Chapter 3

Methods of Parametric Statistics for the Pattern Recognition Problem

§1 The Pattern Recognition Problem

It is required to minimize the functional

$$I(\alpha) = \int (y - F(x, \alpha))^2 P(x, y)\, dx\, dy \tag{3.1}$$

under the conditions when the density $P(x, y)$ is unknown but the sample

$$x_1, y_1; \ldots; x_l, y_l \tag{3.2}$$

is given, based on random independent trials according to $P(x, y)$.

We shall solve this problem applying the following scheme:

(1) Estimate the density from the sample (3.2). Denote the estimated density by $\hat{P}(x, y)$.

(2) Construct the functional

$$I_{\text{emp}}(\alpha) = \int (y - F(x, \alpha))^2 \hat{P}(x, y)\, dx\, dy \tag{3.3}$$

using the estimated density.

(3) Obtain the minimum of this functional, and declare the function $F(x, \alpha_{\text{emp}})$ which yields the minimum of (3.3) to be the solution of the original minimization problem (3.1).

As was pointed out in Chapter 2, this scheme can be successfully carried out only if substantial prior information concerning the density $P(x, y)$ is available (namely, when the density is completely specified up to its parameters). In other words, success can be achieved if the model of the

V. Vapnik, *Estimation of Dependences Based on Empirical Data*

estimated density is known. The model of the required density turns out to be quite different for different problems of estimating dependences.

In this chapter we shall consider the pattern recognition problem. A characteristic feature of this problem is that the unknown probability density† $P(x, \omega)$ can be represented as a union of two densities $P(x|\omega = 0)$ and $P(x|\omega = 1)$ defined on different subspaces $X, 0$ and $X, 1$:

$$P(x, \omega) = P(x|\omega = 0)P(\omega = 0)(1 - \omega) + P(x|\omega = 1)P(\omega = 1)\omega. \tag{3.4}$$

The set of pairs x, ω consists of two nonoverlapping subspaces of dimensionality n, namely

$$X \subset E_n, \quad \omega = 0 \quad \text{and} \quad X \subset E_n, \quad \omega = 1.$$

The formula (3.4) asserts that on the first subspace the density is equal to $P(x|\omega = 0)P(\omega = 0)$, and on the second $P(x|\omega = 1)P(\omega = 1)$. In formula (34) $P(x|\omega = 0)$ and $P(x|\omega = 1)$ are the components of the union; $P(\omega = 0)$ and $P(\omega = 1) = 1 - P(\omega = 0)$ are the proportions.

Let the density $P(x, \omega)$ be known up to a finite number $m_1 + m_2 + 1$ of parameters

$$P(x, \omega) = P_\beta(x|\omega = 0)P(\omega = 0)(1 - \omega) + P_\gamma(x|\omega = 1)P(\omega = 1)\omega, \tag{3.5}$$

where β is an unknown m_1-dimensional vector of parameters of density $P_\beta(x|\omega = 0)$, γ is an unknown m_2-dimensional vector of parameters of the density $P_\gamma(x|\omega = 1)$, and $P(\omega = 0)$ is a scalar parameter.

Now in order to implement our scheme it is necessary to be able to solve two problems:

(1) to find the minimum of functional (3.3) for a given density $P(x, \omega)$;
(2) based on the sample (3.2), to estimate the density of $P(x, \omega)$.

The first problem is referred to in statistics as the *problem of discriminant analysis*; the second is called the *problem of estimating the density in a parametric class of functions*. We now consider these two problems.

§2 Discriminant Analysis

It is required to obtain the minimum of the functional (3.3) for a given density (given components of union $P(x|\omega = 0)$, $P(x|\omega = 1)$ and proportions $P(\omega = 0)$, $P(\omega = 1) = 1 - P(\omega = 0)$).

First consider the simple case: the class of possible decision rules $F(x, \alpha)$ is in no way restricted. In this situation it is easy to construct a minimizing

† We use the letter ω instead of y to emphasize that it takes only the two values 0 and 1.

rule which minimizes the functional (3.3). Indeed, according to Bayes's formula the probability that the vector x belongs to the first (second) class is determined by

$$P(\omega = 0|x) = \frac{P(x|\omega = 0)P(\omega = 0)}{P(x|\omega = 0)P(\omega = 0) + P(x|\omega = 1)P(\omega = 1)}$$
$$\left(P(\omega = 1|x) = \frac{P(x|\omega = 1)(1 - P(\omega = 0))}{P(x|\omega = 0)P(\omega = 0) + P(x|\omega = 1)P(\omega = 1)}\right). \tag{3.6}$$

Minimal loss (the minimum probability of error) can be obtained for the classification in which the vector x is assigned to the first class if its affiliation to the first class is more probable than to the second, i.e., if

$$P(\omega = 0|x) > P(\omega = 1|x).$$

Otherwise the vector x is assigned to the second class. In other words, taking (3.6) into account, the vector x should be assigned to the first class provided the inequality

$$\frac{P(x|\omega = 1)}{P(x|\omega = 0)} < \frac{P(\omega = 0)}{1 - P(\omega = 0)},$$

is fulfilled, or equivalently, the optimal classification of vectors is carried out by means of the indicator function

$$F(x) = \theta\left[\ln P(x|\omega = 1) - \ln P(x|\omega = 0) + \ln \frac{1 - P(\omega = 0)}{P(\omega = 0)}\right], \tag{3.7}$$

where

$$\theta(z) = \begin{cases} 1 & \text{for } z \geq 0, \\ 0 & \text{for } z < 0. \end{cases}$$

Therefore the knowledge of the probability density (composition and proportion of the union (3.5)) allows us to construct an optimal decision rule immediately.

However, the problem of finding an optimal decision rule becomes substantially more complex if the class of admissible decision rules $F(x, \alpha)$ is restricted. In particular, the problem of finding an optimal linear decision rule of the form

$$F(x, \alpha) = \theta[\alpha^T x + \alpha_0] \tag{3.8}$$

is a difficult one. The vector $\alpha = (\alpha_1, \ldots, \alpha_n)^T$ determines the direction of a linear discriminant function, and the parameter α_0 its threshold value. The problem of finding the minimum of (3.3) in the class (3.8) is called the problem of *linear discriminant analysis.*

In the thirties R. A. Fisher proposed as the direction of the linear discriminant function a direction along which the maximum of the relative distance between the mathematical expectations of projections of vectors of different classes is obtained, i.e., the direction α along which the maximum of

$$T(\alpha) = \frac{(m_1(\alpha) - m_2(\alpha))^2}{\sigma_1^2(\alpha) + \sigma_2^2(\alpha)}, \tag{3.9}$$

where

$$m_1(\alpha) = \int \alpha^T x P_\beta(x|\omega = 0)\, dx,$$

$$m_2(\alpha) = \int \alpha^T x P_\gamma(x|\omega = 1)\, dx,$$

$$\sigma_1^2(\alpha) = \int (\alpha^T x - m_1(\alpha))^2 P_\beta(x|\omega = 0)\, dx,$$

$$\sigma_2^2(\alpha) = \int (\alpha^T x - m_2(\alpha))^2 P_\gamma(x|\omega = 1)\, dx,$$

$$\alpha^T \alpha = 1$$

is attained.

The determination of the maximum of (3.9) for arbitrary densities is a very difficult problem. Therefore basic investigations in the area of linear discriminant analysis were directed first toward verifying for specific types of densities that Fisher's linear discriminant function indeed determines a solution of linear discriminant analysis, and secondly toward finding algorithms for computing the discriminant function. The basic result was that for the union of two normal laws

$$P(x|\omega = 0) = N(\mu_1, \Delta_1), \qquad P(x|\omega = 1) = N(\mu_2, \Delta_2)$$

(μ_1 is the mean vector, Δ_1 is the covariance matrix for the first multivariate normal distribution, and μ_2, Δ_2 are the analogous parameters for the second distribution), taken in proportions $P(\omega = 0)$ and $1 - P(\omega = 0)$, the optimal linear discriminant function is given by the direction vector

$$\alpha_{t^*} = (\mu_1 - \mu_2)^T (t^* \Delta_1 + (1 - t^*)\Delta_2)^{-1}, \tag{3.10}$$

where $0 \leq t^* \leq 1$. The value t^* is determined as the root of the so-called *resolvent function*

$$f(t) = t\sigma_1^2(\alpha_t) + (1 - t)\sigma_2^2(\alpha_t) - \ln\left(\frac{P(\omega = 0)}{1 - P(\omega = 0)} \cdot \frac{\sigma_2^2(\alpha_t)}{\sigma_1^2(\alpha_t)}\right). \tag{3.11}$$

For $P(\omega = 0) = \frac{1}{2}$ the direction (3.10) of the linear discriminant function maximizes the functional

$$I(\alpha) = \frac{(m_1(\alpha) - m_2(\alpha))^2}{t^*\sigma_1^2(\alpha) + (1 - t^*)\sigma_2^2(\alpha)}.$$

The calculation of the roots of the resolvent equation (3.11) is quite a difficult task. Therefore in practice when constructing a linear discriminant function it is assumed that $t^* = \frac{1}{2}$, and Fisher's linear discriminant is taken to be the solution of the problem. (More details are given in [71].)

Thus problems arising in discriminant analysis are due to the fact that the class of possible decision rules on which the minimum of functional (3.3) is to be determined is bounded. Therefore it may seem that the problem of discriminant analysis is artificial. Indeed, if it is possible to estimate probability density, what is the need for seeking a decision rule which yields a conditional minimum of the functional, when it is easy to find a decision rule (cf. (3.7)) which yields an absolute minimum for the functional (3.3)?

The fact of the matter is that if the density is estimated imprecisely, then the value of the guaranteed deviation of the minimum for the empirical functional from the minimum for the expected risk functional becomes larger for a function chosen from a wider class. Therefore it may happen that the smaller value of the guaranteed expected risk will be achieved, not at a function yielding the absolute minimum for the empirical functional, but rather on a function belonging to a narrower class and yielding the conditional minimum.

This result is connected with the effect of the second procedure for minimizing the expected risk (cf. Chapter 2, Section 4). The idea of narrowing the class of decision rules in order to obtain a smaller guaranteed value of the expected risk will be implemented below in Chapters 8 and 9. In the present chapter we shall consider parametric methods of estimating densities. In view of (3.7), the knowledge of the density immediately leads to the construction of a decision rule yielding the absolute minimum for (3.3).

§3 Decision Rules in Problems of Pattern Recognition

Algorithms of pattern recognition based on estimation of the density (gives components of the union (mixture) $P(x|\omega = 0)$ and $P(x|\omega = 1)$ and its proportion $P(\omega = 0)$) are traditionally associated with two classes of distributions.

3.1 First Class of Distributions

The probability distribution $P_\omega(x) = P(x|\omega)$ is such that coordinates of the vector $x = (x^1, \ldots, x^n)^T$ are statistically independent, i.e.,

$$P_\omega(x) = P_\omega(x^1) \cdots P_\omega(x^n), \qquad \omega = 0, 1, \tag{3.12}$$

and moreover each coordinate x^i of the vector x can take on only a fixed number of values. Let us assume that each coordinate x^i takes on τ_i values $c^i(1), \ldots, c^i(\tau_i)$. Thus in the case under consideration the distribution laws of random variables $P_{\omega=0}(x)$ and $P_{\omega=1}(x)$ are defined by the expression

(3.12), where $P_\omega(x^i)$ can be written as

$$P_\omega(x^i) = \begin{cases} p^i_\omega(1) & \text{for } x^i = c^i(1), \\ \vdots & \\ p^i_\omega(\tau_i) & \text{for } x^i = c^i(\tau_i), \end{cases} \tag{3.13}$$

$$\sum_{j=1}^{\tau_i} p^i_\omega(j) = 1.$$

Here $p^i_\omega(j)$ is the probability that for a vector belonging to the class $\omega = \{0, 1\}$ the value of the x^ith coordinate equals $c^i(j)$. To estimate the probability distribution for such a union means to find values of

$$r = 2\sum_{i=1}^{n} \tau_i + 1$$

parameters ($\sum_{i=1}^{n} \tau_i$ parameters for estimating each distribution $P_\omega(x)$, and one parameter—the proportion of the union).

According to (3.7) an optimal decision rule for the mixture formed by the two distributions (3.12) will be the following linear discriminant function:

$$F(x) = \theta\left(\sum_{i=1}^{n} \ln \frac{P_{\omega=1}(x^i)}{P_{\omega=0}(x^i)} - \ln \frac{p}{1-p}\right),$$

where p, $1 - p$ are proportions of the union.

3.2 Second Class of Distributions

Here in each class $\omega = \{0, 1\}$ vectors x are distributed according to the multivariate normal distribution

$$P_\omega(x) = \frac{1}{(2\pi)^{n/2}|\Delta_\omega|^{1/2}} \exp\{-\tfrac{1}{2}(x - \mu_\omega)^{\mathrm{T}} \Delta_\omega^{-1}(x - \mu_\omega)\},$$

where μ_ω is the vector of mean values and Δ_ω is the covariance matrix.

It follows from (3.7) that the optimal decision rule in this case becomes the quadratic discriminant function

$$F(x) = \theta[\tfrac{1}{2}(x - \mu_0)^{\mathrm{T}} \Delta_0^{-1}(x - \mu_0) - \tfrac{1}{2}(x - \mu_1)^{\mathrm{T}} \Delta_1^{-1}(x - \mu_1) + \ln \frac{|\Delta_0|}{|\Delta_1|} - \ln \frac{p}{1-p}], \tag{3.14}$$

where $\mu_0, \Delta_0; \mu_1, \Delta_1$ are parameters of the normal distributions forming the union (3.5) and p, $1 - p$ are the corresponding proportions. In the particular case when $\Delta_0 = \Delta_1 = \Delta$ the quadratic discriminant function (3.14) reduces to a linear one:

$$F(x) = \theta\left[(\mu_1 - \mu_0)^{\mathrm{T}} \Delta^{-1} x + \tfrac{1}{2}(\mu_0^{\mathrm{T}} \Delta^{-1} \mu_0 - \mu_1^{\mathrm{T}} \Delta^{-1} \mu_1) - \ln \frac{p}{1-p}\right].$$

§4 Evaluation of Qualities of Algorithms for Density Estimation

Thus the construction of a discriminant function based on empirical data reduces to an estimation of the probability distributions $P(x|\omega = 0)$ and $P(x|\omega = 1)$ and of the parameter p. The parameter p determines the fraction of pairs x, ω with $\omega = 0$ and may be estimated by the quantity $\tilde{p} = m/l$, where m is the number of pairs in the sample with $\omega = 0$ and l is the sample size.†

What are the algorithms that one should utilize for estimating the probability densities $P(x|\omega = 0)$, $P(x|\omega = 1)$? To answer this question one should first agree on the method of assessing qualities of estimating algorithms on the basis of samples of fixed size.

The quality of specific algorithm A which estimates the density $P(x, \alpha_0)$ from a sample $x_1, \ldots, x_l$ is naturally defined as the distance between the density and the estimated function $P_A(x|x_1, \ldots, x_l)$, i.e., by the quantity

$$\rho(P(x, \alpha_0), P_A(x|x_1, \ldots, x_i)) = \rho_{\alpha_0, A}(x_1, \ldots, x_l).$$

We shall define the closeness of densities in terms of the L^2 metric, i.e.,

$$\rho_{\alpha_0, A}(x_1, \ldots, x_l) = \left(\int (P(x, \alpha_0) - P_A(x|x_1, \ldots, x_l))^2 \, dx\right)^{1/2}. \quad (3.15)$$

Since the choice of the density $P_A(x|x_1, \ldots, x_l)$ depends on the sample $x_1, \ldots, x_l$, the quantity $\rho_{\alpha_0, A}(x_1, \ldots, x_l)$ is a random variable. We shall characterize the quality of the algorithm A by the mathematical expectation of $\rho^2_{\alpha_0, A}(x_1, \ldots, x_l)$:

$$R(\alpha_0, A) = \int \rho^2_{\alpha_0, A}(x_1, \ldots, x_l) P(x_1) \cdots P(x_l) \, dx_1 \cdots dx_l.$$

The smaller $R(\alpha_0, A)$ is, the better the algorithm is for estimating the density $P(x, \alpha_0)$ from a sample of size l.

Thus we have determined how the quality of an algorithm A designed for estimating a specific density $P(x, \alpha_0)$ should be measured. It is now necessary to agree on how to measure the quality of an algorithm earmarked for estimating an arbitrary density belonging to a given class $P(x, \alpha)$ (in our case the class of densities is defined up to values of a vector of parameters α). Two principles are used in statistical decision theory in such a situation: Bayes's principle and the minimax principle.

Bayes's principle asserts that the quality of an algorithm should be estimated as the mean quality over all the estimated densities. In order to estimate the mean value of an algorithm it is necessary to know how often any particular density belonging to $P(x, \alpha)$ will be estimated, i.e., in our case

† It will be shown in Section 6 that $\tilde{\tilde{p}} = (m + 1)/(l + 2)$ is a more precise estimator.

it is necessary to have information about the probability density $P(\alpha)$ of the vector of parameters α. In that case the quality of an algorithm is defined as

$$R_{\text{B}}(A) = \int R(\alpha, A)P(\alpha)\, d\alpha. \tag{3.16}$$

The smaller $R_{\text{B}}(A)$ is, the better the algorithm.

The minimax principle asserts that one must estimate the quality of an algorithm on the basis of the most unfavorable probability density $P(x, \alpha^*)$ for this algorithm. Here the densities which may be encountered in practice are not taken into account. It may therefore turn out that the quality of the algorithm is determined by a case which will never occur. The quality of an algorithm according to the minimax principle is defined as

$$R_{\text{mnx}}(A) = \sup_{\alpha} R(\alpha, A). \tag{3.17}$$

The smaller the value of $R_{\text{mnx}}(A)$, the better the algorithm.

§5 The Bayesian Algorithm for Density Estimation

We shall determine the structure of algorithms which assure the solution of the Bayesian estimation of density, i.e., which minimize the functional

$$R_{\text{B}}(A) = \int R(\alpha, A)P(\alpha)\, d\alpha.$$

From a sample $x_1, \ldots, x_l$, let a density which belongs to the class $P(x, \alpha)$ be estimated and the prior probability density $P(\alpha)$ be given. Utilizing Bayes's formula, we obtain

$$P(\alpha | x_1, \ldots, x_l) = \frac{P(x_1, \ldots, x_l | \alpha)P(\alpha)}{P(x_1, \ldots, x_l)},$$

which is the density of posterior probabilities $P(\alpha | x_1, \ldots, x_l)$ which characterizes the possibilities of realizations of various values of parameters α after the information about the sample $x_1, \ldots, x_l$ has been added to the prior information $P(\alpha)$. Here $P(x_1, \ldots, x_l | \alpha)$ is the conditional and $P(x_1, \ldots, x_l)$ is the unconditional density of occurrence of the sample $x_1, \ldots, x_l$:

$$P(x_1, \ldots, x_l) = \int P(x_1, \ldots, x_l | \alpha)P(\alpha)\, d\alpha.$$

Below we shall show that the posterior mean, i.e., the function

$$P_{\text{B}}(x | x_1, \ldots, x_l) = \int P(x, \alpha)P(\alpha | x_1, \ldots, x_l)\, d\alpha \tag{3.18}$$

is the solution of the Bayesian problem.

In general the density $P_B(x|x_1, \ldots, x_l)$ obtained as a result of averaging functions $P(x, \alpha)$ with respect to the measure $P(\alpha|x_1, \ldots, x_l)$ need not belong to the parametric family $P(x, \alpha)$ under consideration. Therefore, strictly speaking, the method for constructing the posterior mean (3.18) cannot actually be called the estimation of a function belonging to the class $P(x, \alpha)$.

Thus we obtain a function $\pi(x; x_1, \ldots, x_l)$ which minimizes the functional

$$R_B(\pi) = \int (P(x|\alpha) - \pi(x; x_1, \ldots, x_l))^2 \times P(x_1, \ldots, x_l|\alpha)P(\alpha)\, d\alpha\, dx\, dx_1 \cdots dx_l. \tag{3.19}$$

Denote

$$r(x; x_1, \ldots, x_l) = \int (P(x|\alpha) - \pi(x; x_1, \ldots, x_l))^2 P(x_1, \ldots, x_l|\alpha)P(\alpha)\, d\alpha.$$

Interchanging the order of integration in (3.19), we arrive at

$$R_B(\pi) = \int r(x; x_1, \ldots, x_l)\, dx\, dx_1 \cdots dx_l. \tag{3.20}$$

We now transform the function $r(x; x_1, \ldots, x_l)$:

$$\begin{aligned} r(x; x_1, \ldots, x_l) = &\int P^2(x|\alpha)P(x_1, \ldots, x_l|\alpha)P(\alpha)\, d\alpha \\ &- 2\pi(x; x_1, \ldots, x_l) \int P(x|\alpha)P(x_1, \ldots, x_l|\alpha)P(\alpha)\, d\alpha \\ &+ \pi^2(x; x_1, \ldots, x_l) \int P(x_1, \ldots, x_l|\alpha)P(\alpha)\, d\alpha. \end{aligned} \tag{3.21}$$

Denote

$$\hat{P}(x|x_1, \ldots, x_l) = \frac{\int P(x|\alpha)P(x_1, \ldots, x_l|\alpha)P(\alpha)\, d\alpha}{P(x_1, \ldots, x_l)},$$

where

$$P(x_1, \ldots, x_l) = \int P(x_1, \ldots, x_l|\alpha)P(\alpha)\, d\alpha,$$

and rewrite the equality (3.21) in the form

$$\begin{aligned} r(x; x_1, \ldots, x_l) = &\int P^2(x|\alpha)P(x_1, \ldots, x_l|\alpha)P(\alpha)\, d\alpha \\ &- \hat{P}^2(x|x_1, \ldots, x_l)P(x_1, \ldots, x_l) \\ &+ [\hat{P}(x|x_1, \ldots, x_l) - \pi(x; x_1, \ldots, x_l)]^2 P(x_1, \ldots, x_l). \end{aligned}$$

Substitute the expression for $r(x; x_1, \ldots, x_l)$ into (3.20). This results in a functional which can be expressed as the sum of two summands

$$R_B(\pi) = R_1 + R_2(\pi),$$

where

$$R_1 = \int [P^2(x|\alpha)P(x_1, \ldots, x_l|\alpha)P(\alpha)\, d\alpha - P(x_1, \ldots, x_l)\hat{P}^2(x|x_1, \ldots, x_l)]\, dx\, dx_1 \cdots dx_l,$$

$$R_2(\pi) = \int [\hat{P}(x|x_1, \ldots, x_l) - \pi(x; x_1, \ldots, x_l)]^2\, dx\, dx_1 \cdots dx_l.$$

The first summand does not depend on $\pi(x; x_1, \ldots, x_l)$. Therefore minimization of $R_B(\pi)$ is equivalent to the minimization of the second summand $R_2(\pi)$. The minimum of this summand is zero and is attained if

$$\pi(x; x_1, \ldots, x_l) = \hat{P}(x|x_1, \ldots, x_l) \equiv P_B(x|x_1, \ldots, x_l).$$

In succeeding sections, for prior distributions $P(\alpha)$ Bayesian approximations of densities will be obtained. The construction of a Bayesian approximation for a fixed prior distribution $P(\alpha)$ depends on whether the expression (3.18) can be integrated analytically.

§6 Bayesian Estimators of Discrete Probability Distributions

In Section 3 the probability distribution function of the discrete independent features (3.12) and (3.13) was introduced. Here we shall show that, under minimal prior information concerning the values of the parameters $p^i(j)$, namely: for each i the parameters $p^i(1), \ldots, p^i(\tau_i)$ are uniformly distributed on the simplex

$$C_i = \left\{ p : \sum_{j=1}^{\tau_i} p^i(j) = 1, p^i(j) \geq 0 \right\}.$$

The Bayesian estimator of the probability distribution of discrete independent features equals

$$P_B(x) = \prod_{i=1}^{n} P_B(x^i),$$

where

$$P_B(x^i) = \begin{cases} \hat{p}^i(1) = \dfrac{m_i(1) + 1}{l + \tau_i}, \\ \vdots \\ \hat{p}^i(\tau_i) = \dfrac{m_i(\tau_i) + 1}{l + \tau_i}. \end{cases}$$

$m_i(j)$ is the number of vectors in the sample such that the rth coordinate takes the jth value, τ_i is the number of values taken by the ith coordinate, and l is the sample size.

We now obtain the Bayesian estimator of the probability distribution of the discrete independent features. For this purpose we compute the function

$$P_{\mathrm{B}}(x^i) = \frac{\int P(x^i|p)P(x_1^i, \ldots, x_l^i|p)P(p)\,dp}{\int P(x_1^i, \ldots, x_l^i|p)P(p)\,dp}. \tag{3.22}$$

In our case

$$P(x^i|p) = \begin{cases} p^i(1) & \text{for } x^i = c^i(1), \\ \vdots & \\ p^{(i)}(\tau_i) & \text{for } x^i = c^i(\tau_i). \end{cases}$$

First compute the denominator of (3.22). Since the sample is random and independent, we obtain

$$\int P(x_1^i, \ldots, x_l^i|p)P(p)\,dp = \frac{1}{v}\int_{C_i} \prod_{j=1}^{\tau_i} [p^i(j)]^{m_i(j)}\,dp^i(1)\cdots dp^i(\tau_i), \tag{3.23}$$

where v is the volume of the simplex C_i. It is known (see, e.g. [52]) that the definite integral (3.23) may be computed analytically:

$$P(x_1^i, \ldots, x_l^i) = \frac{1}{v}\frac{\Gamma(m_i(1)+1)\cdots\Gamma(m_i(\tau_i)+1)}{\Gamma(m_i(1)+\cdots+m_i(\tau_i)+\tau_i)}, \tag{3.24}$$

where $\Gamma(n)$ is the gamma function. For integer n this function is given by

$$\Gamma(n) = (n-1)!.$$

We now derive the numerator of the expression (3.22) for the case $x^i = c^i(k)$:

$$\begin{aligned} I_k^i &= \int_{C_i} P(x^i = c^i(k)|p)P(x_1^i, \ldots, x_l^i|p)P(p)\,dp \\ &= \frac{1}{v}\int_{C_i} p^i(k)\prod_{j=1}^{\tau_i}[p^i(j)]^{m_i(j)}\,dp^i(1)\cdots dp^i(\tau_i). \end{aligned}$$

The definite integral I_k^i is equal to (cf. [52])

$$I_k^i = \frac{1}{v}\frac{\Gamma(m_i(1)+1)\cdots\Gamma(m_i(\tau_i)+1)\Gamma(m_i(k)+2)}{\Gamma(m_i(1)+\cdots+m_i(\tau_i)+\tau_i+1)\Gamma(m_i(k)+1)}. \tag{3.25}$$

Dividing (3.25) by (3.24), we obtain

$$P_{\mathrm{B}}(x^i = c^i(k)) = \frac{\Gamma(m_i(k)+2)\Gamma(l+\tau_i)}{\Gamma(m_i(k)+1)\Gamma(l+\tau_i+1)} = \frac{m_i(k)+1}{l+\tau_i}.$$

Thus

$$P_{\mathrm{B}}(x^i) = \begin{cases} \hat{p}^i(1) = \dfrac{m_i(1)+1}{l+\tau_i} & \text{for } x^i = c^i(1), \\ \vdots & \\ \hat{p}^i(\tau_i) = \dfrac{m_i(\tau_i)+1}{l+\tau_i} & \text{for } x^i = c^i(\tau_i). \end{cases}$$

Since the features are independent, we have $P_{\mathrm{B}}(x) = \prod_{i=1}^{n} P_{\mathrm{B}}(x^i)$.

§7 Bayesian Estimators for the Gaussian (Normal) Density

We shall now obtain Bayesian estimators for the Gaussian (normal) density in some special cases of the prior distribution on the parameters. First we shall obtain Beyesian estimator for the univariate normal distribution $N(\mu, \sigma^2)$ under the assumption that the parameters μ and σ of this distribution are distributed uniformly on the rectangular region $0 \leq \sigma \leq \Pi$, $-T \leq \mu \leq T$. It turns out that if Π and T are sufficiently large, then the Bayesian estimators are equal to

$$P_{\text{B}}(x) = \frac{E(l)}{\sigma_{\text{emp}}}\left[1 + \frac{(x - x_{\text{emp}})^2}{(l+1)\sigma_{\text{emp}}^2}\right]^{-(l-1)/2}, \tag{3.26}$$

where

$$E(l) = \frac{\Gamma\left(\frac{l-1}{2}\right)}{\sqrt{(l+1)\pi}\,\Gamma\left(\frac{l}{2} - 1\right)},$$

$$x_{\text{emp}} = \frac{1}{l}\sum_{i=1}^{l} x_i, \qquad \sigma_{\text{emp}}^2 = \frac{1}{l}\sum_{i=1}^{l}(x_i - x_{\text{emp}})^2.$$

(See the derivation below.)

Next we shall obtain the Bayesian estimators for the n-dimensional normal distribution for a special prior distribution on parameters μ and Δ (μ is an n-dimensional vector of the means and Δ is an $n \times n$ covariance matrix). It turns out that in this case the Bayesian approximation equals

$$P_{\text{B}}(x) = \frac{\bar{E}(l)}{|S|^{1/2}}\left[1 + \frac{(x - x_{\text{emp}})^{\text{T}} S^{-1}(x - x_{\text{emp}})}{l+1}\right]^{-(l+n)/2}, \tag{3.27}$$

where

$$\bar{E}(l) = \frac{\Gamma\left(\frac{l+n}{2}\right)}{((l+1)\pi)^{n/2}\Gamma(l/2)},$$

the vector x_{emp} is an estimator for the vector of the means:

$$x_{\text{emp}} = \frac{1}{l}\sum_{i=1}^{l} x_i,$$

and S is the empirical covariance matrix:

$$S = \frac{1}{l}\sum_{i=1}^{l}(x_i - x_{\text{emp}})(x_i - x_{\text{emp}})^{\text{T}}.$$

Note that neither of the estimators (3.26) and (3.27) belongs to the normal class. However, it is easy to verify that in both cases

$$P_{\mathrm{B}}(x) \underset{l\to\infty}{\longrightarrow} N(\mu, \Delta).$$

as $l \to \infty$.

Yet another remark: In order to calculate explicitly the Bayesian estimators of a multidimensional normal distribution (see Section 7.2 below) it was necessary to consider a special prior distribution on the parameters which differs from the uniform one (used in the univariate case; see Section 7.1 below). However, the Bayesian estimators for the univariate case obtained from (3.27) by setting $n = 1$ is close to the one obtained assuming the uniform distribution on the parameters in the univariate case (3.26).

7.1 Bayesian Estimator for the Univariate Normal Distribution

Let the variable x be distributed according to the normal distribution

$$P(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}.$$

Moreover, let the prior distribution of parameters μ and σ be uniform in the rectangle $0 \le \sigma \le \Pi$, $-T \le \mu \le T$; since the sample $x_1, \ldots, x_l$ is random and independent, we have

$$P(x_1, \ldots, x_l; \mu, \sigma) = \frac{1}{(2\pi)^{l/2}\sigma^l} \exp\left\{-\frac{\sum_{i=1}^{l}(x_i-\mu)^2}{2\sigma^2}\right\}.$$

In view of (3.18) the Bayesian estimator of the probability density is equal to

$$P_{\mathrm{B}}(x) = \left(\frac{1}{2T\Pi}\frac{1}{(2\pi)^{(l+1)/2}}\int_{-T}^{T}\int_{0}^{\Pi}\frac{1}{\sigma^{l+1}}\exp\left\{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{l}(x_i-\mu)^2+(x-\mu)^2\right)\right\}d\mu\, d\sigma\right)$$

$$\times\left(\frac{1}{2T\Pi}\frac{1}{(2\pi)^{l/2}}\int_{-T}^{T}\int_{0}^{\Pi}\frac{1}{\sigma^{l}}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{l}(x_i-\mu)^2\right\}d\mu\, d\sigma\right)^{-1}. \tag{3.28}$$

We shall assume that the intervals $[-T, T]$ and $[0, \Pi]$ are so large that the limits of integration in (3.28) may be extended to $(-\infty, \infty)$ and $(0, \infty)$ respectively. This can evidently be done if $l \ge 2$. In this case the integrals in (3.28) are convergent. We compute the numerator of (3.28):

$$I(x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\int_{0}^{\infty}\frac{1}{\sigma^{l+1}}\exp\left\{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{l}(x_i-\mu)^2+(x-\mu)^2\right)\right\}d\mu\, d\sigma. \tag{3.29}$$

Denote

$$T(\mu) = \sum_{i=1}^{l}(x_i-\mu)^2 + (x-\mu)^2, \qquad y = \frac{\sqrt{T(\mu)}}{\sigma}.$$

Then the integral (3.29) becomes

$$I(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \int_{0}^{\infty} \frac{y^{l-1}}{T^{l/2}(\mu)} \exp\{-\tfrac{1}{2}y^2\}\, dy\, d\mu$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \frac{d\mu}{T^{l/2}(\mu)} \int_{0}^{\infty} y^{l-1} \exp\left\{-\frac{y^2}{2}\right\} dy.$$

Denoting

$$c(l) = \int_{0}^{\infty} y^{l-1} \exp\left\{-\frac{y^2}{2}\right\} dy,$$

where $c(l)$ depends on neither μ nor on σ, this integral can be rewritten as

$$I(x) = \frac{c(l)}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \frac{d\mu}{T^{l/2}(\mu)}.$$

We now transform the expression for $T(\mu)$. For this purpose we note that

$$\sum_{i=1}^{l} (x_i - \mu)^2 = l\sigma_{\text{emp}}^2 + l(\mu - x_{\text{emp}})^2,$$

where

$$x_{\text{emp}} = \frac{1}{l} \sum_{i=1}^{l} x_i, \qquad \sigma_{\text{emp}}^2 = \frac{1}{l} \sum_{i=1}^{l} (x_i - x_{\text{emp}})^2.$$

The expression for $T(\mu)$ is transformed analogously to yield

$$T(\mu) = l\sigma_{\text{emp}}^2 + l(\mu - x_{\text{emp}})^2 + (x - \mu)^2.$$

Now set

$$\bar{x} = \frac{x_{\text{emp}} l + x}{l + 1}$$

and rewrite $T(\mu)$:

$$T(\mu) = l\sigma_{\text{emp}}^2 + \frac{l}{l+1}(x - x_{\text{emp}})^2 + (\bar{x} - \mu)^2(l + 1).$$

We now write the integral $I(x)$ in the form

$$I(x) = \frac{c(l)}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \frac{d\mu}{\left[l\sigma_{\text{emp}}^2 + \frac{l}{l+1}(x - x_{\text{emp}})^2 + (\bar{x} - \mu)^2(l + 1)\right]^{l/2}}$$

$$= \frac{c(l)}{\sqrt{2\pi(l+1)}} \left(l\sigma_{\text{emp}}^2 + \frac{l(x - x_{\text{emp}})^2}{(l+1)}\right)^{-(l-1)/2} \int_{-\infty}^{\infty} \frac{dz}{(1 + z^2)^{l/2}}.$$

Observe now that the integrand is independent of the parameters. We thus have

$$I(x) = c'(l, \sigma_{\text{emp}}) \left(1 + \frac{(x - x_{\text{emp}})^2}{(l+1)\sigma_{\text{emp}}^2}\right)^{-(l-1)/2} \tag{3.30}$$

To obtain a Bayesian estimator it is required only to normalize the expression (3.30):

$$P_{\mathrm{B}}(x) = \frac{I(x)}{\int_{-\infty}^{\infty} I(x)\,dx}.$$

It is known (cf. [52]) that the integral in the denominator equals the following expression:

$$\int_{-\infty}^{+\infty} I(x)\,dx = c''(l, \sigma_{\mathrm{emp}}) \int_{-\infty}^{+\infty} \frac{dx}{\left(1 + \dfrac{(x - x_{\mathrm{emp}})^2}{(l+1)\sigma_{\mathrm{emp}}^2}\right)^{(l-1)/2}}$$

$$= \frac{c''(l, \sigma_{\mathrm{emp}})\,\sigma_{\mathrm{emp}}\sqrt{l+1}\cdot\Gamma\left(\dfrac{1}{2}\right)\Gamma\left(\dfrac{l}{2} - 1\right)}{\Gamma\left(\dfrac{l-1}{2}\right)}.$$

Denote

$$E(l) = \frac{\Gamma\left(\dfrac{l-1}{2}\right)}{\sqrt{l+1}\,\Gamma\left(\dfrac{1}{2}\right)\Gamma\left(\dfrac{l}{2} - 1\right)} = \frac{\Gamma\left(\dfrac{l-1}{2}\right)}{\sqrt{\pi(l+1)}\,\Gamma\left(\dfrac{l}{2} - 1\right)}.$$

Thus

$$P_{\mathrm{B}}(x) = \frac{E(l)}{\sigma_{\mathrm{emp}}}\left(1 + \frac{(x - x_{\mathrm{emp}})^2}{(l+1)\sigma_{\mathrm{emp}}^2}\right)^{-(l-1)/2}.$$

7.2 Bayesian Estimator for the n-dimensional Normal Distribution

To obtain the Bayesian estimator for the n-dimensional normal distribution, the following two facts from the theory of multidimensional normal distributions are used:

(1) The convolution of two multidimensional normal distributions $N(0, \Delta)$ and $N(\mu, \gamma\Delta)$, where γ is a positive number, is the normal distribution $N(\mu, (1+\gamma)\Delta)$. In other words the equality

$$\int_{E_n} N(\mu - t, \gamma\Delta) \cdot N(t, \Delta)\,dt = N(\mu, (1+\gamma)\Delta)$$

is valid (see [4]).

(2) The distribution of empirical estimators S of the covariance matrix Δ given by the formula

$$S = \frac{1}{l}\sum_{i=1}^{l}(x_i - x_{\mathrm{emp}})(x_i - x_{\mathrm{emp}})^{\mathrm{T}}, \qquad x_{\mathrm{emp}} = \frac{1}{l}\sum_{i=1}^{l} x_i,$$

is expressed by the Wishart distribution (see [5]):

$$W_{l,n}(S; \Delta) = \begin{cases} C_{n,l}|\Delta|^{-(l-1)/2}|S|^{(l-n-2)/2}\exp\left\{-\dfrac{l}{2}\mathrm{Sp}[\Delta^{-1}S]\right\} & \text{for } |S| > 0, \\ 0 & \text{for } |S| \le 0, \end{cases}$$

where it is assumed that $l > n + 1$, $\mathrm{Sp}\|a_{ij}\| = \sum_{i=1}^{n} a_{ii}$. The quantity $C_{n,l}$ is a constant and equals

$$C_{n,l} = \left(\left(\frac{l}{2}\right)^{-(l-1)n/2} \pi^{n(n-1)/4} \prod_{i=1}^{n} \Gamma\left(\frac{l-i}{2}\right)\right)^{-1}. \tag{3.31}$$

Since the Wishart distribution sums to 1, we have

$$\int_{|S|>0} |S|^{(l-n-2)/2} \exp\left\{-\frac{l}{2}\mathrm{Sp}[\Delta^{-1}S]\right\} dS = \frac{1}{C_{n,l}} |\Delta|^{(l-1)/2}. \tag{3.32}$$

We now derive the Bayesian estimator. Denote the matrix Δ^{-1} by $\mathscr{D}$. Clearly $|\Delta| = 1/|\mathscr{D}|$. Let the prior distribution of parameters μ and Δ of an n-dimensional normal distribution $N(\mu, \Delta)$ be defined in the form

$$P_{a,A}(\mu, \mathscr{D}) = P_a(\mu|\mathscr{D}) \cdot P_A(\mathscr{D}),$$

where the vector μ is distributed according to the normal distribution

$$P_a(\mu|\mathscr{D}) = c_1 |\mathscr{D}|^{1/2} \exp\left\{-\frac{\omega}{2}(\mu - a)^{\mathrm{T}}\mathscr{D}(\mu - a)\right\};$$

here c_1 is a constant, $\omega > 0$ is a number, a is a vector, and $\mathscr{D}$ is a matrix distributed according to the Wishart distribution:

$$P_A(\mathscr{D}) = \begin{cases} C_{n,v}|\omega A|^{(v-1)/2}|\mathscr{D}|^{(v-n-2)/2} \exp\left\{-\dfrac{v\omega}{2}\mathrm{Sp}[A\mathscr{D}]\right\} & \text{for } |\mathscr{D}| > 0, \\ 0 & \text{for } |\mathscr{D}| \le 0. \end{cases}$$

Here $v > n + 2$ is a constant, A is a matrix. Observe that

$$\mathrm{Sp}[\mathscr{D}xx^{\mathrm{T}}] = \mathrm{Sp}[xx^{\mathrm{T}}\mathscr{D}] = x^{\mathrm{T}}\mathscr{D}x, \tag{3.33}$$

where $\mathscr{D}$ is a symmetric matrix and x is a column vector. We now write the joint density $P(x_1, \ldots, x_l|\mu, \mathscr{D})$ for a random independent sample $x_1, \ldots, x_l$:

$$\begin{aligned} P(x_1, \ldots, x_l|\mu, \mathscr{D}) &= c_2|\mathscr{D}|^{l/2} \exp\left\{\frac{-\sum_{i=1}^{l}(x_i - \mu)^{\mathrm{T}}\mathscr{D}(x_i - \mu)}{2}\right\} \\ &= c_2|\mathscr{D}|^{l/2} \exp\{-\tfrac{1}{2}\mathrm{Sp}[l\mathscr{D}S + l\mathscr{D}(x_{\mathrm{emp}} - \mu)(x_{\mathrm{emp}} - \mu)^{\mathrm{T}}]\}. \end{aligned}$$

Here and below c_0, c_1, c_2, and c_3 are constants which are determined by normalizing conditions. In view of Bayes's formula the posterior density $P(\mu, \mathscr{D}|x_1, \ldots, x_l)$ equals

$$P(\mu, \mathscr{D}|x_1, \ldots, x_l) = c_0 P(x_1, \ldots, x_l|\mu, \mathscr{D}) P_a(\mu|\mathscr{D}) P_A(\mathscr{D}). \tag{3.34}$$

Compute the right-hand side of (3.34):

$$\begin{aligned} &P(\mu, \mathscr{D}|x_1, \ldots, x_l) \\ &\quad = c_0|\mathscr{D}|^{l/2} \exp\{-\tfrac{1}{2}\mathrm{Sp}[l\mathscr{D}S + l\mathscr{D}(x_{\mathrm{emp}} - \mu)(x_{\mathrm{emp}} - \mu)^{\mathrm{T}}]\} \\ &\qquad \times c_1|\mathscr{D}|^{l/2} \exp\{-\tfrac{1}{2}\mathrm{Sp}[\mathscr{D}\omega(\mu - a)(\mu - a)^{\mathrm{T}}]\} \\ &\qquad \times c_2 \cdot C_{n,v}|\mathscr{D}|^{(v-n-2)/2}|\omega A|^{(v-1)/2} \exp\{-\tfrac{1}{2}\mathrm{Sp}[v\mathscr{D}A\omega]\} \\ &\quad = c_3|\mathscr{D}|^{(l+v-n-1)/2} \exp\{-\tfrac{1}{2}\mathrm{Sp}[l\mathscr{D}S + l\mathscr{D}(x_{\mathrm{emp}} - \mu)(x_{\mathrm{emp}} - \mu)^{\mathrm{T}} \\ &\qquad\qquad + \omega\mathscr{D}(\mu - a)(\mu - a)^{\mathrm{T}} + v\omega\mathscr{D}A]\}. \end{aligned} \tag{3.35}$$

Transforming the expression in the exponent of (3.35), we obtain

$$\mathscr{D}(lS + l(x_{\text{emp}} - \mu)(x_{\text{emp}} - \mu)^{\mathrm{T}} + \omega(\mu - a)(\mu - a)^{\mathrm{T}} + v\omega A) = \mathscr{D}[(l + \omega)(\mu - b)(\mu - b)^{\mathrm{T}} + (l + v)B],$$

where the notation

$$b = \frac{lx_{\text{emp}} + a\omega}{l + \omega}, \qquad B = \frac{\left(lS + \omega v A + \dfrac{l\omega}{l + \omega}(x_{\text{emp}} - a)(x_{\text{emp}} - a)^{\mathrm{T}}\right)}{l + v} \tag{3.36}$$

is used. Using this notation we rewrite (3.35):

$$P(\mu, \mathscr{D} | x_1, \ldots, x_l) = c_3 |\mathscr{D}|^{(l+v-n-1)/2} \times \exp\{-\tfrac{1}{2}\operatorname{Sp}[\mathscr{D}((l + \omega)(\mu - b)(\mu - b)^{\mathrm{T}} + (l + v)B)]\}. \tag{3.37}$$

The normalizing condition allows us to determine the constant c_3:

$$\begin{aligned} c_3^{-1} &= \int |\mathscr{D}|^{(l+v-n-2)/2} \exp\left\{-\frac{l + v}{2}\operatorname{Sp}[\mathscr{D}B]\right\} d\mathscr{D} \\ &\quad \times \int |\mathscr{D}|^{1/2} \exp\left\{-\frac{l + \omega}{2}\operatorname{Sp}[\mathscr{D}(\mu - b)(\mu - b)^{\mathrm{T}}]\right\} d\mu \\ &= \left(\frac{2\pi}{l + \omega}\right)^{n/2} (C_{n,l+v} |(l + v)B|^{(l+v-1)/2})^{-1}. \end{aligned}$$

The outer integral was computed utilizing equality (3.32). Finally we obtain the Bayesian estimator

$$\begin{aligned} P_{\mathrm{B}}(x) &= \int P(x | \mu, \mathscr{D}) P(\mu, \mathscr{D} | x_1, \ldots, x_l)\, d\mu\, d\mathscr{D} \\ &= \int (2\pi)^{-n/2} |\mathscr{D}|^{1/2} \exp\left\{-\frac{1}{2}(x - \mu)^{\mathrm{T}}\mathscr{D}(x - \mu)\right\} c_3 |\mathscr{D}|^{(l+v-n-1)/2} \\ &\quad \times \exp\left\{-\frac{l + \omega}{2}(\mu - b)^{\mathrm{T}}\mathscr{D}(\mu - b)\right\} \exp\left\{-\frac{l + v}{2}\operatorname{Sp}[\mathscr{D}B]\right\} d\mu\, d\mathscr{D} \\ &= \left(\frac{2\pi}{l + \omega}\right)^{n/2} \int c_3 |\mathscr{D}|^{(l+v-n-2)/2} \exp\left\{-\frac{l + v}{2}\operatorname{Sp}[\mathscr{D}B]\right\} d\mathscr{D} \\ &\quad \times \int (2\pi)^{-n/2} |\mathscr{D}|^{1/2} (2\pi)^{-n/2} |(l + \omega)\mathscr{D}|^{1/2} \\ &\quad \times \exp\left\{-\frac{1}{2}(x - \mu)^{\mathrm{T}}\mathscr{D}(x - \mu)\right\} \exp\left\{-\frac{l + \omega}{2}(\mu - b)^{\mathrm{T}}\mathscr{D}(\mu - b)\right\} d\mu. \end{aligned}$$

Observe that the inner integral with respect to μ is a convolution of two normal distributions; we thus obtain

$$P_{\mathrm{B}}(x) = c_3 \int (l + \omega + 1)^{-n/2} |\mathscr{D}|^{(l+v-n-1)/2} \times \exp\left\{-\tfrac{1}{2}\operatorname{Sp}\left[\mathscr{D}\left(B(l + v) + \frac{l + \omega}{l + \omega + 1}(x - b)(x - b)^{\mathrm{T}}\right)\right]\right\} d\mathscr{D}. \tag{3.38}$$

In view of (3.32) we have

$$P_B(x) = \frac{c_3(l+\omega+1)^{-n/2}}{C_{n,l+v+1}} \times \left|(l+v)B + \frac{l+\omega}{l+\omega+1}(x-b)(x-b)^T\right|^{-(l+v)/2}$$

$$= \left(\frac{l+\omega}{l+\omega+1}\frac{1}{2\pi}\right)^{n/2} \frac{C_{n,l+v}}{C_{n,l+v+1}} \times \frac{|(l+v)B|^{(l+v-1)/2}}{\left|(l+v)B + \frac{l+\omega}{l+\omega+1}(x-b)(x-b)^T\right|^{(l+v)/2}}. \tag{3.39}$$

We now transform the expression (3.39):

$$P_B(x) = \left(\frac{1}{\pi}\frac{l+\omega}{l+\omega+1}\right)^{n/2} \frac{\Gamma\left(\frac{l+v}{2}\right)}{\Gamma\left(\frac{l+v-n}{2}\right)} \times \frac{|(l+v)\cdot B|^{-1/2}}{\left|I + \frac{l+\omega}{l+\omega+1}\frac{1}{l+v}(x-b)(x-b)^T B^{-1}\right|^{(l+v)/2}}. \tag{3.40}$$

In the denominator of this expression I is the unit matrix. Observe that the matrix $(x-b)(x-b)^T$ and hence the matrix $(x-b)(x-b)^T B^{-1}$ are of rank 1. Thus only one of its eigenvalues is different from zero, which implies that the denominator of (3.40) is equal to

$$\left|I + \frac{l+\omega}{l+\omega+1}\frac{1}{l+v}(x-b)(x-b)^T B^{-1}\right|^{(l+v)/2} = \left(1 + \frac{l+\omega}{l+\omega+1}\frac{1}{l+v}(x-b)^T B^{-1}(x-b)\right)^{(l+v)/2}.$$

Thus we finally obtain

$$P_B(x) = \left(\frac{1}{\pi}\frac{l+\omega}{l+\omega+1}\right)^{n/2} \frac{\Gamma\left(\frac{l+v}{2}\right)}{\Gamma\left(\frac{l+v-n}{2}\right)} \times \frac{|(l+v)B|^{-1/2}}{\left(1 + \frac{l+\omega}{l+\omega+1}\frac{1}{l+v}(x-b)^T B^{-1}(x-b)\right)^{(l+v)/2}}.$$

We now assign specific values for v and ω in order that under the conditions of the scheme we shall obtain the most general (undetermined) prior conditions:

(1) $v = n + \varepsilon$ ($\varepsilon > 0$). This condition is necessary for integrating Wishart's distribution.

(2) $\omega \to 0$, $\varepsilon \to 0$. This condition assures that each of the elements of the matrix A tends to zero.

Then in view of (3.36) we obtain that $b \to x_{\text{emp}}$, $(l + v)B \to lS$, whence

$$P_{\text{B}}(x) = \left(\frac{1}{(l+1)\pi}\right)^{n/2} \frac{\Gamma\left(\frac{l+n}{2}\right)}{\Gamma\left(\frac{l}{2}\right)} \frac{|S|^{-1/2}}{\left(1 + \frac{1}{l+1}(x - x_{\text{emp}})^{\text{T}} S^{-1}(x - x_{\text{kmp}})\right)^{(l+n)/2}}.$$

Finally for the one-dimensional case (setting $n = 1$) we obtain

$$P_{\text{B}}(x) = \sqrt{\frac{1}{\pi(l+1)}} \frac{1}{\sigma_{\text{emp}}} \frac{\Gamma\left(\frac{l+1}{2}\right)}{\Gamma\left(\frac{l}{2}\right)} \frac{1}{\left(1 + \frac{1}{l+1}\frac{(x - x_{\text{emp}})^2}{\sigma_{\text{emp}}^2}\right)^{(l+1)/2}}.$$

§8 Unbiased Estimators

In the preceding sections, the Bayesian estimators of a probability density for special prior distributions on parameters were obtained. However, in practical problems the prior distribution is usually unknown. The minimax scheme of estimating the density may lead to overly imprecise results. It would therefore be desirable to find a sufficiently reliable method of estimating densities which is not connected with the Bayesian approach. How can this be done?

Assume that there exists a method of estimating densities which is best not only on the average (this corresponds to the Bayesian criterion), but also the best for estimating each specific density. For this uniformly best method to exist it must be independent of the prior distribution imposed on the density.

Unfortunately there is no such (uniformly best) method of estimation in the class of all possible methods. Indeed there exists a trivial algorithm which estimates the density to have the same fixed values of parameters independently of the sample. Such an algorithm estimates a single density with complete precision, while it is a poor estimator for all the other ones. This estimator is of course the best for its own density.

However, while there is no uniformly best method in the class of all possible estimation methods, there may perhaps exist such a method in a more restricted class. This prompts the idea of restricting the class of all possible methods of density estimation and attempting to find the best method within the class. It turns out that if we restrict the class of estimators to the so-called *unbiased estimators of density*, then the problem of finding a uniformly best one admits a solution.

Definition. We say that the function $\pi(x; x_1, \ldots, x_l)$ is an *unbiased estimator* of the density $P(x, \alpha^*)$ belonging to the class $P(x, \alpha)$ constructed from a sample $x_1, \ldots, x_l$ of size l obtained according to distribution $P(x, \alpha^*)$ if

the mathematical expectation of the estimator $\pi(x_1, \ldots, x_l)$ equals the density $P(x, \alpha^*)$, i.e., if for any $P(x, \alpha^*)$ belonging to $P(x, \alpha)$ the equality

$$M_{\alpha^*}\pi(x; x_1, \ldots, x_l) = P(x, \alpha^*)$$

is valid.

Note that the unbiasedness property has no value on its own and it is introduced solely to narrow down the class of possible estimators. The reason why the class of unbiased estimators is widely used in statistics is that this class is accessible to analysis.

What is the meaning of this accessibility? We write once again the definition of an unbiased estimator:

$$\int \pi(x; x_1, \ldots, x_l)P(x_1, \ldots, x_l; \alpha)\, dx_1 \cdots dx_l = P(x, \alpha). \tag{3.41}$$

This expression not only determines unbiased density estimators, but indicates a method for their construction: the set of unbiased estimators is the set of solutions of Fredholm's equation of type I. However, to obtain a solution of Equation (3.41) is usually a difficult problem. It was shown in Chapter 1 that even in the case when the solution of Fredholm's equation is unique, its numerical solution is an ill-posed problem. Therefore one can obtain unbiased estimators of the density $P(x, \alpha)$ only if Equation (3.41) can be solved analytically.

In Section 10 an optimal unbiased estimator of density for a multivariate normal distribution will be derived. Before proceeding to construct this estimator, we note that in Chapter 2 a more general problem of density estimation in the class of continuous functions was also reduced to a solution of Fredholm's equation of type I. In this case a special problem—obtaining an unbiased estimator of a density known up to its parameters—is reduced to Fredholm's equation.

The substantial difference between these two situations is that in the general case considered in Chapter 2 the right-hand side of Fredholm's equation of type I is known up to the error term. Here, however, it is given precisely.

§9 Sufficient Statistics

The construction of the optimal unbiased estimator is possible in terms of the so-called *sufficient statistics*. Up until now, when studying estimators we assumed that the estimator of a density is of the form $\pi(x; x_1, \ldots, x_l)$, i.e., the estimator is a function of $l + 1$ vectors: the vector x and l vector-valued variables $x_1, \ldots, x_l$. Fixing the last l variables we obtained a specific form of the estimated density.

However, such a method of expressing the density estimator is not quite convenient. Evidently $\pi(x; x_1, \ldots, x_l)$ should not depend on the order of

the vectors $x_1, \ldots, x_l$ of the sample. Moreover, for another sample size, say $l + 1$, it is necessary to give a new function (of dimensionality $l + 2$).

Therefore it would be desirable to find k characteristics of the sample

$$t_i = f_i(x_1, \ldots, x_l), \qquad i = 1, \ldots, k,$$

such that, first of all, the information concerning the density contained in the sample $x_1, \ldots, x_l$ would be included in these k numbers, and secondly, that the number of necessary characteristics k would depend not on the sample size but on the features of the class of estimated densities. It would be desirable to obtain an unbiased estimator $\pi^*(x; t_1, \ldots, t_k)$ in terms of these characteristics of the sample. Sufficient statistics indeed serve this purpose (see [58]).

Definition. We say that the functions $t_i = f_i(x_1, \ldots, x_l)$ are *sufficient statistics* for the density $P(x, \alpha)$ if the joint density $P(x_1, \ldots, x_l; \alpha)$ of the sample can be represented in the form

$$P(x_1, \ldots, x_l; \alpha) = P_1(t_1, \ldots, t_k; \alpha) P_2(x_1, \ldots, x_l).$$

In other words, the joint density $P(x_1, \ldots, x_l; \alpha)$ is decomposed into the product of two terms. One of them, $P_2(\cdot)$, does not depend on the parameter α, while the other involving α depends only on the values $t_1, \ldots, t_k$ (but not on the sample $x_1, \ldots, x_l$).

It is easy to verify that for an n-dimensional normal distribution the following $n(n + 3)/2$ quantities serve as sufficient statistics:

$$t = \frac{1}{l} \sum_{j=1}^{l} x_j, \quad t = (t_1, \ldots, t_n)^{\mathrm{T}} \qquad (n \text{ values});$$

$$\|t_{ij}\| = \sum_{r=1}^{l} (x_r - t)(x_r - t)^{\mathrm{T}} \qquad \left(\frac{n(n+1)}{2} \text{ values}\right).$$

Indeed, for an n-dimensional normal distribution we have

$$\begin{aligned}
&P(x_1, \ldots, x_l; \mu, \Delta) \\
&\quad = \frac{1}{(2\pi)^{nl/2} |\Delta|^{l/2}} \exp\left\{ -\frac{1}{2} \sum_{i=1}^{l} (x_i - \mu)^{\mathrm{T}} \Delta^{-1} (x_i - \mu) \right\} \\
&\quad = (2\pi)^{-nl/2} |\Delta|^{-l/2} \exp\left\{ -\tfrac{1}{2} \operatorname{Sp}\left[\Delta^{-1} \sum_{i=1}^{l} (x_i - \mu)(x_i - \mu)^{\mathrm{T}} \right\} \right] \\
&\quad = (2\pi)^{-nl/2} |\Delta|^{-l/2} \\
&\qquad \times \exp\left\{ -\tfrac{1}{2} \operatorname{Sp}\left[\Delta^{-1} \left(\sum_{i=1}^{l} (x_i - t)(x_i - t)^{\mathrm{T}} + l(t - \mu)(t - \mu)^{\mathrm{T}} \right) \right] \right\} \\
&\quad = (2\pi)^{-nl/2} |\Delta|^{-l/2} \exp\{ -\tfrac{1}{2} \operatorname{Sp}[\Delta^{-1}(\|t_{ij}\| + l(t - \mu)(t - \mu)^{\mathrm{T}})] \}.
\end{aligned}$$

In the derivation the equality $z^{\mathrm{T}} B z = \operatorname{Sp}[z z^{\mathrm{T}} B]$ was used.

Thus we seek an estimator of the density as a function of sufficient statistics.

The remarkable feature of unbiased estimators $\pi^*(x; t_1, \ldots, t_k)$ is that they are in some sense always at least as good as the estimators $\pi(x; x_1, \ldots, x_l)$.

Theorem (cf. [35, 58]). *For any estimator $\pi(x; x_1, \ldots, x_l)$ there exists an estimator $\pi^*(x; t_1, \ldots, t_k)$ such that for any density belonging to $P(x, \alpha)$ the mathematical expectations of the estimators are the same:*

$$M\pi^*(x; t_1, \ldots, t_k) = M\pi(x; x_1, \ldots, x_l) = \pi(x),$$

but the variance $\pi^*(x; t_1, \ldots, t_k)$ is not larger than the variance of the estimator $\pi^*(x; x_1, \ldots, x_l)$, i.e.,

$$M(\pi(x) - \pi^*(x; t_1, \ldots, t_k))^2 \leq M(\pi(x) - \pi(x; x_1, \ldots, x_l))^2.$$

It follows from this theorem that the class of unbiased estimators—expressed in terms of a sufficient statistic—contains the best one.

§10 Computing the Best Unbiased Estimator

We shall construct the best unbiased estimator of the density for a multidimensional normal distribution. Here we utilize the fact that for distributions of the exponential type there exists a unique unbiased estimator expressed in terms of sufficient statistics [26, 35]. In other words there exists a unique solution for Fredholm's equation of type I,

$$\int \pi^*(x; t_1, \ldots, t_k) P(t_1, \ldots, t_k; \alpha)\, dt_1, \ldots, dt_k = P(x, \alpha), \tag{3.42}$$

where $P(x, \alpha)$ is the normal distribution and $P(t_1, \ldots, t_k; \alpha)$ is the probability density of its sufficient statistics.

According to the theorem cited in the preceding section, the solution of Equation (3.42), in view of its uniqueness, is the best unbiased estimator of the density of a multidimensional normal distribution.

We shall show that an unbiased estimator of an n-dimensional normal density is

$$P_{\text{unb}}(x) = \frac{\Gamma\left(\frac{l-1}{2}\right)}{[(l-1)\pi]^{n/2}\Gamma\left(\frac{l-n-1}{2}\right)|S|^{1/2}} \times \left[1 - \frac{(x - x_{\text{emp}})^{\mathrm{T}} S^{-1}(x - x_{\text{emp}})}{l-1}\right]_+^{(l-n-3)/2}.$$

Here $x_{\text{emp}} = (1/l)\sum_{i=1}^{l} x_i$ is the vector of the means,

$$S = \frac{1}{l}\sum_{i=1}^{l}(x_i - x_{\text{emp}})(x_i - x_{\text{emp}})^{\mathrm{T}}$$

is the empirical estimator of the covariance matrix Δ, and $[z]_+$ denotes

$$[z]_+ = \begin{cases} z & \text{for } z \geq 0, \\ 0 & \text{for } z < 0. \end{cases}$$

In deriving the best unbiased estimator of an n-dimensional density we shall utilize Bayes's formula

$$\varphi(x_l|t) = \frac{q(x_l, t; \alpha)}{P(t; \alpha)}, \tag{3.43}$$

where $t = (t_1, \ldots, t_k)^{\mathrm{T}}$, $x_l = (x_l^1, \ldots, x_l^n)^{\mathrm{T}}$, the density $q(x_l, t; \alpha)$ defines the distribution of statistics x_l and t, $P(t, \alpha)$ is the distribution of t, and $\varphi(x_l|t)$ is the conditional density. We shall show that the conditional density (3.43) is an unbiased estimator of the density $P(x, \alpha)$. Indeed,

$$\int \varphi(x_l|t)P(t; \alpha)\, dt = \int q(x, t; \alpha)\, dt = P(x, \alpha).$$

And since the unbiased estimator expressed in terms of sufficient statistics is unique, $\varphi(x|t)$ is the best unbiased estimator.

We now compute $\varphi(x|t)$. First we shall find $q(x, t; \alpha)$. For a normal distribution of the occurrence of vector x we have

$$q(x, t; \alpha) = q(x, x_{\text{emp}}, S; \mu, \Delta),$$

where

$$x = x_l, \qquad x_{\text{emp}} = \frac{1}{l}\sum_{i=1}^{l} x_i, \qquad S = \frac{1}{l}\sum_{i=1}^{l}(x_i - x_{\text{emp}})(x_i - x_{\text{emp}})^{\mathrm{T}}.$$

Let the vectors $x_1, \ldots, x_l$ which form the triples x, x_{emp}, S appear randomly and independently according to the density $N(\mu, \Delta)$.

Consider vectors $y_1, \ldots, y_l$ obtained from $x_1 - \mu, \ldots, x_l - \mu$ by an orthogonal transformation

$$\mathscr{L} = \begin{bmatrix} c_{11} & \cdots & c_{1\,l-1} & 0 \\ \vdots & & \vdots & \\ c_{l-2\,1} & \cdots & c_{l-2\,l-1} & 0 \\ \dfrac{1}{\sqrt{l-1}} & \cdots & \dfrac{1}{\sqrt{l-1}} & 0 \\ 0 & \cdots & 0 & 1 \end{bmatrix}.$$

Vectors $y_1, \ldots, y_l$ are distributed independently according to the $N(0, \Delta)$ distribution. The following relations are valid:

$$x_l = y_l + \mu, \qquad x_{\text{emp}} = \frac{\sqrt{l-1}}{l}\, y_{l-1} + \frac{y_l}{l} + \mu.$$

We now express the matrix S in terms of the vectors $y_1, \ldots, y_l$. For this purpose we utilize the representation

$$
\begin{aligned}
S = {} & \frac{1}{l}\sum_{i=1}^{l-1}(x_i-\mu)(x_i-\mu)^{\mathrm{T}} \\
& + \frac{(x_l-\mu)(x_l-\mu)^{\mathrm{T}}}{l} - \frac{l-1}{l}\left[\sum_{i=1}^{l-1}\frac{x_i-\mu}{\sqrt{l-1}}\right]\left[\sum_{i=1}^{l-1}\frac{x_i-\mu}{\sqrt{l-1}}\right]^{\mathrm{T}} \\
& - \frac{l-1}{l}\left[\sum_{i=1}^{l-1}\frac{x_i-\mu}{\sqrt{l-1}}\right](x_l-\mu)^{\mathrm{T}} - \frac{l-1}{l}(x_l-\mu)\left[\sum_{i=1}^{l-1}\frac{x_i-\mu}{\sqrt{l-1}}\right]^{\mathrm{T}} \\
& - \frac{1}{l^2}(x_l-\mu)(x_l-\mu)^{\mathrm{T}},
\end{aligned}
$$

and the fact that the transformation $\mathscr{L}$ satisfies

$$
\sum_{i=1}^{l-1}(x_i-\mu)(x_i-\mu)^{\mathrm{T}} = \sum_{i=1}^{l-1} y_i y_i^{\mathrm{T}}.
$$

We thus obtain

$$
S = \frac{1}{l}\sum_{i=1}^{l-2} y_i y_i^{\mathrm{T}} + \left(\frac{y_{l-1}-\sqrt{l-1}\,y_l}{l}\right)\cdot\left(\frac{y_{l-1}-\sqrt{l-1}\,y_l}{l}\right)^{\mathrm{T}}.
$$

Denote

$$
\mathscr{D} = \frac{1}{l}\sum_{i=1}^{l-2} y_i y_i^{\mathrm{T}}.
$$

Observe that vectors $y_1, \ldots, y_l$ are distributed according to the normal distribution $N(0, \Delta)$. Moreover the variables y_{l-1}, y_l, and $\mathscr{D}$ are independent. Since y_{l-1}, y_l are distributed according to the normal distribution and $\mathscr{D}$ has a Wishart distribution, the joint distribution $P(y_{l-1}, y_l, \mathscr{D}; 0, \Delta)$ equals

$$
P(y_{l-1}, y_l, \mathscr{D}; 0, \Delta) = P(y_{l-1}; 0, \Delta)\, P(y_l; 0, \Delta)\, W_{l-1}(\mathscr{D}; \Delta), \tag{3.44}
$$

where $W_{l-1}(\mathscr{D}, \Delta)$ is the Wishart distribution:

$$
W_{l-1}(\mathscr{D}, \Delta) = \begin{cases} C_{n,l-1}\dfrac{|\mathscr{D}|^{(l-n-3)/2}\exp\{-\frac{1}{2}\operatorname{Sp}[\Delta^{-1}\mathscr{D}]\}}{|\Delta|^{(l-2)/2}} & \text{for } |\mathscr{D}| > 0, \\ 0 & \text{for } |\mathscr{D}| \le 0, \end{cases}
$$

and $C_{n,l}$ is a constant defined in (3.31).

We now express $P(y_{l-1}, y_l, \mathscr{D}; 0, \Delta)$ in terms of x_l, x_{emp}, and S. First observe that

$$
y_l = x_l - \mu, \qquad y_{l-1} = \frac{l}{\sqrt{l-1}}(x_{\text{emp}} - \mu) - \frac{(x_l-\mu)}{\sqrt{l-1}},
$$

$$
\mathscr{D} = lS - \frac{l}{l-1}(x_l - x_{\text{emp}})(x_l - x_{\text{emp}})^{\mathrm{T}}. \tag{3.45}
$$

Taking into account that the Jacobian of the transformation (3.45) equals $l^{n(n+3)/2}/(l-1)^{n/2}$, and substituting (3.45) into (3.44), we obtain

$$q(x_l, x_{\text{emp}}, S; \mu, \Delta) = \frac{l^{n(n+3)/2}}{(l-1)^{n/2}} P\left(\frac{l}{\sqrt{l-1}}(x_{\text{emp}} - \mu) - \frac{(x_l - \mu)}{\sqrt{l-1}}; 0, \Delta\right)$$
$$\times P(x_l - \mu; 0, \Delta) W_{l-1}\left(lS - \frac{l}{l-1}(x_l - x_{\text{emp}})(x_l - x_{\text{emp}})^{\mathrm{T}}; \Delta\right),$$

whence

$$q(x_l, x_{\text{emp}}, S; \mu, \Delta)$$
$$= \begin{cases} \dfrac{l^{n(n+3)/2} C_{n,l-1} \left| lS - \dfrac{l(x_l - x_{\text{emp}})(x_l - x_{\text{emp}})^{\mathrm{T}}}{l-1} \right|^{(l-n-3)/2} |\mathscr{D}|^{l/2}}{(2\pi)^n (l-1)^{n(l-1)/2} |\Delta|^{l/2} \exp\left\{\dfrac{l}{2} \operatorname{Sp}[\Delta^{-1}(S + (x_{\text{emp}} - \mu)(x_{\text{emp}} - \mu)^{\mathrm{T}})]\right\}}, \\ \qquad \text{if } \left| S - \dfrac{(x_l - x_{\text{emp}})(x_l - x_{\text{emp}})^{\mathrm{T}}}{l-1} \right| > 0, \\ 0, \qquad \text{if } \left| S - \dfrac{(x_l - x_{\text{emp}})(x_l - x_{\text{emp}})^{\mathrm{T}}}{l-1} \right| = 0. \end{cases} \tag{3.46}$$

We shall now determine the denominator of the expression (3.43). For a normal distribution of vectors x, the statistics x_{emp} and lS are distributed independently:

$$P(x_{\text{emp}}, S; \mu, \Delta) = P(x_{\text{emp}}; \mu, \Delta) P(S; \Delta), \tag{3.47}$$

where x_{emp} is normally distributed with $N(\mu, (1/l)\Delta)$, and lS has the Wishart distribution $W_l(S; \Delta)$. This implies that

$$P(x_{\text{emp}}, S; \mu, \Delta) = \frac{C_{n,l}}{(2\pi)^{n/2}} \frac{l^{n/2} |S|^{(l-n-2)/2}}{|\Delta|^{l/2} \exp\left\{\dfrac{l}{2} \operatorname{Sp}[\Delta^{-1}(S + (x_{\text{emp}} - \mu)(x_{\text{emp}} - \mu)^{\mathrm{T}})]\right\}}, \tag{3.48}$$

if $|S| \geq 0$ and zero otherwise. $C_{n,l}$ is a constant defined in (3.31).

Substituting (3.46) and (3.48) into (3.43) we obtain

$$\varphi(x|t) = \frac{\Gamma\left(\dfrac{l-1}{2}\right)[(l-1)\pi]^{-n/2}}{\Gamma\left(\dfrac{l-n-1}{2}\right)|S|^{1/2}} \left(\frac{\left| S - \dfrac{(x - x_{\text{emp}})(x - x_{\text{emp}})^{\mathrm{T}}}{l-1} \right|}{|S|} \right)^{(l-n-3)/2}$$

in the case when $|S| > 0$ and $|S - [(x - x_{\text{emp}})(x - x_{\text{emp}})^{\mathrm{T}}/(l-1)]| \geq 0$. Observe that

$$\frac{\left| S - \dfrac{(x - x_{\text{emp}})(x - x_{\text{emp}})^{\mathrm{T}}}{l-1} \right|}{|S|} = \left(1 - \frac{(x - x_{\text{emp}})^{\mathrm{T}} S^{-1} (x - x_{\text{emp}})}{l-1}\right).$$

Hence we finally obtain

$$\varphi(x|x_{\text{emp}}, S) = \frac{\Gamma\left(\frac{l-1}{2}\right)}{[(l-1)\pi]^{n/2}\Gamma\left(\frac{l-n-1}{2}\right)|S|^{1/2}}\left[1 - \frac{(x-x_{\text{emp}})^{\mathrm{T}}S^{-1}(x-x_{\text{emp}})}{l-1}\right]_{+}^{(l-n-3)/2},$$

where

$$[z]_{+} = \begin{cases} z & \text{for } z \geq 0, \\ 0 & \text{for } z < 0. \end{cases}$$

§11 The Problem of Estimating the Parameters of a Density

It would thus seem that we have succeeded in achieving our goal of constructing a Bayesian estimator of a density and computing the best unbiased estimator. However, the methods which were utilized in obtaining these estimators substantially utilize special properties of the estimated density. Therefore the methods studied above are not the common ones for estimating densities of various types.

It is therefore of interest to study methods which perhaps do not yield such precise approximations as those studied above but which are regular, i.e., which could be used for estimating densities belonging to different parametric classes.

To obtain these methods we shall reformulate our problem. We shall assume that our purpose is the estimation of parameters of a density rather than density estimation. We also assume that if one succeeds in solving the intermediate problem of obtaining a "nice" estimator for the parameters of the density, then the density itself can be satisfactorily estimated by choosing as an approximation the density function $P(x, \alpha^*)$, where α^* are the estimated values of the parameters.

Observe that when the normal (Gaussian) distribution is estimated, neither the Bayes approximation nor the unbiased estimator of the density belongs to the class of normal distributions. In the case when the density is "assessed" by estimating its parameters, the approximations obtained belong to the Gaussian class. (This fact of itself is of no value. It only indirectly indicates how far the solution obtained may be from, say, the Bayesian one.)

Thus we shall estimate the parameters α_0 of the density $P(x, \alpha_0)$. The quantity

$$d(\alpha_0, \hat{\alpha}; x_1, \ldots, x_l) = |\alpha_0 - \hat{\alpha}(x_1, \ldots, x_l)|^2$$

will serve as the measure of the quality of the estimator $\hat{\alpha} = \hat{\alpha}(x_1, \ldots, x_l)$ of the vector of parameters $\alpha = \alpha_0$ based on the sample $x_1, \ldots, x_l$. The mathematical expectation of the quantity $d(\alpha_0, \hat{\alpha}; x_1, \ldots, x_l)$, i.e.,

$$d(\alpha_0, \hat{\alpha}, l) = \int d(\alpha_0, \hat{\alpha}; x_1, \ldots, x_l) P(x_1, \ldots, x_l; \alpha_0)\, dx_1 \cdots dx_l$$

serves as the measure of the quality of estimators of $\alpha = \alpha_0$ based on samples of size l (where $P(x_1, \ldots, x_l; \alpha_0)$ is the probability density of the sample $x_1, \ldots, x_l$).

Finally the quality of an estimator used for estimating the parameter α under the prior distribution $P(\alpha)$ will be measured by

$$R_B(\hat{\alpha}, l) = \int d(\alpha, \hat{\alpha}, l) P(\alpha)\, d\alpha. \tag{3.49}$$

The estimator $\hat{\alpha}$ which yields the minimum of the functional (3.49) is called a *Bayesian estimator of parameters.*

As in the case of density estimation, the prior distribution $P(\alpha)$ of parameters α is usually unknown; therefore, as before, the minimax criterion

$$R_{\text{mnx}}(\hat{\alpha}, l) = \sup_{\alpha} d(\alpha, \hat{\alpha}, l)$$

makes sense. The vector $\hat{\alpha}$ which yields the minimum of $R_{\text{mnx}}(\hat{\alpha}, l)$ forms the *minimax estimator of parameters.* However, the construction of a regular method for parameter estimation of a density is associated with the idea of the best unbiased estimation rather than with the Bayesian or minimax estimation.

Definition. We say that estimator $\hat{\alpha} = \hat{\alpha}(x_1, \ldots, x_l)$ is an *unbiased estimator* of the vector of parameters α_0 if

$$\int \hat{\alpha}(x_1, \ldots, x_l) P(x_1, \ldots, x_l; \alpha_0)\, dx_1 \cdots dx_l = \alpha_0.$$

Consider first the case when the probability density $P(x, \alpha_0)$ depends only on a scalar parameter α_0. Then for the class of unbiased estimators, the remarkable *Rao–Cramèr inequality* is valid:

$$\int (\alpha_0 - \hat{\alpha}(x_1, \ldots, x_l))^2 P(x_1, \ldots, x_l; \alpha_0)\, dx_1 \cdots dx_l \geq \frac{1}{I_\Phi}, \tag{3.50}$$

where

$$I_\Phi = -\int \frac{d^2 \ln P(x_1, \ldots, x_l; \alpha_0)}{d\alpha^2} P(x_1, \ldots, x_l; \alpha_0)\, dx_1 \cdots dx_l.$$

The quantity I_Φ is called *Fisher's information quantity*. For an independent sample it equals

$$I_\Phi = -l \int \frac{d^2 \ln P(x, \alpha_0)}{d\alpha^2} P(x, \alpha_0)\, dx.$$

A derivation of the Rao–Cramèr inequality is given in practically all modern texts in statistics (see, e.g., [35, 49, 58]). The meaning of this inequality is that the variance of an unbiased estimator (and this variance measures the precision of estimation in the case of unbiased estimators) is never less than the inverse of the Fisher's information quantity.

Thus the right-hand side of the inequality (3.50) determines the limiting precision of unbiased estimation of a parameter. An estimator for which the inequality (3.50) becomes an equality is called *efficient*. The problem is to obtain a regular method for constructing efficient estimators of parameters for various parametric classes of densities.

An inequality analogous to (3.50) may be obtained also for simultaneous unbiased estimation of several parameters. In this case the *Fisher information matrix I* whose elements are

$$I_{ij} = -\int \frac{\partial^2 \ln P(x_1, \ldots, x_l; \alpha_0)}{\partial \alpha_i\, \partial \alpha_j} P(x_1, \ldots, x_l; \alpha_0)\, dx_1 \cdots dx_l,$$

$$i, j = 1, 2, \ldots, n,$$

serves as an analog of the information quantity.

For an independent sample $x_1, \ldots, x_l$ the elements I_{ij} are equal to

$$I_{ij} = -l \int \frac{\partial^2 \ln P(x, \alpha)}{\partial \alpha_i\, \partial \alpha_j} dx.$$

Let the Fisher information matrix I be nonsingular, and let the estimators $\hat{\alpha}_1(x_1, \ldots, x_l), \ldots, \hat{\alpha}_n(x_1, \ldots, x_l)$ be unbiased estimators of the parameters $\alpha_1^0, \ldots, \alpha_n^0$. Consider for these estimators a covariance matrix B, i.e., a matrix with the elements

$$b_{ij} = M(\alpha_i^0 - \hat{\alpha}_i(x_1, \ldots, x_l))(\alpha_j^0 - \hat{\alpha}_j(x_1, \ldots, x_l)).$$

Then a multidimensional analog of the Rao–Cramèr inequality is the following assertion: for any vector z and any unbiased estimators $\hat{\alpha}_1(x_1, \ldots, x_l), \ldots, \hat{\alpha}_n(x_1, \ldots, x_l)$, the inequality

$$z^{\mathrm{T}} B z \geq z^{\mathrm{T}} I^{-1} z \tag{3.51}$$

is valid. The meaning of this inequality is as follows: let the quality of the joint estimator of n parameters $\alpha_1^0, \ldots, \alpha_n^0$ be determined by the square of weighted sums of deviation (with weights $z = (z, \ldots, z_n)^{\mathrm{T}}$, $z_i \geq 0$) over all the estimated parameters:

$$T(x_1, \ldots, x_l) = \left(\sum_{i=1}^{n} z_i(\alpha_i^0 - \hat{\alpha}_i(x_1, \ldots, x_l)) \right)^2.$$

Then the mathematical expectation of $T(x_1, \ldots, x_n)$ is bounded from below by the quantity $z^T I^{-1} z$. In other words, no matter how the quality of the joint unbiased estimation of n parameters is measured (i.e., for any weights z_i), the bound

$$MT(x_1, \ldots, x_l) \geq z^T I^{-1} z$$

is valid. In particular it follows from the inequality (3.51) that the variance of the estimator with respect to each parameter separately satisfies the inequality (3.50). Indeed, (3.50) is obtained from (3.51) for the specific vector $z = (0, \ldots, 0, 1, 0, \ldots, 0)^T$.

Estimation methods which yield equality in (3.51) for all z are called *jointly efficient*. When estimating several parameters our goal is to find jointly efficient estimators.

§12 The Maximum-Likelihood Method

Unfortunately there is no "regular" method to obtain efficient estimators of parameters of density based on a sample of a fixed size. There is only a method which allows us to construct asymptotically efficient estimators. This is the *maximum-likelihood method* developed by R. A. Fisher [58]. However, before considering this method we shall introduce several notions which are necessary for classifying estimators obtained from samples of large size.

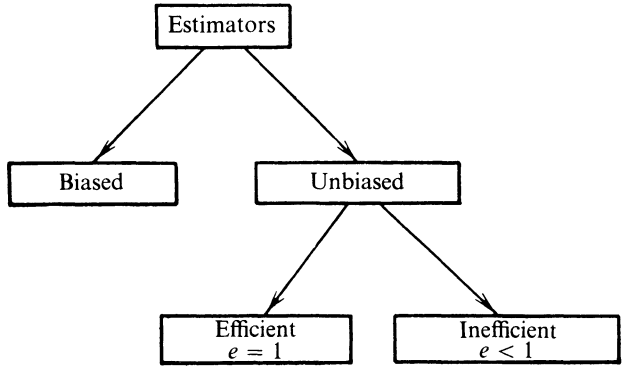

Figure 3

In the preceding section the classification presented here in Figure 3 was introduced for the characterization of estimators of parameters of a distribution based on samples of a fixed size. In this figure a measure of the efficiency of an unbiased estimator of parameters α_0 is also shown. In the

case of a single parameter this measure is given by

$$e_l = \frac{1}{M(\alpha_0 - \hat{\alpha}(x_1, \ldots, x_\alpha))^2 I_\Phi}. \tag{3.52}$$

In the case of joint estimation of several parameters the measure of efficiency is defined by

$$e_l = \frac{v(B, l)}{v(I, l)}, \tag{3.53}$$

which equals the ratio of the volume $v(B, l)$ of the ellipsoid

$$z^T B z = 1$$

to the volume of the ellipsoid

$$z^T I^{-1} z = 1.$$

For sample of large size a somewhat different classification is used which incorporates the notions of asymptotically unbiased, consistent, and asymptotically efficient estimators. Estimators satisfying

$$M_{\alpha_0} \hat{\alpha}(x_1, \ldots, x_l) \underset{l \to \infty}{\longrightarrow} \alpha_0$$

are called *asymptotically unbiased.* Estimators satisfying

$$P_{\alpha_0}\{|\hat{\alpha}(x_1, \ldots, x_l) - \alpha_0| > \varepsilon\} \underset{l \to \infty}{\longrightarrow} 0$$

for all $\varepsilon > 0$ are called *consistent.* Asymptotic unbiased estimators satisfying

$$e_l \underset{l \to \infty}{\longrightarrow} 1$$

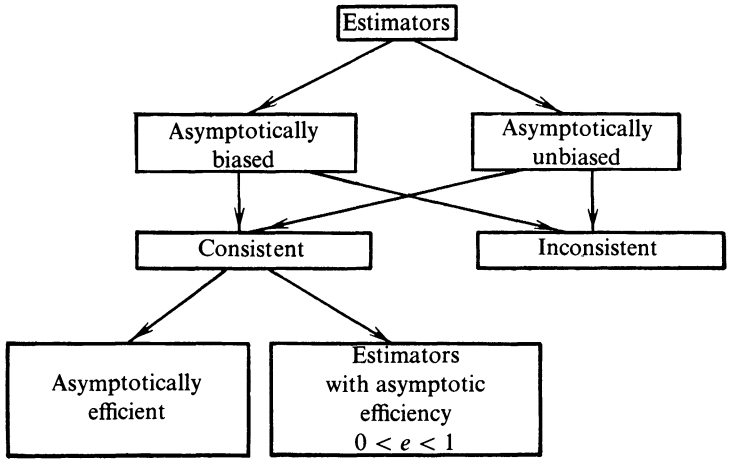

Figure 4

are called *asymptotically efficient*. Here e_l is given by (3.52) in the case of a single parameter α and by (3.53) when several parameters are jointly estimated. This classification is presented in Figure 4.

The method of maximum likelihood involves examining the likelihood function $P(x_1, \ldots, x_l; \alpha)$. In our case, when the sample $x_1, \ldots, x_l$ is obtained as a result of random independent observations according to the density $P(x, \alpha)$, the likelihood function can be represented as

$$P(x_1, \ldots, x_l; \alpha) = \prod_{i=1}^{l} P(x_i, \alpha). \tag{3.54}$$

The method of maximum likelihood chooses as the estimator those α which yield the maximum for (3.54). Along with the likelihood function (3.54) it is common to consider the function

$$\ln P(x_1, \ldots, x_l; \alpha) = \sum_{i=1}^{l} \ln P(x_i, \alpha). \tag{3.55}$$

The maxima of the functions (3.54) and (3.55) are the same, and hence to obtain maximum-likelihood estimators we need to solve the system of equations

$$\frac{\partial P(x_1, \ldots, x_l; \alpha)}{\partial \alpha_i} = 0, \qquad i = 1, 2, \ldots, n, \tag{3.56}$$

or the system of equations

$$\frac{\partial \ln P(x_1, \ldots, x_l; \alpha)}{\partial \alpha_i} = 0, \qquad i = 1, 2, \ldots, n. \tag{3.57}$$

The theory of maximum-likelihood estimation, which is well developed, aims to justify the applicability of this method. The substance of this theory is that for certain classes $P(x, \alpha)$ (to which all the classes of densities considered in this book belong) the maximum-likelihood method assures the asymptotic efficiency of the estimators (cf. [24, 58]).

We also remark that in the case of maximum-likelihood estimation the problem is reduced here to a simpler one than the one encountered in Bayesian estimation (multiple integration) or in constructing unbiased estimators (solution of Fredholm's equations of type I).

To implement the maximum-likelihood method it is necessary to solve the system of equations (3.56) or (3.57). Although this is not always a linear system, its numerical solution is not usually too difficult, and moreover, for a wide class of functions there exists a unique solution.

§13 Estimation of Parameters of the Probability Density Using the Maximum-Likelihood Method

In this section, utilizing the maximum-likelihood method, we shall obtain estimators for parameters of the distribution

$$P(x^i) = \begin{cases} p^i(1), & \text{for } x^i = c(1), \\ \vdots & \\ p^i(\tau_i), & \text{for } x^i = c(\tau_i), \end{cases} \qquad i = 1, 2, \ldots, n,$$

$$\sum_{j=1}^{\tau_j} p^i(j) = 1, \qquad i = 1, 2, \ldots, n,$$

as well as for parameters of the normal distribution

$$N(\mu, \Delta) = \frac{1}{(2\pi)^{n/2}|\Delta|^{1/2}} \exp\{-\tfrac{1}{2}(x - \mu)^{\mathrm{T}} \Delta^{-1}(x - \mu)\}.$$

It turns out that for the distribution $P(x^i)$ the estimators are given by

$$\hat{P}(x^i) = \begin{cases} \hat{p}^i(1) = \dfrac{m_i(1)}{l} & \text{for } x^i = c^i(1), \\ \vdots & \\ \hat{p}^i(\tau_i) = \dfrac{m_i(\tau_i)}{l} & \text{for } x^i = c^i(\tau_i), \end{cases} \tag{3.58}$$

where $m_i(j)$ is the number of vectors in the sample with the ith coordinate taking on the value $x^i = c^i(j)$.

Maximum-likelihood estimators of parameters of a multidimensional normal distribution are given by

$$x_{\text{emp}} = \frac{1}{l} \sum_{i=1}^{l} x_i,$$

$$S = \frac{1}{l} \sum_{i=1}^{l} (x_i - x_{\text{emp}})(x_i - x_{\text{emp}})^{\mathrm{T}}.$$

Thus we obtain the following estimator of the normal density:

$$\hat{P}(x) = \frac{1}{(2\pi)^{n/2}|S|^{1/2}} \exp\{-\tfrac{1}{2}(x - x_{\text{emp}})^{\mathrm{T}} S (x - x_{\text{emp}})\}. \tag{3.59}$$

13.1 Derivation in the Discrete Case

We estimate the parameters of the distribution $P(x^i)$. First we form the likelihood function:

$$P(x_1, \ldots, x_l; p) = \prod_{j=1}^{l} \prod_{i=1}^{n} P(x_j^i, p^i),$$

where x_j^i is the value of the ith coordinate of the j-vector in the sample.

Interchanging the order of the factors, we have

$$P(x_1, \ldots, x_l; p) = \prod_{i=1}^{n} \prod_{j=1}^{l} P(x_j^i, p^i).$$

We now proceed to the function

$$\ln P(x_1, \ldots, x_l; p) = \sum_{i=1}^{n} \sum_{j=1}^{l} \ln P(x_j^i, p^i).$$

Consider the quantity

$$\sum_{j=1}^{l} \ln P(x_j^i, p^i).$$

It can be represented in the form

$$\sum_{j=1}^{l} \ln P(x_j^i, p^i) = \sum_{r=1}^{\tau_i} m_i(r) \ln p^i(r),$$

where $m_i(r)$ is the number of vectors in the sample such that the ith coordinate takes the value $x^i = c^i(r)$. Thus

$$\ln P(x_1, \ldots, x_l; p) = \sum_{i=1}^{n} \sum_{r=1}^{\tau_i} m_i(r) \ln p^i(r). \tag{3.60}$$

We now obtain the maximum with respect to $p^i(r)$ of function (3.60) subject to $\sum_{r=1}^{\tau_i} p^i(r) = 1$, $i = 1, 2, \ldots, n$. For this purpose the method of Lagrange multipliers will be used. We form the Lagrange function

$$L(p, \lambda) = \sum_{i=1}^{n} \sum_{r=1}^{\tau_i} (m_i(r) \ln p^i(r) - \lambda_i p^i(r)), \tag{3.61}$$

where the λ_i are the Lagrange multipliers. The vector p^i which yields the maximum of $L(p, \lambda)$ is determined by the system of equations

$$\frac{\partial L(p^i, \lambda)}{\partial p^i(r)} = \frac{m_i(r)}{p^i(r)} - \lambda_i = 0, \qquad i = 1, \ldots, n. \tag{3.62}$$

From (3.62), taking the condition

$$\sum_{r=1}^{\tau_i} p^i(r) = 1,$$

into account, we obtain

$$\hat{p}^i(r) = \frac{m_i(r)}{l}.$$

Observe that here the maximum-likelihood estimators turn out to be unbiased.

13.2 Derivation in the Normal Case

We now estimate the parameters μ and Δ of the normal distribution:

$$P(x;\mu,\Delta)=\frac{1}{(2\pi)^{n/2}|\Delta|^{1/2}}\exp\{-\tfrac{1}{2}(x-\mu)^{\mathrm{T}}\Delta^{-1}(x-\mu)\}.$$

We form the likelihood function

$$P(x_1,\ldots,x_l;\mu,\mathscr{D})=\frac{|\mathscr{D}|^{l/2}}{(2\pi)^{ln/2}}\exp\left\{-\frac{1}{2}\sum_{i=1}^{l}(x_i-\mu)^{\mathrm{T}}\mathscr{D}(x_i-\mu)\right\},$$

where $\Delta^{-1}=\mathscr{D}$. We obtain its logarithm

$$\ln P(x_1,\ldots,x_l;\mu,\mathscr{D})=-\frac{nl}{2}\ln 2\pi+\frac{l}{2}\ln|\mathscr{D}|-\frac{1}{2}\sum_{i=1}^{l}(x_i-\mu)^{\mathrm{T}}\mathscr{D}(x_i-\mu).$$

Write

$$\frac{\partial P(x_1,\ldots,x_l;\mu,\mathscr{D})}{\partial\mu}=\mathscr{D}\left(\sum_{i=1}^{l}x_i-l\mu\right)=0, \tag{3.63}$$

$$\frac{\partial P(x_1,\ldots,x_l;\mu,\mathscr{D})}{\partial\mathscr{D}}=\frac{l}{2}\mathscr{D}^{-1}-\frac{1}{2}\sum_{i=1}^{l}(x_i-\mu)(x_i-\mu)^{\mathrm{T}}=0. \tag{3.64}$$

Here we have used the relationship

$$\frac{d\ln|A|}{dA}=A^{-1}.$$

From Equations (3.63) and (3.64) we obtain

$$x_{\mathrm{emp}}=\frac{1}{l}\sum_{i=1}^{l}x_i,\qquad S=\mathscr{D}^{-1}=\frac{1}{l}\sum_{i=1}^{l}(x_i-x_{\mathrm{emp}})(x_i-x_{\mathrm{emp}})^{\mathrm{T}}.$$

The estimator of the covariance matrix is biased.

§14 Remarks on Various Methods for Density Estimation

Three types of estimation for densities defined up to parameters were considered in this chapter: Bayesian, best unbiased, and those obtained using the maximum-likelihood method. For our specific problems of estimating densities of two classes (3.58) and (3.59), all three estimators were obtained. Which one is preferable for use in practice, then—which one should be substituted into (3.7) to obtain decision rules in a pattern recognition problem?

Theoretically the Bayesian is undoubtedly the preferable one. This estimator optimizes a functional which defines the quality of the estimator

in a reasonable manner. However, in order to obtain a Bayes estimator the prior distribution of parameters of the density must be known, i.e., a distribution which determines how often in practice a particular density is estimated. Usually this information is not available.

In Sections 6 and 7 Bayesian estimators were obtained for prior distributions which on the one hand contain fairly indefinite information but on the other yield a maximal simplification of calculations. How much confidence should be given to a Bayesian estimator based on one prior distribution if in practice another distribution is implemented? Only a qualitative answer is available to this question. As the sample size increases, the effect of the prior information on the Bayesian estimator decreases. Thus the use of the Bayesian estimator is justified by the belief that in practice the inconsistency in the choice of a prior distribution has little effect.

When constructing the best unbiased estimator of a density there is no need to take prior information into account. In this class of estimators there exists a best estimator which is independent of a particular estimated density belonging to this class. It would seem that no risk is involved in choosing the best unbiased estimator in such a situation. Actually this is not the case. It does not follow at all that the class of unbiased estimators contains sufficiently "nice" estimators. It has already been mentioned that the unbiasedness by itself is of no value and is introduced only to restrict the class of estimators. The class of unbiased estimators is a narrow one (for example, an unbiased estimator of the normal distribution expressible in terms of sufficient statistics is unique). It is not excluded that the narrow class of unbiased estimators consists only of rather "inferior" estimators and then the choice of the best one in this class does not assure that the estimator is satisfactory.

The example suggested by C. Stein indicates that this indeed is quite possible: when estimating the mean vector μ of the n-dimensional ($n > 2$) normal distribution with unit covariance matrix I, the biased estimator

$$\hat{x}_{\text{emp}} = \left(1 - \frac{n-2}{l x_{\text{emp}}^{\mathrm{T}} x_{\text{emp}}}\right) x_{\text{emp}}$$

turns out to be a uniformly better estimator than the arithmetic mean

$$x_{\text{emp}} = \frac{1}{l} \sum_{i=1}^{l} x_i,$$

which is the best unbiased one. (More details on Stein-type estimators are given in Chapter 5.) Stein's example is remarkable in that it is constructed for the simplest problems of parameter estimation and even here uniformly better biased estimators exist.

Thus the choice of the best unbiased estimator can be justified only by the belief that the class of unbiased estimators contains an adequate one.

Finally, the theory of maximum-likelihood estimators provides no answers to the question concerning the properties of estimators for samples

of finite size. The theory only guarantees that the maximum-likelihood estimators approach the efficient ones as the sample size increases, i.e., with an increase in sample size, the quality of a maximum-likelihood estimator approaches that of the best unbiased estimator.

Due to a lucky contingency, we were able in this chapter to find Bayesian estimators explicitly, i.e., to carry out the analytic integration of a multiple integral (numerical integration of multiple integrals of high dimensions is troublesome) to obtain explicitly the best unbiased estimator of the density. That is, we were able to arrive at an analytic solution of Fredholm's Type I equation (whereas a numerical solution of this equation is an ill-posed problem). This result is due to a specific feature of the parametric class of densities.

In general, however, such approximations can hardly be anticipated. In this respect the maximum-likelihood method has an advantage in that it can be used for diverse classes of densities. This property of the maximum-likelihood method is due to the fact that it reduces to the solution of algebraic equations, i.e., to a problem for which efficient computer methods exist.

Yet another remark: The methods for estimating densities discussed in this chapter make sense only if the density under consideration belongs to a given parametric family of densities. In practice, however, the prior information which would allow us to select a parametric family of functions containing the unknown one is not available. It turns out, in fact, that not only the choice of a particular method of density estimation, but also the choice of a *parametric* formulation of the problem of estimating dependences from empirical data, is largely a matter of belief.