§4. Nonparametric and Robust Detection

Previously, we have assumed that the probability distribution of the data is known under each hypothesis. But in practice, it may not be realistic to assume that these distributions are known exactly and sometimes it can not even be assumed known approximately. Without the assumption, the previously developed techniques may need to modify/change. In this section, two design philosophies that can be applied, nonparametric and robust detections, will be discussed.

Consider the following general composite binary hypothesis testing problem based on an i.i.d. observation sequence

$$H_0: Y_k \sim P \in \mathcal{P}_0, \quad k = 1, \cdots, n \qquad (*)$$

vs.

$$H_1: Y_k \sim P \in \mathcal{P}_1, \quad k = 1, 2, \cdots, n.$$

where $\mathcal{P}_0$ and $\mathcal{P}_1$ are two non overlapping classes of possible marginal distributions for the observations. This problem is said to be a parametric hypothesis-testing problem if the classes $\mathcal{P}_0$ and $\mathcal{P}_1$ can be parameterized by a real or vector parameter. For example, the composite hypothesis-testing problem we discussed before are parametric problems.

Otherwise, the problem (*) is said to be a nonparametric problem, it may be because that $\mathcal{P}_0$ and $\mathcal{P}_1$ are too broad to be parameterized by a finite-dimensional parameter vector.

An example of a nonparametric hypothesis-testing problem is the location-testing probem:

$$H_0: Y_k = N_k, \quad k=1, \cdots, n,$$

vs.

$$H_1: Y_k = N_k + \theta, \quad k=1, \cdots, n,$$

in which $\{N_k\}_{k=1}^n$ is an i.i.d. sequence whose marginal distribution is known only to be symmetric about zero.

§4.1. Nonparametric Detection

A nonparametric test is usually on wide classes $\mathcal{P}_0$ and $\mathcal{P}_1$ with some invariant performance characteristics. These tests are usually simple and use rough information about the data (such as, signs, ranks, etc) rather than the exact values of the data.

The performance characteristic that is to be kept invariant in nonparametric problems is usually the false-alarm probability. The standard definition of a nonparametric test (or detector) for (*) is

one whose false-alarm probability is constant over $\mathcal{P}_0$.

For a sequence of tests $\{\delta_n(y_1, \cdots, y_n)\}_{n=1}^{\infty}$, they are called asymptotically nonparametric if $\lim_{n\to\infty} P_F(\delta_n)$ is a constant for all $P \in \mathcal{P}_0$.

Nonparametric tests and detectors have found many applications in areas such as radar and sonar. In these applications, nonparametric detectors are sometimes called constant-false-alarm-rate (CFAR) detectors.

* The Sign Test

Suppose that we have a sequence $Y_1, \cdots, Y_n$ of i.i.d. real-valued observations. Define the parameter p by $p = P(Y_1 > 0)$.

Consider the hypothesis pair

$$H_0 : p = \frac{1}{2} \qquad (**)$$

$$\text{vs.} \quad H_1 : \frac{1}{2} < p < 1$$

(**) is the hypothesis that Y_k's have zero median vs. the hypothesis that the median of Y_k's is greater than zero.

The hypotheses (**) are nonparametric as follows. In terms of (*),

$$\mathcal{P}_0 = \{ P \in \mathcal{M} \mid P((0,\infty)) = \tfrac{1}{2} \}$$
$$\mathcal{P}_1 = \{ P \in \mathcal{M} \mid 1 > P((0,\infty)) > \tfrac{1}{2} \},$$

where $\mathcal{M}$ denotes the class of all distributions on $(\mathbb{R}, \mathcal{B})$. Neither of these classes can be parameterized by a finite dimensional parameter vector.

To derive an optimum test for (**), we first choose an arbitrary distribution Q_1 in $\mathcal{P}_1$. For convenience, assume Q_1 has a density g_1. Define

$$g_1^+(x) = \begin{cases} g_1(x) & \text{if } x > 0 \\ 0 & \text{if } x \le 0 \end{cases}$$

and

$$g_1^-(x) = \begin{cases} 0 & \text{if } x > 0 \\ g_1(x) & \text{if } x \le 0 \end{cases}$$

and define a density g_0 on $(\mathbb{R}, \mathcal{B})$ by

$$g_0(x) = \frac{g_1^+(x)}{2\int_0^\infty g_1(t)\,dt} + \frac{g_1^-(x)}{2\int_{-\infty}^0 g_1(t)\,dt}.$$

Then, the distribution Q_0 corresponding to the desity g_0 is in $\mathcal{P}_0$ since

$$Q_0((0,\infty)) = \int_0^\infty g_0(x)\,dx = \frac{\int_0^\infty g_1^+(x)\,dx}{2\int_0^\infty g_1(x)\,dx} = \frac{1}{2}.$$

Consider the simple hypothesis pair:

H_0' : $Y_k \sim Q_0$, $k = 1, 2, \cdots, n$

v.s.

H_1' : $Y_k \sim Q_1$, $k = 1, 2, \cdots, n$.

By the Neyman–Pearson Lemma, a most powerful α-level test of H_0' v.s. H_1' is the likelihood ratio test based on comparison of the statistic

$$L(\underline{y}) = \prod_{k=1}^{n} \frac{g_1(y_k)}{g_0(y_k)}$$

to a threshold. Note that

$$\frac{g_1(y_k)}{g_0(y_k)} = \begin{cases} 2Q_1^+ & \text{if } y_k > 0 \\ 2(1-Q_1^+) & \text{if } y_k \le 0 \end{cases}$$

where $Q_1^+ \triangleq Q_1((0, \infty))$. Thus,

$$L(\underline{y}) = 2^n [1-Q_1^+]^n \left[Q_1^+/(1-Q_1^+) \right]^{t(\underline{y})}$$

where $t(\underline{y}) = \sum_{k=1}^{n} u(y_k)$

$$u(x) = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{if } x \le 0 \end{cases}$$

$t(\underline{y})$ is the number of the observed y_k's that are positive.

By hypothesis $\frac{1}{2} < Q_1^+ < 1$, so that $Q_1^+/(1-Q_1^+) > 1$. Thus, $L(\underline{y}) \nearrow$ as $t(\underline{y}) \nearrow$

and a most-powerful α-level test of H_0' vs. H_1' is

$$\tilde{\delta}_s(\underline{y}) = \begin{cases} 1 & \text{if } t(\underline{y}) > \tau \\ \gamma & \vdots = \\ 0 & \vdots < \end{cases}$$

where γ and τ are the randomization and the threshold for false-alarm probability α.

$t(\underline{Y})$ is a binomial random variable with parameters (n, Q_1^+) under H_1' and $(n, \frac{1}{2})$ under H_0'. Thus, for size α, the threshold τ is the smallest such that

$$2^{-n} \sum_{k=\tau}^{n} \frac{n!}{(n-k)!\,k!} \leq \alpha$$

and the randomization constant is

$$\gamma = \frac{\alpha - 2^{-n} \sum_{k=\tau+1}^{n} \frac{n!}{(n-k)!\,k!}}{2^{-n} \frac{n!}{(n-\tau)!\,\tau!}}$$

Since the distribution of $t(\underline{y})$ is binomial with parameters $(n, \frac{1}{2})$ for any $Q_0 \in \mathcal{P}_0$, the above test $\tilde{\delta}_s(\underline{y})$ has size α for the entire class $\mathcal{P}_0$ and thus it is nonparametric. This test does not depend on the choice of Q_1, so it is a uniformly most powerful α-level test of H_0 vs. H_1.

Since $t(\underline{y})$ is binomial (n, p) under H_1, the detection probability of test $\tilde{\delta}_s(\underline{y})$ is given by

$$P_D = \sum_{k=\tau+1}^{n} \frac{n!}{(n-k)!\,k!} p^k (1-p)^{n-k}$$

$$+ \gamma \frac{n!}{(n-\tau)!\,\tau!} p^{\tau} (1-p)^{n-\tau}.$$

Although the false-alarm probability of $\tilde{\delta}_S$ is constant under H_0, its detection probability depends on p and this is not independent of the choice $P \in \mathcal{P}_1$. It can be shown that P_D increases monotonically from α to 1 as p increases from $\frac{1}{2}$ to 1.

The above test $\tilde{\delta}_S$ only uses the signs of the observations $y_1, \cdots, y_n$ and so it is known as the sign test. Although the sign test is α-level UMP for $\mathcal{P}_0$ vs. $\mathcal{P}_1$, we could do better than the sign test if we knew the exact distribution of the observations under the two hypotheses by using the likelihood ratio test between those two distributions. Or, the sign test is not a UMP α-level test for a particular $P \in \mathcal{P}_0$ vs. the class $\mathcal{P}_1$.

How much performance do we lose by assuming nothing about the distribution other than the very coarse assumptions made in the hypotheses (**)?

As a partial answer to this question, we consider an asymptotic ($n \to \infty$) analysis based on the Pitman

asymptotic relative efficiency (ARE) studied before. ARE for one detection relative to another is a measure of the relative number of samples that one needs to achieve the same performance as the other in the limit as the number of samples increases without bound.

To do so, let us consider again the model

$$H_0 : Y_k = N_k, \quad k = 1, 2, \cdots, n,$$

vs. (***)

$$H_1 : Y_k = N_k + \theta, \quad k = 1, 2, \cdots, n$$

where $N_1, \cdots, N_n$ are i.i.d. with zero mean. (***) is the problem of detecting a constant signal in i.i.d. additive noise.

If the noise distribution in (***) is $N(0, \sigma^2)$ with σ^2 unknown, then it can be shown (see, Lehmann 1986) that a UMP (among all unbiased tests) α-level test of H_0 vs. H_1 is given by

$$\tilde{\delta}_t(y) = \begin{cases} 1 & \text{if } \dfrac{\bar{y}}{[\bar{s}^2]^{1/2}} > \tau \\ \gamma & \cdots \quad = \cdots \\ 0 & \cdots \quad < \cdots \end{cases}$$

where $\bar{y}$ is the sample mean, i.e. $\bar{y} \triangleq \frac{1}{n}\sum_{k=1}^{n} y_k$, and $\bar{s}^2$ is the sample variance $\bar{s}^2 = \frac{1}{n}\sum_{k=1}^{n}(y_k - \bar{y})^2$.

The test $\tilde{\delta}_t$ is known as the t-test.

By choosing the threshold $\tau = \Phi^{-1}(1-\alpha)/\sqrt{n}$ and the randomization γ arbitrarily, the t-test is asymptotically nonparametric at $P_F = \alpha$ for any noise distribution with zero mean and finite variance:

$$P_F(\tilde{\delta}_t) = P_0(\bar{Y}/(\bar{S^2})^{\frac{1}{2}} > \tau) + \gamma P_0(\bar{Y}/(\bar{S^2})^{\frac{1}{2}} = \tau)$$

$$= P_0\left(\frac{1}{\sqrt{n}}\sum_{k=1}^{n} Y_k/(\bar{S^2})^{\frac{1}{2}} > \Phi^{-1}(1-\alpha)\right)$$

$$+ \gamma P_0\left(\frac{1}{\sqrt{n}}\sum_{k=1}^{n} Y_k/(\bar{S^2})^{\frac{1}{2}} = \Phi^{-1}(1-\alpha)\right).$$

By the weak law of large numbers, $\bar{S^2}$ converges in probability to $\mathrm{Var}(N_1)$, and by the central limit theorem, $n^{-\frac{1}{2}}\sum Y_k/(\bar{S^2})^{\frac{1}{2}}$ converges in distribution to an $N(0,1)$ random variable under H_0, so

$$\lim_{n\to\infty} P_F(\tilde{\delta}_t) = \frac{1}{\sqrt{2\pi}}\int_{\Phi^{-1}(1-\alpha)}^{\infty} e^{-x^2/2}\,dx = \alpha$$

(Note that γ is irrelavant since the limiting distribution is continuous).

The t-test with Gaussian noise is optimal and asymptotically nonparametric with finite variance noise.

If we impose the additional constraint that the noises have zero median in addition to zero mean, then the t-test corresponds to testing a subset of the distributions for the hypothesis problem (**).

It is of interest to compare the sign test and the t-test.

If we assume that the noise (H*) has a pdf f that has zero mean, variance $\sigma^2 < \infty$, and that is continuous, then it follows straightforwardly from the Pitman–Noether theorem that the asymptotic efficiency of the sign test relative to the t-test under (H*) is given by $ARE_{s,t} = 4\sigma^2 f^2(0)$.

For the particular case of Gaussian noise, in which f is the $N(0, \sigma^2)$ density,

$$ARE_{s,t} = 4\sigma^2 \left(\frac{1}{\sqrt{2\pi}\,\sigma}\right)^2 = \frac{2}{\pi} \cong 0.64$$

So that the t-test requires 64% of the samples required by an equivalent sign test.

For the Laplacian noise case, $f(x) = \frac{\beta}{2} e^{-\beta|x|}$, we have $\sigma^2 = 2/\beta^2$ and

$$ARE_{s,t} = \frac{8}{\beta^2}\left(\frac{\beta}{2}\right)^2 = 2.$$

Thus, the t-test requires twice as many samples as the equivalent sign test. It should be noted that the sign test is optimum in terms of asymptotic efficiency for the Laplacian noise case.

For any symmetric unimodal density $f(x) = f(-x)$

and $f(|x_1|) > f(|x_2|)$ if $|x_2| > |x_1|$, then

$$ARE_{s,t} \geq 1/3.$$

Thus, the t-test requires at least one-third of the numbers of samples required by an equivalent sign test.

Both the sign test and the t-test are used quite frequently in applications such as a CFAR radar detection.

* Rank Tests

Suppose we replace the sign test statistic

$$t(\underline{y}) = \sum_{k=1}^{n} u(y_k)$$

in $\tilde{\delta}_s$ with a weighted version

$$\sum_{k=1}^{n} \lambda_k \, u(y_k)$$

where λ_k is the rank of y_k in the samples $y_1, \ldots, y_n$ when reordered in increasing order of absolute values. Suppose we rank $y_1, \ldots, y_n$ as $y_{k_1}, \ldots, y_{k_n}$ where $|y_{k_1}| \leq |y_{k_2}| \leq \cdots \leq |y_{k_n}|$, and perform a threshold test based on the statistic

$$t_W(\underline{y}) = \sum_{i=1}^{n} i \, u(y_{k_i}).$$

The resulting test is known as the Wilcoxon test and it is an example of a rank test.

The Wilcoxon test statistic $t_W(\underline{y})$ can be rewritten as

$$t_W(\underline{y}) = \sum_{k=1}^{n} \sum_{j=1}^{k} u(y_k + y_j)$$

It can be shown that the Wilcoxon test is nonparametric for the hypothesis that $Y_1, \ldots, Y_n$ are i.i.d. with symmetric marginal distribution, i.e., $F_{Y_k}(b) = 1 - F_{Y_k}(-b)$ for all real b. Note that this is a smaller class of models than the class all distributions with zero median.

The asymptotic efficiency of the Wilcoxon test relative to the t-test is given by the Pitman-Noether theorem as

$$ARE_{W,t} = 12\sigma^2 \left[\int_{-\infty}^{\infty} f^2(x)\,dx\right]^2.$$

For the Gaussian noise, $N(0, \sigma^2)$,

$$ARE_{W,t} = 3/\pi = 0.995$$

$\Rightarrow$ The Wilcoxon test is nearly optimum for Gaussian noise.

For the Laplacian case, $ARE_{W,t} = 1.5$, i.e., it has a loss in efficiency of 25% relative to the sign test.

It can be shown by minimizing $\int_{-\infty}^{\infty} f^2(x)\,dx$ subject to the constraint $\int_{-\infty}^{\infty} x^2 f^2(x)\,dx = \sigma^2$,

that $ARE_{W,t} \geq 0.864$

for any symmetric noise density. Thus, the Wilcoxon

test is never less than 86.4% as efficient as the t-test. Since there is no corresponding upper bound on ARE, the Wilcoxon test offers substantial advantage over the t-test. However, a disadvantage of the Wilcoxon test is that all samples must be stored in order to compute its test statistic, which is not true for either the sign test or the t-test.

§4.2. Robust Detection

The problem of testing between two possible marginal distributions P_0 and P_1 for an i.i.d. sequence $Y_1, \cdots, Y_n$ has the optimum tests based on the likelihood ratio

$$L(\underline{y}) = \prod_{k=1}^{n} \frac{P_1(y_k)}{P_0(y_k)}.$$

To calculate this likelihood ratio may incur problem when $P_1(y_k) >> P_0(y_k)$ or $P_0(y_k) >> P_1(y_k)$ and P_0 and P_1 are not precisely known but with errors. $L(\underline{y})$ is unbounded. This is known as lack of robustness.

One obvious way of stabilizing the performance of the likelihood ratio test is to replace the likelihood ratio P_1/P_0 with a version that is limited from above or below as follows. Let $l \triangleq P_1/P_0$ and

$$[l]_a^b(y) = \begin{cases} b & \text{if } l(y) > b, \\ l(y) & \text{if } a \le l(y) \le b, \\ a & \text{if } l(y) < a, \end{cases}$$

where $0 < a < b < \infty$. For properly chosen a and b, a test based on $\prod_{k=1}^{n} [l]_a^b(y_k)$ would not exhibit the difficulties for unbounded l.

Recall that if P_0 is the true marginal distribution of the Y_k, the false-alarm probability of a test δ is

$$P_F(\delta, P_0) = \int_\Gamma \delta(\underline{y}) \left[\prod_{k=1}^{n} p_0(y_k) \right] \mu(d\underline{y})$$

where p_0 is the marginal density for P_0. Similarly, the miss probability when P_1 is the true marginal distribution is $P_M(\delta, P_1) = \int_\Gamma [1 - \delta(\underline{y})] \left[\prod_{k=1}^{n} p_1(y_k) \right] \mu(d\underline{y})$.

Assuming for simplicity that costs are uniform, the three usual criteria for simple binary hypothesis testing are:

(i) $\min_\delta [\pi_0 P_F(\delta, P_0) + \pi_1 P_M(\delta, P_1)]$ (Bayes)

(ii) $\min_\delta [\max \{ P_F(\delta, P_0), P_M(\delta, P_1) \}]$ (Minimax)

(iii) $\min_\delta P_M(\delta, P_1)$ subject to $P_F(\delta, P_0) \le \alpha$ (Neyman-Pearson).

Instead of assuming that the marginal distribution of Y_k is exactly P_0 or P_1, we assume that the marginal lies either in a neighborhood $\mathcal{P}_0$ of P_0 or in a neighborhood $\mathcal{P}_1$ of P_1 such as

$$(1-\varepsilon) P_j + \varepsilon M_j, \quad j = 0, 1 \qquad (**)$$

for unknown and arbitrary "contaminating"

distribution M_0, M_1, $0 < \varepsilon < 1$ is a number representing the degree of uncertainty to be placed on the model.

By replacing $P_F(\delta, P_0)$ and $P_M(\delta, P_1)$ in (i)–(iii) with their worst-case values:

$$P_F(\delta, \mathcal{P}_0) \triangleq \sup_{P \in \mathcal{P}_0} P_F(\delta, P)$$

$$P_M(\delta, \mathcal{P}_1) \triangleq \sup_{P \in \mathcal{P}_1} P_M(\delta, P),$$

we arrive at the alternative design problems:

(i)' $\min_{\delta} [\pi_0 P_F(\delta, \mathcal{P}_0) + \pi_1 P_M(\delta, \mathcal{P}_1)]$

(ii)' $\min_{\delta} [\max\{P_F(\delta, \mathcal{P}_0), P_M(\delta, \mathcal{P}_1)\}]$

(iii)' $\min_{\delta} P_M(\delta, \mathcal{P}_1)$ subject to $P_F(\delta, \mathcal{P}_0) \leq \alpha$.

Solutions to (i)–(iii) will have the best worst-case performance (over the neighborhoods $\mathcal{P}_0$ and $\mathcal{P}_1$) of all possible tests. Of course, it is possible that such tests might be overly conservative and certainly would be true if $\mathcal{P}_0$ and $\mathcal{P}_1$ were too large.

Although there is a general approach to solving problems (i)'–(iii') due to Huber and Strassen 1973, it is too involved. We instead will focus on the solutions to (i')–(iii') for the particular uncertainty

neighborhoods described in (*) known as ε-contaminated mixtures.

It turns out that the solutions to (i')–(iii') for ε-contaminated mixtures are the corresponding optimum tests for (i)–(iii) when $\mathcal{P}_0$ and $\mathcal{P}_1$ in (i)–(iii) are replaced by a pair $Q_0 \in \mathcal{P}_0$ and $Q_1 \in \mathcal{P}_1$ of <u>least favorable</u> distribution. Q_0 and Q_1 are given in terms of their densities by

(1) $$q_0(y_k) = \begin{cases} (1-\varepsilon)\, p_0(y_k) & \text{if } p_1(y_k) < c'' p_0(y_k) \\ \frac{1-\varepsilon}{c''}\, p_1(y_k) & \text{if } p_1(y_k) \geq c'' p_0(y_k) \end{cases}$$

and

(2) $$q_1(y_k) = \begin{cases} (1-\varepsilon)\, p_1(y_k) & \text{if } p_1(y_k) > c' p_0(y_k) \\ c'(1-\varepsilon)\, p_0(y_k) & \text{if } p_1(y_k) \leq c' p_0(y_k), \end{cases}$$

where $0 < c' < 1 < c'' < \infty$ are two constants chosen so that Q_0 and Q_1 are probability distributions, i.e., so that their total probability equals 1. This condition is given by

$$(1-\varepsilon)\left[P_0(l(Y_k) < c'') + P_1(l(Y_k) \geq c'')/c''\right] = 1$$

and $$(1-\varepsilon)\left[P_1(l(Y_k) > c') + c' P_0(l(Y_k) \leq c')\right] = 1.$$

Since each of (i)–(iii) is solved by a likelihood-ratio test, the solutions to (i')–(iii') are likelihood-ratio tests between Q_0 and Q_1, namely, they are based on the likelihood ratio,

$$\prod_{k=1}^{n} \frac{q_1(y_k)}{q_0(y_k)}.$$

Using (1), we have

$$\frac{q_1(y_k)}{q_0(y_k)} = \begin{cases} c' & \text{if } l(y_k) < c' \\ l(y_k) & \text{if } c' \le l(y_k) \le c'' \\ c'' & \text{if } l(y_k) > c''. \end{cases}$$

Thus, the solutions to (i')–(iii') are threshold tests based on $\prod_{k=1}^{n} [l]_{c'}^{c''}(y_k)$.

This implies the the worst-case performance of δ_R is in fact its performance at the pair of distributions (Q_0, Q_1) for which it is optimum.

* Remarks

1. When ε is too large, Q_0 and Q_1 will have overlap. In this case, $c' = c'' = 1$ and Q_0 and Q_1 will be equal.

2. Even for small ε, the difference in performance between the tests based on l and $[l]_{\varepsilon}^{c''}$ can be quite dramatic. Let us see an example.

Consider the Bayesian formulation (i) with $\pi_0 = \pi_1 = \frac{1}{2}$. Then, for either test, the threshold is unity and the randomization is arbitrary.

Suppose $0 < l(y_k) < \infty$ and $\sup_{y_k \in \mathbb{R}} l(y_k) = \infty$. Since M_0 is arbitrary, it can be chosen to put all of its probability on a value of y_k for which $l(y_k)$ is arbitrarily large. In this way, any of the observations can cause δ_0 to commit a false-alarm with probability ε, where δ_0 is the test based on l.

Since there are n observations, this implies that

$$P_F(\delta_0, \mathcal{P}_0) \geq 1 - (1-\varepsilon)^n.$$

Similarly, if $\inf_{y_k \in \mathbb{R}} l(y_k) = 0$, the miss probability of δ_0 over $\mathcal{P}_1$ satisfies

$$P_M(\delta_0, \mathcal{P}_1) \geq 1 - (1-\varepsilon)^n.$$

So, the worst-case average error probability of δ_0 satisfies

$$\sup P_e(\delta_0) = \tfrac{1}{2} P_F(\delta_0, \mathcal{P}_0) + \tfrac{1}{2} P_M(\delta_0, \mathcal{P}_1) \geq 1 - (1-\varepsilon)^n,$$

where the supremum is taken over $\mathcal{P}_0$ and $\mathcal{P}_1$.

For any $\varepsilon > 0$, $\lim_{n \to \infty} (1-\varepsilon)^n = 0 \Rightarrow \lim_{n \to \infty} [\sup P_e(\delta_0)] = 1$.

So, the performance of δ_0 can be arbitrarily bad, and in fact it can be worse than simply guessing at the hypothesis on the basis of a coin toss, since guessing would cause an error probability of $1/2$.

In constrast, consider the test δ_R based on

8./80.

$$\sup P_e(\delta_R) = \frac{1}{2} P_F(\delta_R, Q_0) + \frac{1}{2} P_M(\delta_R, Q_1)$$

Applying the Chernoff bound to the right-hand side, we have

$$\sup P_e(\delta_R) \leq \frac{1}{2} \left[\int [Q_0 Q_1]^{\frac{1}{2}} \right]^n$$

If $Q_0 \neq Q_1$, we have

$$\int [Q_0 Q_1]^{1/2} < 1$$

so that

$$\lim_{n \to \infty} [\sup P_e(\delta_R)] = 0.$$

From the above, we see that in terms of worst-case performance, $P_e(\delta_B)$ converges exponentially to unity while $P_e(\delta_R)$ converges exponentially to zero.

It should be noted that under nomial conditions, $P_e(\delta_0)$ also converges exponentially to zero. Thus, δ_R achieves, in its worst case, behavior similar to the nomial behavior of δ_0, while δ_0 in its worst case behaves radically differently.

3. Solutions to (i')-(iii') are known for a number of uncertainty models other than the ϵ-contaminated mixture. For example, several interesting types of neighborhoods that can be treated in this context

are of the form $\mathcal{P}_0 = \{P \mid \rho(P, P_0) \le \varepsilon_0\}$ and
$\mathcal{P}_1 = \{P \mid \rho(P, P_1) \le \varepsilon_1\}$
for some measure ρ of distance between probability distributions.

The ε-contaminated model can also be generalized to allow for time-varying nominals and ε's. In this case, $\prod_{k=1}^{n} [l]_{c'}^{c''}(y_k)$ is similarly replaced by

$$\prod_{k=1}^{n} [l_k]_{c_k}^{c_k''}(y_k),$$

where l_k is the nominal likelihood ratio for the kth sample.

* Example: The Correlator—Limiter

Consider the coherent signal detection problem:

$$H_0: Y_k = N_k, \quad k = 1, 2, \cdots, n, \qquad (*)$$

vs.

$$H_1: Y_k = N_k + \theta s_k, \quad k = 1, 2, \cdots, n,$$

where $s_1, s_2, \cdots, s_n$ is a known signal sequence; $N_1, \cdots, N_n$ is an i.i.d. sequence of $N(0, 1)$ noise samples; and θ is a known positive amplitude.

The kth sample likelihood ratio for this problem is given before by

$$l_k(y_k) = \exp[\theta s_k(y_k - \theta s_k/2)],$$

and $$\log L(\underline{y}) = \theta \sum_{k=1}^{n} s_k(y_k - \theta s_k/2)$$

which leads to the correlation detector

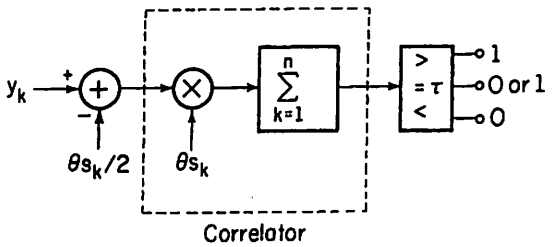


FIGURE III.E.1. Nominally optimum detector for coherent signals in Gaussian noise.

In practice, the above model (*) can only be assumed to be approximately correct. In particular, the noise distribution is unlikely to be exactly $N(0,1)$, and the measurement model in which signal and noise are assumed to be additive is not completely accurate due to nonlinearities in the observation mechanism. Thus, to have a more realistic model, we could modify (*) to

$$H_0: Y_k \sim (1-\varepsilon)P_0 + \varepsilon M_0, \quad k = 1, 2, \cdots, n$$

vs. (**)

$$H_1: Y_k \sim (1-\varepsilon)P_1^{(k)} + \varepsilon M_1^{(k)}, \quad k = 1, 3 \cdots n$$

where P_0 is the $N(0,1)$ distribution, $P_1^{(k)}$ is the $N(\theta s_k, 1)$ distribution, M_0 and $M_1^{(k)}$, $k = 1, 2, \cdots, n$, are arbitrary (Note: M_0 could also be allowed to change with k; this would not change the solution given below), and ε is between 0 and 1.

Note that $l_k(y_k) = \exp[\theta s_k (y_k - \theta s_k/2)]$, we have

$$0 < l_k(y_k) < \infty \text{ and } \sup_{y_k \in \mathbb{R}} l_k(y_k) = \infty,$$

$$\inf_{y_k \in \mathbb{R}} l_k(y_k) = 0.$$

Thus, the correlation detector will suffer the performan degradation discussed above in the presence of uncertainty modeled as in (**).

Huber's robust likelihood ratio test δ_R is thus performed. The logarithm of the robust likelihood ratio

$$L_R(\underline{y}) = \prod_{k=1}^{n} [l_k]_{c_k'}^{c_k''}(y_k)$$

can be written as

$$\log L_R(\underline{y}) = \sum_{k=1}^{n} [\theta s_k (y_k - \theta s_k/2)]_{d_k'}^{d_k''},$$

where $d_k' \triangleq \log c_k'$ and $d_k'' \triangleq \log c_k''$.

Using the symmetry properties of the $N(0,1)$ distribution and of the likelihood ratio, it can be shown that $d_k' = -d_k''$ and d_k'' is the solution to

$$\underline{\Phi}\left(\frac{d_k''}{\theta|s_k|} + \frac{\theta|s_k|}{2}\right) + e^{-d_k''}\left[1 + \underline{\Phi}\left(\frac{d_k''}{\theta|s_k|} - \frac{\theta|s_k|}{2}\right)\right]$$

$$= (1-\varepsilon)^{-1}$$

where Φ is the $N(0,1)$ cumulative distribution function. The detector is depicted below:

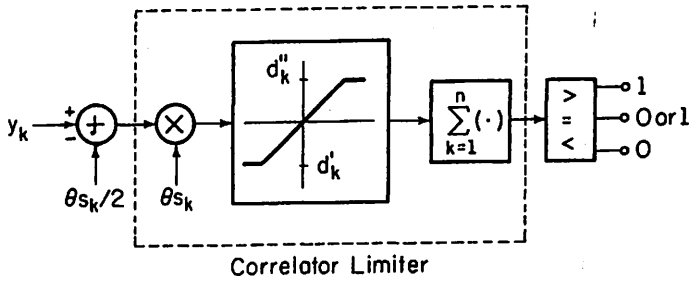
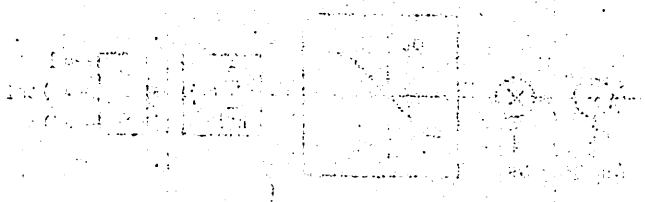


FIGURE III.E.2. Robust detector for coherent signals in nominally Gaussian noise.

This is identical to the correlation receiver of Fig. III.E.1 except there is a (time-varying) limiter with limits $\pm d_k'$ placed between the multiplier and the accumulator in the receiver.

This detector structure is called a <u>correcter-limiter</u> and it was originally derived and analyzed by Martin-Schwartz (1971).



Chapter IV. Elements of Parameter Estimation

§4.1. Introduction

In the signal detection we studied previously, we considered how to optimally decide between two possible statistical situations. In many applications, we are interested not in making a choice between two (or several) discrete situations, but rather in making a choice among a continuum of possible states of nature.

We wish to determine as accurately as possible the actual value of the parameter from the observation. Such problems are known as parameter estimation problems. This chapter is to discuss the basic ideas on to design of optimum procedures for estimating parameters.

§4.2. Bayesian Parameter Estimation.

Model: A family of distributions for the random observation Y, indexed by a parameter θ taking values in a parameter set Λ: $\{P_\theta; \theta \in \Lambda\}$, where P_θ denotes a distribution on the observation space $(\Gamma, \mathcal{G})$, Λ is a subset of $\mathbb{R}^m$ for some m,

Goal: find a function $\hat{\theta}: \Gamma \to \Lambda$ such that $\hat{\theta}(y)$ is the "best" guess of the true value of θ, i.e., the value of θ for which $Y \sim P_\theta$, based on the observation $Y = y$.

The solution to this problem depends on the criterion of goodness by which we measure estimation performan

Cost: $C: \Lambda \times \Lambda \to \mathbb{R}$ such that $C[a, \theta]$ is the cost of estimating a true value of θ as a.

Conditional risk: $R_\theta(\hat{\theta}) = E_\theta\{C[\hat{\theta}(Y), \theta]\}$,
Cost averaged over Y for each $\theta \in \Lambda$.

Bayes risk: $r(\hat{\theta}) \triangleq E\{R_\Theta(\hat{\theta})\}$
By interpreting the actual parameter value θ is the realization of a random variable Θ.

The appropriate design goal is to find an estimator minimizing $r(\hat{\theta})$. Such as estimator is known as a Bayes estimate of θ.

Noting that $R_\theta(\hat{\theta}) = E\{C[\hat{\theta}(Y), \Theta] \mid \Theta = \theta\}$, we have

$$r(\hat{\theta}) = E\{C[\hat{\theta}(Y), \Theta]\}$$
$$= E\{E\{C[\hat{\theta}(Y), \Theta] \mid Y\}\}$$

$\Rightarrow$ The Bayes estimate of θ can be found (if it exists) by minimizing, for each $y \in \Gamma$, the posterior cost given $Y = y$:

$$E\{C[\hat{\theta}(y), \Theta] \mid Y = y\}.$$

This is the same procedure as that followed in the Bayes hypothesis testing problem. If we assume that Θ has a conditional density $w(\theta|y)$ given $Y = y$ for each $y \in \Gamma$, then the Bayes estimate $\hat{\theta}(y)$ corresponding to $y \in \Gamma$ can be sought by minimizing

$$\int_{\Lambda} C[\hat{\theta}(y), \theta]\, w(\theta|y)\, \mu(d\theta).$$

* Case 4.2.1: Minimum-Mean-Squared-Error (MMSE) Estimation

$\Lambda = \mathbb{R}$ and $E[\Theta^2] < \infty$. A commonly used cost function is

$$C[a, \theta] = (a - \theta)^2, \quad (a, \theta) \in \mathbb{R}^2.$$

This measures the square of the estimation error $\hat{\theta}(y) - \theta$. The Bayes risk here is $E\{(\hat{\theta}(Y) - \Theta)^2\}$ known as the mean-squared error (MSE). In this case, the Bayes estimate is a minimum-mean-squared-error (MMSE) estimator.

The posterior cost given $Y = y$ is, in this case,

$$E\{(\hat{\theta}(y)-\Theta)^2 \mid Y=y\}$$
$$= E\{[\hat{\theta}(y)]^2 \mid Y=y\} - 2E\{\hat{\theta}(y)\Theta \mid Y=y\} + E\{\Theta^2 \mid Y=y\}$$
$$= [\hat{\theta}(y)]^2 - 2\hat{\theta}(y)\, E\{\Theta \mid Y=y\} + E\{\Theta^2 \mid Y=y\}.$$

By setting the derivative to be zero above, we have

$$\hat{\theta}_{MMSE}(y) = E\{\Theta \mid Y=y\},$$

i.e., the MMSE estimate of Θ given $Y=y$ is the conditional mean of Θ given $Y=y$. It is sometimes called the conditional mean estimate (CME).

* Case 4.2.2: Minimum-Mean-Absolute-Error (MMAE) Estimation

Another cost function is, when $\Lambda = \mathbb{R}$,

$$C[a,\theta] = |a-\theta|, \quad (a,\theta) \in \mathbb{R}^2.$$

The Bayes risk here is $E\{|\hat{\theta}(Y)-\Theta|\}$.

⇒ The mean-absolute error and the corresponding Bayes estimate is known as the minimum-mean-absolute-error (MMAE) estimate.

Note that, if $P(X \geq 0) = 1$, then

$$E\{X\} = \int_0^\infty P(X > x)\, dx.$$

$$\Rightarrow E\{|\hat{\theta}(y)-\theta| \,|\, \overline{Y}=y\}$$

$$=\int_0^\infty P(|\hat{\theta}(y)-\theta|>x \,|\, \overline{Y}=y)\,dx$$

$$=\int_0^\infty P(\theta>x+\hat{\theta}(y) \,|\, \overline{Y}=y)\,dx$$

$$+\int_0^\infty P(\theta<-x+\hat{\theta}(y) \,|\, \overline{Y}=y)\,dx$$

$$=\int_{\hat{\theta}(y)}^\infty P(\theta>t \,|\, \overline{Y}=y)\,dt$$

$$+\int_{-\infty}^{\hat{\theta}(y)} P(\theta<t \,|\, \overline{Y}=y)\,dt$$

$$\Rightarrow \frac{\partial}{\partial \hat{\theta}(y)} E\{|\hat{\theta}(y)-\theta| \,|\, \overline{Y}=y\}$$

$$= P(\theta<\hat{\theta}(y) \,|\, \overline{Y}=y\} - P(\theta>\hat{\theta}(y) \,|\, \overline{Y}=y\}.$$

It is a non-decreasing function of $\hat{\theta}(y)$ that approaches -1 as $\hat{\theta}(y)\to-\infty$ and $+1$ as $\hat{\theta}(y)\to\infty$.

$\Rightarrow E\{|\hat{\theta}(y)-\theta| \,|\, \overline{Y}=y\}$ achieves its minimum over $\hat{\theta}(y)$ at the point (or the set of points) where its derivative changes sign.

The Bayes estimate in this case, denoted by $\hat{\theta}_{ABS}(y)$, is any point such that

$$P(\Theta < t \mid Y = y) \le P(\Theta > t \mid Y = y), \quad t < \hat{\theta}_{ABS}(y)$$

and

$$P(\Theta < t \mid Y = y) \ge P(\Theta > t \mid Y = y), \quad t > \hat{\theta}_{ABS}(y).$$

A point $\hat{\theta}_{ABS}(y)$ satisfies the above property is a median of the conditional distribution of Θ given $Y = y$. Thus, the MMAE estimate is a conditional median estimate.

* Case 4.2.3. Maximum A Posteriori Probability (MAP) Estimation

$\Lambda = \mathbb{R}$, and Consider the uniform cost function

$$C[a, \theta] = \begin{cases} 0, & \text{if } |a - \theta| \le \Delta, \\ 1, & \text{if } |a - \theta| > \Delta, \end{cases}$$

where $\Delta > 0$. For an estimator $\hat{\theta}$, the average posterior cost given $Y = y$:

$$E\{C[\hat{\theta}(y), \Theta] \mid Y = y\}$$

$$= P[|\hat{\theta}(y) - \Theta| > \Delta \mid Y = y]$$

$$= 1 - P(|\hat{\theta}(y) - \Theta| \le \Delta \mid Y = y).$$

* First let us consider Θ is a discrete random variable taking values in $\Lambda = \{\theta_0, \cdots, \theta_{M-1}\}$ with $|\theta_i - \theta_j| > \Delta$ for $i \ne j$. Then,

$$E\{C[\hat{\theta}(y), \Theta] | Y=y\} = 1 - P(\Theta = \hat{\theta}(y) | Y=y)$$

$$= 1 - \omega(\hat{\theta}(y)|y), \quad \text{for } \hat{\theta}(y) \in \Lambda.$$

where $\omega(\theta|y)$ is the conditional probability mass function of Θ given $Y=y$.

$\Rightarrow$ The Bayes estimate in this case is given for each $y \in \Gamma$ by any value of θ that maximizes $\omega(\theta|y)$ over $\theta \in \Lambda$, i.e., the Bayes estimate is the value of Θ that has the maximum a posteriori probability of occuring given $Y=y$.

* Now if $\Lambda = \mathbb{R}$ and Θ is a conditional random variable with conditional density function $\omega(\theta|y)$ given $Y=y$. Then,

$$E\{C[\hat{\theta}(y), \Theta] | Y=y\}$$

$$= 1 - \int_{\hat{\theta}(y)-\Delta}^{\hat{\theta}(y)+\Delta} \omega(\theta|y)\, d\theta$$

Thus, for small Δ and smooth $\omega(\theta|y)$,

$$\int_{\hat{\theta}(y)-\Delta}^{\hat{\theta}(y)+\Delta} \omega(\theta|y)\, d\theta \cong 2\Delta\, \omega(\theta|y)\Big|_{\theta=\hat{\theta}(y)}$$

and the right hand side is maximized by choosing $\hat{\theta}(y)$ to be the value of θ maximizing $\omega(\theta|y)$ over Λ.

* In the above two cases, the uniform cost criterion leads to the procedure for estimating θ as that value maximizing the a posteriori (discrete or continuous) density $w(\theta|y)$ [similarly, with θ discrete but taking on infinitely many values, it can be argued that the conditional risk function is minimized approximately by choosing $\hat{\theta}(y) \neq o$ maximizing the conditional mass function $w(\theta|y)$]

This estimate is known as the maximum a posteriori (MAP) estimate and is denoted by

$$\hat{\theta}_{MAP}$$

Although this estimate often only approximates the Bayes estimate for uniform cost with small Δ, the MAP criterion is widely used.

Note that a point at which a density achieves its maximum value is termed a mode of the corresponding probability distribution. Thus, since $\hat{\theta}_{MAP}$ estimates θ by the mode of its conditional distribution, it is a conditional mode estimate.

* From the above three cases, we see that Bayes estimates are determined from the

conditional distribution of the parameter given the observations. In particular, the MMSE, MMAE, and MAP estimater are the mean, median, and mode of this distributions, respectively.

As in the case of hypothesis testing, we can think of the observation as a means for converting the prior distribution of the parameter into a posterior distribution. In general, Bayes estimators are features of this posterior distribution.

With a parameterized probability distribution family $\{P_\theta, \theta \in \Lambda\}$ with density family $\{p_\theta, \theta \in \Lambda\}$,

$$w(\theta|y) = \frac{p_\theta(y)\, w(\theta)}{\int_\Lambda p_\theta(y) w(\theta) \mu(d\theta)}$$

$\|$

$p(y)$, the unconditional density of Y.

Then, the Bayes estimates for the three cases above can be obtained straightforwardly

$$\hat{\theta}_{MAP}(y) = \arg \max_{\theta \in \Lambda} p_\theta(y)\, w(\theta)$$

$$= \arg \max_{\theta \in \Lambda} (\log p_\theta(y) + \log w(\theta))$$

⇒ If Θ is a continuous random variable given $Y=y$, then for sufficiently smooth p_θ and w, a necessary condition for this maximization is

$$\frac{\partial}{\partial\theta}\log p_\theta(y)\Big|_{\theta=\hat{\theta}_{MAP}(y)} = -\frac{\partial}{\partial\theta}\log w(\theta)\Big|_{\theta=\hat{\theta}_{MAP}(y)}$$

↑ MAP equation.

* Example 1: Estimation of the Parameter of an Exponential Distribution.

Consider $\Lambda=(0,\infty)$, $\Gamma=\mathbb{R}$, the observations have the following conditional probability function given $\Theta=\theta$:

$$p_\theta(y)=\begin{cases}\theta e^{-\theta y} & \text{if } y\geq 0\\ 0 & \text{if } y<0\end{cases}$$

This is the exponential density with parameter θ: Modelling the time intervals between successive events occuring randomly in time, such as, messages or data packets arriving at a communications switching station; vehicles arriving at an intersection of roads; photons emitting from a coherent light source; or devices failing in a logic circuit.

The parameter θ in this model can be interpreted

as the rate of each occurences.

Suppose that our prior information about Θ is that it also has an exponential distribution with density

$$w(\theta) = \begin{cases} \alpha e^{-\alpha\theta} & \text{if } \theta \geq 0 \\ 0 & \text{if } \theta < 0, \end{cases}$$

where $\alpha > 0$ is known.

Then, the posterior distribution of Θ given $Y = y$ is

$$w(\theta|y) = \frac{\alpha\theta e^{-(\alpha+y)\theta}}{\int_0^\infty \alpha\theta e^{-(\alpha+\theta)\theta} d\theta}$$

$$= (\alpha+y)^2 \theta e^{-\theta(\alpha+y)}$$

for $\theta \geq 0$ and $y \geq 0$ and $w(\theta|y) = 0$ otherwise.

$$\hat{\theta}_{MMSE}(y) = \int_0^\infty \theta w(\theta|y) d\theta$$

$$= (\alpha+y)^2 \int_0^\infty \theta^2 e^{-\theta(\alpha+y)} d\theta$$

$$= \frac{2}{\alpha+y}.$$

For a fixed α, this estimate of Θ varies inversely with y, which is intuitively reasonable: Since a large inter- arrival time (large y)

would be evidence of a low rate (small θ).

$$MMSE = r(\hat{\theta}_{MMSE})$$

$$= E\{E\{(\hat{\theta}_{MMSE}(Y) - \Theta)^2 | Y\}\}$$

$$= E\{E\{(\Theta - E\{\Theta|Y\})^2 | Y\}\}$$

$$= E\{Var(\Theta|Y)\}.$$

Since $Var(\Theta|Y=y) = E\{\Theta^2|Y=y\} - E^2\{\Theta|Y=y\}$,

we have $Var(\Theta|Y=y) = \int_0^\infty \theta^2 w(\theta|y)\,d\theta - (\hat{\theta}_{MMSE}(y))^2$

$$= (\alpha+y)^2 \int_0^\infty \theta^3 e^{-\theta(\alpha+y)}\,d\theta - \frac{4}{(\alpha+y)^2}$$

$$= \frac{2}{(\alpha+y)^2}.$$

Thus,

$$MMSE = E\left\{\frac{2}{(\alpha+y)^2}\right\} = \int_0^\infty \frac{2}{(\alpha+y)^2} p(y)\,dy$$

$$= \int_0^\infty \frac{2\alpha}{(\alpha+y)^4}\,dy = \frac{2}{3\alpha^2}$$

where $p(y) = \int_0^\infty \alpha\theta e^{-(\alpha+y)}\,d\theta = \frac{\alpha}{(\alpha+y)^2}$.

The MMAE estimate, $\hat{\theta}_{ABS}(y)$, is the median of $w(\theta|y)$.

Since θ is continuous given $Y=y$, we can find $\hat{\theta}_{ABS}(y)$ by solving

$$\int_{\hat{\theta}_{ABS}(y)}^{\infty} w(\theta|y)\,d\theta = \frac{1}{2}$$

$$\Rightarrow [1+(\alpha+y)\hat{\theta}_{ABS}(y)]\, e^{-(\alpha+y)\hat{\theta}_{ABS}(y)} = \frac{1}{2}$$

$$\Rightarrow \hat{\theta}_{ABS}(y) = \frac{T_0}{\alpha+y},$$

where T_0 is the solution to $(1+T_0)e^{-T_0} = \frac{1}{2}$, i.e., $T_0 \cong 1.68$.

Comparing the MMSE and MMAE, they only differ by a constant.

The MAP estimate of θ can be found below:

$$\frac{\partial}{\partial\theta}[\log p_\theta(y) + \log w(\theta)] = \frac{\partial}{\partial\theta}[\log\theta - \theta y + \log\alpha - \alpha\theta]$$

$$= \theta^{-1} - (\alpha+y)$$

and

$$\frac{\partial^2}{\partial \theta^2}\left[\log p_\theta(y) + \log w(\theta)\right] = -\theta^{-2} < 0$$

We know that $w(\theta|y)$ has its unique maximum at

$$\hat{\theta}_{MAP}(y) = \frac{1}{\alpha + y}.$$

Again, it only differs with MMSE and MMAE with a scale factor.

Which one is what you need has to be decided by a particular application problem for which cost function fits better.

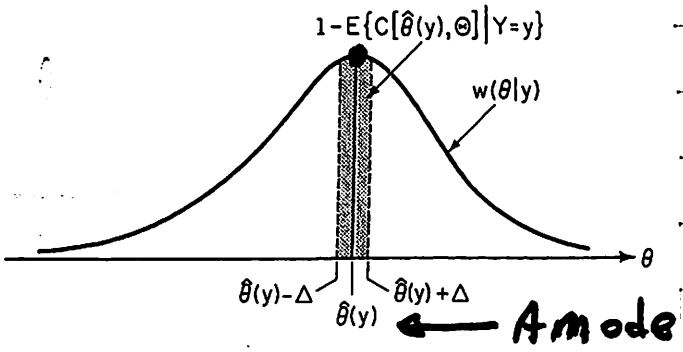


FIGURE IV.B.1. Illustration of MAP estimation.

IV.F Exercises

1. Suppose Θ is a random parameter and that, given $\Theta = \theta$, the real observation Y has density

$$p_\theta(y) = (\theta/2)e^{-\theta|y|}, \quad y \in \mathbb{R}.$$

Suppose further that Θ has prior density

$$w(\theta) = \begin{cases} 1/\theta, & 1 \le \theta \le e \\ 0, & \text{otherwise.} \end{cases}$$

(a) Find the MAP estimate of Θ based on Y.

(b) Find the MMSE estimate of Θ based on Y.

* Example 2: Estimation of Signal Amplitude

Consider $\Gamma = \mathbb{R}^n$ and $\Lambda = \mathbb{R}$ with

$$Y_k = N_k + \Theta s_k, \quad k = 1, 2, \cdots, n,$$

where $\underline{N} \sim N(0, \Sigma)$, $\underline{s}$ is known, $\Theta \sim N(\mu, v^2)$, and $\underline{N}$ and Θ are independent.

Given $\Theta = \theta$, we have $\underline{Y} \sim N(\theta \underline{s}, \Sigma)$. Thus the posteriori density for Θ is

$$w(\theta|\underline{y}) = \frac{\frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma|^{1/2}} e^{-\frac{1}{2}(\underline{y}-\theta\underline{s})^T\Sigma^{-1}(\underline{y}-\theta\underline{s})} \frac{1}{\sqrt{2\pi}\,v} e^{-(\theta-\mu)^2/(2v^2)}}{\int_{-\infty}^{\infty} \frac{1}{(2\pi)^{n/2}|\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(\underline{y}-\theta\underline{s})^T\Sigma^{-1}(\underline{y}-\theta\underline{s})} \frac{1}{\sqrt{2\pi}\,v} e^{-(\theta-\mu)^2/(2v^2)} d\theta}$$

$$= K(\underline{y}) \exp\left\{-\frac{\theta^2}{2}\left(d^2 + \frac{1}{v^2}\right) + \theta\left(\underline{s}^T\Sigma^{-1}\underline{y} + \frac{\mu}{v^2}\right)\right\},$$

where $d^2 = \underline{s}^T\Sigma^{-1}\underline{s}$ and $K(\underline{y})$ is a function depending on $\underline{y}$ but not on θ.

Note $w(\theta|\underline{y})$ is a Gaussian density. So if $w(\theta|\underline{y})$ were $N(m, q^2)$, we would have

$$w(\theta|\underline{y}) = \frac{1}{\sqrt{2\pi}\,q} e^{-(\theta-m)^2/(2q^2)} = \frac{e^{-m^2/(2q^2)}}{\sqrt{2\pi}\,q} e^{-\theta^2/(2q^2) + \theta m/q^2}$$

$\Rightarrow$ Given $\underline{Y} = \underline{y}$, $\Theta \sim N(m, q^2)$ with

$$q^2 = \left(d^2 + \frac{1}{v^2}\right)^{-1}$$

$$m = \left(d^2 + \frac{1}{v^2}\right)^{-1}\left(\underline{s}^T\Sigma^{-1}\underline{y} + \frac{\mu}{v^2}\right)$$

and $K(\underline{y})$ becomes $e^{-m^2/(2\sigma^2)} / (\sqrt{2\pi}\,\sigma)$.

$\hat{\theta}_{MMSE}(\underline{y})$ is the mean of $W(\theta|\underline{y})$

$$\Rightarrow \hat{\theta}_{MMSE}(\underline{y}) = \frac{\underline{s}^T \Sigma^{-1} \underline{y} + \frac{\mu}{v^2}}{d^2 + \frac{1}{v^2}}$$

$$= \frac{v^2 d^2 \hat{\theta}_1(\underline{y}) + \mu}{v^2 d^2 + 1}$$

where $\hat{\theta}_1(\underline{y}) \triangleq \underline{s}^T \Sigma^{-1} \underline{y} / d^2$

$$MMSE = E\{Var(\Theta|\underline{Y})\} = \frac{1}{d^2 + \frac{1}{v^2}} = \frac{v^2}{v^2 d^2 + 1}$$

Since $Var(\Theta|Y) = (d^2 + \frac{1}{v^2})^{-1}$ that does not depend on $\underline{Y}$.

Since the Gaussian density is symmetric about its mean and it achieves its maximum at it mean, the conditional median and conditional mode equal the conditional mean, i.e.,

$$\hat{\theta}_{ABS} = \hat{\theta}_{MAP} = \hat{\theta}_{MMSE}.$$

* The behavior of this estimator well illustrates the nature of Bayesian estimation.

v^2 determines the accuray of our prior knowledge about Θ; that is, the smaller v^2 is, the more

accurately we know θ in the absence of observation. From what we studied in the previous chapter, we know that d^2 is a measure of the quality with which $\underline{s}$ can be distinguished from the $N(0, \Sigma^2)$ noise, that is, d^2 is a measure of the accuracy of our observations in terms of producing information about the signal — large d^2 components to high-quality observations and small d^2 to low-quality observations.

$\Rightarrow$ If $v^2 d^2$ is very small relative to the other quantities in the estimate, we have $\hat{\theta}_{MMSE}(\underline{y}) \cong \mu$. This occurs when the prior knowledge is very accurate relative to the observations (i.e., v^2 is small relative to $\frac{1}{d^2}$), so the estimator ignores the observation and chooses the mean of the prior distribution as its estimate. In this case, the MMSE is approximately v^2, the prior variance.

If $v^2 d^2$ is large, then, $\hat{\theta}_{MMSE}(\underline{y}) \cong \hat{\theta}_1(\underline{y})$, an estimate that depends only on the observations and does not incorporate the prior information at all. In this case, the MMSE is approximately $\frac{1}{d^2}$.

* When $\Sigma = \sigma^2 I$ and $\underline{s} = (1, 1, \cdots, 1)^T \triangleq \underline{1}$. Then

$$Y_k = N_k + \theta, \quad k = 1, 2, \cdots, n$$

with $N_1, \cdots, N_n$ i.i.d. $N(0, \sigma^2)$. The quantity $v^2 d^2 = n v^2/\sigma^2$ and $\hat{\theta}_1(\underline{y}) = \bar{y} = \frac{1}{n}\sum_{k=1}^{n} y_k$, the sample mean.

If we have no observations, i.e., $n=0$, we simply estimate θ as its prior mean μ. As we take more observations (increase n), the sample mean $\bar{y}$ becomes more reliable and we place more weight on it. In the limit as $n \to \infty$, we disregard the prior mean entirely and adopt the sample mean as our estimate.

* Case 4.2.4: Estimation of Vector Parameters

Consider $\Lambda = \mathbb{R}^m$. A cost function $C: \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}$. We consider C of the following form:

$$C[\underline{a}, \underline{\theta}] = \sum_{i=1}^{m} C_i[a_i, \theta_i] \qquad (*)$$

where C_i is a cost function associated with the estimation of the ith component of the parameter. Then, the conditional posterior cost for an estimate $\underline{\hat{\theta}}$ is

$$E\{C[\underline{\hat{\theta}}(y), \underline{\theta}] \mid Y = y\} = \sum_{i=1}^{m} E\{C_i[\hat{\theta}_i(y), \theta_i] \mid Y = y\}$$

So that we essentially have m scalar estimation problems to solve. That is, $\hat{\theta}_i(y)$, the ith

component of $\hat{\underline{\theta}}(y)$, is chosen to minimize

$$E\{C_i[\hat{\theta}_i(y), \Theta_i] \mid Y = y\}.$$

An example of a useful cost function that decomposes as in (*) is the square of the Euclidean norm of the error

$$C[\underline{a}, \underline{\theta}] = \|\underline{a} - \underline{\theta}\|^2 = \sum_{i=1}^{m} (a_i - \theta_i)^2.$$

It follows from Case 4.2.1 that for this cost function the ith component of the Bayes estimate is $E\{\Theta_i \mid Y = y\}$, i.e., the Bayes estimate is

$$\hat{\underline{\theta}}_B(y) = E\{\underline{\Theta} \mid Y = y\},$$

the conditional mean of $\underline{\Theta}$ given $Y = y$.

Another example of a cost function satisfying (*) is the following

$$C[\underline{a}, \underline{\theta}] = \sum_{i=1}^{n} |a_i - \theta_i|.$$

From Case 4.2.2, we see that this cost function leads to the estimate whose ith component is the conditional median of Θ_i given $Y = y$.

To extend the MAP estimation to vector parameters, we might consider a cost function of the form (*), in which $C_i[a_i, \theta_i]$ is the uniform cost function. This leads to the vector estimator that

has as its ith component the conditional mode of Θ_i given $\underline{Y} = \underline{y}$. However, this decomposed cost function is not the most meaningful extension of the uniform cost function to the vector case. More meaningful one is

$$C[\underline{a}, \underline{\theta}] = \begin{cases} 1 & \text{if } \max_{1 \le i \le m} |a_i - \theta_i| > \Delta \\ 0 & \text{if } \max_{1 \le i \le m} |a_i - \theta_i| \le \Delta \end{cases}$$

for which we have

$$E\{C[\underline{\hat{\theta}}(\underline{y}), \underline{\Theta}] \mid \underline{Y} = \underline{y}\}$$

$$= 1 - P[|\hat{\Theta}_1(\underline{Y}) - \Theta_1| \le \Delta, \ldots, |\hat{\Theta}_m(\underline{Y}) - \Theta_m| \le \Delta \mid \underline{Y} = \underline{y}].$$

The above means that the approximate optimality of estimating $\underline{\Theta}$ as its conditional mode given $\underline{Y} = \underline{y}$, which may differ from the vector whose ith component is the conditional mode of Θ_i given $\underline{Y} = \underline{y}$ obtained from decomposing the cost.

A further useful cost function of interest in estimating vector parameters is a generalization of the squared-error norm:

$$C[\underline{a}, \underline{\theta}] = (\underline{a} - \underline{\theta})^T A (\underline{a} - \underline{\theta})$$

where A is a symmetric, positive-definite matrix. This cost function allows for joint weightings of errors in different parameters.

To derive the Bayes estimate, we write

$$E\{(\underline{\hat{\theta}}(y)-\underline{\theta})^T A(\underline{\hat{\theta}}(y)-\underline{\theta}) \mid Y=y\}$$

$$= [\underline{\hat{\theta}}(y)]^T A\,\underline{\hat{\theta}}(y) - 2[\underline{\hat{\theta}}(y)]^T A E\{\underline{\theta} \mid Y=y\}$$

$$+ E\{\underline{\theta}^T A\,\underline{\theta} \mid Y=y\}.$$

It is a quadratic function of $\underline{\hat{\theta}}(y)$ and thus achieves its minimum at the point at which its gradient with respect to $\underline{\hat{\theta}}(y)$ vanishes. That is

$$\nabla_{\underline{\hat{\theta}}(y)} E\{C[\underline{\hat{\theta}}(y), \underline{\theta}] \mid Y=y\}$$

$$= 2A\,\underline{\hat{\theta}}(y) - 2AE\{\underline{\theta} \mid Y=y\} = 0.$$

So, the Bayes estimate, $\underline{\hat{\theta}}_B$, satisfies

$$2A\,\underline{\hat{\theta}}_B(y) = 2A\,E\{\underline{\theta} \mid Y=y\}.$$

$$\Rightarrow \quad \underline{\hat{\theta}}_B(y) = E\{\underline{\theta} \mid Y=y\}$$

This means that the general quadratic cost criterion yeilds the conditional mean vector as a Bayes estimate regardless of choice of A. The resulting Bayes risk does depend on A:

$$r(\hat{\theta}_B) = tr\{A\,E\{Cov(\theta \mid Y)\}\},$$

where $tr\{\cdot\}$ denotes the trace operator.

* Example 3: Estimation of a Gaussian Vector from a Jointly Gaussian Observation

Consider $\Gamma = \mathbb{R}^n$, $\Lambda = \mathbb{R}^m$, and $\underline{Y}$ and $\underline{\Theta}$ are jointly Gaussian with mean vectors $\underline{\mu}_Y$ and $\underline{\mu}_\Theta$, covariance matrices Σ_Y and Σ_Θ, and cross-covariance matrix

$$\Sigma_{Y\Theta} \triangleq E\{(\underline{Y} - \underline{\mu}_Y)(\underline{\Theta} - \underline{\mu}_\Theta)^T\}, \text{ i.e.,}$$

$$\begin{pmatrix} \underline{Y} \\ \underline{\Theta} \end{pmatrix} \sim N\left(\begin{pmatrix} \underline{\mu}_Y \\ \underline{\mu}_\Theta \end{pmatrix}, \begin{pmatrix} \Sigma_Y & \Sigma_{Y\Theta} \\ \Sigma_{\Theta Y} & \Sigma_\Theta \end{pmatrix} \right)$$

with $\Sigma_{\Theta Y} = \Sigma^T_{Y\Theta}$

It is not hard to see that the conditional distribution of $\underline{\Theta}$ given $\underline{Y} = \underline{y}$ is also Gaussian, with conditional mean $\hat{\underline{\mu}}(\underline{y})$ given by

$$\hat{\underline{\mu}}(\underline{y}) = \underline{\mu}_\Theta + \Sigma_{\Theta Y} \Sigma_Y^{-1} (\underline{y} - \underline{\mu}_Y)$$

and with conditional covariance matrix $\hat{\Sigma}$:

$$\hat{\Sigma} = \Sigma_\Theta - \Sigma_{\Theta Y} \Sigma_Y^{-1} \Sigma_{Y\Theta}.$$

With these properties, we may find all the optimum estimates in this section.

The conditional mean estimate is $\hat{\underline{\mu}}(\underline{y})$ shown above. Since the multivariate Gaussian density has its mode at its mean, the MAP estimate is

$\hat{\underline{\mu}}(\underline{y})$ as well. Moreover, since $\underline{\Theta}$ being Gaussian given $\underline{Y}=\underline{y}$ implies that Θ_i is marginally Gaussian given $\underline{Y}=\underline{y}$, the marginal mode and median of Θ_i given $\underline{Y}=\underline{y}$ occur at $\hat{\mu}_i(\underline{y})$, the ith component of $\hat{\underline{\mu}}(\underline{y})$. Thus, $\hat{\underline{\mu}}(\underline{y})$ provides the optimum estimate in all the senses discussed before. Note that this estimate is linear in $\underline{y}$.

Since $\mathrm{Cov}(\underline{\Theta}|\underline{Y})=\hat{\Sigma}$ that is independent of $\underline{Y}$. Thus, $E\{\mathrm{Cov}(\underline{\Theta}|\underline{Y})\}=\hat{\Sigma}$ and the minimum Bayes risk is

$$r(\hat{\underline{\theta}}_B)=\mathrm{tr}\{A\hat{\Sigma}\}=\mathrm{tr}(A\Sigma_\Theta)-\mathrm{tr}(A\Sigma_{\Theta Y}\Sigma_Y^{-1}\Sigma_{Y\Theta})$$

Note also $\hat{\Sigma}=E\{(\underline{\Theta}-\hat{\underline{\theta}}_B(\underline{Y}))(\underline{\Theta}-\hat{\underline{\theta}}_B(\underline{Y}))^T\}$.

* Linear observation model:

$$\underline{Y}=H\underline{\Theta}+\underline{N},$$

where $\underline{\Theta}\sim N(\underline{\mu}_\Theta,\Sigma_\Theta)$, $\underline{N}\sim N(\underline{0},\Sigma)$, H is a fixed $n\times m$ matrix, and $\underline{\Theta}$ and $\underline{N}$ are independent.

If we think of $\Theta_1,\cdots,\Theta_m$ as being samples of a stochastic signal, then

$$Y_k=\sum_{j=1}^{m}h_{k,j}\Theta_j+N_k,\quad k=1,\cdots,n,$$

is an observation sequence consisting of linearly filtered signal plus additive noise. This may occur when a signal is observed through a channel with finite bandwidth or other linearly distorting characteristic. In this case, the estimation of $\underline{\Theta}$ is known as the problem of equalizing the channel. A further application of this model will be discussed in the next chapter in Kalman–Bucy filtering.

In this model, it is easy to show that $\underline{Y}$ and $\underline{\Theta}$ are jointly Gaussian, with $\underline{\mu}_\Theta$ and Σ_Θ given,

$$\underline{\mu}_Y = H\underline{\mu}_\Theta, \quad \Sigma_Y = H\Sigma_\Theta H^T + \Sigma \text{ and}$$

$$\Sigma_{\Theta Y} = \Sigma_\Theta H^T.$$

Thus, we get the Bayes estimate

$$\hat{\underline{\mu}}(\underline{y}) = \underline{\mu}_\Theta + \Sigma_\Theta H^T (H\Sigma_\Theta H^T + \Sigma)^{-1}(\underline{y} - H\underline{\mu}_\Theta)$$

and the error covariance matrix

$$\hat{\Sigma} = \Sigma_\Theta - \Sigma_\Theta H^T (H\Sigma_\Theta H^T + \Sigma)^{-1} H\Sigma_\Theta.$$

In the above computations, we note that it involves the inversion of an $n \times n$ matrix, whose complexity is of the order of n^3 in general. This complexity can sometimes be reduced by making use of the following simple matrix identity:

$$\Sigma_\Theta H^T (H\Sigma_\Theta H^T + \Sigma)^{-1}$$
$$= (H^T \Sigma^{-1} H + \Sigma_\Theta^{-1})^{-1} H^T \Sigma^{-1},$$

If Σ^{-1} is known, e.g. if $\Sigma = \sigma^2 I$, and $m < n$, the matrix on the right-hand side is easier to compute than that on the left.

When $m=1$, $H = \underline{s}$, $\underline{\mu}_\Theta = \mu$, and $\Sigma_\Theta = v^2$, putting these into the above equation, we have

$$\hat{\underline{\mu}}(\underline{y}) = \mu + \left(\underline{s}^T \Sigma^{-1} \underline{s} + \frac{1}{v^2}\right)^{-1} \underline{s}^T \Sigma^{-1} (\underline{y} - \underline{s}\mu)$$
$$= \frac{v^2 d^2 \hat{\theta}_1(\underline{y}) + \mu}{v^2 d^2 + 1}$$

and $r(\hat{\underline{\mu}}) = \hat{\Sigma} = v^2 - (\underline{s}^T \Sigma^{-1} \underline{s} + v^{-2})^{-1} \underline{s}^T \Sigma^{-1} \underline{s}\, v^2$

$$= \frac{v^2}{v^2 d^2 + 1}$$

which is back to Example 2.

2. Suppose we have a real observation Y given by

$$Y = N + \Theta S$$

where $N \sim \mathcal{N}(0,1)$, $P(S=1) = P(S=-1) = 1/2$, and Θ has pdf

$$w(\theta) = \begin{cases} Ke^{\theta^2/2}, & 0 \le \theta \le 1 \\ 0, & \text{otherwise} \end{cases}$$

where $K = [\int_0^1 e^{\theta^2/2} d\theta]^{-1}$. Assume that N, Θ, and S are independent.

(a) Find the MMSE estimate of Θ given $Y = y$.

(b) Find the MAP estimate of Θ given $Y = y$.

4. Suppose we have a single observation y of a random variable Y given by

$$Y = N + \Theta$$

where N is a Gaussian random variable with mean zero and variance σ^2. The parameter Θ is a random variable, independent of N, with probability mass function

$$w(\theta) = P(\Theta = \theta) = \begin{cases} \frac{1}{2}, & \theta = -1 \\ \frac{1}{2}, & \theta = +1. \end{cases}$$

(a) Find $\hat{\theta}_{MMSE}$ and $\hat{\theta}_{MAP}$. (You may consider the parameter set Λ to be $\mathbb{R}$.)

(b) Under what conditions are the two estimates in (a) approximately equal?

5. Suppose Θ is a random parameter with prior density

$$w(\theta) = \begin{cases} e^{-\theta}, & \theta \geq 0 \\ 0, & \theta < 0, \end{cases}$$

and that Y has conditional density

$$p_\theta(y) = \tfrac{1}{2} e^{-|y-\theta|}, \quad -\infty < y < \infty.$$

Find $\hat{\theta}_{MMSE}$ and $\hat{\theta}_{MAP}$.

7. Suppose Θ is uniformly distributed on the interval $(0, 1)$ and that we observe $Y = N + \Theta$ where N is a random variable, independent of Θ, with density

$$p_N(n) = \begin{cases} e^{-n}, & n \geq 0 \\ 0, & n < 0. \end{cases}$$

Find $\hat{\theta}_{MMSE}, \hat{\theta}_{ABS}$, and $\hat{\theta}_{MAP}$.

8. (a) Consider the observation model of Exercise 7 but with the prior of Exercises 5 (i.e., N and Θ both have the unit exponential distribution). Find the MMSE and MMAE estimates of Θ based on Y.

(b) Find the minimum mean-squared error for (a).

(c) Consider now the observation model

$$Y_k = N_k + \Theta, \quad k = 1, \ldots, n,$$

where $N_1, N_2, \ldots, N_n$, and Θ are i.i.d random variables with the unit exponential distribution. Find the MAP estimate of Θ based on $Y_1, Y_2, \ldots, Y_n$.