Introduction to Bayesian Concepts and Methods

Cornelius L. Pieterse
cornelius.pieterse@npl.co.uk

National Centre of Excellence in Mass Spectrometry Imaging
National Physical Laboratory
Teddington, UK

Max Planck Institute for the Structure and Dynamics of Matter
Center for Free-Electron Laser Science
Hamburg, Germany

Chris Engelbrecht Summer School
January 2018
Introduction

Course overview

This course will provide an introduction to the Bayesian foundational concepts and methods with the primary emphasis to discuss the fundamental conceptual mindset shift:

*probability as logic rather than a relative frequency.*
Introduction

Outline of the course

- **Introduction**
  - Background and expectations
  - Bayesian approach and corollaries

- **Parameter Estimation**
  - Fair coin example
  - Effect of different priors
  - Best estimates and error bars
  - Gaussian noise and averages

- **Parameter Estimation II**
  - Presence of background
  - Marginalization of background
  - Practical numerical considerations

More relevant applications will be presented by the other speakers
The course almost exclusively follows Sivia and Skilling.

A paperback copy is available for reference/consultation.

Statistics education

- **Undergraduate (BSc Physics and Applied Mathematics @ SU):** courses such as *Probability Theory and Statistics 114*

- **Postgraduate (MEng in Electronic Engineering @ SU):** statistics was non-existent and rarely considered

- **Postgraduate (PhD in Physics @ MPSD):** statistics tend to *complicate* things...

C.L. Pieterse (National Physical Laboratory)  
January 2018 · Chris Engelbrecht Summer School
Why do we care?

- NPL has launched the 3DOrbiSIMS: a new molecular imaging technology with the highest reported simultaneous spatial and mass resolutions.

- To the pharmaceutical industry one of the major challenges is measurement of the intracellular drug concentration.

- This instrument can help identify where drugs go at the cellular level, answering long-standing questions about whether drug concentrations are sufficiently high in the right places to have an effect.
There are three kinds of lies: lies, damned lies and statistics.

(Mark Twain, 1924)

The basics

- A surprisingly large amount of scientists feel uneasy about statistics
- Bayes and Laplace provided us with the more logical approach
- This is an introductory tutorial to the Bayesian approach
Introduction

Deductive Logic versus Plausible Reasoning

(a) Cause

(b) Possible causes

Effects or outcomes

Effects or observations

**Introduction**

**Notation**

\[ X \equiv \text{Logical statement} \]
\[ I \equiv \text{Background information} \]
\[ | \equiv \text{Vertical bar read as given} \]

**Example**

Assign plausibility: \( P(X|I) \) where \( X \equiv \text{It will rain tomorrow} \)
Rules of Reasoning

- **Logical NOT:**
  \[
  \text{Not } X = \overline{X}
  \]

- **Logical AND:**
  \[
  X \text{ and } Y = X, Y
  \]

- **Logical Sum Rule:**
  \[
  P(X|I) + P(\overline{X}|I) = 1
  \]

- **Logical Product Rule:**
  \[
  P(X, Y|I) = P(X|Y, I) \times P(Y|I)
  \]
Assigning Probabilities

- **Important**: An absolute probability does not exist!
- We have made the probabilities $P(X|I)$ conditional on $I$
- Where $I$ denotes the background information and assumptions

The failure to explicitly include all of the relevant *background information* and *assumptions* is often the cause of disagreement about data analysis

**Example**

- $I_1 \equiv$ The newspaper predicted thunderstorms
- $I_2 \equiv$ There are dark clouds in the sky
- $I_3 \equiv$ It is already raining heavily
Corollary I (Bayes theorem)

- Replace $X = H \equiv \text{Hypothesis}$ and $Y = D \equiv \text{Data}$

- Substitute the above and apply the product rule:


$$P(H|D, I) = \frac{P(D|H, I)P(H|I)}{P(D|I)} \quad \text{Bayes theorem}$$

- Often normalisation is not required (hence omitting the evidence)

- Note that the evidence does not explicitly depend on the hypothesis
Corollary II (Marginalization)

Can we integrate over a logical statement, as $Y$ is a proposition?

$$P(X|I) = \int P(X, Y|I) \, dY = P(X, Y|I) + P(X, \overline{Y}|I)$$

These two terms can be expanded with the product rule:

$$P(X, Y|I) = P(Y, X|I) = P(Y|X, I)P(X|I)$$
$$P(X, \overline{Y}|I) = P(\overline{Y}, X|I) = P(\overline{Y}|X, I)P(X|I)$$

Summing the above two equations:

$$P(X, Y|I) + P(X, \overline{Y}|I) = \left( \frac{P(Y|X, I) + P(\overline{Y}|X, I)}{P(X|I)} \right) P(X|I)$$

Unity from Sum Rule
Corollary III (Marginalization)

- Consider a set of propositions:
  \[ \{ Y_k \} = Y_1, Y_2, Y_3, \ldots, Y_M \]

- Decompose over the entire discrete set:
  \[ P(X|I) = \sum_{k=1}^{M} P(X, Y_k|I) \]

- A generalization which requires that the set is normalized:
  \[ \sum_{k=1}^{M} P(Y_k|X, I) = 1 \]

- **Therefore:** the set is required to be mutually exclusive and exhaustive
Corollary III (Marginalization)

- Arbitrarily large number of propositions $\rightarrow$ continuum limit
- A mutually exclusive and exhaustive set of propositions should have
  - intervals chosen to have a common/contiguous border
  - the intervals cover a big enough range of values

Normalization condition within the continuum limit:

$$\int P(Y|X, I) \, dY = 1$$

Therefore: here $Y$ represents the parameter of interest
- Marginalization allows nuisance parameters to be tortured
- Unwanted background signals are golden nuisance parameters
Corollary III (Marginalization)

Laplace considered the mass $M$ of Saturn, given orbital data $D$:

$$\int P(M | D, I) \, dM = 1$$

---

Fig. 1.2 A schematic illustration of the result of Laplace' analysis of the mass of Saturn.

---

Introduction

Bayes, Laplace and some orthodox statistics

- Bernoulli, Bayes and Laplace pondered a lack of certainty
- They interpreted probability as a *degree-of-belief* or *plausibility*

- Unfortunately, other scholars considered this too vague and subjective
- They preferred the *relative frequency* at which the event occurred
- However, this approach assumes infinitely many repeated trials
Bayes, Laplace and some orthodox statistics

- The *frequentist* definition has a limited range of validity/use
- **Example**: mass of Saturn is constant and thus not a random variable
- **Therefore**: no frequency distribution (cannot use probability theory)

---

\[\text{\textsuperscript{2}XKCD.} \textit{Frequentists versus Bayesians.}\]
Coin example

- How can we determine if this coin is fair?

- Consider several hypotheses for the bias-weighting $H$
  - $H = 0$ represents a coin which produces a tail on every flip
  - $H = 1$ represents a coin which produces a head on every flip

- As an example, we proposed the hypotheses:
  - $0.00 \leq H_1 < 0.25$
  - $0.25 \leq H_2 < 0.49$
  - $0.49 \leq H_3 < 0.51$
  - $0.51 \leq H_4 < 0.75$
  - $0.75 \leq H_5 < 1.00$

- A fair coin would therefore be represented by hypothesis $H_3$

- Fairness of the coin is summarized by $P(H|D, I) \, dH$
We can use Bayes Theorem to determine the posterior:

\[ P(H|D, I) \propto P(D|H, I) P(H|I) \]

- Notice that we omitted the constant evidence
- Often the evidence does not depend on \( H \)
- If required, we can determine it using marginalization:
  \[ \int_0^1 P(H|D, I) \, dH = 1 \]

**Therefore**: we need to determine the likelihood and prior
Coin example (prior)

- The prior represents what we know given only the information $I$.
- As a start, we are ignorant about the nature of this coin:

$$P(H|I) = \begin{cases} 
1 & \text{for } 0 \leq H \leq 1, \\
0 & \text{for otherwise.}
\end{cases}$$
Coin example (likelihood)

- Our ignorance is modified by the data through the likelihood
- A measure of the chance that we would have obtained the data that we actually observed, if the value of the bias-weighting was known.
- Probability of obtaining the data $R$ heads in $N$ flips is given:

$$P(D|H, I) \propto H^R (1 - H)^{N-R}$$
Coin example (posterior)

- Helpful to see how the posterior evolves as we obtain more data
  - Left-hand corner figure indicates the result of the previous flip
  - Right-hand corner figure shows the total number of trails
- Probability of obtaining the data \( R \) heads in \( N \) flips is given:
  
  \[
P(D|H, I) \propto H^R (1 - H)^{N-R}
  \]

---

Coin example (posterior)

- Helpful to see how the posterior evolves as we obtain more data
  - Left-hand corner figure indicates the result of the previous flip
  - Right-hand corner figure shows the total number of trials
- Probability of obtaining the data $R$ heads in $N$ flips is given:

$$P(D|H,I) \propto H^R(1-H)^{N-R}$$

---

Parameter Estimation

Coin example (posterior)

- The width of the posterior becomes narrower with more data
- The position of the posterior maximum eventually settles

---

The uniform prior was chosen for its simplicity. **In other words:** it represented our initial ignorance. For illustrative purposes, consider two additional priors:

- Our subconscious believe that most coins are fair ($H = 0.5$)
- Reliable gossip that the coin is heavily biased ($H = 0$ and $H = 1$)
**Coin example (different prior)**

- **Few data:** resulting posteriors are quite different in detail
- **Ample data:** all become sharp and converge to the same answer

---

Coin example (different prior)

- **Few data**: resulting posteriors are quite different in detail
- **Ample data**: all become sharp and converge to the same answer

---

Consider a set of data \( \{ D_k \} = \{ D_1, D_2, D_3, \cdots, D_N \} \) for \( N \) coin flips.

Previously, we considered this to be an one-step process:
\[
P(H|D_2, D_1, I) \propto P(D_2, D_1|H, I)P(H|I) \quad \text{for} \quad N = 2
\]

However, we can also consider a sequential process:
\[
P(H|D_1, I) \propto P(D_1|H, I)P(H|I)
\]
\[
P(H|D_2, D_1, I) \propto P(D_2|H, D_1, I)P(H|D_1, I)
\]

We, therefore, used the posterior as the subsequent prior.

This only works if we have logical independence:
\[
P(D_2|H, I) = P(D_2|H, D_1, I).
\]

Be careful not to use the resulting posterior on the same data set.
Parameter Estimation

Reliabilities: best estimates, error bars and confidence intervals

- Posterior encodes our inference about the value of a parameter \( X \), given the data \( Y \) and the relevant background information \( I \).
- Convenient to summarize the posterior with two numbers:
  - best estimate (maximum)
  - measure of its reliability (width)

Fig. 1.2 A schematic illustration of the result of Laplace’ analysis of the mass of Saturn.

---

Reliabilities: best estimates, error bars and confidence intervals

- The **best estimate** $X_0$ is therefore easy to determine:
  \[
  \frac{dP}{dX} \bigg|_{X_0} = 0 \quad \text{and} \quad \frac{d^2P}{dX^2} \bigg|_{X_0} \leq 0.
  \]

- The posterior is peaky, so we consider the logarithm $L$ and expand it:
  \[
  L = L(X_0) + \frac{1}{2} \frac{d^2L}{dX^2} \bigg|_{X_0} (X - X_0)^2 + \cdots
  \]

- Posterior is approximated by the Gaussian distribution (not shown)
- Quadratic term is the dominant factor determining the width:
  \[
  \sigma = \left( -\frac{d^2L}{dX^2} \bigg|_{X_0} \right)^{-1/2}
  \]
- Our inference is conveyed very concisely: $X = X_0 \pm \sigma$
The probability that the true value of $X$ lies within $\pm \sigma$ of $X = X_0$ can be determined by integration:

$$P(X_0 - \sigma \leq X < X_0 + \sigma | D, I) = \int_{X_0 - \sigma}^{X_0 + \sigma} P(X, D | I) \, dX \approx 0.67$$

---

Fig. 2.3 The Gaussian, or normal, distribution. It is symmetric with respect to the maximum, at $x = \mu$, and has a full width at half maximum (FWHM) of about $2.35 \sigma$.

---

Reliabilities: best estimates, error bars and confidence intervals

- The probability that the true value of $X$ lies within $\pm \sigma$ of $X = X_0$ can be determined by integration:

$$\int_{X_0 - \sigma}^{X_0 + \sigma} P(X|D, I) \, dX \approx 0.67$$

---


---

Fig. 2.3 The Gaussian, or normal, distribution. It is symmetric with respect to the maximum, at $x = \mu$, and has a full width at half maximum (FWHM) of about $2.35 \sigma$. 
Let us take a look again at the coin flipping example, for which we found the probability of obtaining the data $R$ heads in $N$ flips:

$$P(D|H, I) \propto H^R (1 - H)^{N-R}$$

To reduce the gradient, we take the natural logarithm:

$$L = \text{constant} + R \log(H) + (N - R) \log(1 - H)$$

With this result we can subsequently determine the derivatives:

$$\frac{dL}{dH} = \frac{R}{H} - \frac{(N - R)}{(1 - H)}$$

and

$$\frac{d^2L}{dH^2} = -\frac{R}{H^2} - \frac{(N - R)}{(1 - H)^2}$$

Finally, we equate the first derivative to zero:

$$\frac{dL}{dH} \bigg|_{H_0} = \frac{R}{H_0} - \frac{(N - R)}{(1 - H_0)} = 0$$
Coin example (reliabilities)

- The **best estimate** and associated **error-bar** is then determined as:

\[
H_0 = \frac{R}{N} \quad \text{and} \quad \sigma = \sqrt{\frac{H_0(1 - H_0)}{N}}
\]

**Fig. 2.4** The Gaussian, or quadratic, approximation (dashed line) to the posterior pdf for the bias-weighting of the coin, given 9 heads in 32 flips.

---

Asymmetric posteriors

- The error-bar is not appropriate for asymmetric posteriors
- More reliable to determine a **confidence interval** (solve $X_1$ and $X_2$):

\[
\int_{X_1}^{X_2} P(X|D, I) \, dX \approx 0.95
\]

---

**Fig. 2.5** The shortest 95% confidence interval, shown by the shaded region.

---

Asymmetric posteriors

- The best estimate becomes vague for asymmetric posteriors
- More representative to determine the **expected value**:

\[
\langle X \rangle = \int X \, P(X|D, I) \, dX
\]

---

**Fig. 2.5** The shortest 95% confidence interval, shown by the shaded region.

---

Multimodal posteriors

- The most honest approach is just to display the posterior itself
- **Otherwise**: provide several best estimates and error bars

![Diagram](image)

**Fig. 2.6** A multimodal posterior pdf. Since its shape cannot be summarized by just a couple of numbers, the concept of a best estimate and an error-bar is inappropriate.

---

Parameter Estimation

Gaussian noise and averages

- As another example, let us estimate the mean of a Gaussian process
- The Gaussian distribution is often used as a model to describe the noise associated with experimental data (for now, *we just believe*)
- As such, the probability of the \( k \)th trail having a value \( x_k \) is

\[
P(x_k|\mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp \left[ -\frac{(x_k - \mu)^2}{2\sigma^2} \right]
\]

- **Question**: given a set of data \( \{x_k\} \), what is \( \mu \) assuming \( \sigma \) is given?

- We (obviously) use Bayes Theorem to determine the posterior:

\[
P(\mu|\{x_k\}, \sigma, I) \propto P(\{x_k\}|\mu, \sigma, I)P(\mu|\sigma, I)
\]
Parameter Estimation

Gaussian noise and averages

- Assuming that the data is **independent**, we have for the likelihood:

  \[
P(\{x_k\}|\mu, \sigma, I) = \prod_{k=1}^{N} P(x_k|\mu, \sigma, I)
  \]

- The above result follows from the repeated use of the product rule:

  \[
P(X, Y|I) = P(X|I)P(Y|I)
  \]

- This step is allowed since the measurements are **independent**:

  \[
P(X|Y, I) = P(X|I)
  \]

- **Note**: this is often not the case!
Gaussian noise and averages

- Knowledge of the Gaussian width tells us nothing about its position.
- As before, we are ignorant about the nature of this process:

\[
P(\mu|\sigma, I) = P(\mu|I) = \begin{cases} 
A & \text{for } \mu_{\text{min}} \leq \mu \leq \mu_{\text{max}}, \\
0 & \text{for otherwise}.
\end{cases}
\]

- The normalization constant \( A \) is determined by the range.
- We obtain the following for the posterior logarithm:

\[
L = \log [P(\mu|\{x_k\}, \sigma, I)] = \text{constant} - \sum_{k=1}^{N} \left[ \frac{(x_k - \mu)^2}{2\sigma^2} \right]
\]

- The constant includes all terms not involving \( \mu \).
Parameter Estimation

Gaussian noise and averages

- As before, with this result we determine the derivatives:

\[
\frac{dL}{d\mu} \bigg|_{\mu_0} = \sum_{k=1}^{N} \left[ \frac{x_k - \mu}{\sigma^2} \right] = 0
\]

- Since \( \sigma \) is independent of \( k \), we can rearrange the sum as follows:

\[
\sum_{k=1}^{N} x_k = \sum_{k=1}^{N} \mu_0 = N \mu_0
\]

- The **best estimate** is therefore just the arithmetic mean:

\[
\mu_0 = \frac{1}{N} \sum_{k=1}^{N} x_k
\]

- **Note**: the best estimate is independent of the measurement error
Parameter Estimation

Gaussian noise and averages

- Next, using the previous result we determine the second derivative:

\[
\frac{d^2 L}{d\mu^2} \bigg|_{\mu_0} = - \sum_{k=1}^{N} \frac{1}{\sigma^2} = - \frac{N}{\sigma^2}
\]

- Obviously, the error-bar depends on the measurement error.

- Our inference is therefore conveyed very concisely (should be a familiar result):

\[
\mu = \mu_0 \pm \frac{\sigma}{\sqrt{N}}
\]

- **Note**: the reliability is proportional to the number of measurements

- The above is not an approximation: it is the exact solution
Sometimes the magnitude of the error-bars are not the same for each datum. We then obtain the following the posterior logarithm:

\[
L = \log P(\mu|\{x_k\}, \sigma_k, I) = \text{constant} - \sum_{k=1}^{N} \left[ \frac{(x_k - \mu_k)^2}{2\sigma^2} \right]
\]

The best estimate is now slightly more complicated than before:

\[
\mu_0 = \frac{\sum_{k=1}^{N} w_k x_k}{\sum_{k=1}^{N} w_k} \quad \text{where} \quad w_k = \frac{1}{\sigma_k^2}
\]

Our inference is however still conveyed very concisely:

\[
\mu = \mu_0 \pm \left( \sum_{k=1}^{N} w_k \right)^{-1/2}
\]
A lighthouse is somewhere off a piece of straight coastline at a position $\alpha$ along the shore and a distance $\beta$ out at sea.

It emits a series of flashes at random intervals (therefore also random azimuths), which are intercepted on the coast by photodetectors.

If $N$ flashes were noted at positions $\{x_k\}$, where is the lighthouse?

---

Lighthouse problem

- After trigonometry and change of variables, the likelihood is found:

\[
P(x_k | \alpha, \beta, I) = \frac{\beta}{\pi \left[ \beta^2 + (x_k - \alpha)^2 \right]}
\]

- **Cauchy distribution**: frequently encountered in physics.

---

**Lighthouse problem**

- Fix the offshore distance $\beta$ (limit problem to one parameter).
- We are ignorant about the location along the shoreline:
  
  $$P(\alpha|\beta, I) = P(\alpha|I) = \begin{cases} A & \text{for } \alpha_{\text{min}} \leq \alpha \leq \alpha_{\text{max}}, \\ 0 & \text{for otherwise.} \end{cases}$$

- The detectors do not influence one another (thus independence):
  
  $$P(\{x_k\}|\alpha, \beta, I) = \prod_{k=1}^{N} P(x_k|\alpha, \beta, I)$$

- We obtain the following for the posterior logarithm:
  
  $$L = \text{constant} - \sum_{k=1}^{N} \log \left[ \beta^2 + (x_k - \alpha)^2 \right]$$
Lighthouse problem

- Best estimate of the position $\alpha_0$ is given by the posterior maximum:

$$\frac{dL}{d\alpha}\bigg|_{\alpha_0} = 2 \sum_{k=1}^{N} \frac{x_k - \alpha_0}{\beta^2 + (x_k - \alpha_0)^2} = 0$$

- An analytical solution will confound us (use numerical brute force)
- Visual representation of our inference about the lighthouse
Lighthouse problem

- Visual representation of our inference about the lighthouse
- Evolution of the posterior as the number of flashes detected increases
  - The position of the flashes are marked by the open circles
  - The number of data analyzed are shown in the corner

---

Parameter Estimation

Lighthouse problem

- Visual representation of our inference about the lighthouse
- Evolution of the posterior as the number of flashes detected increases
  - The position of the flashes are marked by the open circles
  - The number of data analyzed are shown in the corner

---

Central limit theorem

- Arithmetic mean of the data sounds like a good idea...
- Averages based on the central limit theorem (not always valid)
- For example, the mean is not appropriate for the Cauchy distribution

**Moral of the story**: the sample mean is not always a useful number. When in doubt, just show the complete posterior distribution.

---

Presence of background

- Previously we estimated only a single parameter
- Let us now consider multiple and nuisance parameters
- **Example**: often our signal is obscured by background noise:

![Graph showing a signal peak of amplitude $A$ sitting on a flat background of magnitude $B$.](image)

**Fig. 3.1** A signal peak of amplitude $A$ sitting on a flat background of magnitude $B$.

---

Presence of background

- **Simplest case**: consider a flat background of unknown magnitude $B$
- Signal has an amplitude $A$ with a peak of known shape and position
- **Example**: Number of photons detected at a given wavelength

Counts are proportional to the sum of the signal and background

Peak is assumed to be Gaussian, located at $x_0$ and having width $\nu$

$$D_k = n_0 \left[ Ae^{-(x_k-x_0)^2/2\nu^2} + B \right]$$

- The constant $n_0$ is related to the experimental conditions
- The trail/datum $D_k$ is therefore not an integer (an approximation)
Presence of background

- Poisson distribution is usually called upon for counting experiments:

\[ P(N|D) = \frac{D^N e^{-D}}{N!} \]

- The expectation value can be shown to be \( \langle N \rangle = D \)

- **Example**: these fluctuations are due to shot noise

---

**Fig. 3.2** The Poisson distribution \( \text{prob}(N|D) \) for: (a) \( D = 1.7 \), and (b) \( D = 12.5 \).

---

Presence of background

- We, therefore, have an assignment for the likelihood:
  \[ P(N_k|A, B, I) = \frac{D_k^{N_k}e^{-D_k}}{N_k!} \]
  
- **Note**: the background information \( I \) contains a lot of knowledge.

- Assume that channels do not influence each other (independence):
  \[ P(\{N_k\}|A, B, I) = \prod_{k=1}^{M} P(N_k|A, B, I) \]

- Posterior contains our inference about the **signal and background**:
  \[ P(A, B|\{N_k\}, I) \propto P(\{N_k\}|A, B, I)P(A, B|I) \]
Parameter Estimation II

Presence of background

- Amplitude of neither the signal nor the background can be negative:

\[ P(A, B|I) = \begin{cases} 
\text{constant} & \text{for } A \geq 0 \text{ and } B \geq 0, \\
0 & \text{for otherwise.} 
\end{cases} \]

- **Note**: this is a very powerful and not the most ignorant prior.

- We obtain the following for the posterior logarithm:

\[ L = \text{constant} + \sum_{k=1}^{M} \left[ N_k \log (D_k) - D_k \right] \]

- **Therefore**: maximize \( L \) for both the signal and the background

- We will need to perform a numerical estimation (**very easy example**)
Presence of background

- Data was simulated with a Poisson random number generator:

\[
P(N|D) = \frac{D^Ne^{-D}}{N!}
\]

- Signal is located at the origin with contours shown for \( A \) and \( B \)

---

Parameter Estimation II

Presence of background

- Data was simulated with a Poisson random number generator:

\[ P(N|D) = \frac{D^N e^{-D}}{N!} \]

- Signal is located at the origin with contours shown for A and B

---

Marginal distributions

The posterior contains **all of the information** regarding $A$ and $B$.

- We are often only interested in the signal information $A$:

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

- However, we might be interested in the background information $B$:

  $$P(B|\{N_k\}, I) = \int_0^\infty P(A, B|\{N_k\}, I) \, dA$$

- **Important**: marginal and conditional probabilities are not equal!

  $$P(A|\{N_k\}, I) \neq P(A|\{N_k\}, B, I)$$
Marginal distributions

- Comparison of the marginal and conditional probabilities are shown
- Note the narrowing for the conditional (dotted curve) scenario
- Calibration not required for large signal-to-noise ratios

---

Parameter Estimation II

Marginal distributions

- Comparison of the marginal and conditional probabilities are shown
- Note the narrowing for the conditional (dotted curve) scenario
- Calibration not required for large signal-to-noise ratios

\[\text{prob}(A|\{N_k\}, I)\]

\[\text{prob}(B|\{N_k\}, I)\]

Amplitude \( A \)  

Background \( B \)

---

If we bin the data, we should *technically* do the integral:

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

Although the data appears noisier, the effect of binning is cosmetic.

---

Reliabilities: best estimates, correlations and error-bars

- Prefer to summarize the posterior: best estimates and reliabilities.
- We have seen now that the posterior can contain several parameters.

For quantities of interest \( \{X_j\} \) with a posterior \( P \), the best estimate of their values are given by a solution to the system of equations:

\[
\left. \frac{\partial P}{\partial X_i} \right|_{X_{0j}} = 0
\]

**Note**: we need to avoid minima and saddle-points.

We will not consider the reliability here (similar as before).

As an example, consider the specific case of two variables \( X \) and \( Y \):

\[
\left. \frac{\partial L}{\partial X} \right|_{X_0, Y_0} = 0 \quad \text{and} \quad \left. \frac{\partial L}{\partial Y} \right|_{X_0, Y_0} = 0
\]
Reliabilities: best estimates, correlations and error-bars

Although we can determine reliability expressions like $\sigma_x$ and $\sigma_y$, they often do not provide the complete picture regarding the error bar.

The **variance** is formally defined as the expectation value of the square of the deviations from the mean:

$$\text{Var}(X) = \langle (X - \mu)^2 \rangle = \int (X - \mu)^2 P(X|D, I) \, dX$$

For the one-dimensional Gaussian distribution we find:

$$\langle (X - \mu)^2 \rangle = \sigma^2$$

We can therefore also extend this to more than one dimension:

$$\sigma_x^2 = \langle (X - X_0)^2 \rangle = \int (X - X_0)^2 P(X, Y|D, I) \, dX \, dY$$
Reliabilities: best estimates, correlations and error-bars

- Variance be further extended to consider simultaneous variations:
  \[ \sigma_{xy}^2 = \langle (X - X_0)(Y - Y_0) \rangle \]

- Called the **covariance** and measures **correlation** of the parameters:
  \[
  \begin{pmatrix}
    \sigma_x^2 & \sigma_{xy}^2 \\
    \sigma_{xy}^2 & \sigma_y^2
  \end{pmatrix}
  = - \begin{pmatrix} A & C \\ C & B \end{pmatrix}^{-1}
  \]

---

Numerical interlude

- Constructing the posterior is easy, *finding that maximum often not*.
- Brute force and ignorance is a very limited approach.