ESTIMATION THEORY

It is often the case in control and communication systems that the mathematical model describing a particular phenomenon is completely specified, except some unknown quantities. These quantities must be estimated. Identification, adaptive control, learning systems, and the like, provide examples. Exact answers are often difficult, expensive, or merely impossible to obtain. However, approximate answers that are likely to be close to the exact answers may be fairly easily obtainable. Estimation theory provides a general guide for obtaining such answers; above all, it makes mathematically precise such phrases as "likely to be close," "this estimator is optimal (better than others)," and so forth.

Though estimation theory originated from certain practical problems, only the mathematical aspects of the subject are considered below. The practical usefulness of the concepts used is not comprehensively discussed. One can refer to the treatises (3) and (9) for thorough motivations of these concepts from the application point of view.

What follows considers only the *statistical* framework, that is, it is supposed that the noisy environment, where observations are taken, is of a *stochastic* (random) nature. Situations when this assumption does not hold are addressed by *minimax estimation* methods.

Depending on how much prior information about the system to be identified is available, one may distinguish between two cases:

1. The system can be specified up to an unknown parameter of finite dimension. Then the problem is called the *parametric* estimation problem. For instance, such a problem arises when the parameters of a linear system of bounded dimension are to be estimated.
2. However, rather often, one has to infer relationships between input and output data of a system, when very little prior knowledge is available. In engineering practice, this problem is known as black-box modeling. Linear system of infinite dimension and general nonlinear systems, when the input/output relation cannot be defined in terms of a fixed number of parameters, provide examples. In estimation theory, these problems are referred to as those of *nonparametric* estimation.

Consider now some simple examples of mathematical statements of estimation problems.

Example 1. Let $X_1, \ldots, X_n$ be independent random variables (or observations) with a common unknown distribution $\mathcal{P}$ on the real line. One can consider several estimates (i.e., functions of the observations (X_i), $i = 1, \ldots, n$) of the mean $\theta = \int x d\mathcal{P}$:

1. The empirical mean
$$\tilde{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{1}$$
2. The empirical median $m = \text{median}(X_1, \ldots, X_n)$, which is constructed as follows: Let $Z_1, \ldots, Z_n$ be an increasing rearrangement of $X_1, \ldots, X_n$. Then $m = Z_{[(n+1)/2]}$ for n odd and $m = (Z_{n/2} + Z_{n/2+1})/2$ for n even (here $[x]$ stands for the integer part of x).
3. $g = (\max_{1\le i\le n}(X_i) + \min_{1\le i\le n}(X_i))/2$

Example 2. The (linear regression model). The variables y_i, X_i^k, $i = 1, \ldots, n$, $k = 1, \ldots, d$ are observed, where

$$y_i = \theta_1 X_i^1 + \cdots + \theta_d X_i^d + e_i$$

The e_i are random disturbances and $\theta_1, \ldots, \theta_d$ should be estimated. Let us denote $X_i = (X_i^1, \ldots, X_i^d)^T$, $\theta = (\theta_1, \ldots, \theta_d)^T$. The estimate

$$\hat{\theta}_n = \arg\min_{\theta} \sum_{i=1}^{n} (y_i - \theta^T X_i)^2$$

of θ is referred to as the *least squares estimate.*

Example 3. Let $f(x)$ be an unknown signal, observed at the points, $X_i = i/n$, $i = 1, \ldots, n$ with additive noise:

$$y_i = f(X_i) + e_i, \qquad i = 1, \ldots, n \tag{2}$$

This problem is referred to as nonparametric regression. Suppose that f is square-integrable and periodic on [0,1]. Then one can develop f into Fourier series

$$f(x) = \sum_{k=0}^{\infty} c_k \phi_k(x)$$

where, for instance, $\phi_0(x) = 0$, $\phi_{2l-1}(x) = \sqrt{2}\sin(2\pi lx)$, and $\phi_{2l}(x) = \sqrt{2}\cos(2\pi lx)$ for $l = 1, 2, \ldots$. Then one can compute the empirical coefficients

$$\hat{c}_k = \frac{1}{n}\sum_{k=1}^{n} y_i \phi_k(X_i) \tag{3}$$

and construct an estimate $\hat{f}_n$ of f by substituting the estimates of the coefficients in the Fourier sum of the length M:

$$\hat{f}_n(x) = \sum_{k=1}^{M} \hat{c}_k \phi_k(x) \tag{4}$$

Examples 1 and 2 above are simple parametric estimation problems. Example 3 is a nonparametric problem. Typically, one chooses the order M of the Fourier approximation as a function of total number of observations n. This way, the problem of function estimation can be seen as that of parametric estimation, though the number of parameters to be estimated is not bounded beforehand and can be large.

The basic ideas of estimation theory will now be illustrated, using parametric estimation examples. Later, it shall be seen how they can be applied in the nonparametric estimation.

BASIC CONCEPTS

Note the abstract statement of the estimation problem. It is assumed an observation of X is a random element, whose unknown distribution belongs to a given family of distributions P. The family can always be parametrized and written in the form $\{\mathscr{P}_\theta: \theta \in \Theta\}$. Here the form of dependence on the parameter and the set Θ are assumed to be known. The problem of estimation of an unknown parameter θ or of the value $g(\theta)$ of a function g at the point θ consists of constructing a function $\hat{\theta}(X)$ from the observations, which gives a sufficiently good approximation of θ (or of $g(\theta)$).

A commonly accepted approach to *comparing estimators,* resulting from A. Wald's contributions, is as follows: consider a quadratic *loss function* $q(\hat{\theta}(X) - \theta)$ (or, more generally, a nonnegative function $w(\hat{\theta}(X), \theta)$), and given two estimators $\hat{\theta}_1(X)$ and $\hat{\theta}_1(X)$, the estimator for which the expected loss (risk) $Eq(\hat{\theta}_i(X) - \theta)$, $i = 1, 2$ is the smallest is called the better, with respect to the quadratic loss function q (or to w).

Obviously, such a method of comparison is not without its defects. For instance, the estimator that is good for one value of the parameter θ may be completely useless for other values. The simplest example of this kind is given by the "estimator" $\hat{\theta}_0 \equiv \theta_0$, for some fixed θ_0 (independent of observations). Evidently, the estimator $\hat{\theta}^*$ possessing the property

$$Eq(\hat{\theta}^*(X) - \theta) \le Eq(\hat{\theta}(X) - \theta), \qquad \text{for any } \theta \in \Theta$$

for any estimate $\hat{\theta}$ may be considered as optimal. The trouble is that such estimators do not exist (indeed, any "reasonable" estimator cannot stand the comparison with the "fixed" estimator $\hat{\theta}_0$ at θ_0). Generally, in this method of comparing the quality of estimators, many estimators prove to be incomparable. Estimators can be compared by their behavior at "worst" points: an estimator $\hat{\theta}^*$ of θ is called *minimax estimator* relative to the quadratic loss function $q(\cdot)$ if

$$\sup_{\theta\in\Theta} Eq(\hat{\theta}^*(X) - \theta) = \inf_{\hat{\theta}} \sup_{\theta\in\Theta} Eq(\hat{\theta}(X) - \theta)$$

where the lower bound is taken over all estimators $\hat{\theta}$ of θ.

In the Bayesian formulation of the problem the unknown parameter is considered to represent values of the random variables with prior distribution Q on Θ. In this case, the best estimator $\hat{\theta}^*$ relative to the quadratic loss is defined by the relation

$$\begin{aligned} Eq(\hat{\theta}^*(X) - \theta) &= \int_\Theta Eq(\hat{\theta}^*(X) - \theta) Q(d\theta) \\ &= \inf_{\hat{\theta}} \int_\Theta Eq(\hat{\theta}(X) - \theta) Q(d\theta) \end{aligned}$$

and the lower bound is taken over all estimators $\hat{\theta}$.

As a rule, it is assumed that in parametric estimation problems the elements of the parametric family $\{\mathscr{P}_\theta: \theta \in \Theta\}$ possess the density $p(x, \theta)$. If the density is sufficiently smooth function of θ and the *Fisher information* matrix

$$I(\theta) = \int \frac{dp}{d\theta}(x,\theta)\left(\frac{dp}{d\theta}(x,\theta)\right)^T \frac{dx}{p(x,\theta)}$$

exists. In this case, the estimation problem is said to be *regular,* and the accuracy of estimation can be bounded from below by the *Cramér-Rao* inequality: if $\theta \in \mathbf{R}$, then for any estimator $\hat{\theta}$,

$$E|\hat{\theta} - \theta|^2 \ge \frac{\left[1 + \left(\frac{db}{d\theta}\right)(\theta)\right]^2}{I(\theta)} + b^2(\theta) \tag{5}$$

where $b(\theta) = E\hat{\theta} - \theta$ is the bias of the estimate $\hat{\theta}$. An analogous inequality holds in the case of multidimensional parameter θ. Note that if the estimate θ is *unbiased,* that is, $E\hat{\theta} = \theta$, then

$$E|\hat{\theta} - \theta|^2 \ge I^{-1}(\theta)$$

Moreover, the latter inequality typically holds asymptotically, even for biased estimators when $I(\theta) = I$ does not depend on θ. It can be easily verified that for independent observations $X_1, \ldots, X_n$ with common regular distribution $\mathscr{P}_\theta$, if $I(\theta)$ is the Fisher information on one observation, then the Fisher information on the whole sample $I_n(\theta) = nI(\theta)$, and the Cramér–Rao inequality takes the form

$$E|\hat{\theta} - \theta|^2 \geq \frac{\left[1 + \left(\frac{db}{d\theta}\right)(\theta)\right]^2}{nI(\theta)} + b^2(\theta)$$

where $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$.

Return to Example 1. Let X_i be normal random variables with distribution density

$$\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-\theta)^2}{2\sigma^2}\right)$$

If σ^2 is known, then the estimator $\bar{X}$ is an unbiased estimator of θ, and $E(\bar{X} - \theta)^2 = \sigma^2/n$. On the other hand, the Fisher information of the normal density $I(\theta) = \sigma^{-2}$. Thus $\bar{X}$ is in this situation the best unbiased estimator of θ.

If, in the same example, the distribution $\mathscr{P}$ possesses the Laplace density

$$\frac{1}{2a} \exp\left(-\frac{|x-\theta|}{a}\right)$$

then the Fisher information on one observation $I(\theta) = a^{-1}$. In this case $E(\bar{X} - \theta)^2 = 2a/n$. However, the median estimator m, as n grows to infinity, satisfies $nE(m - \theta)^2 \to a$. Therefore, one can suggest that m is an asymptotically better estimator of θ, in this case.

The error $\hat{\theta}_n - \theta$ of the least-squares estimator $\hat{\theta}$ in Example 2, given the observations $y_1, X_1, \ldots, y_n, X_n$, has the covariance matrix

$$E(\hat{\theta}_n - \theta)(\hat{\theta} - \theta)^T = \sigma^2 \left(\sum_{i=1}^n X_i X_i^T\right)^{-1}$$

This estimator is the best unbiased estimator of θ if the disturbances e_i obey normal distribution with zero mean and variance σ^2.

Note that, if the Fisher information $I(\theta)$ is infinite, the estimation with the better rate than $1/n$ is possible. For instance, if in Example 1 the distribution $\mathscr{P}$ is uniform over $[\theta - 1/2, \theta + 1/2]$, then the estimate g satisfies

$$E(g - \theta)^2 = \frac{1}{2(n+1)(n+2)}$$

ASYMPTOTIC BEHAVIOR OF ESTIMATORS

Accepting the stochastic model in estimation problems makes it possible to use the power of limit theorems (the law of large numbers, the central limit theorem, etc.) of probability theory, in order to study the properties of the estimation methods. However, these results holds *asymptotically,* that is, when certain parameters of the problem tend to limiting values (e.g., when the sample size increases indefinitely, the intensity of the noise approaches zero, etc.). On the other hand, the solution of nonasymptotic problems, although an important task in its own right, cannot be a subject of a sufficiently general mathematical theory: the correspondent solution depends heavily on the specific noise distribution, sample size, and so on. As a consequence, for a long time there have been attempts to develop a general procedure of constructing estimates which are not necessarily optimal for a given finite amount of data, but which approach optimality asymptotically (when the sample size increases or the signal-to-noise ratio goes to zero).

For the sake of being explicit, a problem such as in Example 2 is examined, in which $\Theta \in \mathbf{R}^d$. It is to be expected that, when $n \to \infty$, "good" estimators will get infinitely close to the parameter being estimated. Let P_θ denote the distribution of observations $y_1, X_1, \ldots$ for a fixed parameter θ. A sequence of estimators $\hat{\theta}_n$ is called *a consistent sequence of estimators* of θ, if $\hat{\theta}_n \to \theta$ in the probability P_θ for all $\theta \in \Theta$. Note that the estimators, proposed for Examples 1 and 2 above, are consistent.

Note that the notion of the minimax estimator can be refined when the asymptotic framework is concerned. An estimator $\hat{\theta}_n$, for which the quantity

$$\lim_{n\to\infty} \sup \sup_{\theta\in\Theta} Eq(\hat{\theta}_n - \theta)$$

is minimized is referred to as the *asymptotically minimax* estimator in Θ, relative to the quadratic loss q. At first glance, this approach seems to be excessively "cautious": if the number n of observations n is large, a statistician can usually localize the value of parameter θ with sufficient reliability in a small interval around some θ_0. In such a situation, it would seem unnecessary to limit oneself to the estimators that "behave nicely" for values θ that are far away from θ_0. Thus one may consider locally asymptotic minimax estimators at a given point θ_0, that is, estimators that become arbitrarily close to the asymptotically minimax estimators in a small neighborhood of θ_0. However, it is fortunate that, in all interesting cases, asymptotically minimax estimators in Θ are also asymptotically minimax in any nonempty open subset of Θ. Detailed study of asymptotic properties of statistical estimators is a subject of *asymptotic theory* of estimation. Refer to (7) and (10) for exact statements and thorough treatment of correspondent problems.

METHODS OF PRODUCING ESTIMATORS

Let $p(X, \theta)$ stand for the density of the observation measure $\mathscr{P}_\theta$. The most widely used maximum-likelihood method recommends that the estimator $\hat{\theta}(X)$ be defined as the maximum point of the random function $p(X, \theta)$. Then $\hat{\theta}(X)$ is called the *maximum-likelihood estimator*. When the parameter set $\Theta \leq \mathbf{R}^d$, the maximum-likelihood estimators are to be found among the roots of the *likelihood equation*

$$\frac{d}{d\theta} \ln p(X, \theta) = 0$$

if these roots are inner points of Θ and $p(X, \theta)$ is continuously differentiable. In Example 1, $\bar{X}$ in (1) is the maximum-likelihood estimator if the distribution $\mathscr{P}$ is Gaussian. In Example 2, if the disturbances e_i have Laplace density, the maximum-likelihood estimator m_n satisfies

$$m_n = \arg\min_m \sum_{i=1}^n |y_i - m^T X_i|$$

Another approach consists to suppose that the parameter θ obeys a prior distribution Q on Θ. Then one can take a *Bayesian estimator* $\hat{\theta}$ relative to Q, although the initial formulation is not Bayesian. For example, if $\Theta = \mathbf{R}^d$, it is possible to estimate θ by means

$$\frac{\int_{\mathbf{R}^d} \theta p(X,\theta)\, d\theta}{p(X,\theta)\, d\theta}$$

This is a Bayesian estimator relative to the uniform prior distribution.

The basic merit of maximum-likelihood and Bayesian estimators is that, given certain general conditions, they are consistent, asymptotically efficient, and asymptotically normally distributed. The latter means that is $\hat{\theta}$ is an estimator, then the normalized error $I(\theta)^{1/2}(\hat{\theta} - \theta)$ converges in distribution to a Gaussian random variable with zero mean, and the identity covariance matrix.

The advantages of the maximum-likelihood estimators justify the amount of computation involved in the search for the maximum of the *likelihood function* $p(X, \theta)$. However, this can be a hard task. In some situations, the *least-squares method* can be used instead. In Example 1, it recommends that the minimum point of the function

$$\sum_{i=1}^{n} (X_i - \theta)^2$$

be used as the estimator. In this case, $\bar{X}$ in Eq. (1) is the least-squares estimate. In Example 2, the least squares estimator $\hat{\theta}_n$ coincides with the maximum-likelihood solution if the noises e_i are normally distributed.

Often, the exact form of density $p(X, \theta)$ of observations is unknown. However, the information that $p(X, \theta)$ belongs to some convex class P is available. The *robust approach* estimation recommends to find the density $p^*(X, \theta)$, which maximizes the risk of the least-squares estimate on P, and then to take

$$\hat{\theta}^*(X) = \arg\min_{\theta} p^*(X,\theta)$$

as the estimator. The $\hat{\theta}^*$ is referred to as the *robust estimate.* Suppose, for instance, that in Example 1 the distribution $\mathscr{P}$ satisfies $\int (x - \theta)\mathscr{P}(dx) \le \sigma^2$. Then the empirical mean $\bar{X}$ is the robust estimate. If $p(x - \theta)$ is the density of $\mathscr{P}$, and it is known that $p(\cdot)$ is unimodal and for some $a > 0$ $p(0) \ge (2a)^{-1}$, then the median m is the robust estimator of θ [for more details, refer to (5)].

NONPARAMETRIC ESTIMATION

Consider the problem of nonparametric estimation. To be concrete, consider Eq. (2) in Example 3 above. There are two factors that limit the accuracy with which the signal f can be recovered. First, only a finite number of observation points $(X_i)_{i=1}^n$ are available. This suggests that $f(x)$, at other points x than those which are observed, must be obtained from the observed points by interpolation or extrapolation. Second, as in the case of parametric estimation, at the points of observation, X_i, $i = 1, \ldots, n$, $f(X_i)$ is observed with an additive noise $e_i = y_i - f(X_i)$. Clearly, the observation noises e_i introduce a random component in the estimation error. A general approach to the problem is the following: one first chooses an approximation method, that is, substitutes the function in question by its approximation. For instance, in Example 3, the approximation with a Fourier sum is chosen (it is often referred to as the projection approximation, since the function f is approximated by its projection on a final-dimensional subspace, generated by M first functions in the Fourier basis). Then one estimates the parameters involved in this approximation. This way the problem of function estimation is reduced to that of parametric estimation, though the number of parameters to be estimated is not fixed beforehand and can be large. To limit the number of parameters some *smoothness* or *regularity* assumptions have to be stated concerning f. Generally speaking, smoothness conditions require that the unknown function f belongs to a restricted class, such that, given an approximation technique, any function from the class can be “well” approximated, using a limited number of parameters. The choice of the approximation method is crucial for the quality of estimation and heavily depends on the prior information available about the unknown function f [refer to Ref. (8) for a more extensive discussion]. Now see how the basic ideas of estimation theory can be applied to the nonparametric estimation problems.

Performance Measures for Nonparametric Estimators

The following specific issues are important:

1. What plays the role of Cramér-Rao bound and Fisher Information Matrix in this case? Recall that the Cramér-Rao bound [Eq. (5)] reveals the best performance one can expect in identifying the unknown parameter θ from sample data arising from some parameterized distribution P_θ, $\theta \in \Theta$, where Θ is the domain over which the unknown parameter θ ranges. In the nonparametric (as well as in the parametric) case, lower bounds for the best achievable performance are provided by *minimax risk functons.* These lower bounds will be introduced and associated notions of optimality will be discussed.
2. For parametric estimation problems, a quadratic loss function is typical to work with. In functional estimation, however, the choice is much wider. One can be interested in the behavior of the estimate at one particular point x_0, or in the global behavior of the estimate. Different distance measures should be used in these two different cases.

In order to compare different nonparametric estimators, it is necessary to introduce suitable figures of merit. It seems first reasonable to build on the mean square deviation (or mean absolute deviation) of some semi-norm of the error, it is denoted by $\| \hat{f}_N - f \|$. A semi-norm is a norm, except it does not satisfy the condition: $\| f \| = 0$ implies $f = 0$. The following semi-norms are commonly used: $\| f \| = (\int f^2(x)dx)^{1/2}$, ($L_2$-norm), $\| f \| = \sup_x |f(x)|$ (uniform norm, C- or L_∞-norm), $\| f \| = |f(x_0)|$ (absolute value at a fixed point x_0). Then consider the *risk function*

$$R_{a_N}(\hat{f}_N, f) = E[a_N^{-1} \| \hat{f}_N - f \|]^2 \tag{6}$$

where a_N is a normalizing positive sequence. Letting a_N decrease as fast as possible so that the risk still remains bounded yields a notion of a convergence rate. Let $\mathscr{F}$ be a set of functions that contains the "true" regression function f, then the maximal risk $r_{a_N}(\hat{f}_N)$ of estimator $\hat{f}_N$ on $\mathscr{F}$ is defined as follows:

$$r_{a_N}(\hat{f}_N) = \sup_{f\in\mathscr{F}} R_{a_N}(\hat{f}_N, f)$$

If the maximal risk is used as a figure of merit, the optimal estimator $\hat{f}^*_N$ is the one for which the maximal risk is minimized, that is, such that

$$r_{a_N}(\hat{f}^*_N) = \min_{\hat{f}_N} \sup_{f\in\mathscr{F}} R_{a_N}(\hat{f}_N, f)$$

$\hat{f}^*_N$ is called *the minimax estimator* and the value

$$\min_{\hat{f}_N} \sup_{f\in\mathscr{F}} R_{a_N}(\hat{f}_N, f)$$

is called *the minimax risk* on $\mathscr{F}$. Notice that this concept is consistent with the minimax concept used in the parametric case.

The construction of minimax nonparametric regression estimators for different sets $\mathscr{F}$ is a difficult problem. However, letting a_N decrease as fast as possible so that the minimax risk still remains bounded yields a notion of a best achievable convergence rate, similar to that of parametric estimation. More precisely, one may state the following definition:

1. The positive sequence a_N is a lower rate of convergence for the set $\mathscr{F}$ in the semi-norm $\|\cdot\|$ if

$$\liminf_{N\to\infty} r_{a_N}(\hat{f}^*_N) = \liminf_{N\to\infty} \inf_{\hat{f}_N} \sup_{f\in\mathscr{F}} E[a_N^{-1}\|\hat{f}_N - f\|]^2 \ge C_0 \tag{7}$$

   for some positive C_0.
2. The positive sequence a_N is called minimax rate of convergence for the set $\mathscr{F}$ in semi-norm $\|\cdot\|$, if it is a lower rate of convergence, and if, in addition, there exists an estimator $\hat{f}^*_N$ achieving this rate, that is, such that

$$\limsup_{N\to\infty} r_{a_N}(\hat{f}^*_N) < \infty$$

The inequality [Eq. (7)] is a kind of negative statement that says that no estimator of function f can converge to f faster than a_N. Thus, a coarser, but easier approach consists in assessing the estimators by their convergence rates. In this setting, by definition, optimal estimators reach the lower bound as defined in Eq. (7) (recall that the minimax rate is not unique: it is defined to within a constant).

It holds that the larger the class of functions, the slower the convergence rate. Generally, it can be shown that no "good" estimator can be constructed on too rich functional class which is "too rich" [refer to (4)]. Note, however, that convergence can sometimes be proved without any smoothness assumption, though the convergence can be arbitrary slow, depending on the unknown function f to be estimated.

Consider Example 3. The following result can be acknowledged; refer to (7): Consider the Sobolev class $W^s(L)$ on [0, 1], which is the family of periodic functions $f(x)$, $x \in [0, 1]$, such that

$$\sum_{j=0}^{\infty} (1 + j^{2s})|c_j|^2 \le L^2 \tag{8}$$

(here c_j are the Fourier coefficients of f). If

$$\|g\| = \left(\int |g(x)|^2\,dx\right)^{1/2}, \qquad \text{or} \quad \|g\| = |g(x_0)|$$

then $n^{-s/2s+d}$ is a lower rate of convergence for the class $W^s(L)$ in the semi-norm $\|\cdot\|$.

On the other hand, one can construct an estimate $\hat{f}_n$ [refer to (2)], such that uniformly, over $f \in W^s(L)$,

$$E\|\hat{f}_n - f\|_2^2 \le O(L, \sigma)n^{-2s/(2s+1)} \tag{9}$$

Note that the condition [Eq. (8)] on f means that the function can be "well" approximated by a finite Fourier sum. Indeed, due to the Parseval equality, Eq. (8) implies that if

$$\overline{f}(x) = \sum_{j=1}^{M} c_j\phi_j(x)$$

then $\|\overline{f} - f\|_2^2 = O(M^{-2s})$. The upper bound, Eq. (9), appears rather naturally if one considers the following argument: If one approximates the coefficients c_j by their empirical estimates $\hat{c}_j$ in Eq. (3), the quadratic error in each j is $O(n^{-1})$. Thus, if the sum, Eq. (4) of M terms of the Fourier series is used to approximate f, the "total" stochastic error is order of M/n. The balance between the approximation (the bias) and the stochastic error gives the best choice $M = O(n^{1/(2s+1)})$ and the quadratic error $O(n^{-2s/(2s+1)})$. This simple argument can be used to analyze other nonparametric estimates.

MODEL SELECTION

So far the concern has been with estimation problems when the model structure has been fixed. In the case of parametric estimation, this corresponds to the fixed (a priori known) model order; in functional estimation this corresponds to the known functional class $\mathscr{F}$, which defines the exact approximation order. However, rather often, this knowledge is not accessible beforehand. This implies that one should be able to provide methods to retrieve this information from the data, in order to make estimation algorithms "implementable." One should distinguish between two statements of the model (order) selection problem: the first one arises typically in the *parametric* setting, when one suppose that the exact structure of the model is known up to unknown dimension of the parameter vector; the second one is essentially *nonparametric,* when it is assumed that the true model is of infinite dimension, and the order of a finite-dimensional approximation is to be chosen to minimize a prediction error (refer to the choice of the approximation order M in Eq. (4) of Example 3). These two approaches are illustrated in a simple example.

Example 4. Consider the following problem:

1. Let $\theta = (\theta_0, . . . \theta_{d-1})^T$ be coefficients of a digital filter of unknown order d, that is,

$$y_i = \sum_{k=0}^{d-1} \theta_i x_{i-k+1} + e_i$$

   We assume that x_i are random variables. The problem is to retrieve θ from the noisy observations (y_i, x_i), $i = 1, . . ., n$. If one denotes $X_i = (x_i, . . ., x_{i-d+1})^T$, then the estimation problem can be reformulated as that of the linear regression in Example 2. If the exact order d was known, then the least-squares estimate $\hat{\theta}_n$ could be used to recover θ from the data. If d is unknown, it should be estimated from the data.
2. A different problem arises when the true filter is of infinite order. However, all the components of the vector θ of infinite dimension cannot be estimated. In this case one can approximate the parameter θ of infinite dimension by an estimate $\hat{\theta}_{d,n}$ which has only finite number d of nonvanishing entries:

$$\hat{\theta}_n = (\hat{\theta}_n^{(1)} \ldots, \hat{\theta}_n^{(d)}, 0, 0 \ldots)^T$$

   Then the "estimate order" d can be seen as a nuisance parameter to be chosen, in order to minimize, for instance, the mean prediction error $E[(\hat{\theta}_{d,n} - \theta)^T X_n]^2$.

Suppose that e_i are independent and Gaussian random variables. Denote $S_{d,n}^2 = n^{-1}\sum_{i=1}^{n}(y_i - \hat{\theta}_{d,n}^T X_i)^2$. If d is unknown, one cannot minimize $S_{d,n}^2$ with respect to d directly: the result of such a brute-force procedure would give an estimate $\hat{\theta}_{d,n}(x)$, which perfectly fits the noisy data (this is known as "overfitting" in the neural network literature). The reason is that $S_{d,n}^2$ is a biased estimate of $E(y_i - \hat{\theta}_{d,n}^T X_i)^2$. The solution rather consists to modify $S^2(d, n)$ to obtain an unbiased estimate of the prediction error. This can be achieved by introducing a penalty which is proportional to the model order d:

$$AIC(d,n) = \left(S_{d,n}^2 + \frac{2\sigma_e^2 d}{n}\right)$$

which is an unbiased (up to the terms of the higher order) estimate of the error up to terms that do not depend on d, $AIC(d, n)$. One can look for d_n such that

$$d_n = \arg\min_{d<n} AIC(d,n)$$

This technique leads to the celebrated Mallows–Akaike criterion (1, 11):

Unfortunately, d_n is not a consistent estimate of d. Thus it does not give a solution to the first problem of Example 4 above. On the other hand, it is shown in (6) that minimization over d of the criterion

$$HIC(d,n) = \left(S_{d,n}^2 + \frac{2\sigma_e^2 \lambda(n) d}{n}\right)$$

where

$$\liminf_n \frac{\lambda(n)}{\log\log n} > 1 \quad \text{and} \quad \frac{\lambda(n)}{n} \to 0$$

gives a consistent estimate of the true dimension d in the problem 1 of Example 4.

Another approach is proposed in (12) and (14). It consists to minimize, with respect to d, the total length of the incoding of the sequence y_i, X_i (MML—minimum message length, or MDL—minimum description length). This code length should also take into account the incoding of $\hat{\theta}_{d,n}$. This leads to the criterion (the first-order approximation)

$$d_n = \arg\min_{d\le n} BIC(d,n)$$

where

$$BIC(d,n) = \left(S_{d,n}^2 + \frac{2\sigma_e^2 d \log(n)}{n}\right)$$

As was shown in (13), the Bayesian approach (MAP—maximum a posteriori probability) leads to the minimization of $BIC(d, n)$, independently of the distribution of the parameter d.

BIBLIOGRAPHY

1. H. Akaike, Statistical predictor identification, *Ann. Inst. Math. Statist.,* **22**: 203–217, 1970.
2. N. Cencov, Statistical decision rules and optimal inference, *Amer. Math. Soc. Transl.,* **53**: 1982.
3. H. Cramer, *Mathematical Methods of Statistics,* Princeton, NJ: Princeton University Press, 1946.
4. L. Devroye and L. Györfi, *Nonparametric Density Estimation L_1 View,* New York: Wiley, 1985.
5. P. Huber, *Robust Statistics,* New York: Wiley, 1981.
6. E. J. Hannan, Estimation of the order of an ARMA process, *Ann. Stat.,* **8**: 339–364, 1980.
7. I. A. Ibragimov and R. Z. Khas'minskii, *Statistical Estimation: Asymptotic Theory,* New York: Springer, 1981.
8. A. Juditsky et al., Nonlinear black-box modelling in system identification: Mathematical foundations, *Automatica,* **31** (12): 1725–1750, 1995.
9. M. G. Kendall and A. Stuart, *The Advanced Theory of Statistics,* Griffin, 1979.
10. L. LeCam, *Asymptotic Methods in Statistical Decision Theory,* Springer Series in Statistics, Vol. 26, New York: Springer-Verlag, 1986.
11. C. Mallows, Some comments on Cp, *Technometrics,* **15**: 661–675, 1973.
12. J. Rissanen, *Stochastic Complexity in Statistical Inquiry,* Series in Computer Science, Vol. 15, World Scientific, 1989.
13. G. Schwartz, Estimating the dimension of a model, *Ann. Stat.,* **6** (2): 461–464, 1978.
14. C. S. Wallace and P. R. Freeman, Estimation and inference by compact coding, *J. Royal Stat. Soc., Ser. B,* **49** (3): 240–265, 1987.

Anatoli Juditsky
INRIA Rhone-Alpes

ESTIMATION THEORY. See CORRELATION THEORY; KALMAN FILTERS.