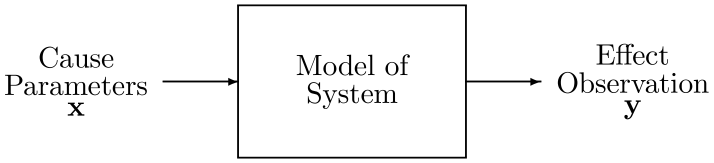
Chapter 5 Bayesian statistical inference and parameter estimation

5.1 Forward and inverse probability

Bayesian statistics provides a theory of inference which enables us to relate the results of observation with theoretical predictions. Consider the process of trying to understand some physical system. Theoretical physics constructs a model which tells us what observations to expect if certain causes are present. Abstractly, a set of causes can be represented by a parameter vector $\mathbf{x}$, and the result of the observations by an observation vector $\mathbf{y}$.



Theory tells us $p(\mathbf{y}|\mathbf{x})$, the conditional probability of the observation given the cause. This is usually called the **forward probability** density. Based on observations (and possibly control of some of the parameters), experimental physics involves trying to deduce the values of the parameters, under the assumption that the model is valid. Experimentalists want $p(\mathbf{x}|\mathbf{y})$, which is the conditional probability of the possible causes, given that some effect has been observed. This **inverse probability** represents our state of knowledge of $\mathbf{x}$ after measuring $\mathbf{y}$. In the context of inverse problem theory, $\mathbf{x}$ is the image and $\mathbf{y}$ is the data.

Examples:

1. A rod of length x is measured with precision σ. If we assume Gaussian errors in the measurement model, theory predicts that an observation y has the probability density

$$p(y|x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{y-x}{\sigma}\right)^2\right] \tag{5.1}$$

   Given one or more measurements y_i, what is our state of knowledge of x?

2. A radioactive solid with a long half-life decays at rate x disintegrations per second so that in time T, the probability of obtaining y counts is Poisson distributed, namely

$$p(y|x) = \frac{\exp(-xT)\,(xT)^y}{y!} \tag{5.2}$$

   In various intervals each of duration T, y_i counts were observed. What can we say about x?

3. A photograph $\mathbf{y}$ is taken of a scene $\mathbf{x}$ with an out-of-focused camera so that

$$\mathbf{y} = F(\mathbf{x}) + \mathbf{n} \tag{5.3}$$

   where F denotes a "blurring operator" and $\mathbf{n}$ denotes a noise vector. The forward probability density is given by

$$p(\mathbf{y}|\mathbf{x}) = p(\mathbf{n} = \mathbf{y} - F(\mathbf{x})) \tag{5.4}$$

   where the noise statistics are assumed to be known. Given $\mathbf{y}$, how can we process it to recover $\mathbf{x}$, and how confident can we be of the result?

5.2 Bayes' theorem

The central result that helps us solve all of these problems is Bayes' theorem, which is based on the relationship between joint and conditional probabilities.

Given events A and B,

$$p(A,B) = p(A|B)p(B) = p(B|A)p(A) \tag{5.5}$$

Hence,

$$p(A|B) = \frac{1}{p(B)}p(B|A)p(A) \tag{5.6}$$

This (and extensions to include more than 2 events) is called Bayes' theorem. It relates forward probabilities $p(B|A)$ to inverse probabilities $p(A|B)$.

In terms of parameters and observations,

$$p(\mathbf{x}|\mathbf{y}) = \frac{1}{p(\mathbf{y})}p(\mathbf{y}|\mathbf{x})p(\mathbf{x}) \tag{5.7}$$

We shall interpret this equation as telling us how we should change our state of knowledge of $\mathbf{x}$ as a result of making an observation which yields the result $\mathbf{y}$.

- On the right-hand side, $p(\mathbf{x})$ is the probability function which represents what we know (or believe) about the parameter $\mathbf{x}$ **before** making the observation. It is known as the **prior probability** and summarizes our initial state of knowledge about the parameter.
- On the left-hand side, $p(\mathbf{x}|\mathbf{y})$ is the probability function which tells us what we know about the parameter $\mathbf{x}$ **after** making the observation. It is called the **posterior probability**.
- The way we change the prior probability into the posterior probability is to multiply by two factors. One of these is $p(\mathbf{y}|\mathbf{x})$ which is just the forward probability function which is determined by theoretical physics. Notice that in this application we think about this as a function of $\mathbf{x}$ for a fixed observation $\mathbf{y}$. When viewed in this way, the forward probability is called the **likelihood function**.
- The remaining factor $p(\mathbf{y})^{-1}$ can be determined by normalization since we know that the sum of the posterior probability distribution function over all possible causes must be equal to one.

We now consider a specific example to illustrate the operation of this single most important result in statistical inference and data analysis.

5.2.1 The tramcar problem

A town has n tramcars labelled from 1 to n. On k separate occasions, I see trams numbered m_1, m_2,..., m_k. Based on this information, what can I say about the number of tramcars in the town?

At first sight, all that one can say as a result of seeing tramcar m is that there are at least m tramcars in the town. However one intuitively feels that if one has lived in the town for some time the highest numbered tramcar that one has ever seen will be close to the actual number of tramcars. We shall now see how Bayes' theorem can make these considerations precise.

Let us start off by analyzing the effect of the first observation, which is of tramcar m_1. The parameter we wish to estimate is n so writing down Bayes' theorem we obtain

$$p(n|m_1) = \frac{1}{p(m_1)} \times p(m_1|n) \times p(n) \tag{5.8}$$

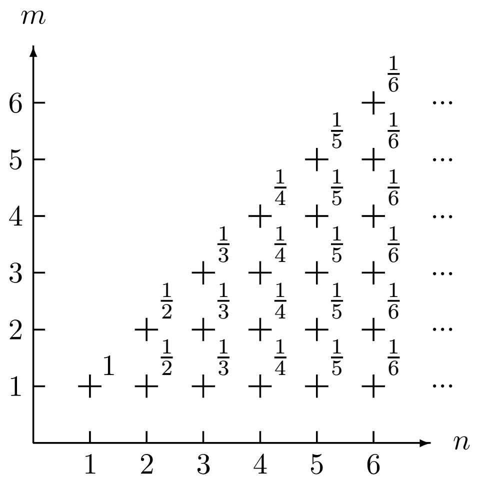
Prior: Before making any observations, let us suppose that I believe that it is equally probable that there are any number between 1 and N tramcars. Later we shall consider the effect of changing N. This corresponds to the choice

$$p(n) = \begin{cases} 1/N & \text{if } n \leq N \\ 0 & \text{otherwise} \end{cases} \tag{5.9}$$

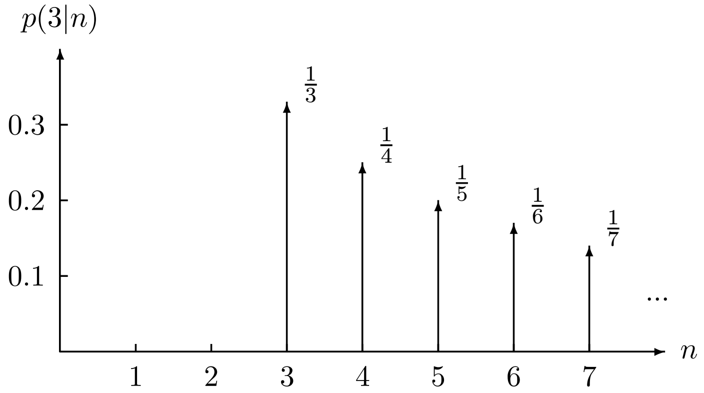
Likelihood: This is just the forward probability function. If there happen to be n tramcars in town, the probability that I see tramcar m is $1/n$ if $m \leq n$ and it is impossible to see a tramcar numbered $> n$. Thus

$$p(m|n) = \begin{cases} 1/n & \text{if } m \leq n \\ 0 & \text{otherwise} \end{cases} \tag{5.10}$$

This likelihood function can be represented by the following diagram. Each cross denotes particular values of n and m and the function value associated with each cross is written next to it.



We consider the posterior probability as a function of n for a fixed observation m_1. From the expression (5.8) for the posterior probability, this involves looking across a row of the likelihood function. For example, if $m_1 = 3$, the section through the likelihood function is $p(3|n)$ which appears as shown.



The posterior probability function is found by multiplying together the prior and the likelihood. The normalization term $1/p(m_1)$ does not depend on n and so we can say that $p(n|m_1)$ is proportional to

$$p(n|m_1) \propto \begin{cases} 1/(nN) & \text{if } m_1 \le n \le N \\ 0 & \text{otherwise} \end{cases} \tag{5.11}$$

Since the sum over all n must be unity, we easily determine $p(m_1) = \sum_{n=m_1}^{N} (nN)^{-1}$. From the posterior probability, we notice that

- It is not possible for there to be fewer than m_1 tramcars,
- The posterior probability is a sampled section of a hyperbola from m_1 to N,
- If we approximate the discrete posterior probability function by a continuous one, the mean of the posterior probability function is

$$\mu \approx \int_{m_1}^{N} n\, p(n|m_1)\, dn = \frac{N - m_1}{\log N - \log m_1} \tag{5.12}$$

At this stage the mean tends to infinity if we let N tend to infinity. This indicates that the posterior probability function is still strongly dependent on the prior probability that we chose. With only one observation, we have not yet learnt very much and are still highly influenced by our prior suppositions.

5.3 Multiple Observations

Now let us suppose that we see tramcar number m_2. We want to compute $p(n|m_1, m_2)$. Again using Bayes' theorem, assuming that the observations are independent,

$$\begin{aligned} p(n|m_1, m_2) &= \frac{p(n, m_1, m_2)}{p(m_1, m_2)} = \frac{p(m_2|n, m_1)\, p(n, m_1)}{p(m_2)\, p(m_1)} \\ &= \frac{1}{p(m_2)}\, p(m_2|n)\, p(n|m_1) \end{aligned} \tag{5.13}$$

So for independent observations, we can use the posterior probability for the first observation $p(n|m_1)$ as the prior probability for the second observation. In effect, the likelihood functions are multiplied together. For k independent observations,

$$p(n|m_1, m_2, ..., m_k) = \frac{1}{p(m_1)p(m_2)...p(m_k)} \times p(m_k|n)\, p(m_{k-1}|n)...p(m_1|n) \times p(n) \tag{5.14}$$

This formalizes the process of how we learn from a succession of independent observations.

For the tramcar problem, after k observations we find that the posterior probability function is

$$p(n|m_1, m_2, ..., m_k) = \begin{cases} \mathcal{N}/n^k & \text{if } \max(m_1, m_2, ..., m_k) \le n \le N \\ 0 & \text{otherwise} \end{cases} \tag{5.15}$$

The normalization constant $\mathcal{N}$ is chosen so these probabilities sum to one. We see that

- $M = \max(m_1, m_2, ..., m_k)$ is a lower bound for the number of tramcars.
- With more observations (k large), the posterior probability falls off more sharply with n
- In the approximation in which the discrete probability function is approximated by a continuous probability density, the mean of the posterior probability is

  $$\mu = \left(\frac{k-1}{k-2}\right) M \left[\frac{1-(M/N)^{k-2}}{1-(M/N)^{k-1}}\right] \qquad \text{for } k > 2 \tag{5.16}$$

  As $N \to \infty$, $\mu \to (k-1)M/(k-2)$ which is just slightly larger than M. We see that as k is increased, the cutoff N in the prior makes little difference. This means that the observations are becoming more important to the inference than our initial presuppositions.

- For large N, the variance of the posterior probability is approximately

  $$\sigma^2 = \left(\frac{k-1}{k-3}\right)\left[\frac{M}{k-2}\right]^2 \tag{5.17}$$

  This becomes small as k increases. We can thus be more confident of the total number of tramcars as we see more of them.

- The likelihood function for k observations $p\,(m_1, m_2, ..., m_k|n)$ depends on the observations $m_1, ..., m_k$ only through the two quantities k and $M = \max(m_1, m_2, ..., m_k)$. When a likelihood function is completely determined by a set of quantities, these quantities are said to form a set of **sufficient statistics** for the estimation process. In other words, a set of sufficient statistics summarizes all the information present in the set of data which is relevant for the estimation of the desired quantities.

- In the Bayesian viewpoint, our **complete state of knowledge** of the parameter(s) after the observations is given by the posterior probability density function. Often, we are asked for a "best estimate" of the parameters rather than the entire representation of our state of knowledge. The procedure for selecting such an estimate is **not** part of the framework and can sometimes be problematical. Various choices are to use the MAP (maximum *à posteriori*) estimate, the mean of the posterior probability, the maximum likelihood estimate or some other *ad hoc* estimator.

Exercises

1. In one of three boxes there are two gold coins, in another there are two silver coins and in the third there are one gold coin and one silver coin. I go to one of the boxes at random and extract a coin. Given that this coin is gold, what is the probability that the other coin in that box is also gold? Repeat the problem if there are a hundred coins in each box, all gold in the first, all silver in the second and one gold and ninety-nine silver in the third.
2. In a town, 80 percent of all taxis are blue and the other 20 percent are green. One night, an accident involving a taxi occurred, and a witness who is able to identify the colour of a taxi under the lighting conditions with 75 percent accuracy testifies that the colour of the taxi involved was green. Compute the posterior probability of the colour of the taxi after receiving the testimony.
3. A bag contains 80 fair coins and 20 double-headed coins. A coin is withdrawn at random and tossed, yielding a head. What is the probability that the coin is fair? How many times must the coin be tossed before we can be 99 percent sure that it is double-headed?

We shall now embark on a series of examples showing how these principles may be applied to a variety of problems.

5.4 Estimating a quantity with Gaussian measurement errors

Consider the problem of measuring the length x of a rod using a measuring instrument for which

$$p(y|x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(y-x)^2}{2\sigma^2}\right) \tag{5.18}$$

This means that each observation comes from a Gaussian of standard deviation σ centred about the true value of x.

Experimentally, we measure the length several times and obtain an independent sequence of measurements $\{y_1, y_2, ..., y_N\}$. The likelihood function for these is

$$p(y_1, y_2, ..., y_N|x) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left(-\sum_{i=1}^{N} \frac{(y_i-x)^2}{2\sigma^2}\right) \tag{5.19}$$

The posterior probability for x starting with a prior probability of the form $p(x)$ is

$$p(x|y_1, ..., y_N) = \frac{\mathcal{N}}{(2\pi\sigma^2)^{N/2}} \exp\left(-\sum_{i=1}^{N} \frac{(y_i-x)^2}{2\sigma^2}\right) p\left(x\right) \tag{5.20}$$

where $\mathcal{N} = 1/p(y_1, ..., y_N)$ is a normalization constant. We are interested in seeing the right-hand side as a function of x. This is facilitated by expanding the squares and collecting terms in various powers of x yielding

$$p(x|y_1, ..., y_N) = \frac{\mathcal{N}}{(2\pi\sigma^2)^{N/2}} \exp\left[-\left(\frac{N}{2\sigma^2}\right)x^2 + \left(\frac{1}{\sigma^2}\right)\sum_{i=1}^{N} y_i x - \left(\frac{1}{2\sigma^2}\right)\sum_{i=1}^{N} y_i^2\right] p(x) \tag{5.21}$$

$$\propto \exp\left[-\frac{N}{2\sigma^2}\left(x - \frac{1}{N}\sum_{i=1}^{N} y_i\right)^2\right] p\left(x\right) \tag{5.22}$$

where all terms not explicitly dependent on x have been included in the proportionality. We see that the effect of collecting the data is to multiply the prior $p(x)$ by a Gaussian of mean $\sum y_i/N$ and standard deviation $\sigma/\sqrt{N}$. If the prior probability $p(x)$ before making the measurement is approximately uniform in a sufficiently large interval around $\sum y_i/N$, the posterior probability function will be almost completely determined by the data. For a Gaussian posterior probability density, the mean, median and mode all coincide, so there is little doubt as to what should be quoted as the estimate. The variance of the posterior probability represents our confidence in the result. We thus quote the mean of the data points $m = \sum y_i/N$ as the best estimate for x and give its uncertainty as $\sigma/\sqrt{N}$.

5.4.1 Estimating the measurement error as well as the quantity

In the above we assumed that the error in each datum σ is known. Often it is unknown and we seek to estimate σ as well as x from the data. This can be readily done within the Bayesian formulation by considering σ as an additional parameter. We can write

$$p(x,\sigma|y_1,y_2,...,y_N) = \mathcal{N}\, p(y_1,y_2,...,y_N|x,\sigma)\, p(x,\sigma) \tag{5.23}$$

where the likelihood $p(y_1,y_2,...,y_N|x,\sigma)$ has the same form as (5.19) above. Evaluating this yields

$$p(x,\sigma|y_1,y_2,...,y_N) = \frac{\mathcal{N}}{(2\pi\sigma^2)^{N/2}} \exp\left(-\frac{N}{2\sigma^2}\left[(x-m)^2+s^2\right]\right) p(x,\sigma) \tag{5.24}$$

where $m = (1/N)\sum y_i$ and $s^2 = (1/N)\sum y_i^2 - m^2$. This is a joint distribution for x and σ. Again if the prior probability is approximately constant, the first factor (essentially the likelihood function) determines the posterior probability. In the Bayesian framework, this posterior probability density summarizes all that we know about the parameters after making the measurement. In the following, we shall find the following integral useful

$$\int_0^\infty \frac{1}{\sigma^k} \exp\left(-\frac{A}{\sigma^2}\right) d\sigma = \frac{\Gamma\left(\frac{k-1}{2}\right)}{2\sqrt{A^{k-1}}}. \tag{5.25}$$

For a flat prior, the normalized form of the posterior probability is given by

$$p(x,\sigma|y_1,y_2,...,y_N) = \sqrt{\frac{8}{N\pi}}\left(\frac{Ns^2}{2}\right)^{N/2} \frac{1}{s^2\Gamma\left(\frac{1}{2}N-1\right)\sigma^N} \exp\left(-\frac{1}{2}\frac{N}{\sigma^2}\left[(x-m)^2+s^2\right]\right). \tag{5.26}$$

The peak of this function is given by

$$x_{\mathrm{MAP}} = m \tag{5.27}$$

$$\sigma_{\mathrm{MAP}} = s \tag{5.28}$$

where "MAP" stands for **maximum à posteriori estimate**, i.e., the mode of the posterior probability density. This is also the **maximum likelihood estimate,** since the prior is assumed to be flat.

From the joint posterior probability density, we can find the marginal probability densities by integrating over the variable(s) which we do not wish to consider. The results are

$$p\left(x|y_1,y_2,...,y_N\right) = \frac{\Gamma\left(\frac{1}{2}N-\frac{1}{2}\right)}{\Gamma\left(\frac{1}{2}N-1\right)} \frac{s^{N-2}}{\pi^{\frac{1}{2}}\left[(x-m)^2+s^2\right]^{\frac{N-1}{2}}}, \tag{5.29}$$

$$p\left(\sigma|y_1,y_2,...,y_N\right) = \frac{2}{\Gamma\left(\frac{1}{2}N-1\right)}\left(\frac{Ns^2}{2}\right)^{N/2-1} \frac{1}{\sigma^{N-1}} \exp\left(-\frac{1}{2}\frac{N}{\sigma^2}s^2\right). \tag{5.30}$$

In Figures 5.1 and 5.2 we show the joint and marginal posterior probability densities for the cases of $N = 3$ and $N = 50$ measured data points. These graphs are plotted in terms of the variables $(x - m)/s$ and σ/s in order to make them independent of m and s. The joint posterior probability density is shown as a contour diagram. The contour label λ indicates the contour at which the joint posterior probability density has fallen to a value of $\exp\left(-\lambda^2/2\right)$ of the peak value.

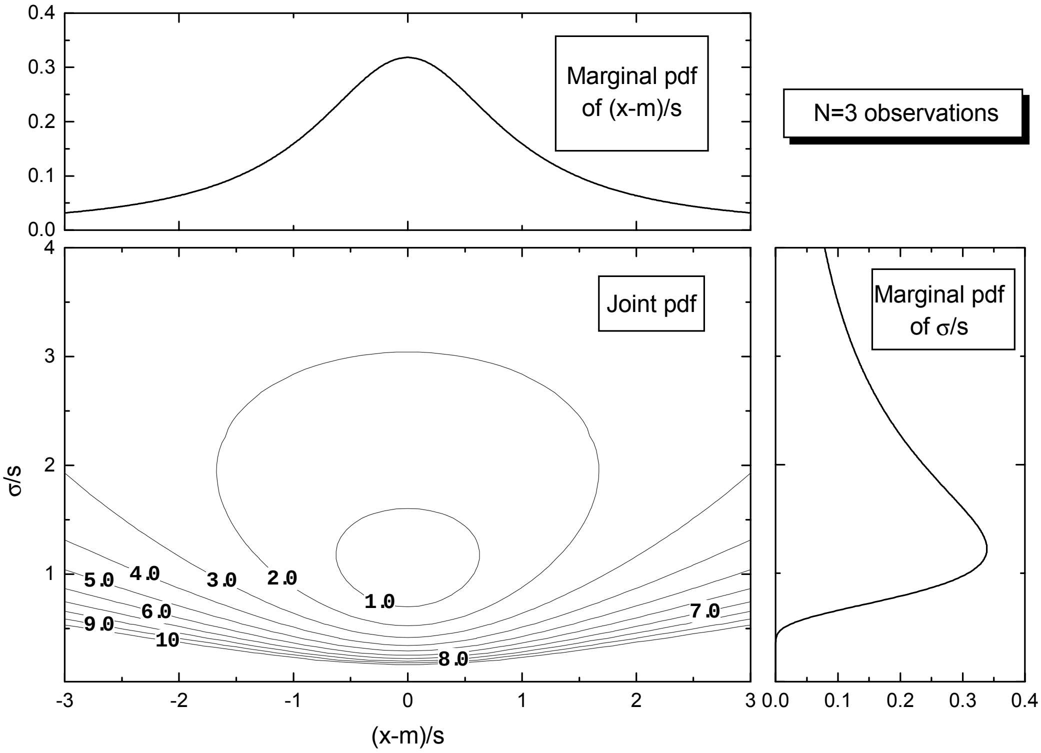


Figure 5.1 Posterior probability densities after $N = 3$ observations.

When giving estimates of x and σ, the peak of the posterior probability may not be a good representative of the probability distribution. This is especially the case for σ when N is small, since the posterior probability of σ is quite asymmetrical. It is possible (with some effort) to find analytic expressions for the mean of these probability densities,

$$\mathrm{E}\left[x|\mathbf{y}\right] = m, \tag{5.31}$$

$$\mathrm{E}\left[\sigma|\mathbf{y}\right] = \sqrt{\frac{N}{2}}\frac{\Gamma\left(\frac{1}{2}N - \frac{3}{2}\right)}{\Gamma\left(\frac{1}{2}N - 1\right)}s \sim \left(1 + \frac{7}{4N} + \frac{145}{32N^2} + ...\right)s. \tag{5.32}$$

For finite N, the asymmetry in the posterior probability for σ pushes the mean higher than the

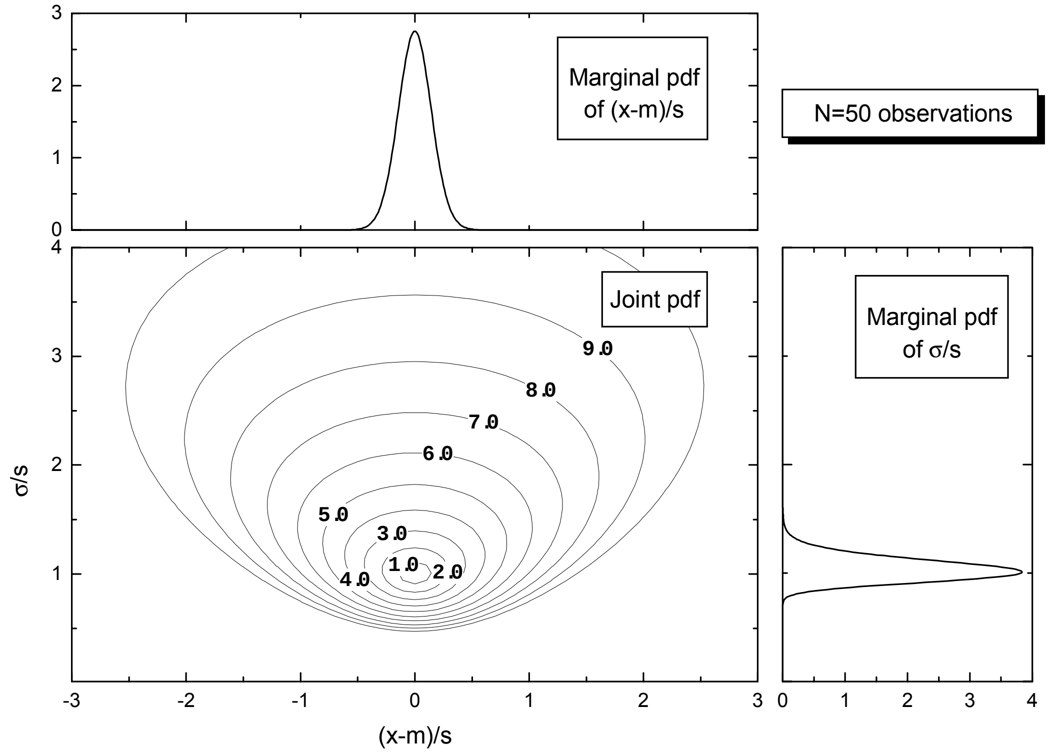


Figure 5.2 Posterior probability densities after $N = 50$ observations

mode which is at s. The covariance of the posterior probability distribution is

$$\mathrm{E}\left[(\Delta x)^2|\mathbf{y}\right] = \mathrm{E}\left[x^2|\mathbf{y}\right] - (\mathrm{E}\left[x|\mathbf{y}\right])^2 = \frac{s^2}{N-4} \sim \left(\frac{1}{N} + \frac{4}{N^2} + ...\right) s^2, \tag{5.33}$$

$$\mathrm{E}\left[(\Delta x)(\Delta \sigma)|\mathbf{y}\right] = \mathrm{E}\left[x\sigma|\mathbf{y}\right] - \mathrm{E}\left[x|\mathbf{y}\right]\mathrm{E}\left[\sigma|\mathbf{y}\right] = 0, \tag{5.34}$$

$$\begin{aligned}\mathrm{E}\left[(\Delta \sigma)^2|\mathbf{y}\right] &= \mathrm{E}\left[\sigma^2|\mathbf{y}\right] - (\mathrm{E}\left[\sigma|\mathbf{y}\right])^2 \\ &= \left[\frac{1}{N-4} - \frac{1}{2}\left(\frac{\Gamma\left(\frac{1}{2}N-\frac{3}{2}\right)}{\Gamma\left(\frac{1}{2}N-1\right)}\right)^2\right] N s^2 &(5.35) \\ &\sim \left(\frac{1}{2N} + \frac{31}{8N^2} + ...\right) s^2. &(5.36)\end{aligned}$$

The asymptotic approximations hold for large values of N and are based on Stirling's approximation for the Γ function, namely

$$\log\Gamma(z) \sim \frac{1}{2}\log(2\pi) - z + \left(z - \frac{1}{2}\right)\log z + \frac{1}{12z} + \mathrm{O}\left(\frac{1}{z^3}\right)$$

from which we deduce that for large z,

$$z^{b-a}\frac{\Gamma(z+a)}{\Gamma(z+b)} \sim 1+\frac{(a-b)(a+b-1)}{2z} \tag{5.37}$$

$$+\frac{1}{12}\frac{(a-b)(a-b-1)\left(3(a+b-1)^2-a+b-1\right)}{2z^2}+\mathrm{O}\left(\frac{1}{z^3}\right). \tag{5.38}$$

We see that it is not until $N > 4$ that these probability densities have finite variances. Using the above expressions, we can give estimates of x and σ as well as the uncertainties in these quantities. Notice that the uncertainty of our estimate of x still falls as $N^{-1/2}$ for large N just as when the value of σ is known. However, for smaller values of N, this uncertainty is larger than when σ is known.

5.5 Estimating radioactive source strength and half-life

Suppose that a radioactive source has strength which decays with time according to the law

$$S(t)=S_0\exp(-\alpha t) \tag{5.39}$$

where the half-life is $(\ln 2)/\alpha$. At time $t=0$, we turn on an ideal Geiger counter and record the counts which occur until time T. From the record of counts, how should we estimate the values of S_0 and of α?

In order to carry out a Bayesian analysis, we want the posterior probability of the parameters S_0 and α given a particular record of counts. The forward problem requires us to find the probability of getting the particular sequence of counts given values of S_0 and α. We note that we can record the times at which the counts occur. Let us divide the interval $[0,T]$ into short subintervals of duration Δt starting at t_0, $t_1, ..., t_k, ..., t_{T/\Delta t-1}$. Each of these intervals is assumed to be so short that there is at most one count in an interval. The probability that a count occurs in the subinterval starting at t_k is $S(t_k)\Delta t$.

Let us suppose that there were a total of N counts in the interval $[0,T]$ and that they occured in the subintervals starting at $t_{k_1}, t_{k_2}, ..., t_{k_N}$. The probability of this particular record is the product of the probabilities that counts *did* occur in the specified subintervals and that they *did not* occur in the others. This is

$$\Pr(t_{k_1}, t_{k_2}, ..., t_{k_N}|S_0,\alpha)=(\Delta t)^N\left[\prod_{i=1}^N S(t_{k_i})\right]\prod_{k\neq k_i}\left[1-S(t_k)\Delta t\right] \tag{5.40}$$

By Bayes' theorem,

$$p(S_0,\alpha|t_{k_1},...,t_{k_N})\propto p(S_0,\alpha)\left[\prod_{i=1}^N S_0e^{-\alpha t_{k_i}}\right]\prod_{k\neq k_i}\left[1-S_0e^{-\alpha t_k}\Delta t\right] \tag{5.41}$$

$$\begin{aligned}\log p(S_0,\alpha|t_{k_1},...,t_{k_N}) &= \text{const}+\log p(S_0,\alpha)+N\log S_0\\ &\quad -\alpha\left(\sum_{i=1}^N t_{k_i}\right)+\sum_{k\neq k_i}\log\left[1-S_0e^{-\alpha t_k}\Delta t\right]\end{aligned} \tag{5.42}$$

As Δt becomes small, we can expand the last logarithm and retain only the linear term

$$\sum_{k\neq k_i} \log\left[1 - S_0 e^{-\alpha t_k}\Delta t\right] \approx -S_0 \sum_{k\neq k_i} e^{-\alpha t_k}\Delta t \to -S_0 \int_0^T e^{-\alpha t}\,dt = -\frac{S_0}{\alpha}\left(1 - e^{-\alpha T}\right) \tag{5.43}$$

and so

$$\log p\left(S_0, \alpha | t_{k_1}, ..., t_{k_N}\right) = \text{const} + \log p\left(S_0, \alpha\right) + N \log S_0 - \alpha\left(\sum_{i=1}^N t_{k_i}\right) - \frac{S_0}{\alpha}\left(1 - e^{-\alpha T}\right). \tag{5.44}$$

The *log likelihood function* consists of all terms on the right-hand side excluding that containing the prior probability density. From the form of this function, it is clear that the sufficient statistics for this problem are N, the number of counts in the interval and $\sum t_{k_i}$ which is the *sum* of the decay times. For any data set, we only need to calculate these sufficient statistics in order to completely determine the likelihood function.

If we assume that the prior is flat, we may examine how the log likelihood varies with the parameters S_0 and α. One possible strategy for estimating S_0 and α is to maximize the likelihood function for the given data, this gives the so-called *maximum likelihood* estimator. In this example the maximum likelihood estimator of α is the solution of

$$\frac{1 - e^{-\alpha T}\left(1 + \alpha T\right)}{\alpha\left(1 - e^{-\alpha T}\right)} = \frac{1}{N}\sum_{i=1}^N t_{k_i} \tag{5.45}$$

and having found α, the estimate for S_0 is

$$S_0 = \frac{N\alpha}{1 - e^{-\alpha T}}. \tag{5.46}$$

We can check that this result is reasonable by considering the limit in which the source half life is very long compared to the measurement time T. In this case, the rate of counts is constant over the interval and we expect the mean of the count times on the right hand side of (5.45) to be equal to $T/2$. It is easy to check that the solution for α in this situation is $\alpha = 0$. Substituting into (5.46) gives

$$S_0 = \frac{N}{T}. \tag{5.47}$$

So the maximum likelihood estimate for the source strength when the decay is negligible is just the total number of counts divided by the counting time, as we might have expected.

In Figure (5.3), we show the contours of the log likelihood function for a source with $S_0 = 100$ and $\alpha = 0.5$. The counting time was 10 sec, during which time $N = 215$ counts were detected and the sum of the decay times was 435.2. If the likelihood is approximated by a Gaussian, the $n\sigma$ level corresponds to the probability density falling to $\exp\left(-n^2/2\right)$ of the peak. In the figure, contour lines are drawn at the maximum value of the log likelihood minus $n^2/2$. Recall that the quoted standard error is the *projection* of the 1σ uncertainty ellipse onto the coordinate axes. Notice that in this example there is a positive correlation between the values of S_0 and of α. This means that if we increase our estimate of S_0, it is also necessary to increase our estimate of the decay rate in order to maintain the same data misfit.

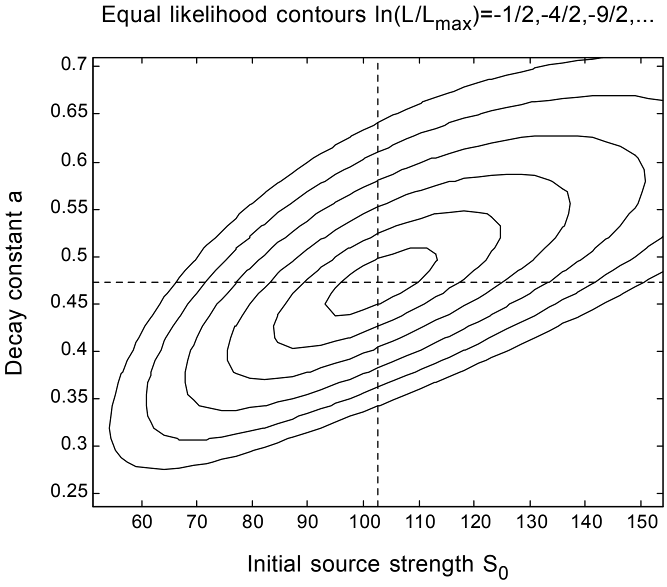


Figure 5.3 Contour plot of log likelihood for radioactive decay problem

5.6 Approximation of unimodal probability densities by Gaussians

It is often inconvenient to have to display an entire posterior probability density. If the posterior probability has a single well-defined peak and falls away from the maximum in an approximately Gaussian fashion, it is usual to approximate the probability density by a Gaussian. The key advantage of this is that a Gaussian is *completely determined* when one specifies the mean vector and the covariance matrix. Often, the user is not interested in the off-diagonal elements of the covariance matrix and only the diagonal elements (namely the variances) are required.

A Gaussian probability density is unimodal and has the property that its logarithm is a quadratic function of the variables. The maximum of this quadratic form gives the position of the mean and the *curvature* (second derivative) at the maximum gives information about the variance. In order to approximate a unimodal probability density $p(\mathbf{x})$ by a Gaussian $g(\mathbf{x})$, we adopt the following procedure:

1. Find the logarithm of the probability density $\log p(\mathbf{x})$ and find the position of its maximum $\hat{\mathbf{x}}$. This gives the mean of the approximate Gaussian.

2. Expand $\log p(\mathbf{x})$ using a Taylor series to second order about the point $\hat{\mathbf{x}}$. The linear terms vanish since $\hat{\mathbf{x}}$ is an extremum. Thus we find

$$\log p(\mathbf{x}) = \log p(\hat{\mathbf{x}}) - \frac{1}{2}(\mathbf{x} - \hat{\mathbf{x}})^t \mathbf{Q}(\mathbf{x} - \hat{\mathbf{x}}) \tag{5.48}$$

where $\mathbf{Q}$ is the negative of the matrix of second derivatives with components of $\log p(\mathbf{x})$, i.e.,

$$Q_{ij} = -\frac{\partial^2 \log p}{\partial x_i \partial x_j} \tag{5.49}$$

3. The approximating Gaussian is

$$g(\mathbf{x}) = \mathcal{N} \exp\left[-\frac{1}{2}(\mathbf{x}-\hat{\mathbf{x}})^t \mathbf{Q}(\mathbf{x}-\hat{\mathbf{x}})\right] \tag{5.50}$$

where $\mathcal{N}$ is a normalization factor. The covariance matrix for the approximating Gaussian is $\mathbf{Q}^{-1}$.

5.6.1 Joint estimation of quantity and measurement error problem

We now return to the posterior probability function for the problem of measuring a constant with Gaussian distributed errors

$$p(x,\sigma|y_1,y_2,...,y_N) = \frac{\mathcal{N}}{(2\pi\sigma^2)^{N/2}} \exp\left(-\frac{N}{2\sigma^2}\left[(x-m)^2+s^2\right]\right) p(x,\sigma)$$

If we suppose that the prior $p(x,\sigma)$ is flat,

$$\log p(x,\sigma|y_1,y_2,...,y_N) = \text{const } - N\log\sigma - \frac{N}{2\sigma^2}\left[(x-m)^2+s^2\right] \tag{5.51}$$

We find that

$$\frac{\partial \log L}{\partial x} = -\frac{N}{\sigma^2}(x-m) \tag{5.52}$$

$$\frac{\partial \log L}{\partial \sigma} = -\frac{N}{\sigma} + \frac{N}{\sigma^3}\left[(x-m)^2+s^2\right] \tag{5.53}$$

and so the maximum occurs at $\hat{x} = m$ and $\hat{\sigma} = s$. Calculating the negative of the matrix of second derivatives and xevaluating this at $(\hat{x},\hat{\sigma})$ yields

$$\mathbf{Q} = \begin{pmatrix} N/s^2 & 0 \\ 0 & 2N/s^2 \end{pmatrix} \tag{5.54}$$

The approximating Gaussian thus has mean (m,s) and a diagonal covariance matrix $\mathbf{Q}^{-1}$ with variance s^2/N in x and variance $s^2/(2N)$ in σ. Comparing these with the mean and covariance of the actual joint posterior probability density given above in equations (5.31) through (5.36), we see that the answers approach each other when N is large, but there are significant differences for small N.

5.6.2 Radioactive decay problem

The logarithm of the posterior probability in this case for a flat prior was

$$\log\, p\,(S_0,\alpha|t_{k_1},...,t_{k_N}) = \text{const} + N\log S_0 - \alpha\left(\sum_{i=1}^N t_{k_i}\right) - \frac{S_0}{\alpha}\left(1-e^{-\alpha T}\right). \tag{5.55}$$

The second derivatives are

$$\frac{\partial^2}{\partial S_0^2}\left(N\log S_0 - \alpha\left(\sum_{i=1}^N t_{k_i}\right) - \frac{S_0}{\alpha}\left(1-e^{-\alpha T}\right)\right) = -\frac{N}{S_0^2}$$

$$\frac{\partial^2}{\partial S_0 \partial\alpha}\left(N\log S_0 - \alpha\left(\sum_{i=1}^N t_{k_i}\right) - \frac{S_0}{\alpha}\left(1-e^{-\alpha T}\right)\right) = -\frac{1}{\alpha^2}\left[(1+\alpha T)\,e^{-\alpha T}-1\right]$$

$$\frac{\partial^2}{\partial \alpha^2}\left(N\log S_0 - \alpha\left(\sum_{i=1}^N t_{k_i}\right) - \frac{S_0}{\alpha}\left(1-e^{-\alpha T}\right)\right) = -\frac{S_0}{\alpha^3}\left[2-e^{-\alpha T}\left(T^2\alpha^2+2T\alpha+2\right)\right]$$

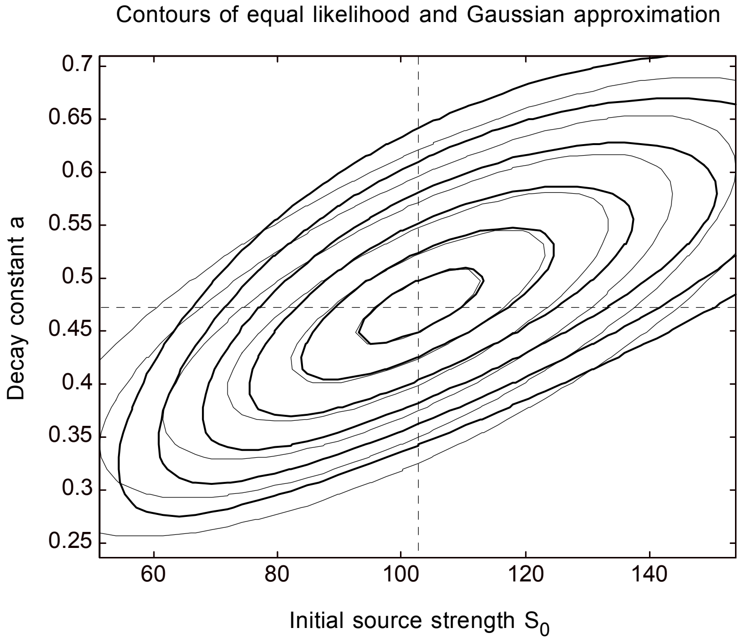


Figure 5.4 Gaussian approximation to likelihood function for radioactive decay problem

For the data given above, the inverse covariance matrix is

$$\mathbf{Q} = \begin{pmatrix} 0.0204 & -4.24 \\ -4.24 & 1650 \end{pmatrix}, \tag{5.56}$$

and the covariance matrix is

$$\mathbf{Q}^{-1} = \begin{pmatrix} 105 & 0.271 \\ 0.271 & 0.00130 \end{pmatrix}. \tag{5.57}$$

The square roots of the diagonal elements give the standard errors in the estimates. Thus for the data set in the example

$$S_0 = 103 \pm 10 \tag{5.58}$$

$$\alpha = 0.47 \pm 0.04. \tag{5.59}$$

The graph of the Figure 5.4 shows the contours of the approximate Gaussian posterior probability density (thin lines) superimposed upon the actual posterior probability (thick lines).

5.6.3 Interpretation of the covariance matrix

Figure 5.5 shows contours of equal probability of the bivariate Gaussian

$$p(x_1, x_2) = \frac{1}{2\pi\sqrt{\det(\mathbf{R})}} \exp\left[-\frac{1}{2} \begin{pmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{pmatrix}^t \begin{pmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{pmatrix} \begin{pmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{pmatrix} \right] \tag{5.60}$$

where

$$\mathbf{Q} = \begin{pmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{pmatrix} \tag{5.61}$$

is the inverse covariance matrix and

$$\mathbf{R} = \mathbf{Q}^{-1} = \begin{pmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{pmatrix} \tag{5.62}$$

is the covariance matrix.

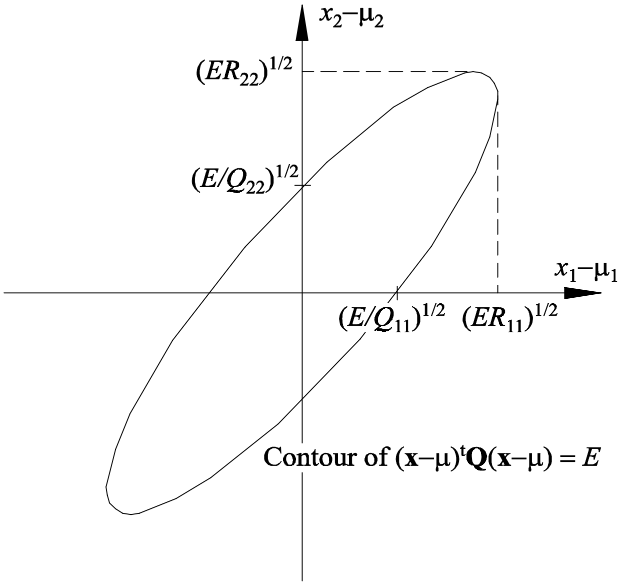


Figure 5.5 The error ellipse

Remember that this is a representation of our **state of knowlege** of the parameters x_1 and x_2. We note the following

- Corresponding to a range of parameter values such as $\mu - k\sigma$ to $\mu + k\sigma$ for a one-dimensional Gaussian, we have an **error ellipsoid** in the parameter space. These are bounded by contours of $E = (\mathbf{x} - \boldsymbol{\mu})^t\mathbf{Q}(\mathbf{x} - \boldsymbol{\mu})$ and the probability that $E \leq \chi^2$ is given by the χ^2 distribution with ν degrees of freedom where ν is the number of components in $\mathbf{x}$.

- The diagonal components of the inverse covariance matrix gives information about the intersections of the error ellipsoid with the axes. In general, these are **not** useful as estimates of the error in the parameter values.

- If we calculate the **marginal** probability density of one of the parameters, say x_k, the **variance** of x_k is given by the diagonal element R_{kk} of the covariance matrix. The standard error of x_k is $\sqrt{R_{kk}}$ and this is given by the projection of the error ellipsoid $E = 1$ on the k axis.

- The directions of the principal axes of the error ellipsoids are given by the eigenvectors of $\mathbf{R}$ (or of $\mathbf{Q}$). The lengths of the principal axes are related to the eigenvalues of $\mathbf{R}$.

You should be able to prove the above results.

5.7 Estimators and parameter estimation

As mentioned previously, in the Bayesian framework, our state of knowledge of a parameter or set of parameters $\mathbf{x}$ is given by the posterior probability density $p(\mathbf{x}|\mathbf{y})$. However, we are sometimes asked to give a single **estimate** of the quantity of interest. We have seen various ways of generating such estimates such as the **maximum à posteriori** estimate, the mean of the posterior probability, the maximum likelihood estimate and so on. Besides these estimators which are based on Bayesian ideas, we may consider any other method of generating an estimate. Abstractly, an **estimator** is a function $\hat{\mathbf{X}} : \mathbf{y} \to \hat{\mathbf{X}}(\mathbf{y})$ that converts the data vector into a number (or vector) which is our estimate of the quantity of interest.

In the Bayesian approach, we focus on the data that have been collected and try to discover the parameter values which best account for these data. In more conventional approaches to statistics, we decide on the **estimator** and then work out how well this estimator performs well in the long run. More specifically, we first suppose that the true value of the parameters $\mathbf{x}$ are given and calculate the **sampling distribution** $p(\mathbf{y}|\mathbf{x})$ of the possible data sets. Using our estimator, we calculate for each $\mathbf{y}$ the estimate $\hat{\mathbf{X}}(\mathbf{y})$. Then by the rule for transformation of variables, we can find the probability density of the estimator

$$p(\hat{\mathbf{x}}|\mathbf{x}) = \int \delta\left(\hat{\mathbf{x}} - \hat{\mathbf{X}}(\mathbf{y})\right) p(\mathbf{y}|\mathbf{x})\, d\mathbf{y}. \tag{5.63}$$

If the estimator $\hat{\mathbf{X}}$ is good, we would expect that $p(\hat{\mathbf{x}}|\mathbf{x})$ is strongly peaked around $\hat{\mathbf{x}} = \mathbf{x}$. We would like the estimator to be good for all the parameters $\mathbf{x}$ that we are likely to encounter. It is usual to quantify the distribution of $\hat{\mathbf{x}}$ about $\mathbf{x}$ in terms of the first few moments. We define

- The **bias** of the estimator for a given true parameter vector $\mathbf{x}$ by

$$\begin{aligned} \mathcal{B}_{\hat{\mathbf{X}}}(\mathbf{x}) &= \mathrm{E}[\hat{\mathbf{x}}|\mathbf{x}] - \mathbf{x} & (5.64)\\ &= \int \left(\hat{\mathbf{X}}(\mathbf{y}) - \mathbf{x}\right) p(\mathbf{y}|\mathbf{x})\, d\mathbf{y}. & (5.65)\end{aligned}$$

  This gives the difference between the mean of the estimator over $p(\mathbf{y}|\mathbf{x})$and the true value of the parameters $\mathbf{x}$. An **unbiassed estimator** is one for which the bias is zero.

- The **mean-square error** of an estimator for a given true parameter vector $\mathbf{x}$ by

$$\begin{aligned} \text{m.s.e.}_{\hat{\mathbf{X}}}(\mathbf{x}) &= \mathrm{E}\left[(\hat{\mathbf{x}} - \mathbf{x})(\hat{\mathbf{x}} - \mathbf{x})^t \middle| \mathbf{x}\right] & (5.66)\\ &= \int \left(\hat{\mathbf{X}}(\mathbf{y}) - \mathbf{x}\right)\left(\hat{\mathbf{X}}(\mathbf{y}) - \mathbf{x}\right)^t p(\mathbf{y}|\mathbf{x})\, d\mathbf{y}. & (5.67)\end{aligned}$$

  For a scalar parameter x we have

$$\begin{aligned} \text{m.s.e.}_{\hat{X}}(x) &= \mathrm{E}\left[(\hat{x} - x)^2 \middle| x\right] & (5.68)\\ &= \int \left(\hat{X}(\mathbf{y}) - x\right)^2 p(\mathbf{y}|x)\, d\mathbf{y}. & (5.69)\end{aligned}$$

- The **variance** of an estimator for a given true parameter vector $\mathbf{x}$ by

$$\begin{aligned} \text{var}_{\hat{\mathbf{X}}}(\mathbf{x}) &= \mathrm{E}\left[(\hat{\mathbf{x}} - \mathrm{E}[\hat{\mathbf{x}}|\mathbf{x}])(\hat{\mathbf{x}} - \mathrm{E}[\hat{\mathbf{x}}|\mathbf{x}])^t \middle| \mathbf{x}\right] & (5.70)\\ &= \int \left(\hat{\mathbf{X}}(\mathbf{y}) - \mathrm{E}[\hat{\mathbf{x}}|\mathbf{x}]\right)\left(\hat{\mathbf{X}}(\mathbf{y}) - \mathrm{E}[\hat{\mathbf{x}}|\mathbf{x}]\right)^t p(\mathbf{y}|\mathbf{x})\, d\mathbf{y}. & (5.71)\end{aligned}$$

  For a scalar parameter x we have

$$\operatorname{var}_{\hat{X}}(x) = \mathrm{E}\left[(\hat{x} - \mathrm{E}[\hat{x}|x])^2 \middle| \mathbf{x}\right] \tag{5.72}$$

$$= \int \left(\hat{X}(\mathbf{y}) - \mathrm{E}[\hat{x}|x]\right)^2 p(\mathbf{y}|\mathbf{x})\, d\mathbf{y}. \tag{5.73}$$

It is easy to show that (check this as an exercise) for any estimator $\hat{\mathbf{X}}$ and for any $\mathbf{x}$,

$$\text{m.s.e.}_{\hat{\mathbf{X}}}(\mathbf{x}) = \operatorname{var}_{\hat{\mathbf{X}}}(\mathbf{x}) + \mathcal{B}_{\hat{\mathbf{X}}}(\mathbf{x})\, \mathcal{B}_{\hat{\mathbf{X}}}(\mathbf{x})^t. \tag{5.74}$$

and for a scalar parameter x, this reduces to

$$\text{m.s.e.}_{\hat{X}}(x) = \operatorname{var}_{\hat{X}}(x) + \mathcal{B}_{\hat{X}}(x)^2 \tag{5.75}$$

Of course, the best estimators are those which have small bias, variance and mean-square errors.

5.7.1 Examples

1. Suppose that we have samples $\mathbf{y}$ drawn from a uniformly distributed random variable which extends from zero to x. We wish to estimate x from the data $y_1, ..., y_N$. Let us first discuss the properties of the estimator

$$\hat{X}(\mathbf{y}) = \frac{2}{N}(y_1 + y_2 + ... + y_N). \tag{5.76}$$

Since this is just a linear combination of the data, it is easy to calculate the moments of $\hat{X}$ in terms of the moments of the data. For a uniform random variable Y extending from zero to x, we know that

$$\mathrm{E}[y|x] = \frac{1}{2}x, \tag{5.77}$$

$$\mathrm{E}\left[y^2|x\right] = \frac{1}{3}x^2 \Rightarrow \operatorname{var}[y] = \frac{1}{12}x^2 \tag{5.78}$$

Hence when N of these independent random variables are added together, the mean and variances for each variable are just added together. Thus

$$\mathrm{E}\left[\hat{X}|x\right] = \frac{2}{N}\left(\frac{N}{2}x\right) = x \tag{5.79}$$

$$\operatorname{var}\left[\hat{X}|x\right] = \frac{4}{N^2}\left(\frac{N}{12}x^2\right) = \frac{x^2}{3N} \tag{5.80}$$

The estimate is **unbiassed**, and the variance and mean square error are both equal to $x^2/(3N)$.

2. Using the same data as above, let us consider instead the estimator

$$\hat{X}(\mathbf{y}) = \max(y_1, y_2, ..., y_N) \tag{5.81}$$

which happens to be the maximum likelihood estimator of x. In order to find the probability density of $\hat{X}$, we make use of the result in the previous chapter for the maximum of a set of independent identically distributed random variables. This is

$$p(\hat{x}|x) = N p_Y(\hat{x}|x) \left(\int_{-\infty}^{\hat{x}} p_Y(y|x)\, dy\right)^{N-1} \tag{5.82}$$

$$= \frac{N}{x}\left(\frac{\hat{x}}{x}\right)^{N-1} \text{ for } 0 \le \hat{x} \le x \tag{5.83}$$

The mean of this distribution is

$$\mathrm{E}\left[\hat{X}\middle|\, x\right] = \int_0^x \hat{x}\frac{N}{x}\left(\frac{\hat{x}}{x}\right)^{N-1} d\hat{x} = \frac{Nx}{N+1} \tag{5.84}$$

The variance of the distribution is

$$\mathrm{E}\left[\left(\hat{X} - \frac{Nx}{N+1}\right)^2 \middle|\, x\right] = \frac{Nx^2}{(N+2)(N+1)^2} \tag{5.85}$$

We see that the estimator is biassed and that the bias is

$$\mathcal{B}_{\hat{X}}(x) = \left(\frac{N}{N+1}\right)x - x = -\frac{x}{N+1} \tag{5.86}$$

The mean square error is

$$\begin{aligned}\text{m.s.e.}_{\hat{X}}(x) &= \frac{Nx^2}{(N+2)(N+1)^2} + \left(-\frac{x}{N+1}\right)^2 \\ &= \frac{2x^2}{(N+1)(N+2)}\end{aligned} \tag{5.87}$$

Note that as N becomes large, the mean-square error of this estimator is much smaller than that for the estimator which is twice the mean of the y_k.

Exercise: Consider the estimator $\hat{X}(\mathbf{y}) = \left(\frac{N+1}{N}\right)\max(y_1, y_2, ..., y_N)$. Show that this is unbiassed and that its variance and mean-square error are $x^2/\left[N(N+2)\right]$. The variance of this estimator is larger than the maximum likelihood estimator but its mean-square error is smaller.

5.8 Optimal Estimators

5.8.1 The minimum mean-square error estimator

As defined above, the value of the mean-square error is a function of the true value of the parameters $\mathbf{x}$. If we have a prior probability density $p(\mathbf{x})$ which describes how we believe the parameters are distributed, we can consider the problem of finding the estimator which minimizes the prior probability weighted average of the mean-square errors, i.e., we choose the estimator so as to minimize

$$E = \int p(\mathbf{x})\, \text{m.s.e.}_{\hat{\mathbf{X}}}(\mathbf{x})\, d\mathbf{x}. \tag{5.88}$$

Substituting the definition of the mean-square error, we see that

$$\begin{aligned} E &= \int\int \left\|\hat{\mathbf{X}}(\mathbf{y}) - \mathbf{x}\right\|^2 p(\mathbf{y}|\mathbf{x})\, p(\mathbf{x})\, d\mathbf{y}\, d\mathbf{x} \qquad (5.89)\\ &= \int\int \left\|\hat{\mathbf{X}}(\mathbf{y}) - \mathbf{x}\right\|^2 p(\mathbf{x},\mathbf{y})\, d\mathbf{y}\, d\mathbf{x}. \qquad (5.90)\end{aligned}$$

To minimize this, we adopt a variational approach. We consider perturbing the estimator function

$$\hat{\mathbf{X}}(\mathbf{y}) \to \hat{\mathbf{X}}(\mathbf{y}) + \varepsilon\hat{\mathbf{F}}(\mathbf{y}), \tag{5.91}$$

so that

$$E(\varepsilon) = \int\int \left\| \hat{\mathbf{X}}(\mathbf{y}) + \varepsilon \hat{\mathbf{F}}(\mathbf{y}) - \mathbf{x} \right\|^2 p(\mathbf{x}, \mathbf{y})\, d\mathbf{y}\, d\mathbf{x}. \tag{5.92}$$

For the optimal estimator

$$0 = \left[\frac{\partial E}{\partial \varepsilon} \right]_{\varepsilon=0} = 2 \int\int \hat{\mathbf{F}}(\mathbf{y})^t \left[\hat{\mathbf{X}}(\mathbf{y}) - \mathbf{x} \right] p(\mathbf{x}, \mathbf{y})\, d\mathbf{y}\, d\mathbf{x}. \tag{5.93}$$

Since this has to hold for every choice of perturbing function $\hat{\mathbf{F}}(\mathbf{y})$, we see that

$$2 \int \left[\hat{\mathbf{X}}(\mathbf{y}) - \mathbf{x} \right] p(\mathbf{x}, \mathbf{y})\, d\mathbf{x} = \mathbf{0}. \tag{5.94}$$

Thus

$$\hat{\mathbf{X}}(\mathbf{y})\, p(\mathbf{y}) = \int \mathbf{x} p(\mathbf{x}, \mathbf{y})\, d\mathbf{x}, \tag{5.95}$$

or

$$\hat{\mathbf{X}}(\mathbf{y}) = \int \mathbf{x} \frac{p(\mathbf{x}, \mathbf{y})}{p(\mathbf{y})}\, d\mathbf{x} = \int \mathbf{x} p(\mathbf{x}|\mathbf{y})\, d\mathbf{x}. \tag{5.96}$$

The optimal estimator in this sense is just the mean over the posterior probability density.

5.8.2 The Cramér-Rao lower bound

The Cramér-Rao lower bound is a relation which gives the minimum variance that an estimator can have for a given bias. It does not give a construction for such minimum-variance estimators, but is useful for evaluating how near an estimator is to the ideal. The bound is expressed completely in terms of the forward probability for the data given the parameter $p(\mathbf{y}|x)$ and makes no reference to the ideas of prior probability. For simplicity, let us consider the case of a single **scalar** parameter x.

The result is based on the Cauchy-Schwarz inequality which may be stated in the form that if $\mathbf{y}$ is a vector-valued random variable and if $F(\mathbf{y})$ and $G(\mathbf{y})$ are scalar-valued functions of $\mathbf{y}$, then

$$\left| \mathrm{E}\left[F(\mathbf{y})\, G(\mathbf{y}) \right] \right|^2 \leq \mathrm{E}\left[F(\mathbf{y})^2 \right] \mathrm{E}\left[\mathbf{G}(\mathbf{y})^2 \right] \tag{5.97}$$

We suppose that we have some estimator $\hat{X}(\mathbf{y})$whose variance we wish to bound. Let us consider the following choice of F and G and suppose that the expectation values are being taken over the probability density $p(\mathbf{y}|x)$ for some fixed x.

$$F(\mathbf{y}) = \hat{X}(\mathbf{y}) - \mathrm{E}\left[\hat{X}|x \right] \tag{5.98}$$

$$G(\mathbf{y}) = \frac{\partial}{\partial x} \log p(\mathbf{y}|x) \tag{5.99}$$

Then

$$F(\mathbf{y})\, G(\mathbf{y}) = \left(\hat{X}(\mathbf{y}) - \mathrm{E}\left[\hat{X}|x \right] \right) \frac{\partial}{\partial x} \log p(\mathbf{y}|x) \tag{5.100}$$

and

$$\begin{aligned}
\mathrm{E}\left[F\left(\mathbf{y}\right)G\left(\mathbf{y}\right)\right] &= \int\left\{\left(\hat{X}\left(\mathbf{y}\right)-\mathrm{E}\left[\hat{X}|x\right]\right)\frac{\partial}{\partial x}\log p\left(\mathbf{y}|\mathbf{x}\right)\right\}p\left(\mathbf{y}|\mathbf{x}\right)\,d\mathbf{y} \\
&= \int\left(\hat{X}\left(\mathbf{y}\right)-\mathrm{E}\left[\hat{X}|x\right]\right)\frac{\partial}{\partial x}p\left(\mathbf{y}|\mathbf{x}\right)\,d\mathbf{y}\text{ since }\frac{\partial}{\partial x}\log p\left(\mathbf{y}|x\right)=\frac{\frac{\partial}{\partial x}p\left(\mathbf{y}|x\right)}{p\left(\mathbf{y}|x\right)} \\
&= \int\hat{X}\left(\mathbf{y}\right)\frac{\partial}{\partial x}p\left(\mathbf{y}|x\right)\,d\mathbf{y}\text{ since }\frac{\partial}{\partial x}\int p\left(\mathbf{y}|x\right)\,d\mathbf{y}=\mathbf{0} \\
&= \frac{\partial}{\partial x}\left(\int\hat{X}\left(\mathbf{y}\right)p\left(\mathbf{y}|x\right)\,d\mathbf{y}\right)=\frac{\partial\mathrm{E}\left[\hat{X}|x\right]}{\partial x} \qquad (5.101)
\end{aligned}$$

Further, we see that

$$\mathrm{E}\left[F\left(\mathbf{y}\right)^2\right]=\mathrm{E}\left[\left(\hat{X}\left(\mathbf{y}\right)-\mathrm{E}\left[\hat{X}|x\right]\right)^2\middle|\,\mathbf{x}\right]=\mathrm{var}_{\hat{X}}\left(x\right) \qquad (5.102)$$

$$\mathrm{E}\left[G\left(\mathbf{y}\right)^2\right]=\mathrm{E}\left[\left(\frac{\partial}{\partial x}\log p\left(\mathbf{y}|x\right)\right)^2\middle|\,x\right] \qquad (5.103)$$

and so substituting into the Cauchy-Schwarz inequality we have

$$\left(\frac{\partial\mathrm{E}\left[\hat{X}|x\right]}{\partial x}\right)^2\le\mathrm{var}_{\hat{X}}\left(x\right)\mathrm{E}\left[\left(\frac{\partial}{\partial x}\log p\left(\mathbf{y}|x\right)\right)^2\middle|\,x\right] \qquad (5.104)$$

By the definition of the bias,

$$\mathcal{B}_{\hat{X}}\left(x\right)=\mathrm{E}\left[\hat{X}|x\right]-x \qquad (5.105)$$

$$\mathrm{var}_{\hat{X}}\left(x\right)\ge\frac{\left(1+\mathcal{B}'_{\hat{X}}\left(x\right)\right)^2}{\mathrm{E}\left[\left(\frac{\partial}{\partial x}\log p\left(\mathbf{y}|x\right)\right)^2|x\right]} \qquad (5.106)$$

This is the statement of the **Cramér-Rao** lower bound (CRLB). Notice that it gives a minimum possible value for the variance of any estimator for a given value of x. For the special case of an **unbiassed** estimator, we have that

$$\mathrm{var}_{\hat{X}}\left(x\right)\ge\frac{1}{\mathrm{E}\left[\left(\frac{\partial}{\partial x}\log p\left(\mathbf{y}|x\right)\right)^2|x\right]} \qquad (5.107)$$

An alternative form of the denominator is often convenient. Consider

$$\frac{\partial^2}{\partial x^2}\log p\left(\mathbf{y}|x\right)=\frac{\partial}{\partial x}\left(\frac{1}{p}\frac{\partial p}{\partial x}\right)=\frac{1}{p}\frac{\partial^2 p}{\partial x^2}-\frac{1}{p^2}\left(\frac{\partial p}{\partial x}\right)^2=\frac{1}{p}\frac{\partial^2 p}{\partial x^2}-\left(\frac{\partial}{\partial x}\log p\right)^2 \qquad (5.108)$$

So taking the expectation over $p\left(\mathbf{y}|x\right)$ we get

$$\mathrm{E}\left[\frac{\partial^2}{\partial x^2}\log p\left(\mathbf{y}|x\right)\middle|\,x\right]=-\mathrm{E}\left[\left(\frac{\partial}{\partial x}\log p\left(\mathbf{y}|x\right)\right)^2\middle|\,x\right] \qquad (5.109)$$

since

$$\mathrm{E}\left[\frac{1}{p}\frac{\partial^2 p\left(\mathbf{y}|x\right)}{\partial x^2}\middle|\,x\right]=\int\frac{\partial^2 p\left(\mathbf{y}|x\right)}{\partial x^2}d\mathbf{y}=\frac{\partial^2}{\partial x^2}\int p\left(\mathbf{y}|x\right)d\mathbf{y}=0. \qquad (5.110)$$

Thus, the CRLB may also be written as

$$\operatorname{var}_{\hat{X}}(x) \geq \frac{\left(1+\mathcal{B}'_{\hat{X}}(x)\right)^2}{\mathrm{E}\left[-\frac{\partial^2}{\partial x^2}\log p(\mathbf{y}|x)\Big| x\right]} \tag{5.111}$$

This has an appealing interpretation since it states that the bound is related to the expectation value of the **curvature** of the **log likelihood** function. The term in the denominator is large when the likelihood function is sharply peaked. For such likelihood functions, it is possible for estimators to achieve a lower variance than for situations in which the likelihood function is broad.

Note that in order for us to be able to use the Cramér-Rao lower bound, the function $\log p(\mathbf{y}|x)$ must be differentiable with respect to x and not have any singularities in the derivative.

5.8.3 Examples

1. Let us consider estimating the variance v from a set of N Gaussian random variables with

$$p(y_1,...,y_N|v) = \frac{1}{(2\pi v)^{N/2}}\exp\left(-\frac{1}{2v}\sum_{k=1}^N (y_k-\mu)^2\right). \tag{5.112}$$

From this we see that

$$\begin{aligned}\frac{\partial^2}{\partial v^2}\log p(\mathbf{y}|v) &= \frac{\partial^2}{\partial v^2}\log\left(\frac{1}{(2\pi v)^{N/2}}\exp\left(-\frac{1}{2v}\sum_{k=1}^N (y_k-\mu)^2\right)\right)\\ &= \frac{1}{2v^3}\left(Nv-2\sum_{k=1}^N (y_k-\mu)^2\right)\end{aligned} \tag{5.113}$$

Taking the expectation value over $p(\mathbf{y}|v)$ yields

$$\begin{aligned}\mathrm{E}\left[\frac{\partial^2}{\partial v^2}\log p(\mathbf{y}|v)\Big| v\right] &= \frac{N}{2v^2}-\frac{1}{v^3}\sum_{k=1}^N \mathrm{E}\left[(y_k-\mu)^2\right]\\ &= \frac{N}{2v^2}-\frac{N}{v^2} = -\frac{N}{2v^2}\end{aligned} \tag{5.114}$$

So the CRLB for the estimator is

$$\operatorname{var}_{\hat{X}}(x) \geq \frac{2v^2}{N}\left|1+\mathcal{B}'_{\hat{X}}(x)\right|^2$$

For the sample variance estimator discussed above,

$$\mathcal{B}_{\hat{V}}(v) = -\frac{v}{N} \tag{5.115}$$

and so by the CRLB

$$\operatorname{var}_{\mathrm{CRLB}} \geq \frac{2v^2}{N}\left(1-\frac{1}{N}\right)^2 = \frac{2(N-1)^2 v^2}{N^3}$$

The actual variance of the estimator was found above to be

$$\operatorname{var}_{\hat{V}}(v) = \frac{2(N-1)v^2}{N^2} \tag{5.116}$$

The ratio of the CRLB to the actual variance is called the **efficiency** of the estimate. In this case

$$\frac{\text{var}_{\text{CRLB}}}{\text{var}_{\hat{V}}(v)} = \frac{N-1}{N} \tag{5.117}$$

As N becomes large, this efficiency approaches unity, and the estimator is said to be **asymptotically efficient.**

2. One **cannot** apply the CRLB to the estimators associated with finding the width of a uniform distribution since the log likelihood function is $-\infty$ in certain regions, and there are discontinuities at which it fails to be differentiable.

Note that it is possible to evaluate the variance of an estimator numerically by simulation and to compare the result with that given by the Cramér-Rao lower bound.

5.9 Data Modelling

We now wish to address some of the practical issues involved in *data modelling*, which may be regarded as a way of summarizing data $\mathbf{y}$ by fitting it to a "model" which depends on a set of adjustable parameters $\mathbf{x}$. This model may result from some underlying theory that the data are supposed to arise from, or it may simply be a member of a convenient class of functions (such as a polynomial, or sum of sinusoids of unknown amplitude, frequency and phase). We have seen that the Bayesian approach is to calculate the posterior probability density for $\mathbf{x}$ by using the familiar rule

$$p(\mathbf{x}|\mathbf{y}) \propto p(\mathbf{y}|\mathbf{x})\, p(\mathbf{x}) \tag{5.118}$$

we then estimate $\mathbf{x}$ by choosing some measure of the "centre" of the posterior probability function. Although this is straightforward in principle, it is often difficult to display the posterior probability density function or to calculate its statistics, because the number of parameters $\mathbf{x}$ is often rather large, and the topology of the posterior probability function may be complicated and have many local maxima. A variety of approximate methods are often employed, some of which we shall consider.

By taking the logarithm of the above equation, we may write

$$\log p(\mathbf{x}|\mathbf{y}) = \text{const } -\frac{1}{2}\mathcal{E}(\mathbf{x};\mathbf{y}) + \frac{1}{2}\mathcal{S}(\mathbf{x}) \tag{5.119}$$

where $\mathcal{S}(\mathbf{x}) = 2\log p(\mathbf{x})$ and $\mathcal{E}(\mathbf{x};\mathbf{y}) = -2\log p(\mathbf{y}|\mathbf{x})$. The quantity

$$\mathcal{E}(\mathbf{x};\mathbf{y}) - \mathcal{S}(\mathbf{x}) \tag{5.120}$$

may be regarded as a **figure-of-merit function** (or merit function, for short) which is small when the posterior probability is large. This merit function has two terms, the first $\mathcal{E}(\mathbf{x};\mathbf{y})$ depends both on the data and the parameters, and may be interpreted naturally as a measure of the **misfit** between the actual data and the predictions of the model. The second term $\mathcal{S}(\mathbf{x})$ which depends on the prior may be interpreted as a **preference** function, which is large when the parameters conform to our preconceptions. It should be clear that finding $\mathbf{x}$ which minimizes $\mathcal{E}(\mathbf{x};\mathbf{y})$ alone corresponds to the maximum-likelihood estimate while minimizing $\mathcal{E}(\mathbf{x};\mathbf{y}) - \mathcal{S}(\mathbf{x})$ corresponds to the maximum *à posteriori* estimate.

Besides estimating the **values** of the parameters, there are two additional important issues. One is to assess whether or not the model is appropriate for explaining the data — this involves testing the **goodness of fit** against some statistical standard, and the other is to obtain an indication of the **uncertainties** in the estimated parameter values.

5.10 Least-Squares for Parameter Estimation

Let us suppose that the noise process is **additive** and **Gaussian** distributed so that the actual data may be written as

$$\mathbf{y} = \hat{\mathbf{y}}(\mathbf{x}) + \mathbf{n}$$

where $\hat{\mathbf{y}}(\mathbf{x})$ is the **mock data** which would have been generated in the absence of noise if the true parameter vector was $\mathbf{x}$ and $\mathbf{n}$ represents the noise. The likelihood function is

$$p(\mathbf{y}|\mathbf{x}) = p(\mathbf{n} = \mathbf{y} - \hat{\mathbf{y}}(\mathbf{x})) = \frac{1}{(2\pi)^{N/2}\sqrt{\det\Gamma}} \exp\left[-\frac{1}{2}(\mathbf{y} - \hat{\mathbf{y}}(\mathbf{x}))^t \Gamma^{-1} (\mathbf{y} - \hat{\mathbf{y}}(\mathbf{x}))\right], \qquad (5.121)$$

where the noise is assumed to have zero mean and covariance matrix Γ. Ignoring a constant, the misfit function is given by

$$\mathcal{E}(\mathbf{x};\mathbf{y}) = (\mathbf{y} - \hat{\mathbf{y}}(\mathbf{x}))^t \Gamma^{-1} (\mathbf{y} - \hat{\mathbf{y}}(\mathbf{x})). \qquad (5.122)$$

If the noise samples are independent, the matrix Γ is diagonal with the diagonal elements being given by the variances. If further all the variances are equal to σ^2, then the likelihood function has the particularly simple form

$$p(\mathbf{y}|\mathbf{x}) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left[-\frac{1}{2\sigma^2}(\mathbf{y} - \hat{\mathbf{y}}(\mathbf{x}))^t (\mathbf{y} - \hat{\mathbf{y}}(\mathbf{x}))\right] \qquad (5.123)$$

and the misfit is

$$\mathcal{E}(\mathbf{x};\mathbf{y}) = \frac{(\mathbf{y} - \hat{\mathbf{y}}(\mathbf{x}))^t (\mathbf{y} - \hat{\mathbf{y}}(\mathbf{x}))}{\sigma^2} = \sum_{k=1}^{N} \frac{1}{\sigma^2}(y_k - \hat{y}_k(\mathbf{x}))^2 \qquad (5.124)$$

which is simply a sum of squares. We shall investigate the process of minimizing this misfit, or equivalently, maximizing the likelihood function. Thus we see that

$$\boxed{\text{Least squares} \equiv \text{maximum likelihood with independent Gaussian noise}}$$

In order to illustrate this process, we consider some specific examples.

5.10.1 Estimating amplitude of a known signal in additive Gaussian noise

Let us suppose that the data consist of N samples from the signal

$$y(t) = As(t) \tag{5.125}$$

taken at times $t_1, t_2, ..., t_N$ which need not necessarily be evenly spaced. We shall assume that $s(t)$ is known but that the amplitude A is to be determined. The components of the data vector $\mathbf{y} \in \mathbb{R}^N$ are given by

$$y_k = As(t_k) + n_k, \tag{5.126}$$

where n_k are samples of the noise. The mock data is

$$\hat{y}_k(A) = As(t_k) \equiv As_k \tag{5.127}$$

If we suppose that the noise samples are independent and each have variance σ^2, the misfit function for a given data vector $\mathbf{y}$ is

$$\mathcal{E}(\mathbf{y}|A) = \frac{(\mathbf{y} - \hat{\mathbf{y}}(A))^t (\mathbf{y} - \hat{\mathbf{y}}(A))}{\sigma^2} = \sum_{k=1}^{N} \frac{1}{\sigma^2} (y_k - As_k)^2 \tag{5.128}$$

$$= \frac{1}{\sigma^2} \left[\left(\sum_{k=1}^{N} s_k^2 \right) \left(A - \frac{\sum_{k=1}^{N} y_k s_k}{\sum_{k=1}^{N} s_k^2} \right)^2 + \left(\sum_{k=1}^{N} y_k^2 - \frac{\left(\sum_{k=1}^{N} y_k s_k \right)^2}{\sum_{k=1}^{N} s_k^2} \right) \right] \tag{5.129}$$

where we have completed the square in order to show more clearly the dependence on A. The maximum-likelihood estimate is given by maximizing the exponent. This leads to the estimate

$$\hat{A} = \frac{\sum_{k=1}^{N} y_k s_k}{\sum_{k=1}^{N} s_k^2} \tag{5.130}$$

We see that in order to obtain the estimate of A, the only function of the data that needs to be calculated is

$$\sum_{k=1}^{N} y_k s_k \tag{5.131}$$

This may be interpreted as multiplying the measured data by a copy of the known signal and summing the result (or integrating, in the continuous case). This is the basis of correlation detectors or lock-in amplifiers.

5.10.2 Estimating parameters of two sinusoids in noise

Let us suppose that the data consist of N samples from the signal

$$y(t) = A_1 \cos \omega_1 t + B_1 \sin \omega_1 t + A_2 \cos \omega_2 t + B_2 \sin \omega_2 t, \tag{5.132}$$

taken at times $t_1, t_2, ..., t_N$ which need not be evenly spaced. The quantities A_1, A_2, B_1, B_2, ω_1 and ω_2 are regarded as the unknown parameters which we wish to estimate. We shall refer to these parameters collectively by the vector $\mathbf{x} \in \mathbb{R}^M$ and the components of the data vector $\mathbf{y} \in \mathbb{R}^N$ are given by

$$y_k = A_1 \cos \omega_1 t_k + B_1 \sin \omega_1 t_k + A_2 \cos \omega_2 t_k + B_2 \sin \omega_2 t_k + n_k, \tag{5.133}$$

where n_k are samples of the noise. The mock data is

$$\hat{y}_k\left(A_1, A_2, B_1, B_2, \omega_1, \omega_2\right) = A_1 \cos \omega_1 t_k + B_1 \sin \omega_1 t_k + A_2 \cos \omega_2 t_k + B_2 \sin \omega_2 t_k. \tag{5.134}$$

If we suppose that the noise samples are independent and each have variance σ^2, the misfit function for a given data vector $\mathbf{y}$ is

$$\mathcal{E}\left(\mathbf{y}|\mathbf{x}\right) = \frac{\left(\mathbf{y} - \hat{\mathbf{y}}\left(\mathbf{x}\right)\right)^t \left(\mathbf{y} - \hat{\mathbf{y}}\left(\mathbf{x}\right)\right)}{\sigma^2} = \sum_{k=1}^{N} \frac{1}{\sigma^2} \left(y_k - \hat{y}_k\left(A_1, A_2, B_1, B_2, \omega_1, \omega_2\right)\right)^2. \tag{5.135}$$

Minimizing this as a function of the $M = 6$ parameters may be regarded as an exercise in multidimensional non-linear optimization. A variety of iterative methods are available, which generate a sequence of iterates $\mathbf{x}_1, \mathbf{x}_2, ...$ which converge to $\arg\min_{\mathbf{x}} \mathcal{E}\left(\mathbf{x}; \mathbf{y}\right)$. Some of these are

1. Non-linear simplex methods: These work on the principle of evaluating the merit function on a set of $M+1$ points called an M simplex. An M simplex is the simplest entity which encloses a "volume" in M dimensions (e.g., a 2–simplex is a triangle and a 3–simplex is a tetrahedron), and the idea is to try to enclose the position of the minimum of the merit function within the volume of the simplex. By applying a series of transformations based on the function values at the vertices, the simplex moves downhill and (hopefully) shrinks until it is small enough to specify the minimum position to the desired accuracy. The main advantages of the simplex method are its simplicity, in that it only requires function evaluations, and its robustness to non-smooth merit functions. The main disadvantage is its often prohibitively slow speed when M is moderate or large. The Matlab routine `fmins` uses the simplex algorithm.

2. Gradient-based methods: Besides using function evaluations, these methods also require the user to supply the **gradient** or first derivative of the merit function so that the algorithm knows how to proceed downhill. The naïve **steepest descents** algorithm simply computes the direction of steepest descents at the current $\mathbf{x}_n$ and proceeds as far along this direction, i.e., along the line

   $$\mathbf{x}_{n+1} = \mathbf{x}_n - \beta_n \boldsymbol{\nabla}\mathcal{E}\left(\mathbf{x}_n; \mathbf{y}\right) \tag{5.136}$$

   until β_n is such that a minimum is achieved, i.e.,

   $$\beta_n = \arg\min_{\beta>0} \mathcal{E}\left(\mathbf{x}_n - \beta\boldsymbol{\nabla}\mathcal{E}\left(\mathbf{x}_n; \mathbf{y}\right)\right). \tag{5.137}$$

   It then repeats this procedure starting at $\mathbf{x}_{n+1}$ to find $\mathbf{x}_{n+2}$ and so on. Although the merit function keeps decreasing on each iterate, this algorithm is **extremely inefficient** when the contours of the merit function are long ellipses.

3. Conjugate gradient algorithm: The steepest descents method is an example of an iterative method which is based on a sequence of line searches along vectors $\mathbf{p}_n$ called "search directions", i.e.,

   $$\mathbf{x}_{n+1} = \mathbf{x}_n - \beta_n \mathbf{p}_n \tag{5.138}$$

   Ideally, we would like the search directions to be such that the sequence of minimizations starting from $\mathbf{x}_0$ along the directions $\mathbf{p}_0, ..., \mathbf{p}_{K-1}$ should give the **same result** as a multidimensional minimization over the space

   $$S_K = \left\{\mathbf{x}_0 + b_0\mathbf{p}_0 + ... + b_{K-1}\mathbf{p}_{K-1} : b_0, ..., b_{K-1} \in \mathbb{R}\right\} \tag{5.139}$$

Unfortunately this is not the case unless the search directions $\mathbf{p}_n$ form a **mutually conjugate set.** In the conjugate gradient algorithm, the new search direction on the iteration $n+1$ is not simply $-\boldsymbol{\nabla}\mathcal{E}\left(\mathbf{x}_{n+1}\right)$. Instead, this negative gradient is combined with a multiple of the **previous** search direction so that the new search direction is conjugate to the last

$$\mathbf{p}_{n+1} = -\boldsymbol{\nabla}\mathcal{E}\left(\mathbf{x}_{n+1};\mathbf{y}\right) + \gamma_n \mathbf{p}_n \tag{5.140}$$

where γ_n is chosen to give conjugacy. It turns out that if the merit function happens to be exactly quadratic, this procedure ensures that **all** the $\mathbf{p}_n$ are mutually conjugate and so the algorithm reaches the minimum in no more than M iterations. In practise, inexact arithmetic and the non-quadratic nature of the merit function mean that this is not always achieved. The advantage of the conjugate gradient algorithm is that it can be implemented without using very much memory even for large problems. Some minor disadvantages are the need to carry out line searches and to calculate the derivatives.

4. Newton based methods: If we expand the function we wish to minimize in a Taylor series about the current iterate $\mathbf{x}_n$, we find

$$\mathcal{E}\left(\mathbf{x};\mathbf{y}\right) \approx \mathcal{E}\left(\mathbf{x}_n;\mathbf{y}\right) + \boldsymbol{\nabla}\mathcal{E}\left(\mathbf{x}_n;\mathbf{y}\right)^t\left(\mathbf{x}-\mathbf{x}_n\right) + \frac{1}{2}\left(\mathbf{x}-\mathbf{x}_n\right)^t \boldsymbol{\nabla}\boldsymbol{\nabla}\mathcal{E}\left(\mathbf{x}_n;\mathbf{y}\right)\left(\mathbf{x}-\mathbf{x}_n\right) + ... \tag{5.141}$$

where $\boldsymbol{\nabla}\boldsymbol{\nabla}\mathcal{E}\left(\mathbf{x}_n;\mathbf{y}\right)$ denotes the Hessian matrix of second derivatives taken with respect to $\mathbf{x}$. The minimum of the quadratic form found by truncating the Taylor series at the quadratic term is where

$$\mathbf{0} = \boldsymbol{\nabla}\mathcal{E}\left(\mathbf{x};\mathbf{y}\right) \approx \boldsymbol{\nabla}\mathcal{E}\left(\mathbf{x}_n;\mathbf{y}\right) + \boldsymbol{\nabla}\boldsymbol{\nabla}\mathcal{E}\left(\mathbf{x}_n;\mathbf{y}\right)\left(\mathbf{x}-\mathbf{x}_n\right), \tag{5.142}$$

This has the solution

$$\mathbf{x} = \mathbf{x}_n - \left[\boldsymbol{\nabla}\boldsymbol{\nabla}\mathcal{E}\left(\mathbf{x}_n;\mathbf{y}\right)\right]^{-1}\boldsymbol{\nabla}\mathcal{E}\left(\mathbf{x}_n;\mathbf{y}\right) \tag{5.143}$$

Since this is only an approximation to the minimum, given that merit functions are in general non-quadratic, this is used as the next iterate $\mathbf{x}_{n+1}$. Newton methods converge quadratically in a neighbourhood of a minimum, but can give bad results far from a minimum where $\boldsymbol{\nabla}\boldsymbol{\nabla}\mathcal{E}\left(\mathbf{x}_n;\mathbf{y}\right)$ may be nearly singular. It is preferable to use a method which has the features of the steepest descents algorithm (which is guaranteed to reduce the merit function on each iterate) while still far from the minimum, but which switches to a Newton based method near the minimum. One such algorithm is called the **Levenberg-Marquardt** method which sets

$$\mathbf{x}_{n+1}= \mathbf{x}_n - \left[\lambda\mathbf{D}\left(\mathbf{x}_n;\mathbf{y}\right) + \boldsymbol{\nabla}\boldsymbol{\nabla}\mathcal{E}\left(\mathbf{x}_n;\mathbf{y}\right)\right]^{-1}\boldsymbol{\nabla}\mathcal{E}\left(\mathbf{x}_n;\mathbf{y}\right)$$

where the quantity λ is chosen to be small once we are near the minimum. $\mathbf{D}$ is a diagonal matrix, usually chosen to be the diagonal part of $\boldsymbol{\nabla}\boldsymbol{\nabla}\mathcal{E}\left(\mathbf{x}_n;\mathbf{y}\right)$. This ensures that for large λ, the algorithm takes a small step in a direction which decreases $\mathcal{E}$. Note that Newton based methods require the storage and inversion of an $M \times M$ matrix on each iteration, which may be prohibitive if M is large.

Computing the gradient and Hessian matrix of $\mathcal{E}\left(\mathbf{x};\mathbf{y}\right)$ is straightforward for the least-squares problem. We see that if

$$\mathcal{E}\left(\mathbf{x};\mathbf{y}\right) = \sum_{k=1}^{N}\frac{1}{\sigma^2}\left(y_k - \hat{y}_k\left(\mathbf{x}\right)\right)^2, \tag{5.144}$$

then

$$\frac{\partial \mathcal{E}}{\partial x_r} = -2\sum_{k=1}^{N} \frac{[y_k - \hat{y}_k(\mathbf{x})]}{\sigma^2} \frac{\partial \hat{y}_k(\mathbf{x})}{\partial x_r} \tag{5.145}$$

and

$$\frac{\partial^2 \mathcal{E}}{\partial x_r \partial x_s} = 2\sum_{k=1}^{N} \frac{1}{\sigma^2} \left[\frac{\partial \hat{y}_k(\mathbf{x})}{\partial x_r} \frac{\partial \hat{y}_k(\mathbf{x})}{\partial x_s} - [y_k - \hat{y}_k(\mathbf{x})] \frac{\partial^2 \hat{y}_k(\mathbf{x})}{\partial x_r \partial x_s} \right] \tag{5.146}$$

In practise, the second term in the sum is small compared to the first, either because the model is weakly non-linear or because the $y_k - \hat{y}_k(\mathbf{x})$ is essentially noise and so they tend to be as often positive as negative, leading to cancellation when the sum is taken over k. Thus it is usual to use

$$\frac{\partial^2 \mathcal{E}}{\partial x_r \partial x_s} \approx 2\sum_{k=1}^{N} \frac{1}{\sigma^2} \left[\frac{\partial \hat{y}_k(\mathbf{x})}{\partial x_r} \frac{\partial \hat{y}_k(\mathbf{x})}{\partial x_s} \right] \tag{5.147}$$

for Newton based algorithms such as the Levenberg-Marquardt method. We see that both first and second derivatives of $\mathcal{E}$ may be formed from the Jacobian matrix

$$(\mathbf{J_x}\hat{\mathbf{y}})_{ij} = \frac{\partial \hat{y}_i(\mathbf{x})}{\partial x_j}$$

5.10.2.1 Linear and Non-linear parameters

In all of the methods for minimizing the merit function, the time needed for finding the minimum increases as the number of parameters is increased. It is therefore sometimes advantageous to try to reduce the size of the numerical minimization problem by doing a part of the minimization analytically. If we return to the problem of the two sinusoids in noise, we see that the parameters may be divided into two classes. In the first class we have A_1, A_2, B_1 and B_2 on which $\hat{\mathbf{y}}$ depends linearly and in the second class we have ω_1 and ω_2 on which $\hat{\mathbf{y}}$ depends non-linearly. It turns out to be possible to carry out the minimization over the linear parameters **analytically,** thus reducing the size of the search space for our optimization routines to the number of non-linear parameters alone.

We may write the dependence of $\hat{\mathbf{y}}$ on the parameters as in Eq (5.134) in the form

$$\hat{y}_k = \begin{pmatrix} \cos\omega_1 t_k & \sin\omega_1 t_k & \cos\omega_2 t_k & \sin\omega_2 t_k \end{pmatrix} \begin{pmatrix} A_1 \\ A_2 \\ B_1 \\ B_2 \end{pmatrix} \tag{5.148}$$

or

$$\begin{pmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_N \end{pmatrix} = \begin{pmatrix} \cos\omega_1 t_1 & \sin\omega_1 t_1 & \cos\omega_2 t_1 & \sin\omega_2 t_1 \\ \cos\omega_1 t_2 & \sin\omega_1 t_2 & \cos\omega_2 t_2 & \sin\omega_2 t_2 \\ \vdots & \vdots & \vdots & \vdots \\ \cos\omega_1 t_N & \sin\omega_1 t_N & \cos\omega_2 t_N & \sin\omega_2 t_N \end{pmatrix} \begin{pmatrix} A_1 \\ A_2 \\ B_1 \\ B_2 \end{pmatrix} \tag{5.149}$$

$$\hat{\mathbf{y}} = \mathbf{C}(\mathbf{x}_{\text{nonlin}})\,\mathbf{x}_{\text{lin}} \tag{5.150}$$

where $\mathbf{x}_{\text{lin}}$ represents the linear parameters $\{A_1, A_2, B_1, B_2\}$ and $\mathbf{x}_{\text{nonlin}}$ represents the nonlinear parameters $\{\omega_1, \omega_2\}$. The misfit function is

$$\begin{aligned} \mathcal{E}\left(\mathbf{x};\mathbf{y}\right) = \mathcal{E}\left(\mathbf{x}_{\text{lin}},\mathbf{x}_{\text{nonlin}};\mathbf{y}\right) &= \frac{1}{\sigma^2}\left(\mathbf{y}-\hat{\mathbf{y}}\left(\mathbf{x}\right)\right)^t\left(\mathbf{y}-\hat{\mathbf{y}}\left(\mathbf{x}\right)\right) \\ &= \frac{1}{\sigma^2}\left(\mathbf{y}-\mathbf{C}\left(\mathbf{x}_{\text{nonlin}}\right)\mathbf{x}_{\text{lin}}\right)^t\left(\mathbf{y}-\mathbf{C}\left(\mathbf{x}_{\text{nonlin}}\right)\mathbf{x}_{\text{lin}}\right) \end{aligned} \tag{5.151}$$

The derivatives with respect to the linear parameters may be computed and set equal to zero. This leads to the following simultaneous equations for $\mathbf{x}_{\text{lin}}$

$$\mathbf{C}^t\mathbf{C}\mathbf{x}_{\text{lin}} = \mathbf{C}^t\mathbf{y} \tag{5.152}$$

or

$$\mathbf{x}_{\text{lin}} = \left(\mathbf{C}^t\mathbf{C}\right)^{-1}\mathbf{C}^t\mathbf{y} \tag{5.153}$$

where $\mathbf{C}$ is evaluated at the value of the non-linear parameters. Having found the linear parameters for a particular choice of the non-linear parameters, we can write the misfit function as

$$\mathcal{E}\left(\mathbf{x}_{\text{lin}}\left(\mathbf{x}_{\text{nonlin}}\right),\mathbf{x}_{\text{nonlin}};\mathbf{y}\right) = \frac{\mathbf{y}^t\mathbf{y}-\mathbf{y}^t\mathbf{C}\left(\mathbf{x}_{\text{nonlin}}\right)\left[\mathbf{C}\left(\mathbf{x}_{\text{nonlin}}\right)^t\mathbf{C}\left(\mathbf{x}_{\text{nonlin}}\right)\right]^{-1}\mathbf{C}\left(\mathbf{x}_{\text{nonlin}}\right)^t\mathbf{y}}{\sigma^2} \tag{5.154}$$

this is only a function of the non-linear parameters which often makes it more convenient for optimization. We use `fmins` or some other convenient method of finding $\mathbf{x}_{\text{nonlin}}$ and then calculate the linear parameters from this.

The Matlab code below shows how this process is applied to the problem of estimating the angular frequencies of the two sinusoids and then finding the amplitudes.

```
% List of times at which data are measured
tlist = linspace(0,6,41)';
% Synthesize some data with linear parameters xlin = [A_1;B_1;A_2;B_2]
%  and non-linear parameters w1 and w2
xlin = [1;1;2;0]; w1 = 1.5; w2 = 2.5;
C = makeC([w1;w2],tlist); yhat = C * xlin;
%  and add noise
sigma = 0.5; y = yhat + sigma*randn(size(yhat));
% Use fmins to find the optimal parameters
xnlbest = fmins('misfit',[1;2],1,[],tlist,y);
C = makeC(xnlbest,tlist); xlinbest = (C'*C)\(C'*y);
Ebest = misfit(xnlbest,tlist,y)
figure(1); plot(tlist,y,'x',tlist,C*xlinbest);
xlabel('Time'); ylabel('Samples and best fit');
% Grid of points for calculating misfit
nl = 30;
wlist = linspace(0.1,3.0,nl);
[w1,w2] = meshgrid(wlist,wlist+1e-3);
E = zeros(nl,nl);
for k = 1:nl
   w1t = w1(1,k);
   for l = 1:nl
      w2t = w2(l,1);
      E(1,k) = misfit([w1t;w2t],tlist,y);
   end
end
figure(2); contour(w1,w2,E,min(min(E))+[1:2:40]);
hold on
plot(xnlbest(1),xnlbest(2),'+',xnlbest(2),xnlbest(1),'+');
hold off
xlabel('w1'); ylabel('w2');
save fitdata tlist y xnlbest xlinbest Ebest sigma
```

This code requires the two supporting functions listed below

```
function E = misfit(xnl,tlist,y)
%
C = makec(xnl,tlist);
xlin = (C'*C)\(C'*y);
E = sum(abs(y-C*xlin).^2);

function C = makec(xnl,tlist)
%
w1 = xnl(1); w2 = xnl(2);
C = [cos(w1*tlist) sin(w1*tlist) cos(w2*tlist) sin(w2*tlist)];
```

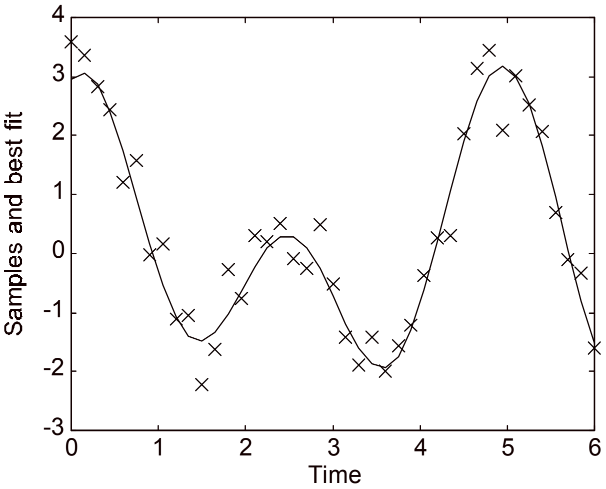


Figure 5.6 Samples of the sum of two sinusoids and best fitting model

In Figure 5.6, the noisy samples are drawn together with the best fitting model. Note that the standard deviation of the noise at each point is taken to be $\sigma = 0.5$. The true values of the parameters and those of the best fitting model are

	ω_1	ω_2	A_1	B_1	A_2	B_2
True value	1.5	2.5	1	1	2	0
Estimate	1.40	2.57	1.25	0.73	1.70	0.30

In Figure 5.7, the merit surface as a function of the non-linear parameters ω_1 and ω_2 is shown to give an idea of the topology of the function for which the minimization is carried out. We see that even in this simple example, the surface is quite complicated and there is the danger of missing the minimum if we start off with an inaccurate initial estimate. At the two (equal) minima, we find that $\mathcal{E}_{\min} = 7.4$

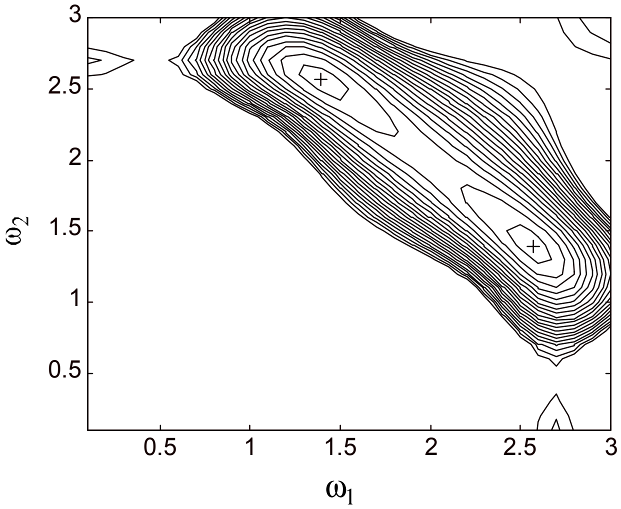
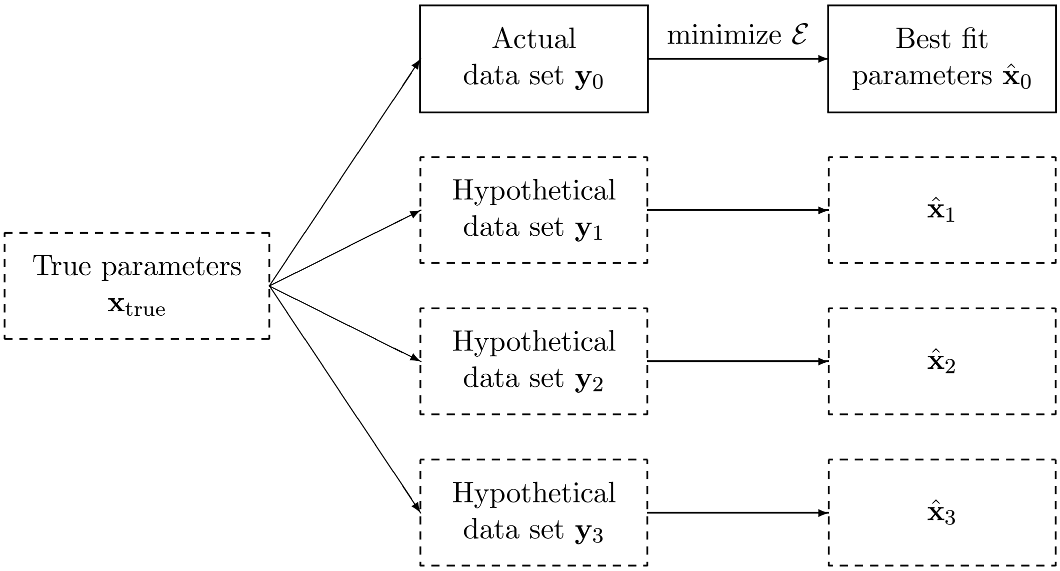


Figure 5.7 Merit function surface for problem of fitting two sinusoids.

5.10.3 Determining the adequacy of the model and error estimates for the parameters

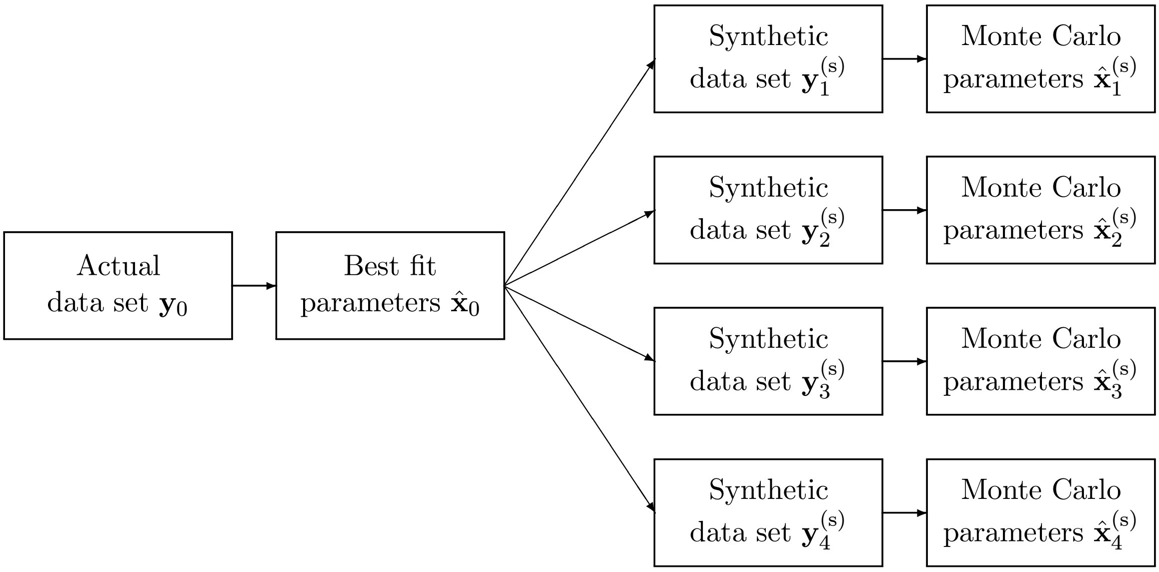
In the Bayesian formalism, we should investigate the posterior probability function (which is related to the merit function) to see how the probability for the parameter estimates change away from the point at which the merit function is minimized. In particular, we are concerned that the posterior probability density may be such that its maximum is not a good representative of the function, for example because there may be multiple maxima, or because most of the probability may in fact be located in a low, broad peak away from the maximum. Since the posterior probability (or the likelihood) is a function of many variables ($M = 6$ in our example), it is often very difficult to visualize, unless there is a simple analytic form for the result. One solution is to try to find an algorithm which generates samples of M variables which are drawn from the posterior probability function. This is a technique which is possible for some classes of posterior probability functions, and will be discussed in more detail in the third part of this course.

Often, however, we are unable to handle the posterior probability function because of its high dimensionality but still wish to make a statement about the quality of our parameter estimates. To do this, we adopt a different point of view introduced at the end of the last chapter. Instead on focussing on the **particular data set** that we collected, we ask what is the **likely range** of possible data sets given that the parameters have **specified** values.


Actual data set $\mathbf{y}_0$
minimize $\mathcal{E}$
Best fit parameters $\hat{\mathbf{x}}_0$
True parameters $\mathbf{x}_{\text{true}}$
Hypothetical data set $\mathbf{y}_1$
$\hat{\mathbf{x}}_1$
Hypothetical data set $\mathbf{y}_2$
$\hat{\mathbf{x}}_2$
Hypothetical data set $\mathbf{y}_3$
$\hat{\mathbf{x}}_3$

In the above diagram, we show the conceptual framework. There are some "true" parameter values $\mathbf{x}_{\text{true}}$ which generate the data using the model defined by the forward probability $p(\mathbf{y}|\mathbf{x}_{\text{true}})$. The actual data set $\mathbf{y}_0$ we collected is a **particular sample** from this forward probability. However, since the noise could have been different, other possible (hypothetical) data sets are $\mathbf{y}_1$, $\mathbf{y}_2, \ldots$, as shown. From a data set, we have an algorithm for estimating the parameters $\hat{\mathbf{x}}$ which may involve defining a merit function and minimizing this, just as described above. Using this algorithm, we compute the estimate $\hat{\mathbf{x}}_0$ from the actual data set $\mathbf{y}_0$. Conceptually, however, we can also compute estimates $\hat{\mathbf{x}}_1$, $\hat{\mathbf{x}}_2, \ldots$ from the hypothetical data sets $\mathbf{y}_1$, $\mathbf{y}_2, \ldots$ We now look at the **width** of the **distributions** of the estimates $\hat{\mathbf{x}}_0$, $\hat{\mathbf{x}}_1$, $\hat{\mathbf{x}}_2, \ldots$ about $\mathbf{x}_{\text{true}}$ in order to quantify the accuracy of the estimates.

Unfortunately, this strategy requires us to know the value of $\mathbf{x}_{\text{true}}$, which is of course unavailable. What we do instead is to first calculate $\hat{\mathbf{x}}_0$ from the actual data set $\mathbf{y}_0$ and pretend that this is the true value of $\mathbf{x}$. We then construct "synthetic data sets" $\mathbf{y}_1^{(\text{s})}$, $\mathbf{y}_2^{(\text{s})}, \ldots$ by drawing from the forward probability function $p(\mathbf{y}|\hat{\mathbf{x}}_0)$. This requires us to have a reasonable idea of the noise process. From each of the synthetic data sets, we carry out the estimation process to find $\hat{\mathbf{x}}_1^{(\text{s})}$, $\hat{\mathbf{x}}_2^{(\text{s})}, \ldots$ which we call **Monte Carlo parameters.** This process is shown in the diagram.


Actual data set $\mathbf{y}_0$
Best fit parameters $\hat{\mathbf{x}}_0$
Synthetic data set $\mathbf{y}_1^{(s)}$
Synthetic data set $\mathbf{y}_2^{(s)}$
Synthetic data set $\mathbf{y}_3^{(s)}$
Synthetic data set $\mathbf{y}_4^{(s)}$
Monte Carlo parameters $\hat{\mathbf{x}}_1^{(s)}$
Monte Carlo parameters $\hat{\mathbf{x}}_2^{(s)}$
Monte Carlo parameters $\hat{\mathbf{x}}_3^{(s)}$
Monte Carlo parameters $\hat{\mathbf{x}}_4^{(s)}$

We then study the distribution of $\hat{\mathbf{x}}_1^{(s)}$, $\hat{\mathbf{x}}_2^{(s)}$, ... about $\hat{\mathbf{x}}_0$ to tell us about the accuracy of estimation process. When forming each Monte Carlo parameter estimate, we obtain a value for the merit function at the minimum. By plotting a histogram of these minima, we can see whether $\mathcal{E}\left(\hat{\mathbf{x}}_0; \mathbf{y}_0\right) = \min \mathcal{E}\left(\mathbf{x}; \mathbf{y}_0\right)$ is reasonable. This gives an indication of the **model adequacy.** Note that we are making the (possibly big) **assumption** that the distribution of $\hat{\mathbf{x}}_k^{(s)}$ about $\hat{\mathbf{x}}_0$ is not too different from the (inaccessible) distribution of $\hat{\mathbf{x}}_k$ about $\mathbf{x}_{\text{true}}$.

The Matlab program below illustrates how Monte Carlo simulation may be carried out for the fitting problem

```
clear
load fitdata
Nmc = 100;
ymock = makeC(xnlbest,tlist) * xlinbest;
fid = fopen('fitmc.dat','w');

for k = 1:Nmc
% Make synthetic data
   ysyn = ymock + sigma*randn(size(y));
   xnlMC = fmins('misfit',xnlbest,0,[],tlist,ysyn);
   xnlMC = sort(xnlMC);
   C = makeC(xnlMC,tlist);
   xlinMC = (C'*C)\(C'*ysyn);
   EMC = misfit(xnlMC,tlist,ysyn);
   fprintf(fid,'%f  ',xnlMC,xlinMC,EMC);
   fprintf(fid,'\n');
end
fclose(fid);
```

The results of the simulation are stored in the file `fitmc.dat` and are summarized in the following table. Each row corresponds to a new synthetic data set, and the estimates of the parameters and the minimum value of the merit function are tabulated for that data.

	$\omega_1^{(s)}$	$\omega_2^{(s)}$	$A_1^{(s)}$	$B_1^{(s)}$	$A_2^{(s)}$	$B_2^{(s)}$	$\mathcal{E}_{\min}$
Data realization $\mathbf{y}_1^{(s)}$	1.33	2.63	1.32	0.50	1.71	0.49	8.61
Data realization $\mathbf{y}_2^{(s)}$	1.36	2.58	1.38	0.61	1.91	0.31	7.50
Data realization $\mathbf{y}_3^{(s)}$	1.45	2.48	1.40	0.81	1.86	−0.22	14.66
	⋮	⋮	⋮	⋮	⋮	⋮	⋮
Data realization $\mathbf{y}_{100}^{(s)}$	1.40	2.60	1.25	0.93	1.70	0.61	12.14
Mean	1.38	2.57	1.27	0.71	1.71	0.29	8.73
Standard deviation	0.08	0.07	0.16	0.34	0.16	0.35	2.19

From this table, we can place error bars on the estimates found previously. We write

$$\begin{aligned}
\omega_1 &= 1.40 \pm 0.08\\
\omega_2 &= 2.57 \pm 0.07\\
A_1 &= 1.2 \pm 0.2\\
B_1 &= 0.7 \pm 0.3\\
A_2 &= 1.7 \pm 0.2\\
B_2 &= 0.3 \pm 0.4
\end{aligned}$$

Since the means over the Monte Carlo runs lie within these error bars, we have no evidence that the estimates are biassed. Notice that the minimum value of $\mathcal{E}$ obtained in the original data fit is 7.4, which is well within the range of $\mathcal{E}_{\min}$ obtained during the Monte Carlo simulations. Note that the distribution of $\mathcal{E}_{\min}$ is **not** Gaussian, in general, and so we would usually not be too alarmed even if the value of $\mathcal{E}_{\min}$ that we obtained in the original fit is several standard deviations larger than the average in the Monte Carlo simulations. We would become concerned about using an inadequate model only if the probability of getting the value of $\mathcal{E}_{\min}$ found in the original fit is less than about 10^{-3}.

5.10.4 The Joint Gaussian Approximation

The above method of using Monte Carlo simulation of many data sets is very general and allows us to consider non-linear models and non-Gaussian noise processes. However, the need to produce many synthetic data sets can be quite time consuming. If the posterior probability function can be approximated by a multivariate Gaussian, it is possible to obtain error estimates on the parameters by using the technique described previously. Since we are approximating the posterior probability by

$$\mathcal{N} \exp\left(-\frac{1}{2}\left[\mathcal{E}\left(\mathbf{x};\mathbf{y}\right) - \mathcal{S}\left(\mathbf{x}\right)\right]\right) \tag{5.155}$$

or the likelihood function by

$$\mathcal{N} \exp\left(-\frac{1}{2}\mathcal{E}\left(\mathbf{x};\mathbf{y}\right)\right) \tag{5.156}$$

the **formal covariance matrix** is $\mathbf{Q}^{-1}$ where

$$Q_{ij}=\frac{\partial^2 L}{\partial x_i \partial x_j}, \tag{5.157}$$

and $L(\mathbf{x})$ is either $\mathcal{E}(\mathbf{x};\mathbf{y}) - \mathcal{S}(\mathbf{x})$ for the MAP estimator or is $\mathcal{E}(\mathbf{x};\mathbf{y})$ for the maximum likelihood estimator. The formal variances of the individual parameters are found by taking the diagonal elements of $\mathbf{Q}^{-1}$.

In order to assess the model adequacy, it can be shown that in the special case of **additive Gaussian noise** of covariance Γ and a purely linear model $\hat{\mathbf{y}}(\mathbf{x}) = \mathbf{C}\mathbf{x}$, for some $N \times M$ matrix $\mathbf{C}$, it can be shown if we define

$$\mathcal{E}(\mathbf{x};\mathbf{y}) = (\mathbf{y} - \hat{\mathbf{y}}(\mathbf{x}))^t \, \Gamma^{-1} (\mathbf{y} - \hat{\mathbf{y}}(\mathbf{x}))$$

then for a given data set $\mathbf{y}$, the distribution of

$$\mathcal{E}_{\min} \equiv \min_{\mathbf{x}} \mathcal{E}(\mathbf{x};\mathbf{y})$$

is χ^2 with $N - M$ degrees of freedom. By using tables of the χ^2 distribution, one can judge if the value of $\mathcal{E}_{\min}$ obtained in the fit is reasonable.

These results are often used in practise even for non-linear fitting problems, although this can be dangerous in pathological cases. In particular, the quantitative confidence levels for Gaussian distributions will generally differ from those of the true posterior probability, even though the shapes of the contours of equal posterior probability (which are assumed to be ellipsoids) may be reasonable near the optimum parameter values.