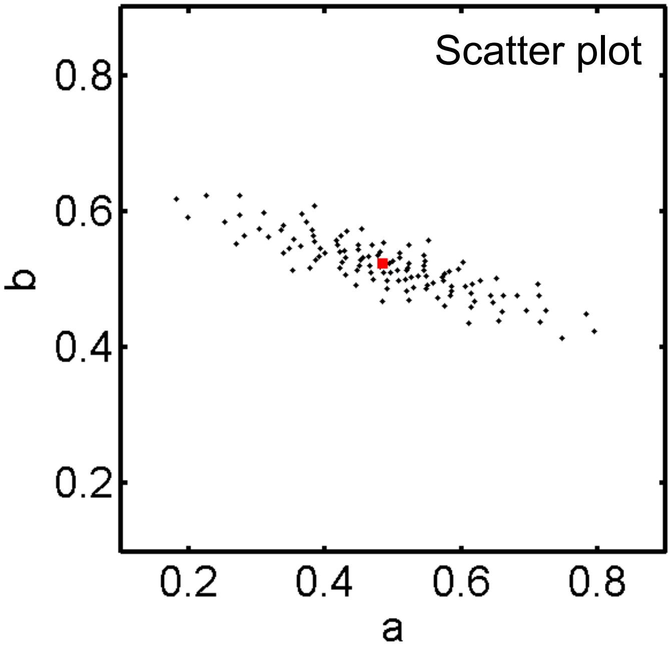
$$\hat{a} = 0.484 \quad \sigma_a = 0.127$$

$$\hat{b} = 0.523 \quad \sigma_b = 0.044$$

$$\mathbf{R} = \begin{bmatrix} 1 & -0.867 \\ -0.867 & 1 \end{bmatrix}$$

- Strong correlations between parameters a and b





Sampling from correlated normal distribution

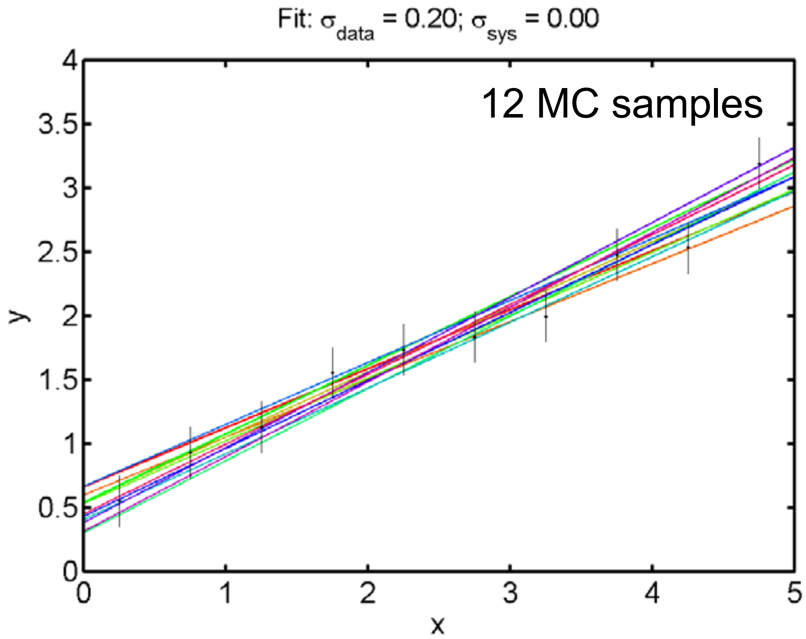
- Want to draw samples **x** from multi-variate normal distribution with known covariance $\mathbf{C_x}$
- Important to include correlations among uncertainties, i.e., off-diagonal elements
- Algorithm:
  - ► perform eigenanalysis of covariance matrix of *d* dimensions

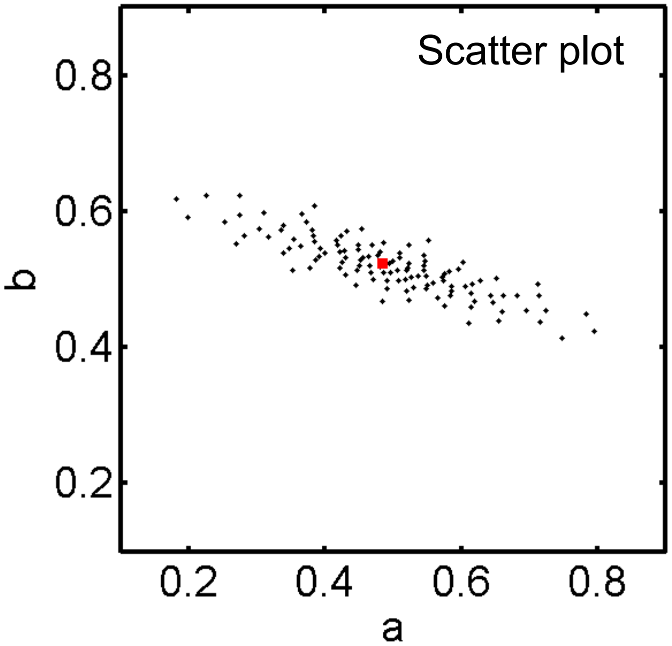
    $$\mathbf{C_x} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^{\mathrm{T}}$$

    where **U** is orthogonal matrix of eigenvectors and
    Λ is the diagonal matrix of eigenvalues
  - ► draw *d* samples from unit variance normal distribution, ξ_i
  - ► scale this vector by $\lambda_{\mathrm{i}}^{½}$
  - ► transform vector into parameter space using the eigenvector matrix
  - ► to summarize: $\mathbf{x} = \mathbf{U}\mathbf{\Lambda}^{1/2}\boldsymbol{\xi}$

Linear fit – uncertainty visualization

- Uncertainties in parameters are represented by Gaussian pdf in 2-D parameter space
  - correlations evidenced by tilt in scatter plot
  - points are samples from pdf
- Should focus on implied uncertainties in physical domain
  - model realizations drawn from parameter uncertainty pdf
  - these appear plausible – called **model checking**
  - this comparison to the original data confirms model adequacy
  - called **predictive distribution**



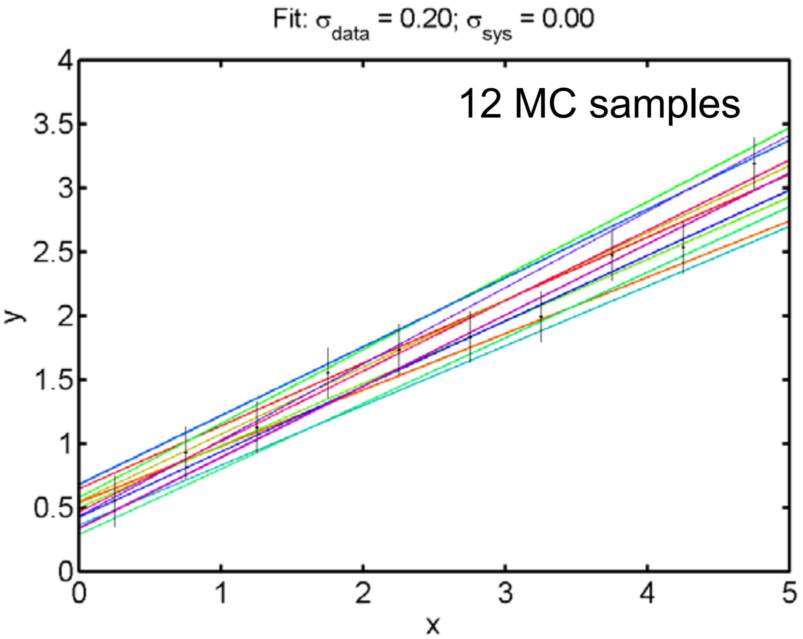


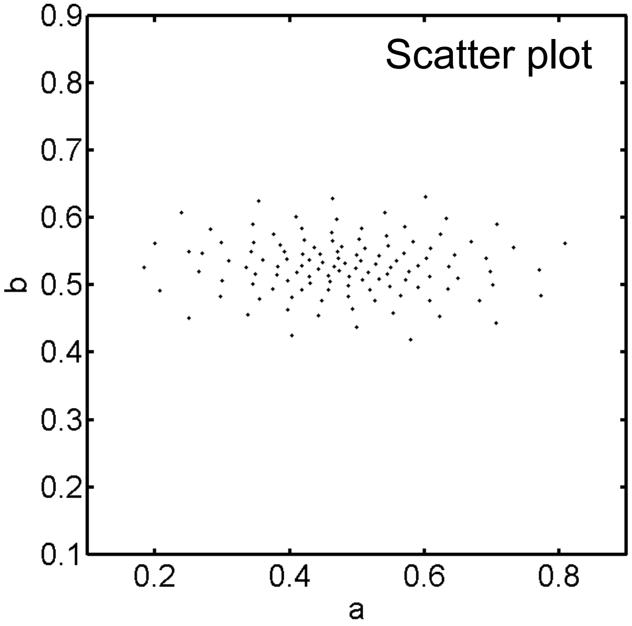
Linear fit – correlations are important

- Plots show what happens if off-diagonal terms of covariance matrix are ignored
- Correlation matrix is

$$\mathbf{R} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

- Model realizations show much wider dispersion than consistent with uncertainties in data
- No tilt in scatter plot – uncorrelated
- Correlations are important !





Probabilistic model for additive error

- Represent systematic additive uncertainty in measurements by common additive offset Δ: $y_i = a + bx_i + \varepsilon_i + \Delta = f(x_i; a,b) + \varepsilon_i + \Delta$
  - where the ε_i represent the random fluctuations
- Bayes law gives joint pdf for all the parameters

$$p(a,b,\Delta \,|\, \boldsymbol{y}, \boldsymbol{x}) = p(\boldsymbol{y} \,|\, a,b,\Delta,\boldsymbol{x})\, p(a)\, p(b)\, p(\Delta)$$

where priors $p(a)$, $p(b)$ are uniform and $p(\Delta)$ assumed normal

- Writing $p(a,b,\Delta \,|\, \boldsymbol{y}, \boldsymbol{x}) \propto \exp\{-\varphi\}$ and assuming normal distributions

$$2\varphi = \sum \frac{\left(y_i \text{ - } f(x_i; a,b) - \Delta\right)^2}{\sigma_i^2} + \frac{\Delta^2}{\sigma_\Delta^2}$$

- Pdf for x obtained by integration: $p(a,b \,|\, \boldsymbol{y}, \boldsymbol{x}) = \int p(a,b,\Delta \,|\, \boldsymbol{y}, \boldsymbol{x})\, \mathrm{d}\Delta$

- This model equivalent to standard least-squares approach by including Δ in fit, and using just results for a and b

Linear fit – systematic uncertainty

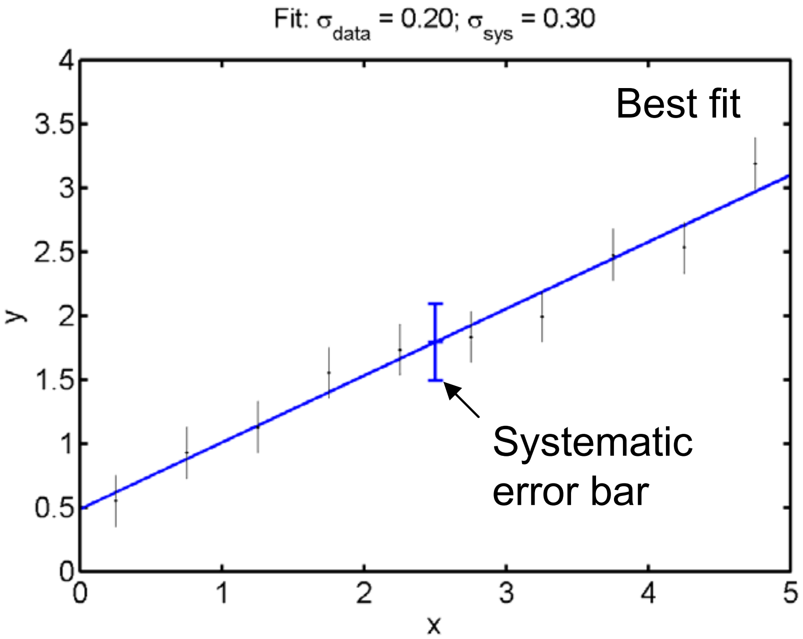
- Introduce systematic offset Δ with uncertainty $\sigma_\Delta = 0.3$
- Linear model: $y = a + bx + \Delta$
- Determine parameters, *a*, *b*, and offset Δ by minimizing chi-squared (standard least-squares analysis)
- Result: $\hat{\Delta} = 0$

$$\hat{a} = 0.484 \qquad \sigma_a = 0.326$$

$$\hat{b} = 0.523 \qquad \sigma_b = 0.044$$

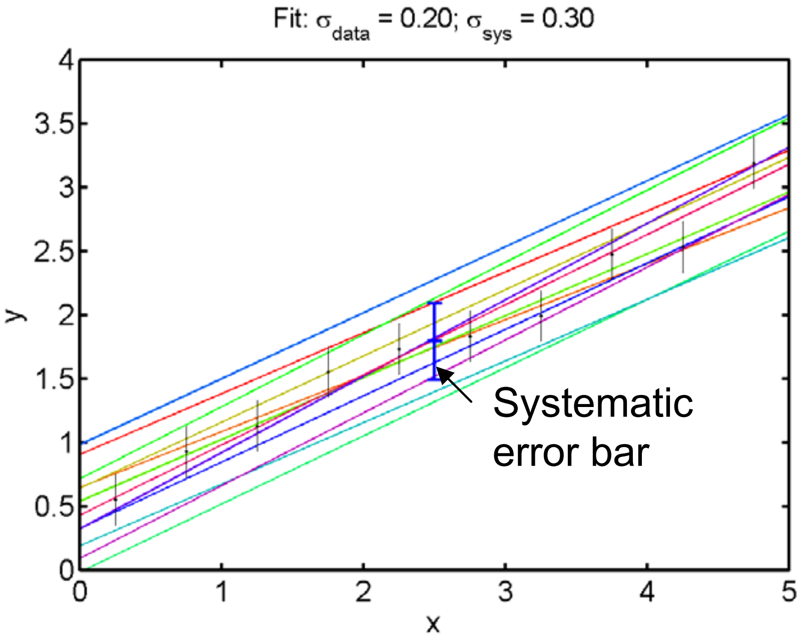
$$\mathbf{R} = \begin{bmatrix} 1 & -0.338 \\ -0.338 & 1 \end{bmatrix}$$

- Same parameters, but σ_a much larger



Linear fit – systematic uncertainty

- Show uncertainties in inferred models
  - colored lines are model realizations drawn from parameter uncertainty pdf
  - these appear plausible, considering additional systematic uncertainty, $\sigma_{\Delta} = 0.3$



Role of simulated data

- Simulated data are crucially important for testing algorithms
  - treat simulated data as is actual measurements
  - can compare algorithmic results with known true values
  - can test how well algorithm copes with specific data deficiencies
  - aid in debugging computer code, underlying ideas
- Important to mimic real data
  - characteristics of measurement fluctuations (noise)
  - limited resolution (blur) of signal
  - systematic effects

Linear fit to many data

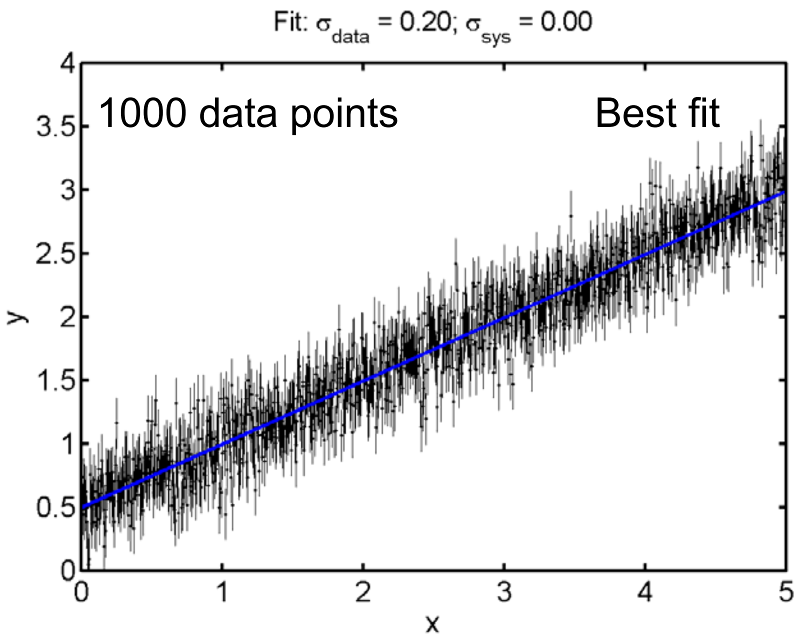
- Linear model: $y = a + bx$
- Simulate 1000 data points, $\sigma_y = 0.2$
  exact values: $a = 0.5 \quad b = 0.5$
- Determine parameters by minimizing chi-squared
- Result: $\chi^2_{\min} = 972.0 \quad p = 0.717$

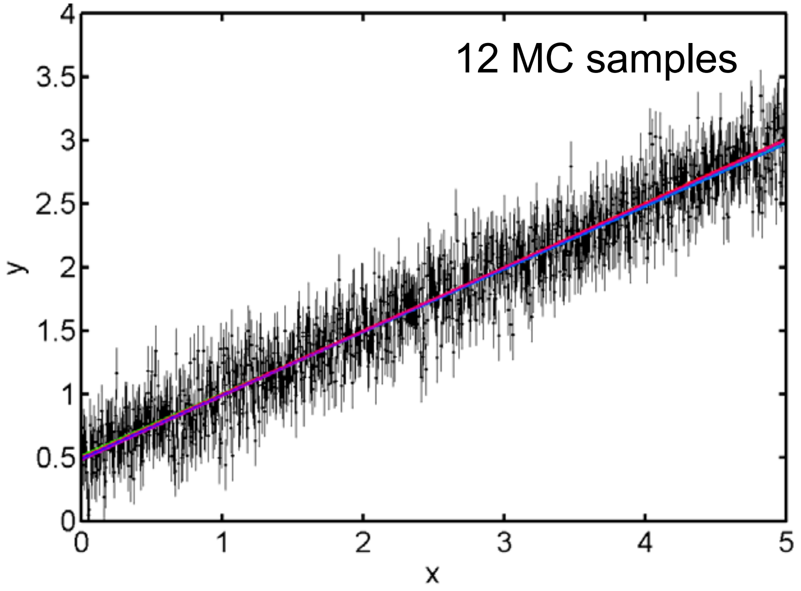
$$\hat{a} = 0.496 \quad \sigma_a = 0.0126$$

$$\hat{b} = 0.499 \quad \sigma_b = 0.0044$$

$$\mathbf{R} = \begin{bmatrix} 1 & -0.866 \\ -0.866 & 1 \end{bmatrix}$$

- Standard errors are reduced by factor of 10 through data averaging
- Is this reasonable?





Linear fit to many data - systematic uncertainty

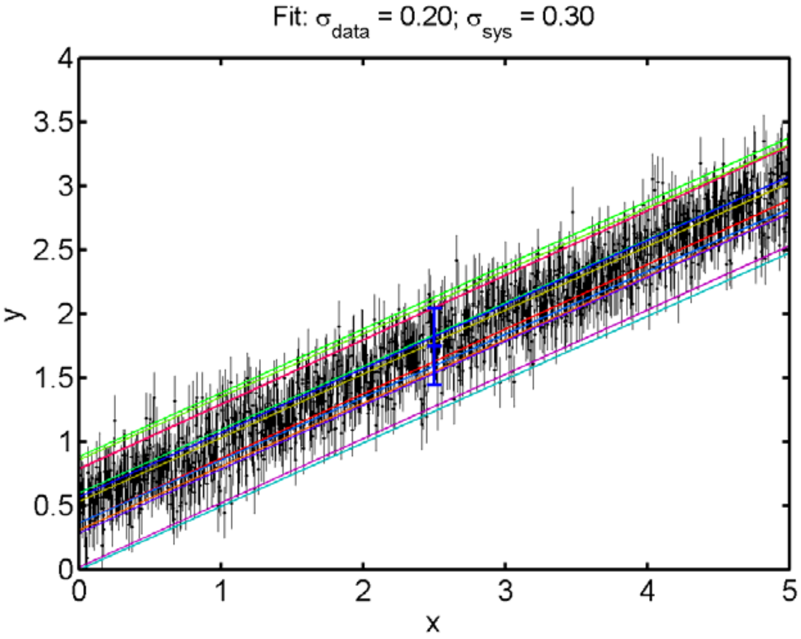
- Introduce systematic offset Δ with uncertainty $\sigma_\Delta = 0.3$
- Linear model: $y = a + bx + \Delta$
- Determine parameters, *a*, *b*, and offset Δ by minimizing chi-squared (standard least-squares analysis)
- Result: $\hat{\Delta} = 0$

$$\hat{a} = 0.496 \qquad \sigma_a = 0.300$$

$$\hat{b} = 0.499 \qquad \sigma_b = 0.0044$$

$$\mathbf{R} = \begin{bmatrix} 1 & -0.036 \\ -0.036 & 1 \end{bmatrix}$$

- Same fit, but σ_a dominated by σ_Δ
- Uncertainty in slope still small

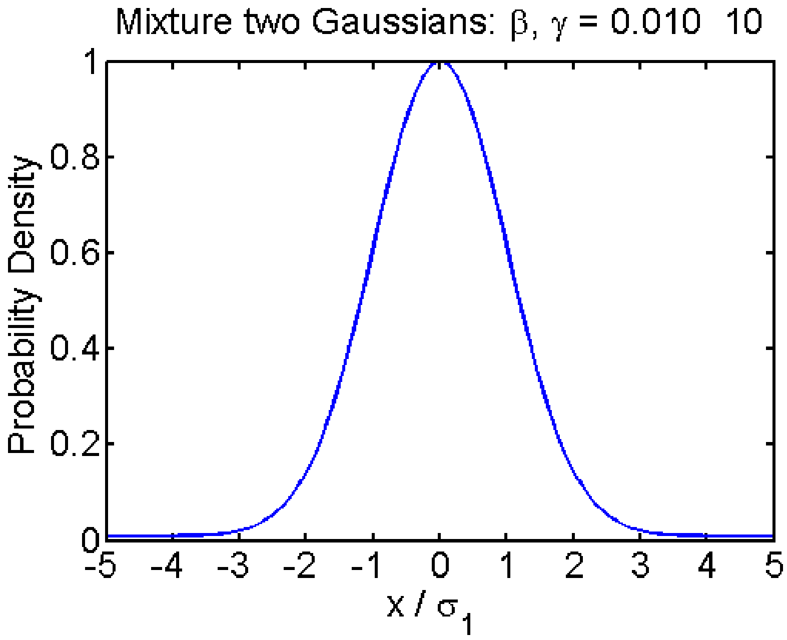


Outliers

- Measurements that differ from true value by more than expected
- Often caused by mistakes
  - every experimenter knows mistakes happen!
- Can accommodate in likelihood function by including long tail
- Simple model: likelihood is mixture of two Gaussians

$$(1-\beta)\exp\left\{-\frac{(x-m)^2}{2\sigma^2}\right\}+\beta\exp\left\{-\frac{(x-m)^2}{2\gamma\sigma^2}\right\}$$

- Long tail includes possibility of large deviation from true value
- Outlier-tolerant analysis generally called “robust estimation”



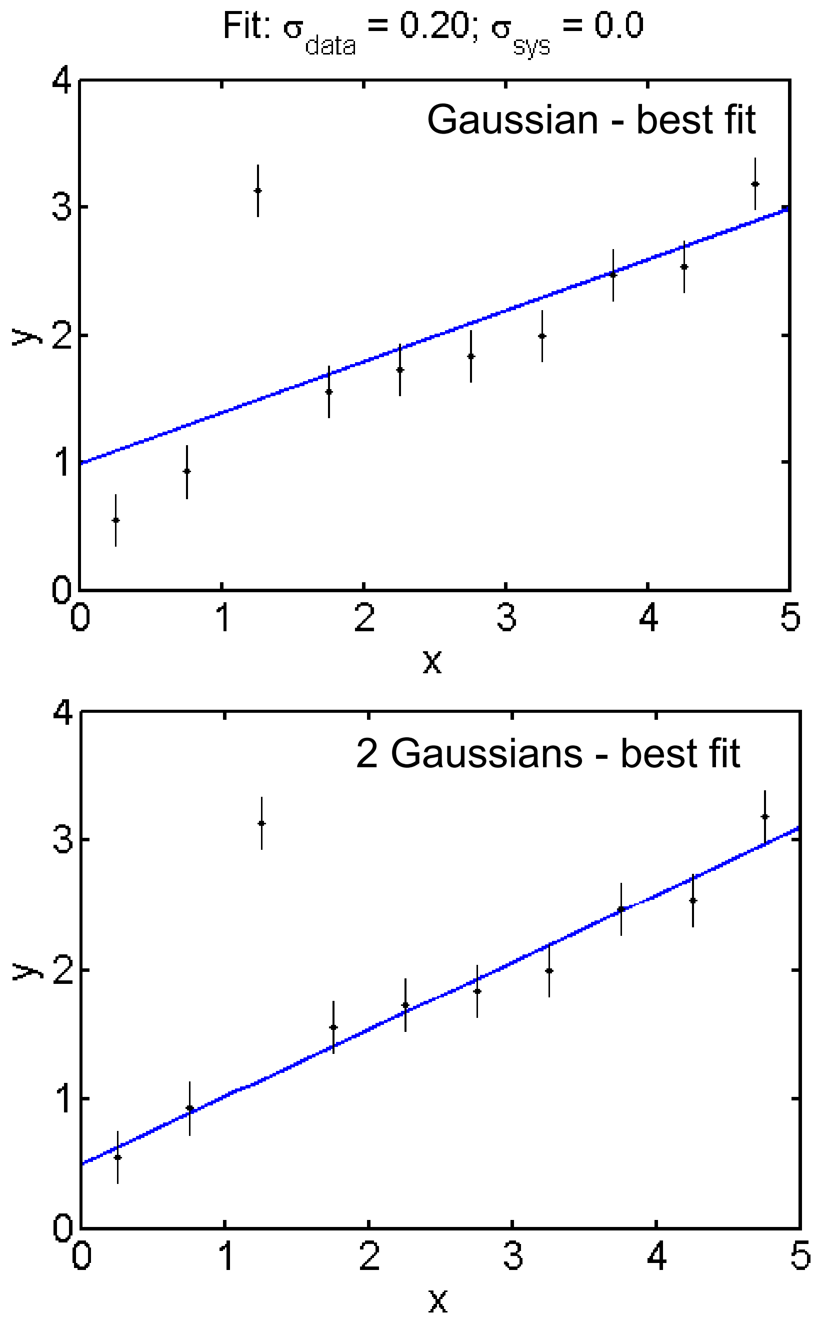
Linear fit – outliers

- Outliers pose significant problem for min χ^2 algorithm
- Create outlier by artificially perturbing third point
- Min-χ^2 results in large shift of fitted line: $\chi^2_{\min} = 85.6 \quad p = 10^{-15}$

$$\hat{a} = 0.987 \qquad \sigma_a = 0.180$$
$$\hat{b} = 0.402 \qquad \sigma_b = 0.062$$

- Two-Gaussian likelihood handles outlier very well
  - fit is nearly the same as before

$$\hat{a} = 0.494 \qquad \sigma_a = 0.140$$
$$\hat{b} = 0.520 \qquad \sigma_b = 0.043$$

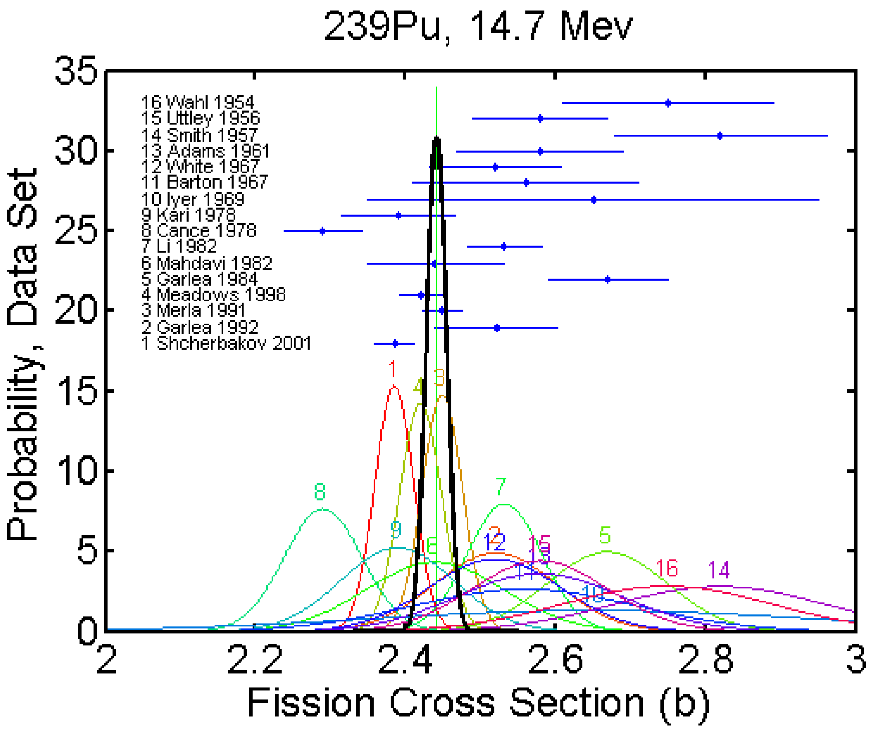


^{239}Pu cross sections – Gaussian likelihood

- With Gaussian likelihood (min χ^2) yields
  - $\chi^2 = 44.7$, $p = 0.009\%$ for 15 DOF
    2.441 ± 0.013
  - implausibly small uncertainty given three smallest uncerts. ≈ 0.027
- Each datum reduces the standard error of result, even if it does not agree with it!
  - consequence of Gaussian likelihood

$$\sigma^{-2} = \sum_{i=1}^{n} \sigma_i^{-2}$$

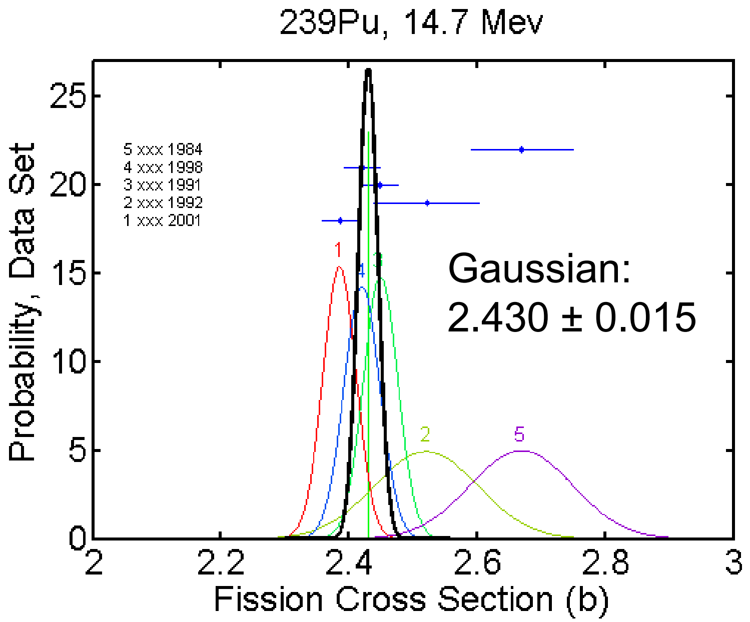
  - independent of where data lie!
    which doesn’t make sense



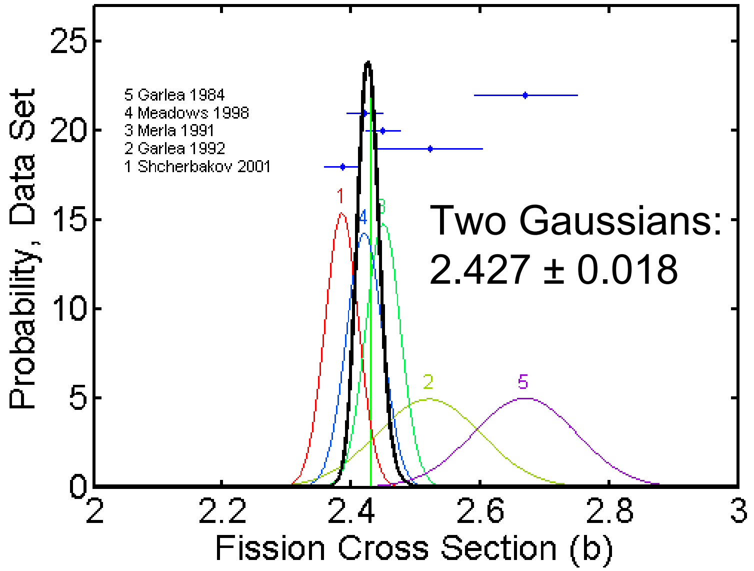
Gaussian: 2.441 ± 0.013

^{239}Pu cross sections – outlier-tolerant likelihood

- Use just latest five measurements
- Compare results from alternative likelihoods:
  - Gaussian: 2.430 ± 0.015
    $\chi^2 = 13.88$, $p = 0.8\%$ for 4 DOF
  - two Gaussians: 2.427 ± 0.018
- For two-Gaussian likelihood:
  - result not pulled as hard by outlier
  - σ is not as small, seemingly taking into account discrepant nature of data



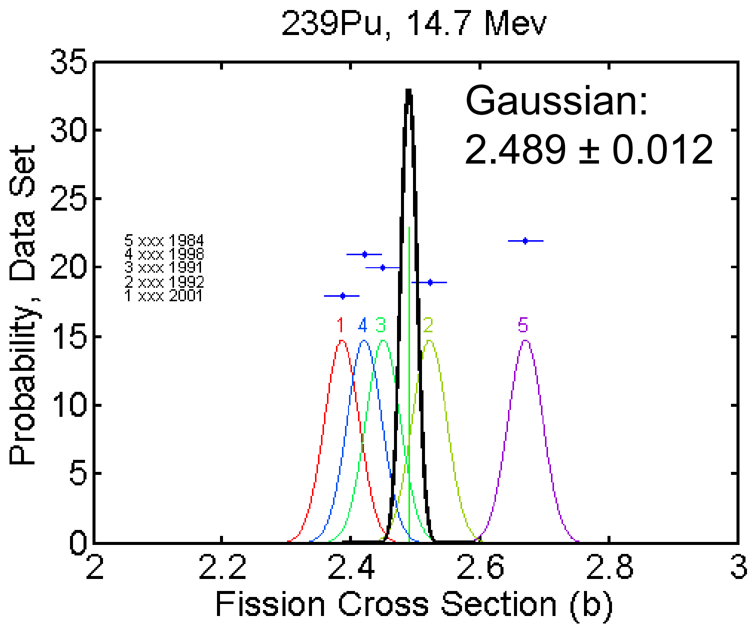
Gaussian:
2.430 ± 0.015

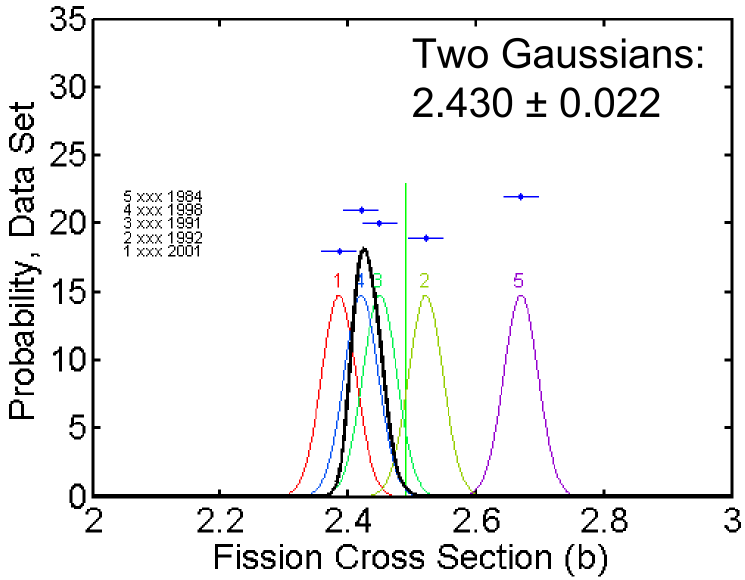


Two Gaussians:
2.427 ± 0.018

^{239}Pu cross sections – outlier-tolerant likelihood

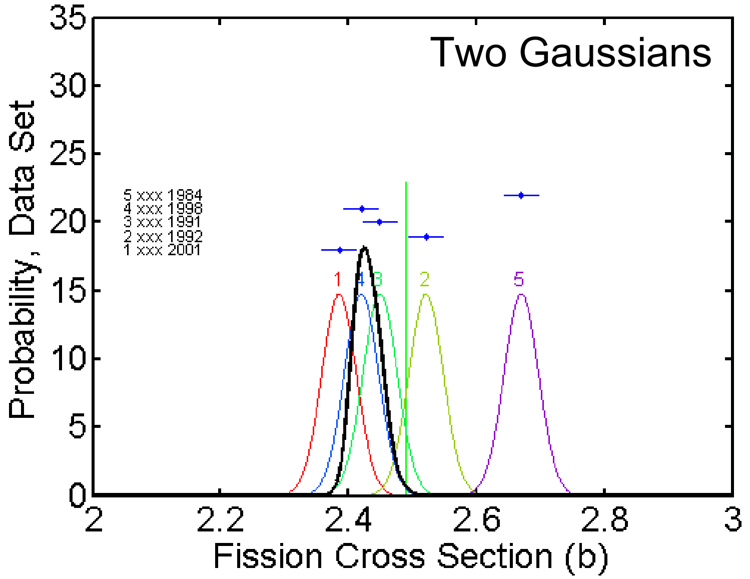
- Use just latest five measurements
- To exaggerate outlier problem, set all standard errors = 0.027
- Compare results from alternative likelihoods:
  - Gaussian: 2.489 ± 0.012
    $\chi^2 = 69.9$, $p = 2\times10^{-14}$ for 4 DOF
  - two Gaussians: 2.430 ± 0.022
- For two-Gaussian likelihood:
  - result is close to cluster of three points; outliers have little effect
  - uncertainty is plausible

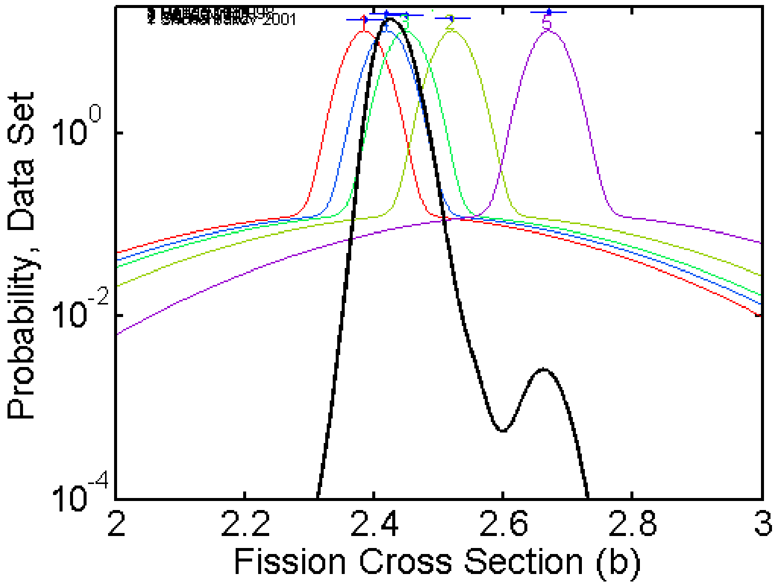




^{239}Pu cross sections – outlier-tolerant likelihood

- To exaggerate outlier problem, set all standard errors = 0.027, using just latest five measurements
- Plot shows pdfs on log scale, which shows what is going on with two-Gaussian likelihood
  - long tail of likelihood function for outlier does not influence peak shape near cluster of three measurements; for single Gaussian, it would make it narrower
  - long tails of likelihood functions from cluster allows outlier to produce a small secondary peak; has little effect on posterior mean





Hierarchical model – scale uncertainties

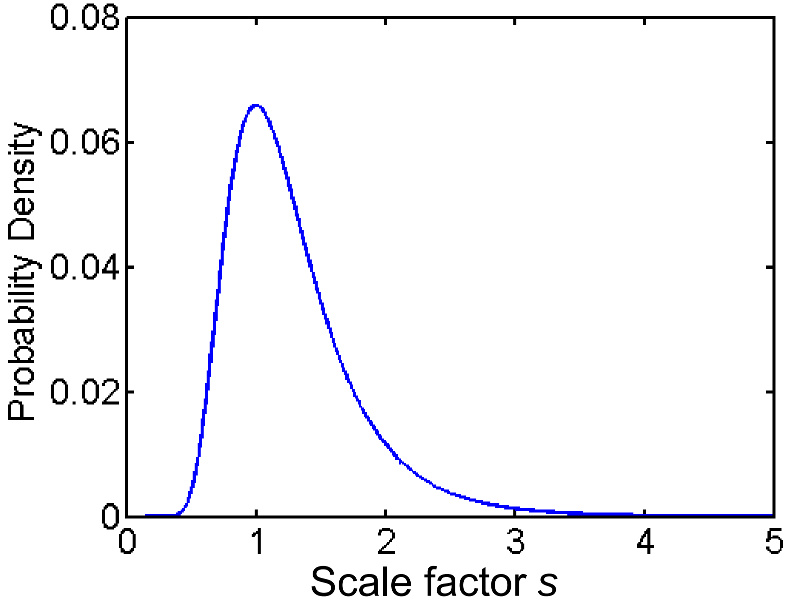
- When data disagree a lot, we may question whether quoted standard errors are correct
- Scale all σ by factor *s*: $\sigma = s\,\sigma_0$
- Then marginalize over *s*

$$p(\boldsymbol{a}\,|\,\boldsymbol{d}) = \int p(\boldsymbol{a}, s\,|\,\boldsymbol{d})\,\mathrm{d}s$$

$$p(\boldsymbol{a}\,|\,\boldsymbol{d}) \propto \int p(\boldsymbol{d}\,|\,\boldsymbol{a}, s)\,p(\boldsymbol{a}, s)\,\mathrm{d}s$$

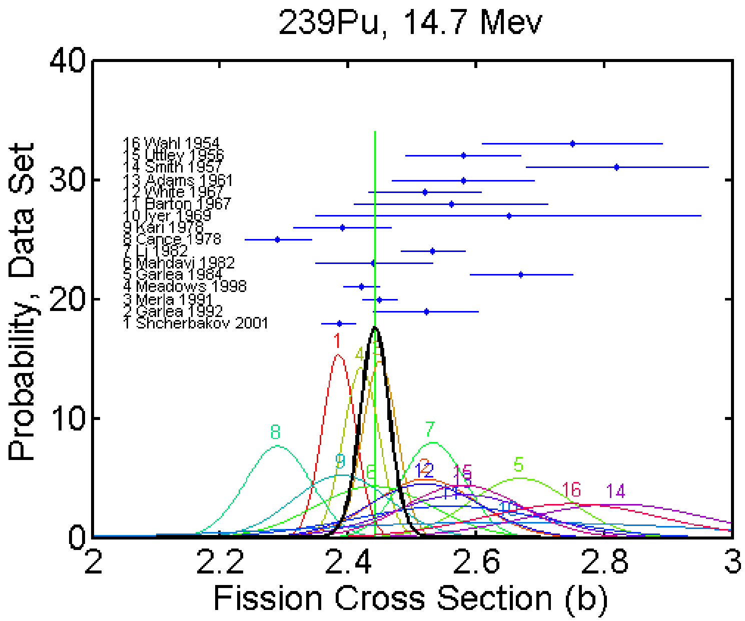
$$p(\boldsymbol{a}\,|\,\boldsymbol{d}) \propto \int p(\boldsymbol{d}\,|\,\boldsymbol{a}, s)\,p(\boldsymbol{a})\,p(s)\,\mathrm{d}s$$

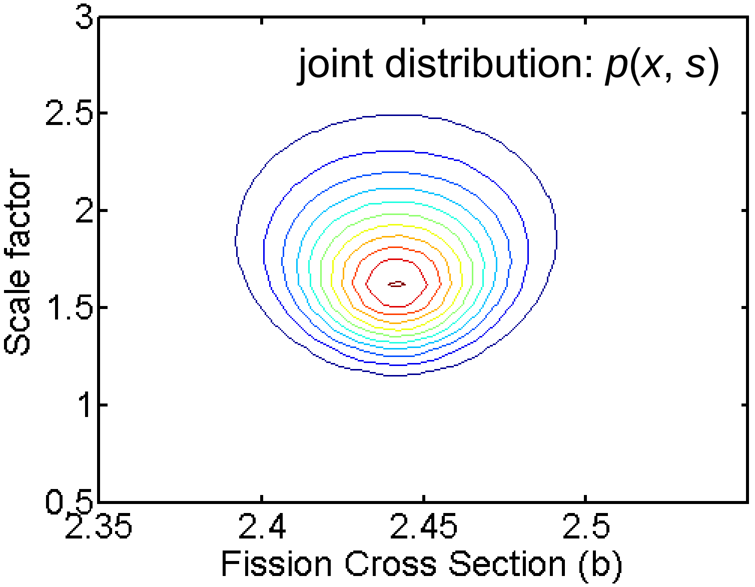
- For prior *p*(*s*), either use noninformative (flat in log(*s*)) or one like shown in plot
- Let the data decide!
- This is called **hierarchical model** because properties of one pdf, the likelihood, are specified by another pdf



^{239}Pu cross sections – scale uncertainties

- Accommodate large dispersion in data by scaling all σ by factor *s*:
  $\sigma = s\,\sigma_0\,;\;\; \sigma_0\;$ = quoted stand. err.
- For likelihood, use Gaussian with scaled σ
  $$p(\boldsymbol{d}\,|\,x,s) \propto \frac{1}{s^n}\exp\left(-\frac{\chi_0^2}{2s^2}\right)$$
- For prior *p*(*s*), use non-informative prior for scaling parameter $p(s) \propto 1/s$
- Bottom plot shows joint posterior pdf
- Marginalize over *s*:
  $$p(x\,|\,\boldsymbol{d}) \propto \int p(\boldsymbol{d}\,|\,x,s)\,p(x)\,p(s)\,\mathrm{d}s$$
  to get posterior for *x* (top plot)
- Result is: 2.441 ± 0.024;
  very plausible uncertainty



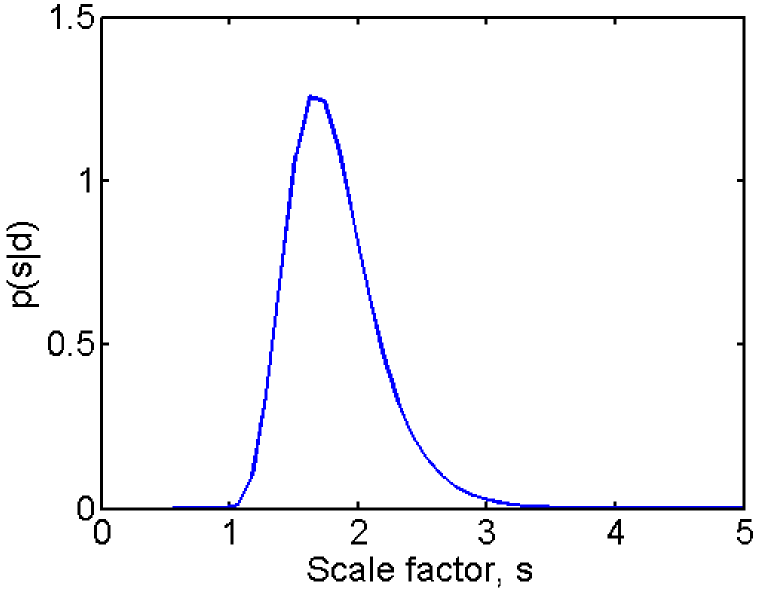


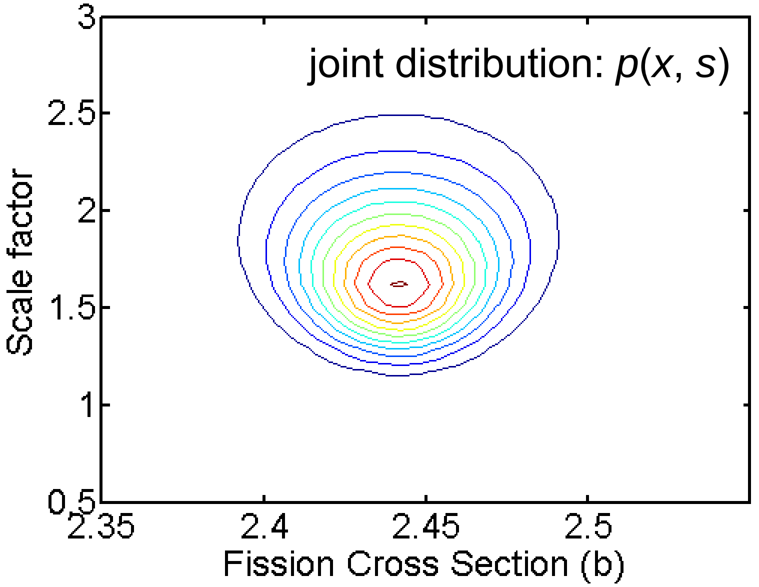
^{239}Pu cross sections – scale uncertainties

- To obtain the posterior for the scaling parameter *s*, marginalize joint posterior over *x*:

$$p(s\,|\,\boldsymbol{d}) \propto \int p(\boldsymbol{d}\,|\,x,s)\,p(x)\,p(s)\,\mathrm{d}x$$

- Plot (top) shows result
  - maximum at about 1.7, $\approx \sqrt{\dfrac{\chi^2}{\mathrm{DOF}}}$ for original fit
  - however, this result is different from just scaling σ to make χ^2 per DOF unity
  - it allows for a distribution in *s*, taking into account that *s* is uncertain
- This model can be extended to allow each σ_i to be scaled separately
  - prior on s_i could reflect our confidence in quoted σ_i for each experiment





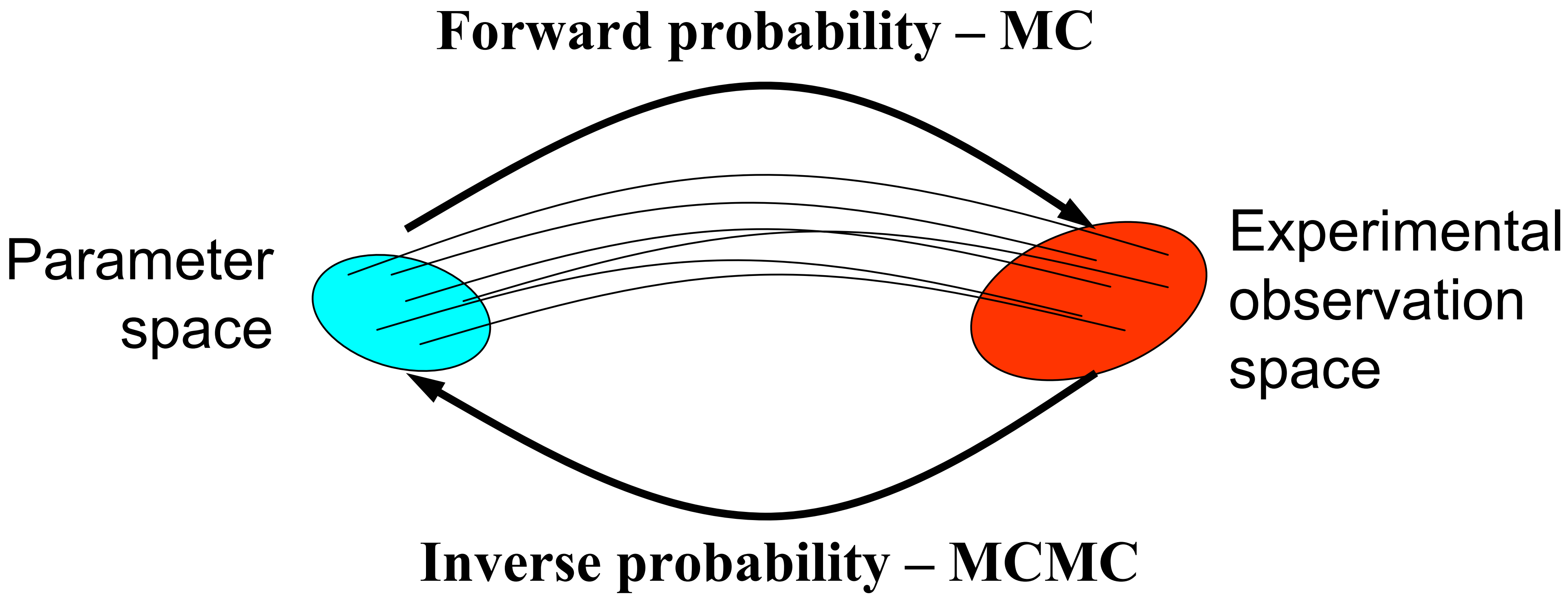
Summary

In this tutorial:

- Types of uncertainties in measurements – random and systematic
- Uniform prior ⇒ likelihood analysis ⇒ χ^2 analysis
- Used straight line fit to illustrate various Bayesian concepts and models
  - posterior sampling; predictive distribution and model checking
  - systematic uncertainties
  - averaging over many measurements
  - outliers
- Studied Pu cross-section data at 14.7 MeV
  - outlier-tolerant likelihood
  - scaling of quoted standard errors using a distribution of scales, which is determined by input data

Tutorial 4
Bayesian calculations

Forward and inverse probability



- Forward probability - determine uncertainties in observables resulting from model parameter uncertainties; use Monte Carlo
- Inverse probability - infer model parameter uncertainties from uncertainties in observables; use Markov chain Monte Carlo

MCMC - problem statement

- Parameter space of *n* dimensions represented by vector **x**
- Given an “arbitrary” **target** probability density function (pdf), $q(\mathbf{x})$, draw a set of samples $\{\mathbf{x}_k\}$ from it
- Only requirement typically is that, given **x**, one be able to evaluate $Cq(\mathbf{x})$, where C is an unknown constant, that is, $q(\mathbf{x})$ need not be normalized
- Although focus here is on continuous variables, MCMC applies to discrete variables as well

- It all started with seminal paper:
  - N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, “Equations of state calculations by fast computing machine,” *J. Chem. Phys.* **21**, pp. 1087–1091 (1953)
    - MANIAC: 5 KB RAM, 100 KHz, 1 KHz multiply, 50 KB disc

Uses of MCMC

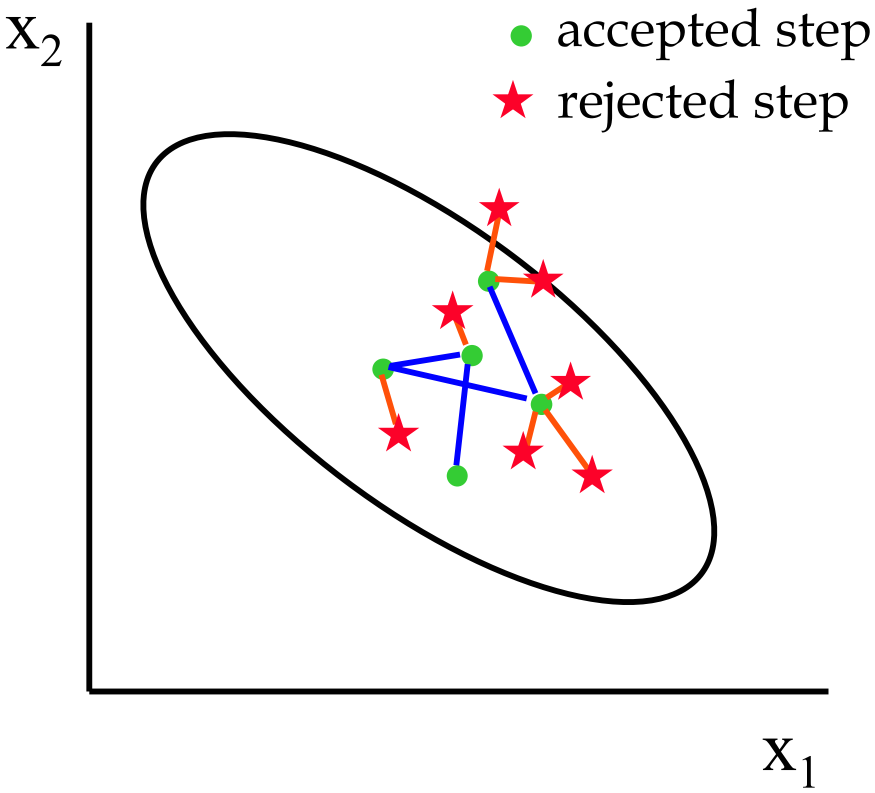
- Permits evaluation of the expectation values of functions of **x**, e.g.,
  $$\langle f(\mathbf{x}) \rangle = \int f(\mathbf{x})\, q(\mathbf{x})\, d\mathbf{x} \cong (1/\mathrm{K})\, \Sigma_{\mathrm{k}}\, f(\mathbf{x}_{\mathrm{k}})$$
  - typical use is to calculate mean $\langle \mathbf{x} \rangle$ and variance $\langle (\mathbf{x} - \langle \mathbf{x} \rangle)^2 \rangle$
- Useful for evaluating integrals, such as the partition function for properly normalizing the pdf
- Dynamic display of sequences provides visualization of uncertainties in model and range of model variations
- Automatic marginalization; when considering any subset of parameters of an MCMC sequence, the remaining parameters are marginalized over (integrated out)

Markov Chain Monte Carlo

Generates sequence of random samples from an arbitrary probability density function

- Metropolis algorithm:
  - draw trial step from symmetric pdf, i.e., $t(\Delta\, \mathbf{x}) = t(-\Delta\, \mathbf{x})$
  - accept or reject trial step
  - simple and generally applicable
  - relies only on calculation of target pdf for any **x**

Probability(x_1, x_2) = q(**x**)



Metropolis algorithm

- Target pdf is $q(\mathbf{x})$
- Select initial parameter vector $\mathbf{x}_0$
- Iterate as follows: at iteration number k

  (1) create new trial position $\mathbf{x}^* = \mathbf{x}_k + \Delta\mathbf{x}$,

  where $\Delta\mathbf{x}$ is randomly chosen from $t(\Delta\mathbf{x})$

  (2) calculate ratio $r = q(\mathbf{x}^*)/q(\mathbf{x}_k)$

  (3) accept trial position, i.e. set $\mathbf{x}_{k+1} = \mathbf{x}^*$

  if $r \geq 1$ or with probability r, if $r < 1$

  otherwise stay put, $\mathbf{x}_{k+1} = \mathbf{x}_k$

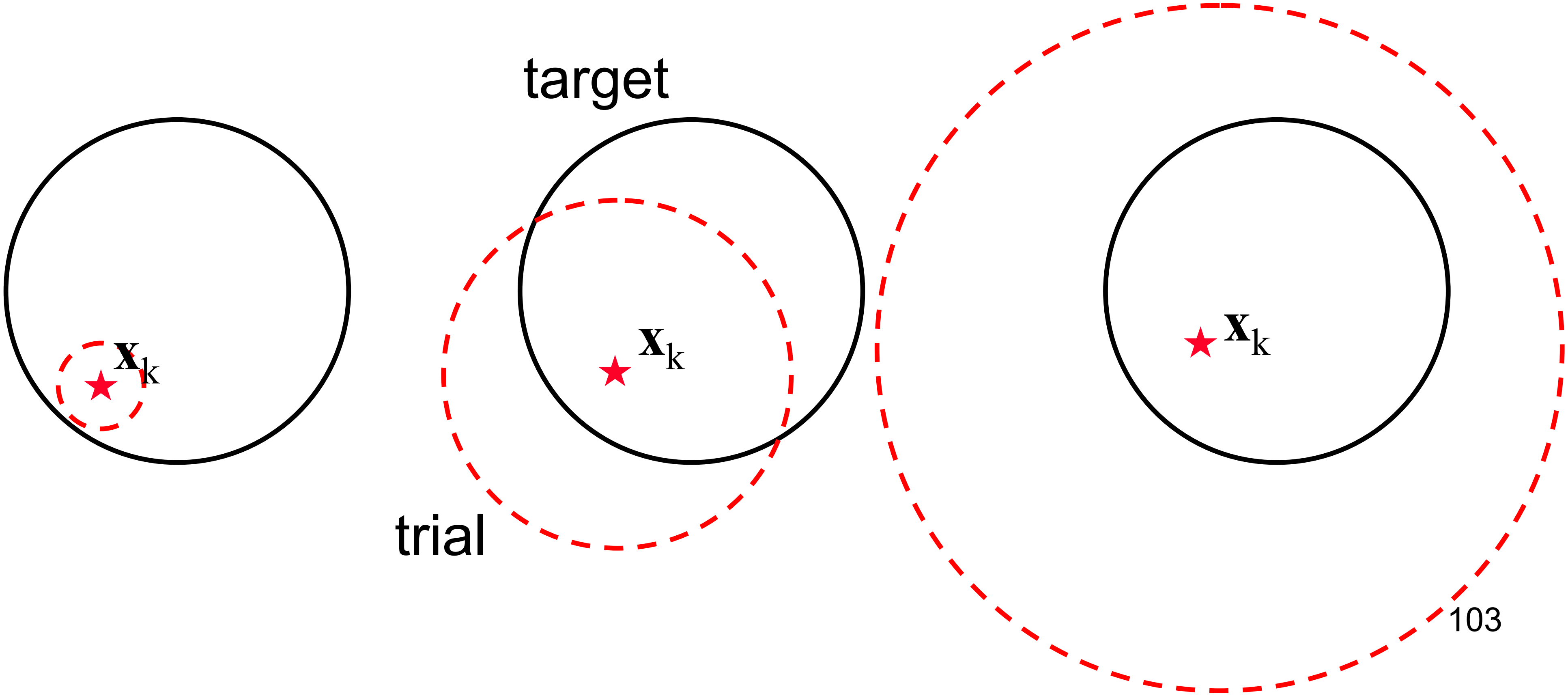
- Requires only computation of $cq(\mathbf{x})$, where c is a constant
- Trail distribution must be symmetric: $t(\Delta\ \mathbf{x}) = t(-\Delta\ \mathbf{x})$
- Maintains detailed balance: $p(\mathbf{x}_k \rightarrow \mathbf{x}_{k+1}) = p(\mathbf{x}_{k+1} \rightarrow \mathbf{x}_k)$
- “Markov chain” since $\mathbf{x}_{k+1}$ depends probabilistically only on $\mathbf{x}_k$

Choice of trial distribution

- Algorithm places loose requirements on trial distribution *t*()
  - stationary; independent of position
- Often used functions include
  - *n*-D Gaussian, isotropic and uncorrelated
  - *n*-D Cauchy, isotropic and uncorrelated
- Choose width to “optimize” MCMC efficiency
  - rule of thumb: aim for acceptance fraction of about 25%

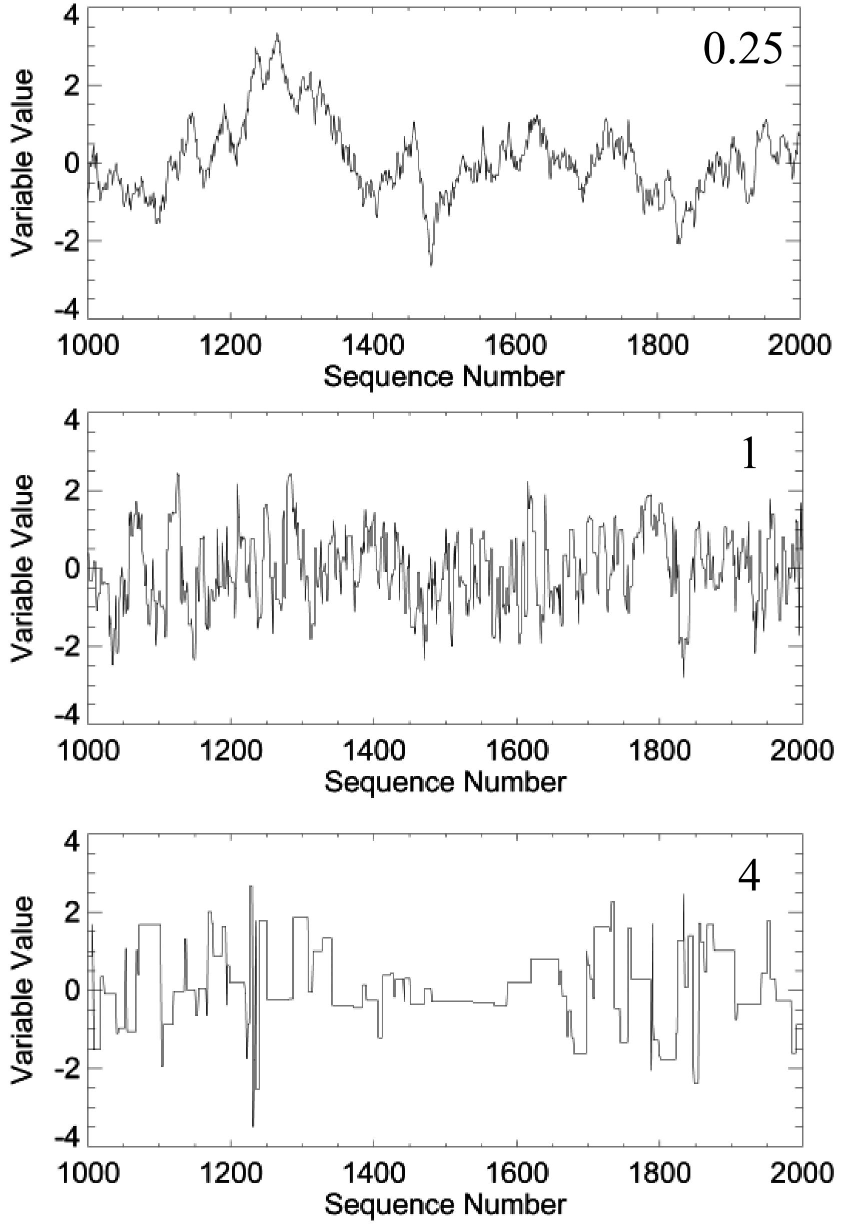
Choice of trial distribution – experiments

- Target distribution $q(\mathbf{x})$ is n dimensional Gaussian
  - uncorrelated, univariate (isotropic with unit variance)
  - most generic case
- Trial distribution $t(\Delta\mathbf{x})$ is n dimensional Gaussian
  - uncorrelated, equivariate; various widths



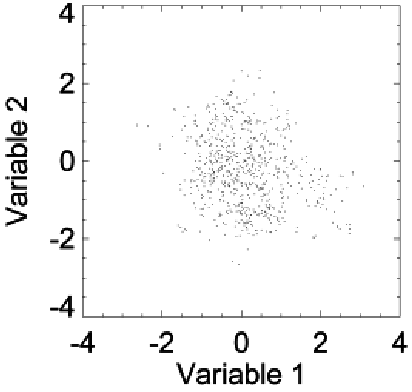
MCMC sequences for 2D Gaussian

- Results of running Metropolis with ratios of width of trial pdf to target pdf of 0.25, 1, and 4
- When trial pdf is much smaller than target pdf, movement across target pdf is slow
- When trial width same as target, samples seem to better sample target pdf
- When trial width much larger than target, trials stay put for long periods, but jumps are large

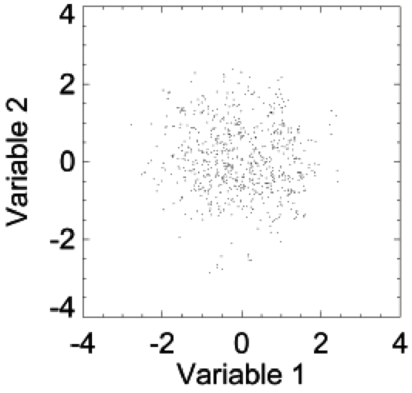


MCMC sequences for 2D Gaussian

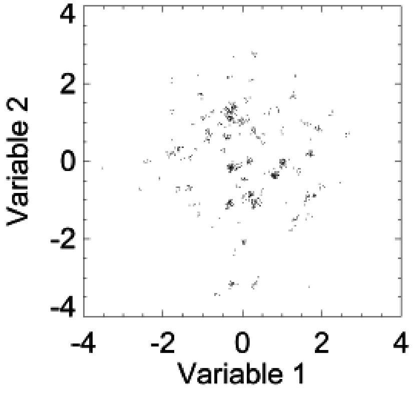
- Results of running Metropolis with ratios of width of trial pdf to target pdf of 0.25, 1, and 4
- Display accumulated 2D distribution for 1000 trials
- Viewed this way, it is difficult to see difference between top two images
- When trial pdf much larger than target, fewer splats, but further apart



0.25



1



4

MCMC - autocorrelation and efficiency

- In MCMC sequence, subsequent parameter values are usually correlated
- Degree of correlation quantified by autocorrelation function:

$$\rho(l) = \frac{1}{N}\sum_{i=1}^{N} y(i)\,y(i-l)$$

  - where *y(x)* is the sequence and *l* is lag

- For Markov chain, expect exponential

$$\rho(l) = \exp[-\left|\frac{l}{\lambda}\right|]$$

- Sampling efficiency is

$$\eta = [1 + 2\sum_{l=1}^{\infty}\rho(l)]^{-1} = \frac{1}{1+2\lambda}$$

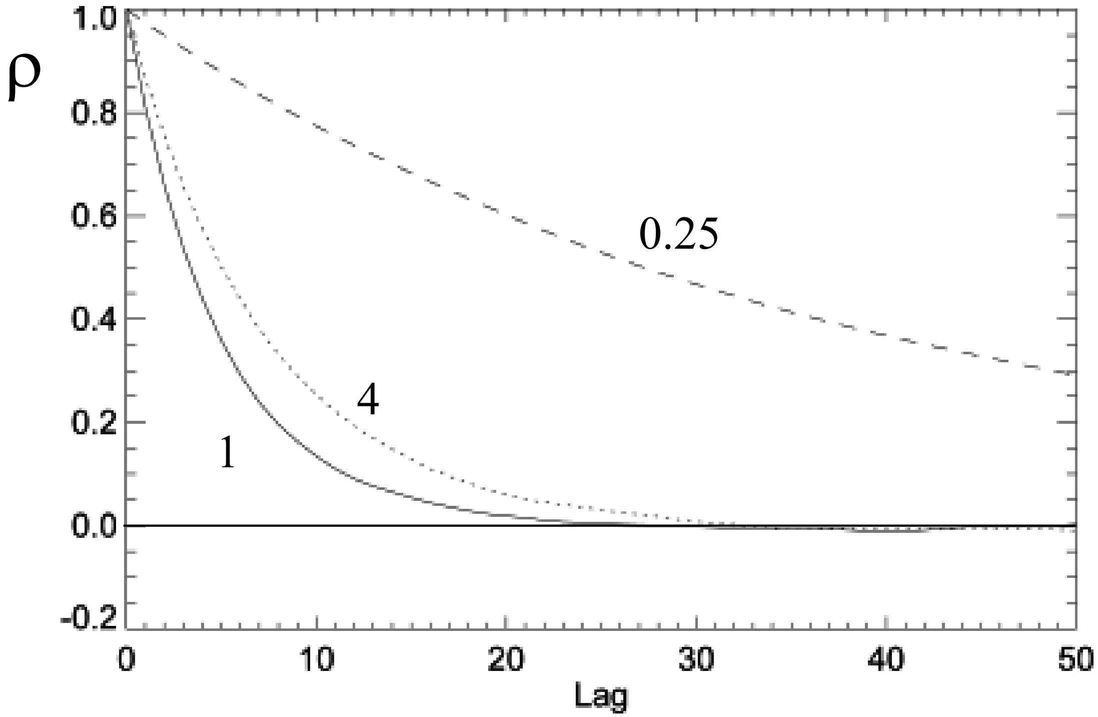
- In other words, η^{-1} iterates required to achieve one statistically independent sample

Autocorrelation for 2D Gaussian

- Plot confirms that the autocorrelation drops slowly when the trial width is much smaller than the target width; MCMC efficiency is poor
- Sampling efficiency is

$$\eta = \frac{1}{1+2\lambda}$$

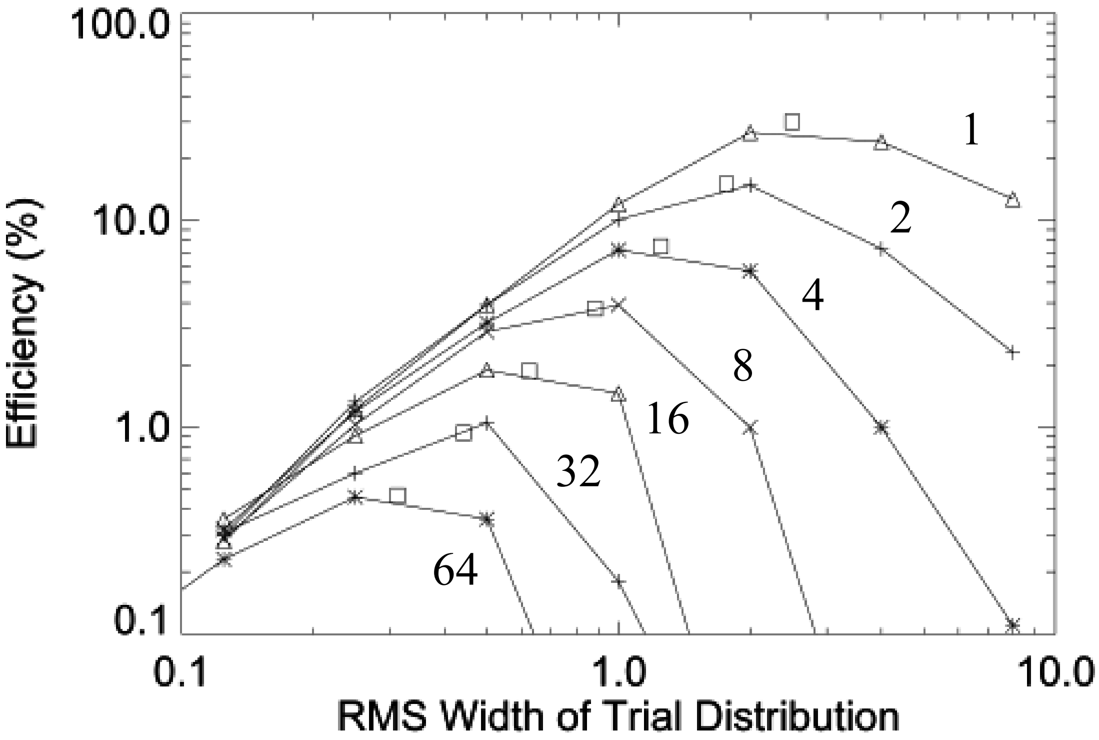
- Best efficiency occurs when trial about same size as target (for 2D)



Normalized autocovariance for various widths of trial pdf relative to target: 0.25, 1, and 4

Efficiency as function of width of trial pdf

- for univariate, uncorrelated Gaussians, with 1 to 64 dimensions
- efficiency as function of width of trial distributions
- boxes are predictions of optimal efficiency from diffusion theory [A. Gelman, et al., 1996]
- efficiency drops reciprocally with number of dimensions

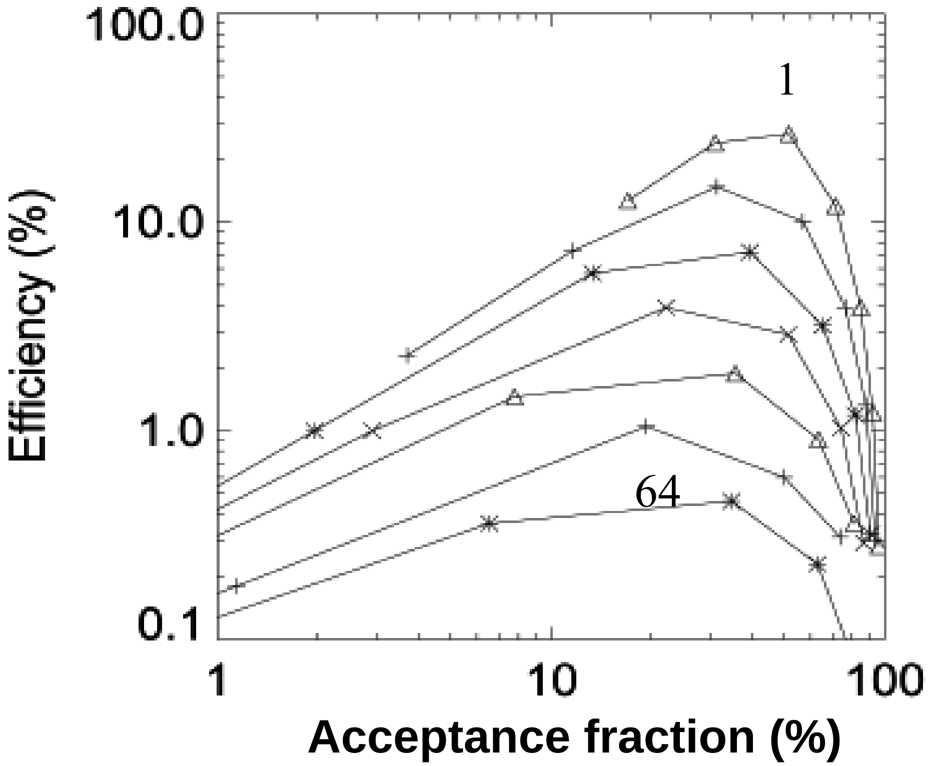


Efficiency as function of acceptance fraction

- For univariate Gaussians, with 1 to 64 dimensions
- Efficiency as function of acceptance fraction
- Best efficiency is achieved when about 25% of trials are accepted for moderate number of dimensions
- Optimal statistical efficiency:

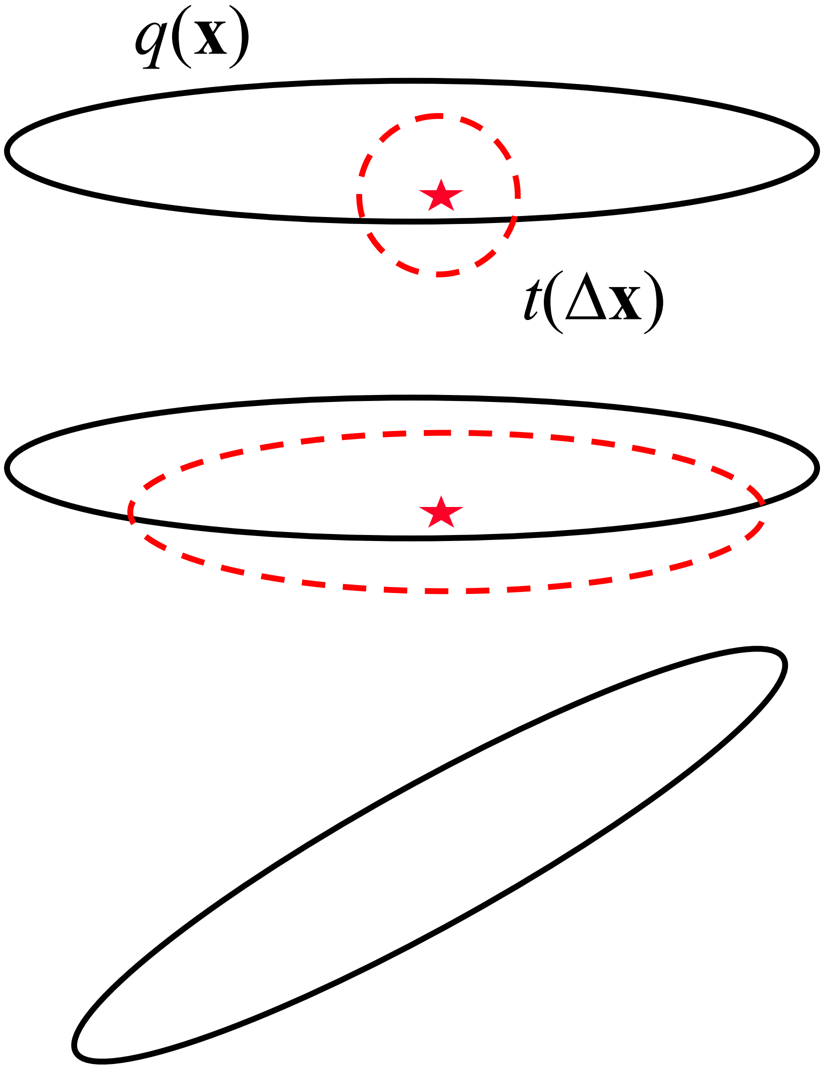
  $$\eta \sim 0.3/n$$

  - for uncorrelated, equivariate Gaussian
  - generally decreases correlation and variable variance
  - consistent with diffusion theory derivation [A. Gelman, et al., 1996]



Further considerations

- When target distribution $q(\mathbf{x})$ not isotropic
  - difficult to accommodate with isotropic $t(\Delta\mathbf{x})$
  - each parameter can have different efficiency
  - desirable to vary width of different $t(\mathbf{x})$ to approximately match $q(\mathbf{x})$
  - recovers efficiency of univariate case
- When $q(\mathbf{x})$ has correlations
  - $t(\mathbf{x})$ should match shape of $q(\mathbf{x})$

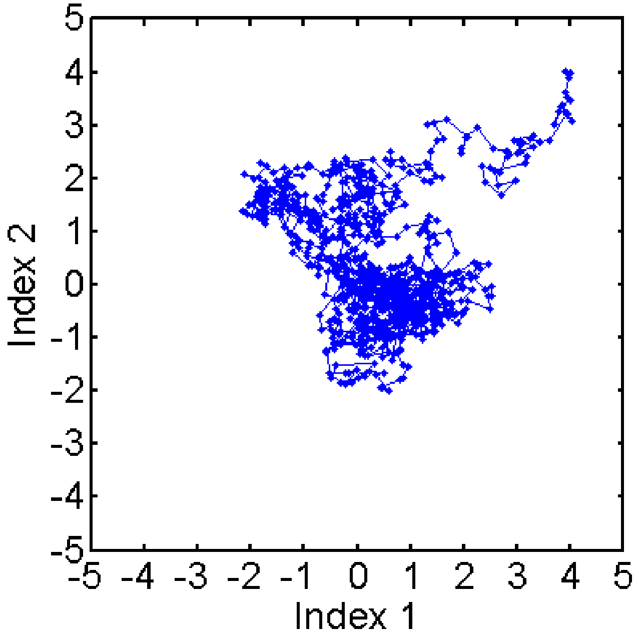


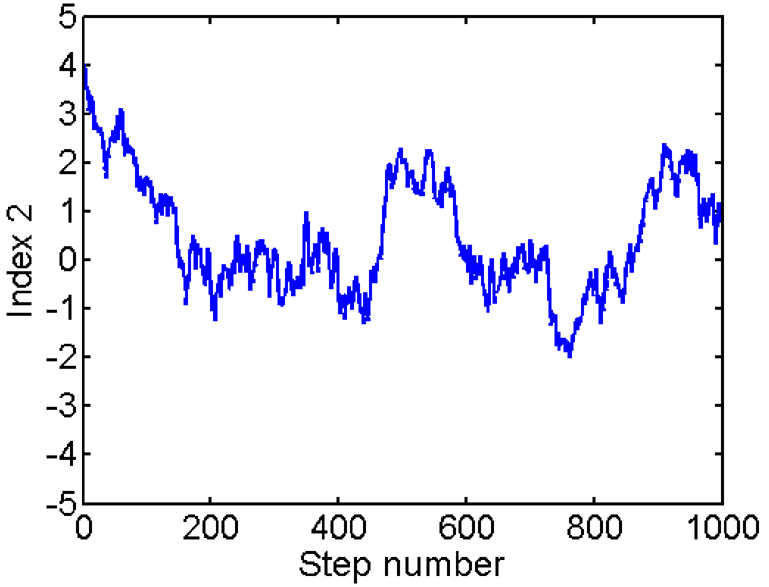
MCMC - Issues

- Identification of convergence to target pdf
  - is sequence in thermodynamic equilibrium with target pdf?
  - validity of estimated properties of parameters (covariance)
- Burn in
  - at beginning of sequence, may need to run MCMC for awhile to achieve convergence to target pdf
- Use of multiple sequences
  - different starting values can help confirm convergence
  - natural choice when using computers with multiple CPUs
- Accuracy of estimated properties of parameters
  - related to efficiency, described above
- Optimization of efficiency of MCMC

MCMC – convergence and burn in

- Example: sequence obtained for 2 D unit-variance Gaussian pdf
  - Metropolis algorithm
  - starting point is (4, 4)
  - trial pdf is Gaussian, σ = 0.2
  - 1000 steps
  - avg acceptance = 0.87
- Observe:
  - large number of steps required before sequence has converged to core region (burn in)
  - hard to tell whether sequence has converged, either from 2D plot or by looking at individual coordinate (convergence)



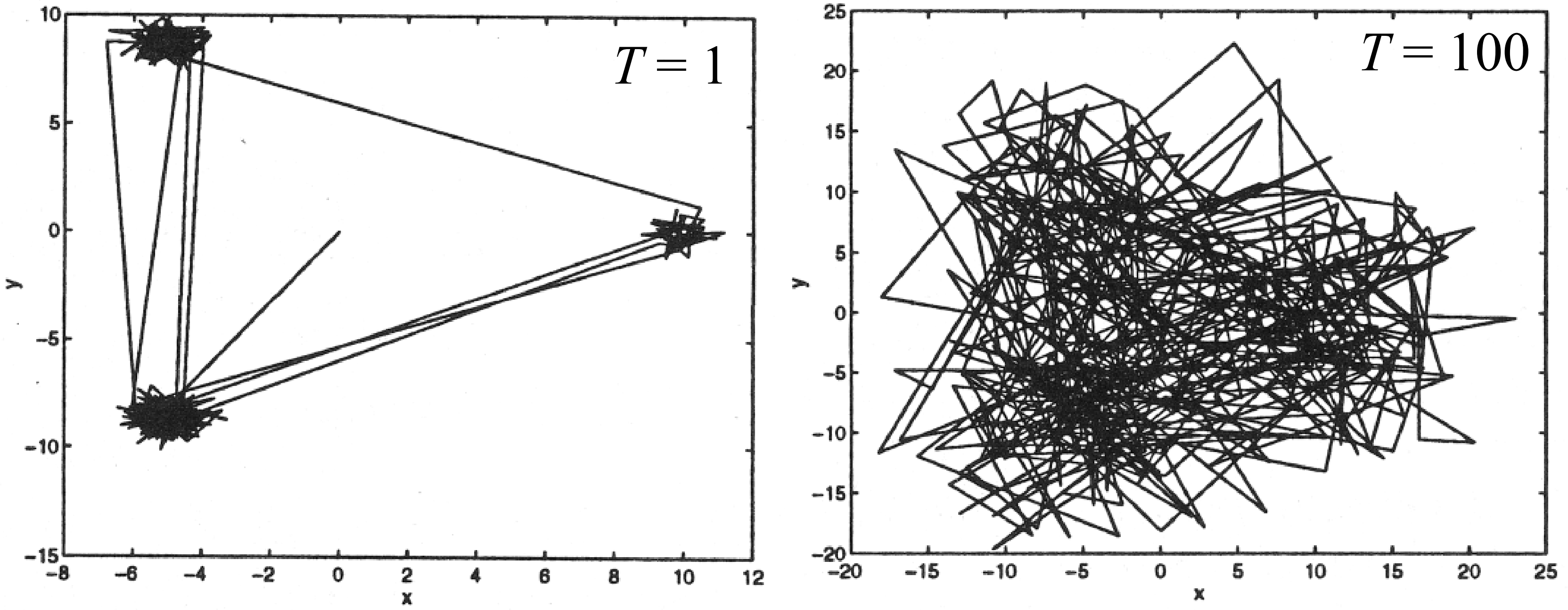


Annealing

- Introduction of fictitious temperature
  - define functional $\varphi(\mathbf{x})$ as minus-logarithm of target probability
    $$\varphi(\mathbf{x}) = - \log(q(\mathbf{x}))$$
  - scale φ by an inverse "temperature" to form new pdf
    $$q'(\mathbf{x}, T) = \exp[- \varphi(\mathbf{x})/ T]$$
  - $q'(\mathbf{x}, T)$ is flatter than $q(\mathbf{x})$ for $T > 1$ (called annealing)
- Uses of annealing (also called tempering)
  - allows MCMC to move between multiple peaks in $q(\mathbf{x})$
  - simulated-annealing optimization algorithm (takes lim $T \rightarrow 0$)

Annealing helps handle multiple peaks

- Scale minus-log-prob: $q'(\mathbf{x}, T) = \exp[-\,\varphi(\mathbf{x})/T]$, T = temperature
- Example: target distribution is three narrow, well separated peaks
- For original distribution ($T = 1$), an MCMC run of 10000 steps rarely moves between peaks
- At temperature $T = 100$ (right), MCMC moves easily between peaks and through surrounding regions



from M-D Wu and W. J. Fitzgerald, *Maximum Entropy and Bayesian Methods* (1996)

Other MCMC algorithms

- Gibbs
  - vary only one component of **x** at a time
  - draw new value of x_j from conditional $q(x_j | x_1 x_2 \ldots x_{j-1} x_{j+1} \ldots)$
- Metropolis-Hastings
  - allows use of nonsymmetric trial functions, $t(\Delta\mathbf{x}; \mathbf{x}_\mathrm{k})$
  - uses acceptance criterion $r = [t(\Delta\mathbf{x}; \mathbf{x}_\mathrm{k})\, q(\mathbf{x}^*)] \,/\, [t(-\Delta\mathbf{x}; \mathbf{x}^*)\, q(\mathbf{x}_\mathrm{k})]$
- Langevin technique
  - variation of Metropolis-Hastings approach
  - uses gradient* of minus-log-prob to shift trial function towards regions of higher probability
- Hamiltonian hybrid algorithm
  - based on particle dynamics; requires gradient* of minus-log-prob
  - provides potentially higher efficiency for large number of variables
- Many others

* adjoint differentiation affords efficient gradient calculation

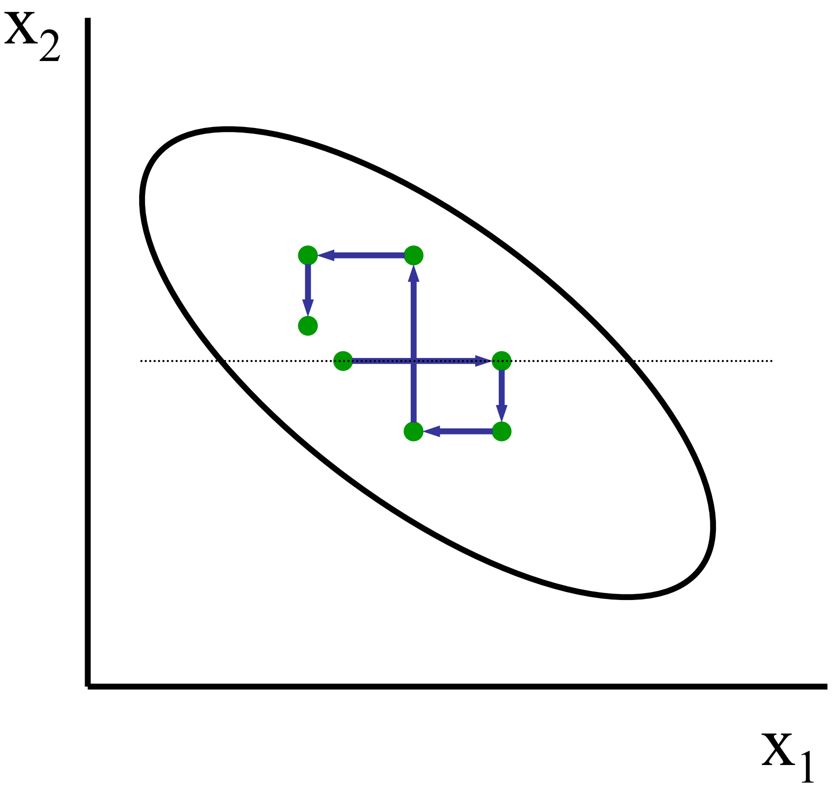
Gibbs algorithm

- Vary only one component of **x** at a time
- Draw new value of x_j from conditional pdf

  $q(x_j| x_1 x_2 \ldots x_{j-1} x_{j+1} \ldots)$

  - algorithm typically used only when draws from q are relatively easy to do
- Cycle through all components

Probability(x_1, x_2)



Hamiltonian hybrid algorithm

- Hamiltonian hybrid algorithm
  - called hybrid because it alternates Gibbs & Metropolis steps
  - associate with each parameter x_i a momentum p_i
  - define a Hamiltonian
    $H = \varphi(\mathbf{x}) + \Sigma\, p_i^{2}/(2\, m_i)$; where $\varphi = -\log\,(q\,(\mathbf{x}\,))$
  - new pdf:
    $$q'(\mathbf{x}, \mathbf{p}) = \exp(-\,H(\mathbf{x}, \mathbf{p})) = q(\mathbf{x})\,\exp(-\Sigma\, p_i^{2}/(2\, m_i))$$
  - can easily move long distances in ($\mathbf{x}$, $\mathbf{p}$) space at constant H using Hamiltonian dynamics, so Metropolis step is very efficient
  - uses gradient* of φ (minus-log-prob)
  - Gibbs step in constant **p** is easy
  - efficiency may be better than Metropolis for large dimensions

* adjoint differentiation affords efficient gradient calculation

Hamiltonian algorithm

- Gibbs step: randomly sample momentum distribution
- Follow trajectory of constant *H* using leapfrog algorithm:

$$p_i(t+\frac{\tau}{2}) = p_i(t) - \frac{\tau}{2}\frac{\partial\varphi}{\partial x_i}\bigg|_{\mathbf{x}(t)}$$

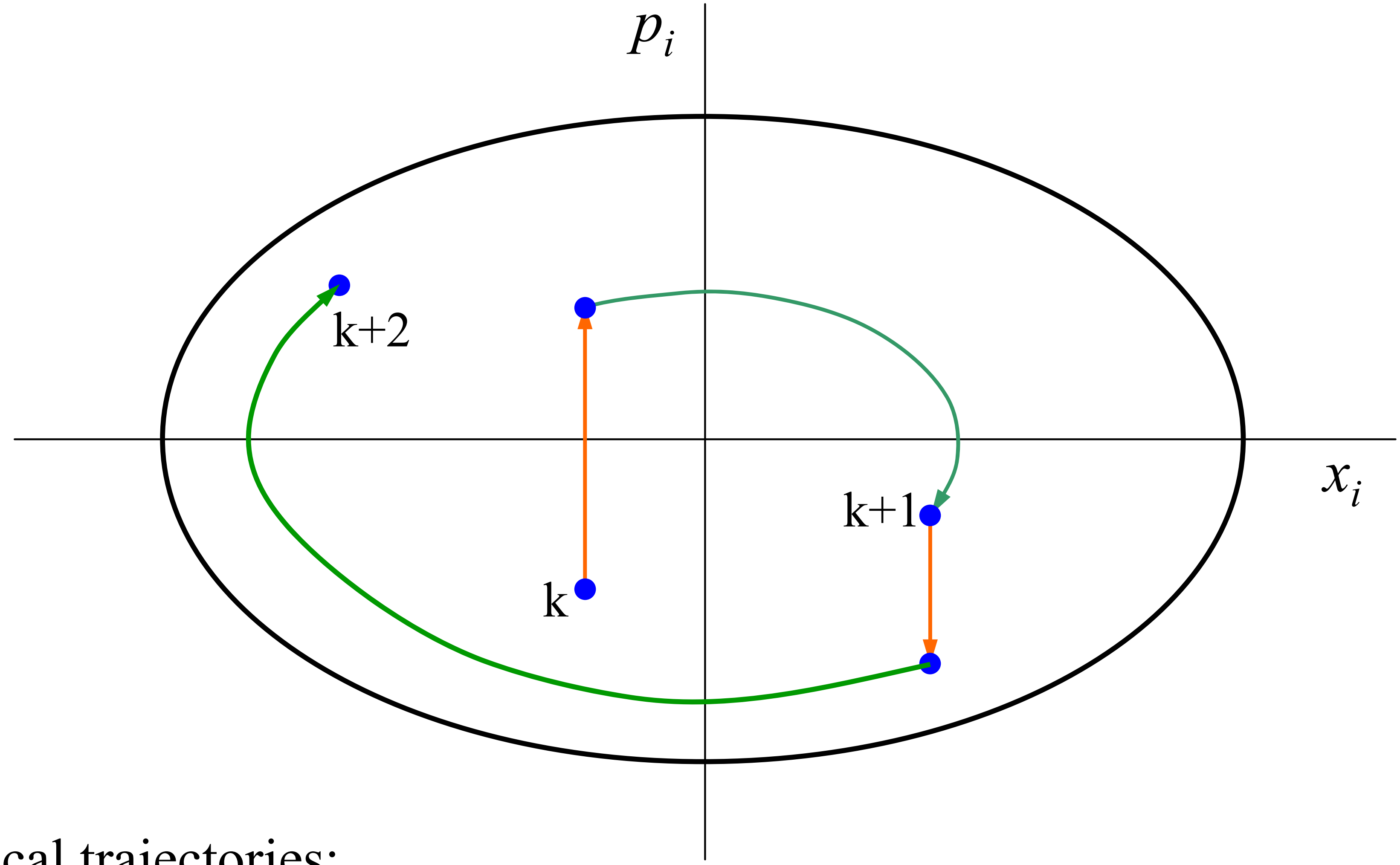
$$x_i(t+\tau) = x_i(t+\tau) + \frac{\tau}{m_i}p_i(t+\frac{\tau}{2})$$

$$p_i(t+\tau) = p_i(t+\frac{\tau}{2}) - \frac{\tau}{2}\frac{\partial\varphi}{\partial x_i}\bigg|_{\mathbf{x}(t+\tau)}$$

where τ is leapfrog time step.

- Repeat leapfrog a predetermined number of times
- Metropolis step: accept or reject on basis of *H* at beginning and end of H trajectory

Hamiltonian hybrid algorithm



Typical trajectories:

red path - Gibbs sample from momentum distribution

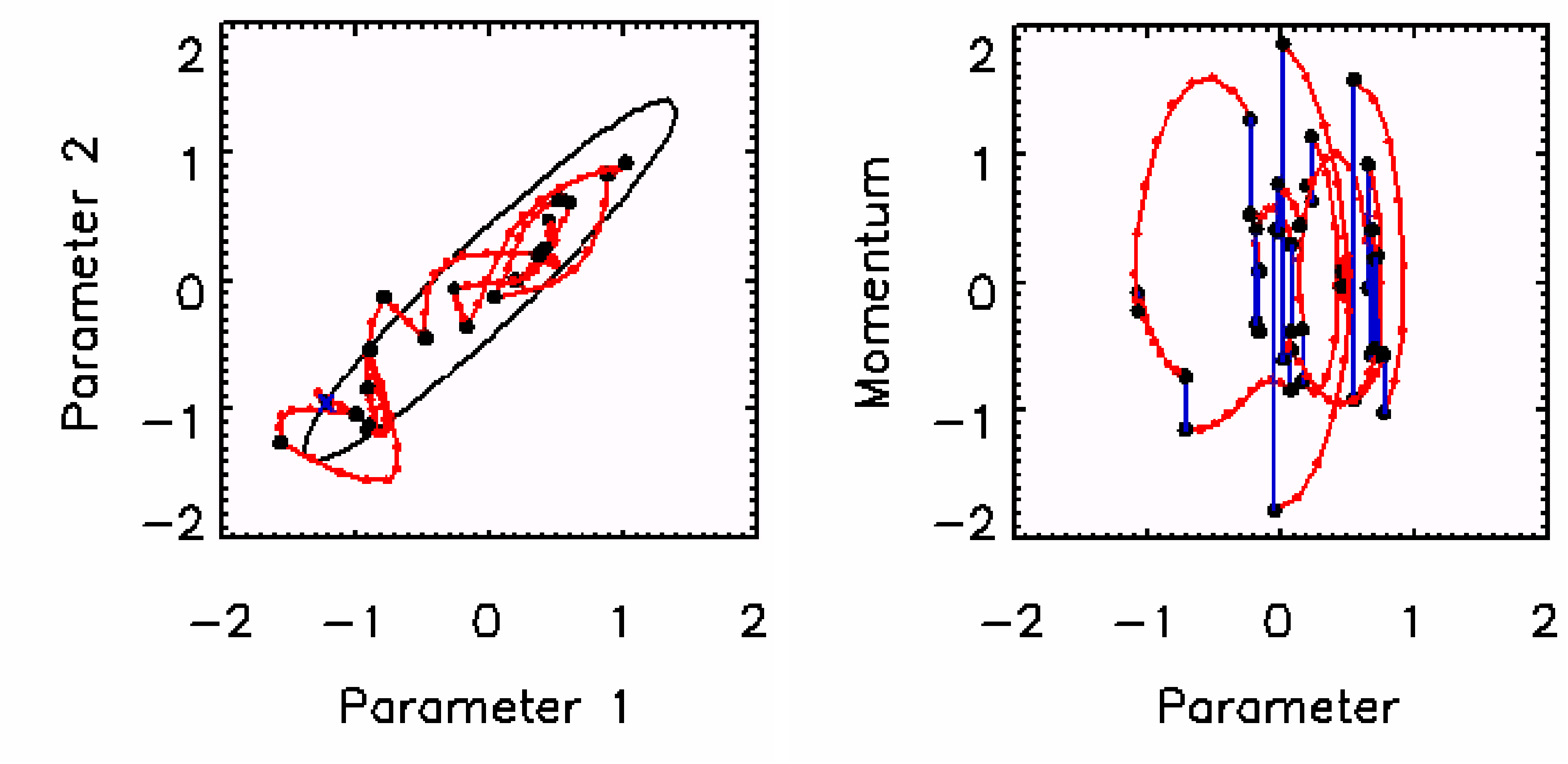
green path - trajectory with constant *H*, follow by Metropolis

Hamiltonian algorithm

- Gibbs step - easy because draws are from uncorrelated Gaussian
- H trajectories followed by several leapfrog steps permit long jumps in (**x**, **p**) space, with little change in *H*
  - specify total time = *T* ; number of leapfrog steps = T/τ
  - randomize *T* to avoid coherent oscillations
  - reverse momenta at end of H trajectory to guarantee that it is symmetric process (condition for Metropolis step)
- Metropolis step - no rejections if *H* is unchanged

- Adjoint differentiation efficiently provides gradient

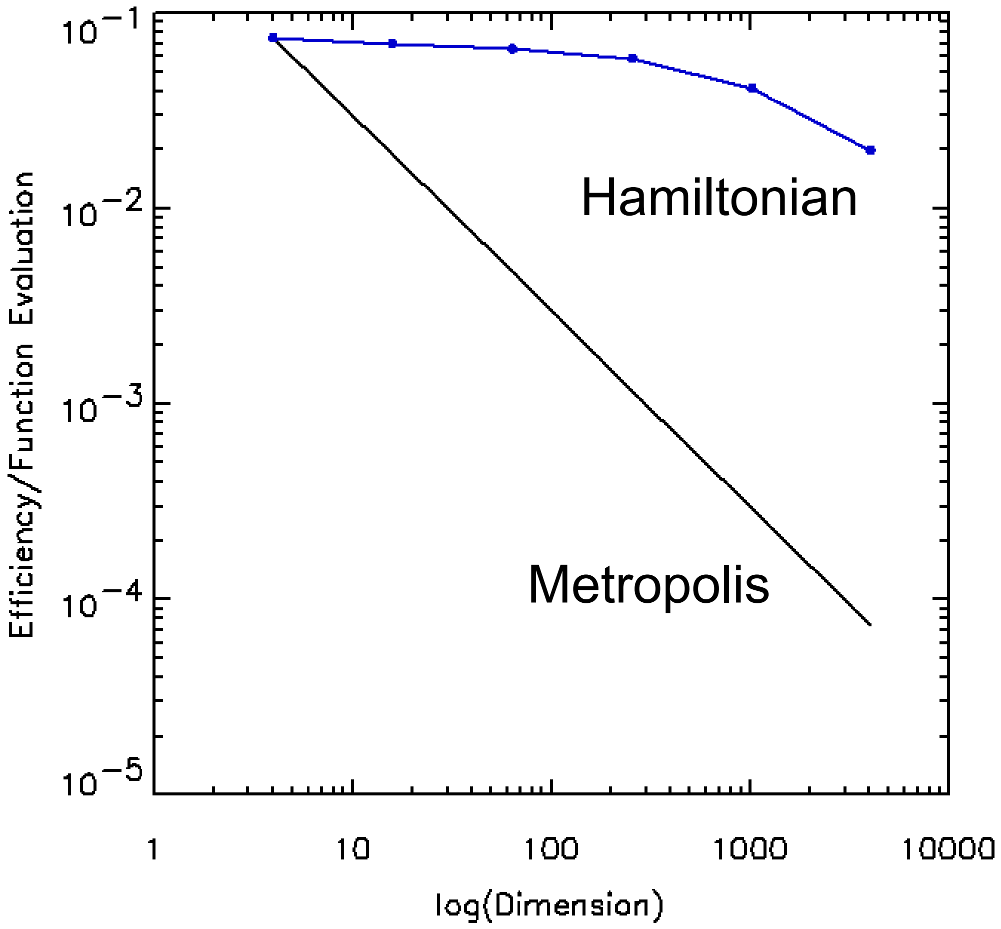
2D correlated Gaussian distribution



- 2D Gaussian pdf with high correlation (r =0.95)
- Length of H trajectories randomized

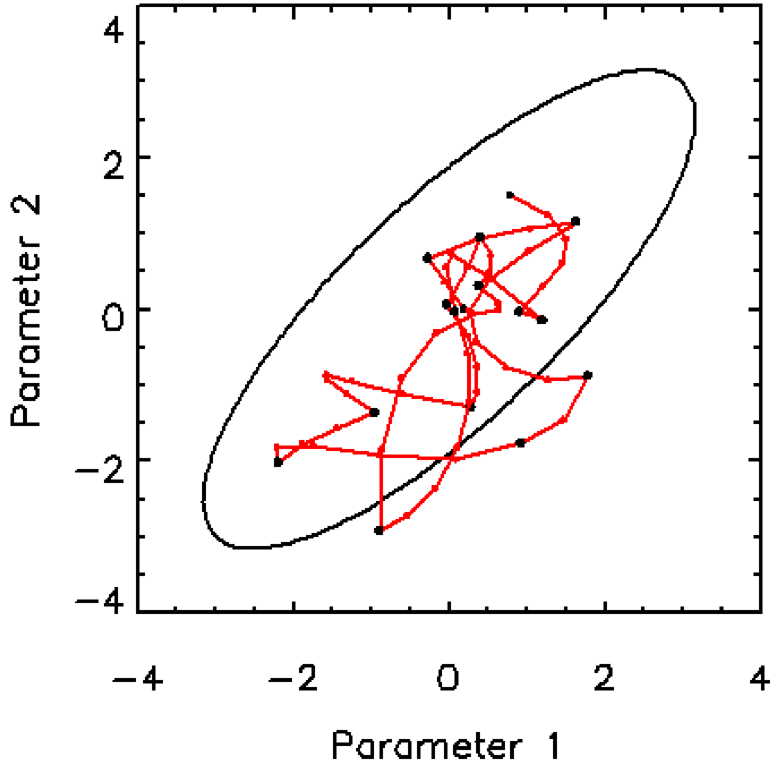
n-D isotropic Gaussian distributions

- Assume that gradient of φ are calculated as quickly as φ itself (e.g., using adjoint differentiation)
- MCMC efficiency versus number dimensions
  - ▸ Hamiltonian method: drops little
  - ▸ Metropolis method: goes as 0.3/*n*



- Hamiltonian method much more efficient at high dimensions

16D correlated Gaussian distribution



- 16D Gaussian pdf related to smoothness prior based on integral of L2 norm of second derivative
- Efficiency/(function evaluation) =
  2.2% (Hamiltonian algorithm)
  0.11% or 1.6% (Metropolis; without and with covariance adaptation)

Conclusions – Hamiltonian MCMC

- MCMC provides good tool for exploring the Bayesian posterior and hence for drawing inferences about models and parameters
- Hamiltonian method
  - based on Hamiltonian dynamics
  - efficiency for isotropic Gaussians is about 7% per function evaluation, independent of number of dimensions
  - caveat – must be able to calculate gradient of minus-log-posterior in time comparable to the posterior itself (e.g., through adjoint differentiation)
  - much better efficiency than Metropolis for large dimensions
  - more robust to correlations among parameters than Metropolis

Conclusions – MCMC

- MCMC provides good tool for exploring the posterior and hence for drawing inferences about models and parameters
- For valid results, care must be taken to
  - verify convergence of the sequence
  - exclude early part of sequence, before convergence reached
  - be wary of multiple peaks that need to be sampled
- For good efficiency with Metropolis alg., care must be taken to
  - adjust the size and shape of the trial distribution; rule of thumb is to aim for 25% trial acceptance for $5 < n < 100$

- A lot of MCMC research is going on
- Software libraries for MCMC are available for most computer languages, or as stand-alone applications, e.g., OpenBUGS (formerly WinBUGS)

Rossi analysis – example of MCMC

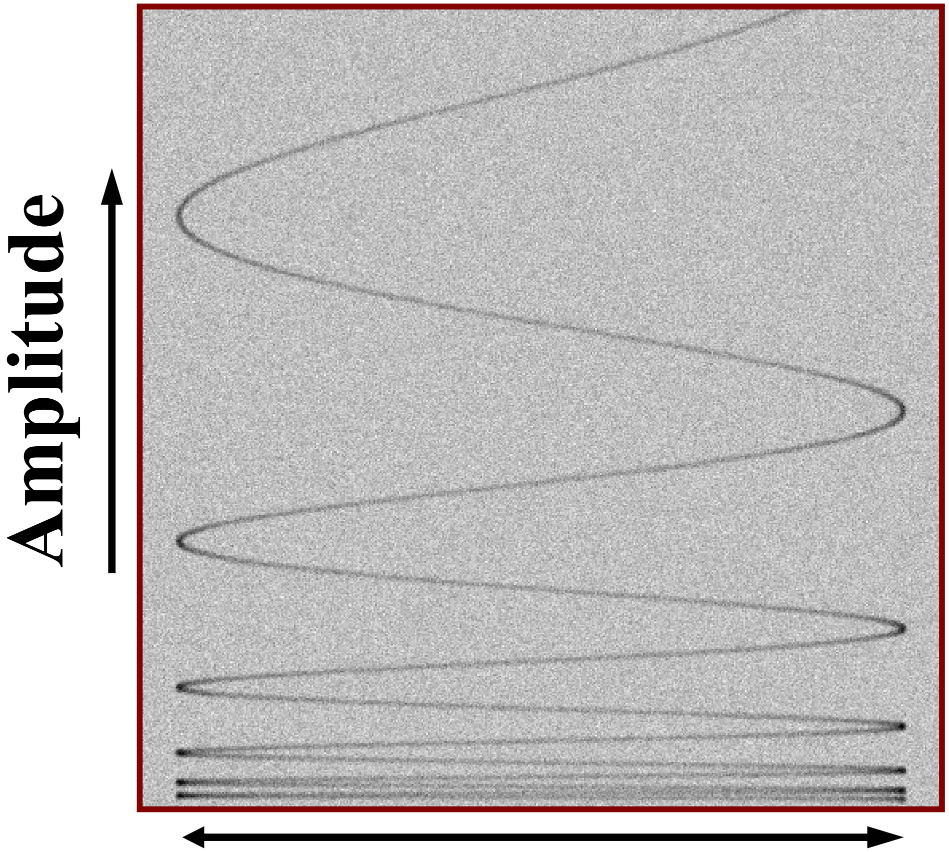
- Goal: measure flux as function of time, $\Phi(t)$, to obtain alpha, a measure of criticality, versus time

$$\alpha(t) = \frac{1}{\Phi}\frac{d\Phi}{dt} = \frac{d(\ln\Phi)}{dt}$$

- Experimental issues
  - measurements made using Rossi technique
  - signal displayed on oscilloscope, photographed, read
  - recorded signal is band limited
- Analysis complicated by intricate error model for measurements

The Rossi technique

- Rossi technique - photograph oscilloscope screen
  - horizontal sweep is driven sinusoidally in time
  - signal amplitude vertical
- Records rapidly increasing signal while keeping trace in middle of CRT, which minimizes oscilloscope nonlinearities



$$x = x_R \cos(2\pi f_R t + \phi_0)$$

Bayesian analysis of an experiment

- The pdf describing uncertainties in model parameter vector **a**, called **posterior**:
  - ► $p(\mathbf{a}|\mathbf{d}) \sim p(\mathbf{d}|\mathbf{d}^*)\, p(\mathbf{a})$ (Bayes law)

    where **d** is vector of measurements, and

    d*(**a**) is measurement vector predicted by model
  - ► $p(\mathbf{d}|\mathbf{d}^*)$ is likelihood, probability of measurements **d** given the values **d*** predicted by simulation of experiment
  - ► $p(\mathbf{a})$ is prior; summarizes previous knowledge of **a**
  - ► “best” parameters estimated by
    - maximizing posterior (called MAP solution)
    - mean of posterior
  - ► uncertainties in **a** are fully characterized by $p(\mathbf{a}|\mathbf{d})$

Cubic spline expansion of alpha curve

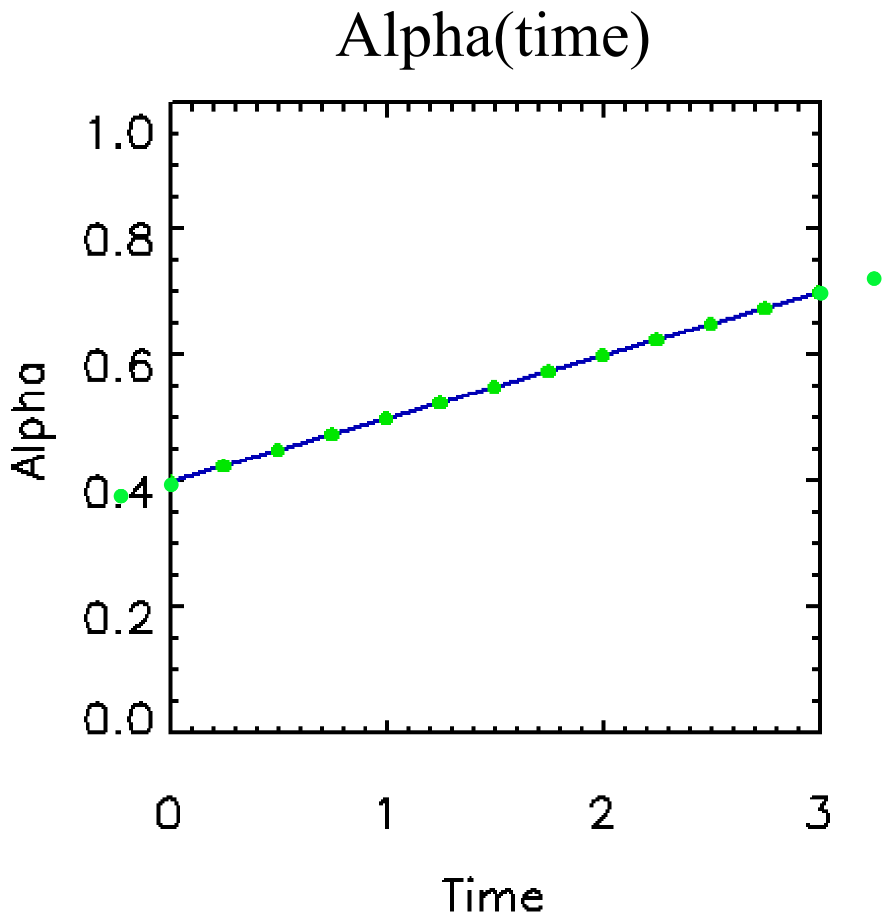
- Expand $\alpha(t)$ in terms of basis functions:

$$\alpha(t) = \sum_k a_k \, \phi\left[\frac{t - t_k}{\Delta t}\right]$$

where

  - a_k is the expansion coefficient,
  - ϕ is a spline basis function,
  - t_k is the position of the kth knot
  - Δt is the knot spacing

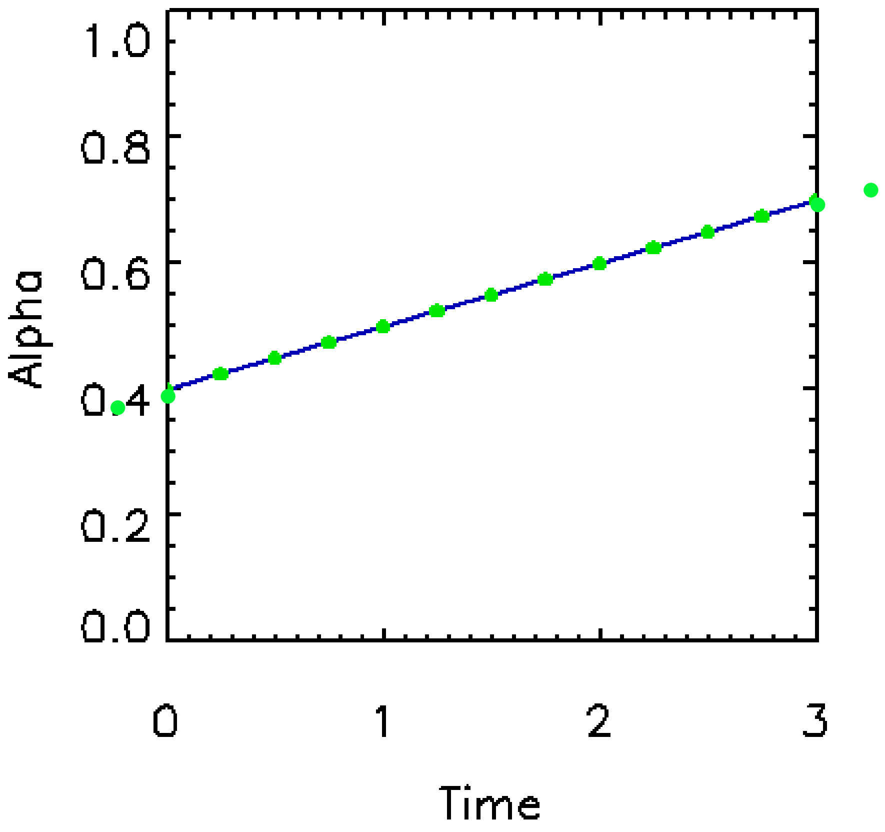
- Use 15 evenly-space knots
  - spacing chosen on basis of limited bandwidth of signal y
  - two are outside data interval to handle end conditions
- Parameters a_k are to be determined



Modeling the Rossi data

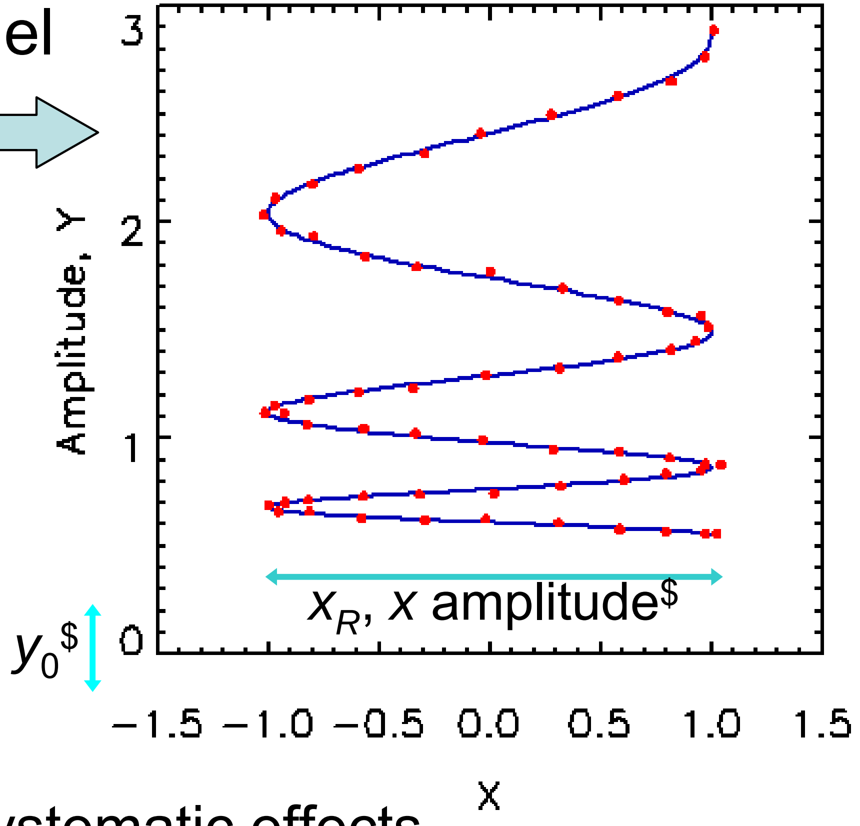
- $\alpha(t)$ represented as cubic spline
- measurement model predicts data
- can include systematic effects of measurement system

Alpha(time)



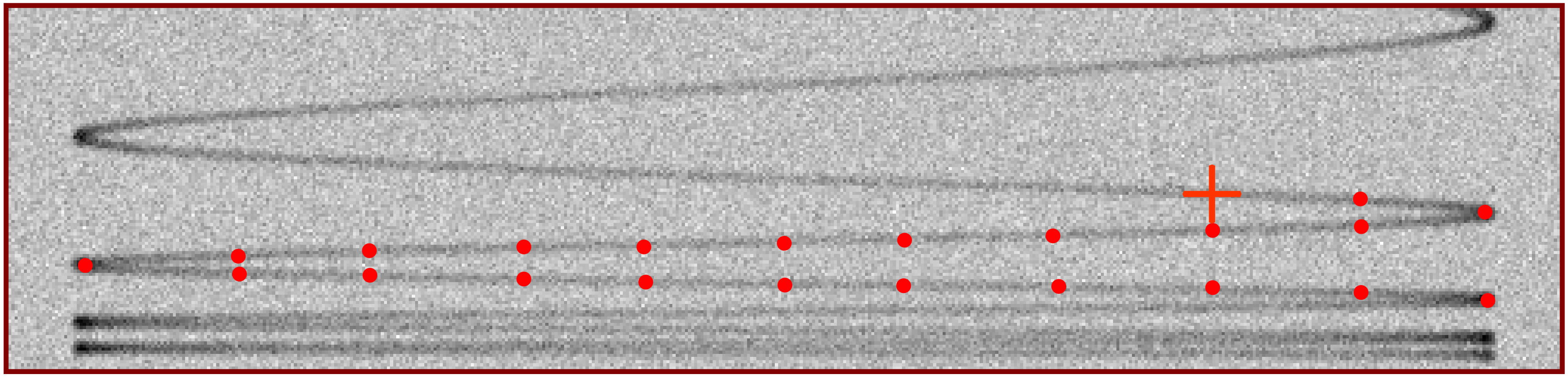
Measurement Model

x-y data (used in calc.)



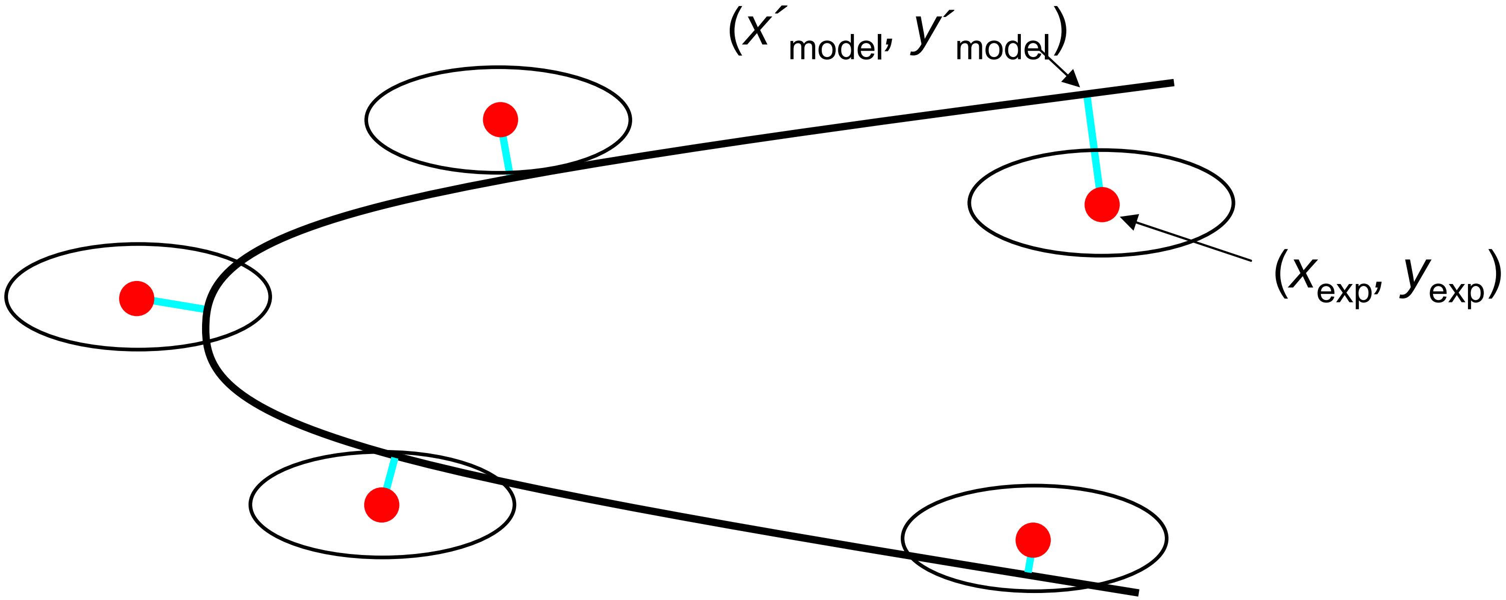
$systematic effects

Reading a Rossi trace



- Technician reads points by centering cross hairs of a reticule on trace; computer records positions, $\{x_i, y_i\}$
- Points are read with intent to:
  - place point at peaks
  - achieve otherwise arbitrary placement along curve with even spacing along trace

Likelihood model - uncertainties in Rossi data



- minus-log-likelihood, p(**d**|**a**), for measured point ($x_{\exp}$, $y_{\exp}$):

$$\Delta\frac{\chi^2}{2} = \frac{(x_{\exp} - x'_{\text{model}})^2}{2\sigma_x^2} + \frac{(y_{\exp} - y'_{\text{model}})^2}{2\sigma_y^2}$$

where $(x'_{\text{model}}, y'_{\text{model}})$ is the model point closest to ($x_{\exp}$, $y_{\exp}$)

Smoothness constraint

- Cubic splines tend to oscillate in some applications
- Smoothness of $\alpha(t)$ can be controlled by minimizing

$$S(\alpha) = T^3 \int \left| \frac{d^2 \alpha}{dt^2} \right|^2 dt$$

where T is the time interval; T^3 factor removes T dependence

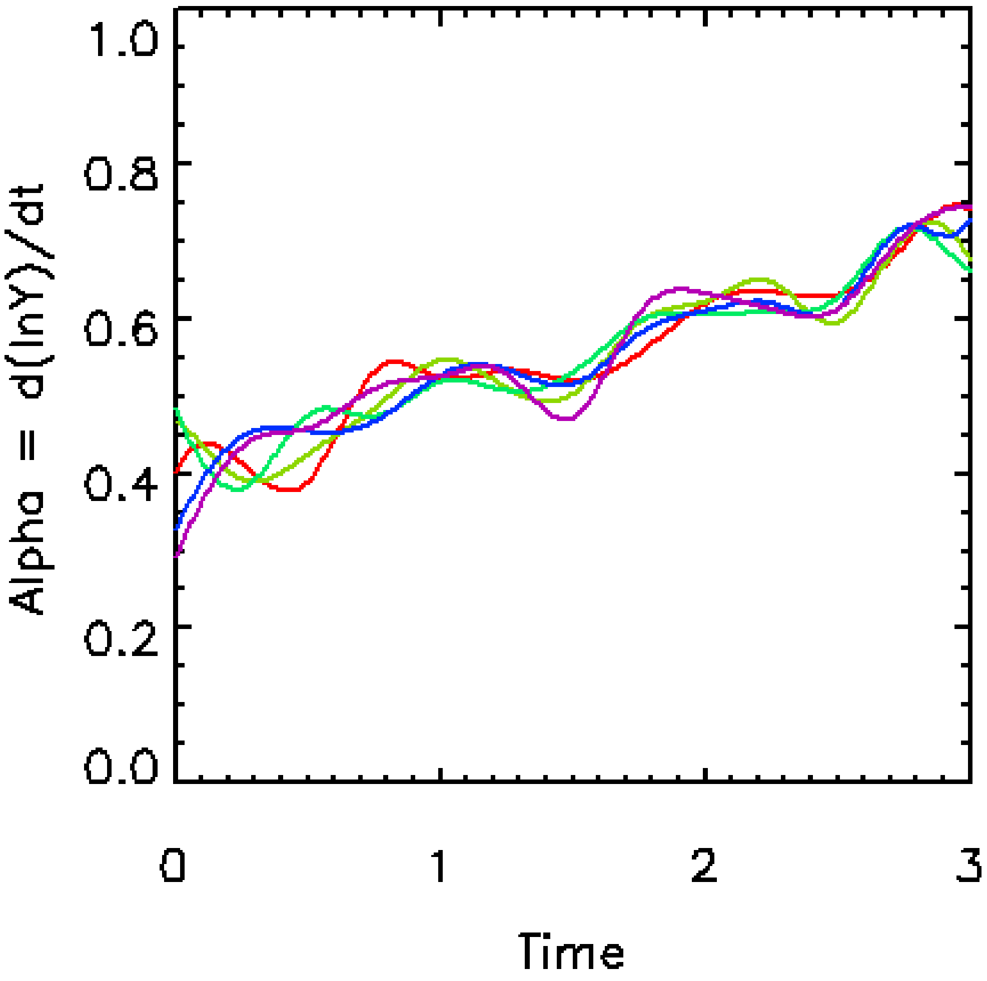
- Smoothness can be incorporated in Bayesian context by setting prior on spline coefficients to

$$-\log p(\mathbf{a}) = \lambda\, S(\alpha(\mathbf{a}))$$

- Hyperparameter λ can be determined in Bayesian approach by maximizing $p(\lambda|\mathbf{d})$

MCMC - alpha uncertainty

- MCMC samples from posterior
  - plot shows several $\alpha(t)$ curves consistent with data
  - uncertainties in model visualized as variability among curves
- Smoothness parameter, $\lambda = 0.4$



MCMC – estimation of λ

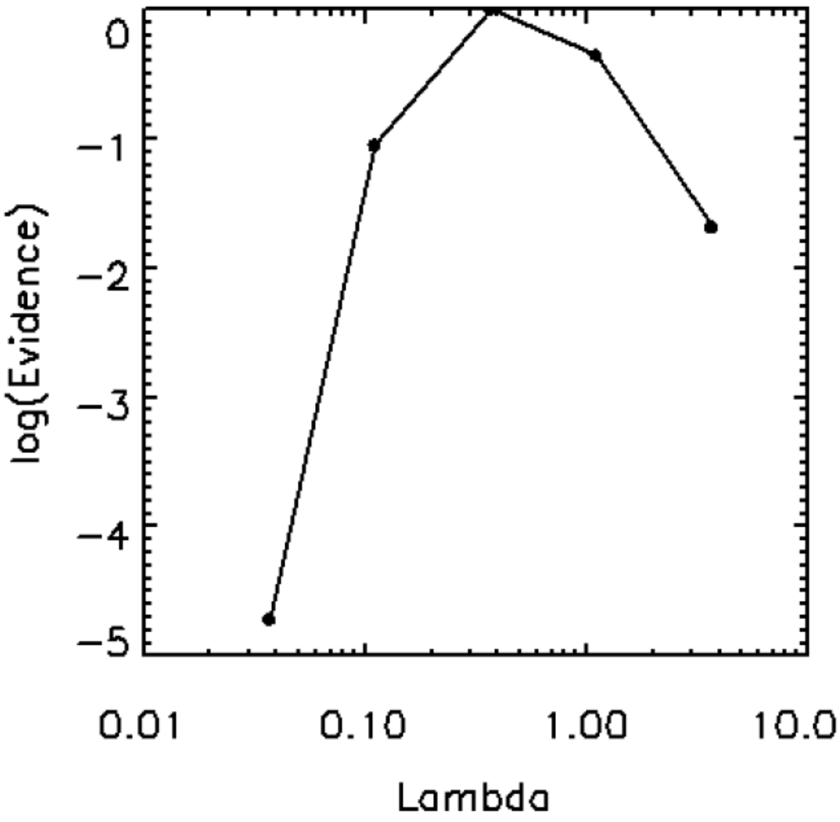
- Strength of smoothness prior given by λ
- Determine λ using Bayes law

$$p(\lambda \mid \boldsymbol{d}) = \int p(\boldsymbol{a}, \lambda \mid \boldsymbol{d})\, d\boldsymbol{a}$$

$$\propto \int p(\boldsymbol{d} \mid \boldsymbol{a}, \lambda)\, p(\boldsymbol{a}, \lambda)\, d\boldsymbol{a}$$

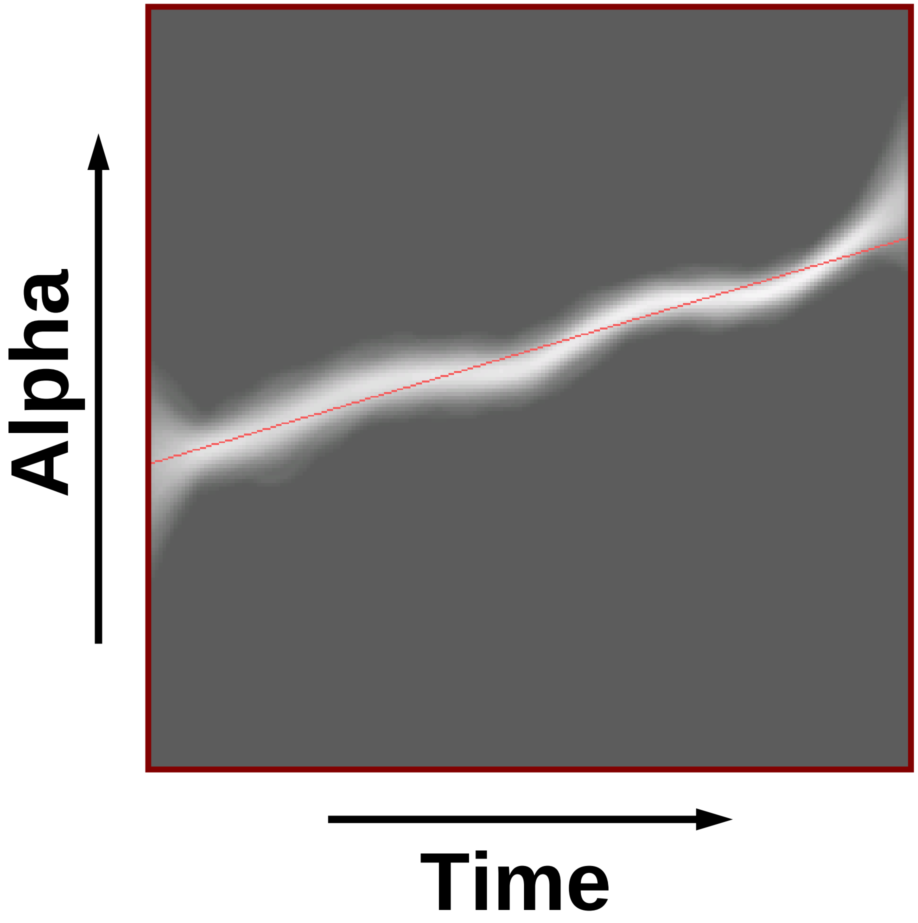
$$= p(\lambda) \int p(\boldsymbol{d} \mid \boldsymbol{a}, \lambda)\, p(\boldsymbol{a})\, d\boldsymbol{a}$$

- Last integral, called **evidence**, is estimated as value of integrand at its peak times its volume
- Volume given by determinant of covariance matrix of **a**, estimated using MCMC sequence
- At maximum λ = 0.4



MCMC - Alpha

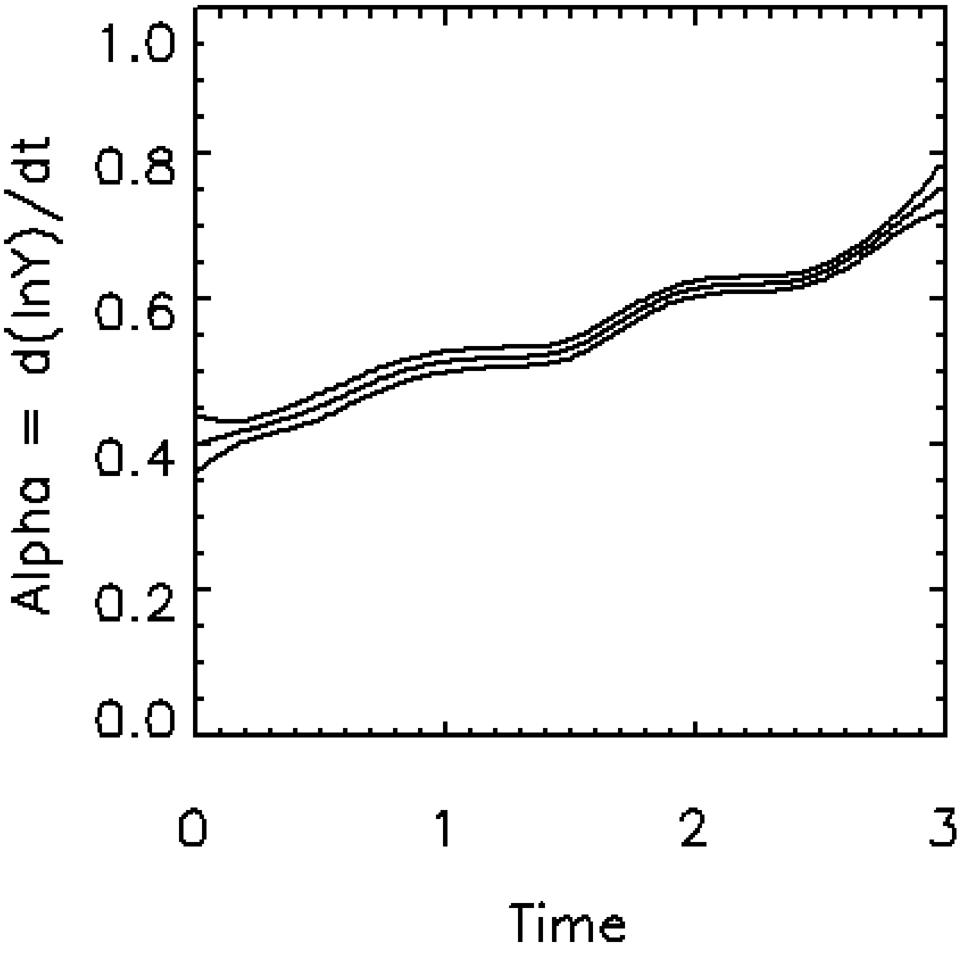
- For MCMC sequence with 10^5 samples, image shows accumulated MCMC curves in alpha domain
- Effectively shows PDF for uncertainty distribution in alpha, estimated from data
- However, does **not** show correlations between uncertainties at two different times, as do individual MCMC samples



$\lambda = 0.4$ (best value)

MCMC - Alpha

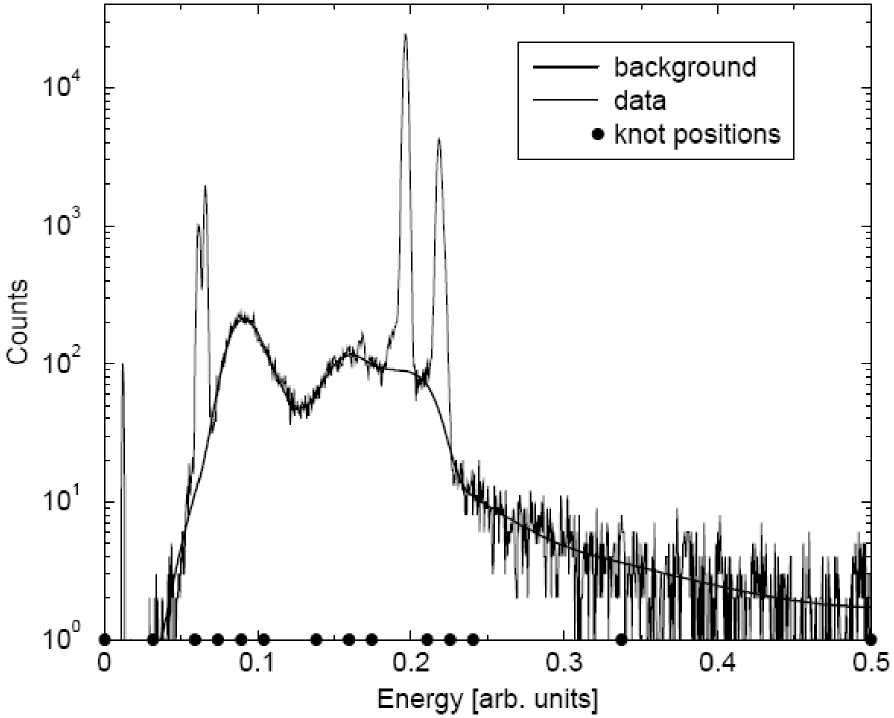
- Interpreting accumulated alpha curve as a PDF, one can estimate $\alpha(t)$ in terms of
  - posterior mean
  - posterior max. (MAP estimate)
- Or characterize uncertainties
  - standard deviations
  - covariance matrix (correlations)
  - credible intervals (envelope)
- Plot on right shows
  - posterior mean
  - posterior mean +/- standard dev. (one standard dev. envelope)

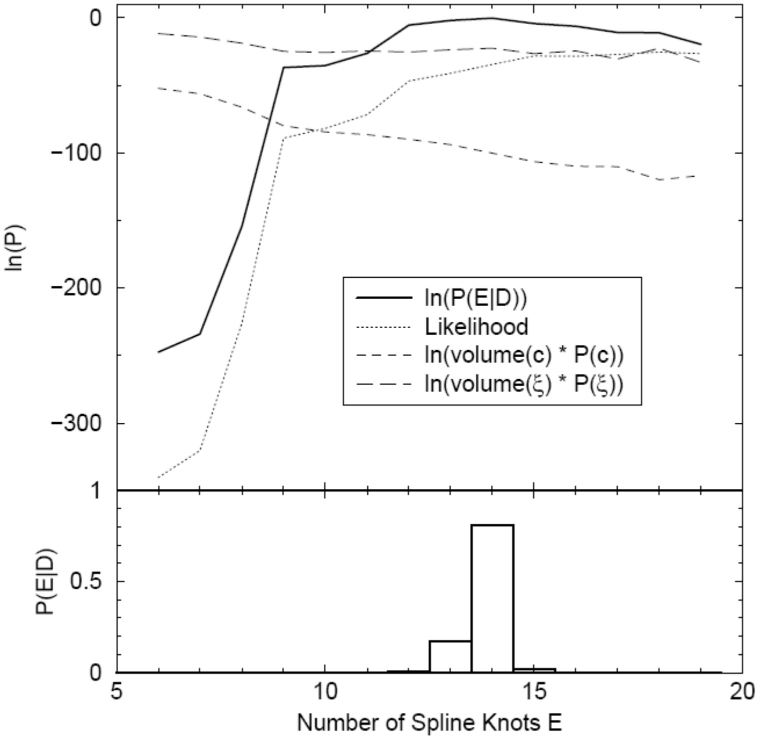


λ = 0.4 (best value)

Background estimation in spectral data

- Problem: estimate background for PIXE spectrum
- Approach is based on assuming background is smooth and treating resonances as outlying data
- Fully Bayesian calculation using MCMC to estimate spline parameters, their knot positions, and number of knots





from Fischer et al., Phys. Rev. E **61**, 1152 (2000)

Summary

In this tutorial:

- MCMC provides random draws from calculational pdf
- Metropolis algorithm
  - ▸ choosing the trial function
  - ▸ diagnositics
- Hamiltonian (hybrid) algorithm
  - ▸ potentially more efficient than Metropolis, provided ∇φ can be calculated as quickly as φ
- Examples:
  - ▸ analysis of Rossi traces; complex likelihood function
    - possibility of elaborating on model to include systematic effects
  - ▸ background estimation using splines and treating signal as outliers