Optimization

Given $f(\underline{x})$, find minimum value of f as a function of $\underline{x}$

Example

i) $f(\underline{x}) = x_1^2 + x_2^2$ (continuous variables)

minimum $x_1 = x_2 = 0$

ii) Let $i = 1 \ldots N$ label cities

$\eta_i p = 1$ if traveller at city i at time step $p = 1 \ldots N$

d_{ij} = distance from city i to city j. Minimize

$$D = \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{p=1}^{N-1} d_{ij}\, \eta_{ip}\, \eta_{jp+1}$$

Travelling Salesman Problem

(NP Complete) Discrete Variables

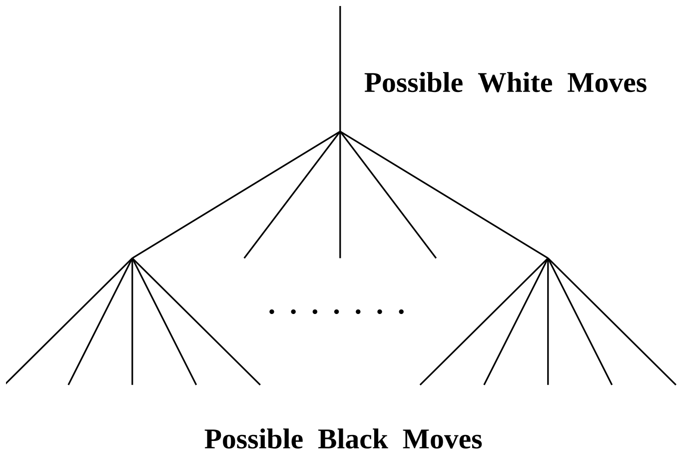
iii) Given a collection of molecules, find its ground state (minimize its energy as a function of atomic positions) and see if it will make a nice drug.

iv) Given a set of pixels, forming an image (as seen by your eyes), optimize your life based on information—pattern recognition/(artificial) intelligence (e.g., image contains a lion, run away image contains football thrown by Jim Kelly, catch it and run to goal line).

v) Given a set of ELINT (electronic intelligence = signals as a function of position), decide on best steps to win the battle—spatial reasoning or situation assessment.

vi) Minimize $\int \left(\nabla \oint\right)^2 d^3x$ which is variational principle that derives finite element method

vii) Given position on chess board, decide on best move. Write down a tree



and expand several "plies." Assign values to final position. Chosen move is one that leads to highest (best) value of worst possible scenario—minimax problem.

viii) Transportation

Take M warehouses and N retail outlets. Consider one product. We wish to minimize cost of shipping products from warehouses to outlets.

a_i units at warehouse i

b_j units needed at outlet j

x_{ij} units shipped from i to j at cost c_{ij} each

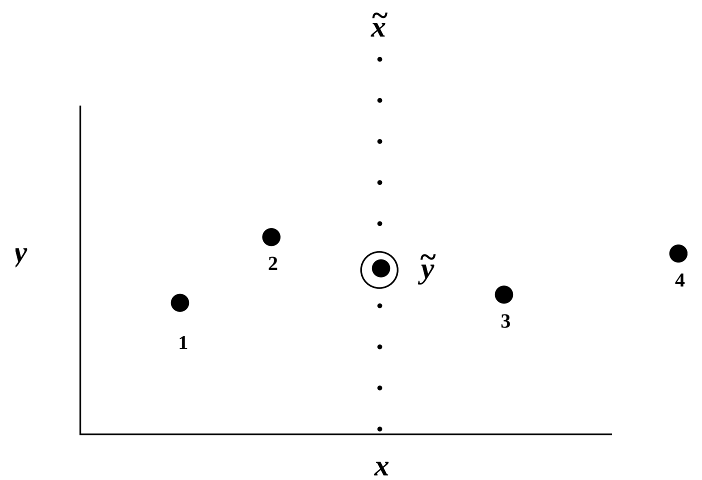
$$\text{constraints} \begin{cases} \sum_{j=1}^{N} x_{ij} & \leq a_i \\ \sum_{i=1}^{M} x_{ij} & = b_j \end{cases}$$

Minimize $\sum_{i=1}^{N} \sum_{j=1}^{M} c_{ij}\, x_{ij}$ objective function.

Linear Programming—objective function and constraints are linear

ix) Interpolation

Given a set of values $y = y_i$ at $x = x_i$, find the best estimate of $y = \tilde{y}$ at $x \geq \tilde{x}$.



Typically do by assuming $y(x)$ is a polynomial—maybe of degree 3 in this example.

Note that interpolation, like many optimization problems, is mathematically undefined. One can find polynomials of degree 4 taking any given values of x_1, x_2, x_3, x_4, and $\tilde{x}$. We can only determine value at $\tilde{x}$ by assuming polynomial of degree 3 (or other limiting assumption).

A Collection of Methods for Optimization

Maximum Likelihood

χ^2

Maximum Entropy

Combinational Optimization

α-β Pruning

Branch and Bound

Branch and Cut

Linear Programming

Nonlinear Programming

Integer Programming

Heuristic

Greedy Algorithm

Conjugate Gradient

Simulated Annealing

Simulated Tempering

Neural Networks

Deterministic Annealing

Genetic Algorithms

We will start with Maximum Likelihood and χ^2 illustrated by data analysis problem

x) Analysis of Experimental Data

Interpolation involves finding function values given function values at other points. This naturally extends to case where given function values have errors. We consider the general case where we have a set of observations x_1, $x_2 \ldots x_n$ and we wish to deduce something from it.

(a) x_i could be number of people claiming they will vote for Clinton and i could label either geographic position or time sample taken.

Goal: Deduce who will win Presidential election and what strategy changes are appropriate to affect result.

(b) x_i are measurements of particle positions in a detector at the proposed SSC—Superconducting Supercollidor

Goal: Deduce which is the correct theory of the fundamental interactions. What particles exist?

(c) x_i are separate measurements of the length of a rod which differ due to uncertainities in measurement process (measurer drunk, rod quivering in earthquake).

Goal: What is best estimate for length of rod and error therein.

In each case, the x_i's are related to "goal" (which can be qualitative or quantitative) by probability distributions. Why?

(a) Voting

We can associate with the voting process, a function V

$$\begin{aligned} V &= 0 && \text{doesn't vote} \\ &= 1 && \text{votes for Clinton} \\ &= 2 && \text{votes for Bush} \\ &= 3 && \text{votes for Perot, etc.} \end{aligned}$$

Each person is described by a set of attributes $\underline{\alpha}$

$$\alpha_1 = \text{age}$$
$$\alpha_2 = \text{income}$$
$$\alpha_3 = \text{place they live, etc.}$$

with a big enough size vector $\underline{\alpha}$

$$V \equiv V(\underline{\alpha})$$

Each person $\rightarrow \underline{\alpha}_i$ a particular vector.

If we were mathematicians, we could find $V(\underline{\alpha})$ from survey (determine parameters in it) and so predict election result.

But society is a "complex system." Impossible to write down some function V even though actions might be described by some sort of function.

$\Rightarrow$ use probabilistic approach.

(b) High Energy Physics

The events recorded are fundamentally probabilistic because quantum theory describes events this way, not deterministically. There are no "hidden variables."

We actually also see issues described under (a) and (c) here as well.

(c) Length of Rod

This is simplest case—should have been first example! Here, the length is well defined and uncertainties come from variations in measurement process itself.

The Computational and Numerical issues fall into three areas.

(1) The underlying mathematics is that of random variables. The key basic theorem is the CENTRAL LIMIT THEOREM—this is at the level of CPS615.

(2) There is a set of methods where we believe that we KNOW something precise about associated probability distributions.

LEAST SQUARES METHOD
MAXIMUM LIKELIHOOD METHOD

(3) There are a set of techniques for data exploration where we search a large dataset for expected or unexpected correlations and clusterings which are qualitative and usually not describable by mathematical functions.

HISTOGRAMS
SCATTER PLOTS

Data Analysis at the "SSC"—a typical problem which is rich enough to exhibit all these issues.

1)

$$\begin{array}{ccl} P \ \ \overline{P} & \rightarrow & N \text{ particles} \\ \rightarrow \ \leftarrow & & N = 2 \rightarrow 1000? \end{array}$$

Particles labelled by

$$\begin{array}{ccl} \text{type} & & \pi,\ V_2,\ P,\ \overline{P}\ldots \\ \text{momentum} & & \underline{P},\ E \\ \text{and energy} & & \\ \text{mass}^2 & = & E^2 - \underline{P}^2 \\ \uparrow & & \end{array}$$

essentially determined by type

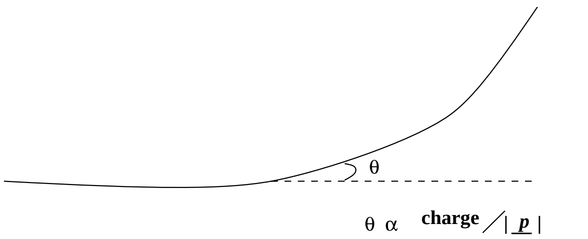
2) A typical experiment might record 10 million such events on some storage medium (e.g., ~ 3 events per second for a year).

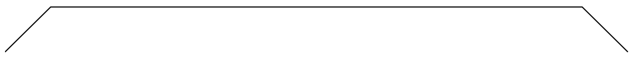
3) The events might be "unbiased", i.e., all possible events recorded or they may be selected by a "trigger."

The events are specified by a set of parameters.

N = number of particles for each particle, parameters defining them. This is a vector $\underline{y}$. Each event described by $\underline{y}_i$ produced with probability $P\left(\underline{y}_i\right)$

Any realistic detector is imperfect. "Image Processing" used to find tracks of charged particles. Magnets used to bend charged particles $\Rightarrow$ size of bend $\Rightarrow$ sign of charge, magnitude of momentum.





Uncertainties

1) Flaws in Image Processing

cf. SDI problem of tracking a bunch of missiles

"Mathematically illposed problem" use neural networks, etc.

2) Geometric problems

$\Rightarrow$ In some regions of $\underline{P}$ particles are "lost" with some probability.

Neutral particles only make tracks in dense material $\Rightarrow$ harder to measure and "destructive"

So first measure charged particles

then measure photons $\left(\pi^0\right)$

then measure $\mu^{\pm}$

cannot measure neutrinos

What you see = probability from Science convoluted (integrated) with probability from measurement probability from trigger.

Probability from Science contains functional dependences which are understood together with those that are not. ⇒ this is also a "complex system".

The observed features =

$$P\left(\underline{y}_i\right) = \int \quad \begin{array}{l} \text{known dependecies} \\ \\ \text{convoluted with unknown} \\ \text{dependencies} \\ \\ \text{PLUS} \\ \text{totally unknown} \\ \text{probabilities.} \end{array}$$

Can MODEL dependencies and probabilities (use suitably inspired = theoretically motivated "guesswork") ⇒ PARAMETERIZE ignorance or use approaches which do NOT involve parameterization.

Maximum Likelihood Principle

(i) Example:

(a) Rod Quivering in an Earthquake

$x_1, x_2 \dots x_n$ are n measurements of the length of a rod. Owing to a Gaussian earthquake (or a Gaussian i.e., normal drunkenness) occurring during measurement, the x_i are scattered about the true value: you visit your favorite theorist (or do yourself after sobering up) who believes that the probability distribution of x_i is Gaussian with mean α_0 = true length and standard deviation σ_0.

We may write down the (relative) probability of each measurement x_i as:

$$P_i\left(x_i\right) = c \exp\left\{-\frac{1}{2\sigma_0^2}\left(x_i - \alpha_0\right)^2\right\}$$

The Likelihood function for whole experiment of n events is defined as:

$$\begin{aligned} L\left(x_1 \dots x_n\right) &= \prod_{i=1}^{n} c \exp\left\{-\frac{1}{2\sigma_0^2}\left(x_i - \alpha_0\right)^2\right\} \\ &= \tilde{c} \exp\left\{-\frac{1}{2\sigma_0^2}\sum_i \left(x_i - \alpha_0\right)^2\right\} \end{aligned}$$

This is—up to a constant—the probability of the x_i given the true length α_0 and certain theoretical assumptions we represent as T.

$$L\left(x_1 \dots x_n\right) = \Pr\left\{x_1 \dots x_n | \alpha_0 , T\right\}$$

in conditional probability notation. T is, for instance, the notion that distribution Gaussian and the value of σ_0 taken (the letter being calculable from the strength of earthquake or amount of beer drunk).

(b) Use of Bayes Formula

Now what we really want is the probability distribution of an estimate α of α_0. We can formally get this by use of Bayes law: First extend $P(x_i) = \exp\left\{-\frac{1}{2\sigma_0^2}(x_i-\alpha)^2\right\}$ to be defined for any value of α: not necessarily $\alpha =$ true value α_0. Introduce:

$P(\alpha|x_i\,,T) =$ Probability true length is α given measurements x_i and predjudice T. Then Bayes law of conditional probability is

$$P(\alpha|x_i\,,T) = P(x_i|\alpha\,,T)\,\frac{P(\alpha|T)}{P(x_i|T)} \qquad (*)$$

This is the essence of the maximum likelihood principle. $(*)$ converts the likelihood $L(x_1 \ldots x_n) = P(x_i|\alpha\,,T)$ which is a function probability of x_i for fixed $\alpha\,,T$ into $P(\alpha|x_i\,,T)$ which is the desired probability distribution of α for given measurements x_i and predjudice T.

(c) Interpretation of $(*)$

(α) We can ignore $P(x_i|T)$ because it is independent of α and fixed for our given x_i and T. It has the practical point that Likelihood is never normalized sensibly.

(β) $P(\alpha|T)$ philosophically represents a priori knowledge of α before experiment (measurement of x_i) performed. This is shaky—not because we don't really know how to define what $|T$ ("given T") means—but, for instance, suppose we wish to parameterize "complete ignorance" of α before experiment performed.

Natural is $P(\alpha|T) =$ constant

But suppose we put $\beta = \alpha^2$ and then β has a priori probability $\varphi(\beta|T)$

$$\varphi(\beta|T)d\beta = P(\alpha|T)d\alpha$$

i.e., $\varphi(\beta|T) = \frac{1}{2\sqrt{\beta}} \neq$ constant.

So the concept of "complete ignorance" is not well defined. Jeffrey (a big bad English-Oxford-book) spends 100's of pages trying to avoid this difficulty. His efforts are so much garbage.

Fortunately, there is no trouble for large n because of both an obvious numerical reason (see next section (i) (d)) and a not so obvious mathematical reason.

(d) The Likelihood Maximizes

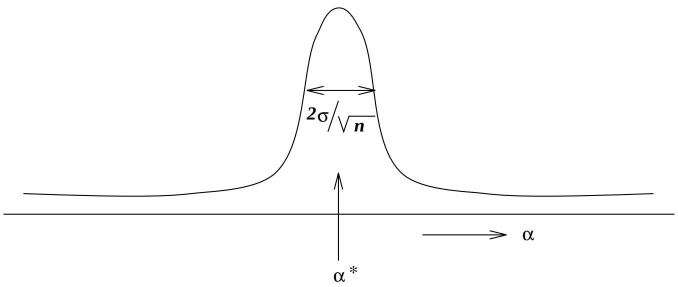
Consider, again, the explicit form of the Likelihood of a rod quivering in the earthquake.

$$L(x_1 \ldots x_n) = c\,\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\alpha)\right\}^2$$

According to (D4.1), we now regard L as a function of α and so rewrite it:

$$L(\alpha) = c \exp \left\{ -\frac{n}{2\sigma^2} \left(\alpha - \frac{1}{n} \sum_{i=1}^{n} x_i \right)^2 - \frac{1}{2\sigma^2} \left[\sum_i x_i^2 - \frac{1}{n} \left(\sum_i x_i \right)^2 \right] \right\}$$

The last term is a constant independent of α, and so as a function of α, L is just a Gaussian which maximizes at $\alpha = \frac{1}{n} \sum\limits_{i=1}^{n} x_i$ and has width $\frac{\sigma}{\sqrt{n}}$.



So as $L = \Pr(\alpha | x_i, T)$ we say that maximum likelihood estimate of α is $\alpha^* = \frac{1}{n} \sum\limits_{i=1}^{n} x_i$ and its standard deviation is $\sigma / \sqrt{n}$.

Note that whatever $P(\alpha)$ you put in the shape of $LP(\alpha|T)$ is essentially independent of P as $n \to \infty$ and we get a Gaussian peaked at α^* riding on top of constant value $P(\alpha^*|T)$.

(ii) General Maximum Likelihood Method

The above is glib in many ways—especially in its treatment of errors. We will repair this later with real live theorems. First, we state the obvious generalization.

The general maximum likelihood method is to put $P(\alpha|T) = 1$ in (B1) and form $L(\alpha) = \prod_{i=1}^{n} P(x_i, \alpha)$ for any measurement x_i whose probability $P(x_i, \alpha)$ depends on value x_i and theoretical parameter(s) α.

The maximum likelihood estimate of α is $\alpha = \alpha^*$ where $L(\alpha^*)$ is a maximum. The error $\Delta\alpha$ is the standard deviation of the distribution of L as a function of α, i.e.,

$$L(\alpha) = \exp \left\{ \text{const.} - \frac{1}{2\Delta\alpha^2} (\alpha - \alpha^*)^2 \right\}$$

Note usually manipulate

$$-\ln L = -\sum_{i=1}^{n} \ln P(\underline{x}_i, \underline{\alpha})$$

Lifetime Example

Physics texts give the example of determining the lifetime τ_0 of a particle from the observation of decay times $t_1 \ldots t_n$ of n decays.

The probability distribution is Poisson:

$$p(t)dt = \frac{1}{\tau}\exp\left(-\frac{t}{\tau}\right)dt$$

Now dt is meaningless—it is independent of τ and only affects normalization of likelihood. The unknown $P(\alpha|T)$, $P(x_i|T)$ terms in (B1) made this arbitrary anyway!

Likelihood $L(\tau) \propto \frac{1}{\tau}e^{-\frac{t_1}{\tau}} \ldots \frac{1}{\tau}e^{-\frac{t_n}{\tau}}$ or $L(\tau) = \exp\left[-\frac{1}{\tau}\sum_{i=1}^{n} t_i - n\ln\tau\right]$ where we lazily convert proportionality into an $=$ sign. Maximum is

$$\frac{dL}{d\tau} = 0 \qquad \text{or}$$

$$\frac{1}{\tau^2}\sum_{i=1}^{n} t_i - \frac{n}{\tau^*} = 0$$

$$\underline{\text{or}} \qquad \tau^* = \frac{1}{n}\sum_{i=1}^{n} t_i$$

or the maximum likelihood estimate is just the mean: such a simple result is only true because of particular Poisson example.

In this special case, we can illustrate that maximum likelihood method does give results which "converge" as $n \to \infty$.

According to the central limit theorem, the variable $y = \frac{1}{n}\sum_{i=1}^{n} t_i$ is a sum over independent random variables and has mean $\langle y\rangle = \tau_0$ and standard deviation σ

$$\begin{aligned} \sigma^2 &= \frac{1}{n}\left\{\left\langle t_i^2\right\rangle - \langle t_i\rangle^2\right\} \\ &= \frac{1}{n}\left\{\frac{1}{\tau_0}\int t^2 e^{-\frac{t}{\tau_0}}\,dt - \tau_0^2\right\} \\ \text{or}\quad \sigma &= \frac{\tau_0}{\sqrt{n}} \end{aligned}$$

(v) Multiplication of Experiments

If we have two experiments that are independent but depend on same theoretical parameter α, then

$$L_1 = \prod_{i=1}^{n_1} p\left(x_i\,,\alpha\right) \qquad \text{is likelihood of experiment \#1}$$

$$L_2 = \prod_{i=1}^{n_2} q\left(y_i\,,\alpha\right) \qquad \text{is likelihood of experiment \#2}$$

and clearly the combined likelihood is

$$L = L_1\ L_2$$

i.e., one should multiply likelihoods when combining experiments. We will now apply this to case when each L_i <u>is</u> Gaussian in α.

(vi) General Theorem

Note that one can prove that Maximum Likelihood is asymptotically the "best algorithm," i.e., in limit $n \to \infty$ (n total events, there is NO method which will give a smaller standard deviation).

Also, Maximum Likelihood method will give one the "right answer," i.e., Bias = Correct Estimate—Maximum Likelihood Estimate $\to 0$ (like $1/n$) as $n \Rightarrow \infty$.

χ^2 Method

This is standard and fully, correct (best) method for combining results from several experiments. It is in fact used (because it's easy conceptually and for computer) even when it's inapplicable (e.g., if there are correlations between supposed "independent" random variables or if the basic experiments have too few events to be Gaussian).

The problem is: Given N experimental measurements x_i and errors σ_i plus estimates ξ_i of the x_i which are functions of n theoretical parameters. Then the probability of each observation is

$$P_i = \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left\{ -\frac{(x_i - \xi_i)^2}{2\sigma^2} \right\}$$

The Likelihood

$$\begin{aligned} L &= \prod_{i=1}^{N} P_i \\ &\propto \exp\left\{ -\sum_{i=1}^{N} \frac{(x_i - \xi_i)^2}{2\sigma_i^2} \right\} \end{aligned}$$

so, the principle of maximizing the likelihood is equivalent to minimizing χ^2 where

$$\chi^2 = \sum_{i=1}^{N} \frac{(x_i - \xi_i)^2}{\sigma_i^2}$$

with respect to n theoretical parameters in ξ_i.

(i) Notice that χ^2 is the sum $\sum\limits_{i=1}^{N} v_i^2$ where v_i are independent Gaussian random variables of unit standard deviation and zero mean. We <u>can</u> calculate χ^2 distribution exactly, but in practice this is rarely necessary. Thus, we can easily show that—for fixed ξ_i—that

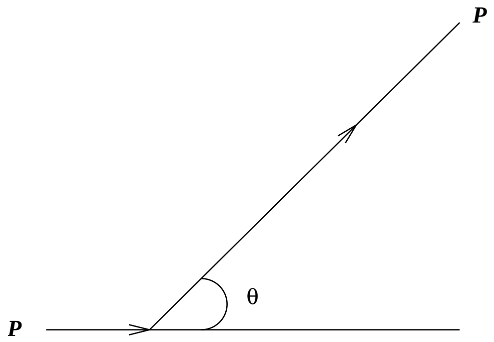
$$\begin{aligned} \left\langle \chi^2 \right\rangle &= N \\ \sigma_{\chi^2}^2 &= 2N \end{aligned} \qquad (**)$$

and this is sufficient for large N as central limit theorem says χ^2 will be Gaussian with above mean and standard deviation.

(ii) Now (**) is not quite correct because if we determine the theoretical parameters to minimize χ^2 then implicitly ξ_i are also random variables that are functions of x_j. It is not too hard to show that in this case, one should just replace N by $N - n$ (number of degrees of freedom) in (**). In any case, (**) is very important because it allows an easy criterion for the goodness of fit. Thus, if value of χ^2 is many (say > 2) standard deviations away from its mean, then the theoretical form is in doubt. The usual maximum likelihood method does not have this advantage: Remember, L was unnormalized and so <u>its</u> value had no significance for the goodness of fit. This is quite a difficult problem.

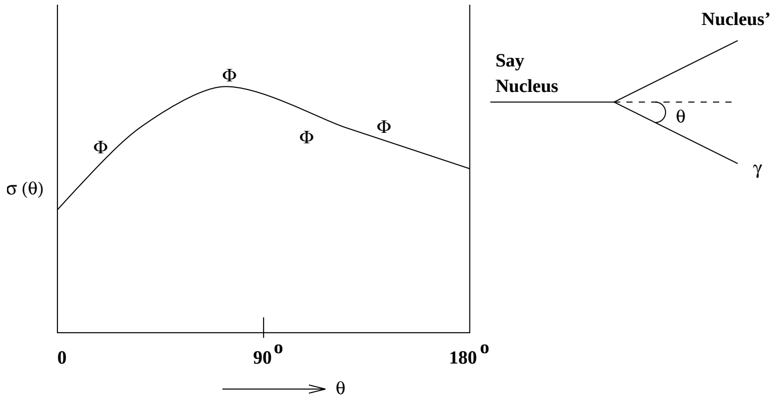
Example:

A physics experiment measures a scattering cross-section σ as a function of angle θ.



In the case of decay (of nucleus), we know rigorously that angular $\sigma(\theta)$ is a polynomial in $\cos\theta$ with highest degree $\leq 2J$ (spin of nucleus)

$$\sigma(\theta) = a_0 + a_1 \cos\theta + a_2 \cos^2\theta + a_3 \cos^3\theta$$



The measurements x_i's are cross sections and the theoretical predictions ξ_i are linear in unknown parameters a_j.

$$\xi_i = \sum_{m=1}^{n} c_{im}\, a_m$$

$$\text{minimize: } \sum_{i=1}^{N} \frac{(x_i - \xi_i)^2}{\sigma_i^2}$$

$$x_i = \text{const. } N_i$$

$$\sigma_i = \text{const. } \sqrt{N_i} \qquad \sigma_i = \frac{x_i}{\sqrt{N_i}}$$

where ith bin has N_i events.

$$\sum_{i=1}^{N} \frac{(x_i - \xi_i)}{\sigma_i^2} \frac{\partial \xi_i}{\partial a_m} = 0 \qquad \text{or}$$

$$\sum_l A_{ml}\, a_l^* = b_m$$

* denotes χ^2 or maximum likelihood estimate.

$$\begin{aligned} \chi^2 \text{ matrix} &= \sum_i \frac{C_{im}\, C_{il}}{\sigma_i^2} \qquad \text{Positive Symmetric Matrix} \\ &= \sum_i \frac{(\partial \xi_i / \partial a_m)\ (\partial \xi_i / \partial a_l)}{\sigma_i^2} \\ b_m &= \sum_i \frac{x_i C_{im}}{\sigma_i^2} = \sum_i \frac{x_i}{\sigma_i^2} \frac{\partial \xi_i}{\partial a_m} \end{aligned}$$

Therefore, $\underline{a}^* = A^{-1} \underline{b}$.

We can find errors in either of the two ways

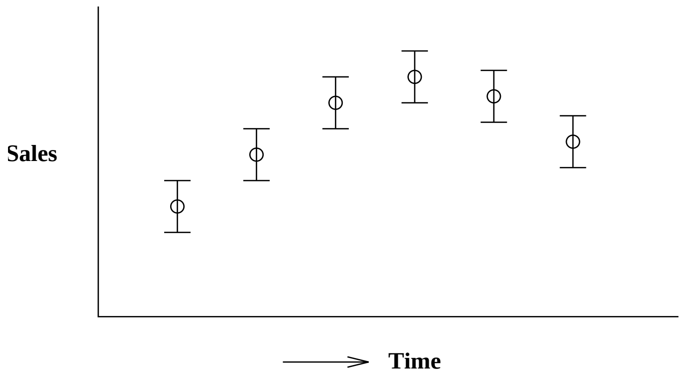
a) regard $\underline{a}^*$ as a function of x_i and calculate standard deviation from central limit theorem.

b) shape of likelihood

$$L = \exp\left(-\frac{1}{2}\chi^2\right)$$

We have – as far as shape of L is concerned. $L = \exp$ (Independent of a plus no terms linear in $a - a^*$ due to "maximum" condition $-1/2\,(\underline{a} - \underline{a}^*)^T A\,(\underline{a} - \underline{a}^*)$), i.e., A is error matrix for a.

In typical regression analysis the ξ_i are also linear in the unknowns, e.g.,



$$\text{Sales}(t) = \sum_{i=1}^{m} x_i P_i(t)$$

where x_i is unknown and P_i is known expansion function.

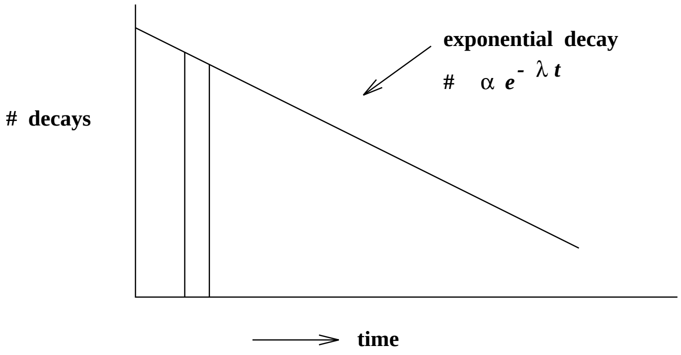
e.g., $P_1(t) = 1$

$P_2(t) = (t - 1992)$

$P_3(t) = (t - 1992)^2$, etc.

$P_k(t)$ isn't linear in t, but Sales(t) is linear in $P_k(t)$ as this is what counts!

However, we consider nuclear decay



e_i = number of decays observed in time interval $t_0 + (i-1)\delta t$ to $t_0 + i\,\delta t$.

$$\xi_i = c\,\exp\left(-\lambda[t_o + (i-1/2)\delta t\right)$$

is a function of two unknowns, c and λ. ξ is linear in c but nonlinear in λ. Note we can analyze nuclear decays in two ways

a) Maximum Likelihood—

i labels events $\quad i = 1 \ldots n_{ev}$

Minimize—$\ln L$

$$L = \prod_{i=1}^{n_{ev}} \frac{1}{\tau} c\exp\left(-\lambda t_i\right)$$

b) χ^2

i labels bins $\quad i = 1 \ldots N_{\text{bin}}$

Minimize χ^2

$$\chi^2 = \sum_{i=1}^{N_{\text{bin}}} \frac{(\xi_i N_i - \xi_i)^2}{\sigma_i^2}$$

where $\sigma_i \sim \sqrt{N_i}$ and N_i is number of events in ith bin $(t_0 + (i-1)\delta t \leq t \leq t_0)$. If N_i large, N_i will be Gaussian, whatever distribution of individual events.

Which is best? Likelihood method is "best," but χ^2 method is computationally easier as $N_{\text{bin}} << N_{\text{ev}}$.

Typically, as number of events in each bin must be large

$$N_{\text{bin}} \sim \frac{N_{\text{ev}}}{50}\,.$$

χ^2 method will give an estimate with a larger standard deviation, but the difference will not be large if $\lambda\delta t$ is small, so probabilities $\exp(-\lambda t)$ do not vary across bin. So, can choose correctly

Small total number of events—no good method.

Large number of events–Maximum Likelihood works and is best.

Very large number of events—Maximum Likelihood and χ^2 works, but χ^2 will give a good answer with less computational complexity.

Generally, we analyze

$$\begin{aligned}\chi^2 &= \sum_{i=1}^{N} \left(\frac{e_i - \xi_i}{\sigma_i}\right)^2 \\ &= \sum_{i=1}^{N} \mathrm{res}_i^2 \qquad \leftarrow \text{scaled residual}\end{aligned}$$

which shows why χ^2 is also called least squares fit. Expand ξ_i about $\underline{x}_0$

$$\begin{aligned}\xi\left(\underline{x}\right) = \xi\left(\underline{x}_0\right) + \left(\underline{x} - \underline{x}_0\right)^T \frac{\partial \xi}{\partial \underline{x}} \\ + \text{Higher order in}\left(\underline{x} - \underline{x}_0\right)\end{aligned}$$

In this approximation, we now get "standard linear formulae."

$$\begin{aligned}\chi^2\left(\underline{x}\right) = \chi^2\left(\underline{x}_0\right) + \frac{1}{2}\left(\underline{x} - \underline{x}_0\right)^T A\left(\underline{x} - \underline{x}_0\right) \\ - \left(\underline{x} - \underline{x}_0\right)^T \underline{b} + \text{corrections}\end{aligned}$$

Minimize χ^2 ignoring corrections gives

$$\underline{\underline{A}}\left(\underline{x} - \underline{x}_0\right) = \underline{b}$$

which is same equation we found for iterative solution of P.D.E. **BUT** only exact when ξ_i LINEAR in unknown parameters $\underline{x}$.

Nonlinear χ^2 Minimization

$\underline{\underline{A}}$ is symmetric positive definite

$$\xi_i = \xi_i\left(\underline{x}_0\right) + \sum_{d=1}^{m} t_{id}\left(x - x_0\right)_d$$

$$\text{where } t_{id} = \frac{\partial \xi_i}{\partial x_d}$$

$$A_{dd'} = \sum_i t_{id}\, t_{id'}$$

Basic idea:

Think of initial $\underline{x}_0$

$$\underline{\underline{A}}\left(\underline{x}_{n+1} - \underline{x}_n\right) = \underline{b}$$

Calculate $\underline{\underline{A}}$ and $\underline{b}$ using $\underline{x} = \underline{x}_n$.

$$\underline{x}_{n+1} = \underline{x}_n + A^{-1} b$$

As A positive definite, inverse always exists.

Why doesn't this work?

$$\underbrace{\begin{bmatrix} \lambda_1 & & & & \\ & \cdot & & & \\ & & \cdot & & \\ & & & \cdot & \\ & & & & \lambda_m \end{bmatrix}}_{\text{Diagonalize}} \overbrace{\begin{bmatrix} \delta\tilde{x}_1 \\ \\ \\ \\ \delta\tilde{x}_m \end{bmatrix}}^{\text{Transformed}} = \begin{bmatrix} b_1 \\ \\ \\ \\ b_m \end{bmatrix}$$

$\lambda_1 > \ldots > \lambda_m$.

My experience at $m \sim 30$, $\lambda_m \sim 10^{-8}\lambda_1$.

But we made approximation and so λ_m term swamped by corrections neglected in expanding ξ_i linearly. Equation gives big

$$\delta\tilde{x}_m = \lambda_m^{-1} b_m$$

Better

$$\delta\tilde{x}_m = 0$$

And, $\chi^2\left(\underline{x}_{n+1}\right) >> \chi^2\left(\underline{x}_n\right)$. Not $\chi^2\left(x_{n+1}\right) < \chi^2\left(\underline{x}_n\right)$

Solutions to this problem are

1) Conjugate Gradient, but note matrix no longer sparse and so iterative approach not so natural.
2) Marquardt's method

$$\underline{\underline{A}} \rightarrow \underline{\underline{A}} + QI$$

where Q is Marquardt's Parameter

$$\lambda_i \rightarrow \lambda_i + Q$$

Suppose, $Q \sim \sqrt{\lambda_1 \lambda_m}$. Then shifts $\delta\tilde{x}_d$ are unchanged for large λ_d, small for small λ_d.

Increase Q if χ^2 increases.

Decrease Q if χ^2 decreases.

Linear Programming

Minimize $z = \sum_{j=1}^{r} c_j x_j$ where $x_j \geq 0$ and there are Linear Constraints

$$\sum_{j=1}^{r} a_{ij}\, x_j \qquad \leq, \geq, \text{or} = b_i \qquad i = 1 \ldots m$$

Introduce new variables to convert **all** constraints to equalities.

Suppose $i = 1$ is $\geq$. Write $i = 1$ constraint as

$$x_{r+1} = \sum_{j=1}^{r} a_{ij} x_j - b_1 \qquad \geq 0$$

$$\sum_{j=1}^{r} a_{ij} x_j - x_{r+1} = b_1$$

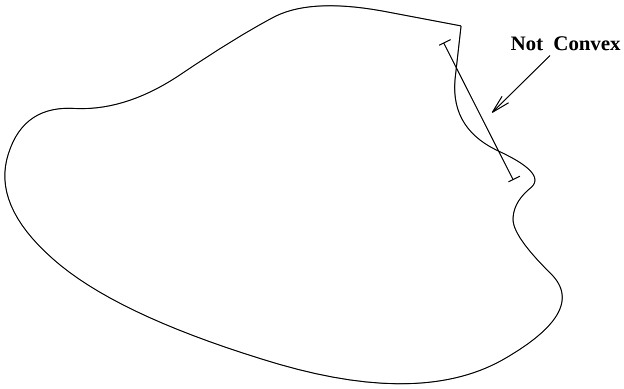
$$a_{1r+1} = -1$$

$$c_{r+1} = 0$$

So linear programming becomes:

Minimize $z = \underline{cx}$

Subject to $\underline{\underline{A}}\underline{x} = \underline{b}$ all components of $\underline{x}$ are positive. Set of $\underline{x}$ satisfying constraints are convex region.



$\underline{x}_1 \underline{x}_2$ in region.

$\lambda \underline{x}_1 + (1-\lambda)\underline{x}_2$ in region.

Extreme point of convex region is one where

$$x_{\text{extreme}} \neq \lambda \underline{x}_1 + (1-\lambda)\underline{x}_2$$

where $\lambda \neq 0, 1$.

One can show that the optimal solution (if it exists) is an extreme point of the convex region.

Consider $\underline{\underline{A}}\underline{x} = \underline{b}$. $\underline{\underline{A}}$ is $m \times r$ matrix. Usually $r >> m$.

Can show that extreme points satisfy

m nonzero elements in $\underline{x}$

$n - m$ zero elements in $\underline{x}$

Choose columns so non-zero elements are first m entries in $\underline{x}$

$$\underline{\underline{A}}^{\text{reduced}} \underline{x} = \underline{b}$$

where $\underline{\underline{A}}$ is $m \times m$.

Simplex method is a strategy for moving from one solution $\underline{x}$ extreme to another $\underline{x}$ extreme by changing one column of $\underline{\underline{A}}^{\text{reduced}}$ to another that is not being used.

There is another class of methods—associated with Karmarkar at AT and T which are so called interior point methods. That is, you do not go from vertex to vertex of simplex but rather wander around inside simplex. Both simplex and interior points are important.

Computational Issues in χ^2 and Maximum Likelihood

Maximum Likelihood

1. Complex nonlinear (in dependence of parameters) function to maximize
2. Each function value
$$-\ln L = -\sum_{i=1}^{N} \ln p\left(\underline{x}_i, \underline{\alpha}\right)$$
(where $\underline{x}_i$ determines events and $\underline{\alpha}$ are unknown parameters) is sum of N independent calculations which is natural form of parallelism
3. Parallelism in optimization (in space of $\underline{\alpha}$) depends on particular minimization approach adopted. Also depends on complexity of calculation
$$\ln p\left(\underline{x}_i, \underline{\alpha}\right)$$
$$\frac{\partial}{\partial \underline{\alpha}} \ln p\left(\underline{x}_i, \underline{\alpha}\right)$$
versus complexity of manipulation of $-LnL\left(\underline{\alpha}\right)$ as a function of $\underline{\alpha}$

   χ^2 and Likelihood are similar in aspects of analysis.

The arithmetic in "$\underline{\alpha}$ world" is that of full matrices which are $n \times n$ where n is number of parameters.

In "linear" case (ξ_i linear in components of $\underline{\alpha}$, χ^2 quadratic in components), minimization requires single solution of equations
$$\underline{\underline{M}}\alpha = \underline{b} \qquad (**)$$

In "nonlinear" case (ξ_i nonlinear in α), one must iteratively solve (**).

Note that one expands ξ_i up to linear term which is a quadratic approximations to χ^2. Thus, need to calculate value
$$\frac{\partial}{\partial \alpha_k} \text{ value} \qquad \text{NOT } \frac{\partial^2}{\partial \alpha_k \, \partial \alpha_j} \text{ value}$$

$$\begin{aligned} \text{value} &= \ln p\left(\underline{x}_i\right) \text{Likelihood} \\ &= \xi_i\left(\underline{x}_i\right) \chi^2 \end{aligned}$$

Often, values are formed inside loops. Inside loop over events

```
      DO 1  iparm = 1, nparm
1             deriv(iparm) = 0
              value = 0
      DO 2  ipart = 1


              value = value + part (ipart)
      DO 2  iparm = 1, nparm
2     deriv(iparm) = deriv(iparm) + ∂/∂σ_iparm part(ipart)
Inside loop over events
```

This code can get rather tedious. `ADIFOR` formally differentiates "existing" FORTRAN code and uses "chain rule" to generate derivative calculations.

Unclear how generally this works.

Maximum Likelihood VERY VERY expensive as

$$\sum_{i=1}^{\#\text{ events}}$$

remember that the number of events could be 10^7 at "SSC". χ^2 is cheaper as

$$\sum_{i=1}^{\#\text{ experiments}}$$

Usually, number of experiments measured in $10 \rightarrow 10^3$.

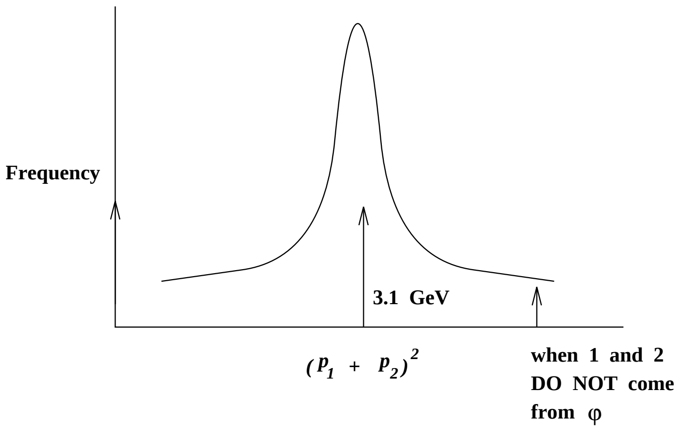
Histograms

High energy physicists like to find new particles.

For example, Nobel prize was given for the discover of ψ particle which is formed from "charm" and "anticharm" quarks.

Mass of $\varphi = 3.1$ GeV $= (p_1 + p_2)^2$ where φ decays into particle 1 plus particle 2.

We measure p_1 and p_2 as well as lots of other things



There are an accumulation of events with $(p_1 + p_2)^2$ near 3.1 GeV.

We can see with a histogram. Let

$$M_0^2 \leq (p_1 + p_2)^2 \leq M_1^2$$

be a region of interest.

We have B Bins

$$\Delta M^2 = \frac{\left(M_1^2 - M_0^2\right)}{B}$$

Let l th bin lie between $M_0^2 + (l_1)\,\Delta M^2$ and $M_0^2 + l\Delta M^2$. $l = 1 \ldots B$.
Define Random Variable

$$\begin{aligned} \eta_l\ (\text{event}) &= 1 \ \ \text{event lies in } l \text{ th bin} \\ &= 0 \ \ \text{event lies outside } l \text{ th bin} \end{aligned}$$

Let $\langle \eta_l \rangle = \mu_l =$ probability that an event will fall in bin.

$$\begin{aligned} \langle \eta_l \rangle &= (1 - \mu_l) \cdot \mathrm{O} + \mu_l \cdot 1 = \mu_l \\ \langle \eta_l^2 \rangle &= (1 - \mu_l) \cdot \mathrm{O}^2 + \mu_l \cdot 1^2 = \mu_l \\ \langle \eta_l^2 \rangle - \langle \eta_l \rangle^2 &= \mu_l - \mu_l^2 \\ &= \mu_l \quad \text{if} \quad \mu_l << 1 \end{aligned}$$

Standard deviation

$$\sigma_l = \sqrt{\mu_l} = \sqrt{\text{mean}_l}$$

Define

$$\begin{aligned} N_l &= \sum_{i=1}^{\#\ \text{events}} \eta_l \ \ (\text{event } i) \\ &= \#\ \text{events in } l\text{th bin} \end{aligned}$$

The central limit applies to

$$\begin{aligned} f_l &= \frac{N_l}{N_{\text{tot}}} \leftarrow \#\ \text{events} \\ \langle f_l \rangle &= \mu_l \\ \langle \sigma_{fl} \rangle &= \frac{\langle \sigma_{\eta l} \rangle}{\sqrt{N_{\text{tot}}}} \end{aligned}$$

N_l is

a) GAUSSIAN

b) Standard deviation

$$\begin{aligned} &= \sqrt{N_{\text{tot}}} \langle \sigma_{\eta_l} \rangle = \sqrt{N_{\text{tot}}\ \text{mean}\ (\eta_l)} \\ &= \sqrt{\langle N_l \rangle} \end{aligned}$$

This is a very very important rule in all SAMPLING EXPERIMENTS.

e.g.,

$l = 1$ vote for Clinton

$l = 2$ vote for Perot, etc.

In a poll of 1,000 people, one finds that 25% are in lth bin (i.e., will vote for xyz).

$$\begin{aligned} N_l &= 250 \\ \sqrt{N_l} &\approx 16 \end{aligned}$$

Thus, error in 25% is <u>16%</u> STATISTICAL.

As we have mentioned, the systematic bias (is "1,000" chosen people representative) is a harder problem! As $0.25 = \mu_l$ not that small, should reduces statistical error by

$$\sqrt{\frac{0.25 - 0.25^2}{0.25}} = \sqrt{\frac{3}{4}} = 0.86$$

Histograms are unbiased ways of analyzing data. One does NOT need to know dependence of data on $(p_1 + p_2)^2$.

General problem is data exploration. We have a set of events in a space of many dimensions

High Energy Physics	p_1	p_2	p_3	p_4	p_5	$\cdots$
Particle Type	1	2	3	4	5	etc.

Political Pools

(Unknown) parameters affecting a person's choice for president.

In general histogram,

1. Select subspace

   e.g., all events with > 30 particles or all events with a lepton.

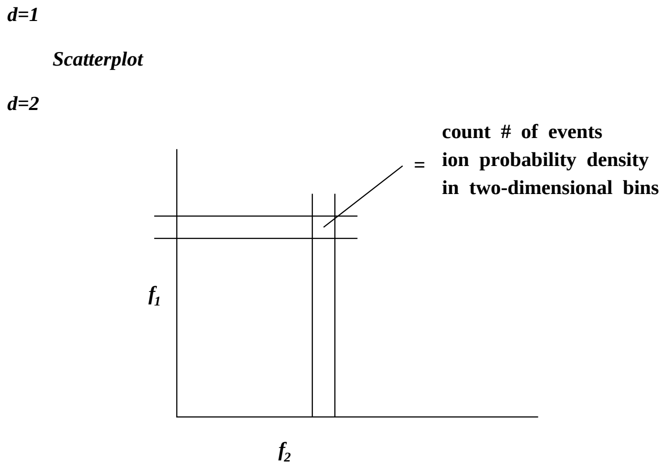
   e.g., all people in phone book whose surname starts with Kr.

2. Choose some d functions f_d (event).

$$d = 1 \qquad f_1 = (p_1 + p_2)^2$$
$$\text{or } f_1 = \text{choice for President}$$

3. Set up set of bins—convectionally uniform in each f_d.

Histogram



e.g., f_1 = Religion and f_2 = # of z's in name would be correlated (Catholic Church strong in Poland?)

e.g., $f_2 =$ Maximum transverse momentum and $f_1 = (p_1 + p_2)^2$ would be correlated.

ψ's tend to be produced in high transverse momentum (violent) collisions.

$d - 3$. The future with VR!

Projects—Maximum Likelihood

I Produce a general χ^2 or maximum likelihood data analysis system/machine in either explicit message passing "CM Fortran—FortranD—HPF". Apply to a simple example. You can use Marquardt's method for minimization (see CPS615 last semester, gcf records (mine performed successfully $10^6 \chi^2$ fits), Numerical Recipes).

Initialize Parameters

For each parameter choice
Initialize data structures
Loop over events
CALL USER1 (I'th event)
accumulate in data structures
Find new parameter values
Are we finished?

Separately or as follow-on to I.

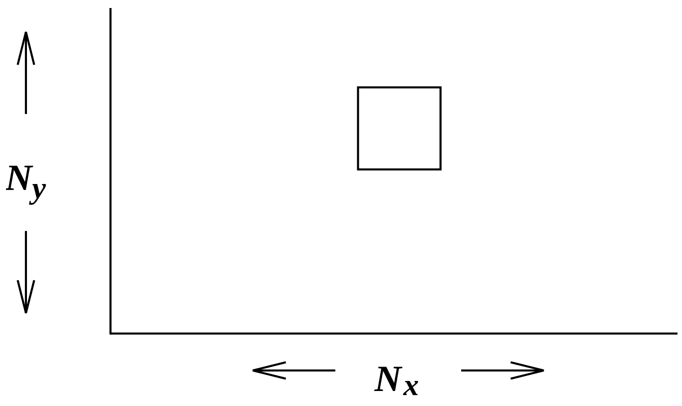
II Perform fit to specific dataset. Simplest: Generate "fake" data, e.g.,

$$f(x,y) = A \exp\left[-\frac{(x-\alpha)^2}{2\gamma^2} - \frac{(y-\beta)^2}{2\delta^2}\right]$$

Parameters A, α, β, γ, δ events—choose x_i, y_i randomly.

$$P\,(\underline{x}) \;\propto\; F(x,y)$$

χ^2 — divide two-dimensional plane into subsquares (see figure) $N_x N_y$ bins. Label i



Number of events in each bin is

$$
\begin{aligned}
N_{\text{expected}} &\propto \int_{\text{bin}} f(x,y)\,dx\,dy \\
&\sim f\left(x_{\text{central}}, y_{\text{central}}\right) dx\,dy
\end{aligned}
$$

$$\text{error} \quad \sqrt{N_{\text{expected}}}$$

Assume Total number of events large. Put experimental measurement as what you get from Gaussian distribution

mean N_{expected}

error $\sqrt{N_{\text{expected}}}$

using chosen values A, α, β, γ, δ. Fit to general form and see that you find "chosen" values.

III Find an interesting data set

— High Energy Physics

— Stock Option Pricing tapes from Kim Mills/Gang Cheng

Simplest: each event/χ^2 bin on a separate node of a MIMD parallel machine.

Complex: allow parallelism in calculation of USER1, e.g., run later on DECmpp

IV Compare χ^2 with maximum likelihood

Performance vs. Accuracy Tradeoff

V Integrate into AVS

VI Evaluate High Performance Fortran

Projects—Data Exploration

I Produce a generic histogram/scatterplot/other "robust" data analysis methods (to be discussed later in course) package on a parallel machine.

This should be structured

Initialize Datastructures

(User defines histograms, etc.)

Loop over events

CALL USER2

This has well defined ways of storing data in histograms/scatterplots

e.g., CALL HIST (x, Fred) deposits value x in histogram whose name is Fred

End Loop

Print out histograms

A Parallel Histogram is a simple database problem. Replicate in each node and combine at end. Scatterplots may be too big? Implement parallelism for histograms and parallelism for event generation.

II Test approach with some examples

“Fake data” as with $f(x, y)$ before

Real data as before

Theoretical calculation as in “Quantum Monte Carlo” (Cornell expertise) in Chemistry and Physics.

III Evaluate High Performance Fortran

IV Integrate into AVS

V Integrate with visual data exploration—user is walking through a multi-dimensional space looking for expected or unexpected structure.

Histograms/SP $\rightarrow$ 1 or 2 dimensions. There is a natural 3D extension which deserves consideration. What world to be explored has much higher dimensionality?

IV Use Mathematica as front end to define selection of subspaces.