

Chapter 2

A Decision-Theoretic Approach to Forecasting

Abstract Statistical forecasting is prediction of future states of a certain process based on the available stochastic observations as well as the available prior model assumptions made about this process. This chapter describes a general (universal) approach to statistical forecasting based on mathematical decision theory, including a brief discussion of discriminant analysis. The following fundamental notions are introduced: optimal and suboptimal forecasts, loss function, risk functional, minimax, admissible, and Bayesian decision rules (BDRs), Bayesian forecast density, decision rule randomization, plug-in principle.

2.1 The Mathematical Model of Decision Making

A generalized mathematical model of decision making has been formulated by Abraham Wald [14] as a generalization of the models used for hypothesis testing and parameter estimation to obtain an adequate description of settings that include stochastic processes. The high degree of uncertainty present in most applied forecasting problems makes the decision-making approach extremely relevant to statistical forecasting.

A general decision-making model contains two abstract objects: the environment (**E**) and the decision maker (**DM**), as well as the following six mathematical objects:

$$(\Theta, \mathcal{Y}, \mathcal{X}, w(\cdot), F(\cdot), D).$$

Here $\Theta \subseteq \mathbb{R}^m$ is the parameter space containing all possible states $\theta \in \Theta$ of the environment **E**, which includes a certain “actual state of **E**” denoted as $\theta^0 \in \Theta$ (this actual state is assumed to be unknown to the **DM** at the moment when the decision is made); $\mathcal{Y} \subseteq \mathbb{R}^m$ is the decision space (each element $Y \in \mathcal{Y}$ is a possible decision of the **DM**); $w(\cdot)$ is the loss function

$$w = w(\theta, Y), \quad \theta \in \Theta, \quad Y \in \mathcal{Y}, \quad w \in \mathbb{R}^1,$$

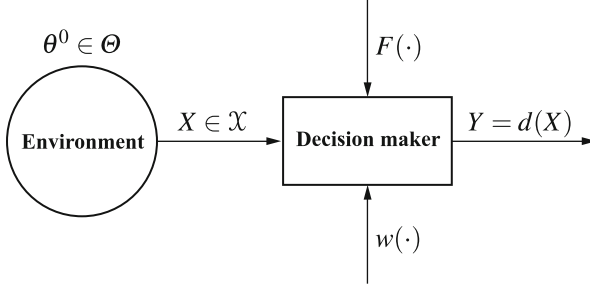


Fig. 2.1 The process of decision making

where w is the loss taken by the **DM** for $\theta^0 = \theta$ and the decision Y ; the function $u = u(\theta, Y) = -w(\theta, Y)$ is usually called the utility function; $\mathcal{X} = \mathcal{B}(\mathbb{R}^N)$ is the sample space (a Borel σ -algebra defined over an N -dimensional Euclidean space) where statistical data is observed; the random N -vector of observations $X \in \mathcal{X}$ is defined over the probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and $F(X; \theta^0) : \mathcal{X} \times \Theta \rightarrow [0, 1]$ is the N -dimensional distribution function of X which depends on the parameter θ^0 ; D is the decision rule space consisting of all Borel maps $d(\cdot) : \mathcal{X} \rightarrow \mathcal{Y}$:

$$D = \{Y = d(X) : X \in \mathcal{X}, Y \in \mathcal{Y}\}.$$

Decision making within this model is illustrated in Fig. 2.1. At the moment when the decision $Y \in \mathcal{Y}$ is being made, the **DM** doesn't know the "actual state of **E**" $\theta^0 \in \Theta$, and therefore the actual loss $w(\theta^0, Y)$ is also unknown. However, the **DM** knows the possible loss $w = w(\theta, Y)$ for every possible situation $(\theta, Y) \in \Theta \times \mathcal{Y}$. In order to reduce the uncertainty of θ^0 , the **DM** collects statistical data in the form of an observation $X \in \mathcal{X}$, which has the probability distribution defined by θ^0 . Based on the knowledge of the loss function $w(\cdot)$, the distribution function $F(\cdot)$, and the collected statistical data X , the **DM** uses a certain performance criterion to choose the optimal decision rule $d_{opt}(\cdot) \in D$ and to make the best possible decision $\hat{Y} = d_{opt}(X)$ by following this rule.

2.2 Minimax, Admissible, and Bayesian Families of Decision Rules

Under a generalized decision-making model presented in Sect. 2.1, consider the problem of constructing the optimal decision rule $d_{opt}(\cdot) \in D$. Let us define a criterion of decision rule optimality [10].

Definition 2.1. The conditional risk of a decision rule $d(\cdot) \in D$ for $\theta^0 = \theta$ is defined as the conditional expectation of the loss function:

$$r = r(d(\cdot); \theta) = \mathbb{E}_\theta\{w(\theta, d(X))\} = \int_{\mathbb{R}^N} w(\theta, d(X)) dF(X; \theta), \quad \theta \in \Theta, \quad r \in \mathbb{R}^1. \quad (2.1)$$

Smaller values of the functional (2.1) correspond to more effective decision rules. It follows from the definition that the uncertainty of the value θ^0 complicates the minimization of the risk functional.

Definition 2.2. A minimax decision rule $Y = d^*(X)$ is defined as a decision rule minimizing the supremum of the risk functional (2.1):

$$r_+(d^*(\cdot)) = \inf_{d(\cdot) \in D} r_+(d(\cdot)), \quad r_+(d(\cdot)) = \sup_{\theta \in \Theta} r(d(\cdot); \theta) \quad (2.2)$$

where $r_+(d(\cdot))$ is the guaranteed (upper) risk, i.e., the maximum possible value of the risk functional for the decision rule $d(\cdot)$.

The guaranteed risk corresponds to the least favorable state of the environment \mathbf{E} , and thus the minimax decision rule (2.2) is often called “pessimistic.”

Another popular approach to decision making is the Bayesian approach [10] which is based on the assumption that there exists an a priori known m -dimensional probability distribution function $G(\theta)$ of the random vector $\theta^0 \in \Theta$ defining the state of the environment \mathbf{E} .

Definition 2.3. Under the assumptions of the decision-making model defined earlier, let $\theta^0 \in \Theta \subseteq \mathbb{R}^m$ be a random m -vector characterized by a prior distribution function $G(\theta)$. Then the Bayesian (unconditional) decision risk is defined as the following functional:

$$r = r(d(\cdot)) = \mathbb{E}\{r(d(\cdot); \theta^0)\} = \int_{\mathbb{R}^m} r(d(\cdot); \theta) dG(\theta), \quad d(\cdot) \in D, \quad r \in \mathbb{R}^1, \quad (2.3)$$

or equivalently

$$r = r(d(\cdot)) = \mathbb{E}\{w(\theta^0, d(X))\} = \int_{\mathbb{R}^m} \int_{\mathbb{R}^N} w(\theta, d(X)) dF(X; \theta) dG(\theta),$$

which follows from (2.1) and the total expectation formula.

Definition 2.4. A Bayesian decision rule (BDR) is a decision rule $Y = d_0(X)$ that minimizes the Bayesian risk (2.3):

$$r(d_0(\cdot)) = \inf_{d(\cdot) \in D} r(d(\cdot)). \quad (2.4)$$

Let us introduce the last type of decision rules that will be discussed in this chapter—the admissible decision rules.

Definition 2.5. It is said that a decision rule $d'(\cdot)$ dominates a decision rule $d''(\cdot)$, where $d'(\cdot), d''(\cdot) \in D$, if

$$r(d'(\cdot); \theta) \leq r(d''(\cdot); \theta) \quad \forall \theta \in \Theta, \quad (2.5)$$

and there exists a $\theta \in \Theta$ such that the inequality in (2.5) is strict. A decision rule $\tilde{d}(\cdot)$ is said to be admissible if no other decision rule $d(\cdot) \in D$ dominates $\tilde{d}(\cdot)$.

Definition 2.6. Decision rules $d_1(\cdot), d_2(\cdot) \in D$ are said to be equivalent w.r.t. the Bayesian decision risk if they have the same Bayesian risk values:

$$r(d_1(\cdot)) = r(d_2(\cdot)).$$

Equivalence w.r.t. the guaranteed risk is defined similarly:

$$r_+(d_1(\cdot)) = r_+(d_2(\cdot)).$$

Let us establish some properties of the above decision rules (see [1] for a more systematic treatment).

Properties of Bayesian, Minimax, and Admissible Decision Rules

Property 2.1. A BDR $d_0(\cdot)$ minimizes the posterior mean loss $w(Y | X)$:

$$\hat{Y} = d_0(X) = \arg \min_{Y \in \mathcal{Y}} w(Y | X), \quad (2.6)$$

where

$$w(Y | X) = \mathbb{E}\{w(\theta^0, Y) | X\} = \int_{\mathbb{R}^m} w(\theta, Y) dG(\theta | X), \quad (2.7)$$

and $G(\theta | X)$ is the posterior probability distribution function of the random parameter θ^0 given the observation X .

Proof. Using (2.3) and the total expectation formula, let us rewrite the Bayesian risk as follows:

$$r(d(\cdot)) = \mathbb{E}\{w(\theta^0, d(X))\} = \mathbb{E}\{\mathbb{E}\{w(\theta^0, d(X)) | X\}\},$$

where the outer expectation is computed w.r.t. the unconditional distribution of the random vector X with a distribution function

$$F(X) = \int_{\mathbb{R}^m} F(X; \theta) dG(\theta);$$

the inner conditional expectation defines the posterior loss (2.7). Thus,

$$r(d(\cdot)) = \mathbb{E}\{w(d(X) | X)\} = \int_{\mathbb{R}^N} w(d(X) | X) dF(X) \geq r_0 ::= \int_{\mathbb{R}^N} \min_{Y \in \mathcal{Y}} w(Y | X) dF(X),$$

and it is obvious that the lower bound r_0 is attained for the decision rule defined by (2.6), (2.7). From Definition 2.4, this decision rule is a BDR. \square

Property 2.2. If a BDR $Y = d_0(X)$ is unique, it is also admissible.

Proof. The statement will be proved by contradiction. Suppose that the BDR $d_0(\cdot)$ is not admissible. Then, by Definition 2.5, there exists a decision rule $d'(\cdot) \in D$, $d'(\cdot) \neq d_0(\cdot)$, such that

$$r(d'(\cdot); \theta) \leq r(d_0(\cdot); \theta) \quad \forall \theta \in \Theta, \quad \exists \theta' \in \Theta : \quad r(d'(\cdot); \theta') < r(d_0(\cdot); \theta).$$

Integrating the first inequality over the probability distribution $G(\theta)$ of θ and applying (2.3) yields the inequality

$$r(d'(\cdot)) \leq r(d_0(\cdot)).$$

Strictness of this inequality would contradict the definition of a BDR given in (2.4). However, an equality is also impossible, since in that case $d'(\cdot)$ would be a different BDR, contradicting the uniqueness of the BDR. This contradiction concludes the proof. \square

Property 2.3. Given that the parameter space is finite, $\Theta = \{\theta^{(1)}, \dots, \theta^{(K)}\}$, $K < \infty$, and that the prior probability distribution of $\theta^0 \in \Theta$ is nonsingular,

$$p_k = \mathbb{P}\{\theta^0 = \theta^{(k)}\} > 0, \quad k = 1, \dots, K,$$

the BDR $Y = d_0(X)$ is admissible.

Proof. Assume the opposite: $d_0(\cdot)$ is not an admissible decision rule. Then (2.5) implies that there exist a decision rule $d(\cdot) \in D$ and an index $i^* \in \{1, \dots, K\}$ such that

$$r(d(\cdot); \theta_k) \leq r(d_0(\cdot); \theta_k), \quad k \neq i^*; \quad r(d(\cdot); \theta_{i^*}) < r(d_0(\cdot); \theta_{i^*}).$$

Multiplying both sides of these inequalities by $p_k > 0$ and $p_{i^*} > 0$, respectively, taking a sum, and applying the equality (2.3) yield

$$r(d(\cdot)) = \sum_{k=1}^K p_k r(d(\cdot); \theta^{(k)}) < \sum_{k=1}^K p_k r(d_0(\cdot); \theta^{(k)}) = r(d_0(\cdot)).$$

This inequality contradicts the definition of the BDR (2.4). \square

Property 2.4. If the parameter space is finite, $\Theta = \{\theta^{(1)}, \dots, \theta^{(K)}\}$, $K < \infty$, and $d(\cdot)$ is an admissible decision rule, then there exists a prior distribution

$$p_k = P\{\theta^{(0)} = \theta^{(k)}\}, \quad k = 1, \dots, K,$$

such that the decision rule $d(\cdot)$ is the BDR w.r.t. the prior distribution $\{p_k\}$. In other words, in this case the set of admissible decision rules is included in the set of BDRs.

Proof. The proof can be obtained by repeating the argument of the previous proof. \square

Property 2.5. If a minimax decision rule $d^*(\cdot)$ is unique, then it is also admissible.

Proof. Let us assume the opposite: there exists a different decision rule $d(\cdot) \in D$, $d(\cdot) \neq d^*(\cdot)$, such that

$$r(d(\cdot); \theta) \leq r(d^*(\cdot); \theta) \quad \forall \theta \in \Theta, \quad \exists \theta' \in \Theta : \quad r(d(\cdot); \theta') < r(d^*(\cdot); \theta').$$

From (2.2), this also yields the inequality

$$r_+(d(\cdot)) \leq r_+(d^*(\cdot)).$$

This contradicts the condition that $d^*(\cdot)$ is a unique minimax decision rule. \square

Property 2.6. Given that $d(\cdot)$ is an admissible decision rule and that the corresponding risk function (2.1) doesn't depend on $\theta \in \Theta$, i.e., $r(d(\cdot); \theta) = \text{const.}$, the decision rule $d(\cdot)$ is also a minimax decision rule.

Proof. Assume that the minimax condition isn't satisfied for the decision rule $d(\cdot)$ and that there exists a different minimax decision rule $d'(\cdot) \neq d(\cdot)$:

$$r_+(d'(\cdot)) < r_+(d(\cdot)).$$

However, we have also assumed that $r_+(d(\cdot)) \equiv r_+(d(\cdot); \theta)$, and thus

$$r(d'(\cdot); \theta) < r(d(\cdot); \theta) \quad \forall \theta \in \Theta,$$

which contradicts the admissibility of $d(\cdot)$. \square

2.3 The Bayesian Forecast Density

Randomization of the decision rule is a commonly used decision-theoretic technique of reducing the decision risk [1].

Definition 2.7. A *randomized decision rule* is a family of random variables

$$Y = d(X, \omega) : \mathcal{X} \times \Omega \rightarrow \mathcal{Y},$$

lying in the basic probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and defined by a *critical function* $\pi(Y; X)$. For a discrete decision space \mathcal{Y} , the function $\pi(Y; X)$ is defined as

$$\pi = \pi(Y; X) ::= \mathbb{P}\{d(X, \omega) = Y \mid X\}, \quad Y \in \mathcal{Y},$$

and we have

$$0 \leq \pi(Y; X) \leq 1, \quad \sum_{Y \in \mathcal{Y}} \pi(Y; X) \equiv 1.$$

In the continuous case, where $\mathcal{Y} \subset \mathbb{R}^M$, and the Lebesgue measure $\text{mes}_M(\mathcal{Y})$ is positive, $\pi = \pi(Y; X)$ is defined as the M -dimensional probability density of the random variable $d(X, \omega)$, and we have

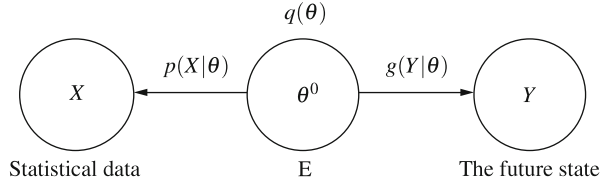
$$\pi(Y; X) \geq 0, \quad \int_{\mathbb{R}^M} \pi(Y; X) dY = 1.$$

Let us consider applications of randomized decision rules in statistical forecasting. Assume that a forecast is constructed for a random M -vector $Y \in \mathcal{Y}$ that describes an unknown future state of the process or the phenomenon that is being investigated. Its probability density $g(Y \mid \theta)$ depends on a parameter $\theta \in \Theta \subseteq \mathbb{R}^m$ with an unknown true value $\theta^0 \in \Theta$. Following the Bayesian paradigm, it is assumed that θ^0 is a random m -vector with a given prior probability density $q(\theta)$. Let $X \in \mathcal{X} \subseteq \mathbb{R}^N$ be statistical data describing past and current states of the process with a conditional probability density $p(X \mid \theta)$ given $\theta^0 = \theta$. Thus, the random parameter vector θ^0 is stochastically dependent not only on the past and current states X but also on the future states Y . This allows forecasting of Y based on the collected statistical data X under prior probabilistic uncertainty of θ^0 .

The problem of constructing a forecast for Y based on X using the randomized decision rule $\hat{Y} = d(X, \omega)$ lies in finding the critical function $\pi(Y; X)$. Following the Bayesian approach outlined above, one of the methods of constructing the critical function is the use of the posterior probability density of Y given the observation X :

$$\pi(Y; X) = p(Y \mid X), \quad Y \in \mathcal{Y}. \quad (2.8)$$

Fig. 2.2 Stochastic dependence between X , θ^0 , and Y



The conditional probability density (2.8) used in forecasting is called the *Bayesian forecast density*.

Following the accepted stochastic model (Fig. 2.2) of the dependence between X , θ^0 , and Y , Bayes formulae, together with certain well-known properties of multivariate probability densities, imply that

$$\pi(Y; X) = \int_{\mathbb{R}^m} g(Y | \theta) p(\theta | X) d\theta, \quad (2.9)$$

where

$$p(\theta | X) = p(X | \theta) q(\theta) \left(\int_{\mathbb{R}^m} p(X | \theta') q(\theta') d\theta' \right)^{-1} \quad (2.10)$$

is the posterior probability density of the random vector θ^0 given a fixed value of the random vector X .

The Bayesian forecast density (2.9), (2.10) allows us not only to compute the randomized forecast $\hat{Y} \in \mathcal{Y}$ as a result of simulating a random M -vector with the probability density $\pi(Y; X)$, $Y \in \mathcal{Y}$ but also to compute the traditional (nonrandomized) point and interval forecasts. Numerical characteristics of the Bayesian forecast density $\pi(Y; X)$ can be used as point forecasts of Y :

- Posterior expected forecast

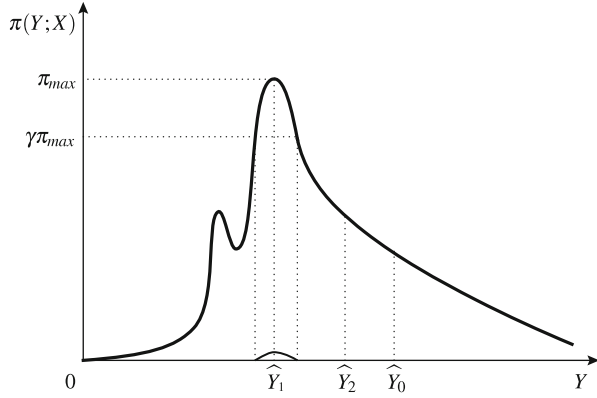
$$\hat{Y}_0 = \int_{\mathbb{R}^m} Y \pi(Y; X) dY; \quad (2.11)$$

- Posterior mode forecast

$$\hat{Y}_1 = \arg \max_{Y \in \mathcal{Y}} \pi(Y; X); \quad (2.12)$$

- Posterior median forecast (for $M = 1$): \hat{Y}_2 is defined as a root of the equation

Fig. 2.3 Construction of point and interval forecasts from the Bayesian forecast density



$$\int_{-\infty}^{\hat{Y}_2} \pi(Y; X) dY = \int_{\hat{Y}_2}^{+\infty} \pi(Y; X) dY. \quad (2.13)$$

Figure 2.3 above presents an example of using the Bayesian forecast density to construct point and interval forecasts in the univariate case ($M = 1$).

The following two techniques can be proposed for *set (interval) forecasting*. Let the *domain of γ -maximal Bayesian forecast density* be a subset of the possible forecasts defined as

$$\mathcal{Y}_\gamma = \{Y \in \mathcal{Y} : \pi(Y; X) \geq \gamma \pi_{\max}\}, \quad (2.14)$$

where $\pi_{\max} = \max_{Y \in \mathcal{Y}} \pi(Y; X)$ and the parameter $\gamma \in (0.5, 1)$ defines the size of the domain \mathcal{Y}_γ , i.e., its M -dimensional volume $\text{mes}_M(\mathcal{Y}_\gamma)$. Following the theory of statistical interval estimation, let us define the posterior γ -confidence region \mathcal{Y}_γ as the solution of the following conditional minimization problem:

$$\text{mes}_M(\mathcal{Y}_\gamma) \rightarrow \min, \quad \int_{\mathcal{Y}_\gamma} \pi(Y; X) dY = \gamma. \quad (2.15)$$

In order to simplify the computations, it is often advisable to consider a parametric family of possible confidence regions.

It should be noted that computation of Bayesian density forecasts (2.9), (2.10) is often complicated by the necessity of multiple integration over $\theta \in \Theta \subseteq \mathbb{R}^m$. If analytic computation of the integrals in (2.9) and (2.10) appears to be unfeasible, Monte Carlo numeric integration may be used:

$$\pi(Y; X) \approx \frac{1}{K} \sum_{i=1}^K g(Y | \theta^{(i)}).$$

Here $\theta^{(1)}, \dots, \theta^{(K)} \in \Theta$ is a sample of K independent random vectors with the probability density function $p(\theta \mid X)$, which can be simulated by using standard algorithms. As the number of Monte Carlo trials K increases to infinity, the mean square error of this approximation decreases as K^{-1} .

To illustrate the notions and methods of this section, let us consider a problem of forecasting a financial company's income.

Assume that the income Y over the next business day is a random variable depending on the average increment $\theta^0 \in \mathbb{R}^1$ of a certain currency exchange rate over the previous day:

$$Y = \mu_0 + k \theta^0 + \varepsilon,$$

where $\mu_0 \in \mathbb{R}^1$ is the (known) guaranteed mean income that doesn't depend on the currency exchange market, $k \theta^0$ is the income depending on θ^0 (here $k > 0$ is a known proportionality coefficient); ε is a random variation of the income modeled by a normally distributed random variable,

$$\mathcal{L}\{\varepsilon\} = N(0, \Delta^2),$$

with a known variance Δ^2 . The parameter θ^0 is unknown, but statistical data x_1, \dots, x_N representing the 1-day exchange rate increments over the previous day (offered by N commercial banks) has been collected, where $X = (x_i) \in \mathbb{R}^N$ is assumed to be a random sample of size N taken from a normal probability distribution, $\mathcal{L}\{x_i\} = N(0, \sigma^2)$, with a known variance σ^2 .

We would like to make point and interval forecasts of the income Y based on statistical data X , the above model assumptions, and a prior assumption that θ^0 is uniformly distributed over a given interval $[a, b]$ (for example, we can assume that the minimum and maximum exchange rates a and b have been set by a central bank).

Model assumptions yield the following expressions:

$$g(Y \mid \theta) = \frac{1}{\sqrt{2\pi}\Delta} \exp\left(-\frac{1}{2\Delta^2} (Y - \mu_0 - k\theta)^2\right),$$

$$p(X \mid \theta) = (2\pi\sigma^2)^{-N/2} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^N (x_i - \theta)^2\right),$$

$$q(\theta) = \frac{1}{b-a} \mathbf{1}_{[a,b]}(\theta),$$

where $\mathbf{1}_A(\theta) = \{1, \theta \in A; 0, \theta \notin A\}$ is the indicator function of the set A . Applying (2.10) results in the Bayesian forecast density

$$p(\theta \mid X) = \frac{\exp(-N(2\sigma^2)^{-1}(\theta - \bar{x})^2)}{\sqrt{2\pi} \frac{\sigma}{\sqrt{N}} \left(\Phi\left(\sqrt{N} \frac{b-\bar{x}}{\sigma}\right) - \Phi\left(\sqrt{N} \frac{a-\bar{x}}{\sigma}\right) \right)} \mathbf{1}_{[a,b]}(\theta),$$

which is the normal probability density function $N(\bar{x}, \sigma^2/N)$ constrained to $[a, b]$, where $\bar{x} = N^{-1} \sum_{i=1}^N x_i$ is the sample mean. This, together with (2.9), leads to the equation

$$\pi(Y; X) = \left(2\pi \frac{\sigma \Delta}{\sqrt{N}} \left(\Phi \left(\sqrt{N} \frac{b - \bar{x}}{\sigma} \right) - \Phi \left(\sqrt{N} \frac{a - \bar{x}}{\sigma} \right) \right) \right)^{-1} \times \\ \times \int_a^b \exp \left(-\frac{1}{2} \left(\frac{N}{\sigma^2} (\theta - \bar{x})^2 + \frac{1}{\Delta^2} (k\theta + \mu_0 - Y)^2 \right) \right) d\theta,$$

where the right-hand side integral can be rewritten using the standard normal distribution function $\Phi(\cdot)$ by performing a substitution of the variables. Applying this equation to (2.11)–(2.15) yields the desired forecasts.

2.4 Forecasting Discrete States by Discriminant Analysis

2.4.1 The Mathematical Model

In applications, the underlying process can often be described by a discrete stochastic model [2, 5–8, 11]:

$$\nu \in S = \{1, 2, \dots, L\},$$

where ν is the future unknown state of the system and $2 \leq L < +\infty$ is the number of possible values of ν (i.e., the number of possible forecasts). Let us consider some examples.

Example 2.1. A bank scores a prospective client (a certain company) applying for a loan. The financial circumstances of the client are characterized by N business indicators $X = (x_i) \in \mathbb{R}^N$ (for instance, x_1 is the total annual income, x_2 is the demand for the products made by the company, and x_3 characterizes the dynamics of the company's bank accounts). Based on statistical data X , the bank makes a forecast $\hat{\nu} = d(X)$, where $\hat{\nu} = 1$ stands for a “reliable client” bringing a profit to the bank and $\hat{\nu} = 2$ —an “unreliable client” failing to repay the loan and causing a loss (in this example, $L = 2$).

Example 2.2. Let $\hat{\nu} = d(X)$ be a success forecast for a certain clinical treatment based on a patient's medical test results $X \in \mathbb{R}^N$; $\hat{\nu} = 0$ means that the patient's health will remain unchanged, $\hat{\nu} = 1$ corresponds to a health improvement, and $\hat{\nu} = 2$ —a health deterioration (in this example, $L = 3$).

The statistical classification model or, to be precise, the discriminant analysis model [3, 9, 11] can be used to solve this type of applied problems. Discriminant analysis is a branch of statistical data analysis devoted to models and methods of identifying the observed data as belonging to one of the given populations (classes, patterns, etc.), i.e., classification of statistical observations.

Let us interpret a classification problem as a forecasting problem defined earlier. Assume that a random observation $x = (x_k) \in \mathbb{R}^N$ belongs to one of the $L \geq 2$ classes $\Omega_1, \dots, \Omega_L$, and let the possible forecasts be the indices of these classes: a forecast $v = i$ corresponds to the class Ω_i and vice versa. Let an observation belonging to the class Ω_i be a random N -vector $X_i \in \mathbb{R}^N$ with a conditional probability density $p_i^0(x)$, $i \in S$. We are going to assume the knowledge of prior class probabilities π_1, \dots, π_L :

$$\pi_i = \mathbb{P}\{v = i\} > 0, \quad \sum_{i \in S} \pi_i = 1.$$

We also assume prior knowledge of the $(L \times L)$ forecasting (classification) loss matrix $W = (w_{il})$, where $w_{il} \geq 0$ is the loss taken if an observation belonging to the class Ω_i is classified as belonging to the class Ω_l , i.e., if $v = i$, but $\hat{v} = l$. For example, a $(0-1)$ loss matrix W is defined as follows:

$$w_{il} = 1 - \delta_{il}, \quad i, l \in S, \quad (2.16)$$

where δ_{il} is the Kronecker delta.

Under this model, optimal forecasting, as defined in Sects. 2.1 and 2.3, is equivalent to constructing a BDR

$$\hat{v} = d_0(x) : \mathbb{R}^N \rightarrow S, \quad (2.17)$$

that minimizes the mean loss resulting from the forecast. This problem is solved differently depending on the available prior knowledge of probabilistic characteristics of the classes $\{\pi_i, p_i^0(\cdot)\}$.

2.4.2 Complete Prior Knowledge of $\{\pi_i, p_i^0(\cdot)\}$

Let us introduce the following notation:

$$f_j(x; \{p_i^0(\cdot)\}) = \sum_{i \in S} \pi_i p_i^0(x) w_{ij}; \quad p(x) = \sum_{i \in S} \pi_i p_i^0(x). \quad (2.18)$$

Here $p(x)$ is the unconditional probability density function of the random observation $X \in \mathbb{R}^N$ determined by the stochastic model of the investigated process. From the Bayes formula,

$$\mathbb{P}\{v = j \mid X = x\} = \frac{\pi_j p_j^0(x)}{p(x)}, \quad (2.19)$$

and thus (2.18) can be rewritten as

$$\frac{f_j(x; \{p_i^0(\cdot)\})}{p(x)} = \sum_{i \in S} \mathbb{P}\{v = i \mid X = x\} w_{ij} = \mathbb{E}\{w_{vj} \mid X = x\}. \quad (2.20)$$

The relation (2.20) means that, to a multiplier $p(x)$ not depending on j , the function f_j in (2.18) defines the posterior mean forecast loss $\hat{v} = j$ given the observation vector $X = x$.

Theorem 2.1. *Under prior knowledge of the probability distributions $\{\pi_i, p_i^0(\cdot)\}$, assume that for all $i, k, l \in S$, $k \neq l$, the condition*

$$\mathbb{P}_{\theta^0}\{f_k(X; \{p_j^0(\cdot)\}) - f_l(X; \{p_j^0(\cdot)\}) = 0\} = 0 \quad (2.21)$$

is satisfied. Then the BDR (2.17) is unique and, up to a set of Lebesgue measure zero, has the form

$$\hat{v} = d = d_0(x) = \arg \min_{j \in S} f_j(x; \{p_i^0(\cdot)\}), \quad x \in \mathbb{R}^N, \quad d \in S. \quad (2.22)$$

This BDR minimizes the mean loss of forecasting (the Bayesian risk):

$$r_0 = \int_{\mathbb{R}^N} \min_{j \in S} f_j(x; \{p_i^0(\cdot)\}) dx. \quad (2.23)$$

Proof. Taking into account (2.3) and (2.18), the Bayesian risk functional can be rewritten as

$$r = r(d(\cdot); \{p_i^0(\cdot)\}) = \sum_{i \in S} \pi_i \int_{\mathbb{R}^N} p_i^0(x) w_{id(x)} dx = \int_{\mathbb{R}^N} f_{d(x)}(x; \{p_i^0(\cdot)\}) dx, \quad d(\cdot) \in D. \quad (2.24)$$

Looking at the form of the functional (2.24), it is easy to find the lower bound of the Bayesian risk over all possible decision rules $d(\cdot) \in D$:

$$r(d(\cdot); \{p_i^0(\cdot)\}) \geq \min_{d(\cdot) \in D} \int_{\mathbb{R}^N} f_{d(x)}(x; \{p_i^0(\cdot)\}) dx \geq \int_{\mathbb{R}^N} \min_{j \in S} f_j(x; \{p_i^0(\cdot)\}) dx = r_0.$$

Therefore, the above inequality becomes an equality after substituting the decision rule (2.22). \square

Observe that, as in Sect. 2.2 (see Property 2.1), the obtained BDR (2.22) minimizes the posterior mean loss.

Corollary 2.1. *For a (0–1) loss matrix (2.16), the Bayesian risk can be interpreted as the probability of making an incorrect forecast*

$$r_0 = \inf_{d(\cdot) \in D} \mathbb{P}\{\hat{v} \neq v\} = 1 - \int_{\mathbb{R}^N} \max_{i \in S} (\pi_i p_i^0(x)) dx,$$

and the expression for the BDR can be written in a simplified form:

$$\hat{v} = d = d_0(x) = \arg \max_{i \in S} (\pi_i p_i^0(x)), \quad x \in \mathbb{R}^N, \quad d \in S. \quad (2.25)$$

Proof. Let us substitute (2.16) into (2.18), (2.22), and (2.24). Taking normalization into account, we obtain

$$\begin{aligned} f_j(x; \{p_i^0(\cdot)\}) &= p(x) - \pi_j p_j^0(x), \\ r(d(\cdot); \{p_i^0(\cdot)\}) &= 1 - \int_{\mathbb{R}^N} \pi_{d(x)} p_{d(x)}^0(x) dx, \end{aligned}$$

and (2.25) follows immediately. \square

Corollary 2.2. *Under the assumptions of Corollary 2.1, let the observations be described by an N -dimensional normal (Gaussian) model:*

$$p_i^0(x) = n_N(x \mid \mu_i, \Sigma_i) = (2\pi)^{-\frac{N}{2}} |\Sigma_i|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(x - \mu_i)' \Sigma_i^{-1} (x - \mu_i)\right), \quad (2.26)$$

where $\mu_i = (\mu_{ij}) \in \mathbb{R}^N$ is the mean vector and $\Sigma_i = (\sigma_{ijk}) \in \mathbb{R}^{N \times N}$ is the nonsingular covariance matrix of the random vector $X_i \in \mathbb{R}^N$. Then the BDR is quadratic:

$$\hat{v} = d = d_0(x) = \arg \min_{i \in S} \left((x - \mu_i)' \Sigma_i^{-1} (x - \mu_i) + \ln(|\Sigma_i|/\pi_i^2) \right). \quad (2.27)$$

Proof. Substitute (2.26) into (2.25) and perform the obvious transformations. \square

Corollary 2.3. *Under the assumptions of Corollary 2.1, take Fisher's model [3]:*

$$\Sigma_1 = \Sigma_2 = \dots = \Sigma_L = \Sigma \quad (2.28)$$

with two classes ($L = 2$). In that case the BDR is linear:

$$\begin{aligned}
 \hat{v} = d = d_0(x) &= \mathbf{1}(l(x)) + 1, \quad x \in \mathbb{R}^N, \\
 l(x) &= b'x + \beta, \\
 b &= \Sigma^{-1}(\mu_2 - \mu_1), \\
 \beta &= (\mu_1' \Sigma^{-1} \mu_1 - \mu_2' \Sigma^{-1} \mu_2)/2 + \ln(\pi_2/\pi_1), \\
 r_0 &= 1 - \left(\pi_1 \Phi \left(\frac{\Delta}{2} + \frac{1}{\Delta} \ln \frac{\pi_1}{\pi_2} \right) + \pi_2 \Phi \left(\frac{\Delta}{2} - \frac{1}{\Delta} \ln \frac{\pi_1}{\pi_2} \right) \right),
 \end{aligned} \tag{2.29}$$

where $\Phi(\cdot)$ is the standard normal $N(0, 1)$ distribution function and

$$\Delta = \sqrt{(\mu_2 - \mu_1)' \Sigma^{-1} (\mu_2 - \mu_1)} \geq 0$$

is the so-called Mahalanobis distance between classes [3].

Proof. Rewriting the BDR as (2.27) for $L = 2$ and taking into account the notation (2.28), (2.29) yields

$$\hat{v} = d = d_0(x) = \arg \min_{i \in S} \left(-\mu_i' \Sigma^{-1} x + \frac{1}{2} \mu_i' \Sigma^{-1} \mu_i - \ln \pi_i \right) \equiv \mathbf{1}(l(x)) + 1, \quad x \in \mathbb{R}^N,$$

which is the first expression of (2.29).

Now let us compute the Bayesian risk (i.e., the unconditional probability of a forecast error) for the BDR (2.29):

$$r_0 = \pi_1 P_1 + \pi_2 P_2, \tag{2.30}$$

where

$$P_i = \mathbb{P}\{\hat{v} \neq i \mid v = i\}, \quad i \in S,$$

is the conditional probability of a forecast error given that the true number of the class equals $v = i$. Due to (2.29), we have

$$P_1 = \mathbb{P}\{\hat{v} = 2 \mid v = 1\} = \mathbb{P}\{l(X_1) \geq 0\} = 1 - F_{l_1}(0), \tag{2.31}$$

where $l_i = l(X_i) = b'X_i + \beta$ is a random variable and $F_{l_i}(z)$, $z \in \mathbb{R}^1$, is the distribution function of the random variable l_i , $i \in S$. From the condition (2.26), the probability distribution of X_i can be written as

$$\mathcal{L}\{X_i\} = \mathcal{N}_N(\mu_i, \Sigma_i),$$

and the linear transformation theorem for normal random vectors [3] yields

$$\begin{aligned}\mathcal{L}\{l_i\} &= \mathcal{N}_1(m_i, \zeta_i), \\ m_i &= b' \mu_i + \beta = (-1)^i \Delta^2/2 + \ln \frac{\pi_2}{\pi_1}, \\ \zeta_i &= b' \Sigma b = (\mu_2 - \mu_1)' \Sigma^{-1} (\mu_2 - \mu_1) = \Delta^2,\end{aligned}$$

where the variance $\zeta_i = \Delta^2$ doesn't depend on $i \in S$. Therefore,

$$F_{l_i}(z) = \Phi\left(\frac{z - m_i}{\Delta}\right), \quad i \in S.$$

Substituting this equality into (2.31) results in the expression

$$P_1 = 1 - \Phi\left(\frac{\Delta}{2} + \frac{1}{\Delta} \ln \frac{\pi_1}{\pi_2}\right).$$

Similarly, we have

$$P_2 = 1 - \Phi\left(\frac{\Delta}{2} - \frac{1}{\Delta} \ln \frac{\pi_1}{\pi_2}\right).$$

Substituting P_1, P_2 into (2.30) yields (2.29). \square

Definition 2.8. Fisher's linear discriminant function is defined as $l(x) = b'x + \beta$ (as implied by (2.29), its sign determines the forecast). The set

$$\Gamma_0 = \{x : b'x + \beta = 0\} \subset \mathbb{R}^N$$

is called Fisher's discriminant hyperplane.

Figure 2.4 illustrates Fisher's linear decision rule for $N = 2$.

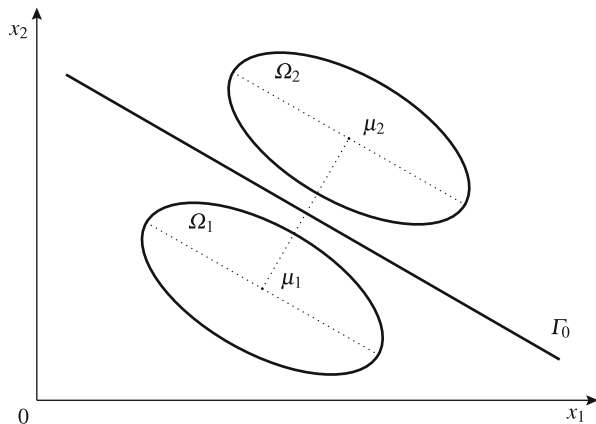
To conclude this subsection, let us observe that, as in Sect. 2.3, it is possible to construct a randomized decision rule $\hat{v} = \tilde{d}(x, \omega)$ which is going to be described by a Bayesian forecast distribution (2.9), (2.10) defined on the set of possible forecasts $A = S = \{1, 2, \dots, L\}$:

$$\pi(i; x) = \mathbb{P}\{\hat{v} = i \mid X = x\} = \frac{\pi_i p_i^0(x)}{p(x)}, \quad i \in S.$$

The nonrandomized forecast (2.25) is, in fact, equal to the posterior mode (2.12).

Some applications require interval forecasts $H_\gamma \subseteq S$ defined by (2.15). In the discrete case, this definition can be rewritten as

$$\sum_{i \in H_\gamma} \pi(i, x) \geq \gamma, \quad |H_\gamma| \rightarrow \min. \quad (2.32)$$

Fig. 2.4 Fisher's linear decision rule

Interval forecasts defined by (2.32) become very useful in the rather common case, where the number of classes is large ($L \gg 1$).

2.4.3 Prior Uncertainty

Consider a setting with a priori unknown conditional probability densities of the observations $\{p_i^0(\cdot)\}$. To overcome this prior uncertainty, we can use a so-called classified training sample $Z \subset \mathbb{R}^N$ of total size $n = n_1 + \dots + n_L$, which consists of L independent subsamples:

$$Z = \bigcup_{i \in S} Z_i, \quad Z_i \cap Z_j = \emptyset, \quad j \neq i.$$

Here

$$Z_i = \{z_{ij} \in \mathbb{R}^N : j = 1, \dots, n_i\}$$

is a random subsample of size n_i taken from the class Ω_i (i.e., a subset of statistical data corresponding to the forecast value $v = i$).

Let us start by considering the case of *parametric prior uncertainty*, where the densities $\{p_i^0(\cdot)\}$, $i \in S$, belong to a given family of probability distributions, but the distribution parameters remain unknown:

$$p_i^0(x) = q(x; \theta_i^0), \quad x \in \mathbb{R}^N, \quad \theta_i^0 \in \Theta,$$

where

$$Q = \{q(x; \theta), \quad x \in \mathbb{R}^N : \theta \in \Theta \subseteq \mathbb{R}^m\}$$

is some given m -parametric family of N -dimensional probability densities. Forecasting under parametric uncertainty is usually based on one of the two approaches described below. Recall that we are constructing a forecast $\hat{v} \in S$ based on the collected statistical data Z and the recorded observation $x \in \mathbb{R}^N$.

A. Construction of plug-in decision rules (PDRs).

Definition 2.9. A PDR is defined as the decision rule obtained from a BDR (2.22) by substituting consistent statistical estimators $\{\hat{\theta}_i\}$ for the unknown true values of the parameters $\{\theta_i^0\}$ based on the training sample Z :

$$\tilde{v} = d_1(x; Z) = \arg \min_{j \in S} f_j(x; \{q(x; \hat{\theta}_i)\}), \quad x \in \mathbb{R}^N, \quad \tilde{v} \in S, \quad (2.33)$$

where functions $\{f_j(\cdot)\}$ are defined by (2.18).

The estimators $\{\hat{\theta}_i\}$ are usually the maximum likelihood estimators (MLEs):

$$\hat{\theta}_i = \arg \max_{\theta \in \Theta} \frac{1}{n_i} \sum_{j=1}^{n_i} \ln q(z_{ij}; \theta), \quad i \in S. \quad (2.34)$$

Theorem 2.2. *If the parametric family Q of probability densities satisfies the classical regularity conditions [4], then the forecast \tilde{v} defined by the PDR (2.33), (2.34), converges in probability to the forecast \hat{v} defined by the BDR (2.22):*

$$d_1(x; Z) \xrightarrow{\mathbf{P}} d_0(x), \quad x \in \mathbb{R}^N, \quad (2.35)$$

for

$$n_0 = \min_{i \in S} n_i \rightarrow \infty.$$

Proof. Regularity conditions together with certain well-known asymptotic properties of MLEs [4] imply that

$$\hat{\theta}_i \xrightarrow{\mathbf{P}} \theta_i^0, \quad i \in S.$$

Notation (2.18) and well-known results on functional transformations of convergent sequences [4, 13] yield the relations

$$\begin{aligned} q(x; \hat{\theta}_i) &\xrightarrow{\mathbf{P}} q(x; \theta_i^0), \\ f_j(x; \{q(x; \hat{\theta}_i)\}) &\xrightarrow{\mathbf{P}} f_j(x; \{q(x; \theta_i^0)\}), \quad i, j \in S, \quad x \in \mathbb{R}^N. \end{aligned}$$

Since S is a finite set, the convergence of the objective functions

$$f_j(\cdot), \quad j \in S,$$

implies that the minimum points also converge:

$$\arg \min_{j \in S} f_j(x; \{q(x; \hat{\theta}_i)\}) \xrightarrow{\mathbf{P}} \arg \min_{j \in S} f_j(x; \{q(x; \theta_i^0)\}), \quad x \in \mathbb{R}^N.$$

This, together with (2.33) and (2.22), proves the convergence in (2.35). \square

Let us define the unconditional Bayesian risk of a PDR $d_1(\cdot)$ similarly to (2.24):

$$r(d_1(\cdot)) = \mathbb{E} \left\{ \int_{\mathbb{R}^N} f_{d_1(x; Z)}(x; \{p_i^0(x)\}) dx \right\}, \quad (2.36)$$

where the expectation $\mathbb{E}\{\cdot\}$ is computed w.r.t. the probability distribution of the random sample Z . Known asymptotic expansions of the deviations $\{\hat{\theta}_i - \theta_i^0\}$ [9, 13] lead to the following asymptotic expansion for the unconditional risk (2.36):

$$r(d_1(\cdot)) = r_0 + \sum_{i=1}^L \frac{\varrho_i}{n_i} + O(n_0^{-3/2}), \quad (2.37)$$

where the coefficients $\{\varrho_i\}$ above satisfy the condition

$$\varrho_i = \varrho_i(N, \{\pi_i\}, \{q(\cdot; \theta_i^0)\}, \{w_{ij}\}) \geq 0.$$

It is easy to see from (2.37) that for $n_0 \rightarrow \infty$ the PDR risk (2.33) converges to the minimal Bayesian risk (2.23):

$$r(d_1) \rightarrow r_0, \quad (2.38)$$

and therefore in practice (2.33) is often called the *suboptimal decision rule*.

In practical applications, it is important to choose sufficiently large training sample sizes n_1, n_2, \dots, n_L that guarantee a minor relative increase in the forecast risk due to the uncertainty of $\{\theta_i^0\}$. The relation (2.37) can be used to evaluate this increment:

$$\frac{r(d_1) - r_0}{r_0} \approx \sum_{i=1}^L \frac{\varrho_i}{n_i r_0} \leq \delta. \quad (2.39)$$

B. Using the Bayesian forecast distribution.

Define an Lm -dimensional composite column vector of parameters for the probability distributions

$$p_i^0 = q(x; \theta_i^0), \quad x \in \mathbb{R}^N, \quad \theta_i^0 \in \Theta \subseteq \mathbb{R}^m,$$

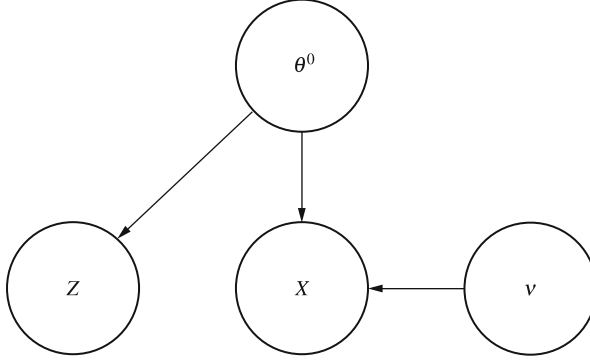


Fig. 2.5 Stochastic dependence between model components

where $i \in S$, as

$$\theta^0 = (\theta_1^{0'} : \theta_2^{0'} : \dots : \theta_L^{0'})' \in \mathbb{R}^{Lm},$$

and assume that θ^0 is a random vector with an a priori given probability density function $q(\theta)$, $\theta \in \mathbb{R}^{Lm}$.

To define a Bayesian forecast distribution and construct a randomized decision rule

$$\tilde{v} = d_2(x; Z, \omega),$$

we are going to use the diagram of the stochastic dependence between v , X , θ^0 , and Z presented in Fig. 2.5.

The Bayesian forecast distribution is defined on the decision space S in the following way:

$$\pi(i; x, Z) = \mathbb{P}\{v = i \mid X = x, Z\} = \int_{\mathbb{R}^{Lm}} \frac{\pi_i q(x; \theta_i)}{\sum_{j \in S} \pi_j q(x; \theta_j)} p(\theta \mid Z) d\theta, \quad (2.40)$$

$$p(\theta \mid Z) = \frac{q(\theta) \prod_{i=1}^L \prod_{j=1}^{n_i} q(z_{ij}; \theta_i)}{\int_{\mathbb{R}^{Lm}} q(\theta') \prod_{i=1}^L \prod_{j=1}^{n_i} q(z_{ij}; \theta'_i) d\theta'}.$$

As in Sect. 2.3, (2.40) can be used to construct point and interval forecasts of v .

To conclude the section, let us briefly discuss the case of models with *nonparametric prior uncertainty*, where the N -dimensional probability densities lie in a distribution family \mathcal{P} which doesn't allow for a finite parameterization:

$$p_1^0(\cdot), p_2^0(\cdot), \dots, p_L^0(\cdot) \in \mathcal{P}.$$

In this setting, the approach A is still valid, requiring only a modified construction of admissible estimators $\{\hat{p}_i(\cdot)\}$ for $\{p_i^0(\cdot)\}$.

Two types of nonparametric estimators $\{\hat{p}_i(\cdot)\}$ are the most relevant to applications: the Rosenblatt–Parzen estimators and the k -Nearest-Neighbor estimators.

A *nonparametric (kernel) Rosenblatt–Parzen estimator* [12] of the density $p_i^0(\cdot)$ based on the sample Z_i is defined as the statistic

$$\hat{p}_i(x) = \frac{1}{n_i |H_i|} \sum_{j=1}^{n_i} K(H_i^{-1}(x - z_{ij})), \quad x = (x_l) \in \mathbb{R}^N. \quad (2.41)$$

In this definition,

$$K(x) = \prod_{l=1}^N K_l(x_l)$$

is an N -dimensional kernel, and each $K_l(y)$, $y \in \mathbb{R}^1$, is a *one-dimensional kernel*—a nonnegative bounded differentiable even function such that $K_l(|y|)$ is nonincreasing in $|y|$, the conditions

$$\int_0^{+\infty} y^m K_l(y) dy < +\infty \quad (m > 0), \quad \int_{-\infty}^{+\infty} y^2 K_l(y) dy = 1;$$

are satisfied, and $H_i = \text{diag}\{h_{il}\}$ is a diagonal $(N \times N)$ -matrix. Diagonal elements of H_i are known as smoothing coefficients; they are strictly positive, $h_{il} > 0$. Given the convergence

$$h_{il} = h_{il}(n_i) \rightarrow 0, \quad n_i |H_i| \rightarrow \infty$$

as $n_i \rightarrow \infty$, the estimator (2.41) is consistent [12].

It has been proved [9] that if \mathcal{P} is a family of thrice differentiable densities, the sizes of training samples $\{n_i\}$ are comparable:

$$n_0 = \min_{i \in S} n_i \rightarrow \infty, \quad n_i = c_i n_0, \quad 1 \leq c_i < +\infty,$$

where $\{c_i\}$ are certain constants, and the smoothing coefficients can be written asymptotically as

$$h_{il} = b_{il} n_i^{-\frac{1}{N+4}}, \quad l = 1, \dots, N,$$

then the unconditional Bayesian risk (2.36) satisfies an asymptotic expansion similar to (2.37):

$$r(d_1(\cdot)) = r_0 + \frac{q}{n_0^{4/(N+4)}} + o\left(n_0^{-4/(N+4)}\right), \quad (2.42)$$

where $q = q(N, \{\pi_i\}, \{p_i^0(\cdot)\}, \{w_{ij}\})$ is a known coefficient of the asymptotic formula. The asymptotic expansion (2.42) implies that a PDR $d_1(\cdot)$ constructed from Rosenblatt–Parzen estimators satisfies (2.38) and is, therefore, suboptimal. Similarly to the parametric case (2.39), the asymptotic expansion (2.42) yields the following explicit relation between n_0 and δ :

$$n_0 \geq \left(\frac{q}{r_0 \delta}\right)^{\frac{N}{4}+1}. \quad (2.43)$$

A *generalized nonparametric k -Nearest-Neighbor (k -NN) estimator* of the density $p_i^0(\cdot)$ based on the sample Z_i is defined as the following statistic [9]:

$$\hat{p}_i(x) = \frac{1}{n_i \varrho_i^N} \sum_{j=1}^{n_i} L_i\left(\frac{x - z_{ij}}{\varrho_i}\right), \quad x \in \mathbb{R}^N, \quad i \in S, \quad (2.44)$$

where $\varrho_i = \varrho_i(x; Z_i) > 0$ is the Euclidean distance between a point $x \in \mathbb{R}^N$ and the k_i th nearest neighbor of the point x in Z_i ; each number of neighbors k_i , $2 \leq k_i \leq n_i$, is a positive integer parameter of the estimator; the function $L_i(u)$, $u = (u_k) \in \mathbb{R}^N$, is a bounded integrable weight function such that

$$\int_{\mathbb{R}^N} L_i(u) du = 1, \quad \int_{\mathbb{R}^N} |u|^3 L_i(u) du < \infty, \quad \int_{\mathbb{R}^N} u_k L_i(u) du = 0, \quad k = 1, \dots, N.$$

In applications, a uniform weight function is used most frequently:

$$L_i(u) = (2\pi^{N/2})^{-1} N \Gamma(N/2) \mathbf{1}_{[0,1]}(|u|),$$

where $\Gamma(\cdot)$ is the gamma function. Assuming the convergences $k_i = k_i(n_i) \rightarrow \infty$, $k_i(n_i)/n_i \rightarrow 0$ as $n_i \rightarrow \infty$, the estimator (2.44) is consistent.

It is known [9] that if \mathcal{P} is a family of thrice differentiable densities, then a PDR based on k -NN estimators (2.44) with the coefficients k_i defined as

$$k_i = \lceil b_i n_i^{4/(N+4)} \rceil, \quad i = 1, \dots, N,$$

where $\{b_i\}$ are some constants, is suboptimal. The unconditional Bayesian risk of this PDR satisfies the asymptotic formula (2.42) with a different value of the coefficient q .

The above analysis shows that parametric prior uncertainty leads to a risk increment that can be estimated as $O(n_0^{-1})$, and nonparametric uncertainty results in a much larger increment— $O\left(n_0^{-4/(N+4)}\right)$. The difference between risks of nonparametric and parametric forecasting becomes higher as N (the number of dimensions of the observation space, or, equivalently, the number of quantities characterizing the investigated process) increases.

To conclude the chapter, let us note that an even higher level of prior uncertainty may be considered, where the training sample Z is assumed to be unclassified. In that case, a forecasting algorithm can be constructed by applying methods of cluster analysis [9].

References

1. Aitchison, J., Dunsmore, J.: Statistical Prediction Analysis. CUP, Cambridge (1975)
2. Amagor, H.: A Markov analysis of DNA sequences. *J. Theor. Biol.* **104**, 633–642 (1983)
3. Anderson, T.: An Introduction to Multivariate Statistical Analysis. Wiley, Hoboken (2003)
4. Borovkov, A.: Mathematical Statistics. Gordon & Breach, Amsterdam (1998)
5. Collet, D.: Modeling Binary Data. Chapman & Hall, London (2002)
6. Fokianos, K., Fried, R.: Interventions in ingarch models. *J. Time Ser. Anal.* **31**, 210–225 (2010)
7. Fokianos, K., Kedem, B.: Prediction and classification of nonstationary categorical time series. *J. Multivar. Anal.* **67**, 277–296 (1998)
8. Fokianos, K., Kedem, B.: Regression theory for categorical time series. *Stat. Sci.* **18**, 357–376 (2003)
9. Kharin, Yu.: Robustness in Statistical Pattern Recognition. Kluwer Academic, Dordrecht (1996)
10. Lloyd, E.: Handbook of Applicable Mathematics, vol. 6. Wiley, Chichester (1994)
11. McLachlan, G.: Discriminant Analysis and Statistical Pattern Recognition. Wiley, New York (1992)
12. Parzen, E.: On the estimation of a probability density function and the mode. *Ann. Math. Stat.* **40**, 1063–1076 (1962)
13. Serfling, R.: Approximation Theorems of Mathematical Statistics. Wiley, New York (1980)
14. Wald, A.: Sequential Analysis. Wiley, New York (1947)



<http://www.springer.com/978-3-319-00839-4>

Robustness in Statistical Forecasting

Kharin, Y.

2013, XVI, 356 p. 47 illus., Hardcover

ISBN: 978-3-319-00839-4