Copyright 2022. All Rights Reserved. These notes cannot be copied or distributed in print or electronically without the written consent of Cambridge University Press and the author. The notes are distributed to students attending the course EE566: Adaptation and Learning taught by the author at EPFL during Spring 2022. The notes originate from the text: A. H. Sayed, *Inference and Learning from Data*, Vol. II: *Inference*, Cambridge University Press, 2022.

31 Maximum Likelihood

The maximum-likelihood (ML) formulation is one of the most formidable tools for the solution of inference problems in modern statistical analysis. It allows the estimation of unknown parameters in order to fit probability density functions onto data measurements. We introduce the ML approach in this chapter and limit our discussions to properties that will be relevant for the future developments in the text. The presentation is not meant to be exhaustive but targets key concepts that will be revisited in later chapters. We also avoid anomalous situations and focus on the main features of ML inference that are generally valid under some reasonable regularity conditions.

The ML approach is one notable example of the non-Bayesian viewpoint to inference whereby the unknown quantity to be estimated is modeled as a deterministic *unknown but fixed parameter*, rather than as a random variable. This viewpoint is very relevant when we attempt to fit probability density models onto data. We will comment at the end of the chapter, as well as in later chapters, on the relation to the Bayesian approach to inference problems. In this latter case, both the unknown *and* observations are treated as random variables.

31.1 PROBLEM FORMULATION

Consider a random variable $\boldsymbol{y}$ with probability density function denoted by $f_{\boldsymbol{y}}(y)$. In statistical inference, this pdf is also called the *evidence* of $\boldsymbol{y}$. We assume that $f_{\boldsymbol{y}}(y)$ is dependent on some parameters that are denoted generically by the letter θ. For emphasis, we will write $f_{\boldsymbol{y}}(y;\theta)$ instead of $f_{\boldsymbol{y}}(y)$. For example, the pdf $f_{\boldsymbol{y}}(y)$ could be a Gaussian distribution, in which case θ would refer to its mean or variance or both. In this case, we write $f_{\boldsymbol{y}}(y;\mu,\sigma_y^2)$.

Given an observation y, the maximum-likelihood formulation deals with the problem of estimating θ by maximizing the likelihood function:

$$\widehat{\theta} \;=\; \underset{\theta}{\text{argmax}}\; f_{\boldsymbol{y}}(y;\theta) \tag{31.1}$$

That is, it selects the value of θ that maximizes the likelihood of the observation. The pdf, $f_{\boldsymbol{y}}(y;\theta)$, is called the *likelihood* function and its logarithm is called the *log-likelihood function*:

$$\ell(y;\theta) \overset{\Delta}{=} \ln f_{\boldsymbol{y}}(y;\theta) \tag{31.2}$$

Since the logarithm function is monotonically increasing, the maximum-likelihood estimate can also be determined by solving instead:

$$\widehat{\theta} = \underset{\theta}{\text{argmax}}\; \ell(y;\theta) \tag{31.3}$$

Usually, in the context of maximum-likelihood estimation, we observe N independent and identically distributed realizations, $\{y_n\}$, and use them to estimate θ by maximizing the likelihood function corresponding to these joint observations:

$$\widehat{\theta} = \underset{\theta}{\text{argmax}}\; \ell(y_1,y_2,\ldots,y_N;\theta) = \underset{\theta}{\text{argmax}}\; \ln\left(\prod_{n=1}^{N} f_{\boldsymbol{y}}(y_n;\theta)\right) \tag{31.4}$$

so that

$$\boxed{\widehat{\theta}_{\text{ML}} = \underset{\theta}{\text{argmax}} \left\{ \sum_{n=1}^{N} \ln f_{\boldsymbol{y}}(y_n;\theta) \right\}} \tag{31.5}$$

where we are adding the ML subscript for clarity. Clearly, the ML estimate need not exist; it also need not be unique. We will sometimes write $\widehat{\theta}_N$, with a subscript N, to indicate that the computation of the estimate is based on N measurements.

It is important to realize that the estimate $\widehat{\theta}_{\text{ML}}$ is dependent on the observations $\{y_n\}$. A different collection of N observations arising from the *same* underling true distribution $f_{\boldsymbol{y}}(y)$ will generally lead to a different value for the estimate $\widehat{\theta}_{\text{ML}}$. For this reason, we treat the ML solution as a random variable and introduce the ML *estimator*, $\widehat{\boldsymbol{\theta}}_{\text{ML}}$, which we denote in boldface notation. From this perspective, every estimate $\widehat{\theta}_{\text{ML}}$ corresponds to a realization for the random variable $\widehat{\boldsymbol{\theta}}_{\text{ML}}$. We introduce the estimation error

$$\widetilde{\boldsymbol{\theta}}_{\text{ML}} \overset{\Delta}{=} \theta - \widehat{\boldsymbol{\theta}}_{\text{ML}} \tag{31.6}$$

where θ represents the true unknown parameter. We associate three measures of quality with the ML estimator, namely, its bias, variance, and mean-square error defined by

$$\textbf{bias}: \; \text{bias}(\widehat{\boldsymbol{\theta}}_{\text{ML}}) \overset{\Delta}{=} \theta - \mathbb{E}\,\widehat{\boldsymbol{\theta}}_{\text{ML}} = \mathbb{E}\,\widetilde{\boldsymbol{\theta}}_{\text{ML}} \tag{31.7a}$$

$$\textbf{variance}: \; \text{var}(\widehat{\boldsymbol{\theta}}_{\text{ML}}) \overset{\Delta}{=} \mathbb{E}\,(\widehat{\boldsymbol{\theta}}_{\text{ML}} - \mathbb{E}\,\widehat{\boldsymbol{\theta}}_{\text{ML}})^2 \tag{31.7b}$$

$$\textbf{mean-square-error}: \; \text{MSE}(\widehat{\boldsymbol{\theta}}_{\text{ML}}) \overset{\Delta}{=} \mathbb{E}\,(\widehat{\boldsymbol{\theta}}_{\text{ML}} - \theta)^2 = \mathbb{E}\,\widetilde{\boldsymbol{\theta}}^2_{\text{ML}} \tag{31.7c}$$

where the expectation is relative to the true distribution, $f_{\boldsymbol{y}}(y;\theta)$. When the estimator is unbiased, the mean-square error coincides with its variance. The bias measures how far the estimator is on average from the true parameter, θ. The variance measures how well concentrated the distribution of the estimator is around its mean, whereas the mean-square error measures how well concentrated the same distribution is around the true parameter, θ. Ideally, we would like the error to have zero mean, in which case we say that the ML estimator is *unbiased.* We would also like the estimator to have a small mean-square error (or variance). We will explain in the sequel that the ML estimator has two useful properties for large measurement sizes, N. Specifically, it will be asymptotically unbiased, i.e.,

$$\lim_{N\to\infty} \left\{ \mathbb{E}\,\widehat{\boldsymbol{\theta}}_N \right\} = \theta \tag{31.8}$$

as well as asymptotically *efficient*, meaning that it will attain the smallest variance (and mean-square error) possible:

$$\lim_{N\to\infty} \left\{ \text{var}(\widehat{\boldsymbol{\theta}}_{\text{ML}}) \right\} \;=\; \text{smallest value it can be} \tag{31.9}$$

We will quantify the value of this smallest mean-square error by means of the Cramer-Rao bound.

Example 31.1 **(Bias-variance relation)** It is not always the case that unbiased estimators are preferred. Consider an unknown parameter θ whose estimator is $\widehat{\boldsymbol{\theta}}$ with mean denoted by $\bar{\theta} = \mathbb{E}\,\widehat{\boldsymbol{\theta}}$. The mean-square error of the estimator is given by

$$\begin{aligned}
\text{MSE} &\overset{\Delta}{=} \mathbb{E}\,(\theta - \widehat{\boldsymbol{\theta}})^2 \\
&= \mathbb{E}\,(\theta - \bar{\theta} + \bar{\theta} - \widehat{\boldsymbol{\theta}})^2 \\
&= (\theta - \bar{\theta})^2 + \mathbb{E}\,(\bar{\theta} - \widehat{\boldsymbol{\theta}})^2 + 2(\theta - \bar{\theta})\cancelto{0}{\mathbb{E}\,(\bar{\theta} - \widehat{\boldsymbol{\theta}})} \\
&= (\theta - \bar{\theta})^2 + \mathbb{E}\,(\bar{\theta} - \widehat{\boldsymbol{\theta}})^2 \\
&= \text{bias}^2(\widehat{\boldsymbol{\theta}}) \;+\; \text{var}(\widehat{\boldsymbol{\theta}})
\end{aligned} \tag{31.10}$$

In other words, the MSE is the sum of two components: the squared bias and the variance of the estimator. This means that one may still employ a biased estimator as long as the sum of both components remains small. We commented on the bias-variance relation earlier in Sec. 27.4.

Example 31.2 **(Comparing ML and the Bayesian MAP approach)** The ML formulation treats the parameter θ as some unknown *constant*, and parameterizes the pdf of the observation $\boldsymbol{y}$ in terms of θ by writing $f_{\boldsymbol{y}}(y;\theta)$. This same pdf can be rewritten in the suggestive conditional form $f_{\boldsymbol{y}|\boldsymbol{\theta}}(y|\theta)$ to emphasize that we are referring to the distribution of $\boldsymbol{y}$ given that the parameter $\boldsymbol{\theta}$ is *fixed* at the value $\boldsymbol{\theta} = \theta$. The value of θ is then estimated by maximizing the likelihood function:

$$\widehat{\theta}_{\text{ML}} \;=\; \underset{\theta}{\text{argmax}}\; f_{\boldsymbol{y}|\boldsymbol{\theta}}(y|\theta) \tag{31.11}$$

It is instructive to compare this formulation with the Bayesian MAP approach where both $\boldsymbol{\theta}$ and $\boldsymbol{y}$ are treated as random variables. Returning to (28.11), and using Bayes rule (3.39), we find that the MAP estimator (28.11) corresponds to solving:

$$\widehat{\theta}_{\text{MAP}} \;=\; \underset{\theta}{\text{argmax}} \left\{ f_{\boldsymbol{\theta}}(\theta) f_{\boldsymbol{y}|\boldsymbol{\theta}}(y|\theta) \right\} \tag{31.12}$$

where we are ignoring the marginal pdf, $f_{\boldsymbol{y}}(y)$, because it does not depend on the unknown θ. Observe from the term on the right-hand side of (31.12) that, in contrast to (31.11), the MAP formulation incorporates information about the prior distribution for $\boldsymbol{\theta}$ into the problem statement.

Example 31.3 **(Comparing ML and minimum-variance unbiased estimation)** We discussed the Gauss-Markov theorem in Sec. 29.6, where we considered observation vectors $\boldsymbol{y}$ generated by a linear model of the form $\boldsymbol{y} = H\theta + \boldsymbol{v}$. The parameter $\theta \in \mathbb{R}^M$ is unknown and the perturbation $\boldsymbol{v}$ has zero-mean and covariance matrix $R_v > 0$. The minimum-variance unbiased estimator for θ, i.e., the unbiased estimator with the smallest mean square error was found to be

$$\widehat{\boldsymbol{\theta}}_{\text{MVUE}} = (H^{\mathsf{T}} R_v^{-1} H)^{-1} H^{\mathsf{T}} R_v^{-1} \boldsymbol{y} \tag{31.13}$$

In this example, we wish to explain the relation to maximum likelihood estimation. Although we are dealing now with *vector* quantities $\{\theta, \boldsymbol{y}\}$, the same ML construction applies: we form the log-likelihood function and maximize it over θ.

For the ML derivation, however, we will assume additionally that $\boldsymbol{v}$ is Gaussian distributed. It follows from the model $\boldsymbol{y} = H\theta + \boldsymbol{v}$ that $\boldsymbol{y}$ is Gaussian distributed with mean vector $\bar{y} = H\theta$ and covariance matrix

$$R_y \stackrel{\Delta}{=} \mathbb{E}\,(\boldsymbol{y} - \bar{y})(\boldsymbol{y} - \bar{y})^{\mathsf{T}} = \mathbb{E}\,\boldsymbol{v}\boldsymbol{v}^{\mathsf{T}} = R_v \tag{31.14}$$

In other words, the probability density function of $\boldsymbol{y}$ is given by

$$f_{\boldsymbol{y}}(y;\theta) = \frac{1}{\sqrt{(2\pi)^N}} \frac{1}{\sqrt{\det R_v}} \exp\left\{ -\frac{1}{2}(y - H\theta)^{\mathsf{T}} R_v^{-1} (y - H\theta) \right\} \tag{31.15}$$

The corresponding log-likelihood function is

$$\ell(y;\theta) = -\frac{1}{2}(y - H\theta)^{\mathsf{T}} R_v^{-1} (y - H\theta) + \text{cte} \tag{31.16}$$

where terms independent of θ are grouped into the constant factor. Maximizing $\ell(y;\theta)$ over θ amounts to minimizing the weighted least-squares cost:

$$\widehat{\theta} = \underset{\theta \in \mathbb{R}^M}{\text{argmin}} \left\{ (y - H\theta)^{\mathsf{T}} R_v^{-1} (y - H\theta) \right\} \tag{31.17}$$

Differentiating with respect to θ we find that the minimizer occurs at

$$\widehat{\theta}_{\text{ML}} = (H^{\mathsf{T}} R_v^{-1} H)^{-1} H^{\mathsf{T}} R_v^{-1} y \tag{31.18}$$

which has the same form as (31.13). The main difference though is that the ML derivation assumes the noise component to be Gaussian-distributed and seeks to maximize the log-likelihood function, while the Gauss-Markov theorem is independent of the distribution of the noise and minimizes the mean-square-error.

31.2 GAUSSIAN DISTRIBUTION

We illustrate the ML construction by considering the problem of estimating the mean and variance of a Gaussian distribution. Thus, consider a collection of N independent and identically distributed Gaussian observations, $\{y_n\}$, with unknown mean μ and variance σ_y^2. The joint pdf (or likelihood function) of the observations is given by

$$f_{\boldsymbol{y}_1,\ldots,\boldsymbol{y}_N}(y_1,\ldots,y_N;\mu,\sigma_y^2) = \prod_{n=1}^{N} \frac{1}{(2\pi\sigma_y^2)^{1/2}} e^{-\frac{1}{2\sigma_y^2}(y_n-\mu)^2} \tag{31.19}$$

so that

$$\ell(y_1,\ldots,y_N;\mu,\sigma_y^2) \;=\; -\frac{N}{2}\ln(2\pi\sigma_y^2) \;-\; \frac{1}{2\sigma_y^2}\sum_{n=1}^{N}(y_n-\mu)^2 \tag{31.20}$$

Differentiating this log-likelihood function relative to μ and σ_y^2 and setting the derivatives to zero, we obtain two equations in the unknowns $(\widehat{\mu},\widehat{\sigma}_y^2)$:

$$\frac{1}{\widehat{\sigma}_y^2}\sum_{n=1}^{N}(y_n-\widehat{\mu}) = 0 \tag{31.21a}$$

$$-N\widehat{\sigma}_y^2 \;+\; \sum_{n=1}^{N}(y_n-\widehat{\mu})^2 = 0 \tag{31.21b}$$

Solving these equations leads to the ML *estimates*:

$$\widehat{\mu}_{\rm ML} = \frac{1}{N}\sum_{n=1}^{N} y_n \tag{31.22a}$$

$$\widehat{\sigma}_{y,\rm ML}^2 = \frac{1}{N}\sum_{n=1}^{N}(y_n-\widehat{\mu}_{\rm ML})^2 \tag{31.22b}$$

as well as to similar expressions for the ML *estimators*, where all variables are treated as random variables and expressed in boldface notation:

$$\widehat{\boldsymbol{\mu}}_{\rm ML} = \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{y}_n \tag{31.23a}$$

$$\widehat{\boldsymbol{\sigma}}_{y,\rm ML}^2 = \frac{1}{N}\sum_{n=1}^{N}(\boldsymbol{y}_n-\widehat{\boldsymbol{\mu}}_{\rm ML})^2 \tag{31.23b}$$

It is straightforward to verify from these expressions that one of the estimators is unbiased while the other is biased; see Prob. 31.1 where it is shown that

$$\mathbb{E}\,\widehat{\boldsymbol{\mu}}_{\rm ML} = \mu, \quad \mathbb{E}\,\widehat{\boldsymbol{\sigma}}_{y,\rm ML}^2 = \left(\frac{N-1}{N}\right)\sigma_y^2 \tag{31.24}$$

Although the variance estimator is biased, it nevertheless becomes asymptotically unbiased as $N \to \infty$. This does not mean that we cannot construct an unbiased estimator for σ_y^2 for finite N. Actually, the rightmost expression in (31.24) suggests the following construction:

$$\widehat{\boldsymbol{\sigma}}_{y,\rm unbiased}^2 = \frac{1}{N-1}\sum_{n=1}^{N}(\boldsymbol{y}_n-\widehat{\boldsymbol{\mu}}_{\rm ML})^2 \tag{31.25}$$

where the scaling by $1/N$ in (31.23b) is replaced by $1/(N-1)$ so that

$$\mathbb{E}\,\widehat{\boldsymbol{\sigma}}_{y,\rm unbiased}^2 = \sigma_y^2 \tag{31.26}$$

This second construction, however, is not an ML estimator.

What about the mean-square error performance? In this case, we can construct yet another estimator for σ_y^2 with a smaller mean-square error than $\widehat{\boldsymbol{\sigma}}^2_{y,\mathrm{ML}}$. To see this, assume we pose the problem of searching for an estimator for σ_y^2 of the following form:

$$\widehat{\boldsymbol{\sigma}}^2_{y,\mathrm{MSE}} \overset{\Delta}{=} \alpha \sum_{n=1}^{N} (\boldsymbol{y}_n - \widehat{\boldsymbol{\mu}}_{\mathrm{ML}})^2 \tag{31.27}$$

for some scalar $\alpha > 0$ chosen to minimize the resulting mean-square error:

$$\alpha^o = \underset{\alpha}{\mathrm{argmin}}\ \mathbb{E}\,(\sigma_y^2 - \widehat{\boldsymbol{\sigma}}^2_{y,\mathrm{MSE}})^2 \tag{31.28}$$

We show in Prob. 31.2 that $\alpha^o = 1/(N+1)$ so that the third estimator is:

$$\widehat{\boldsymbol{\sigma}}^2_{y,\mathrm{MSE}} = \frac{1}{N+1} \sum_{n=1}^{N} (\boldsymbol{y}_n - \widehat{\boldsymbol{\mu}})^2 \tag{31.29}$$

Obviously, this estimator is biased. It agrees with neither the ML estimator (31.23b), which is scaled by $1/N$, nor the unbiased estimator (31.25), which is scaled by $1/(N-1)$.

Example 31.4 **(Fitting Gaussian and Beta distributions)** The top row in Figure 31.1 shows on the left a histogram distribution for the serum cholesterol level measured in mg/dl for $N = 297$ patients. The vertical axis measures absolute frequencies. The plot uses 15 bins of width 30mg/dl each, and shows how many patients fall within each bin. The same plot is normalized on the right by dividing each bin value by $N = 297$ measurements and by the bin width — recall the explanation given in Remark 6.1. By doing so, the result is an approximate probability density function. A Gaussian pdf is fitted on top of the normalized data. The mean and variance of the Gaussian distribution are determined by using expressions (31.23a) and (31.23b). If we denote the cholesterol level by the random variable $\boldsymbol{y}$, then the sample mean and variance values are found to be

$$\widehat{\mu}_{\mathrm{cholesterol}} = \frac{1}{N} \sum_{n=1}^{N} y_n \approx 247.35 \tag{31.30a}$$

$$\widehat{\sigma}^2_{\mathrm{cholesterol}} = \frac{1}{N-1} \sum_{n=1}^{N} (y_n - \bar{y})^2 \approx 2703.7 \tag{31.30b}$$

where y_n refers to the $n-$th cholesterol measurement.

The bottom row in Figure 31.1 repeats the same construction for the maximal heart rate of a patient measured in beats per minute (bpm) from the same dataset. If we denote the heart rate by the random variable $\boldsymbol{z}$, then the sample mean and variance values are found to be

$$\widehat{\mu}_{\mathrm{heartrate}} = \frac{1}{N} \sum_{n=1}^{N} z_n \approx 149.60 \tag{31.31a}$$

$$\widehat{\sigma}^2_{\mathrm{heartrate}} = \frac{1}{N-1} \sum_{n=1}^{N} (z_n - \bar{z})^2 \approx 526.32 \tag{31.31b}$$

where z_n refers to the $n-$th heart beat measurement. By examining the rightmost lower plot in Figure 31.1, it appears that the histogram distribution is skewed to the right. This observation motivates us to consider fitting a different distribution onto the data in order to better capture the skewness in the histogram; this is not possible if we persist with the Gaussian distribution due to its symmetry.

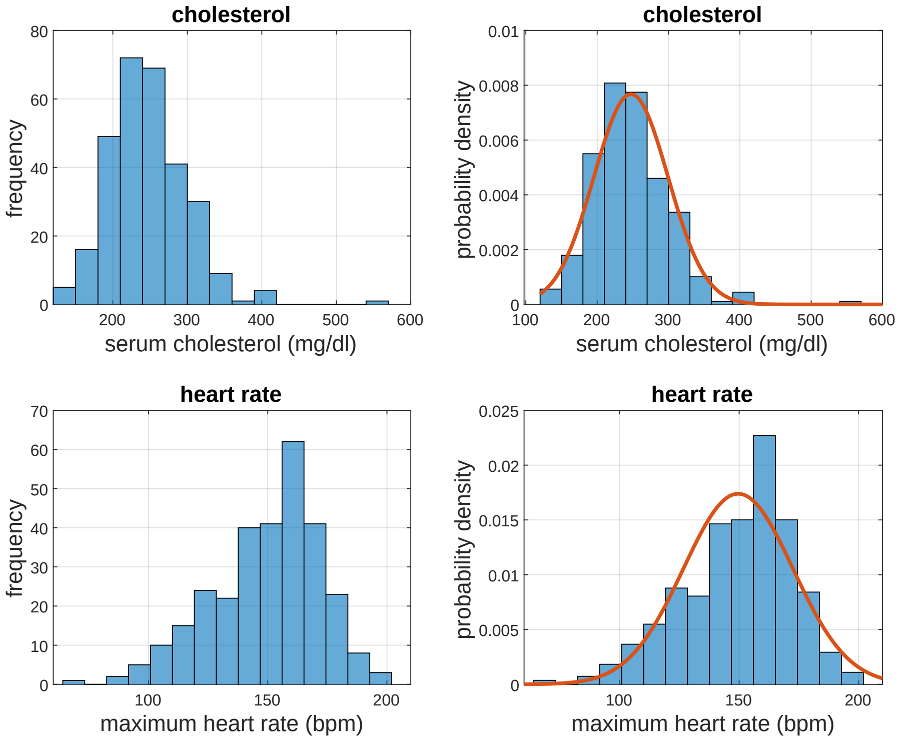

Figure 31.1 (*Top*) Histogram distribution of the serum cholesterol level measured in mg/dl on the left using 15 bins of width 30mg/dl each, and its normalized version on the right where each bin value is divided $N = 297$ measurements and by the bin width. By doing so, the result is an approximate probability density function. A Gaussian pdf is fitted on top of the normalized data. (*Bottom*) Similar construction for the maximum heart rate of a patient measured in beats per minute (bpm). A Gaussian pdf is fitted on top of the normalized data. The data is derived from the processed Cleveland dataset from the site `https://archive.ics.uci.edu/ml/datasets/heart+Disease`.

First, we normalize the heart rate variable so that it is confined to the interval $[0, 1)$. We do so by dividing $\boldsymbol{z}$ by (a slightly larger number than) the maximum heart rate in the data, which is 202. We denote this normalized variable by $\boldsymbol{t}$. We have access to $N = 297$ measurements $\{t_n\}$, obtained by normalizing the heart rates z_n by $\epsilon + \max z_n$ (for a small ϵ; this ensures that all values of t_n are strictly less than one so that logarithms of $1 - t_n$ will be well-defined further ahead in (31.34)). Next, we consider fitting a *Beta distribution* onto the data $\{t_n\}$. The pdf of a Beta distribution has the form:

$$f_{\boldsymbol{t}}(t; a, b) = \begin{cases} \dfrac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\, t^{a-1}(1-t)^{b-1}, & 0 \le t \le 1 \\ 0, & \text{otherwise} \end{cases} \tag{31.32}$$

where $\Gamma(x)$ denotes the Gamma function defined earlier in Prob. 4.3. Different choices for (a,b) result in different behavior for the distribution $f_{\boldsymbol{t}}(t)$. We need to estimate the shape parameters (a,b).

We have a collection of N independent measurements $\{t_n\}$. The likelihood function of these observations is given by

$$f_{\boldsymbol{t}_1,\ldots,\boldsymbol{t}_N}(t_1,\ldots,t_N;a,b)=\left(\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\right)^N\prod_{n=1}^{N}t_n^{a-1}(1-t_n)^{b-1} \tag{31.33}$$

so that, in the log domain,

$$\begin{aligned}\ell(t_1,\ldots,t_N;a,b)=N\ln\,\Gamma(a+b)-N\ln\,\Gamma(a)-N\ln\,\Gamma(b)\,+&\\ (a-1)\sum_{n=1}^{N}\ln t_n\;+\;(b-1)\sum_{n=1}^{N}\ln(1-t_n)&\end{aligned} \tag{31.34}$$

Differentiating with respect to a and b gives

$$\partial\ell(t_1,\ldots,t_N;a,b)/\partial a=N\left(\frac{\Gamma'(a+b)}{\Gamma(a+b)}-\frac{\Gamma'(a)}{\Gamma(a)}\right)+\sum_{n=1}^{N}\ln t_n \tag{31.35}$$

$$\partial\ell(t_1,\ldots,t_N;a,b)/\partial b=N\left(\frac{\Gamma'(a+b)}{\Gamma(a+b)}-\frac{\Gamma'(b)}{\Gamma(b)}\right)+\sum_{n=1}^{N}\ln(1-t_n) \tag{31.36}$$

where $\Gamma'(x)$ denotes the derivative of the $\Gamma-$function. Two complications arise here. First, we need to know how to compute ratios of the form $\psi(x)=\Gamma'(x)/\Gamma(x)$ for the Gamma function; this ratio is known as the *digamma* function and it is equal to the derivative of $\ln\Gamma(x)$. The computation of the digamma function is not straightforward. As was mentioned earlier in (5.62), and based on properties of the Gamma function, it is known that

$$\psi(x)\;\stackrel{\Delta}{=}\;\frac{\Gamma'(x)}{\Gamma(x)}\approx-0.577215665+\sum_{m=0}^{\infty}\left(\frac{1}{1+m}-\frac{1}{x+m}\right) \tag{31.37}$$

The expression on the right-hand side can be used to approximate $\Gamma'(x)/\Gamma(x)$ by replacing the infinite series by a finite sum. Second, even then, if we set the derivatives (31.35)–(31.36) to zero, the resulting equations will not admit a closed-from solution for the parameters (a,b).

Another way to seek values $(\widehat{a},\widehat{b})$ that maximize the likelihood function is to employ a gradient-ascent recursion of the following form for $n\geq 0$ (along the lines discussed in Chapter 12 on gradient-descent algorithms):

$$a_n=a_{n-1}+\mu\left\{\frac{\Gamma'(a_{n-1}+b_{n-1})}{\Gamma(a_{n-1}+b_{n-1})}-\frac{\Gamma'(a_{n-1})}{\Gamma(a_{n-1})}+\frac{1}{N}\sum_{n=1}^{N}\ln t_n\right\} \tag{31.38}$$

$$b_n=b_{n-1}+\mu\left\{\frac{\Gamma'(a_{n-1}+b_{n-1})}{\Gamma(a_{n-1}+b_{n-1})}-\frac{\Gamma'(b_{n-1})}{\Gamma(b_{n-1})}+\frac{1}{N}\sum_{n=1}^{N}\ln(1-t_n)\right\} \tag{31.39}$$

where μ is a small step-size parameter. These recursions need to be initialized from a good starting point. In this example, we repeat the iterations for a total of 10,000 times using $\mu=0.001$. We use the construction explained next to determine good initial conditions (a_{-1},b_{-1}). Actually, the construction provides yet another method to fit a Beta distribution onto the data, albeit one that does not need to run the gradient-ascent recursion altogether.

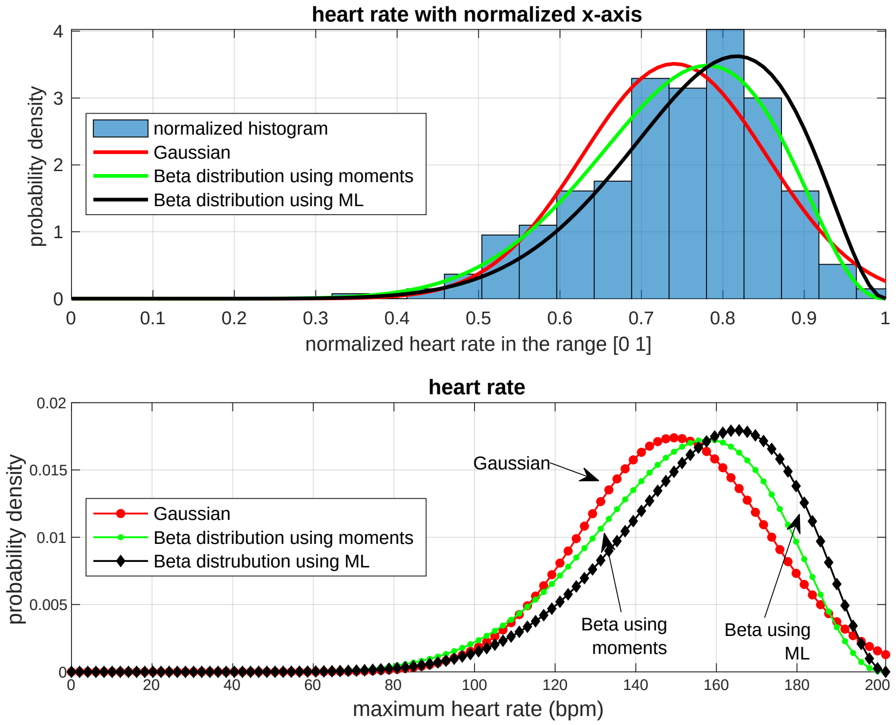

Figure 31.2 (*Top*) Normalized histogram for the scaled heart rate variables $\{t_n\}$, along with three probability density functions: a Gaussian fit, a Beta distribution fit obtained from the moment matching method, and a Beta distribution fit obtained from a gradient-ascent iteration for maximum-likelihood. (*Bottom*) The same probability distributions with the horizontal axis returned to the original heart rate scale (obtained by multiplying the horizontal axis of the top figure by the maximum heart rate, as well as scaling the vertical axis down by the same value to ensure that the area under each of the probability distributions stays normalized to one.)

Indeed, the mean and variance of a Beta distribution with shape parameters a and b are given by

$$\bar{t} = \frac{a}{a+b}, \qquad \sigma_t^2 = \frac{ab}{(a+b)^2(a+b+1)} \tag{31.40}$$

We can solve these equations in terms of a and b and find that

$$a = \bar{t}\left(\frac{\bar{t}(1-\bar{t})}{\sigma_t^2} - 1\right) \tag{31.41}$$

$$b = (1-\bar{t})\left(\frac{\bar{t}(1-\bar{t})}{\sigma_t^2} - 1\right) \tag{31.42}$$

These expressions suggest another method (called a *moment matching method*) to fit the Beta distribution to data measurements. We estimate the mean and variance of the distribution from the data, say, as

$$\widehat{t} = \frac{1}{N}\sum_{n=1}^{N} t_n, \qquad \widehat{\sigma}_t^2 = \frac{1}{N-1}\sum_{n=1}^{N}(t_n - \widehat{t}\,)^2 \tag{31.43}$$

and then use these values in (31.41)–(31.42) to estimate a and b. Using this construction we obtain

$$\widehat{a} = 10.2900, \quad \widehat{b} = 3.6043, \qquad (\textbf{moment matching}) \tag{31.44}$$

This is of course not a maximum likelihood solution. Using these values as initial conditions for the gradient-ascent iterations (31.38)–(31.39) we arrive at a second set of estimates for a and b:

$$\widehat{a} = 10.2552, \quad \widehat{b} = 3.0719, \qquad (\textbf{ML method}) \tag{31.45}$$

The resulting Beta distributions are shown in Figure 31.2 along with the Gaussian distribution from the earlier figure for comparison purposes.

Example 31.5 **(Fitting the empirical data distribution – discrete case)** Let us return to the ML formulation (31.5) and provide two useful interpretations for it in terms of what is known as the *empirical data distribution*. The interpretations are easier to describe for discrete random variables, $\boldsymbol{y}$. For convenience, we will continue to denote the probability mass function (pmf) for $\boldsymbol{y}$ by the notation $f_{\boldsymbol{y}}(y;\theta)$ so that

$$f_{\boldsymbol{y}}(y;\theta) \text{ stands for } \mathbb{P}(\boldsymbol{y} = y;\theta) \tag{31.46}$$

In this notation, the pmf is dependent on the parameter θ. The observations for $\boldsymbol{y}$ arise from a discrete set $\mathcal{Y}$ representing the support of its pmf, i.e.,

$$\boldsymbol{y} \in \mathcal{Y} \;\stackrel{\Delta}{=}\; \{o_1, o_2, \ldots, o_L\} \tag{31.47}$$

where we are denoting the individual elements of $\mathcal{Y}$ by $\{o_\ell\}$.

We consider a collection of N independent realizations $\{y_n\}$. The maximum-likelihood problem for estimating θ does not change if we scale the likelihood function by $1/N$ so that:

$$\boxed{\widehat{\theta}_{\text{ML}} = \underset{\theta}{\text{argmax}} \left\{ \frac{1}{N}\sum_{n=1}^{N} \ln f_{\boldsymbol{y}}(y_n;\theta) \right\}} \qquad (\textbf{log-likelihood}) \tag{31.48}$$

If we now examine the observations $\{y_n\}$, some of them may assume repeated values. Let p_ℓ denote the relative frequency for realization o_ℓ in the observed set (this is a measure of how often o_ℓ appears within the observation set). In particular, if o_ℓ appears a_ℓ times within the N observations, then

$$p_\ell \;\stackrel{\Delta}{=}\; a_\ell/N, \qquad (\textbf{relative frequency for } o_\ell) \tag{31.49}$$

In this way, we end up constructing an *empirical distribution* (or histogram) with the observed data defined by

$$\widehat{f}_{\boldsymbol{y}}(y = o_\ell) = p_\ell \tag{31.50}$$

Figure 31.3 compares the parameterized pmf $f_{\boldsymbol{y}}(y;\theta)$ and the empirical distribution $\widehat{f}_{\boldsymbol{y}}(y)$, which corresponds to a normalized histogram with all relative frequencies $\{p_\ell\}$ adding up to one.

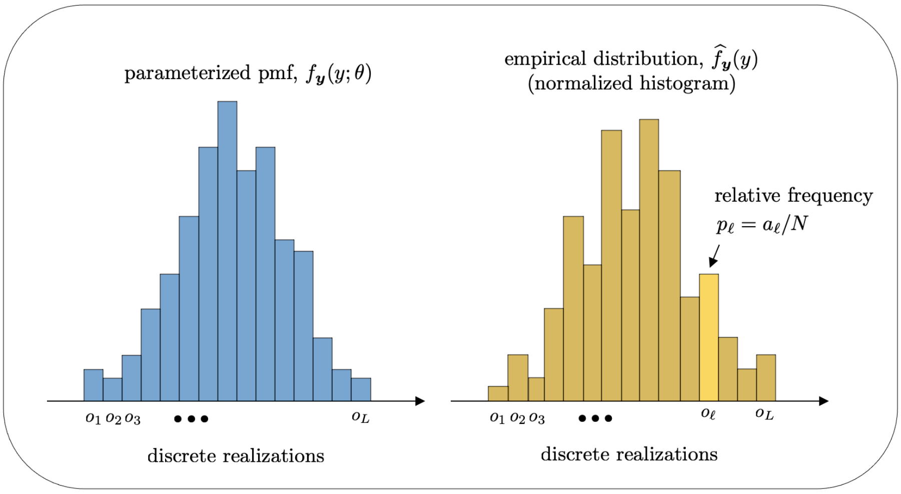

Figure 31.3 Parameterized pmf $f_{\boldsymbol{y}}(y;\theta)$ on the left versus the empirical distribution, $\widehat{f}_{\boldsymbol{y}}(y)$, on the right, which amounts to a normalized histogram with the relative frequencies $\{p_\ell\}$ adding up to one.

Now, using expression (6.43), the KL divergence between the empirical and actual pmfs is given by

$$\begin{aligned} D_{\rm KL}(\widehat{f}_y \| f_y) &\triangleq \mathbb{E}_{\widehat{f}_y} \ln \widehat{f}_{\boldsymbol{y}}(y) \;-\; \mathbb{E}_{\widehat{f}_y} \ln f_{\boldsymbol{y}}(y;\theta) \\ &= \sum_{\ell=1}^{L} p_\ell \ln p_\ell \;-\; \sum_{\ell=1}^{L} p_\ell \ln f_{\boldsymbol{y}}(y=o_\ell;\theta) \\ &= \sum_{\ell=1}^{L} p_\ell \ln p_\ell \;-\; \frac{1}{N}\sum_{\ell=1}^{L} a_\ell \ln f_{\boldsymbol{y}}(y=o_\ell;\theta) \\ &= \sum_{\ell=1}^{L} p_\ell \ln p_\ell \;-\; \frac{1}{N}\sum_{n=1}^{N} \ln f_{\boldsymbol{y}}(y_n;\theta) \end{aligned} \tag{31.51}$$

where the expectations are computed relative to the empirical distribution. It follows that the ML solution, which maximizes the rightmost term in the above equation, is effectively minimizing the KL divergence between the empirical pmf $\widehat{f}_{\boldsymbol{y}}(y)$ and the fitted model $f_{\boldsymbol{y}}(y;\theta)$:

$$\boxed{\widehat{\theta}_{\rm ML} = \underset{\theta}{\rm argmin}\; D_{\rm KL}(\widehat{f}_y \,\|\, f_y)} \qquad \textbf{(KL divergence)} \tag{31.52}$$

Another useful interpretation for the maximum-likelihood solution follows if we appeal to the conclusion from Example 6.10, which relates the cross-entropy between two distributions to their KL divergence. The cross-entropy between the empirical pmf

$\widehat{f}_{\boldsymbol{y}}(y)$ and the fitted model $f_{\boldsymbol{y}}(y;\theta)$ is defined by

$$\begin{aligned} H(\widehat{f}_y, f_y) &\stackrel{\Delta}{=} -\mathbb{E}_{\widehat{f}_y} \ln f_{\boldsymbol{y}}(y;\theta) \\ &= -\sum_{\ell=1}^{L} p_\ell \ln f_{\boldsymbol{y}}(y=o_\ell;\theta) \\ &= -\frac{1}{N}\sum_{n=1}^{N} \ln f_{\boldsymbol{y}}(y_n;\theta) \end{aligned} \tag{31.53}$$

and, hence, it also holds that

$$\boxed{\widehat{\theta}_{\rm ML} = \underset{\theta}{\text{argmin}}\ H(\widehat{f}_y, f_y)} \qquad \textbf{(cross-entropy)} \tag{31.54}$$

In summary, the following interpretations hold:

$$\begin{cases} \widehat{\theta}_{\rm ML} \text{ maximizes the log-likelihood function (31.48)} \\ \widehat{\theta}_{\rm ML} \text{ minimizes the KL divergence (31.52)} \\ \widehat{\theta}_{\rm ML} \text{ minimizes the cross-entropy (31.54)} \end{cases} \tag{31.55}$$

Example 31.6 **(Fitting the data distribution – continuous case)** We extend the conclusions of the previous example to *continuous* random variables $\boldsymbol{y}$ as follows. We assume the realizations for $\boldsymbol{y}$ arise from a *true* pdf denoted by $f_{\boldsymbol{y}}(y;\theta^o)$ and parameterized by some unknown parameter θ^o. We again consider a collection of N independent realizations $\{y_n\}$ arising from this distribution. The maximum-likelihood formulation fits a model $f_{\boldsymbol{y}}(y;\theta)$ by seeking a value for θ that solves:

$$\widehat{\theta}_{\rm ML} = \underset{\theta}{\text{argmax}} \left\{ \frac{1}{N}\sum_{n=1}^{N} \ln f_{\boldsymbol{y}}(y_n;\theta) \right\} \tag{31.56}$$

Under ergodicity, and for large enough $N \to \infty$, the above problem can be replaced by

$$\boxed{\widehat{\theta}_{\rm ML} = \underset{\theta}{\text{argmax}}\ \mathbb{E}_{\theta^o}\Big(\ln f_{\boldsymbol{y}}(\boldsymbol{y};\theta)\Big)} \qquad \textbf{(log-likelihood)} \tag{31.57}$$

where the expectation is over the true distribution, $\boldsymbol{y} \sim f_{\boldsymbol{y}}(y;\theta^o)$. We are highlighting this fact by writing $\mathbb{E}_{\theta^o}$, with a subscript θ^o.

Now note from expression (6.43) that the KL divergence between the true and fitted pdfs is given by

$$D_{\rm KL}\Big(f_{\boldsymbol{y}}(y;\theta^o)\,\|\,f_{\boldsymbol{y}}(y;\theta)\Big) \stackrel{\Delta}{=} \mathbb{E}_{\theta^o}\Big(\ln f_{\boldsymbol{y}}(\boldsymbol{y};\theta^o)\Big) - \mathbb{E}_{\theta^o}\Big(\ln f_{\boldsymbol{y}}(\boldsymbol{y};\theta)\Big) \tag{31.58}$$

where the first term is independent of θ. Since the ML solution maximizes the rightmost term in the above equation, we conclude that it is also minimizing the KL divergence between the true and fitted models, namely, for large enough N:

$$\boxed{\widehat{\theta}_{\rm ML} = \underset{\theta}{\text{argmin}}\ D_{\rm KL}\Big(f_{\boldsymbol{y}}(y;\theta^o)\,\|\,f_{\boldsymbol{y}}(y;\theta)\Big)} \qquad \textbf{(KL divergence)} \tag{31.59}$$

We encountered one application for this result earlier in Example 28.7 while motivating the logistic risk. Another interpretation for the maximum-likelihood solution can be obtained if we refer to the conclusion of Example 6.10, which relates the cross-entropy between two distributions to their KL divergence. The cross-entropy between the true and fitted models is defined by

$$H\Big(f_{\boldsymbol{y}}(y;\theta^o), f_{\boldsymbol{y}}(y;\theta)\Big) \;\overset{\Delta}{=}\; -\mathbb{E}_{\theta^o}\Big(\ln f_{\boldsymbol{y}}(y;\theta)\Big) \tag{31.60}$$

and, hence, it also holds that for N large enough:

$$\boxed{\widehat{\theta}_{\rm ML} = \underset{\theta}{\operatorname{argmin}}\; H\Big(f_{\boldsymbol{y}}(y;\theta^o), f_{\boldsymbol{y}}(y;\theta)\Big)} \qquad (\textbf{cross-entropy}) \tag{31.61}$$

In summary, the following interpretations hold:

$$\begin{cases} \widehat{\theta}_{\rm ML} \text{ maximizes the log-likelihood function (31.57)} \\ \widehat{\theta}_{\rm ML} \text{ minimizes the KL divergence (31.59)} \\ \widehat{\theta}_{\rm ML} \text{ minimizes the cross-entropy (31.61)} \end{cases} \tag{31.62}$$

31.3 MULTINOMIAL DISTRIBUTION

We continue to illustrate the ML construction by considering next the problem of estimating the parameters defining a multinomial distribution, which is a generalization of the binomial distribution. Recall that the binomial distribution is useful to model the outcome of an experiment involving the tossing of a coin N times. Each experiment consists of only two possible outcomes: “heads” or “tails”. The binomial distribution then allows us to assess the likelihood of observing y heads in N tosses by means of the following expression defined in terms of the factorial operation:

$$\begin{aligned} \mathbb{P}(y \text{ heads in } N \text{ tosses}) &= \binom{N}{y} p^y(1-p)^{N-y} \\ &= \frac{N!}{y!(N-y)!}p^y(1-p)^{N-y} \end{aligned} \tag{31.63}$$

where $p \in [0,1]$ denotes the probability of observing a “head” in any given toss of the coin. The multinomial distribution generalizes this setting and allows each experiment to involve more than two outcomes, say, $L \geq 2$ of them. This situation arises, for example, if we toss a dice with L faces with each face ℓ having a probability p_ℓ of being observed with the probabilities satisfying the obvious normalization:

$$\sum_{\ell=1}^{L} p_\ell = 1 \tag{31.64}$$

The multinomial distribution then allows us to assess the likelihood of observing y_1 times the first face, y_2 times the second face, and so on, in a total of N tosses, by means of the following expression:

$$\begin{aligned}\mathbb{P}(y_1, y_2, \ldots, y_L \text{ in } N \text{ tosses}) &= \frac{N!}{y_1! y_2! \ldots y_L!} p_1^{y_1} p_2^{y_2} \ldots p_L^{y_L} \\ &= N! \left(\prod_{\ell=1}^{L} \frac{p_\ell^{y_\ell}}{y_\ell!} \right) \\ &\stackrel{\Delta}{=} f_{\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_L}(y_1, y_2, \ldots, y_L) \qquad (31.65)\end{aligned}$$

where the last line introduces a compact notation for the probability expression. It is clear that the multinomial distribution is parameterized by the scalars $\{p_1, p_2, \ldots, p_L\}$. In the next two examples, we consider special cases and delay the general case to Sec. 31.4 where we study the exponential family of distributions.

Example 31.7 (**Elephants, horses, and cars**) Consider a multinomial distribution with $L = 3$ outcomes with probabilities $\{p_1, p_2, p_3\}$. The same arguments and derivation that follow can be extended to an arbitrary number L of outcomes. Thus, assume we are dealing with an experiment involving a box with $L = 3$ types of images in it: type #1 are images of horses, type #2 are images of elephants, and type #3 are images of cars. The probability of selecting any given type ℓ is p_ℓ; this situation is illustrated schematically in Fig. 31.4, where we are assuming for this example that the probabilities $\{p_1, p_2, p_3\}$ are parameterized in terms of some unknown scalar parameter θ as follows:

$$p_1 = \frac{1}{4}, \quad p_2 = \frac{1}{4} + \theta, \quad p_3 = \frac{1}{2} - \theta \qquad (31.66)$$

Assume we repeat the experiment a total of N independent times, and write down the number of times, $\{y_1, y_2, y_3\}$, that images of types $\#1, \#2$, and $\#3$ are observed. In view of the multinomial distribution (31.65), the probability of observing each type ℓ a number y_ℓ times is given by

$$f_{\boldsymbol{y}_1, \boldsymbol{y}_2, \boldsymbol{y}_3}(y_1, y_2, y_3) = \frac{N!}{y_1! y_2! y_3!} p_1^{y_1} p_2^{y_2} p_3^{y_3} \qquad (31.67)$$

where the y_ℓ assume integer values and satisfy:

$$y_\ell \in \{0, 1, \ldots, N\}, \qquad y_1 + y_2 + y_3 = N \qquad (31.68)$$

Expression (31.67) allows us to assess the likelihood of observing an "elephant" y_1 times, a "horse" y_2 times, and a "car" y_3 times.

Now, using the observations $\{y_1, y_2, y_3\}$, we are interested in determining the maximum likelihood estimate for the parameter θ (and, consequently, for the probabilities $\{p_1, p_2, p_3\}$). This can be done by maximizing directly the log-likelihood function, which in this case is given by

$$\begin{aligned}\ell(y_1, y_2, y_3; \theta) &\stackrel{\Delta}{=} \ln \left\{ \frac{N!}{y_1! y_2! y_3!} \left(\frac{1}{4}\right)^{y_1} \left(\frac{1}{4} + \theta\right)^{y_2} \left(\frac{1}{2} - \theta\right)^{y_3} \right\} \qquad (31.69) \\ &= \ln\left(\frac{N!}{y_1! y_2! y_3!}\right) + y_1 \ln\left(\frac{1}{4}\right) + y_2 \ln\left(\frac{1}{4} + \theta\right) + y_3 \ln\left(\frac{1}{2} - \theta\right)\end{aligned}$$

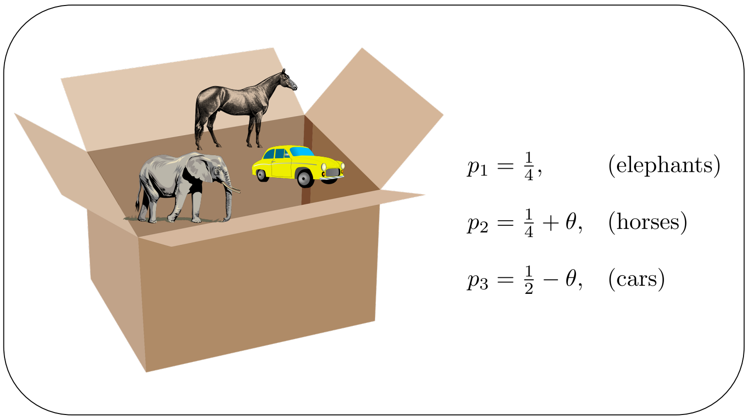
$p_1 = \frac{1}{4}$, (elephants)
$p_2 = \frac{1}{4} + \theta$, (horses)
$p_3 = \frac{1}{2} - \theta$, (cars)

Figure 31.4 The box contains $L = 3$ types of images (elephants, horses, and cars). The probability of picking with replacement an image of type ℓ is p_ℓ; these probabilities are parameterized by the scalar θ in (31.66). The source of the individual images is www.pixabay.com, where images are free to use.

Differentiating with respect to θ and setting the derivative to zero at $\theta = \widehat{\theta}$ gives

$$\left.\frac{y_2}{\frac{1}{4}+\theta} - \frac{y_3}{\frac{1}{2}-\theta}\right|_{\theta=\widehat{\theta}} = 0 \Longrightarrow \boxed{\widehat{\theta}_{|y_1,y_2,y_3} = \frac{\frac{1}{2}y_2 - \frac{1}{4}y_3}{y_2+y_3}} \tag{31.70}$$

where the subscript added to $\widehat{\theta}$ is meant to indicate that this estimate is based on the measurements y_1, y_2, and y_3.

Example 31.8 **(Partial information)** The solution in the previous example assumes that we have access to the number of outcomes $\{y_1, y_2, y_3\}$. Let us now consider a scenario where we only have access to *partial* information. Assume that all we know is the number of times that an "animal" image has been observed and the number of times that a "car" image has been observed. That is, we only know the quantities $y_1 + y_2$ and y_3. We are still interested in estimating θ from this information. This particular problem can still be solved directly using the maximum-likelihood (ML) formulation. To do so, we need to determine the likelihood function for the random variables $\{\boldsymbol{y}_1 + \boldsymbol{y}_2, \boldsymbol{y}_3\}$, which can be found from the following calculation (note that the variables (s, y_3) below satisfy $s + y_3 = N$):

$$\begin{aligned}
f_{\boldsymbol{s},\boldsymbol{y}_3}(y_1+y_2=s,\, y_3;\theta) &= \sum_{m=0}^{s} f_{\boldsymbol{y}_1,\boldsymbol{y}_2,\boldsymbol{y}_3}(y_1=m,\, y_2=s-m,\, y_3;\theta)\\
&= \sum_{m=0}^{s} \frac{N!}{m!(s-m)!y_3!}\left(\frac{1}{4}\right)^m \left(\frac{1}{4}+\theta\right)^{s-m}\left(\frac{1}{2}-\theta\right)^{y_3}\\
&= \frac{N!}{s!y_3!}\left(\frac{1}{2}-\theta\right)^{y_3}\sum_{m=0}^{s}\frac{s!}{m!(s-m)!}\left(\frac{1}{4}\right)^m\left(\frac{1}{4}+\theta\right)^{s-m}\\
&= \frac{N!}{s!y_3!}\left(\frac{1}{2}-\theta\right)^{y_3}\sum_{m=0}^{s}\begin{pmatrix} s\\ m\end{pmatrix}\left(\frac{1}{4}\right)^m\left(\frac{1}{4}+\theta\right)^{s-m}
\end{aligned} \tag{31.71}$$

We now call upon the binomial theorem, which states that, for any integer s and real numbers a and b:

$$(a+b)^s = \sum_{m=0}^{s} \binom{s}{m} a^m b^{s-m} \tag{31.72}$$

and use it to simplify (31.71) as

$$\begin{aligned} f_{\boldsymbol{s},\boldsymbol{y}_3}(y_1+y_2=s,\, y_3;\theta) &= \frac{N!}{s!y_3!}\left(\frac{1}{2}-\theta\right)^{y_3}\left(\frac{1}{2}+\theta\right)^{s} \\ &\overset{(a)}{=} \frac{N!}{s!(N-s)!}\left(\frac{1}{2}-\theta\right)^{N-s}\left(\frac{1}{2}+\theta\right)^{s} \\ &= \binom{N}{s}\left(\frac{1}{2}+\theta\right)^{s}\left(\frac{1}{2}-\theta\right)^{N-s} \end{aligned} \tag{31.73}$$

where step (a) follows from the fact that the value of y_3 is fixed at $y_3 = N-s$. Note that expression (31.73) shows that the sum variable $\boldsymbol{s} = \boldsymbol{y}_1 + \boldsymbol{y}_2$ follows a binomial distribution with success rate equal to $\frac{1}{2}+\theta$. It follows that the log-likelihood function is given by:

$$\ln f_{\boldsymbol{s},\boldsymbol{y}_3}(y_1+y_2=s,\, y_3;\theta) \;=\; \ln\binom{N}{s} \;+\; s\ln\left(\frac{1}{2}+\theta\right) \;+\; y_3\ln\left(\frac{1}{2}-\theta\right) \tag{31.74}$$

Differentiating with respect to θ and setting the derivative to zero at $\theta=\widehat{\theta}$ leads to:

$$\left.\frac{s}{\frac{1}{2}+\theta}-\frac{y_3}{\frac{1}{2}-\theta}\right|_{\theta=\widehat{\theta}} = 0 \Longrightarrow \boxed{\widehat{\theta}_{|y_1+y2,y_3} = \frac{\frac{1}{2}(y_1+y_2)-\frac{1}{2}y_3}{N}} \tag{31.75}$$

where the subscript added to $\widehat{\theta}$ is meant to indicate that this estimate is based on the measurements y_1+y_2 and y_3.

31.4 EXPONENTIAL FAMILY OF DISTRIBUTIONS

We illustrate next the application of the ML formulation to the exponential family of probability distributions. We showed in Chapter 5 that this family includes many other distributions as special cases, such as the Gaussian distribution, the binomial distribution, the multinomial distribution, the Dirichlet distribution, the Gamma distribution, and others.

Thus, consider a *vector* random variable $\boldsymbol{y} \in \mathbb{R}^P$ that follows the exponential distribution described by (5.2) in its natural or canonical form, namely,

$$\boxed{f_{\boldsymbol{y}}(y;\theta) \;=\; h(y)e^{\theta^{\mathsf{T}}T(y)-a(\theta)}} \tag{31.76}$$

where the pdf is parameterized by $\theta \in \mathbb{R}^M$, and the functions $\{h(y), T(y), a(\theta)\}$ satisfy

$$h(y) \geq 0 : \mathbb{R}^P \rightarrow \mathbb{R}, \qquad T(y) : \mathbb{R}^P \rightarrow \mathbb{R}^M, \qquad a(\theta) : \mathbb{R}^M \rightarrow \mathbb{R} \tag{31.77}$$

Note that we are allowing the parameter θ to be vector-valued, as well as the observation $\boldsymbol{y}$. We refer to $h(y)$ as the *base function*, to $a(\theta)$ as the *log-partition function*, and to $T(y)$ as the sufficient statistic. We showed in Table 5.1 how different distributions can be obtained as special cases of (31.76) through the selection of $\{h(y), T(y), a(\theta)\}$. We continue with the general description (31.76) and derive the maximum-likelihood estimator for θ.

Assume we have a collection of N independent and identically distributed realizations $\{y_n\}$, arising from the exponential distribution (31.76) with unknown parameter vector, θ. The joint pdf (or likelihood function) of the observations is given by

$$\begin{aligned} f_{\boldsymbol{y}_1,\ldots,\boldsymbol{y}_N}(y_1,\ldots,y_N;\theta) &= \prod_{n=1}^{N} h(y_n) e^{\theta^{\mathsf{T}} T(y_n) - a(\theta)} \\ &= \left(\prod_{n=1}^{N} h(y_n) \right) e^{-Na(\theta)} \exp\left\{ \sum_{n=1}^{N} \theta^{\mathsf{T}} T(y_n) \right\} \end{aligned} \tag{31.78}$$

so that the log-likelihood function is

$$\ell(y_1,\ldots,y_N;\theta) \;=\; \sum_{n=1}^{N} \ln h(y_n) \;-\; Na(\theta) \;+\; \sum_{n=1}^{N} \theta^{\mathsf{T}} T(y_n)) \tag{31.79}$$

It was argued after (5.87) that this function is concave in the parameter θ. Computing the gradient vector relative to θ and setting it to zero at $\theta = \widehat{\theta}$ gives

$$-\,N \nabla_{\theta^{\mathsf{T}}}\, a(\widehat{\theta}) + \sum_{n=1}^{N} T(y_n) = 0 \tag{31.80}$$

and, hence, the ML estimate $\widehat{\theta}$ is found by solving the equation:

$$\boxed{\nabla_{\theta^{\mathsf{T}}}\, a(\widehat{\theta}_{\text{ML}}) = \frac{1}{N} \sum_{n=1}^{N} T(y_n)} \tag{31.81}$$

This is an important conclusion. It shows that in order to estimate the parameter θ, it is sufficient to know the average of the values of $\{T(y_n)\}$; the individual measurements $\{y_n\}$ are not needed. We say that the function $T(\boldsymbol{y})$ plays the role of a *sufficient* statistic for $\boldsymbol{y}$, or that the sample average of the values $\{T(y_n)\}$ is sufficient knowledge for the problem of estimating θ — recall the comments on the concept of a sufficient statistics at the end of Chapter 5. We comment further on this concept in Example 31.10.

Example 31.9 **(Gaussian distribution)** Let us illustrate how construction (31.81) reduces to known results by considering the first row of Table 5.1, which corresponds to the Gaussian distribution. In this case we have

$$\theta = \begin{bmatrix} \mu/\sigma_y^2 \\ -1/2\sigma_y^2 \end{bmatrix}, \quad T(y) = \begin{bmatrix} y \\ y^2 \end{bmatrix}, \quad a(\theta) = -\frac{1}{2}\ln(-2\theta_2) - \frac{\theta_1^2}{4\theta_2} \tag{31.82}$$

so that

$$\nabla_{\theta^{\mathsf{T}}}\, a(\theta) \;=\; \begin{bmatrix} \partial a(\theta)/\partial\theta_1 \\ \partial a(\theta)/\partial\theta_2 \end{bmatrix} \;=\; \begin{bmatrix} -\dfrac{\theta_1}{2\theta_2} \\ -\dfrac{1}{2\theta_2} + \frac{\theta_1^2}{4\theta_2^2} \end{bmatrix} \;=\; \begin{bmatrix} \mu \\ \sigma_y^2 + \mu^2 \end{bmatrix} \tag{31.83}$$

It follows from (31.81) that

$$\begin{bmatrix} \widehat{\mu}_{\rm ML} \\ \widehat{\sigma}^2_{y,{\rm ML}} + \widehat{\mu}^2_{\rm ML} \end{bmatrix} \;=\; \begin{bmatrix} \dfrac{1}{N}\displaystyle\sum_{n=1}^{N} y_n \\ \dfrac{1}{N}\displaystyle\sum_{n=1}^{N} y_n^2 \end{bmatrix} \tag{31.84}$$

These expressions agree with the earlier result (31.23b) since they lead to the estimates:

$$\widehat{\mu}_{\rm ML} \;=\; \frac{1}{N}\sum_{n=1}^{N} y_n, \qquad \widehat{\sigma}^2_{y,{\rm ML}} \;=\; \frac{1}{N}\sum_{n=1}^{N} y_n^2 \;-\; \widehat{\mu}^2_{\rm ML} \tag{31.85}$$

Example 31.10 **(Sufficient statistic)** The function $T(y)$ that appears in (31.81) plays the important role of a *sufficient statistic*. Generally, any function of an observation, $T(y)$, is a statistic. However, this concept is most relevant when the statistic happens to be *sufficient*. The statistic is said to be sufficient for estimating a parameter θ from a measurement $\boldsymbol{y}$ if the conditional distribution of $\boldsymbol{y}$ given $T(\boldsymbol{y})$ does not depend on θ:

$$\begin{aligned} &f_{\boldsymbol{y}|T(\boldsymbol{y})}(y|T(y)) \;\text{ does not depend on } \theta \\ &\Longrightarrow\; T(y) \text{ is a sufficient statistic} \end{aligned} \tag{31.86}$$

A key factorization theorem in statistics states that $T(y)$ is a sufficient statistic for θ if, and only if, the pdf $f_{\boldsymbol{y}}(y;\theta)$ can be factored in the form — recall expression (5.138) and Prob. 5.11:

$$f_{\boldsymbol{y}}(y;\theta) \;=\; h(y)\, g(T(y);\theta) \tag{31.87}$$

That is, the pdf can be factored as the product of two nonnegative functions, $h(\cdot)$ and $g(\cdot;\theta)$, such that $h(y)$ depends solely on y and $g(T(y);\theta)$, which depends on θ, depends on the observation only through $T(y)$. If we examine the form of the exponential distribution (31.76) we observe that it can be written in this form with

$$g(T(y);\theta) = e^{\theta^{\mathsf{T}} T(y) - a(\theta)} \tag{31.88}$$

More commonly, one often discusses sufficiency in the context of estimating θ from a *collection* of independent and identically distributed observations $\{y_n\}$ arising from the distribution $f_{\boldsymbol{y}}(y)$. If we consider N such observations then their joint pdf is given by

$$\begin{aligned} f_{\boldsymbol{y}_1,\boldsymbol{y}_2,\ldots,\boldsymbol{y}_N}(y_1,y_2,\ldots,y_N;\theta) &= \prod_{n=1}^{N} h(y_n) e^{\theta^{\mathsf{T}} T(y_n) - a(\theta)} \\ &= \left(\prod_{n=1}^{N} h(y_n)\right) \exp\left\{\theta^{\mathsf{T}}\left(\sum_{n=1}^{N} T(y_n)\right) - N a(\theta)\right\} \end{aligned} \tag{31.89}$$

which is seen to be of the desired factored form (31.87) with

$$g\left(\sum_{n=1}^{N} T(y_n);\theta\right) = \exp\left\{\theta^{\mathsf{T}}\left(\sum_{n=1}^{N} T(y_n)\right) - Na(\theta)\right\} \tag{31.90}$$

Therefore, the maximum likelihood estimate for θ from a collection of independent and identically distributed observations $\{y_1,\ldots,y_N\}$ can be determined by relying solely on knowledge of the sufficient statistic given by the sum

$$\text{sufficient statistic} \triangleq \sum_{n=1}^{N} T(y_n) \tag{31.91}$$

rather than on the individual observations. We already observe this feature in expression (31.81).

31.5 CRAMER-RAO LOWER BOUND

The derivation in Sec. 31.2 provides examples of ML estimators that can be biased or unbiased. It also shows examples of estimators that differ in their mean-square error. Ideally, we would like our estimators to be unbiased and to have the smallest mean-square error (or variance) possible. The Cramer-Rao bound is a useful result in that regard. It provides a lower bound on the variance for any estimator (whether of ML-type or not). Estimators that meet the Cramer-Rao bound are said to be *efficient* since no other estimator will be able to deliver a smaller variance or mean-square error. Efficient estimators that are also unbiased belong to the class of minimum-variance unbiased estimators (MVUE) because they are *both* unbiased and attain the smallest variance. It turns out that, under certain regularity conditions, the ML estimators are *asymptotically* unbiased and efficient as the number of observations grows, i.e., as $N \to \infty$.

Thus, consider the problem of estimating an unknown constant parameter θ, which may be scalar or vector-valued, from an observation vector $\boldsymbol{y}$. For generality, we describe the Cramer-Rao bound for the case of vector parameters, $\theta \in \mathbb{R}^M$. We denote the individual entries of θ by $\{\theta_m\}$ and denote the corresponding estimation error by

$$\widetilde{\boldsymbol{\theta}}_m \triangleq \theta_m - \widehat{\boldsymbol{\theta}}_m \tag{31.92}$$

where the estimators $\{\widehat{\boldsymbol{\theta}}_m\}$ are all assumed to be *unbiased*, i.e., $\mathbb{E}\,\widehat{\boldsymbol{\theta}}_m = \theta_m$. The error covariance matrix is denoted by

$$R_{\tilde{\theta}} = \mathbb{E}\,(\theta - \widehat{\boldsymbol{\theta}})(\theta - \widehat{\boldsymbol{\theta}})^{\mathsf{T}} = \mathbb{E}\,\widetilde{\boldsymbol{\theta}}\widetilde{\boldsymbol{\theta}}^{\mathsf{T}} \tag{31.93}$$

where $\widehat{\boldsymbol{\theta}} = \text{col}\{\widehat{\boldsymbol{\theta}}_m\}$ and $\widetilde{\boldsymbol{\theta}} = \theta - \widehat{\boldsymbol{\theta}}$.

31.5.1 Fisher Information Matrix

We associate with the inference problem an $M \times M$ *Fisher information matrix*, whose entries are constructed as follows. The $(n,m)-$th entry is defined in terms of the (negative) expectation of the second-order partial derivative of the log-likelihood function relative to the parameter entries:

$$\boxed{[F(\theta)]_{n,m} \stackrel{\Delta}{=} -\mathbb{E}\left(\frac{\partial^2 \ln f_{\boldsymbol{y}}(y;\theta)}{\partial\theta_n\partial\theta_m}\right), \quad n,m = 1,2,\ldots,M} \tag{31.94}$$

or, in matrix form and using the Hessian matrix notation:

$$F(\theta) \stackrel{\Delta}{=} -\mathbb{E}\,\nabla_\theta^2 \ln f_{\boldsymbol{y}}(y;\theta) \tag{31.95}$$

The derivatives in these expressions are evaluated at the true value for the parameter θ. The Fisher information matrix helps reflect how much information the distribution of $\boldsymbol{y}$ conveys about θ. The above expressions define the Fisher information matrix relative to a *single* observation $\boldsymbol{y}$.

31.5.2 Score Function

Expression (31.94) assumes that the log-likelihood function is twice-differentiable with respect to θ. Under a regularity condition that integration and differentiation operations are exchangeable (recall the discussion in Appendix 16.A), there is an equivalent form for the Fisher matrix as the covariance matrix of the so-called score function, namely,

$$\boxed{F(\theta) = \mathbb{E}\,\mathcal{S}(\theta)\mathcal{S}^{\mathsf{T}}(\theta)} \tag{31.96}$$

where the *score function* is defined in terms of the gradient vector with respect to θ) (i.e.,only first-order derivatives are involved):

$$\mathcal{S}(\theta) \stackrel{\Delta}{=} \nabla_{\theta^{\mathsf{T}}} \ln f_{\boldsymbol{y}}(y;\theta) \tag{31.97}$$

Proof of (31.96): Note first that

$$\begin{aligned}
\nabla_\theta^2 \ln f_{\boldsymbol{y}}(y;\theta) &= \nabla_{\theta^{\mathsf{T}}}\left(\nabla_\theta \ln f_{\boldsymbol{y}}(y;\theta)\right)\\
&= \nabla_{\theta^{\mathsf{T}}}\left(\frac{\nabla_\theta f_{\boldsymbol{y}}(y;\theta)}{f_{\boldsymbol{y}}(y;\theta)}\right)\\
&= \frac{f_{\boldsymbol{y}}(y;\theta)\,\nabla_\theta^2 f_{\boldsymbol{y}}(y;\theta) - \nabla_{\theta^{\mathsf{T}}} f_{\boldsymbol{y}}(y;\theta)\,\nabla_\theta f_{\boldsymbol{y}}(y;\theta)}{f_{\boldsymbol{y}}^2(y;\theta)}\\
&= \frac{\nabla_\theta^2 f_{\boldsymbol{y}}(y;\theta)}{f_{\boldsymbol{y}}(y;\theta)} - \mathcal{S}(\theta)\mathcal{S}^{\mathsf{T}}(\theta)
\end{aligned} \tag{31.98}$$

Consequently, if we denote the domain of y by $y \in \mathcal{Y}$,

$$\begin{aligned}
\mathbb{E}\,\mathcal{S}(\theta)\mathcal{S}^{\mathsf{T}}(\theta) &\overset{(31.98)}{=} -\mathbb{E}\,\nabla_{\theta}^2 \ln f_{\boldsymbol{y}}(y;\theta) + \mathbb{E}\left(\frac{\nabla_{\theta}^2 f_{\boldsymbol{y}}(y;\theta)}{f_{\boldsymbol{y}}(y;\theta)}\right) \\
&= -\mathbb{E}\,\nabla_{\theta}^2 \ln f_{\boldsymbol{y}}(y;\theta) \;+\; \int_{y\in\mathcal{Y}} \nabla_{\theta}^2 f_{\boldsymbol{y}}(y;\theta)dy \\
&\overset{(a)}{=} -\mathbb{E}\,\nabla_{\theta}^2 \ln f_{\boldsymbol{y}}(y;\theta) + \nabla_{\theta}^2 \underbrace{\left(\int_{y\in\mathcal{Y}} f_{\boldsymbol{y}}(y;\theta)dy\right)}_{=1} \\
&= -\mathbb{E}\,\nabla_{\theta}^2 \ln f_{\boldsymbol{y}}(y;\theta) \qquad (31.99)
\end{aligned}$$

where step (a) assumes that the operations of integration and differentiation can be exchanged.

■

It is explained in Appendix 31.A that under two *regularity* conditions stated in (31.229)–(31.230), the log-likelihood function satisfies for any θ:

$$\mathbb{E}\left(\frac{\partial \ln f_{\boldsymbol{y}}(y;\theta)}{\partial \theta_m}\right) = 0, \quad m = 1, 2, \ldots, M \qquad (31.100)$$

which implies that the score function exists, is bounded, and has zero mean

$$\mathbb{E}\,\mathcal{S}(\theta) = \mathbb{E}\,\nabla_{\theta} \ln f_{\boldsymbol{y}}(y;\theta) = 0 \qquad (31.101)$$

It follows that the Fisher information matrix defined by (31.96) is the actual covariance matrix of the score function. We obtain from (31.96) that we also have:

$$[F(\theta)]_{n,m} \;=\; \mathbb{E}\left(\frac{\partial \ln f_{\boldsymbol{y}}(y;\theta)}{\partial \theta_n}\,\frac{\partial \ln f_{\boldsymbol{y}}(y;\theta)}{\partial \theta_m}\right), \quad n, m = 1, 2, \ldots, M \qquad (31.102)$$

This expression again defines the Fisher information matrix relative to a *single* observation.

Example 31.11 **(Diagonal covariance matrix)** Consider a vector Gaussian distribution with mean $\mu \in \mathbb{R}^P$ and diagonal covariance matrix

$$\Sigma_{\boldsymbol{y}} = \text{diag}\left\{\sigma_1^2, \sigma_2^2, \ldots, \sigma_P^2\right\} \qquad (31.103)$$

We denote the individual entries of the mean vector by

$$\mu = \text{col}\left\{\mu_1, \mu_2, \ldots, \mu_P\right\} \qquad (31.104)$$

We can write the pdf in the form:

$$f_{\boldsymbol{y}}(y) = \frac{1}{\sqrt{(2\pi)^P}}\,\frac{1}{\sqrt{\prod_{p=1}^{P}\sigma_p^2}}\,\prod_{p=1}^{P}\exp\left\{-\frac{1}{2\sigma_p^2}(y_p-\mu_p)^2\right\} \qquad (31.105)$$

We wish to evaluate the Fisher information matrix of this distribution relative to its mean and variance parameters. First, note that the likelihood function is given by

$$\ln f_{\boldsymbol{y}}(y) = -\frac{P}{2}\ln(2\pi) - \frac{1}{2}\sum_{p=1}^{P}\ln\sigma_p^2 - \sum_{p=1}^{P}\frac{1}{2\sigma_p^2}(y_p-\mu_p)^2 \qquad (31.106)$$

so that

$$\partial \ln f_{\boldsymbol{y}}(y)/\partial \mu_p = \frac{1}{\sigma_p^2}(y_p - \mu_p) \tag{31.107}$$

$$\partial \ln f_{\boldsymbol{y}}(y)/\partial \sigma_p^2 = -\frac{1}{2\sigma_p^2} + \frac{1}{2\sigma_p^4}(y_p - \mu_p)^2 \tag{31.108}$$

The Fisher information matrix in this case has dimensions $2P \times 2P$. Let us order the parameters in θ with the $\{\mu_p\}$ coming first followed by the $\{\sigma_p^2\}$, for $p = 1, 2, \ldots, P$:

$$\theta = \Big\{\mu_1, \ldots, \mu_P, \sigma_1^2, \ldots, \sigma_P^2\Big\} \tag{31.109}$$

Then, the diagonal entries of the Fisher information matrix are given by

$$\begin{aligned}
[F(\mu, \Sigma)]_{p,p} &= \mathbb{E}\left(\frac{\partial \ln f_{\boldsymbol{y}}(y)}{\partial \mu_p}\,\frac{\partial \ln f_{\boldsymbol{y}}(y)}{\partial \mu_p}\right)\\
&= \frac{1}{\sigma_p^4}\mathbb{E}\,(\boldsymbol{y}_p - \mu_p)^2\\
&= 1/\sigma_p^2
\end{aligned} \tag{31.110}$$

and

$$\begin{aligned}
[F(\mu, \Sigma)]_{p+P,p+P} &= \mathbb{E}\left(\frac{\partial \ln f_{\boldsymbol{y}}(y)}{\partial \sigma_p^2}\,\frac{\partial \ln f_{\boldsymbol{y}}(y)}{\partial \sigma_p^2}\right)\\
&= \mathbb{E}\left(-\frac{1}{2\sigma_p^2} + \frac{1}{2\sigma_p^4}(\boldsymbol{y}_p - \mu_p)^2\right)^2\\
&= \frac{1}{4\sigma_p^4} + \frac{1}{4\sigma_p^8}\mathbb{E}\,(\boldsymbol{y}_p - \mu_p)^4 - \frac{1}{2\sigma_p^6}\mathbb{E}\,(\boldsymbol{y}_p - \mu_p)^2\\
&= \frac{1}{4\sigma_p^4} + \frac{3\sigma_p^4}{4\sigma_p^8} - \frac{\sigma_p^2}{2\sigma_p^6}\\
&= \frac{1}{2\sigma_p^4}
\end{aligned} \tag{31.111}$$

where we used the fact that, for a Gaussian random variable $\boldsymbol{x}$ with mean $\bar{x}$ and variance σ_x^2, it holds that $\mathbb{E}\,(\boldsymbol{x} - \bar{x})^4 = 3\sigma_x^4$. On the other hand, for $p \neq p' = 1, 2, \ldots, P$, the off-diagonal entries of the Fisher information matrix are given by

$$\begin{aligned}
[F(\mu, \Sigma)]_{p,p'} &= \mathbb{E}\left(\frac{\partial \ln f_{\boldsymbol{y}}(y)}{\partial \mu_p}\,\frac{\partial \ln f_{\boldsymbol{y}}(y)}{\partial \mu_p'}\right)\\
&= \frac{1}{\sigma_p^2\sigma_{p'}^2}\mathbb{E}\,(\boldsymbol{y}_p - \mu_p)(\boldsymbol{y}_{p'} - \mu_{p'})\\
&= 0
\end{aligned} \tag{31.112}$$

since $\{\boldsymbol{y}_p, \boldsymbol{y}_{p'}\}$ are uncorrelated (actually independent) random variables due to the diagonal covariance structure. Likewise, for $p \neq p'$:

$$\begin{aligned}
[F(\mu,\Sigma)]_{p+P,p'+P} &= \mathbb{E}\left(\frac{\partial \ln f_{\boldsymbol{y}}(y)}{\partial \sigma_p^2}\,\frac{\partial \ln f_{\boldsymbol{y}}(y)}{\partial \sigma_{p'}^2}\right)\\
&= \mathbb{E}\left(-\frac{1}{2\sigma_p^2}+\frac{1}{2\sigma_p^4}(\boldsymbol{y}_p-\mu_p)^2\right)\left(-\frac{1}{2\sigma_{p'}^2}+\frac{1}{2\sigma_{p'}^4}(\boldsymbol{y}_{p'}-\mu_{p'})^2\right)\\
&= \frac{1}{4\sigma_p^2\sigma_{p'}^2}-\frac{\sigma_{p'}^2}{4\sigma_p^2\sigma_{p'}^4}-\frac{\sigma_p^2}{4\sigma_p^4\sigma_{p'}^2}+\frac{\sigma_p^2\sigma_{p'}^2}{4\sigma_p^4\sigma_{p'}^4}\\
&= \frac{1}{4\sigma_p^2\sigma_{p'}^2}-\frac{1}{4\sigma_p^2\sigma_{p'}^2}-\frac{1}{4\sigma_p^2\sigma_{p'}^2}+\frac{1}{4\sigma_p^2\sigma_{p'}^2}\\
&= 0
\end{aligned}\tag{31.113}$$

while

$$\begin{aligned}
[F(\mu,\Sigma)]_{p,p'+P} &= \mathbb{E}\left(\frac{\partial \ln f_{\boldsymbol{y}}(y)}{\partial \mu_p}\,\frac{\partial \ln f_{\boldsymbol{y}}(y)}{\partial \sigma_{p'}^2}\right)\\
&= \frac{1}{\sigma_p^2\sigma_{p'}^2}\mathbb{E}\left\{(\boldsymbol{y}_p-\mu_p)\left(-\frac{1}{\sigma_{p'}^2}+\frac{1}{2\sigma_{p'}^4}(\boldsymbol{y}_{p'}-\mu_{p'})^2\right)\right\}\\
&= 0
\end{aligned}\tag{31.114}$$

and

$$\begin{aligned}
[F(\mu,\Sigma)]_{p+P,p'} &= \mathbb{E}\left(\frac{\partial \ln f_{\boldsymbol{y}}(y)}{\partial \sigma_p^2}\,\frac{\partial \ln f_{\boldsymbol{y}}(y)}{\partial \mu_{p'}}\right)\\
&= \frac{1}{\sigma_{p'}^2\sigma_p^2}\mathbb{E}\left\{(\boldsymbol{y}_{p'}-\mu_{p'})\left(-\frac{1}{\sigma_p^2}+\frac{1}{2\sigma_p^4}(\boldsymbol{y}_p-\mu_p)^2\right)\right\}\\
&= 0
\end{aligned}\tag{31.115}$$

We conclude that the Fisher information matrix in this case is diagonal and given by

$$F(\mu,\Sigma) = \text{diag}\left\{\frac{1}{\sigma_1^2},\frac{1}{\sigma_2^2},\ldots,\frac{1}{\sigma_P^2},\frac{1}{2\sigma_1^4},\frac{1}{2\sigma_2^4},\ldots,\frac{1}{2\sigma_P^4}\right\}\tag{31.116}$$

31.5.3 Cramer-Rao Bound

We list the Cramer-Rao bound for the two situations of unbiased and biased estimators, and also show the correction that needs to be made to the Fisher information matrix when a multitude of observations are used to determine the ML estimator rather than a single observation.

Unbiased estimators

The Cramer-Rao lower bound for *unbiased* estimators amounts to the statement that — see Appendix 31.A for one derivation:

$$\mathbb{E}\,\widetilde{\boldsymbol{\theta}}_m^2 \geq \left[F^{-1}(\theta)\right]_{m,m}\tag{31.117}$$

in terms of the $m-$the diagonal entry of the inverse of the Fisher matrix. The result can also be rewritten in terms of the variance of the individual entries as follows:

$$\text{var}(\widehat{\boldsymbol{\theta}}_m) \;\geq [F^{-1}(\theta)]_{m,m} \tag{31.118}$$

or in terms of the aggregate covariance matrix of the estimator

$$\boxed{R_{\widehat{\theta}} \;\geq\; F^{-1}(\theta)} \qquad (\textbf{unbiased vector estimators}) \tag{31.119}$$

where the notation $A \geq B$ for two nonnegative-definite matrices means that $A - B \geq 0$. In the special case in which θ happens to be a scalar, the Cramer-Rao lower bound (31.117) can be rewritten equivalently in the forms:

$$\mathbb{E}\,\widetilde{\boldsymbol{\theta}}^2 \geq -\left(\mathbb{E}\,\frac{\partial^2 \ln f_{\boldsymbol{y}}(y;\theta)}{\partial^2\theta}\right)^{-1} \tag{31.120a}$$

$$= \left(\mathbb{E}\left(\frac{\partial \ln f_{\boldsymbol{y}}(y;\theta)}{\partial\theta}\right)^2\right)^{-1} \tag{31.120b}$$

or, more compactly,

$$\boxed{\text{var}(\widehat{\boldsymbol{\theta}}_m) \;\geq 1/F(\theta)} \qquad (\textbf{unbiased scalar estimators}) \tag{31.121}$$

Biased estimators

When the estimator is *biased*, with mean denoted by $g(\theta) = \mathbb{E}\,\widehat{\boldsymbol{\theta}}$ for some function of θ, statement (31.120a) for the Cramer-Rao bound for scalar parameters is modified as follows — see Appendix 31.A:

$$\mathbb{E}\,(g(\theta) - \widehat{\boldsymbol{\theta}})^2 \geq -\left(\mathbb{E}\,\frac{\partial^2 \ln f_{\boldsymbol{y}}(y;\theta)}{\partial^2\theta}\right)^{-1}\left(\frac{\partial g(\theta)}{\partial\theta}\right)^2 \tag{31.122}$$

or, equivalently, in terms of the variance of the estimator:

$$\boxed{\text{var}(\widehat{\boldsymbol{\theta}}) \geq \frac{(\partial g(\theta)/\partial\theta)^2}{F(\theta)}} \qquad (\textbf{biased scalar estimators}) \tag{31.123}$$

The special case $g(\theta) = \theta$ reduces this expression to (31.120a). For vector parameters θ, the corresponding relation becomes

$$\mathbb{E}\,(g(\theta) - \widehat{\boldsymbol{\theta}})(g(\theta) - \widehat{\boldsymbol{\theta}})^{\mathsf{T}} \geq (\nabla_{\theta^{\mathsf{T}}} g(\theta))\; F^{-1}(\theta)\, \nabla_{\theta}\, g(\theta) \tag{31.124}$$

or, equivalently, in terms of the covariance matrix of the estimator:

$$\boxed{R_{\widehat{\theta}} \;\geq\; \Big(\nabla_{\theta^{\mathsf{T}}}\, g(\theta)\Big)\, F^{-1}(\theta)\, \Big(\nabla_{\theta}\, g(\theta)\Big)} \qquad (\textbf{biased vector estimators}) \tag{31.125}$$

Multiple observations

The Cramer-Rao bounds listed so far are expressed in terms of the Fisher information matrix for a *single* observation. If a collection of N independent and identically distributed realizations $\{y_n\}$ are used to determine the estimator $\widehat{\boldsymbol{\theta}}$, as is usually the case, then $F(\theta)$ will need to be scaled by N. This is because the Fisher information matrix that is associated with N observations will be

$$\begin{aligned} F_N(\theta) &\overset{\Delta}{=} -\mathbb{E}\,\nabla_\theta^2\left(\ln\prod_{n=1}^{N} f_{\boldsymbol{y}}(y_n;\theta)\right) \\ &= -\mathbb{E}\,\nabla_\theta^2\left(\sum_{n=1}^{N}\ln f_{\boldsymbol{y}}(y_n;\theta)\right) \\ &= -\sum_{n=1}^{N}\mathbb{E}\left(\nabla_\theta^2\ln f_{\boldsymbol{y}}(y_n;\theta)\right) \\ &= -N\mathbb{E}\left(\nabla_\theta^2\ln f_{\boldsymbol{y}}(y_n;\theta)\right) \end{aligned} \tag{31.126}$$

where in the last step we used the fact that the observations are identically distributed. It follows that

$$\boxed{F_N(\theta) \;=\; NF(\theta)} \qquad \text{(when } N \text{ observations are used)} \tag{31.127}$$

31.5.4 Efficiency and Consistency

The maximum likelihood estimator exhibits several important properties, which have been studied at great length in the literature. We list three classical properties here without proof; derivations can be found in the references listed at the end of the chapter. It can be shown, again under some reasonable regularity conditions, that the ML estimator satisfies the following three conclusions:

(a) (**Consistency**). An estimator $\widehat{\boldsymbol{\theta}}_N$, based on N observations, for some unknown θ is said to be consistent if $\widehat{\boldsymbol{\theta}}_N$ converges to θ in probability, meaning that for any $\epsilon > 0$:

$$\lim_{N\to\infty}\ \mathbb{P}(|\theta - \widehat{\boldsymbol{\theta}}_N| > \epsilon) = 0, \qquad \text{(convergence in probability)} \tag{31.128}$$

Maximum-likelihood estimators satisfy this property and are therefore *consistent.*

(b) (**Asymptotic normality**). The random variable $\sqrt{N}\,\widetilde{\boldsymbol{\theta}}_N$ converges *in distribution* to a Gaussian pdf with mean zero and covariance matrix $F^{-1}(\theta)$, written as:

$$\boxed{\sqrt{N}(\theta - \widehat{\boldsymbol{\theta}}_N) \ \xrightarrow{\mathrm{d}}\ \mathcal{N}_{\widetilde{\boldsymbol{\theta}}_N}\left(0,\, F^{-1}(\theta)\right), \quad \text{as } N\to\infty} \tag{31.129}$$

where θ is the true unknown parameter. It follows that the ML estimator is asymptotically unbiased as well.

(c) (**Efficiency**). Maximum-likelihood estimators are asymptotically efficient, meaning that their covariance matrix $R_{\widehat{\theta}}$ attains the Cramer-Rao bound (31.119) in the limit as $N \rightarrow \infty$. There are also situations in which maximum-likelihood estimators attain the bound even for finite sample sizes, N — see the next example. The property of asymptotic efficiency follows from asymptotic normality since the latter implies that

$$\widehat{\boldsymbol{\theta}}_N \xrightarrow{\text{d}} \mathcal{N}_{\widehat{\boldsymbol{\theta}}_N}\left(\theta,\, F_N^{-1}(\theta)\right) \tag{31.130}$$

where $F_N(\theta) = NF(\theta)$.

Example 31.12 (**Estimating a DC level**) Consider a collection of N independent and identically distributed random measurements:

$$\boldsymbol{y}_n = \theta + \boldsymbol{v}_n \tag{31.131}$$

where $\boldsymbol{v}_n$ is Gaussian noise with zero mean and variance σ_v^2, while θ is an unknown constant parameter (which amounts to the mean of $\boldsymbol{y}_n$). The likelihood function is given by

$$f_{\boldsymbol{y}_1,\ldots,\boldsymbol{y}_N}(y_1,\ldots,y_N;\theta) = \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi\sigma_v^2}} \exp\left\{-\frac{1}{2\sigma_v^2}(y_n-\theta)^2\right\} \tag{31.132}$$

so that the log-likelihood function is

$$\ell(y_1,\ldots,y_N;\theta) \;=\; -\frac{N}{2}\ln(2\pi\sigma_v^2) \;-\; \frac{1}{2\sigma_v^2}\sum_{n=1}^{N}(y_n-\theta)^2 \tag{31.133}$$

We conclude from setting the derivative to zero that the maximum-likelihood estimate for θ is given by

$$\widehat{\theta}_{\text{ML}} \;=\; \frac{1}{N}\sum_{n=1}^{N} y_n \tag{31.134}$$

which is clearly unbiased with error variance given by

$$\begin{aligned}
\mathbb{E}\,(\theta - \widehat{\boldsymbol{\theta}}_{\text{ML}})^2 &= \mathbb{E}\left(\theta - \frac{1}{N}\sum_{n=1}^{N}\boldsymbol{y}_n\right)^2 \\
&= \mathbb{E}\left(\theta - \frac{1}{N}\sum_{n=1}^{N}(\theta+\boldsymbol{v}_n)\right)^2 \\
&= \frac{1}{N^2}\sum_{n=1}^{N}\mathbb{E}\,\boldsymbol{v}_n^2 \;=\; \sigma_v^2/N
\end{aligned} \tag{31.135}$$

It can be verified that the two regularity conditions (31.229)–(31.230) hold in this case — see Prob. 31.18. We compute next the Cramer-Rao lower bound. For this purpose, we first evaluate

$$\frac{\partial^2 \ell(y_1,\ldots,y_N;\theta)}{\partial^2\theta} \;=\; -N/\sigma_v^2 \tag{31.136}$$

so that the lower bound is given by

$$-\left(\mathbb{E}\,\frac{\partial^2 \ln f(y_1,\ldots,y_N;\theta)}{\partial^2\theta}\right)^{-1} = \sigma_v^2/N \tag{31.137}$$

which agrees with the error variance found in (31.135). Therefore, the maximum-likelihood estimator in this case is *efficient.*

Observe from (31.135) that the error variance in this example decays at the rate of $1/N$, in inverse proportion to the sample size. There are situations, where the error variance can decay exponentially fast. One such example is given in the commentaries at the end of the chapter in (31.206); that example deals with the same problem of recovering θ from noisy measurements under Gaussian noise, except that the unknown θ is constrained to being an integer.

Example 31.13 (**Bias, efficiency, and consistency**) Let θ be an unknown scalar parameter that we wish to estimate. Let $\widehat{\boldsymbol{\theta}}_N$ be an estimator for θ that is based on N observations. We encountered in our presentation of ML estimators three important properties related to the notions of bias, efficiency, and consistency. These properties apply to other types of estimators as well. We describe them together in this example for ease of comparison:

(a) An estimator $\widehat{\boldsymbol{\theta}}_N$ is said to be *unbiased* if $\mathbb{E}\,\widehat{\boldsymbol{\theta}}_N = \theta$. It is said to be asymptotically unbiased if this equality holds in the limit as $N \to \infty$. Thus, the notion of bias relates to a property about the mean (or the first-order moment) of the distribution of the random variable $\theta - \widehat{\boldsymbol{\theta}}_N$.

(b) An estimator $\widehat{\boldsymbol{\theta}}_N$ is said to be *efficient* if its variance attains the Cramer-Rao bound (31.123). The estimator is said to be asymptotically efficient if it attains the Cramer-Rao bound in the limit as $N \to \infty$. Thus, the notion of efficiency relates to a property about the second-order moment of the distribution of the random variable $\theta - \widehat{\boldsymbol{\theta}}_N$.

(c) An estimator $\widehat{\boldsymbol{\theta}}_N$ is said to be *consistent* if $\widehat{\boldsymbol{\theta}}_N$ converges to θ in probability, meaning that for any $\epsilon > 0$:

$$\lim_{N\to\infty} \mathbb{P}(|\theta - \widehat{\boldsymbol{\theta}}_N| > \epsilon) = 0, \qquad \text{(convergence in probability)} \tag{31.138}$$

Thus, the notion of consistency relates to a property about the limiting distribution of the random variable $\theta - \widehat{\boldsymbol{\theta}}_N$, which tends to concentrate around θ.

31.6 MODEL SELECTION

Once a family of density distributions $f_{\boldsymbol{y}}(y;\theta)$ is selected, parameterized by some θ, the ML formulation allows us to estimate θ from observations $\{y_n\}$. This construction presumes that the designer has already selected what family of distributions to use.

In many situations of interest, the designer will be faced with the additional task of selecting the family of distributions from among a collection of possible choices. Each family k will be parameterized by its own θ_k. In these cases, the designer will need to **(a)** select the best family of distributions from among the available choices and, moreover, **(b)** estimate the optimal parameter θ for the selected family.

There are several criteria that can be used to solve this problem, with the log-likelihood function and the ML estimate playing a prominent role in the solution. In this section, we describe the Bayesian Information Criterion (BIC), the Akaike Information Criterion (AIC), and the minimum description length (MDL) criterion for selecting among density models. They all serve as "goodness-of-fit" tests and guide the selection of the best model. In Sec. 31.6.5 we describe another method for choosing among models that is based on the *cross-validation* technique; this technique has found widespread application in learning and inference problems and leads to good performance, often under weaker conditions than needed for the BIC, AIC, and MDL methods of this section.

31.6.1 Motivation and Overfitting

Consider a collection of N data measurements $\{y_1, y_2, \ldots, y_N\}$ and a collection of K probability density models with parameters $\{\theta_1, \theta_2, \ldots, \theta_K\}$. Each model (or class or family) θ_k amounts to assuming a probability density function parameterized by θ_k for the observation y, say, according to

$$f_{\boldsymbol{y}}(y;\theta_k), \quad (k\text{−th pdf model with parameter } \theta_k) \tag{31.139}$$

For each k, the size of θ_k is denoted by M_k and it can vary over k. For example, θ_1 could be representing a Gaussian distribution with unknown mean μ_1 but known variance σ_1^2, in which case

$$f_{\boldsymbol{y}}(y;\theta_1) \sim \mathcal{N}_{\boldsymbol{y}}(\mu_1, \sigma_1^2), \quad \theta_1 = \{\mu_1\} \tag{31.140}$$

This corresponds to a problem where the number of parameters to be selected is $M_1 = 1$. The second model θ_2 could be representing a second Gaussian distribution with both unknown mean *and* variance:

$$f_{\boldsymbol{y}}(y;\theta_2) \sim \mathcal{N}_{\boldsymbol{y}}(\mu_2, \sigma_2^2), \quad \theta_2 = \{\mu_2, \sigma_2^2\} \tag{31.141}$$

In this case, we would need to learn two parameters with $M_2 = 2$. Likewise, θ_3 could correspond to a third model where we are trying to fit the sum of two Gaussian distributions, say,

$$f_{\boldsymbol{y}}(y;\theta_3) \sim \pi\mathcal{N}_{\boldsymbol{y}}(\mu_a, \sigma_a^2) + (1-\pi)\mathcal{N}_{\boldsymbol{y}}(\mu_b, \sigma_b^2) \tag{31.142a}$$

$$\theta_3 = \{\mu_a, \mu_b, \sigma_a^2, \sigma_b^2, \pi\} \tag{31.142b}$$

where $\pi \in ()0, 1)$. In this case, we would need to learn five parameters with $M_3 = 5$.

Overfitting

Generally, the more complex the model θ_k is, the more parameters it will involve (i.e., the larger the value of M_k will be). While complex models can be expected to fit the data better because of the degree of freedom that results from using a larger number of parameters, they are nevertheless less desirable in practice. We are going to learn later in this text that complex models lead to *overfitting*;

a property that we should avoid. Overfitting essentially amounts to using more complex models to fit the data than is necessary. This can be illustrated by means of an example. Assume each y_n is a scalar measurement that arises from small perturbations to a quadratic function of the form:

$$y = ax^2 + bx + c + \textsf{small noise}, \qquad (\textbf{true model}) \tag{31.143}$$

where x is given and y is the response. For each given x_n, we measure the corresponding noisy y_n according to this model. We could then use the N data points $\{x_n, y_n\}$ to fit a quadratic model to the data. This can be done by estimating the parameter vector $\theta = \{a, b, c\}$ of size $M = 3$ by solving a least-squares problem of the form:

$$\{\widehat{a}, \widehat{b}, \widehat{c}\} \;=\; \underset{\{a,b,c\}}{\text{argmin}} \left\{ \sum_{n=1}^{N} (y_n - ax_n^2 - bx_n - c)^2 \right\} \tag{31.144}$$

Each term in the above cost function penalizes the squared error between the noisy measurement y_n and its quadratic fit. It is straightforward to differentiate the above cost relative to $\{a, b, c\}$ and determine expressions for their estimates — see Prob. 31.23. The expressions are not relevant for the discussion here but once they are determined, they can be used, for example, to compute predictions for future values x_m by using the fitted model:

$$\widehat{y}_m = \widehat{a}\, x_m^2 + \widehat{b}\, x_m + \widehat{c}, \qquad (\textbf{prediction}) \tag{31.145}$$

If the model parameters have been learned well, one would expect $\widehat{y}_m$ to provide a good prediction for the noiseless value of y_m that would have been observed under the true model (a, b, c), namely,

$$y_m = ax_m^2 + bx_m + c \tag{31.146}$$

This situation is illustrated in the left plot of Fig. 31.5. The red curve shows $N = 21$ noisy measurements resulting from the parameters

$$\{a, b, c\} = \{-0.2883, 0.3501, -1.8359\}, \qquad \sigma_v^2 = 3 \tag{31.147}$$

The location of the measurements are indicated on the red line by the filled circles; the horizontal axis shows the values of x with the range $x \in [-5, 5]$ in increments of one. The blue curve with squares shows the same measurements without the noise component. The black line shows the fitted curve (31.145) resulting from the following estimated parameters for this particular simulation:

$$\{\widehat{a}, \widehat{b}, \widehat{c}\} \;=\; \{-0.2115, 0.0803, -2.3376\} \tag{31.148}$$

The quality of these estimated parameters would be better and their values would be closer to the true (a, b, c) if we use larger N and have less noise. We continue with the values (31.148) to illustrate the main idea and to facilitate the visualization of the resulting effects. Using the fitted curve (31.145) we can predict values for the non-noisy curve for any given x. For example, for $x = -1.3$, we get

$$x = -1.3 \quad \Longrightarrow \left\{ \begin{array}{rl} ax^2 + bx + c \approx -2.7781 & \text{(non-noisy measurement)} \\ \widehat{y} = \widehat{a}\,x^2 + \widehat{b}\,x + \widehat{c} \approx -2.7993 & \text{(prediction)} \end{array} \right. \tag{31.149}$$

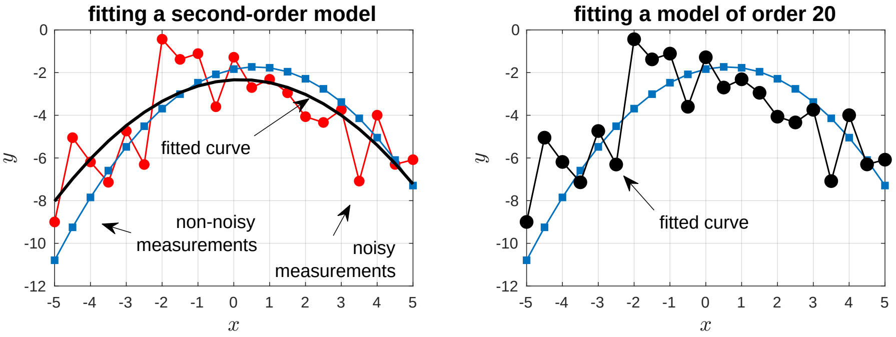

Figure 31.5 The plot on the left shows the result of fitting a second-order model onto the measurements, while the plot on the right shows the result of fitting a model of order 20.

Now, given the same N data points $\{x_n, y_n\}$, we could consider fitting a higher-order model; one that weaves more closely through the $\{x_n, y_n\}$ points in the $2D$ plane. For instance, we could consider fitting a fifth-order model with parameters $\theta = \{a, b, c, d, e, f\}$ instead of the second-order model, such as:

$$\widehat{y}_m = \widehat{a}\,x_m^5 + \widehat{b}\,x_m^4 + \widehat{c}\,x_m^3 + \widehat{d}\,x_m^2 + \widehat{e}\,x_m + \widehat{f} \tag{31.150}$$

Doing so would amount to overfitting (fitting a more complex model than necessary since the data originates from a second-order model to begin with). While the fifth-order model may fit the given data points $\{x_n, y_n\}$ better than the second-order model, the higher-order model will perform poorly on predicting future samples y_m. Poor performance means that if we were to substitute x_m into the higher-order model, the predicted value $\widehat{y}_m$ will generally be far from the value y_m that would result from the true model. This situation is illustrated in the right plot in Fig. 31.5. The black curve shows the same $N = 20$ noisy measurements from before, while the blue curve shows the same non-noisy measurements. We now fit a model of order 20 even though the data was generated from a second-order model. We observe in this case that the fitted curve lies on top of the black curve. In other words, the fitting now is so good that the fitted curve weaves through the measurement points and even accommodates the presence of noise in the measurements. This property is undesirable because it will generally lead to bad prediction performance. For instance, for the same point $x = -1.3$, the new fitted curve now predicts:

$$x = -1.3 \quad \Longrightarrow \widehat{y} \approx -0.4718 \tag{31.151}$$

which is further away from the true value at approximately -2.7781.

31.6.2 Akaike Information Criterion

The AIC formulation discourages overfitting (i.e., it discourages overly complex models) by penalizing the number of parameters in the model. From a collection of candidate models, it selects the optimal model as follows:

$$(k^\star, \theta^\star) \;=\; \underset{k,\theta_k}{\text{argmin}} \Big\{ 2M_k - 2\ell(y_1, y_2, \ldots, y_N; \theta_k) \Big\} \tag{31.152}$$

where $\ell(\cdot)$ is the log-likelihood function of the observations, assumed independent of each other:

$$\ell(y_1, y_2, \ldots, y_N; \theta_k) = \ln\left(\prod_{n-1}^{N} f_{\boldsymbol{y}}(y_n; \theta_k) \right) \tag{31.153}$$

In other words, the AIC formulation selects the model as follows:

$$\boxed{(k^\star, \theta^\star) \;=\; \underset{k,\theta_k}{\text{argmin}} \left\{ 2M_k \;-\; 2\sum_{n=1}^{N} \ln f_{\boldsymbol{y}}(y_n; \theta_k) \right\}} \quad (\textbf{AIC}) \tag{31.154}$$

The first factor $2M_k$ penalizes the complexity of the model, while the second term (also known as the "goodness-of-fit" measure) quantities how well the model θ_k fits the data by calculating its log-likelihood value. We can of course remove the factor 2 from both terms; it is kept for "historical" reasons to match the original formulation. We provide one motivation for the cost function (31.154) in Appendix 31.B. Since only the second term depends on θ_k, we find that the AIC solution can be determined as follows:

(a) For each model class θ_k, we determine its maximum-likelihood (ML) estimate by solving

$$\widehat{\theta}_k = \underset{\theta_k}{\text{argmax}} \left\{ \sum_{n=1}^{N} \ln f_{\boldsymbol{y}}(y_n; \theta_k) \right\}, \quad k = 1, 2, \ldots, K \tag{31.155}$$

(b) We assign an AIC score to each model k:

$$\text{AIC}(k) \;\stackrel{\Delta}{=}\; 2M_k - 2\sum_{n=1}^{N} \ln f_{\boldsymbol{y}}(y_n; \widehat{\theta}_k) \tag{31.156}$$

(c) We select the model class with the smallest AIC score:

$$k^\star = \underset{1 \le k \le K}{\text{argmin}} \; \text{AIC}(k) \;\Longrightarrow\; \theta^\star = \widehat{\theta}_{k^\star} \tag{31.157}$$

It is explained in Appendix 31.B that the AIC formulation seeks the model that minimizes the KL divergence between the true model and the ML models $\{\widehat{\theta}_1, \ldots, \widehat{\theta}_K\}$. Since the true model is unknown, AIC ignores the entropy of the true distribution in expression (31.238) in the appendix. For this reason, the AIC score is a relative measure of the "distance" from the true model. The lower the AIC score is, the closer the selected model will be to the true model. For this reason, in practice, these scores are handled as follows:

(a') We determine the model with the lowest AIC score and denote it by $\theta^\star$, as already explained above.

(b') We associate with each model k a (non-negative) delta score computed as follows:

$$\delta(k) \;\overset{\Delta}{=}\; \text{AIC}(k) - \text{AIC}(k^\star) \tag{31.158}$$

which measures how far model k is from the optimal model $k^\star$.

(c') We associate a probability distribution with the models conditioned on the observations (also known as a *softmax* mapping) as follows:

$$\pi(k|y_1,\ldots,y_N) \;\overset{\Delta}{=}\; \frac{e^{-\delta(k)/2}}{\sum_{k'=1}^{K} e^{-\delta(k')/2}}, \quad k=1,2,\ldots,K \tag{31.159}$$

where the division by 2 removes the "unnecessary" scaling that appears in the AIC score expression. The ratios $\pi(k|y_1,\ldots,y_N)$ lie in the interval $[0,1]$ and they add up to one. Therefore, they can be interpreted as probability values. In this way, for an arbitrary model k, the value $\pi(k|y_1,\ldots,y_N)$ indicates how likely it is for model k to be the best model based on the observed measurements.

31.6.3 Bayesian Information Criterion

The BIC formulation is closely related to AIC. The main difference between both criteria is the manner by which they penalize the complexity of the model; AIC penalizes the model selection less strongly than BIC, which selects the model as follows:

$$\boxed{(k^\star,\theta^\star) \;=\; \underset{k,\theta_k}{\text{argmin}}\left\{ M_k \ln N - 2\sum_{n=1}^{N} \ln f_{\boldsymbol{y}}(y_n;\theta_k)\right\}} \quad (\textbf{BIC}) \tag{31.160}$$

with an additional factor $\ln N$ multiplying M_k. We motivate this cost function in Appendix 31.C. Since only the second term depends on θ_k, we find that the BIC solution can be determined as follows:

(a) For each model class θ_k, we determine its maximum-likelihood estimate by solving:

$$\widehat{\theta}_k = \underset{\theta_k}{\text{argmax}}\left\{\sum_{n=1}^{N} \ln f_{\boldsymbol{y}}(y_n;\theta_k)\right\}, \quad k=1,2,\ldots,K \tag{31.161}$$

(b) We assign a BIC score to each model k:

$$\text{BIC}(k) \;\stackrel{\Delta}{=}\; M_k \ln N - 2\sum_{n=1}^{N} \ln f_{\boldsymbol{y}}(y_n; \widehat{\theta}_k) \tag{31.162}$$

(c) We select the model class with the smallest BIC score:

$$k^\star = \underset{1\le k\le K}{\text{argmin}}\ \text{BIC}(k) \implies \theta^\star = \widehat{\theta}_{k^\star} \tag{31.163}$$

It is explained in Appendix 31.C that the BIC formulation maximizes the *a posteriori* probability of the model selection given the observations. Thus, the lower the BIC score is, the more likely the selected model is a good approximation for the true model. Specifically, from expression (31.279) in the appendix we deduce that the likelihood of selecting model k given the observations satisfies:

$$\pi(k|y_1, y_2, \ldots, y_N) \approx \frac{e^{\text{BIC}(k)/2}}{\sum_{k'=1}^{K} e^{\text{BIC}(k')/2}} \tag{31.164}$$

Example 31.14 **(Illustrating BIC and AIC procedures)** We illustrate the BIC and AIC procedures by considering the problem of fitting a Gaussian distribution into a collection of N data points $\{y_n\}$. We wish to select the best fit among two models for the data. The first model is a Gaussian distribution with known variance σ^2 but unknown mean μ_1. That is, $\theta_1 = \{\mu_1\}$ and $M_1 = 1$:

<u>**model** θ_1</u> :

$$\theta_1 = \{\mu_1\} \tag{31.165a}$$

$$f_{\boldsymbol{y}}(y;\theta_1) \sim \mathcal{N}_{\boldsymbol{y}}(\theta_1, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y-\theta_1)^2} \tag{31.165b}$$

The log-likelihood function in this case is

$$\ell(\theta_1) \;\stackrel{\Delta}{=}\; \ln\left(\prod_{n=1}^{N} f_{\boldsymbol{y}}(y_n;\theta_1)\right) \;=\; -\frac{N}{2}\ln(2\pi\sigma^2) - \sum_{n=1}^{N}\frac{1}{2\sigma^2}(y_n-\theta_1)^2 \tag{31.166}$$

Differentiating relative to θ_1 and setting the derivative to zero leads to the ML estimate:

$$\widehat{\theta}_1 = \frac{1}{N}\sum_{n=1}^{N} y_n, \quad \text{with ``goodness-to-fit'' measure } \ell(\widehat{\theta}_1) \tag{31.167}$$

The second model is also a Gaussian distribution albeit with unknown variance and mean. That is, $\theta_2 = \{\mu_2, \sigma_2^2\}$ and $M_2 = 2$:

<u>**model** θ_2</u> :

$$\theta_2 = \{\mu_2, \sigma_2^2\} \tag{31.168a}$$

$$f_{\boldsymbol{y}}(y;\theta_1) \sim \mathcal{N}_{\boldsymbol{y}}(\theta_2, \sigma_2^2) = \frac{1}{\sqrt{2\pi\sigma_2^2}} e^{-\frac{1}{2}(y-\mu_2)^2} \tag{31.168b}$$

The log-likelihood function in this case is

$$\ell(\theta_2) = -\frac{N}{2}\ln(2\pi\sigma_2^2) - \sum_{n=1}^{N}\frac{1}{2\sigma_2^2}(y_n-\mu_2)^2 \tag{31.169}$$

Differentiating relative to (μ_2, σ_2^2) and setting the derivative to zero leads to the same ML estimates we encountered before in (31.23a)–(31.23b):

$$\widehat{\mu}_2 = \frac{1}{N}\sum_{n=1}^{N} y_n, \qquad \widehat{\sigma}_2^2 = \frac{1}{N}\sum_{n=1}^{N}(y_n - \widehat{\mu}_2)^2, \qquad \widehat{\theta}_2 = \{\widehat{\mu}_2, \widehat{\sigma}_2^2\} \tag{31.170}$$

with “goodness-to-fit” measure $\ell(\widehat{\theta}_2)$. The BIC and AIC scores for the two models are given by

$$\text{BIC}(1) = \ln N - 2\ell(\widehat{\theta}_1) \tag{31.171a}$$

$$\text{BIC}(2) = 2\ln N - 2\ell(\widehat{\theta}_2) \tag{31.171b}$$

$$\text{AIC}(1) = 2 - 2\ell(\widehat{\theta}_1) \tag{31.171c}$$

$$\text{AIC}(2) = 4 - 2\ell(\widehat{\theta}_2) \tag{31.171d}$$

Example 31.15 **(Moving average model)** We studied regression problems in Chapter 29. Here we consider a motivating example. Assume we collect N independent and identically-distributed scalar observations $\{\boldsymbol{\gamma}(n), \boldsymbol{h}(n)\}$. We wish to model $\boldsymbol{\gamma}(n)$ by a linear model of the form:

$$\boldsymbol{\gamma}(n) = \boldsymbol{h}_n^{\mathsf{T}} w + \boldsymbol{v}(n) \tag{31.172}$$

where $w \in \mathbb{R}^M$ is a parameter vector to be determined, and $\boldsymbol{h}_n$ is an observation vector consisting of M delayed samples $\boldsymbol{h}(n)$, namely,

$$\boldsymbol{h}_n \;\overset{\Delta}{=}\; \text{col}\Big\{\boldsymbol{h}(n),\, \boldsymbol{h}(n-1),\, \ldots,\, \boldsymbol{h}(n-M+1)\Big\} \tag{31.173}$$

Moreover, the term $\boldsymbol{v}(n)$ represents some small zero-mean discrepancy assumed to be Gaussian-distributed:

$$f_{\boldsymbol{v}}(v) \;=\; \frac{1}{\sqrt{2\pi\sigma_v^2}} e^{-\frac{v^2}{2\sigma_v^2}} \tag{31.174}$$

In this way, expression (31.172) is attempting to fit a linear regression model into the data by representing $\boldsymbol{\gamma}(n)$ as a combination of current and delayed samples $\{\boldsymbol{h}(m)\}$. The order of the model is M because it uses M samples $\{\boldsymbol{h}(n), \ldots, \boldsymbol{h}(n-M+1)\}$. The parameter w plays the role of the model θ, and M is its size. We wish to determine the optimal size M.

It is straightforward to see that the log-likelihood function of the observations $\{\gamma(n), h_n\}$ for a particular model w is

$$\begin{aligned}
\ell(w) &\overset{\Delta}{=} \ln f_{\boldsymbol{\gamma}_1,\ldots,\boldsymbol{\gamma}_N,\boldsymbol{h}_1,\ldots,\boldsymbol{h}_N}\Big(h_1, \ldots, h_N, \gamma(1), \ldots, \gamma(N); w\Big) \\
&= \ln f_{\boldsymbol{v}_1,\ldots,\boldsymbol{v}_N}(v(1), \ldots, v(N); w) \\
&= \ln \left\{ \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi\sigma_v^2}} e^{-\frac{\left(\gamma(n) - h_n^{\mathsf{T}} w\right)^2}{2\sigma_v^2}} \right\} \\
&= -\frac{N}{2}\ln(2\pi\sigma_v^2) \;-\; \frac{1}{2\sigma_v^2}\sum_{n=1}^{N}(\gamma(n) - h_n^{\mathsf{T}} w)^2
\end{aligned} \tag{31.175}$$

It follows that the maximum of the log-likelihood function over w is obtained by solving the least-squares problem:

$$w_M^\star = \underset{w\in\mathbb{R}^M}{\text{argmin}} \left\{ \sum_{n=1}^{N}(\gamma(n) - h_n^\mathsf{T} w)^2 \right\}, \text{ with “goodness-of-fit” measure } \ell(w_M^\star) \tag{31.176}$$

Differentiating relative to w and setting the gradient to zero we find the following expressions for the minimizer and the corresponding minimum cost — see Prob. 31.26:

$$w_N^\star = \left(\sum_{n=1}^{N} h_n h_n^\mathsf{T}\right)^{-1} \sum_{n=1}^{N} \gamma(n) h_n \tag{31.177a}$$

$$\mathcal{E}_M \triangleq \sum_{n=1}^{N}(\gamma(n) - h_n^\mathsf{T} w)^2 \Big|_{w=w_M^\star} = \sum_{n=1}^{N} \gamma(n)(\gamma(n) - h_n^\mathsf{T} w_M^\star) \tag{31.177b}$$

so that

$$\ell(w_M^\star) = -\frac{N}{2}\ln(2\pi\sigma_v^2) - \frac{1}{2\sigma_v^2}\mathcal{E}_M \tag{31.178}$$

The BIC and AIC scores for models of order M are then given by

$$\text{BIC}(M) = M \ln N - 2\ell(w_M^\star) \tag{31.179a}$$

$$\text{AIC}(M) = 2M - 2\ell(w_M^\star) \tag{31.179b}$$

Either of these scores can now be minimized over M to select the best model order.

31.6.4 Minimum Description Length

The minimum description length (MDL) criterion is based on the principle that the best model fit is one that compresses the data the most, i.e., a solution where the representation of both the model and the data requires the smallest number of bits. This is based on the intuition that the more we are able to compress the data (i.e., the easier it is for us to describe it), the more we would have learned about its inherent structure.

Let $\mathcal{B}(\theta)$ represent the number of bits that are needed to represent a generic model θ. For example, if the model θ corresponds to choosing the means of two Gaussian distributions in a mixture model of the form $\frac{1}{2}\mathcal{N}_{\boldsymbol{y}}(\mu_a, 1) + \frac{1}{2}\mathcal{N}_{\boldsymbol{y}}(\mu_b, 1)$, and if the means are known to be one of only four possibilities:

$$(\mu_a, \mu_b) \in \Big\{(\mu_{a1}, \mu_{b1}),\ (\mu_{a2}, \mu_{b2}),\ (\mu_{a3}, \mu_{b3}),\ (\mu_{a4}, \mu_{b4})\Big\} \tag{31.180}$$

then, for this example, $\mathcal{B}(\theta) = 2$ bits. In a second example, assume θ has M dimensions and lies within a bounded region. We can discretize every dimension into $\sqrt{N}$ small segments (where N is the number of data points); this choice is motivated by the fact that the size of the error in estimating each entry of θ is on the order of $1/\sqrt{N}$ as suggested by (31.129). Then, we would need roughly

$\log_2 \sqrt{N}$ bits to represent the value of θ along each of its dimensions so that $\mathcal{B}(\theta)$ can be approximated by

$$\mathcal{B}(\theta) \propto \frac{1}{2} M \ln N \tag{31.181}$$

which is similar to the term that appears in the BIC objective function in (31.160).

In other situations, it is justified to treat θ as a realization for some random variable $\boldsymbol{\theta}$ and to assign a distribution for $\boldsymbol{\theta}$ (also called its *prior*), say, $f_{\boldsymbol{\theta}}(\theta)$. For instance, if θ corresponds to the mean of a Gaussian distribution $f_{\boldsymbol{y}}(y;\theta) \sim \mathcal{N}_{\boldsymbol{y}}(\mu, \sigma^2)$, then one could assume that the unknown mean μ is a realization that arises from some exponential distribution $\boldsymbol{\theta} \sim \lambda e^{-\lambda\theta}$. When the model parameters are treated in this manner as random variables with priors assigned to them, we can appeal to our earlier discussion on information and entropy from Chapter 6 to deduce that the number of bits that are needed to code θ is on the order of (recall expression (6.1)):

$$\mathcal{B}(\theta) = -\ln f_{\boldsymbol{\theta}}(\theta) \tag{31.182}$$

The MDL approach exploits this connection between code lengths and pdfs to great effect. Since we are using the natural logarithm in (31.182), the units should be listed as *nats* instead of *bits*.

Let $\mathcal{B}(y_1, y_2, \ldots, y_n; \theta)$ represent the number of bits that are needed to represent the N data points under model θ. The MDL criterion then selects the model as follows:

$$(k^\star, \theta^\star) \;=\; \underset{k,\theta_k}{\text{argmin}} \left\{ \mathcal{B}(\theta_k) + \mathcal{B}(y_1, y_2, \ldots, y_n; \theta_k) \right\} \tag{31.183}$$

Again, motivated by (31.182) we can select

$$\mathcal{B}(y_1, y_2, \ldots, y_n; \theta) = -\ln \left\{ \prod_{n=1}^{N} f_{\boldsymbol{y}}(y_n; \theta) \right\} \tag{31.184}$$

in terms of the natural logarithm of the likelihood function and rewrite the MDL formulation as

$$\boxed{(k^\star, \theta^\star) \;=\; \underset{k,\theta_k}{\text{argmin}} \left\{ \mathcal{B}(\theta_k) - \sum_{n=1}^{N} \ln f_{\boldsymbol{y}}(y_n; \theta_k) \right\}} \quad (\textbf{MDL}) \tag{31.185}$$

One useful interpretation for the MDL objective arises when the choice (31.182) is used, i.e., when

$$(k^\star, \theta^\star) \;=\; \underset{k,\theta_k}{\text{argmax}} \left\{ \ln f_{\boldsymbol{\theta}}(\theta_k) + \sum_{n=1}^{N} \ln f_{\boldsymbol{y}}(y_n; \theta_k) \right\} \tag{31.186}$$

where we replaced argmin by argmax and removed the negative signs. In this case, MDL can be shown to correspond to choosing among a collection of *maximum a-posteriori* (MAP) estimators. This can be seen as follows. When $\boldsymbol{\theta}$ is treated as random, each term $f_{\boldsymbol{y}}(y_n;\theta_k)$ in (31.185) has the interpretation of the conditional distribution of $\boldsymbol{y}$ given the realization for $\boldsymbol{\theta}_k$:

$$f_{\boldsymbol{y}}(y_n;\theta_k) \;=\; f_{\boldsymbol{y}|\boldsymbol{\theta}_k}(y_n|\theta_k) \tag{31.187}$$

Then, using (31.182), we can rewrite the cost that is being optimized in (31.186) in the form:

$$\begin{aligned}
&\ln f_{\boldsymbol{\theta}}(\theta_k) + \sum_{n=1}^{N} \ln f_{\boldsymbol{y}}(y_n;\theta_k) \qquad (31.188)\\
&= \ln\left\{ f_{\boldsymbol{\theta}}(\theta_k) \prod_{n=1}^{N} f_{\boldsymbol{y}|\boldsymbol{\theta}_k}(y_n|\theta_k)\right\}\\
&= \ln\left\{ f_{\boldsymbol{\theta}}(\theta_k)\, f_{\boldsymbol{y}_1,\ldots,\boldsymbol{y}_N|\boldsymbol{\theta}_k}(y_1,\ldots,y_N|\theta_k)\right\}\\
&= \ln\, f_{\boldsymbol{y}_1,\ldots,\boldsymbol{y}_N,\boldsymbol{\theta}_k}(y_1,y_2,\ldots,y_N,\theta_k)\\
&= \ln\Big\{ f_{\boldsymbol{y}_1,\ldots,\boldsymbol{y}_N}(y_1,\ldots,y_n) \times f_{\boldsymbol{\theta}_k|\boldsymbol{y}_1,\ldots,\boldsymbol{y}_N}(\theta_k|y_1,\ldots,y_N)\Big\}\\
&= \ln\Big\{ f_{\boldsymbol{\theta}_k|\boldsymbol{y}_1,\ldots,\boldsymbol{y}_N}(\theta_k|y_1,\ldots,y_N)\Big\} \;+\; \text{term independent of } \theta_k
\end{aligned}$$

It follows that the MDL criterion (31.186) is the best fit from among a collection of MAP estimators $\{\widehat{\boldsymbol{\theta}}_k\}$:

$$(k^\star,\theta^\star) \;=\; \underset{k,\theta_k}{\mathrm{argmax}} Bigl\{, f_{\boldsymbol{\theta}_k|\boldsymbol{y}_1,\ldots,\boldsymbol{y}_N}(\theta_k|y_1,\ldots,y_N)\Big\} \tag{31.189}$$

This construction involves the following steps:

(a) For each model class θ_k, we determine its maximum *a-posteriori* (MAP) estimate by solving:

$$\widehat{\theta}_k = \underset{\theta_k}{\mathrm{argmax}}\;\Big\{ f_{\boldsymbol{\theta}_k|\boldsymbol{y}_1,\ldots,\boldsymbol{y}_N}(\theta_k|y_1,\ldots,y_N)\Big\},\quad k=1,2,\ldots,K \tag{31.190}$$

(b) We assign an MDL score to each class k:

$$\mathrm{MDL}(k) \;\stackrel{\Delta}{=}\; f_{\boldsymbol{\theta}_k|\boldsymbol{y}_1,\ldots,\boldsymbol{y}_N}(\widehat{\theta}_k|y_1,\ldots,y_N) \tag{31.191}$$

(c) We select the model class with the largest MDL score:

$$k^\star = \underset{1\le k\le K}{\mathrm{argmax}}\;\mathrm{MDL}(k) \;\Longrightarrow\; \theta^\star = \widehat{\theta}_{k^\star} \tag{31.192}$$

Example 31.16 **(MDL with model prior)** Consider a collection of N independent and identically distributed realizations $\{y_n\}$. We wish to select the best fit among two models for the data. The first model is a Gaussian distribution with known variance

σ_1^2 and unknown mean μ_1. That is, $\theta_1 = \{\mu_1\}$. We model $\boldsymbol{\theta}_1$ as a random variable and assume it is Gaussian distributed with zero mean and unit variance, i.e.,

model θ_1 :

$$f_{\boldsymbol{\mu}_1}(\boldsymbol{\mu}_1) \sim \mathcal{N}_{\boldsymbol{\mu}_1}(0,1) \;=\; \frac{1}{\sqrt{2\pi}} e^{-\mu_1^2/2} \tag{31.193a}$$

$$\boldsymbol{\theta}_1 = \boldsymbol{\mu}_1 \tag{31.193b}$$

$$f_{\boldsymbol{y}}(y;\theta_1) \sim \mathcal{N}_{\boldsymbol{y}}(\theta_1, \sigma^2) \;=\; \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma_y^2}(y-\theta_1)^2} \tag{31.193c}$$

If we denote the cost that is maximized in the MDL formulation (31.186) by $J(\theta_k)$, then it is given by the following expression for model $\theta_1 = \mu_1$:

$$J(\mu_1) = -\frac{1}{2}\mu_1^2 - \frac{1}{2\sigma_1^2}\sum_{n=1}^{N}(y_n - \mu_1)^2 \;+\; \text{cte} \tag{31.194}$$

where constant terms that are independent of $\theta_1 = \mu_1$ are separated.

The second model is also Gaussian albeit with unknown variance and mean. That is, $\theta_2 = \{\mu_2, \sigma_2^2\}$. We assume the two components of $\boldsymbol{\theta}_2$ are independent of each other, with $\boldsymbol{\mu}$ being Gaussian-distributed with zero mean and unit variance and $\boldsymbol{\sigma}^2$ being exponentially-distributed with parameter $\lambda = 1$, i.e.,

model θ_2 :

$$f_{\boldsymbol{\mu}_2}(\mu_2) \sim \mathcal{N}_{\boldsymbol{\mu}_2}(0,1) \;=\; \frac{1}{\sqrt{2\pi}} e^{-\mu_2^2/2} \tag{31.195a}$$

$$f_{\boldsymbol{\sigma}_2^2}(\sigma_2^2) \;=\; e^{-\sigma_2^2} \tag{31.195b}$$

$$f_{\boldsymbol{\theta}_2}(\theta_2) \;=\; \frac{1}{\sqrt{2\pi}} e^{-\mu^2/2} \times e^{-\sigma_2^2} \tag{31.195c}$$

$$f_{\boldsymbol{y}}(y;\theta_2) \sim \mathcal{N}_{\boldsymbol{y}}(\mu_2, \sigma_2^2) \;=\; \frac{1}{\sqrt{2\pi\sigma_2^2}} e^{-\frac{1}{2\sigma_2^2}(y-\mu_2)^2} \tag{31.195d}$$

The cost in (31.186) that corresponds to this model is given by

$$J(\mu_2, \sigma_2^2) = -\frac{1}{2}\mu_2^2 - \sigma_2^2 - \frac{1}{2\sigma_2^2}\sum_{n=1}^{N}(y_n - \mu_2)^2 \;+\; \text{cte} \tag{31.196}$$

where again constant terms that are independent of $\theta_2 = \{\mu_2, \sigma_2^2\}$ are separated. This example is pursued further in Probs. 31.24 and 31.25.

31.6.5 Cross Validation Method

We describe next an alternative method for choosing among models that is based on a popular technique known as *cross validation*; the method leads to good performance often under weaker conditions than needed for the other (AIC, BIC, MDL) methods described before.

Consider again a collection of N independent and identically-distributed data measurements $\{y_1, y_2, \ldots, y_N\}$ arising from an underlying unknown pdf, $f_{\boldsymbol{y}}(y)$. Introduce K models with parameters $\{\theta_1, \theta_2, \ldots, \theta_K\}$, where each θ_k defines a probability density function for the observations. Cross validation splits the N

data points $\{y_n\}$ into L segments of size N/L each. At each iteration of the construction described below, we use $L-1$ segments for estimation purposes and the last segment for a supporting role — see Fig. 31.6. To simplify the notation, we let $E = (L-1)N/L$ denote the total number of samples used for estimation from the $L-1$ segments and $T = N/L$ denote the number of samples used for the support role from the remaining segment, so that $N = E + T$. The objective is to select the "best fit" model from among the $\{\theta_1, \ldots, \theta_K\}$.

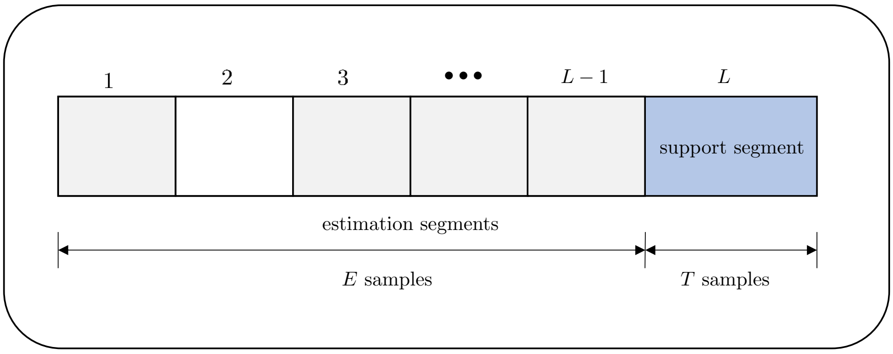

Figure 31.6 The data is divided into L segments, with $L-1$ of them used for estimation purposes for a total of E samples and the remaining segment with T samples used for a supporting role.

The first step is to use the E data points to estimate θ_k by maximizing the corresponding log-likelihood function:

$$\widehat{\theta}_k \;=\; \underset{\theta_k}{\operatorname{argmax}} \left\{ \sum_{e=1}^{E} \ln f_{\boldsymbol{y}}(y_e|\theta_k) \right\}, \quad k = 1, 2, \ldots, K \tag{31.197}$$

where the subscript e is used to index the samples from the E collection. This step generates K estimated models $\{\widehat{\theta}_k\}$. Next, we need to select a "best fit" model from among these models. One way to achieve this task is to minimize the KL divergence between the true (unknown) pdf and its approximations, namely,

$$k^\star \;=\; \underset{1\le k\le K}{\operatorname{argmin}} \left\{ D_{\rm KL}\Big(f_{\boldsymbol{y}}(y) \,\|\, f_{\boldsymbol{y}}(y;\widehat{\theta}_k) \Big) \right\} \;\Longrightarrow\; \theta^\star = \widehat{\theta}_{k^\star} \tag{31.198}$$

where

$$\begin{aligned} & D_{\rm KL}\Big(f_{\boldsymbol{y}}(y) \,\|\, f_{\boldsymbol{y}}(y;\widehat{\theta}_k) \Big) \\ & = \int_{y\in\mathcal{Y}} f_{\boldsymbol{y}}(y) \ln f_{\boldsymbol{y}}(y) dy \;-\; \int_{y\in\mathcal{Y}} f_{\boldsymbol{y}}(y) \ln f_{\boldsymbol{y}}(y;\widehat{\theta}_k) dy \end{aligned} \tag{31.199}$$

The first term on the right-hand side is independent of k. Therefore, problem

(31.198) for selecting the optimal model k reduces to

$$k^\star = \operatorname*{argmax}_{1\leq k\leq K}\left\{\int_{y\in\mathcal{Y}} f_{\boldsymbol{y}}(y)\ln f_{\boldsymbol{y}}(y;\widehat{\theta}_k)dy\right\} = \operatorname*{argmax}_{1\leq k\leq K}\left\{\mathbb{E}_{\boldsymbol{y}}\ \ln f_{\boldsymbol{y}}(y;\widehat{\theta}_k)\right\} \tag{31.200}$$

The quantity that is being maximized is the mean of $\ln f_{\boldsymbol{y}}(y;\widehat{\theta}_k)$, where the expectation is computed relative to the true distribution, $f_{\boldsymbol{y}}(y)$. The expectation cannot be computed since $f_{\boldsymbol{y}}(y)$ is unknown. One approximation is derived in Appendix 31.B in the form of expression (31.240) and used to motivate the AIC method. Here, we pursue a different approach based on cross validation.

In the cross validation approach, the quantity $\mathbb{E}_{\boldsymbol{y}}\, f(y;\widehat{\theta}_\ell)$ is estimated by using the T samples from the support segment, which were not involved in the estimation of the $\{\widehat{\theta}_k\}$. That is, we use

$$\mathbb{E}_{\boldsymbol{y}}\ \ln f_{\boldsymbol{y}}(y;\widehat{\theta}_k) \approx \frac{1}{T}\sum_{t=1}^{T}\ln f_{\boldsymbol{y}}(y_t;\widehat{\theta}_k) \tag{31.201}$$

where the samples $\{y_t\}$ in this expression arise from the T collection and are independent from the samples $\{y_e\}$ used to estimate $\widehat{\theta}_k$. Therefore, the sample average estimator on the right-hand side of (31.201) is an *unbiased* estimator for the quantity of interest, namely,

$$\mathbb{E}\left\{\underbrace{\frac{1}{T}\sum_{t=1}^{T}\ln f_{\boldsymbol{y}}(y_t;\widehat{\theta}_k)}_{\triangleq\,\widehat{X}_k}\right\} = \underbrace{\mathbb{E}_{\boldsymbol{y}}\ \ln f_{\boldsymbol{y}}(y;\widehat{\theta}_k)}_{\triangleq\,X_k} \tag{31.202}$$

To simplify the notation, we denote the variable that we wish to approximate by X_k and its sample approximation by $\widehat{X}_k$; we use the subscript k to indicate that these values relate to model k.

So far we have computed the estimate $\widehat{X}_k$ by considering a single pass over the L data segments. More generally, cross validation performs L passes over these segments. During each pass, one segment is chosen for support and the remaining $L-1$ segments for estimation. Each pass ℓ generates an estimate $\widehat{X}_{k,\ell}$ as above by computing the ensemble average over the samples of the support segment for that pass. Subsequently, the final estimate for X_k is obtained by averaging these multiple pass estimates as follows:

$$\widehat{X}_k \triangleq \frac{1}{L}\sum_{\ell=1}^{L}\widehat{X}_{k,\ell}, \qquad (\textbf{cross validation score}) \tag{31.203}$$

Cross validation generates a score $\widehat{X}_k$ in this manner for each model θ_k and then selects the model with the highest score in view of (31.200):

$$k^\star = \operatorname*{argmax}_{1\leq k\leq K}\left\{\widehat{X}_k\right\} \implies \theta^\star = \theta_{k^\star} \tag{31.204}$$

We explain in Prob. 31.29 that, under certain conditions, the cross validation construction is able to discover the best model with high likelihood. We will discuss cross validation further in Chapter 59 and provide comments on its history and application in the context of inference and learning methods.

31.7 COMMENTARIES AND DISCUSSION

Maximum-likelihood. The maximum-likelihood approach was developed by the English statistician **Ronald Fisher (1890–1962)** in the works by Fisher (1912,1922,1925) — see the presentations by Pratt (1976), Savage (1976), and Aldrich (1997). The approach does not assume any prior distribution for the parameter θ and estimates it from observations of the random variable $\boldsymbol{y}$ by maximizing the likelihood function defined by (31.1). Since its inception, the maximum likelihood technique has grown to become one of the most formidable tools in modern statistical analysis, motivated largely by the foundational works of Fisher (1922,1956) and also by the efficiency of this class of estimators. As already noted by (31.129), maximum-likelihood estimators are asymptotically efficient in that their mean-square errors approach the Cramer-Rao bound as the number of observations grows. For additional information on ML estimators, and for more details on the Carmer-Rao bound, its ramifications, and the asymptotic efficiency and normality of ML solutions, readers may refer to the texts by Zacks (1971), Box and Tiao (1973), Scharf (1991), Kay (1993,1998), Lehmann (1998), Cassella and Berger (2002), Cox (2006), Hogg and McKean (2012), and Van Trees (2013).

Cramer-Rao bound. This important bound, which is due to Rao (1945) and Cramer (1946), provides a lower limit on the achievable mean-square error for any unbiased (and also biased) estimator of unknown constant parameters. The lower bound is determined by the inverse of the Fisher information matrix which, although named after Fisher (1922,1956), was actually advanced in the works by Edgeworth (1908a,b,c) — see the expositions by Savage (1976) and Pratt (1976). The entries of the Fisher matrix reflect the amount of information that the observations convey about the unknown parameter — see Frieden (2004). In Appendix 31.A we provide one derivation for the Cramer-Rao bound for the case of scalar parameters by following an argument similar to Cassella and Berger (2002), Frieden (2004), and Van Trees (1968,2013).

Expression (31.129), and the result in Example 31.12, suggest that the mean-square error decays at the rate of $1/N$, in inverse proportion to the sample size. There are situations, however, where the rate of decay of the mean-square error can be exponentially fast in N. These situations arise, for example, when estimating unknown parameters θ that are restricted in certain ways as discussed by Hammersley (1950). To illustrate this possibility, let us reconsider Example 31.12 where we are still interested in determining the maximum likelihood estimate for θ, except that θ is now constrained to being an *integer*. It is shown in Hammersley (1950) that, in this case — see Prob. 31.20:

$$\widehat{\theta}_{\rm ML} = \text{round}\left(\frac{1}{N}\sum_{n=1}^{N} y_n\right) \tag{31.205}$$

where the function round(x) denotes the integer value that is closest to x. The corresponding mean-square error behaves asymptotically as

$$\mathbb{E}\,\widetilde{\boldsymbol{\theta}}^2_{\rm ML} \;=\; \left(\frac{8\sigma_v^2}{\pi N}\right)^{1/2} \times e^{-N/8\sigma_v^2}, \quad \text{as } N \to \infty \tag{31.206}$$

which decays exponentially with N.

Sufficient statistics. We commented on sufficient statistics at the end of Chapter 5. Let $\{y_1, y_2, \ldots, y_N\}$ denote realizations for a random variable $\boldsymbol{y}$ whose pdf is parameterized by some θ (scalar or vector), written as $f_{\boldsymbol{y}}(y;\theta)$. Let the short-hand notation $S(y)$ denote any function of these realizations i.e.,

$$S(y) \;\stackrel{\Delta}{=}\; S(y_1, y_2, \ldots, y_N) \tag{31.207}$$

We refer to $S(y)$ as a *statistic*. The statistic is said to be *sufficient* for θ if the conditional distribution of $\{\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_N\}$ given $S(y)$ does not depend on θ. This concept plays a key role in maximum-likelihood estimation theory and was introduced by Fisher (1922) in his development of theoretical statistics. In a way, the sufficient statistic, when it exists, contains all the information that is embedded in the observations about θ so that the observations can be discarded and replaced by $S(y)$ for estimation purposes. This step amounts to compressing the observations down to $S(y)$. Let us consider the following classical example.

Let $\boldsymbol{y}$ denote the outcome of a Bernoulli experiment with success rate p, i.e., $y = 1$ with probability p and $y = 0$ with probability $1 - p$. The variable p plays the role of the parameter θ. Assume we perform N experiments and observe the outcomes $\{y_1, \ldots, y_N\}$. Define the function

$$S(y) \;\stackrel{\Delta}{=}\; \sum_{n=1}^{N} y_n \tag{31.208}$$

which counts the number of successes. The conditional pdf of $\{\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_N\}$ given $S(y)$ can be shown to be independent of p. Indeed, it is left as an exercise for Prob. 31.16 to show that:

$$\mathbb{P}\Big(\boldsymbol{y}_1 = y_1, \ldots, \boldsymbol{y}_N = y_N \,|\, S(y) = s\Big) \;=\; \begin{cases} \dfrac{s!(N-s)!}{N!}, & \text{if } \displaystyle\sum_{n=1}^{N} y_n = s \\ 0, & \text{if } \displaystyle\sum_{n=1}^{N} y_n \neq s \end{cases} \tag{31.209}$$

which shows that the conditional distribution is independent of the parameter p. Therefore, the statistic $S(y)$ defined by (31.208) is sufficient for p.

The following result explains how, starting from some initial crude estimator for a parameter θ that is not necessarily optimal, we can construct better estimators for it by conditioning on a sufficient statistic for θ — see Prob. 31.17 and also Caines (1988), Scharf (1991), and Kay (1993,1998). Observe how the conditional mean plays a useful role in constructing estimators.

(Rao-Blackwell theorem) (Rao (1945) and Blackwell (1947)): *Let $\widehat{\boldsymbol{\theta}}_1$ denote an unbiased estimator for a parameter θ given observations of a variable $\boldsymbol{y}$ and assume $\mathbb{E}\,\widehat{\boldsymbol{\theta}}_1^2 < \infty$. Let $S(\boldsymbol{y})$ denote a sufficient statistic for θ and construct the estimator:*

$$\widehat{\boldsymbol{\theta}}_2 = \mathbb{E}\left(\widehat{\boldsymbol{\theta}}_1 \,|\, S(\boldsymbol{y})\right) \tag{31.210}$$

Then, $\widehat{\boldsymbol{\theta}}_2$ is also an unbiased estimator for θ with at most the same mean-square error (or variance), namely,

$$\mathbb{E}\,(\theta - \widehat{\boldsymbol{\theta}}_2)^2 \;\leq \mathbb{E}\,(\theta - \widehat{\boldsymbol{\theta}}_1)^2 \tag{31.211}$$

The inequality holds strictly, except when the estimator $\widehat{\boldsymbol{\theta}}_1$ is a function of $S(\boldsymbol{y})$.

Gamma function. We encountered the Gamma function in Example 31.4 while fitting a Beta distribution onto measured data. The function is defined by the integral expression

$$\Gamma(z) = \int_0^\infty s^{z-1} e^{-s} ds, \quad z > 0 \tag{31.212}$$

and has several useful properties such as $\Gamma(1/2) = \sqrt{\pi}$, $\Gamma(z+1) = z\Gamma(z)$ for any $z > 0$, and $\Gamma(n+1) = n!$ for any integer $n \geq 0$. This last property shows that the Gamma function can be viewed as an extension of the factorial operation to real (and even complex) numbers. In the example, we needed to evaluate the digamma function $\psi(z) = \Gamma'(z)/\Gamma(z)$, which arises often in applications, involving the derivative of the Gamma function. This ratio is sometimes referred to as the *polygamma function* of order zero and is known to satisfy the relation:

$$\frac{\Gamma'(z)}{\Gamma(z)} = -c + \sum_{m=0}^{\infty} \left(\frac{1}{1+m} - \frac{1}{z+m} \right) \tag{31.213}$$

where

$$c \stackrel{\Delta}{=} -\lim_{z \to 1} \Gamma'(z) \approx 0.577215665 \tag{31.214}$$

is Euler's constant (often denoted instead by the letter γ); it appears in many other problems in mathematical analysis. For more information on the Gamma function and its properties, the reader may consult the works by Davis (1959), Abramowitz and Stegun (1965), Lebedev (1972), Temme (1996), and Artin (2015).

Method of moments. We discussed in Example 31.4 two methods to fit a Beta distribution onto data measurements. One method was based on the maximum-likelihood formulation and required an iterative procedure to learn the shape parameters (a, b), while the second method estimated these parameters by matching the first and second-order moments (mean and variance) of the resulting Beta distribution to the sample mean and sample variance computed from the data. It is a historical curiosity that the maximum-likelihood approach to fitting a Beta distribution was favored by the English statistician **Ronald Fisher (1890–1962)**, while the moment matching approach was favored by the English statistician **Karl Pearson (1857–1936)** — see Pearson (1936) and the account by Bowman and Shenton (2007). Both Fisher and Pearson were giants in their field and are credited with establishing the modern field of mathematical statistics.

Akaike and Bayesian information criteria. The Akaike information criterion (AIC) is due to the Japanese statistician **Hirotugu Akaike (1927–2009)** and appeared in the work by Akaike (1974). Since its inception, it has flourished to become one of the main tools in statistical analysis. The criterion is based on information-theoretic concepts and seeks the "best fit" that minimizes the KL divergence relative to the true (unknown) distribution of the data. We explain in Appendix 31.B that AIC achieves this goal by constructing an "*unbiased*" estimate for the mean log-likelihood function defined in (31.243) as follows:

$$L(\theta) \stackrel{\Delta}{=} \int_{y \in \mathcal{Y}} f(y) \ln f(y;\theta) dy \;=\; \mathbb{E}_{\boldsymbol{y}} \ln f(\boldsymbol{y};\theta) \tag{31.215}$$

The expectation is relative to the true distribution of $\boldsymbol{y}$. Since $L(\theta)$ is unavailable, AIC approximates it from data measurements by using (31.240), namely,

$$\mathbb{E}_{\boldsymbol{y}} \ln f(y;\widehat{\theta}_k) \approx \frac{1}{N} \sum_{n=1}^{N} \ln f(y_n;\widehat{\theta}_k) \;-\; \frac{M_k}{N} \tag{31.216}$$

where the correction by M_k/N is necessary to remove the bias from the first term.

It is explained in the survey article by Cavanaugh and Neath (2019) that the AIC formulation is *efficient*. This means that the selected model $\theta^\star$ will generate predictions $\widehat{\boldsymbol{y}}$ for $\boldsymbol{y}$ that have the lowest mean-square-error, $\mathbb{E}\,(\boldsymbol{y}-\widehat{\boldsymbol{y}})^2$. In other words, AIC favors model selections that are good predictors. The approximation (31.216) is known to perform well for large N but its performance degrades for small or moderate-size data sets as explained in Linhart and Zucchini (1986) and McQuarrie and Tsai (1998). To address this difficulty, Hurvich and Tsai (1989) suggested one correction to the AIC score based on replacing the original expression (31.156) by the following corrected score for sample sizes satisfying $N < 40M_k$:

$$\text{AIC}(k) \;\stackrel{\Delta}{=}\; 2M_k + \frac{2M_k(M_k+1)}{N-k-1} - 2\sum_{n=1}^{N} \ln f_{\boldsymbol{y}}(y_n;\widehat{\theta}_k), \quad (\textbf{corrected AIC}) \qquad (31.217)$$

The text by Burnham and Anderson (2002) provides a theoretical justification for the superior performance of corrected AIC in relation to BIC.

The Bayesian information criterion (BIC) was proposed by Schwarz (1978); it bears a strong resemblance to AIC except that it penalizes the model selection more strongly by scaling the model order M_k by $\ln N$ instead of 1 as happens with AIC. BIC adopts a Bayesian approach and assigns a prior $\pi(k)$ to the model variable $\boldsymbol{k} \in \{1, 2, \ldots, K\}$, as well as a prior to the parameter $\boldsymbol{\theta}_k$ under each class k. It then maximizes the posterior $\pi(k|y_1, y_2, \ldots, y_N)$ given the observations. One of the main features of the BIC formulation is that its criterion is *consistent*. This means that if the true unknown distribution that generated the observations happens to belong to the collection of candidate models $\{\theta_1, \ldots, \theta_K\}$, then the BIC solution is guaranteed to select it with probability one in the limit of large datasets — see, e.g., Claeskens and Hjort (2008). This property may also explain the observation in McQuarrie and Tsai (1998) that BIC outperforms AIC for moderate-size datasets in the sense that BIC tends to select the true model more frequently.

It is clear from the derivations in Appendices 31.B and 31.C that the AIC and BIC formulations are based on some approximations, especially asymptotically as the sample size N tends to infinity. This reflects on their behavior. For instance, BIC penalizes the model complexity more heavily than AIC; it uses the penalty term $M_k \ln N$ versus M_k for AIC. For this reason, AIC is more likely to favor more complex models, whereas BIC favors simpler models. For more information on AIC and BIC, their derivations, and applications, readers may refer to Linhart and Zucchini (1986), Ghosh, Delampady, and Samanta (2006), Claeskens and Hjort (2008), Konishi and Kitagawa (2008), Hastie, Tibshirani and Friedman (2009), and Neath and Cavanaugh (2012).

Minimum description length. The minimum length description (MDL) criterion is due to Rissanen (1978,1986). It selects solutions that can represent the model and the data in the most compressed form (in terms of bit representation). This line of reasoning is consistent with what is generally referred to as the *Occam razor principle*. The principle basically states that simpler explanations or hypotheses should be preferred over more complex explanations or hypotheses. We will encounter it again in future Sec. 64.5 when we discuss the issue of overfitting by complex models and the bias-variance trade-off. MDL relies on information-theoretic and coding theory arguments, and exploits to great effect the connection between code lengths and the pdfs of the variables through the notion of entropy. Specifically, $-\log_2 f_{\boldsymbol{x}}(x)$ bits are needed to represent realizations x for the random variable with distribution $f_{\boldsymbol{x}}(x)$. The MDL approach exploits this connection to formulate a design criterion. We explained in the body of the chapter that when prior distributions are assigned to the models, the MDL solution reduces to selecting the best fit from among a collection of MAP estimators. We also explained that MDL and BIC are closely related. In particular, if we use the bit representation (31.182), then the MDL formulation in (31.185) will reduce to BIC. A good overview of MDL is given by Hansen and Yu (2001). For more information, the reader may consult Barron and Cover (1991), Barron, Rissanen, and Yu (1998), and Grunwald (2007).

Frequentist view. The maximum-likelihood approach of this chapter treats the parameter θ as an *unknown but fixed* quantity and does not attach any probability distribution to it. This approach is reminiscent of the *frequentist* viewpoint to probability and statistics. In the frequentist approach, the notion of probability is defined as the long-term *frequency of occurrence* of events, evaluated from repeated experimentation or observation. For example, the probability of landing heads (H) in a coin toss can be determined by repeatedly tossing the coin N times and counting how many heads are observed, say, M times. The ratio M/N then approaches the probability of the event, $\mathbb{P}(H)$, as $N \to \infty$.

We can re-examine the ML formulation in light of this description, namely,

$$\widehat{\theta}_N = \underset{\theta}{\text{argmax}} \left\{ \sum_{n=1}^{N} \ln f_{\boldsymbol{y}}(y_n;\theta) \right\} \tag{31.218}$$

where we are adding the subscript N to indicate that $\widehat{\theta}_N$ is computed from N measurements. This solution leads to a mapping from the observations to the estimate:

$$\left\{ y_n \right\}_{n=1}^{N} \xrightarrow{\text{ML}} \widehat{\theta}_N \tag{31.219}$$

But since the observations $\{y_n\}$ are realizations of a random process, the randomness will reflect itself on the ML estimate as well. Specifically, $\widehat{\theta}_N$ will vary with the measurements $\{y_n\}$: two different collections of N measurements, each arising from the *same* data distribution $f_{\boldsymbol{y}}(y)$, will generally lead to two different values for $\widehat{\theta}_N$. For this reason, we will denote the ML estimate in boldface and write $\widehat{\boldsymbol{\theta}}_N$ to highlight its random nature. Thus, although the ML formalism models θ as an unknown but fixed parameter, its estimator $\widehat{\boldsymbol{\theta}}_N$ is a random quantity with mean and variance. In particular, the Cramer-Rao bound (31.117) provides a lower bound on the expected mean-square error in terms of the inverse of the Fisher information matrix. Obviously, the bound is useful when the Fisher matrix can be evaluated in closed-form. This is often challenging since, as can be seen from expressions (31.102) and (31.99), the designer will need to compute certain expectations.

As befits the frequentist approach, one useful alternative to assess the performance of ML estimators is to resort to a *bootstrap* calculation. The term "bootstrap" is commonly used in the statistical literature to refer to a technique where a statistic (such as mean or variance) is estimated by re-sampling with replacement from existing measurements. We will encounter this approach in other contexts in this text, e.g., when studying bagging classifiers in Chapter 62 and also temporal-difference techniques in reinforcement learning in Chapter 46. Under bootstrap, we re-sample the original N measurements $\{y_n\}$ *with replacement* and obtain another collection of N samples, denoted by $\{y_n^{(1)}\}$. These sample values arise from the *same* original set and some values may appear repeated due to resampling with replacement. We then compute the ML estimate again from this new collection and denote it by:

$$\left\{ y_n^{(1)} \right\}_{n=1}^{N} \xrightarrow{\text{ML}} \widehat{\theta}_N^{(1)} \tag{31.220}$$

We repeat the re-sampling operation multiple times. Each time b leads to a new ML estimate, $\widehat{\theta}_N^{(b)}$. The main advantage of carrying out these bootstrap calculations is that, without collecting any additional data, the estimates $\{\widehat{\theta}_N^{(b)}\}$ lead to a histogram distribution that approximates the pdf for the estimator $\widehat{\boldsymbol{\theta}}_N$. We can subsequently use the estimated pdf to deduce useful statistics about the ML estimator, such as its sample mean, variance, and confidence interval, as illustrated in Fig. 31.7.

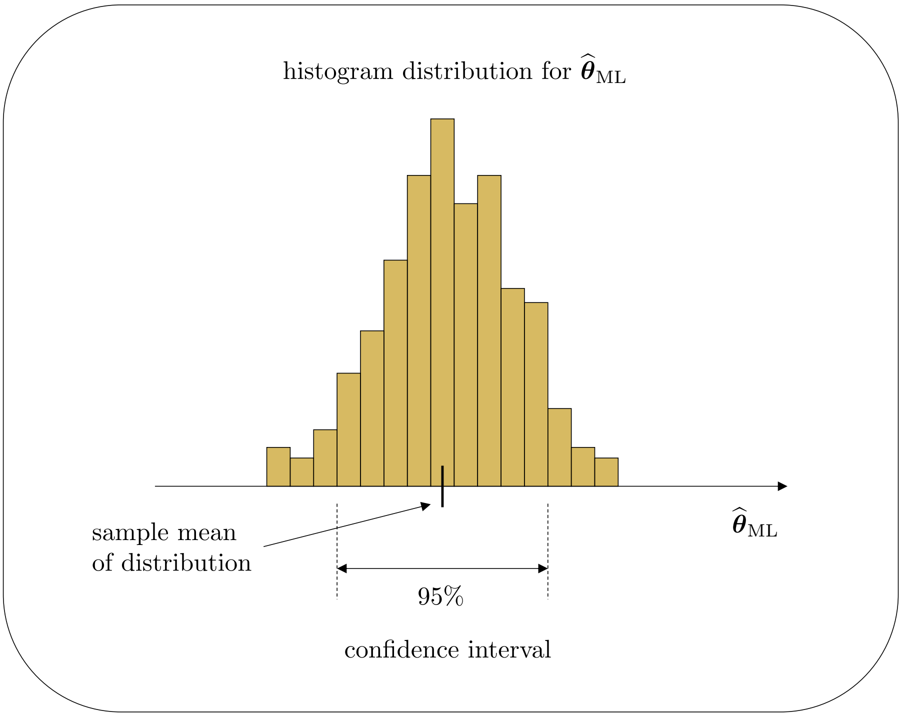

Figure 31.7 Illustration of a histogram constructed for the distribution of a maximum-likelihood estimator $\widehat{\boldsymbol{\theta}}_{\rm ML}$ obtained by applying the bootstrap method.

Bayesian view. In contrast to the frequentist approach, the Bayesian view to inference treats the concept of probability as a measure of *uncertainty* rather than frequency of occurrence. In this case, the probability of an event is a subjective measure and provides an indication of the belief we have about its occurrence. This point of view is particularly useful to model events that do not occur frequently and are therefore difficult to capture by long-term frequency calculations. The early proponent of this interpretation for the notion of probability was the British philosopher and mathematician **Frank Ramsey (1903–1930)** in the work by Ramsey (1931) published posthumously — see also the accounts by Sahlin (2008) and Misak (2020).

Under the Bayesian paradigm, it is possible to assign probabilities to events that are not necessarily repeatable. For example, consider again the case of a likelihood function that is parameterized by some θ, as in (31.218). The objective continues to be finding an estimate for θ. However, in many instances, the designer may have available some prior information about which values for θ are more or less likely. This information can be codified into a probability density function. One can model the unknown as a random variable, $\boldsymbol{\theta}$, and associate a pdf with it, $f_{\boldsymbol{\theta}}(\theta)$. This pdf is called the *prior* and it can be interpreted as a weighting function. The prior models the amount of uncertainty we have about $\boldsymbol{\theta}$: regions of high confidence will have higher likelihood of occurring than regions of lower confidence. For example, assume $\boldsymbol{\theta}$ is a scalar and we know that its value lies in the range $\boldsymbol{\theta} \in [0,1]$. If we adopt a uniform prior over this interval, then we are codifying that all values in this range are equally likely. If, on the other hand, $\boldsymbol{\theta}$ can assume values over $(-\infty,\infty)$ but its true value lies somewhere within $[-1,1]$, then we could perhaps select a Gaussian prior with zero mean and unit standard deviation to reflect this knowledge. Once a prior is selected for $\boldsymbol{\theta}$, the Bayesian approach then seeks the estimate for θ by maximizing the *posterior* likelihood of $\boldsymbol{\theta}$ given the observations. Using Bayes rule, this posterior is given by

$$f_{\boldsymbol{\theta}|\boldsymbol{y}_1,\ldots,\boldsymbol{y}_N}(\theta|y_1,\ldots,y_N) \;=\; \frac{f_{\boldsymbol{\theta}}(\theta)\;\times\;f_{\boldsymbol{y}_1,\ldots,\boldsymbol{y}_N|\boldsymbol{\theta}}(\boldsymbol{y}_1,\ldots,\boldsymbol{y}_N|\boldsymbol{\theta})}{f_{\boldsymbol{y}_1,\ldots,\boldsymbol{y}_N}(y_1,\ldots,y_N)} \tag{31.221}$$

The evidence in the denominator is independent of θ and can be ignored in the maximization so that we can write

$$\underbrace{f_{\boldsymbol{\theta}|\boldsymbol{y}_1,\ldots,\boldsymbol{y}_N}(\theta|y_1,\ldots,y_N)}_{\textsf{posterior}} \;\propto\; \underbrace{f_{\boldsymbol{\theta}}(\theta)}_{\textsf{prior}} \times \underbrace{f_{\boldsymbol{y}_1,\ldots,\boldsymbol{y}_N|\boldsymbol{\theta}}(\boldsymbol{y}_1,\ldots,\boldsymbol{y}_N|\boldsymbol{\theta})}_{\textsf{likelihood}} \tag{31.222}$$

or, equivalently, in the log domain (to bring forth the analogy with the ML approach)

$$\ln f_{\boldsymbol{\theta}|\boldsymbol{y}_1,\ldots,\boldsymbol{y}_N}(\theta|y_1,\ldots,y_N) = \ln f_{\boldsymbol{\theta}}(\theta) \;+\; \sum_{n=1}^{N}\ln f_{\boldsymbol{y}}(y_n;\theta) \;+\; \text{constant} \tag{31.223}$$

The Bayesian approach then seeks the estimate that solves:

$$\widehat{\theta}_N \;=\; \underset{\theta}{\text{argmax}}\left\{\ln f_{\boldsymbol{\theta}}(\theta) \;+\; \sum_{n=1}^{N}\ln f_{\boldsymbol{y}}(y_n;\theta)\right\} \tag{31.224}$$

This construction leads to the *maximum a-posteriori* (MAP) estimate for θ, which we have already encountered in Chapter 28. We find another instance of it in our derivation of the Bayesian information criterion (BIC) in Appendix 31.C. Comparing with the ML formulation (31.218) we observe that the main difference is the appearance of the additional term $\ln f_{\boldsymbol{\theta}}(\theta)$ originating from the prior. We therefore find that the frequentist approach ignores the prior and employs only the likelihood function to arrive at the ML estimate $\widehat{\theta}_N$, while the Bayesian approach keeps the prior and uses it to arrive at the MAP estimate. Table 31.1 compares the main features of the frequentist and Bayesian approaches to inference: the former focuses on finding a θ that best fits the likelihood model to the data, while the latter focuses on finding a θ that best fits the posterior distribution.

Table 31.1 Comparing frequentist and Bayesian approaches.

frequentist inference	**Bayesian inference**
1. ML is prime example	1. MAP is prime example
2. Probability is long-term frequency	2. Probability is measure of uncertainty
3. Model unknown but fixed	3. Model random with uncertainty
4. Does not use a *prior* for the model	4. Uses a *prior* for the model
5. Uses likelihood of data given model	5. Uses likelihood of data given model
6. Finds best-fit model for the data	6. Finds best model for the parameter
7. Usually less complex	7. Finding evidence is challenging

The additional term $\ln f_{\boldsymbol{\theta}}(\theta)$ in (31.224) has another useful interpretation. Consider an example where θ is $M-$dimensional with a Gaussian prior of the form:

$$f_{\boldsymbol{\theta}}(\theta) = \frac{1}{(2\pi)^{M/2}}e^{-\frac{1}{2}\|\theta\|^2} \tag{31.225}$$

then

$$\ln f_{\boldsymbol{\theta}}(\theta) = -\frac{1}{2}\|\theta\|^2 \;+\; \text{constant} \tag{31.226}$$

and the cost that is being maximized in (31.224) becomes

$$\widehat{\theta} \;=\; \underset{\theta}{\operatorname{argmax}} \left\{ -\frac{1}{2}\|\theta\|^2 \;+\; \sum_{n=1}^{N} \ln f_{\boldsymbol{y}}(y_n;\theta) \right\} \tag{31.227}$$

We see that the effect of the additional term is to discourage large values for $\widehat{\theta}$. We refer to $\ln f_{\boldsymbol{\theta}}(\theta)$ in (31.224) as a *regularization* term, and we will study its effect more closely in Chapter 51 where we will also consider other choices for the regularization factor. We will find out then that different choices for this factor help infuse into $\widehat{\theta}$ certain desirable properties such as forcing them to have small norm or sparse structure.

The frequentist and Bayesian approaches lead to different but related results in general. To illustrate the difference, we consider in Prob. 31.8 a random variable $\boldsymbol{y}$ that follows a binomial distribution with parameters N and p, i.e., the probability of observing $y = k$ successes in N trials is given by:

$$\mathbb{P}(\boldsymbol{y} = k) = \binom{N}{k} p^k (1-p)^{N-k}, \quad k = 0, 1, \ldots, N \tag{31.228a}$$

We show in that problem that having observed $\boldsymbol{y} = y$, the maximum-likelihood estimate for the probability of success p is

$$\widehat{p}_{\text{ML}} = y/N \tag{31.228b}$$

whereas in the earlier Prob. 28.11 we modeled $\boldsymbol{p}$ as a random variable that follows a beta distribution with shape parameters $(2, 1)$. We found then that the MAP estimate for p is given by

$$\widehat{p}_{\text{MAP}} \;=\; \frac{y+1}{N+1}, \quad \text{assuming } \boldsymbol{p} \sim \text{Beta}(2,1) \tag{31.228c}$$

Observe that the expressions for both estimates are different, although they tend to each other for large sample sizes, N. The fact that the expressions are different should not come as a surprise. After all, while the ML formulation is using solely the observation y to estimate p, the Bayesian formulation is using one additional piece of information represented by the assumption of a Beta distribution for $\boldsymbol{p}$.

We will encounter repeated instances of the Bayesian formulation in our treatment. One of the main difficulties that arises in this technique is the following. Referring back to expression (31.221), we indicated that the evidence in the denominator is independent of θ and can therefore be ignored in the process of seeking $\widehat{\theta}$. However, in many instances, we still desire to know the resulting posterior distribution that appears on the left-hand side. We will explain in future chapters that the posterior quantity is useful in many cases, for example, to predict future values for $\boldsymbol{y}$ from past observations — see Chapter 33. The difficulty lies in computing the evidence that appears in the denominator. We will describe several techniques in later chapters for this purpose including variational inference methods and Markov chain Monte Carlo (MCMC) methods.

In summary, the Bayesian formulation is anchored on the Bayes rule for mapping priors to posteriors. Although the frequentist approach was popular in the 20th Century, the Bayesian approach has become more prominent in recent years due to several theoretical and computational advances described in later chapters. Its main challenge continues to be selecting a suitable prior that conforms to the physical reality of $\boldsymbol{\theta}$. This is actually one of the main criticisms directed at the Bayesian approach: the prior is often selected in a subjective manner and for the convenience of mathematical tractability; it need not represent a faithful codification of the truth about the unknown, $\boldsymbol{\theta}$. Moreover, different priors will lead to different MAP estimates even for the same measurements.

This old-age debate about the merits of the frequentist and Bayesian approaches is likely to continue for decades to come. However, both approaches have merit and have proven useful in many contexts. It is better to view them as complementary rather than competing or conflicting approaches. While both methodologies assume knowledge of the likelihood function, we should note that the likelihood expressions are not exact but rather approximations in most cases anyway. In this way, the ML and Bayesian inference approaches seek, in their own ways, parameter estimates that "best" explain the observed data and they lead to good performance in many applications of interest. One may view the Bayesian formulation as a "regularized" frequentist formulation, in which case more commonalities link these two approaches than actual differences. The main distinction between them then becomes one of interpreting what their respective costs mean. For more discussion on the frequentist and Bayesian approaches, the reader may refer to Savage (1954), de Finetti (1974), Samaniego and Reneau (1994), Barnett (1999), Samaniego (2010), Wakefield (2013), and VanderPlas (2014).

Heart disease Cleveland dataset. Figure 31.1 relies on data derived from the heart-disease Cleveland dataset. The dataset consists of 297 samples that belong to patients with and without heart disease. It is available on the UCI Machine Learning Repository at https://archive.ics.uci.edu/ml/datasets/heart+Disease. The investigators responsible for the collection of the data are the leading 4 co-authors of the article by Detrano *et al.* (1989).

PROBLEMS

31.1 Establish relations (31.24).
31.2 Establish the validity of (31.29).
31.3 Let $\boldsymbol{y}(n) = x + \boldsymbol{v}(n)$, where x is an unknown scalar constant and $\boldsymbol{v}(n)$ is zero-mean white noise with power σ_v^2. An estimator for x is constructed recursively in the following manner:

$$\widehat{\boldsymbol{x}}(n) \;=\; (1-\alpha)\widehat{\boldsymbol{x}}(n-1) \;+\; \alpha\,\boldsymbol{y}(n), \quad n \geq 0$$

starting from $\widehat{\boldsymbol{x}}(-1) = 0$ and where $0 < \alpha < 1$. Determine the steady-state mean and variance of $\widehat{\boldsymbol{x}}(n)$ as $n \to \infty$. Any optimal choice for α?
31.4 Consider a vector-valued Gaussian distribution $\boldsymbol{y} \sim \mathbb{N}_{\boldsymbol{y}}(\mu, R_y)$ where $\boldsymbol{y} \in \mathbb{R}^M$. Follow the same maximum-likelihood arguments from Sec. 31.1 to motivate the following unbiased estimates for (μ, R_y) from N independent realizations $\{y_n\}$:

$$\widehat{\mu} = \frac{1}{N}\sum_{n=1}^{N} y_n, \qquad \widehat{R}_y = \frac{1}{N-1}\sum_{n=1}^{N}(y_n - \widehat{\mu})(y_n - \widehat{\mu})^{\mathsf{T}}$$

31.5 Consider the same setting of Prob. 28.10. Assume we collect N independent realizations $\{y_n\}$ for $n = 1, 2, \ldots, N$.
(a) Verify that

$$f_{\boldsymbol{y}_1,\ldots,\boldsymbol{y}_N}(y_1,\ldots,y_N;\lambda) = \frac{\lambda^S e^{-N\lambda}}{\prod_{n=1}^{N} y_n!}$$

where $S = \sum_{n=1}^{N} y_n$, and conclude that the maximum-likelihood estimate of λ is given by $\widehat{\lambda}_{\text{ML}} = \frac{1}{N}\sum_{n=1}^{N} y_n$. Is the estimator unbiased?
(b) Show that $T(y) = \frac{1}{N}\sum_{n=1}^{N} y_n$ is a sufficient statistic for λ.
31.6 A random variable $\boldsymbol{y}$ is uniformly distributed over the interval $0 \leq y \leq a$. We observe N independent and identically distributed realizations $\{\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_N\}$ and we wish to determine the maximum-likelihood estimate for a.

(a) Verify that

$$f_{\boldsymbol{y}_1,\ldots,\boldsymbol{y}_N}(y_1,\ldots,y_N;a) = \frac{1}{a^N}\,\mathbb{I}\left[0 \leq \max_{1\leq n\leq N} y_n \leq a\right]$$

where $\mathbb{I}[x]$ denotes the indicator function and was defined earlier in (4.164).

(b) Conclude that $\widehat{a}_{\rm ML} = \max_{1\leq n\leq N} y_n$.

(c) Show that the estimator is biased, namely, establish that $\mathbb{E}\,\widehat{\boldsymbol{a}}_{\rm ML} = Na/(N+1)$.

(d) Show that $T(y) = \max_{1\leq n\leq N} y_n$ is a sufficient statistic for a.

31.7 Assume $\boldsymbol{y}$ is an exponentially-distributed random variable with rate $\lambda > 0$, i.e., $f_{\boldsymbol{y}}(y;\lambda) = \lambda e^{-\lambda y}$ for $y \geq 0$. Verify that the maximum-likelihood estimate of λ given N independent and identically distributed realizations $\{y_1, y_2, \ldots, y_N\}$ is $\widehat{\lambda} = N/\sum_{n=1}^{N} y_n$. Show that

$$\mathbb{E}\,\widehat{\boldsymbol{\lambda}} = \frac{N\lambda}{N-1}, \quad \text{var}(\widehat{\boldsymbol{\lambda}}) = \frac{N^2\lambda^2}{(N-1)^2(N-2)}$$

Is the ML estimator efficient in this case?

31.8 A random variable $\boldsymbol{y}$ follows a binomial distribution with parameters N and p, i.e., the probability of observing k successes in N trials is given by:

$$\mathbb{P}(\boldsymbol{y} = k) = \binom{N}{k} p^k (1-p)^{N-k}, \quad k = 0, 1, \ldots, N$$

(a) Having observed $\boldsymbol{y} = y$, show that the maximum-likelihood estimate of p when N is known is given by $\widehat{p}_{\rm ML} = y/N$.

(b) Having observed $\boldsymbol{y} = y$, show that the maximum-likelihood estimate of N when p is known is given by the smallest value of $\widehat{N}$ that satisfies $\widehat{N}_{\rm ML} + 1 \geq y/p$.

(c) Under part (b), assume y/p is an integer. Conclude that there are two ML estimates given by $\widehat{N}_{\rm ML} = y/p$ and $\widehat{N}_{\rm ML} = (y/p) - 1$.

Remark. We compare in Probs. 28.11 and 28.12 the above ML solution to the MAP (maximum a-posteriori) and MMSE (minimum mean-square-error) solutions.

31.9 Derive expressions (31.41)–(31.42).

31.10 Is the unbiased estimator (31.25) efficient? That is, does it attain the Cramer-Rao lower bound?

31.11 Let $\boldsymbol{y}$ be a Bernoulli random variable that is equal to one with probability p and equal to zero with probability $1-p$. Show that the Fisher information value for p is given by $F(p) = 1/p(1-p)$.

31.12 Let $\boldsymbol{y}$ be distributed according to a Poisson distribution with mean $\lambda \geq 0$, i.e.,

$$\mathbb{P}(\boldsymbol{y} = k) = \frac{\lambda^k e^{-\lambda}}{k!}, \quad k = 0, 1, 2, \ldots$$

where λ is the average number of events occurring in an interval of time. Show that the Fisher information value for λ is given by $F(\lambda) = 1/\lambda$. Show further that the maximum-likelihood estimate of λ from N independent and identically distributed observations $\{\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_N\}$ is efficient.

31.13 Let $\boldsymbol{y}$ be distributed according to the beta distribution

$$f_{\boldsymbol{y}}(y;\theta) = ay^{a-1}, \quad y \in (0,1), \quad a > 0$$

Assume we collect N independent and identically distributed observations $\{\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_N\}$.

(a) Show that the maximum-likelihood estimate of a is given by

$$\widehat{a}_{\rm ML} \;=\; -\left(\frac{1}{N}\sum_{n=1}^{N} \ln y_n\right)^{-1}$$

(b) Verify that the variance of $\widehat{a}_{\rm ML}$ is given by

$$\sigma^2_{\hat{a}_{\rm ML}} = \frac{N^2a^2}{(N-1)^2(N-2)}$$

(c) Show that the Fisher information value for a is $F(a) = N/a^2$.
(d) Is the maximum-likelihood estimator efficient?

31.14 Consider the vector-valued Gaussian distribution (31.105) with a diagonal covariance matrix. Introduce the vector of parameters

$$\theta = \text{col}\{\mu_1, \mu_2, \ldots, \mu_P, \ln\sigma_1, \ln\sigma_2, \ldots, \ln\sigma_P\}$$

Compute the Fisher information matrix $I(\theta)$ relative to these parameters.

31.15 Conclude from (31.122) that the mean-square error for any biased estimator satisfies

$$\mathbb{E}\,\widetilde{\boldsymbol{\theta}}^2 \;\geq\; (\theta - g(\theta))^2 \;-\; \left(\mathbb{E}\,\frac{\partial^2 \ln f_{\boldsymbol{y}}(y;\theta)}{\partial^2\theta}\right)^{-1}\left(\frac{\partial g(\theta)}{\partial\theta}\right)^2$$

where $\theta - g(\theta)$ is the bias.

31.16 Establish expression (31.209) for the conditional distribution of the observations given the statistic $T(y)$.

31.17 Establish the validity of inequality (31.211) given by the Rao-Blackwell theorem.

31.18 Refer to the log-likelihood function (31.133) and verify that the two regularity conditions (31.229)–(31.230) are satisfied.

31.19 Let $\boldsymbol{y}$ denote a random variable that is uniformly distributed over the interval $[0,\theta]$, where $\theta > 0$ is an unknown parameter, i.e., $f_{\boldsymbol{y}}(y;\theta) = 1/\theta$ for $0 \leq y \leq \theta$.
(a) Argue that $\widehat{\boldsymbol{\theta}} = 2\boldsymbol{y}$ is an unbiased estimator for θ, i.e., $\mathbb{E}\,\widehat{\boldsymbol{\theta}} = \theta$.
(b) Verify that the second regularity condition (31.230) fails in this case.
(c) Verify that

$$\mathbb{E}\,\widetilde{\boldsymbol{\theta}}^2 = \theta^2/3, \qquad \mathbb{E}\left(\frac{\partial \ln f_{\boldsymbol{y}}(y;\theta)}{\partial\theta}\right)^2 = 1/\theta^2$$

and conclude that the Cramer-Rao bound result does not hold in this case.

31.20 Consider the problem of estimating the mean of a sequence of N observations $\{\boldsymbol{y}(n)\}$ drawn from a Gaussian distribution with mean μ and known variance σ_y^2.
(a) Assume that $\mu \in \mathbb{R}$. Let $\widehat{\boldsymbol{\mu}}_{\rm real}$ denote the maximum-likelihood estimator for μ from the observations $\{\boldsymbol{y}(n)\}$. What is the variance of $\widehat{\boldsymbol{\mu}}_{\rm real}$? Is the estimator $\widehat{\boldsymbol{\mu}}_{\rm real}$ unbiased?
(b) Continue assuming that $\mu \in \mathbb{R}$. What is the Cramér-Rao bound (lower bound on $\text{var}(\widehat{\boldsymbol{\mu}}_{\rm real})$) for the parameter μ as a function of σ_y^2 and N. Is the estimator $\widehat{\boldsymbol{\mu}}_{\rm real}$ efficient?
(c) Now assume that $\mu \in \mathbb{Z}$, where $\mathbb{Z}$ denotes the set of integers. Consider the estimator $\widehat{\boldsymbol{\mu}}_{\rm integer} = \text{round}(\widehat{\boldsymbol{\mu}}_{\rm real})$, where $\text{round}(x) : \mathbb{R} \rightarrow \mathbb{Z}$ returns the nearest-integer to the real number x (where $\text{round}(x-1/2) = x$ when x itself is an integer). Compute the probability mass function of $\widehat{\boldsymbol{\mu}}_{\rm integer}$ and express it using the standard Gaussian cumulative distribution function (CDF) $Q(x)$ defined earlier in part (b) of Prob. 4.31. Based on the probability mass function, is the estimator $\widehat{\boldsymbol{\mu}}_{\rm integer}$ unbiased?
(d) Using the fact that the variance of $\widehat{\boldsymbol{\mu}}_{\rm integer}$ can be written as:

$$\text{var}(\widehat{\boldsymbol{\mu}}_{\rm integer}) = Q\left(\frac{1}{2}\frac{\sqrt{N}}{\sigma_y}\right) - 2Q\left(\frac{N}{\sigma_y}\right)$$

where

$$Q(x) \sim \sqrt{\frac{2}{\pi}}\,\frac{1}{x}\sum_{y=1}^{\infty} e^{-\frac{1}{2}y^2x^2}, \qquad \text{as } x \rightarrow \infty$$

show that the variance of the rounded estimator $\widehat{\boldsymbol{\mu}}_{\text{integer}}$ is of the form:

$$\text{var}(\widehat{\boldsymbol{\mu}}_{\text{integer}}) \sim \sqrt{\frac{8\sigma_y^2}{\pi N}} e^{-\frac{N}{8\sigma_y^2}}, \quad \text{as } N/\sigma_y^2 \to \infty$$

(e) Compare the variances of parts (a) and (d). How fast does the variance of each estimator decay as a function of N? Is this a surprising result?

Remark. For more details on these results, the reader may refer to Hammersley (1950).

31.21 Refer to the expressions in Table 5.1 for several traditional probability distributions. Use result (31.81) to determine the ML estimates for the parameter θ in each case. Express your results in terms of estimates for the original parameters for the various distributions.

31.22 Refer to expression (31.79) for the log-likelihood function of an exponential distribution. Establish the identity

$$\frac{1}{N}\nabla_{\theta^{\mathsf{T}}} \ln f_{\boldsymbol{y}_1,\ldots,\boldsymbol{y}_N}(y_1,\ldots,y_N;\theta) = \frac{1}{N}\sum_{n=1}^{N} T(y_n) \; - \; \mathbb{E}\,T(\boldsymbol{y})$$

31.23 Refer to the least-squares problem (31.144) with unknown scalar parameters $\{a,b,c\}$. Determine the estimates $\{\widehat{a},\widehat{b},\widehat{c}\}$.

31.24 Refer to expressions (31.194)–(31.196) derived under an MDL formulation for selecting the best of two models.

(a) Maximize $J(\mu_1)$ over the parameter μ_1 and determine an expression for $\widehat{\mu}_1$. Ignore the constant term and evaluate $J(\widehat{\mu}_1)$.

(b) Maximize $J(\mu_2,\sigma_2^2)$ over the parameters (μ_2,σ_2^2) and determine expressions for $(\widehat{\mu}_2,\widehat{\sigma}_2^2)$. Ignore the constant term and evaluate $J(\widehat{\mu}_2,\widehat{\sigma}_2^2)$.

(c) Which model would you choose according to the MDL criterion?

31.25 We continue with Prob. 31.24 but assume now that the number of bits needed to represent each model is computed based on (31.181). In this case, the cost functions $J(\mu_1)$ and $J(\mu_2,\sigma_2^2)$ will be replaced by

$$J(\mu_1) = -\frac{1}{2}\ln(N) - \frac{1}{2\sigma_1^2}\sum_{n=1}^{N}(y_n-\mu_1)^2 \; + \; \text{cte}$$

$$J(\mu_2,\sigma_2^2) = -\ln(N) - \frac{1}{2\sigma_2^2}\sum_{n=1}^{N}(y_n-\mu_2)^2 \; + \; \text{cte}$$

Repeat the derivations of Prob. 31.24 for this case.

31.26 Refer to the least-squares problem (31.176). Differentiate the cost with respect to w and establish the validity of the solution $w_M^\star$ and the corresponding minimum cost $\mathcal{E}_M$ given by(31.177a)–(31.177b).

31.27 Consider N independent measurements $\{y_1,y_2,\ldots,y_N\}$ and introduce their sample mean and sample variance estimates:

$$\widehat{\mu} = \frac{1}{N}\sum_{n=1}^{N} y_n, \qquad \widehat{\sigma}^2 = \frac{1}{N}\sum_{n=1}^{N}(y_n-\widehat{\mu})^2$$

We wish to use the data to decide between two models, $\mathcal{N}_{\boldsymbol{y}}(0,1)$ and $\mathcal{N}_{\boldsymbol{y}}(a,1)$. One of the models has $\theta=0$ and $M_\theta=0$ while the other model has $\theta=a$ and $M_\theta=1$. Here, M_θ denotes the number of parameters under model θ. Show that the AIC scores for both models are given by

$$\text{AIC}(0) = N\ln(2\pi) + N\widehat{\sigma}^2 + N\widehat{\mu}^2$$

$$\text{AIC}(a) = 2 + N\ln(2\pi) + N\widehat{\sigma}^2$$

Conclude that model $\mathcal{N}_{\boldsymbol{y}}(a, 0)$ is selected whenever $\widehat{\mu}^2 > 2/N$. Verify that under BIC the condition changes to $\widehat{\mu}^2 > (\ln N)/N$.

31.28 Use approximation (31.279) for large sample size N in the BIC formulation to justify the expression

$$\pi(k|y_1 \ldots, y_N) \;=\; \frac{e^{\text{BIC}(k)/2}}{\sum_{k'=1}^{K} e^{\text{BIC}(k')/2}}$$

31.29 Refer to the cross validation construction (31.204). Assume, for simplicity, that cross validation performs a single pass over the data so that, for each model θ_k, a single estimate $\widehat{X}_{k,1}$ is generated using $\widehat{X}_{k,1} = \frac{1}{T}\sum_{t=1}^{T} \ln f_{\boldsymbol{y}}(y_t; \widehat{\theta}_k)$. Assume further that the log-likelihood functions are bounded, say, $0 \leq \ln f_{\boldsymbol{y}}(y; \widehat{\theta}_\ell) \leq b$ for all y.

(a) For any small $\delta > 0$, use Hoeffding inequality (3.232b) to verify that $\mathbb{P}(|\widehat{\boldsymbol{X}}_{k,1} - X_k| \geq \delta) \leq 2e^{-2T\delta^2/b^2}$.

(b) Show how to select δ to ensure that with high probability $1 - \epsilon$:

$$\max_{1 \leq k \leq K} |\widehat{X}_{k,1} - X_k| < \left(\frac{b^2 \ln(2/\epsilon)}{2T}\right)^{1/2}$$

(c) Explain that

$$\max_{1 \leq k \leq K} \widehat{X}_{k,1} \text{ is close to } \max_{1 \leq k \leq K} X_k \text{ with high probability}$$

Quantify the likely distance between $\max_k \widehat{X}_{k,1}$ and $\max_k X_k$.

31.30 This example is extracted from Eddy (2004) and VanderPlas (2014). Alice and Bob are present in a room with a billiard table; they cannot see the table. Their friend Carol rolls a ball down the table and marks the location where it lands. Subsequently, Carol rolls more balls. If the ball lands to the left of the mark, a point is given to Alice. Otherwise, a point is given to Bob. The first person to reach six points wins the game. After 8 throws, Alice has 5 points and Bob has 3 points. We wish to evaluate the chance of Bob winning the game using the frequentist and Bayesian approaches.

(a) Based on the frequentist approach, the probability of a ball throw favoring Alice is $\widehat{p} = 5/8$. The same probability for Bob is $1 - \widehat{p} = 3/8$. Use these values to estimate the likelihood that Bob will win the game.

(b) Based on the Bayesian approach, we model the unknown success probability for Alice as a random variable $\boldsymbol{p}$. Assume a uniform prior for $\boldsymbol{p}$, i.e., $f_{\boldsymbol{p}}(p) = 1$ for $p \in [0, 1]$. The measurements in this problem are $y_a = 5$ and $y_b = 3$ (the number of balls successfully assigned to Alice and Bob, respectively). Use Bayes' rule to show that

$$\mathbb{P}\Big(\text{Bob wins} \,|\, y_a = 5, y_b = 3\Big) = \frac{\displaystyle\int_0^1 p^5(1-p)^6 dp}{\displaystyle\int_0^1 p^5(1-p)^3 dp}$$

Evaluate the expression and compare your result with part (a). Comment on the difference.

31.31 Let $\{\boldsymbol{y}_n\}$ denote a collection of N random variables, for $n = 1, 2, \ldots, N$, with each variable $\boldsymbol{y}_n$ distributed according to its own individual pdf, denoted by $f_n(y)$, with its own mean μ_n. In many problems of interest (e.g., in economics, decision-making, and in reinforcement learning studied later in Example **??**), the objective is to estimate the maximum mean value of the variables, i.e., to solve problems of the type:

$$\mu_{\max} \;\stackrel{\Delta}{=}\; \max_{1 \leq n \leq N} \; \mathbb{E}\,\boldsymbol{y}_n$$

If the means $\{\mu_n\}$ were known beforehand, then the answer is obviously the largest value among the $\{\mu_n\}$, i.e., $\mu_{\max} = \max_{1\leq n\leq N} \mu_n$. The challenge is to estimate $\mu_{\max}$ directly from observations. Thus, assume we collect M_n independent and identically distributed observations for each variable $\boldsymbol{y}_n$. We denote the observations by $\{y_{n,m},\ m = 1,\ldots,M_n\}$. These observations can be used to determine unbiased estimates for the $\{\mu_n\}$ using the sample mean calculation $\widehat{\mu}_n = (1/M_n)\sum_{m=1}^{M_n} y_{n,m}$. Now consider the estimate construction:

$$\widehat{\mu}_{\max} \overset{\Delta}{=} \max_{1\leq n\leq N} \widehat{\mu}_n$$

(a) Show that the estimator $\widehat{\boldsymbol{\mu}}_{\max}$ constructed in this manner for $\mu_{\max}$ is biased. It is sufficient to provide an example.

(b) Let $\widehat{\boldsymbol{y}}_n$ denote any unbiased estimator for $\boldsymbol{y}_n$, i.e., $\mathbb{E}\,\widehat{\boldsymbol{y}}_n = \mathbb{E}\,\boldsymbol{y}_n$. Show that $\widehat{\boldsymbol{\mu}}_{\max}$ is an unbiased estimator for the alternative problem $\mu_{\max} = \mathbb{E}\,\max_{1\leq n\leq N}\ \widehat{\boldsymbol{y}}_n$ (observe how the maximization and expectation operations are switched relative to the original problem).

Remark. The bias of the maximum sample mean estimator is well-studied in the literature, with ramifications in the fields of economics and management — see, e.g., the works by Capen, Clapp, and Campbell (1971), Smith and Winkler (2006), Thaler (1988), and Van den Steen (2004).

31.32 We continue with the setting of Prob. 31.31. Assume we collect, for each variable $\boldsymbol{y}_n$, two disjoint sets of independent and identically distributed observations with $M_n^{(a)}$ and $M_n^{(b)}$ samples in each. We use these measurements to compute two sample means for $\boldsymbol{y}_n$, namely,

$$\widehat{\mu}_n^{(a)} = \frac{1}{M_n^{(a)}}\sum_{m=1}^{M_n^{(a)}} y_{n,m}, \qquad \widehat{\mu}_n^{(b)} = \frac{1}{M_n^{(b)}}\sum_{m=1}^{M_n^{(b)}} y_{n,m}$$

where we are using the superscripts (a) and (b) to distinguish between the two sample means. The observations $\{y_{n,m}\}$ used in each expression originate from the corresponding set of measurements. Both sample means $\{\widehat{\mu}_n^{(a)}, \widehat{\mu}_n^{(b)}\}$ are unbiased estimators for the *same* mean, μ_n, i.e.,

$$\mathbb{E}\,\widehat{\boldsymbol{\mu}}_n^{(a)} = \mathbb{E}\,\widehat{\boldsymbol{\mu}}_n^{(b)} = \mathbb{E}\,\boldsymbol{y}_n$$

Let $\mathcal{N}$ be the set of indexes that maximize the expected values of the $\{\boldsymbol{y}_n\}$:

$$\mathcal{N} = \left\{ n^\star \;\middle|\; n^\star = \underset{1\leq n\leq N}{\text{argmax}}\ \mathbb{E}\,\boldsymbol{y}_n \right\}$$

Let n^a be an index that maximizes $\{\widehat{\mu}_n^{(a)}\}$, i.e.,

$$n^a = \underset{1\leq n\leq N}{\text{argmax}}\ \widehat{\mu}_n^{(a)}$$

and consider the sample mean from set (b) that corresponds to this same index, namely, $\widehat{\mu}_{n^a}^{(b)}$ (that is, we maximize over one sample mean set and consider the corresponding sample mean from the other set). Verify that $\widehat{\boldsymbol{\mu}}_{n^a}^{(b)}$ continues to be a biased estimator for $\mu_{\max}$ but that it does not overestimate it in the sense that

$$\mathbb{E}\,\widehat{\boldsymbol{\mu}}_{n^a}^{(b)} \;\leq\; \mu_{\max} \overset{\Delta}{=} \max_{1\leq n\leq N}\ \mathbb{E}\,\boldsymbol{y}_n$$

Show further that the inequality is strict if, and only if, $\mathbb{P}(n^a \notin \mathcal{N}) > 0$. *Remark.* For further motivation and discussion on this construction, the reader may refer to van Hasselt (2010).

31.A DERIVATION OF CRAMER-RAO BOUND

We provide in this appendix one derivation for the Cramer-Rao bounds (31.120a)–(31.120b) and (31.122) for the case of scalar parameters by following an argument similar to Cassella and Berger (2002), Frieden (2004), and Van Trees (1968,2013); the argument can be extended with proper adjustments to the vector case. The derivation relies on two regularity conditions on the density function, namely,

(a) The score function exists and is finite, i.e., for every y where $f_{\boldsymbol{y}}(y;\theta) > 0$, it should hold that

$$\mathcal{S}(\theta) \;\triangleq\; \frac{\partial \ln f_{\boldsymbol{y}}(\boldsymbol{y};\theta)}{\partial\theta} < \infty \tag{31.229}$$

(b) It is possible to exchange the operations of integration and differentiation in the equality below involving $\widehat{\theta}$ and the distribution of the observation, i.e.,

$$\frac{\partial}{\partial\theta}\left(\int_{-\infty}^{\infty}\widehat{\theta} f_{\boldsymbol{y}}(y;\theta)dy\right) \;=\; \int_{-\infty}^{\infty}\widehat{\theta}\left(\frac{\partial}{\partial\theta}f_{\boldsymbol{y}}(y;\theta)\right)dy \tag{31.230}$$

where the estimate $\widehat{\theta}$ does not depend on θ.

It is straightforward to verify that these conditions imply that the score function has zero mean since

$$\begin{aligned}
\mathbb{E}\,\mathcal{S}(\theta) &= \mathbb{E}\left(\frac{\partial \ln f_{\boldsymbol{y}}(\boldsymbol{y};\theta)}{\partial\theta}\right)\\
&= \int_{-\infty}^{\infty}\left(\frac{\partial \ln f_{\boldsymbol{y}}(y;\theta)}{\partial\theta}\right)f_{\boldsymbol{y}}(y;\theta)dy\\
&= \int_{-\infty}^{\infty}\frac{1}{f_{\boldsymbol{y}}(y;\theta)}\left(\frac{\partial f_{\boldsymbol{y}}(y;\theta)}{\partial\theta}\right)f_{\boldsymbol{y}}(y;\theta)dy\\
&= \int_{-\infty}^{\infty}\frac{\partial f_{\boldsymbol{y}}(y;\theta)}{\partial\theta}dy\\
&\overset{(31.230)}{=} \frac{\partial}{\partial\theta}\left(\underbrace{\int_{-\infty}^{\infty}f_{\boldsymbol{y}}(y;\theta)dy}_{=1}\right)\\
&= 0
\end{aligned} \tag{31.231}$$

The derivation that follows is not limited to unbiased estimators $\widehat{\boldsymbol{\theta}}$ and will lead to the more general statement (31.122) when $\mathbb{E}\,\widehat{\boldsymbol{\theta}} = g(\theta)$, for some function $g(\cdot)$. Thus, consider the following sequence of calculations involving the correlation between the score function and $\widehat{\boldsymbol{\theta}}$, namely,

$$\begin{aligned}
\mathbb{E}\,(g(\theta)-\widehat{\boldsymbol{\theta}})\mathcal{S}(\theta) &= \mathbb{E}\,g(\theta)\mathcal{S}(\theta) \;-\; \mathbb{E}\,\widehat{\boldsymbol{\theta}}\mathcal{S}(\theta) \\
&= g(\theta)\underbrace{\mathbb{E}\,\mathcal{S}(\theta)}_{=0} \;-\; \mathbb{E}\,\widehat{\boldsymbol{\theta}}\mathcal{S}(\theta) \\
&= -\mathbb{E}\,\widehat{\boldsymbol{\theta}}\mathcal{S}(\theta) \\
&= -\int_{-\infty}^{\infty}\widehat{\theta}\left(\frac{\partial \ln f_{\boldsymbol{y}}(y;\theta)}{\partial\theta}\right)f_{\boldsymbol{y}}(y;\theta)dy \\
&= -\int_{-\infty}^{\infty}\widehat{\theta}\frac{1}{f_{\boldsymbol{y}}(y;\theta)}\left(\frac{\partial f_{\boldsymbol{y}}(y;\theta)}{\partial\theta}\right)f_{\boldsymbol{y}}(y;\theta)dy \\
&= -\int_{-\infty}^{\infty}\widehat{\theta}\left(\frac{\partial f_{\boldsymbol{y}}(y;\theta)}{\partial\theta}\right)dy \\
&\overset{(31.230)}{=} -\frac{\partial}{\partial\theta}\left(\int_{-\infty}^{\infty}\widehat{\theta}f_{\boldsymbol{y}}(y;\theta)dy\right) \\
&= -\frac{\partial}{\partial\theta}\,\mathbb{E}_{\boldsymbol{y}}\,\widehat{\boldsymbol{\theta}} \\
&= -\frac{\partial g(\theta)}{\partial\theta}
\end{aligned} \tag{31.232}$$

We now call upon the Cauchy-Schwarz inequality for random variables, which states that — recall Prob. 3.12:

$$\left\{\mathbb{E}\,(g(\theta)-\widehat{\boldsymbol{\theta}})\mathcal{S}(\theta)\right\}^2 \le \mathbb{E}\,(g(\theta)-\widehat{\boldsymbol{\theta}})^2\,\mathbb{E}\,(\mathcal{S}(\theta))^2 \tag{31.233}$$

so that, using (31.232), we have

$$\left(\frac{\partial g(\theta)}{\partial\theta}\right)^2 \le \mathbb{E}\,(g(\theta)-\widehat{\boldsymbol{\theta}})^2\,\mathbb{E}\,(\mathcal{S}(\theta))^2 \tag{31.234}$$

or, equivalently,

$$\mathbb{E}\,(g(\theta)-\widehat{\boldsymbol{\theta}})^2 \ge \frac{1}{\mathbb{E}\,(\mathcal{S}(\theta))^2}\left(\frac{\partial g(\theta)}{\partial\theta}\right)^2 \;=\; \frac{1}{F(\theta)}\left(\frac{\partial g(\theta)}{\partial\theta}\right)^2 \tag{31.235}$$

where we used the fact that the information value, $F(\theta)$, coincides with the variance of the score function. The above inequality coincides with (31.122).

31.B DERIVATION OF THE AIC FORMULATION

In this appendix we motivate the cost function that is optimized in (31.154) by the AIC criterion by following the argument from Cavanaugh (1997) adjusted to our notation and conventions. The objective is to devise a criterion that selects the “best” fit from a collection of K models $\{\theta_1,\ldots,\theta_K\}$ using data measurements $\{y_1,y_2,\ldots,y_N\}$.

Let $f_{\boldsymbol{y}}(y)$ denote the *true* unknown pdf of the variable $\boldsymbol{y}$. We will simplify the notation and write $f(y)$ without the subscript. Let $f(y;\theta_k)$ denote a model for the unknown pdf that is parameterized by θ_k of size M_k. We estimate θ_k by maximizing the log-likelihood function:

$$\widehat{\theta}_k \;=\; \underset{\theta_k}{\text{argmax}}\left\{\sum_{n=1}^{N}\ln f(y_n|\theta_k)\right\} \tag{31.236}$$

One way to select a "best fit" model from among the K models is to minimize the KL divergence between the true model and its approximation, namely,

$$k^\star \;=\; \underset{1\le k\le K}{\operatorname{argmin}} \left\{ D_{\rm KL}\Big(f(y)\,\|\,f(y;\widehat{\theta}_k)\Big)\right\} \;\;\Longrightarrow \theta^\star = \widehat{\theta}_{k^\star} \tag{31.237}$$

where

$$D_{\rm KL}\Big(f(y)\,\|\,f(y;\widehat{\theta}_k)\Big) \;=\; \underbrace{\int_y f(y)\ln f(y)dy}_{-\textbf{entropy}} \;-\; \underbrace{\int_y f(y)\ln f(y;\widehat{\theta}_k)dy}_{-\textbf{cross-entropy}} \tag{31.238}$$

The first term on the right-hand side is independent of k. Therefore, the problem of selecting the optimal k reduces to

$$k^\star \;=\; \underset{1\le k\le K}{\operatorname{argmax}} \left\{ \int_y f(y)\ln f(y;\widehat{\theta}_k)dy \right\} \;=\; \underset{1\le k\le K}{\operatorname{argmax}} \left\{ \mathbb{E}_{\boldsymbol{y}} \ln f(y;\widehat{\theta}_k)\right\} \tag{31.239}$$

The quantity that is being maximized is the mean of $\ln f(y;\widehat{\theta}_k)$ over the true distribution of $\boldsymbol{y}$; in this case, the argument y of $f(y;\widehat{\theta}_k)$ is treated as a random variable. The expectation is not available since $f(y)$ is unknown. The AIC formulation uses the available data measurements to construct an *unbiased* estimate for the mean as follows:

$$\mathbb{E}_{\boldsymbol{y}} \ln f(y;\widehat{\theta}_k) \approx \frac{1}{N}\sum_{n=1}^{N} \ln f(y_n;\widehat{\theta}_k) \;-\; \frac{M_k}{N} \tag{31.240}$$

The first term on the right-hand side is a sample mean approximation; however, for this problem, it leads to a *biased* estimate for $\mathbb{E}_{\boldsymbol{y}} \ln f(y;\widehat{\theta}_k)$. The reason for the bias is because the measurements are used twice: once as an argument of $f(y_n;\widehat{\theta}_k)$ and the other in the computation of $\widehat{\theta}_k$. The bias can be "removed" by subtracting M_k/N. This result needs proof as we proceed to explain. Once established, we would then arrive at the AIC formulation (31.154), namely, (where we are scaling (31.240) by multiplying by $2N$):

$$k^\star \;=\; \underset{1\le k\le K}{\operatorname{argmin}} \left\{ 2M_k - 2\sum_{n=1}^{N} \ln f(y_n;\widehat{\theta}_k) \right\} \tag{31.241}$$

We thus need to establish that the expression on the right-hand side of (31.240) provides an "unbiased" estimate for $\mathbb{E}_{\boldsymbol{y}} \ln f(\boldsymbol{y};\widehat{\theta}_k)$. That is, we need to verify that

$$\mathbb{E}_{\boldsymbol{y}} \left\{ \frac{1}{N}\sum_{n=1}^{N} \ln f(y_n;\widehat{\theta}_k) \;-\; \frac{M_k}{N} \right\} \approx \mathbb{E}_{\boldsymbol{y}} \ln f(y;\widehat{\theta}_k) \tag{31.242}$$

As the argument will show, the derivation is based on a couple of approximations and the result is valid for large sample sizes, N.

Proof of (31.242): Motivated by (31.239), we introduce first the following notation for the mean log-likelihood function that we are trying to approximate from the data:

$$L(\theta) \;\stackrel{\Delta}{=}\; \int_y f(y)\ln f(y;\theta)dy \;=\; \mathbb{E}_{\boldsymbol{y}} \ln f(\boldsymbol{y};\theta) \tag{31.243}$$

Its maximizer is denoted by $\theta^\star$. We perform a second-order Taylor series expansion for $L(\widehat{\theta}_k)$ around $\theta^\star$ to find

$$L(\widehat{\theta}_k) = L(\theta^\star) + \frac{1}{2}(\theta^\star - \widehat{\theta}_k)^{\sf T} J_k(\theta^\star - \widehat{\theta}_k) \;+\; o(\|\theta^\star - \widehat{\theta}_k\|^2) \tag{31.244}$$

in terms of the Hessian matrix:

$$\begin{aligned} J_k &\overset{\Delta}{=} \nabla_\theta^2\, L(\theta)\Big|_{\theta=\theta^\star} \\ &= \nabla_\theta^2\, \mathbb{E}\ \ln f(y;\theta)\Big|_{\theta=\theta^\star} \\ &= \mathbb{E}\,\nabla_\theta^2\ \ln f(y;\theta)\Big|_{\theta=\theta^\star} \\ &\overset{(31.99)}{=} -F(\theta^\star) \end{aligned} \tag{31.245}$$

where $F(\theta^\star)$ is the Fisher information matrix at $\theta^\star$. In the third equality we assumed that we can exchange the differentiation and expectation operations; this is usually justified by a result known as the *dominated convergence theorem* in analysis; we commented on this topic in some detail in Appendix 16.A. Ignoring higher-order error terms, we conclude that

$$\boxed{L(\widehat{\theta}_k) \approx L(\theta^\star) - \frac{1}{2}(\theta^\star - \widehat{\theta}_k)^\mathsf{T} F(\theta^\star)(\theta^\star - \widehat{\theta}_k)} \tag{31.246}$$

In a similar vein, consider the empirical log-likelihood function used in the approximation (31.240), namely,

$$\ell(\theta) = \sum_{n=1}^{N} \ln f(y_n;\theta) \tag{31.247}$$

and let us perform a second-order Taylor series expansion around $\theta^\star$:

$$\ell(\widehat{\theta}_k) \approx \ell(\theta^\star) + (\widehat{\theta}_k - \theta^\star)^\mathsf{T} S_k + \frac{1}{2}(\theta^\star - \widehat{\theta}_k)^\mathsf{T} H_k(\theta^\star - \widehat{\theta}_k) \tag{31.248}$$

in terms of the gradient vector and Hessian matrix quantities:

$$S_k \overset{\Delta}{=} \nabla_{\theta^\mathsf{T}}\ell(\theta)\Big|_{\theta=\theta^\star} = \sum_{n=1}^{N} \nabla_{\theta^\mathsf{T}} \ln f(y_n;\theta)\Big|_{\theta=\theta^\star} \tag{31.249a}$$

$$H_k \overset{\Delta}{=} \nabla_\theta^2\,\ell(\theta)\Big|_{\theta=\theta^\star} = \sum_{n=1}^{N} \nabla_\theta^2 \ln f(y_n;\theta)\Big|_{\theta=\theta^\star} \tag{31.249b}$$

We rework the expressions for S_k and H_k by noting that the second-order Taylor series approximations are valid for $\widehat{\theta}_k$ close to $\theta^\star$ (which in turn requires large N):

$$\begin{aligned}
S_k &= \sum_{n=1}^{N} \nabla_{\theta^\mathsf{T}} \ln f(y_n;\theta^\star), \quad \text{(by definition)}\\
&\overset{(a)}{=} \sum_{n=1}^{N} \nabla_{\theta^\mathsf{T}} \ln f(y_n;\theta^\star) \;-\; \underbrace{\sum_{n=1}^{N} \nabla_{\theta^\mathsf{T}} \ln f(y_n;\widehat{\theta}_k)}_{=0}\\
&= \sum_{n=1}^{N} \left\{ \nabla_{\theta^\mathsf{T}} \ln f(y_n;\theta^\star) \;-\; \nabla_{\theta^\mathsf{T}} \ln f(y_n;\widehat{\theta}_k) \right\}\\
&\overset{(b)}{=} \sum_{n=1}^{N} \left(\int_0^1 \nabla_\theta^2 \, \ln f(y_n;\widehat{\theta}_k + t(\theta^\star - \widehat{\theta}_k)dt \right) (\theta^\star - \widehat{\theta}_k)\\
&\overset{(c)}{\approx} \sum_{n=1}^{N} \nabla_\theta^2 \, \ln f(y_n;\theta^\star)(\theta^\star - \widehat{\theta}_k)\\
&= N\left(\frac{1}{N} \sum_{n=1}^{N} \nabla_\theta^2 \, \ln f(y_n;\theta^\star) \right)(\theta^\star - \widehat{\theta}_k)\\
&\approx -NF(\theta^\star)(\theta^\star - \widehat{\theta}_k), \quad N \text{ large enough}
\end{aligned} \tag{31.250}$$

where step (a) is because $\widehat{\theta}_k$ is the ML estimate that solves (31.236), step (b) uses the mean-value theorem (10.5), and step (c) is because $\theta^\star$ and $\widehat{\theta}_k$ are close to each other. In the last step we used the sample average as an approximation for the Fisher information matrix. Similarly, for the Hessian matrix we can verify that

$$\frac{1}{N} H_k \;\overset{\Delta}{=}\; \frac{1}{N} \sum_{n=1}^{N} \nabla_\theta^2 \ln f(y_n;\theta^\star) \;\approx\; -F(\theta^\star) \tag{31.251}$$

We can also write by the law of large numbers:

$$\begin{aligned}
\ell(\theta^\star) &= \sum_{n=1}^{N} \ln f(y_n;\theta^\star), \quad \text{(by definition)}\\
&= N\left(\frac{1}{N} \sum_{n=1}^{N} \ln f(y_n;\theta^\star) \right)\\
&\approx NL(\theta^\star), \quad \text{for } N \text{ large enough}
\end{aligned} \tag{31.252}$$

Substituting into (31.248) we find that

$$\boxed{\frac{1}{N}\ell(\widehat{\theta}_k) \approx L(\theta^\star) + \frac{1}{2}(\theta^\star - \widehat{\theta}_k)^\mathsf{T} F(\theta^\star)(\theta^\star - \widehat{\theta}_k)} \tag{31.253}$$

According to (31.242) we need to verify that

$$\mathbb{E}_{\boldsymbol{y}}\left\{ \frac{1}{N}\ell(\widehat{\theta}_k) \;-\; \frac{M_k}{N} \right\} \overset{?}{\approx} L(\widehat{\theta}_k) \tag{31.254}$$

or, equivalently,

$$\mathbb{E}_{\boldsymbol{y}}\left\{ \frac{1}{N}\ell(\widehat{\theta}_k) - L(\widehat{\theta}_k) \right\} \overset{?}{\approx} M_k/N \tag{31.255}$$

To establish that the relation is valid, we first note using (31.246) and (31.253) that the difference on the left-hand side is given

$$\begin{aligned}\mathbb{E}_{\boldsymbol{y}}\left\{\frac{1}{N}\ell(\widehat{\theta}_k)-L(\widehat{\theta}_k)\right\} &\approx \mathbb{E}\left\{(\theta^\star-\widehat{\theta}_k)^{\mathsf{T}}F(\theta^\star)(\theta^\star-\widehat{\theta}_k)\right\} \qquad (31.256)\\ &= \frac{1}{N}\mathbb{E}\left\{\sqrt{N}(\theta^\star-\widehat{\theta}_k)^{\mathsf{T}}F(\theta^\star)(\theta^\star-\widehat{\theta}_k)\sqrt{N}\right\}\end{aligned}$$

To facilitate the derivation from this point, we assume initially that $\theta^\star$ is close to the true (unknown) model that generated the data. For all practical purposes, this assumption amounts to the "strong" requirement that the true model is included in the set of models parameterized by the $\{\theta_k\}$; we will remove this condition afterwards. Under this assumption, we can appeal to property (31.129) to note that approximately:

$$\sqrt{N}(\theta^\star-\widehat{\boldsymbol{\theta}}_k) \xrightarrow{\text{d}} \mathcal{N}\Big(0,\, F^{-1}(\theta^\star)\Big) \qquad (31.257)$$

Next, introduce the eigendecomposition of $F(\theta^\star)$, say,

$$F(\theta^\star)=U\Lambda U^{\mathsf{T}} \qquad (31.258)$$

where U is orthogonal and Λ is $M_k\times M_k$ diagonal with positive entries $\{\lambda_j\}$. Then, transformation by $\Lambda^{1/2}U^{\mathsf{T}}$ does not destroy Gaussianity:

$$z \;\stackrel{\Delta}{=}\; \sqrt{N}\Lambda^{1/2}U^{\mathsf{T}}(\theta^\star-\widehat{\boldsymbol{\theta}}_k) \xrightarrow{\text{d}} \mathcal{N}(0,\, I) \qquad (31.259)$$

where $\Lambda^{1/2}$ denotes a diagonal matrix with the positive square-roots of the $\{\lambda_j\}$. It follows that

$$\mathbb{E}_{\boldsymbol{y}}\left\{\frac{1}{N}\ell(\widehat{\theta}_k)-L(\widehat{\theta}_k)\right\} \approx \frac{1}{N}\mathbb{E}\left\{\sum_{j=1}^{M_k}\boldsymbol{z}_j^2\right\} \qquad (31.260)$$

where the term on the right-hand side involves the sum of M_k independent Gaussian-distributed random variables with zero mean and unit variance each. It is well-known that such a sum is Chi-square distributed with M_k degrees of freedom – recall Prob. 4.3:

$$\sum_{j=1}^{M_k}\boldsymbol{z}_j^2 \;\sim\; \chi^2(M_k) \qquad (31.261)$$

Since the mean of a Chi-square distributed random variable is equal to its degree, we conclude that

$$\mathbb{E}_{\boldsymbol{y}}\left\{\frac{1}{N}\ell(\widehat{\theta}_k)-L(\widehat{\theta}_k)\right\} \approx M_k/N \qquad (31.262)$$

as claimed.

The previous argument assumed that the true unknown model is included in the set of models parameterized by the $\{\theta_k\}$; this assumption enabled the use of property (31.257) where the covariance matrix of the Gaussian distribution is given by $F^{-1}(\theta^\star)$. Consider now the more general case where the true model that generated the data need not be included in the set of candidate models. Let us denote this true model by θ^o. We verify that the covariance matrix in (31.257) will need to be adjusted (while the remainder of the argument will continue to be the same). To see this, introduce the score variable:

$$\boldsymbol{\mathcal{S}}(\theta) \;\stackrel{\Delta}{=}\; \nabla_{\theta^{\mathsf{T}}}\ln f(\boldsymbol{y},\theta) \qquad (31.263)$$

We know that its mean is zero and its covariance matrix at $\theta=\theta^o$ is the Fisher information matrix, $F(\theta^o)$; recall (31.97). The covariance matrix will be different at other values for θ. In particular, let V denote the covariance matrix at $\theta = \theta^\star$:

$$V(\theta^\star) \;\overset{\Delta}{=}\; \mathbb{E}\,\mathbf{\mathcal{S}}(\theta^\star)\mathbf{\mathcal{S}}^{\mathsf{T}}(\theta^\star) \tag{31.264}$$

If we examine the definition for S_k from the first line of (31.250), we find that it is the sum of N independent realizations for $\mathbf{\mathcal{S}}(\theta^\star)$, each with zero mean and covariance matrix V. Therefore, from the central limit theorem, we have

$$\sqrt{N}\left(\frac{1}{N}\sum_{n=1}^{N}\nabla_{\theta^{\mathsf{T}}}\ln f(y_n;\theta^*)\right) \;\xrightarrow{d}\; \mathcal{N}(0, V(\theta^\star)) \tag{31.265}$$

It follows from the last line of (31.250) that — compare with (31.257):

$$\sqrt{N}(\theta^\star - \widehat{\boldsymbol{\theta}}_k) \;\xrightarrow{\text{d}}\; \mathcal{N}\Big(0, F^{-1}(\theta^\star)V(\theta^\star)F^{-1}(\theta^\star)\Big) \tag{31.266}$$

The argument continues in the same manner from here by introducing the eigendecomposition:

$$F^{-1}(\theta^\star)V(\theta^\star)F^{-1}(\theta^\star) \;=\; U\Lambda U^{\mathsf{T}} \tag{31.267}$$

to arrive at the same expression (31.262).

31.C DERIVATION OF THE BIC FORMULATION

In this appendix we motivate the cost function that is optimized in (31.160) by the BIC criterion by following the argument from Neath and Cavanaugh (2012) and Ghosh, Delampady, and Samanta (2006) adjusted to our notation and conventions. The objective is to devise a criterion that selects the "best" fit from a collection of K models $\{\theta_1, \ldots, \theta_K\}$. Each model θ_k has M_k parameters.

So far in the chapter, we have modeled the θ_k as unknown (deterministic) parameters that we wish to estimate. The Bayesian information criterion, however, follows an alternative paradigm (which we encountered in Chapter 28 under Bayesian inference). Specifically, as befits a Bayesian formulation, each variable θ_k will be modeled as as a random variable in its own right (rather than as a deterministic variable that is unknown). Given a class k, the Bayesian formulation assumes a distribution $f_{\boldsymbol{\theta}_k|k}(\theta_k|k)$ for the variable $\boldsymbol{\theta}_k$. This essentially amounts to giving more or less weights to different values for θ_k, with some values being preferred over other values as dictated by the assumed pdf.

We assign a *prior* probability mass distribution to the model indexes $\{k\}$ as well, denoted by:

$$\pi(k) \;\overset{\Delta}{=}\; \mathbb{P}(\boldsymbol{k} = k), \quad \sum_{k=1}^{K}\pi(k) = 1 \tag{31.268}$$

This pmf is not going to influence the final solution; it can be thought of as assigning more or less relevance to some models based on prior knowledge or experience. The prior $\pi(k)$ could, for example, be chosen as the uniform distribution, $\pi(k) = 1/K$, in which case all models are treated equally. We will compute the *posterior* distribution given the observations, namely, $\pi(k|y_1, y_2, \ldots, y_N)$, and then show that selecting the k that maximizes this posterior leads to the BIC formulation.

For simplicity of notation in this appendix, we will write $f(z)$ instead of $f_{\boldsymbol{z}}(z)$ to refer to the pdf of a random variable $\boldsymbol{z}$ without the subscript. As such, we will write $f(\theta_k|k)$ without the subscripts, which are understood from the arguments. We are already using this notation in the distribution $\pi(\cdot)$, and will be using it below for joint and conditional distributions involving the variables $\{\boldsymbol{k}, \boldsymbol{y}_1, \ldots, \boldsymbol{y}_N\}$. From Bayes rule we have:

$$\begin{aligned}
\pi(k|y_1, y_2, \ldots, y_N) &= \frac{f(k, y_1, y_2, \ldots, y_N)}{f(y_1, y_2, \ldots, y_N)} \\
&\overset{(a)}{\propto} f(k, y_1, y_2, \ldots, y_N) \\
&= \pi(k)\, f(y_1, y_2, \ldots, y_N|k) \qquad (31.269)
\end{aligned}$$

where step (a) is because the denominator in the first equality is independent of k. To evaluate the last conditional pdf we marginalize over θ_k and apply Bayes rule again:

$$\begin{aligned}
f(y_1, y_2, \ldots, y_N|k) &= \int_{\theta_k \in \Theta_k} f(y_1, y_2, \ldots, y_N, \theta_k|k) d\theta_k \\
&= \int_{\theta_k \in \Theta_k} f(y_1, y_2, \ldots, y_N|\theta_k, k)\, f(\theta_k|k) d\theta_k \qquad (31.270)
\end{aligned}$$

where are using $\theta_k \in \Theta_k$ to denote the domain of θ_k. Given k and θ_k, the first term under integration is related to the log-likelihood function over the data, which we denote by

$$\ell(\theta_k) \;\stackrel{\Delta}{=}\; \ln f(y_1, y_2, \ldots, y_N|\theta_k, k) \;=\; \sum_{n=1}^{N} \ln f(y_n|\theta_k) \qquad (31.271)$$

Now, let $\widehat{\theta}_k$ be the ML estimate obtained by maximizing this function:

$$\widehat{\theta}_k \;=\; \underset{\theta_k}{\text{argmax}}\; \ell(\theta_k) \;\Longrightarrow\; \nabla_{\theta_k^{\mathsf{T}}} \ell(\theta_k)\Big|_{\theta_k = \widehat{\theta}_k} = 0 \qquad (31.272)$$

We perform a second-order Taylor series expansion of $\ell(\theta_k)$ around $\widehat{\theta}_k$ to find

$$\ell(\theta_k) = \ell(\widehat{\theta}_k) + \frac{1}{2}(\theta_k - \widehat{\theta}_k)^{\mathsf{T}} H_k (\theta_k - \widehat{\theta}_k) \;+\; o(\|\theta_k - \widehat{\theta}_k\|^2) \qquad (31.273)$$

in terms of the Hessian matrix

$$H_k \;\stackrel{\Delta}{=}\; \nabla_\theta^2\, \ell(\theta_k)\Big|_{\theta_k = \widehat{\theta}_k} \;=\; \sum_{n=1}^{N} \nabla_{\theta_k}^2 \ln f(y_n|\theta_k)\Big|_{\theta_k = \widehat{\theta}_k} \qquad (31.274)$$

Comparing with (31.97) and (31.99) we find that for large enough N, the above Hessian matrix is well approximated by the Fisher information matrix, namely,

$$\frac{1}{N} H_k \;\approx\; -F(\widehat{\theta}_k) \qquad (31.275)$$

Ignoring higher-order terms and substituting (31.273) into (31.271) we determine the following expression for the likelihood function:

$$f(y_1, y_2, \ldots, y_N|\theta_k, k) \approx e^{\ell(\widehat{\theta}_k)} \times \exp\left\{ -\frac{N}{2}(\theta_k - \widehat{\theta}_k)^{\mathsf{T}} F(\widehat{\theta}_k)(\theta_k - \widehat{\theta}_k) \right\} \qquad (31.276)$$

Substituting further into (31.270) we obtain:

$$
\begin{aligned}
&f(y_1, y_2, \ldots, y_N|k) \\
&\approx e^{\ell(\widehat{\theta}_k)} \times \int_{\theta_k \in \Theta_k} \exp\Big\{-\frac{1}{2}(\theta_k - \widehat{\theta}_k)^{\mathsf{T}}\Big[\frac{1}{N}F^{-1}(\widehat{\theta}_k)\Big]^{-1}(\theta_k - \widehat{\theta}_k)\Big\}\, f(\theta_k|k)d\theta_k \\
&\approx \Big(\frac{2\pi}{N}\Big)^{M_k/2} \times \Big(\det F(\widehat{\theta}_k)\Big)^{-1/2} \times e^{\ell(\widehat{\theta}_k)} \times \int_{\theta_k \in \Theta_k} \mathcal{N}_{\boldsymbol{\theta}_k}\Big(\widehat{\theta}_k, \frac{1}{N}F^{-1}(\widehat{\theta}_k)\Big)\, f(\theta_k|k)d\theta_k
\end{aligned}
\tag{31.277}
$$

where in the last equality we scaled the expression inside the integral to transform it into a Gaussian distribution. The above expression is valid for large values of N and for values of θ_k close to $\widehat{\theta}_k$. At this stage it is customary in the literature to appeal to approximate arguments in order to evaluate the integral expression. In one case, it is assumed that $f(\theta_k|k) = 1$ in the vicinity of $\widehat{\theta}_k$ (which corresponds to an uninformative or *flat* prior) so that the integral evaluates to the mean of the Gaussian distribution, which is $\widehat{\theta}_k$. In a second case, one notes that the variance of the Gaussian distribution shrinks as $N \to \infty$ and uses this observation to approximate the integral by $f(\widehat{\theta}_k|k)$. Either approximation leads to the same final result. We continue with this second approximation. Substituting into (31.269) we obtain (where we are further multiplying both sides of the equality by 2 for convenience to arrive at the same form as the BIC):

$$
\begin{aligned}
2\ln \pi(k|y_1, y_2, \ldots, y_N) = &- M_k \ln N \; + \; 2\sum_{n=1}^{N} \ln f(y_n; \widehat{\theta}_k) \\
&+2\ln \pi(k) \; - \; \ln\det F(\widehat{\theta}_k) + \ln f(\widehat{\theta}_k|k) + \text{cte}
\end{aligned}
\tag{31.278}
$$

The terms on the second line remain bounded as $N \to \infty$; we can ignore them in the maximization of $\ln \pi(k|y_1, y_2, \ldots, y_N)$, which leads to the BIC formulation (31.160).

REFERENCES

Abramowitz, M. and I. Stegun (1965), *Handbook of Mathematical Functions*, Dover, NY.

Akaike, H. (1974), “A new look at the statistical model identification,” *IEEE Trans. Automatic Control*, vol. 19, no. 6, pp. 716–723.

Aldrich, J. (1997), “R. A. Fisher and the making of maximum likelihood 1912–1922,” *Statistical Science*, vol. 12, no. 3, pp. 162–176.

Artin, E. (2015), *The Gamma Function*, Dover, NY.

Barnett, V. (1999), *Comparative Statistical Inference*, Wiley, UK.

Barron, A. and T. Cover (1991), “Minimum complexity density estimation,” *IEEE Trans. Information Theory*, vol. 37, no.4, pp. 1034–1054.

Barron A., J. Rissanen, and B. Yu (1998), “The minimum description length principle in coding and modeling,” *IEEE Trans. Information Theory*, vol. 44, pp. 2734–2760.

Blackwell, D. (1947), “Conditional expectation and unbiased sequential estimation,” *Annals Math. Statis.*, vol. 18, no. 1, pp.105–110.

Bowman, K. O. and L. R. Shenton (2007), “The beta distribution, moment method, Karl Pearson and R.A. Fisher,” *Far East J. Theo. Stat.*, vol. 23, no. 2, pp. 133–164.

Box, G. E. P. and G. C. Tiao (1973), *Bayesian Inference in Statistical Analysis*, Addison-Wesley, Reading, MA.

Burnham, K. P. and D. R. Anderson (2002), *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*, 2nd edition, Springer-Verlag, NY

Caines, P. E. (1988), *Linear Stochastic Systems*, Wiley, NY.

Capen, E., R. Clapp, and T. Campbell (1971), "Bidding in high risk situations," *Journal of Petroleum Technology*, vol. 23, pp. 641–653.

Cassella, G. and R. L. Berger (2002), *Statistical Inference*, Duxbury, CA.

Cavanaugh, J. E. (1997), "Unifying the derivations of the Akaike and corrected Akaike information criteria," *Statistics and Probability Letters*, vol. 33, pp. 201–208.

Cavanaugh, J. E. and A. A. Neath (2019), "The Akaike information criterion: Background, derivation, properties, application, interpretation, and refinements," *WIRE Comput. Stat.*, pp. 1–11.

Claeskens, G. and N. L. Hjort (2008), *Model Selection and Model Averaging*, Cambridge University Press.

Cox, D. R. (2006), *Principles of Statistical Inference*, Cambridge University Press.

Cramer, H. (1946), *Mathematical Methods of Statistics*, Princeton University Press, NJ.

Davis, P. J. (1959), "Leonhard Euler's integral: A historical profile of the Gamma function," *American Mathematical Monthly*, vol. 66, no. 10, pp. 849–869.

de Finetti, B. (1974), *Theory of Probability*, Wiley, NY.

Detrano, R., A. Janosi, W. Steinbrunn, M. Pfisterer, J. Schmid, S. Sandhu, K. Guppy, S. Lee, and V. Froelicher (1989), "International application of a new probability algorithm for the diagnosis of coronary artery disease," *American J. Cardiology*, vol. 64, pp. 304–310.

Eddy, S. (2004), "What is Bayesian statistics?" *Nature Biotechnology*, vol. 22, pp. 1177–1178.

Edgeworth, F. Y. (1908a), "On the probable errors of frequency-constants," *J. Royal Statist. Society*, vol. 71, no. 2, pp. 381–397.

Edgeworth, F. Y. (1908b), "On the probable errors of frequency-constants (continued)," *J. Royal Statist. Society*, vol. 71, no. 3, pp. 499–512.

Edgeworth, F. Y. (1908c), "On the probable errors of frequency-constants (continued)," *J. Royal Statist. Society*, vol. 71, no. 4, pp. 651–678.

Fisher, R. A. (1912), "On an absolute criterion for fitting frequency curves," *Messeg. Math.*, vol. 41, pp. 155–160.

Fisher, R. A. (1922), "On the mathematical foundations of theoretical statistics," *Philos. Trans. Roy. Soc. London Ser. A.*, vol. 222, pp. 309–368.

Fisher, R. A. (1925), "Theory of statistical estimation," *Proc. Cambridge Philos. Soc.*, vol. 22, pp. 700–725.

Fisher, R. A. (1956), *Statistical Methods and Scientific Inference*, Oliver and Boyd, Edinburgh.

Frieden, B. R. (2004), *Science from Fisher Information: A Unification*, Cambridge University Press.

Ghosh, J. K., M. Delampady, and T. Samanta (2006), *An Introduction to Bayesian Analysis: Theory and Methods*, Springer-Verlag, NY.

Grunwald, P. (2007), *The Minimum Description Length Principle*, MIT Press, MA.

Hammersley, J. M. (1950), "On estimating restricted parameters," *J. Royal Stat. Soc.,* Series B, vol. 12, no. 2, pp. 192–240.

Hansen, M. and B. Yu (2001), "Model selection and the principle of minimum description length," *J. American Stat. Assoc.*, vol. 96, no. 454, pp. 746–774.

Hastie, T., R. Tibshirani, and J. Friedman (2009), *The Elements of Statistical Learning*, 2nd edition, Springer, NY.

Hogg, R. V. and J. McKean (2012), *Introduction to Mathematical Statistics*, 7th edition, Pearson.

Hurvich, C. M. and C. L. Tsai (1989), "Regression and time series model selection in small samples," *Biometrika*, vol. 76, no. 2, pp. 297–307.

Kay, S. (1993), *Fundamentals of Statistical Signal Processing: Estimation Theory*, Prentice Hall, NJ.

Kay, S. (1998), *Fundamentals of Statistical Signal Processing: Detection Theory*, Prentice Hall, NJ.

Konishi, S. and G. Kitagawa (2008), *Information Criteria and Statistical Modeling*, Springer.

Lebedev, N. N. (1972), *Special Functions and their Applications*, Dover, NY. Translated by R. A. Silverman.

Lehmann, E. L. (1998), *Elements of Large-Sample Theory*, Springer, NY.

Linhart, H. and W. Zucchini (1986), *Model Selection*, Wiley, NY.

McQuarrie A. D. R. and C. L. Tsai (1998), *Regression and Time Series Model Selection*, World Scientific, NJ.

Misak, C. (2020), *Frank Ramsey: A Sheer Excess of Powers*, Oxford University Press.

Neath, A. A. and J. E. Cavanaugh (2012), "The Bayesian information criterion: Background, derivation, and applications," *WIREs Comput. Stat.*, vol. 41, pp. 199-203.

Pearson, K. (1936), "Method of moments and method of maximum likelihood," *Biometrika*, vol. 28, nos. 1/2, p. 34.

Pratt, J. W. (1976), "F. Y. Edgeworth and R. A. Fisher on the efficiency of maximum likelihood estimation," *Annals of Statistics*, vol. 4, no. 3, pp. 501–514.

Ramsey, F. P. (1931), "Truth and Probability," appears reprinted in *Philosophical Papers*, D. H. Mellor, *Editor*, Cambridge University Press, 1990, pp. 52–94. Original 1931 article published posthumously.

Rao, C. R. (1945), "Information and accuracy attainable in the estimation of statistical parameters," *Bullet. Calcutta Math. Soc.*, vol. 37, no. 3, pp. 81–91.

Rissanen, J. (1978), "Modeling by shortest data description," *Automatica*, vol. 14, no. 5, pp. 465–471.

Rissanen, J. (1986), "Stochastic complexity and modeling," *Ann. Stat.*, vol. 14, no. 3, pp. 1080–1100.

Sahlin, N.-E. (2008), *The Philosophy of F. P. Ramsey*, reprint edition, Cambridge University Press.

Samaniego, F. J. (2010), *A Comparison of the Bayesian and Frequentist Approaches to Estimation*, Springer, NY.

Samaniego, F. J. and D. M. Reneau (1994), "Toward a reconciliation of the Bayesian and frequentist approaches to point estimation," *J. Amer. Stat. Assoc.*, vol. 89, no. 427, pp. 947–957.

Savage, L. J. (1954), *The Foundations of Statistics*, Wiley, NY.

Savage, L. J. (1976), "On rereading R. A. Fisher," *Ann. Statist.*, vol. 4, pp. 441–500.

Scharf, L. L. (1991), *Statistical Signal Processing: Detection, Estimation, and Time-Series Analysis*, Addison-Wesley, Reading, MA.

Schwarz, G. E. (1978), "Estimating the dimension of a model," *Annals of Statistics*, vol. 6, no. 2, pp. 461–464.

Smith, J. E. and R. L. Winkler (2006), "The optimizer's curse: Skepticism and post-decision surprise in decision analysis," *Management Science*, vol. 52, no. 3, pp. 311–322.

Temme, N. M. (1996), *Special Functions: An Introduction to Classical Functions of Mathematical Physics*, Wiley, NY.

Thaler, R. H. (1988), "Anomalies: The winner's curse," *Journal of Economic Perspectives*, vol. 2, no. 1, pp. 191–202.

Van den Steen, E. (2004), "Rational overoptimism (and other biases)," *American Economic Review,*, vol. 94, no. 4, pp. 1141–1151.

van Hasselt, H. (2010), "Double $Q-$learning," *Proc. Advances in Neural Processing Systems* (NIPS), pp. 1–9, Vancouver, Canada.

VanderPlas, J. (2014), "Frequentism and Bayesianism: A Python-driven primer," *Proc. 13th Python in Science Conf.* (SCIPY), pp. 85–93, Austin, TX. Also available online at https://arxiv.org/pdf/1411.5018.pdf

Van Trees, H. L. (1968), *Detection, Estimation, and Modulation Theory*, Wiley, NY.

Van Trees, H. L. (2013), *Detection, Estimation, and Modulation Theory*, Part I, 2nd edition, Wiley, NY.

Wakefield, J. (2013), *Bayesian and Frequentist Regression Methods*, Springer.

Zacks, S. (1971), *The Theory of Statistical Inference*, Wiley, NY.