5 Classical Inference II: Optimization

In Chapter 4, we introduced the notions of statistics and estimators, discussed their basic properties, and examined specific basic estimators, most particularly those of moments and centered moments, used in classical inference. Although of obvious interest, moments are often insufficient to fully characterize the probability distribution governing a particular phenomenon or dataset. It is often the case that the general functional form of a distribution is known, but not fully specified. Indeed, while the functional dependence on a random variable X may be known, the function might have dependencies on finitely many model parameters $\vec{\theta} = (\theta_1, \theta_2, \ldots, \theta_m)$, which are a priori unknown or unspecified. For instance, it might be known that a particular dataset behaves according to a log-normal distribution, but the parameters μ and σ that determine the specific shape of the distribution could be unknown. Alternatively, one might be interested in determining the values of model parameters governing the dependence between dependent and independent variables. One is thus in need of methods to determine the parameters $\vec{\theta}$ of a distribution that best describe or match a set of measured data. Several such "fitting" methods exist, including the method of maximum likelihood, the extended method of maximum likelihood, various variants of least-squares methods, and Kalman filter methods. These are presented in §5.1, §5.1.7, §5.2, and §5.6, respectively. An excellent discussion of the relatively less used method of moments is presented in ref. [67].

The maximum likelihood and the least-squares methods are optimization problems: both involve the optimization of a **goal function**, often called **objective function** or **merit function**. The former involves *maximization* of a **likelihood function** while the latter requires *minimization* of a χ^2 **function**. Both techniques involve a search in (model) parameter space for an optimum value, that is, an extremum of an objective function. While such searches are relatively simple when the model involves only a few parameters, they may become particularly challenging when models involve a very large number of parameters. Several techniques exist to handle searches in multiparameter space and it is clearly not possible to cover them all in this introductory text. We thus focus our discussion on the general principles of the maximum likelihood and the least-squares methods in following sections of this chapter, and present a selection of optimization techniques pertaining to both methods in §7.6 after the introduction of Bayesian inference methods in §§7.2–7.4.

Once an optimum is achieved, one wishes to establish how good the fit really is. One thus requires a measure of the goodness-of-fit. Such a measure is discussed in §5.3 on the basis of the likelihood and chi-square functions. One is then particularly interested in evaluating errors on the parameters obtained in the optimization. This and related matters are discussed in §5.3. With parameter values and error estimates in hand, it becomes possible to extend or extrapolate the results predicted by the model. Techniques to evaluate the

https://doi.org/10.1017/9781108241922.007 Published online by Cambridge University Press

errors on such extrapolations are presented in §5.4, whereas §5.5 presents a technique to average the results (i.e., parameter values) obtained by two or more experimental studies.

5.1 The Method of Maximum Likelihood

Let $p(x|\vec{\theta})\,dx$ determine the probability a random variable X be found in the interval $[x, x + dx]$ given m parameters $\vec{\theta}$. Let us assume that the functional form of $p(x|\vec{\theta})$ is known but not the values of the parameters $\vec{\theta}$. Our goal is thus to obtain estimators of these parameters $\vec{\theta}$ based on the likelihood of a set $\{x_i\}$, $i = 1, \ldots, n$, of measured data.

5.1.1 Basic Principle of the $\mathbb{ML}$ Method

Let us assume the measurement of an observable X is repeated n times, thereby yielding a set of values $\{x_i\}$, $i = 1, \ldots, n$. If the parameters $\vec{\theta}$ were known, the data probability model embodied in the probability density function (PDF) $p(x|\vec{\theta})$ would give us the probability to obtain the value x_1 in the interval $[x_1, x_1 + dx]$, the value x_2 in the interval $[x_2, x_2 + dx]$, and so on. Assuming all n measurements are independent and yield uncorrelated results, the probability of measuring specifically the values $\vec{x} = (x_1, x_2, \ldots, x_n)$ is then simply the product of their respective probabilities:

$$p(x_1|\vec{\theta})\,dx_1 \times p(x_2|\vec{\theta})\,dx_2 \times \cdots \times p(x_n|\vec{\theta})\,dx_n = \prod_{i=1}^{n} p(x_i|\vec{\theta})\,dx_i. \tag{5.1}$$

If the PDF $p(x|\vec{\theta})$ is a good model of the data, one would expect the foregoing probability to be relatively large. Conversely, a poor model of the data should yield a low probability. Obviously, since $p(x|\vec{\theta})$ is a function of $\vec{\theta} = (\theta_1, \theta_2, \ldots, \theta_m)$, the value of the above probability must explicitly depend on the values of these m parameters. Well-chosen values of $\vec{\theta}$ should lead to a high probability, whereas a poor choice should result in a low probability. Since the intervals dx_i feature no dependency on the parameters $\vec{\theta}$, it is then sensible to seek an extremum of the product $\prod_{i=1}^{n} p(x_i|\vec{\theta})$ which corresponds to the **likelihood function** $L(\vec{x}|n, \vec{\theta})$ already introduced in Eq. (4.6):

$$L(\vec{x}|\vec{\theta}) \equiv \prod_{i=1}^{n} f(x_i|\vec{\theta}). \tag{5.2}$$

Nominally, $L(\vec{x}|\vec{\theta})$ corresponds to the joint PDF of the measured x_i given the parameters $\vec{\theta}$, but in this context, the data points are considered fixed (i.e., constant) and one seeks the values $\vec{\theta}$ that maximize the likelihood given the data points; in other words, one seeks the parameter values $\vec{\theta}$ such that the measured data points are the most probable.

The $\mathbb{ML}$ method specifically consists in seeking an extremum (a maximum actually) of $L(\vec{x}|\vec{\theta})$ relative to a variation of the m parameters $\vec{\theta}$:

$$\frac{\partial L(\vec{\theta})}{\partial \theta_i} = 0 \quad \text{for} \quad i = 1, \ldots, m, \tag{5.3}$$

https://doi.org/10.1017/9781108241922.007 Published online by Cambridge University Press

Simultaneous solution of these equations yields the $\mathbb{ML}$ estimators $\hat{\theta}_i$ of the parameters θ_i. A valid solution exists provided the function L is differentiable with respect to the parameters θ_i, and the extremum is not on a boundary of the parameters' range. The $\mathbb{ML}$ estimator $\hat{\theta}$ thus corresponds to the most likely value of $\vec{\theta}$ based on the n data points x_i.

Conceivably, depending on the dataset, the form of the functional, and the number of parameters, several solutions may exist that correspond to **local maxima**. Great care must then be taken to fully explore the parameter space in order to find the parameters $\vec{\theta}$ with the largest likelihood L.

We emphasize that given that the solution of $\partial L(\vec{\theta})/\partial\theta_i = 0$ is obtained for a specific set of data points, the parameter values θ_i corresponding to the extremum thus constitute estimators of the actual values. As such, it is convenient to write the solution(s) with a hat, $\hat{\theta}_i$, which indicate they are estimators to be distinguished from the true values $\vec{\theta}_i$.

Proof that the $\mathbb{ML}$ method produces consistent and unbiased estimators may be found, for instance, in ref. [116].

The $\mathbb{ML}$ method is relatively easy to use and does not require data to be binned or histogrammed. We consider a few examples of application of the method in the following sections.

5.1.2 Example 1: $\mathbb{ML}$ Estimator of the Rate Parameter of the Poisson Distribution

Very massive stars end their existence in spectacular explosions known as supernovae. Supernovae are a relatively rare phenomenon taking place randomly in this and other galaxies. Imagine disposing of a large aperture telescope (several meters) equipped with a high-efficiency and high-resolution camera. You might then be interested in characterizing the rate of supernovae explosions according to the type of galaxy where they take place. Being a rare phenomenon, the number of observations n per time period (e.g., per night) may be modeled according to a Poisson distribution

$$p(n|\mu) = \frac{e^{-\mu}\mu^n}{n!}, \tag{5.4}$$

where μ represents the average rate of explosions (per night). Several nights of observation will be required to carry out the measurement. For the sake of simplicity, let us assume it is possible to observe the same region of the sky and for the same exact duration during $N = 100$ nights, with equal observational conditions. The number of observations made nightly are labeled n_i, $i = 1, \ldots, N$.

Our goal is to determine the rate μ using an $\mathbb{ML}$ estimator. We thus need an expression for the likelihood of the data $\{n_i\}$, $i = 1, \ldots, N$, given a specific value of μ. Equation (5.4) provides the probability of observing n explosions in one night given a mean μ. The likelihood of the data amounts to the probability of a sequence $n_1, n_2, \ldots, n_N$ and is thus simply the product of the probabilities $p(n_i|\mu)$ of observing n_i explosions during

https://doi.org/10.1017/9781108241922.007 Published online by Cambridge University Press

nights $i = 1, \ldots, N$:

$$L(\mu) = \prod_{i=1}^{N} \frac{e^{-\mu}\mu^{n_i}}{n_i!}, \tag{5.5}$$

$$= \left(\prod_{i=1}^{N} n_i!\right)^{-1} \exp\left(-N\mu\right) \mu^{\left(\sum_{i=1}^{N} n_i\right)}. \tag{5.6}$$

Given the exponential factors, it seems simpler to maximize the log of the likelihood, $\ln L$, rather than L. Indeed, since $\ln L$ is a monotonically increasing function of its argument L, an extremum of this argument shall also correspond to an extremum of the log, and conversely. The technique is then referred to as log-maximum-likelihood, or simply $\mathbb{LML}$ method. We proceed to seek an extremum of

$$\ln L(\mu) = -N\mu + \left(\sum_{i=1}^{N} n_i\right) \ln \mu - \sum_{i=1}^{N} \ln n_i, \tag{5.7}$$

by taking a derivative relative to μ

$$0 = \frac{\partial \ln L}{\partial \mu} = -N + \frac{\sum_{i=1}^{N} n_i}{\mu}, \tag{5.8}$$

which yields

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} n_i. \tag{5.9}$$

We conclude that the $\mathbb{LML}$ estimator for the rate parameter μ of the Poisson distribution is equal to the arithmetic mean of the observations $\{n_i\}$, $i = 1, \ldots, N$.

As a concrete example of the method, we consider a simulation of a measurement of the number of supernova over a span of 100 nights. The number n_i of explosions observed nightly are generated with a Poisson random number generator, with a rate parameter $\mu = 2.5$, and shown in Figure 5.1 as a histogram. The rate parameter is assumed unknown in the remainder of the analysis. We then proceed to calculate the likelihood of the simulated data, $L(\mu)$, as function of the nightly rate μ in the range $[0, 4]$ according to Eq. (5.5). The likelihood $L(\mu)$ is vanishingly small for most values of μ but clearly peaks near $\mu = 2.5$. The estimate obtained with the $\mathbb{ML}$ method thus corresponds to an extremum (mode) of the likelihood function $L(\mu)$, which is indeed very close to the actual value of the parameter used in the simulation. We will see, later in this chapter, that the width of the likelihood function may be used to assess an error on the estimate.

5.1.3 Example 2: $\mathbb{LML}$ Estimator of the Decay Constant of the Exponential PDF

Consider a radiological experiment reporting n values $t_1, t_2, \ldots, t_n$ corresponding to decay times of some radioactive isotope X. If the production of this isotope involves no feed down from heavier isotopes, it is then legitimate to assume the data may be represented by

https://doi.org/10.1017/9781108241922.007 Published online by Cambridge University Press

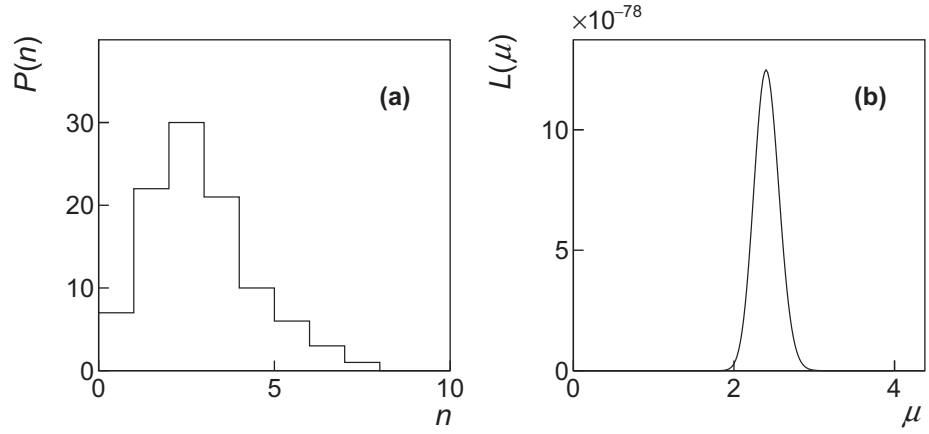

Fig. 5.1 (a) Distribution of the number of supernovae explosion per night in a simulated measurement spanning 100 nights for a rate parameter of $\mu = 2.5$. (b) Likelihood of the simulated data plotted as a function of μ (assumed unknown).

an exponential PDF:

$$p(t|\tau) = \frac{1}{\tau} e^{-t/\tau}. \tag{5.10}$$

Our goal is to determine the mean lifetime of the isotope, that is, find the parameter value τ that best represents the collected data and can then be used to characterize the isotope X.

The likelihood function is here the product of n exponential factors

$$L(\tau) = \prod_{i=1}^{n} p(t_i|\tau) = \prod_{i=1}^{n} \frac{1}{\tau} e^{-t_i/\tau}, \tag{5.11}$$

which readily transforms into the exponential of a sum

$$L(\tau) = \tau^{-n} \exp\left(-\sum_{i=1}^{n} t_i/\tau\right). \tag{5.12}$$

Here again, it is simpler to seek an extremum of $\ln L$ rather than L. We find

$$0 = \frac{d \ln L}{d\tau} = \left[-n\tau^{-1} + \tau^{-2} \sum_{i=1}^{n} t_i\right] \tau^{-n} \exp\left(-\sum_{i=1}^{n} t_i/\tau\right). \tag{5.13}$$

The lifetime τ and exponential are nonzero. The factor in square brackets must consequently be null. Solving for τ thus yields the $\mathbb{LML}$ estimator $\hat{\tau}$

$$\hat{\tau} = \frac{1}{n} \sum_{i=1}^{n} t_i, \tag{5.14}$$

as the arithmetic mean of the measured decay times. It is easily verified that $\hat{\tau}$ is an unbiased estimator of the mean lifetime τ (see Problem 5.1).

As a specific example of application of the estimator (5.14), imagine an experiment involving the measurement of decays of the unstable isotope ^{11}C, with a mean lifetime

https://doi.org/10.1017/9781108241922.007 Published online by Cambridge University Press

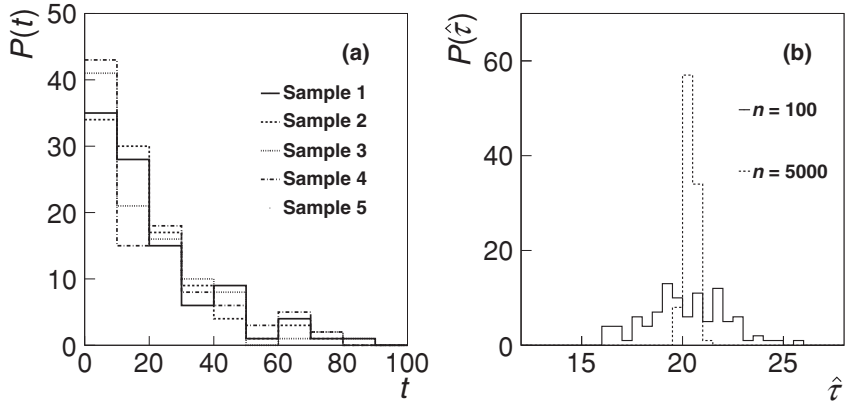

Fig. 5.2 (a) Histograms of decay times of ^{11}C measured in five different simulated experiments (samples). (b) Histogram of measured estimates $\hat{\tau}$ obtained from 100 samples based on measurements of 100 (solid line) and 5000 (dashed line) decay times.

of 20.334 s. We simulated such an experiment using the exponential random generator introduced in §13.3.3 and produced a set of hundred sequences of 100 decay time values. Five of these sequences were used to fill the histograms plotted in Figure 5.2a. We used Eq. (5.14) to calculate hundred estimates of the ^{11}C lifetime on the basis of these sequences. The values were used to fill a histogram of the estimators, shown in Figure 5.2b. We find that the estimates are broadly distributed about the value of 20.334 s used to generate the random numbers. We then repeated the simulations, but this time with datasets consisting of 5,000 points each. Estimates are once again computed on the basis of Eq. (5.14) and plotted in Figure 5.2b as the dotted histogram. As expected, we find these estimates are more narrowly concentrated about the mean. The estimates obtained with a sample size of 100 average to 20.19 s, whereas the estimates obtained with the sample size of 5,000 average to 20.35 s, which is closer to the value of 20.334 s used in their generation.

5.1.4 Example 3: $\mathbb{ML}$ Estimators of the Mean and Variance of a Gaussian PDF

As a third example of application of the $\mathbb{ML}$ method, we determine estimators of the mean and variance of a Gaussian PDF. Let us assume that n measured values x_i are distributed according to a Gaussian PDF with unknown mean μ and standard deviation σ. Given the exponential nature of the Gaussian PDF, it is once again convenient to use the logarithm of the likelihood function L:

$$\ln L(\mu, \sigma^2) = \ln \left(\prod_{i=1}^{n} p(x_i; \mu, \sigma^2) \right), \tag{5.15}$$

$$= \ln \left(\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp \left(-\frac{(x_i - \mu)^2}{2\sigma^2} \right) \right). \tag{5.16}$$

https://doi.org/10.1017/9781108241922.007 Published online by Cambridge University Press

Transforming the log of the product as a sum of logs, and rearranging the terms, one gets

$$\ln L(\mu, \sigma^2) = \sum_{i=1}^{n} \left(-\ln \sqrt{2\pi} - \frac{1}{2} \ln \sigma^2 - \frac{(x_i - \mu)^2}{2\sigma^2} \right), \tag{5.17}$$

$$= -n \ln \sqrt{2\pi} - \frac{n}{2} \ln \sigma^2 - \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2}. \tag{5.18}$$

The extremum of $\ln(L)$ is found in the usual way by equating its derivatives with respect to μ and σ to zero. Let us first consider an estimator of the mean:

$$0 = \frac{\partial \ln L(\mu, \sigma^2)}{\partial \mu} = -\sum_{i=1}^{n} \frac{\partial}{\partial \mu} \left(\frac{(x_i - \mu)^2}{2\sigma^2} \right). \tag{5.19}$$

Solving for μ yields the $\mathbb{ML}$ estimator $\hat{\mu}$:

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i, \tag{5.20}$$

which, again in this case, is the arithmetic mean of the sampled values. We already showed, in §4.5.1, that the arithmetic mean of a sample is in general an unbiased estimator of true mean μ. It is thus simple to verify that this conclusion applies specifically to the Gaussian distribution also (see Problem 5.7).

We next proceed to find an estimator for the variance σ^2 of the distribution:

$$0 = \frac{\partial \ln L(\mu, \sigma^2)}{\partial \sigma^2} = -\frac{n}{2} \frac{1}{\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{n} (x_i - \mu)^2. \tag{5.21}$$

Solution for σ^2 yields the estimator

$$\widehat{\sigma^2} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\mu})^2, \tag{5.22}$$

which based on our generic discussion of the variance of estimators, in §4.5.3, is known to be an *asymptotically unbiased* estimator of the variance (also see Problem 5.2) with the expectation value

$$\mathrm{E}[\widehat{\sigma^2}] = (n-1)\sigma^2/n. \tag{5.23}$$

Also recall, from §4.5.3, that the estimator s^2 given by Eq. (4.39) is an unbiased estimator of the variance of distributions. We then expect that it also constitutes an unbiased estimator of the variance of a Gaussian PDF (see Problem 5.9). However, s^2 is *not* the $\mathbb{ML}$ estimator of σ^2 for a Gaussian PDF.

5.1.5 Errors

There are limited instances of $\mathbb{ML}$ estimators whose variance can be calculated analytically. For instance, the variance of the estimator (5.14) of the mean of an exponential decay,

https://doi.org/10.1017/9781108241922.007 Published online by Cambridge University Press

$\hat{\tau} = \frac{1}{n}\sum_{i=1}^{n}$, can be computed by direct substitution in the expression of the variance. One gets

$$\mathrm{Var}[\hat{\tau}] = \mathrm{E}[\hat{\tau}^2] - \mathrm{E}[\hat{\tau}]^2 = \frac{\tau^2}{n}, \tag{5.24}$$

which is identical in form to the expression (4.27) of the variance of the sample mean. The variance of $\hat{\tau}$ is a function of τ, the true mean of the PDF considered. This might seem rather problematic, since τ is a priori unknown: how indeed does one report the standard error based on an unknown quantity? However, we have found that $\hat{\tau}$ is an unbiased estimator of τ obtained by finding an extremum of the likelihood function L. One can thus write

$$\frac{\partial L}{\partial \tau} = \frac{\partial L}{\partial \hat{\tau}}\frac{\partial \hat{\tau}}{\partial \tau} = 0. \tag{5.25}$$

Since $\partial\hat{\tau}/\partial\tau \neq 0$, one concludes that an extremum for τ yields an extremum for $\hat{\tau}$, and conversely. It is thus legitimate to use $\hat{\tau}$ in lieu of τ to get an estimate of the variance. One can then report a measurement of τ as $\hat{\tau} \pm \hat{\tau}/\sqrt{n}$. This implies that if an experiment was repeated several times, with the same number of measurements n, one would expect the standard deviations of the results (i.e., estimates) to be $\tau/\sqrt{n}$. Note, however, that in cases where the distribution of estimates significantly deviates from a Gaussian distribution, it is more meaningful (and common) to report an error corresponding to the 68.3% confidence interval (see §6.1.2).

Analytical computation of the variance $\mathrm{Var}[\hat{\theta}]$ of estimators of "complicated" observables may be tedious, difficult, or even impossible. Fortunately, a number of alternative computation techniques exist, few of which we briefly examine in the following. See also ref. [67] for an in-depth discussion of this technical topic.

The most basic technique, commonly applied, to estimate the variance of an estimator consists in carrying out several experiments yielding the quantity of interest. One repeats the experimental procedure several times and obtains several sets of measurements, $\{x_i\}_k$, where k is an index used to identify the different datasets. Estimators $\hat{\theta}_k$ are computed for each dataset k. One can then calculate the expectation values and variance of these estimators. Obviously, it may not always be possible to repeat a given experiment because of cost, lack of time, or because the observed phenomena might be unique by its very nature (e.g., observation of neutrinos from a particular supernova explosion). It may, however, be possible to split the dataset into several subsamples, each of which can be used to obtain distinct $\mathbb{ML}$ estimates. The variance of these estimates relative to the $\mathbb{ML}$ estimate of the full sample can thus be used to evaluate the variance of the estimator.

Whenever repetition or splitting of the data sample produced by an experiment is not an option, one may resort to a detailed simulation of the experiment to artificially create repeated evaluations of the estimator of interest. The idea is to replicate the conditions and procedure of a measurement in a Monte Carlo simulation. While such simulations of experiments will be discussed in detail in Chapter 14, the technique can be summarized as follows. Suppose one wishes to measure a variable X distributed according to a certain PDF $f(x|\theta)$. Once an estimate $\hat{\theta}$ of the parameter θ is obtained experimentally, one carries out repeated simulations of the experiment by generating instances of x based on $f(x|\hat{\theta})$,

https://doi.org/10.1017/9781108241922.007 Published online by Cambridge University Press

making sure experimental effects and correction procedures are properly taken into account. The simulated experiments yield estimates $\hat{\theta}_i$ of the parameter θ. Given a sufficient number of replications of the experiment, it is then possible to compute the variance of the values $\hat{\theta}_i$ and consequently obtain an estimate of the true variance of the estimator.

Yet another technique commonly applied to determine the variance of $\mathbb{ML}$ estimators is to use the minimal variance bound based on the Rao–Cramer–Frechet (RCF) inequality given by Eq. (4.17), which for several parameters $\vec{\theta} = (\theta_1, \ldots, \theta_m)$, may be written

$$(V^{-1})_{ij} = -\left\langle \frac{\partial^2 \ln L}{\partial\theta_i \partial\theta_j} \right\rangle = -\left. \frac{\partial^2 \ln L}{\partial\theta_i \partial\theta_j} \right|_{\vec{\theta}=\hat{\theta}}. \tag{5.26}$$

As an example, consider the calculation of the variance of estimators $\hat{\mu}$ and $\hat{\sigma}^2$ of the mean and variance of the Gaussian PDF. Using the log of the likelihood function (5.17), it is easy to calculate (see Problem 5.10) the second-order derivatives with respect to μ and σ^2:

$$\left\langle \frac{\partial^2 \ln L}{\partial\mu^2} \right\rangle = -\frac{n}{\sigma^2}, \tag{5.27}$$

$$\left\langle \frac{\partial^2 \ln L}{\partial\sigma^2} \right\rangle = -\frac{2n}{\sigma^2}, \tag{5.28}$$

$$\left\langle \frac{\partial^2 \ln L}{\partial\mu\partial\sigma} \right\rangle = 0. \tag{5.29}$$

The matrix V^{-1} is diagonal and is trivially inverted to yield the variances

$$\mathrm{Var}[\mu] = \left\langle \frac{\partial^2 \ln L}{\partial\mu^2} \right\rangle^{-1} = \frac{\sigma^2}{n}, \tag{5.30}$$

$$\mathrm{Var}[\sigma] = \left\langle \frac{\partial^2 \ln L}{\partial\sigma^2} \right\rangle^{-1} = \frac{\sigma^2}{2n}. \tag{5.31}$$

One finds that the variances $\mathrm{Var}[\mu]$ and $\mathrm{Var}[\sigma]$ are proportional to the variance σ^2 of the Gaussian PDF and inversely proportional to the size n of the data sample used to carry out the estimate. This is a rather general property that holds in the large n limit for most estimators. It expresses the well-known result that statistical errors are inversely proportional to the square root of n, the sample size.

The variance of $\mathbb{ML}$ estimators may also be determined using a graphical technique. The technique is based on the RCF bound discussed earlier and the fact that, near an extremum of the likelihood function L, one can expand the function in a Taylor series about the $\mathbb{ML}$ estimate $\hat{\theta}$:

$$\ln L(\theta) = \ln L(\hat{\theta}) + \frac{1}{2}\left[\frac{\partial^2 \ln L}{\partial\theta^2}\right]_{\theta=\hat{\theta}} (\theta - \hat{\theta})^2 + O(3), \tag{5.32}$$

where we omitted the first-order term in $\partial \ln L/\partial\theta$, which by construction, vanishes at the extremum. Based on Eq. (5.26), this can be written

$$\ln L(\theta) \approx \ln L_{\max} - \frac{(\theta - \hat{\theta})^2}{2\hat{\sigma}_\theta^2}, \tag{5.33}$$

https://doi.org/10.1017/9781108241922.007 Published online by Cambridge University Press

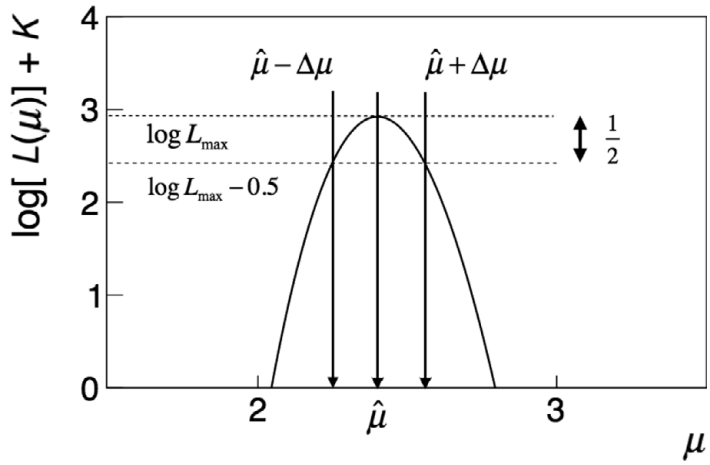

Fig. 5.3 Determination of a parameter error with the log-likelihood graphical method: The solid curve is the log of the likelihood function of the simulated data presented in Figure 5.1. An arbitrary constant K was added to $\ln L$ for convenience of presentation. The log-likelihood function is approximately parabolic in the vicinity of the extremum located at $\hat{\mu} = 2.4$ and the value $\log L - 1/2$ is found at 2.56. The standard error $\Delta\mu$ thus approximately amounts to $2.56 - 2.4 = 0.16$.

which means the error can be estimated graphically on the basis of

$$\ln L(\hat{\theta} \pm \sigma_\theta) = \ln L_{\max} - \frac{1}{2}. \tag{5.34}$$

The principle of the graphical method is illustrated in Figure 5.3, which presents a graph of the logarithm of the likelihood function of the rate parameter μ corresponding to simulated measurements, discussed in §5.1.2, of the number of supernovae observations detected nightly over a span of 100 days. The standard error on the parameter μ is here obtained by graphically finding the values $\hat{\mu} - \Delta\mu$ and $\hat{\mu} + \Delta\mu$ corresponding to $log[L_{\max}] - 1/2$, where $L_{\max}$ is the maximum likelihood observed.

5.1.6 Maximum Likelihood Fit of Binned Data

The LML method enables a relatively straightforward estimation of parameters for PDFs such as exponential or Gaussian distributions. However, it becomes impractical when the number of observations of random variable x becomes excessively large. This could be the case, for instance, if there is insufficient memory to store all the measured values. There are also issues of numerical accuracy for very large data samples. It is then often desirable, or more convenient, to carry out a *fit* based on a histogram of the data (§4.6). The range of interest $[x_{\min}, x_{\max}]$ is partitioned into m bins, chosen to permit identification of the relevant PDF features while accounting for the finite resolution of the measurements, the size of the sample, and the memory available. The bins are not required to be of equal size; one can use arbitrary bin boundaries $[x_{\min,i}, x_{\max,i}]$ for bins $i = 1, \ldots, m$. Consider a sample consisting of N measured values histogrammed into m bins. Each bin will contain

https://doi.org/10.1017/9781108241922.007 Published online by Cambridge University Press

a number of entries n_i (with $i = 1, \ldots, m$) representative of the PDF that characterizes the measured data. The number of entries in the whole histogram may then be expressed as a vector $\vec{n} = (n_1, n_2, \ldots, n_m)$. Obviously, each value n_i is subject to statistical fluctuations. However, the expectation value of the number of entries in bin i, noted ν_i, can be expressed as a function of the unknown parameters $\vec{\theta}$ of the PDF, given by the following expression:

$$\nu_i(\vec{\theta}) = N \int_{x_{\min,i}}^{x_{\max,i}} p(x|\vec{\theta})\, dx. \tag{5.35}$$

It is convenient to denote the expectation values ν_i in a vector form $\vec{\nu} = (\nu_1, \nu_2, \ldots, \nu_m)$ also. It is the purpose of the fit to determine the parameter(s) $\vec{\theta}$ most consistent with the measured values, in other words, the values that yield an extremum of the likelihood function $L(\vec{\theta})$.

The vector (histogram) $\vec{n}$ may be viewed as a single measurement of the m-dimensional vector $\vec{x}$. If all measurements of x are independent of one another, the values are uncorrelated. It implies that the number of entries in the m bins are uncorrelated (although they are obviously determined by the PDF). In this context, the measurement of the vector $\vec{n}$ with a total number of entries N amounts to a random partition of N draws into m bins, each with expectation $\nu_i(\vec{\theta})$. This corresponds to a multinomial PDF. The joint probability to measure the vector $\vec{n}$, given a total number of entries N and expectations $\vec{\nu}$, is then given by

$$p_{\text{joint}}(\vec{n}|\vec{\nu}) = \frac{N!}{n_1!n_2!\cdots n_m!}\left(\frac{\nu_1}{N}\right)^{n_1}\left(\frac{\nu_2}{N}\right)^{n_2}\cdots\left(\frac{\nu_m}{N}\right)^{n_m}, \tag{5.36}$$

where the ratio ν_i/N expresses the probability of getting entries in bin i. The logarithm of this expression yields the log-likelihood function

$$\ln L(\theta) = \ln\left(\frac{N!}{n_1!n_2!\cdots n_m!}\right) + \sum_{i=1}^{m} n_i \ln\left(\frac{\nu_i}{N}\right). \tag{5.37}$$

Clearly, the first term of the right-hand side involves variables that are independent of the PDF parameter(s) $\vec{\theta}$. Since we are seeking an extremum of L, it is unnecessary to keep track of these constants and the search for values $\vec{\theta}$ that maximize L can thus proceed on the basis of the second term alone by whatever optimization method is available or practical (see §7.6 for examples of such methods). When the number of bins is very large, $m \gg N$, then the binned method becomes equivalent to the standard $\mathbb{ML}$ method. This method is thus insensitive to the presence of null bins in the histograms, in stark contrast to the least-squares method discussed in §5.2.

5.1.7 Extended Maximum Likelihood Method

We saw in previous sections that the $\mathbb{ML}$ method is useful to determine unknown parameters $\vec{\theta} = (\theta_1, \theta_2, \ldots, \theta_n)$ of a PDF $p(\vec{x}|\vec{\theta})$ given a set of n measured values $\vec{x} = (x_1, x_2, \ldots, x_n)$. But what if the number of observations n is itself a random variable determined by the process or system under observation. This is the case, for instance, in measurements of scattering cross section where the number of produced particles is itself

https://doi.org/10.1017/9781108241922.007 Published online by Cambridge University Press

a random variable, or in observations and counting of radioactive decays of an unknown substance over a fixed period of time.

Let us consider processes in which the fluctuations of n are determined by a Poisson distribution with mean λ. The likelihood of observing n values $\vec{x}$ may then be described by an **extended likelihood function** as follows:

$$L(\lambda; \vec{\theta}) = \frac{\lambda^n e^{-\lambda}}{n!} \prod_{i=1}^{n} p(x_i|\vec{\theta}). \tag{5.38}$$

In general, λ may be regarded as a function of $\vec{\theta}$, $\lambda \equiv \lambda(\vec{\theta})$, or conversely $\vec{\theta} \equiv \vec{\theta}(\lambda)$. The log of the likelihood function may then be written

$$\ln L = n \ln \lambda(\vec{\theta}) - \lambda(\vec{\theta}) - \ln(n!) + \sum_{i=1}^{n} \ln p(x_i|\vec{\theta}), \tag{5.39}$$

$$= -\lambda(\vec{\theta}) + \sum_{i=1}^{n} \ln(\lambda(\theta) p(x_i|\vec{\theta})), \tag{5.40}$$

where in the second line, we have dropped unnecessary constants and use the fact that the log of a product of distinct factors equals the sum of their logs. Maximization of $\ln L$ is thus, in general, dependent on both the observed value of n as well as the measured values $\vec{x}$. The observed value n thus constrain the model parameters $\vec{\theta}$ and conversely, the observed values x_i constrain λ.

The maximization of $\ln L$ greatly simplifies, of course, if λ and $\vec{\theta}$ are independent. Derivatives of $\ln L$ relative to these variables yield independent conditions:

$$\frac{\partial \ln L}{\partial \lambda} = \frac{\partial}{\partial \lambda} (n \ln \lambda - \lambda) = 0, \tag{5.41}$$

$$\frac{\partial \ln L}{\partial \theta} = \frac{\partial}{\partial \theta} \sum_{i=1}^{n} \ln p(x_i|\vec{\theta}) = 0. \tag{5.42}$$

The first line yields $\hat{\lambda} = n$ whereas the second line amounts to the regular likelihood method, which is independent of λ. Use of the extended maximum likelihood thus appears to provide little gain over the regular method in this case.

The extended method remains of interest, nonetheless, for cases where the function $p(x_i|\vec{\theta})$ may be expressed as a linear combination of several (linearly independent) elementary functions

$$p(x|\vec{\theta}) = \sum_{k=1}^{m} \theta_k p_k(x), \tag{5.43}$$

with

$$\int p_k(x)\, dx = 1. \tag{5.44}$$

Such a situation arises when an observable can be expressed as a combination of finitely many signals, each with their distinct PDF $p_k(x)$, and unknown relative probability describable with parameters θ_k. A specific example of this situation involves the energy deposition

https://doi.org/10.1017/9781108241922.007 Published online by Cambridge University Press

of charged particles in a detector volume. One finds that the energy loss is a stochastic process that depends on particle types. The energy loss profile of particles at a given momentum (observed in a specific scattering experiment) thus depends on the relative probabilities of the different species (e.g., electron, pion, kaon, and so forth). The functions $p_k(x)$ could thus represent the energy loss profiles of different particle types while the parameters θ_k would represent their relative probabilities constrained by

$$\sum_{k=1}^{m} \theta_k = 1. \tag{5.45}$$

Evidently, one could treat this case as a regular application of the maximum likelihood method with $m-1$ independent parameters (i.e., with $\theta_m = 1 - \sum_{k=1}^{m-1} \theta_k$) but it is advantageous to carry out the search for an extremum of the likelihood function using all parameters θ_k and λ simultaneously.

Substituting the linear combination (5.43) for $p(x_i|\vec{\theta})$ in Eq. (5.40) yields

$$\ln L = -\lambda + \sum_{i=1}^{n} \ln \left(\sum_{k=1}^{m} \lambda \theta_k p_k(x_i) \right). \tag{5.46}$$

Let $\mu_k = \lambda \theta_k$ represent the mean value of the number of instances of type k. Using Eq. (5.45), $\ln L$ may then be written

$$\ln L(\vec{\mu}) = -\sum_{k=1}^{m} \mu_k + \sum_{i=1}^{n} \ln \left(\sum_{k=1}^{m} \mu_k p_k(x_i) \right), \tag{5.47}$$

where the parameters μ_k are not subjected to any constraints. The total number of events n may then be treated as a sum of independent Poisson variables μ_i and optimization of $\ln L$ thus yields estimates $\hat{\mu}_k$ of the mean of each of the types k. While mathematically equivalent to the independent optimization of λ and θ_k, this approach involves the advantage that all parameters are treated equally and one obtains the contributions of each type k directly.

5.2 The Method of Least-Squares

The $\mathbb{ML}$ and $\mathbb{LML}$ methods enable the determination of parameters that best characterize a data set given a specific PDF assumption. Although these methods are powerful, they may become fastidious, inconvenient, or impractical in many cases. However, an alternative, called the **Least-Squares** ($\mathbb{LS}$) method, is available and applicable in a wide range of situations.

We show in §5.2.1 how the $\mathbb{LS}$ method can be formally derived from the $\mathbb{ML}$ method in cases where the measured values can be considered Gaussian random variables. The $\mathbb{LS}$ method is, however, commonly applicable to problems of parameter estimation in which the Gaussian variable hypothesis is not strictly valid.[1]

[1] The $\mathbb{LS}$ method typically yields reasonable results whether or not fluctuations of the dependent variable y are Gaussian. However, one must be careful with the interpretation of error estimates obtained with non-Gaussian deviates.

https://doi.org/10.1017/9781108241922.007 Published online by Cambridge University Press

The formal derivation and definition of the $\mathbb{LS}$ method presented in §5.2.1 may be skipped in a first reading. Section 5.2.2 discusses the simple case of straight line fit and linear regression, which may be skipped by readers already familiar with these basic notions. Polynomial fits and progressively more complex minimization problems are presented in §§5.2.3–5.2.7.

5.2.1 Derivation of the $\mathbb{LS}$ Method

Consider a set of observations where a variable y is measured as a function of a variable x. Let us assume there exists, at least in principle, some relation between the two quantities. A basic measurement shall consist of n pairs (x_i, y_i) where the values x_i are assumed to be the control variable (whether explicitly controlled or not) and the values y_i are assumed to be functions of x_i. In this context, the variables y and x are also commonly called **dependent** and **independent** variables, respectively. For simplicity's sake, we will here assume the x_i are known without error. The $\mathbb{LS}$ method can, however, be generalized for cases where both x and y carry measurement errors. We will further assume the y_i are Gaussian random variables, that is, variables with a Gaussian PDF, $p_G(y_i|\mu_i, \sigma_i)$, defined by Eq. (3.124). This is meant to imply that if it were possible to repeat the measurement several times at the same value x_i, the measured values y_i would be distributed according to a Gaussian PDF of definite mean, μ_i and width σ_i. The values μ_i are assumed to depend on x_i in some manner we model with a function $f(x|\vec{\theta})$ that depends on one or more parameters, $\vec{\theta} = (\theta_1, \theta_2, \ldots \theta_m)$, of a priori unknown value:

$$\mu_i \equiv f(x_i|\vec{\theta}). \tag{5.48}$$

This type of measurement is rather general. Consider as a simple example, a measurement of the position, x, vs. time, t, of a car subjected to some unknown but constant acceleration (see example in §5.2.2). An $\mathbb{LS}$ fit of the data might then yield the value of this acceleration. Alternatively, one could measure the temperature (dependent variable) along a bar of metal (position) when the extremities of the bar are submitted to finite temperature differences, and a model of heat conduction could be used to describe the temperature profile along the bar. The possibilities are endless. One can envision measuring any physical quantity as a function of some other variable, be it time, space, currents, or electric and magnetic fields, and seek to model the relation between them. All one needs is a function, $y = f(x|\vec{\theta})$, modeling the relationship between y and x based on some "free" parameters, that is, model parameters $\vec{\theta}$ of unknown or unspecified value. We will introduce the $\mathbb{LS}$ method for problems involving a single independent variable, x, and one dependent variable, y, but it can be readily extended to an arbitrary number of independent and dependent variables.[2]

The goal of the $\mathbb{LS}$ method is to find the value(s) $\vec{\theta}$ that maximize the probability of getting the measured values y_i. But since the n variables y_i are by assumption distributed according to Gaussian PDFs $p_G(y_i|\mu_i, \sigma_i)$ of mean μ_i and width σ_i, we will apply the $\mathbb{LML}$ method to determine the value(s) of the parameters $\vec{\theta}$ that maximize the probability

[2] In this context, the phrase *independent variable* implies that x is a control variable, i.e., its values can be selected, or controlled, while the variable y adopts values possibly determined by x and is, as such, dependent on x.

https://doi.org/10.1017/9781108241922.007 Published online by Cambridge University Press

of measuring the values y_i at the given x_i. We use the vector notations $\vec{x}$ and $\vec{y}$ to denote all values x_i and y_i, respectively. The likelihood L of the measured values $\vec{y}$ depends on the joint probability, $g_{\text{joint}}(\vec{y}, \vec{x}|\vec{\theta})$. Assuming the n points (x_i, y_i) are measured independently and are thus uncorrelated, g_{joint} reduces to the product of the probabilities of all pairs (x_i, y_i):

$$g_{\text{joint}}(\vec{y}|\vec{\mu}, \vec{\sigma}) = \prod_{i=1}^{N} p_G(y_i|x_i; \mu_i, \sigma_i), \tag{5.49}$$

$$= \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{(y_i - \mu_i(x_i))^2}{2\sigma_i^2}\right), \tag{5.50}$$

where the values $\mu_i(x_i)$ are determined by the model function, $f(x|\vec{\theta})$, which is dependent on the unknown or unspecified parameter(s) $\vec{\theta}$[3]. We will assume there exists an estimate for the widths σ_i, that is, that values σ_i can be inferred either from the data directly, or on the basis of some theoretical considerations. The likelihood function of the values y_i may then be written:

$$L(\vec{y}|\vec{x}, \vec{\sigma}, \vec{\theta}) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{(y_i - f(x_i|\vec{\theta}))^2}{2\sigma_i^2}\right). \tag{5.51}$$

Note that this expression of the likelihood function assumes the data points are uncorrelated. Extension to a case in which the data points are correlated is relatively simple and will be discussed in §5.2.4.

Given its formulation as a product of exponentials, it is convenient to consider the logarithm of the likelihood function. Since the logarithm, $\ln(x)$, grows monotonically with x, a search for an extremum of the log of the likelihood function shall yield parameter values that maximize the likelihood function itself:

$$\ln L = -\frac{1}{2}\sum_{i=1}^{N} \ln\left(2\pi\sigma_i^2\right) - \frac{1}{2}\sum_{i=1}^{N} \frac{\left(y_i - f(x_i|\vec{\theta})\right)^2}{\sigma_i^2}. \tag{5.52}$$

The first term is not a function of the parameter(s) $\vec{\theta}$ and can be ignored in the search for an extremum of the log of the likelihood function. Consequently, maximization of $\ln(L)$ only involves the second term of Eq. (5.52) and it is thus convenient to define a **chi-square function** as follows:

$$\chi^2(\vec{\theta}) = \sum_{i=1}^{N} \frac{\left[y_i - f(x_i|\vec{\theta})\right]^2}{\sigma_i^2}. \tag{5.53}$$

Given the negative sign in front of the sum in Eq. (5.52), maximization of $\ln L$ then amounts to a minimization of the χ^2 function. This forms the basis of the $\mathbb{LS}$ method.

Proof that the method of least-squares produces consistent and unbiased estimators may be found for instance in ref. [116].

[3] The values x_i are taken as given, i.e., selected a priori. One consequently does not consider the probability of having such values. Only the y_i are considered random variables and assigned a probability.

https://doi.org/10.1017/9781108241922.007 Published online by Cambridge University Press

The fact that the minimization of the $\chi^2(\vec{\theta})$ function is equivalent to the maximization of the log-likelihood is based on the assumption that the random variables y_i are Gaussian distributed. Although this is not always true in practice, we note that by virtue of the central limit theorem, the multitude of random phenomena that produce the random character of the y_i, implies their distributions are nearly Gaussian in general. The $\mathbb{LS}$ method thus usually constitutes a reasonable approximation of the $\mathbb{ML}$ method. There are nonetheless cases where this approximation is not valid, and one must use the $\mathbb{ML}$ method, rather than the $\mathbb{LS}$ method, to carry out a search for optimal parameter(s) $\vec{\theta}$.

The parameters that minimize the χ^2 function are called $\mathbb{LS}$ estimators and are noted $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_m$ or simply $\hat{\theta}$ with the understanding that there are m such parameters. Additionally, the function is commonly called "chi-square" even in cases where the individual measurements y_i do not have Gaussian PDFs.

The χ^2 function, Eq. (5.53), was obtained based on the additional assumption that the N variables y_i are uncorrelated. When this assumption is not valid, and there are significant correlations among the variables y_i, one must transform Eq. (5.53) in terms of variables that are independent. In §5.2.4, we will introduce a technique to accomplish this in the context of the $\mathbb{LS}$ method. However, we first consider two cases of χ^2 minimization that do not involve such correlations: straight-line fits are presented in §5.2.2 while more general polynomial fits of order n are discussed in §5.2.3.

5.2.2 Straight-Line Fit and Linear Regression

Arguably the simplest and most common case of application of the $\mathbb{LS}$ method is for the determination of the parameters of a **straight line**, applicable when a variable y is known, or believed, to depend linearly on an independent variable x:

$$y = f(x|a_0, a_1) = a_0 + a_1 x. \tag{5.54}$$

Given a set of n measured points (x_i, y_i), with $i = 1, \ldots, n$, our goal is to find the values of the slope a_1 and the ordinate at the origin a_0 that minimize the chi-square function, χ^2, defined by Eq. (5.53). By virtue of our choice of model, the χ^2 is now a function of the parameters a_0 and a_1, which can be written

$$\chi^2 = \sum_{i=1}^{N} \frac{[y_i - f(x_i|a_0, a_1)]^2}{\sigma_i^2}. \tag{5.55}$$

We substitute the expression (5.54) for $f(x|a_0, a_1)$ in Eq. (5.55) to obtain a χ^2 function that explicitly depends on a_0 and a_1:

$$\chi^2(a_0, a_1) = \sum_{i=1}^{N} \frac{(y_i - a_0 - a_1 x_i)^2}{\sigma_i^2}. \tag{5.56}$$

We seek an extremum of this function (a minimum, actually) relative to variations of a_0 and a_1. This is accomplished by finding values of these two parameters for which derivatives

https://doi.org/10.1017/9781108241922.007 Published online by Cambridge University Press

with respect to a_0 and a_1 are null simultaneously:

$$\frac{\partial \chi^2(a_0, a_1)}{\partial a_0} = 0, \tag{5.57}$$

$$\frac{\partial \chi^2(a_0, a_1)}{\partial a_1} = 0. \tag{5.58}$$

Computation of these two derivatives yields

$$0 = \frac{\partial \chi^2(a_0, a_1)}{\partial a_0} = -2 \sum_{i=1}^{N} \frac{(y_i - a_0 - a_1 x_i)}{\sigma_i^2}, \tag{5.59}$$

$$0 = \frac{\partial \chi^2(a_0, a_1)}{\partial a_1} = -2 \sum_{i=1}^{N} \frac{(y_i - a_0 - a_1 x_i) x_i}{\sigma_i^2}. \tag{5.60}$$

Dropping the common multiplicative factors and rearranging, we get

$$a_0 \sum_{i=1}^{N} \frac{1}{\sigma_i^2} + a_1 \sum_{i=1}^{N} \frac{x_i}{\sigma_i^2} = \sum_{i=1}^{N} \frac{y_i}{\sigma_i^2}, \tag{5.61}$$

$$a_0 \sum_{i=1}^{N} \frac{x_i}{\sigma_i^2} + a_1 \sum_{i=1}^{N} \frac{x_i^2}{\sigma_i^2} = \sum_{i=1}^{N} \frac{x_i y_i}{\sigma_i^2}. \tag{5.62}$$

It is convenient to define the following quantities:

$$\begin{aligned} S &\equiv \sum_{i=1}^{N} \frac{1}{\sigma_i^2} & S_x &\equiv \sum_{i=1}^{N} \frac{x_i}{\sigma_i^2} \\ S_{xx} &\equiv \sum_{i=1}^{N} \frac{x_i^2}{\sigma_i^2} & S_y &\equiv \sum_{i=1}^{N} \frac{y_i}{\sigma_i^2} \\ S_{xy} &\equiv \sum_{i=1}^{N} \frac{x_i y_i}{\sigma_i^2} & \Delta &\equiv S_{xx} S - S_x^2 \end{aligned} \tag{5.63}$$

Eq. (5.61) may then be rewritten

$$a_0 S + a_1 S_x = S_y, \tag{5.64}$$

$$a_0 S_x + a_1 S_{xx} = S_{xy}, \tag{5.65}$$

or equivalently in matrix form:

$$\alpha \vec{a} = \vec{b}, \tag{5.66}$$

where we introduced the matrix α as well as the vectors $\vec{a}$ and $\vec{b}$ defined as follows:

$$\alpha = \begin{pmatrix} S & S_x \\ S_x & S_{xx} \end{pmatrix} \vec{a} = \begin{pmatrix} a_0 \\ a_1 \end{pmatrix} \vec{b} = \begin{pmatrix} S_y \\ S_{xy} \end{pmatrix}. \tag{5.67}$$

To solve for $\vec{a}$, we multiply both sides of Eq. (5.66) by the inverse α^{-1}:

$$\vec{a} = \alpha^{-1} \vec{b}. \tag{5.68}$$

https://doi.org/10.1017/9781108241922.007 Published online by Cambridge University Press

The inverse of α is

$$\alpha^{-1} = \frac{1}{\Delta}\begin{pmatrix} S_{xx} & -S_x \\ -S_x & S \end{pmatrix}. \tag{5.69}$$

The estimators $\hat{a}_0$ and $\hat{a}_1$, which minimize the χ^2 function, may thus be written

$$\hat{a}_0 = \frac{1}{\Delta}\left(S_y S_{xx} - S_x S_{xy}\right), \tag{5.70}$$

$$\hat{a}_1 = \frac{1}{\Delta}\left(S S_{xy} - S_x S_y\right). \tag{5.71}$$

The calculation of the coefficients S, S_x, S_{xx}, and so on relies exclusively on the points (x_i, y_i). Equations (5.70 and 5.71) then yield estimators $\hat{a}_0$ and $\hat{a}_1$ of the straight-line parameters that best fit the measured data points. Note that if the values y_i are computed without estimates of their standard deviations, σ_i, it suffices to set all values σ_i equal to unity in the foregoing calculations to obtain estimates of $\hat{a}_0$ and $\hat{a}_1$. However, the interpretation of the χ^2 of the fit in terms of a χ^2-distribution is not strictly possible in this case, nor are meaningful estimates of the errors on $\hat{a}_0$ and $\hat{a}_1$.

The same mathematical procedure applies whether one considers a straight-line fit or a linear regression. The term **fit** is, however, usually reserved for problems (or systems) where a model is used to infer a linear relationship between the dependent and independent variables. For instance, Hubble's law ($V = Hz$, with $z = \Delta\lambda/\lambda$) states there is a linear relation between the receding velocity, V, and the redshift, z, of galaxies. A linear fit carried out on a set of measured points, (z_i, V_i), consequently yields an estimate of the Hubble constant H. By contrast, the term **linear regression** is typically used for cases where no model is known a priori, or whenever large variances characterize both variables. The foregoing procedure thus yields an estimate of the trend between the variables, akin to an estimate of correlation. A linear regression can, for instance, be used to characterize the relationship between the height and weight of humans in a given population (see Figure 5.4).

The fact that the measurements y_i each carry an error σ_i implies the model parameters a_0 and a_1 are known with limited precision only. We can estimate their respective errors using the error propagation technique introduced in §2.11, Eqs. (2.222, 2.223). Given that only the coefficients S_y and S_{xy} are functions of y_i, we can write

$$\frac{\partial a_0}{\partial y_j} = \frac{1}{\Delta}\left(S_{xx}\frac{\partial S_y}{\partial y_j} - S_x\frac{\partial S_{xy}}{\partial y_j}\right), \tag{5.72}$$

$$\frac{\partial a_1}{\partial y_j} = \frac{1}{\Delta}\left(S\frac{\partial S_{xy}}{\partial y_j} - S_x\frac{\partial S_y}{\partial y_j}\right). \tag{5.73}$$

The derivatives of S_y and S_{xy} with respect to y_j yield $1/\sigma_j^2$ and x_j/σ_j^2, respectively. Inserting these values in the preceding expressions, we get

$$\frac{\partial a_0}{\partial y_j} = \frac{1}{\Delta}\left(S_{xx}\frac{1}{\sigma_j^2} - S_x\frac{x_j}{\sigma_j^2}\right), \tag{5.74}$$

$$\frac{\partial a_1}{\partial y_j} = \frac{1}{\Delta}\left(S\frac{x_j}{\sigma_j^2} - S_x\frac{1}{\sigma_j^2}\right). \tag{5.75}$$

https://doi.org/10.1017/9781108241922.007 Published online by Cambridge University Press

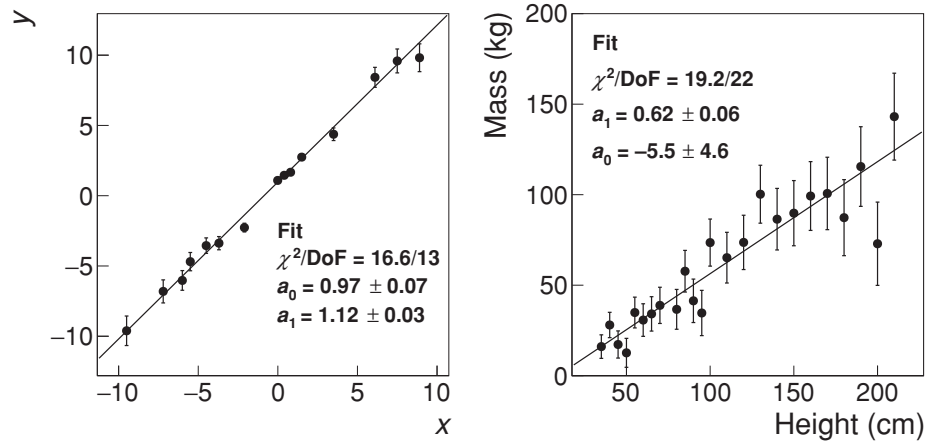

Fig. 5.4 (a) Straight-line fit on simulated data. (b) Example of a linear regression between the weight and height of children in an arbitrary population sample.

Assuming the y_i are independent, the variances $\sigma^2_{a_0}$ and $\sigma^2_{a_1}$ may now be estimated using Eq. (2.222), which we rewrite here in terms of a_0, a_1 as function of y_i:

$$\sigma^2_{a_0} = \sum_{j=1}^{N} \left[\frac{\partial a_0}{\partial y_j}\right]^2 \sigma_j^2, \tag{5.76}$$

$$\sigma^2_{a_1} = \sum_{j=1}^{N} \left[\frac{\partial a_1}{\partial y_j}\right]^2 \sigma_j^2. \tag{5.77}$$

Substituting the derivatives (5.74) in the preceding expressions, we get after simplification (see Problem 5.11):

$$\sigma^2_{a_0} = \frac{S_{xx}}{\Delta}, \tag{5.78}$$

$$\sigma^2_{a_1} = \frac{S}{\Delta}. \tag{5.79}$$

The variances $\sigma^2_{a_0}$ and $\sigma^2_{a_1}$ correspond respectively to the elements $(\alpha^{-1})_{11}$ and $(\alpha^{-1})_{22}$ of the inverse of matrix α. This is no mere accident and in fact derives from a general result we will discuss in §5.2.5.

5.2.3 $\mathbb{LS}$ Fit of a Polynomial

The $\mathbb{LS}$ method introduced in the previous section for linear fits is readily extended to polynomial fits of any order. For instance, let us assume the data may be represented by a polynomial of order m:

$$f(x) = a_o + a_1 x + \ldots + a_m x^m = \sum_{j=0}^{m} a_j x^j. \tag{5.80}$$

https://doi.org/10.1017/9781108241922.007 Published online by Cambridge University Press

For notational convenience, we represent the $m+1$ parameters a_i as a vector $\vec{a} = (a_0, a_1, \ldots, a_n)$. Once again, we assume the measurements y_i are independent. The χ^2 function may then be written

$$\chi^2(\vec{a}) = \sum_{i=1}^{N} \frac{\left(y_i - \sum_{k=0}^{m} a_k x_i^k\right)^2}{\sigma_i^2}. \tag{5.81}$$

We seek the values of the parameters a_j, $j = 0, \ldots, m$, that yield a minimum χ^2. This is accomplished by setting all derivatives of χ^2 with respect to the parameters a_j equal to zero simultaneously:

$$\frac{\partial \chi^2}{\partial a_j} = -2 \sum_{i=1}^{N} \frac{\left(y_i - \sum_{k=0}^{m} a_k x_i^k\right)}{\sigma_i^2} \frac{\partial}{\partial a_j} \left(\sum_{k=0}^{m} a_k x_i^k\right) = 0. \tag{5.82}$$

The derivative of $\sum_{k=0}^{m} a_k x_i^k$ with respect to a_j yields x^j. Equation (5.82) thus simplifies to

$$\frac{\partial \chi^2}{\partial a_j} = -2 \sum_{i=1}^{N} \frac{\left(y_i - \sum_{k=0}^{m} a_k x_i^k\right) x_i^j}{\sigma_i^2} = 0. \tag{5.83}$$

We rearrange and separate the terms to get

$$\sum_{i=1}^{N} \frac{y_i x_i^j}{\sigma_i^2} = \sum_{i=1}^{N} \sum_{k=0}^{m} \frac{a_k x_i^k x_i^j}{\sigma_i^2} = \sum_{k=0}^{m} a_k \sum_{i=1}^{N} \frac{x_i^k x_i^j}{\sigma_i^2}. \tag{5.84}$$

The index j takes values from 0 to m, and the index i runs from 1 to N. Equation (5.84) hence corresponds to $m+1$ equations that must be solved simultaneously. It is convenient to define a matrix α and a column vector $\vec{b}$ with the elements

$$\alpha_{kj} = \sum_{i=1}^{N} \frac{x_i^k x_i^j}{\sigma_i^2}, \tag{5.85}$$

$$b_j = \sum_{i=1}^{N} \frac{y_i x_i^j}{\sigma_i^2}. \tag{5.86}$$

Equation (5.84) may then be written as a linear equation:

$$\alpha \vec{a} = \vec{b}, \tag{5.87}$$

which yields the solution

$$\hat{a} = \alpha^{-1} \vec{b}, \tag{5.88}$$

in which we use the notation $\hat{a}$ to emphasize that the preceding expression yields estimators of the model parameters a_i, $i = 1, \ldots, m$.

A polynomial fit may thus be accomplished with the following three steps: (1) calculation of the matrix α, (2) calculation of the column vector $\vec{b}$, and (3) inversion of the matrix

https://doi.org/10.1017/9781108241922.007 Published online by Cambridge University Press

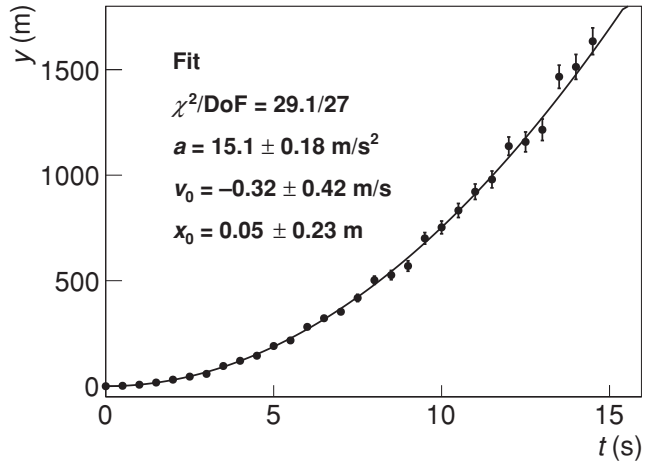

Fig. 5.5 Quadratic fit of simulated data (t_i, y_i) where the y_i are measured altitudes vs. times t_i of a rocket subjected to a constant acceleration a.

α and multiplication by $\vec{b}$ according to Eq. (5.88). Techniques and programs to invert matrices and solve linear equations are described in various textbooks and are available in various software packages.

Figure 5.5 displays a quadratic fit of simulated data describing the altitude y of a rocket as a function of time. The data were generated with the constant acceleration model $y(t) = 0.5at^2 + v_o t + y_o$, with values $a = 15$ m/s^2, $v_o = 0$, and $y_o = 0$. Measurement errors were simulated with Gaussian deviates with widths $\sigma_i = 5.00 + 0.03 * y$. The fit was carried out using the same quadratic model and assumed knowledge of the measurement errors. Trajectory parameters obtained from the fit are within statistical errors of the values used for the generation of the simulated trajectory.

We will show in §5.2.5 that the variances of the estimators $\hat{a}$ are given by the diagonal elements of the inverse matrix α. However, for polynomials of high-order m, the matrix α is prone to become **ill conditioned**, and its inversion may become numerically unstable. It is possible to partly remedy this problem by using **orthogonal polynomials**. Indeed, orthogonal polynomials, or any other complete basis of orthogonal functions, enable a straightforward and unique decomposition of arbitrary (continuous) functions. As such, they typically produce fit coefficients, for each element of the basis, that are nearly independent of one another.

5.2.4 $\mathbb{LS}$ Fit for Correlated Variables y_i

In §5.2.1, we showed that minimization of the χ^2 function, defined by Eq. (5.53), is equivalent to the maximum of the likelihood function L of measuring the values y_i. The derivation assumed the values y_i are mutually independent. There are, however, several classes of measurements that yield correlated variables y_i, that is, with nonvanishing covariances,

https://doi.org/10.1017/9781108241922.007 Published online by Cambridge University Press

$\mathrm{Cov}[y_i, y_j] \neq 0$ for $i \neq j$. Given the existence of such correlations, one expects the minimization of Eq. (5.53) to yield incorrect results because the values will be given inappropriate weights. Proper weights may be restored if one can transform the variables y_i, with nonzero covariances $\mathrm{Cov}[y_i, y_j]$, into a set of variables z_i with $\mathrm{Cov}[z_i, z_j] = 0$ for $i \neq j$. This can be readily accomplished by using the inverse of the covariance matrix of the variables y_i. Given $V_{ij} = \mathrm{Cov}[y_i, y_j] = \mathrm{E}[y_i y_j] - \mathrm{E}[y_i]\mathrm{E}[y_j]$, the χ^2 function may then be written

$$\chi^2(\vec{\theta}) = \sum_{i,j=1}^{N} (y_i - f(x_i|\vec{\theta}))(V^{-1})_{ij}(y_j - f(x_j|\vec{\theta})), \tag{5.89}$$

in which expectation values $\mu_i(x_i|\vec{\theta})$ were replaced by model functions $f(x_i|\vec{\theta})$. Note that if the variables y_i are uncorrelated, then the covariance matrix V_{ij} is diagonal, its inverse is a diagonal matrix with coefficients $\left(2\sigma_i^2\right)^{-1}$, and Eq. (5.53) is thus recovered. The parameters that minimize the function χ^2 are called $\mathbb{LS}$ estimators and are noted $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_m$ or simply $\hat{\theta}$. As in all other cases, minimization of χ^2 proceeds by equating the derivatives of Eq. (5.89) with respect to θ_i to zero. We discuss general implementations of the method for linear and nonlinear models in §§5.2.5 and 5.2.7 respectively, and an implementation for binned data in §5.2.6.

5.2.5 Generalized Linear Least-Squares Fit

The $\mathbb{LS}$ fit method is applicable for fits of any function $f(x|\vec{\theta})$ but is particularly well suited and considerably simplifies when f is a linear function of its parameters a_j, $j = 1, \ldots, m$

$$f(x_i|\vec{a}) = \sum_{j=1}^{m} a_j f_j(x_i), \tag{5.90}$$

where the coefficients $f_j(x_i)$ may be arbitrary functions of x, not just powers of x, as in the case of simple polynomial fits discussed in §5.2.3. However, the functions $f_j(x_i)$ must be linearly independent and may not depend on the model parameters a_j. **Orthogonal functions**, in particular, present the advantage that the coefficients a_j are not correlated. Commonly used functions, beside powers of x, include Fourier decompositions, orthogonal polynomials, and Legendre polynomials.

We repeat the steps carried out for fits with polynomials to obtain the estimates $\hat{a}_j$. However, we include the possibility, discussed in the previous section, that the data y_i might be correlated. For notational convenience, we introduce coefficients F_{ij} defined as follows:

$$F_{ij} = f_j(x_i). \tag{5.91}$$

The χ^2 function becomes

$$\chi^2 = \sum_{i,k=1}^{N} \left(y_i - \sum_{j=1}^{m} F_{ij} a_j \right) (V^{-1})_{ik} \left(y_k - \sum_{j'=1}^{m} F_{kj'} a_{j'} \right), \tag{5.92}$$

https://doi.org/10.1017/9781108241922.007 Published online by Cambridge University Press

which may also be formulated in a convenient matrix form

$$\chi^2 = (\vec{y} - \mathbf{F}\vec{a})^T V^{-1} (\vec{y} - \mathbf{F}\vec{a}), \tag{5.93}$$

in which $\vec{y} - \mathbf{F}\vec{a}$ is considered an $N \times 1$ column vector, $\vec{y}$ being a column vector representing all N entries y_i. $\mathbf{F}$ is an $N \times m$ matrix with elements equal to the coefficients F_{ij}, and $\vec{a}$ is an $m \times 1$ column vector containing the parameters a_j. The notation $\mathbf{O}^T$ is used to denote the transpose of matrix $\mathbf{O}$.

We find the minimum of χ^2 by differentiating with respect to a_p, with $p = 0, \ldots, m$:

$$0 = \frac{\partial \chi^2}{\partial a_p}, \tag{5.94}$$

$$= -\sum_{i,k=1}^{N} \left(\sum_{j=1}^{m} F_{ij}\delta_{jp} \right) (V^{-1})_{ik} \left(y_k - \sum_{j'=1}^{m} F_{kj'} a_{j'} \right)$$

$$- \sum_{i,k=1}^{N} \left(y_k - \sum_{j=1}^{m} F_{kj} a_j \right) (V^{-1})_{ik} \left(\sum_{j'=1}^{m} F_{ij'}\delta_{j'p} \right), \tag{5.95}$$

$$= -2 \sum_{i,k=1}^{N} F_{ip} (V^{-1})_{ik} \left(y_k - \sum_{j'=1}^{m} F_{kj'} a_{j'} \right). \tag{5.96}$$

On the second line, we used $\partial a_j / \partial a_p = \delta_{jp}$, in which δ_{jp} is the Kroenecker symbol:

$$\delta_{ij} = \begin{matrix} 1 & \text{for i} = \text{j}, \\ 0 & \text{for i} \neq \text{j}. \end{matrix}$$

We next took the sum $\sum_{j=1}^{m} F_{ij}\delta_{jp} = F_{ip}$ and made use of the fact that the inverse of matrix $\mathbf{V}$ is symmetric. The preceding expression is succinctly expressed in matrix form:

$$0 = \mathbf{F}^T V^{-1} (\vec{y} - \mathbf{F}\vec{a}), \tag{5.97}$$

which we rewrite as

$$\mathbf{F}^T V^{-1} \mathbf{F}\vec{a} = \mathbf{F}^T V^{-1} \vec{y}. \tag{5.98}$$

As in prior sections, it is convenient to introduce the matrix α and column vector $\vec{b}$, defined as

$$\alpha = \mathbf{F}^T V^{-1} \mathbf{F}, \tag{5.99}$$

$$\vec{b} = \mathbf{F}^T V^{-1} \vec{y}. \tag{5.100}$$

Equation (5.98) becomes

$$\alpha \vec{a} = \vec{b}. \tag{5.101}$$

Solving for $\vec{a}$, we get

$$\hat{a} = \alpha^{-1} \vec{b}. \tag{5.102}$$

https://doi.org/10.1017/9781108241922.007 Published online by Cambridge University Press

This expression provides estimates $\hat{a} = (\hat{a}_0, \ldots, \hat{a}_m)$ that are linear functions of the measurements $\vec{y}$. It can thus be computed analytically. The inversion of large matrices, however, becomes rather tedious for large values of N or m and is thus best carried out numerically on a computer. In practice, it is often most efficient or simply convenient to use numerical algorithms, such as the one described in §5.2.7. It can be shown that the foregoing estimates $\hat{a}_j$ have zero bias and minimum variance.

Errors (variances) on the parameters may be obtained using the error propagation technique introduced in Eqs. (2.222, 2.223) and used in §5.2.2. The covariance matrix U_{ij} of the fit estimators $\hat{a}_i$ and $\hat{a}_j$may be written

$$U_{ij} = \sum_{k,k'=1}^{N} \frac{\partial a_i}{\partial y_k} V_{kk'} \frac{\partial a_j}{\partial y_{k'}}. \tag{5.103}$$

In order to compute the derivatives $\partial a_i / \partial y_k$, we note that Eqs. (5.100) and (5.102) may be combined to obtain

$$\vec{a} = \left(\mathbf{F}^T V^{-1} \mathbf{F}\right)^{-1} \mathbf{F}^T V^{-1} \vec{y}. \tag{5.104}$$

We thus get

$$\frac{\partial \vec{a}}{\partial \vec{y}} = \left(\mathbf{F}^T V^{-1} \mathbf{F}\right)^{-1} \mathbf{F}^T V^{-1}. \tag{5.105}$$

The covariance matrix $\mathbf{U}$ of the estimators $\hat{a}_i$ may then be written

$$\mathbf{U} = \left(\mathbf{F}^T V^{-1} \mathbf{F}\right)^{-1} \mathbf{F}^T V^{-1} V \left[\left(\mathbf{F}^T V^{-1} \mathbf{F}\right)^{-1} \mathbf{F}^T V^{-1}\right]^T. \tag{5.106}$$

The matrix $\left(\mathbf{F}^T V^{-1} \mathbf{F}\right)$ and its inverse are by construction symmetric. Equation (5.106) thus simplifies to

$$\mathbf{U} = \left(\mathbf{F}^T V^{-1} \mathbf{F}\right)^{-1}. \tag{5.107}$$

By construction, $\mathbf{U}$ is an $m \times m$ symmetric matrix. Its diagonal elements U_{jj} correspond to the variances, $\mathrm{Var}[\hat{a}_j]$, of the estimators a_j and as such should provide estimates of the errors on each of the fit parameters. However, the nondiagonal elements U_{ij}, corresponding to covariances $\mathrm{Cov}[\hat{a}_i, \hat{a}_j]$ of the estimators a_i and a_i are in general non-null, even if the matrix $\mathbf{V}$ is itself diagonal. The errors on the parameters a_j are correlated and thus cannot be specified independently, that is, for each parameter individually.

It is instructive to consider the covariance matrix $\mathbf{U}$ in terms of second-order derivatives of the χ^2 function. Toward this end, we will calculate the second-order derivatives of the χ^2 function based on Eq. (5.96).

$$\left.\frac{\partial^2 \chi^2}{\partial a_r \partial a_s}\right|_{\vec{a}=\hat{a}} = -2 \frac{\partial}{\partial a_r} \sum_{i,i'=1}^{N} F_{is} \left(V^{-1}\right)_{ii'} \left(y_{i'} - \sum_{j'=0}^{m} F_{i'j} a_j\right), \tag{5.108}$$

$$= 2 \sum_{i,i'=1}^{N} F_{is} \left(V^{-1}\right)_{ii'} F_{i'r}, \tag{5.109}$$

$$= 2 \left(F^T V^{-1} F\right)_{sr}. \tag{5.110}$$

https://doi.org/10.1017/9781108241922.007 Published online by Cambridge University Press

The order of the derivatives is inconsequential and the matrix $F^T V^{-1} F$ is symmetric, that is, equal to its transpose:

$$\left(F^T V^{-1} F\right)^T = F^T \left(V^{-1}\right)^T F = F^T V^{-1} F, \tag{5.111}$$

since V and V^{-1} are symmetric matrices. But as per Eq. (5.107), we found that the covariance matrix $\mathbf{U}$ is equal to $\left(F^T V^{-1} F\right)^{-1}$. We thus obtain the interesting and useful result:

$$\left.\frac{\partial^2 \chi^2}{\partial a_r \partial a_s}\right|_{\vec{a}=\hat{a}} = 2\left(U^{-1}\right)_{sr} \tag{5.112}$$

In the vicinity of the solution $\hat{a}$ (i.e., near the minimum of the χ^2-function), it is legitimate to write

$$\chi^2(\vec{a}) = \chi^2(\hat{a}) + \frac{1}{2}\sum_{i,j=0}^{m} \left.\frac{\partial^2 \chi^2}{\partial a_i \partial a_j}\right|_{\vec{a}=\hat{a}} (a_i - \hat{a}_i)\left(a_j - \hat{a}_j\right) + O(3) \tag{5.113}$$

Note that the absence of first derivatives stems from the fact that the series expansion is carried out at the minimum in which first-order derivatives vanish implicitly. Substituting the expression (5.112) for the second-order derivative, we obtain

$$\chi^2(\vec{a}) = \chi^2(\hat{a}) + \sum_{i,j=0}^{m} \left(U^{-1}\right)_{ij} \delta a_i \delta a_j, \tag{5.114}$$

in which $\delta a_i = a_i - \hat{a}_i$. In order to interpret this expression, consider for illustrative purposes a case in which $\mathbf{U}$ is a diagonal with elements σ_i^2. The inverse $\mathbf{U}^{-1}$ thus has diagonal elements $1/\sigma_i^2$ and Eq. (5.114) therefore implies that the χ^2 shall increase by one unit when deviations $\delta a_i = \sigma_i$ from $\hat{a}$ are considered. This is a rather generic result. It tells us that the χ^2 increases by one unit when a fit parameter is varied away from its optimal value by one standard deviation (while all other coefficients are kept constant and equal to their value at the χ^2 minimum).

Equation (5.114) provides numerical and graphical techniques to estimate and visualize the errors on the estimators $\hat{a}_i$ as illustrated in Figure 5.6 for cases involving one- and two-parameter fits. Panel (a) displays 15 simulated measurements of a constant but noisy signal of amplitude $y = 10$ fitted with a constant polynomial $y(x) = a_0$. The fit yields a value $a_0 = 9.67 \pm 0.81$ with a minimum $\chi^2 = 19.6$ for 14 degrees of freedom. Panel (c) shows the dependence of the fit χ^2 on a_0 and displays how the errors on a_0 may be obtained by increasing the χ^2 by one unit. Panel (b) displays a noisy linear signal fitted with a first order polynomial, $y(x) = a_0 + a_1 x$. Panel (d) displays iso-contours of the fit χ^2 plotted as a function of a_0 and a_1 for values $\chi^2 = \chi^2_{\min} + 1$ and $\chi^2 = \chi^2_{\min} + 2$. The symmetric and circular aspects of the contour indicate the parameters a_0 and a_1 are essentially uncorrelated. Their errors are thus independent and shown as one standard deviation errors in panel (b) based on the $\chi^2 = \chi^2_{\min} + 1$ contour.

https://doi.org/10.1017/9781108241922.007 Published online by Cambridge University Press

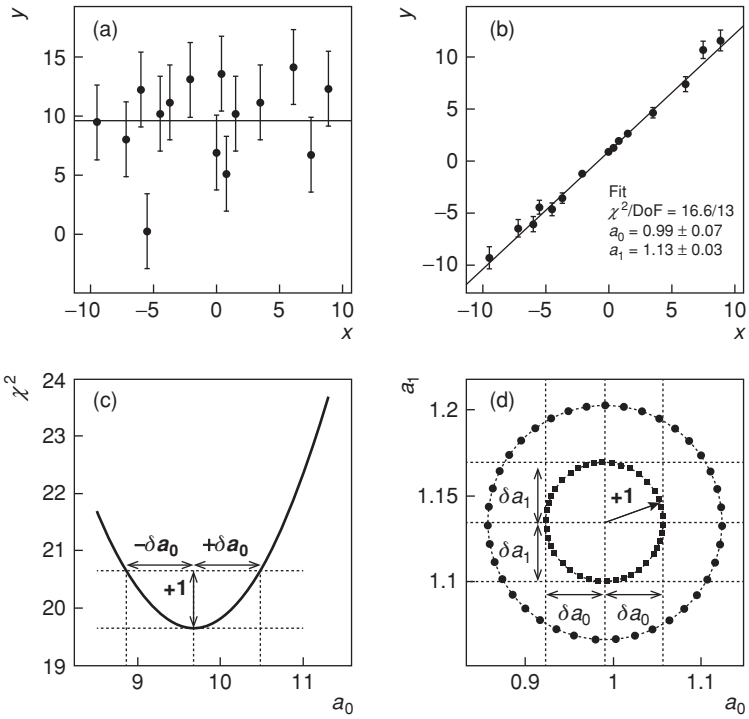

Fig. 5.6 Graphical interpretation of fit χ^2 for one- and two-parameter fits. (a) Noisy constant signal; (b) linear signal dependence; (c) χ^2 of the constant signal fit vs. the fit parameter a_0; (d) χ^2 iso-contours at $\chi^2_{\text{min}} + 1$ and $\chi^2_{\text{min}} + 2$ as a function of the fit parameters a_0 and a_1.

5.2.6 $\mathbb{LS}$ Fit with Binned Data

The $\mathbb{LS}$ method discussed in prior sections enables fits with arbitrary functions (models) for datasets containing an arbitrary number N of data points (x_i, y_i). However, it is not always possible or practical to handle all data points (x_i, y_i) individually. Often, one wishes to fit a model to data that have been binned into a histogram. We saw in §5.1.6 that the $\mathbb{ML}$ method readily enables model parameter estimation with histograms. But the $\mathbb{ML}$ method can be cumbersome and one consequently wishes to extend the $\mathbb{LS}$ method for fits of binned data also.

Consider a measurement in which data have been binned into a histogram consisting of n bins. Let H_i represent the number of entries in bin i, with $i = 1, \ldots, n$. We wish to fit the data with some function $f(x|\vec{\theta})$ determined by $m \geq 1$ parameter(s) $\vec{\theta}$, which have yet to be estimated. Let N be the total number of entries in the histogram. The function $f(x|\vec{\theta})$ is used as a model of the data. One thus expects the number of entries in bin i to be given by

$$\mu_i(\vec{\theta}) = N p_i(\vec{\theta}) = N \int_{x_{i,\text{min}}}^{x_{i,\text{max}}} f(x|\vec{\theta})\, dx, \tag{5.115}$$

https://doi.org/10.1017/9781108241922.007 Published online by Cambridge University Press

in which we have defined the probability $p_i(\vec{\theta})$ that there will be entries in bin i. The parameters $\vec{\theta}$ are determined by minimization of the χ^2 function:

$$\chi^2(\vec{\theta}) = \sum_{i=1}^{N} \frac{\left(H_i - \mu_i(\vec{\theta})\right)^2}{\sigma_i^2}, \tag{5.116}$$

in which the variance σ_i^2 of the number of entries H_i must be estimated either from the data or from the model. There are thus few options as to how to proceed with the minimization of χ^2.

The $\mathbb{LS}$ fit can be somewhat simplified if the number of bins n is very large and such that there are just a few entries in each bin. In this case, the content of each bin may be reasonably well described by Poisson distributions and the variance of the number of entries in bin i is equal to the mean number of entries "predicted" by the model, $\mu_i(\vec{\theta})$. The χ^2 function may then be written as

$$\chi^2(\vec{\theta}) = \sum_{i=1}^{N} \frac{\left(H_i - \mu_i(\vec{\theta})\right)^2}{\mu_i(\vec{\theta})} = \sum_{i=1}^{N} \frac{\left(H_i - Np_i(\vec{\theta})\right)^2}{Np_i(\vec{\theta})}. \tag{5.117}$$

The functions $p_i(\vec{\theta})$ are integrals of $f(x|\vec{\theta})$ dependent on the unknown parameters $\vec{\theta}$ and the boundaries of each bins. Minimization of the χ^2 by analytical methods, described in prior sections, thus become intractable and one must then use numerical techniques such as those presented in §5.2.7.

An alternative approach consists in approximating the variances σ_i^2 by the number of entries H_i in each bin. Such a substitution is reliable if the bin contents H_i are uncorrelated and sufficiently large to provide a reliable estimate of the fluctuations. This leads to the **Modified Least-Squares** ($\mathbb{MLS}$) fit method based on the minimization of the χ^2 function defined as

$$\chi^2(\vec{\theta}) = \sum_{i=1}^{n} \frac{\left(H_i - Np_i(\vec{\theta})\right)^2}{H_i}. \tag{5.118}$$

Again in this case, the minimization of the χ^2 involves integrals $p_i(\vec{\theta})$ and as such is best handled by numerical methods. However, note that since the denominator includes the bin content H_i, the method will fail whenever the number of entries in one or more bin is null. It may be possible to remedy this problem by rebinning, that is, grouping bins together, or by using variable size bins.

Recall that in the case of the $\mathbb{ML}$ method, a multinomial function is used to estimate the expectation value of the number of entries per bin μ_i rather than a Poisson PDF. One can show that the variance of the $\mathbb{ML}$ estimate converges faster to the minimum variance bound than the $\mathbb{LS}$ or $\mathbb{MLS}$ estimates. This implies that for fits of binned data, it is preferable to use $\mathbb{ML}$ estimators whenever feasible.

https://doi.org/10.1017/9781108241922.007 Published online by Cambridge University Press

5.2.7 Numerical $\mathbb{LS}$ Methods

We have so far focused on applications of the $\mathbb{LS}$ method for fits of linear models of the form given by Eq. (5.90). However, scientific analyses and data parameterization commonly involve nonlinear models that cannot be readily linearized. Examples of nonlinear functions commonly used include the Gaussian and Breit–Wigner distributions. It is also often desirable to add or mix elementary functions. For instance, one might wish to represent a signal and its background as a sum of a Gaussian distribution and a polynomial of order m as follows:

$$f(x|a_i, N, \mu, \sigma) = \sum_{i=0}^{m} a_i x^m + N \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right). \tag{5.119}$$

Clearly, this function's dependence on μ and σ is nonlinear, and given the addition of the "background" terms $\sum a_i x^m$ cannot be linearized by taking a logarithm of the function. Nonlinear models abound in physics, and in science in general. It is thus particularly important to consider model parameter estimation in the context of such models.

Numerical χ^2 minimization methods are a particular case of the more general case of optimization (or extremum finding) encountered both in classical and Bayesian inference problems. Rather than discussing numerical approaches piecemeal in this and following chapters, we present a systematic and comprehensive discussion of numerical techniques and algorithms in §7.6.

5.3 Determination of the Goodness-of-Fit

As we saw in §5.2.1, the $\mathbb{LS}$ method is derived from and therefore strictly equivalent to the $\mathbb{ML}$ method whenever the measurements y_i are Gaussian random variables. When this condition is met, the $\mathbb{LS}$ estimators $\vec{\theta}$ obtained by χ^2 minimization consequently coincide with those obtained by the $\mathbb{ML}$ method. Once estimators are known, and for a given set of data points, one can then seek the probability of getting a certain χ^2 value.

It is convenient to introduce **normalized deviates**, denoted $z_i(\vec{\theta})$, and defined as

$$z_i(\vec{\theta}) = \big(y_i - \mu_i(\hat{\theta})\big)/\sigma_i, \tag{5.120}$$

in which $\hat{\theta}$ are the estimator values obtained in the fit. By construction, a normalized deviate $z_i(\vec{\theta})$ measures the deviation between an observed value y_i and the value $\mu_i(\hat{\theta})$ predicted by the model for x_i, according to parameters $\hat{\theta}$, and relative to the standard deviation σ_i. As such, the normalized deviates provide a measure of the level of agreement or compatibility between the data and the model (obtained from the fit), relative to the errors σ_i. The deviates z_i should be distributed according to the standard normal distribution if the y_i are Gaussian distributed. Additionally, with known variances σ_i^2, a function $\mu_i(\hat{\theta})$ linear in the parameters $\vec{\theta}$, and with proper functional form (i.e., an appropriate representation of the data), one expects the minimum $\chi^2 = \sum_i z_i^2$ obtained by the $\mathbb{LS}$ method should be

https://doi.org/10.1017/9781108241922.007 Published online by Cambridge University Press

distributed according to the χ^2 PDF with $N - m$ degrees of freedom, as shown in §3.13.3. It is consequently appropriate to use the χ^2 value obtained in a fit of data with Gaussian deviates to evaluate the **goodness** of the fit.

Recall from §3.13 that the expectation value of a random variable with a χ^2 PDF is equal to the number of degrees of freedom, N_{DoF}. It is thus convenient (and customary) to quote the χ^2 divided by the number of degrees of freedom, N_{DoF}, as a measure of the goodness of a fit. The number of degrees of freedom N_{DoF} is equal to $N - m$, N being the number of data points or bins, and m the number of fit parameters. If the value χ^2/N_{DoF} is much smaller than one, or near zero, then the fit is much better than expected, on average, given the size of the data set and the number of fit parameters. Very small χ^2/N_{DoF} values are not impossible but have rather low probability of occurring. Fits yielding "very" small χ^2/N_{DoF} values might then signal that the errors (standard deviations), σ_i, used in the fit are overestimated or correlated. Large values of χ^2/N_{DoF}, on the other hand, indicate that the fit is very poor. This is an indication that the model $f(x|\vec{\theta})$ used to fit the data has a very small likelihood of yielding the measured data. The hypothesis that this particular model constitutes an accurate representation of the data (and the phenomenon considered) is thus regarded as having a low probability. Alternatively, it is possible that the errors σ_i are much underestimated and consequently yield a rather large χ^2.

It is customary to report the goodness of a fit in terms of its **significance level** or p-value. The significance level corresponds to the probability that the model hypothesis[4] would lead to a χ^2 value worse (i.e., larger) than that actually achieved:

$$p = \int_{\chi^2}^{\infty} p_\chi(z|N_{\mathrm{DoF}})\, dz, \tag{5.121}$$

where $p_{\chi^2}(z|N_{\mathrm{DoF}})$ is the χ^2-distribution for N_{DoF} degrees of freedom. Integrals of $p_{\chi^2}(z|N_{\mathrm{DoF}})$ are best calculated with numerical routines available in software packages such as Mathematica®, MATLAB®, or ROOT®.

It is important to realize that the choice of minimum p-value used toward the rejection of model hypotheses is rather subjective and may very well depend on the purpose of a particular measurement. See §6.6.5 for a more extensive discussion of this issue. Additionally, it is also important to acknowledge that the errors σ_i may be under- or overestimated, thereby resulting in too large or too small a value of χ^2, respectively. The use of a χ^2 test as a measure of the goodness of a fit may thus be completely unwarranted if the errors have not been properly calibrated.

5.4 Extrapolation

Having obtained estimates of model parameters with the ML, LML, LS, or related methods, it is often necessary to use the estimates to determine values predicted by the model in

[4] The notions of hypothesis and hypothesis testing are discussed at great length in Chapter 6.

https://doi.org/10.1017/9781108241922.007 Published online by Cambridge University Press

regions where no measurements exist. Doing this is fairly easy: one needs only to plug in the parameter estimates, $\hat{\theta}$, obtained from the fit into the model formula, and calculate the function values $y = f(x|\theta)$ at the relevant values of x. Somewhat less easy, however, is the calculation of error estimates δy on values $y = f(x|\theta)$ predicted by or extrapolated from the model. The model parameters are in general not independent. One must then use the covariance matrix U of the model parameters to determine the error δy on an extrapolated value according to

$$\delta y^2 = \left(\frac{\partial y}{\partial \vec{\theta}}\right)^T \mathbf{U} \left(\frac{\partial y}{\partial \vec{\theta}}\right). \tag{5.122}$$

As a practical example, let us consider the implementation of the foregoing formula for a polynomial of order m with coefficients a_i, $i = 0, \ldots, m$. We must first calculate the derivatives $\partial y/\partial a_j$:

$$\frac{\partial y}{\partial \vec{a}_j} = \sum_{k=0}^{m} \frac{\partial a_k}{\partial a_j} x^k = x^j. \tag{5.123}$$

The expression (5.122) may then be written:

$$\delta y^2 = \begin{pmatrix} 1 & x & \cdots & x^m \end{pmatrix} \begin{pmatrix} U_{00} & U_{01} & \cdots & U_{0m} \\ U_{10} & U_{11} & \cdots & U_{1m} \\ \vdots & \vdots & \ddots & \vdots \\ U_{m0} & U_{m1} & \cdots & U_{mm} \end{pmatrix} \begin{pmatrix} 1 \\ x \\ \vdots \\ x^m \end{pmatrix}. \tag{5.124}$$

For a straight-line fit, $y = a_0 + a_1 x$, this expression reduces to

$$\delta y^2 = U_{00} + 2U_{01}x + U_{11}x^2, \tag{5.125}$$

in which $U_{00} = S_{xx}/\Delta$ and $U_{11} = S/\Delta$ are the variances of a_0 and a_1, respectively, while $U_{01} = S_x/\Delta$ is the covariance of a_0 and a_1.

It is important to stress that neglect of the off-diagonal terms, which may be quite large relative to the diagonal terms, can lead to gross misrepresentations of the errors δy on extrapolated values (see Problem 5.12).

5.5 Weighted Averages

It is the hallmark of science that measurements should be reproducible and that advances in technology typically lead to improved measurements of physical quantities. It is often the case that different groups of scientists conduct distinct measurements of a given observable, for instance, the mass of a particle, the value of a fundamental constant such as the speed of light, and so on. Distinct experiments generally have different degrees of reliability, accuracy, precision, and yield different measurement results with distinct errors. Given historical trends, one might expect that more modern experiments yield more accurate and precise results. One might thus be tempted to rely only on the latest or most

https://doi.org/10.1017/9781108241922.007 Published online by Cambridge University Press

precise results. But why give up the valuable information provided by other experiments? Why not combine the information gathered to obtain a **world average** that accounts for all data available? This can be accomplished with the **weighted average** ($\mathbb{WA}$) method, which is introduced here as a special case of the $\mathbb{LS}$ method. Weighted averages may additionally be obtained with the $\mathbb{ML}$ method and within the Bayesian inference paradigm discussed in Chapter 7.

Suppose the quantity of interest, with a true value θ, has been measured N times, yielding N independent values (estimates) $\hat{\theta}_i$ and errors σ_i, with $i = 1, 2, \ldots N$. Since a single phenomenon is being considered, it is reasonable to expect all measurements should yield the same value θ. It is thus acceptable to combine them to get a better estimate of θ. This can be readily accomplished with the $\mathbb{LS}$ method by minimization of the χ^2 objective function for a model $f(x|\theta) = \theta$:

$$\chi^2(\theta) = \sum_{i=1}^{N} \frac{\left(\hat{\theta}_i - \theta\right)^2}{\sigma_i^2}. \tag{5.126}$$

We set the derivative of Eq. (5.126) with respect to θ equal to zero to seek an extremum that yields the value θ most compatible with the existing measurements $\hat{\theta}_i$:

$$\frac{d}{d\theta}\chi^2(\theta) = -2\sum_{i=1}^{N} \frac{(\theta_i - \theta)}{\sigma_i^2} = 0. \tag{5.127}$$

Solving for θ yields

$$\hat{\theta}_{\text{WA}} = \sum_{i=1}^{N} \frac{\theta_i}{\sigma_i^2} \Bigg/ \sum_{i=1}^{N} \frac{1}{\sigma_i^2}. \tag{5.128}$$

We use the subscript "WA" to indicate that the preceding estimate is equal to the sum of the estimates θ_i weighted by their respective variances and as such corresponds to a special case of a weighted average procedure defined as

$$\hat{\theta}_{\text{WA}} = \sum_{i=1}^{N} \omega_i \theta_i \Bigg/ \sum_{i=1}^{N} \omega_i, \tag{5.129}$$

with weights $w_i = 1/\sigma_i^2$.

The weights ω_i determine the importance given to estimates θ_i in the average. Measurements with a smaller variance σ_i^2 have a larger weight and thus contribute more to the weighted average $\hat{\theta}_{\text{WA}}$. Note that the factor $\sum_i w_i$ is needed for proper normalization of the weights, unless they are already normalized, in other words, if $\sum_i w_i = 1$.

The second-order derivative of the χ^2-function with respect to θ yields the variance of the estimate

$$\text{Var}\left[\hat{\theta}_{\text{WA}}\right] = \left(\sum_{i=1}^{N} \sigma_i^{-2}\right)^{-1}, \tag{5.130}$$

which amounts to the inverse of the sums of all the weights. The variance $\text{Var}[\hat{\theta}_{WA}]$ is, by construction, smaller than the individual variances σ_i^2. Consequently, there is an obvious

https://doi.org/10.1017/9781108241922.007 Published online by Cambridge University Press

advantage in combining the results of several measurements. Three special cases are of interest: first, if all measurements have equal errors, $\sigma_i = \sigma$, then the variance of the estimate simplifies to

$$\mathrm{Var}\left[\hat{\theta}_{\mathrm{WA}}\right] = \left(\sum_{i=1}^{N} \sigma^{-2}\right)^{-1} = \frac{\sigma^2}{N}, \tag{5.131}$$

and the error σ_{WA} on the estimate $\hat{\theta}_{WA}$ equals the measurement error divided by the square root of the number N of measurements:

$$\sigma_{\mathrm{WA}} = \frac{\sigma}{\sqrt{N}} \qquad \text{(equal errors)}. \tag{5.132}$$

Second, if one measurement has a much smaller error than the others, it will dominate both the mean and its variance. Third, if a measurement has a much larger error than the others, it will play a negligible role in the evaluation of the mean and its variance.

The foregoing weighted procedure can be generalized to situations in which the measurements θ_i are not independent. This would be the case, for instance, if some or all of the estimates are based in part on the same data. One must then first determine the covariance V_{ij} of the measurements. One can then verify (see Problem 5.13) that the WA is given by Eq. (5.129) with weights replaced by

$$w_j = \frac{\sum_{i=1}^{N} (V^{-1})_{ij}}{\sum_{k,m=1}^{N} (V^{-1})_{km}}. \tag{5.133}$$

Clearly, Eq. (5.133) reduces to Eq. (5.129) if the covariance matrix is diagonal, that is, if the measurements $\hat{\theta}_i$ are uncorrelated.

Averaging of experimental results may also be achieved with products of likelihood functions (as well as sums of log of likelihood functions) of combined datasets [67] and Bayesian inference techniques discussed in Chapter 7.

5.6 Kalman Filtering

Kalman filtering ($\mathbb{KF}$) is a technique that was initially designed and used for radar signal processing. It is quite general, however, and is used nowadays in many applications, including signal processing, signal fitting, pattern recognition, as well as navigation and control.

By design, a Kalman filter operates recursively on one or multiple streams of noisy input data to produce statistically optimal estimates of the underlying state of a physical system. The technique is named after Rudolf E. Kalman[5] (b. 1930), one of the early and primary developers of the theory [122]. It was introduced in high-energy physics by Billoir as a

[5] Hungarian-born American electrical engineer, mathematician, and inventor.

https://doi.org/10.1017/9781108241922.007 Published online by Cambridge University Press

progressive method of track-fitting [41]. The equivalence between progressive methods and Kalman filters was established by Fruhwirth [88].

In nuclear and high-energy physics, Kalman filtering is commonly applied toward track reconstruction in complex detectors, where it usually involves a linear, recursive method of track-fitting shown to be equivalent to a global least-squares minimization procedure (see §5.6.6). It is therefore an optimal linear estimator of track parameters. Provided the track model is truly linear and measurement errors are Gaussian, the Kalman filter is also efficient. It was formally shown that no nonlinear estimator can do better. Extensions and generalizations of the method to nonlinear systems, known as **extended Kalman filters (**$\mathbb{EKF}$**)**, have also been developed.

Kalman filters have the following attractive features that make them preferable over global least-squares methods under appropriate circumstances:

1. A Kalman filter is recursive and is thus well suited for progressive signal processing, particularly track finding and fitting in large and complex detection systems.
2. A Kalman filter can be extended into a **smoother** and thereby provides for optimal estimates of signals throughout the evolution of a system.
3. A Kalman filter readily enables efficient resolution and removal of outlier points.
4. In contrast to least-squares methods, a Kalman filter does not involve the manipulation or inversion of large matrices.

We motivate and introduce the notion of recursive fitting (filtering) in §5.6.1. The linear Kalman filter algorithm is outlined in §5.6.2. A detailed derivation of the expression of the Kalman gain matrix is presented in §5.6.5, whereas a proof of the equivalence between the Kalman filter method and the least-squares method is sketched in §5.6.6. An example of application for charged particle track reconstruction in complex detectors is presented in §9.2.2. The Kalman filtering techniques presented in this section constitute a small subset of the field of optimal estimation and control theory covered in more specialized works (e.g., see [70, 93, 168, 199]).

5.6.1 Recursive Least-Squares Fitting and Filtering

The least-squares method discussed in §5.2 is ideal for the estimation of model parameters when all data have been acquired and can be fitted all at once. However, there are applications in which a progressive and recursive knowledge of a system's or model's parameters are required. In other words, rather than waiting for the whole dataset to become available, one wishes to obtain an estimate on the basis of existing data and progressively improve the estimate as additional data are collected. Such a task is the domain of recursive least-squares filtering methods.

We saw in §4.5 that the arithmetic mean constitutes an unbiased estimator of the mean of a set of data. One can alternatively obtain the arithmetic mean as the least-squares estimator of a data model involving a zeroth-order polynomial. Indeed, for a dataset

https://doi.org/10.1017/9781108241922.007 Published online by Cambridge University Press

$\vec{x} = (x_1, x_2, \ldots, x_k)$ with measurement errors $\vec{\sigma} = (\sigma_1, \sigma_2, \ldots, \sigma_k)$, the χ^2 of a zeroth-order polynomial is

$$\chi^2 = \sum_{i=1}^{k} \frac{(x_i - a)^2}{\sigma_i^2}. \tag{5.134}$$

Setting the derivative of χ^2 with respect to a equal to zero yields the least-squares estimate:

$$\hat{a} = \frac{\sum_{i=1}^{k} x_i/\sigma_i^2}{\sum_{i=1}^{k} 1/\sigma_i^2}. \tag{5.135}$$

For the sake of simplicity, let us consider a case where all errors σ_i are equal. One writes

$$\hat{a}_k = \frac{1}{k} \sum_{i=1}^{k} x_i, \tag{5.136}$$

where $\hat{a}_k$ denotes the estimate of a obtained with k values x_i. It corresponds to the arithmetic mean of a sample of k values x_i. Adding one value to the sample, one can obviously write

$$\hat{a}_{k+1} = \frac{1}{k+1} \sum_{i=1}^{k+1} x_i. \tag{5.137}$$

It is convenient to formulate the estimate $\hat{a}_{k+1}$ in terms of the estimate $\hat{a}_k$, as follows:

$$\begin{aligned} \hat{a}_{k+1} &= \frac{1}{k+1} \left(k \frac{1}{k} \sum_{i=1}^{k} x_i + x_{k+1} \right) \\ &\quad \frac{1}{k+1} \left(k\hat{a}_k + x_{k+1} \right) \end{aligned} \tag{5.138}$$

where in the second line, we have used the expression (5.136) of the estimator $\hat{a}_k$. Defining $\hat{a}_0 = 0$ and $\hat{a}_1 = x_1$, we shift the indices of Eq. (5.138) by one unit and obtain

$$\begin{aligned} \hat{a}_k &= \frac{1}{k} \left((k-1)\hat{a}_{k-1} + x_k \right) \\ &= \hat{a}_{k-1} + \frac{1}{k} \left(x_k - \hat{a}_{k-1} \right) \end{aligned} \tag{5.139}$$

We find that the *new estimate* $\hat{a}_k$ is equal to the *prior estimate* $\hat{a}_{k-1}$, plus a "correction" proportional to the difference between the new measurement x_k and the prior estimate, known as the kth residue. The weight given to the correction is determined by the factor $1/k$, which we call the *gain* of the filter and denote $K_k^{(1)}$. We thus can write

$$\hat{a}_k = \hat{a}_{k-1} + K_k^{(1)} \left(x_k - \hat{a}_{k-1} \right), \tag{5.140}$$

with the filter gain

$$K_k^{(1)} = \frac{1}{k}. \tag{5.141}$$

Equation (5.140) provides us with a recursive formula to estimate the arithmetic mean of a growing sample of values, x_i. Setting $a_0 = 0$, for $k = 1$, one has $K_1^{(1)} = 1$, and the first

https://doi.org/10.1017/9781108241922.007 Published online by Cambridge University Press

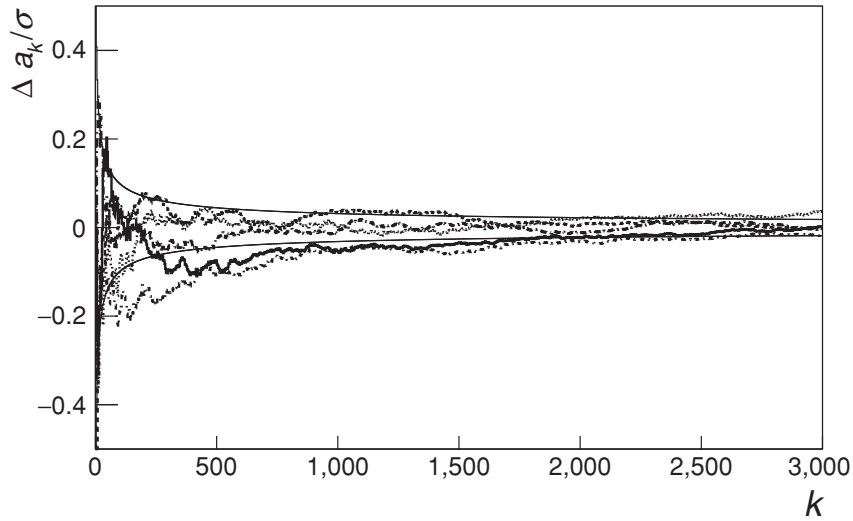

Fig. 5.7 Evolution of the error Δa_k of the estimate $\hat{a}_k$ of five constant signals with Gaussian noise of standard deviation σ. Solid lines show the 68% confidence interval $(1/\sqrt{k})$ of the sample mean for a data sample of size k.

measured value is given maximal weight

$$\hat{a}_1 = K_1^{(1)} x_1 = x_1, \tag{5.142}$$

whereas for increasing values of k, one finds the gain $K_k^{(1)}$ decreases monotonically and vanishes for $k \to \infty$: additional values are progressively given a smaller weight in the calculation of the estimate $\hat{a}$.

Equation (5.140) epitomizes the concept of recursive filtering and fitting. One starts with no information and the first measurement is given maximal weight in the determination of the first estimate $\hat{a}_1$. This and subsequent estimates serve as priors toward the recursive determination of posterior estimates, which are expected to progressively converge toward the true value of the observable. The gain K determines the importance given to new information provided by measurements x_k. It is initially large but tends to decrease as the number of sampled values progressively increases.

Figure 5.7 illustrates how estimates $\hat{a}_k$ of a constant value a progressively converge toward the true value while the gain tends to zero.

We saw in §4.5 that the estimator of the mean is given by Eq. (5.136). This can be verified also for the estimators $\hat{a}_k$ as follows:

$$\mathrm{E}[\hat{a}_k] = \frac{1}{k} \sum_{i=1}^{k} \mathrm{E}[x_i] = \frac{1}{k} \sum_{i=1}^{k} a = a. \tag{5.143}$$

In order to examine the variance of estimators $\hat{a}_k$, it is convenient to express the measurements x_k in terms of their expectation value a and a signal noise v_k:

$$x_k = a + v_k, \tag{5.144}$$

https://doi.org/10.1017/9781108241922.007 Published online by Cambridge University Press

where the term v_k represents random numbers with null expectation value, $\mathrm{E}[v_k] = 0$, and variance $\mathrm{E}\left[v_k^2\right] = \sigma^2$. One can then calculate the residue

$$a - \hat{a}_k = a - \hat{a}_{k-1} - \frac{1}{k}\left(a + v_k - \hat{a}_{k-1}\right) \tag{5.145}$$

$$= \left(a - \hat{a}_{k-1}\right)\left(1 - \frac{1}{k}\right) - \frac{1}{k}v_k \tag{5.146}$$

Squaring and taking the expectation value, we obtain the variance of the estimator $\hat{a}_k$:

$$\mathrm{Var}\left[\hat{a}_k\right] = \mathrm{E}\left[(a - \hat{a}_k)^2\right] = \left(1 - \frac{1}{k}\right)^2 \mathrm{Var}\left[\hat{a}_{k-1}\right]$$

$$- 2\frac{1}{k}\left(1 - \frac{1}{k}\right)\mathrm{E}\left[\left(a - \hat{a}_{k-1}\right)v_k\right] + \frac{1}{k^2}\mathrm{E}\left[v_k^2\right] \tag{5.147}$$

The expectation value $\mathrm{E}\left[v_k^2\right]$ is the variance σ^2 of the noise v_k while the expectation value $\mathrm{E}\left[\left(a - \hat{a}_{k-1}\right)v_k\right]$ vanishes because the noise v_k is assumed to be uncorrelated to that of prior measurements. The preceding expression thus provides the variance $\mathrm{E}\left[(a - \hat{a}_k)^2\right]$ in terms of the variance at step $k - 1$ and a second term depending on the signal noise. Defining the "covariance matrix," $S_k = \mathrm{E}\left[(a - \hat{a}_k)^2\right]$, we can then rewrite the preceding as

$$S_k = (1 - K_k)^2 S_{k-1} + K_k^2 \sigma^2, \tag{5.148}$$

which gives us an equation for the evolution of the covariance S_k of the estimate $\hat{a}_k$ in terms of prior values S_{k-1} and the variance σ^2 of the measurement noise. Note that for small values of k, the gain is near unity, $K \approx 1$, and the evolution of S_k is dominated by the signal noise, whereas for large k the gain nearly vanishes and the covariance S_k becomes approximately constant. One can verify by simple substitution that the expectation value of the covariances S_k scales as

$$S_k = \frac{\sigma^2}{k^2}. \tag{5.149}$$

The foregoing recursive formalism is readily applicable to polynomial or linear function of all orders. We consider a general extension to all linearizable models in §5.6.2.

5.6.2 The Kalman Filter Algorithm

In the framework of the Kalman filter, a physical system (e.g., a radio signal, a charged particle track traversing a magnetic spectrometer) is represented by a set of n_s parameters, called the Kalman state vector, $\vec{s} = (s_1, s_2, \ldots, s_{n_s})$, which is allowed to vary as a function of some independent variable, t. For live-feed signal processing applications, time is obviously a convenient choice of independent variable. However, other choices are also possible or appropriate depending on the specificities of physical systems under study. For instance, in the case of charged particle reconstruction inside a magnetic

https://doi.org/10.1017/9781108241922.007 Published online by Cambridge University Press

spectrometer, the independent variable may be taken as the track length or the position across the spectrometer.

The state of the system is usually regarded as dynamically evolving as a function of the independent parameter t. While primarily deterministic, the dynamic process may also involve a stochastic component, usually called **process noise**. Process noise may arise because of background processes or through the dynamical evolution of the system. For instance, a track propagating through a detector is likely to interact with materials composing the detector. Interactions may lead to energy loss and scattering. The instantaneous properties of the track (e.g., momentum and direction) are thus likely to change stochastically due to such interactions.

Kalman filtering typically involves recursive measurements of n_m dependent parameters, $\vec{m}_k = (m_1, m_2, \ldots, m_{n_m})_k$, determined by the state of the system and the properties of the measurement device. Measurements typically involve fluctuations associated with the granularity and geometry of the devices as well as, ultimately, the intrinsically stochastic nature of the measurement process. Uncertainties associated with the measuring process are commonly known as **measurement noise** (see §12.1 for a more in-depth discussion of process and measurement noises).

Measurements of the dependent parameters are achieved recursively by sampling the system's signal(s). Depending on the applications and systems considered, such sampling might be carried out repeatedly at an arbitrarily large frequency or through finitely many steps. Either way, it is usually the case that both the measurements and the state of the system are expressed as functions of the independent variable t. The frequency of the sampling process being finite by its very nature, it is convenient to discretize all variables of interests in terms of t steps measured with an arbitrary index k. Thus, $\vec{s}_k$ and $\vec{m}_k$ represent the state of the system and measured values at "step" t_k. Some applications involve recursively unlimited measurement of samples and thus have unbound values of step t_k. Spectrometers used in particle physics for measurements of charged particle momenta, however, involve finitely many detection planes and thus feature data processing with finitely many values of (time) steps t_k.

Basic Kalman filters assume that knowledge of the state at step k, noted $\hat{s}_{k|k-1}$ and known as *prior*, can be predicted based on knowledge of the state at t_{k-1} according to a linear model:

$$\vec{s}_{k|k-1} = \mathbf{F}_k \vec{s}_{k-1} + \vec{w}_k, \tag{5.150}$$

where $\mathbf{F}$ is a linear function (an $n_s \times n_s$ matrix) describing the evolution of the state vector between the two times t_{k-1} and t_k. In practical situations, the evolution of the system between two steps may be nonlinear, and one should instead write

$$\vec{s}_{k|k-1} = \phi_k(\vec{s}_{k-1}) + \vec{w}_k, \tag{5.151}$$

where $\phi_k(\vec{s}_{k-1})$ is a nonlinear function of the state vector at step k. The principle of the method remains the same, however, and leads to extended Kalman filters (EKFs) (see §5.6.4).

The vector $\vec{w}_k$ corresponds to process noise and amounts to stochastic variations of the signal (state) associated with background processes accumulated during the interval

https://doi.org/10.1017/9781108241922.007 Published online by Cambridge University Press

t_k–t_{k-1}. One assumes the process noise to be unbiased, that is, such that

$$\mathrm{E}\left[\vec{w}_k\right] = 0, \tag{5.152}$$

and characterized by a predictable $n_s \times n_s$ covariance matrix $\mathbf{W}_k$ with elements

$$(\mathbf{W}_k)_{ij} = \mathrm{Cov}\left[(w_k)_i, (w_k)_j\right], \tag{5.153}$$

where $(w_k)_i$ and $(w_k)_j$ are noise values of state parameters $(s_k)_i$ and $(s_k)_j$, $i, j = 1, \ldots, n_s$ at step k.

Uncertainties of the components of the state vector $\vec{s_k}$ are likewise described by an $n_s \times n_s$ covariance matrix denoted $\mathbf{S}_k$[6]:

$$(\mathbf{S}_k)_{ij} = \mathrm{Cov}\left[(s_k)_i, (s_k)_j\right]. \tag{5.154}$$

One represents measurements performed at step k (layer k) with a vector noted $\vec{m}_k$. The dimensionality n_m of $\vec{m}_k$ is smaller or equal to that of the state vector ($n_m \leq n_s$) and is generally limited to just a few parameters, that is, $n_m \ll n_s$. For instance, in a magnetic spectrometer, straw tube chambers would provide a single measurement of position, u or v, yielding $n_m = 1$, while pad chambers or continuous devices such as Time Projection Chambers would yield measurements of two coordinates, y and z, and be represented by a two-element vector, $\vec{z} = (y, z)$. One further assumes that it is possible, given an estimate of the Kalman state, $\vec{s_k}$, to project (predict) a measurement by means of a linear function, $\mathbf{H}_k$, according to

$$\vec{m}_k = \mathbf{H}_k \vec{s_k} + \vec{v}_k. \tag{5.155}$$

The function H_k represents an $n_m \times n_s$ matrix that projects the vector $\vec{s_k}$ onto the measurement coordinates $\vec{m}_k$. It is often possible and convenient to choose some parameters of the state model $\vec{s_k}$ to be values of the measurements $\vec{m}_k$. The matrix H_k thus simplifies trivially. For instance, for $n_m = 2$ and $n_s = 6$, one could write

$$\mathbf{H}_k = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \end{bmatrix}. \tag{5.156}$$

However, such a simplification is not always possible or desirable. There are also cases where a linear function is not available.

The vector $\vec{v}$ represents the measurement noise associated with the determination of the measurement vector $\vec{m}$. The measurement noise is assumed to be unbiased, in other words, with null expectation value such that

$$\mathrm{E}\left[\vec{v}_k\right] = 0. \tag{5.157}$$

It is characterized by a known error covariance matrix, $\mathbf{V}_k$ defined according to:

$$(\mathbf{V}_k)_{ij} = \mathrm{Cov}\left[(v_k)_i, (v_k)_j\right]. \tag{5.158}$$

[6] In this and following sections, we use lowercase letters for the physical quantities and corresponding capital letters for their respective covariance matrix: $s \to S$; $w \to W$; $v \to V$

https://doi.org/10.1017/9781108241922.007 Published online by Cambridge University Press

It is generally further assumed that the process and measurement noises are strictly independent, and that successive measurement noises are also uncorrelated.

$$\mathrm{Cov}\left[(s_k)_i, (v_k)_j\right] = 0 \tag{5.159}$$

$$\mathrm{Cov}\left[(v_k)_i, (v_{k'})_j\right] = 0 \ \text{ for } \ k \neq k' \tag{5.160}$$

The Kalman filter algorithm requires an initial estimate of the system state $\vec{s_0}$ at t_0. This and subsequent estimates of the state vector at t_{k-1}, noted $\vec{s}_{k-1|k-1}$ with $k > 1$, are used to predict the state of the system at the next measurement step t_k.

$$\vec{s}_{k|k-1} = \mathbf{F}_k \vec{s}_{k-1|k-1}. \tag{5.161}$$

One also predicts (projects) the covariance matrix of the state vector as

$$\mathbf{S}_{k|k-1} = \mathbf{F}_k \mathbf{S}_{k-1|k-1} \left(\mathbf{F}_k\right)^T + \mathbf{W}_k. \tag{5.162}$$

The measurement $\vec{m}_k$ is then used to update and improve the knowledge of the Kalman state with

$$\vec{s_k} \equiv \vec{s}_{k|k} = \vec{s}_{k|k-1} + \mathbf{K}_k \left(\vec{m}_k - \mathbf{H}_k \vec{s}_{k|k-1}\right) \tag{5.163}$$

where the quantity $\mathbf{K}_k$ is called the **Kalman gain matrix**. It can be calculated as follows (for a derivation of this result, see §5.6.5):

$$\mathbf{K}_k = \mathbf{S}_{k|k-1} \left(\mathbf{H}_k\right)^T \left(\mathbf{V}_k + \mathbf{H}_k \mathbf{S}_{k|k-1} \left(\mathbf{H}_k\right)^T\right)^{-1}. \tag{5.164}$$

The matrix inversion involved in Eq. (5.164) is typically rather simple because the measurement error covariance matrix has a small dimensionality. In some cases, when $\mathbf{H}_k$ is diagonal, the inversion may even become trivial.

To operate the Kalman filter, one first initializes the covariance matrix $\mathbf{S}_{0|0}$ with large diagonal values and null off-diagonal elements. As the filter progresses from step to step, more information about the system is acquired by added measurements $\vec{m}_k$. The diagonal elements reduce to values representative of the uncertainty on the system parameters. Initially, $\mathbf{S}_{k|k-1}$ dominates the factor $\left(\mathbf{V}_k + \mathbf{H}_k \mathbf{S}_{k|k-1} \left(\mathbf{H}_k\right)^T\right)^{-1}$ so the gain $\mathbf{K}_k$ is near unity. As the number of sampled measurements increases, this factor becomes increasingly dominated by the covariant matrix $\mathbf{V}_k$, and the Kalman $\mathbf{K}_k$ gain becomes progressively smaller. With a large Kalman gain, the addition of a new measurement has a significant impact on the updated system parameters. As the gain reduces, the addition of new measurements has a progressively smaller impact on the updated state of the system.

The filtered (updated) covariance is given by

$$\mathbf{S}_{k|k} = \left(\mathbf{I} - \mathbf{K}_k \mathbf{H}_k\right) \mathbf{S}_{k|k-1}, \tag{5.165}$$

where $\mathbf{I}$ denotes an $n_s \times n_s$ identity matrix. Again, one finds that initially, the Kalman gain being large, the covariance matrix rapidly decreases in magnitude. As more information is added, the gain diminishes and added measurements have diminishing impact on the state covariance.

The recursive operation of the filter is schematically illustrated in Figure 5.9. An example of an application for charged particle track reconstruction in a complex detector is schematically illustrated in Figure 5.8 and discussed in detail in §9.2.3.

https://doi.org/10.1017/9781108241922.007 Published online by Cambridge University Press

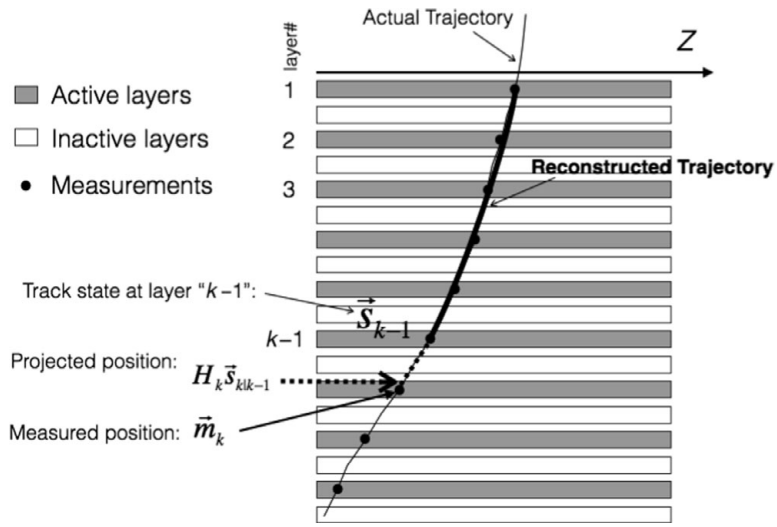

Fig. 5.8 A particle detector may be represented as a succession of material volumes or layers. Measurements $\vec{m}_k$ are carried out in active layers whereas passive volumes contribute no information to the knowledge of the track and may in fact produce degradation of information through stochastic processes such as differential energy loss and multiple Coulomb scattering. The state vector is therefore defined at finitely many layers only. Starting at some base layer *i*, one proceeds iteratively to predict and measure the state at successive layers. Given the knowledge of a track state $\vec{s}_{k-1}$ at layer $k-1$, one predicts its state $\vec{s}_k$ at layer *k* using a linear function. The measurement $\vec{m}_k$ is then used to update and improve the knowledge of the state of the track. The process is repeated iteratively until all (relevant) layers have been traversed.

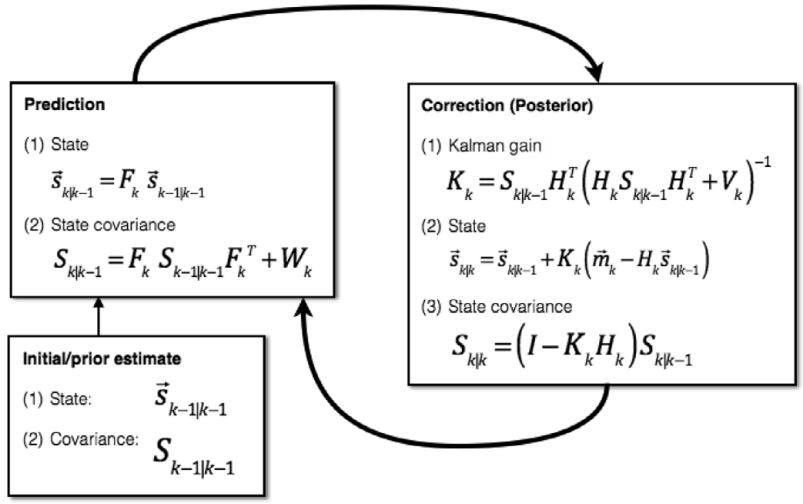

Fig. 5.9 Schematic illustration of the components of the Kalman filter algorithm.

https://doi.org/10.1017/9781108241922.007 Published online by Cambridge University Press

5.6.3 Kalman "Smoother"

In cases where it might be useful to have optimal information on the system at all steps t_k, one can carry out a "smoothing" pass on the data. The smoothing pass begins with the first step $k = n$ and proceeds recursively backward from the last measurement $k = n$. The smoothed state vector at t_k is based on all n measurements steps and is calculated as follows:

$$\vec{s}_{k|n} = \vec{s}_{k|k} + \mathbf{A}_k \left(\vec{s}_{k+1|n} - \vec{s}_{k|n} \right), \tag{5.166}$$

with the smoother gain matrix

$$\mathbf{A}_k = \mathbf{S}_{k|k} \left(\mathbf{F}_k \right)^T + \left(\mathbf{S}_{k+1|k} \right)^{-1}. \tag{5.167}$$

The covariance matrix of the smoothed state vector is

$$\mathbf{S}_{k|n} = \mathbf{S}_{k|k} + \mathbf{A}_k \left(\mathbf{S}_{k+1|n} - \mathbf{S}_{k+1|k} \right) \left(\mathbf{A}_k \right)^T. \tag{5.168}$$

Multiple extensions and variants of the foregoing algorithm are documented in the scientific and engineering literature.

5.6.4 The Extended Kalman Filter

The propagation of the state vector $\vec{s}$ with Eq. (5.150) assumes the evolution of the system may be described with a linear function of the state parameters. This is a rather limiting assumption and, in practice, one often deals with nonlinear state evolution equations such as

$$\vec{s}_k = \vec{f}_k \left(\vec{s}_{k-1}, \vec{w}_{k-1} \right), \tag{5.169}$$

as well as nonlinear state-to-measurement "projection" equations such as

$$\vec{m}_k = \vec{h}_k (\vec{s}_k, \vec{v}_k). \tag{5.170}$$

It may be possible, however, to linearize Eqs. (5.169) and (5.170) and obtain an **extended Kalman filter** ($\mathbb{EKF}$).

Let us define matrices $\mathbf{A}_k$ and $\mathbf{B}_k$ as derivatives of the functions $(\vec{f}_k \left(\vec{s}_{k-1}, \vec{w}_{k-1} \right))_i$, $i = 1, \ldots, n_s$, with respect to j components ($j = 1, \ldots, n_s$) of the state vector $\vec{s}_k$ and process noise $\vec{w}_k$, respectively:

$$(\mathbf{A}_k)_{ij} = \frac{\partial (\vec{f}_k)_i}{\partial (\vec{s}_k)_j} \left(\vec{s}_{k-1}, 0 \right), \tag{5.171}$$

$$(\mathbf{B}_k)_{ij} = \frac{\partial (\vec{f}_k)_i}{\partial (\vec{w}_k)_j} \left(\vec{s}_{k-1}, 0 \right). \tag{5.172}$$

Let us additionally define matrices $\mathbf{P}_k$ and $\mathbf{Q}_k$ as derivatives of the functions $(\vec{h}_k \left(\vec{s}_{k-1}, \vec{v}_{k-1} \right))_i$, $i = 1, \ldots, n_m$, with respect to j components of the state vector $\vec{s}_k$ and

https://doi.org/10.1017/9781108241922.007 Published online by Cambridge University Press

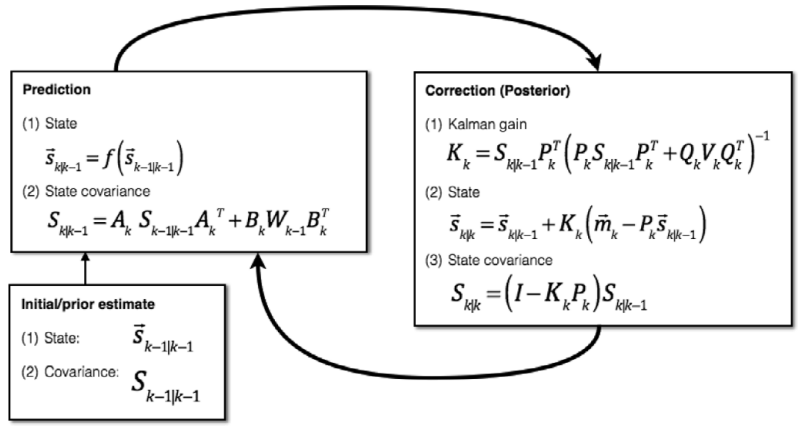

Fig. 5.10 Schematic illustration of the extended Kalman filter algorithm. The algorithm is essentially identical to the discrete algorithm, but the matrices are here replaced by nonlinear functions for the evolution of the state $\vec{s}_k$ and derivatives of these functions for the evolution of the covariance matrix S_k.

measurement noise $\vec{v}_k$, respectively:

$$(\mathbf{P}_k)_{ij} = \frac{\partial(\vec{h}_k)_i}{\partial(\vec{s}_k)_j}(\vec{s}_{k-1}, 0)\,, \tag{5.173}$$

$$(\mathbf{Q}_k)_{ij} = \frac{\partial(\vec{h}_k)_i}{\partial(\vec{v}_k)_j}(\vec{s}_{k-1}, 0)\,. \tag{5.174}$$

The state and state covariance evolution equations may then be written

$$\vec{s}_{k|k-1} = \vec{f}\,(\vec{s}_{k-1}, 0)\,, \tag{5.175}$$

$$\mathbf{S}_{k|k-1} = \mathbf{A}_k \mathbf{S}_{k-1} \mathbf{A}_k^T + \mathbf{B}_k \mathbf{W}_{k-1} \mathbf{B}_k^T\,, \tag{5.176}$$

while the measurement update equations are

$$\mathbf{K}_k = \mathbf{S}_{k|k-1} \mathbf{P}_k^T \left(\mathbf{P}_k \mathbf{S}_{k|k-1} \mathbf{P}_k^T + \mathbf{Q}_k \mathbf{V}_k \mathbf{Q}_k^T\right)^{-1}\,, \tag{5.177}$$

$$\vec{s}_k = \vec{s}_{k|k-1} + \mathbf{K}_k \left(\vec{m} - \vec{h}(s_{k|k-1}, 0)\right)\,, \tag{5.178}$$

$$\mathbf{S}_k = (\mathbf{I} - \mathbf{K_k P_k})\, \mathbf{S}_{k|k-1}\,. \tag{5.179}$$

As illustrated in Figure 5.10, the extended Kalman filter is quite similar to the regular discrete Kalman filter and proceeds recursively through steps of state update and measurement update based on Eqs. (5.175) and (5.177), respectively.

However, it is important to note, in closing this section, that the linearization of the evolution equations implies that the random variables are no longer Gaussian distributed after undergoing their respective nonlinear transformations.

https://doi.org/10.1017/9781108241922.007 Published online by Cambridge University Press

5.6.5 Derivation of the Kalman Gain Matrix

In this section, we derive the expression (5.164) of the Kalman gain $\mathbf{K}_k$. It may be skipped in a first reading of the material.

Let us assume the state of a system can be represented by an n-elements state vector $\hat{s}_k$ and that its evolution can be described by the linear equation

$$\hat{s}_k = \mathbf{F}_k \hat{s}_{k-1} + \mathbf{K}_k \left(\vec{m}_k - \mathbf{H}_k \mathbf{F}_k \hat{s}_{k-1}\right), \tag{5.180}$$

where the matrix $\mathbf{F}_k$ determines the evolution of the system in the absence of noise, $\hat{s}_{k-1}$ is the prior information on the state of the system, K_k the Kalman gain matrix we wish to determine, $\mathbf{H}_k$ a matrix that projects the state $\hat{s}_k$ onto a prediction of a measurement, and $\vec{m}_k$ represents an actual measurement. A measurement $\vec{m}_k$ of the system may be represented as $\vec{m}_k = \mathbf{H}_k \vec{s}_k + \vec{v}_k$, where $\vec{s}_k$ is the actual value of the system's state at step k and $\vec{v}_k$ a random vector representing the process noise incurred in the evolution of the state from step $k-1$ to step k. We will assume that the state, process noise, and measurement noise are uncorrelated:

$$\mathrm{Cov}\left[\Delta\vec{s}_{k-1}, \vec{w}_k^T\right] = 0, \tag{5.181}$$

$$\mathrm{Cov}\left[\Delta\vec{s}_{k-1}, \vec{v}_k^T\right] = 0, \tag{5.182}$$

$$\mathrm{Cov}\left[\vec{w}_k, \vec{v}_k^T\right] = 0. \tag{5.183}$$

Our goal is to calculate the covariance matrix $\mathbf{S}_k$ of the system state $\vec{s}_k$ and determine the gain matrix $\mathbf{K}$ that simultaneously minimizes all elements of this covariance matrix. In order to obtain the covariance matrix, let us first calculate the residue $\Delta\vec{s}_k$ at each step k of filtering as follows:

$$\begin{aligned}\Delta\vec{s}_k &= \vec{s}_k - \hat{s}_k \\ &= \vec{s}_k - \mathbf{F}_k \hat{s}_{k-1} - \mathbf{K}_k \left(\vec{m}_k - \mathbf{H}_k \mathbf{F}_k \hat{s}_{k-1}\right).\end{aligned} \tag{5.184}$$

We next replace the value of the measurement $\vec{m}_k$ by the sum of the projection of the state $\vec{s}_k$ and measurement noise $\vec{v}_k$:

$$\Delta\vec{s}_k = \vec{s}_k - \mathbf{F}_k \hat{s}_{k-1} - \mathbf{K}_k \left(\mathbf{H}_k \vec{s}_k + \vec{v}_k - \mathbf{H}_k \mathbf{F}_k \hat{s}_{k-1}\right). \tag{5.185}$$

We next also replace $\vec{s}_k$ on the righthand side by its value in terms of the previous state and process noise:

$$\begin{aligned}\Delta\vec{s}_k = {} & \mathbf{F}_k \vec{s}_{k-1} + \vec{w}_k - \mathbf{F}_k \hat{s}_{k-1} \\ & - \mathbf{K}_k \left(\mathbf{H}_k \left(\mathbf{F}_k \vec{s}_{k-1} + \vec{w}_k\right) + \vec{v}_k - \mathbf{H}_k \mathbf{F}_k \hat{s}_{k-1}\right).\end{aligned} \tag{5.186}$$

Noting that the difference $\vec{s}_{k-1} - \hat{s}_{k-1}$ corresponds to the residue at step $k-1$, we obtain after a simple reorganization of Eq. (5.186)

$$\Delta\vec{s}_k = \left(1 - \mathbf{K}_k \mathbf{H}_k\right) \mathbf{F}_k \Delta\vec{s}_{k-1} + \left(1 - \mathbf{K}_k \mathbf{H}_k\right) \vec{w}_k - \mathbf{K}_k \vec{v}_k, \tag{5.187}$$

https://doi.org/10.1017/9781108241922.007 Published online by Cambridge University Press

which expresses the residue $\Delta\vec{s}_k$ in terms of the residue $\Delta\vec{s}_{k-1}$, plus two terms that account for the process and measurement noises.

We define the state vector covariance matrix $\mathbf{S}_k$ as

$$\mathbf{S}_k = \mathrm{E}\left[\Delta\vec{s}_k \Delta\vec{s}_k^T\right], \tag{5.188}$$

where $\Delta\vec{s}_k$ is represented as a column vector while $\Delta\vec{s}_k^T$ corresponds to its transpose, a row vector. The covariance of the process noise and measurement noise are noted $\mathbf{W}_k$ and $\mathbf{V}_k$, respectively:

$$\mathbf{W}_k = \mathrm{E}\left[\Delta\vec{w}_k \Delta\vec{w}_k^T\right], \tag{5.189}$$

$$\mathbf{V}_k = \mathrm{E}\left[\Delta\vec{v}_k \Delta\vec{v}_k^T\right]. \tag{5.190}$$

Substituting the expression (5.187) in the definition (5.188) of the covariance $\mathbf{S}_k$, we obtain after some algebraic manipulations

$$\begin{aligned}\mathbf{S}_k = {} & \left((1 - \mathbf{K}_k\mathbf{H}_k)\,\mathbf{F_k}\right)\mathrm{E}\left[\Delta s_{k-1}\Delta s_{k-1}^T\right]\left((1 - \mathbf{K}_k\mathbf{H}_k)\,\mathbf{F_k}\right)^T \\ & + (1 - \mathbf{K}_k\mathbf{H}_k)\,\mathrm{E}\left[w_k w_k^T\right](1 - \mathbf{K}_k\mathbf{H}_k)^T \\ & + \mathbf{K}_k\mathrm{E}\left[\vec{v}_k\vec{v}_k^T\right]\mathbf{K}_k^T + \text{cross terms}\end{aligned}$$

where the "cross terms" are proportional to expectation values $\mathrm{E}[\vec{\Delta}s_{k-1}\vec{w}_k]$, $\mathrm{E}[\vec{\Delta}s_{k-1}\vec{v}_k]$, and $\mathrm{E}[\vec{w}_k\vec{v}_k^T]$ and thus null by hypothesis, while the factors $\mathrm{E}[\Delta s_{k-1}\Delta s_{k-1}^T]$, $\mathrm{E}[w_k w_k^T]$, and $\mathrm{E}[\vec{v}_k\vec{v}_k^T]$ correspond to $\mathbf{S}_{k-1}$, $\mathbf{W}_k$, and $\mathbf{V}_k$, respectively. We thus obtain

$$\begin{aligned}\mathbf{S}_k = {} & (1 - \mathbf{K}_k\mathbf{H}_k)\,\mathbf{F_k}\mathbf{S}_{k-1}\left((1 - \mathbf{K}_k\mathbf{H}_k)\,\mathbf{F_k}\right)^T \\ & + (1 - \mathbf{K}_k\mathbf{H}_k)\,\mathbf{W}_k\,(1 - \mathbf{K}_k\mathbf{H}_k)^T + \mathbf{K}_k\mathbf{V}_k\mathbf{K}_k^T.\end{aligned} \tag{5.191}$$

Noting that the transpose of a product of matrices $(\mathbf{AB})^T$ equals the product of their transposes in reverse order, $\mathbf{B}^T\mathbf{A}^T$, we define the matrix $\mathbf{S}_{k|k-1}$ as

$$\mathbf{S}_{k|k-1} = \mathbf{F}_k\mathbf{S}_{k-1}\mathbf{F}_k^T + \mathbf{W}_k. \tag{5.192}$$

The foregoing expression for the covariance matrix $\mathbf{S}_k$ may thus be written

$$\mathbf{S}_k = (1 - \mathbf{K}_k\mathbf{H}_k)\,\mathbf{S}_{k|k-1}\,(1 - \mathbf{K}_k\mathbf{H}_k)^T + \mathbf{K}_k\mathbf{V}_k\mathbf{K}_k^T, \tag{5.193}$$

which provides us with a rather complicated formula for the evolution of the state covariance matrix. We will see in the following that this expression greatly simplifies. But first, let us find the value of the Kalman gains $\mathbf{K}_k$ that minimize the covariance $\mathbf{S}_k$. This is readily accomplished by setting derivatives of $\mathbf{S}_k$ with respect to $\mathbf{K}_k$ equal to zero and solving for $\mathbf{K}_k$. Noting that $\mathbf{S}_k$, $\mathbf{S}_{k|k-1}$, and $\mathbf{V}_k$ are symmetric matrices by construction, the derivatives of $\mathbf{S}_k$ readily simplify to

$$0 = \frac{\partial\mathbf{S}_k}{\partial\mathbf{K}_k} = -2\,(1 - \mathbf{K}_k\mathbf{H}_k)\,\mathbf{S}_{k|k-1}\mathbf{H}_k^T + 2\mathbf{K}_k\mathbf{V}_k \tag{5.194}$$

which further simplifies to

https://doi.org/10.1017/9781108241922.007 Published online by Cambridge University Press

$$0 = -\mathbf{S}_{k|k-1}\mathbf{H}_k^T + \mathbf{K}_k\left(\mathbf{H}_k\mathbf{S}_{k|k-1}\mathbf{H}_k^T + \mathbf{V}_k\right) \tag{5.195}$$

The optimal Kalman gain matrix is thus

$$\mathbf{K}_k = \mathbf{S}_{k|k-1}\mathbf{H}_k^T\left(\mathbf{H}_k\mathbf{S}_{k|k-1}\mathbf{H}_k^T + \mathbf{V}_k\right)^{-1}. \tag{5.196}$$

Inserting the foregoing expression in Eq. (5.193), we find after some simple manipulations

$$\mathbf{S}_k = (1 - \mathbf{K}_k\mathbf{H}_k)\,\mathbf{S}_{k|k-1}, \tag{5.197}$$

or substituting the definition (5.192) of $\mathbf{S}_{k|k-1}$, we obtain

$$\mathbf{S}_k = (1 - \mathbf{K}_k\mathbf{H}_k)\left(\mathbf{F}_k\mathbf{S}_{k-1}\mathbf{F}_k^T + \mathbf{W}_k\right). \tag{5.198}$$

The covariance $\mathbf{S}_k$ is determined by the linear projection of the prior $\mathbf{S}_{k-1}$ and the process noise determined by the Kalman gain $\mathbf{K}_k$. The gain is initially large and gives more weight to measurements $\vec{m}_k$. It progressively decreases, however, with the addition of data and eventually vanishes for large values of k. The covariance matrix thus tends toward

$$\mathbf{S}_k = \mathbf{F}_k\mathbf{S}_{k-1}\mathbf{F}_k^T + \mathbf{W}_k \tag{5.199}$$

in the limit where the gain vanishes.

5.6.6 Kalman Filter as a Least-Squares Fitter

We demonstrate, in this section, that the Kalman filter technique is equivalent to the least-squares method.

The χ^2 function is defined on the basis of the measurements $\vec{m}_i$, the measurement covariance matrix $\mathbf{V}_i$, and a model $h(t;\vec{s})$ with state parameters $\vec{s}$ to be determined by the fitting procedure:

$$\chi^2 = \sum_{i=1}^{k}\left(\vec{m}_i - \vec{h}(t_i,\vec{s}_i)\right)\mathbf{V}_i^{-1}\left(\vec{m}_i - \vec{h}(t_i,\vec{s}_i)\right)^T \tag{5.200}$$

Rather than trying to evaluate Eq. (5.200) for all values of i, let us consider the contribution of the "last" measurement k and those of the $k-1$ prior measurements. Let χ_k^2 represent the contribution of the last measured point:

$$\chi_k^2 = \left(\vec{m}_k - \vec{h}(t_k,\vec{s}_k)\right)\mathbf{V}_k^{-1}\left(\vec{m}_k - \vec{h}(t_k,\vec{s}_k)\right)^T, \tag{5.201}$$

where $\vec{m}_k$ and $\mathbf{V}_k$ stand for the kth measurement and its covariance matrix, while $\vec{h}(t_k,\vec{s}_k)$ represents the expected measurement value for step t_k and model parameters $\vec{s}_k$. The

https://doi.org/10.1017/9781108241922.007 Published online by Cambridge University Press

contribution of all $k-1$ prior measurements is encapsulated in the covariance matrix $\mathbf{S}_{k-1}$ of the state. One can thus write

$$\chi^2_{k-1} = (\vec{s}_{k-1} - \hat{s}_{k-1})^T \, \mathbf{S}'^{-1}_{k-1} \, (\vec{s}_{k-1} - \hat{s}_{k-1}) \, . \tag{5.202}$$

The total χ^2 is then simply the sum $\chi^2_{k-1} + \chi^2_k$, which one seeks to minimize in order to determine the optimal state (model) parameters $\vec{s}$. We thus take the derivative of χ^2 relative to $\vec{s}$:

$$\frac{d\chi^2}{d\vec{s}} = 2\mathbf{S}'^{-1}_k \, (\vec{s}_{k-1} - \hat{s}_{k-1}) - \nabla_s \vec{h} \mathbf{V}_k^{-1} \left(\vec{m}_k - \vec{h}(t_k, \vec{s}_k) \right) . \tag{5.203}$$

Equation (5.203) has a dependency on the unknown parameters $\vec{s}_k$. We thus replace $\vec{s}_k$ by $\hat{s}_k + \Delta\vec{s}_k$ with $\Delta\vec{s}_k$ defined as the residue:

$$\Delta\vec{s}_k = \vec{s}_k - \hat{s}_k \tag{5.204}$$

For small residues, one can expand $\vec{h}(t_k, \vec{s}_k)$ as a Taylor series

$$\vec{h}(t_k, \hat{x}_k + \Delta\vec{s}_k) = \vec{h}(t_k, \hat{s}_k) + \Delta\vec{s}_k \nabla_s \vec{h}(t_k, \hat{s}_k), \tag{5.205}$$

and obtain

$$\begin{aligned} \frac{d\chi^2}{d\vec{s}} &= 2\mathbf{S}'^{-1}_k \Delta\vec{s}_{k-1} \\ &\quad - \nabla_s \vec{h}(\hat{s}) \mathbf{V}_k^{-1} \left(\vec{m}_k - \vec{h}(\hat{s}_k) - \Delta\vec{s}_k \nabla_s \vec{h}(\hat{s}) \right) , \end{aligned} \tag{5.206}$$

where for brevity we have omitted the dependence on t_k. The first $\nabla_s \vec{h}$ in Eq. (5.206) is a function of the true value $\vec{s}_k$ but it should be legitimate to use a gradient evaluated at $\hat{s}_k$ instead. Defining $\mathbf{H}_k = \nabla_s \vec{h}$, we thus obtain

$$\frac{d\chi^2}{d\vec{s}} = 2\mathbf{S}'^{-1}_k \Delta\vec{s}_{k-1} \tag{5.207}$$

$$+ \mathbf{H}_k^T \mathbf{V}_k^{-1} \mathbf{H}_k \Delta\vec{s}_k - 2\mathbf{H}_k^T \mathbf{V}_k^{-1} \left[\vec{m}_k - \vec{h}(\hat{s}_k) \right] . \tag{5.208}$$

Setting this derivative to zero and solving for $\Delta\vec{s}_k$, one gets

$$\Delta\vec{s}_k = \left[\mathbf{S}'^{-1}_k + \mathbf{H}_k^T \mathbf{V}_k^{-1} \mathbf{H}_k \right]^{-1} \mathbf{H}_k^T \mathbf{V}_k^{-1} \left[m_k - \vec{h}(\hat{s}) \right] \tag{5.209}$$

which one readily rewrites

$$\vec{s}_k = \hat{s}_k + \left[\mathbf{S}'^{-1}_k + \mathbf{H}_k^T \mathbf{V}_k^{-1} H \right]^{-1} \mathbf{H}_k^T \mathbf{V}_k^{-1} \left[m_k - \vec{h}(\hat{s}) \right] \tag{5.210}$$

to obtain an expression of the form of Eq. (5.180) but with a seemingly different gain matrix

$$\mathbf{K}_k = \left[\mathbf{S}'^{-1}_k + \mathbf{H}_k^T \mathbf{V}_k^{-1} H \right]^{-1} \mathbf{H}_k^T \mathbf{V}_k^{-1} . \tag{5.211}$$

https://doi.org/10.1017/9781108241922.007 Published online by Cambridge University Press

In order to demonstrate this expression is equivalent to Eq. (5.164), we must first show that the inverse of the updated covariance matrix $\mathbf{S}_k$ may be written as

$$\mathbf{S}_k^{-1} = \mathbf{S'}_k^{-1} + \mathbf{H}_k \mathbf{V}_k^{-1} \mathbf{H}_k^T. \tag{5.212}$$

To verify this statement, it suffices to demonstrate that $\mathbf{S}_k \mathbf{S}_k^{-1} = I$ using Eq. (5.197) for $\mathbf{S}_k$ and the preceding expressions for $\mathbf{S}_k^{-1}$ and $\mathbf{K}$.

$$\begin{aligned}
\mathbf{S}_k \mathbf{S}_k^{-1} &= (1 - \mathbf{K}_k \mathbf{H}_k) \mathbf{S'}_k \left(\mathbf{S'}_k^{-1} + \mathbf{H}_k \mathbf{V}_k^{-1} \mathbf{H}_k^T \right) \\
&= 1 + \mathbf{S'}_k \mathbf{V}_k^{-1} \mathbf{H}_k^T - \mathbf{S'}_k \mathbf{H}_k^T \left(\mathbf{H}_k \mathbf{S'}_k \mathbf{H}_k^T + \mathbf{V}_k \right)^{-1} \mathbf{H}_k \mathbf{S'}_k \mathbf{S'}_k^{-1} \\
&\quad - \mathbf{S'}_k \mathbf{H}_k^T \left(\mathbf{H}_k \mathbf{S'}_k \mathbf{H}_k^T + \mathbf{V}_k \right)^{-1} \mathbf{H}_k \mathbf{S'}_k \mathbf{H}_k \mathbf{V}_k^{-1} \mathbf{H}_k^T,
\end{aligned}$$

which can be shown to indeed simplify to unity after a modest amount of matrix algebra (see Problem 9.1). We can then use Eq. (5.212) for $\mathbf{S}_k$ to obtain an alternative expression of the gain matrix. Starting from Eq. (5.211), we write

$$\mathbf{K}_k = \mathbf{S'}_k \mathbf{H}_k^T \left(\mathbf{H}_k \mathbf{S'}_k \mathbf{H}_k^T + \mathbf{V}_k \right)^{-1} \tag{5.213}$$

and insert $\mathbf{S}_k \mathbf{S}_k^{-1}$ and $\mathbf{V}_k^{-1} \mathbf{V}_k$ judiciously in Eq. (5.213):

$$\begin{aligned}
\mathbf{K}_k &= \mathbf{S}_k \mathbf{S}_k^{-1} \mathbf{S'}_k \mathbf{H}_k^T \mathbf{V}_k^{-1} \mathbf{V}_k \left(\mathbf{H}_k \mathbf{S'}_k \mathbf{H}_k^T + \mathbf{V}_k \right)^{-1} \\
&= \mathbf{S}_k \mathbf{S}_k^{-1} \mathbf{S'}_k \mathbf{H}_k^T \mathbf{V}_k^{-1} \left(\mathbf{H}_k \mathbf{S'}_k \mathbf{H}_k^T \mathbf{V}_k^{-1} + \mathbf{I} \right)^{-1}.
\end{aligned}$$

Inserting the expression (5.212) for $\mathbf{S}_k^{-1}$, one gets

$$\begin{aligned}
\mathbf{K}_k &= \mathbf{S}_k \left(\mathbf{S'}_k^{-1} + \mathbf{H}_k \mathbf{V}_k^{-1} H^T \right) \mathbf{S'}_k \mathbf{H}_k^T \mathbf{V}_k^{-1} \mathbf{V}_k \left(\mathbf{H}_k \mathbf{S'}_k \mathbf{H}_k^T + \mathbf{V}_k \right)^{-1} \\
&= \mathbf{S}_k \left(\mathbf{I} + \mathbf{H}_k^T \mathbf{V}_k^{-1} H \mathbf{S'}_k \right) \mathbf{H}_k^T \mathbf{V}_k^{-1} \left(\mathbf{H}_k \mathbf{S'}_k \mathbf{H}_k^T \mathbf{V}_k^{-1} + \mathbf{I} \right)^{-1} \\
&= \mathbf{S}_k \mathbf{H}_k^T \mathbf{V}_k^{-1} \left(\mathbf{I} + \mathbf{H}_k \mathbf{S'}_k \mathbf{H}_k^T \mathbf{V}_k^{-1} \right) \left(\mathbf{H}_k \mathbf{S'}_k \mathbf{H}_k^T \mathbf{V}_k^{-1} + \mathbf{I} \right)^{-1} \\
&= \mathbf{S}_k \mathbf{H}_k^T \mathbf{V}_k^{-1}.
\end{aligned}$$

Finally, substituting the expression (5.212) for $\mathbf{S}_k$, we get

$$\mathbf{K}_k = \left[\mathbf{S'}_k^{-1} + \mathbf{H}_k \mathbf{V}_k^{-1} H^T \right]^{-1} \mathbf{H}_k^T \mathbf{V}_k^{-1}, \tag{5.214}$$

which is the expression (5.211) we sought to demonstrate is equivalent to Eq. (5.164) for the Kalman gain. We have thus established that the Kalman filter is formally equivalent to the least-squares method.

Note that while the foregoing expression for the Kalman gain is equivalent to Eq. (5.164), its calculation involves two matrix inversions and is thus more computationally intensive. Use of Eq. (5.164) is thus preferred in general.

https://doi.org/10.1017/9781108241922.007 Published online by Cambridge University Press

Exercises

5.1 Show by direct calculation that the estimator $\hat{\tau}$ given by Eq. (5.14) is an unbiased estimator of the lifetime τ of the exponential PDF, Eq. (5.10). Next, determine whether the estimator $\hat{\tau}$ is biased, unabised, or asymptotically unbiased.

5.2 Show that the expectation value of $\widehat{\sigma^2}$ defined by Eq. (5.22) is $E[\widehat{\sigma^2}] = (n-1)\sigma^2/n$ for a Gaussian distribution, and that $\widehat{\sigma^2}$ is consequently a biased estimator of the variance of this distribution.

5.3 Imagine a researcher is studying the behavior of system and finds it can be represented by a variable x with some PDF $f(x|\vec{\theta})$ in which $\vec{\theta}$ represents a single or multiple unknown parameters. It is often the case that an "experiment" returns a random number, n, of values $\{x_i\}$, with $i = 1, \ldots, n$. Assuming the number n obeys a Poisson distribution of mean ν, show that one can extend the $\mathbb{ML}$ method to obtain an extended likelihood function, $L(\nu, \vec{\theta})$, defined as

$$L(\nu, \vec{\theta}) = \frac{\nu^n}{n!} e^{-\nu} \prod_{i=1}^{n} f(x_i|\vec{\theta}) = \frac{e^{-\nu}}{n!} \prod_{i=1}^{n} \nu f(x_i|\vec{\theta}). \tag{5.215}$$

5.4 Maximize the function $L(\nu, \vec{\theta})$ derived in Problem 5.3 to first show that the estimator $\hat{\nu}$ has an expected value of n. Next, consider the application of this extended likelihood function for the determination of the estimator, $\hat{\theta}$, in experiments where n is a random variable.

5.5 Devise an exponential generator and test the convergence of the estimator $\hat{\tau}$ for various combinations of sample sizes and number of samples. Repeat the exercise for a Gaussian PDF.

5.6 Show that the following expression yields an asymptotically unbiased estimator $\hat{\lambda}$ of the decay constant $\lambda = 1/\tau$:

$$\hat{\lambda} = \frac{1}{\hat{\tau}} = n\left(\sum_{i=1}^{n} t_i\right)^{-1}. \tag{5.216}$$

5.7 Show that the estimator $\hat{\mu}$ obtained in Eq. (5.20) is an unbiased estimator of the mean μ of a Gaussian PDF.

5.8 Show that the estimator $\widehat{\sigma^2}$, given by Eq. (5.22), is an asymptotically unbiased estimator of the variance σ^2 of a Gaussian PDF.

5.9 Verify by direct calculation that s^2, given by Eq. (5.22), is an unbiased estimator of the variance σ^2 of a Gaussian PDF.

5.10 Verify that Eqs. (5.27) and (5.30) yield the variances of estimators of the mean and standard deviation of a Gaussian distribution, respectively.

5.11 Derive Eq. (5.78), providing the variance of the coefficients a_0 and a_1 of linear fits obtained with the LS method.

5.12 Imagine a linear fit $y = a_0 + a_1 y$ of 103 measured points $\{x_i, y_i\}$ has produced coefficients $a_0 = 0.55$ and $a_1 = 10.04$ with an error matrix $U_{00} = 0.04$, $U_{11} = 0.11$,

https://doi.org/10.1017/9781108241922.007 Published online by Cambridge University Press

and $U_{01} = -0.03$. Extrapolate the value y predicted by the model at $x = 5$ and compare the errors δy on this extrapolation obtained while excluding and including the off-diagonal elements of the error matrix.

5.13 Show that if correlations exist between measurements $\hat{\theta}_i$ and their covariance matrix V_{ij} is known, then the weighted average of these measurements involves the weights given by Eq. (5.133).

5.14 Show that the sum of the weights given by Eq. (5.133) equals unity.

https://doi.org/10.1017/9781108241922.007 Published online by Cambridge University Press