3. Parameter Estimation II – Several Variables

Last lecture, we estimated one parameter in cases where all other parameters were fixed. Now, we will solve problems where more than one variable is unknown.

3.1. Example 4: The amplitude of a signal in presence of a flat background

Suppose you measure an integer quantity like the number of photons from a quasar at certain wavelengths in the ~~bsto~~ presence of flat background noise.

Suppose further that your detector has several (M) wavelength channels x_k and that the signal is gaussian around the wavelength x_0 with some width w.

The ideal datum is then

amplitude of signal

Fig 3.1. $$D_K = n_0 \left[A e^{-(x_K - x_0)^2 / 2w^2} + B \right]$$

constant, e.g. duration of measurement, sensitivity

flat Background in all channels

Yet, while D_K is a real number, the # photons in each channel is an integer.

So use Poisson distribution

$$\text{prob}(N \mid D) = \frac{D^N}{N!} e^{-D}$$

Please note that this is normalized = 1, because

$$\sum_{N=0}^{\infty} \text{prob}(N \mid D) = \sum_{N=0}^{\infty} \frac{D^N}{N!} e^{-D} = e^{-D} \sum_{N=0}^{\infty} \frac{D^N}{N!} = e^{-D} e^{D} = 1$$

It is easy to show and I leave it as an exercise that for the Poisson distribution

$$\langle N \rangle = \sum_{N=0}^{\infty} N \, \text{prob}(N \mid D) = D .$$

So the probability of finding N_k photons in channel x_k is

$$\text{prob}(N_k \mid A, B, I) = \frac{D_k^{N_k} e^{-D_k}}{N_k!}$$

If the data are independent, the total prob. factorizes:

$$\text{prob}(\{N_k\} \mid A, B, I) = \prod_{k=1}^{M} \text{prob}(N_k \mid A, B, I)$$

To proceed, we assume that W, x_0 and x_k's are given and we would like to infer A and B.

As usual, we use Bayes theorem:

$$\text{prob}(A, B \mid \{N_k\}, I) \propto \text{prob}(\{N_k\} \mid A, B, I) \, \text{prob}(A, B \mid I)$$

includes x_0 etc

Prior: $\mathrm{prob}(A,B|I) = \begin{cases} \text{const} & ; A \geq 0 \wedge B \geq 0 \\ 0 & ; \text{otherwise} \end{cases}$

Put everything together to find

Figure 3.3! $L = \ln \mathrm{prob}(A,B|\{N_k\},I) = \text{const} + \sum_{k=1}^{M} \left(N_k \ln D_k - D_k \right)$

[For figure 3.3: ~~FWHM~~ FWHM $= 2.35\,\sigma$ for gaussians.
$\Rightarrow$ FWHM $= 5 \Rightarrow \sigma = w = 2.12$]

Marginal distributions

Often, we are not interested in one of the variables, e.g. the background B. So we need to marginalize:

$$\mathrm{prob}(A|\{N_k\},I) = \int_0^\infty \mathrm{prob}(A,B|\{N_k\},I)\,dB$$

Fig. 3.4

Caution: $\mathrm{prob}(A|\{N_k\},I) \neq \mathrm{prob}(A|\{N_k\},B,I)$
because in the latter, B is known. In former, we are ignorant about B.

And what if w and x_0 were unknown?
Then we have to marginalize over w and x_0:

$$\mathrm{prob}(A,B|\{N_k\},I) = \int \underbrace{\mathrm{prob}(A,B,w,x_0|\{N_k\},I)}_{\text{use Bayes:}}\,dx_0\,dw$$

$$\text{prob}(A, B, \omega, x_0 \mid \{N_k\}, I) \propto \text{prob}(\{N_k\} \mid A, B, \omega, x_0, I) \times \text{prob}(A, B, \omega, x_0 \mid I)$$

Likelihood funct. known from above

prior

Decompose prior:

$$\text{prob}(A, B, \omega, x_0 \mid I) = \text{prob}(A, B \mid I)\, \text{prob}(\omega, x_0 \mid I)$$

To recover situation where x_0 and ω are known, use δ-functions:

$$\text{prob}(\omega, x_0 \mid I) = \delta(\omega - 2.12)\, \delta(x_0)$$

Plug this into margialization to recover original situation:

$$\text{prob}(A, B \mid \{N_k\}, I) \propto \text{prob}(\{N_k\} \mid A, B, \omega = 2.12, x_0 = 0, I) \times \text{prob}(A, B \mid I)$$

Binning the data:

Experimental channel has finite width, so what we really measure is

$$D_k = \int_{x_k - \frac{\Delta}{2}}^{x_k + \Delta/2} n_0 \left\{ A e^{-\frac{(x - x_0)^2}{2\omega^2}} + B \right\} dx$$

using box-approximation for slim bins gives

$$D_k = \underbrace{n_0 \Delta}_{!} \left\{ A e^{-\frac{(x_k - x_0)^2}{2\omega^2}} + B \right\}$$

So Δ is just a redefinition of time spend,
i.e. n_0 is time spend x collection area (sensitivity)

<u>Note:</u>

- If broader bins, data quality looks better, but that's cosmetics!

- Don't make bins to broad $\Rightarrow$ lose information!

3.2 Reliabilities : best estimates, correlations, error bars

Suppose you have several quantities of interest $\{x_i\}$ with posterior prob $(\{x_i\} \mid \{data\}, I)$

As usual, use $L = \ln \text{prob}(\{x_i\} \mid \{data\}, I)$

and best estimate

$$\left.\frac{\partial L}{\partial x_i}\right|_{\{x_i^0\}} = 0 \qquad \forall i$$

Expanding around the maximum yields

$$L = L(\vec{x}_0) + \text{linear} + \frac{1}{2}(x_i - x_i^0)\frac{\partial^2 L}{\partial x_i \partial x_j}(x_j - x_j^0) + \ldots$$

$$= L(\vec{x}_0) + \frac{1}{2}(\vec{x} - \vec{x}_0)^T L^{(2)} (\vec{x} - \vec{x}_0) + \ldots$$

where $\left[L^{(2)}\right]_{ij} \equiv \frac{\partial^2 L}{\partial x_i \partial x_j}$ is a matrix

For a gaussian integral with source $\vec{J}$, the following result holds which is actually at the heart of perturbation theory in QFT:

$$\frac{\sqrt{\det L^{(2)}}}{(2\pi)^{n/2}} \int dx_1 \dots dx_n \, e^{\frac{1}{2}[\vec{x}-\vec{x}_0]^T L^{(2)} [\vec{x}-\vec{x}_0] + [\vec{x}_0 - \vec{x}_0]^T \vec{J}}$$

$$= \frac{\sqrt{\det L^{(2)}}}{(2\pi)^{n/2}} \int dx_1 \dots dx_n \, e^{\frac{1}{2}[x_i - x_i^0] L_{ij}^{(2)} [x_j - x_j^0] + [x_i - x_i^0] J_i}$$

$$= e^{-\frac{1}{2} J_i [[L^{(2)^{-1}}]]_{ij} J_j} = e^{-\frac{1}{2} \vec{J}^T L^{(2)^{-1}} \vec{J}}$$

$\Uparrow$ <u>very important relation!</u>

As an example, consider the case of 2 variables, α and β.

So $\vec{x} = \begin{pmatrix} \alpha \\ \beta \end{pmatrix}$ and call $L^{(2)} = \begin{pmatrix} A & C \\ C & B \end{pmatrix}$

$\vec{J} = \begin{pmatrix} J_\alpha \\ J_\beta \end{pmatrix}$

e.g. $\frac{\partial^2 L}{\partial \beta^2}$

The <u>Variance</u> is the square of the standard deviation or (r.m.s.) errors and it is defined as

$$\mathrm{Var}(\alpha) = \sigma_\alpha^2 = \langle (\alpha - \alpha_0)^2 \rangle$$

$$= \int (\alpha - \alpha_0)^2 \, \mathrm{prob}(\alpha, \beta \mid \{\mathrm{data}\}, I) \, d\alpha \, d\beta$$

For our gaussian approximation, we can evaluate this easily because

$$\mathrm{prob}(\alpha, \beta \mid \{\mathrm{data}\}, I) = \frac{\sqrt{\det L^{(2)}}}{(2\pi)^{n/2}} \, e^{\frac{1}{2}[\vec{x}-\vec{x}_0]^T L^{(2)} [\vec{x}-\vec{x}_0]}$$

and the integral is then

$$\langle(\alpha-\alpha_0)^2\rangle = \sigma_\alpha^2 =$$

$$= \frac{\sqrt{\det L^{(2)}}}{(2\pi)^{n/2}} \int d\alpha\, d\beta\, (\alpha-\alpha_0)^2\, e^{\frac{1}{2}[\vec{x}-\vec{x}_0]^T L^{(2)} [\vec{x}-\vec{x}_0]}$$

$$= \frac{\sqrt{\det L^{(2)}}}{(2\pi)^{n/2}} \frac{d^2}{dJ_\alpha^2} \int d\alpha\, d\beta\, e^{\frac{1}{2}[\vec{x}-\vec{x}_0]^T L^{(2)} [\vec{x}-\vec{x}_0] + \vec{x}^T \vec{J}} \Big|_{\vec{J}=0}$$

$$= \frac{d^2}{dJ_\alpha^2} e^{-\frac{1}{2}\vec{J}^T L^{(2)^{-1}} \vec{J}} \Big|_{\vec{J}=0} \qquad ; \vec{J} = \begin{pmatrix} J_\alpha \\ J_\beta \end{pmatrix}$$

$$= \frac{d^2}{dJ_1^2} e^{-\frac{1}{2} J_i \left[L^{(2)^{-1}}\right]_{ij} J_j}$$

$$= -\left[L^{(2)^{-1}}\right]_{11}$$

$$L^{(2)^{-1}} = \frac{1}{AB-C^2} \begin{pmatrix} B & -C \\ -C & A \end{pmatrix}$$

Hence

$$\sigma_\alpha^2 = \frac{-B}{AB-C^2} = \frac{B}{C^2-AB}$$

As you can see easily, the same holds for all combinations, i.e.

$$\begin{pmatrix} \sigma_\alpha^2 & \sigma_{\alpha\beta}^2 \\ \sigma_{\alpha\beta}^2 & \sigma_\beta^2 \end{pmatrix} = -L^{(2)^{-1}}$$

$\Uparrow$ COVARIANCE MATRIX

Here $\sigma_{\alpha\beta}^2$ is the covariance of α and β which measures the correlation between α and β.

For the two-variable case with α and β, the second order likelihood term

$$(\alpha-\alpha_0,\ \beta-\beta_0)\begin{pmatrix} A & C \\ C & B \end{pmatrix}\begin{pmatrix} \alpha-\alpha_0 \\ \beta-\beta_0 \end{pmatrix} \equiv Q$$

defines ellipses $Q = k$.
The directions of principal axis is given by eigenvectors of $L^{(2)}$ and eigenvalues related to length.

Fig 3.6

Fig 3.7.

Fig 3.8.

3.5 Approximations: Maximum Likelihood and Least Squares

Denote: Data $\vec{D}$; $\dim \vec{D} = N$

Parameters $\vec{X}$; $\dim \vec{x} = M$

$$\text{prob}(\vec{X} \mid \vec{D}, I) \propto \text{prob}(\vec{D} \mid \vec{x}, I)\ \text{prob}(\vec{X} \mid I)$$

If we are fairly ignorant, we might use flat prior

$\text{prob}(\vec{X} \mid I) = \text{const.}$

Absorb this into proportionality:

$$\text{prob}(\vec{x} \mid D, I) \propto \text{prob}(\vec{D} \mid \vec{x}, I)$$

Hence, <u>best estimate</u> $\vec{X}_0$ given by maximum of posterior is <u>equal</u> to <u>maximum</u> of the <u>likelihood</u>.

Hence, $\vec{X}_0$ is calle "maximum likelihood" estimate

Suppose Data independent, then decompose

$$\text{prob}(\vec{D} \mid \vec{X}, I) = \prod_{K=1}^{N} \text{prob}(D_k \mid \vec{X}, I)$$

If in addition noise of experiment is gaussian, i.e.

$$\text{prob}(D_K \mid \vec{x}, I) = \frac{1}{\sqrt{2\pi}\,\sigma_K} \exp\left[- \frac{(F_K - D_k)^2}{2\sigma_K^2} \right]$$

$$F_k = f(\vec{x}, k) \quad \text{noiseless data}$$

$$\Rightarrow \text{prob}(\vec{D}|\vec{X},I) \propto \exp\left[-\frac{1}{2}\sum \frac{(F_k-D_k)^2}{\sigma_k^2}\right]$$

$$= \exp\left[-\frac{\chi^2}{2}\right]$$

where

$$\chi^2 \equiv \sum_k^N \left(\frac{F_k - D_k}{\sigma_k}\right)^2$$

and we have

$$L = \ln \text{prob}(\vec{D}|\vec{X},I) = \text{const} - \frac{\chi^2}{2}$$

Maximum of posterior $\Leftrightarrow$ χ^2 smallest so corresponding $\vec{X}_0$ is called "least squares" estimate

Maximum likelihood and least squares are among the most frequently used in data analysis. From bayesian point of view, nothing magic, as we have seen.

Popularity of χ^2 is that it is really simple to apply.

Even more simple, when F_k is a linear function of $\vec{X}$, i.e.

$$F_k = \sum_{j=1}^{M} O_{kj} X_j + C_k \Rightarrow O\vec{X} + \vec{C} = \vec{F}$$

as $L = \text{const} - \frac{\chi^2}{2}$, use chain rule to differentiate

$$\frac{\partial L}{\partial x_j} = -\frac{1}{2}\frac{\partial \chi^2}{\partial x_j} = -\sum_{k=1}^{N} \frac{(F_k - D_k)}{\sigma_k^2}\frac{\partial F_k}{\partial x_j}$$

let's differentiate again

$$\frac{\partial^2 L}{\partial x_i \partial x_j} = -\sum_{k=1}^{N} \frac{\partial F_k}{\partial x_i}\frac{\partial F_k}{\partial x_j}\frac{1}{\sigma_k^2} + \frac{(F_k - D_k)}{\sigma_k^2}\underbrace{\frac{\partial^2 F_k}{\partial x_i \partial x_j}}_{=\emptyset}$$

$$= -\sum_{k=1}^{N} \frac{Q_{ki} Q_{kj}}{\sigma_k^2} = \text{const}$$

So posterior pdf completely determined by optimal solution $\vec{x}_0$ and covariance matrix

$$\langle (x_i - x_i^0)(x_j - x_j^0) \rangle = -(L^{(2)^{-1}})_{ij}$$

$$= 2((\vec{\nabla}^2 \chi^2)^{-1})_{ij}$$

It is only true that the covariance matrix is constant here, because we have assumed that the noise is gaussian and F_k linear.