Statistics

Lecture 2

August 7, 2000

Frank Porter

Caltech

The plan for these lectures:

- The Fundamentals; Point Estimation
- Maximum Likelihood, Least Squares and All That
- What is a Confidence Interval?
- Interval Estimation
- Monte Carlo Methods

Additional topics will be covered by Roger Barlow and Norman Graf.

Starting Comments

Context of today's discussion will be what is now called "Classical Statistics", with a frequency interpretation.

This is the appropriate context for summarizing information content. Thus, we will define an "Information Number" associated with a statistic.

Today's discussion continues from last time on the subject of point estimation. We'll take up intervals tomorrow.

Likelihood Function

Likelihood function: If an experiment has been performed to obtain a measurement x, drawn from some probability distribution with population parameter θ, the Likelihood Function, $L(x;\theta)$, for that experiment is defined as the probability (density) evaluated at the observed value of x.

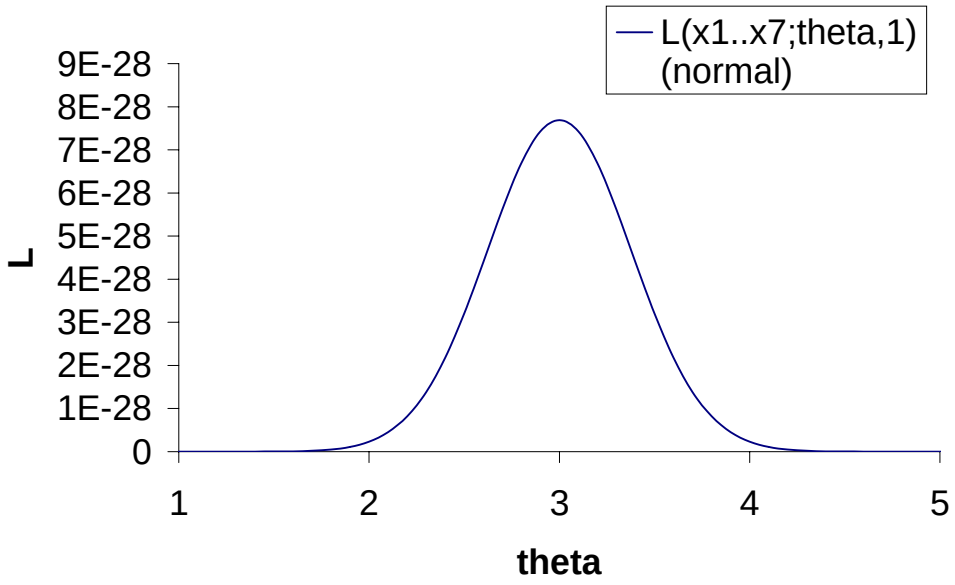
The likelihood function is often denoted by $\mathcal{L}(\theta;x)$, and is treated as a function of θ in some algorithms for making inferences concerning θ.

Considerations of Bayesian vs. classical statistics are irrelevant in this definition of the likelihood function.

Likelihood Function – Examples

Example I: Let **x** represent n independent samplings from a normal distribution, with mean θ and standard deviation σ. The likelihood function for such an "experiment" is:

$$L(\mathbf{x};\theta,\sigma) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\theta)^2\right].$$



Likelihood Function – Examples (cont.)

Example II: Suppose we do a "counting" experiment, in which we have a set (e.g., an array) of k independent counters which we run for some time. The observations then are $\{n_1, n_2, \ldots, n_k\}$, the numbers of counts in each detector. Assuming independence, and that the processes yielding counts are really Poisson, the likelihood function is:

$$L(\mathbf{n};\boldsymbol{\theta}) = \prod_{i=1}^{k} \frac{\theta_i^{n_i} e^{-\theta_i}}{n_i!}.$$

Of course, we may have a relation which reduces the number of independent parameters, but this is the general form, and the θ_i may be expressed in terms of the reduced set as appropriate.

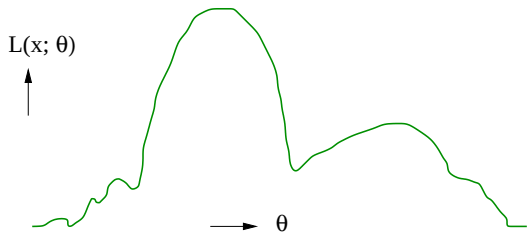
Information

Fisher Information: If $L(\mathbf{x};\theta)$ is a likelihood function which depends on some parameter θ, the Fisher Information Number, corresponding to θ, is:

$$I(\theta) = \langle \left(\frac{\partial \ln L}{\partial \theta} \right)^2 \rangle .$$

Intuition: If L varies rapidly with θ, the experimental sampling distribution will be very sensitive to θ. Hence, a measurement will contain a lot of "information" relevant to θ. It may be helpful to note that (you show):

$$\langle \left(\frac{\partial \ln L}{\partial \theta} \right)^2 \rangle = -\langle \frac{\partial^2 \ln L}{\partial \theta^2} \rangle .$$



Note that $\partial_\theta \ln L$ is known as the "Score Function"

Rao-Cramer-Frechet Inequality

We are ready for this important result which gives us a bound on the best possible efficiency for a given bias.

We suppose that we have an estimator $\hat{\theta}$ for a parameter θ, with a bias function $b(\theta)$: $\hat{\theta} = \hat{\theta}(\mathbf{x})$, where $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ are n measurements, with likelihood function $L(\mathbf{x}; \theta, \text{other parameters})$.

With this understanding, we have the theorem on the next slide:

Rao-Cramer-Frechet Inequality (cont.)

Theorem: (Rao-Cramer-Frechet) Assume:

(1) The range of $\mathbf{x}$ is independent of θ.

(2) The variance of $\hat{\theta}$ is finite, for any θ.

(3) $\partial_\theta \int_{-\infty}^{\infty} f(\mathbf{x}) L(\mathbf{x};\theta) d\mathbf{x} = \int_{-\infty}^{\infty} f(\mathbf{x}) \partial_\theta L(\mathbf{x};\theta) d\mathbf{x}$,

where $f(\mathbf{x})$ is any statistic of finite variance.

Then:

$$\sigma^2_{\hat{\theta}} \geq \frac{\left(1 + \partial_\theta b\right)^2}{I(\theta)}.$$

Proof: Exercise. Sketch: First, show that

$$I(\theta) = \mathrm{Var}\left(\partial_\theta \ln L\right).$$

Next, find the linear correlation parameter, ρ, between the score function and $\hat{\theta}$. Finally, note that $\rho^2 \leq 1$.

Efficient Estimators

Interesting question: Under what (if any) circumstances, can the minimum variance bound be achieved? If an unbiased estimator achieves the minimum variance, it is called efficient.

Theorem: (Efficiency) An efficient (perhaps biased) estimator for θ exists iff:

$$\frac{\partial \ln L(\mathbf{x};\theta)}{\partial \theta} = [f(\mathbf{x}) - h(\theta)]g(\theta).$$

An unbiased efficient estimator exists iff we further have:

$$h(\theta) = \theta.$$

Proof: Exercise. Hint: The RCF bound made use of the linear correlation coefficient, in which equality holds iff there is a linear relation:

$$\partial_\theta \ln L(\mathbf{x};\theta) = a(\theta)\hat{\theta} + c(\theta).$$

Exponential Family

An important class of probability distributions is the Exponential Family:

$$L(\mathbf{x};\theta) = \exp\big[A(\theta)D(\mathbf{x}) + B(\theta) + C(\mathbf{x})\big].$$

Exercise: Show that $\hat{\theta} = D(\mathbf{x})$ is an efficient estimator for θ.

Exercise: If x is a sample from a normal distribution of known variance, show that x is an unbiased efficient estimator for the mean.

Substitution (Moment) Method

Suppose we have pdf $p(x;\theta)$ for RV x, depending on unknown parameter θ. Any statistic $f(x)$ computed from x has expectation value:

$$\langle f \rangle = \int f(x) p(x;\theta) dx = \phi_f(\theta).$$

If ϕ_f is invertible, we have:

$$\theta = \phi_f^{-1}[\langle f \rangle].$$

Thus, we can define a plausible estimator for θ if we substitute the sample average $\overline{f} = \frac{1}{n}\sum_i^n f(x_i)$ for the expectation value of f:

$$\hat{\theta} = \phi_f^{-1}[\overline{f}].$$

Example: Angular Distribution

A common application of the moment method is in the estimation of parent angular distributions. For example, suppose we want to estimate a in an assumed angular distribution of the form:

$$\frac{d\sigma}{d\Omega} = A(1 + a\cos\vartheta),$$

where our measurement consists of the n samplings, $\{x_1, \ldots, x_n\}$ of $x = \cos\vartheta$.

Taking $f(x) = x$, we have

$$\langle x \rangle = \int x(1 + ax)dx/2 = a/3 = \phi_f(a).$$

This is readily invertible, and we obtain the estimator for a:

$$\hat{a} = \frac{3}{n}\sum_{i=1}^{n} x_i.$$

Exercises - Moment Method

Efficiency and Bias: What is the bias of this estimator for $\hat{a}$? Compare its efficiency with the minimum bound.

Generalization: Generalize this example to an estimator for the strength of an arbitrary $Y_{\ell m}$ moment.

Consistency: Is the moment method always consistent?

Example: CP violation via mixing

In the measurement of CP violation via mixing at the $\Upsilon(4S)$, the pdf for the time random variable may be written:

$$p(t; A) = \frac{1}{2} e^{-|t|}(1 + A \sin xt),$$

where $t \in (-\infty, \infty)$, $x = \Delta m/\Gamma$ is known, and A is the CP asymmetry parameter of interest.

In the early days, there was some small dispute concerning the importance of a "dilution factor" in what amounts to a moment method. We have the tools to analyze this now.

Example: CP violation (continued)

The simplified analysis under discussion was to simply count the number of times $t < 0$, n_-, and the number of times $t > 0$, n_+. The expectation value of the difference between these, for a total sample size $n = n_- + n_+$, is:

$$\langle n_+ - n_- \rangle = n \frac{xA}{1 + x^2}.$$

This is readily inverted, leading to the estimator:

$$\hat{A}_\pm = d^{-1} \frac{n_+ - n_-}{n},$$

where $d = x/(1 + x^2)$ is (or once was) known as a "dilution factor".

Example: CP violation (continued)

We note that $\hat{A}_\pm$ is by definition an unbiased estimator for A. The question is, how efficient is it? In particular, we are throwing away detailed time information - does that matter very much, assuming our time resolution isn't too bad?

First, what is the variance of $\hat{A}_\pm$? For a given n, we may treat the sampling of $n_\pm$ as a binomial process, giving:

$$\delta\hat{A}_\pm = d^{-1}\sqrt{(1-d^2A^2)/n}.$$

Second, how well can we do, at least in principle, if we do our best? Let's use the RCF bound to estimate this (and argue that, at least asymptotically, we can achieve this bound, e.g., with the maximum likelihood estimator):

Example: CP violation (RCF bound)

For n independent time samplings, the RCF bound on the variance of any unbiased estimator for A is:

$$\delta^2 \hat{A} \geq 1/\langle \left[\frac{\partial}{\partial A} \sum_1^n \log p(t_i; A) \right]^2 \rangle$$

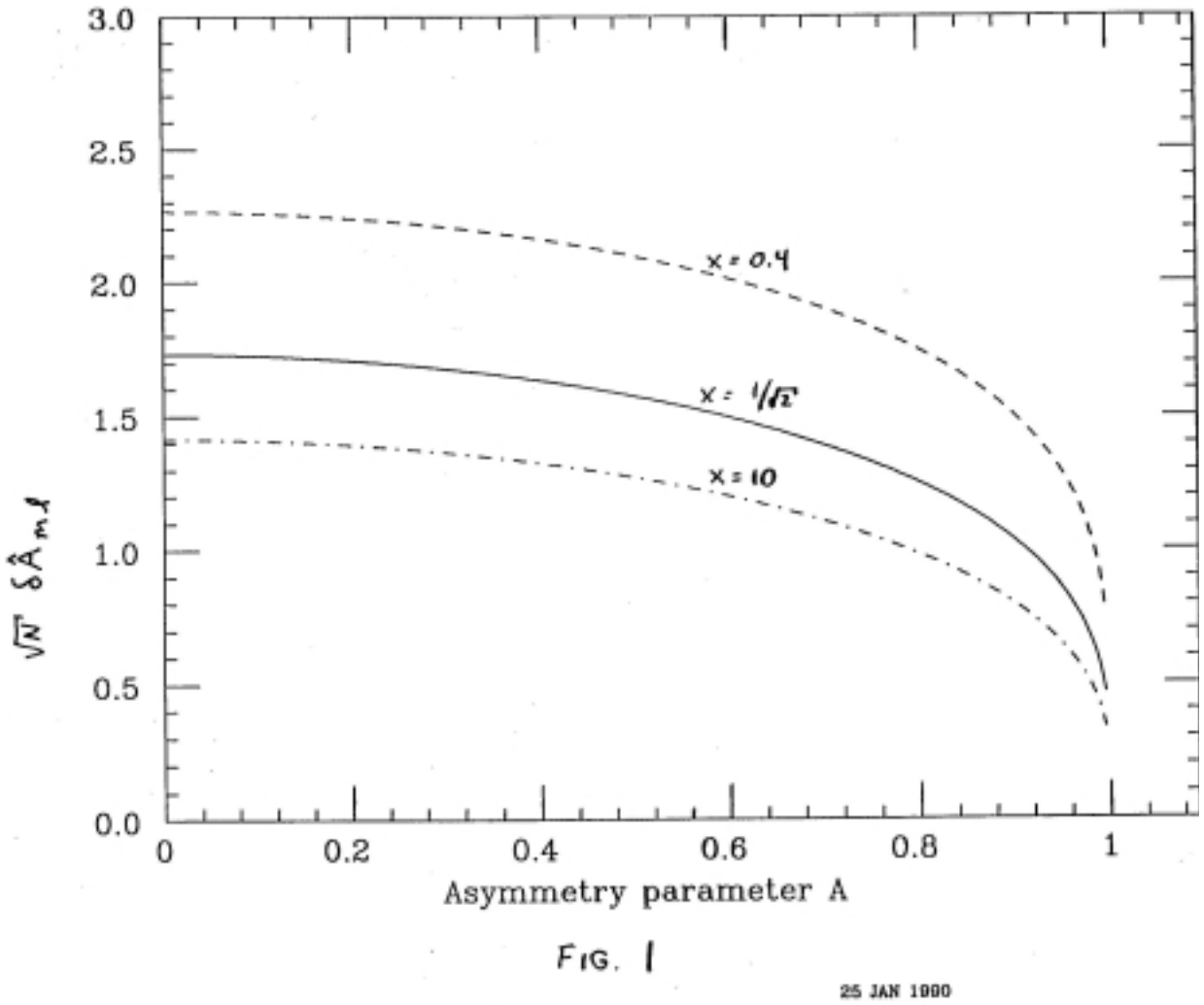
$$\geq 1/n \langle \left(\frac{\sin xt}{1 + A \sin xt} \right)^2 \rangle .$$

Performing the integral gives:

$$\delta^2 \hat{A} = \frac{1}{n} \sum_{k=1}^{\infty} A^{2(k-1)} \frac{x^{2k}(2k)!}{[1 + (2x)^2][1 + (4x)^2] \cdots [1 + (2kx)^2]} .$$

We may graph this function, and graphically compare the moment method variance with this bound:

RCF Bound on Error in Asymmetry Parameter Estimators



Compare Moment Method with RCF Bound on Variance in Asymmetry Parameter Estimates

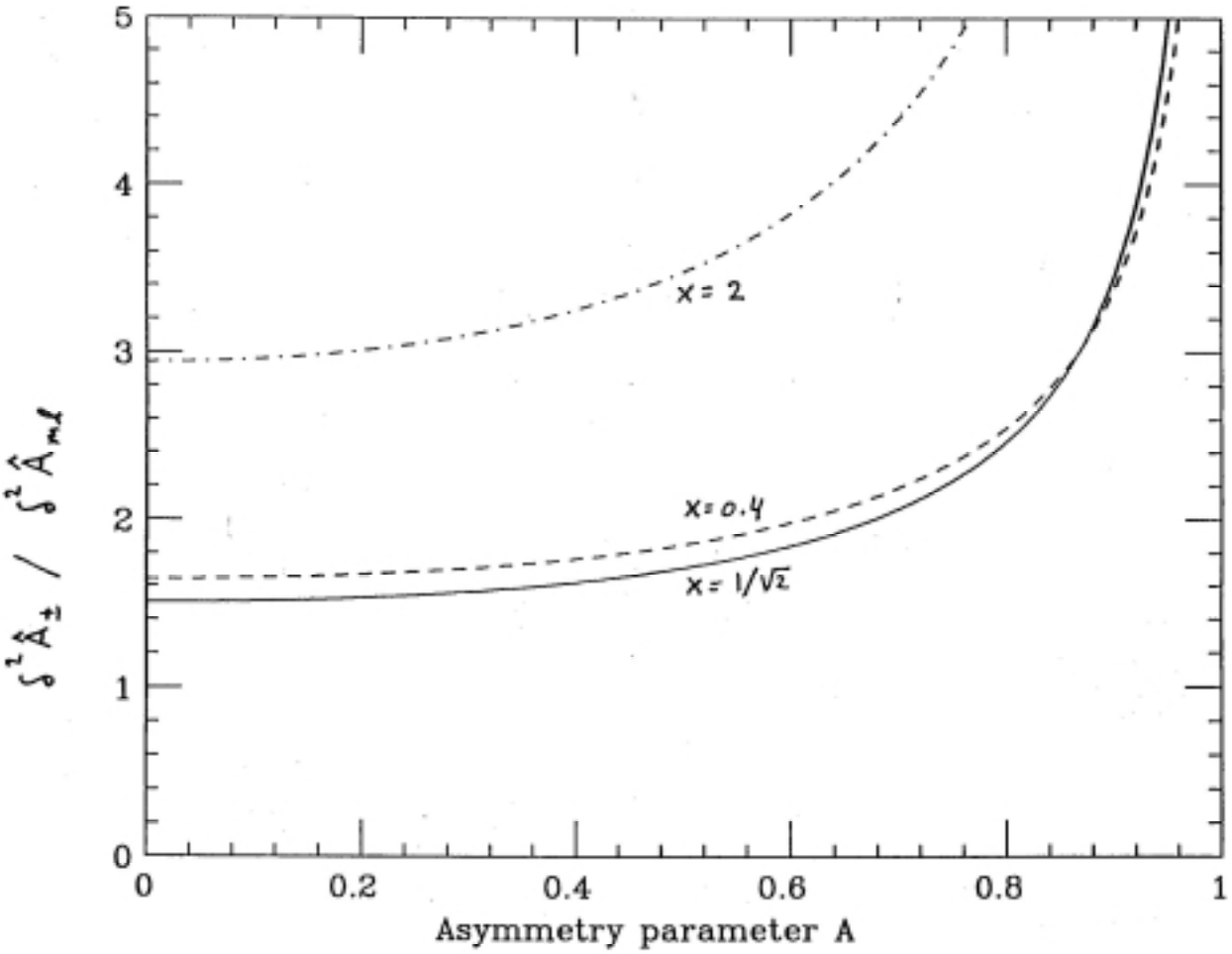


FIG. 2

Maximum Likelihood Method

A popular method for parameter estimation with many desirable properties is the Maximum Likelihood Method:

Maximum Likelihood Estimator: Given measurements $\mathbf{x}$, the Maximum Likelihood Estimator (MLE), $\hat{\theta}$, for a parameter θ, is the value of the parameter for which the likelihood function, $L(\mathbf{x};\theta)$, is maximized:

$$L(\mathbf{x};\hat{\theta}) = \max_{\theta} L(\mathbf{x};\theta).$$

Maximum Likelihood Method (continued)

Intuition: The MLE is that value of the parameter which would make the actual observed data values the most likely observation (compared with other possible parameter values). This isn't the same as saying that it is somehow the "most likely" value of θ (a statement outside of classical statistics).

Let us examine some of the properties of this oft-misunderstood estimator.

MLE Efficiency, Bias

Theorem: (MLE efficiency) The MLE will be unbiased and efficient, if an unbiased efficient estimator exists.

Proof: The maximum likelihood prescription (assuming no "endpoint" troubles) corresponds to:

$$\left.\frac{\partial \ln L(\mathbf{x};\theta)}{\partial\theta}\right|_{\theta=\hat{\theta}} = 0.$$

Suppose an efficient, unbiased estimator exists. Then

$$\frac{\partial \ln L}{\partial\theta} = [f(\mathbf{x}) - \theta]g(\theta),$$

and, hence:

$$\left.\frac{\partial \ln L}{\partial\theta}\right|_{\hat{\theta}} = [f(\mathbf{x}) - \hat{\theta}]g(\hat{\theta}) = 0.$$

Thus, $\hat{\theta} = f(\mathbf{x})$ is the maximum likelihood estimator, and is unbiased and efficient.

But note that the MLE is otherwise biased/inefficient!

MLE in Asymptopia

Theorem: (MLE asymptotic property) The MLE is asymptotically (i.e., as the sample size $n \to \infty$) efficient, unbiased, consistent, and normal (assuming the sampling space does not depend on the parameter value).

Proof: Exercise. Hint: Make a Taylor series expansion of the ML condition in terms of $\log L$ about the true parameter value. Use the Central Limit Theorem.

MLE - Other Comments

1. The MLE is parameterization-independent. Given function $\alpha(\theta)$,

$$\hat{\alpha}_{\rm ML} = \alpha(\hat{\theta}_{\rm ML}).$$

   Note that, since $\langle f(x)\rangle \neq f(\langle x\rangle)$ in general, this means that MLE are "typically" biased.

2. The MLE is sufficient, if a sufficient statistic exists.

3. The MLE may not be robust.

4. The MLE does not provide a test for "goodness of fit". (But likelihood ratio test is available).

5. Watch out for multiple maxima!

6. This method is sometimes difficult to carry out.

MLE – Poisson example

Given likelihood function:

$$L(n;\theta) = \frac{e^{-\theta-b}(\theta+b)^n}{n!},$$

The MLE for θ is conveniently found by taking:

$$\begin{aligned}\partial_\theta \log L &= \partial_\theta[-\theta - b + n\log(\theta+b) - \log n!] \\ &= -1 + n/(\theta+b).\end{aligned}$$

Setting this to zero gives the MLE:

$$\hat{\theta} = n - b,$$

which is intuitive! (You consider $n = 0$ case...)

Note that:

$$\langle\hat{\theta}\rangle = \langle n - b\rangle = (\theta + b) - b = \theta,$$

so this estimator is unbiased.

MLE – Poisson example (continued)

Further,

$$-\langle\frac{\partial^2 \log L}{\partial\theta^2}\rangle = \langle\frac{n}{(\theta+b)^2}\rangle$$
$$= 1/(\theta+b).$$

Hence, the minimum variance bound is $\theta + b$.

What is the variance of our MLE? It is:

$$\sigma^2_{\hat{\theta}} = \langle(n-b)^2\rangle - \langle n-b\rangle^2.$$

Noting that

$$\langle n(n-1)\cdots(n-k)\rangle = (\theta+b)^{k+1},$$

we obtain:

$$\sigma^2_{\hat{\theta}} = \theta + b,$$

which is the minimum bound.

Unbiased and efficient, even for small Poisson samples.

MLE – Sample Exercise

Angular Distribution: Consider the simple angular distribution problem we have discussed in terms of the moment method already, with pdf:

$$\frac{d\sigma}{d\Omega} = A(1 + a\cos\vartheta),$$

where our measurement consists of the n samplings, $\{x_1, \ldots, x_n\}$ of $x = \cos\vartheta$.

Find the MLE for a, and compare its properties with the estimator from the moment method.

Example – Analysis of Bias in m_τ

The BES experiment made a precision measurement of m_τ, by measuring the $e^+e^- \to \tau^+\tau^-$ cross section near threshold. To optimize running time, they used a "data-driven" algorithm to update the energy setting of the storage ring in real time, to try to run at the energy where the cross section is most sensitive to the mass.

The final mass value is obtained by a maximum likelihood fit to the observed cross section. The likelihood function used is:

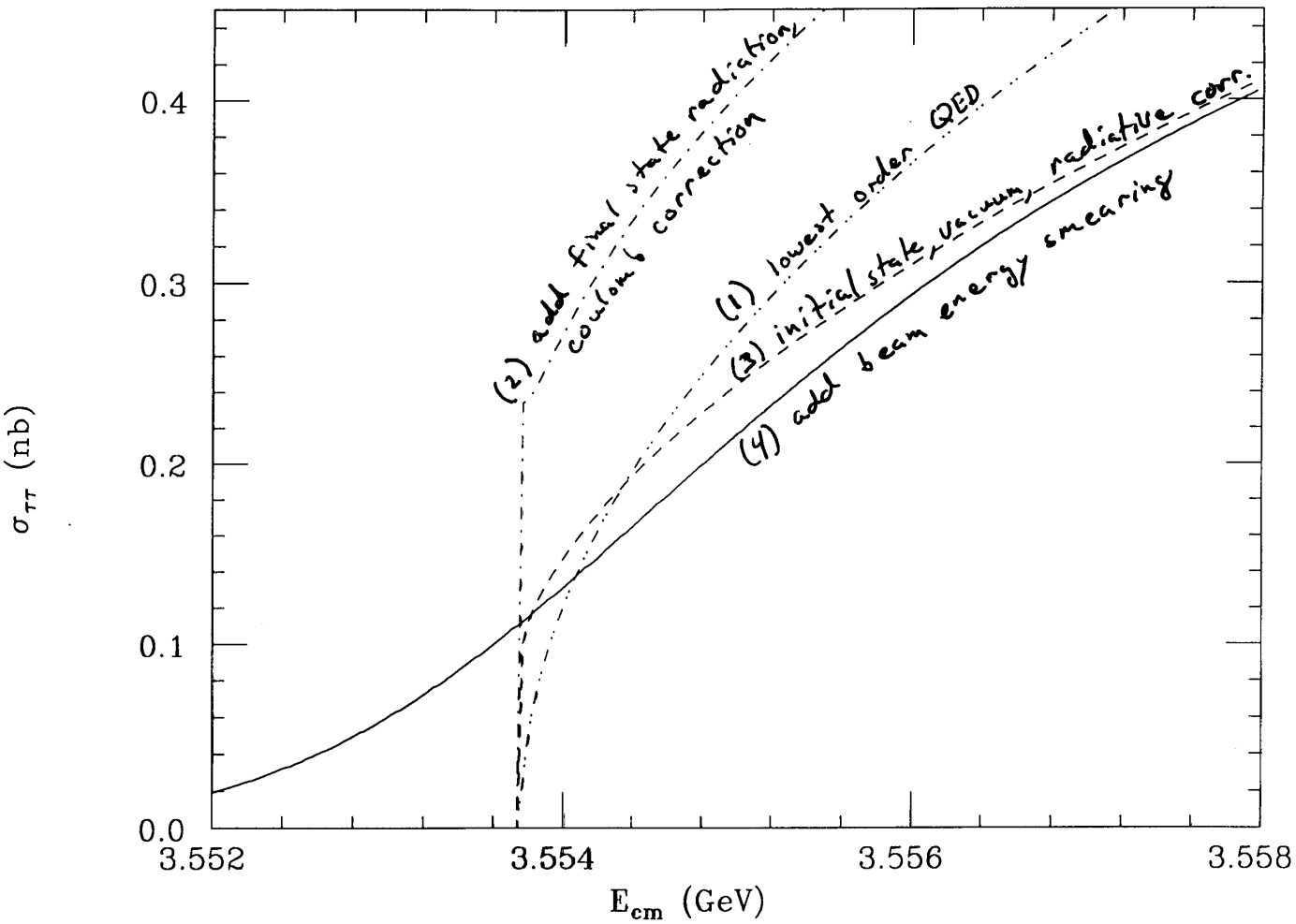
$$L(\mathbf{n}; m) = \prod_{i=1}^{k} \frac{e^{-\theta_i(m)} \theta_i(m)^{n_i}}{n_i!},$$

where k is the number of energy points.

Is this method of measurement biased?

Yes, in general.

BES $e^+e^- \to \tau^+\tau^-$ Cross Section



Distribution of Number of $e\mu$ Events

The $e\mu$ channel is the one which drives the scan. Hence, this is the channel where we might expect to see the greatest sensitivity of the bias to the scan algorithm. As a quick check on the simulation, Figure 1 shows the distribution of the number of events in an experiment. It may be noted that the number we observed in the BES scan is in the region of the peak.

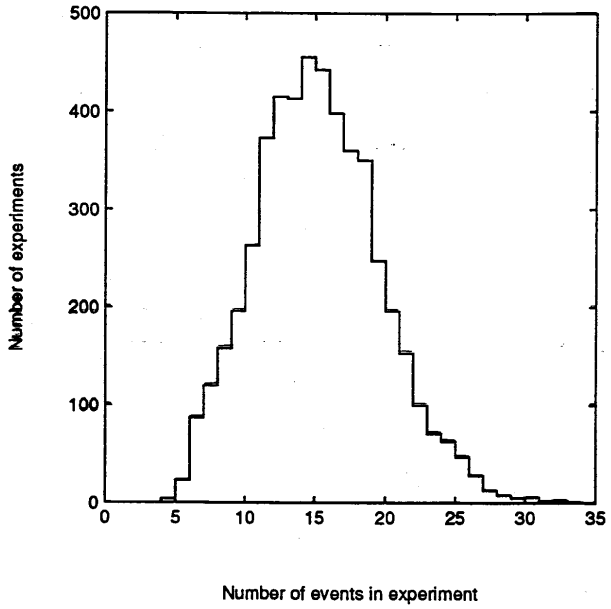
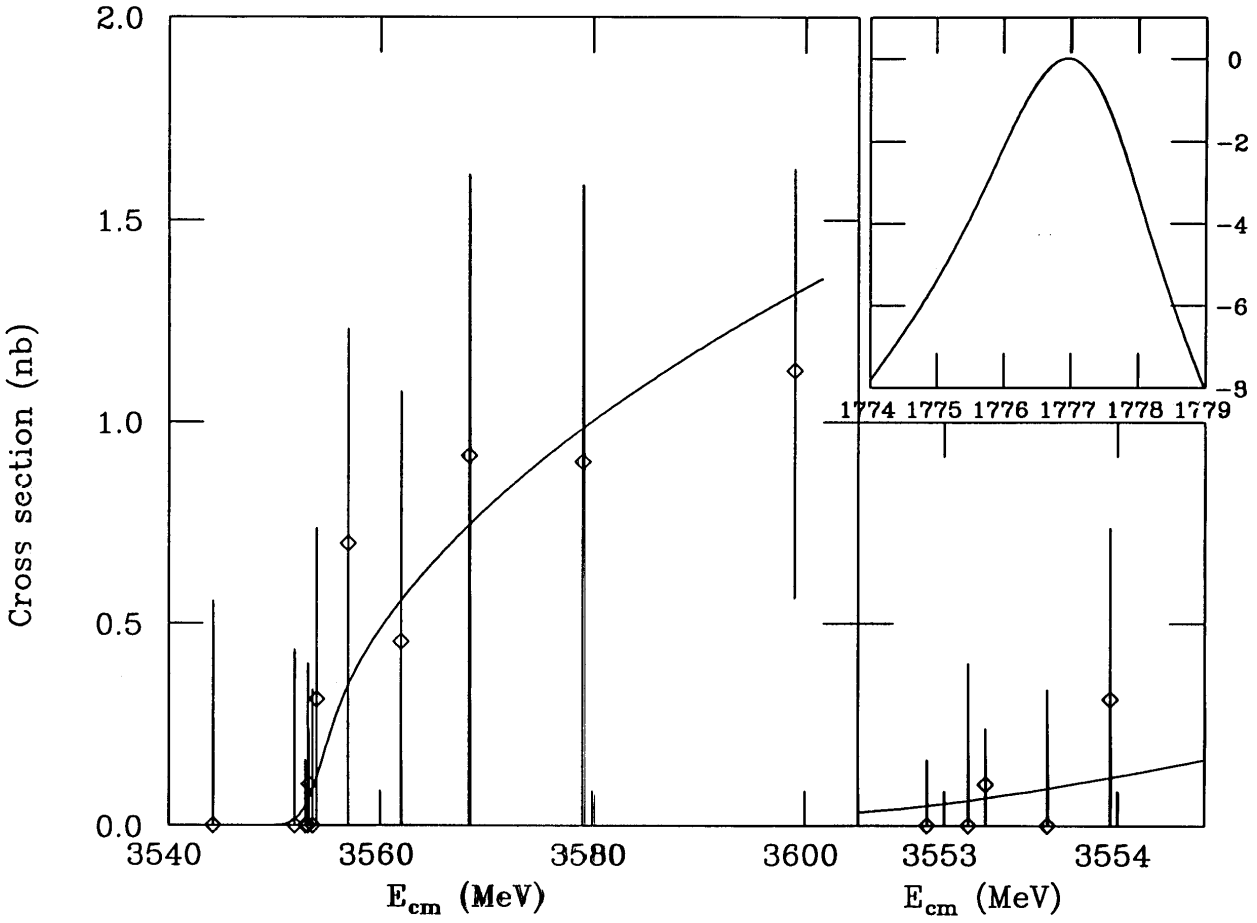
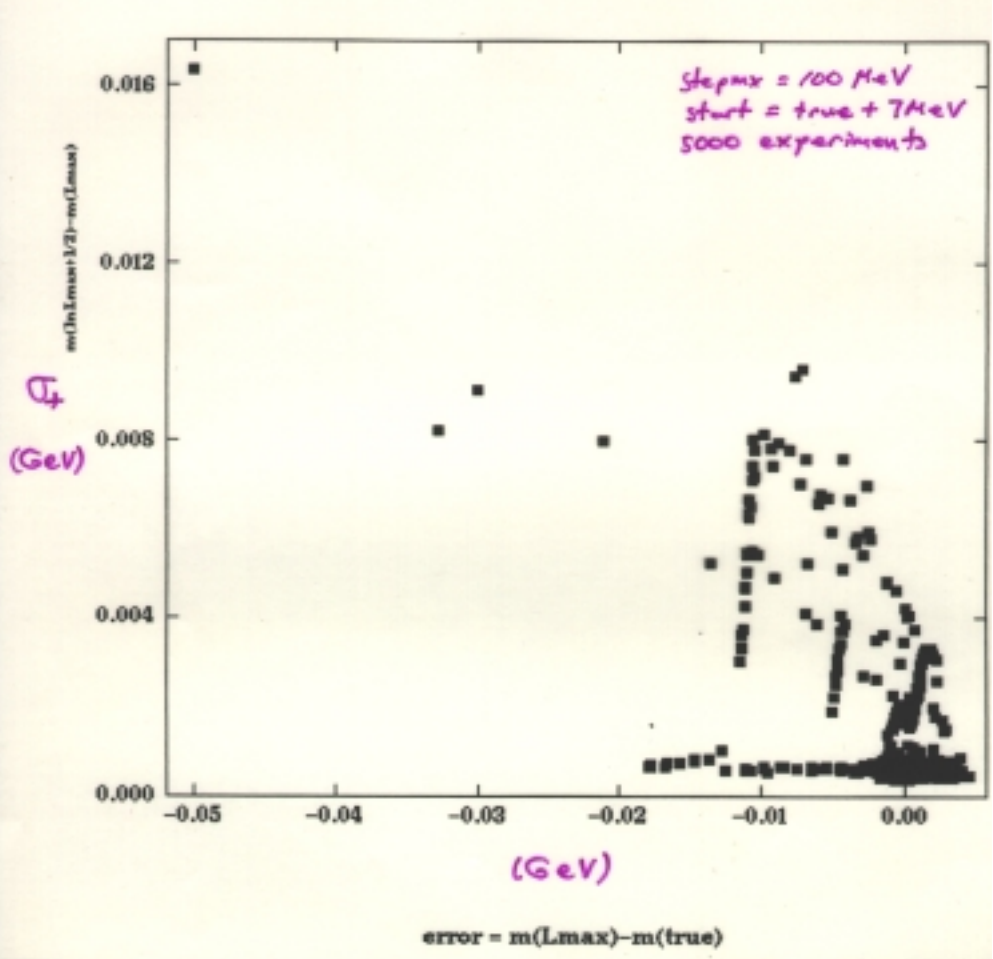


Figure 1. Distribution of the number of $e\mu$ events in an experiment, for a starting error of `start`=7 MeV, and a maximum step size of 100 MeV.

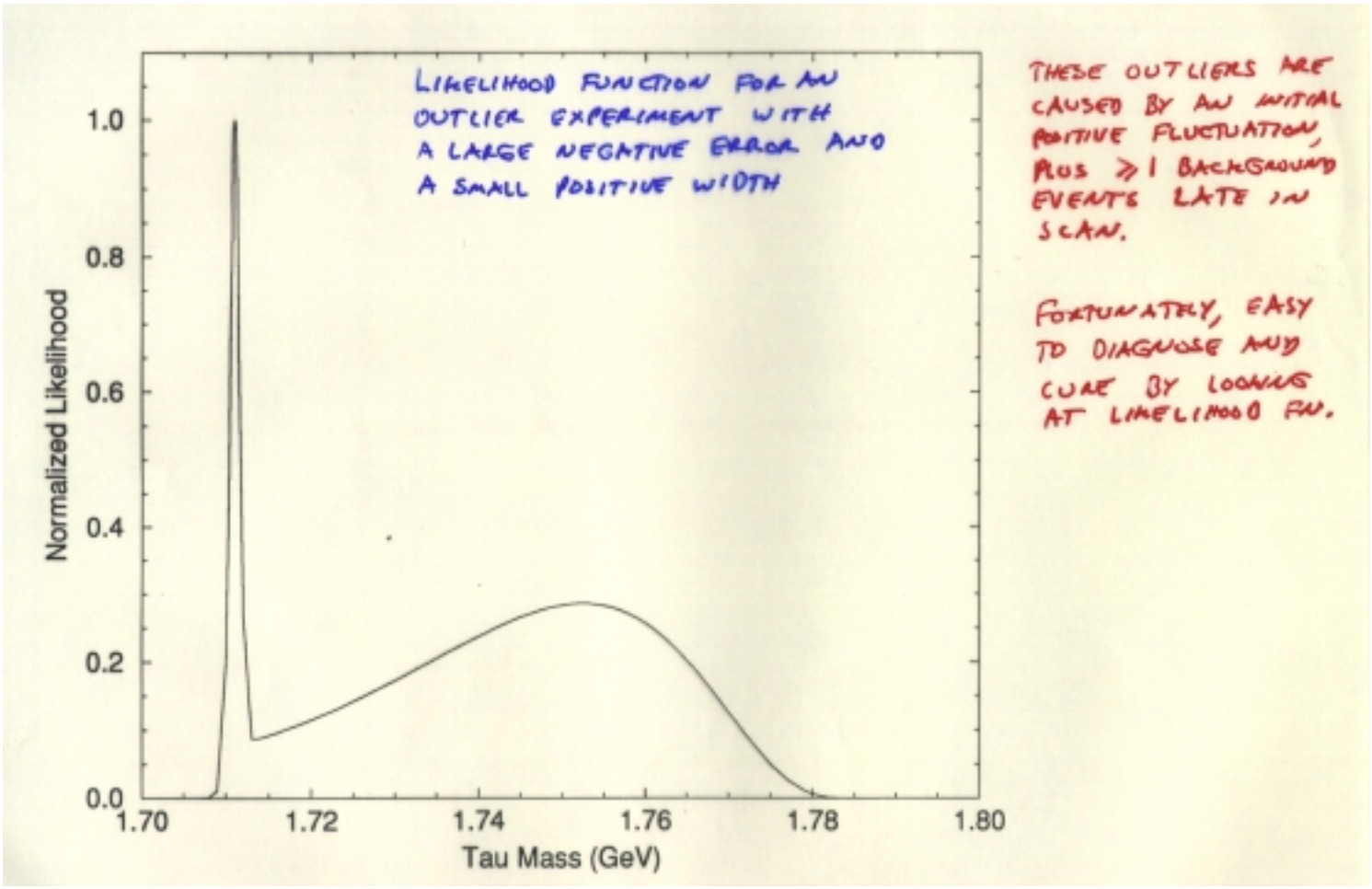
BES $\tau^+\tau^-$ Likelihood, Data Points, Cross Section at Maximum Likelihood



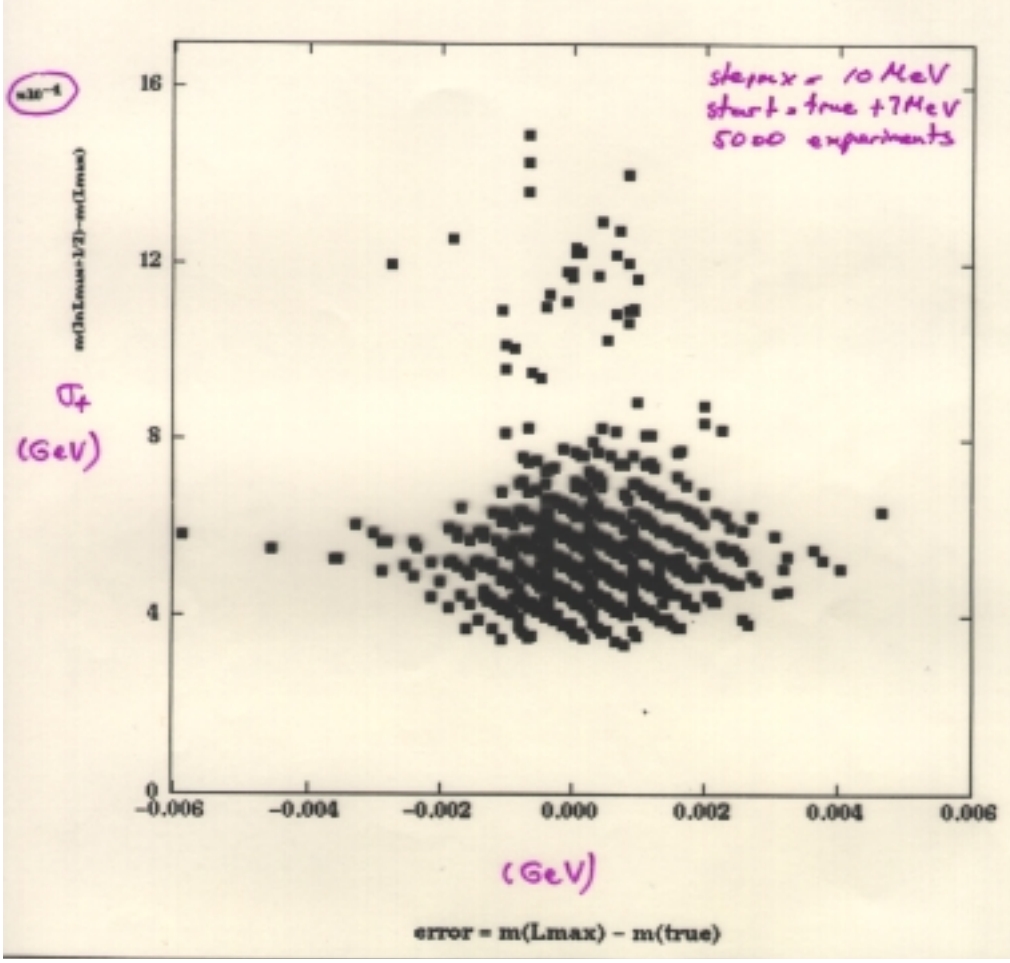
Error Estimate vs Actual Error (Simulation)



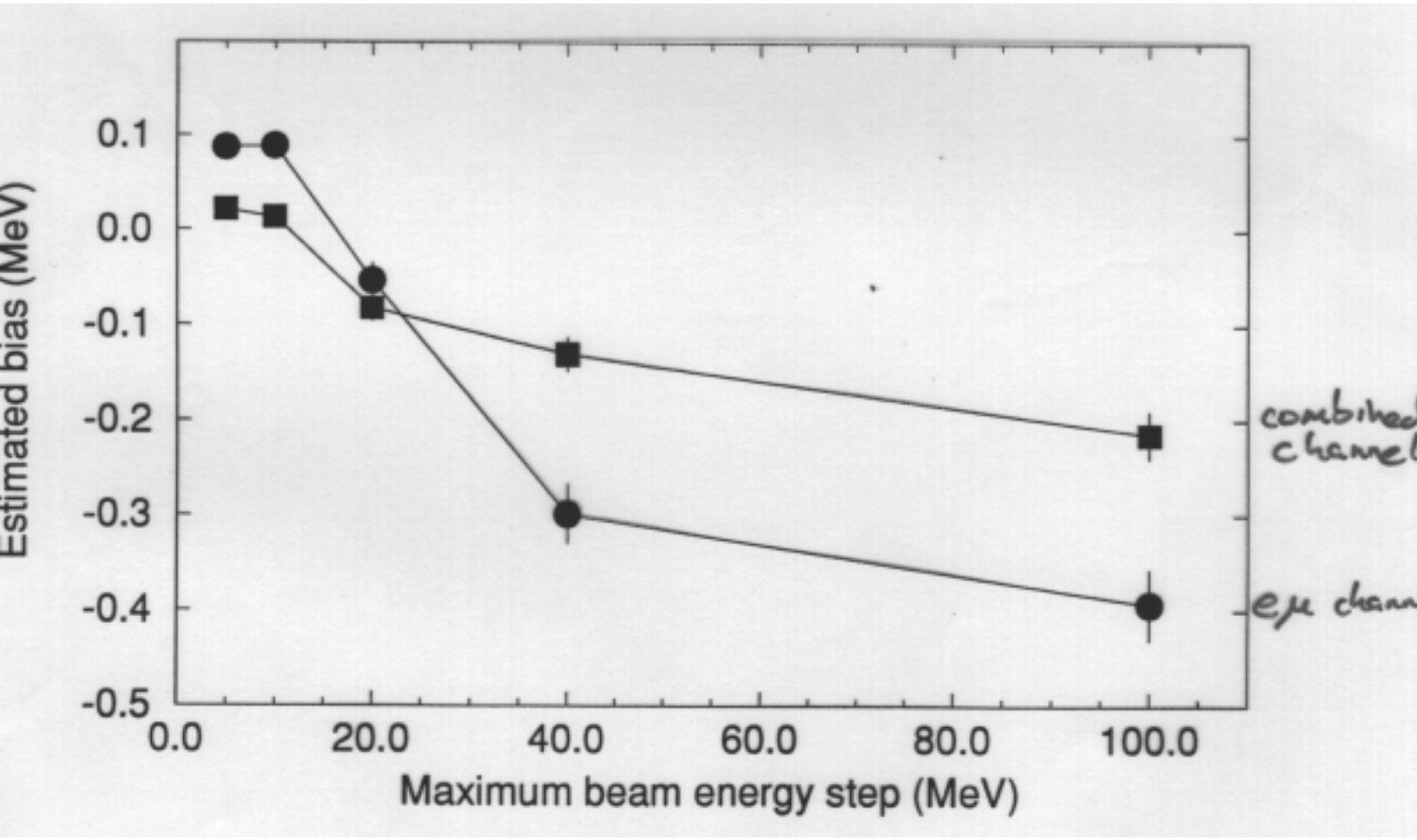
Sample Outlier Likelihood Function (Simulation)



Error Estimate vs Actual Error
stepmx=10MeV (Simulation)



BES Bias vs Energy Maximum Step Size (Simulation)

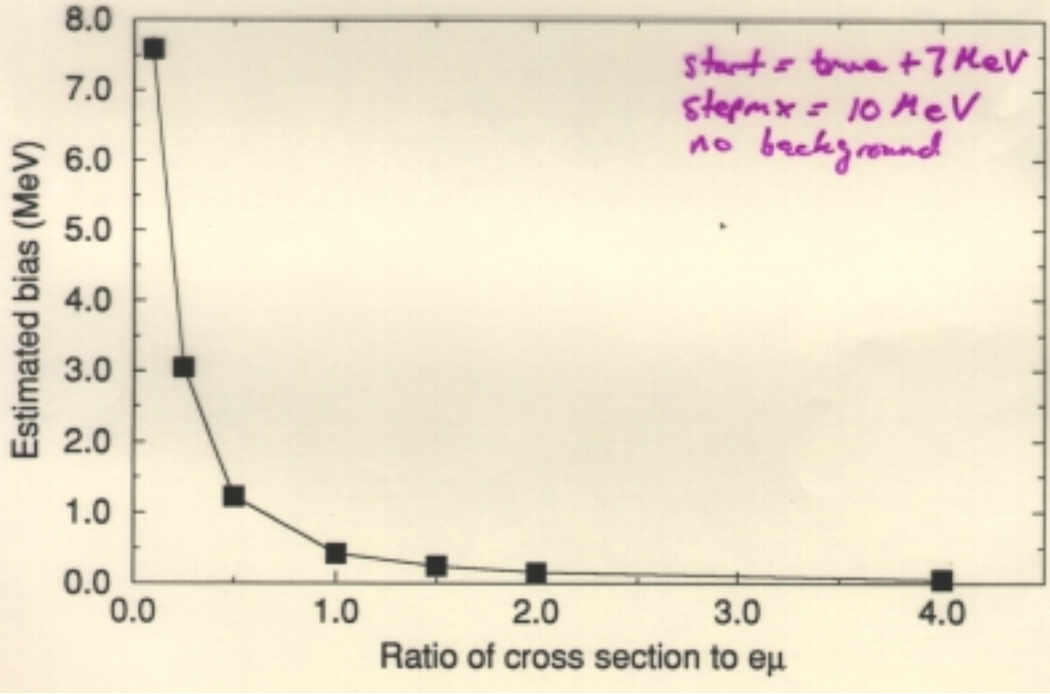


Bias vs Cross Section of Non-driving Channel (Simulation)

$M_{\tilde{\chi}}$ MEASUREMENT

Dependence of bias on channel cross section (non-driving channel)

Asymptotic behavior → consistency



Least Squares Method

Reference: F. T. Solmitz, Ann. Rev. Nucl. Sci., vol 14, 375-402 (1964).

A third popular method is the method of Least Squares Estimation:

Least Squares Estimate: Given a set of observations $\{x_1, \ldots, x_n\}$, with expectation values $\{g_1(\boldsymbol{\theta}) = \langle x_1 \rangle, \ldots, g_n(\boldsymbol{\theta})\}$ and covariance (moment) matrix M, then the set of parameter values $\hat{\boldsymbol{\theta}}$ which minimizes the quantity:

$$S = (\mathbf{x} - \mathbf{g})^T M^{-1} (\mathbf{x} - \mathbf{g})$$

is called the Least Squares Estimate (LSE) for $\boldsymbol{\theta}$.

LSE - Motivation

If the x_i are sampled from a multi-variate normal distribution, then the LSE is the same as the MLE, and we also have that S is distributed according to a χ^2 distribution with $n - r$ degrees of freedom, where r is the number of independent parameters being estimated. This provides us with a test for "Goodness of fit".

Even if the observations are not normally distributed, the LSE may be useful, in particular, if the distribution is approximately normal.

LSE - Sample Application

Suppose our data consists of a histogram which we wish to fit to some model, including the estimation of some parameters.

In general, the histogram bin contents may be described by a Poisson distribution, rather than a normal. However, if the contents are large, the normal approximation may suffice.

Rule-of-thumb: The normal approximation is typically reasonable (and the χ^2 goodness of fit valid) if each bin has at least 7 counts. Note that it is quite permissible to combine bins until this is satisfied. The bin width need not be constant.

Linear Least Squares Methodology

Suppose the expectation values g_i for x_i are (independent) linear functions of the r parameters $\boldsymbol{\theta}$:

$$\langle \mathbf{x} \rangle = \mathbf{g} = \mathbf{g_0} + F\boldsymbol{\theta},$$

where F is an $n \times r$ matrix.

Exercise: Show that the LSE for $\boldsymbol{\theta}$ is:

$$\hat{\boldsymbol{\theta}} = H^{-1} F^T M^{-1} (\mathbf{x} - \mathbf{g_0}),$$

where $H = F^T M^{-1} F$ is an $r \times r$ matrix equal to the inverse of the moment matrix for $\hat{\boldsymbol{\theta}}$. Show further that the estimator is unbiased.

Non-linear Least Squares

In general, we are not lucky enough to have a linear problem. In this case:

First, see whether it is equivalent to a linear problem.

Second, if you don't need to do it often, plug it into a general-purpose minimizer.

Or, third, especially if you need to do it many times (e.g., track fitting or kinematic fitting) it may be a good approximation to linearize the problem via a Taylor series expansion about some starting value for the parameters. The process is iterated until convergence is (hopefully) attained.

LSE – Comments

Constraints may be incorporated into a least squares fit via the method of Lagrange multipliers. This is often much easier than solving the constraint equation to eliminate a parameter. (e.g., kinematic fitting)

The LSE is efficient and unbiased if the observations are normal and the parameter functions are linear.

Theorem: (Gauss-Markov) If the observations are not normal, the LSE still gives the most efficient unbiased linear estimators (if unbiased linear estimators exist).

Proof: Exercise.

LSE – Comments (continued)

Since, at the correct moment matrix M, the (linear case) estimators are efficient, small errors in M enter at second order. Hence, an approximation for M may be adequate.

The "pulls" (or normalized residuals), are a handy way to tell whether the fit assumptions (e.g., M) are reasonable:

$$\text{pull}_i = \frac{x_i - g_i(\hat{\boldsymbol{\theta}})}{\sqrt{M_{ii} - \left(FH^{-1}F^T\right)_{ii}}}.$$

If all is well, the pulls should be $N(0,1)$ distributed. (Exercise)

LSE – Exercise

Angular Distribution: Consider the simple angular distribution problem we have discussed in terms of the moment method already, with pdf:

$$\frac{d\sigma}{d\Omega} = A(1 + a\cos\vartheta),$$

where our measurement consists of the n samplings, $\{x_1, \ldots, x_n\}$ of $x = \cos\vartheta$.

Find the LSE for a, and compare its properties with the estimator from the other methods.