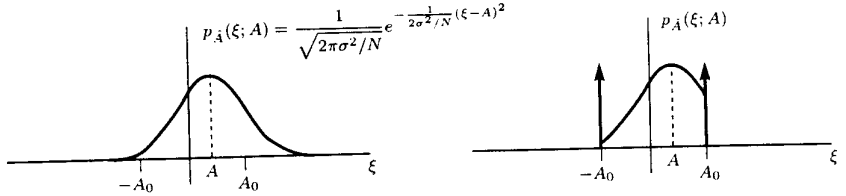


(a) PDF of sample mean **(b)** PDF of truncated sample mean

Figure 10.1 Improvement of estimator via prior knowledge

estimate the realization of a random variable based on data is described in Section 10.5 as being due to the correlation between the random variables. Theorem 10.2 summarizes the result that a jointly Gaussian PDF yields a conditional PDF that is also Gaussian, having a mean (10.24) and a covariance (10.25). This is then applied to the Bayesian linear model of (10.26) to yield the posterior PDF as summarized in Theorem 10.3. As will be shown in Chapter 11, the mean of the posterior PDF (10.28) is the minimum MSE estimator for a vector parameter. Section 10.7 discusses nuisance parameters from a Bayesian viewpoint, while Section 10.8 describes the potential difficulties of using a Bayesian estimator in a classical estimation problem.

10.3 Prior Knowledge and Estimation

It is a fundamental rule of estimation theory that the use of prior knowledge will lead to a more accurate estimator. For example, if a parameter is constrained to lie in a known interval, then any good estimator should produce only estimates within that interval. In Example 3.1 it was shown that the MVU estimator of A is the sample mean $\bar{x}$. However, this assumed that A could take on any value in the interval $-\infty < A < \infty$. Due to physical constraints it may be more reasonable to assume that A can take on only values in the finite interval $-A_0 \leq A \leq A_0$. To retain $\hat{A} = \bar{x}$ as the best estimator would be undesirable since $\hat{A}$ may yield values outside the known interval. As shown in Figure 10.1a, this is due to noise effects. Certainly, we would expect to improve our estimation if we used the *truncated* sample mean estimator

$$\check{A} = \begin{cases} -A_0 & \bar{x} < -A_0 \\ \bar{x} & -A_0 \leq \bar{x} \leq A_0 \\ A_0 & \bar{x} > A_0 \end{cases}$$

which would be consistent with the known constraints. Such an estimator would have the PDF

$$\begin{aligned} p_{\check{A}}(\xi;A) &= \Pr\{\bar{x} \leq -A_0\}\delta(\xi + A_0) \\ &\quad + p_{\hat{A}}(\xi;A)[u(\xi + A_0) - u(\xi - A_0)] \\ &\quad + \Pr\{\bar{x} \geq A_0\}\delta(\xi - A_0) \end{aligned} \tag{10.1}$$

where $u(x)$ is the unit step function. This is shown in Figure 10.1b. It is seen that $\check{A}$ is a biased estimator. However, if we compare the MSE of the two estimators, we note that for any A in the interval $-A_0 \leq A \leq A_0$

$$\begin{aligned} \text{mse}(\hat{A}) &= \int_{-\infty}^{\infty} (\xi - A)^2 p_{\hat{A}}(\xi;A)\,d\xi \\ &= \int_{-\infty}^{-A_0} (\xi - A)^2 p_{\hat{A}}(\xi;A)\,d\xi + \int_{-A_0}^{A_0} (\xi - A)^2 p_{\hat{A}}(\xi;A)\,d\xi \\ &\quad + \int_{A_0}^{\infty} (\xi - A)^2 p_{\hat{A}}(\xi;A)\,d\xi \\ &> \int_{-\infty}^{-A_0} (-A_0 - A)^2 p_{\hat{A}}(\xi;A)\,d\xi + \int_{-A_0}^{A_0} (\xi - A)^2 p_{\hat{A}}(\xi;A)\,d\xi \\ &\quad + \int_{A_0}^{\infty} (A_0 - A)^2 p_{\hat{A}}(\xi;A)\,d\xi \\ &= \text{mse}(\check{A}). \end{aligned}$$

Hence, $\check{A}$, the truncated sample mean estimator, is better than the sample mean estimator in terms of MSE. Although $\hat{A}$ is still the MVU estimator, we have been able to reduce the *mean square error* by allowing the estimator to be biased. In as much as we have been able to produce a better estimator, the question arises as to whether an *optimal* estimator exists for this problem. (The reader may recall that in the classical case the MSE criterion of optimality usually led to unrealizable estimators. We shall see that this is not a problem in the Bayesian approach.) We can answer affirmatively but only after reformulating the data model. Knowing that A must lie in a known interval, we suppose that the true value of A has been chosen from that interval. We then model the process of choosing a value as a random event to which a PDF can be assigned. With knowledge only of the interval and no inclination as to whether A should be nearer any particular value, it makes sense to assign a $\mathcal{U}[-A_0, A_0]$ PDF to the *random variable* A. The overall data model then appears as in Figure 10.2. As shown there, the act of choosing A according to the given PDF represents the departure of the Bayesian approach from the classical approach. The problem, as always, is to estimate the *value* of A or the *realization* of the random variable. However, now we can incorporate our knowledge of *how* A was chosen. For example, we might attempt to find an estimator $\hat{A}$ that would minimize the *Bayesian* MSE defined as

$$\text{Bmse}(\hat{A}) = E[(A - \hat{A})^2]. \tag{10.2}$$

We choose to define the error as $A - \hat{A}$ in contrast to the classical estimation error of $\hat{A} - A$. This definition will be useful later when we discuss a vector space interpretation of the Bayesian estimator. In (10.2) we emphasize that since A is a random variable, the expectation operator is with respect to the *joint PDF* $p(\mathbf{x}, A)$. This is a fundamentally different MSE than in the classical case. We distinguish it by using the Bmse notation. To appreciate the difference compare the classical MSE

$$\text{mse}(\hat{A}) = \int (\hat{A} - A)^2 p(\mathbf{x};A)\,d\mathbf{x} \tag{10.3}$$

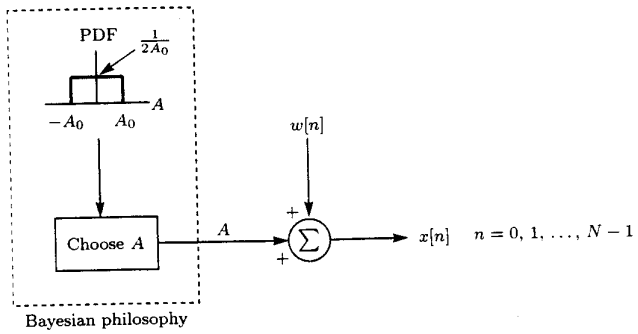


Figure 10.2 Bayesian approach to data modeling

to the Bayesian MSE

$$\text{Bmse}(\hat{A}) = \int\int (A - \hat{A})^2 p(\mathbf{x}, A)\, d\mathbf{x}\, dA. \tag{10.4}$$

Even the underlying experiments are different, as is reflected in the averaging PDFs. If we were to assess the MSE performance using a Monte Carlo computer simulation, then for the classical approach we would choose a realization of $w[n]$ and add it to a *given* A. This procedure would be repeated M times. Each time we would add a new realization of $w[n]$ to the *same* A. In the Bayesian approach, for each realization we would choose A according to its PDF $\mathcal{U}[-A_0, A_0]$ and then generate $w[n]$ (assuming that $w[n]$ is independent of A). We would then repeat this procedure M times. In the classical case we would obtain a MSE for each assumed value of A, while in the Bayesian case the single MSE figure obtained would be an average over the PDF of A. Note that whereas the classical MSE will depend on A, and hence estimators that attempt to minimize the MSE will usually depend on A (see Section 2.4), the Bayesian MSE will not. In effect, we have integrated the parameter dependence away! It should be clear that comparing classical and Bayesian estimators is like comparing "apples and oranges." The reader who is tempted to do so may become thoroughly confused. Nonetheless, at times, the *forms* of the estimators will be identical (see Problem 10.1).

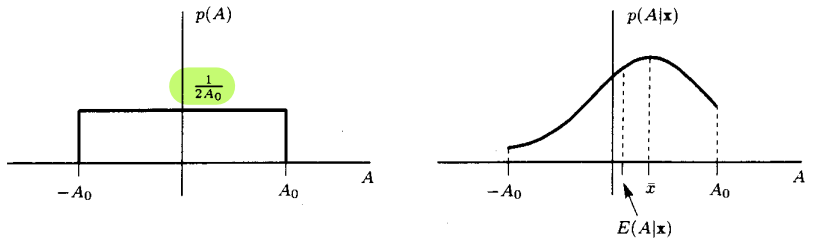
To complete our example we now derive the estimator that minimizes the Bayesian MSE. First, we use Bayes' theorem to write

$$p(\mathbf{x}, A) = p(A|\mathbf{x})p(\mathbf{x})$$

so that

$$\text{Bmse}(\hat{A}) = \int \left[\int (A - \hat{A})^2 p(A|\mathbf{x})\, dA\right] p(\mathbf{x})\, d\mathbf{x}.$$

Now since $p(\mathbf{x}) \geq 0$ for all $\mathbf{x}$, if the integral in brackets can be minimized for each $\mathbf{x}$, then the Bayesian MSE will be minimized. Hence, fixing $\mathbf{x}$ so that $\hat{A}$ is a scalar variable



(a) Prior PDF **(b)** Posterior PDF

Figure 10.3 Comparison of prior and posterior PDFs

(as opposed to a general function of $\mathbf{x}$), we have

$$\begin{aligned}\frac{\partial}{\partial \hat{A}} \int (A - \hat{A})^2 p(A|\mathbf{x})\, dA &= \int \frac{\partial}{\partial \hat{A}} (A - \hat{A})^2 p(A|\mathbf{x})\, dA \\ &= \int -2(A - \hat{A}) p(A|\mathbf{x})\, dA \\ &= -2\int A p(A|\mathbf{x})\, dA + 2\hat{A}\int p(A|\mathbf{x})\, dA\end{aligned}$$

which when set equal to zero results in

$$\hat{A} = \int A p(A|\mathbf{x})\, dA$$

or finally

$$\hat{A} = E(A|\mathbf{x}) \tag{10.5}$$

since the conditional PDF must integrate to 1. It is seen that the optimal estimator in terms of minimizing the Bayesian MSE is the *mean* of the *posterior* PDF $p(A|\mathbf{x})$ (see also Problem 10.5 for an alternative derivation). The posterior PDF refers to the PDF of A *after* the data have been observed. In contrast, $p(A)$ or

$$p(A) = \int p(\mathbf{x}, A)\, d\mathbf{x}$$

may be thought of as the prior PDF of A, indicating the PDF *before* the data are observed. We will henceforth term the estimator that minimizes the Bayesian MSE the minimum mean square error (MMSE) estimator. Intuitively, the effect of observing data will be to concentrate the PDF of A as shown in Figure 10.3 (see also Problem 10.15). This is because knowledge of the data should reduce our uncertainty about A. We will return to this idea later.

In determining the MMSE estimator we first require the posterior PDF. We can use Bayes' rule to determine it as

$$\begin{aligned} p(A|\mathbf{x}) &= \frac{p(\mathbf{x}|A)p(A)}{p(\mathbf{x})} \\ &= \frac{p(\mathbf{x}|A)p(A)}{\int p(\mathbf{x}|A)p(A)\,dA}. \end{aligned} \tag{10.6}$$

Note that the denominator is just a normalizing factor, independent of A, needed to ensure that $p(A|\mathbf{x})$ integrates to 1. If we continue our example, we recall that the prior PDF $p(A)$ is $\mathcal{U}[-A_0, A_0]$. To specify the conditional PDF $p(\mathbf{x}|A)$ we need to further assume that the choice of A via $p(A)$ does not affect the PDF of the noise samples or that $w[n]$ is independent of A. Then, for $n = 0, 1, \ldots, N-1$

$$\begin{aligned} p_x(x[n]|A) &= p_w(x[n] - A|A) \\ &= p_w(x[n] - A) \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(x[n] - A)^2\right] \end{aligned}$$

and therefore

$$p(\mathbf{x}|A) = \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp\left[-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}(x[n] - A)^2\right]. \tag{10.7}$$

It is apparent that the PDF is identical *in form* to the usual classical PDF $p(\mathbf{x}; A)$. In the Bayesian case, however, the PDF is a *conditional* PDF, hence the "|" separator, while in the classical case, it represents an unconditional PDF, albeit parameterized by A, hence the separator ";" (see also Problem 10.6). Using (10.6) and (10.7), the posterior PDF becomes

$$p(A|\mathbf{x}) = \begin{cases} \dfrac{\frac{1}{2A_0(2\pi\sigma^2)^{\frac{N}{2}}} \exp\left[-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}(x[n] - A)^2\right]}{\int_{-A_0}^{A_0} \frac{1}{2A_0(2\pi\sigma^2)^{\frac{N}{2}}} \exp\left[-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}(x[n] - A)^2\right] dA} & |A| \le A_0 \\ 0 & |A| > A_0. \end{cases}$$

But

$$\begin{aligned} \sum_{n=0}^{N-1}(x[n] - A)^2 &= \sum_{n=0}^{N-1} x^2[n] - 2NA\bar{x} + NA^2 \\ &= N(A - \bar{x})^2 + \sum_{n=0}^{N-1} x^2[n] - N\bar{x}^2 \end{aligned}$$

so that we have

$$p(A|\mathbf{x}) = \begin{cases} \dfrac{1}{c\sqrt{2\pi\frac{\sigma^2}{N}}} \exp\left[-\dfrac{1}{2\frac{\sigma^2}{N}}(A - \bar{x})^2\right] & |A| \le A_0 \\ 0 & |A| > A_0 \end{cases} \tag{10.8}$$

The factor c is determined by the requirement that $p(A|\mathbf{x})$ integrate to 1, resulting in

$$c = \int_{-A_0}^{A_0} \frac{1}{\sqrt{2\pi\frac{\sigma^2}{N}}} \exp\left[-\frac{1}{2\frac{\sigma^2}{N}}(A - \bar{x})^2\right] dA.$$

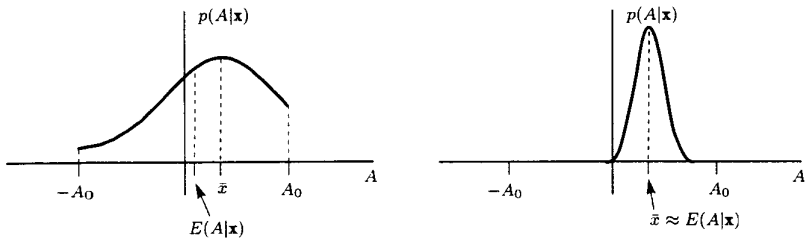
The PDF is seen to be a truncated Gaussian, as shown in Figure 10.3b. The MMSE estimator, which is the mean of $p(A|\mathbf{x})$, is

$$\begin{aligned} \hat{A} &= E(A|\mathbf{x}) \\ &= \int_{-\infty}^{\infty} Ap(A|\mathbf{x})\,dA \\ &= \frac{\int_{-A_0}^{A_0} A\frac{1}{\sqrt{2\pi\frac{\sigma^2}{N}}} \exp\left[-\frac{1}{2\frac{\sigma^2}{N}}(A - \bar{x})^2\right] dA}{\int_{-A_0}^{A_0} \frac{1}{\sqrt{2\pi\frac{\sigma^2}{N}}} \exp\left[-\frac{1}{2\frac{\sigma^2}{N}}(A - \bar{x})^2\right] dA}. \end{aligned} \tag{10.9}$$

Although this cannot be evaluated in closed form, we note that $\hat{A}$ will be a function of $\bar{x}$ as well as of A_0 and σ^2 (see Problem 10.7). The MMSE estimator will not be $\bar{x}$ due to the truncation shown in Figure 10.3b unless A_0 is so large that there is effectively no truncation. This will occur if $A_0 \gg \sqrt{\sigma^2/N}$. Otherwise, the estimator will be "biased" towards zero as opposed to being equal to $\bar{x}$. This is because the prior knowledge embodied in $p(A)$ would in the absence of the data $\mathbf{x}$ produce the MMSE estimator (see Problem 10.8)

$$\hat{A} = E(A) = 0.$$

The effect of the data is to position the posterior mean between $A = 0$ and $A = \bar{x}$ in a compromise between the prior knowledge and that contributed by the data. To further appreciate this weighting consider what happens as N becomes large so that the data knowledge becomes more important. As shown in Figure 10.4, as N increases, we have from (10.8) that the posterior PDF becomes more concentrated about $\bar{x}$ (since σ^2/N decreases). Hence, it becomes nearly Gaussian, and its mean becomes just $\bar{x}$. The MMSE estimator relies less and less on the prior knowledge and more on the data. It is said that the data "swamps out" the prior knowledge.



(a) Short data record **(b)** Large data record

Figure 10.4 Effect of increasing data record on posterior PDF

The results of this example are true in general and are now summarized. The Bayesian approach to parameter estimation assumes that the parameter to be estimated is a realization of the random variable θ. As such, we assign a prior PDF $p(\theta)$ to it. After the data are observed, our state of knowledge about the parameter is summarized by the posterior PDF $p(\theta|\mathbf{x})$. An optimal estimator is defined to be the one that minimizes the MSE when averaged over all realizations of θ and $\mathbf{x}$, the so-called Bayesian MSE. This estimator is the mean of the posterior PDF or $\hat{\theta} = E(\theta|\mathbf{x})$. The estimator is determined explicitly as

$$\hat{\theta} = E(\theta|\mathbf{x}) = \int \theta p(\theta|\mathbf{x})\, d\theta. \tag{10.10}$$

The MMSE estimator will in general depend on the prior knowledge as well as the data. If the prior knowledge is weak relative to that of the data, then the estimator will ignore the prior knowledge. Otherwise, the estimator will be "biased" towards the prior mean. As expected, the use of prior information always improves the estimation accuracy (see Example 10.1).

The choice of a prior PDF is critical in Bayesian estimation. The wrong choice will result in a poor estimator, similar to the problems of a classical estimator designed with an incorrect data model. Much of the controversy surrounding the use of Bayesian estimators stems from the inability in practice to be able to justify the prior PDF. Suffice it to say that unless the prior PDF can be based on the physical constraints of the problem, then classical estimation is more appropriate.

10.4 Choosing a Prior PDF

As shown in the previous section, once a prior PDF has been chosen, the MMSE estimator follows directly from (10.10). There is no question of existence as there is with the MVU estimator in the classical approach. The only practical stumbling block that remains, however, is whether or not $E(\theta|\mathbf{x})$ can be determined in closed form. In

the introductory example the posterior PDF $p(A|\mathbf{x})$ as given by (10.8) could not be found explicitly due to the need to normalize $p(\mathbf{x}|A)p(A)$ so that it integrates to 1. Additionally, the posterior mean could not be found, as evidenced by (10.9). We would have to resort to numerical integration to actually implement the MMSE estimator. This problem is compounded considerably in the vector parameter case. There the posterior PDF becomes

$$p(\boldsymbol{\theta}|\mathbf{x}) = \frac{p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})\, d\boldsymbol{\theta}}$$

which requires a p-dimensional integration over $\boldsymbol{\theta}$. Additionally, the mean needs to be evaluated (see Chapter 11), requiring further integration. For practical MMSE estimators we need to be able to express them in closed form. The next example illustrates an important case where this is possible.

Example 10.1 - DC Level in WGN - Gaussian Prior PDF

We now modify our prior knowledge for the introductory example. Instead of assuming the uniform prior PDF

$$p(A) = \begin{cases} \frac{1}{2A_0} & |A| \leq A_0 \\ 0 & |A| > A_0 \end{cases}$$

which led to an intractable integration, consider the Gaussian prior PDF

$$p(A) = \frac{1}{\sqrt{2\pi\sigma_A^2}} \exp\left[-\frac{1}{2\sigma_A^2}(A-\mu_A)^2\right].$$

The two prior PDFs clearly express different prior knowledge about A, although with $\mu_A = 0$ and $3\sigma_A = A_0$ the Gaussian prior PDF could be thought of as incorporating the knowledge that $|A| \leq A_0$. Of course, values of A near zero are thought to be more probable with the Gaussian prior PDF. Now if

$$\begin{aligned} p(\mathbf{x}|A) &= \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp\left[-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}(x[n]-A)^2\right] \\ &= \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp\left[-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}x^2[n]\right] \exp\left[-\frac{1}{2\sigma^2}(NA^2 - 2NA\bar{x})\right] \end{aligned}$$

we have

$$
\begin{aligned}
p(A|\mathbf{x}) &= \frac{p(\mathbf{x}|A)p(A)}{\int p(\mathbf{x}|A)p(A)\,dA} \\
&= \frac{\frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}\sqrt{2\pi\sigma_A^2}}\exp\left[-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}x^2[n]\right]\exp\left[-\frac{1}{2\sigma^2}(NA^2-2NA\bar{x})\right]}{\int_{-\infty}^{\infty}\frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}\sqrt{2\pi\sigma_A^2}}\exp\left[-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}x^2[n]\right]\exp\left[-\frac{1}{2\sigma^2}(NA^2-2NA\bar{x})\right]} \\
&\qquad \cdot \frac{\exp\left[-\frac{1}{2\sigma_A^2}(A-\mu_A)^2\right]}{\exp\left[-\frac{1}{2\sigma_A^2}(A-\mu_A)^2\right]}\,dA \\
&= \frac{\exp\left[-\frac{1}{2}\left(\frac{1}{\sigma^2}(NA^2-2NA\bar{x})+\frac{1}{\sigma_A^2}(A-\mu_A)^2\right)\right]}{\int_{-\infty}^{\infty}\exp\left[-\frac{1}{2}\left(\frac{1}{\sigma^2}(NA^2-2NA\bar{x})+\frac{1}{\sigma_A^2}(A-\mu_A)^2\right)\right]dA} \\
&= \frac{\exp\left[-\frac{1}{2}Q(A)\right]}{\int_{-\infty}^{\infty}\exp\left[-\frac{1}{2}Q(A)\right]dA}.
\end{aligned}
$$

Note, however, that the denominator does not depend on A, being a normalizing factor, and the argument of the exponential is quadratic in A. Hence, $p(A|\mathbf{x})$ must be a Gaussian PDF whose mean and variance depend on $\mathbf{x}$. Continuing, we have for $Q(A)$

$$
\begin{aligned}
Q(A) &= \frac{N}{\sigma^2}A^2 - \frac{2NA\bar{x}}{\sigma^2} + \frac{A^2}{\sigma_A^2} - \frac{2\mu_A A}{\sigma_A^2} + \frac{\mu_A^2}{\sigma_A^2} \\
&= \left(\frac{N}{\sigma^2}+\frac{1}{\sigma_A^2}\right)A^2 - 2\left(\frac{N}{\sigma^2}\bar{x}+\frac{\mu_A}{\sigma_A^2}\right)A + \frac{\mu_A^2}{\sigma_A^2}.
\end{aligned}
$$

Let

$$
\begin{aligned}
\sigma_{A|x}^2 &= \frac{1}{\frac{N}{\sigma^2}+\frac{1}{\sigma_A^2}} \\
\mu_{A|x} &= \left(\frac{N}{\sigma^2}\bar{x}+\frac{\mu_A}{\sigma_A^2}\right)\sigma_{A|x}^2.
\end{aligned}
$$

Then, by completing the square we have

$$
Q(A) = \frac{1}{\sigma_{A|x}^2}\left(A^2 - 2\mu_{A|x}A + \mu_{A|x}^2\right) - \frac{\mu_{A|x}^2}{\sigma_{A|x}^2} + \frac{\mu_A^2}{\sigma_A^2}
$$

$$
= \frac{1}{\sigma_{A|x}^2}(A-\mu_{A|x})^2 - \frac{\mu_{A|x}^2}{\sigma_{A|x}^2} + \frac{\mu_A^2}{\sigma_A^2}
$$

so that

$$
\begin{aligned}
p(A|\mathbf{x}) &= \frac{\exp\left[-\frac{1}{2\sigma_{A|x}^2}(A-\mu_{A|x})^2\right]\exp\left[-\frac{1}{2}\left(\frac{\mu_A^2}{\sigma_A^2}-\frac{\mu_{A|x}^2}{\sigma_{A|x}^2}\right)\right]}{\int_{-\infty}^{\infty}\exp\left[-\frac{1}{2\sigma_{A|x}^2}(A-\mu_{A|x})^2\right]\exp\left[-\frac{1}{2}\left(\frac{\mu_A^2}{\sigma_A^2}-\frac{\mu_{A|x}^2}{\sigma_{A|x}^2}\right)\right]dA} \\
&= \frac{1}{\sqrt{2\pi\sigma_{A|x}^2}}\exp\left[-\frac{1}{2\sigma_{A|x}^2}(A-\mu_{A|x})^2\right]
\end{aligned}
$$

where the last step follows from the requirement that $p(A|\mathbf{x})$ integrate to 1. The posterior PDF is also Gaussian, as claimed. (This result could also have been obtained by using Theorem 10.2 since $A, \mathbf{x}$ are jointly Gaussian.) In this form the MMSE estimator is readily found as

$$
\begin{aligned}
\hat{A} &= E(A|\mathbf{x}) \\
&= \mu_{A|x} \\
&= \frac{\frac{N}{\sigma^2}\bar{x}+\frac{\mu_A}{\sigma_A^2}}{\frac{N}{\sigma^2}+\frac{1}{\sigma_A^2}}
\end{aligned}
$$

or finally, the MMSE estimator is

$$
\begin{aligned}
\hat{A} &= \frac{\sigma_A^2}{\sigma_A^2+\frac{\sigma^2}{N}}\bar{x} + \frac{\frac{\sigma^2}{N}}{\sigma_A^2+\frac{\sigma^2}{N}}\mu_A \\
&= \alpha\bar{x} + (1-\alpha)\mu_A
\end{aligned} \tag{10.11}
$$

where

$$
\alpha = \frac{\sigma_A^2}{\sigma_A^2+\frac{\sigma^2}{N}}.
$$

Note that α is a weighting factor since $0 < \alpha < 1$. Using a Gaussian prior PDF, we are able to determine the MMSE estimator explicitly. It is interesting to examine the interplay between the prior knowledge and the data. When there is little data so that $\sigma_A^2 \ll \sigma^2/N$, α is small and $\hat{A} \approx \mu_A$, but as more data are observed so that $\sigma_A^2 \gg \sigma^2/N$, $\alpha \approx 1$ and $\hat{A} \approx \bar{x}$. The weighting factor α depends directly on our confidence in the prior knowledge or σ_A^2 and the data knowledge or σ^2/N. (The quantity σ^2/N is interpreted as the *conditional* variance or $E[(\bar{x}-A)^2|A]$). Alternatively, we may view this process by examining the posterior PDF as N increases. Referring to Figure 10.5, as the data record length N increases, the posterior PDF becomes narrower. This is because the

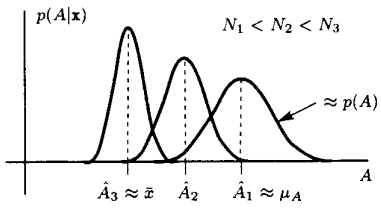


Figure 10.5 Effect of increasing data record length on posterior PDF

posterior variance

$$\text{var}(A|\mathbf{x}) = \sigma_{A|x}^2 = \frac{1}{\frac{N}{\sigma^2} + \frac{1}{\sigma_A^2}} \tag{10.12}$$

will decrease. Also, the posterior mean (10.11) or $\hat{A}$ will also change with increasing N. For small N it will be approximately μ_A but will approach $\bar{x}$ for increasing N. In fact, as $N \to \infty$, we will have $\hat{A} \to \bar{x}$, which in turn approaches the true value of A chosen. Observe that if there is no prior knowledge, which can be modeled by letting $\sigma_A^2 \to \infty$, then $\hat{A} \to \bar{x}$ for any data record length. The "classical" estimator is obtained. Finally, it was originally claimed that by using prior knowledge we could improve the estimation accuracy. To see why this is so recall that

$$\text{Bmse}(\hat{A}) = E[(A - \hat{A})^2]$$

where we evaluate the expectation with respect to $p(\mathbf{x}, A)$. But

$$\begin{aligned} \text{Bmse}(\hat{A}) &= \int\int (A - \hat{A})^2 p(\mathbf{x}, A)\, d\mathbf{x}\, dA \\ &= \int\int (A - \hat{A})^2 p(A|\mathbf{x})\, dA p(\mathbf{x})\, d\mathbf{x}. \end{aligned}$$

Since $\hat{A} = E(A|\mathbf{x})$, we have

$$\begin{aligned} \text{Bmse}(\hat{A}) &= \int\int [A - E(A|\mathbf{x})]^2 p(A|\mathbf{x})\, dA p(\mathbf{x})\, d\mathbf{x} \\ &= \int \text{var}(A|\mathbf{x}) p(\mathbf{x})\, d\mathbf{x}. \end{aligned} \tag{10.13}$$

We see that *the Bayesian MSE is just the variance of the posterior PDF when averaged over the PDF of* $\mathbf{x}$. As such, we have

$$\begin{aligned} \text{Bmse}(\hat{A}) &= \int \sigma_{A|x}^2 p(\mathbf{x})\, d\mathbf{x} \\ &= \frac{1}{\frac{N}{\sigma^2} + \frac{1}{\sigma_A^2}} \end{aligned}$$

since $\sigma_{A|x}^2$ does not depend on $\mathbf{x}$. This can be rewritten as

$$\text{Bmse}(\hat{A}) = \frac{\sigma^2}{N}\left(\frac{\sigma_A^2}{\sigma_A^2 + \frac{\sigma^2}{N}}\right) \tag{10.14}$$

so that finally we see that

$$\text{Bmse}(\hat{A}) < \frac{\sigma^2}{N}$$

where σ^2/N is the minimum MSE obtained when no prior knowledge is available (let $\sigma_A^2 \to \infty$). Clearly, any prior knowledge *when modeled in the Bayesian sense* will improve our *Bayesian estimator.* ◇

Gaussian prior PDFs are quite useful in practice due to their mathematical tractability, as illustrated by the previous example. The basic property that makes this so is the *reproducing property.* If $p(\mathbf{x}, A)$ is Gaussian, then $p(A)$, being the marginal PDF, is Gaussian, as is the posterior PDF $p(A|\mathbf{x})$. Hence, the form of the PDF remains the same, as it is conditioned on $\mathbf{x}$. Only the mean and variance change. Another example of a PDF sharing this property is given in Problem 10.10. Furthermore, Gaussian prior PDFs occur naturally in many practical problems. For the previous example we can envision the problem of measuring the DC voltage of a power source by means of a DC voltmeter. If we set the power source to 10 volts, for example, we would probably be willing to assume that the true voltage is close to 10 volts. Our prior knowledge then might be modeled as $A \sim \mathcal{N}(10, \sigma_A^2)$, where σ_A^2 would be small for a precision power source and large for a less reliable one. Next, we could take N measurements of the voltage. Our model for the measurements could be $x[n] = A + w[n]$, where the voltmeter error $w[n]$ is modeled by WGN with variance σ^2. The value of σ^2 would reflect our confidence in the quality of the voltmeter. The MMSE estimator of the true voltage would be given by (10.11). If we repeated the procedure for an ensemble of power sources and voltmeters with the same error characteristics, then our estimator would minimize the Bayesian MSE.

10.5 Properties of the Gaussian PDF

We now generalize the results of the previous section by examining the properties of the Gaussian PDF. The results of this section will be needed for the derivations of Bayesian estimators in the next chapter. The bivariate Gaussian PDF is first investigated to illustrate the important properties. Then, the corresponding results for the general multivariate Gaussian PDF are described. The remarkable property that we shall exploit is that the posterior PDF is also Gaussian, although with a different mean and variance. Some physical interpretations are stressed.

Consider a jointly Gaussian random vector $[x\, y]^T$ whose PDF is

$$p(x,y) = \frac{1}{2\pi \det^{\frac{1}{2}}(\mathbf{C})} \exp\left[-\frac{1}{2}\begin{bmatrix} x - E(x) \\ y - E(y) \end{bmatrix}^T \mathbf{C}^{-1} \begin{bmatrix} x - E(x) \\ y - E(y) \end{bmatrix}\right]. \tag{10.15}$$

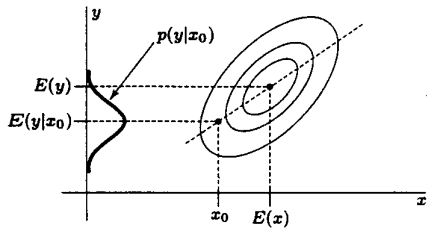


Figure 10.6 Contours of constant density for bivariate Gaussian PDF

This is also termed the *bivariate* Gaussian PDF. The mean vector and covariance matrix are

$$E\left(\begin{bmatrix} x \\ y \end{bmatrix}\right) = \begin{bmatrix} E(x) \\ E(y) \end{bmatrix}$$

$$\mathbf{C} = \begin{bmatrix} \text{var}(x) & \text{cov}(x,y) \\ \text{cov}(y,x) & \text{var}(y) \end{bmatrix}.$$

Note that the marginal PDFs $p(x)$ and $p(y)$ are also Gaussian, as can be verified by the integrations

$$p(x) = \int_{-\infty}^{\infty} p(x,y)\,dy = \frac{1}{\sqrt{2\pi \text{var}(x)}} \exp\left[-\frac{1}{2\text{var}(x)}(x-E(x))^2\right]$$

$$p(y) = \int_{-\infty}^{\infty} p(x,y)\,dx = \frac{1}{\sqrt{2\pi \text{var}(y)}} \exp\left[-\frac{1}{2\text{var}(y)}(y-E(y))^2\right].$$

The contours along which the PDF $p(x,y)$ is constant are those values of x and y for which

$$\begin{bmatrix} x-E(x) \\ y-E(y) \end{bmatrix}^T \mathbf{C}^{-1} \begin{bmatrix} x-E(x) \\ y-E(y) \end{bmatrix}$$

is a constant. They are shown in Figure 10.6 as elliptical contours. Once x, say x_0, is observed, the conditional PDF of y becomes

$$p(y|x_0) = \frac{p(x_0,y)}{p(x_0)} = \frac{p(x_0,y)}{\int_{-\infty}^{\infty} p(x_0,y)\,dy}$$

so that the conditional PDF of y is that of the cross section shown in Figure 10.6 when suitably normalized to integrate to 1. It is readily seen that since $p(x_0,y)$ (where x_0 is a fixed number) has the Gaussian form in y (from (10.15) the exponential argument is quadratic in y), the conditional PDF must also be Gaussian. Since $p(y)$ is also Gaussian, we may view this property as saying that if x and y are jointly Gaussian, the prior PDF $p(y)$ and posterior PDF $p(y|x)$ are both Gaussian. In Appendix 10A we derive the exact PDF as summarized in the following theorem.

Theorem 10.1 (Conditional PDF of Bivariate Gaussian) *If x and y are distributed according to a bivariate Gaussian PDF with mean vector $[E(x)\,E(y)]^T$ and covariance matrix*

$$\mathbf{C} = \begin{bmatrix} \text{var}(x) & \text{cov}(x,y) \\ \text{cov}(y,x) & \text{var}(y) \end{bmatrix}$$

so that

$$p(x,y) = \frac{1}{2\pi \det^{\frac{1}{2}}(\mathbf{C})} \exp\left[-\frac{1}{2}\begin{bmatrix} x-E(x) \\ y-E(y) \end{bmatrix}^T \mathbf{C}^{-1} \begin{bmatrix} x-E(x) \\ y-E(y) \end{bmatrix}\right],$$

then the conditional PDF $p(y|x)$ is also Gaussian and

$$E(y|x) = E(y) + \frac{\text{cov}(x,y)}{\text{var}(x)}(x-E(x)) \tag{10.16}$$

$$\text{var}(y|x) = \text{var}(y) - \frac{\text{cov}^2(x,y)}{\text{var}(x)}. \tag{10.17}$$

We can view this result in the following way. Before observing x, the random variable y is distributed according to the prior PDF $p(y)$ or $y \sim \mathcal{N}(E(y), \text{var}(y))$. After observing x, the random variable y is distributed according to the posterior PDF $p(y|x)$ given in Theorem 10.1. Only the mean and variance have changed. Assuming that x and y are not independent and hence $\text{cov}(x,y) \neq 0$, the posterior PDF becomes more concentrated since there is less uncertainty about y. To verify this, note from (10.17) that

$$\text{var}(y|x) = \text{var}(y)\left[1 - \frac{\text{cov}^2(x,y)}{\text{var}(x)\text{var}(y)}\right]$$

$$= \text{var}(y)(1-\rho^2) \tag{10.18}$$

where

$$\rho = \frac{\text{cov}(x,y)}{\sqrt{\text{var}(x)\text{var}(y)}} \tag{10.19}$$

is the correlation coefficient satisfying $|\rho| \leq 1$. From our previous discussions we also realize that $E(y|x)$ is the MMSE estimator of y after observing x, so that from (10.16)

$$\hat{y} = E(y) + \frac{\text{cov}(x,y)}{\text{var}(x)}(x-E(x)). \tag{10.20}$$

In normalized form (a random variable with zero mean and unity variance) this becomes

$$\frac{\hat{y}-E(y)}{\sqrt{\text{var}(y)}} = \frac{\text{cov}(x,y)}{\sqrt{\text{var}(x)\text{var}(y)}} \frac{x-E(x)}{\sqrt{\text{var}(x)}}$$

or

$$\hat{y}_n = \rho x_n. \tag{10.21}$$

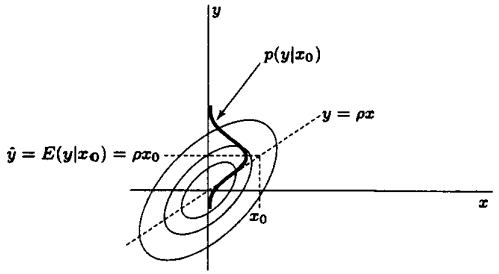


Figure 10.7
Contours of constant density of normalized bivariate PDF

The correlation coefficient then acts to scale the normalized observation x_n to obtain the MMSE estimator of the normalized realization of the random variable y_n. If the random variables are already normalized ($E(x) = E(y) = 0$, $\text{var}(x) = \text{var}(y) = 1$), the constant PDF contours appear as in Figure 10.7. The locations of the peaks of $p(x, y)$, when considered as a function of y for each x, is the dashed line $y = \rho x$, and it is readily shown that $\hat{y} = E(y|x) = \rho x$ (see Problem 10.12). *The MMSE estimator therefore exploits the correlation between the random variables to estimate the realization of one based on the realization of the other.*

The minimum MSE is, from (10.13) and (10.18),

$$\begin{aligned} \text{Bmse}(\hat{y}) &= \int \text{var}(y|x)p(x)\,dx \\ &= \text{var}(y|x) \\ &= \text{var}(y)(1-\rho^2) \end{aligned} \tag{10.22}$$

since the posterior variance does not depend on x ($\text{var}(y)$ and ρ depend on the covariance matrix only). Hence, the quality of our estimator also depends on the correlation coefficient, which is a measure of the statistical dependence between x and y.

To generalize these results consider a jointly Gaussian vector $[\mathbf{x}^T\ \mathbf{y}^T]^T$, where $\mathbf{x}$ is $k \times 1$ and $\mathbf{y}$ is $l \times 1$. In other words, $[\mathbf{x}^T\ \mathbf{y}^T]^T$ is distributed according to a multivariate Gaussian PDF. Then, the conditional PDF of $\mathbf{y}$ for a given $\mathbf{x}$ is also Gaussian, as summarized in the following theorem (see Appendix 10A for proof).

Theorem 10.2 (Conditional PDF of Multivariate Gaussian) *If $\mathbf{x}$ and $\mathbf{y}$ are jointly Gaussian, where $\mathbf{x}$ is $k \times 1$ and $\mathbf{y}$ is $l \times 1$, with mean vector $[E(\mathbf{x})^T\ E(\mathbf{y})^T]^T$ and partitioned covariance matrix*

$$\mathbf{C} = \begin{bmatrix} \mathbf{C}_{xx} & \mathbf{C}_{xy} \\ \mathbf{C}_{yx} & \mathbf{C}_{yy} \end{bmatrix} = \begin{bmatrix} k \times k & k \times l \\ l \times k & l \times l \end{bmatrix} \tag{10.23}$$

so that

$$p(\mathbf{x}, \mathbf{y}) = \frac{1}{(2\pi)^{\frac{k+l}{2}} \det^{\frac{1}{2}}(\mathbf{C})} \exp\left[-\frac{1}{2}\left(\begin{bmatrix} \mathbf{x} - E(\mathbf{x}) \\ \mathbf{y} - E(\mathbf{y}) \end{bmatrix}\right)^T \mathbf{C}^{-1} \left(\begin{bmatrix} \mathbf{x} - E(\mathbf{x}) \\ \mathbf{y} - E(\mathbf{y}) \end{bmatrix}\right)\right],$$

then the conditional PDF $p(\mathbf{y}|\mathbf{x})$ is also Gaussian and

$$\begin{aligned} E(\mathbf{y}|\mathbf{x}) &= E(\mathbf{y}) + \mathbf{C}_{yx}\mathbf{C}_{xx}^{-1}(\mathbf{x} - E(\mathbf{x})) && (10.24) \\ \mathbf{C}_{y|x} &= \mathbf{C}_{yy} - \mathbf{C}_{yx}\mathbf{C}_{xx}^{-1}\mathbf{C}_{xy}. && (10.25) \end{aligned}$$

Note that the covariance matrix of the conditional PDF does not depend on $\mathbf{x}$, although this property is not generally true. This will be useful later. As in the bivariate case, the prior PDF $p(\mathbf{y})$ is Gaussian, as well as the posterior PDF $p(\mathbf{y}|\mathbf{x})$. The question may arise as to when the jointly Gaussian assumption may be made. In the next section we examine an important data model for which this holds, the Bayesian linear model.

10.6 Bayesian Linear Model

Recall that in Example 10.1 the data model was

$$x[n] = A + w[n] \qquad n = 0, 1, \ldots, N-1$$

where $A \sim \mathcal{N}(\mu_A, \sigma_A^2)$, and $w[n]$ is WGN independent of A. In terms of vectors we have the equivalent data model

$$\mathbf{x} = \mathbf{1}A + \mathbf{w}.$$

This appears to be of the form of the linear model described in Chapter 4 except for the assumption that A is a random variable. It should not then be surprising that a Bayesian equivalent of the general linear model can be defined. In particular let the data be modeled as

$$\mathbf{x} = \mathbf{H}\boldsymbol{\theta} + \mathbf{w} \tag{10.26}$$

where $\mathbf{x}$ is an $N \times 1$ data vector, $\mathbf{H}$ is a known $N \times p$ matrix, $\boldsymbol{\theta}$ is a $p \times 1$ random vector with prior PDF $\mathcal{N}(\boldsymbol{\mu}_\theta, \mathbf{C}_\theta)$, and $\mathbf{w}$ is an $N \times 1$ noise vector with PDF $\mathcal{N}(\mathbf{0}, \mathbf{C}_w)$ and independent of $\boldsymbol{\theta}$. This data model is termed the *Bayesian general linear model.* It differs from the classical general linear model in that $\boldsymbol{\theta}$ is modeled as a random variable with a Gaussian prior PDF. It will be of interest in deriving Bayesian estimators to have an explicit expression for the posterior PDF $p(\boldsymbol{\theta}|\mathbf{x})$. From Theorem 10.2 we know that if $\mathbf{x}$ and $\boldsymbol{\theta}$ are jointly Gaussian, then the posterior PDF is also Gaussian. Hence, it only remains to verify that this is indeed the case. Let $\mathbf{z} = [\mathbf{x}^T\ \boldsymbol{\theta}^T]^T$, so that from (10.26) we have

$$\begin{aligned} \mathbf{z} &= \begin{bmatrix} \mathbf{H}\boldsymbol{\theta} + \mathbf{w} \\ \boldsymbol{\theta} \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{H} & \mathbf{I} \\ \mathbf{I} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{\theta} \\ \mathbf{w} \end{bmatrix} \end{aligned}$$

where the identity matrices are of dimension $N \times N$ (upper right) and $p \times p$ (lower left), and $\mathbf{0}$ is an $N \times N$ matrix of zeros. Since $\boldsymbol{\theta}$ and $\mathbf{w}$ are independent of each other and each one is Gaussian, they are jointly Gaussian. Furthermore, because $\mathbf{z}$ is a linear

transformation of a Gaussian vector, it too is Gaussian. Hence, Theorem 10.2 applies directly, and we need only determine the mean and covariance of the posterior PDF. We identify $\mathbf{x}$ as $\mathbf{H}\boldsymbol{\theta}+\mathbf{w}$ and $\mathbf{y}$ as $\boldsymbol{\theta}$ to obtain the means

$$\begin{aligned} E(\mathbf{x}) &= E(\mathbf{H}\boldsymbol{\theta}+\mathbf{w}) = \mathbf{H}E(\boldsymbol{\theta}) = \mathbf{H}\boldsymbol{\mu}_\theta \\ E(\mathbf{y}) &= E(\boldsymbol{\theta}) = \boldsymbol{\mu}_\theta \end{aligned}$$

and covariances

$$\begin{aligned} \mathbf{C}_{xx} &= E\left[(\mathbf{x}-E(\mathbf{x}))(\mathbf{x}-E(\mathbf{x}))^T\right] \\ &= E\left[(\mathbf{H}\boldsymbol{\theta}+\mathbf{w}-\mathbf{H}\boldsymbol{\mu}_\theta)(\mathbf{H}\boldsymbol{\theta}+\mathbf{w}-\mathbf{H}\boldsymbol{\mu}_\theta)^T\right] \\ &= E\left[(\mathbf{H}(\boldsymbol{\theta}-\boldsymbol{\mu}_\theta)+\mathbf{w})(\mathbf{H}(\boldsymbol{\theta}-\boldsymbol{\mu}_\theta)+\mathbf{w})^T\right] \\ &= \mathbf{H}E\left[(\boldsymbol{\theta}-\boldsymbol{\mu}_\theta)(\boldsymbol{\theta}-\boldsymbol{\mu}_\theta)^T\right]\mathbf{H}^T + E(\mathbf{w}\mathbf{w}^T) \\ &= \mathbf{H}\mathbf{C}_\theta\mathbf{H}^T + \mathbf{C}_w \end{aligned}$$

recalling that $\boldsymbol{\theta}$ and $\mathbf{w}$ are independent. Also, the cross-covariance matrix is

$$\begin{aligned} \mathbf{C}_{yx} &= E\left[(\mathbf{y}-E(\mathbf{y}))(\mathbf{x}-E(\mathbf{x}))^T\right] \\ &= E\left[(\boldsymbol{\theta}-\boldsymbol{\mu}_\theta)(\mathbf{H}(\boldsymbol{\theta}-\boldsymbol{\mu}_\theta)+\mathbf{w})^T\right] \\ &= E\left[(\boldsymbol{\theta}-\boldsymbol{\mu}_\theta)(\mathbf{H}(\boldsymbol{\theta}-\boldsymbol{\mu}_\theta))^T\right] \\ &= \mathbf{C}_\theta\mathbf{H}^T. \end{aligned}$$

We can now summarize our results for the Bayesian general linear model.

Theorem 10.3 (Posterior PDF for the Bayesian General Linear Model) *If the observed data* $\mathbf{x}$ *can be modeled as*

$$\mathbf{x} = \mathbf{H}\boldsymbol{\theta}+\mathbf{w} \tag{10.27}$$

where $\mathbf{x}$ *is an* $N\times 1$ *data vector,* $\mathbf{H}$ *is a known* $N\times p$ *matrix,* $\boldsymbol{\theta}$ *is a* $p\times 1$ *random vector with prior PDF* $\mathcal{N}(\boldsymbol{\mu}_\theta,\mathbf{C}_\theta)$*, and* $\mathbf{w}$ *is an* $N\times 1$ *noise vector with PDF* $\mathcal{N}(\mathbf{0},\mathbf{C}_w)$ *and independent of* $\boldsymbol{\theta}$*, then the posterior PDF* $p(\boldsymbol{\theta}|\mathbf{x})$ *is Gaussian with mean*

$$E(\boldsymbol{\theta}|\mathbf{x}) = \boldsymbol{\mu}_\theta + \mathbf{C}_\theta\mathbf{H}^T(\mathbf{H}\mathbf{C}_\theta\mathbf{H}^T+\mathbf{C}_w)^{-1}(\mathbf{x}-\mathbf{H}\boldsymbol{\mu}_\theta) \tag{10.28}$$

and covariance

$$\mathbf{C}_{\theta|x} = \mathbf{C}_\theta - \mathbf{C}_\theta\mathbf{H}^T(\mathbf{H}\mathbf{C}_\theta\mathbf{H}^T+\mathbf{C}_w)^{-1}\mathbf{H}\mathbf{C}_\theta. \tag{10.29}$$

In contrast to the classical general linear model, $\mathbf{H}$ need not be full rank to ensure the invertibility of $\mathbf{H}\mathbf{C}_\theta\mathbf{H}^T+\mathbf{C}_w$. We illustrate the use of these formulas by applying them to Example 10.1.

Example 10.2 - DC Level in WGN - Gaussian Prior PDF (continued)

Since $x[n] = A+w[n]$ for $n=0,1,\ldots,N-1$ with $A\sim\mathcal{N}(\mu_A,\sigma_A^2)$ and $w[n]$ is WGN with variance σ^2 and independent of A, we have the Bayesian general linear model

$$\mathbf{x} = \mathbf{1}A+\mathbf{w}.$$

According to Theorem 10.3, $p(A|\mathbf{x})$ is Gaussian and

$$E(A|\mathbf{x}) = \mu_A + \sigma_A^2\mathbf{1}^T(\mathbf{1}\sigma_A^2\mathbf{1}^T+\sigma^2\mathbf{I})^{-1}(\mathbf{x}-\mathbf{1}\mu_A).$$

Using Woodbury's identity (see Appendix 1)

$$\left(\mathbf{I}+\frac{\sigma_A^2}{\sigma^2}\mathbf{1}\mathbf{1}^T\right)^{-1} = \mathbf{I}-\frac{\frac{\sigma_A^2}{\sigma^2}\mathbf{1}\mathbf{1}^T}{1+N\frac{\sigma_A^2}{\sigma^2}} \tag{10.30}$$

so that

$$\begin{aligned} E(A|\mathbf{x}) &= \mu_A + \frac{\sigma_A^2}{\sigma^2}\mathbf{1}^T\left(\mathbf{I}-\frac{\mathbf{1}\mathbf{1}^T}{N+\frac{\sigma^2}{\sigma_A^2}}\right)(\mathbf{x}-\mathbf{1}\mu_A) \\ &= \mu_A + \frac{\sigma_A^2}{\sigma^2}\left(\mathbf{1}^T-\frac{N}{N+\frac{\sigma^2}{\sigma_A^2}}\mathbf{1}^T\right)(\mathbf{x}-\mathbf{1}\mu_A) \\ &= \mu_A + \frac{\sigma_A^2}{\sigma^2}\left(1-\frac{N}{N+\frac{\sigma^2}{\sigma_A^2}}\right)(N\bar{x}-N\mu_A) \\ &= \mu_A + \frac{N}{N+\frac{\sigma^2}{\sigma_A^2}}(\bar{x}-\mu_A) \\ &= \mu_A + \frac{\sigma_A^2}{\sigma_A^2+\frac{\sigma^2}{N}}(\bar{x}-\mu_A). \end{aligned} \tag{10.31}$$

It is interesting to note that in this form the MMSE estimator resembles a "sequential"-type estimator (see Section 8.7). The estimator with no data or $\hat{A}=\mu_A$ is corrected by the error between the data estimator $\bar{x}$ and the "previous" estimate μ_A. The "gain factor" $\sigma_A^2/(\sigma_A^2+\sigma^2/N)$ depends on our confidence in the previous estimate and the current data. We will see later that a sequential MMSE estimator can be defined and will have just these properties. Finally, with some more algebra (10.11) can be obtained.

To find the posterior variance we use (10.29), so that

$$\text{var}(A|\mathbf{x}) = \sigma_A^2 - \sigma_A^2\mathbf{1}^T(\mathbf{1}\sigma_A^2\mathbf{1}^T+\sigma^2\mathbf{I})^{-1}\mathbf{1}\sigma_A^2.$$

Using (10.30) once again, we obtain

$$\begin{aligned}\text{var}(A|\mathbf{x}) &= \sigma_A^2 - \frac{\sigma_A^2}{\sigma^2}\mathbf{1}^T\left(\mathbf{I} - \frac{\mathbf{1}\mathbf{1}^T}{N + \frac{\sigma^2}{\sigma_A^2}}\right)\mathbf{1}\sigma_A^2 \\ &= \sigma_A^2 - \frac{\sigma_A^4}{\sigma^2}\left(N - \frac{N^2}{N + \frac{\sigma^2}{\sigma_A^2}}\right) \\ &= \frac{\frac{\sigma^2}{N}\sigma_A^2}{\sigma_A^2 + \frac{\sigma^2}{N}}\end{aligned}$$

which is just (10.12). ◇

In the succeeding chapters we will make extensive use of the Bayesian linear model. For future reference we point out that the mean (10.28) and covariance (10.29) of the posterior PDF can be expressed in alternative forms as (see Problem 10.13)

$$E(\boldsymbol{\theta}|\mathbf{x}) = \boldsymbol{\mu}_\theta + \left(\mathbf{C}_\theta^{-1} + \mathbf{H}^T\mathbf{C}_w^{-1}\mathbf{H}\right)^{-1}\mathbf{H}^T\mathbf{C}_w^{-1}(\mathbf{x} - \mathbf{H}\boldsymbol{\mu}_\theta) \tag{10.32}$$

and

$$\mathbf{C}_{\theta|x} = \left(\mathbf{C}_\theta^{-1} + \mathbf{H}^T\mathbf{C}_w^{-1}\mathbf{H}\right)^{-1} \tag{10.33}$$

or

$$\mathbf{C}_{\theta|x}^{-1} = \mathbf{C}_\theta^{-1} + \mathbf{H}^T\mathbf{C}_w^{-1}\mathbf{H}. \tag{10.34}$$

The latter expression is particularly interesting. For the previous example we would have

$$\begin{aligned}\frac{1}{\text{var}(A|\mathbf{x})} &= \frac{1}{\sigma_A^2} + \mathbf{1}^T(\sigma^2\mathbf{I})^{-1}\mathbf{1} \\ &= \frac{1}{\sigma_A^2} + \frac{1}{\frac{\sigma^2}{N}}.\end{aligned}$$

This form lends itself to the interpretation that the "information" or reciprocal of the variance of the prior knowledge $1/\sigma_A^2$ and the "information" of the data $1/(\sigma^2/N)$ *add* to yield the information embodied in the posterior PDF.

10.7 Nuisance Parameters

Many estimation problems are characterized by a set of unknown parameters, of which we are really interested only in a subset. The remaining parameters, which serve only

to complicate the problem, are referred to as *nuisance* parameters. Such would be the case if for a DC level in WGN, we were interested in estimating σ^2 but A was unknown. The DC level A would be the nuisance parameter. If we assume that the parameters are deterministic, as in the classical estimation approach, then in general we have no alternative but to estimate σ^2 *and* A. In the Bayesian approach we can rid ourselves of nuisance parameters by "integrating them out." Suppose the unknown parameters to be estimated are $\boldsymbol{\theta}$ and some additional nuisance parameters $\boldsymbol{\alpha}$ are present. Then, if $p(\boldsymbol{\theta}, \boldsymbol{\alpha}|\mathbf{x})$ denotes the posterior PDF, we can determine the posterior PDF of $\boldsymbol{\theta}$ *only* as

$$p(\boldsymbol{\theta}|\mathbf{x}) = \int p(\boldsymbol{\theta}, \boldsymbol{\alpha}|\mathbf{x})\, d\boldsymbol{\alpha}. \tag{10.35}$$

We can also express this as

$$p(\boldsymbol{\theta}|\mathbf{x}) = \frac{p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})\, d\boldsymbol{\theta}} \tag{10.36}$$

where

$$p(\mathbf{x}|\boldsymbol{\theta}) = \int p(\mathbf{x}|\boldsymbol{\theta}, \boldsymbol{\alpha})p(\boldsymbol{\alpha}|\boldsymbol{\theta})\, d\boldsymbol{\alpha}. \tag{10.37}$$

If we furthermore assume that the nuisance parameters are independent of the desired parameters, then (10.37) reduces to

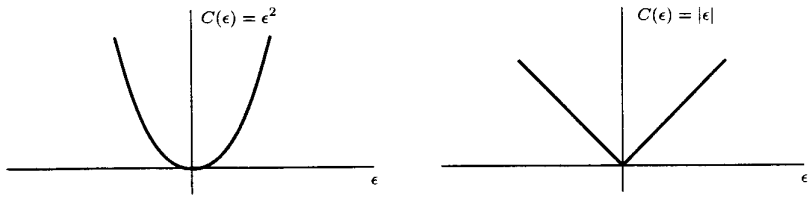
$$p(\mathbf{x}|\boldsymbol{\theta}) = \int p(\mathbf{x}|\boldsymbol{\theta}, \boldsymbol{\alpha})p(\boldsymbol{\alpha})\, d\boldsymbol{\alpha}. \tag{10.38}$$

We observe that the nuisance parameters are first integrated out of the conditional PDF $p(\mathbf{x}|\boldsymbol{\theta}, \boldsymbol{\alpha})$ and then the posterior PDF is found as usual by Bayes' theorem. If a MMSE estimator is desired, we need only determine the mean of the posterior PDF. The nuisance parameters no longer enter into the problem. Of course, their presence will affect the final estimator since from (10.38) $p(\mathbf{x}|\boldsymbol{\theta})$ depends on $p(\boldsymbol{\alpha})$. From a theoretical viewpoint, the Bayesian approach does not suffer from the problems of classical estimators in which nuisance parameters may invalidate an estimator. We now illustrate the approach with an example.

Example 10.3 - Scaled Covariance Matrix

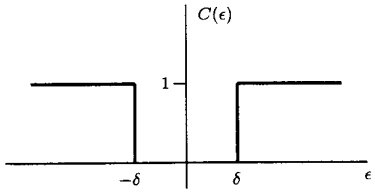
Assume that we observe the $N \times 1$ data vector $\mathbf{x}$ whose conditional PDF $p(\mathbf{x}|\boldsymbol{\theta}, \sigma^2)$ is $\mathcal{N}(\mathbf{0}, \sigma^2\mathbf{C}(\boldsymbol{\theta}))$. (The reader should not confuse $\mathbf{C}(\boldsymbol{\theta})$, the scaled covariance matrix of $\mathbf{x}$, with $\mathbf{C}_\theta$, the covariance matrix of $\boldsymbol{\theta}$.) The parameter $\boldsymbol{\theta}$ is to be estimated, and σ^2 is to be regarded as a nuisance parameter. The covariance matrix depends on $\boldsymbol{\theta}$ in some unspecified manner. We assign the prior PDF to σ^2 of

$$p(\sigma^2) = \begin{cases} \frac{\lambda \exp\left(-\frac{\lambda}{\sigma^2}\right)}{\sigma^4} & \sigma^2 > 0 \\ 0 & \sigma^2 < 0 \end{cases} \tag{10.39}$$



(a) Quadratic error

(b) Absolute error



(c) Hit-or-miss error

Figure 11.1 Examples of cost function

an error ellipse as discussed in Example 11.7. Finally, Theorem 11.1 summarizes the MMSE estimator and its performance for the important Bayesian linear model.

11.3 Risk Functions

Previously, we had derived the MMSE estimator by minimizing $E[(\theta - \hat{\theta})^2]$, where the expectation is with respect to the PDF $p(\mathbf{x}, \theta)$. If we let $\epsilon = \theta - \hat{\theta}$ denote the error of the estimator for a particular realization of $\mathbf{x}$ and θ, and also let $C(\epsilon) = \epsilon^2$, then the MSE criterion minimizes $E[C(\epsilon)]$. The deterministic function $C(\epsilon)$ as shown in Figure 11.1a is termed the *cost function*. It is noted that large errors are particularly costly. Also, the average cost or $E[C(\epsilon)]$ is termed the *Bayes risk* $\mathcal{R}$ or

$$\mathcal{R} = E[C(\epsilon)] \tag{11.1}$$

and measures the performance of a given estimator. If $C(\epsilon) = \epsilon^2$, then the cost function is quadratic and the Bayes risk is just the MSE. Of course, there is no need to restrict ourselves to quadratic cost functions, although from a mathematical tractibility standpoint, they are highly desirable. Other possible cost functions are shown in Figures 11.1b and 11.1c. In Figure 11.1b we have

$$C(\epsilon) = |\epsilon|. \tag{11.2}$$

This cost function penalizes errors proportionally. In Figure 11.1c the "hit-or-miss" cost function is displayed. It assigns no cost for small errors and a cost of 1 for all

errors in excess of a threshold error or

$$C(\epsilon) = \begin{cases} 0 & |\epsilon| < \delta \\ 1 & |\epsilon| > \delta \end{cases} \tag{11.3}$$

where $\delta > 0$. If δ is small, we can think of this cost function as assigning the same penalty for any error (a "miss") and no penalty for no error (a "hit"). Note that in all three cases the cost function is symmetric in ϵ, reflecting the implicit assumption that positive errors are just as bad as negative errors. Of course, in general this need not be the case.

We already have seen that the Bayes risk is minimized for a quadratic cost function by the MMSE estimator $\hat{\theta} = E(\theta|\mathbf{x})$. We now determine the optimal estimators for the other cost functions. The Bayes risk $\mathcal{R}$ is

$$\begin{aligned} \mathcal{R} &= E[C(\epsilon)] \\ &= \int\int C(\theta - \hat{\theta}) p(\mathbf{x}, \theta)\, d\mathbf{x}\, d\theta \\ &= \int \left[\int C(\theta - \hat{\theta}) p(\theta|\mathbf{x})\, d\theta \right] p(\mathbf{x})\, d\mathbf{x}. \end{aligned} \tag{11.4}$$

As we did for the MMSE case in Chapter 10, we will attempt to minimize the inner integral for each $\mathbf{x}$. By holding $\mathbf{x}$ fixed $\hat{\theta}$ becomes a scalar variable. First, considering the absolute error cost function, we have for the inner integral of (11.4)

$$\begin{aligned} g(\hat{\theta}) &= \int |\theta - \hat{\theta}| p(\theta|\mathbf{x})\, d\theta \\ &= \int_{-\infty}^{\hat{\theta}} (\hat{\theta} - \theta) p(\theta|\mathbf{x})\, d\theta + \int_{\hat{\theta}}^{\infty} (\theta - \hat{\theta}) p(\theta|\mathbf{x})\, d\theta. \end{aligned}$$

To differentiate with respect to $\hat{\theta}$ we make use of Leibnitz's rule:

$$\begin{aligned} \frac{\partial}{\partial u} \int_{\phi_1(u)}^{\phi_2(u)} h(u, v)\, dv &= \int_{\phi_1(u)}^{\phi_2(u)} \frac{\partial h(u, v)}{\partial u}\, dv + \frac{d\phi_2(u)}{du} h(u, \phi_2(u)) \\ &\quad - \frac{d\phi_1(u)}{du} h(u, \phi_1(u)). \end{aligned}$$

Letting $h(\hat{\theta}, \theta) = (\hat{\theta} - \theta) p(\theta|\mathbf{x})$ for the first integral, we have

$$h(u, \phi_2(u)) = h(\hat{\theta}, \hat{\theta}) = (\hat{\theta} - \hat{\theta}) p(\hat{\theta}|\mathbf{x}) = 0$$

and $d\phi_1(u)/du = 0$ since the lower limit does not depend on u. Similarly, for the second integral the corresponding terms are zero. Hence, we can differentiate the integrand *only* to yield

$$\frac{dg(\hat{\theta})}{d\hat{\theta}} = \int_{-\infty}^{\hat{\theta}} p(\theta|\mathbf{x})\, d\theta - \int_{\hat{\theta}}^{\infty} p(\theta|\mathbf{x})\, d\theta = 0$$

or

$$\int_{-\infty}^{\hat{\theta}} p(\theta|\mathbf{x})\,d\theta = \int_{\hat{\theta}}^{\infty} p(\theta|\mathbf{x})\,d\theta.$$

By definition $\hat{\theta}$ is the *median* of the posterior PDF or the point for which $\Pr\{\theta \leq \hat{\theta}|\mathbf{x}\} = 1/2$.

For the "hit-or-miss" cost function we have $\mathcal{C}(\epsilon) = 1$ for $\epsilon > \delta$ and $\epsilon < -\delta$ or for $\theta > \hat{\theta} + \delta$ and $\theta < \hat{\theta} - \delta$, so that the inner integral in (11.4) is

$$g(\hat{\theta}) = \int_{-\infty}^{\hat{\theta}-\delta} 1 \cdot p(\theta|\mathbf{x})\,d\theta + \int_{\hat{\theta}+\delta}^{\infty} 1 \cdot p(\theta|\mathbf{x})\,d\theta.$$

But

$$\int_{-\infty}^{\infty} p(\theta|\mathbf{x})\,d\theta = 1,$$

yielding

$$g(\hat{\theta}) = 1 - \int_{\hat{\theta}-\delta}^{\hat{\theta}+\delta} p(\theta|\mathbf{x})\,d\theta.$$

This is minimized by maximizing

$$\int_{\hat{\theta}-\delta}^{\hat{\theta}+\delta} p(\theta|\mathbf{x})\,d\theta.$$

For δ arbitrarily small this is maximized by choosing $\hat{\theta}$ to correspond to the location of the *maximum* of $p(\theta|\mathbf{x})$. The estimator that minimizes the Bayes risk for the "hit-or-miss" cost function is therefore the *mode* (location of the maximum) of the posterior PDF. It is termed the *maximum a posteriori* (MAP) estimator and will be described in more detail later.

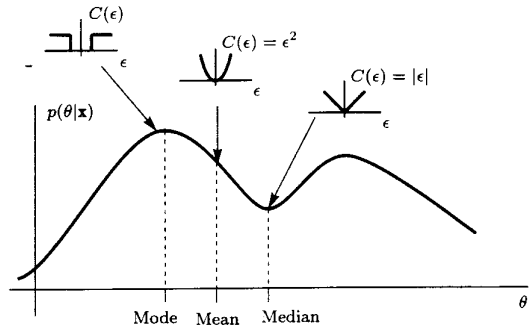
In summary, the estimators that minimize the Bayes risk for the cost functions of Figure 11.1 are the *mean*, *median*, and *mode* of the posterior PDF. This is illustrated in Figure 11.2a. For some posterior PDFs these three estimators are identical. A notable example is the Gaussian posterior PDF

$$p(\theta|\mathbf{x}) = \frac{1}{\sqrt{2\pi\sigma_{\theta|x}^2}} \exp\left[-\frac{1}{2\sigma_{\theta|x}^2}(\theta - \mu_{\theta|x})^2\right].$$

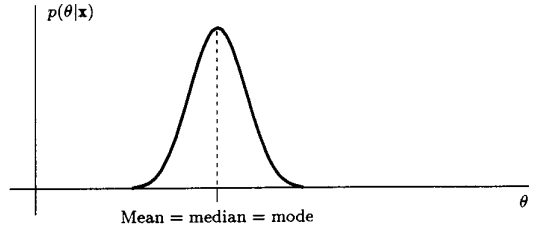
The mean $\mu_{\theta|x}$ is identical to the median (due to the symmetry) and the mode, as illustrated in Figure 11.2b. (See also Problem 11.2.)

11.4 Minimum Mean Square Error Estimators

In Chapter 10 the MMSE estimator was determined to be $E(\theta|\mathbf{x})$ or the mean of the posterior PDF. For this reason it is also commonly referred to as the *conditional mean*



(a) General posterior PDF



(b) Gaussian posterior PDF

Figure 11.2 Estimators for different cost functions

estimator. We continue our discussion of this important estimator by first extending it to the vector parameter case and then studying some of its properties.

If $\boldsymbol{\theta}$ is a vector parameter of dimension $p \times 1$, then to estimate θ_1, for example, we may view the remaining parameters as nuisance parameters (see Chapter 10). If $p(\mathbf{x}|\boldsymbol{\theta})$ is the conditional PDF of the data and $p(\boldsymbol{\theta})$ the prior PDF of the vector parameter, we may obtain the posterior PDF for θ_1 as

$$p(\theta_1|\mathbf{x}) = \int \cdots \int p(\boldsymbol{\theta}|\mathbf{x})\,d\theta_2 \cdots d\theta_p \tag{11.5}$$

where

$$p(\boldsymbol{\theta}|\mathbf{x}) = \frac{p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})\,d\boldsymbol{\theta}}. \tag{11.6}$$

Then, by the same reasoning as in Chapter 10 we have

$$\begin{aligned}\hat{\theta}_1 &= E(\theta_1|\mathbf{x}) \\ &= \int \theta_1 p(\theta_1|\mathbf{x})\, d\theta_1\end{aligned}$$

or in general

$$\hat{\theta}_i = \int \theta_i p(\theta_i|\mathbf{x})\, d\theta_i \qquad i = 1, 2, \ldots, p. \tag{11.7}$$

This is the MMSE estimator that minimizes

$$E[(\theta_i - \hat{\theta}_i)^2] = \int (\theta_i - \hat{\theta}_i)^2 p(\mathbf{x}, \theta_i)\, d\mathbf{x}\, d\theta_i \tag{11.8}$$

or the squared error when averaged with respect to the *marginal* PDF $p(\mathbf{x}, \theta_i)$. Thus, the MMSE estimator for a vector parameter does not entail anything new but only a need to determine the posterior PDF for each parameter. Alternatively, we can express the MMSE estimator for the first parameter from (11.5) as

$$\begin{aligned}\hat{\theta}_1 &= \int \theta_1 p(\theta_1|\mathbf{x})\, d\theta_1 \\ &= \int \theta_1 \left[\int \cdots \int p(\boldsymbol{\theta}|\mathbf{x})\, d\theta_2 \ldots d\theta_p \right] d\theta_1 \\ &= \int \theta_1 p(\boldsymbol{\theta}|\mathbf{x})\, d\boldsymbol{\theta}\end{aligned}$$

or in general

$$\hat{\theta}_i = \int \theta_i p(\boldsymbol{\theta}|\mathbf{x})\, d\boldsymbol{\theta} \qquad i = 1, 2, \ldots, p.$$

In vector form we have

$$\begin{aligned}\hat{\boldsymbol{\theta}} &= \begin{bmatrix} \int \theta_1 p(\boldsymbol{\theta}|\mathbf{x})\, d\boldsymbol{\theta} \\ \int \theta_2 p(\boldsymbol{\theta}|\mathbf{x})\, d\boldsymbol{\theta} \\ \vdots \\ \int \theta_p p(\boldsymbol{\theta}|\mathbf{x})\, d\boldsymbol{\theta} \end{bmatrix} \\ &= \int \boldsymbol{\theta} p(\boldsymbol{\theta}|\mathbf{x})\, d\boldsymbol{\theta} \qquad (11.9) \\ &= E(\boldsymbol{\theta}|\mathbf{x}) \qquad (11.10)\end{aligned}$$

where the expectation is with respect to the posterior PDF of the *vector parameter* or $p(\boldsymbol{\theta}|\mathbf{x})$. Note that the vector MMSE estimator $E(\boldsymbol{\theta}|\mathbf{x})$ *minimizes the MSE for each component of the unknown vector parameter*, or $[\hat{\boldsymbol{\theta}}]_i = [E(\boldsymbol{\theta}|\mathbf{x})]_i$ minimizes $E[(\theta_i - \hat{\theta}_i)^2]$. This follows from the derivation.

As discussed in Chapter 10, the minimum Bayesian MSE for a scalar parameter is the posterior PDF variance when averaged over the PDF of $\mathbf{x}$ (see (10.13)). This is because

$$\text{Bmse}(\hat{\theta}_1) = E[(\theta_1 - \hat{\theta}_1)^2] = \int (\theta_1 - \hat{\theta}_1)^2 p(\mathbf{x}, \theta_1)\, d\theta_1 d\mathbf{x}$$

and since $\hat{\theta}_1 = E(\theta_1|\mathbf{x})$, we have

$$\begin{aligned}\text{Bmse}(\hat{\theta}_1) &= \int \left[\int (\theta_1 - E(\theta_1|\mathbf{x}))^2 p(\theta_1|\mathbf{x})\, d\theta_1 \right] p(\mathbf{x})\, d\mathbf{x} \\ &= \int \text{var}(\theta_1|\mathbf{x}) p(\mathbf{x})\, d\mathbf{x}.\end{aligned}$$

Now, however, the posterior PDF can be written as

$$p(\theta_1|\mathbf{x}) = \int \cdots \int p(\boldsymbol{\theta}|\mathbf{x})\, d\theta_2 \ldots d\theta_p$$

so that

$$\text{Bmse}(\hat{\theta}_1) = \int \left[\int (\theta_1 - E(\theta_1|\mathbf{x}))^2 p(\boldsymbol{\theta}|\mathbf{x})\, d\boldsymbol{\theta} \right] p(\mathbf{x})\, d\mathbf{x}. \tag{11.11}$$

The inner integral in (11.11) is the variance of θ_1 for the posterior PDF $p(\boldsymbol{\theta}|\mathbf{x})$. This is just the [1,1] element of $\mathbf{C}_{\theta|x}$, the covariance matrix of the posterior PDF. Hence, in general we have that the minimum Bayesian MSE is

$$\text{Bmse}(\hat{\theta}_i) = \int [\mathbf{C}_{\theta|x}]_{ii} p(\mathbf{x})\, d\mathbf{x} \qquad i = 1, 2, \ldots, p \tag{11.12}$$

where

$$\mathbf{C}_{\theta|x} = E_{\theta|x}\left[(\boldsymbol{\theta} - E(\boldsymbol{\theta}|\mathbf{x}))(\boldsymbol{\theta} - E(\boldsymbol{\theta}|\mathbf{x}))^T \right]. \tag{11.13}$$

An example follows.

Example 11.1 - Bayesian Fourier Analysis

We reconsider Example 4.2, but to simplify the calculations we let $M = 1$ so that our data model becomes

$$x[n] = a\cos 2\pi f_0 n + b \sin 2\pi f_0 n + w[n] \qquad n = 0, 1, \ldots, N-1$$

where f_0 is a multiple of $1/N$, excepting 0 or $1/2$ (for which $\sin 2\pi f_0 n$ is identically zero), and $w[n]$ is WGN with variance σ^2. It is desired to estimate $\boldsymbol{\theta} = [a\, b]^T$. We depart from the classical model by assuming a, b are random variables with prior PDF

$$\boldsymbol{\theta} \sim \mathcal{N}(\mathbf{0}, \sigma_\theta^2 \mathbf{I})$$

and $\boldsymbol{\theta}$ is independent of $w[n]$. This type of model is referred to as a *Rayleigh fading sinusoid* [Van Trees 1968] and is frequently used to represent a sinusoid that has

propagated through a dispersive medium (see also Problem 11.6). To find the MMSE estimator we need to evaluate $E(\boldsymbol{\theta}|\mathbf{x})$. The data model is rewritten as

$$\mathbf{x} = \mathbf{H}\boldsymbol{\theta} + \mathbf{w}$$

where

$$\mathbf{H} = \begin{bmatrix} 1 & 0 \\ \cos 2\pi f_0 & \sin 2\pi f_0 \\ \vdots & \vdots \\ \cos[2\pi f_0(N-1)] & \sin[2\pi f_0(N-1)] \end{bmatrix}$$

which is recognized as the Bayesian linear model. From Theorem 10.3 we can obtain the mean as well as the covariance of the posterior PDF. To do so we let

$$\begin{aligned} \boldsymbol{\mu}_\theta &= \mathbf{0} \\ \mathbf{C}_\theta &= \sigma_\theta^2 \mathbf{I} \\ \mathbf{C}_w &= \sigma^2 \mathbf{I} \end{aligned}$$

to obtain

$$\begin{aligned} \hat{\boldsymbol{\theta}} &= E(\boldsymbol{\theta}|\mathbf{x}) = \sigma_\theta^2 \mathbf{H}^T (\mathbf{H}\sigma_\theta^2 \mathbf{H}^T + \sigma^2 \mathbf{I})^{-1} \mathbf{x} \\ \mathbf{C}_{\theta|x} &= \sigma_\theta^2 \mathbf{I} - \sigma_\theta^2 \mathbf{H}^T (\mathbf{H}\sigma_\theta^2 \mathbf{H}^T + \sigma^2 \mathbf{I})^{-1} \mathbf{H}\sigma_\theta^2. \end{aligned}$$

A somewhat more convenient form is given by (10.32) and (10.33) as

$$\begin{aligned} \hat{\boldsymbol{\theta}} &= E(\boldsymbol{\theta}|\mathbf{x}) = \left(\frac{1}{\sigma_\theta^2}\mathbf{I} + \mathbf{H}^T \frac{1}{\sigma^2}\mathbf{H}\right)^{-1} \mathbf{H}^T \frac{1}{\sigma^2}\mathbf{x} \\ \mathbf{C}_{\theta|x} &= \left(\frac{1}{\sigma_\theta^2}\mathbf{I} + \mathbf{H}^T \frac{1}{\sigma^2}\mathbf{H}\right)^{-1}. \end{aligned}$$

Now, because the columns of $\mathbf{H}$ are orthogonal (due to the choice of frequency), we have (see Example 4.2)

$$\mathbf{H}^T\mathbf{H} = \frac{N}{2}\mathbf{I}$$

and

$$\begin{aligned} \hat{\boldsymbol{\theta}} &= \left(\frac{1}{\sigma_\theta^2}\mathbf{I} + \frac{N}{2\sigma^2}\mathbf{I}\right)^{-1} \frac{\mathbf{H}^T\mathbf{x}}{\sigma^2} \\ &= \frac{\frac{1}{\sigma^2}}{\frac{1}{\sigma_\theta^2} + \frac{N}{2\sigma^2}} \mathbf{H}^T\mathbf{x} \end{aligned}$$

or the MMSE estimator is

$$\begin{aligned} \hat{a} &= \frac{1}{1 + \frac{2\sigma^2/N}{\sigma_\theta^2}} \left[\frac{2}{N}\sum_{n=0}^{N-1} x[n]\cos 2\pi f_0 n\right] \\ \hat{b} &= \frac{1}{1 + \frac{2\sigma^2/N}{\sigma_\theta^2}} \left[\frac{2}{N}\sum_{n=0}^{N-1} x[n]\sin 2\pi f_0 n\right]. \end{aligned}$$

The results differ from the classical case only in the scale factor, and if $\sigma_\theta^2 \gg 2\sigma^2/N$, the two results are identical. This corresponds to little prior knowledge compared to the data knowledge. The posterior covariance matrix is

$$\mathbf{C}_{\theta|x} = \frac{1}{\frac{1}{\sigma_\theta^2} + \frac{N}{2\sigma^2}}\mathbf{I}$$

which does not depend on $\mathbf{x}$. Hence, from (11.12)

$$\begin{aligned} \text{Bmse}(\hat{a}) &= \frac{1}{\frac{1}{\sigma_\theta^2} + \frac{1}{2\sigma^2/N}} \\ \text{Bmse}(\hat{b}) &= \frac{1}{\frac{1}{\sigma_\theta^2} + \frac{1}{2\sigma^2/N}}. \end{aligned}$$

◇

It is interesting to note that in the absence of prior knowledge in the Bayesian linear model the MMSE estimator yields the same form as the MVU estimator for the classical linear model. Many fortuitous circumstances enter into making this so, as described in Problem 11.7. To verify this result note that from (10.32)

$$\hat{\boldsymbol{\theta}} = E(\boldsymbol{\theta}|\mathbf{x}) = \boldsymbol{\mu}_\theta + (\mathbf{C}_\theta^{-1} + \mathbf{H}^T\mathbf{C}_w^{-1}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{C}_w^{-1}(\mathbf{x} - \mathbf{H}\boldsymbol{\mu}_\theta).$$

For no prior knowledge $\mathbf{C}_\theta^{-1} \to \mathbf{0}$, and therefore,

$$\hat{\boldsymbol{\theta}} \to (\mathbf{H}^T\mathbf{C}_w^{-1}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{C}_w^{-1}\mathbf{x} \tag{11.14}$$

which is recognized as the MVU estimator for the general linear model (see Chapter 4).

The MMSE estimator has several useful properties that will be exploited in our study of Kalman filters in Chapter 13 (see also Problems 11.8 and 11.9). First, it commutes over linear (actually affine) transformations. Assume that we wish to estimate $\boldsymbol{\alpha}$ for

$$\boldsymbol{\alpha} = \mathbf{A}\boldsymbol{\theta} + \mathbf{b} \tag{11.15}$$

where $\mathbf{A}$ is a known $r \times p$ matrix and $\mathbf{b}$ is a known $r \times 1$ vector. Then, $\boldsymbol{\alpha}$ is a random vector for which the MMSE estimator is

$$\hat{\boldsymbol{\alpha}} = E(\boldsymbol{\alpha}|\mathbf{x}).$$

Because of the linearity of the expectation operator

$$\begin{aligned} \hat{\boldsymbol{\alpha}} &= E(\mathbf{A}\boldsymbol{\theta} + \mathbf{b}|\mathbf{x}) \\ &= \mathbf{A}E(\boldsymbol{\theta}|\mathbf{x}) + \mathbf{b} \\ &= \mathbf{A}\hat{\boldsymbol{\theta}} + \mathbf{b}. \end{aligned} \tag{11.16}$$

This holds regardless of the joint PDF $p(\mathbf{x}, \boldsymbol{\theta})$. A second important property focuses on the MMSE estimator based on two data vectors $\mathbf{x}_1, \mathbf{x}_2$. We assume that $\boldsymbol{\theta}, \mathbf{x}_1, \mathbf{x}_2$ are jointly Gaussian and the data vectors are independent. The MMSE estimator is

$$\hat{\boldsymbol{\theta}} = E(\boldsymbol{\theta}|\mathbf{x}_1, \mathbf{x}_2).$$

Letting $\mathbf{x} = [\mathbf{x}_1^T \, \mathbf{x}_2^T]^T$, we have from Theorem 10.2

$$\hat{\boldsymbol{\theta}} = E(\boldsymbol{\theta}|\mathbf{x}) = E(\boldsymbol{\theta}) + \mathbf{C}_{\theta x}\mathbf{C}_{xx}^{-1}(\mathbf{x} - E(\mathbf{x})). \tag{11.17}$$

Since $\mathbf{x}_1, \mathbf{x}_2$ are independent,

$$\begin{aligned} \mathbf{C}_{xx}^{-1} &= \begin{bmatrix} \mathbf{C}_{x_1x_1} & \mathbf{C}_{x_1x_2} \\ \mathbf{C}_{x_2x_1} & \mathbf{C}_{x_2x_2} \end{bmatrix}^{-1} \\ &= \begin{bmatrix} \mathbf{C}_{x_1x_1} & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_{x_2x_2} \end{bmatrix}^{-1} \\ &= \begin{bmatrix} \mathbf{C}_{x_1x_1}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_{x_2x_2}^{-1} \end{bmatrix} \end{aligned}$$

and also

$$\mathbf{C}_{\theta x} = E\left[\boldsymbol{\theta}\begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix}^T\right] = \begin{bmatrix} \mathbf{C}_{\theta x_1} & \mathbf{C}_{\theta x_2} \end{bmatrix}.$$

It follows from (11.17) that

$$\begin{aligned} \hat{\boldsymbol{\theta}} &= E(\boldsymbol{\theta}) + \begin{bmatrix} \mathbf{C}_{\theta x_1} & \mathbf{C}_{\theta x_2} \end{bmatrix} \begin{bmatrix} \mathbf{C}_{x_1x_1}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_{x_2x_2}^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{x}_1 - E(\mathbf{x}_1) \\ \mathbf{x}_2 - E(\mathbf{x}_2) \end{bmatrix} \\ &= E(\boldsymbol{\theta}) + \mathbf{C}_{\theta x_1}\mathbf{C}_{x_1x_1}^{-1}(\mathbf{x}_1 - E(\mathbf{x}_1)) + \mathbf{C}_{\theta x_2}\mathbf{C}_{x_2x_2}^{-1}(\mathbf{x}_2 - E(\mathbf{x}_2)). \end{aligned}$$

We may interpret the estimator as composed of the prior estimator $E(\boldsymbol{\theta})$ as well as that due to the independent data sets. The MMSE is seen to have an additivity property for *independent* data sets. This result is useful in deriving the sequential MMSE estimator (see Chapter 13).

Finally, in the jointly Gaussian case the MMSE is linear in the data, as can be seen from (11.17). This will allow us to easily determine the PDF of the error, as described in Section 11.6.

11.5 Maximum A Posteriori Estimators

In the MAP estimation approach we choose $\hat{\theta}$ to maximize the posterior PDF or

$$\hat{\theta} = \arg\max_{\theta} p(\theta|\mathbf{x}).$$

This was shown to minimize the Bayes risk for a "hit-or-miss" cost function. In finding the maximum of $p(\theta|\mathbf{x})$ we observe that

$$p(\theta|\mathbf{x}) = \frac{p(\mathbf{x}|\theta)p(\theta)}{p(\mathbf{x})}$$

so an equivalent maximization is of $p(\mathbf{x}|\theta)p(\theta)$. This is reminiscent of the MLE except for the presence of the prior PDF. Hence, the MAP estimator is

$$\hat{\theta} = \arg\max_{\theta} p(\mathbf{x}|\theta)p(\theta) \tag{11.18}$$

or, equivalently,

$$\hat{\theta} = \arg\max_{\theta} \left[\ln p(\mathbf{x}|\theta) + \ln p(\theta)\right]. \tag{11.19}$$

Before extending the MAP estimator to the vector case we give some examples.

Example 11.2 - Exponential PDF

Assume that

$$p(x[n]|\theta) = \begin{cases} \theta\exp(-\theta x[n]) & x[n] > 0 \\ 0 & x[n] < 0 \end{cases}$$

where the $x[n]$'s are conditionally IID, or

$$p(\mathbf{x}|\theta) = \prod_{n=0}^{N-1} p(x[n]|\theta)$$

and the prior PDF is

$$p(\theta) = \begin{cases} \lambda\exp(-\lambda\theta) & \theta > 0 \\ 0 & \theta < 0. \end{cases}$$

Then, the MAP estimator is found by maximizing

$$\begin{aligned} g(\theta) &= \ln p(\mathbf{x}|\theta) + \ln p(\theta) \\ &= \ln\left[\theta^N \exp\left(-\theta\sum_{n=0}^{N-1} x[n]\right)\right] + \ln\left[\lambda\exp(-\lambda\theta)\right] \\ &= N\ln\theta - N\theta\bar{x} + \ln\lambda - \lambda\theta \end{aligned}$$

for $\theta > 0$. Differentiating with respect to θ produces

$$\frac{dg(\theta)}{d\theta} = \frac{N}{\theta} - N\bar{x} - \lambda$$

and setting it equal to zero yields the MAP estimator

$$\hat{\theta} = \frac{1}{\bar{x} + \frac{\lambda}{N}}.$$

12.3 Linear MMSE Estimation

We begin our discussion by assuming a scalar parameter θ is to be estimated based on the data set $\{x[0], x[1], \ldots, x[N-1]\}$ or in vector form $\mathbf{x} = [x[0]\, x[1] \ldots x[N-1]]^T$. The unknown parameter is modeled as the realization of a random variable. We do not assume any specific form for the joint PDF $p(\mathbf{x}, \theta)$, but as we shall see shortly, only a knowledge of the first two moments. That θ may be estimated from $\mathbf{x}$ is due to the assumed statistical dependence of θ on $\mathbf{x}$ as summarized by the joint PDF $p(\mathbf{x}, \theta)$, and in particular, for a linear estimator we rely on the correlation between θ and $\mathbf{x}$. We now consider the class of all linear (actually affine) estimators of the form

$$\hat{\theta} = \sum_{n=0}^{N-1} a_n x[n] + a_N \tag{12.1}$$

and choose the weighting coefficients a_n's to minimize the Bayesian MSE

$$\text{Bmse}(\hat{\theta}) = E\left[(\theta - \hat{\theta})^2\right] \tag{12.2}$$

where the expectation is with respect to the PDF $p(\mathbf{x}, \theta)$. The resultant estimator is termed the *linear minimum mean square error (LMMSE) estimator*. Note that we have included the a_N coefficient to allow for nonzero means of $\mathbf{x}$ and θ. If the means are both zero, then this coefficient may be omitted, as will be shown later.

Before determining the LMMSE estimator we should keep in mind that the estimator will be suboptimal unless the MMSE estimator happens to be linear. Such would be the case, for example, if the Bayesian linear model applied (see Section 10.6). Otherwise, better estimators will exist, although they will be nonlinear (see the introductory example in Section 10.3). Since the LMMSE estimator relies on the correlation between random variables, a parameter uncorrelated with the data cannot be linearly estimated. Consequently, the proposed approach is not always feasible. This is illustrated by the following example. Consider a parameter θ to be estimated based on the single data sample $x[0]$, where $x[0] \sim \mathcal{N}(0, \sigma^2)$. If the parameter to be estimated is the power of the $x[0]$ *realization* or $\theta = x^2[0]$, then a perfect estimator will be

$$\hat{\theta} = x^2[0]$$

since the minimum Bayesian MSE will be zero. This estimator is clearly nonlinear. If, however, we attempt to use a LMMSE estimator or

$$\hat{\theta} = a_0 x[0] + a_1,$$

then the optimal weighting coefficients a_0 and a_1 can be found by minimizing

$$\begin{aligned} \text{Bmse}(\hat{\theta}) &= E\left[(\theta - \hat{\theta})^2\right] \\ &= E\left[(\theta - a_0 x[0] - a_1)^2\right]. \end{aligned}$$

We differentiate this with respect to a_0 and a_1 and set the results equal to zero to produce

$$\begin{aligned} E\left[(\theta - a_0 x[0] - a_1) x[0]\right] &= 0 \\ E(\theta - a_0 x[0] - a_1) &= 0 \end{aligned}$$

or

$$\begin{aligned} a_0 E(x^2[0]) + a_1 E(x[0]) &= E(\theta x[0]) \\ a_0 E(x[0]) + a_1 &= E(\theta). \end{aligned}$$

But $E(x[0]) = 0$ and $E(\theta x[0]) = E(x^3[0]) = 0$, so that

$$\begin{aligned} a_0 &= 0 \\ a_1 &= E(\theta) = E(x^2[0]) = \sigma^2. \end{aligned}$$

Therefore, the LMMSE estimator is $\hat{\theta} = \sigma^2$ and does not depend on the data. This is because θ and $x[0]$ are uncorrelated. The minimum MSE is

$$\begin{aligned} \text{Bmse}(\hat{\theta}) &= E\left[(\theta - \hat{\theta})^2\right] \\ &= E\left[(\theta - \sigma^2)^2\right] \\ &= E\left[(x^2[0] - \sigma^2)^2\right] \\ &= E(x^4[0]) - 2\sigma^2 E(x^2[0]) + \sigma^4 \\ &= 3\sigma^4 - 2\sigma^4 + \sigma^4 \\ &= 2\sigma^4 \end{aligned}$$

as opposed to a minimum MSE of zero for the nonlinear estimator $\theta = x^2[0]$. Clearly, the LMMSE estimator is inappropriate for this problem. Problem 12.1 explores how to modify the LMMSE estimator to make it applicable.

We now derive the optimal weighting coefficients for use in (12.1). Substituting (12.1) into (12.2) and differentiating

$$\frac{\partial}{\partial a_N} E\left[\left(\theta - \sum_{n=0}^{N-1} a_n x[n] - a_N\right)^2\right] = -2E\left[\theta - \sum_{n=0}^{N-1} a_n x[n] - a_N\right].$$

Setting this equal to zero produces

$$a_N = E(\theta) - \sum_{n=0}^{N-1} a_n E(x[n]) \tag{12.3}$$

which as asserted earlier is zero if the means are zero. Continuing, we need to minimize

$$\text{Bmse}(\hat{\theta}) = E\left\{\left[\sum_{n=0}^{N-1} a_n (x[n] - E(x[n])) - (\theta - E(\theta))\right]^2\right\}$$

over the remaining a_n's, where a_N has been replaced by (12.3). Letting $\mathbf{a} = [a_0\, a_1 \ldots a_{N-1}]^T$, we have

$$\begin{aligned} \text{Bmse}(\hat{\theta}) &= E\left\{\left[\mathbf{a}^T(\mathbf{x} - E(\mathbf{x})) - (\theta - E(\theta))\right]^2\right\} \\ &= E\left[\mathbf{a}^T(\mathbf{x} - E(\mathbf{x}))(\mathbf{x} - E(\mathbf{x}))^T\mathbf{a}\right] - E\left[\mathbf{a}^T(\mathbf{x} - E(\mathbf{x}))(\theta - E(\theta))\right] \\ &\quad - E\left[(\theta - E(\theta))(\mathbf{x} - E(\mathbf{x}))^T\mathbf{a}\right] + E\left[(\theta - E(\theta))^2\right] \\ &= \mathbf{a}^T\mathbf{C}_{xx}\mathbf{a} - \mathbf{a}^T\mathbf{C}_{x\theta} - \mathbf{C}_{\theta x}\mathbf{a} + C_{\theta\theta} \end{aligned} \tag{12.4}$$

where $\mathbf{C}_{xx}$ is the $N \times N$ covariance matrix of $\mathbf{x}$, and $\mathbf{C}_{\theta x}$ is the $1 \times N$ cross-covariance vector having the property that $\mathbf{C}_{\theta x}^T = \mathbf{C}_{x\theta}$, and $C_{\theta\theta}$ is the variance of θ. Making use of (4.3) we can minimize (12.4) by taking the gradient to yield

$$\frac{\partial \text{Bmse}(\hat{\theta})}{\partial \mathbf{a}} = 2\mathbf{C}_{xx}\mathbf{a} - 2\mathbf{C}_{x\theta}$$

which when set to zero results in

$$\mathbf{a} = \mathbf{C}_{xx}^{-1}\mathbf{C}_{x\theta}. \tag{12.5}$$

Using (12.3) and (12.5) in (12.1) produces

$$\begin{aligned} \hat{\theta} &= \mathbf{a}^T\mathbf{x} + a_N \\ &= \mathbf{C}_{x\theta}^T\mathbf{C}_{xx}^{-1}\mathbf{x} + E(\theta) - \mathbf{C}_{x\theta}^T\mathbf{C}_{xx}^{-1}E(\mathbf{x}) \end{aligned}$$

or finally the *LMMSE estimator* is

$$\hat{\theta} = E(\theta) + \mathbf{C}_{\theta x}\mathbf{C}_{xx}^{-1}(\mathbf{x} - E(\mathbf{x})). \tag{12.6}$$

Note that it is identical in form to the MMSE estimator for jointly Gaussian $\mathbf{x}$ and θ, as can be verified from (10.24). This is because in the Gaussian case the MMSE estimator happens to be linear, and hence our constraint is automatically satisfied. If the means of θ and $\mathbf{x}$ are zero, then

$$\hat{\theta} = \mathbf{C}_{\theta x}\mathbf{C}_{xx}^{-1}\mathbf{x}. \tag{12.7}$$

The minimum Bayesian MSE is obtained by substituting (12.5) into (12.4) to yield

$$\begin{aligned} \text{Bmse}(\hat{\theta}) &= \mathbf{C}_{x\theta}^T\mathbf{C}_{xx}^{-1}\mathbf{C}_{xx}\mathbf{C}_{xx}^{-1}\mathbf{C}_{x\theta} - \mathbf{C}_{x\theta}^T\mathbf{C}_{xx}^{-1}\mathbf{C}_{x\theta} \\ &\quad - \mathbf{C}_{\theta x}\mathbf{C}_{xx}^{-1}\mathbf{C}_{x\theta} + C_{\theta\theta} \\ &= \mathbf{C}_{\theta x}\mathbf{C}_{xx}^{-1}\mathbf{C}_{x\theta} - 2\mathbf{C}_{\theta x}\mathbf{C}_{xx}^{-1}\mathbf{C}_{x\theta} + C_{\theta\theta} \end{aligned}$$

or finally

$$\text{Bmse}(\hat{\theta}) = C_{\theta\theta} - \mathbf{C}_{\theta x}\mathbf{C}_{xx}^{-1}\mathbf{C}_{x\theta}. \tag{12.8}$$

Again this is identical to that obtained by substituting (10.25) into (11.12). An example follows.

Example 12.1 - DC Level in WGN with Uniform Prior PDF

Consider the introductory example in Chapter 10. The data model is

$$x[n] = A + w[n] \qquad n = 0, 1, \ldots, N-1$$

where $A \sim \mathcal{U}[-A_0, A_0]$, $w[n]$ is WGN with variance σ^2, and A and $w[n]$ are independent. We wish to estimate A. The MMSE estimator cannot be obtained in closed form due to the integration required (see (10.9)). Applying the LMMSE estimator, we first note that $E(A) = 0$, and hence $E(x[n]) = 0$. Since $E(\mathbf{x}) = \mathbf{0}$, the covariances are

$$\begin{aligned} \mathbf{C}_{xx} &= E(\mathbf{x}\mathbf{x}^T) \\ &= E\left[(A\mathbf{1} + \mathbf{w})(A\mathbf{1} + \mathbf{w})^T\right] \\ &= E(A^2)\mathbf{1}\mathbf{1}^T + \sigma^2\mathbf{I} \\ \mathbf{C}_{\theta x} &= E(A\mathbf{x}^T) \\ &= E\left[A(A\mathbf{1} + \mathbf{w})^T\right] \\ &= E(A^2)\mathbf{1}^T \end{aligned}$$

where $\mathbf{1}$ is an $N \times 1$ vector of all ones. Hence, from (12.7)

$$\begin{aligned} \hat{A} &= \mathbf{C}_{\theta x}\mathbf{C}_{xx}^{-1}\mathbf{x} \\ &= \sigma_A^2\mathbf{1}^T(\sigma_A^2\mathbf{1}\mathbf{1}^T + \sigma^2\mathbf{I})^{-1}\mathbf{x} \end{aligned}$$

where we have let $\sigma_A^2 = E(A^2)$. But the form of the estimator is identical to that encountered in Example 10.2 if we let $\mu_A = 0$, so that from (10.31)

$$\hat{A} = \frac{\sigma_A^2}{\sigma_A^2 + \frac{\sigma^2}{N}}\bar{x}.$$

Since $\sigma_A^2 = E(A^2) = (2A_0)^2/12 = A_0^2/3$, the LMMSE estimator of A is

$$\hat{A} = \frac{\frac{A_0^2}{3}}{\frac{A_0^2}{3} + \frac{\sigma^2}{N}}\bar{x}. \tag{12.9}$$

As opposed to the original MMSE estimator which required integration, we have obtained the LMMSE estimator in closed form. Also, note that we did not really need to know that A was uniformly distributed but only its mean and variance, or that $w[n]$ was Gaussian but only that it is white and its variance. Likewise, independence of A and $\mathbf{w}$ was not required, only that they were uncorrelated. In general, all that is required to determine the LMMSE estimator are the first two moments of $p(\mathbf{x}, \theta)$ or

$$\begin{bmatrix} E(\theta) \\ E(\mathbf{x}) \end{bmatrix}, \begin{bmatrix} C_{\theta\theta} & \mathbf{C}_{\theta x} \\ \mathbf{C}_{x\theta} & \mathbf{C}_{xx} \end{bmatrix}.$$

However, we must realize that the LMMSE of (12.9) will be *suboptimal* since it has been *constrained to be linear*. The optimal estimator for this problem is given by (10.9). ◇

threshold is now determined by the prior probabilities of the hypotheses. For equal prior probabilities of the hypotheses, the detector becomes the maximum likelihood detector of (3.14). More generally, the optimal decision rule that minimizes the probability of error is given by the maximum a posteriori probability detector of (3.16). A generalization of the minimum probability of error criterion is the Bayes risk as discussed in Section 3.7 with the detector given by (3.18). For multiple hypothesis testing the decision rule for minimizing the Bayes risk is given by (3.21). Specializing the result to the minimum probability of error criterion leads to the maximum a posteriori probability detector of (3.22) and the maximum likelihood detector of (3.24) for multiple hypothesis testing.

3.3 Neyman-Pearson Theorem

In discussing the Neyman-Pearson (NP) approach to signal detection we will center our discussion around a simple example of hypothesis testing. Assume that we observe a realization of a random variable whose PDF is either $\mathcal{N}(0,1)$ or $\mathcal{N}(1,1)$. The notation $\mathcal{N}(\mu,\sigma^2)$ denotes a Gaussian PDF with mean μ and variance σ^2. We must therefore determine if $\mu = 0$ or $\mu = 1$ based on a single observation $x[0]$. Each possible value of μ can be thought of as a hypothesis so that our problem is to choose among two competing hypotheses. These are summarized as follows:

$$\begin{aligned} \mathcal{H}_0 &: \mu = 0 \\ \mathcal{H}_1 &: \mu = 1 \end{aligned} \tag{3.1}$$

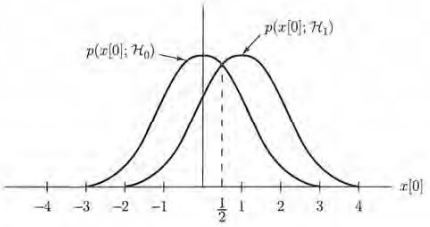


Figure 3.1. PDFs for hypothesis testing problem.

where $\mathcal{H}_0$ is referred to as the *null hypothesis* and $\mathcal{H}_1$ as the *alternative hypothesis*. This problem is known as a *binary* hypothesis test since we must choose between *two* hypotheses. The PDFs under each hypothesis are shown in Figure 3.1, with the difference in means causing the PDF under $\mathcal{H}_1$ to be shifted to the right. On the basis of a single sample it is difficult to determine which PDF generated it.

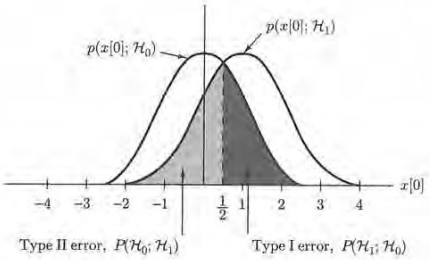


Figure 3.2. Possible hypothesis testing errors and their probabilities.

However, a reasonable approach might be to decide $\mathcal{H}_1$ if $x[0] > 1/2$. This is because if $x[0] > 1/2$, the observed sample is more *likely* if $\mathcal{H}_1$ is true. Or if $x[0] > 1/2$, we have from Figure 3.1 that $p(x[0];\mathcal{H}_1) > p(x[0];\mathcal{H}_0)$. Our detector then compares the observed datum value with $1/2$, the latter being called the *threshold*. Note that with this scheme we can make two types of errors. If we decide $\mathcal{H}_1$ but $\mathcal{H}_0$ is true, we make a *Type I error*. On the other hand, if we decide $\mathcal{H}_0$ but $\mathcal{H}_1$ is true, we make a *Type II error*. These errors are illustrated in Figure 3.2. The notation $P(\mathcal{H}_i;\mathcal{H}_j)$ indicates the probability of deciding $\mathcal{H}_i$ when $\mathcal{H}_j$ is true. For example, $P(\mathcal{H}_1;\mathcal{H}_0) = \Pr\{x[0] > 1/2;\mathcal{H}_0\}$ and is shown as the darker area. These two errors are unavoidable to some extent but may be traded off against each other. To do so we need only change the threshold as shown in Figure 3.3. Clearly, the Type I error probability ($P(\mathcal{H}_1;\mathcal{H}_0)$) is decreased at the expense of increasing the Type II error probability ($P(\mathcal{H}_0;\mathcal{H}_1)$). *It is not possible to reduce both error probabilities simultaneously.* A typical approach then in designing an optimal detector is to hold one error probability fixed while minimizing the other. We choose to constrain $P(\mathcal{H}_1;\mathcal{H}_0)$ to a fixed value, say α. If we view the problem of (3.1) as an attempt to distinguish between the hypotheses

$$\begin{aligned} \mathcal{H}_0 &: x[0] = w[0] \\ \mathcal{H}_1 &: x[0] = s[0] + w[0] \end{aligned}$$

where $s[0] = 1$ and $w[0] \sim \mathcal{N}(0,1)$, then we have the signal detection problem. Deciding $\mathcal{H}_1$ when $\mathcal{H}_0$ is true can be thought of as a false alarm. As a result, $P(\mathcal{H}_1;\mathcal{H}_0)$ is referred to as the *probability of false alarm* and is denoted by P_{FA}. Usually this is a small value, say 10^{-8}, in keeping with the disastrous effects that may ensue. For example, if we falsely say an enemy aircraft is present, we may initiate an attack. To design the optimal detector we then seek to minimize the other error $P(\mathcal{H}_0;\mathcal{H}_1)$ or equivalently to maximize $1 - P(\mathcal{H}_0;\mathcal{H}_1)$. The latter is just

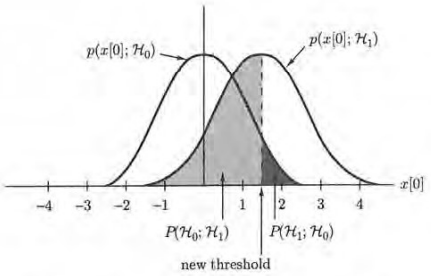


Figure 3.3. Trading off errors by adjusting threshold.

$P(\mathcal{H}_1;\mathcal{H}_1)$ and in keeping with the signal detection problem is called the *probability of detection.* It is denoted by P_D. This setup is termed the *Neyman-Pearson* (NP) approach to hypothesis testing or to signal detection. In summary, we wish to *maximize* $P_D = P(\mathcal{H}_1;\mathcal{H}_1)$ subject to the constraint $P_{FA} = P(\mathcal{H}_1;\mathcal{H}_0) = \alpha$.

Returning to the previous example we can constrain P_{FA} by choosing the threshold γ since

$$\begin{aligned} P_{FA} &= P(\mathcal{H}_1;\mathcal{H}_0) \\ &= \Pr\{x[0] > \gamma;\mathcal{H}_0\} \\ &= \int_\gamma^\infty \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}t^2\right) dt \\ &= Q(\gamma). \end{aligned}$$

As an example, if $P_{FA} = 10^{-3}$, we have $\gamma = 3$. We therefore decide $\mathcal{H}_1$ if $x[0] > 3$. Furthermore, with this choice we have

$$\begin{aligned} P_D &= P(\mathcal{H}_1;\mathcal{H}_1) \\ &= \Pr\{x[0] > \gamma;\mathcal{H}_1\} \\ &= \int_\gamma^\infty \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}(t-1)^2\right] dt \\ &= Q(\gamma - 1) = Q(2) = 0.023. \end{aligned}$$

The question arises as to whether $P_D = 0.023$ is the maximum P_D for this problem. Our choice of the detector that decides $\mathcal{H}_1$ if $x[0] > \gamma$ was just a guess. Might there be a better approach?

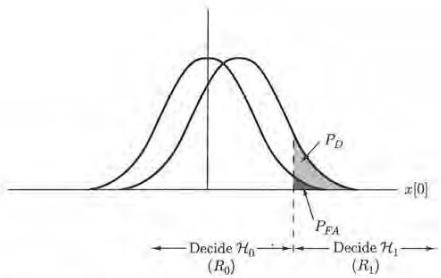


Figure 3.4. Decision regions and probabilities.

Before answering this question we first describe the operation of a detector in more general terms. The goal of a detector is to decide either $\mathcal{H}_0$ or $\mathcal{H}_1$ based on an observed set of data $\{x[0], x[1], \ldots, x[N-1]\}$. This is a mapping from each possible data set value into a decision. For the previous example the *decision regions* are shown in Figure 3.4. A detector then may be thought of as a mapping from the data values into a decision. In particular, let R_1 be the set of values in R^N that map into the decision $\mathcal{H}_1$ or

$$R_1 = \{\mathbf{x} : \text{decide } \mathcal{H}_1 \text{ or reject } \mathcal{H}_0\}.$$

This region is termed the *critical region* in statistics. The set of points in R^N that map into the decision $\mathcal{H}_0$ is the complement set of R_1 or $R_0 = \{\mathbf{x} : \text{decide } \mathcal{H}_0 \text{ or reject } \mathcal{H}_1\}$. Clearly, $R_0 \cup R_1 = R^N$ since R_0 and R_1 *partition* the data space. For the previous example the critical region was $x[0] > 3$. The P_{FA} constraint then becomes

$$P_{FA} = \int_{R_1} p(\mathbf{x};\mathcal{H}_0) d\mathbf{x} = \alpha. \tag{3.2}$$

In statistics, α is termed the *significance level* or *size* of the test. Now there are many sets R_1 that satisfy (3.2) (see Problem 3.2). Our goal is to choose the one that maximizes

$$P_D = \int_{R_1} p(\mathbf{x};\mathcal{H}_1) d\mathbf{x}.$$

In statistics, P_D is called the *power* of the test and the critical region that attains the maximum power is the *best critical region.* See Table 3.1 for a summary of the statistical terminology and the engineering equivalents.

The NP theorem tells us how to choose R_1 if we are given $p(\mathbf{x};\mathcal{H}_0)$, $p(\mathbf{x};\mathcal{H}_1)$, and α.

Statisticians	Engineers
Test statistic ($T(\mathbf{x})$) and threshold (γ)	Detector
Null hypothesis ($\mathcal{H}_0$)	Noise only hypothesis
Alternative hypothesis ($\mathcal{H}_1$)	Signal + noise hypothesis
Critical region	Signal present decision region
Type I error (decide $\mathcal{H}_1$ when $\mathcal{H}_0$ true)	False alarm (FA)
Type II error (decide $\mathcal{H}_0$ when $\mathcal{H}_1$ true)	Miss (M)
Level of significance or size of test (α)	Probability of false alarm (P_{FA})
Probability of Type II error (β)	Probability of miss (P_M)
Power of test ($1-\beta$)	Probability of detection (P_D)

Table 3.1. Cross-Reference of Statistical Terms for Binary Hypothesis Testing

Theorem 3.1 (Neyman-Pearson) *To maximize P_D for a given $P_{FA} = \alpha$ decide $\mathcal{H}_1$ if*

$$L(\mathbf{x}) = \frac{p(\mathbf{x};\mathcal{H}_1)}{p(\mathbf{x};\mathcal{H}_0)} > \gamma \tag{3.3}$$

where the threshold γ is found from

$$P_{FA} = \int_{\{\mathbf{x}:L(\mathbf{x})>\gamma\}} p(\mathbf{x};\mathcal{H}_0)d\mathbf{x} = \alpha.$$

The proof is given in Appendix 3A. The function $L(\mathbf{x})$ is termed the *likelihood ratio* since it indicates for each value of $\mathbf{x}$ the likelihood of $\mathcal{H}_1$ versus the likelihood of $\mathcal{H}_0$. The entire test of (3.3) is called the *likelihood ratio test* (LRT). We next illustrate the NP test with some examples.

Example 3.1 - Introductory Example (continued)

For the hypothesis test of (3.1) we can easily find the NP test. Assume that we require $P_{FA} = 10^{-3}$. Then, from (3.3) we decide $\mathcal{H}_1$ if

$$\frac{p(\mathbf{x};\mathcal{H}_1)}{p(\mathbf{x};\mathcal{H}_0)} = \frac{\frac{1}{\sqrt{2\pi}}\exp\left[-\frac{1}{2}(x[0]-1)^2\right]}{\frac{1}{\sqrt{2\pi}}\exp\left[-\frac{1}{2}x^2[0]\right]} > \gamma$$

or

$$\exp\left[-\frac{1}{2}(x^2[0]-2x[0]+1-x^2[0])\right] > \gamma$$

or finally

$$\exp\left(x[0]-\frac{1}{2}\right) > \gamma. \tag{3.4}$$

At this point we could determine γ from the false alarm constraint

$$P_{FA} = \Pr\left\{\exp\left(x[0]-\frac{1}{2}\right) > \gamma;\mathcal{H}_0\right\} = 10^{-3}.$$

This would require us to find the PDF of $\exp(x[0]-1/2)$. A much simpler approach is to note that the inequality of (3.4) is not changed if we take logarithms of both sides. This is because the logarithm is a monotonically increasing function (see Problem 3.3). Alternatively, since $\gamma > 0$, we can let $\gamma = \exp(\beta)$ so that we decide $\mathcal{H}_1$ if

$$\exp\left(x[0]-\frac{1}{2}\right) > \exp(\beta)$$

or

$$x[0] > \beta + \frac{1}{2} = \ln\gamma + \frac{1}{2}.$$

Letting $\gamma' = \ln\gamma + 1/2$ we decide $\mathcal{H}_1$ if $x[0] > \gamma'$. To explicitly find γ' (or equivalently γ) we use the P_{FA} constraint

$$\begin{aligned} P_{FA} = \Pr\{x[0] > \gamma';\mathcal{H}_0\} &= 10^{-3} \\ \int_{\gamma'}^{\infty} \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}t^2\right)dt &= 10^{-3} \end{aligned}$$

so that $\gamma' = 3$. The NP test is to decide $\mathcal{H}_1$ if $x[0] > 3$. Thus, the detector of the previous example is indeed optimum in the NP sense in that it maximizes P_D. As before we find P_D as follows

$$\begin{aligned} P_D &= \Pr\{x[0] > 3;\mathcal{H}_1\} \\ &= \int_3^{\infty} \frac{1}{\sqrt{2\pi}}\exp\left[-\frac{1}{2}(t-1)^2\right]dt = 0.023. \end{aligned}$$

Note that the detection performance is poor. Although we have satisfied our false alarm constraint, we will only detect the signal a small fraction of the time. To improve the detection performance we can increase P_{FA}, employing the usual tradeoff. For example, if $P_{FA} = 0.5$, then the threshold is found from

$$0.5 = \int_{\gamma'}^{\infty} \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}t^2\right)dt$$

as $\gamma' = 0$. Then

$$\begin{aligned} P_D &= \int_{\gamma'}^{\infty} \frac{1}{\sqrt{2\pi}}\exp\left[-\frac{1}{2}(t-1)^2\right]dt \\ &= \int_0^{\infty} \frac{1}{\sqrt{2\pi}}\exp\left[-\frac{1}{2}(t-1)^2\right]dt \end{aligned}$$

$$= Q\left(\frac{0-1}{1}\right) = Q(-1)$$
$$= 1 - Q(1) = 0.84.$$

(Recall that if $x \sim \mathcal{N}(\mu, \sigma^2)$, the right-tail probability for a threshold γ' is $Q((\gamma' - \mu)/\sigma)$. See Chapter 2.) By changing the threshold we can trade off P_{FA} and P_D. This point is discussed further in the next section.

$\diamond$

Example 3.2 - DC Level in WGN

Now consider the more general signal detection problem

$$\begin{array}{ll} \mathcal{H}_0 : x[n] = w[n] & n = 0, 1, \ldots, N-1 \\ \mathcal{H}_1 : x[n] = A + w[n] & n = 0, 1, \ldots, N-1 \end{array}$$

where the signal is $s[n] = A$ for $A > 0$ and $w[n]$ is WGN with variance σ^2. The previous example is just a special case where $A = 1$, $N = 1$, and $\sigma^2 = 1$. Also, note that the current problem is actually a test of the mean of a multivariate Gaussian PDF. This is because under $\mathcal{H}_0$, $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$ while under $\mathcal{H}_1$, $\mathbf{x} \sim \mathcal{N}(A\mathbf{1}, \sigma^2\mathbf{I})$, where $\mathbf{1}$ is the vector of all ones. Hence, we have equivalently

$$\begin{array}{l} \mathcal{H}_0 : \boldsymbol{\mu} = \mathbf{0} \\ \mathcal{H}_1 : \boldsymbol{\mu} = A\mathbf{1}. \end{array}$$

We will often use this *parameter test of the PDF* interpretation in describing a signal detection problem. Now the NP detector decides $\mathcal{H}_1$ if

$$\frac{\frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp\left[-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}(x[n]-A)^2\right]}{\frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp\left[-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}x^2[n]\right]} > \gamma.$$

Taking the logarithm of both sides results in

$$-\frac{1}{2\sigma^2}\left(-2A\sum_{n=0}^{N-1}x[n] + NA^2\right) > \ln\gamma$$

which simplifies to

$$\frac{A}{\sigma^2}\sum_{n=0}^{N-1}x[n] > \ln\gamma + \frac{NA^2}{2\sigma^2}$$

Since $A > 0$, we have finally

$$\frac{1}{N}\sum_{n=0}^{N-1}x[n] > \frac{\sigma^2}{NA}\ln\gamma + \frac{A}{2} = \gamma'. \tag{3.5}$$

The NP detector compares the *sample mean* $\bar{x} = (1/N)\sum_{n=0}^{N-1}x[n]$ to a threshold γ'. This is intuitively reasonable since $\bar{x}$ may be thought of as an estimate of A. If the estimate is large and positive, then the signal is probably present. How large the estimate must be before we are willing to declare that a signal is present depends upon our concern that noise only may cause a large estimate. To avoid this possibility we adjust γ' to control P_{FA}, with larger threshold values reducing P_{FA} (as well as P_D).

To determine the detection performance we first note that the test statistic $T(\mathbf{x}) = (1/N)\sum_{n=0}^{N-1}x[n]$ is Gaussian under each hypothesis. The means and variances are

$$\begin{aligned} E(T(\mathbf{x}); \mathcal{H}_0) &= E\left(\frac{1}{N}\sum_{n=0}^{N-1}w[n]\right) \\ &= \frac{1}{N}\sum_{n=0}^{N-1}E(w[n]) \\ &= 0. \end{aligned}$$

Similarly, $E(T(\mathbf{x}); \mathcal{H}_1) = A$ and

$$\begin{aligned} \mathrm{var}(T(\mathbf{x}); \mathcal{H}_0) &= \mathrm{var}\left(\frac{1}{N}\sum_{n=0}^{N-1}w[n]\right) \\ &= \frac{1}{N^2}\sum_{n=0}^{N-1}\mathrm{var}(w[n]) \\ &= \frac{\sigma^2}{N}. \end{aligned}$$

Similarly, $\mathrm{var}(T(\mathbf{x}); \mathcal{H}_1) = \sigma^2/N$ where we have noted that the noise samples are uncorrelated. Thus,

$$T(\mathbf{x}) \sim \begin{cases} \mathcal{N}(0, \frac{\sigma^2}{N}) & \text{under } \mathcal{H}_0 \\ \mathcal{N}(A, \frac{\sigma^2}{N}) & \text{under } \mathcal{H}_1. \end{cases}$$

We have then

$$\begin{aligned} P_{FA} &= \Pr\{T(\mathbf{x}) > \gamma'; \mathcal{H}_0\} \\ &= Q\left(\frac{\gamma'}{\sqrt{\sigma^2/N}}\right) \end{aligned} \tag{3.6}$$

and

$$\begin{aligned} P_D &= \Pr\{T(\mathbf{x}) > \gamma'; \mathcal{H}_1\} \\ &= Q\left(\frac{\gamma' - A}{\sqrt{\sigma^2/N}}\right). \end{aligned} \tag{3.7}$$

We can relate P_D to P_{FA} more directly by noting that the Q function is monotonically decreasing since $1 - Q$ is a CDF, which is monotonically increasing. Thus, Q has an inverse that we denote as Q^{-1}. As a result, the threshold is found from (3.6) as

$$\gamma' = \sqrt{\frac{\sigma^2}{N}} Q^{-1}(P_{FA})$$

and

$$\begin{aligned} P_D &= Q\left(\frac{\sqrt{\sigma^2/N}Q^{-1}(P_{FA}) - A}{\sqrt{\sigma^2/N}}\right) \\ &= Q\left(Q^{-1}(P_{FA}) - \sqrt{\frac{NA^2}{\sigma^2}}\right). \end{aligned} \tag{3.8}$$

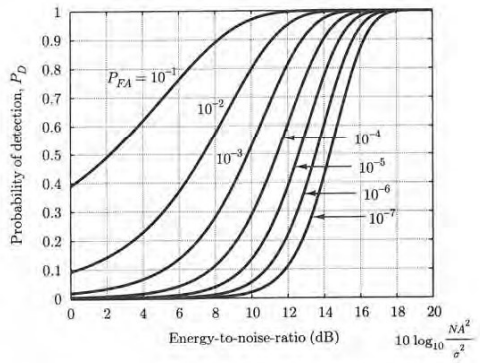


Figure 3.5. Detection performance for DC level in WGN.

It is seen that for a given P_{FA} the detection performance increases monotonically with NA^2/σ^2, which is the *signal energy-to-noise ratio* (ENR). An alternative interpretation is explored in Problem 3.5. The detection performance is shown in Figure 3.5 for various values of P_{FA}. It is sometimes convenient to display the detection curves on normal probability paper (see Chapter 2). This has the effect of straightening the curves when plotted versus $\sqrt{\text{ENR}}$ as shown in Figure 3.6. The advantage is an easier reading of the required ENR for a given P_D, especially for P_D's near one. The disadvantage is that the abscissa values are not in decibels (dB), which is customary in engineering. We will usually employ the former approach.

◇

The previous example illustrates a particularly useful hypothesis testing problem called the *mean-shifted Gauss-Gauss* problem. We observe the value of a test statistic T and decide $\mathcal{H}_1$ if $T > \gamma'$ and $\mathcal{H}_0$ otherwise. The PDF of T is assumed to be

$$T \sim \begin{cases} \mathcal{N}(\mu_0, \sigma^2) & \text{under } \mathcal{H}_0 \\ \mathcal{N}(\mu_1, \sigma^2) & \text{under } \mathcal{H}_1 \end{cases}$$

where $\mu_1 > \mu_0$. Hence, we wish to decide between the two hypotheses that differ by a shift in the mean of T. In the previous example $T = \bar{x}$. For this type of detector

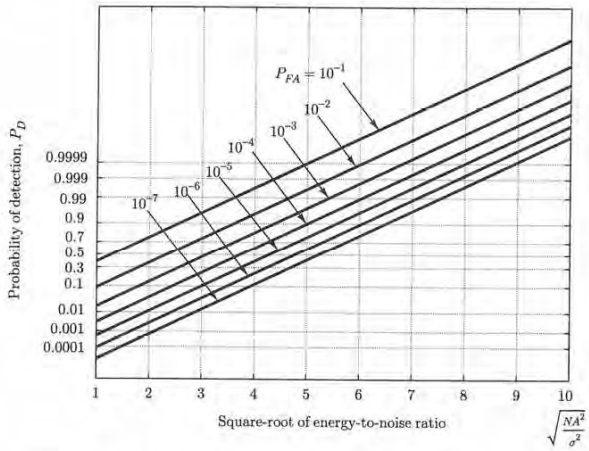


Figure 3.6. Detection performance for DC level in WGN-normal probability paper.

the detection performance is totally characterized by the *deflection coefficient* d^2. This is defined as

$$\begin{aligned} d^2 &= \frac{(E(T;\mathcal{H}_1) - E(T;\mathcal{H}_0))^2}{\mathrm{var}(T;\mathcal{H}_0)} \\ &= \frac{(\mu_1 - \mu_0)^2}{\sigma^2}. \end{aligned} \tag{3.9}$$

In the case when $\mu_0 = 0$, $d^2 = \mu_1^2/\sigma^2$ may be interpreted as a signal-to-noise ratio (SNR). To verify the dependence of detection performance on d^2 we have that

$$\begin{aligned} P_{FA} &= \Pr\{T > \gamma'; \mathcal{H}_0\} \\ &= Q\left(\frac{\gamma' - \mu_0}{\sigma}\right) \\ P_D &= \Pr\{T > \gamma'; \mathcal{H}_1\} \\ &= Q\left(\frac{\gamma' - \mu_1}{\sigma}\right) \\ &= Q\left(\frac{\mu_0 + \sigma Q^{-1}(P_{FA}) - \mu_1}{\sigma}\right) \\ &= Q\left(Q^{-1}(P_{FA}) - \left(\frac{\mu_1 - \mu_0}{\sigma}\right)\right) \end{aligned}$$

and using (3.9) we have

$$P_D = Q\left(Q^{-1}(P_{FA}) - \sqrt{d^2}\right) \tag{3.10}$$

since $\mu_1 > \mu_0$. *The detection performance is therefore monotonic with the deflection coefficient.* We end this section with another example.

Example 3.3 - Change in Variance

This hypothesis testing example illustrates that a change in the variance of a Gaussian statistic can be used to distinguish between two hypotheses. We observe $x[n]$ for $n = 0, 1, \ldots, N-1$, where the $x[n]$'s are independent and identically distributed (IID). The latter qualification means that the first-order PDF for each $x[n]$ is the same. Assume that $x[n] \sim \mathcal{N}(0, \sigma_0^2)$ under $\mathcal{H}_0$ and $x[n] \sim \mathcal{N}(0, \sigma_1^2)$ under $\mathcal{H}_1$, where $\sigma_1^2 > \sigma_0^2$. Then the NP test is to decide $\mathcal{H}_1$ if

$$\frac{\frac{1}{(2\pi\sigma_1^2)^{\frac{N}{2}}} \exp\left(-\frac{1}{2\sigma_1^2}\sum_{n=0}^{N-1} x^2[n]\right)}{\frac{1}{(2\pi\sigma_0^2)^{\frac{N}{2}}} \exp\left(-\frac{1}{2\sigma_0^2}\sum_{n=0}^{N-1} x^2[n]\right)} > \gamma.$$

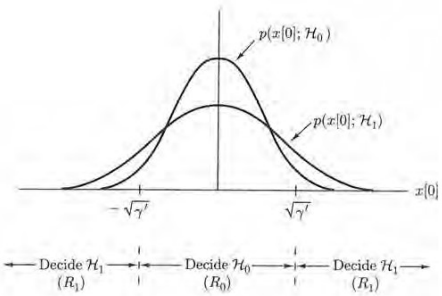


Figure 3.7. Decision regions for change in variance hypothesis test.

Taking logarithms of both sides we have

$$-\frac{1}{2}\left(\frac{1}{\sigma_1^2} - \frac{1}{\sigma_0^2}\right)\sum_{n=0}^{N-1} x^2[n] > \ln\gamma + \frac{N}{2}\ln\frac{\sigma_1^2}{\sigma_0^2}.$$

Since $\sigma_1^2 > \sigma_0^2$, we have

$$\frac{1}{N}\sum_{n=0}^{N-1} x^2[n] > \gamma'$$

where

$$\gamma' = \frac{\frac{2}{N}\ln\gamma + \ln\frac{\sigma_1^2}{\sigma_0^2}}{\frac{1}{\sigma_0^2} - \frac{1}{\sigma_1^2}}.$$

The test statistic is just an estimate of the variance. We decide $\mathcal{H}_1$ if the power in the observed samples is large enough. In particular, if $N = 1$ we have a detector that decides $\mathcal{H}_1$ if $x^2[0] > \gamma'$ or equivalently if $|x[0]| > \sqrt{\gamma'}$. The decision regions are shown in Figure 3.7 and are seen to be plausible. The performance of this detector is examined in Problem 3.9 for $N = 2$ and in more generality in Chapter 5, where we discuss the energy detector.

◇

Note that for the DC level in WGN and the change in variance examples we distinguish between two hypotheses whose PDFs have different parameter values. We do so by estimating the parameter and comparing the estimated value to a threshold. This is not merely a coincidence but is due to the presence of a sufficient statistic [Kay-I 1993, Chapter 5]. In particular, assume that we observe

$\mathbf{x} = [x[0]\, x[1] \ldots x[N-1]]^T$ from a PDF that is parameterized by θ. The PDF is denoted by $p(\mathbf{x};\theta)$. (In the DC level in WGN example $\theta = A$.) We wish to test for the value of θ as

$$\begin{aligned} \mathcal{H}_0 &: \theta = \theta_0 \\ \mathcal{H}_1 &: \theta = \theta_1. \end{aligned}$$

If a sufficient statistic exists for θ, then by the Neyman-Fisher factorization theorem [Kay-I 1993, Chapter 5] we can express the PDF as

$$p(\mathbf{x};\theta) = g(T(\mathbf{x}),\theta)h(\mathbf{x})$$

where $T(\mathbf{x})$ is a sufficient statistic for θ. The NP test, which is

$$\frac{p(\mathbf{x};\theta_1)}{p(\mathbf{x};\theta_0)} > \gamma$$

then becomes

$$\frac{g(T(\mathbf{x}),\theta_1)}{g(T(\mathbf{x}),\theta_0)} > \gamma.$$

Clearly, the test will depend on the data only through $T(\mathbf{x})$. In the DC level in WGN example it can be shown (see Problem 3.10) that the sufficient statistic is $T(\mathbf{x}) = (1/N)\sum_{n=0}^{N-1} x[n]$ while in the change in variance example $T(\mathbf{x}) = (1/N)\sum_{n=0}^{N-1} x^2[n]$ is a sufficient statistic. In essence the sufficient statistic summarizes all the relevant information in the data about θ that is needed to make a decision. (See also Problem 3.11.) Furthermore, if $T(\mathbf{x})$ is an unbiased estimator of θ, then the detector will be based on an *estimate* of the unknown parameter. Unfortunately, sufficient statistics do not always exist, as our final example illustrates.

Example 3.4 - DC Level in NonGaussian Noise

Assume that under $\mathcal{H}_0$ we observe N IID samples $x[n] = w[n]$ for $n = 0, 1, \ldots, N-1$ from the noise PDF $p(w[n])$ while under $\mathcal{H}_1$ we observe $x[n] = A + w[n]$ for $n = 0, 1, \ldots, N-1$. Thus, under $\mathcal{H}_0$ we have

$$p(\mathbf{x};\mathcal{H}_0) = \prod_{n=0}^{N-1} p(x[n])$$

and under $\mathcal{H}_1$ we have

$$p(\mathbf{x};\mathcal{H}_1) = \prod_{n=0}^{N-1} p(x[n]-A).$$

The NP detector decides $\mathcal{H}_1$ if

$$\frac{\prod_{n=0}^{N-1} p(x[n]-A)}{\prod_{n=0}^{N-1} p(x[n])} > \gamma.$$

If the PDF of the noise is a *Gaussian mixture*

$$p(w[n]) = \frac{1}{2}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}w^2[n]\right) + \frac{1}{2}\frac{1}{\sqrt{4\pi}}\exp\left(-\frac{1}{4}w^2[n]\right)$$

then the detector becomes

$$\frac{\prod_{n=0}^{N-1}\frac{1}{2}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}(x[n]-A)^2\right) + \frac{1}{2}\frac{1}{\sqrt{4\pi}}\exp\left(-\frac{1}{4}(x[n]-A)^2\right)}{\prod_{n=0}^{N-1}\frac{1}{2}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}x^2[n]\right) + \frac{1}{2}\frac{1}{\sqrt{4\pi}}\exp\left(-\frac{1}{4}x^2[n]\right)} > \gamma.$$

No further simplification is possible due to the lack of a sufficient statistic for A. We will explore the nonGaussian detection problem further in Chapter 10. ◇

3.4 Receiver Operating Characteristics

An alternative way of summarizing the detection performance of a NP detector is to plot P_D versus P_{FA}. As an example, for the DC level in WGN we have from (3.6), (3.7), and (3.8)

$$\begin{aligned} P_{FA} &= Q\left(\frac{\gamma'}{\sqrt{\sigma^2/N}}\right) \\ P_D &= Q\left(\frac{\gamma' - A}{\sqrt{\sigma^2/N}}\right) \end{aligned}$$

and

$$P_D = Q\left(Q^{-1}(P_{FA}) - \sqrt{d^2}\right)$$

where $d^2 = NA^2/\sigma^2$. The latter is shown in Figure 3.8 for $d^2 = 1$. Each point on the curve corresponds to a value of (P_{FA}, P_D) for a *given* threshold γ'. By adjusting γ' any point on the curve may be obtained. As expected as γ' increases, P_{FA} decreases but so does P_D and vice-versa. This type of performance summary is called the *receiver operating characteristic* (ROC). The ROC should always be above the 45° line (shown dashed in Figure 3.8). This is because the 45° ROC can be attained by a detector that bases its decision on flipping a coin, ignoring all the data. Consider the detector that decides $\mathcal{H}_1$ if a flipped coin comes up a head, where $\Pr\{\text{head}\} = p$. For a tail outcome we decide $\mathcal{H}_0$. Then,

$$\begin{aligned} P_{FA} &= \Pr\{\text{head};\mathcal{H}_0\} \\ P_D &= \Pr\{\text{head};\mathcal{H}_1\}. \end{aligned}$$

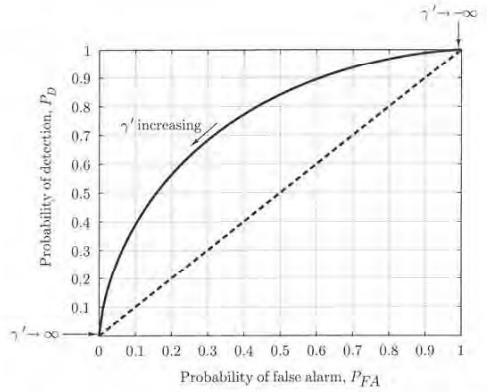


Figure 3.8. Receiver operating characteristics for DC level in WGN ($d^2 = 1$).

But the probability of obtaining a head does not depend upon which hypothesis is true and so $P_{FA} = P_D = p$. This detector then generates the point (p, p) on the ROC. To generate the other points on the 45° line we need only use coins with different p.

As the deflection coefficient increases, a family of ROCs is generated as shown in Figure 3.9. For $d^2 \to \infty$ the ideal ROC is attained or $P_D = 1$ for any P_{FA} (see also Problem 3.12). For $d^2 \to 0$, the 45° lower bound is attained. Other properties of the ROC are discussed in Problem 3.13.

3.5 Irrelevant Data

In many signal detection problems one must decide which data are relevant to the detection problem and which may be discarded. As an example, for a DC level in WGN assume that we observe some extra or reference noise samples $w_R[n]$ for $n = 0, 1, \ldots, N-1$. This could be the output of a second sensor, which is incapable of passing the DC signal. Hence, the observed data set is $\{x[0], x[1], \ldots, x[N-1], w_R[0], w_R[1], \ldots, w_R[N-1]\}$ or in vector form $[\mathbf{x}^T \ \mathbf{w}_R^T]^T$. It might at first appear that $\mathbf{w}_R$ is irrelevant to the detection problem, but that could be a hasty conclusion. If, for example, $x[n] = w[n]$ under $\mathcal{H}_0$, $x[n] = A + w[n]$ for $A > 0$ under $\mathcal{H}_1$, and $w_R[n] = w[n]$ under either hypothesis, then the reference noise samples $w_R[n]$ could be used to *cancel* the corrupting noise $w[n]$. In particular, a detector that decides

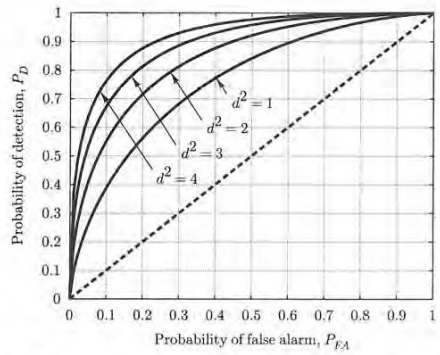


Figure 3.9. Family of receiver operating characteristics for DC level in WGN.

$\mathcal{H}_1$ if

$$T = x[0] - w_R[0] > \frac{A}{2}$$

would yield perfect detection. Under $\mathcal{H}_0$, $T = 0$, while under $\mathcal{H}_1$, $T = A$. Of course, this is an extreme case of perfect statistical dependence. At the other extreme, if $\mathbf{w}_R$ is independent of $\mathbf{x}$ under either hypothesis, then $\mathbf{w}_R$ is irrelevant to the problem. An example of this condition is encountered in the following problem. We observe $\{x[0], x[1], \ldots, x[N-1], x[N], \ldots, x[2N-1]\}$ or $\mathbf{x} = [\mathbf{x}_1^T \ \mathbf{x}_2^T]^T$ where $\mathbf{x}_1$ denotes the first N samples and $\mathbf{x}_2$ the remaining ones. Then, consider the problem

$$\begin{aligned} \mathcal{H}_0 : x[n] &= w[n] \qquad n = 0, 1, \ldots, 2N-1 \\ \mathcal{H}_1 : x[n] &= \begin{cases} A + w[n] & n = 0, 1, \ldots, N-1 \\ w[n] & n = N, N+1, \ldots, 2N-1 \end{cases} \end{aligned}$$

where $w[n]$ is WGN with variance σ^2. The noise samples outside the signal interval $[0, N-1]$ are irrelevant and can be discarded since they are independent of the data samples within the interval. This may also be verified by examining the NP test that decides $\mathcal{H}_1$ if

$$\frac{p(\mathbf{x}_1, \mathbf{x}_2; \mathcal{H}_1)}{p(\mathbf{x}_1, \mathbf{x}_2; \mathcal{H}_0)} > \gamma$$

which becomes

$$\frac{\prod_{n=0}^{N-1}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{1}{2\sigma^2}(x[n]-A)^2\right]\prod_{n=N}^{2N-1}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{1}{2\sigma^2}x^2[n]\right]}{\prod_{n=0}^{N-1}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{1}{2\sigma^2}x[n]^2\right]\prod_{n=N}^{2N-1}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{1}{2\sigma^2}x^2[n]\right]} > \gamma$$

or finally

$$\frac{p(\mathbf{x}_1;\mathcal{H}_1)}{p(\mathbf{x}_1;\mathcal{H}_0)} > \gamma$$

so that $\mathbf{x}_2$ is irrelevant to the detection problem. Thus, in practice, *for detection of signals in WGN we can limit the observation interval to the signal interval.* If, however, the noise is correlated, then for best performance we should also include noise samples from outside the signal interval in our detector.

The preceding discussion can be generalized using the NP theorem. The likelihood ratio is

$$\begin{aligned} L(\mathbf{x}_1,\mathbf{x}_2) &= \frac{p(\mathbf{x}_1,\mathbf{x}_2;\mathcal{H}_1)}{p(\mathbf{x}_1,\mathbf{x}_2;\mathcal{H}_0)} \\ &= \frac{p(\mathbf{x}_2|\mathbf{x}_1;\mathcal{H}_1)p(\mathbf{x}_1;\mathcal{H}_1)}{p(\mathbf{x}_2|\mathbf{x}_1;\mathcal{H}_0)p(\mathbf{x}_1;\mathcal{H}_0)}. \end{aligned}$$

It follows that if

$$p(\mathbf{x}_2|\mathbf{x}_1;\mathcal{H}_1) = p(\mathbf{x}_2|\mathbf{x}_1;\mathcal{H}_0) \tag{3.11}$$

then $L(\mathbf{x}_1,\mathbf{x}_2) = L(\mathbf{x}_1)$ and $\mathbf{x}_2$ is irrelevant to the detection problem. A special case occurs when $\mathbf{x}_1$ and $\mathbf{x}_2$ are independent under either hypothesis and the PDF of $\mathbf{x}_2$ does not depend on the hypothesis. Then, (3.11) holds since $p(\mathbf{x}_2;\mathcal{H}_1) = p(\mathbf{x}_2;\mathcal{H}_0)$. The DC level in WGN with extra noise samples is an example. See also Problems 3.14 and 3.15.

3.6 Minimum Probability of Error

In some detection problems one can reasonably assign probabilities to the various hypotheses. In doing so, we express a prior belief in the likelihood of the hypotheses. An example is in digital communications in which the transmission of a "0" or "1" is equally likely. Then, it is reasonable to assign equal probabilities to $\mathcal{H}_0$ ("0" sent) and $\mathcal{H}_1$ ("1" sent). We say that $P(\mathcal{H}_0) = P(\mathcal{H}_1) = 1/2$, where $P(\mathcal{H}_0)$, $P(\mathcal{H}_1)$ are the *prior probabilities* of the respective hypotheses. In other applications, such as sonar or radar, this is not possible. If one is attempting to detect an enemy submarine, then the likelihood of its appearance can usually not be determined. This type of approach, where we assign prior probabilities, is the Bayesian approach

to hypothesis testing. It is completely analogous to the Bayesian philosophy of estimation theory in which a prior PDF is assigned to an unknown parameter.

With the Bayesian paradigm we can define a *probability of error* P_e as

$$\begin{aligned} P_e &= \Pr\{\text{decide } \mathcal{H}_0,\ \mathcal{H}_1 \text{ true}\} + \Pr\{\text{decide } \mathcal{H}_1,\ \mathcal{H}_0 \text{ true}\} \\ &= P(\mathcal{H}_0|\mathcal{H}_1)P(\mathcal{H}_1) + P(\mathcal{H}_1|\mathcal{H}_0)P(\mathcal{H}_0) \end{aligned} \tag{3.12}$$

where $P(\mathcal{H}_i|\mathcal{H}_j)$ is the *conditional* probability that indicates the probability of deciding $\mathcal{H}_i$ when $\mathcal{H}_j$ is true. Note the slight distinction between $P(\mathcal{H}_i;\mathcal{H}_j)$ of the NP approach and $P(\mathcal{H}_i|\mathcal{H}_j)$ of the Bayesian approach. The former is the probability of deciding $\mathcal{H}_i$ if $\mathcal{H}_j$ *is true* with no probabilistic meaning assigned to the likelihood that $\mathcal{H}_j$ is true. The latter assumes that the outcome of a probabilistic experiment is observed to be $\mathcal{H}_j$ and that the probability of deciding $\mathcal{H}_i$ is *conditioned* on that outcome. Using the P_e criterion, the two errors are weighted appropriately to yield an overall error measure. Our goal will be to design a detector that minimizes P_e.

The derivation for the minimum P_e detector as a special case of the more general Bayesian detector is given in Appendix 3B. It is shown there that we should decide $\mathcal{H}_1$ if

$$\frac{p(\mathbf{x}|\mathcal{H}_1)}{p(\mathbf{x}|\mathcal{H}_0)} > \frac{P(\mathcal{H}_0)}{P(\mathcal{H}_1)} = \gamma. \tag{3.13}$$

Similar to the NP test we compare the *conditional* likelihood ratio to a threshold. Here, however, the threshold is determined by the prior probabilities. If, as is commonly the case, the prior probabilities are equal, we decide $\mathcal{H}_1$ if

$$p(\mathbf{x}|\mathcal{H}_1) > p(\mathbf{x}|\mathcal{H}_0). \tag{3.14}$$

Equivalently, we choose the hypothesis with the larger conditional likelihood or the one that maximizes $p(\mathbf{x}|\mathcal{H}_i)$ for $i = 0, 1$. This is called the *maximum likelihood* (ML) detector. (Actually, we should term this the maximum *conditional* likelihood. We defer to common usage in not doing so.) An example follows.

Example 3.5 - DC Level in WGN - Minimum P_e Criterion

We have the detection problem

$$\begin{aligned} \mathcal{H}_0 &: x[n] = w[n] \qquad n = 0, 1, \ldots, N-1 \\ \mathcal{H}_1 &: x[n] = A + w[n] \quad n = 0, 1, \ldots, N-1 \end{aligned}$$

where $A > 0$ and $w[n]$ is WGN with variance σ^2. If this is a digital communication problem where we transmit either $s_0[n] = 0$ or $s_1[n] = A$ (called an on-off keyed (OOK) communication system), it is reasonable to assume $P(\mathcal{H}_0) = P(\mathcal{H}_1) = 1/2$. The receiver that minimizes P_e is given by (3.13) with $\gamma = 1$. Hence, we decide $\mathcal{H}_1$

if

$$\frac{\frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}}\exp\left[-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}(x[n]-A)^2\right]}{\frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}}\exp\left[-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}x^2[n]\right]} > 1.$$

Taking logarithms yields

$$-\frac{1}{2\sigma^2}\left(-2A\sum_{n=0}^{N-1}x[n] + NA^2\right) > 0$$

or we decide $\mathcal{H}_1$ if $\bar{x} > A/2$. This is the same form of the detector as for the NP criterion except for the threshold (and, of course, the performance). To determine P_e we use (3.12) and note that

$$\bar{x} \sim \begin{cases} \mathcal{N}(0, \frac{\sigma^2}{N}) & \text{conditioned on } \mathcal{H}_0 \\ \mathcal{N}(A, \frac{\sigma^2}{N}) & \text{conditioned on } \mathcal{H}_1. \end{cases}$$

Thus

$$\begin{aligned} P_e &= \frac{1}{2}\left[P(\mathcal{H}_0|\mathcal{H}_1) + P(\mathcal{H}_1|\mathcal{H}_0)\right] \\ &= \frac{1}{2}\left[\Pr\{\bar{x} < A/2|\mathcal{H}_1\} + \Pr\{\bar{x} > A/2|\mathcal{H}_0\}\right] \\ &= \frac{1}{2}\left[\left(1 - Q\left(\frac{A/2-A}{\sqrt{\sigma^2/N}}\right)\right) + Q\left(\frac{A/2}{\sqrt{\sigma^2/N}}\right)\right] \end{aligned}$$

and since $Q(-x) = 1 - Q(x)$, we have finally

$$P_e = Q\left(\sqrt{\frac{NA^2}{4\sigma^2}}\right). \tag{3.15}$$

The probability of error decreases monotonically with NA^2/σ^2, which is, of course, the deflection coefficient.

◇

Another form of the minimum P_e detector follows directly from (3.13). We decide $\mathcal{H}_1$ if

$$p(\mathbf{x}|\mathcal{H}_1)P(\mathcal{H}_1) > p(\mathbf{x}|\mathcal{H}_0)P(\mathcal{H}_0).$$

But from Bayes rule we have that

$$P(\mathcal{H}_i|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{H}_i)P(\mathcal{H}_i)}{p(\mathbf{x})}$$

where the denominator $p(\mathbf{x})$ does not depend on the true hypothesis. In fact, $p(\mathbf{x})$ is just a normalizing factor that can be written as

$$p(\mathbf{x}) = p(\mathbf{x}|\mathcal{H}_0)P(\mathcal{H}_0) + p(\mathbf{x}|\mathcal{H}_1)P(\mathcal{H}_1).$$

As a result we decide $\mathcal{H}_1$ if

$$P(\mathcal{H}_1|\mathbf{x}) > P(\mathcal{H}_0|\mathbf{x}) \tag{3.16}$$

or we choose the hypothesis whose a posteriori (after the data are observed) probability is maximum. This detector, which minimizes P_e for any prior probability, is termed the *maximum a posteriori probability* (MAP) detector. Of course, for equal prior probabilities, the MAP detector reduces to the ML detector. The decision regions for the DC level in WGN with $N = 1$, $A = 1$, $\sigma^2 = 1$ are shown in Figure 3.10 for various prior probabilities (see Problem 3.16).

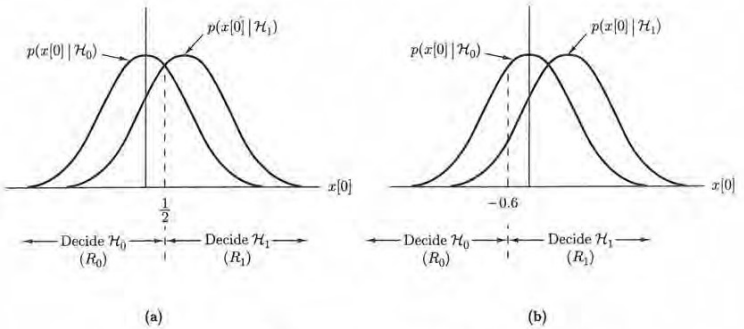


Figure 3.10. Effect of prior probability on decision regions (a) MAP detector with $P(\mathcal{H}_0) = P(\mathcal{H}_1) = 1/2$ (b) MAP detector with $P(\mathcal{H}_0) = 1/4$, $P(\mathcal{H}_1) = 3/4$.

3.7 Bayes Risk

A generalization of the minimum P_e criterion assigns costs to each type of error. Suppose that we wish to design a system to automatically inspect a machine part. The result of the inspection is either to use the part in a product if it is deemed satisfactory or else to discard it. We could set up the hypothesis test

$\mathcal{H}_0$: part is defective
$\mathcal{H}_1$: part is satisfactory

and assign costs to the errors. Let C_{ij} be the cost if we decide $\mathcal{H}_i$ but $\mathcal{H}_j$ is true. For example, we would probably want $C_{10} > C_{01}$. If we decide the part is satisfactory but it proves to be defective, the entire product may be defective and we incur a large cost (C_{10}). If, however, we decide that the part is defective when it is not, we incur the smaller cost of the part only (C_{01}). Once costs have been assigned, the decision rule is based on minimizing the expected cost or *Bayes risk* $\mathcal{R}$ defined as

$$\mathcal{R} = E(C) = \sum_{i=0}^{1}\sum_{j=0}^{1} C_{ij}P(\mathcal{H}_i|\mathcal{H}_j)P(\mathcal{H}_j). \tag{3.17}$$

Usually, if no error is made, we do not assign a cost so that $C_{00} = C_{11} = 0$. However, for convenience we will retain the more general form. Also, note that if $C_{00} = C_{11} = 0$, $C_{10} = C_{01} = 1$, then $\mathcal{R} = P_e$.

Under the reasonable assumption that $C_{10} > C_{00}$, $C_{01} > C_{11}$, the detector that minimizes the Bayes risk is to decide $\mathcal{H}_1$ if

$$\frac{p(\mathbf{x}|\mathcal{H}_1)}{p(\mathbf{x}|\mathcal{H}_0)} > \frac{(C_{10} - C_{00})P(\mathcal{H}_0)}{(C_{01} - C_{11})P(\mathcal{H}_1)} = \gamma. \tag{3.18}$$

See Appendix 3B for the proof. Once again, the conditional likelihood ratio is compared to a threshold.

3.8 Multiple Hypothesis Testing

We now consider the case where we wish to distinguish between M hypotheses, where $M > 2$. Such a problem arises quite frequently in communications, in which one of M signals must be detected. Also, pattern recognition systems make extensive use of the results in distinguishing between different patterns. In addition to signal detection, this problem also goes by the name of *classification* or *discrimination*. Although an NP criterion can be formulated for the M-ary hypothesis test, it seems to seldom be used in practice. The interested reader should consult [Lehmann 1959] for further details. More commonly the minimum P_e criterion or its generalization, the Bayes risk, is employed. We now consider the latter.

Assume that we now wish to decide among the M possible hypotheses $\{\mathcal{H}_0, \mathcal{H}_1, \ldots, \mathcal{H}_{M-1}\}$. The cost assigned to the decision to choose $\mathcal{H}_i$ when $\mathcal{H}_j$ is true is denoted by C_{ij}. The expected cost or Bayes risk becomes

$$\mathcal{R} = \sum_{i=0}^{M-1}\sum_{j=0}^{M-1} C_{ij}P(\mathcal{H}_i|\mathcal{H}_j)P(\mathcal{H}_j). \tag{3.19}$$

For the particular assignment

$$C_{ij} = \begin{cases} 0 & i = j \\ 1 & i \neq j \end{cases} \tag{3.20}$$

we have that $\mathcal{R} = P_e$. The decision rule that minimizes $\mathcal{R}$ is derived in Appendix 3C. There it is shown that we should choose the hypothesis that *minimizes*

$$C_i(\mathbf{x}) = \sum_{j=0}^{M-1} C_{ij}P(\mathcal{H}_j|\mathbf{x}) \tag{3.21}$$

over $i = 0, 1, \ldots, M-1$. To determine the decision rule that minimizes P_e we use (3.20). Then

$$\begin{aligned} C_i(\mathbf{x}) &= \sum_{\substack{j=0 \\ j\neq i}}^{M-1} P(\mathcal{H}_j|\mathbf{x}) \\ &= \sum_{j=0}^{M-1} P(\mathcal{H}_j|\mathbf{x}) - P(\mathcal{H}_i|\mathbf{x}). \end{aligned}$$

Since the first term is independent of i, $C_i(\mathbf{x})$ is minimized by *maximizing* $P(\mathcal{H}_i|\mathbf{x})$. Thus, the minimum P_e decision rule is to decide $\mathcal{H}_k$ if

$$P(\mathcal{H}_k|\mathbf{x}) > P(\mathcal{H}_i|\mathbf{x}) \qquad i \neq k. \tag{3.22}$$

As in the binary case we seek to maximize the a posteriori probability. This is the M-ary maximum a posteriori probability (MAP) decision rule. If, however, the prior probabilities are equal, then

$$\begin{aligned} P(\mathcal{H}_i|\mathbf{x}) &= \frac{p(\mathbf{x}|\mathcal{H}_i)P(\mathcal{H}_i)}{p(\mathbf{x})} \\ &= \frac{p(\mathbf{x}|\mathcal{H}_i)\frac{1}{M}}{p(\mathbf{x})} \end{aligned} \tag{3.23}$$

and to maximize $P(\mathcal{H}_i|\mathbf{x})$ we need only maximize $p(\mathbf{x}|\mathcal{H}_i)$. Hence, for equal prior probabilities we decide $\mathcal{H}_k$ if

$$p(\mathbf{x}|\mathcal{H}_k) > p(\mathbf{x}|\mathcal{H}_i) \qquad i \neq k. \tag{3.24}$$

This is the M-ary maximum likelihood (ML) decision rule.

Finally, observe from (3.23) that to maximize $P(\mathcal{H}_i|\mathbf{x})$ we can equivalently maximize $p(\mathbf{x}|\mathcal{H}_i)P(\mathcal{H}_i)$ since $p(\mathbf{x})$ does not depend on i. Equivalently then, the MAP rule maximizes

$$\ln p(\mathbf{x}|\mathcal{H}_i) + \ln P(\mathcal{H}_i).$$

An example follows.

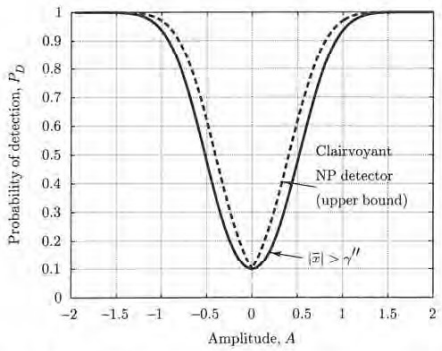


Figure 6.4. Detection performance comparison of clairvoyant NP detector and realizable detector.

that (see Problem 6.5)

$$P_D = Q\left(Q^{-1}\left(\frac{P_{FA}}{2}\right) - \sqrt{\frac{NA^2}{\sigma^2}}\right) + Q\left(Q^{-1}\left(\frac{P_{FA}}{2}\right) + \sqrt{\frac{NA^2}{\sigma^2}}\right). \quad (6.9)$$

For $N = 10$, $\sigma^2 = 1$, $P_{FA} = 0.1$, this is plotted in Figure 6.4, along with the clairvoyant detector bound (shown in Figure 6.3). It is observed that the proposed detector has performance close to the bound. This detector then would appear to be a good compromise in that it is *robust* to the sign of A. In fact, the proposed detector is an example of a more general approach to composite hypothesis testing, the generalized likelihood ratio test, which is described in the next section (see Example 6.4).

6.4 Composite Hypothesis Testing Approaches

There are two major approaches to composite hypothesis testing. The first is to consider the unknown parameters as realizations of random variables and to assign a prior PDF. The second is to estimate the unknown parameters for use in a likelihood ratio test. We will term the first method the Bayesian approach and the second, the generalized likelihood ratio test (GLRT). The Bayesian approach employs the philosophy described in [Kay-I 1993, Chapter 10], where it is applied to parameter estimation. It requires prior knowledge of the unknown parameters whereas the

GLRT does not. In practice, the GLRT appears to be more widely used due to its ease of implementation and less restrictive assumptions. The Bayesian approach requires multidimensional integration, which is usually not possible in closed form. For these reasons our emphasis in succeeding chapters will be on the GLRT.

The general problem is to decide between $\mathcal{H}_0$ and $\mathcal{H}_1$ when the PDFs depend on a set of unknown parameters. These parameters may or may not be the same under each hypothesis. Under $\mathcal{H}_0$ we assume that the vector parameter $\boldsymbol{\theta}_0$ is unknown while under $\mathcal{H}_1$ the vector parameter $\boldsymbol{\theta}_1$ is unknown. We first discuss the Bayesian approach.

6.4.1 Bayesian Approach

The Bayesian approach assigns prior PDFs to $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_1$. In doing so it models the unknown parameters as *realizations* of a vector random variable. If the prior PDFs are denoted by $p(\boldsymbol{\theta}_0)$ and $p(\boldsymbol{\theta}_1)$, respectively, the PDFs of the data are

$$\begin{aligned} p(\mathbf{x};\mathcal{H}_0) &= \int p(\mathbf{x}|\boldsymbol{\theta}_0;\mathcal{H}_0)p(\boldsymbol{\theta}_0)d\boldsymbol{\theta}_0 \\ p(\mathbf{x};\mathcal{H}_1) &= \int p(\mathbf{x}|\boldsymbol{\theta}_1;\mathcal{H}_1)p(\boldsymbol{\theta}_1)d\boldsymbol{\theta}_1 \end{aligned}$$

where $p(\mathbf{x}|\boldsymbol{\theta}_i;\mathcal{H}_i)$ is the conditional PDF of $\mathbf{x}$, conditioned on $\boldsymbol{\theta}_i$, assuming $\mathcal{H}_i$ is true. The unconditional PDFs $p(\mathbf{x};\mathcal{H}_0)$ and $p(\mathbf{x};\mathcal{H}_1)$ are now completely specified, no longer dependent on the unknown parameters. With the Bayesian approach the optimal NP detector decides $\mathcal{H}_1$ if

$$\frac{p(\mathbf{x};\mathcal{H}_1)}{p(\mathbf{x};\mathcal{H}_0)} = \frac{\int p(\mathbf{x}|\boldsymbol{\theta}_1;\mathcal{H}_1)p(\boldsymbol{\theta}_1)d\boldsymbol{\theta}_1}{\int p(\mathbf{x}|\boldsymbol{\theta}_0;\mathcal{H}_0)p(\boldsymbol{\theta}_0)d\boldsymbol{\theta}_0} > \gamma. \quad (6.10)$$

The required integrations are multidimensional with dimension equal to the unknown parameter dimension. The choice of the prior PDFs can prove difficult. If, indeed, one does have some prior knowledge, then one should use it. If not, then a noninformative prior (see [Kay-I 1993, pg. 332]) should be used. A noninformative prior is a PDF that is as "flat" as possible. As an example, if we wish to detect a sinusoid with unknown phase ϕ and we have no reason to favor any particular value of ϕ over another, then $\phi \sim \mathcal{U}[0, 2\pi]$ would be consistent with our lack of prior knowledge (see also Problem 7.22). A uniform PDF, then, would be a good choice, although the integration may prove difficult. If, however, the parameter takes on values over an infinite interval, say $-\infty < A < \infty$ for the DC level, then we cannot assign a uniform PDF. More commonly, a Gaussian PDF is chosen with $A \sim \mathcal{N}(0, \sigma_A^2)$ and we let $\sigma_A^2 \to \infty$ to reflect our lack of prior knowledge. This approach is illustrated next.

Example 6.3 - DC Level in WGN with Unknown Amplitude - Bayesian Approach

Assume that for the DC level in WGN A is unknown and can take on values $-\infty < A < \infty$. We assign the prior PDF $A \sim \mathcal{N}(0, \sigma_A^2)$, where A is independent of $w[n]$. Note that as $\sigma_A^2 \to \infty$, the PDF becomes a noninformative prior (see also [Kay-I 1993, Problem 10.17]). The conditional PDF under $\mathcal{H}_1$ is assumed to be

$$p(\mathbf{x}|A;\mathcal{H}_1) = \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp\left[-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}(x[n]-A)^2\right].$$

Under $\mathcal{H}_0$ the PDF is completely known. Hence, according to (6.10) with $\boldsymbol{\theta}_1 = A$, the NP detector decides $\mathcal{H}_1$ if

$$\frac{p(\mathbf{x};\mathcal{H}_1)}{p(\mathbf{x};\mathcal{H}_0)} = \frac{\int_{-\infty}^{\infty} p(\mathbf{x}|A;\mathcal{H}_1)p(A)dA}{p(\mathbf{x};\mathcal{H}_0)} > \gamma.$$

But

$$\begin{aligned} p(\mathbf{x};\mathcal{H}_1) &= \int_{-\infty}^{\infty} p(\mathbf{x}|A;\mathcal{H}_1)p(A)dA \\ &= \int_{-\infty}^{\infty} \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp\left[-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}(x[n]-A)^2\right] \\ &\quad \cdot \frac{1}{\sqrt{2\pi\sigma_A^2}} \exp\left(-\frac{1}{2\sigma_A^2}A^2\right) dA. \end{aligned}$$

Letting

$$Q(A) = \frac{1}{\sigma^2}\sum_{n=0}^{N-1}(x[n]-A)^2 + \frac{A^2}{\sigma_A^2}$$

we have upon completing the square in A

$$\begin{aligned} Q(A) &= \frac{1}{\sigma^2}\sum_{n=0}^{N-1}x^2[n] - \frac{2N}{\sigma^2}\bar{x}A + \frac{N}{\sigma^2}A^2 + \frac{A^2}{\sigma_A^2} \\ &= \underbrace{\left(\frac{N}{\sigma^2}+\frac{1}{\sigma_A^2}\right)}_{1/\sigma_{A|x}^2} A^2 - \frac{2N}{\sigma^2}\bar{x}A + \frac{1}{\sigma^2}\sum_{n=0}^{N-1}x^2[n] \\ &= \frac{A^2}{\sigma_{A|x}^2} - \frac{2N\sigma_{A|x}^2\bar{x}A}{\sigma^2\sigma_{A|x}^2} + \frac{1}{\sigma^2}\sum_{n=0}^{N-1}x^2[n] \end{aligned}$$

$$= \frac{1}{\sigma_{A|x}^2}\left(A - \frac{N\bar{x}\sigma_{A|x}^2}{\sigma^2}\right)^2 - \frac{N^2\bar{x}^2}{\sigma^4}\sigma_{A|x}^2 + \frac{1}{\sigma^2}\sum_{n=0}^{N-1}x^2[n]$$

so that

$$\begin{aligned} \frac{p(\mathbf{x};\mathcal{H}_1)}{p(\mathbf{x};\mathcal{H}_0)} &= \frac{\frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}}\frac{1}{\sqrt{2\pi\sigma_A^2}}\int_{-\infty}^{\infty}\exp\left[-\frac{1}{2}Q(A)\right]dA}{\frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}}\exp\left(-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}x^2[n]\right)} \\ &= \frac{1}{\sqrt{2\pi\sigma_A^2}}\sqrt{2\pi\sigma_{A|x}^2}\exp\left(\frac{N^2\bar{x}^2\sigma_{A|x}^2}{2\sigma^4}\right) > \gamma. \end{aligned}$$

Taking logarithms of both sides and retaining only the data dependent terms we decide $\mathcal{H}_1$ if

$$(\bar{x})^2 > \gamma'$$

or

$$|\bar{x}| > \sqrt{\gamma'}. \tag{6.11}$$

This is similar to the detector we proposed in the previous section. In order to set the threshold we do not need knowledge of σ_A^2. However, if we had assumed $A \sim \mathcal{N}(\mu_A, \sigma_A^2)$, then we would need to know μ_A and σ_A^2 (see Section 7.4.2) to implement the Bayesian detector. The performance of the detector of (6.11) is derived in Problem 6.7.

◇

6.4.2 Generalized Likelihood Ratio Test

The GLRT replaces the unknown parameters by their maximum likelihood estimates (MLEs). Although there is no optimality associated with the GLRT, in practice, it appears to work quite well. (Asymptotically, it can be shown that the GLRT is UMP among all tests that are *invariant* [Lehmann 1959].) In general, the GLRT decides $\mathcal{H}_1$ if

$$L_G(\mathbf{x}) = \frac{p(\mathbf{x};\hat{\boldsymbol{\theta}}_1,\mathcal{H}_1)}{p(\mathbf{x};\hat{\boldsymbol{\theta}}_0,\mathcal{H}_0)} > \gamma \tag{6.12}$$

where $\hat{\boldsymbol{\theta}}_1$ is the MLE of $\boldsymbol{\theta}_1$ assuming $\mathcal{H}_1$ is true (maximizes $p(\mathbf{x};\boldsymbol{\theta}_1,\mathcal{H}_1)$), and $\hat{\boldsymbol{\theta}}_0$ is the MLE of $\boldsymbol{\theta}_0$ assuming $\mathcal{H}_0$ is true (maximizes $p(\mathbf{x};\boldsymbol{\theta}_0,\mathcal{H}_0)$). The approach also provides information about the unknown parameters since the first step in determining $L_G(\mathbf{x})$ is to find the MLEs. We now continue the DC level in WGN example.

Example 6.4 - DC Level in WGN with Unknown Amplitude - GLRT

In this case we have $\boldsymbol{\theta}_1 = A$ and there are no unknown parameters under $\mathcal{H}_0$. The hypothesis test becomes

$$\begin{aligned}\mathcal{H}_0 &: A = 0 \\ \mathcal{H}_1 &: A \neq 0.\end{aligned}$$

Thus, the GLRT decides $\mathcal{H}_1$ if

$$L_G(\mathbf{x}) = \frac{p(\mathbf{x}; \hat{A}, \mathcal{H}_1)}{p(\mathbf{x}; \mathcal{H}_0)} > \gamma.$$

The MLE of A is found by maximizing

$$p(\mathbf{x}; A, \mathcal{H}_1) = \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp\left[-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}(x[n] - A)^2\right].$$

This was done in [Kay-I 1993, pp. 163–164] with the result that $\hat{A} = \bar{x}$. Thus,

$$L_G(\mathbf{x}) = \frac{\frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp\left[-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}(x[n] - \bar{x})^2\right]}{\frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp\left(-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}x^2[n]\right)}.$$

Taking logarithms we have

$$\begin{aligned}\ln L_G(\mathbf{x}) &= -\frac{1}{2\sigma^2}\left(\sum_{n=0}^{N-1}x^2[n] - 2\bar{x}\sum_{n=0}^{N-1}x[n] + N\bar{x}^2 - \sum_{n=0}^{N-1}x^2[n]\right) \\ &= -\frac{1}{2\sigma^2}(-2N\bar{x}^2 + N\bar{x}^2) \\ &= \frac{N\bar{x}^2}{2\sigma^2}\end{aligned}$$

or we decide $\mathcal{H}_1$ if

$$|\bar{x}| > \gamma'. \tag{6.13}$$

The performance has already been given by (6.9). When compared to the clairvoyant detector, there is only a slight loss as shown in Figure 6.4. ◇

The GLRT can also be expressed in another form, which is sometimes more convenient. Since $\hat{\boldsymbol{\theta}}_i$ is the MLE under $\mathcal{H}_i$, it maximizes $p(\mathbf{x}; \boldsymbol{\theta}_i, \mathcal{H}_i)$ or

$$p(\mathbf{x}; \hat{\boldsymbol{\theta}}_i, \mathcal{H}_i) = \max_{\boldsymbol{\theta}_i} p(\mathbf{x}; \boldsymbol{\theta}_i, \mathcal{H}_i).$$

Hence, (6.12) can be written as

$$L_G(\mathbf{x}) = \frac{\max_{\boldsymbol{\theta}_1} p(\mathbf{x}; \boldsymbol{\theta}_1, \mathcal{H}_1)}{\max_{\boldsymbol{\theta}_0} p(\mathbf{x}; \boldsymbol{\theta}_0, \mathcal{H}_0)}. \tag{6.14}$$

For the special case where the PDF under $\mathcal{H}_0$ is completely known

$$\begin{aligned}L_G(\mathbf{x}) &= \frac{\max_{\boldsymbol{\theta}_1} p(\mathbf{x}; \boldsymbol{\theta}_1, \mathcal{H}_1)}{p(\mathbf{x}; \mathcal{H}_0)} \\ &= \max_{\boldsymbol{\theta}_1} \frac{p(\mathbf{x}; \boldsymbol{\theta}_1, \mathcal{H}_1)}{p(\mathbf{x}; \mathcal{H}_0)}\end{aligned}$$

or we maximize the likelihood ratio over $\boldsymbol{\theta}_1$ so that

$$L_G(\mathbf{x}) = \max_{\boldsymbol{\theta}_1} L(\mathbf{x}; \boldsymbol{\theta}_1). \tag{6.15}$$

See also Problem 6.14. We now illustrate the GLRT approach with a slightly more complicated example.

Example 6.5 - DC Level in WGN with Unknown Amplitude and Variance - GLRT

Consider the detection problem

$$\begin{aligned}\mathcal{H}_0 &: x[n] = w[n] \quad n = 0, 1, \ldots, N-1 \\ \mathcal{H}_1 &: x[n] = A + w[n] \quad n = 0, 1, \ldots, N-1\end{aligned}$$

where A is unknown with $-\infty < A < \infty$ and $w[n]$ is WGN with *unknown variance* σ^2. A UMP test does not exist because the equivalent parameter test is

$$\begin{aligned}\mathcal{H}_0 &: A = 0, \sigma^2 > 0 \\ \mathcal{H}_1 &: A \neq 0, \sigma^2 > 0\end{aligned} \tag{6.16}$$

which is two-sided. In this example the hypothesis test contains a *nuisance parameter*, which is σ^2. Although we are not directly concerned with σ^2, it enters into the problem since it affects the PDFs under $\mathcal{H}_0$ and $\mathcal{H}_1$. The detector of (6.13) can no longer be implemented since the threshold is dependent upon the PDF under $\mathcal{H}_0$, which in turn depends on σ^2. The variance can take on any value $0 < \sigma^2 < \infty$. The GLRT decides $\mathcal{H}_1$ if

$$L_G(\mathbf{x}) = \frac{p(\mathbf{x}; \hat{A}, \hat{\sigma_1^2}, \mathcal{H}_1)}{p(\mathbf{x}; \hat{\sigma_0^2}, \mathcal{H}_0)} > \gamma$$

orthogonal signaling for communications, which illustrates the concept of channel capacity, and to pattern recognition for images.

4.3 Matched Filters

4.3.1 Development of Detector

We begin our discussion of optimal detection approaches by considering the problem of detecting a *known deterministic* signal in white Gaussian noise (WGN). The Neyman-Pearson (NP) criterion will be used, but as discussed in Chapter 3, the resulting test statistic will be identical to that obtained using the Bayesian risk criterion. Only the threshold and detection performance will differ. The detection problem is to distinguish between the hypotheses

$$\begin{array}{ll} \mathcal{H}_0 : x[n] = w[n] & n = 0, 1, \ldots, N-1 \\ \mathcal{H}_1 : x[n] = s[n] + w[n] & n = 0, 1, \ldots, N-1 \end{array} \tag{4.1}$$

where the signal $s[n]$ is assumed known and $w[n]$ is WGN with variance σ^2. WGN is defined as a zero mean Gaussian noise process with autocorrelation function (ACF)

$$\begin{aligned} r_{ww}[k] &= E(w[n]w[n+k]) \\ &= \sigma^2\delta[k] \end{aligned}$$

where $\delta[k]$ is the discrete-time delta function ($\delta[k] = 1$ if $k = 0$ and $\delta[k] = 0$ for $k \neq 0$). Such a model can be derived from a continuous-time setup as described in [Kay-I 1993, pg. 54].

The NP detector decides $\mathcal{H}_1$ if the likelihood ratio exceeds a threshold or

$$L(\mathbf{x}) = \frac{p(\mathbf{x};\mathcal{H}_1)}{p(\mathbf{x};\mathcal{H}_0)} > \gamma \tag{4.2}$$

where $\mathbf{x} = [x[0]\, x[1] \ldots x[N-1]]^T$. Since

$$\begin{aligned} p(\mathbf{x};\mathcal{H}_1) &= \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp\left[-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}(x[n]-s[n])^2\right] \\ p(\mathbf{x};\mathcal{H}_0) &= \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp\left[-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}x^2[n]\right] \end{aligned}$$

we have

$$L(\mathbf{x}) = \exp\left[-\frac{1}{2\sigma^2}\left(\sum_{n=0}^{N-1}(x[n]-s[n])^2 - \sum_{n=0}^{N-1}x^2[n]\right)\right] > \gamma.$$

Taking the logarithm (a monotonically increasing transformation) of both sides does not change the inequality (see Section 3.3) so that

$$l(\mathbf{x}) = \ln L(\mathbf{x}) = -\frac{1}{2\sigma^2}\left(\sum_{n=0}^{N-1}(x[n]-s[n])^2 - \sum_{n=0}^{N-1}x^2[n]\right) > \ln\gamma.$$

We decide $\mathcal{H}_1$ if

$$\frac{1}{\sigma^2}\sum_{n=0}^{N-1}x[n]s[n] - \frac{1}{2\sigma^2}\sum_{n=0}^{N-1}s^2[n] > \ln\gamma.$$

Since $s[n]$ is known (and thus not a function of the data), we may incorporate the energy term into the threshold to yield

$$T(\mathbf{x}) = \sum_{n=0}^{N-1}x[n]s[n] > \sigma^2\ln\gamma + \frac{1}{2}\sum_{n=0}^{N-1}s^2[n].$$

Calling the right-hand-side of the inequality a new threshold γ', we decide $\mathcal{H}_1$ if

$$T(\mathbf{x}) = \sum_{n=0}^{N-1}x[n]s[n] > \gamma'. \tag{4.3}$$

This is our NP detector and as expected consists of a test statistic $T(\mathbf{x})$ (a function of the data) and a threshold γ', which is chosen to satisfy $P_{FA} = \alpha$ for a given α. We now determine the test statistic for some simple examples.

Example 4.1 - DC Level in WGN

Assume that $s[n] = A$ for some known level A, where $A > 0$. Then from (4.3), $T(\mathbf{x}) = A\sum_{n=0}^{N-1}x[n]$. An equivalent detector divides $T(\mathbf{x})$ by NA to decide $\mathcal{H}_1$ if

$$T'(\mathbf{x}) = \frac{1}{NA}T(\mathbf{x}) = \frac{1}{N}\sum_{n=0}^{N-1}x[n] = \bar{x} > \gamma''$$

where $\gamma'' = \gamma'/NA$. But this is just the sample mean detector discussed in Chapter 3. Its performance has also been described there. Note that if $A < 0$, the inequality reverses and we decide $\mathcal{H}_1$ if $\bar{x} < \gamma''$.

◇

Example 4.2 - Damped Exponential in WGN

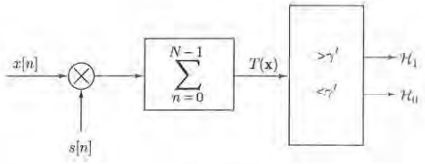
Now let $s[n] = r^n$ for $0 < r < 1$. From (4.3) the test statistic is

$$T(\mathbf{x}) = \sum_{n=0}^{N-1}x[n]r^n$$

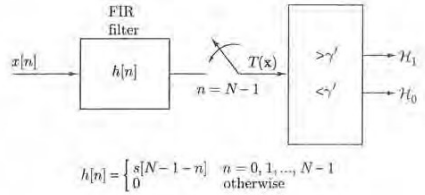
and is seen to weight the earlier samples more heavily than the later ones. This is due to the fact that the signal is decaying as n increases while the noise power remains constant. The signal-to-noise ratio (SNR) for the nth sample is $s^2[n]/\sigma^2 = r^{2n}/\sigma^2$, which decreases with n. The detection performance can easily be determined as described in Section 4.3.2.

◇

In general, the test statistic of (4.3) weights the data samples according to the value of the signal. More weight is reserved for the larger signal samples. Even negative signal samples are weighted in the same manner because by multiplying $x[n]$ by $s[n]$, negative samples yield a positive contribution to the sum. The detector of (4.3) is referred to as a *correlator* or *replica-correlator* since we correlate the received data with a replica of the signal. The detector is shown in Figure 4.1a. An alternative interpretation of the test statistic is based on relating the correlation process to the effect of a finite impulse response (FIR) filter on the data. Specifically, if $x[n]$ is the input to an FIR filter with impulse response $h[n]$, where $h[n]$ is nonzero



(a)



(b)

Figure 4.1. Neyman-Pearson detector for deterministic signal in white Gaussian noise (a) Replica-correlator (b) Matched filter.

for $n = 0, 1, \ldots, N-1$, then the output at time n is

$$y[n] = \sum_{k=0}^{n} h[n-k]x[k] \tag{4.4}$$

for $n \geq 0$. (For $n < 0$ the output is zero since we assume $x[n]$ is also nonzero only over the interval $[0, N-1]$.) If we let the impulse response be a "flipped around" version of the signal or

$$h[n] = s[N-1-n] \qquad n = 0, 1, \ldots, N-1 \tag{4.5}$$

then

$$y[n] = \sum_{k=0}^{n} s[N-1-(n-k)]x[k].$$

Now the output of the filter at time $n = N-1$ is

$$y[N-1] = \sum_{k=0}^{N-1} s[k]x[k]$$

which with a change of summation variable is identical to the replica-correlator. This implementation of the NP detector is shown in Figure 4.1b and is known as a *matched filter*. The filter impulse response is *matched* to the signal. Figure 4.2 shows some examples. The matched filter impulse response is obtained by flipping $s[n]$ about $n = 0$ and shifting to the right by $N-1$ samples.

Although the test statistic is obtained by sampling the matched filter output at time $n = N-1$, it is instructive to view the entire output of the matched filter. For the DC level signal shown in Figure 4.2b, the signal output is given via the convolution sum in Figure 4.3. Note that the signal output attains a maximum at the sampling time. This is true in general (see Problem 4.2). When noise is present, the maximum may be perturbed but it should be intuitively clear that the best detection performance will be obtained by sampling at $n = N-1$. If, however, the signal does not begin at $n = 0$, but we assume that it does and use the corresponding matched filter, poor detection performance may be obtained. An example is given in Problem 4.3. Thus, for signals with *unknown arrival times*, we cannot use the matched filter in its present form. In Chapter 7 we will see how to modify it to circumvent this problem.

The matched filter may also be viewed in the frequency domain. From (4.4) we have

$$y[n] = \int_{-\frac{1}{2}}^{\frac{1}{2}} H(f)X(f)\exp(j2\pi fn)df \tag{4.6}$$

where $H(f), X(f)$ are the discrete-time Fourier transforms of $h[n]$ and $x[n]$, respectively. But from (4.5), $H(f) = \mathcal{F}\{s[N-1-n]\}$, where $\mathcal{F}$ denotes the discrete-time Fourier transform. This may be shown to be

$$H(f) = S^*(f)\exp[-j2\pi f(N-1)] \tag{4.7}$$

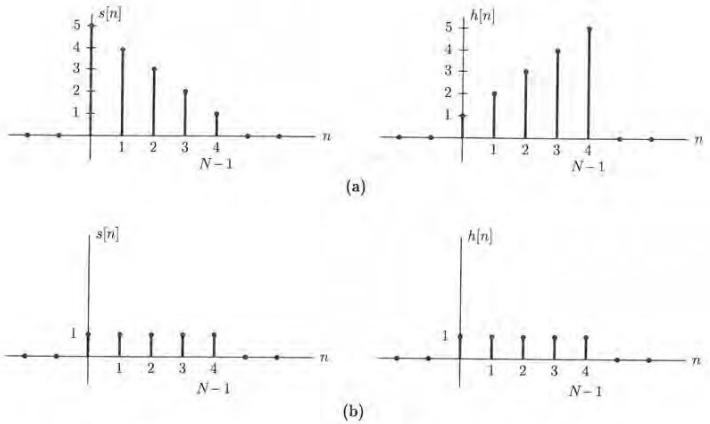


Figure 4.2. Examples of matched filter impulse response ($N = 5$).

so that (4.6) becomes

$$y[n] = \int_{-\frac{1}{2}}^{\frac{1}{2}} S^*(f)X(f)\exp[j2\pi f(n-(N-1))]df. \tag{4.8}$$

Sampling the output at $n = N-1$ produces

$$y[N-1] = \int_{-\frac{1}{2}}^{\frac{1}{2}} S^*(f)X(f)df \tag{4.9}$$

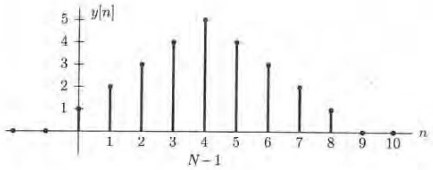


Figure 4.3. Matched filter signal output for DC level signal input ($N = 5$).

which could also have been obtained from the correlator implementation by using Parseval's theorem. Note from (4.7) that the matched filter emphasizes the bands where there is more signal power. Also, from (4.9) (or equivalently from (4.3)), when the noise is absent or $X(f) = S(f)$, the matched filter output is just the signal *energy*.

Another property of the matched filter is that *it maximizes the SNR at the output of an FIR filter*. In other words, we consider all detectors of the form of Figure 4.1b but with an arbitrary $h[n]$ over $[0, N-1]$ and zero otherwise. If we define the output SNR as

$$\begin{aligned}\eta &= \frac{E^2(y[N-1];\mathcal{H}_1)}{\text{var}(y[N-1];\mathcal{H}_0)} \\ &= \frac{\left(\sum_{k=0}^{N-1} h[N-1-k]s[k]\right)^2}{E\left[\left(\sum_{k=0}^{N-1} h[N-1-k]w[k]\right)^2\right]}\end{aligned} \tag{4.10}$$

then the matched filter of (4.5) maximizes (4.10). To show this let $\mathbf{s} = [s[0]\, s[1] \ldots s[N-1]]^T$, $\mathbf{h} = [h[N-1]\, h[N-2] \ldots h[0]]^T$ and $\mathbf{w} = [w[0]\, w[1] \ldots w[N-1]]^T$. Then

$$\begin{aligned}\eta &= \frac{\left(\mathbf{h}^T\mathbf{s}\right)^2}{E\left[(\mathbf{h}^T\mathbf{w})^2\right]} = \frac{\left(\mathbf{h}^T\mathbf{s}\right)^2}{\mathbf{h}^T E(\mathbf{w}\mathbf{w}^T)\mathbf{h}} \\ &= \frac{\left(\mathbf{h}^T\mathbf{s}\right)^2}{\mathbf{h}^T\sigma^2\mathbf{I}\mathbf{h}} = \frac{1}{\sigma^2}\frac{\left(\mathbf{h}^T\mathbf{s}\right)^2}{\mathbf{h}^T\mathbf{h}}.\end{aligned}$$

By the Cauchy-Schwarz inequality (see Appendix 1)

$$(\mathbf{h}^T\mathbf{s})^2 \le (\mathbf{h}^T\mathbf{h})(\mathbf{s}^T\mathbf{s})$$

with equality if and only if $\mathbf{h} = c\mathbf{s}$, where c is any constant. Hence

$$\eta \le \frac{1}{\sigma^2}\mathbf{s}^T\mathbf{s}$$

with equality if and only if $\mathbf{h} = c\mathbf{s}$. The maximum output SNR is attained for (letting $c = 1$)

$$h[N-1-n] = s[n] \qquad n = 0, 1, \ldots, N-1$$

or equivalently

$$h[n] = s[N-1-n] \qquad n = 0, 1, \ldots, N-1$$