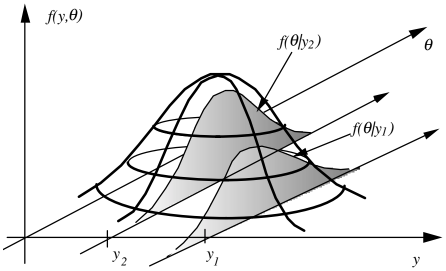
Advanced Digital Signal Processing and Noise Reduction, Second Edition.
Saeed V. Vaseghi
Copyright © 2000 John Wiley & Sons Ltd
ISBNs: 0-471-62692-9 (Hardback): 0-470-84162-1 (Electronic)



4

BAYESIAN ESTIMATION

Bayesian estimation is a framework for the formulation of statistical inference problems. In the prediction or estimation of a random process from a related observation signal, the Bayesian philosophy is based on combining the evidence contained in the signal with prior knowledge of the probability distribution of the process. Bayesian methodology includes the classical estimators such as maximum a posteriori (MAP), maximum-likelihood (ML), minimum mean square error (MMSE) and minimum mean absolute value of error (MAVE) as special cases. The hidden Markov model, widely used in statistical signal processing, is an example of a Bayesian model. Bayesian inference is based on minimisation of the so-called Bayes' risk function, which includes a posterior model of the unknown parameters given the observation and a cost-of-error function. This chapter begins with an introduction to the basic concepts of estimation theory, and considers the statistical measures that are used to quantify the performance of an estimator. We study Bayesian estimation methods and consider the effect of using a prior model on the mean and the variance of an estimate. The estimate–maximise (EM) method for the estimation of a set of unknown parameters from an incomplete observation is studied, and applied to the mixture Gaussian modelling of the space of a continuous random variable. This chapter concludes with an introduction to the Bayesian classification of discrete or finite-state signals, and the K-means clustering method.

4.1 Bayesian Estimation Theory: Basic Definitions

Estimation theory is concerned with the determination of the best estimate of an unknown parameter vector from an observation signal, or the recovery of a clean signal degraded by noise and distortion. For example, given a noisy sine wave, we may be interested in estimating its basic parameters (i.e. amplitude, frequency and phase), or we may wish to recover the signal itself. An estimator takes as the input a set of noisy or incomplete observations, and, using a dynamic model (e.g. a linear predictive model) and/or a probabilistic model (e.g. Gaussian model) of the process, estimates the unknown parameters. The estimation accuracy depends on the available information and on the efficiency of the estimator. In this chapter, the Bayesian estimation of continuous-valued parameters is studied. The modelling and classification of finite-state parameters is covered in the next chapter.

Bayesian theory is a general inference framework. In the estimation or prediction of the state of a process, the Bayesian method employs both the evidence contained in the observation signal and the accumulated prior probability of the process. Consider the estimation of the value of a random parameter vector $\boldsymbol{\theta}$, given a related observation vector $\boldsymbol{y}$. From Bayes' rule the posterior probability density function (pdf) of the parameter vector $\boldsymbol{\theta}$ given $\boldsymbol{y}$, $f_{\Theta|Y}(\boldsymbol{\theta}\,|\,\boldsymbol{y})$, can be expressed as

$$f_{\Theta|Y}(\boldsymbol{\theta}\,|\,\boldsymbol{y}) = \frac{f_{Y|\Theta}(\boldsymbol{y}\,|\,\boldsymbol{\theta})f_{\Theta}(\boldsymbol{\theta})}{f_{Y}(\boldsymbol{y})} \tag{4.1}$$

where for a given observation, $f_Y(\boldsymbol{y})$ is a constant and has only a normalising effect. Thus there are two variable terms in Equation (4.1): one term $f_{Y|\Theta}(\boldsymbol{y}|\boldsymbol{\theta})$ is the likelihood that the observation signal $\boldsymbol{y}$ was generated by the parameter vector $\boldsymbol{\theta}$ and the second term is the prior probability of the parameter vector having a value of $\boldsymbol{\theta}$. The relative influence of the likelihood pdf $f_{Y|\Theta}(\boldsymbol{y}|\boldsymbol{\theta})$ and the prior pdf $f_{\Theta}(\boldsymbol{\theta})$ on the posterior pdf $f_{\Theta|Y}(\boldsymbol{\theta}|\boldsymbol{y})$ depends on the shape of these function, i.e. on how relatively peaked each pdf is. In general the more peaked a probability density function, the more it will influence the outcome of the estimation process. Conversely, a uniform pdf will have no influence.

The remainder of this chapter is concerned with different forms of Bayesian estimation and its applications. First, in this section, some basic concepts of estimation theory are introduced.

4.1.1 Dynamic and Probability Models in Estimation

Optimal estimation algorithms utilise dynamic and statistical models of the observation signals. A dynamic predictive model captures the correlation structure of a signal, and models the dependence of the present and future values of the signal on its past trajectory and the input stimulus. A statistical probability model characterises the random fluctuations of a signal in terms of its statistics, such as the mean and the covariance, and most completely in terms of a probability model. Conditional probability models, in addition to modelling the random fluctuations of a signal, can also model the dependence of the signal on its past values or on some other related process.

As an illustration consider the estimation of a P-dimensional parameter vector $\boldsymbol{\theta}=[\theta_0, \theta_1, ..., \theta_{P-1}]$ from a noisy observation vector $\boldsymbol{y}=[y(0), y(1), ..., y(N-1)]$ modelled as

$$\boldsymbol{y}=h(\boldsymbol{\theta},\boldsymbol{x},\boldsymbol{e})+\boldsymbol{n} \tag{4.2}$$

where, as illustrated in Figure 4.1, the function $h(\cdot)$ with a random input $\boldsymbol{e}$, output $\boldsymbol{x}$, and parameter vector $\boldsymbol{\theta}$, is a predictive model of the signal $\boldsymbol{x}$, and $\boldsymbol{n}$ is an additive random noise process. In Figure 4.1, the distributions of the random noise $\boldsymbol{n}$, the random input $\boldsymbol{e}$ and the parameter vector $\boldsymbol{\theta}$ are modelled by probability density functions, $f_N(\boldsymbol{n})$, $f_E(\boldsymbol{e})$, and $f_\Theta(\boldsymbol{\theta})$ respectively. The pdf model most often used is the Gaussian model. Predictive and statistical models of a process *guide* the estimator towards the set of values of the unknown parameters that are most consistent with both the prior distribution of the model parameters and the noisy observation. In general, the more modelling information used in an estimation process, the better the results, provided that the models are an accurate characterisation of the observation and the parameter process.

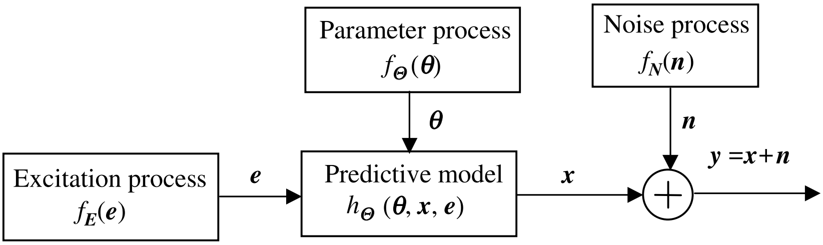


Figure 4.1 A random process $\boldsymbol{y}$ is described in terms of a predictive model $h(\cdot)$, and statistical models $f_E(\cdot)$, $f_\Theta(\cdot)$ and $f_N(\cdot)$.

4.1.2 Parameter Space and Signal Space

Consider a random process with a parameter vector $\boldsymbol{\theta}$. For example, each instance of $\boldsymbol{\theta}$ could be the parameter vector for a dynamic model of a speech sound or a musical note. The parameter space of a process $\boldsymbol{\Theta}$ is the collection of all the values that the parameter vector $\boldsymbol{\theta}$ can assume. The parameters of a random process determine the "character" (i.e. the mean, the variance, the power spectrum, etc.) of the signals generated by the process. As the process parameters change, so do the characteristics of the signals generated by the process. Each value of the parameter vector $\boldsymbol{\theta}$ of a process has an associated signal space $\boldsymbol{Y}$; this is the collection of all the signal realisations of the process with the parameter value $\boldsymbol{\theta}$. For example, consider a three-dimensional vector-valued Gaussian process with parameter vector $\boldsymbol{\theta}=[\boldsymbol{\mu}, \boldsymbol{\Sigma}]$, where $\boldsymbol{\mu}$ is the mean vector and $\boldsymbol{\Sigma}$ is the covariance matrix of the Gaussian process. Figure. 4.2 illustrates three mean vectors in a three-dimensional parameter space. Also shown is the signal space associated with each parameter. As shown, the signal space of each parameter vector of a Gaussian process contains an infinite number of points, centred on the mean vector $\boldsymbol{\mu}$, and with a spatial volume and orientation that are determined by the covariance matrix $\boldsymbol{\Sigma}$. For simplicity, the variances are not shown in the parameter space, although they are evident in the shape of the Gaussian signal clusters in the signal space.

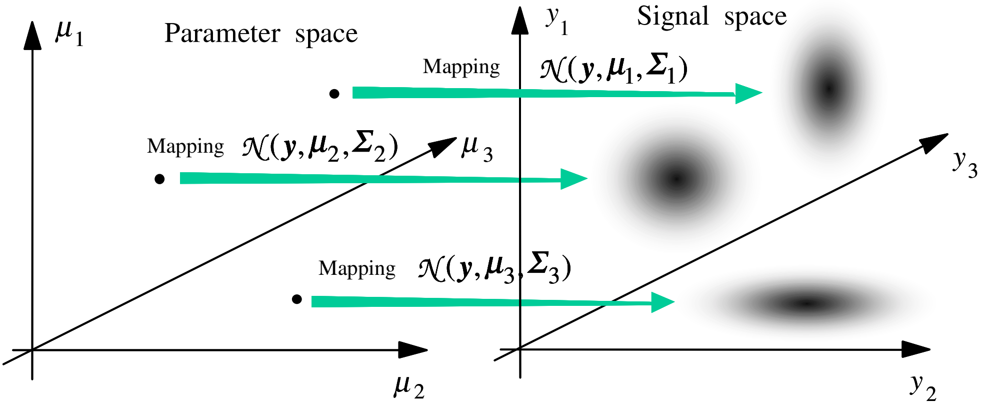

μ_1
Parameter space
y_1
Signal space
Mapping
$\mathcal{N}(\boldsymbol{y},\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1)$
Mapping
$\mathcal{N}(\boldsymbol{y},\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2)$
μ_3
y_3
Mapping
$\mathcal{N}(\boldsymbol{y},\boldsymbol{\mu}_3,\boldsymbol{\Sigma}_3)$
μ_2
y_2

Figure 4.2 Illustration of three points in the parameter space of a Gaussian process and the associated signal spaces, for simplicity the variances are not shown in parameter space.

4.1.3 Parameter Estimation and Signal Restoration

Parameter estimation and signal restoration are closely related problems. The main difference is due to the rapid fluctuations of most signals in comparison with the relatively slow variations of most parameters. For example, speech sounds fluctuate at speeds of up to 20 kHz, whereas the underlying vocal tract and pitch parameters vary at a relatively lower rate of less than 100 Hz. This observation implies that normally more averaging can be done in parameter estimation than in signal restoration.

As a simple example, consider a signal observed in a zero-mean random noise process. Assume we wish to estimate (a) the average of the clean signal and (b) the clean signal itself. As the observation length increases, the estimate of the signal mean approaches the mean value of the clean signal, whereas the estimate of the clean signal samples depends on the correlation structure of the signal and the signal-to-noise ratio as well as on the estimation method used.

As a further example, consider the interpolation of a sequence of lost samples of a signal given N recorded samples, as illustrated in Figure 4.3. Assume that an autoregressive (AR) process is used to model the signal as

$$\boldsymbol{y} = \boldsymbol{X\theta} + \boldsymbol{e} + \boldsymbol{n} \tag{4.3}$$

where $\boldsymbol{y}$ is the observation signal, $\boldsymbol{X}$ is the signal matrix, $\boldsymbol{\theta}$ is the AR parameter vector, $\boldsymbol{e}$ is the random input of the AR model and $\boldsymbol{n}$ is the random noise. Using Equation (4.3), the signal restoration process involves the estimation of both the model parameter vector $\boldsymbol{\theta}$ and the random input $\boldsymbol{e}$ for the lost samples. Assuming the parameter vector $\boldsymbol{\theta}$ is time-invariant, the estimate of $\boldsymbol{\theta}$ can be averaged over the entire N observation samples, and as N becomes infinitely large, a consistent estimate should approach the true

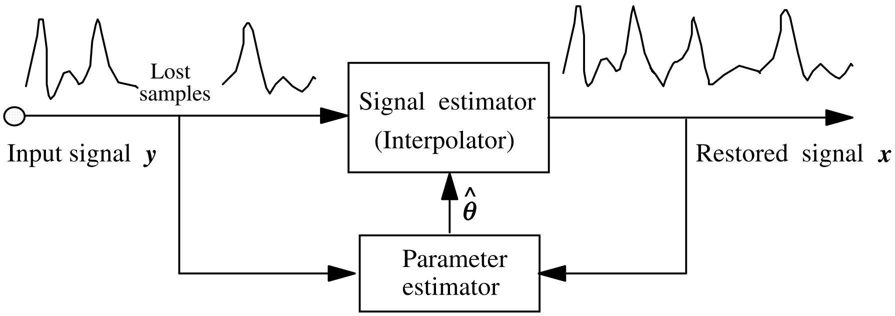


Figure 4.3 Illustration of signal restoration using a parametric model of the signal process.

parameter value. The difficulty in signal interpolation is that the underlying excitation $\boldsymbol{e}$ of the signal $\boldsymbol{x}$ is purely random and, unlike $\boldsymbol{\theta}$, it cannot be estimated through an averaging operation. In this chapter we are concerned with the parameter estimation problem, although the same ideas also apply to signal interpolation, which is considered in Chapter 11.

4.1.4 Performance Measures and Desirable Properties of Estimators

In estimation of a parameter vector $\boldsymbol{\theta}$ from N observation samples $\boldsymbol{y}$, a set of performance measures is used to quantify and compare the characteristics of different estimators. In general an estimate of a parameter vector is a function of the observation vector $\boldsymbol{y}$, the length of the observation N and the process model $\mathcal{M}$. This dependence may be expressed as

$$\hat{\boldsymbol{\theta}} = f(\boldsymbol{y}, N, \mathcal{M}) \tag{4.4}$$

Different parameter estimators produce different results depending on the estimation method and utilisation of the observation and the influence of the prior information. Due to randomness of the observations, even the same estimator would produce different results with different observations from the same process. Therefore an estimate is itself a random variable, it has a mean and a variance, and it may be described by a probability density function. However, for most cases, it is sufficient to characterise an estimator in terms of the mean and the variance of the estimation error. The most commonly used performance measures for an estimator are the following:

(a) *Expected value* of estimate: $\mathcal{E}[\hat{\boldsymbol{\theta}}]$

(b) *Bias* of estimate: $\mathcal{E}[\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}] = \mathcal{E}[\hat{\boldsymbol{\theta}}] - \boldsymbol{\theta}$

(c) *Covariance* of estimate: $\text{Cov}[\hat{\boldsymbol{\theta}}] = \mathcal{E}[(\hat{\boldsymbol{\theta}} - \mathcal{E}[\hat{\boldsymbol{\theta}}])(\hat{\boldsymbol{\theta}} - \mathcal{E}[\hat{\boldsymbol{\theta}}])^{\mathrm{T}}]$

Optimal estimators aim for zero bias and minimum estimation error covariance. The desirable properties of an estimator can be listed as follows:

(a) Unbiased estimator: an estimator of $\boldsymbol{\theta}$ is unbiased if the expectation of the estimate is equal to the true parameter value:

$$\mathcal{E}[\hat{\boldsymbol{\theta}}] = \boldsymbol{\theta} \tag{4.5}$$

An estimator is *asymptotically unbiased* if for increasing length of observations N we have

$$\lim_{N \to \infty} \mathcal{E}\,[\hat{\boldsymbol{\theta}}] = \boldsymbol{\theta} \tag{4.6}$$

(b) Efficient estimator: an unbiased estimator of $\boldsymbol{\theta}$ is an efficient estimator if it has the smallest covariance matrix compared with all other unbiased estimates of $\boldsymbol{\theta}$:

$$\mathrm{Cov}[\hat{\boldsymbol{\theta}}_{\mathrm{Efficient}}] \le \mathrm{Cov}[\hat{\boldsymbol{\theta}}] \tag{4.7}$$

where $\hat{\boldsymbol{\theta}}$ is any other estimate of $\boldsymbol{\theta}$.

(c) Consistent estimator: an estimator is consistent if the estimate improves with the increasing length of the observation N, such that the estimate $\hat{\boldsymbol{\theta}}$ converges probabilistically to the true value $\boldsymbol{\theta}$ as N becomes infinitely large:

$$\lim_{N\to\infty} P[|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}| > \varepsilon] = 0 \tag{4.8}$$

where ε is arbitrary small.

Example 4.1 Consider the bias in the time-averaged estimates of the mean μ_y and the variance σ_y^2 of N observation samples $[y(0), ..., y(N\text{–}1)]$, of an ergodic random process, given as

$$\hat{\mu}_y = \frac{1}{N}\sum_{m=0}^{N-1} y(m) \tag{4.9}$$

$$\hat{\sigma}_y^2 = \frac{1}{N}\sum_{m=0}^{N-1}\left[y(m) - \hat{\mu}_y\right]^2 \tag{4.10}$$

It is easy to show that $\hat{\mu}_y$ is an unbiased estimate, since

$$\mathcal{E}\left[\hat{\mu}_y\right] = \frac{1}{N}\sum_{m=0}^{N-1}\mathcal{E}[y(m)] = \mu_y \tag{4.11}$$

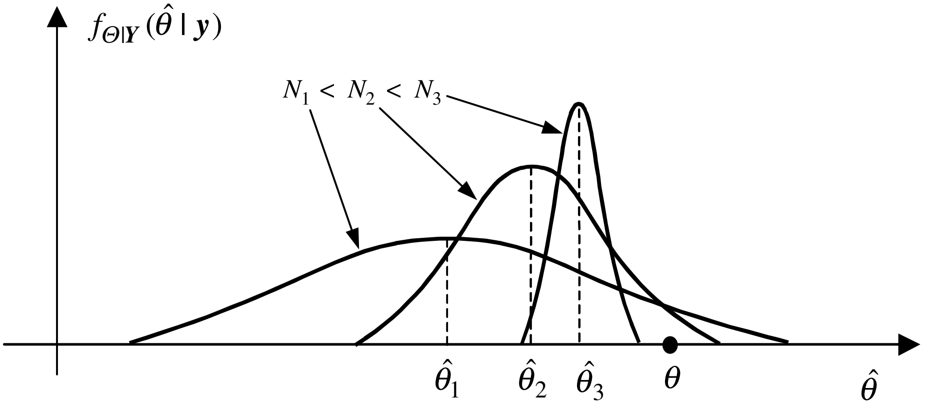


Figure 4.4 Illustration of the decrease in the bias and variance of an asymptotically unbiased estimate of the parameter θ with increasing length of observation.

The expectation of the estimate of the variance can be expressed as

$$\begin{aligned}\mathcal{E}\left[\hat{\sigma}_y^2\right] &= \mathcal{E}\left[\frac{1}{N}\sum_{m=0}^{N-1}\left(y(m)-\frac{1}{N}\sum_{k=0}^{N-1}y(k)\right)^2\right]\\ &= \sigma_y^2 - \frac{2}{N}\sigma_y^2 + \frac{1}{N}\sigma_y^2 \qquad (4.12)\\ &= \sigma_y^2 - \frac{1}{N}\sigma_y^2\end{aligned}$$

From Equation (4.12), the bias in the estimate of the variance is inversely proportional to the signal length *N*, and vanishes as *N* tends to infinity; hence the estimate is asymptotically unbiased. In general, the bias and the variance of an estimate decrease with increasing number of observation samples *N* and with improved modelling. Figure 4.4 illustrates the general dependence of the distribution and the bias and the variance of an asymptotically unbiased estimator on the number of observation samples *N*.

4.1.5 Prior and Posterior Spaces and Distributions

The *prior space* of a signal or a parameter vector is the collection of all possible values that the signal or the parameter vector can assume. The *posterior signal* or *parameter space* is the subspace of all the likely values of a signal or a parameter consistent with *both* the prior information and the evidence in the *observation*. Consider a random process with a parameter

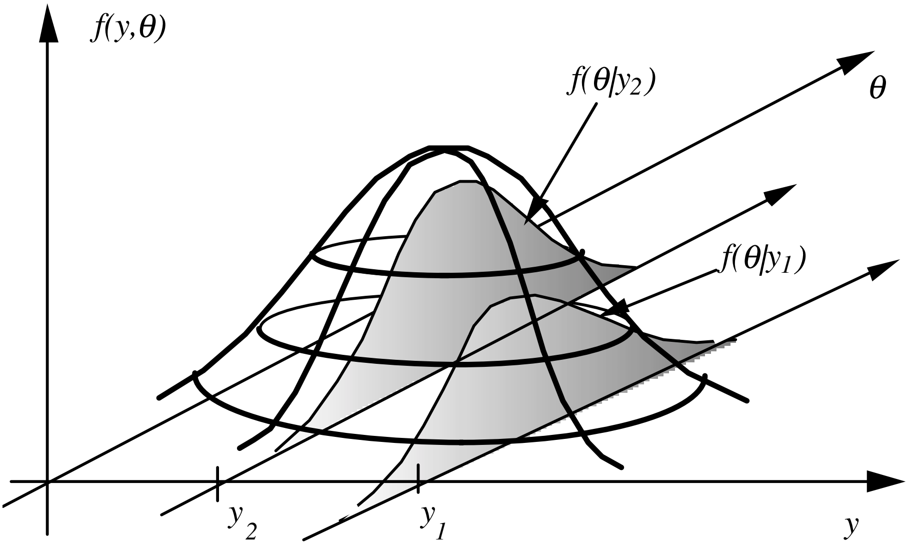


Figure 4.5 Illustration of joint distribution of signal *y* and parameter *θ* and the posterior distribution of *θ* given *y*.

space Θ observation space $\boldsymbol{Y}$ and a joint pdf $f_{Y,\Theta}(\boldsymbol{y},\boldsymbol{\theta})$. From the Bayes' rule the posterior pdf of the parameter vector $\boldsymbol{\theta}$, given an observation vector $\boldsymbol{y}$, $f_{\Theta|Y}(\boldsymbol{\theta}\,|\,\boldsymbol{y})$, can be expressed as

$$
\begin{aligned}
f_{\Theta|Y}(\boldsymbol{\theta}|\boldsymbol{y}) &= \frac{f_{Y|\Theta}(\boldsymbol{y}|\boldsymbol{\theta})f_{\Theta}(\boldsymbol{\theta})}{f_{Y}(\boldsymbol{y})} \\
&= \frac{f_{Y|\Theta}(\boldsymbol{y}|\boldsymbol{\theta})f_{\Theta}(\boldsymbol{\theta})}{\int\limits_{\Theta} f_{Y|\Theta}(\boldsymbol{y}|\boldsymbol{\theta})f_{\Theta}(\boldsymbol{\theta})\,d\boldsymbol{\theta}}
\end{aligned}
\tag{4.13}
$$

where, for a given observation vector $\boldsymbol{y}$, the pdf $f_Y(\boldsymbol{y})$ is a constant and has only a normalising effect. From Equation (4.13), the posterior pdf is proportional to the product of the likelihood $f_{Y|\Theta}(\boldsymbol{y}|\boldsymbol{\theta})$ that the observation $\boldsymbol{y}$ was generated by the parameter vector $\boldsymbol{\theta}$, and the prior pdf $f_{\Theta}(\boldsymbol{\theta})$. The prior pdf gives the unconditional parameter distribution *averaged* over the entire observation space as

$$
f_{\Theta}(\boldsymbol{\theta}) = \int\limits_{Y} f_{Y,\Theta}(\boldsymbol{y},\boldsymbol{\theta})\,d\boldsymbol{y}
\tag{4.14}
$$

For most applications, it is relatively convenient to obtain the likelihood function $f_{Y|\Theta}(\boldsymbol{y}|\boldsymbol{\theta})$. The *prior* pdf *influences* the inference drawn from the likelihood function by weighting it with $f_{\Theta}(\boldsymbol{\theta})$. The influence of the prior is particularly important for short-length and/or noisy observations, where the confidence in the estimate is limited by the lack of a sufficiently long observation and by the noise. The influence of the prior on the bias and the variance of an estimate are considered in Section 4.4.1.

A prior knowledge of the signal distribution can be used to confine the estimate to the prior signal space. The observation then guides the estimator to focus on the posterior space: that is the subspace consistent with both the prior and the observation. Figure 4.5 illustrates the joint pdf of a signal $y(m)$ and a parameter θ. The prior pdf of θ can be obtained by integrating $f_{Y|\Theta}(y(m)|\theta)$ with respect to $y(m)$. As shown, an observation $y(m)$ cuts a posterior pdf $f_{\Theta|Y}(\theta|y(m))$ through the joint distribution.

Example 4.2 A noisy signal vector of length N samples is modelled as

$$\boldsymbol{y}(m)=\boldsymbol{x}(m)+\boldsymbol{n}(m) \tag{4.15}$$

Assume that the signal $\boldsymbol{x}(m)$ is Gaussian with mean vector $\boldsymbol{\mu}_x$ and covariance matrix $\boldsymbol{\Sigma}_{xx}$, and that the noise $\boldsymbol{n}(m)$ is also Gaussian with mean vector $\boldsymbol{\mu}_n$ and covariance matrix $\boldsymbol{\Sigma}_{nn}$. The signal and noise pdfs model the prior spaces of the signal and the noise respectively. Given an observation vector $\boldsymbol{y}(m)$, the underlying signal $\boldsymbol{x}(m)$ would have a likelihood distribution with a mean vector of $\boldsymbol{y}(m) - \boldsymbol{\mu}_n$ and covariance matrix $\boldsymbol{\Sigma}_{nn}$ as shown in Figure 4.6.The likelihood function is given by

$$\begin{aligned} f_{Y|X}\left(\boldsymbol{y}(m)|\boldsymbol{x}(m)\right)&=f_N\left(\boldsymbol{y}(m)-\boldsymbol{x}(m)\right)\\ &=\frac{1}{(2\pi)^{N/2}|\boldsymbol{\Sigma}_{nn}|^{1/2}}\exp\left\{-\frac{1}{2}[\boldsymbol{x}(m)-(\boldsymbol{y}(m)-\boldsymbol{\mu}_n)]^{\mathrm{T}}\boldsymbol{\Sigma}_{nn}^{-1}[\boldsymbol{x}(m)-(\boldsymbol{y}(m)-\boldsymbol{\mu}_n)]\right\} \end{aligned} \tag{4.16}$$

where the terms in the exponential function have been rearranged to emphasize the illustration of the likelihood space in Figure 4.6. Hence the posterior pdf can be expressed as

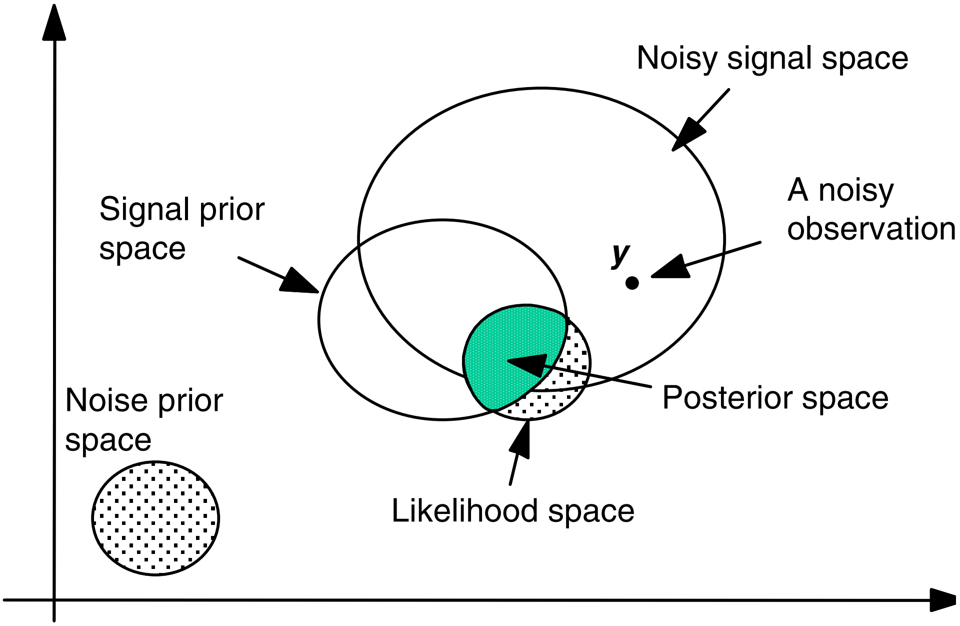


Figure 4.6 Sketch of a two-dimensional signal and noise spaces, and the likelihood and posterior spaces of a noisy observation ***y***.

$$
\begin{aligned}
f_{X|Y}(\boldsymbol{x}(m)|\boldsymbol{y}(m)) &= \frac{f_{Y|X}(\boldsymbol{y}(m)|\boldsymbol{x}(m))f_X(\boldsymbol{x}(m))}{f_Y(\boldsymbol{y}(m))} \\
&= \frac{1}{f_Y(\boldsymbol{y}(m))}\frac{1}{(2\pi)^N|\boldsymbol{\Sigma}_{nn}|^{1/2}|\boldsymbol{\Sigma}_{xx}|^{1/2}} \\
&\times \exp\left(-\frac{1}{2}\left\{[\boldsymbol{x}(m)-(\boldsymbol{y}(m)-\boldsymbol{\mu}_n)]^{\mathrm{T}}\boldsymbol{\Sigma}_{nn}^{-1}[\boldsymbol{x}(m)-(\boldsymbol{y}(m)-\boldsymbol{\mu}_n)]+(\boldsymbol{x}(m)-\boldsymbol{\mu}_x)^{\mathrm{T}}\boldsymbol{\Sigma}_{xx}^{-1}(\boldsymbol{x}(m)-\boldsymbol{\mu}_x)\right\}\right)
\end{aligned}
\tag{4.17}
$$

For a two-dimensional signal and noise process, the prior spaces of the signal, the noise, and the noisy signal are illustrated in Figure 4.6. Also illustrated are the likelihood and posterior spaces for a noisy observation vector ***y***. Note that the centre of the posterior space is obtained by subtracting the noise mean vector from the noisy signal vector. The clean signal is then somewhere within a subspace determined by the noise variance.

4.2 Bayesian Estimation

The Bayesian estimation of a parameter vector $\boldsymbol{\theta}$ is based on the minimisation of a Bayesian risk function defined as an average cost-of-error function:

$$\begin{aligned}\mathcal{R}(\hat{\boldsymbol{\theta}}) &= \mathcal{E}[C(\hat{\boldsymbol{\theta}},\boldsymbol{\theta})]\\ &= \int_{\boldsymbol{\theta}}\int_{\boldsymbol{Y}} C(\hat{\boldsymbol{\theta}},\boldsymbol{\theta})\, f_{Y,\Theta}(\boldsymbol{y},\boldsymbol{\theta})\, d\boldsymbol{y}\, d\boldsymbol{\theta}\\ &= \int_{\boldsymbol{\theta}}\int_{\boldsymbol{Y}} C(\hat{\boldsymbol{\theta}},\boldsymbol{\theta})\, f_{\Theta|Y}(\boldsymbol{\theta}\,|\,\boldsymbol{y})\, f_Y(\boldsymbol{y})\, d\boldsymbol{y}\, d\boldsymbol{\theta}\end{aligned} \tag{4.18}$$

where the cost-of-error function $C(\hat{\boldsymbol{\theta}},\boldsymbol{\theta})$ allows the appropriate weighting of the various outcomes to achieve desirable objective or subjective properties. The cost function can be chosen to associate a high cost with outcomes that are undesirable or disastrous. For a given observation vector $\boldsymbol{y}$, $f_Y(\boldsymbol{y})$ is a constant and has no effect on the risk-minimisation process. Hence Equation (4.18) may be written as a conditional risk function:

$$\mathcal{R}(\hat{\boldsymbol{\theta}}\,|\boldsymbol{y}) = \int_{\boldsymbol{\theta}} C(\hat{\boldsymbol{\theta}},\boldsymbol{\theta})\, f_{\Theta|Y}(\boldsymbol{\theta}\,|\boldsymbol{y})\, d\boldsymbol{\theta} \tag{4.19}$$

The Bayesian estimate obtained as the minimum-risk parameter vector is given by

$$\hat{\boldsymbol{\theta}}_{\text{Bayesian}} = \arg\min_{\hat{\boldsymbol{\theta}}} \mathcal{R}(\hat{\boldsymbol{\theta}}\,|\,\boldsymbol{y}) = \arg\min_{\hat{\boldsymbol{\theta}}}\left[\int_{\boldsymbol{\theta}} C(\hat{\boldsymbol{\theta}},\boldsymbol{\theta})\, f_{\Theta|Y}(\boldsymbol{\theta}\,|\boldsymbol{y})\, d\boldsymbol{\theta}\right] \tag{4.20}$$

Using Bayes' rule, Equation (4.20) can be written as

$$\hat{\boldsymbol{\theta}}_{\text{Bayesian}} = \arg\min_{\hat{\boldsymbol{\theta}}}\left[\int_{\boldsymbol{\theta}} C(\hat{\boldsymbol{\theta}},\boldsymbol{\theta})\, f_{Y|\Theta}(\boldsymbol{y}\,|\,\boldsymbol{\theta})\, f_\Theta(\boldsymbol{\theta})\, d\boldsymbol{\theta}\right] \tag{4.21}$$

Assuming that the risk function is differentiable, and has a well-defined minimum, the Bayesian estimate can be obtained as

$$\hat{\boldsymbol{\theta}}_{\text{Bayesian}} = \arg\operatorname*{zero}_{\hat{\boldsymbol{\theta}}} \frac{\partial\mathcal{R}(\hat{\boldsymbol{\theta}}\,|\,\boldsymbol{y})}{\partial\hat{\boldsymbol{\theta}}} = \arg\operatorname*{zero}_{\hat{\boldsymbol{\theta}}}\left[\frac{\partial}{\partial\hat{\boldsymbol{\theta}}}\int_{\boldsymbol{\theta}} C(\hat{\boldsymbol{\theta}},\boldsymbol{\theta})\, f_{Y|\Theta}(\boldsymbol{y}\,|\,\boldsymbol{\theta})\, f_\Theta(\boldsymbol{\theta})\, d\boldsymbol{\theta}\right] \tag{4.22}$$

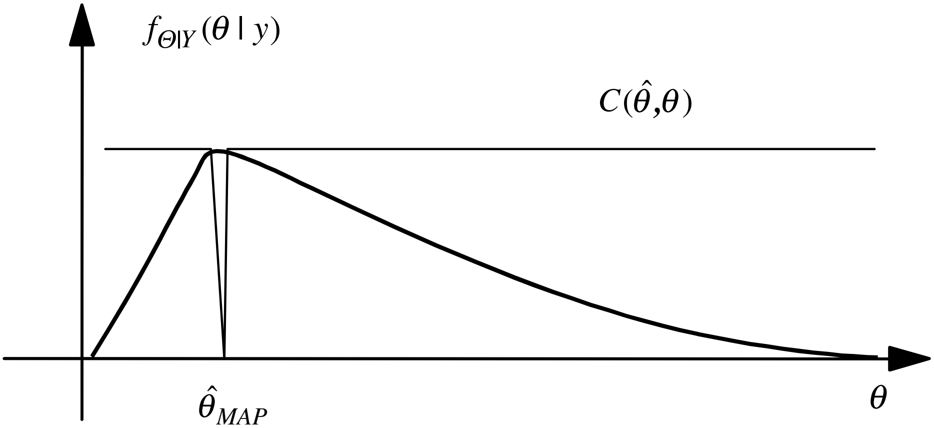


Figure 4.7 Illustration of the Bayesian cost function for the MAP estimate.

4.2.1 Maximum A Posteriori Estimation

The maximum a posteriori (MAP) estimate $\hat{\theta}_{MAP}$ is obtained as the parameter vector that maximises the posterior pdf $f_{\Theta|Y}(\boldsymbol{\theta}|\boldsymbol{y})$. The MAP estimate corresponds to a Bayesian estimate with a so-called uniform cost function (in fact, as shown in Figure 4.7 the cost function is notch-shaped) defined as

$$C(\hat{\boldsymbol{\theta}},\boldsymbol{\theta}) = 1 - \delta(\hat{\boldsymbol{\theta}},\boldsymbol{\theta}) \tag{4.23}$$

where $\delta(\hat{\boldsymbol{\theta}},\boldsymbol{\theta})$ is the Kronecker delta function. Substitution of the cost function in the Bayesian risk equation yields

$$\begin{aligned}\mathcal{R}_{MAP}(\hat{\boldsymbol{\theta}}|\boldsymbol{y}) &= \int_{\boldsymbol{\theta}}[1-\delta(\hat{\boldsymbol{\theta}},\boldsymbol{\theta})] f_{\Theta|Y}(\boldsymbol{\theta}|\boldsymbol{y})\, d\boldsymbol{\theta} \\ &= 1 - f_{\Theta|Y}(\hat{\boldsymbol{\theta}}|\boldsymbol{y})\end{aligned} \tag{4.24}$$

From Equation (4.24), the minimum Bayesian risk estimate corresponds to the parameter value where the posterior function attains a maximum. Hence the MAP estimate of the parameter vector $\boldsymbol{\theta}$ is obtained from a minimisation of the risk Equation (4.24) or equivalently maximisation of the posterior function:

$$\begin{aligned}\hat{\theta}_{MAP} &= \arg\max_{\boldsymbol{\theta}} f_{\Theta|Y}(\boldsymbol{\theta}|\boldsymbol{y}) \\ &= \arg\max_{\boldsymbol{\theta}} [f_{Y|\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{\theta}) f_{\Theta}(\boldsymbol{\theta})]\end{aligned} \tag{4.25}$$

4.2.2 Maximum-Likelihood Estimation

The maximum-likelihood (ML) estimate $\hat{\boldsymbol{\theta}}_{ML}$ is obtained as the parameter vector that maximises the likelihood function $f_{Y|\Theta}(\boldsymbol{y}|\boldsymbol{\theta})$. The ML estimator corresponds to a Bayesian estimator with a uniform cost function and a uniform parameter prior pdf:

$$\begin{aligned}\mathcal{R}_{\mathrm{ML}}(\hat{\boldsymbol{\theta}}|\boldsymbol{y}) &= \int_{\boldsymbol{\theta}} [1-\delta(\hat{\boldsymbol{\theta}},\boldsymbol{\theta})] f_{Y|\Theta}(\boldsymbol{y}|\boldsymbol{\theta}) f_{\Theta}(\boldsymbol{\theta}) d\boldsymbol{\theta} \\ &= \text{const.}[1-f_{Y|\Theta}(\boldsymbol{y}|\hat{\boldsymbol{\theta}})]\end{aligned} \tag{4.26}$$

where the prior function $f_\Theta(\boldsymbol{\theta})$=const. From a Bayesian point of view the main difference between the ML and MAP estimators is that the ML assumes that the prior pdf of $\boldsymbol{\theta}$ is uniform. Note that a uniform prior, in addition to modelling genuinely uniform pdfs, is also used when the parameter prior pdf is unknown, or when the parameter is an unknown constant.

From Equation (4.26), it is evident that minimisation of the risk function is achieved by maximisation of the likelihood function:

$$\hat{\boldsymbol{\theta}}_{ML} = \arg\max_{\boldsymbol{\theta}} f_{Y|\Theta}(\boldsymbol{y}|\boldsymbol{\theta}) \tag{4.27}$$

In practice it is convenient to maximise the log-likelihood function instead of the likelihood:

$$\boldsymbol{\theta}_{ML} = \arg\max_{\boldsymbol{\theta}} \; \log f_{Y|\boldsymbol{\theta}}(\boldsymbol{Y}\,|\,\boldsymbol{\theta}) \tag{4.28}$$

The log-likelihood is usually chosen in practice because:

(a) the logarithm is a monotonic function, and hence the log-likelihood has the same turning points as the likelihood function;
(b) the joint log-likelihood of a set of independent variables is the sum of the log-likelihood of individual elements; and
(c) unlike the likelihood function, the log-likelihood has a dynamic range that does not cause computational under-flow.

Example 4.3 _ML Estimation of the mean and variance of a Gaussian process_ Consider the problem of maximum likelihood estimation of the mean vector $\boldsymbol{\mu}_y$ and the covariance matrix $\boldsymbol{\Sigma}_{yy}$ of a P-dimensional

Gaussian vector process from N observation vectors $\left[\boldsymbol{y}(0), \boldsymbol{y}(1), \ldots, \boldsymbol{y}(N-1)\right]$. Assuming the observation vectors are uncorrelated, the pdf of the observation sequence is given by

$$f_Y\left(\boldsymbol{y}(0), \cdots, \boldsymbol{y}(N-1)\right) = \prod_{m=0}^{N-1} \frac{1}{(2\pi)^{P/2} \left|\boldsymbol{\Sigma}_{yy}\right|^{1/2}} \exp\left\{ -\frac{1}{2}\left[\boldsymbol{y}(m) - \boldsymbol{\mu}_y\right]^{\mathrm{T}} \boldsymbol{\Sigma}_{yy}^{-1}\left[\boldsymbol{y}(m) - \boldsymbol{\mu}_y\right]\right\} \tag{4.29}$$

and the log-likelihood equation is given by

$$\ln f_Y\left(\boldsymbol{y}(0), \ldots, \boldsymbol{y}(N-1)\right) = \sum_{m=0}^{N-1} \left\{ -\frac{P}{2}\ln(2\pi) - \frac{1}{2}\ln\left|\boldsymbol{\Sigma}_{yy}\right| - \frac{1}{2}\left[\boldsymbol{y}(m) - \boldsymbol{\mu}_y\right]^{\mathrm{T}} \boldsymbol{\Sigma}_{yy}^{-1}\left[\boldsymbol{y}(m) - \boldsymbol{\mu}_y\right]\right\} \tag{4.30}$$

Taking the derivative of the log-likelihood equation with respect to the mean vector $\boldsymbol{\mu}_y$ yields

$$\frac{\partial \ln f_Y\left(\boldsymbol{y}(0), \ldots, \boldsymbol{y}(N-1)\right)}{\partial \boldsymbol{\mu}_y} = \sum_{m=0}^{N-1} \left[2\boldsymbol{\Sigma}_{yy}^{-1}\boldsymbol{\mu}_y - 2\boldsymbol{\Sigma}_{yy}^{-1}\boldsymbol{y}(m)\right] = 0 \tag{4.31}$$

From Equation (4.31), we have

$$\hat{\boldsymbol{\mu}}_y = \frac{1}{N}\sum_{m=0}^{N-1} \boldsymbol{y}(m) \tag{4.32}$$

To obtain the ML estimate of the covariance matrix we take the derivative of the log-likelihood equation with respect to $\boldsymbol{\Sigma}_{yy}^{-1}$:

$$\frac{\partial \ln f_Y\left(\boldsymbol{y}(0), \cdots, \boldsymbol{y}(N-1)\right)}{\partial \boldsymbol{\Sigma}_{yy}^{-1}} = \sum_{m=0}^{N-1} \left\{ \frac{1}{2}\boldsymbol{\Sigma}_{yy} - \frac{1}{2}\left[\boldsymbol{y}(m) - \boldsymbol{\mu}_y\right]\left[\boldsymbol{y}(m) - \boldsymbol{\mu}_y\right]^{\mathrm{T}} \right\} = 0 \tag{4.33}$$

From Equation (4.31), we have an estimate of the covariance matrix as

$$\hat{\boldsymbol{\Sigma}}_{yy} = \frac{1}{N}\sum_{m=0}^{N-1} [\boldsymbol{y}(m) - \hat{\boldsymbol{\mu}}_y][\boldsymbol{y}(m) - \hat{\boldsymbol{\mu}}_y]^{\mathrm{T}} \tag{4.34}$$

Example 4.4 ML and MAP Estimation of a Gaussian Random Parameter. Consider the estimation of a P-dimensional random parameter vector $\boldsymbol{\theta}$ from an N-dimensional observation vector $\boldsymbol{y}$. Assume that the relation between the signal vector $\boldsymbol{y}$ and the parameter vector $\boldsymbol{\theta}$ is described by a linear model as

$$\boldsymbol{y} = \boldsymbol{G\theta} + \boldsymbol{e} \tag{4.35}$$

where $\boldsymbol{e}$ is a random excitation input signal. The pdf of the parameter vector $\boldsymbol{\theta}$ given an observation vector $\boldsymbol{y}$ can be described, using Bayes' rule, as

$$f_{\Theta|Y}(\boldsymbol{\theta} \mid \boldsymbol{y}) = \frac{1}{f_Y(\boldsymbol{y})} f_{Y|\Theta}(\boldsymbol{y} \mid \boldsymbol{\theta}) f_\Theta(\boldsymbol{\theta}) \tag{4.36}$$

Assuming that the matrix $\boldsymbol{G}$ in Equation (4.35) is known, the likelihood of the signal $\boldsymbol{y}$ given the parameter vector $\boldsymbol{\theta}$ is the pdf of the random vector $\boldsymbol{e}$:

$$f_{Y|\Theta}(\boldsymbol{y}|\boldsymbol{\theta}) = f_{\boldsymbol{E}}(\boldsymbol{e} = \boldsymbol{y} - \boldsymbol{G\theta}) \tag{4.37}$$

Now assume the input $\boldsymbol{e}$ is a zero-mean, Gaussian-distributed, random process with a diagonal covariance matrix, and the parameter vector $\boldsymbol{\theta}$ is also a Gaussian process with mean of μ_θ and covariance matrix $\boldsymbol{\Sigma}_{\theta\theta}$. Therefore we have

$$f_{Y|\Theta}(\boldsymbol{y} \mid \boldsymbol{\theta}) = f_{\boldsymbol{E}}(\boldsymbol{e}) = \frac{1}{(2\pi\sigma_e^2)^{N/2}} \exp\left[-\frac{1}{2\sigma_e^2} (\boldsymbol{y} - \boldsymbol{G\theta})^{\mathrm{T}} (\boldsymbol{y} - \boldsymbol{G\theta}) \right] \tag{4.38}$$

and

$$f_\Theta(\boldsymbol{\theta}) = \frac{1}{(2\pi)^{P/2} \left| \boldsymbol{\Sigma}_{\theta\theta} \right|^{1/2}} \exp\left[-\frac{1}{2} (\boldsymbol{\theta} - \boldsymbol{\mu}_\theta)^{\mathrm{T}} \boldsymbol{\Sigma}_{\theta\theta}^{-1} (\boldsymbol{\theta} - \boldsymbol{\mu}_\theta) \right] \tag{4.39}$$

The ML estimate obtained from maximisation of the log-likelihood function $\ln\left[f_{Y|\Theta}(\boldsymbol{y} \mid \boldsymbol{\theta})\right]$ with respect to $\boldsymbol{\theta}$ is given by

$$\hat{\boldsymbol{\theta}}_{ML}(\boldsymbol{y}) = \left(\boldsymbol{G}^{\mathrm{T}} \boldsymbol{G}\right)^{-1} \boldsymbol{G}^{\mathrm{T}} \boldsymbol{y} \tag{4.40}$$

To obtain the MAP estimate we first form the posterior distribution by substituting Equations (4.38) and (4.39) in Equation (4.36)

$$f_{\Theta|Y}(\boldsymbol{\theta}\,|\,\boldsymbol{y})=\frac{1}{f_{Y}(\boldsymbol{y})}\,\frac{1}{(2\pi\sigma_e^2)^{N/2}}\,\frac{1}{(2\pi)^{P/2}\left|\boldsymbol{\Sigma}_{\theta\theta}\right|^{1/2}}$$
$$\times\exp\left(-\frac{1}{2\sigma_e^2}(\boldsymbol{y}-\boldsymbol{G\theta})^{\mathrm{T}}(\boldsymbol{y}-\boldsymbol{G\theta})-\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\mu}_{\theta})^{\mathrm{T}}\boldsymbol{\Sigma}_{\theta\theta}^{-1}(\boldsymbol{\theta}-\boldsymbol{\mu}_{\theta})\right) \tag{4.41}$$

The MAP parameter estimate is obtained by differentiating the log-likelihood function $\ln f_{\Theta|Y}(\boldsymbol{\theta}\,|\,\boldsymbol{y})$ and setting the derivative to zero:

$$\hat{\boldsymbol{\theta}}_{MAP}(\boldsymbol{y})=\left(\boldsymbol{G}^{\mathrm{T}}\boldsymbol{G}+\sigma_e^2\boldsymbol{\Sigma}_{\theta\theta}^{-1}\right)^{-1}\left(\boldsymbol{G}^{\mathrm{T}}\boldsymbol{y}+\sigma_e^2\boldsymbol{\Sigma}_{\theta\theta}^{-1}\boldsymbol{\mu}_{\theta}\right) \tag{4.42}$$

Note that as the covariance of the Gaussian-distributed parameter increases, or equivalently as $\boldsymbol{\Sigma}_{\theta\theta}^{-1}\to 0$, the Gaussian prior tends to a uniform prior and the MAP solution Equation (4.42) tends to the ML solution given by Equation (4.40). Conversely as the pdf of the parameter vector $\boldsymbol{\theta}$ becomes peaked, i.e. as $\boldsymbol{\Sigma}_{\theta\theta}\to 0$, the estimate tends towards $\boldsymbol{\mu}_{\theta}$.

4.2.3 Minimum Mean Square Error Estimation

The Bayesian minimum mean square error (MMSE) estimate is obtained as the parameter vector that minimises a mean square error cost function (Figure 4.8) defined as

$$\begin{aligned}\mathcal{R}_{MMSE}(\hat{\boldsymbol{\theta}}\,|\boldsymbol{y})&=\mathcal{E}[(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})^2\,|\,\boldsymbol{y}]\\&=\int_{\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})^2 f_{\theta|Y}(\boldsymbol{\theta}\,|\,\boldsymbol{y})\,d\boldsymbol{\theta}\end{aligned} \tag{4.43}$$

In the following, it is shown that *the Bayesian MMSE estimate is the conditional mean of the posterior pdf.* Assuming that the mean square error risk function is differentiable and has a well-defined minimum, the MMSE solution can be obtained by setting the gradient of the mean square error risk function to zero:

$$\frac{\partial\mathcal{R}_{MMSE}(\hat{\boldsymbol{\theta}}|\boldsymbol{y})}{\partial\hat{\boldsymbol{\theta}}}=2\hat{\boldsymbol{\theta}}\int_{\boldsymbol{\theta}}f_{\Theta|Y}(\boldsymbol{\theta}\,|\,\boldsymbol{y})\,d\boldsymbol{\theta}-2\int_{\boldsymbol{\theta}}\boldsymbol{\theta}\,f_{\Theta|Y}(\boldsymbol{\theta}\,|\,\boldsymbol{y})\,d\boldsymbol{\theta} \tag{4.44}$$

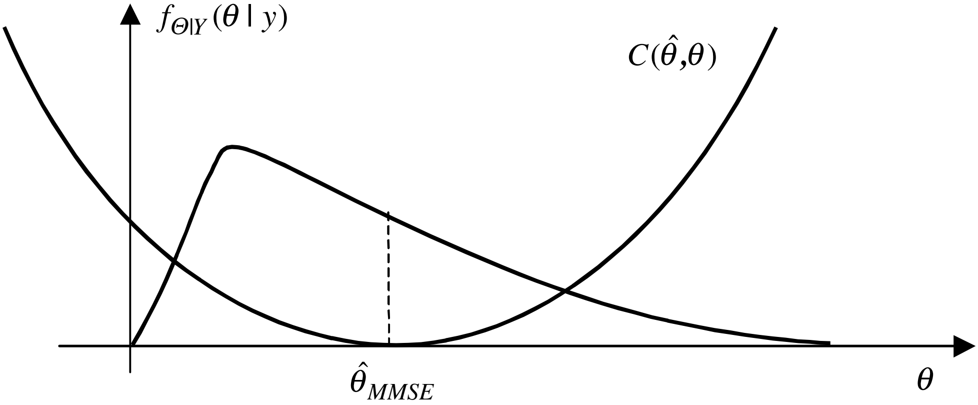


Figure 4.8 Illustration of the mean square error cost function and estimate.

Since the first integral on the right hand-side of Equation (4.42) is equal to 1, we have

$$\frac{\partial R_{MMSE}(\hat{\boldsymbol{\theta}} \mid \boldsymbol{y})}{\partial \hat{\boldsymbol{\theta}}} = 2\hat{\boldsymbol{\theta}} - \int_{\boldsymbol{\theta}} \boldsymbol{\theta}\, f_{\Theta|\boldsymbol{Y}}(d\boldsymbol{\theta} \mid \boldsymbol{y})\, d\boldsymbol{\theta} \tag{4.45}$$

The MMSE solution is obtained by setting Equation (4.45) to zero:

$$\hat{\boldsymbol{\theta}}_{MMSE}(\boldsymbol{y}) = \int_{\boldsymbol{\theta}} \boldsymbol{\theta} f_{\Theta|\boldsymbol{Y}}(\boldsymbol{\theta} \mid \boldsymbol{y}) d\boldsymbol{\theta} \tag{4.46}$$

For cases where we do not have a pdf model of the parameter process, the minimum mean square error (known as the least square error, LSE) estimate is obtained through minimisation of a mean square error function $\mathcal{E}[e^2(\boldsymbol{\theta}|\boldsymbol{y})]$:

$$\hat{\boldsymbol{\theta}}_{LSE} = \arg\min_{\boldsymbol{\theta}} \mathcal{E}[e^2(\boldsymbol{\theta} \mid y)] \tag{4.47}$$

Th LSE estimation of Equation (4.47) does not use any prior knowledge of the distribution of the signals and the parameters. This can be considered as a strength of LSE in situations where the prior pdfs are unknown, but it can also be considered as a weakness in cases where fairly accurate models of the priors are available but not utilised.

Example 4.5 Consider the MMSE estimation of a parameter vector $\boldsymbol{\theta}$ assuming a linear model of the observation $\boldsymbol{y}$ as

$$\boldsymbol{y} = \boldsymbol{G}\boldsymbol{\theta} + \boldsymbol{e} \tag{4.48}$$

The LSE estimate is obtained as the parameter vector at which the gradient of the mean squared error with respect to $\boldsymbol{\theta}$ is zero:

$$\frac{\partial \boldsymbol{e}^{\mathrm{T}}\boldsymbol{e}}{\partial \boldsymbol{\theta}} = \frac{\partial}{\partial \boldsymbol{\theta}}(\boldsymbol{y}^{\mathrm{T}}\boldsymbol{y} - 2\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{G}^{\mathrm{T}}\boldsymbol{y} + \boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{G}^{\mathrm{T}}\boldsymbol{G}\boldsymbol{\theta})\bigg|_{\boldsymbol{\theta}_{LSE}} = 0 \tag{4.49}$$

From Equation (4.49) the LSE parameter estimate is given by

$$\boldsymbol{\theta}_{LSE} = [\boldsymbol{G}^{\mathrm{T}}\boldsymbol{G}]^{-1}\boldsymbol{G}^{\mathrm{T}}\boldsymbol{y} \tag{4.50}$$

Note that for a Gaussian likelihood function, the LSE solution is the same as the ML solution of Equation (4.40).

4.2.4 Minimum Mean Absolute Value of Error Estimation

The minimum mean absolute value of error (MAVE) estimate (Figure 4.9) is obtained through minimisation of a Bayesian risk function defined as

$$\mathcal{R}_{MAVE}(\hat{\theta} | \boldsymbol{y}) = \mathcal{E}[|\hat{\theta} - \theta| \ \boldsymbol{y}] = \int_{\theta} |\hat{\theta} - \theta| f_{\theta|Y}(\theta | \boldsymbol{y})\, d\theta \tag{4.51}$$

In the following it is shown that the minimum mean absolute value estimate is the median of the parameter process. Equation (4.51) can be re-expressed as

$$\mathcal{R}_{MAVE}(\hat{\theta} | \boldsymbol{y}) = \int_{-\infty}^{\hat{\theta}} [\hat{\theta} - \theta] f_{\Theta|Y}(\theta | \boldsymbol{y})\, d\theta + \int_{\hat{\theta}}^{\infty} [\theta - \hat{\theta}] f_{\Theta|Y}(\theta | \boldsymbol{y})\, d\theta \tag{4.52}$$

Taking the derivative of the risk function with respect to $\hat{\theta}$ yields

$$\frac{\partial \mathcal{R}_{MAVE}(\hat{\theta} | \boldsymbol{y})}{\partial \hat{\theta}} = \int_{-\infty}^{\hat{\theta}} f_{\Theta|Y}(\theta | \boldsymbol{y})\, d\theta - \int_{\hat{\theta}}^{\infty} f_{\Theta|Y}(\theta | \boldsymbol{y})\, d\theta \tag{4.53}$$

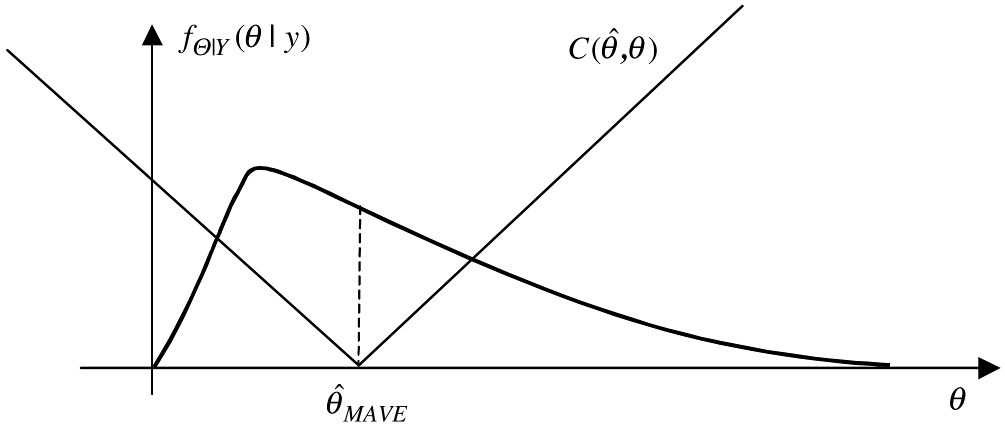


Figure 4.9 Illustration of mean absolute value of error cost function. Note that the MAVE estimate coincides with the conditional median of the posterior function.

The minimum absolute value of error is obtained by setting Equation (4.53) to zero:

$$\int_{-\infty}^{\hat{\theta}_{MAVE}} f_{\Theta|Y}(\theta\,|\,y)\,d\theta = \int_{\hat{\theta}_{MAVE}}^{\infty} f_{\Theta|Y}(\theta\,|\,y)\,d\theta \tag{4.54}$$

From Equation (4.54) we note the MAVE estimate is the median of the posterior density.

4.2.5 Equivalence of the MAP, ML, MMSE and MAVE for Gaussian Processes With Uniform Distributed Parameters

Example 4.4 shows that for a Gaussian-distributed process the LSE estimate and the ML estimate are identical. Furthermore, Equation (4.42), for the MAP estimate of a Gaussian-distributed parameter, shows that as the parameter variance increases, or equivalently as the parameter prior pdf tends to a uniform distribution, the MAP estimate tends to the ML and LSE estimates. In general, for any symmetric distribution, centred round the maximum, the mode, the mean and the median are identical. Hence, for a process with a symmetric pdf, if the prior distribution of the parameter is uniform then the MAP, the ML, the MMSE and the MAVE parameter estimates are identical. Figure 4.10 illustrates a symmetric pdf, an asymmetric pdf, and the relative positions of various estimates.

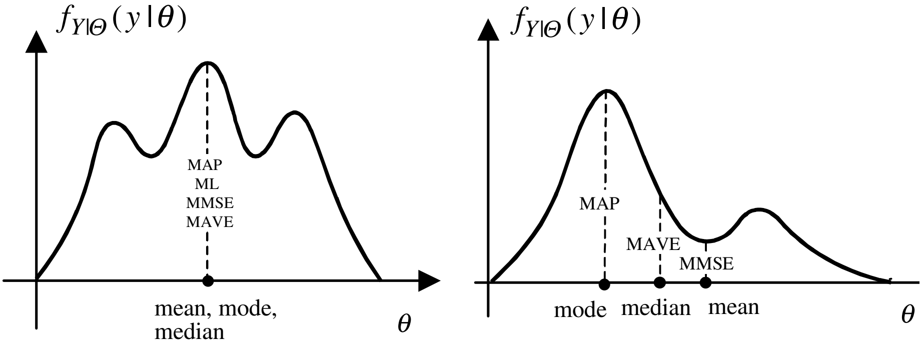


Figure 4.10 Illustration of a symmetric and an asymmetric pdf and their respective mode, mean and median and the relations to MAP, MAVE and MMSE estimates.

4.2.6 The Influence of the Prior on Estimation Bias and Variance

The use of a prior pdf introduces a bias in the estimate towards the range of parameter values with a relatively high prior pdf, and reduces the variance of the estimate. To illustrate the effects of the prior pdf on the bias and the variance of an estimate, we consider the following examples in which the bias and the variance of the ML and the MAP estimates of the mean of a process are compared.

Example 4.6 Consider the ML estimation of a random scalar parameter θ, observed in a zero-mean additive white Gaussian noise (AWGN) $n(m)$, and expressed as

$$y(m) = \theta + n(m), \quad m= 0,..., N-1 \tag{4.55}$$

It is assumed that, for each realisation of the parameter θ, N observation samples are available. Note that, since the noise is assumed to be a zero-mean process, this problem is equivalent to estimation of the mean of the process $y(m)$. The likelihood of an observation vector $\boldsymbol{y}$=[y(0), y(1), …, y(N–1)] and a parameter value of θ is given by

$$\begin{aligned} f_{Y|\Theta}(\boldsymbol{y}|\theta) &= \prod_{m=0}^{N-1} f_N\left(y(m)-\theta\right) \\ &= \frac{1}{(2\pi\sigma_n^2)^{N/2}} \exp\left\{ -\frac{1}{2\sigma_n^2} \sum_{m=0}^{N-1} [y(m)-\theta]^2 \right\} \end{aligned} \tag{4.56}$$

From Equation (4.56) the log-likelihood function is given by

$$\ln f_{Y|\Theta}(\boldsymbol{y}\,|\,\theta)=-\frac{N}{2}\ln(2\pi\sigma_n^2)-\frac{1}{2\sigma_n^2}\sum_{m=0}^{N-1}[y(m)-\theta]^2 \tag{4.57}$$

The ML estimate of θ, obtained by setting the derivative of $\ln f_{Y|\Theta}(\boldsymbol{y}|\theta)$ to zero, is given by

$$\hat{\theta}_{ML}=\frac{1}{N}\sum_{m=0}^{N-1}y(m)=\bar{y} \tag{4.58}$$

where $\bar{y}$ denotes the time average of $y(m)$. From Equation (4.56), we note that the ML solution is an unbiased estimate

$$\mathcal{E}[\hat{\theta}_{ML}]=\mathcal{E}\left(\frac{1}{N}\sum_{m=0}^{N-1}[\theta+n(m)]\right)=\theta \tag{4.59}$$

and the variance of the ML estimate is given by

$$\text{Var}[\hat{\theta}_{ML}]=\mathcal{E}[(\hat{\theta}_{ML}-\theta)^2]=\mathcal{E}\left[\left(\frac{1}{N}\sum_{m=0}^{N-1}y(m)-\theta\right)^2\right]=\frac{\sigma_n^2}{N} \tag{4.60}$$

Note that the variance of the ML estimate decreases with increasing length of observation.

Example 4.7 _Estimation of a uniformly-distributed parameter observed in AWGN._ Consider the effects of using a uniform parameter prior on the mean and the variance of the estimate in Example 4.6. Assume that the prior for the parameter θ is given by

$$f_\Theta(\theta)=\begin{cases}1/(\theta_{\max}-\theta_{\min}) & \theta_{\min}\le\theta\le\theta_{\max}\\ 0 & \text{otherwise}\end{cases} \tag{4.61}$$

as illustrated in Figure 4.11. From Bayes' rule, the posterior pdf is given by

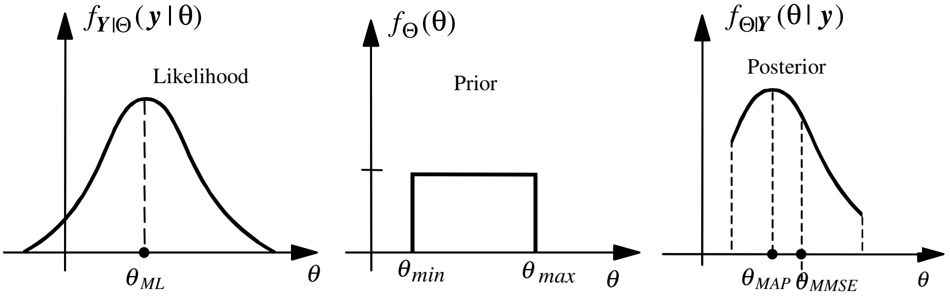


Figure 4.11 Illustration of the effects of a uniform prior.

$$
\begin{aligned}
f_{\Theta|Y}(\theta\,|\,\boldsymbol{y}) &= \frac{1}{f_Y(\boldsymbol{y})} f_{Y|\Theta}(\boldsymbol{y}\,|\,\theta) f_\Theta(\theta) \\
&= \begin{cases} \dfrac{1}{f_Y(\boldsymbol{y})} \dfrac{1}{(2\pi\sigma_n^2)^{N/2}} \exp\left\{ -\dfrac{1}{2\sigma_n^2} \displaystyle\sum_{m=0}^{N-1} [y(m)-\theta]^2 \right\}, & \theta_{\min} \le \theta \le \theta_{\max} \\ 0, & \text{otherwise} \end{cases}
\end{aligned}
\tag{4.62}
$$

The MAP estimate is obtained by maximising the posterior pdf:

$$
\hat{\theta}_{MAP}(\boldsymbol{y}) = \begin{cases} \theta_{\min} & \text{if } \hat{\theta}_{ML}(\boldsymbol{y}) < \theta_{\min} \\ \hat{\theta}_{ML}(\boldsymbol{y}) & \text{if } \theta_{\min} \ge \hat{\theta}_{ML}(\boldsymbol{y}) \ge \theta_{\max} \\ \theta_{\max} & \text{if } \hat{\theta}_{ML}(\boldsymbol{y}) > \theta_{\max} \end{cases}
\tag{4.63}
$$

Note that the MAP estimate is constrained to the range $\theta_{\min}$ to $\theta_{\max}$. This constraint is desirable and moderates the estimates that, due to say low signal-to-noise ratio, fall outside the range of possible values of θ. It is easy to see that the variance of an estimate constrained to a range of $\theta_{\min}$ to $\theta_{\max}$ is less than the variance of the ML estimate in which there is no constraint on the range of the parameter estimate:

$$
\mathrm{Var}[\hat{\theta}_{MAP}] = \int_{\theta_{\min}}^{\theta_{\max}} (\hat{\theta}_{MAP} - \theta)^2 f_{Y|\Theta}(\boldsymbol{y}\,|\,\theta\, d\boldsymbol{y} \le \mathrm{Var}[\hat{\theta}_{ML}] = \int_{-\infty}^{\infty} (\hat{\theta}_{ML} - \theta)^2 f_{Y|\Theta}(\boldsymbol{y}\,|\,\theta)\, d\boldsymbol{y}
\tag{4.64}
$$

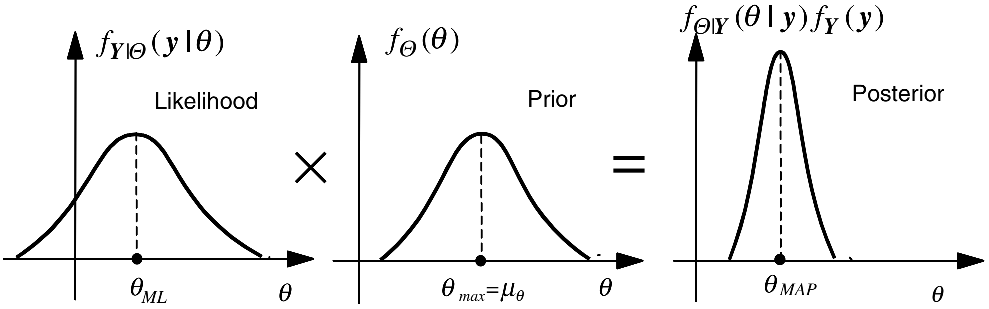


Figure 4.12 Illustration of the posterior pdf as product of the likelihood and the prior.

Example 4.8 Estimation of a Gaussian-distributed parameter observed in AWGN. In this example, we consider the effect of a Gaussian prior on the mean and the variance of the MAP estimate. Assume that the parameter θ is Gaussian-distributed with a mean μ_θ and a variance σ_θ^2 as

$$f_\Theta(\theta) = \frac{1}{(2\pi\sigma_\theta^2)^{1/2}} \exp\left[- \frac{(\theta - \mu_\theta)^2}{2\sigma_\theta^2} \right] \tag{4.65}$$

From Bayes rule the posterior pdf is given as the product of the likelihood and the prior pdfs as:

$$\begin{aligned} f_{\Theta|Y}(\theta|y) &= \frac{1}{f_Y(y)} f_{Y|\Theta}(y|\theta) f_\Theta(\theta) \\ &= \frac{1}{f_Y(y)} \frac{1}{(2\pi\sigma_n^2)^{N/2}(2\pi\sigma_\theta^2)^{1/2}} \exp\left\{ -\frac{1}{2\sigma_n^2}\sum_{m=0}^{N-1}[y(m)-\theta]^2 - \frac{1}{2\sigma_\theta^2}(\theta-\mu_\theta)^2 \right\} \end{aligned} \tag{4.66}$$

The maximum posterior solution is obtained by setting the derivative of the log-posterior function, $\ln f_{\Theta|Y}(\theta|y)$, with respect to θ to zero:

$$\hat{\theta}_{MAP}(y) = \frac{\sigma_\theta^2}{\sigma_\theta^2 + \sigma_n^2/N}\,\bar{y} + \frac{\sigma_n^2/N}{\sigma_\theta^2 + \sigma_n^2/N}\,\mu_\theta \tag{4.67}$$

where $\bar{y} = \sum_{m=0}^{N-1} y(m)/N$.

Note that the MAP estimate is an interpolation between the ML estimate $\bar{y}$ and the mean of the prior pdf μ_θ, as shown in Figure 4.12. The expectation

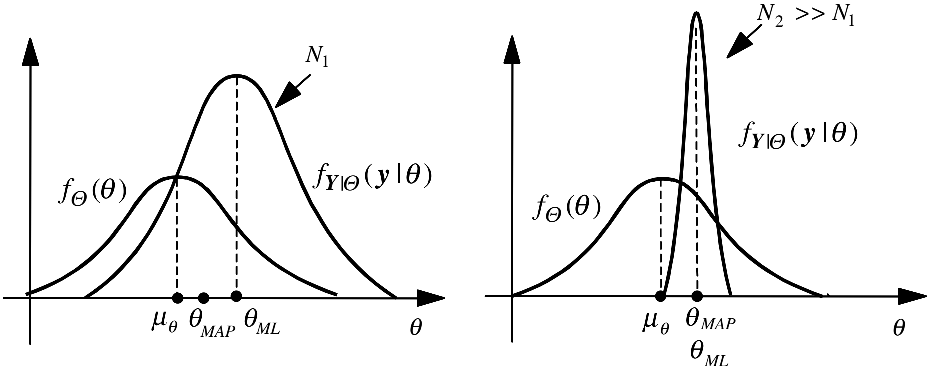


Figure 4.13 Illustration of the effect of increasing length of observation on the variance an estimator.

of the MAP estimate is obtained by noting that the only random variable on the right-hand side of Equation (4.67) is the term $\bar{y}$, and that $\mathcal{E}[\,\bar{y}\,]=\theta$

$$\mathcal{E}[\hat{\theta}_{MAP}(\,\boldsymbol{y})]=\frac{\sigma_\theta^2}{\sigma_\theta^2+\sigma_n^2/N}\theta+\frac{\sigma_n^2/N}{\sigma_\theta^2+\sigma_n^2/N}\mu_\theta \tag{4.68}$$

and the variance of the MAP estimate is given as

$$\text{Var}[\hat{\theta}_{MAP}(\,\boldsymbol{y})]=\frac{\sigma_\theta^2}{\sigma_\theta^2+\sigma_n^2/N}\times\text{Var}[\bar{y}]=\frac{\sigma_n^2/N}{1+\sigma_n^2/N\sigma_\theta^2} \tag{4.69}$$

Substitution of Equation (4.58) in Equation (4.67) yields

$$\text{Var}[\hat{\theta}_{MAP}(\,\boldsymbol{y})]=\frac{\text{Var}[\hat{\theta}_{ML}(\,\boldsymbol{y})]}{1+\text{Var}[\hat{\theta}_{ML}(\,\boldsymbol{y})]/\sigma_\theta^2} \tag{4.70}$$

Note that as σ_θ^2, the variance of the parameter θ, increases the influence of the prior decreases, and the variance of the MAP estimate tends towards the variance of the ML estimate.

4.2.7 The Relative Importance of the Prior and the Observation

A fundamental issue in the Bayesian inference method is the relative influence of the observation signal and the prior pdf on the outcome. The importance of the observation depends on the confidence in the observation, and the confidence in turn depends on the length of the observation and on the signal-to-noise ratio (SNR). In general, as the number of observation samples and the SNR increase, the variance of the estimate and the influence of the prior decrease. From Equation (4.67) for the estimation of a Gaussian distributed parameter observed in AWGN, as the length of the observation N increases, the importance of the prior decreases, and the MAP estimate tends to the ML estimate:

$$\underset{N\to\infty}{\text{limit}}\hat{\theta}_{MAP}(\boldsymbol{y})=\underset{N\to\infty}{\text{limit}}\left(\frac{\sigma_\theta^2}{\sigma_\theta^2+\sigma_n^2/N}\bar{y}+\frac{\sigma_n^2/N}{\sigma_\theta^2+\sigma_n^2/N}\mu_\theta\right)=\bar{y}=\hat{\theta}_{ML} \quad (4.71)$$

As illustrated in Figure 4.13, as the length of the observation N tends to infinity then both the MAP and the ML estimates of the parameter should tend to its true value θ.

Example 4.9 MAP estimation of a signal in additive noise. Consider the estimation of a scalar-valued Gaussian signal $x(m)$, observed in an additive Gaussian white noise $n(m)$, and modelled as

$$y(m)=x(m)+n(m) \quad (4.72)$$

The posterior pdf of the signal $x(m)$ is given by

$$\begin{aligned} f_{X|Y}(x(m)|y(m)) &= \frac{1}{f_Y(y(m))}f_{Y|X}(y(m)|x(m))f_X(x(m)) \\ &= \frac{1}{f_Y(y(m))}f_N(y(m)-x(m))f_X(x(m)) \end{aligned} \quad (4.73)$$

where $f_X(x(m))=\mathcal{N}(x(m),\mu_x,\sigma_x^2)$ and $f_N(n(m))=\mathcal{N}(n(m),\mu_n,\sigma_n^2)$ are the Gaussian pdfs of the signal and noise respectively. Substitution of the signal and noise pdfs in Equation (4.73) yields

$$\begin{aligned} f_{X|Y}(x(m)\,|\,y(m)) = &\frac{1}{f_Y(y(m))}\frac{1}{\sqrt{2\pi}\sigma_n}\exp\left\{-\frac{[y(m)-x(m)-\mu_n]^2}{2\sigma_n^2}\right\} \\ &\times\frac{1}{\sqrt{2\pi}\sigma_x}\exp\left\{-\frac{[x(m)-\mu_x]^2}{2\sigma_x^2}\right\} \end{aligned} \quad (4.74)$$

This equation can be rewritten as

$$f_{X|Y}(x(m)\,|\,y(m)) = \frac{1}{f_Y(y(m))}\frac{1}{2\pi\sigma_n\sigma_x}\exp\left\{-\frac{\sigma_x^2[y(m)-x(m)-\mu_n]^2+\sigma_n^2[x(m)-\mu_x]^2}{2\sigma_x^2\sigma_n^2}\right\} \tag{4.75}$$

To obtain the MAP estimate we set the derivative of the log-likelihood function $\ln f_{X|Y}(x(m)\,|\,y(m))$ with respect to $x(m)$ to zero as

$$\frac{\partial[\ln f_{X|Y}(x(m)\,|\,y(m))]}{\partial\hat{x}(m)} = -\frac{-2\sigma_x^2(y(m)-x(m)-\mu_n)+2\sigma_n^2(x(m)-\mu_x)}{2\sigma_x^2\sigma_n^2}=0 \tag{4.76}$$

From Equation (4.76) the MAP signal estimate is given by

$$\hat{x}(m) = \frac{\sigma_x^2}{\sigma_x^2+\sigma_n^2}[y(m)-\mu_n] + \frac{\sigma_n^2}{\sigma_x^2+\sigma_n^2}\mu_x \tag{4.77}$$

Note that the estimate $\hat{x}(m)$ is a weighted linear interpolation between the unconditional mean of $x(m)$, μ_x, and the observed value $(y(m)–\mu_n)$. At a very poor SNR i.e. when $\sigma_x^2 << \sigma_n^2$ we have $\hat{x}(m) \approx \mu_x$; and, on the other hand, for a noise-free signal $\sigma_n^2 = 0$ and $\mu_n = 0$ and we have $\hat{x}(m) = y(m)$.

Example 4.10 *MAP estimate of a Gaussian–AR process observed in AWGN.* Consider a vector of N samples $\boldsymbol{x}$ from an autoregressive (AR) process observed in an additive Gaussian noise, and modelled as

$$\boldsymbol{y} = \boldsymbol{x} + \boldsymbol{n} \tag{4.78}$$

From Chapter 8, a vector $\boldsymbol{x}$ from an AR process may be expressed as

$$\boldsymbol{e} = \boldsymbol{A}\boldsymbol{x} \tag{4.79}$$

where $\boldsymbol{A}$ is a matrix of the AR model coefficients, and the vector $\boldsymbol{e}$ is the input signal of the AR model. Assuming that the signal $\boldsymbol{x}$ is Gaussian, and that the P initial samples $\boldsymbol{x}_0$ are known, the pdf of the signal $\boldsymbol{x}$ is given by

$$f_X(\boldsymbol{x} \mid \boldsymbol{x}_0) = f_E(\boldsymbol{e}) = \frac{1}{\left(2\pi\sigma_e^2\right)^{N/2}} \exp\left(-\frac{1}{2\sigma_e^2} \boldsymbol{x}^{\mathrm{T}} \boldsymbol{A}^{\mathrm{T}} \boldsymbol{A}\boldsymbol{x} \right) \tag{4.80}$$

where it is assumed that the input signal $\boldsymbol{e}$ of the AR model is a zero-mean uncorrelated process with variance σ_e^2. The pdf of a zero-mean Gaussian noise vector $\boldsymbol{n}$, with covariance matrix $\boldsymbol{\Sigma}_{nn}$, is given by

$$f_N(\boldsymbol{n}) = \frac{1}{(2\pi)^{N/2} |\boldsymbol{\Sigma}_{nn}|^{1/2}} \exp\left(-\frac{1}{2} \boldsymbol{n}^{\mathrm{T}} \boldsymbol{\Sigma}_{nn}^{-1} \boldsymbol{n} \right) \tag{4.81}$$

From Bayes' rule, the pdf of the signal given the noisy observation is

$$f_{X|Y}(\boldsymbol{x} \mid \boldsymbol{y}) = \frac{f_{Y|X}(\boldsymbol{y}|\boldsymbol{x}) f_X(\boldsymbol{x})}{f_Y(\boldsymbol{y})} = \frac{1}{f_Y(\boldsymbol{y})} f_N(\boldsymbol{y} - \boldsymbol{x}) f_X(\boldsymbol{x}) \tag{4.82}$$

Substitution of the pdfs of the signal and noise in Equation (4.82) yields

$$f_{X|Y}(\boldsymbol{x} \mid \boldsymbol{y}) = \frac{1}{f_Y(\boldsymbol{y})(2\pi)^N \sigma_e^{N/2} |\boldsymbol{\Sigma}_{nn}|^{1/2}} \exp\left\{ -\frac{1}{2} \left[(\boldsymbol{y} - \boldsymbol{x})^{\mathrm{T}} \boldsymbol{\Sigma}_{nn}^{-1} (\boldsymbol{y} - \boldsymbol{x}) + \frac{\boldsymbol{x}^{\mathrm{T}} \boldsymbol{A}^{\mathrm{T}} \boldsymbol{A}\boldsymbol{x}}{\sigma_e^2} \right] \right\} \tag{4.83}$$

The MAP estimate corresponds to the minimum of the argument of the exponential function in Equation (4.83). Assuming that the argument of the exponential function is differentiable, and has a well-defined minimum, we can obtain the MAP estimate from

$$\hat{\boldsymbol{x}}_{MAP}(\boldsymbol{y}) = \underset{\boldsymbol{x}}{\arg\,\text{zero}} \left\{ \frac{\partial}{\partial \boldsymbol{x}} \left[(\boldsymbol{y} - \boldsymbol{x})^{\mathrm{T}} \boldsymbol{\Sigma}_{nn}^{-1} (\boldsymbol{y} - \boldsymbol{x}) + \frac{\boldsymbol{x}^{\mathrm{T}} \boldsymbol{A}^{\mathrm{T}} \boldsymbol{A}\boldsymbol{x}}{\sigma_e^2} \right] \right\} \tag{4.84}$$

The MAP estimate is

$$\hat{\boldsymbol{x}}_{MAP}(\boldsymbol{y}) = \left(\mathbf{I} + \frac{1}{\sigma_e^2} \boldsymbol{\Sigma}_{nn} \boldsymbol{A}^{\mathrm{T}} \boldsymbol{A} \right)^{-1} \boldsymbol{y} \tag{4.85}$$

where $\mathbf{I}$ is the identity matrix.

4.3 The Estimate–Maximise (EM) Method

The EM algorithm is an iterative likelihood maximisation method with applications in blind deconvolution, model-based signal interpolation, spectral estimation from noisy observations, estimation of a set of model parameters from a training data set, etc. The EM is a framework for solving problems where it is difficult to obtain a direct ML estimate either because the data is incomplete or because the problem is difficult.

To define the term *incomplete data,* consider a signal $\boldsymbol{x}$ from a random process X with an unknown parameter vector $\boldsymbol{\theta}$ and a pdf $f_{X;\Theta}(\boldsymbol{x};\boldsymbol{\theta})$. The notation $f_{X;\Theta}(\boldsymbol{x};\boldsymbol{\theta})$ expresses the dependence of the pdf of X on the value of the unknown parameter $\boldsymbol{\theta}$. The signal $\boldsymbol{x}$ is the so-called *complete data* and the ML estimate of the parameter vector $\boldsymbol{\theta}$ may be obtained from $f_{X;\Theta}(\boldsymbol{x};\boldsymbol{\theta})$. Now assume that the signal $\boldsymbol{x}$ goes through a many-to-one non-invertible transformation (e.g. when a number of samples of the vector $\boldsymbol{x}$ are lost) and is observed as $\boldsymbol{y}$. The observation $\boldsymbol{y}$ is the so-called incomplete data.
Maximisation of the likelihood of the incomplete data, $f_{Y;\Theta}(\boldsymbol{y};\boldsymbol{\theta})$, with respect to the parameter vector $\boldsymbol{\theta}$ is often a difficult task, whereas maximisation of the likelihood of the complete data $f_{X;\Theta}(\boldsymbol{x};\boldsymbol{\theta})$ is relatively easy. Since the complete data is unavailable, the parameter estimate is obtained through maximisation of the *conditional expectation* of the log-likelihood of the complete data defined as

$$\mathcal{E}[\ln f_{X;\Theta}(\boldsymbol{x};\boldsymbol{\theta})|\,\boldsymbol{y}] = \int_X f_{X|Y;\Theta}(\boldsymbol{x}|\boldsymbol{y};\boldsymbol{\theta}) \ln f_{X;\Theta}(\boldsymbol{x};\boldsymbol{\theta})\,d\boldsymbol{x} \tag{4.86}$$

In Equation (4.86), the computation of the term $f_{X|Y;\Theta}(\boldsymbol{x}|\boldsymbol{y};\boldsymbol{\theta})$ requires an estimate of the unknown parameter vector $\boldsymbol{\theta}$. For this reason, the expectation of the likelihood function is maximised iteratively starting with an initial estimate of $\boldsymbol{\theta}$, and updating the estimate as described in the following.

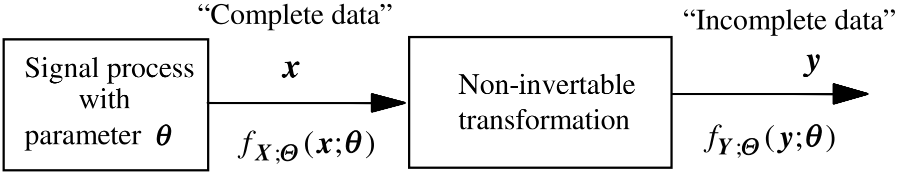

"Complete data"
"Incomplete data"
Signal process with parameter $\boldsymbol{\theta}$
$\boldsymbol{x}$
$f_{X;\Theta}(\boldsymbol{x};\boldsymbol{\theta})$
Non-invertable transformation
$\boldsymbol{y}$
$f_{Y;\Theta}(\boldsymbol{y};\boldsymbol{\theta})$

Figure 4.14 Illustration of transformation of complete data to incomplete data.

EM Algorithm

Step 1: Initialisation Select an initial parameter estimate $\boldsymbol{\theta}_0$, and
for $i = 0, 1, \ldots$ until convergence:

Step 2: Expectation Compute

$$\begin{aligned} U(\boldsymbol{\theta},\hat{\boldsymbol{\theta}}_i) &= E[\ln f_{X;\Theta}(\boldsymbol{x};\boldsymbol{\theta}) \mid \boldsymbol{y};\hat{\boldsymbol{\theta}}_i) \\ &= \int_X f_{X|Y;\Theta}(\boldsymbol{x} \mid \boldsymbol{y};\hat{\boldsymbol{\theta}}_i) \ln f_{X;\Theta}(\boldsymbol{x};\boldsymbol{\theta})\, d\boldsymbol{x} \end{aligned} \tag{4.87}$$

Step 3: Maximisation Select

$$\hat{\boldsymbol{\theta}}_{i+1} = \underset{\boldsymbol{\theta}}{\arg\max}\, U(\boldsymbol{\theta},\hat{\boldsymbol{\theta}}_i) \tag{4.88}$$

Step 4: Convergence test If not converged then go to Step 2.

4.3.1 Convergence of the EM Algorithm

In this section, it is shown that the EM algorithm converges to a maximum of the likelihood of the incomplete data $f_{Y;\Theta}(\boldsymbol{y};\boldsymbol{\theta})$. The likelihood of the complete data can be written as

$$f_{X,Y;\Theta}(\boldsymbol{x},\boldsymbol{y};\boldsymbol{\theta}) = f_{X|Y;\Theta}(\boldsymbol{x}|\boldsymbol{y};\boldsymbol{\theta})\, f_{Y;\Theta}(\boldsymbol{y};\boldsymbol{\theta}) \tag{4.89}$$

where $f_{X,Y;\Theta}(\boldsymbol{x},\boldsymbol{y};\boldsymbol{\theta})$ is the likelihood of $\boldsymbol{x}$ and $\boldsymbol{y}$ with $\boldsymbol{\theta}$ as a parameter. From Equation (4.89), the log-likelihood of the incomplete data is obtained as

$$\ln f_{Y;\Theta}(\boldsymbol{y};\boldsymbol{\theta}) = \ln f_{X,Y;\Theta}(\boldsymbol{x},\boldsymbol{y};\boldsymbol{\theta}) - \ln f_{X|Y;\Theta}(\boldsymbol{x}|\,\boldsymbol{y};\boldsymbol{\theta}) \tag{4.90}$$

Using an estimate $\hat{\boldsymbol{\theta}}_i$ of the parameter vector $\boldsymbol{\theta}$, and taking the expectation of Equation (4.90) over the space of the complete signal $\boldsymbol{x}$, we obtain

$$\ln f_{Y;\Theta}(\boldsymbol{y};\boldsymbol{\theta}) = U(\boldsymbol{\theta};\hat{\boldsymbol{\theta}}_i) - V(\boldsymbol{\theta};\hat{\boldsymbol{\theta}}_i) \tag{4.91}$$

where for a given $\boldsymbol{y}$, the expectation of $\ln f_{Y;\Theta}(\boldsymbol{y};\boldsymbol{\theta})$ is itself, and the function $U(\boldsymbol{\theta};\hat{\boldsymbol{\theta}})$ is the conditional expectation of $\ln f_{X,Y;\Theta}(\boldsymbol{x},\boldsymbol{y};\boldsymbol{\theta})$:

$$\begin{aligned} U(\boldsymbol{\theta},\hat{\boldsymbol{\theta}}_i) &= \mathcal{E}[\ln f_{\boldsymbol{X},\boldsymbol{Y};\Theta}(\boldsymbol{x},\boldsymbol{y};\boldsymbol{\theta}) \mid \boldsymbol{y};\hat{\boldsymbol{\theta}}_i) \\ &= \int\limits_{\boldsymbol{X}} f_{\boldsymbol{X}|\boldsymbol{Y};\Theta}(\boldsymbol{x} \mid \boldsymbol{y};\hat{\boldsymbol{\theta}}_i) \ln f_{\boldsymbol{X};\Theta}(\boldsymbol{x};\boldsymbol{\theta})\, d\boldsymbol{x} \end{aligned} \tag{4.92}$$

The function $V(\boldsymbol{\theta},\hat{\boldsymbol{\theta}})$ is the conditional expectation of $\ln f_{\boldsymbol{X}|\boldsymbol{Y};\Theta}(\boldsymbol{x}|\boldsymbol{y};\boldsymbol{\theta})$:

$$\begin{aligned} V(\boldsymbol{\theta},\hat{\boldsymbol{\theta}}_i) &= \mathcal{E}\Big[\ln f_{\boldsymbol{X}|\boldsymbol{Y};\Theta}(\boldsymbol{x}|\boldsymbol{y};\boldsymbol{\theta})\Big|\boldsymbol{y};\hat{\boldsymbol{\theta}}_i\Big] \\ &= \int\limits_{\boldsymbol{X}} f_{\boldsymbol{X}|\boldsymbol{Y};\Theta}(\boldsymbol{x}|\boldsymbol{y};\hat{\boldsymbol{\theta}}_i) \ln f_{\boldsymbol{X}|\boldsymbol{Y};\Theta}(\boldsymbol{x}|\boldsymbol{y};\boldsymbol{\theta})\, d\boldsymbol{x} \end{aligned} \tag{4.93}$$

Now, from Equation (4.91), the log-likelihood of the incomplete data $\boldsymbol{y}$ with parameter estimate $\hat{\boldsymbol{\theta}}_i$ at iteration i is

$$\ln f_{\boldsymbol{Y};\Theta}(\boldsymbol{y};\hat{\boldsymbol{\theta}}_i) = U(\hat{\boldsymbol{\theta}}_i;\hat{\boldsymbol{\theta}}_i) - V(\hat{\boldsymbol{\theta}}_i;\hat{\boldsymbol{\theta}}_i) \tag{4.94}$$

It can be shown (see Dempster et al., 1977) that the function V satisfies the inequality

$$V(\hat{\boldsymbol{\theta}}_{i+1};\hat{\boldsymbol{\theta}}_i) \le V(\hat{\boldsymbol{\theta}}_i;\hat{\boldsymbol{\theta}}_i) \tag{4.95}$$

and in the maximisation step of EM we choose $\hat{\boldsymbol{\theta}}_{i+1}$ such that

$$U(\hat{\boldsymbol{\theta}}_{i+1};\hat{\boldsymbol{\theta}}_i) \ge U(\hat{\boldsymbol{\theta}}_i;\hat{\boldsymbol{\theta}}_i) \tag{4.96}$$

From Equation (4.94) and the inequalities (4.95) and (4.96), it follows that

$$\ln f_{\boldsymbol{Y};\Theta}(\boldsymbol{y};\hat{\boldsymbol{\theta}}_{i+1}) \ge \ln f_{\boldsymbol{Y};\Theta}(\boldsymbol{y};\hat{\boldsymbol{\theta}}_i) \tag{4.97}$$

Therefore at every iteration of the EM algorithm, the conditional likelihood of the estimate increases until the estimate converges to a local maximum of the log-likelihood function $\ln f_{\boldsymbol{Y};\Theta}(\boldsymbol{y};\boldsymbol{\theta})$.

The EM algorithm is applied to the solution of a number of problems in this book. In Section 4.5, of this chapter the estimation of the parameters of a mixture Gaussian model for the signal space of a recorded process is formulated in an EM framework. In Chapter 5, the EM is used for estimation of the parameters of a hidden Markov model.

4.4 Cramer–Rao Bound on the Minimum Estimator Variance

An important measure of the performance of an estimator is the variance of the estimate with the varying values of the observation signal $\boldsymbol{y}$ and the parameter vector $\boldsymbol{\theta}$. The minimum estimation variance depends on the distributions of the parameter vector $\boldsymbol{\theta}$ and on the observation signal $\boldsymbol{y}$. In this section, we first consider the lower bound on the variance of the estimates of a constant parameter, and then extend the results to random parameters.

The Cramer–Rao lower bound on the variance of estimate of the i^{th} coefficient θ_i of a parameter vector $\boldsymbol{\theta}$ is given as

$$\operatorname{Var}[\hat{\theta}_i(\boldsymbol{y})] \geq \frac{\left(1+\dfrac{\partial \theta_{\text{Bias}}}{\partial \theta_i}\right)^2}{\mathcal{E}\left[\left(\dfrac{\partial \ln f_{Y|\Theta}(\boldsymbol{y}\,|\,\boldsymbol{\theta})}{\partial \theta_i}\right)^2\right]} \tag{4.98}$$

An estimator that achieves the lower bound on the variance is called the minimum variance, or the most efficient, estimator.

Proof The bias in the estimate $\hat{\theta}_i(\boldsymbol{y})$ of the i^{th} coefficient of the parameter vector $\boldsymbol{\theta}$, averaged over the observation space $\boldsymbol{Y}$, is defined as

$$\mathcal{E}\,[\hat{\theta}_i(\boldsymbol{y}) - \theta_i] = \int_{-\infty}^{\infty}[\hat{\theta}_i(\boldsymbol{y}) - \theta_i]\, f_{Y|\Theta}(\boldsymbol{y}|\boldsymbol{\theta})\; d\boldsymbol{y} = \theta_{Bias} \tag{4.99}$$

Differentiation of Equation (4.99) with respect to θ_i yields

$$\int_{-\infty}^{\infty}\left\{[\hat{\theta}_i(\boldsymbol{y})-\theta_i]\frac{\partial f_{Y|\Theta}(\boldsymbol{y}\,|\,\boldsymbol{\theta})}{\partial \theta_i} - f_{Y|\Theta}(\boldsymbol{y}\,|\,\boldsymbol{\theta})\right\}d\boldsymbol{y} = \frac{\partial \theta_{\text{Bias}}}{\partial \theta_i} \tag{4.100}$$

For a probability density function we have

$$\int_{-\infty}^{\infty} f_{Y|\Theta}(\boldsymbol{y}|\boldsymbol{\theta})\; d\boldsymbol{y} = 1 \tag{4.101}$$

Therefore Equation (4.100) can be written as

$$\int_{-\infty}^{\infty} [\hat{\theta}_i(\boldsymbol{y})\theta_i]\frac{\partial f_{\boldsymbol{Y}|\boldsymbol{\Theta}}(\boldsymbol{y}\,|\,\boldsymbol{\theta})}{\partial\theta_i}d\boldsymbol{y} = 1+\frac{\partial\theta_{\text{Bias}}}{\partial\theta_i} \tag{4.102}$$

Now, since the derivative of the integral of a pdf is zero, taking the derivative of Equation (4.101) and multiplying the result by θ_{Bias} yields

$$\theta_{\text{Bias}}\int_{-\infty}^{\infty} \frac{\partial f_{\boldsymbol{Y}|\boldsymbol{\Theta}}(\boldsymbol{y}\,|\,\boldsymbol{\theta})}{\partial\theta_i}\,d\boldsymbol{y} = 0 \tag{4.103}$$

Substituting $\partial f_{\boldsymbol{Y}|\boldsymbol{\Theta}}(\boldsymbol{y}\,|\,\boldsymbol{\theta})/\partial\theta_i = f_{\boldsymbol{Y}|\boldsymbol{\Theta}}(\boldsymbol{y}\,|\,\boldsymbol{\theta})\partial\ln f_{\boldsymbol{Y}|\boldsymbol{\Theta}}(\boldsymbol{y}\,|\,\boldsymbol{\theta})/\partial\theta_i$ into Equation (4.102), and using Equation (4.103), we obtain

$$\int_{-\infty}^{\infty} [\hat{\theta}_i(\boldsymbol{y})-\theta_{\text{Bias}}-\theta_i]\frac{\partial\ln f_{\boldsymbol{Y}|\boldsymbol{\Theta}}(\boldsymbol{y}\,|\,\boldsymbol{\theta})}{\partial\theta_i}f_{\boldsymbol{Y}|\boldsymbol{\Theta}}(\boldsymbol{y}\,|\,\boldsymbol{\theta})\,d\boldsymbol{y} = 1+\frac{\partial\theta_{\text{Bias}}}{\partial\theta_i} \tag{4.104}$$

Now squaring both sides of Equation (4.104), we obtain

$$\left(\int_{-\infty}^{\infty} [\hat{\theta}_i(\boldsymbol{y})-\theta_{\text{Bias}}-\theta_i]\frac{\partial\ln f_{\boldsymbol{Y}|\boldsymbol{\Theta}}(\boldsymbol{y}\,|\,\boldsymbol{\theta})}{\partial\theta_i}f_{\boldsymbol{Y}|\boldsymbol{\Theta}}(\boldsymbol{y}\,|\,\boldsymbol{\theta})\,d\boldsymbol{y}\right)^2 = \left(1+\frac{\partial\theta_{\text{Bias}}}{\partial\theta_i}\right)^2 \tag{4.105}$$

For the left-hand side of Equation (4.105) application of the following Schwartz inequality

$$\left(\int_{-\infty}^{\infty} f(y)g(y)\,dx\right)^2 \le \int_{-\infty}^{\infty}(f(y))^2\,dx \times \int_{-\infty}^{\infty}(g(y))^2\,dy \tag{4.106}$$

yields

$$\left\{ \int_{-\infty}^{\infty} \left([\hat{\theta}_i(\boldsymbol{y}) - \theta_{\text{Bias}} - \theta_i]\, f_{Y|\Theta}^{1/2}(\boldsymbol{y}\,|\,\boldsymbol{\theta})\right)\left(\frac{\partial \ln f_{Y|\Theta}(\boldsymbol{y}\,|\,\boldsymbol{\theta})}{\partial \theta_i} f_{Y|\Theta}^{1/2}(\boldsymbol{y}\,|\,\boldsymbol{\theta})\right) dy \right\}^2 \le$$
$$\left\{\left(\int_{-\infty}^{\infty}\left([\hat{\theta}_i(\boldsymbol{y}) - \theta_{\text{Bias}} - \theta_i]^2\, f_{Y|\Theta}(\boldsymbol{y}\,|\,\boldsymbol{\theta})\right) dy\right)\right\}\left\{\int_{-\infty}^{\infty}\left(\frac{\partial \ln f_{Y|\Theta}(\boldsymbol{y}\,|\,\boldsymbol{\theta})}{\partial \theta_i}\right)^2 f_{Y|\Theta}(\boldsymbol{y}\,|\,\boldsymbol{\theta})\, dy\right\} \tag{4.107}$$

From Equations (4.105) and (4.107), we have

$$\text{Var}[\hat{\theta}_i(\boldsymbol{y})] \times \mathcal{E}\left[\left(\frac{\partial \ln f_{Y|\Theta}(\boldsymbol{y}\,|\,\boldsymbol{\theta})}{\partial \theta_i}\right)^2\right] \ge \left(1 + \frac{\partial \theta_{\text{Bias}}}{\partial \theta_i}\right)^2 \tag{4.108}$$

The Cramer–Rao inequality (4.98) results directly from the inequality (4.108).

4.4.1 Cramer–Rao Bound for Random Parameters

For random parameters the Cramer–Rao bound may be obtained using the same procedure as above, with the difference that in Equation (4.98) instead of the likelihood $f_{Y|\Theta}(\boldsymbol{y}|\boldsymbol{\theta})$ we use the joint pdf $f_{Y,\Theta}(\boldsymbol{y},\boldsymbol{\theta})$, and we also use the logarithmic relation

$$\frac{\partial \ln f_{Y,\Theta}(\boldsymbol{y},\boldsymbol{\theta})}{\partial \theta_i} = \frac{\partial \ln f_{Y|\Theta}(\boldsymbol{y}|\boldsymbol{\theta})}{\partial \theta_i} + \frac{\partial \ln f_{\Theta}(\boldsymbol{\theta})}{\partial \theta_i} \tag{4.109}$$

The Cramer–Rao bound for random parameters is obtained as

$$\text{Var}[\hat{\theta}_i(\boldsymbol{y})] \ge \frac{\left(1 + \dfrac{\partial \theta_{\text{Bias}}}{\partial \theta_i}\right)^2}{\mathcal{E}\left[\left(\dfrac{\partial \ln f_{Y|\Theta}(\boldsymbol{y}\,|\,\boldsymbol{\theta})}{\partial \theta_i}\right)^2 + \left(\dfrac{\partial \ln f_{\Theta}(\boldsymbol{\theta})}{\partial \theta_i}\right)^2\right]} \tag{4.110}$$

where the second term in the denominator of Equation (4.110) describes the effect of the prior pdf of $\boldsymbol{\theta}$. As expected the use of the prior, $f_{\Theta}(\boldsymbol{\theta})$, can result in a decrease in the variance of the estimate. An alternative form of the minimum bound on estimation variance can be obtained by using the likelihood relation

$$\mathcal{E}\left[\left(\frac{\partial \ln f_{\boldsymbol{Y},\Theta}(\boldsymbol{y},\boldsymbol{\theta})}{\partial \theta_i}\right)^2\right] = -\mathcal{E}\left[\frac{\partial^2 \ln f_{\boldsymbol{Y},\Theta}(\boldsymbol{y},\boldsymbol{\theta})}{\partial \theta_i^2}\right] \tag{4.111}$$

as

$$\operatorname{Var}[\hat{\theta}_i\,(\,\boldsymbol{y}\,)] \geq -\frac{\left(1+\dfrac{\partial \theta_{\text{Bias}}}{\partial \theta_i}\right)^2}{\mathcal{E}\left[\dfrac{\partial^2 \ln f_{\boldsymbol{Y}|\Theta}(\boldsymbol{y}\,|\,\boldsymbol{\theta})}{\partial \theta_i^2} + \dfrac{\partial^2 \ln f_{\Theta}(\boldsymbol{\theta})}{\partial \theta_i^2}\right]} \tag{4.112}$$

4.4.2 Cramer–Rao Bound for a Vector Parameter

For real-valued P-dimensional vector parameters, the Cramer–Rao bound for the covariance matrix of an unbiased estimator of $\boldsymbol{\theta}$ is given by

$$\operatorname{Cov}[\hat{\boldsymbol{\theta}}] \geq \boldsymbol{J}^{-1}(\boldsymbol{\theta}) \tag{4.113}$$

where $\boldsymbol{J}$ is the $P \times P$ Fisher information matrix, with elements given by

$$[\boldsymbol{J}(\boldsymbol{\theta})]_{ij} = -\,\mathcal{E}\left[\frac{\partial^2 \ln f_{\boldsymbol{Y},\Theta}(\boldsymbol{y},\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j}\right] \tag{4.114}$$

The lower bound on the variance of the i^{th} element of the vector $\boldsymbol{\theta}$ is given by

$$\operatorname{Var}(\hat{\theta}_i) \geq \left[\boldsymbol{J}^{-1}(\boldsymbol{\theta})\right]_{ii} = \frac{1}{\mathcal{E}\left[\dfrac{\partial^2 \ln f_{\boldsymbol{Y},\Theta}(\boldsymbol{y},\boldsymbol{\theta})}{\partial \theta_i^2}\right]} \tag{4.115}$$

where $(\boldsymbol{J}^{-1}(\boldsymbol{\theta})_{ii})$ is the i^{th} diagonal element of the inverse of the Fisher matrix.

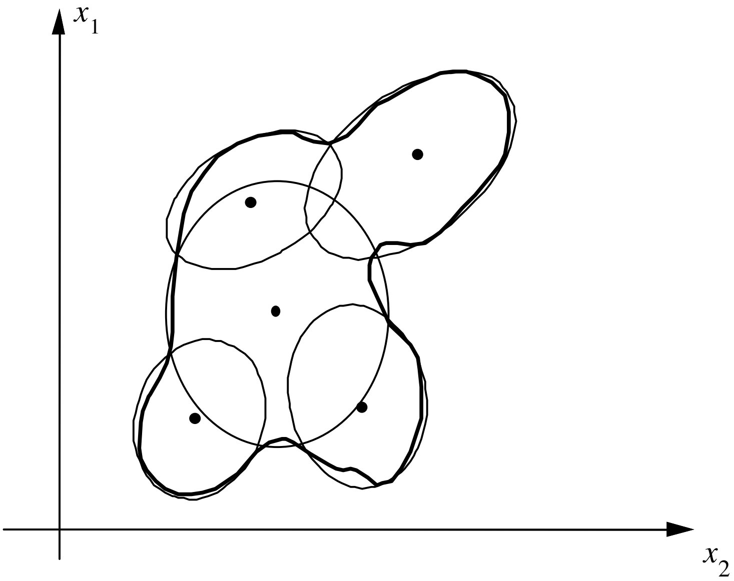


Figure 4.15 Illustration of probabilistic modelling of a two-dimensional signal space with a mixture of five bivariate Gaussian densities.

4.5 Design of Mixture Gaussian Models

A practical method for the modelling of the probability density function of an arbitrary signal space is to fit (or "tile") the space with a mixture of a number of Gaussian probability density functions. Figure 4.15 illustrates the modelling of a two-dimensional signal space with a number of circular and elliptically shaped Gaussian processes. Note that the Gaussian densities can be overlapping, with the result that in an area of overlap, a data point can be associated with different probabilities to different components of the Gaussian mixture.

A main advantage of the use of a mixture Gaussian model is that it results in mathematically tractable signal processing solutions. A mixture Gaussian pdf model for a process $\boldsymbol{X}$ is defined as

$$f_X(\boldsymbol{x}) = \sum_{k=1}^{K} P_k \mathcal{N}_k(\boldsymbol{x};\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{4.116}$$

where $\mathcal{N}_k(\boldsymbol{x};\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ denotes the k^{th} component of the mixture Gaussian pdf, with mean vector $\boldsymbol{\mu}_k$ and covariance matrix $\boldsymbol{\Sigma}_k$. The parameter P_k is the prior probability of the k^{th} mixture, and it can be interpreted as the expected fraction of the number of vectors from the process $\boldsymbol{X}$ associated with the k^{th} mixture.

In general, there are an infinite number of different K-mixture Gaussian densities that can be used to “tile up” a signal space. Hence the modelling of a signal space with a K-mixture pdf space can be regarded as a many-to-one mapping, and the expectation-maximisation (EM) method can be applied for the estimation of the parameters of the Gaussian pdf models.

4.5.1 The EM Algorithm for Estimation of Mixture Gaussian Densities

The EM algorithm, discussed in Section 4.4, is an iterative maximum-likelihood (ML) estimation method, and can be employed to calculate the parameters of a K-mixture Gaussian pdf model for a given data set. To apply the EM method we first need to define the so-called complete and incomplete data sets. As usual the observation vectors [$\boldsymbol{y}(m)$ m=0, ..., N–1] form the incomplete data. The complete data may be viewed as the observation vectors with a *label* attached to each vector $\boldsymbol{y}(m)$ to indicate the component of the mixture Gaussian model that generated the vector. Note that if each signal vector $\boldsymbol{y}(m)$ had a mixture component label attached, then the computation of the mean vector and the covariance matrix of each component of the mixture would be a relatively simple exercise. Therefore the complete and incomplete data can be defined as follows:

The incomplete data $\boldsymbol{y}(m),\ m=0,\dots,N-1$

The complete data $\boldsymbol{x}(m)=[\boldsymbol{y}(m),k]=\boldsymbol{y}_k(m),\ m=0,\dots,N-1, k\in(1,\dots,K)$

The probability of the complete data is the probability that an observation vector $\boldsymbol{y}(m)$ has a label k associating it with the k^{th} component of the mixture density. The main step in application of the EM method is to define the expectation of the complete data, given the observations and a current estimate of the parameter vector, as

$$
\begin{aligned}
U(\boldsymbol{\Theta},\hat{\boldsymbol{\Theta}}_\imath) &= \mathcal{E}[\ln f_{\boldsymbol{Y},K;\boldsymbol{\Theta}}(\boldsymbol{y}(m),k;\boldsymbol{\Theta}) \,|\, \boldsymbol{y}(m);\hat{\boldsymbol{\Theta}}_i] \\
&= \sum_{m=0}^{N-1}\sum_{k=0}^{K} \frac{f_{\boldsymbol{Y},K|\boldsymbol{\Theta}}(\boldsymbol{y}(m),k) \,|\, \hat{\boldsymbol{\Theta}}_i)}{f_{\boldsymbol{Y}|\boldsymbol{\Theta}}(\boldsymbol{y}(m) \,|\, \hat{\boldsymbol{\Theta}}_i)} \ln f_{\boldsymbol{Y},K;\boldsymbol{\Theta}}(\boldsymbol{y}(m),k;\boldsymbol{\Theta})
\end{aligned}
\tag{4.117}
$$

where $\boldsymbol{\Theta}=\{\boldsymbol{\theta}_k=[P_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k], k=1,\dots, K\}$, are the parameters of the Gaussian mixture as in Equation (4.116). Now the joint pdf of $\boldsymbol{y}(m)$ and the k^{th} Gaussian component of the mixture density can be written as

$$\begin{aligned} f_{\boldsymbol{Y},K|\Theta}\left(\boldsymbol{y}(m),k\middle|\hat{\boldsymbol{\theta}}_i\right)&=P_{k_i}\, f_k\left(\boldsymbol{y}(m)\middle|\hat{\boldsymbol{\theta}}_{k_i}\right)\\ &=P_{k_i}\,\mathcal{N}_k\left(\boldsymbol{y}(m);\hat{\boldsymbol{\mu}}_{k_i},\hat{\boldsymbol{\Sigma}}_{k_i}\right) \end{aligned} \tag{4.118}$$

where $\mathcal{N}_k\left(\boldsymbol{y}(m);\hat{\boldsymbol{\mu}}_k,\hat{\boldsymbol{\Sigma}}_k\right)$ is a Gaussian density with mean vector $\boldsymbol{\mu}_k$ and covariance matrix $\boldsymbol{\Sigma}_k$:

$$\mathcal{N}_k\left(\boldsymbol{y}(m);\boldsymbol{\mu}_k,\,\boldsymbol{\Sigma}_k\right)=\frac{1}{(2\pi)^{P/2}|\boldsymbol{\Sigma}_k|^{1/2}}\exp\left(-\frac{1}{2}\left(\boldsymbol{y}(m)-\boldsymbol{\mu}_k\right)^T\boldsymbol{\Sigma}_k^{-1}\left(\boldsymbol{y}(m)-\boldsymbol{\mu}_k\right)\right) \tag{4.119}$$

The pdf of $\boldsymbol{y}(m)$ as a mixture of K Gaussian densities is given by

$$\begin{aligned} f_{\boldsymbol{Y}|\boldsymbol{\theta}}\left(\boldsymbol{y}(m)\middle|\hat{\boldsymbol{\theta}}_i\right)&=\mathcal{N}\left(\boldsymbol{y}(m)\middle|\hat{\boldsymbol{\theta}}_i\right)\\ &=\sum_{k=1}^{K}\hat{P}_{k_i}\,\mathcal{N}_k\left(\boldsymbol{y}(m);\hat{\boldsymbol{\mu}}_{k_i},\hat{\boldsymbol{\Sigma}}_{k_i}\right) \end{aligned} \tag{4.120}$$

Substitution of the Gaussian densities of Equation (4.118) and Equation (4.120) in Equation (4.117) yields

$$U[(\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{P}),(\hat{\boldsymbol{\mu}}_i,\hat{\boldsymbol{\Sigma}}_i,\hat{\boldsymbol{P}}_i)]=\sum_{m=0}^{N-1}\sum_{k=1}^{K}\frac{\hat{P}_{k_i}\mathcal{N}_k(\boldsymbol{y}(m);\hat{\boldsymbol{\mu}}_{k_i},\,\hat{\boldsymbol{\Sigma}}_{k_i})}{\mathcal{N}\left(\boldsymbol{y}(m)\middle|\hat{\Theta}_i\right)}\ln[P_k\mathcal{N}_k(\boldsymbol{y}(m);\boldsymbol{\mu}_k,\,\boldsymbol{\Sigma}_k)]$$

$$=\sum_{m=0}^{N-1}\sum_{k=1}^{K}\left(\frac{\hat{P}_{k_i}\mathcal{N}_k(\boldsymbol{y}(m);\hat{\boldsymbol{\mu}}_{k_i},\,\hat{\boldsymbol{\Sigma}}_{k_i})}{\mathcal{N}\left(\boldsymbol{y}(m)\middle|\hat{\Theta}_i\right)}\ln P_k+\frac{\hat{P}_{k_i}\mathcal{N}_k(\boldsymbol{y}(m);\hat{\boldsymbol{\mu}}_{k_i},\,\hat{\boldsymbol{\Sigma}}_{k_i})}{\mathcal{N}\left(\boldsymbol{y}(m)\middle|\hat{\Theta}_i\right)}\ln\mathcal{N}_k(\boldsymbol{y}_k;\boldsymbol{\mu}_k,\,\boldsymbol{\Sigma}_k)\right) \tag{4.121}$$

Equation (4.121) is maximised with respect to the parameter P_k using the constrained optimisation method. This involves subtracting the constant term $\Sigma P_k=1$ from the right hand side of Equation (4.121) and then setting the derivative of this equation with respect to P_k to zero, this yields

$$\begin{aligned}\hat{P}_{k_{i+1}} &= \arg\max_{P_k} \; U[(\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{P}),(\hat{\boldsymbol{\mu}}_i,\hat{\boldsymbol{\Sigma}}_i,\hat{\boldsymbol{P}}_i)] \\ &= \frac{1}{N}\sum_{m=0}^{N-1} \frac{\hat{P}_{k_i}\mathcal{N}_k(\boldsymbol{y}(m);\hat{\boldsymbol{\mu}}_{k_i},\hat{\boldsymbol{\Sigma}}_{k_i})}{\mathcal{N}\left(\boldsymbol{y}(m)\middle|\hat{\boldsymbol{\Theta}}_i\right)}\end{aligned} \tag{4.122}$$

The parameters $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ that maximise the function U are obtained, by setting the derivative of the function with respect to these parameters to zero:

$$\begin{aligned}\hat{\boldsymbol{\mu}}_{k_{i+1}} &= \arg\max_{\boldsymbol{\mu}_k} \; U[(\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{P}),(\hat{\boldsymbol{\mu}}_i,\hat{\boldsymbol{\Sigma}}_i,\hat{\boldsymbol{P}}_i)] \\ &= \frac{\displaystyle\sum_{m=0}^{N-1} \frac{\hat{P}_{k_i}\mathcal{N}_k(\boldsymbol{y}(m);\hat{\boldsymbol{\mu}}_{k_i},\hat{\boldsymbol{\Sigma}}_{k_i})}{\mathcal{N}\left(\boldsymbol{y}(m)\middle|\hat{\boldsymbol{\Theta}}_i\right)}\,\boldsymbol{y}(m)}{\displaystyle\sum_{m=0}^{N-1} \frac{\hat{P}_{k_i}\mathcal{N}_k(\boldsymbol{y}(m);\hat{\boldsymbol{\mu}}_{k_i},\hat{\boldsymbol{\Sigma}}_{k_i})}{\mathcal{N}\left(\boldsymbol{y}(m)\middle|\hat{\boldsymbol{\Theta}}_i\right)}}\end{aligned} \tag{4.123}$$

and

$$\begin{aligned}\hat{\boldsymbol{\Sigma}}_{k_{i+1}} &= \arg\max_{\boldsymbol{\Sigma}_k} \; U[(\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{P}),(\hat{\boldsymbol{\mu}}_i,\hat{\boldsymbol{\Sigma}}_i,\hat{\boldsymbol{P}}_i)] \\ &= \frac{\displaystyle\sum_{m=0}^{N-1} \frac{\hat{P}_{k_i}\mathcal{N}_k(\boldsymbol{y}(m);\hat{\boldsymbol{\mu}}_{k_i},\hat{\boldsymbol{\Sigma}}_{k_i})}{\mathcal{N}\left(\boldsymbol{y}(m)\middle|\hat{\boldsymbol{\Theta}}_i\right)}\,(\boldsymbol{y}(m)-\hat{\boldsymbol{\mu}}_{k_i})(\boldsymbol{y}(m)-\hat{\boldsymbol{\mu}}_{k_i})^T}{\displaystyle\sum_{m=0}^{N-1} \frac{\hat{P}_{k_i}\mathcal{N}_k(\boldsymbol{y}(m);\hat{\boldsymbol{\mu}}_{k_i},\hat{\boldsymbol{\Sigma}}_{k_i})}{\mathcal{N}\left(\boldsymbol{y}(m)\middle|\hat{\boldsymbol{\Theta}}_i\right)}}\end{aligned} \tag{4.124}$$

Equations (4.122)–(4.124) are the estimates of the parameters of a mixture Gaussian pdf model. These equations can be used in further iterations of the EM method until the parameter estimates converge.

4.6 Bayesian Classification

Classification is the processing and *labelling* of an observation sequence $\{\boldsymbol{y}(m)\}$ with one of M classes of signals $\{C_k;\ k=1, ..., M\}$ that could have generated the observation. Classifiers are present in all modern digital communication systems and in applications such as the decoding of

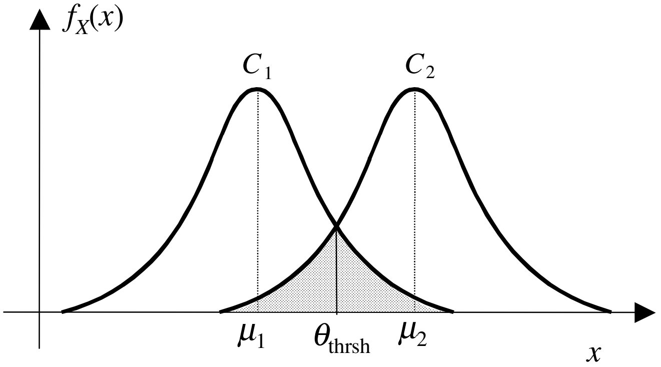


Figure 4.16 – Illustration of the overlap of the distribution of two classes of signals.

discrete-valued symbols in digital communication receivers, speech compression, video compression, speech recognition, image recognition, character recognition, signal/noise classification and detectors. For example, in an M-symbol digital communication system, the channel output signal is classified as one of the M signalling symbols; in speech recognition, segments of speech signals are labelled with one of about 40 elementary phonemes sounds; and in speech or video compression, a segment of speech samples or a block of image pixels are quantised and labelled with one of a number of prototype signal vectors in a codebook. In the design of a classifier, the aim is to reduce the classification error given the constraints on the signal-to-noise ratio, the bandwidth and the computational resources.

Classification errors are due to overlap of the distributions of different classes of signals. This is illustrated in Figure 4.16 for a binary classification problem with two Gaussian distributed signal classes C_1 and C_2. In the shaded region, where the signal distributions overlap, a sample x could belong to either of the two classes. The shaded area gives a measure of the classification error. The obvious solution suggested by Figure 4.16 for reducing the classification error is to reduce the overlap of the distributions. The overlap can be reduced in two ways: (a) by increasing the distance between the mean values of different classes, and (b) by reducing the variance of each class. In telecommunication systems the overlap between the signal classes is reduced using a combination of several methods including increasing the signal-to-noise ratio, increasing the distance between signal patterns by adding redundant error control coding bits, and signal shaping and post-filtering operations. In pattern recognition, where it is not possible to control the signal generation process (as in speech and image recognition), the choice of the pattern features and models affects the classification error. The design of an efficient classification for pattern recognition depends on a number of factors, which can be listed as follows:

(1) Extraction and transformation of a set of discriminative features from the signal that can aid the classification process. The features need to adequately characterise each class and emphasise the difference between various classes.
(2) Statistical modelling of the observation features for each class. For Bayesian classification, a posterior probability model for each class should be obtained.
(3) Labelling of an unlabelled signal with one of the N classes.

4.6.1 Binary Classification

The simplest form of classification is the labelling of an observation with one of two classes of signals. Figures 4.17(a) and 4.17(b) illustrate two examples of a simple binary classification problem in a two-dimensional signal space. In each case, the observation is the result of a random mapping (e.g. signal plus noise) from the binary source to the continuous observation space. In Figure 4.17(a), the binary sources and the observation space associated with each source are well separated, and it is possible to make an error-free classification of each observation. In Figure 4.17(b) there is less distance between the mean of the sources, and the observation signals have a greater spread. This results in some overlap of the signal spaces and classification error can occur. In binary classification, a signal $\boldsymbol{x}$ is labelled with the class that scores the higher a posterior probability:

$$P_{C|X}\left(C_1\middle|\boldsymbol{x}\right) \underset{C_2}{\overset{C_1}{\gtrless}} P_{C|X}\left(C_2\middle|\boldsymbol{x}\right) \tag{4.125}$$

Using Bayes' rule Equation (4.125) can be rewritten as

$$P_C(C_1)f_{X|C}\left(\boldsymbol{x}\middle|C_1\right) \underset{C_2}{\overset{C_1}{\gtrless}} P_C(C_2)f_{X|C}\left(\boldsymbol{x}\middle|C_2\right) \tag{4.126}$$

Letting $P_C(C_1)=P_1$ and $P_C(C_2)=P_2$, Equation (4.126) is often written in terms of a *likelihood ratio test* as

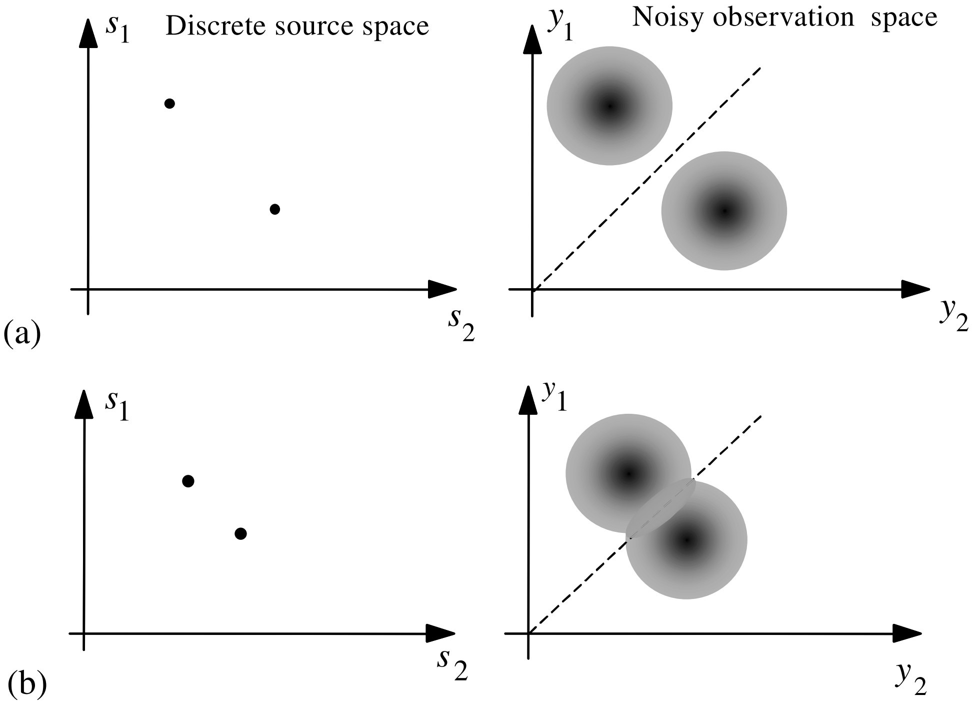


Figure 4.17 Illustration of binary classification: (a) the source and observation spaces are well separated, (b) the observation spaces overlap.

$$\frac{f_{X|C}(\boldsymbol{x}|C_1)}{f_{X|C}(\boldsymbol{x}|C_2)} \underset{C_2}{\overset{C_1}{\gtrless}} \frac{P_2}{P_1} \tag{4.127}$$

Taking the likelihood ratio yields the following discriminant function:

$$h(\boldsymbol{x}) = \ln f_{X|C}(\boldsymbol{x}|C_1) - \ln f_{X|C}(\boldsymbol{x}|C_2) \underset{C_2}{\overset{C_1}{\gtrless}} \ln \frac{P_2}{P_1} \tag{4.128}$$

Now assume that the signal in each class has a Gaussian distribution with a probability distribution function given by

$$f_{X|C}(\boldsymbol{x}|c_i) = \frac{1}{\sqrt{2\pi}\,|\boldsymbol{\Sigma}_i|} \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)^{\mathrm{T}} \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{x}-\boldsymbol{\mu}_i)\right], \quad \text{i=1,2} \tag{4.129}$$

From Equations (4.128) and (4.129), the discriminant function $h(\boldsymbol{x})$ becomes

$$h(\boldsymbol{x}) = -\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_1)^{\mathrm{T}}\boldsymbol{\Sigma}_1^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1)+\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_2)^{\mathrm{T}}\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_2)+\ln\frac{|\boldsymbol{\Sigma}_2|}{|\boldsymbol{\Sigma}_1|} \underset{C_2}{\overset{C_1}{\gtrless}} \ln\frac{P_2}{P_1} \tag{4.130}$$

Example 4.10 For two Gaussian-distributed classes of scalar-valued signals with distributions given by $\mathcal{N}(x(m),\mu_1,\sigma^2)$ and $\mathcal{N}(x(m),\mu_2,\sigma^2)$, and equal class probability P_1=P_2=0.5, the discrimination function of Equation (4.130) becomes

$$h(x(m)) = \frac{\mu_2-\mu_1}{\sigma^2}x(m)+\frac{1}{2}\frac{\mu_2^2-\mu_1^2}{\sigma^2} \underset{C_2}{\overset{C_1}{\gtrless}} 0 \tag{4.131}$$

Hence the rule for signal classification becomes

$$x(m) \underset{C_2}{\overset{C_1}{\lessgtr}} \frac{\mu_1+\mu_2}{2} \tag{4.132}$$

The signal is labelled with class C_1 if $x(m)<(\mu_1+\mu_2)/2$ and as class C_2 otherwise.

4.6.2 Classification Error

Classification errors are due to the overlap of the distributions of different classes of signals. This is illustrated in Figure 4.16 for the binary classification of a scalar-valued signal and in Figure 4.17 for the binary classification of a two-dimensional signal. In each figure the overlapped area gives a measure of classification error. The obvious solution for reducing the classification error is to reduce the overlap of the distributions. This may be achieved by increasing the distance between the mean values of various classes or by reducing the variance of each class. In the binary classification of a scalar-valued variable x, the probability of classification error is given by

$$P(Error|x) = P(C_1)P(x > Thrsh \mid x \in C_1) + P(C_2)P(x > Thrsh \mid x \in C_2) \qquad (4.133)$$

For two Gaussian-distributed classes of scalar-valued signals with pdfs $\mathcal{N}(x(m),\mu_1,\sigma_1^2)$ and $\mathcal{N}(x(m),\mu_2,\sigma_2^2)$, Equation (4.133) becomes

$$\begin{aligned} P(Error|x) = {} & P(C_1) \int_{Thrsh}^{\infty} \frac{1}{\sqrt{2\pi}\,\sigma_1} \exp\left(-\frac{(x-\mu_1)^2}{2\sigma_1^2} \right) dx \\ & + P(C_2) \int_{-\infty}^{Thrsh} \frac{1}{\sqrt{2\pi}\sigma_2} \left(-\frac{(x-\mu_2)^2}{2\sigma_2^2} \right) dx \end{aligned} \qquad (4.134)$$

where the parameter *Thrsh* is the classification threshold.

4.6.3 Bayesian Classification of Discrete-Valued Parameters

Let the set $\Theta=\{\boldsymbol{\theta}_i,\ i=1,\ ...,\ M\}$ denote the values that a discrete P-dimensional parameter vector $\boldsymbol{\theta}$ can assume. In general, the observation space $\boldsymbol{Y}$ associated with a discrete parameter space Θ may be a discrete-valued or a continuous-valued space. Assuming that the observation space is continuous, the pdf of the parameter vector $\boldsymbol{\theta}_i$, given observation vector $\boldsymbol{y}$, may be expressed, using Bayes' rule, as

$$P_{\Theta|\boldsymbol{Y}}(\boldsymbol{\theta}_i|\boldsymbol{y}) = \frac{f_{\boldsymbol{Y}|\Theta}(\boldsymbol{y}|\boldsymbol{\theta}_i)P_{\Theta}(\boldsymbol{\theta}_i)}{f_{\boldsymbol{Y}}(\boldsymbol{y})} \qquad (4.135)$$

For the case when the observation space $\boldsymbol{Y}$ is discrete-valued, the probability density functions are replaced by the appropriate probability mass functions. The Bayesian risk in selecting the parameter vector $\boldsymbol{\theta}_i$ given the observation $\boldsymbol{y}$ is defined as

$$\mathcal{R}(\boldsymbol{\theta}_i \mid \boldsymbol{y}) = \sum_{j=1}^{M} C(\boldsymbol{\theta}_i \mid \boldsymbol{\theta}_j) P_{\Theta|\boldsymbol{Y}}(\boldsymbol{\theta}_j \mid \boldsymbol{y}) \qquad (4.136)$$

where $C(\boldsymbol{\theta}_i|\boldsymbol{\theta}_j)$ is the cost of selecting the parameter $\boldsymbol{\theta}_i$ when the true parameter is $\boldsymbol{\theta}_j$. The Bayesian classification Equation (4.136) can be employed to obtain the maximum a posteriori, the maximum likelihood and the minimum mean square error classifiers.

4.6.4 Maximum A Posteriori Classification

MAP classification corresponds to Bayesian classification with a uniform cost function defined as

$$C(\boldsymbol{\theta}_i \mid \boldsymbol{\theta}_j) = 1 - \delta(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j) \tag{4.137}$$

where $\delta(\cdot)$ is the delta function. Substitution of this cost function in the Bayesian risk function yields

$$\begin{aligned} \mathcal{R}_{MAP}(\boldsymbol{\theta}_i \mid \boldsymbol{y}) &= \sum_{j=1}^{M} [1 - \delta(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j)]\, P_{\Theta|\boldsymbol{y}}(\boldsymbol{\theta}_j \mid \boldsymbol{y}) \\ &= 1 - P_{\Theta|\boldsymbol{y}}(\boldsymbol{\theta}_i \mid \boldsymbol{y}) \end{aligned} \tag{4.138}$$

Note that the MAP risk in selecting $\boldsymbol{\theta}_i$ is the classification error probability; that is the sum of the probabilities of all other candidates. From Equation (4.138) minimisation of the MAP risk function is achieved by maximisation of the posterior pmf:

$$\begin{aligned} \hat{\boldsymbol{\theta}}_{MAP}(\boldsymbol{y}) &= \arg\max_{\boldsymbol{\theta}_i} P_{\Theta|Y}(\boldsymbol{\theta}_i \mid \boldsymbol{y}) \\ &= \arg\max_{\boldsymbol{\theta}_i} P_{\Theta}(\boldsymbol{\theta}_i) f_{Y|\Theta}(\boldsymbol{y} \mid \boldsymbol{\theta}_i) \end{aligned} \tag{4.139}$$

4.6.5 Maximum-Likelihood (ML) Classification

The ML classification corresponds to Bayesian classification when the parameter $\boldsymbol{\theta}$ has a uniform prior pmf and the cost function is also uniform:

$$\begin{aligned} \mathcal{R}_{ML}(\boldsymbol{\theta}_i \mid \boldsymbol{y}) &= \sum_{j=1}^{M} [1 - \delta(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j)] \frac{1}{f_Y(\boldsymbol{y})} f_{Y|\Theta}\big(\boldsymbol{y} \mid \boldsymbol{\theta}_j\big) P_{\Theta}(\boldsymbol{\theta}_j) \\ &= 1 - \frac{1}{f_Y(\boldsymbol{y})} f_{Y|\theta}(\boldsymbol{y} \mid \boldsymbol{\theta}_i) P_{\Theta} \end{aligned} \tag{4.140}$$

where P_Θ is the uniform pmf of $\boldsymbol{\theta}$. Minimisation of the ML risk function (4.140) is equivalent to maximisation of the likelihood $f_{Y|\Theta}(\boldsymbol{y}|\boldsymbol{\theta}_i)$

$$\hat{\boldsymbol{\theta}}_{ML}(\boldsymbol{y}) = \underset{\boldsymbol{\theta}_i}{\arg\max}\, f_{Y|\Theta}(\boldsymbol{y}\,|\,\boldsymbol{\theta}_i) \tag{4.141}$$

4.6.6 Minimum Mean Square Error Classification

The Bayesian minimum mean square error classification results from minimisation of the following risk function:

$$\mathcal{R}_{MMSE}(\boldsymbol{\theta}_i\,|\,\boldsymbol{y}) = \sum_{j=1}^{M} \left|\boldsymbol{\theta}_i - \boldsymbol{\theta}_j\right|^2 P_{\Theta|Y}(\boldsymbol{\theta}_j\,|\,\boldsymbol{y}) \tag{4.142}$$

For the case when $P_{\Theta|Y}(\boldsymbol{\theta}_j|\boldsymbol{y})$ is not available, the MMSE classifier is given by

$$\hat{\boldsymbol{\theta}}_{MMSE}(\boldsymbol{y}) = \underset{\boldsymbol{\theta}_i}{\arg\min} \left|\boldsymbol{\theta}_i - \boldsymbol{\theta}(\boldsymbol{y})\right|^2 \tag{4.143}$$

where $\boldsymbol{\theta}(\boldsymbol{y})$ is an estimate based on the observation $\boldsymbol{y}$.

4.6.7 Bayesian Classification of Finite State Processes

In this section, the classification problem is formulated within the framework of a finite state random process. A finite state process is composed of a probabilistic chain of a number of different random processes. Finite state processes are used for modelling non-stationary signals such as speech, image, background acoustic noise, and impulsive noise as discussed in Chapter 5.

Consider a process with a set of M states denoted as $S=\{s_1, s_2, \ldots, s_M\}$, where each state has some distinct statistical property. In its simplest form, a state is just a single vector, and the finite state process is equivalent to a discrete-valued random process with M outcomes. In this case the Bayesian state estimation is identical to the Bayesian classification of a signal into one of M discrete-valued vectors. More generally, a state generates continuous-valued, or discrete-valued vectors from a pdf, or a pmf, associated with the state. Figure 4.18 illustrates an M-state process, where the output of the ith state is expressed as

$$\boldsymbol{x}(m) = h_i\big(\boldsymbol{\theta}_i, \boldsymbol{e}(m)\big), \quad i = 1, \ldots, M \tag{4.144}$$

where in each state the signal $\boldsymbol{x}(m)$ is modelled as the output of a state-dependent function $h_i(\cdot)$ with parameter $\boldsymbol{\theta}_{i,}$ input $\boldsymbol{e}(m)$ and an input pdf $f_{Ei}(\boldsymbol{e}(m))$. The prior probability of each state is given by

$$P_S(s_i) = \mathcal{E}[N(s_i)] \Bigg/ \mathcal{E}\left[\sum_{j=1}^{M} N(s_j)\right] \tag{4.145}$$

where $\mathcal{E}[N(s_i)]$ is the expected number of observation from state s_i. The pdf of the output of a finite state process is a weighted combination of the pdf of each state and is given by

$$f_X\big(\boldsymbol{x}(m)\big) = \sum_{i=1}^{M} P_S(s_i) f_{X|S}(\boldsymbol{x} \mid s_i) \tag{4.146}$$

In Figure 4.18, the noisy observation $\boldsymbol{y}(m)$ is the sum of the process output $\boldsymbol{x}(m)$ and an additive noise $\boldsymbol{n}(m)$. From Bayes' rule, the posterior probability of the state s_i given the observation $\boldsymbol{y}(m)$ can be expressed as

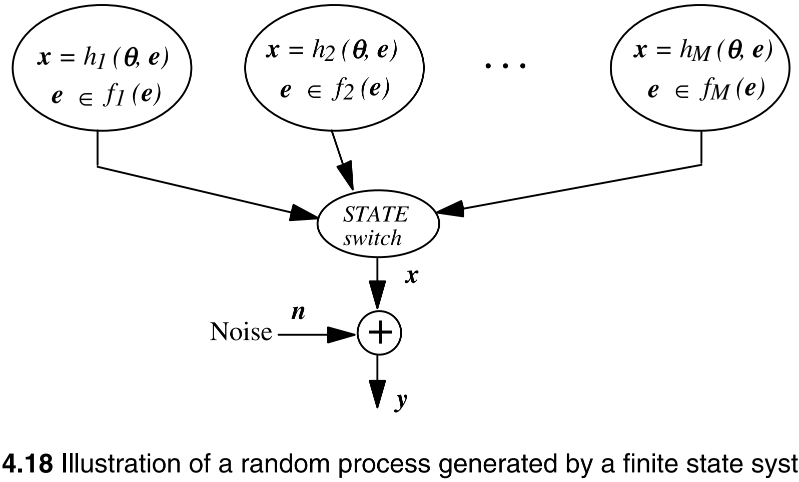


Figure 4.18 Illustration of a random process generated by a finite state system.

$$P_{S|Y}\left(s_i \middle| \boldsymbol{y}(m)\right)=\frac{f_{Y|S}\left(\boldsymbol{y}(m)\middle|s_i\right)P_S(s_i)}{\sum_{j=1}^{M} f_{Y|S}\left(\boldsymbol{y}(m)\middle|s_j\right)P_S(s_j)} \tag{4.147}$$

In MAP classification, the state with the maximum posterior probability is selected as

$$s_{MAP}\left(\boldsymbol{y}(m)\right)=\arg\max_{s_i} P_{S|\boldsymbol{Y}}\left(s_i \middle| \boldsymbol{y}(m)\right) \tag{4.148}$$

The Bayesian state classifier assigns a misclassification cost function $C(s_i|s_j)$ to the action of selecting the state s_i when the true state is s_j. The risk function for the Bayesian classification is given by

$$\mathcal{R}(s_i|\boldsymbol{y}(\mathrm{m}))=\sum_{j=1}^{M} C(s_i\,|s_j)P_{S|\boldsymbol{Y}}(s_j\,|\boldsymbol{y}(\mathrm{m})) \tag{4.149}$$

4.6.8 Bayesian Estimation of the Most Likely State Sequence

Consider the estimation of the most likely state sequence $\boldsymbol{s}=[s_{i_0},s_{i_1},\dots,s_{i_{T-1}}]$ of a finite state process, given a sequence of T observation vectors $\boldsymbol{Y}=[\boldsymbol{y}_0,\boldsymbol{y}_1,\dots,\boldsymbol{y}_{T-1}]$. A state sequence $\boldsymbol{s}$, of length T, is itself a random integer-valued vector process with N^T possible values. From the Bayes rule, the posterior pmf of a state sequence $\boldsymbol{s}$, given an observation sequence $\boldsymbol{Y}$, can be expressed as

$$P_{S|Y}(s_{i_0},\dots,s_{i_{T-1}}\,|\,\boldsymbol{y}_0,\dots,\boldsymbol{y}_{T-1})=\frac{f_{Y|S}(\boldsymbol{y}_0,\dots,\boldsymbol{y}_{T-1}\,|\,s_{i_0},\dots,s_{i_{T-1}})P_S(s_{i_0},\dots,s_{i_{T-1}})}{f_Y(\boldsymbol{y}_0,\dots,\boldsymbol{y}_{T-1})} \tag{4.150}$$

where $P_S(\boldsymbol{s})$ is the pmf of the state sequence $\boldsymbol{s}$, and for a given observation sequence, the denominator $f_Y(\boldsymbol{y}_0,\dots,\boldsymbol{y}_{T-1})$ is a constant. The Bayesian risk in selecting a state sequence $\boldsymbol{s}_i$ is expressed as

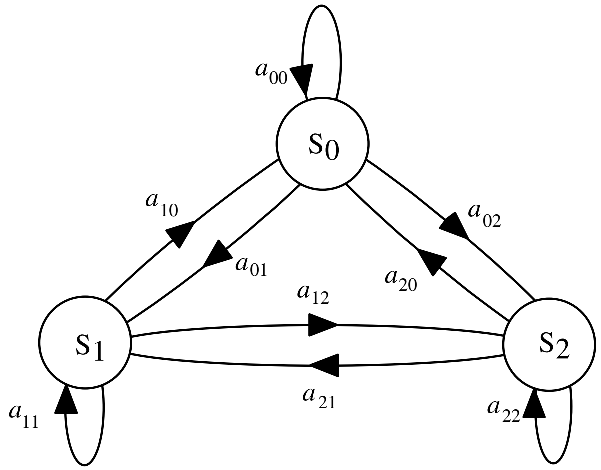


Figure 4.19 A three state Markov Process.

$$\mathcal{R}(\boldsymbol{s}_i|\boldsymbol{y}) = \sum_{j=1}^{N^T} C(\boldsymbol{s}_i \mid \boldsymbol{s}_j) P_{S|Y}(\boldsymbol{s}_j|\boldsymbol{y}) \tag{4.151}$$

For a statistically independent process, the state of the process at any time is independent of the previous states, and hence the conditional probability of a state sequence can be written as

$$P_{S|Y}(s_{i_0},\ldots,s_{i_{T-1}}|\boldsymbol{y}_0,\ldots,\boldsymbol{y}_{T-1}) = \prod_{k=0}^{T-1} f_{Y|S}(\boldsymbol{y}_k|s_{i_k}) P_S(s_{i_k}) \tag{4.152}$$

where s_{ik} denotes state s_i at time instant k. A particular case of a finite state process is the Markov chain where the state transition is governed by a Markovian process such that the probability of the state i at time m depends on the state of the process at time m-1. The conditional pmf of a Markov state sequence can be expressed as

$$P_{S|Y}(s_{i_0},\ldots,s_{i_{T-1}} \mid \boldsymbol{y}_0,\ldots,\boldsymbol{y}_{T-1}) = \prod_{k=0}^{T-1} a_{i_{k-1}i_k} f_{S|Y}(s_{i_k} \mid \boldsymbol{y}_k) \tag{4.153}$$

where $a_{i_{k-1}i_k}$ is the probability that the process moves from state $s_{i_{k-1}}$ to state s_{i_k} Finite state random processes and computationally efficient methods of state sequence estimation are described in detail in Chapter 5.

4.7 Modelling the Space of a Random Process

In this section, we consider the training of statistical models for a database of P-dimensional vectors of a random process. The vectors in the database can be visualised as forming a number of clusters or regions in a P-dimensional space. The statistical modelling method consists of two steps: (a) the partitioning of the database into a number of regions, or clusters, and (b) the estimation of the parameters of a statistical model for each cluster. A simple method for modelling the space of a random signal is to use a set of prototype vectors that represent the centroids of the signal space. This method effectively quantises the space of a random process into a relatively small number of typical vectors, and is known as *vector quantisation* (VQ). In the following, we first consider a VQ model of a random process, and then extend this model to a pdf model, based on a mixture of Gaussian densities.

4.7.1 Vector Quantisation of a Random Process

In vector quantisation, the space of a random vector process $\boldsymbol{X}$ is partitioned into K clusters or regions $[\boldsymbol{X}_1, \boldsymbol{X}_2, ...,\boldsymbol{X}_K]$, and each cluster $\boldsymbol{X}_i$ is represented by a cluster centroid $\boldsymbol{c}_i$. The set of centroid vectors $[\boldsymbol{c}_1, \boldsymbol{c}_2, ...,\boldsymbol{c}_K]$ form a VQ code book model of the process $\boldsymbol{X}$. The VQ code book can then be used to classify an unlabelled vector $\boldsymbol{x}$ with the nearest centroid. The codebook is searched to find the centroid vector with the minimum distance from $\boldsymbol{x}$, then $\boldsymbol{x}$ is labelled with the index of the minimum distance centroid as

$$Label(\boldsymbol{x}) = \arg\min_{i} d(\boldsymbol{x},\boldsymbol{c}_i) \tag{4.154}$$

where $d(\boldsymbol{x}, \boldsymbol{c}_i)$ is a measure of distance between the vectors $\boldsymbol{x}$ and $\boldsymbol{c}_i$. The most commonly used distance measure is the mean squared distance.

4.7.2 Design of a Vector Quantiser: *K*-Means Clustering

The K-means algorithm, illustrated in Figure 4.20, is an iterative method for the design of a VQ codebook. Each iteration consists of two basic steps : (a) Partition the training signal space into K regions or clusters and (b) compute the centroid of each region. The steps in K-Means method are as follows:

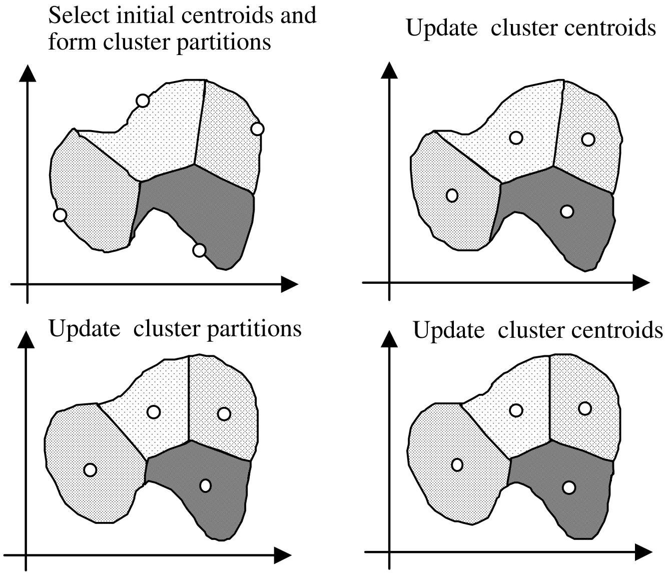

Select initial centroids and form cluster partitions
Update cluster centroids
Update cluster partitions
Update cluster centroids

Figure 4.18 Illustration of the *K*-means clustering method.

Step 1: *Initialisation* Use a suitable method to choose a set of K initial centroids $[\boldsymbol{c}_i]$. For $m = 1, 2, \ldots$

Step 2: *Classification* Classify the training vectors $\{\boldsymbol{x}\}$ into K clusters $\{[\boldsymbol{x}_1], [\boldsymbol{x}_2], \ldots [\boldsymbol{x}_K]\}$ using the so-called nearest-neighbour rule Equation (4.154).

Step 3: *Centroid computation* Use the vectors $[\boldsymbol{x}_i]$ associated with the i^{th} cluster to compute an updated cluster centroid $\boldsymbol{c}_i$, and calculate the cluster distortion defined as

$$D_i(m) = \frac{1}{N_i} \sum_{j=1}^{N_i} d(\boldsymbol{x}_i(j), \boldsymbol{c}_i(m)) \tag{4.155}$$

where it is assumed that a set of N_i vectors $[\boldsymbol{x}_i(j)\ j=0, \ldots, N_i]$ are associated with cluster i. The total distortion is given by

$$D(m) = \sum_{i=1}^{K} D_i(m) \tag{4.156}$$

```
Step 4: Convergence test:
    if
        $D(m-1) - D(m) \geq Threshold$ stop,
    else
            goto Step 2.
```

A vector quantiser models the regions, or the clusters, of the signal space with a set of cluster centroids. A more complete description of the signal space can be achieved by modelling each cluster with a Gaussian density as described in the next chapter.

4.8 Summary

This chapter began with an introduction to the basic concepts in estimation theory; such as the signal space and the parameter space, the prior and posterior spaces, and the statistical measures that are used to quantify the performance of an estimator. The Bayesian inference method, with its ability to include as much information as is available, provides a general framework for statistical signal processing problems. The minimum mean square error, the maximum-likelihood, the maximum a posteriori, and the minimum absolute value of error methods were derived from the Bayesian formulation. Further examples of the applications of Bayesian type models in this book include the hidden Markov models for non-stationary processes studied in Chapter 5, and blind equalisation of distorted signals studied in Chapter 15.

We considered a number of examples of the estimation of a signal observed in noise, and derived the expressions for the effects of using prior pdfs on the mean and the variance of the estimates. The choice of the prior pdf is an important consideration in Bayesian estimation. Many processes, for example speech or the response of a telecommunication channel, are not uniformly distributed in space, but are constrained to a particular region of signal or parameter space. The use of a prior pdf can guide the estimator to focus on the posterior space that is the subspace consistent with both the likelihood and the prior pdfs. The choice of the prior, depending on how well it fits the process, can have a significant influence on the solutions.

The iterative estimate-maximise method, studied in Section 4.3, provides a practical framework for solving many statistical signal processing problems, such as the modelling of a signal space with a mixture Gaussian densities, and the training of hidden Markov models in Chapter 5. In Section 4.4 the Cramer–Rao lower bound on the variance of an estimator was derived, and it was shown that the use of a prior pdf can reduce the minimum estimator variance.

Finally we considered the modelling of a data space with a mixture Gaussian process, and used the EM method to derive a solution for the parameters of the mixture Gaussian model.

Bibliography

ANDERGERG M.R. (1973) Cluster Analysis for Applications. Academic Press, New York.

ABRAMSON N. (1963) Information Theory and Coding. McGraw Hill, New York.

BAUM L.E., PETRIE T., SOULES G. and WEISS N. (1970) A Maximisation Technique occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains. Ann. Math. Stat. **41**, pp.164–171.

BAYES T. (1763) An Essay Towards Solving a Problem in the Doctrine of Changes, Phil. Trans. Royal Society of London, **53**, pp. 370–418, (reprinted in 1958 in Biometrika, **45**, pp. 293–315).

CHOU P. LOOKABAUGH T. and GRAY R. (1989) Entropy-Constrained Vector Quantisation. IEEE Trans. Acoustics, Speech and Signal Processing, **ASSP-37**, pp. 31–42.

BEZDEK J.C. (1981) Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum Press, New York.

CRAMER H. (1974) Mathematical Methods of Statistics. Princeton University Press.

DEUTSCH R. (1965) Estimation Theory. Prentice-Hall, Englewood Cliffs, NJ.

DEMPSTER A.P., LAIRD N.M. and RUBIN D.B. (1977) Maximum Likelihood from Incomplete Data via the EM Algorithm. J. R. Stat. Soc. Ser. B, **39**, pp. 1-38.

DUDA R.O. and HART R.E. (1973) Pattern Classification. Wiley, New York.

FEDER M. and WEINSTEIN E. (1988) Parameter Estimation of Superimposed Signals using the EM algorithm. IEEE Trans. Acoustics, Speech and Signal Processing, **ASSP-36(4)**, pp. 477.489.

FISHER R.A. (1922) On the Mathematical Foundations of the Theoretical Statistics. Phil Trans. Royal. Soc. London, **222**, pp. 309–368.

GERSHO A. (1982) On the Structure of Vector Quantisers. IEEE Trans. Information Theory, **IT-28**, pp. 157–166.

GRAY R.M. (1984) Vector Quantisation. IEEE ASSP Magazine, p. 4-29.

GRAY R.M. and KARNIN E.D (1982), Multiple local Optima in Vector Quantisers. IEEE Trans. Information Theory, **IT-28**, pp. 256–261.

JEFFREY H. (1961) Scientific Inference, 3rd ed. Cambridge University Press.

LARSON H.J. and BRUNO O.S. (1979) Probabilistic Models in Engineering Sciences. **I** and **II.** Wiley, New York.

LINDE Y., BUZO A. and GRAY R.M. (1980) An Algorithm for Vector Quantiser Design. IEEE Trans. Comm. **COM-28**, pp. 84–95.

MAKHOUL J., ROUCOS S., and GISH H. (1985) Vector Quantisation in Speech Coding. Proc. IEEE, **73**, pp. 1551–1588.

MOHANTY N. (1986) Random Signals, Estimation and Identification. Van Nostrand, New York.

RAO C.R. (1945) Information and Accuracy Attainable in the Estimation of Statistical Parameters. Bull Calcutta Math. Soc., **37**, pp. 81–91.

RENDER R.A. and WALKER H.F.(1984) Mixture Densities, Maximum Likelihood and the EM algorithm. SIAM review, **26**, pp. 195–239.

SCHARF L.L. (1991) Statistical Signal Processing: Detection, Estimation, and Time Series Analysis. Addison Wesley, Reading, MA.