Chapter 4

Methods of Parametric Statistics for the Problem of Regression Estimation

§1 The Scheme for Interpreting the Results of Direct Experiments

In the preceding chapter methods of parametric statistics were applied to solve the pattern recognition problem: to minimize the functional

$$I(\alpha) = \int (y - F(x, \alpha))^2 P(x, y)\, dx\, dy \tag{4.1}$$

with unknown density $P(x, y)$, on the basis of empirical data

$$x_1, y_1; \ldots ; x_l, y_l, \tag{4.2}$$

first the density $\hat{P}(x, y)$ was estimated in the parametric class of densities $\{P(x, y)\}$; then, using $\hat{P}(x, y)$, the empirical functional

$$I_{\text{emp}}(\alpha) = \int (y - F(x, \alpha))^2 \hat{P}(x, y)\, dx\, dy \tag{4.3}$$

was constructed; and finally a value α_{emp} was determined which minimizes (4.3).

To implement this process it was essential that the coordinate y take on only two values, zero and one; the set $F(x, \alpha)$ was a set of indicator functions, and the density $P(x, y)$ was a union of two densities. These were characteristic features of the pattern recognition problem. In this chapter we shall implement the same procedure of risk minimization, but in relation to the problem of regression estimation.

For a solution of this problem using methods of parametric statistics a specific model of density which differs from the one discussed in Chapter 3 is

V. Vapnik, *Estimation of Dependences Based on Empirical Data*
© Springer Science+Business Media, Inc. 2006

used. It is assumed that the random variable y and a random vector x are related as follows:

$$y = F(x, \alpha_0) + \xi,$$

where $F(x, \alpha_0)$ is a function which belongs to the class $F(x, \alpha)$ and ξ is a noise independent of x distributed according to the density $P(\xi)$:

$$M\xi = 0, \qquad M\xi^2 < \infty.$$

Thus for any fixed x the distribution $P(\xi)$ induces the conditional density of y,

$$P(y\,|\,x) = P(y - F(x, \alpha_0)). \tag{4.4}$$

The joint density $P(x, y)$ is defined by

$$P(x, y) = P(y\,|\,x)P(x) = P(y - F(x, \alpha_0))P(x), \tag{4.5}$$

where $P(x)$ is the probability density of the vector x.

The problem of regression estimation, $F(x, \alpha_0) \in F(x, \alpha)$, based on a random and independent sample of pairs $x_1, y_1, \ldots, x_l, y_l$, can be interpreted as the estimation of the functional dependence $F(x, \alpha_0)$ in the class $F(x, \alpha)$ based on direct observations which are carried out subject to additive noise at l randomly chosen points. In Chapter 1 this problem was called "interpretation of results of direct experiments".

We shall solve this problem using methods of parametric statistics. First we estimate the density

$$\hat{P}(y\,|\,x) = \hat{P}(y - F(x, \alpha^*)),$$

and then we obtain the minimum point for the empirical functional

$$I_{\text{emp}}(\alpha) = \int (y - F(x, \alpha))^2 \hat{P}(y - F(x, \alpha^*))P(x)\,dx\,dy. \tag{4.6}$$

First we show that the minimum of the functional (4.6) is attained at $\alpha = \alpha^*$. We utilize the identity

$$\begin{aligned} I_{\text{emp}}(\alpha) &= \int (y - F(x, \alpha))^2 \hat{P}(y - F(x, \alpha^*))P(x)\,dx\,dy \\ &= \int (y - F(x, \alpha^*))^2 \hat{P}(y - F(x, \alpha^*))P(x)\,dx\,dy \\ &\quad + \int (F(x, \alpha) - F(x, \alpha^*))^2 P(x)\,dx. \end{aligned} \tag{4.7}$$

Since the first summand on the right-hand side does not depend on α, the minimum of $I_{\text{emp}}(\alpha)$ is attained if the second nonnegative summand vanishes, i.e., at $\alpha = \alpha^*$. Thus the value of the vector $\alpha = \alpha^*$ which defines the conditional density $\hat{P}(y\,|\,x) = \hat{P}(y - F(x, \alpha^*))$ immediately determines the regression. It equals $F(x, \alpha^*)$.

§2 A Remark on the Statement of the Problem of Interpreting the Results of Direct Experiments

In the statement of the problem of interpreting results of direct experiments it is required that the unknown function $F(x, \alpha_0)$ belong to a given parametric family $F(x, \alpha)$. This requirement is imposed because the density $P(y - F(x, \alpha))$ is to be estimated by methods of parametric statistics. However, another formulation is possible according to which the unknown density $P(x, y)$ belongs to a given parametric family of densities $P(x, y; \alpha)$ and the desired dependence $F(x, \alpha_0)$ does not belong to the given set of dependences $f(x, \beta)$. In other words, as the model for interpreting results of direct experiments the following problem may be posed: find the minimum of the functional

$$I(\beta) = \int (y - f(x, \beta))^2 P(y - F(x, \alpha_0))P(x)\, dy\, dx \tag{4.8}$$

from the sample

$$x_1, y_1; \ldots ; x_l, y_l$$

if the joint density $P(x, y) = P(y - F(x, \alpha_0))P(x)$ is unknown, $F(x, \alpha_0) \in F(x, \alpha)$, and the set of functions $f(x, \beta)$ does not necessarily coincide with $F(x, \alpha)$. If $F(x, \alpha_0) \notin f(x, \beta)$, the minimum of the functional (4.8) is attained at a function belonging to $f(x, \beta)$ which is closest to $F(x, \alpha_0)$. The proximity is measured here in the L_P^2 sense:

$$\rho_L(F, f) = \left(\int (F(x, \alpha_0) - f(x, \beta))^2 P(x)\, dx \right)^{1/2}.$$

If however $F(x, \alpha_0) \in f(x, \beta)$, then the minimum coincides with the regression. (This fact also follows immediately from (4.7).) Thus the regression yields an absolute minimum for the functional (4.8).

For a known density $P(x)$ the solution of the minimization problem for the functional (4.8) may also be carried out by means of the methods of parametric statistics: based on sample (4.2), the density $\hat{P}(y - F(x, \alpha))$ is estimated and then the empirical functional

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

is minimized.

Observe that for the problem of pattern recognition the search for a conditional minimum (in the class $f(x, \beta)$) of a functional, rather than the absolute one, was the subject matter of discriminant analysis. As it was pointed out in Section 2 of Chapter 3, the *raison d'être* for this formulation was based on the fact that the sample size is finite and hence the density is estimated only approximately; thus the lower guaranteed minimum for the value of the expected risk can be obtained for a function belonging to a narrower class. An analogous situation arises for the interpretation of results of direct experiments based on finite samples: due to imprecisions in density estimation, the higher guaranteed proximity to regression may be attained at a function belonging to a narrower class $f(x, \beta)$. Methods for contracting classes of desired dependences in order to achieve a lower guaranteed expected risk will be discussed in Chapter 8.

§3 Density Models

Thus in order to estimate regression—under the conditions of the model for interpreting the results of direct experiments—it is sufficient to estimate the density $P(y - F(x, \alpha_0))$ defined up to the value of parameter α. In view of the stipulated model, the parametric family of densities $P(y - F(x, \alpha))$ which contains the desired one is determined firstly by the given parametric family of functions $F(x, \alpha)$ containing the regression $F(x, \alpha_0)$, and secondly by the known probability density for the noise $P(\xi)$.

The assignment of a class of functions $F(x, \alpha)$ containing the regression is an informal step in the formulation of the problem. This class should be assigned *a priori.*

As far as the probability density of errors is concerned, here the choice is in principle arbitrary. However, in the practice of direct experimentation certain typical situations arise connected with common mechanisms which yield observational errors. These mechanisms have been investigated. The following three probability densities are of importance for interpreting results of direct experiments: the uniform density, normal density, and Laplace density.

The *uniform probability density* given by

$$P(\xi) = \begin{cases} \dfrac{1}{2\Delta} & \text{for } |\xi| \leq \Delta, \\ 0 & \text{for } |\xi| > \Delta \end{cases}$$

is used for roundoff errors. For example, let a value of a certain large quantity x be measured up to its integer value. Then the error ξ which arises from the roundoff to the closest integer is often assumed to be distributed according to the distribution

$$P(\xi) = \begin{cases} 1 & \text{for } |\xi| \leq 0.5, \\ 0 & \text{for } |\xi| > 0.5. \end{cases}$$

The *Normal* (*Gaussian*) *density* given by

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

is used to describe errors arising when repeated physical measurements are performed under identical conditions. These conditions determine the value of the variance σ^2. For example, errors resulting in measuring distances by means of a theodolite carried out under the same conditions (the same illumination, humidity, air temperature, degree of atmospheric pollution, etc.) are commonly described by the normal density.

The *Laplace density* given by

$$P(\xi) = \frac{1}{2\Delta} \exp\left\{ -\frac{|\xi|}{\Delta} \right\}$$

is used to describe errors occurring in physical experiments carried out under changing conditions. For example, if measurements of distances take place in unequal cloudiness, at different times, and under different pollution conditions, measurement errors are commonly described by a Laplace distribution.

Each density $P(\xi)$ generates its own parametric set of densities

$$P(y - F(x, \alpha)).$$

In this chapter only the maximum-likelihood method will be used for estimating the density in various parametric families. This method is chosen because its implementation presents no technical difficulties. It is well suited to all the parametric families of densities under consideration.

Thus we shall use the method of maximum likelihood for estimating parameters of the conditional density

$$P(y|x) = P(y - F(x, \alpha_0))$$

from the random independent sample

$$x_1, y_1; \ldots; x_l, y_l$$

distributed according to the density

$$P(x, y) = P(y - F(x, \alpha_0))P(x).$$

For this purpose we write the likelihood function

$$P(x_1, y_1, \ldots, x_l, y_l; \alpha) = \prod_{i=1}^{l} P(y_i - F(x_i, \alpha))P(x_i), \tag{4.9}$$

and then express it as a product of two factors:

$$P_1(\alpha) = \prod_{i=1}^{l} P(y_i - F(x_i, \alpha)), \tag{4.10}$$

which is the likelihood function for the conditional density, and

$$P_2 = \prod_{i=1}^{l} P(x_i).$$

Since the factor P_2 does not depend on the parameter α, (4.9) and (4.10) have the same maximum points. In what follows, the maximization of the function (4.10) will also be called a method of maximum likelihood.

We shall now consider the likelihood function $P_1(\alpha)$ for different distributions of the noise and find the corresponding maximum point.

The likelihood function (4.10) for the uniform distribution of ξ is of the form

$$P_1(\Delta, \alpha) = \prod_{i=1}^{l} \frac{1}{2\Delta} \delta_i(\alpha) = \frac{1}{(2\Delta)^l} \prod_{i=1}^{l} \delta_i(\alpha),$$

where

$$\delta_i(\alpha) = \begin{cases} 1 & \text{for } |y_i - F(x_i, \alpha)| \le \Delta, \\ 0 & \text{for } |y_i - F(x_i, \alpha)| > \Delta. \end{cases}$$

The maximum of the likelihood function is determined by α and Δ for which the minimum of the expression

$$\Delta(\alpha) = \max_{x_i y_i} |y_i - F(x_i, \alpha)| \tag{4.11}$$

is attained, i.e., α is chosen to minimize the largest deviation of $F(x_i, \alpha)$ from y_i.

For the normal density the distribution of the likelihood function is given by the density

$$P_1(\alpha, \sigma) = \frac{1}{(2\pi)^{l/2}\sigma^l} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{l} (y_i - F(x_i, \alpha))^2 \right\},$$

and the maximum-likelihood method is equivalent to the minimization of the functional

$$I_{\text{emp}}(\alpha) = \sum_{i=1}^{l} (y_i - F(x_i, \alpha))^2. \tag{4.12}$$

The method of determining α by means of minimization of functional (4.12) is called the *least-squares method.*

Finally, if the error is distributed according to the Laplace density, then the corresponding likelihood function is

$$P_1(\Delta, \alpha) = \frac{1}{(2\Delta)^l} \exp\left\{ -\frac{1}{\Delta} \sum_{i=1}^{l} |y_i - F(x_i, \alpha)| \right\},$$

and the maximum of the likelihood is attained for the vector α for which the functional

$$I_{\text{emp}}(\alpha) = \sum_{i=1}^{l} |y_i - F(x_i \alpha)| \tag{4.13}$$

is minimized. The method of minimizing the functional (4.13) is called *the method of minimal modules.*

As was indicated in Chapter 3, the method of maximum likelihood is an asymptotically efficient method of estimating parameters; therefore all three algorithms are optimal in a certain sense. Unfortunately each one of them is optimal only under its own conditions (of uniform, normal, or Laplace distributions of errors), and solutions obtained by means of these algorithms may differ significantly.

Indeed, consider the simplest problem of estimating dependences—the determination of the mean value of a random variable y from a sample of size l. This problem is reduced to minimization of the functional

$$I(\alpha) = \int (y - \alpha)^2 P(y)\, dy \tag{4.14}$$

on the basis of a sample $y_1, \ldots, y_l$. Using the method of minimization of the largest deviation (4.11), we obtain the solution

$$\alpha^* = \frac{y_{\min} + y_{\max}}{2}, \tag{4.15}$$

where $y_{\min}$ is the smallest and $y_{\max}$ is the largest sample value; i.e., the estimator is the half range of the sample. The method of least squares (4.12) yields the estimator

$$\alpha^* = \frac{1}{l} \sum_{i=1}^{l} y_i; \tag{4.16}$$

i.e., the estimator is the sample arithmetic mean. Finally, the method of minimal modules (4.13) leads us to the following solution: order the observations according to their magnitude,

$$y_{i_1}, \ldots, y_{i_l},$$

and compute the estimator using the formula

$$\alpha^* = \begin{cases} y_{i_{k+1}} & \text{for } l = 2k + 1, \\ \dfrac{y_{i_k} + y_{i_{k+1}}}{2} & \text{for } l = 2k. \end{cases}$$

§4 Extremal Properties of Gaussian and Laplace Distributions

In the preceding section it was shown that algorithms for estimating regression obtained by methods of parametric statistics depend on the model adopted for the errors. Therefore it is necessary to be able to identify situations in which particular models are to be used. It was pointed out that the uniform distribution is used for describing errors resulting from rounding off, Gaussian distributions for measurement errors under identical conditions, and the Laplace law for measurements under changing conditions. It would be desirable to make these recommendations more precise.

In this section we shall establish certain remarkable properties for the Gaussian and Laplace distributions. We shall see that the Gaussian distribution possesses certain extremal properties under the absolute stability of measuring conditions, while the Laplace distribution possesses analogous extremal properties under "maximal instability" of measuring conditions.

Thus we shall show that among all continuous densities with a given variance, the normal distribution possesses the largest entropy. In other words, the normal distribution is a "noise" distribution in which the size of the measurement is undetermined to the largest possible extent.

We shall estimate the degree of uncertainty of measurements, in the case when errors are determined by the probability density $P(\xi)$, by means of

Shannon's entropy

$$H(P) = -\int_{-\infty}^{\infty} P(\xi) \ln P(\xi)\, d\xi. \tag{4.17}$$

We shall obtain a function $P(\xi)$ obeying the restrictions

$$P(\xi) \geq 0, \tag{4.18}$$

$$\int_{-\infty}^{\infty} P(\xi)\, d\xi = 1, \tag{4.19}$$

$$\int_{-\infty}^{\infty} \xi P(\xi)\, d\xi = 0, \tag{4.20}$$

$$\int_{-\infty}^{\infty} \xi^2 P(\xi)\, d\xi = \sigma^2, \tag{4.21}$$

for which the maximum of the entropy (4.17) is attained. Here the conditions (4.18), (4.19) follow from the definition of the density, (4.20) reflects the unbiasedness of the error, and (4.21) fixes the class of densities of a given variance.

This problem is solved using the standard method of Lagrange multipliers to take the conditions (4.19)–(4.21) into account:

$$L = -(P(\xi) \ln P(\xi) + \lambda_1 P(\xi) + \lambda_2 \xi P(\xi) + \lambda_3 \xi^2 P(\xi)).$$

We then write the Euler condition

$$\frac{\partial L}{\partial P} = -(\ln P(\xi) + 1 + \lambda_1 + \lambda_2 \xi + \lambda_3 \xi^2) = 0. \tag{4.22}$$

The solution of Equation (4.22),

$$P(x) = \exp\{-(\lambda_1 + 1 + \xi\lambda_2 + \xi^2\lambda_3)\},$$

satisfies (4.18) and hence determines the desired density.

To obtain values of the constants λ_1, λ_2, and λ_3 the conditions (4.19)–(4.21) are utilized; we obtain

$$P(\xi) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{\xi^2}{2\sigma^2}\right\}, \tag{4.23}$$

thus the normal density has the largest entropy among all densities with a given variance (i.e., the random variable has the most "uncertain" distribution).

Consider now a somewhat more complicated model for the error term ξ. The value of random variable ξ is a realization of the normal distribution $P_N(\xi | \sigma^2)$ with mean 0 and variance σ^2. However, each time the normal distribution has its own variance. The value of the variance is assigned randomly and independently according to the density $P(\sigma^2)$. Thus we have the distribution

$$P_\Lambda(x) = \int P_N(\xi \,|\, \sigma^2) P(\sigma^2)\, d\sigma^2. \tag{4.24}$$

This model reflects well the practical situation when under fixed conditions of measurements the normal distribution is valid. However, the measurement conditions change randomly and independently, and thus the probability density is a composition of two densities. In the example of measuring distances the factor $P_N(x \,|\, \sigma^2)$ in (4.24) reflects the errors occurring under the same atmospheric conditions. The factor $P(\sigma^2)$ reflects the random nature of the atmospheric conditions. If the measurement conditions do not change (the extreme case when $P(\sigma^2) = \delta(\sigma^2 - \sigma_0^2)$ where $\delta(z)$ is the delta function). then the composition (4.24) defines a normal distribution. We, however, shall consider the other extreme case when the experimental conditions deviate from the mean in the "most uncertain manner", i.e., when the function $P(\sigma^2)$ is such that the maximum of the entropy

$$H(P) = -\int_0^\infty P(\sigma^2) \ln P(\sigma^2)\, d\sigma^2 \tag{4.25}$$

is attained and moreover the conditions

$$P(\sigma^2) \geq 0, \tag{4.26}$$

$$\int P(\sigma^2)\, d\sigma^2 = 1, \tag{4.27}$$

$$\int_0^\infty \sigma^2 P(\sigma^2)\, d\sigma^2 = 2\Delta^2 \tag{4.28}$$

are satisfied. The conditions (4.26) and (4.27) follow from the definition of the probability density. The restriction (4.28) determines the average conditions of conducting the experiment.

We thus derive the maximum of the entropy (4.25) under the conditions (4.26)–(4.28). Writing the corresponding Lagrange function—which takes (4.27) and (4.28) into account

$$L = -(P(\sigma^2) \ln P(\sigma^2) + \lambda_1 P(\sigma^2) + \lambda_2 \sigma^2 P(\sigma^2)),$$

we obtain the Euler equation

$$\frac{\partial L}{\partial P} = -(\ln P(\sigma^2) + 1 + \lambda_1 + \lambda_2 \sigma^2) = 0. \tag{4.29}$$

The solution of Equation (4.29) is

$$P(\sigma^2) = \exp\{-(\lambda_1 + 1 + \lambda_2 \sigma^2)\}$$

which satisfies (4.26) and thus determines the desired density. To find the values of constants λ_1 and λ_2 we substitute solution (4.29) into (4.27) and

(4.28), whence $\lambda_1 + 1 = -\ln 2\Delta^2$ and $\lambda_2 = 1/2\Delta^2$. Thus the "most uncertain" conditions for conducting the experiment are given by density

$$P(\sigma^2) = \frac{1}{2\Delta^2} \exp\left\{-\frac{\sigma^2}{2\Delta^2}\right\}. \tag{4.30}$$

We shall show that the probability density $P_\Delta(\xi)$ given as a composition of densities (4.23) and (4.30) is a Laplace distribution, i.e.,

$$\begin{aligned} P_\Delta(\xi) &= \frac{1}{\sqrt{2\pi}2\Delta^2} \int_0^\infty \frac{1}{\sigma} \exp\left\{-\frac{\xi^2}{2\sigma^2}\right\} \exp\left\{-\frac{\sigma^2}{2\Delta^2}\right\} d\sigma^2 \\ &= \frac{1}{2\Delta} \exp\left\{-\frac{|\xi|}{\Delta}\right\}. \end{aligned} \tag{4.31}$$

In order to compute the integral (4.31) we shall use the following fact, which is valid for any integrable function on $(-\infty, \infty)$:

$$\int_0^\infty f\left[\left(\frac{x}{a} - \frac{b}{x}\right)^2\right] dx = a \int_0^\infty f(y^2)\, dy \qquad (a, b > 0). \tag{4.32}$$

To prove this identity set $y = \dfrac{x}{a} - \dfrac{b}{x}$. Then

$$\begin{aligned} \int_{-\infty}^\infty f(y^2)\, dy &= \int_0^\infty f\left[\left(\frac{x}{a} - \frac{b}{x}\right)^2\right]\left(\frac{1}{a} + \frac{b}{x^2}\right) dx \\ &= \frac{1}{a} \int_0^\infty f\left[\left(\frac{x}{a} - \frac{b}{x}\right)^2\right] dx + b \int_0^\infty f\left[\left(\frac{x}{a} - \frac{b}{x}\right)^2\right] \frac{dx}{x^2}. \end{aligned}$$

Substituting the variable $x = -ab/t$ in the last integral, we arrive at

$$\frac{1}{a} \int_{-\infty}^0 f\left[\left(\frac{t}{a} - \frac{b}{t}\right)^2\right] dt.$$

Thus

$$\int_{-\infty}^\infty f(y^2)\, dy = \frac{1}{a} \int_{-\infty}^\infty f\left[\left(\frac{x}{a} - \frac{b}{x}\right)^2\right] dx.$$

Hence (since the integrand is even) we obtain the identity (4.32).

We now transform the left-hand side of (4.31):

$$\begin{aligned} P_\Delta(\xi) &= \frac{1}{2\sqrt{2\pi\Delta^2}} \int_0^\infty \frac{1}{\sigma} \exp\left\{-\left(\frac{\sigma^2}{2\Delta^2} + \frac{\xi^2}{2\sigma^2}\right)\right\} d\sigma^2 \\ &= \frac{1}{\sqrt{2\pi\Delta^2}} \exp\left\{-\frac{|\xi|}{\Delta}\right\} \int_0^\infty \exp\left\{-\frac{1}{2}\left(\frac{\sigma}{\Delta} - \frac{|\xi|}{\sigma}\right)^2\right\} d\sigma. \end{aligned} \tag{4.33}$$

From (4.33) in view of (4.32) we obtain

$$P_\Delta(\xi) = \frac{1}{\Delta}\exp\left\{-\frac{|\xi|}{\Delta}\right\}\int_0^\infty \frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{y^2}{2}\right\}dy$$

$$= \frac{1}{2\Delta}\exp\left\{-\frac{|\xi|}{\Delta}\right\}. \tag{4.34}$$

In other words the composition (4.31) of a normal distribution and distribution (4.30) results in Laplace density (4.34).

Thus we have shown that under fixed conditions of conducting an experiment the most undetermined (uncertain) result is obtained when the error is normally distributed; if however the conditions of the experiment oscillate around some mean value in the most unfavorable manner, then the most undetermined measurement result is obtained when the error is distributed according to the Laplace law. Thus the choice between a Gaussian and a Laplace law depends on whether the conditions of the experiment are perfectly stable or most unstable.

In practice, however, these two extreme cases seldom occur. Therefore neither Gaussian nor Laplace distributions are usually fulfilled. It is customary to assume that an "intermediate" situation is valid.

Thus we are confronted with a situation where regression is estimated under the assumption that some hypothetical distribution for the error is valid (e.g., Gaussian or Laplace) while actually some other "intermediate" distribution is the correct one. How useful will the algorithms given by (4.11)–(4.13) then be? In other words, to what extent are the algorithms constructed robust as far as the changes in the distribution of errors are concerned, and how should one construct robust algorithms? The answer is given in the succeeding sections.

§5 On Robust Methods of Estimating Location Parameters

Let the probability density of the error be unknown. Suppose it is only known that it belongs to a certain given set of densities $\{P(\xi)\}$. Below we shall define such sets more precisely; for the time being we merely assume that they are convex and that the density functions possess two continuous derivatives and are symmetric around zero. (The symmetry is the basic requirement for the theory discussed below.) The following problem will now be investigated. How should one choose the hypothetical density for the noise from the given class $\{P(\xi)\}$ in order that the possible error shall have the least effect on the estimators of regression parameters if it is known that the true density belongs to $\{P(\xi)\}$?

First consider the simple case: it is required to estimate the mathematical expectation m of a random variable x on the basis of the sample $x_1, \dots, x_l$. If the mathematical expectation m exists the problem is equivalent to estimating the location parameter m of the density $P(x - m)$ (here we utilize the fact that the noise ξ is related to the measurement x by $\xi = x - m$). For a known density $P(\xi)$ the estimator $\hat{m}$ of location parameter m is carried out by the maximum-likelihood method, i.e., by maximizing the expression

$$R(m) = \sum_{i=1}^{l} \ln P(x_i - m). \tag{4.35}$$

In this case the estimator $\hat{m}$ is consistent and asymptotically efficient. However, if the function $P(\cdot)$ in (4.35) does not coincide with the density function of the noise $P(\xi)$, estimator $\hat{m}$ yielding the maximum of (4.35) will in general be neither consistent nor asymptotically efficient.

Denote the value $\hat{m}$ maximizing (4.35) under the assumption that $P(\xi) = P_\Gamma(\xi)$ by $\hat{m} = m(x_1, \dots, x_l; P_\Gamma(\xi))$. We shall now determine how to measure the accuracy of parameter estimation if it is assumed that the noise is distributed according to the distribution $P_\Gamma(\xi) \in \{P(\xi)\}$ while actually the true distribution is $P_0(\xi) \in \{P(\xi)\}$.

It is natural to use the quantity

$$R(P_\Gamma(\xi); x_1, \dots, x_l) = (\hat{m}(x_1, \dots, x_l; P_\Gamma(\xi)) - m)^2$$

as the accuracy of the estimator $\hat{m}$ based on a sample $x_1, \dots, x_l$, assuming that the noise is distributed according to the distribution $P_\Gamma(\xi)$. (This quantity is the square of the deviation of the obtained value of the parameter from the true one.) The accuracy of estimating a location parameter based on samples of size l is naturally measured by the mathematical expectation of the quantity $R(P_\Gamma(\xi); x_1, \dots, x_l)$:

$$\begin{aligned} D(P_0, P_\Gamma) &= MR(P_\Gamma(\xi); x_1, \dots, x_l) \\ &= \int (\hat{m}(x_1, \dots, x_l; P_\Gamma(\xi)) - m)^2 P_0(x_1 - m) \cdots \\ &\quad \times P_0(x_l - m)\, dx_1 \cdots dx_l. \end{aligned} \tag{4.36}$$

The quantity $D(P_0, P_\Gamma)$ depends on two probability densities belonging to the same class $\{P(\xi)\}$: the hypothetical density $P_\Gamma(\xi)$ (according to which the estimator $\hat{m}$ was constructed) and the true density $P_0(\xi)$ (according to which the mean square deviation was computed).

Below we shall utilize the following representation of the function $D(P_0, P_\Gamma)$:

$$D(P_0, P_\Gamma) = \frac{1}{l}\frac{\displaystyle\int\left(\frac{P'_\Gamma(\xi)}{P_\Gamma(\xi)}\right)^2 P_0(\xi)\,d\xi}{\left(\displaystyle\int\left(\frac{P'_\Gamma(\xi)}{P_\Gamma(\xi)}\right)' P_0(\xi)\,d\xi\right)^2}$$

$$= \frac{1}{l}\frac{\displaystyle\int\left(\frac{P'_\Gamma(\xi)}{P_\Gamma(\xi)}\right)^2 P_0(\xi)\,d\xi}{\left(\displaystyle\int\frac{P'_\Gamma(\xi)P'_0(\xi)}{P_\Gamma(\xi)}\,d\xi\right)^2}. \tag{4.37}$$

We shall verify this representation by carrying out a not quite rigorous but intuitively appealing argument. A rigorous theory of robust estimation is presented in [88].

Without loss of generality it may be assumed that the true value of the location parameter m is zero. Denote

$$f(\xi) = \frac{P'_\Gamma(\xi)}{P_\Gamma(\xi)} = (\ln P_\Gamma(\xi))'.$$

Then using the maximum-likelihood method, the estimator $\hat{m}$ of the parameter $m = 0$ is obtained from the condition

$$\left(\sum_{i=1}^{l} \ln P_\Gamma(x_i - \hat{m})\right)' = \sum_{i=1}^{l} f(x_i - \hat{m}) = 0.$$

We now utilize an approximation which is valid for large l and for the symmetric densities considered herein:

$$\sum_{i=1}^{l} f(x_i - \hat{m}) \approx \sum_{i=1}^{l} f(x_i) - \hat{m}\sum_{i=1}^{l} f'(x_i) = 0, \tag{4.38}$$

hence

$$\hat{m} = \frac{\sum_{i=1}^{l} f(x_i)}{\sum_{i=1}^{l} f'(x_i)}.$$

Let l be so large that

$$\hat{m} \approx \frac{\frac{1}{l}\sum_{i=1}^{l} f(x_i)}{\int f'(x)P_0(x)\,dx}.$$

(In the derivation of this relation it was assumed that the integral in the denominator exists. For this purpose it is sufficient that the functions $f'(x)$ be bounded. Below we shall consider only densities satisfying $|(\ln P(\xi))''| < \text{const}$).

Compute now $D(P_0, P_\Gamma) = M\hat{m}^2$:

$$D(P_0, P_\Gamma) = \int \hat{m}^2 P_0(x_1), \ldots, P_0(x_l)\, dx_1, \ldots, dx_l$$

$$= \frac{1}{l^2} \frac{1}{\left[\int f'(x)P_0(x)\, dx\right]^2} \int \sum_{i,j}^{l} f(x_i)f(x_j)$$

$$\times P_0(x_1) \cdots P_0(x_l)\, dx_1 \cdots dx_l.$$

Since the densities $P_0(x)$, $P_\Gamma(x)$ are symmetric, we have

$$\int f(x_i)f(x_j)P_0(x_1) \cdots P_0(x_l)\, dx_1 \cdots dx_l = 0, \qquad i \neq j.$$

Thus we obtain for large l

$$D(P_0, P_\Gamma) = \frac{1}{l^2} \frac{\int \sum_{i=1}^{l} f^2(x_i)P_0(x_i)\, dx_i}{\left(\int f'(x)P_0(x)\, dx\right)^2} = \frac{1}{l} \frac{\int f^2(x)P_0(x)\, dx}{\left(\int f'(x)P_0(x)\, dx\right)^2}.$$

Finally, returning to the original notation we obtain representation (4.37).

We have thus determined a criterion of quality for estimators of location parameters given that the true density is $P_0(\xi)$ and the hypothetized one is $P_\Gamma(\xi)$. Our goal now is to choose a density $P_\Gamma(\xi)$ which minimizes $D(P_0, P_\Gamma)$. It is easy to show (see below) that if the density $P_0(\xi)$ were known, the minimum of $D(P_0, P_\Gamma)$ would be obtained at $P_\Gamma(\xi) = P_0(\xi)$.

The problem is to choose $P_\Gamma(\xi)$ if it is known only that $P_0(\xi)$ belongs to the class $\{P(\xi)\}$. As usual in such situations one of two approaches—the Bayesian or the minimax—is taken.

In the first case, it is assumed that the probability for each density in $\{P(\xi)\}$ to be the true one is known *a priori*, and the measure of quality of estimators is chosen to be the average (with respect to the measure $\mu(P)$) quality, i.e.,

$$D_{\mathrm{B}}(P_\Gamma) = \int D(P_0, P_\Gamma)\, d\mu(P_0).$$

The minimax approach suggests that we choose as a measure of quality the quantity $D(P_0, P_\Gamma)$ evaluated for the least favorable density $P_0(\xi) \in \{P(\xi)\}$, i.e., to evaluate the quality from the condition

$$D_{\mathrm{mnx}}(P_\Gamma) = \max_{P_0} D(P_0, P_\Gamma).$$

Since the construction of a solution optimal in the Bayes sense encounters substantial difficulties here, we shall study only minimax solutions below. Thus we shall judge the quality of an estimator of a location parameter, obtained by means of the hypothetized density $P_\Gamma(\xi)$, by the quantity

$$D_{\mathrm{mnx}}(P_\Gamma) = \max_{P_0} D(P_0, P_\Gamma) = \max_{P_0} \frac{\displaystyle\int \left(\frac{P'_\Gamma(\xi)}{P_\Gamma(\xi)}\right)^2 P_0(\xi)\, d\xi}{\displaystyle l\left(\int \frac{P'_\Gamma(\xi)P'_0(\xi)}{P_\Gamma(\xi)}\, d\xi\right)^2}, \qquad (4.39)$$

and attempt to obtain a hypothetical density $P_\Gamma(\xi)$ minimizing (4.39).

Such a statement of the problem yields a game-theoretic interpretation. Let there be two players—nature and a statistician—who possess the same set of strategies (functions $\{P(\xi)\}$) but opposite goals. The first player (nature) attempts to select a strategy (i.e., assign a true density $P_0(\xi)$) which will maximize the losses of the second player, while the second chooses a strategy (hypothetized density $P_\Gamma(\xi)$) which minimizes his loss. The amount of loss is determined by the functional (4.39).

It is required to obtain the optimal strategy for the second player, i.e., to be able, for a given class of densities, to choose a hypothetized density that will guarantee the minimum losses for the least favorable true density. The density obtained will be called *robust in the class* $\{P(\xi)\}$, and the method of estimation of a location parameter obtained by applying the maximum-likelihood method to the density obtained is called the *method of robust estimation of a location parameter.*

An important fact in the theory of robust estimation of a location parameter is that the game with the loss function (4.39) possesses on the convex set $\{P(\xi)\}$ a saddle point, i.e.,

$$\max_{P_0 \in \{P(\xi)\}} \min_{P_\Gamma \in \{P(\xi)\}} D(P_0, P_\Gamma) = \min_{P_\Gamma \{P(\xi)\}} \max_{P_0 \in \{P(\xi)\}} D(P_0, P_\Gamma).$$

Using this fact one can obtain an optimal strategy against nature.

We now utilize the Cauchy–Schwarz inequality

$$\left(\int a(x)b(x)\, d\mu(x)\right)^2 \le \int a^2(x)\, d\mu(x) \int b^2(x)\, d\mu(x). \qquad (4.40)$$

Using this inequality we rearrange the denominator of (4.37):

$$D(P_0, P_\Gamma) = \frac{1}{l} \frac{\displaystyle\int \left(\frac{P'_\Gamma(\xi)}{P_\Gamma(\xi)}\right)^2 P_0(\xi)\, d\xi}{\displaystyle\left(\int \left(\frac{P'_\Gamma(\xi)}{P_\Gamma(\xi)} \frac{P'_0(\xi)}{P_0(\xi)}\right) P_0(\xi)\, d\xi\right)^2} \ge \frac{1}{\displaystyle l \int \left(\frac{P'_0(\xi)}{P_0(\xi)}\right)^2 P_0(\xi)\, d\xi}. \qquad (4.41)$$

Observe that for $P_\Gamma(\xi) = P_0(\xi)$ the equality

$$D(P_0, P_0) = \frac{1}{l \int \left(\frac{P_0'(\xi)}{P_0(\xi)}\right)^2 P_0(\xi)\, d\xi} \tag{4.42}$$

is valid. It follows from (4.41) and (4.42) that the minimum of (4.39) is attained if $P_\Gamma(\xi) = P_0(\xi)$, i.e., the optimal strategies of nature and the statistician result in the same density. To obtain this density it is necessary to maximize (4.42) over the class $\{P(\xi)\}$ or equivalently to obtain in the class $\{P(\xi)\}$ a density which will minimize the functional

$$I_\Phi(P) = l \int \left(\frac{P'(\xi)}{P(\xi)}\right)^2 P(\xi)\, d\xi.$$

Observe that the functional $I_\Phi(P)$ is the Fisher information quantity (cf. Chapter 3, Section 11).

In Sections 7 and 8 we shall obtain for various classes of probability densities those which minimize the Fisher information quantity and thus find robust estimators (within these classes) of a location parameter. In the next section we shall extend the result obtained here to the case of estimating regression parameters.

§6 Robust Estimation of Regression Parameters

Let it be required to estimate the regression. We shall assume that the class of functions in which the estimation is carried out and to which the regression belongs is represented in the form

$$F(x, \alpha) = \sum_{r=1}^{n} \alpha_r \varphi_r(x),$$

where $\varphi_r(x)$ is a system of linearly independent functions. As above, the true and the hypothetized densities of errors $P_0(\xi)$ and $P_\Gamma(\xi)$ belong to the convex class $\{P(\xi)\}$. The densities are symmetric around zero and have a bounded second logarithmic derivative.

To estimate regression parameters we shall use the maximum-likelihood method, i.e., we shall obtain the vector α which maximizes the expression

$$\ln P_\Gamma(x_1, y_1; \ldots; x_l, y_l; \alpha) = \sum_{i=1}^{l} \ln P_\Gamma\left(y_i - \sum_{r=1}^{n} \alpha_r \varphi_r(x_i)\right). \tag{4.43}$$

Let this vector be $\alpha = \alpha^*$. Consider the vector of deviations of the obtained values of regression parameters α^* from the actual ones α_0:

$$\bar{\alpha} = (\alpha_0 - \alpha^*).$$

Form the covariance matrix B:

$$B = M\bar{\alpha}\cdot\bar{\alpha}^{\mathrm{T}},$$

which determines the quality of estimation of the vector of parameters α (cf. Chapter 3, Section 11).

Below, analogously to (4.37), we shall obtain that for l sufficiently large the equality†

$$B = \frac{1}{l}\frac{\displaystyle\int\left(\frac{P_\Gamma'(\xi)}{P_\Gamma(\xi)}\right)^2 P_0(\xi)\,d\xi}{\left(\displaystyle\int\left(\frac{P_\Gamma'(\xi)}{P_\Gamma(\xi)}\right)' P_0(\xi)\,d\xi\right)^2}\,\|k_{ij}\|^{-1} \tag{4.44}$$

is valid, where

$$k_{ij} = \frac{1}{l}\sum_{t=1}^{l}\varphi_i(x_t)\varphi_j(x_t) \approx \int \varphi_i(x)\varphi_j(x)P_0(x)\,dx.$$

Thus the elements of matrix B are proportional to

$$D(P_0, P_\Gamma) = \frac{1}{l}\frac{\displaystyle\int\left(\frac{P_\Gamma'(\xi)}{P_\Gamma(\xi)}\right)^2 P_0(\xi)\,d\xi}{\left(\displaystyle\int\left(\frac{P_\Gamma'(\xi)}{P_\Gamma(\xi)}\right)' P_0(\xi)\,d\xi\right)^2}.$$

In the representation (4.44) it is important that only the proportionality coefficient $D(P_0, P_\Gamma)$ (and not the matrix $\|k_{ij}\|$) depends on the densities $P_0(\xi)$ and $P_\Gamma(\xi)$. Therefore two quadratic forms $z^{\mathrm{T}}B_1 z$ and $z^{\mathrm{T}}B_2 z$ with the same matrix $\|k_{ij}\|$ but different values of $D(P_0, P_\Gamma)$ correspond to two different hypothetized densities $P_\Gamma(\xi)$ and $\hat{P}_\Gamma(\xi)$. These forms satisfy one of two relations: either

$$z^{\mathrm{T}}B_1 z \geq z^{\mathrm{T}}B_2 z \quad \text{for any } z$$

or

$$z^{\mathrm{T}}B_1 z < z^{\mathrm{T}}B_2 z \quad \text{for any } z,$$

depending on whether $D(P_0, P_\Gamma)$ or $D(P_0, \hat{P}_\Gamma)$ is the largest. It was shown in Section 11 of Chapter 3 that the minimum of the quadratic form $z^{\mathrm{T}}Bz$ defines jointly efficient estimators of the parameters. Thus the value of the coefficient $D(P_0, P_\Gamma)$ determines the quality of estimation of the parameters of a linear regression: the smaller $D(P_0, P_\Gamma)$ is, the better is the quality.

This means that in the case of estimating regression parameters the problem of choosing a robust density leads to a game between nature and the statistician. It was shown in the preceding section that in this game the optimal strategy for the statistician is to choose a density belonging to the

† We assume additionally that the matrix $\|k_{ij}\|$ is not singular.

class of densities $\{P(\xi)\}$ which yields the minimum of Fisher's information quantity

$$I_\Phi(P) = l \int \left(\frac{P'(\xi)}{P(\xi)}\right)^2 P(\xi)\, d\xi. \tag{4.45}$$

Thus, in order to obtain the best hypothetical model for the error in the class $\{P(\xi)\}$ it is necessary to find a function belonging to this class which minimizes (4.45). This density will be used for the determination of regression parameters using the maximum-likelihood method.

It remains to derive the relation (4.44). It is obtained analogously to (4.37). Denote $f(\xi) = P'_\Gamma(\xi)/P_\Gamma(\xi)$. Then the maximum of the likelihood function (4.43) is attained at values of α which satisfy the equations

$$\sum_{i=1}^{l} f\left(\xi_i - \sum_{r=1}^{n} \bar{\alpha}_r \varphi_r(x_i)\right)\varphi_k(x_i) = 0, \qquad k = 1, 2, \ldots, n.$$

Utilizing the approximation (4.38), we have

$$\sum_{i=1}^{l} f\left(\xi_i - \sum_{r=1}^{n} \bar{\alpha}_r \varphi_r(x_i)\right)\varphi_k(x_i)$$
$$\approx \sum_{i=1}^{l} \left[f(\xi_i) - f'(\xi_i) \sum_{r=1}^{n} \bar{\alpha}_r \varphi_r(x_i)\right]\varphi_k(x_i) = 0.$$

Due to the independence of ξ_i and x_i we then obtain, for l sufficiently large,

$$\frac{1}{l}\sum_{i=1}^{l} f(\xi_i)\varphi_k(x_i) - \int f'(\xi)P_0(\xi)\, d\xi \sum_{i=1}^{l}\left(\sum_{r=1}^{n} \bar{\alpha}_r \varphi_r(x_i)\right)\varphi_k(x_i) = 0,$$
$$k = 1, 2, \ldots, n,$$

or in vector form,

$$\|k_{ij}\|\bar{\alpha} \approx \frac{1}{l}\frac{H}{\int f^i(\xi)P_0(\xi)\, d\xi}, \tag{4.46}$$

where H is a column vector with coordinates $h_r = \sum_{i=1}^{l} \varphi_r(x_i) f(\xi_i)$.

It follows from (4.46) that

$$\bar{\alpha} = \frac{1}{l}\frac{1}{\int f'(\xi)P_0(\xi)\, d\xi}\|k_{ij}\|^{-1}H.$$

We now obtain the covariance matrix

$$B = M\bar{\alpha}\bar{\alpha}^{\mathrm{T}} = \frac{1}{l}\frac{\int f^2(\xi_i)P_0(\xi)\, d\xi}{\left(\int f'(\xi)P_0(\xi)\, d\xi\right)^2}\|k_{ij}\|^{-1}.$$

Returning to the original notation, we arrive at (4.44).

§7 Robustness of Gaussian and Laplace Distributions

We shall show that Gaussian and Laplace distributions are robust, each in its own class. As was shown in the preceding section, it is sufficient for this purpose to show that in corresponding classes of densities $\{P(\xi)\}$ the Gaussian and Laplace distributions yield the minimum of Fisher's information quantity (4.45).

For specific classes $\{P(\xi)\}$ which are discussed below this problem becomes a difficult problem in the calculus of variations (the class $\{P(\xi)\}$ is defined by restrictions of the inequality type). Therefore we shall not obtain the hypothetical density here by using a regular method, i.e., by solving nonclassical variational problems, but rather we shall first identify these solutions and then verify that they indeed define a saddle point of the function

$$D(P, P_\Gamma) = \frac{1}{l}\,\frac{\displaystyle\int\left(\frac{P'_\Gamma(\xi)}{P_\Gamma(\xi)}\right)^2 P(\xi)\,d\xi}{\left(\displaystyle\int\left(\frac{P'_\Gamma(\xi)}{P_\Gamma(\xi)}\right)' P(\xi)\,d\xi\right)^2}.$$

In other words it will be required to verify that for a given density $P_\Gamma(\xi)$ the inequalities

$$D(P, P_\Gamma) \le D(P_\Gamma, P_\Gamma) \le D(P_\Gamma, P)$$

are fulfilled. Observe that in view of (4.41) one of the inequalities, namely

$$D(P_\Gamma, P_\Gamma) \le D(P_\Gamma, P)$$

is always valid. Thus in order to prove the optimality of the selected strategy it is sufficient to establish the validity of the inequality

$$D(P, P_\Gamma) \le D(P_\Gamma, P_\Gamma). \tag{4.47}$$

We consider the following classes of densities.

(1) *The class of densities with a bounded variance.* The corresponding variational problem is to minimize the functional (4.45) in the class of functions satisfying the conditions

$$\begin{array}{ll}
(1) & P(\xi) > 0, \\
(2) & \displaystyle\int P(\xi)\,d\xi = 1, \\
(3) & \displaystyle\int \xi P(\xi)\,d\xi = 0, \\
(4) & \displaystyle\int \xi^2 P(\xi)\,d\xi \le \sigma^2.
\end{array} \tag{4.48}$$

Conditions (1), (2), and (3) determine the density of the error term, and condition (4) is a bound on the variance. The solution of this nonclassical problem (in view of (1) and (4)) of the calculus of variations is the density

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

Indeed, substituting

$$P_\Gamma(\xi) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left\{ -\frac{\xi^2}{2\sigma^2} \right\}$$

into the inequality (4.47), we obtain

$$\frac{\displaystyle\int \frac{\xi^2}{\sigma^4} P(\xi)\, d\xi}{\left(\displaystyle\frac{1}{\sigma^2} \int P(\xi)\, d\xi\right)^2} = \int \xi^2 P(\xi)\, d\xi \le \sigma^2. \tag{4.49}$$

This inequality is valid for any density belonging to (4.48), since the class (4.48) consists of densities for which the variance does not exceed σ^2. Thus the normal probability density with zero mean and variance σ^2 is robust in the class of all densities with the variance bounded by σ^2.

2. Now consider the *class of nondegenerate at zero densities.* Densities for which $P(0) \ge 1/2\Delta$ belong to this class. We shall show that the Laplace distribution is robust in this class of densities. For this purpose we substitute

$$P_\Gamma(\xi) = \frac{1}{2\Delta} \exp\left\{ -\frac{|\xi|}{\Delta} \right\}$$

into (4.47). We obtain

$$\frac{\displaystyle\int \left(\frac{\operatorname{sign} \xi}{\Delta}\right)^2 P(\xi)\, d\xi}{\displaystyle\frac{4}{\Delta^2} P^2(0)} = \frac{1}{4P^2(0)} \le \Delta^2,$$

or equivalently

$$P(0) \ge \frac{1}{2\Delta}.$$

And since the densities satisfying $P(0) \ge 1/2\Delta$ are included in the class $\{P(\xi)\}$, the inequality (4.47) is satisfied for any function belonging to this class. Thus the Laplace distribution is robust in the class of densities for which $P(0) \ge 1/2\Delta$.

The robustness of the Gaussian and Laplace densities (each in its own class) is no less remarkable a fact than their extremal properties verified in Section 4.

Although the Gaussian and Laplace densities are robust, the class in which this property is valid often turns out to be exceedingly wide. In such cases a more meaningful statistical model should be constructed on the basis of other, narrower classes of densities.

Below in Sections 8 and 9 we shall consider certain specific classes of densities and obtain robust densities for these classes.

§8 Classes of Densities Formed by a Mixture of Densities

Consider the class H of densities formed by the mixture

$$P(\xi) = g(\xi)(1 - \varepsilon) + \varepsilon h(\xi) \tag{4.50}$$

of a certain fixed density $g(\xi)$ symmetric with respect to the origin and an arbitrary density $h(\xi)$ symmetric with respect to the origin. The weights in the mixture are $1 - \varepsilon$ and ε respectively. For classes of these densities the following theorem is valid.

Theorem 4.1 (Huber). *Let $-\ln g(\xi)$ be a twice continuously differentiable convex function. Then the class H possesses a robust density*

$$P_\Gamma(\xi) = \begin{cases} (1 - \varepsilon)g(\xi_0)\exp\{k(\xi - \xi_0)\}, & \text{for } \xi < \xi_0, \\ (1 - \varepsilon)g(\xi), & \text{for } \xi_0 \le \xi < \xi_1, \\ (1 - \varepsilon)g(\xi_1)\exp\{-k(\xi - \xi_1)\}, & \text{for } \xi \ge \xi_1, \end{cases} \tag{4.51}$$

where ξ_0 and ξ_1 are the end points of the interval $[\xi_0, \xi_1]$ on which a monotone (due to the convexity of $-\ln g(\xi)$) function $g'(\xi)/g(\xi)$ is bounded in absolute value by a constant k determined by the normalization condition

$$1 = (1 - \varepsilon)\int_{\xi_0}^{\xi_1} g(\xi)\,d\xi + \frac{g(\xi_0) + g(\xi_1)}{k}(1 - \varepsilon).$$

Proof. To prove this theorem it is required to show (as in the case of proving robustness of Gaussian and Laplace densities) that functions belonging to the class (4.50) satisfy

$$D(P, P_\Gamma) \le D(P_\Gamma, P_\Gamma) \le D(P_\Gamma, P).$$

As has already been mentioned, the validity of the bound

$$D(P_\Gamma, P_\Gamma) \le D(P_\Gamma, P),$$

follows from the Cauchy–Schwarz inequality (4.40). Therefore to prove the theorem it is sufficient to verify that

$$D(P, P_\Gamma) \le D(P_\Gamma, P_\Gamma)$$

for any function $P(\xi) \in H$.

We represent the density $P_\Gamma(\xi)$ in the form of a mixture of a fixed density $g(\xi)$ and the density $\hat{h}(\xi) = [P_\Gamma(\xi) - (1 - \varepsilon)g(\xi)]/\varepsilon$. We shall write the density $\hat{h}(\xi)$ explicitly taking (4.51) into account:

$$\hat{h}(\xi) = \begin{cases} \dfrac{1-\varepsilon}{\varepsilon}(g(\xi_0)\exp\{k(\xi - \xi_0)\} - g(\xi)) & \text{for } \ \xi < \xi_0, \\ 0 & \text{for } \xi_0 \le \xi < \xi_1, \\ \dfrac{1-\varepsilon}{\varepsilon}(g(\xi_1)\exp\{-k(\xi - \xi_1)\} - g(\xi)) & \text{for } \ \xi \ge \xi_1. \end{cases} \tag{4.52}$$

It is easy to verify that $\hat{h}(\xi)$ is a density. Indeed, $\int \hat{h}(\xi)\,d\xi = 1$, and $\hat{h}(\xi) \ge 0$, since by the assumption of the theorem $-\ln g(\xi)$ is a convex function and hence is situated totally above the tangent:

$$-\ln g(\xi) \ge -\ln g(\xi_i) - (-1)^i k(\xi - \xi_i), \qquad i = 0, 1. \tag{4.53}$$

This inequality is equivalent to the assertion

$$g(\xi) \le g(\xi_i)\exp\{(-1)^i k(\xi - \xi_i)\}, \qquad i = 0, 1.$$

Consider the inequality

$$\frac{\displaystyle\int \left(\frac{P'_\Gamma(\xi)}{P_\Gamma(\xi)}\right)^2 [(1-\varepsilon)g(\xi) + \varepsilon h(\xi)]\,d\xi}{\displaystyle\left(\int \left(\frac{P'_\Gamma(\xi)}{P_\Gamma(\xi)}\right)' [(1-\varepsilon)g(\xi) + \varepsilon h(\xi)]\,d\xi\right)^2} \le \frac{\displaystyle(1-\varepsilon)\int \left(\frac{P'_\Gamma(\xi)}{P_\Gamma(\xi)}\right)^2 g(\xi)\,d\xi + \varepsilon k^2}{\displaystyle(1-\varepsilon)^2\left(\int \left(\frac{P'_\Gamma(\xi)}{P_\Gamma(\xi)}\right)' g(\xi)\,d\xi\right)^2}. \tag{4.54}$$

We shall verify that the right-hand side of this inequality is the least upper bound for the expression appearing in the left-hand side for arbitrary symmetric densities $h(\xi)$. For this purpose we observe that the function $P'_\Gamma(\xi)/P_\Gamma(\xi)$ equals

$$\frac{P'_\Gamma(\xi)}{P_L(\xi)} = \begin{cases} k & \text{for } \ \xi < \xi_0, \\ \dfrac{g'(\xi)}{g(\xi)} & \text{for } \xi_0 \le \xi < \xi_1, \\ -k & \text{for } \ \xi \ge \xi_1, \end{cases}$$

where according to the condition of the theorem $|g'(\xi)/g(\xi)| \le k$, and the function $(P'_\Gamma(\xi)/P_\Gamma(\xi))'$ equals

$$\left(\frac{P'_\Gamma(\xi)}{P_\Gamma(\xi)}\right)' = \begin{cases} 0 & \text{for } \ \xi < \xi_0, \\ \left(\dfrac{g'(\xi)}{g(\xi)}\right)' & \text{for } \xi_0 \le \xi < \xi_1, \\ 0 & \text{for } \ \xi \ge \xi_1. \end{cases}$$

Thus in order to maximize the left-hand side of the inequality it is necessary to choose a density $h(\xi)$ which is situated on the intervals $(-\infty, \xi_0)$ and (ξ_1, ∞). Such a density simultaneously maximizes the numerator and minimizes the denominator of the expression appearing on the left-hand side of the inequality. The value of the expression appearing on the left will then be exactly equal to the value of the right-hand side of the inequality. The density (4.52) indeed belongs to the class of densities concentrated on the intervals $(-\infty, \xi_0)$, (ξ_1, ∞). The theorem is proved. □

This theorem is remarkable in that it allows us to construct various robust densities. In particular, if we choose for $g(\xi)$ the normal density

$$g(\xi) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left\{-\frac{\xi^2}{2\sigma^2}\right\},$$

and consider the class of densities

$$P(\xi) = \frac{1-\varepsilon}{\sqrt{2\pi}\,\sigma} \exp\left\{-\frac{\xi^2}{2\sigma^2}\right\} + \varepsilon h(\xi),$$

then in view of the theorem the density

$$P_{\Gamma}(\xi) = \begin{cases} \dfrac{1-\varepsilon}{\sqrt{2\pi}\,\sigma} \exp\left\{\dfrac{k^2}{2} - \dfrac{k}{\sigma}|\xi|\right\} & \text{for } |\xi| \geq k\sigma, \\ \dfrac{1-\varepsilon}{\sqrt{2\pi}\,\sigma} \exp\left\{-\dfrac{\xi^2}{2\sigma^2}\right\} & \text{for } |\xi| < k\sigma \end{cases}$$

will be robust in this class, where k is determined from the normalization condition

$$1 = \frac{1-\varepsilon}{\sqrt{2\pi}\,\sigma}\left[\int_{-k\sigma}^{k\sigma} \exp\left\{-\frac{\xi^2}{2\sigma^2}\right\} d\xi + \frac{2\exp\left\{-\frac{k^2}{2}\right\}}{k}\right].$$

The density just derived is an intermediate density between Gaussian and Laplace distributions. On the interval $|\xi| < k\sigma$ it coincides up to a normalizing constant with the Gaussian distribution and on the intervals $|\xi| \geq k\sigma$ with the Laplace distribution.

§9 Densities Concentrated on an Interval

We now consider yet another important class of densities and obtain a robust probability density in it.

Consider the class K_p *of densities concentrated on the whole on the interval* $[-A, A]$, i.e., the class of densities $P(\xi)$ for which the condition

$$\int_{-A}^{A} P(\xi)\, d\xi \geq 1 - p$$

is satisfied (where p is a known parameter which defines the class K_p). We shall show that in this class the density

$$P_\Gamma(\xi) = \begin{cases} \dfrac{1}{A}\left(\dfrac{b}{1+b}\cos^2\dfrac{a\xi}{A}\right) & \text{for } \left|\dfrac{\xi}{A}\right| < 1, \\ \dfrac{1}{A}\left(\dfrac{b}{1+b}\cos^2 a\right)\exp\left\{-2b\left(\left|\dfrac{\xi}{A}\right| - 1\right)\right\} & \text{for } \left|\dfrac{\xi}{A}\right| \geq 1 \end{cases} \tag{4.55}$$

is robust, where the parameters a, b are related to the constant p—which determines the class K—by the relations

$$p = 1 - \frac{\cos^2 a}{1+b},$$

$$b = a \tan a, \qquad 0 < a < \frac{\pi}{2}. \tag{4.56}$$

Without loss of generality it will be assumed that $A = 1$ (the class $A \neq 1$ is reduced to the case $A = 1$ by the substitution $z = A\xi$). Thus the problem is to show that in the class of densities satisfying the condition

$$\int_{-1}^{1} P(\xi)\,d\xi \geq 1 - p,$$

the density

$$P_\Gamma(\xi) = \begin{cases} \dfrac{b}{1+b}\cos^2 \xi\alpha & \text{for } |\xi| < 1, \\ \dfrac{b}{1+b}\cos^2 a \exp\{-2b(|\xi| - 1)\} & \text{for } |\xi| \geq 1 \end{cases} \tag{4.57}$$

will be robust. To do this it is sufficient to show that $P_\Gamma(\xi)$ given by (4.57) minimizes in K_p the Fisher functional

$$I_\Phi = l \int \left(\frac{P'(\xi)}{P(\xi)}\right)^2 P(\xi)\,d\xi. \tag{4.58}$$

Instead of directly minimizing the functional (4.58), however, we shall utilize the fact that the necessary and sufficient condition for $P_\Gamma(\xi)$ to be the minimum point for (4.58) is that the functional

$$R(P_\Gamma, P) = l \int (2(-\ln P_\Gamma(\xi))'' - [(\ln P_\Gamma(\xi))']^2)(P(\xi) - P_\Gamma(\xi))\,d\xi \tag{4.59}$$

is nonnegative in K_p. The functional $R(P_\Gamma, P)$ is the derivative with respect to ε of the expression

$$I_\Phi((1 - \varepsilon)P_\Gamma(\xi) + \varepsilon P(\xi)),$$

evaluated at $\varepsilon = 0$, i.e.,

$$\left.\frac{dI_\Phi((1-\varepsilon)P_\Gamma(\xi) + \varepsilon P(\xi))}{d\varepsilon}\right|_{\varepsilon=0} = R(P_\Gamma, P). \tag{4.60}$$

The nonnegativity of derivatives at $\varepsilon = 0$ (in any direction in K_p) for densities $(1-\varepsilon)P_\Gamma(\xi) + \varepsilon P(\xi)$ means that the minimum of I_Φ is attained on $P_\Gamma(\xi)$.

Thus we shall verify that the expression $R(P_\Gamma, P)$ is nonnegative. Since the function under the integral $R(P_\Gamma, P)$ is even, it is sufficient to verify that it is positive on the ray $0 \le \xi < \infty$. First note that (4.57) implies that

$$(-\ln P_\Gamma(\xi))' = \begin{cases} 2a \tan a\xi & \text{for } |\xi| < 1, \\ 2b \operatorname{sign} \xi & \text{for } |\xi| \ge 1. \end{cases} \tag{4.61}$$

Substituting (4.61) into (4.69) and carrying out the calculations, we have

$$R(P_\Gamma, P) = 4a^2 l \int_0^1 (P(\xi) - P_\Gamma(\xi))\, d\xi - 4b^2 l \int_1^\infty (P(\xi) - P_\Gamma(\xi))\, d\xi. \tag{4.62}$$

Transforming (4.62), we have

$$\begin{aligned} R(P_\Gamma, P) &= 4a^2 l \int_0^1 (P(\xi) - P_\Gamma(\xi))\, d\xi - 4b^2 l \int_1^\infty (P(\xi) - P_\Gamma(\xi))\, d\xi \\ &= 4(a^2 + b^2) l \int_0^1 (P(\xi) - P_\Gamma(\xi))\, d\xi. \end{aligned}$$

Thus the expression $R(P_\Gamma, P)$ is nonnegative for all $P(\xi)$ such that

$$\int_{-1}^1 P(\xi)\, d\xi \ge \int_{-1}^1 P_\Gamma(\xi)\, d\xi = 1 - 2\int_1^\infty P_\Gamma(\xi)\, d\xi = 1 - p,$$

i.e., for all functions belonging to K_p.

§10 Robust Methods for Regression Estimation

In preceding sections we have considered several classes of densities and obtained robust densities in these classes. It will now be possible in our scheme for interpreting results of direct experiments to weaken the requirements on prior information concerning the statistical properties of the errors. It is sufficient to know the class of densities to which the errors belong. In this case for estimating parameters of regression using methods of parametric statistics it is possible to use—instead of a true density—a density which is robust in the given class. Obviously this replacement reduces the asymptotic rate of convergence of parameters of the regression. This rate becomes proportional to some quantity I situated in the interval

$$I_{\min} \le I \le I_{\max},$$

where

$$I_{\max} = \sup_{P(\xi) \in \{P(\xi)\}} \frac{1}{l \int \left(\frac{P'(\xi)}{P(\xi)} \right)^2 P(\xi)\, d\xi},$$

instead of being proportional to

$$I_{\min} = \frac{1}{l \int \left(\frac{P_0'(\xi)}{P_0(\xi)} \right)^2 P_0(\xi)\, d\xi},$$

which is the limiting value attainable in the case of unbiased estimation of the location parameter (cf. Chapter 3, Section 11) where $P_0(\xi)$ is the true density of the error. However, if the class $\{P(\xi)\}$ of densities is not too wide, then the possible loss of the rate is not overly large.

The basic constructive result of the theory of robust estimation considered here is the determination of four classes of densities with specified robust density.† We again identify these classes and their densities:

(1) *The class of densities with variance bounded by a constant σ^2.* A robust density in this class is the normal density

$$P_\Gamma(\xi) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left\{ -\frac{\xi^2}{2\sigma^2} \right\}.$$

(2) *The class of nondegenerate densities* (for which $P(0) > 1/2\Delta$). In this class a robust density is

$$P_\Gamma(\xi) = \frac{1}{2\Delta} \exp\left\{ -\frac{|\xi|}{\Delta} \right\}.$$

(3) *The class of densities formed by a mixture of a known density* (for example, a normal $P_N(\xi) = (1/\sqrt{2\pi}\sigma)e^{-\varepsilon^2/2\sigma^2})$ with an arbitrary density in proportion $1 - \varepsilon : \varepsilon$. In this class the density

$$P_\Gamma(\xi) = \begin{cases} c \exp\left\{ -\dfrac{\xi^2}{2\sigma^2} \right\} & \text{for } |\xi| < k\sigma, \\[2ex] c \exp\left\{ \dfrac{k^2}{2} - k \left| \dfrac{\xi}{\sigma} \right| \right\} & \text{for } |\xi| \ge k\sigma \end{cases}$$

is robust (here c and k are constants determined by means of ε and σ).

† There are other classes of densities for which robust densities have been found [46].

(4) *The class of densities concentrated on the whole in the interval* $[-A, A]$ ($\int_{-A}^{A} P(\xi)\, d\xi \geq 1 - p$). A density

$$P_\Gamma(\xi) = \begin{cases} c\cos^2 \dfrac{a\xi}{A} & \text{for } \left|\dfrac{\xi}{A}\right| < 1, \\ c\cos^2 a \exp\left\{-2b\left(\left|\dfrac{\xi}{A}\right| - 1\right)\right\} & \text{for } \left|\dfrac{\xi}{A}\right| \geq 1, \end{cases}$$

where c, a, and b are constants determined via A and p, is robust in this class.

Now suppose instead of the true density for the error $P_0(\xi)$ we choose a robust one in the class $P_\Gamma(\xi)$; determine, by means of it, the density of the conditional probability distribution

$$P_\Gamma\left(y - \sum_{r=1}^{n} \alpha_r \varphi_r(x)\right);$$

and finally utilize the maximum-likelihood method for parameter estimation. Then we arrive at the following algorithm of regression estimation based on the sample

$$x_1, y_1; \ldots; x_l, y_l.$$

One should minimize the functional

$$I_{\text{emp}}(\alpha) = \sum_{i=1}^{l} d\left(y_i - \sum_{r=1}^{n} \alpha_r \varphi_r(x_i)\right),$$

where

$$d(z) = z^2,$$

provided the true density of the error belongs to the class of densities with a bounded variance;

$$d(z) = |z|,$$

provided the true density of the error belongs to the class of nondegenerate densities;

$$d(z) = \begin{cases} \dfrac{z^2}{2\sigma^2} & \text{for } |z| < k\sigma, \\ -\dfrac{k^2}{2} + \dfrac{k}{\sigma}|z| & \text{for } |z| \geq k\sigma, \end{cases}$$

provided the true density is a mixture of a normal density with an arbitrary one;

$$d(z) = \begin{cases} -2 \ln cos \dfrac{a}{A} z & \text{for } |z| < A, \\ b\left(\left|\dfrac{z}{a}\right| - 1\right) - 2 \ln \cos \dfrac{a}{A} z & \text{for } |z| \geq A, \end{cases}$$

provided the true density is concentrated on the whole on the interval $[-A, A]$.

Among these four methods, the least-square method ($d(z) = z^2$) and the method of minimal absolute values ($d(z) = |z|$) do not involve free parameters. The latter method is the most universal—it is determined by a stable density in a wider class of densities.

The other two methods of estimation involve parameters which are computed from the quantities defining the classes of densities. These methods should be used when possible to determine, as precisely as possible, the class of densities containing the desired one.

Thus when estimating regression we were able to remove the condition knowing exactly the error distribution. It is sufficient to know the class of functions which contains the regression and a class of densities to which the error density belongs. However, all of this theory developed for symmetric densities is essentially asymptotic (since in deriving the basic relation (4.37) the law of large numbers was substantially utilized). Therefore the belief that the asymptotic situation will occur rather early is the only guarantee that the algorithms obtained will be workable for samples of limited size.